Fundamentals of CMOS Single-Photon Avalanche Diodes

Matthew W. Fishburn

Fishburn

Fundamentals of Single-Photon Avalanche Diodes

Fundamentals of CMOS Single-Photon Avalanche Diodes

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus prof. ir. K.C.A.M. Luyben,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op maandag 17 september 2012
om 15.00 uur
door

Matthew William FISHBURN

Bachelor of Science in Electrical Engineering
and Computer Science
Massachusetts Institute of Technology
geboren te Verenigde Staten van Amerika.

Dit proefschrift is goedgekeurd door de promotor:

Prof. dr. ir. E. Charbon

Samenstelling promotiecommissie:

Rector Magnificus	voorzitter
Prof. dr. ir. E. Charbon	Technische Universiteit Delft, promotor
Prof. dr. L. K. Nanver	Technische Universiteit Delft
Prof. dr. ir. A. J. P. Theuwissen	Technische Universiteit Delft
Prof. S. Cova	Politecnico di Milano
Prof. dr. S. Kawahito	Shizuoka University
Prof. A. Zeilinger	Universität Wien
Dr. W. W. Moses	Lawrence Berkeley National Labatory
Prof. dr. P. J. French	Technische Universiteit Delft, reserve

ISBN 978-94-91030-29-1

Author email: `fishburn@alum.mit.edu`

To my loving parents

Contents

List of Figures

List of Tables

Nomenclature

ADC	Analog-to-Digital Converter
APD	Avalanche Photodiode
APS	Active Pixel Sensor
CCD	Charge-Coupled Device
CDF	Cumulative Density Function
CMOS	Complementary Metal Oxide Semiconductor
DAC	Digital-to-Analog Converter
DCR	Dark Count Rate (units: Hz)
DNL	Differential Non-Linearity (units: LSB)
ECR	Excess Count Rate (units: Hz)
EMCCD	Electron-Multipying Charge-Coupled Device
FDG	Fludeoxyglucose
FEM	Finite Element Method
FOM	Figure of Merit
FW(1/N)M	Full-Width at (1/N) of Maximum
FWHM	Full-Width at Half of Maximum
GAPD	Geiger-mode Avalanche Photodiode

i.i.d.	Independently and Identically Distributed
IC	Integrated Circuit
INL	Integral Non-Linearity (units: LSB)
LDD	Lightly Doped Drain
LED	Light Emitting Diode
LSB	Least Significant Bit
LW(1/N)M	Left-Width at (1/N) of Maximum
MSB	Most Significant Bit
PDF	Probability Density Function
PET	Positron Emission Tomography
PMT	Photo-multiplying Tube
QE	Quantum Efficiency
QKD	Quantum Key Distribution
RMS	Root-Mean Square
RTS	Random Telegraph Signal
RW(1/N)M	Right-Width at (1/N) of Maximum
SiPM	Silicon Photomultiplier
SPAD	Single-Photon Avalanche Diode
SPI	Serial Peripheral Interface
TAC	Time-to-Analog Converter (sometimes Time-to-Amplitude Converter)
TCP/IP	Transmission Control Protocol/Internet Protocol
TCSPC	Time-Correlated, Single-Photon Counting
TDC	Time-to-Digital Converter
TOF	Time-of-Flight

UTIG	Uniform Time Interval Generator
WLOG	Without Loss of Generality

Chapter 1

Introduction

Photons would have multiple personality disorder if particles could be diagnosed with the condition. Massless yet having momentum, sometimes a particle sometimes a wave, unable to be halved except when a portion of their energy is transferred, photons exhibit a variety of strange and seemingly contradictory behaviors.

Why would anyone want to work with such capricious particles? Under some circumstances, photons will always act in the locally advantageous of the two contradictory ways. This predictability can be exploited in many applications ranging from 3D sensing to health care. When flowing through a diffraction network, photons will act like waves, but when interacting with silicon photons act like particles, allowing the creation of 3D sensors based on diffraction networks above silicon[1, 2]. High-energy photons always deposit charge in packets when interacting with matter; yet the energy in these packets can be split, allowing the creation of many photons for each interaction, a fact which is exploited in positron emission tomography[3]. Photons have no mass nor charge, which is important when moving through a magnetic field, yet photons must have energy for interacting with detectors, a dichotomy important to imaging systems in strong magnetic fields[4]. The unpredictable nature of photon propagation can generate random numbers that, to the knowledge of modern physics, are truly random[5]. These few examples illustrate the importance of single-photon detection. Researchers, engineers, doctors, and consumers all have a vested interest in the capability to detect single photons. This dissertation, motivated by the need for single-photon detectors, advances understanding of the ability to detect, cheaply and under adverse conditions, single photons with a wavelength in the visible spectrum.

In order to advance the state of the art, the ideal detector must first be envisioned. The "holy grail" of photon detection is the ability to detect the arrival time, energy, and spatial path for every photon in some volume of space using a noise-free, inexpensive sensor. Depending on the applications, some constraints may be relaxed. For example, the TimePix[6] and MediPix[7] sensors can, with nearly 100% efficiency, detect the wavelength, spatial position and arrival time of all x-rays incident on a hybrid detector. In microscopy applications, which have set spatial paths and may have specific wavelengths, the requirements of the ideal detector are often reduced to only the 2D position of the photon. However, the ideal detector is not available for all applications, though a great many detectors are slowly converging to this ideal detector. This introduction will briefly discuss sensors targeting visible light and give the rationale for focusing on CMOS singe-photon avalanche diodes.

An important trend moving toward ideal single-photon detectors is the three dimensional integration of functionality in photon sensors. Hybrid sensors[7] and back-side illuminated single-photon sensors bound to secondary read-out chips[8] have both a high probability of detecting incident photons and the ability to integrate complex logic at the cost of increased system complexity. As fabrication techniques further mature for these technologies, the issue of low fill factor in single-photon imagers containing complex logic will disappear, prompting development of such sensors at the present time.

1.1 Electrical Single-Photon Detectors

Before describing single-photon detectors, it is important to discuss what is meant by "single-photon." Due to the packetized nature of photon energy[1], technically all light sensors are single-photon detectors. However, not all sensors operate on the notion of a single-photon. A **single-photon detector** will be defined as a detector capable of measuring one or more characteristics of a single photon with no other present photons. Measured characteristics include, but are not limited to, the photon's arrival time, energy, or spatial path. The human eye is a good test of this definition. While the eye is sensitive to single photons, the triggering threshold for transmission to the brain is much larger than a single photon[9]. The human eye is not considered a single-photon detector under this definition — the characteristics of a single photon are disregarded unless the photon occurs within a larger number of photons. It should be noted that no attempt is made to define, quantitatively, how accurate the measurement of the single photon's characteristic must be. Such accuracy will depend on application demands.

Henceforth the discussion will be limited to single photon detectors capable or expected to be capable of discriminating characteristics of visible light.

1.1.1 PMTs

First constructed in the 1930s, a photo-multiplying tube (PMT) relies on multiplying electrical current, initially seeded from a photon-generated free carrier, to a sensible level using dynodes[10]. Often the initial carrier comes from a photon's interaction with a quartz window. The dynodes may be implemented using microchannels, known as a microchannel plate detector (MCP), allowing an extremely fast timing response[11]. The detection efficiency of a PMT, which is the probability that the PMT will be able to detect a photon incident on its active area, has exceeded 35% for wavelengths in the visible range[12]. PMTs are often coupled to scintillators and used as radiation detectors, especially in positron emission tomography[3].

The accurate timing response and technological maturity are the largest advantages of PMTs. However, PMTs have several drawbacks. They require large operating voltages, typically several hundred volts to several thousand volts, and require mechanical support which is often not compatible with magnetic fields. PMTs tend to be bulky, though recently PMTs that are several square centimeters have been realized.

1.1.2 EMCCDs

Invented in the late 1960s at Bell Telephone Labatories, the charge-coupled device was the first widely used solid-state imager[13]. Most modern CCDs work by transferring photon-generated charges between MOS capacitors. A gain stage can be added between the device and the readout to make the CCD sensitive to single-photon light[14], with such a CCD called an electron-multiplying CCD (EMCCD) or impactron.

There are two main drawback of EMCCDs. First, they are not free-running like a PMT; the use of gating prevents use in a large number of applications, such as positron emission tomography, which requires a free-running device. Second, such devices must usually be cooled to decrease the noise to operable levels.

1.1.3 CMOS APS

An active-pixel sensor (APS) uses an active element in-pixel to amplify a stored, photon-generated signal before it is converted to a digital signal. The stored signal is usually generated by a p-n or p-i-n junction biased in the reverse region[15]. Nowadays the term APS is synonomous with inexpensive CMOS-based imagers. With consumer demand for imagers in mobile phones, the number of APS sensors has seen a meteoric rise between 2000 and 2010, with an associated increase in research activity into such sensors.

Recently, CMOS APS imagers have been realized with almost single-carrier noise levels[16] — it is likely that these detectors will meet the definition of single-photon detectors in the next few years, or be single-photon imagers depending on the definition of the term. Gating currently allows some information about time-resolution to be achieved, though not on the single-photon level[17]. CMOS APS sensors, like EMCCDs, do not currently have sufficient resolution in the temporal domain to work in a free-running mode. However, the sensors do not require cooling, and would be much less expensive in bulk compared to CCDs.

1.1.4 Quantum Dot Detectors

With improvements in molecular beam technology, it has become possible to fabricate quantum dots with exotic materials[18]. Integration with electronics creates single photon detectors, even at telecom wavelengths. Due to cost and fabrication yield issues, however, this method is still in early stages of research.

1.1.5 Superconducting Single-Photon Detectors

When a single photon generates a charge carrier in a cooled superconducting wire, the deposition in energy can shift the device out of superconducting mode, with a sizable increase in wire resistance. As the energy dissipates, the cooled wire will again be superconducting. Superconducting single-photon detectors use this physical effect to sense the impingement of a single photon on a superconducting wire[19]. Because detectors rely on superconductivity, which presently requires low temperatures, these detectors have not seen wide-spread commercial adoption, and are still in the early stages of research.

1.1.6 Linear-Mode Avalanche Photodiodes

When a p-n junction is biased near its breakdown voltage, the high electric field causes ionization, allowing active amplification of photon-generated carriers. Diodes operating in this regime are known as avalanche photodiodes. When the expected number of carriers varies linearly with the impinging photon count, the junction is said to be in linear-mode[20]. Linear-mode avalanche photodiodes (LAPDs) have many drawbacks, including poor timing accuracy and sizable non-uniformities, but are solid-state and can operate in a free-running mode.

1.1.7 Geiger-Mode Avalanche Photodiodes

When an avalanche photodiode is biased far into the breakdown region, ionization occurs following the injection of a carrier until the diode either destroys itself because of heating or external circuitry shifts the diode into the reverse region. Due to the similarity of operation to a Geiger counter, such a diode is called a Geiger-mode avalanche photodiode (GAPD). GAPDs specifically designed to detect photons are known as single-photon avalanche diodes (SPADs). SPADs have been integrated in CMOS technologies, greatly reducing their cost. The drawback of SPADs is generally their poor fill factor.

Silicon Photomultipliers

One or more SPADs in parallel in often called a silicon photo-multiplier (SiPM). The term SiPM conveys the similarity between the interface of such a device and a PMT. In such devices, the term “pixel” refers to groups of SPADs with between one and thousands of SPADs per group. Care will be taken to correctly define pixel whenever the term arises.

1.2 Why CMOS SPADs?

There are a variety of ways to perform single-photon imaging with visible light. However, in biomedical imaging, there are several constraints which eliminate many of the choices. First, the devices must be mass-producible at reasonable costs, and have good yields. This removes, in the short term, detectors based on quantum dots, the superconducting detector, and the linear-mode avalanche photodiode. The remaining four types of detectors — APS, EMCCDs, PMTs and SPADs — are all used in fluourescence lifetime imaging microscopy. For positron emission tomography, the detector must be free-running with a timing accuracy in the single-digit nanoseconds, leaving just SPADs and PMTs. If the constraint is further reduced to detectors with materials compatible with magnetic resonance imaging, such that a dual PET-MRI system can be achieved, only SPADs currently remain a viable option. This thesis focuses on SPADs because they alone show promise for creating inexpensive, simultaneous PET-MRI systems in the next few years. This dissertation shows how understanding the underlying physics allows the creation of better SPAD-based detectors not only for PET-MRI and other types of biomedical imaging, but also applications such as 3D-imaging and QKD.

1.3 Organization

Following discussions of the state of the art, measurement techniques, and distortions from multi-photon triggering in Chs. 2, 3, and 4, Ch. 5 presents SPAD behavior in hostile environments relevant to PET-MRI. The chapter focuses on the identical operation in strong magnetic fields, and noise increases from radiation damage. Noise increases in SPADs can also be observed when the breakdown voltage is electrically controlled, an effect presented in Ch. 6. Mitigating the adverse effects of this noise is discussed in Ch. 7. The content portion of the dissertation concludes with a case study in Ch. 8 examining which figure of merit (FOM) to optimize for time-of-flight (TOF) PET. Ch. 9 concludes the entire thesis with a summary and a listing of contributions.

Chapter 2

Background

2.1 Theory of Operation

When told that a p-n junction can operate above[1] its breakdown voltage, many engineers immediately recall the steady-state behavior of a diode, shown in Fig. 2.1. This curve contains three regions: a forward region, in which the applied voltage is larger than the junction's inherent potential, allowing the flow of current; a reverse region, in which very little current flows, but the electric field magnitudes in the junction increase with the applied voltage; and the breakdown region, with the electric field magnitude so large that impact ionization occurs, once again creating a flow of current. However, there is a transient state when operating the p-n junction at voltages beyond the breakdown voltage. For a short period of time, before the injection of the first carrier into the diode's depletion region, the diode will operate at a voltage above the breakdown voltage, with only leakage current flowing through the junction. The injection of an ionizing carrier into the depletion region creates a self-sustaining avalanche of carriers[20].

If the applied voltage remains too high, in practice the diode will heat up and melt. However, when the diode is coupled with circuitry, it is possible to sense the onset of the avalanche current, lower the applied voltage below the breakdown voltage, wait some time for free carriers to exit the diode, and then raise the voltage above the breakdown voltage again. Diodes specifically designed to operate in this mode of operation are known as Geiger-mode avalanche photodiodes (GAPDs/G-APDs). Such diodes will probabilistically create a current and voltage pulse pair following the injection of a single car-

[1] The term above is used in the literature, though beyond might make more sense to readers not familar with avalanche diodes.

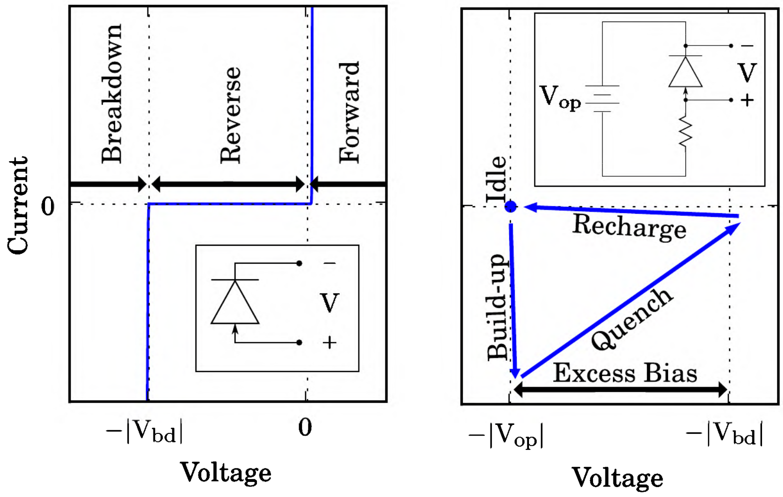


Figure 2.1: **I-V Curves** are shown for a p-n junction's in steady-state (left) and a SPAD's states (right). After [21]

rier into the diode. GAPDs specifically designed to sense carriers injected from single photons are known as single-photon avalanche diodes (SPADs). SPADs have no true steady-state behavior, but will have avalanche phases that are uniform across all quenching schemes: idle; build-up; spread and quench; and recharge. The following section will discuss the avalanche dynamics in detail, relating how SPAD fabrication and the figures of merit both rely on the underlying physics.

2.1.1 Breakdown Voltage and Excess Bias

Qualitatively, a p-n junction is in Geiger-mode when the expected number of carriers following ionization exceeds one. Quantitatively, this condition is met when the mean ionization per free carrier, $\overline{\alpha}$ (units m^{-1}), integrated over the p-n junction's depletion region, z_0 to z_1 as shown in Fig. 2.2 exceeds one:

$$1 < \int_{z_0}^{z_1} \overline{\alpha} dz. \tag{2.1}$$

The ionization rate relates to the mean physical distance the carrier travels before generating another carrier via ionization. If the diode's material has different ionization rates for electrons and holes, which is true for silicon, and these rates are functions of the local position, which also tends to be true in silicon, then the integral becomes

$$1 < \int_{z_0}^{z_1} \alpha_n(z) \cdot \exp\left(\int_{z}^{z_1} [\alpha_p(z') - \alpha_n(z')]\, dz'\right) dz, \tag{2.2}$$

with α_n being the average ionization rate of electrons, and α_p being the average ionization rate of holes[20]. Increasing the applied voltage will increase the electric field strength and hence the ionization rates. The voltage at which the breakdown integral reaches unity is called the breakdown voltage, $\mathrm{V_{bd}}$. The difference between the applied voltage, given by $\mathrm{V_{op}}$ henceforth, and $\mathrm{V_{bd}}$ is termed the excess bias, $\mathrm{V_{eb}}$.

2.1.2 The Drift and Multiplication Regions

A SPAD's depletion region can be separated into two regions with distinct ionization rates: a drift region, where the expected carrier generation is negligible, and the multiplication region, where nearly all of the impact ionization takes place. Quantitatively, the multiplication region is defined as the smallest possible region with 95% of the carrier generation; the drift region is the remaining portion of the depletion region[20]. In SPADs with uniform electric field magnitudes across the depletion region, the multiplication region will occupy most if not nearly all of the depletion region. SPADs lacking such uniform electric fields will have the multiplication region in only a portion of the depletion region. For an abrupt one-sided junction — for example, a p+—n junction with the p+ side so highly doped that order of magnitude shifts in the doping cause little change in the breakdown voltage — the multiplication region will occupy about one third of the depletion region. Fig. 2.2 shows a schematic of these regions for a vertical cross-section of a p+—n junction from a CMOS chip. In discussions of these regions, the depletion region will extend from z_0 to z_1, the multiplication region from z_0 to $z_{1/3}$, and the drift region from $z_{1/3}$ to z_1. There is an additional depletion region around the intersection of the n-well with the silicon substrate around z_w. Carriers entering this junction will be swept towards the substrate, but will not cause ionization since the well-substrate junction's $\mathrm{V_{bd}}$ is larger than that of the p+—n well due to the smaller dopings in the substrate.

The distinction between the drift and multiplication regions is important for a variety of reasons. If the drift region is quite large, but carriers will be generated uniformly over the region, the drift region may introduce sizeable uncertainty into the diode's timing response. During an avalanche, the charge flow across the drift region will act as a small-signal resistor, creating the so-called space-charge resistance[22]. Finally, avalanche propagation will not occur in the drift region once the avalanche begins — the size of

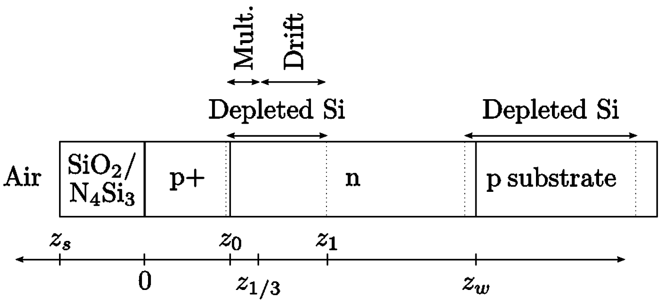


Figure 2.2: **The Multiplication and Drift Regions** — shown is a vertical cross-section of a p+—n junction from a CMOS chip, with the relevant depletion, multiplication, and drift regions labeled. The figure is not to scale

the multiplication region is important when considering how the avalanche propagates.

2.2 Fabrication

SPADs are generally one of two types: thick, reachthrough structures that are at least tens of microns thick; or thin, planar structures with an active region that is a few microns thick. In both cases the multiplication region tends to cover only a few microns of distance; the main difference in structure size occurs in the drift region's size.

2.2.1 The Guard Ring

Whatever a SPAD's thickness, it is necessary to separate its active region from the surrounding area; otherwise, only one large SPAD could be fabricated, with no coupled electronics. The structure responsible for this separation is called the guard ring. With no guard ring, carriers will diffuse into the active region, causing undesirable, spurious avalanches. Additionally, if there is no structure at the edge of the active region, usually the higher surface curvature in the doping near the device edge will cause premature edge breakdown (PEB)[23]. PEB is undesirable because it creates smaller active regions. Many chip fabrication processes also include larger horizontal doping gradients in the p+ implant, which exacerbates PEB.

Fig. 2.3 shows a cross-section of a guard-ring-free device that would exhibit edge breakdown, along with simulations of the electric field magnitude when such a p-n junction has an applied voltage of 20V. The n substrate in this structure is a constant $4 \cdot 10^{16}\text{cm}^{-3}$, with a p+ doping of $5 \cdot 10^{19}\text{cm}^{-3}$.

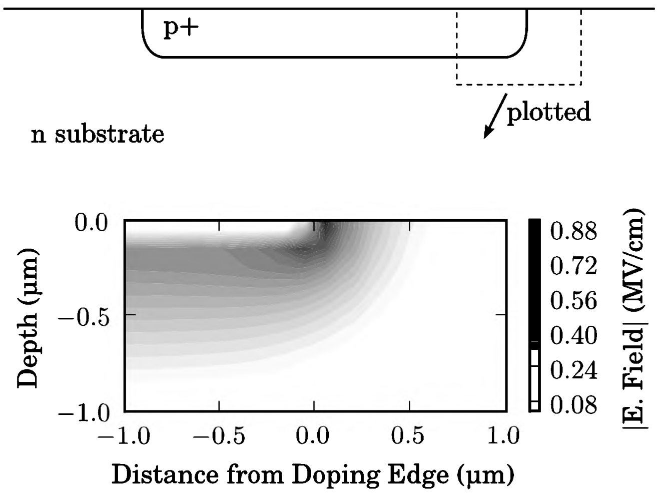


Figure 2.3: **Premature Edge Breakdown** — the electric field magnitude from an $|\vec{E}|$ field simulation using [24] (bottom) is shown for a SPAD without a guard ring (top).

The electric field strength is more than twice as large near the device edge than the center. Such a device will make a poor SPAD, suffering from high noise due to the field strength and active region's surface proximity, and the active region will be small compared to the consumed area.

Theoretically it is possible for a SPAD to exist without a guard ring, but in practice no such SPADs exist. Some SPADs do use "virtual" guard rings[25, 26], with a lack of an extra implant creating doping differences between the structure's outer edge and active region. The guard ring's requirements constrain all SPAD design, and may be the limiting step in CMOS processes with set implants.

2.2.2 Reachthrough SPADs

Reachthrough SPADs are fabricated with structures containing a thick portion of intrinsic silicon, similar to a p-i-n photodiode. In these structures the depletion region spans tens or hundreds of µm. Such structures may use a p-n junction on one side of the intrinsic silicon for the carrier multiplication region. These types of structures can detect nearly every impinging photon of a specific wavelength, dependent on the fabrication material. In silicon, reachthrough diodes are very sensitive to near-infrared (NIR) light. The disadvantage in creating thick junctions are: poor timing response; high noise; a high V_{bd} (e.g. >400V); and incompatibility with standard CMOS processes.

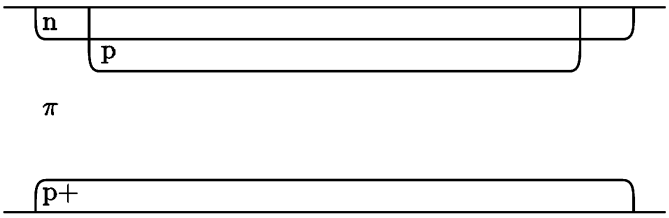


Figure 2.4: **A Reachthrough SPAD's Cross-section** — typically the intrinsic region, π in the figure, is at least 100µm thick, while the implants cover several microns thick at most (figure not to scale). After [27]

Such reachthrough structures are usually fabricated on wafers with several different epitaxial growths, or must have some type of backside implantation, possibly on thinned wafers.

Fig. 2.4 shows a cross-section of a reachthrough diode with a p+-π-p-n structure[27]. This structure's multiplication region will occur at the p-n junction. The intrinsic silicon will be completely depleted, causing injected carriers to drift towards the multiplication region. This structure's guard ring is the lack of the p implant layer near the edge of the p+ layer. The additional p dopants increase the field magnitude toward the center of the device, creating ionization in the middle rather than the edge of the SPAD.

2.2.3 Planar SPADs

Planar SPADs are usually fabricated near a semi-conductor's surface using implantations, though growths are also a possibility. Planar SPADs have depletion regions that are hundreds of nanometers to several microns thick.

Various methods exist for generating the SPAD and guard ring in planar processes. Because most modern CMOS processes use p substrates, a straight-forward SPAD uses the n+ diffusion implant to generate an n+-p junction, with a shallow n-well forming the guard ring[28]. If the CMOS process has a deep n-well, the implant dopants can be reversed with the deep n-well acting as the substrate. With this method, a p+ active region with a p-well forms the guard ring on top of a deep n-well. Fig. 2.5 shows these types of guard ring. Using a deep n-well isolates the SPAD from substrate noise, since the substrate and the deep well form an additional junction that will prevent free carriers in the substrate, which often have a long mean free path, from diffusing into the junction itself. This SPAD structure has been successfully implemented in CMOS processes ranging from a 0.8µm, high-voltage process to 90nm processes, though tunneling noise has been problematic using these techniques in 90nm[29, 30, 31]. SPADs have been created in 65nm CMOS

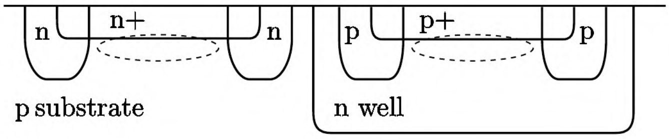


Figure 2.5: **Cross-sections of SPADs with Well-based Guard Rings** — the multiplication region is highlighted with a dashed ellipse. Figure is not to scale.

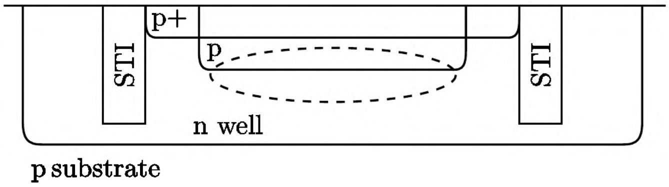


Figure 2.6: **Cross-sections of SPADs with STI-based Guard Rings** — the multiplication region is highlighted with a dashed ellipse. Often, a retrograde n-well or retrograde implant is used to create the junction. Figure is not to scale.

processes, but the work has yet to be widely published[32].

As planar processes have shrunk in feature size, shallow trench isolation (STI) has had an increasingly important role in the guard ring. Early work[33] in fabricating such devices suffered from high noise, due to the injection of free carriers from the trap-filled STI surface near the depletion region. Subsequent work was able to solve this noise problem by using implants near the STI and virtual guard rings to mitigate the problems caused by these carriers[34]. Additionally, rather than using an n-well implant, a retrograde junction has been employed by such devices with great success[25]. A retrograde junction is one in which the implant doping changes slowly, rather than abruptly. This sort of junction creates more uniform electric fields, requiring less physical distance to achieve breakdown. Low-noise SPADs exist using STI-bound structures in CMOS processes with features as small as 90nm[35]. Fig. 2.6 shows an example cross-section of such devices.

More exotic methods have proven effective in creating guard rings[36, 37]. Beveling has been performed in planar processes to form effective guard rings in place of implanting additional layers[38]. Electrodes, deposited above the highly doped implant's edges, use electrical methods to prevent edge breakdown[39]. Even the physical spreading of energetic implants has been used to create guard rings[40]. Due to the complexity and variability of these types of guard rings, they are rarely used when creating SPADs. Chapter 6 will discuss the difficulties with using one of these methods.

The well depths are primary concerns when implementing SPADs in CMOS processes. A lower bound on the minimum depth of a deep n-well is easily obtained by ensuring the depletion region below the guard rings will not meet the depletion region from the p-n junction formed by the substrate with the deep n-well. Such a condition is called punch-through. The depletion regions touching will isolate the deep n-well below the desired multiplication region, preventing current flow and hence observation of the avalanche. For an abrupt one-sided junction, the width of the depletion region, $z_1 - z_0$, will vary as the square root of the applied voltage divided by the well doping, $\mathrm{N_d}$,

$$z_1 - z_0 \approx \sqrt{\frac{2\epsilon_s \mathrm{V_{op}}}{q\mathrm{N_d}}}, \tag{2.3}$$

with ϵ_s being the permittivity of silicon, and q the electron's charge[20]. For a well with a doping of $4 \cdot 10^{16}/\mathrm{cm}^3$ at an applied voltage of 18 volts, which is roughly the breakdown voltage of an abrupt p-n junction in such a well, the depletion region width will be about 760nm. If the guard ring is well-based, a first order approximation in this case of the deep n-well's minimum depth is at least 800nm below the bottom of this guard ring.

2.2.4 Inherent and Fabrication Non-idealities

Dopants in a CMOS process do not have sharp boundaries as often portrayed, and they are not implanted exactly where expected due to diffusion during annealing. The diffusion can cause sizeable distortions to a SPAD's multiplication region, and hence active area. In small diodes, the effect can even distort the expected value of the diode's breakdown voltage. Inactive distances larger than 1.5µm have been reported in the literature, with distortions to the breakdown voltage in diodes with diameters of 6µm[41]. Fig. 2.7 presents this phenomenon.

The effect is similar to the depletion region's infringment on the active area from the guard ring wells[42]; however, the two can be separated by observing whether the inactive distance increases or decreases with the applied voltage. If diffusion of implant dopants causes the inactive distance, then as the applied voltage increases, more of the diode is expected to meet breakdown condition. In contrast, less of the diode will be under breakdown with increasing applied voltage if infringement from the guard ring's well's depletion region causes this inactive distance.

The inactive distance is related to the effects of another fabrication nonideality, the well's ohmic resistance. Due to doped silicon's resistivity, ohmic

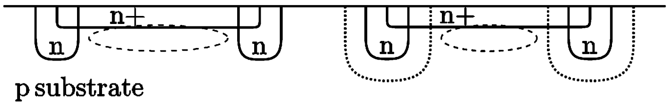


Figure 2.7: **Implant Diffusion Creating the Inactive Distance** — the diffusion of well-based guard rings' implants (dotted curves) can cause sizeable distortions to the multiplication region (dashed curves). Figure is not to scale.

resistances are introduced between the diode itself and contacts to external circuitry. The ohmic resistance R_o is an important factor when attempting to quickly switch the applied voltage on the diode, as the diode has some associated capacitance due to its charge separation. R_o may also play a role when attempting to minimize the SPAD jitter, since it distorts the ability to sense the exact moment of carrier injection into the diode itself.

There are two more inherent non-idealities worth noting. The space-charge resistance was described previously: during an avalanche, carriers take some time to cross the diode's drift region, creating a voltage drop. The small-signal effect is modelled as the space-charge resistance. Also due to this required transit time, the diode will resist instantaneous changes in current, creating a small-signal inductance. Prior to the development of deep submicron CMOS processes, this small signal inductance, coupled with the parasitic capacitances mentioned above, was used in BARIT and IMPATT diodes for the generation of microwave frequencies, often for telecommunication[20]. Ringing can be observed if the diodes quenching circuit is not properly engineered[43]. At the end of this chapter, Table 2.1 shows fabrication tradeoffs for a SPAD with a well-based guard ring in a deep well.

2.3 Avalanche Dynamics

This section will present the details of the four phases of an avalanche in a SPAD: idle; build-up; spread and quench; and recharge.

2.3.1 A Qualitative Description of the Avalanche Phases

Prior to the injection of any free carriers, into the depletion region, there is no charge flowing in the central portions of the depletion region. Near the edge of the depletion region, hole majority carriers from the p+ will diffuse into the top portion of the depletion region, with electron majority carriers from the deep n well diffusing into the bottom. The carriers will not cause an avalanche due to the electric field's orientation, being immediately ejected from the depletion region due to their charge polarity. For this reason, the voltage on the diode itself can be modulated without triggering any spurious avalanches; any current will come from majority carriers, which will not trigger an avalanche.

Following the injection of a free carrier into the diode's depletion region, the carrier will drift with the electric field, possibly causing ionization and triggering an avalanche. This is the build-up phase. The exact probability of triggering an avalanche will be covered later in Sec. 2.4.1. Here, it will be assumed that the impact ionization is a deterministic, rather than a probabilistic, process. With this model, the carrier will cause an avalanche for certain. Two processes govern the changes in the local current: positive feedback from ionization, and negative feedback from drift and coupled resistances, usually dominated by the space-charge resistance. The positive feedback process will cause a rapid increase in local current density, until the current flow across coupled resistances causes the local potential to decrease to the breakdown voltage. In present diodes, these processes occur orders of magnitude faster than the voltage changes across any coupled or parasitic capacitors.

Once the positive and negative feedback processes are balanced in one portion of the diode, the avalanche will spread, via a multiplication-assisted diffusion process, to other portions of the diode. Technically the avalanche will also spread during the build-up process, but the slow spreading speed, on the order of $10-20$µm per ns, limits the spread during the build-up phase to several hundred nanometers. Also during this phase, current into the coupled capacitance will lower the voltage across the diode at the same time, quenching the avalanche. If the diode is small, then the avalanche spreads across the entire diode before the voltages on any coupled capacitances' voltages begin to change. If the diode is large, the avalanche spread and quench will occur at the same time. The avalanche is said to be quenched when all free carriers have exited the depletion region, and voltage across the diode is below the breakdown voltage. Typically the spread and quench phase of the avalanche will take 1ns, but of course the phase will vary with a number of factors.

Finally, following quenching, the diode needs to be restored to its idle state. This is the recharge phase, also called the restoration or reset phase, and is complete when the diode is back to the idle state.

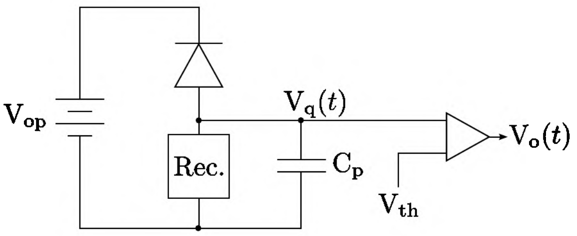


Figure 2.8: **General Recharge Circuit** — The recharge element, boxed "Rec." in the figure, may also take the comparators output as feedback information.

2.3.2 Schemes for Quenching and Recharge

If SPADs were free of correlated noise, the applied voltage would be restored as quickly as possible. However, the afterpulsing phenomenon[44], presented in Sec. 2.4.4, causes noise correlated in time and creates a trade-off when considering the ideal hold-off time, commonly called the dead time, of a SPAD. Several recharge schemes exist, each with different trade-offs. Fig. 2.8 shows a circuit with a generalized recharge element. Some circuits, such as the active recharge elements, will use the comparator's output as feedback to the recharge element.

Passive Recharge

The simplest recharge scheme is to use a large resistor, on the order of $R_q = 100k\Omega$ to $1M\Omega$, as the recharge element in Fig. 2.8[45]. The resistor needs to be large enough in relation to the space-charge resistance that the number of free carriers in the diode will eventually reach zero, possibly aided by some parts of the ionization statistics, thus $R_q >> R_{sc}$. In a CMOS chip, the resistor can be implemented with a transistor whose gate voltage can be externally adjustable. To allow only small current to flow through this resistor, in practice the transistor needs to be placed in weak inversion, with the transistor acting as a current source, instead of a resistor. This scheme's main advantage is its simplicity and low footprint — in applications where fill factor is critical, this scheme allows a large fraction of the surface area to be active area.

There are several disadvantages with passive quench and recharge, all stemming from the fact that the applied voltage across the diode only returns to its operating point slowly. For high event rates, such as those using a high-repetition-rate laser causing an expected avalanche per pulse, an avalanche can occur before the diode is completely recharged to the operating point.

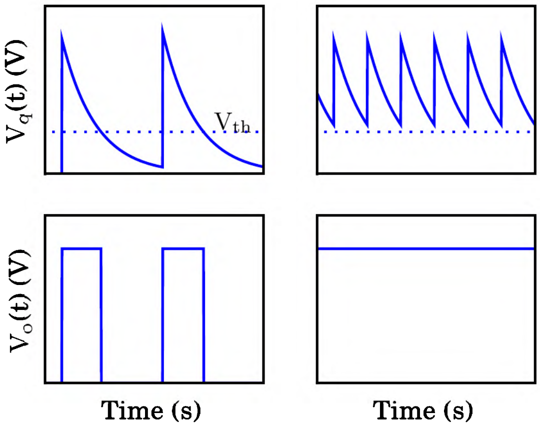


Figure 2.9: **An Issue with Passive Recharge** — a passive recharge circuit normally outputs rising edges that are linear with the number of incident photons (left), but when saturated (right), the output simply remains above the comparator level (dotted line in top plots).

If the applied voltage does not cross the threshold of the coupled comparator, the SPAD will not have appeared to avalanche at all — if this occurs indefinitely, it appears as though the SPAD is continuously in recharge. If the SPAD does avalanche, it may do so before being completely recharged, causing distortion in the timing response. Even in applications with a low repetition rate, afterpulsing can be problematic with passive recharge, causing distortions to the actual dead time if the SPAD re-avalanches too quickly.

Active Recharge

A wide variety of active recharge circuits exist in the literature, many integrated on-chip[46]. The basic idea behind all of these integrated circuits is to place a transistor in series with the SPAD, and turn this transistor strongly on some time after the avalanche using feedback from the comparator's output. This transistor operates differently than the always-on, weakly inverted, passive transistor. The large difference between the active recharge schemes is the method for generating the necessary delay. There is a wide variety of implementations[47], including monostable elements[48], high-threshold comparators in combination with passive quenching[49], and multiple buffers[50]. All of these circuits reduce the afterpulsing, but do so at the cost of additional area. As the feature size shrinks in CMOS, the required area to implement such circuits has decreased. Additionally, active recharge gives more intuitive saturation behavior when the rising edges in the output pulses are being counted, since the number of pulses will saturate with active recharge, but with passive recharge the number of rising edges will decrease at some point, as Fig. 2.9 shows.

Active Quench

In addition to active recharge, some architectures also aid in quenching the avalanche before completion by sensing the avalanche onset and aiding in quenching[48]. In diodes with large parasitic capacitances, such a scheme can reduce the number of free carriers that flow through the diode, decreasing the afterpulsing probability at the cost of additional circuitry.

2.3.3 A Simple, Quantitative Model

This section shows a simple exponential model[29] for a diode in a passive quenching scheme that is applicable if:

1. The avalanche spread is neglected, in other words the avalanche occurs over the entire active region at once because of uniform carrier concentration
2. Electrons and holes are assumed to have the same ionization rate
3. Ionization is assumed to occur uniformly within the multiplication region
4. The ionization rate does not change during the build-up phase

As described at the start of this section, three phenomena govern a SPAD's dynamics: the space-charge resistance, the diode capacitance, and the current from impact ionization. Fig. 2.10 shows a resistor, capacitor, and current source modelling these elements, along with the reponsible locations in the diode itself. Additionally, when the SPAD is placed in a passively quenched scheme, also shown in Fig. 2.10, there are two additional elements that must be modeled: the quenching resistor and parasitic capacitance from coupled elements.

Of all modeled elements, only the current source, I_a, is non-trivial to model. The free carriers in the diode itself form the basis for this current source. Due to the high electrical field, these carriers will travel at the saturation velocity[1] within the junction itself. The current densities for electrons

[1] Appendix A discusses just how rapidly these carriers accelerate.

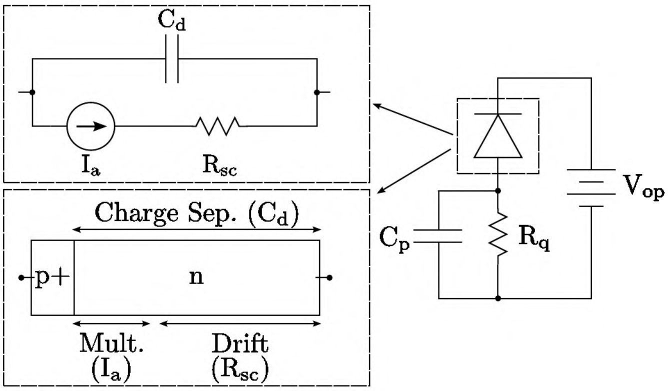


Figure 2.10: **A Simple Model of a SPAD** — A simple model of a SPAD (top left), along with the basis for the constituents in the p-n junction itself (bottom left), is shown in a passively quenched scheme (right). The additional parasitic capacitance, C_p, usually comes from coupled transistors' parasitic capacitances.

and holes will be

$$\vec{j}_n(z,t) = -q \cdot \vec{v}_s \cdot n(z,t), \tag{2.4}$$

$$\vec{j}_p(z,t) = q \cdot (-\vec{v}_s) \cdot p(z,t), \tag{2.5}$$

with the total current density being the sum of the two component densities

$$\vec{j}(z,t) = \vec{j}_n(z,t) + \vec{j}_p(z,t). \tag{2.6}$$

As described in Appendix A, (2.5) and (2.4) can be used with the continuity equations,

$$\frac{\partial n}{\partial t} = G_n - U_n + \frac{1}{q}\nabla \cdot \vec{j}_n, \tag{2.7}$$

$$\frac{\partial p}{\partial t} = G_p - U_p + \frac{1}{q}\nabla \cdot \vec{j}_p, \tag{2.8}$$

and known boundary conditions to extract the current source's governing equation,

$$\frac{\partial c(t)}{\partial t} = c(t)\left(\frac{1}{\tau_p} - \frac{1}{\tau_n}\right). \tag{2.9}$$

where $\tau_p = t_m/(2\overline{\alpha}z_m)$ is the positive feedback from ionization, and $\tau_n = t_m/2$ is the negative feedback from drift. t_m and z_m are the transit time across the multiplication region and the width of the multiplication region, respectively.

The average ionization, $\overline{\alpha}$, is a function of the applied voltage, and at breakdown must be exactly the inverse of the multiplication region's width, $1/z_m$. This is because the assumption that ionization occurs uniformly in the multiplication region allows simplification of (2.1) to

$$\begin{aligned} 1 &= \int_{z_0}^{z_1} \overline{\alpha}(\mathrm{V_{bd}})dz, \\ 1 &= (z_1 - z_0)\overline{\alpha}(\mathrm{V_{bd}}), \\ \overline{\alpha}(\mathrm{V_{bd}}) &= 1/(z_1 - z_0), \end{aligned} \tag{2.10}$$

Appendix A also describes derivation of a formula relating the mean ionization coefficient to the voltage across the current source, termed $\mathrm{V_a}$ here:

$$\overline{\alpha}(\mathrm{V_a}) = \frac{1}{z_m} \int_{z_{1/3}}^{z_0} \left(A \cdot \exp\left(-(a/|\vec{E}(z)|)^m \right) \right) dz, \tag{2.11}$$

with a a fit parameter, and constants A and m.

As the current flowing through the diode increases, the voltage across the current source will drop from the externally applied voltage, $\mathrm{V_{op}}$, by the excess bias to the breakdown voltage. The drop is caused by current flow across the space-charge resistor, with nearly all of the current diverted into the diode and parasitic capacitances. At an excess bias of roughly three volts, the mean ionization rate will be about 2.5 times larger than its value at breakdown, giving an estimate of the current build-up's time constant as

$$\begin{aligned} \tau &= 1/\left(\frac{1}{\tau_p} - \frac{1}{\tau_n} \right), \\ &= \frac{t_m}{2\left(\overline{\alpha} z_m - 1\right)}, \\ &\approx \frac{2.5\mathrm{ps}}{2\left(2.5 - 1\right)}, \qquad (2.12) \\ &\approx 0.8\mathrm{ps}. \qquad (2.13) \end{aligned}$$

When the voltage across the current source increases, $\overline{\alpha}$ decreases, until $\overline{\alpha} z_m = 1$ at breakdown. For the purpose of simplicity, it will be assumed that τ is constant until the voltage drop across $\mathrm{R_{sc}}$ reaches the breakdown voltage; at such time, τ will be set to 0. This gives exponential behavior in the build-up phase.

After the feedback processes from ionization and drift match, these feedback processes will keep the voltage drop across the current source at exactly the breakdown voltage until the number of carriers in the depletion region itself drops to zero. This current source will thus act like a voltage source in this phase, creating RC behavior at the diode's cathode, with time constant $R_{sc}(C_d+C_p)$. Note here we are assuming that $R_{sc} << R_q$.

In this model, the avalanche would theoretically continue forever; however, due to the small-signal inductance[20] of the carriers travelling across the depletion region, the diode is not a true voltage source, and the voltage at the diode's anode will be larger than the excess bias. For this model, it will be assumed that this happens when the number of carriers in the diode itself falls below 10, as the statistics of ionization with this number of carriers will eventually cause the avalanche to cease.

Finally, now that no carriers exist in the depletion region, the current source creates no current, acting as an open circuit. Because we have assumed that $R_{sc} << R_q$, both capacitors are in series to the ground line with a resistance of R_q, and the recharge phase occurs with a time constant that is roughly $R_q(C_d+C_p)$.

Thus a simple model of the avalanche is that it occurs in three phases, each exponential: an exponential build-up of carriers in the diode; RC behavior limited by R_{sc} during the quench phase; and, finally, RC behavior limited by R_q during the recharge phase.

2.3.4 Multiplication-Assisted Diffusion Model

The model presented in the previous section can be used as the basis for a model of a larger diode, if the larger diode is split into smaller diodes that follow the assumptions. Two assumptions must be relaxed. First, the ionization process should not be assumed to follow the single exponential behavior, since the relevant time constants will vary. Second, the avalanche spread cannot be neglected.

There are three possible mechanisms for the avalanche spread: (1) diffusion; (2) free carriers creating local differences in the electric field, causing lateral forces on other carriers; and (3) optical emission causing recombination within the diode itself. The last mechanism, optical emission and recombination, is known to be an important process in avalanche spreading in reachthrough diodes, but is trivial for planar diodes[51], and its effects will be ignored since the present discussion focuses on planar diodes. [52] quantifies the portion of spreading from the first two effects in planar diodes, and shows that the local differences in the electric field are an order of magnitude less important than multiplication-assisted diffusion. Thus, diffusion assisted by ionization will be the dominant factor in the avalanche spread. Under these assumptions, the carrier concentration's governing equation will

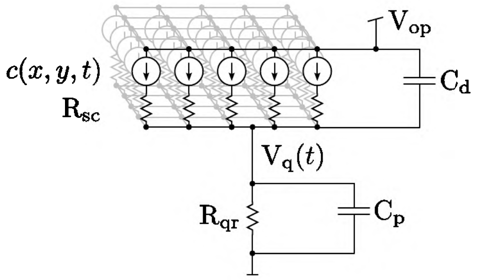


Figure 2.11: **A FEM Model of a SPAD**

change[51] from (2.9) to

$$\frac{\partial c(x,y,t)}{\partial t} = \nabla c(x,y,t) + \frac{c(x,y,t)}{\tau(x,y,t)}, \tag{2.14}$$

with τ being defined as in the simple model, though now the local applied voltage must also be included.

Fig. 2.11 shows the model circuit of a 2µm by 30µm diode coupled to a passive recharge element. To simplify the discussion, the ohmic resistance is not included, though it will be included when modeling actual diodes. The state-containing elements in this model are the carrier concentration in the current sources and the voltage across the capacitor. Using (2.14) in conjunction with the equation for the capacitor's state, $I = C\frac{dV}{dt}$, numerical simulation methods such as the Runge-Kutta methods can be applied to solve for $c(x,y,t)$ and $V_q(t)$. Fig. 2.12 presents results from an abrupt, one-sided junction of dimensions 2µm by 30µm biased at 2.5V above its $\mathrm{V_{bd}}$ of 20V when using a fourth-order Runge-Kutta method[1][53] to solve the component equations.

Evident in the figure are the time scales of the different avalanche phases. The build-up phase lasts 40ps; the quench and spread phase, 1ns; and the recharge phase, almost 1µs.

2.4 Figures of Merit

2.4.1 Photon Detection Probability

The probability that a photon impinging on the active area triggers an avalanche is known as the photon detection probability (PDP)[2]. Due to the

[1] Methods with lower orders showed stability problems with slighter larger step sizes.

[2] Some authors use the term photon detection efficiency (PDE) instead of PDP, whereas others consider the PDE to be the product of the PDP and the fill factor. PDE will not

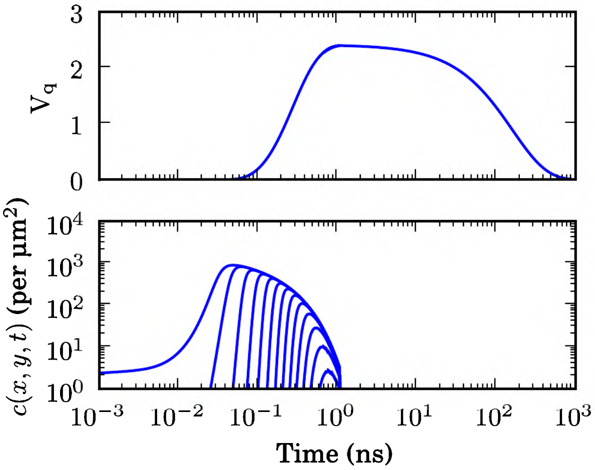


Figure 2.12: **Simulated Avalanche Propagation** — shown are the voltage at the SPAD's quenched node (top), and the free carrier concentration (bottom) every 1µm along the diode. Note the log scale on the abscissa.

statistical natures of impact ionization and light's penetration into silicon, this probability is always less than one. PDP is different than a photodector's QE, in that the QE sometimes includes fill factor effects, whereas the PDP never does. Since PDP allows comparison of singular devices with those inside an array, PDP values rather than QE values are normally quoted for SPADs[21].

To find the PDP, the dead-space-free triggering model from [54] is combined with the distortions of SiO_2 transimission effects from [55]. First, the probability of an injected electron-hole pair triggers an avalanche, signified by $p(z)$, will be quantified as a function of depth. The chance of detecting the electron $p_e(z)$, or the chance of detecting the hole, $p_h(z)$, are the basis for $p(z)$, with $p(z) = p_e(z) + p_h(z) - p_e(z)p_h(z)$. Then, $p(z)$ will be combined with the probability distribution of electron-hole generation by different wavelengths of light, yielding the PDP.

Triggering Probability within the Depletion Region

Let an electron be located at position z into a p-n junction, accelerating in the $+z$ direction. Let dz be set small enough that the probability of more than one ionization between z and $z + dz$ is negligible. There are two possibilities for how this electron at z would cause an avalanche, termed event

be used to avoid confusion.

A: ionization-sourced carriers created between z and $z + dz$ would cause an avalanche, termed event A_i, or the electron would cause an avalanche after moving to $z + dz$, termed event A_e. Note that A_i and A_e are non-exclusive, as both the original carrier and the ionization-generated carriers could go on to cause avalanches if the other were able to be removed from the junction following ionization. In fact, A_i and A_e can be assumed to be independent, since ionization does not significantly alter the local electric field. With these definitions,

$$\begin{aligned} P[A] &= P[A_e \text{ or } A_i], \\ &= P[A_e] + P[\overline{A_e}] \cdot P[A_i|\overline{A_e}], \\ &= P[A_e] + P[\overline{A_e}] \cdot P[A_i]. \end{aligned} \tag{2.15}$$

The probability that the carrier ionizes between z and $z + dz$ is $\alpha_n(z) \cdot dz$, with $P[A_i] = \alpha_n(z) \cdot p(z) \cdot dz$. $P[A] = p_e(z)$ and $P[A_e]$ are simply $p_e(z)$ and $p_e(z + dz)$. Thus the expression above can be simplified to

$$\begin{aligned} P[A] &= P[A_e] + P[\overline{A_e}] \cdot P[A_i], \\ p_e(z) &= p_e(z + dz) + \\ &\quad (1 - p_e(z + dz)) \cdot \alpha_n(z)p(z)dz, \\ (p_e(z + dz) - 1)\alpha_n(z)p(z)dz &= p_e(z + dz) - p_e(z), \\ (p_e(z + dz) - 1)\alpha_n(z)p(z) &= \frac{p_e(z + dz) - p_e(z)}{dz}. \end{aligned} \tag{2.16}$$

Taking the limit of the expression above with $dz \to 0$ yields a differential equation for $p_e(z)$, with a similar analysis possible for holes. Thus, the governing differential equations for these probabilities are

$$\frac{dp_e(z)}{dz} = \alpha_n(z)p(z)\left(p_e(z) - 1\right), \tag{2.17}$$

$$\frac{dp_h(z)}{dz} = \alpha_p(z)p(z)\left(1 - p_h(z)\right). \tag{2.18}$$

Any holes injected at $z = z_0$ will be unable to cause an avalanche due to the depletion region's polarity, and hence $p_h(z_0) = 0$. Similarly for electrons, $p_e(z_1) = 0$. Sweeping the initial value of $p_e(z_0)$ between 0 and 1 such that numerically evaluating (2.18) and (2.17) from z_0 to z_1 gives $p_e(z_1) = 0$, with the solution to these equations following.

Numerical solutions for conditions in an abrupt p-n junction are shown in Fig. 2.13.

Carrier Detection Probability

A visible light photon, or a photon with a similar energy, impinging on a p+—n SPAD inside an n-well will cause one of the following:

1. The photon is reflected, or absorbed in the stack above the silicon
2. The photon generates an electron-hole pair in the p+ region, but outside the depletion region
3. The photon generates an electron-hole pair in the depletion region
4. The photon generates an electron-hole pair in the n-well region, but outside the depletion region
5. The photon generates an electron-hole pair in the substrate
6. The photon passes through the silicon, creating no carrier pairs

The CDP for condition 3, which is that the electron-hole pair is generated in the delpetion region, was discussed in the previous section.

Conditions 1, 5, and 6 will not contribute to the PDP. Conditions 1 and 6 create no carriers for the SPAD to detect. Condition 5's carriers will not be able to trigger the SPAD, as the hole will be blocked from entering the well by the depletion region between the n-well and the substrate, as shown in Fig. 2.2, and the electron will be unable to trigger an avalanche from the well, where it is a majority carrier. Thus, if $z < 0$ or $z > z_w$, $p(z) = 0$.

Similar situations will exist for the majority carriers in conditions 2 and 4 — the majority carrier will be unable to enter the depletion region due to its charge polarity. The minority carriers, however, may enter the depletion region and trigger the avalanche.

In the p+ region, four possibilities exists for the minority electron: the carrier becomes trapped at the surface, the carrier diffuses into depletion region, the carrier leaves the p+ region via a route other than exists considered in the first two possibilities, or the carrier recombines. The surface SiO_2, because of its slight positive charge[56], creates the first possibility, that the carrier becomes trapped at the surface. Thus $p(0) = 0$. The slight positive charge is a mixed blessing; minority carriers generated by surface defects will remain close to the surface, but photon-generated carriers also remain close to the surface.

The second possibility for the electron, that it enters the depletion region, implies that the minority electron will trigger an avalanche with the edge condition given by the depletion region, valued at $p_e(z_0)$.

The third possibility for the electron, that it leaves the p+ region other than through the depletion region or becoming trapped at the surface, will be considered negligible in the present analysis. This assumption is valid for large active area SPADs with no edge effects.

The final possibility for the electron, namely that it recombines, has a probability governed by the diffusion length of the carrier and the junction depth. In a high-voltage, 0.35µm process, the p+ region is doped with silicon that is roughly $2 \cdot 10^{19}$ per cm^3. At these doping levels, the mean free path of the carrier is larger than 2µm[57], roughly 10 times larger than the junction depth between the p+ and the n. Thus the probability of recombination can be neglected.

Assuming the carrier follows a random walk during diffusion, where it is as likely to move towards the depletion region as towards the surface, the probability p_d of entering the depletion region will be,

$$p_d(z) = \frac{1}{2}p_d(z + dz) + \frac{1}{2}p(z - dz). \tag{2.19}$$

With boundary conditions $p_d(0) = 0$ and $p_d(z_0) = 1$, a linear equation $p_d(z) = z/z_o$ meets all relevant criteria. Since the probability of triggering an avalanche at z_0 is $p(z_0)$, when $z < z_0$, $p(z) = \frac{z}{z_o}p(z_0)$.

Similarly, for the condition of the electron-hole pair being generated in the deep n-well, $z_1 < z < z_w$, the probability that the avalanche is triggered will vary as $p(z) = \frac{z-z_w}{z_1-z_w}p(z_1)$. If the diode is an n+—p diode that is directly on the substrate, then the probability of recombination is no longer trivial. Hyperbolic functions govern the probability of the minority carrier reaching the junction when recombination is no longer a trivial factor[38].

A graph of the carrier detection probabilities as a function of electron-hole pair generation depth is shown in Fig. 2.13. The five conditions are clearly visible on the graph. When $z < 0$ or $z > z_w$, the probability is zero. Between $z = 0$ and 0.2µm, the CDP increases linearly, exactly matching $p_e(z)$. In the depleted silicon, between 0.2 and slightly less than 1.0µm, the CDP is governed by the differential equations (2.17) and (2.18), with $p(z)$'s compisition shifting from $p_e(z)$ at the top of the depleted silicon to $p_h(z)$ at the bottom. Like the initial region, between the edge of the depleted silicon $z \approx 1.0$µm and the edge of the well $z = 5.0$µm, the detection probability decreases linearly.

Transmission Effects

Before calculating the PDP, it is necessary to understand the effect of the materials separating the silicon from air. While other materials may be used in the optical stack above the chip, silicon dioxide and silicon nitride are the two most prevalent materials. In thicknesses of several microns, standard values for CMOS processes, these materials absorp very little light; however, depending on the material thickness, they may be highly reflective. Due to

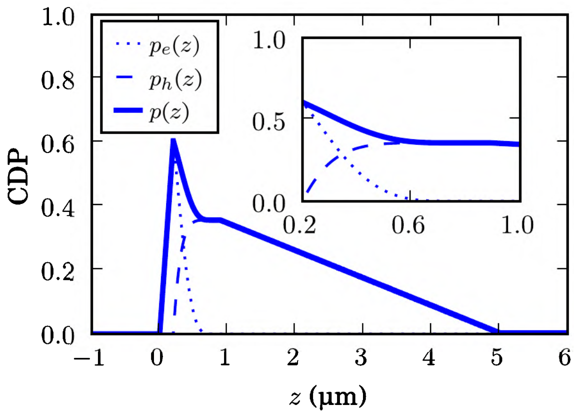


Figure 2.13: **Simulated Carrier Detection Probability vs. Injection Depth** is shown for an abrupt, p+—n junction(main graph), with a zoom of the depletion region itself (inset). The p+—n junction is assumed to have a breakdown voltage of 18V, with an excess bias of 2V. Ionization coefficients were from [58].

the small complex refractive index of silicon dioxide for visible or near visible light, absorption in the silicon dioxide will be ignored.

The net tranmission of light with wavelength λ through a silicon dioxide interface with thickness d_{ox} between air and silicon will add with the reflection to be add,

$$T(\lambda) = 1 - R(\lambda), \tag{2.20}$$

with the reflection given by

$$R(\lambda) = \frac{\left((n_s - 1)\cos(\theta) + \frac{k_s}{n_{ox}} sin(\theta)\right)^2 + \left(\left(\frac{n_s}{n_{ox}} - n_{ox}\right)\sin(\theta) - k_s\cos(\theta)\right)^2}{\left((n_s + 1)\cos(\theta) - \frac{k_s}{n_{ox}}\sin(\theta)\right)^2 + \left(\frac{n_s + n_{ox}}{n_{ox}}\sin(\theta) + k_s\cos(\theta)\right)}, \tag{2.21}$$

where n_s and k_s are the real and imaginary refractive indices of silicon, n_{ox} is the real refractive index of silicon dioxide, and $\theta = \frac{2\pi}{\lambda} n_{ox} d_{ox}$ is the light's phase change through the silicon dioxide[55]. If the light does not move orthogonally through the silicon dioxide, θ must be multiplied by the cosine of the angle the photon propagates through the SiO$_2$, which will be a sizeable distortion for light incident nearly orthogonally to the surface. Note that the θ variable hides the tranmission's dependence on the wavelength and the silicon dioxide thickness; in (2.21), θ is used in place of $\theta(\lambda)$ for purposes of

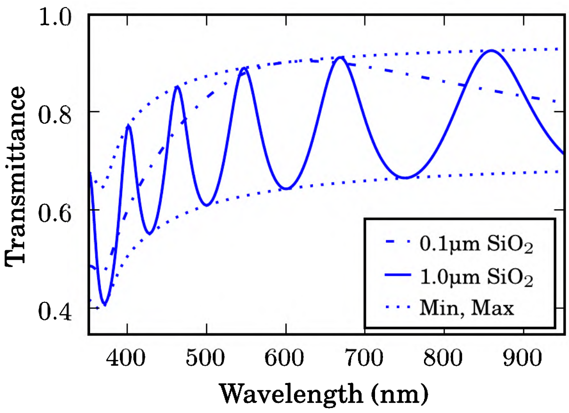


Figure 2.14: **SiO_2 Tranmission Interference Patterns** are shown for various silicon dioxide thicknesses, along with the minimum, maximum, and modeled values. After [55]

brevity.

Fig. 2.14 shows interference patterns that can be seen from the silicon dioxide at two different thicknesses, along with the minimum and maximum transmission.

Photon Detection Probability

The PDF governing light's absorption in silicon as a function of depth follows an exponential process,

$$f_\lambda(z) \begin{cases} 0 & \text{if } z = 0, \\ \mu(\lambda) \exp\left(-\mu(\lambda) z\right) & \text{if } z \geq 0, \end{cases} \tag{2.22}$$

with $\mu(\lambda)$ being the mean penetration depth of the light into the silicon. Because silicon is an indirect bandgap material, μ changes very rapidly from the minimum bandgap, ~ 1 eV, to the maximum bandgap, ~ 4 eV. Below the minimum bandgap, silicon is transparent to light; just above the maximum bandgap, light penetrates only a small distance (tens of atomic layers or less) into silicon. Between these values, optical phonons from the silicon must impart some energy when creating the eletron hole pair[59]. Fig. 2.15 shows the mean penetration depth into silicon as a function of wavelength.

The photon detection probability will be the integral of electron-holes pairs generated at z multiplied by the chance that these carriers cause an

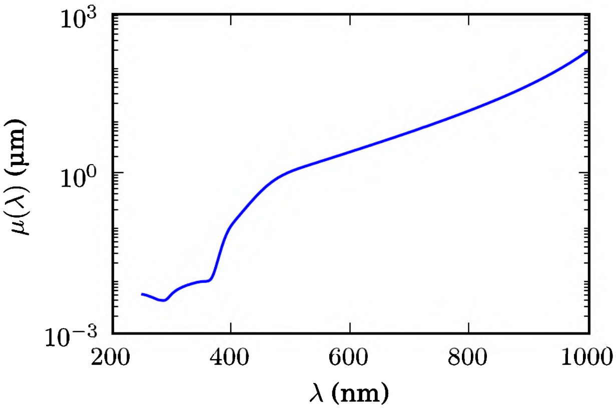


Figure 2.15: **Light's Mean Penetration Depth vs. Wavelength** is shown for visible and near visible light incident on silicon. After [60]

avalanche, which is

$$\mathrm{PDP}(\lambda) = \int_{-\infty}^{\infty} T(\lambda) f_{\lambda}(z) p(z) dz, \tag{2.23}$$

$$= \int_{z_w}^{0} T(\lambda)\mu(\lambda)\exp(-\mu(\lambda)z)p(z)dz. \tag{2.24}$$

Fig. 2.16 shows the PDP for the diode previously given as an example in Fig. 2.13, with transmission effects from a 1µm $\mathrm{SiO_2}$ slab above the silicon. Light with a wavelength less than 400nm generates electron-hole pairs too shallowly for a high PDP; these carriers will become trapped at the surface due to the slight positive charge of the silicon dioxide. Light with a wavelength longer than 500nm generates electron-hole pairs too deeply; carriers will probably be swept into or isolated in the substrate. Between 400nm and 500nm, the light's penetration depth matches the peak values of the CDP, and this light has the best chances of causing an avalanche.

The transmission patterns of the $\mathrm{SiO_2}$ are clearly evident in the figure. Also shown in Fig. 2.16 is a PDP curve with no contribution from the minority carriers generated in the n well. These carriers more than double the expected PDP at wavelengths above 600 nm.

As $\mathrm{V_{eb}}$ increases, there are diminishing returns to the increase in $p(z)$ with the increasing electrical field, causing PDP saturation. Theoretically, the PDP could increase until the entire n-well becomes depleted; practically, noise will limit PDP increases beyond a certain operating point, as well as

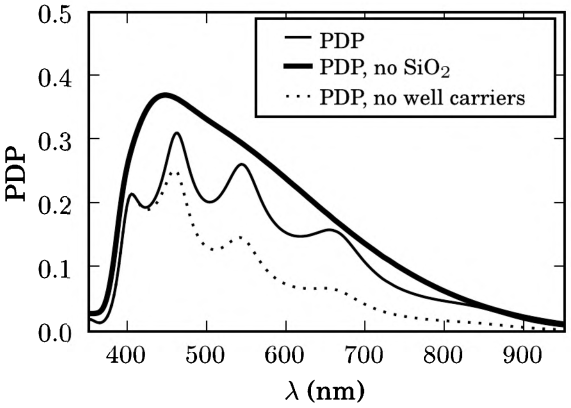


Figure 2.16: **Simulated Photon Detection Probability vs. Wavelength** is shown for visible and near visible light incident on the same junction presented in Fig. 2.13 (thin line). Also shown are the PDP without the SiO_2 (thick line), and without carriers generated in the n well (dashed line)

punch-through that may appear near the guard rings.

2.4.2 Optical Emission

During an avalanche, carrier acceleration, deceleration, and recombination will cause emission of photons. Previous work suggests that the hot carrier braking mechanisms are responsible for the majority of emitted carriers[61]. The majority of emitted carriers interact sparingly with the silicon, as their energy is so low, and these carriers will cause optical crosstalk. This crosstalk has been previously harvested to create opto-couplers[62, 63]. Ch. 3 describes several characterization techniques using the emission. The emitted photons, however, are generally considered unwanted; in quantum key distribution, they even present a security risk[64].

There are two relevant figures of merit (FOMs) for optical emission from SPADs, and both are quite similar to LED FOMs. The first is the number of emitted photons per carrier, which describes the expected number of photons to be generated from a carrier in the SPAD's depletion region. The second FOM is the spectrum of emitted photons.

In silicon SPADs, there are $\sim 10^{-5}$ photons emitted per carrier, with a SPAD's output light efficiency several orders of magnitude lower than most LEDs[61, 51]. The output spectrum peaks in the red and near-IR wavelengths, with the silicon absorbing some of output photons[64]. Despite this seemingly poor output efficiency, optical crosstalk can be sizeable in large arrays, as described in Sec. 2.4.4.

2.4.3 Timing Jitter

Due to the statistical nature of the avalanche build-up, a SPAD's output waveform will not have an identical build-up for every injected carrier[65]. Instead, there will be uncertainty in the waveform following carrier injection, with some of the jitter coming from static effects such as trigger position[66]. Carriers that must diffuse to the depletion region will exacerbate the effect, since an unknown but quantifiable delay will be introduced by the diffusion process[67]. The exponential tail has been observed to be more prominent for lower wavelengths of incident light[68], due to the fact that carriers generated in the p+ region must also diffuse to reach the depletion region; however, this only occurs when working with UV light around 380nm, which has hitherto been uncommon.

The total uncertainty for a SPAD coupled to a time measurement device is the timing jitter. The jitter is often modeled as an Gaussian curve convolved with the sum of a delta function and one or more exponential functions. The delta function represents carriers created directly in the depletion region, with the exponentials modelling the carriers which must diffuse to the depletion region; the sum of these components is the time distribution for the carriers to reach the depletion region. The gaussian component represents the timing uncertainty caused by the statistical nature of the ionization process, justified by the independent nature of the ionizations and the central limit theorem[69].

2.4.4 Noise

There are many undesireable sources of carrier injection in an avalanche diode. Uncorrelated noise is usually divided into tunneling-assisted noise and trap-assisted noise, though noise may be a combination of effects, e.g. trap-assisted tunneling. The dominant types of correlated noise are afterpulsing and crosstalk. Fig. 2.17 shows all of these types of noise in one figure. This section will discuss these sources of noise, beginning with the uncorrelated noise. Both types depend on the generation rate of carriers given specific electric fields, $G(|\vec{E}|)$, as these carriers will cause the spurious avalanches. To achieve the noise rate, the noise rate for a particular point can be found by multiplying that point's CDP times the generation rate; integrating this value over the active volume will yield the total noise rate.

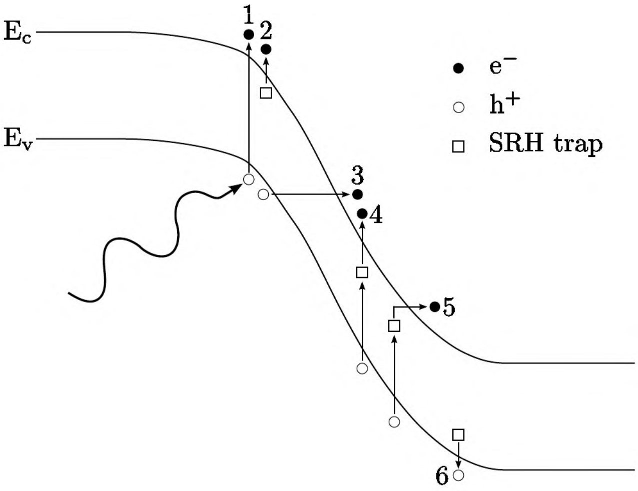


Figure 2.17: **Noise Sources** — (1) Crosstalk from a recombintation-generated photon; (2) Electron afterpulsing; (3) Band-to-band tunneling; (4) Trap-assisted thermal generation; (5) Trap-assisted tunneling; (6) Hole afterpulsing. After [29, 21]

Uncorrelated Noise

In the theory of quantum mechanics, atomic particles can "tunnel" though potential barriers that would classically require more energy than the particles have[70]. The tunneling effect is important for SPADs operating with electric fields above ∼ 1MV/cm; at this electric field strength, tunneling increases dramatically. Tunneling rates are generally modeled by the equation,

$$G_{\mathrm{tu}}(|\vec{E}|) = B \cdot |E(\vec{z})|^{2.5} \exp\left(E_0/|E(\vec{z})|\right), \tag{2.25}$$

with B and E_0 being material constants[71]. The B constant is temperature dependent but E_0 is not; when all factors are considered, tunneling noise increases by a factor of two over a temperature change of more than 100°C. Uncorrelated noise that is nearly temperature independent is usually caused by tunneling.

Trap-assisted noise is more temperature dependent. The Shockley-Read-Hall (SRH) theory models carrier capture and release with lattice defects, or "traps," allowing energy states in silicon's energy gap[72]. The presence of traps greatly increases the probability of thermal carrier generation. The generation and recombination rate in traps depends on the implantation and annealing processes during chip fabrication. Following the assumption that hole and electron trap and release characteristics are the same, the trap-generated carrier rates can be approximated by

$$G_{\mathrm{tr}}(|\vec{E}|) = \frac{n_{tr}(1 + \Gamma(|\vec{E}|))}{2\tau \cosh\left(\frac{E_t - E_i}{kT}\right)}, \tag{2.26}$$

with n_{tr} being the trap concentration, Γ a correction factor based on the electrid field strength, E_t the trap's energy level, E_i the intrinsic Fermi level, and τ the mean capture lifetime[71].

Trap-assisted noise, and trap-assisted tunneling noise, depend strongly on the temperature, with a factor of 2 increase in noise for every 10°C increase in ambient temperature.

The contributions of the two types of noise can be observed in a diode with a simple temperature sweep. If, at a constant excess bias, the noise doubles with a 10°C shift in ambient temperature, it is largely trap-assisted. If the noise varies by less than 10%, then the noise is tunneling-limitied. If the noise changes with a factor between these two quantites, then the noise's causes are distributed between the two types of noise.

Due to the dependence of tunneling on the electric field, tunneling-assisted noise is dominant in SPADs with low breakdown voltages, as these SPADs will have higher electric fields[29, 31].

Across many device structures in different CMOS processes, the distribution of DCRs in an array of SPADs is skewed, with a few pixels causing the vast majority most of the noise[73, 74, 29]. For this reason the median DCR is usually quoted rather than the mean DCR; the two quantities have been reported to vary by at least a factor of 4[73]. Whether the high noise is caused by a lattice defect, a contaminant atom, or a local high field due to implant atoms being within angstorms of one another is unclear.

Correlated Noise: Afterpulsing

Traps allowing energy levels close to the energy bands can capture and hold carriers during the avalanche process. These traps cause a type of correlated noise known as afterpulsing[44]. During an avalanche, these traps can capture and hold carriers, with a release lifetime on the order of nanoseconds. Afterpulsing limits the minimum hold off time, also known as the dead time, of the SPAD. Dead times that are too short allow afterpulsing and introduce correlated noise, though of course if the dead time is set too long then the SPAD spends too much time recovering from the avalanche. Because ambient energy aids in the release of trapped carriers, afterpulsing will be worse at lower temperatures.

Are the expected fraction of events that are afterpulses equal to the probability of an afterpulse per avalanche? Let P_{ap} be the probability of an afterpulse per avalanche. P_{ap} will increase as more charge carriers are trapped, which occurs as the excess bias increases. Increases in the excess bias also make detection of released carriers more likely, again increasing P_{ao}. P_{ap} is not the expected number of afterpulses per avalanche, as the distribution of afterpulses per non-afterpulse avalanche is a geometric distribution[1] with rate parameter $1 - P_{ap}$, having expected value $P_{ap}/(1 - P_{ap})$, not P_{ap}. To extract the total fraction of events that are afterpulses, let the event rate without afterpulses be $e_{\overline{a}}$, with units of Hz. Neglecting dead time distortions, the total number of events per second will be the sum of the uncorrelated noise, and the noise from afterpusling, in this case $e_{\overline{a}} + e_{\overline{a}}P_{ap}/(1 - P_{ap}) = e_{\overline{a}}/(1 - P_{ap})$. The fraction of events that are afterpulses is $[e_{\overline{a}}P_{ap}/(1 - P_{ap})] \,/\, [e/(1 - P_{ap})] = P_{ap}$, meaning that the fraction of events that are afterpulses is equal to the probability that an avalanche will have an afterpulse.

Correlated Noise: Crosstalk

Crosstalk may be split into optical crosstalk and electrical crosstalk. The probability of an avalanche causing one or more other avalanches via crosstalk will be denoted by P_{ct}.

Electrical crosstalk can occur on either the $\mathrm{V_{op}}$ line, or the power supply line. If a large number of SPADs with a shared $\mathrm{V_{op}}$ line fire simultaneously, for example, IR drop on this line can reduce the local value of $\mathrm{V_{op}}$, decreasing the excess bias and causing all the effects such an action would entail. To the author's knowledge, this effect has never been widely reported in the literature.

Optical crosstalk can originate from photons emitted during an avalanche, whose origin Sec. 2.4.2 describes. Due to the red and near-IR wavelengths of these photons, they can travel hundreds of microns in silicon before generating electron-hole pairs. Optical crosstalk has been observed in densely packed SPAD arrays[75], with models being formed to predict the behavior[76].

2.4.5 Dead Time

Following an avalanache, whether noise or not, a SPAD is unable to detect further free carriers for a short period of time termed the dead time. t_d will denote the dead time. Fig. 2.18 shows how the dead time can vary for a passively recharged SPAD; for these devices, either the mean or minimum value of t_d should be considered.

[1] A geometric distribution with output support of $\{0, 1, 2, ...\}$, not $\{1, 2, 3, ...\}$.

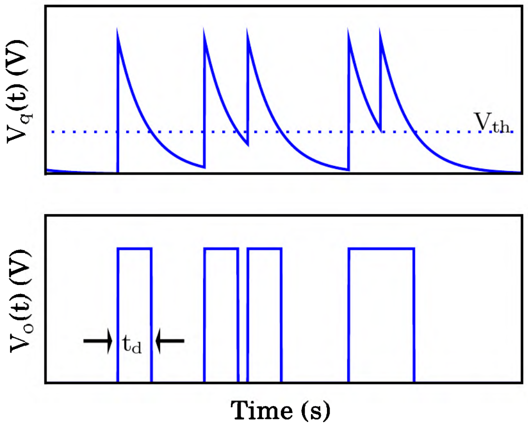


Figure 2.18: **Dead Time**—shown is the SPAD's quench waveform (top) and comparator output (bottom) from Fig. 2.8. Note that the dead time will change in a passive recharge scheme if an avalanche occurs before recharge is complete (last output pulse). The RC time constant comes from the quenching resistance and parasitic capacitance.

2.4.6 Dynamic Range

Let the SNR[1] be defined as

$$\mathrm{SNR} = 20\log_{10}\left(\frac{e_{\text{signal}}}{\sigma}\right), \tag{2.27}$$

$$\approx 20\log_{10}\left(\frac{e - e_{\text{noise}}}{\sqrt{\sigma^2_{\text{noise}} + \sigma^2_{\text{signal}}}}\right), \tag{2.28}$$

$$\approx 20\log_{10}\left(\frac{e - e_{\text{noise}}}{\sqrt{\mu}}\right), \tag{2.29}$$

with μ the count rate and σ the variance[77], both having subscripts denoting signal or noise information, with no subscript meaning all events included.

Neglecting the count rate distortions from the dead time, a SPAD has a theoretical maximum signal count rate of the order of magnitude $t_i/t_d - \mathrm{DCR}$,

[1] It should be noted that the SNR definition used by the imaging community differs with that used by other communities. Technically the definition should be $10\log_{10}(P_{\text{signal}}/P_{\text{noise}})$, not $20\log_{10}(P_{\text{signal}}/P_{\text{noise}})$. Historically the imaging community has used $20\log_{10}$ since imagers measure voltages, and normally the square of the voltage varies with the power. However, in an ideal imager, the voltage will vary linearly with the incident optical power. A coefficient of 20 will be used here for purposes of consistency with the rest of the imaging community.

with an integration period of time t_i. The minimum usable SNR will be assumed to be zero. Using the definition from (2.29) and assuming that shot noise governs the noise, such that the noise's variance is equal to its mean, the maximum dynamic range[77, 78] will be,

$$20 \log_{10} (t_i/t_d) - 10 \log_{10} (t_i/t_d + \mathrm{DCR}) \tag{2.30}$$

In reality, the distortion in count rate from the dead time must be taken into account. This distortion comes from the fact that, for a measured event rate e, the fraction of time the detector is dead is et_d, so the measured event rate must be multiplied by $1/(1 - et_d)$ to compensate for the dead time. If et_d is small, then the distortion will also be small, and the count rate will be a good first order estimate (not taking noise into account). A SPAD's count rate has been linearized to match the input output power with <5% error over six orders of magnitude[79].

2.5 FOM Summary

Table 2.1 shows a summary of the complex trade-offs in fabricating and operating a p+—n SPAD in an n-well with a p-doped guard ring. The variables effecting the most parameters are the deep well doping, the excess bias, and the diode area. Ch. 8 analyzes, quantitatively, some of these trade-offs when considering position emission tomography as an application.

		Indep. Variable										
		Junction depth[1]	Deep well depth[1]	Deep well doping[1]	Guard ring depth[1]	V_{eb}	Temperature	Diode area	Active Quench	Active Recharge[2]	Comparator Threshold[3]	Dead time
Effected Variable	Cost		✗↑		✗↑							
	R_o	✗↑	✓↓	✓↓	✓↓			✗↑				
	C_p			✓↓		✓↓		✗↑				
	V_{bd}			✓↓			✓↓					
	DCR (corr.)			✗↑		✗↑	✗↑	✗↑				
	DCR (unc.)			✗↑		✗↑	✓↓	✗↑	✓↓			✓↓
	Blue PDP	✗↓				✓↑						
	Red PDP	✓↑	✓↑	✗↓		✓↑						
	Jitt. (blue)	✗↑				✓↓	✗↑	✗↑			✗↑	
	Jitt. (red)	✓↓	✗↓	✗↑		✓↓	✗↑	✗↑			✗↑	
	Dyn. Rng.					✗↓						✗↓
	Fill factor							✓↑	✗↓	✗↓	✗↓	

[1] CMOS process variables are fixed for a particular process

[2] Active Recharge gives better saturation behavior, which is not shown.

[3] A higher comparator threshold in the case of a SPAD with changing voltage at the anode

Table 2.1: **FOM Trade-offs** — An up arrow signifies the effect variable increases as the process variable increases, with a down arrow showing the effected variable decreasing. ✗ signifies an undesirable change, with ✓ standing for a desired change.

Chapter 3

Metrology

This chapter compares a number of SPAD characterization techniques, and is meant to serve as a guide for characterizing SPAD performance. Some of the material is original, including the analysis of the afterpulsing methods, the analysis of the breakdown voltage methods, and the methods for measuring the inactive distance. The difference between original and non-original material is not made explicit to aid in readability.

3.1 Breakdown Voltage

As Sec. 2.1.1 describes, $\mathrm{V_{bd}}$ is defined quantitatively for a p-n junction to be the voltage for which the condition in (2.1) is met. However, (2.1) varies as a function of the local doping level, which is not a constant value. Additionally, as Sec. 2.2.4 describes, edge effects can cause the breakdown voltage to increase near the guard ring. When measuring the breakdown voltage, it is important to consider that any single number will be some function of a spatially varying $\mathrm{V_{bd}}$.

There are four ways to measure the *in situ* breakdown voltage that will be discussed:

1. The I-V method
2. The sweep and subtract method
3. Fit to DCR
4. Fit to ECR

Fig. 3.2 shows all of these methods graphically for the circuit in Fig. 3.1. Due to the presence of the quenching resistor, just the diode's I-V characteristic will not be available. If $\mathrm{R_q}$ can be set low enough, then a resistance-limited I-V curve can be acquired by measuring the current output from $\mathrm{V_{op}}$.

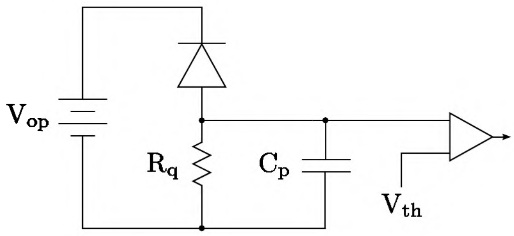


Figure 3.1: **A Passively Quench SPAD**

The resistance must be small enough that an avalanche is not successfully quenched. In this mode of operation, the diode will source enough current to remain at its breakdown voltage, causing the current sourced from V_{op} as $(V_{op} - V_{bd})/R_q$, which is linear with V_{op}. If there are intra-diode spatial variations of the breakdown voltage larger than the step size in the V_{op} sweep, these variations will cause non-linear behavior in the x-intercept of the I-V curve, when some regions of the diode are under breakdown but not others. Applying a linear fit to the higher current region of the curve will yield the mean breakdown voltage in the x-intercept. This method will be called the "I-V fit" method, as shown in the top sub-plot of Fig. 3.2.

There are several disadvantages to this method, the main one being that usually the V_{op} line sources power to more than one diode. If there are variations between the diode's breakdown voltages, then this method will only provide a single estimate for the diodes' breakdown voltages. The measurements will be distorted if there is resistance in V_{op}'s routing that is comparable to R_q. The total current can also cause heating in large arrays, which distorts the breakdown voltage measurement.

Capturing the diode's optical emission, which is proportional to the electrical current flowing through the junction, can be used to estimate the diode's current. However, due to the inefficiency in emission, measuring the current electrically will give better SNR. Still, the emission technique allows a visual check that the breakdown voltage has been correctly estimated, at least within a few hundred mV.

If the comparator's threshold is known and V_{op} can be changed, V_{op} can be swept until pulses appear at the output of the thresholder, with $V_{bd} = V_{op} - V_{th}$. This will be termed the "Sweep and Subtract" method. The measurement must be performed in the dark when the threshold is low; when the SPAD is operating in the linear-mode rather than the Geiger-mode, a large number of simultaneously incident photons can trigger the comparator. The disadvantage to this method is that the comparator's threshold might not be known, or may vary largely from the expected value. For CMOS processes, the intra-chip variation of the transistor's threshold voltages, which is the dominant factor in variance of V_{th}, is normally well controlled and is quite small, less than 100mV. Additionally, this method is based on the minimum breakdown voltage of any region in the diode itself, rather than the mean breakdown voltage across the diode, since the first region exhibiting breakdown will cause the first pulses as V_{op} is increased.

The DCR and excess count rate (ECR), which is the count rate above the DCR, both can also be used to estimate the breakdown voltage. The DCR and ECR will exhibit small-signal linear behavior as the applied voltage is varied, allowing a linear fit to estimate the breakdown voltage. These methods will be the "DCR fit" and "ECR fit" methods, respectively. Distortions will appear in V_{bd}'s estimate as V_{th} and V_{op} increase; PDP saturation will cause underestimation in the ECR, and the exponential behavior[29] of the DCR will cause overestimation. Both methods can estimate the V_{bd} *in situ* for an array, and estimate the mean breakdown voltage of the diode.

Fig. 3.2 shows experimental results of the methods when applied to a single diode coupled to a variable thresholder in an array of four diodes. Fig. 3.2 also lists the various strengths and weaknesses of each of the methods. Two different values of V_{th} were applied to the DCR fit and ECR fit methods to show the over- and underestimation of the breakdown voltage with increasing V_{th}. When V_{th} is set to 100 mV, all four methods produce V_{bd} estimates of 18.6V±0.08V. The I-V fit method, however, shows the highest estimate; because this method measures the mean breakdown voltage of the four diodes, rather than the mean breakdown voltage of a single diode, a larger error is to be expected from this estimate. The sweep and subtract method shows a lower estimate, ostensibly because this method measures the minimum spatial breakdown voltage, rather than the mean. As V_{th} is increased to 1.0V for the ECR and DCR fit methods, under- and overestimations of V_{bd} are produced, with a shift of -0.4V for the ECR fit method and $+0.8$V for the DCR fit method.

3.2 Parasitic Capacitance

The parasitic capacitance at the comparator-connected node is an important factor in the propagation behavior and the timing response of the SPAD. If simulation models are available for transistors, which is almost always the case for CMOS processes, Cadence Spectre[80] or SPICE simulations of the diode's recharge time can estimate the total capacitance at this node (see Fig. 2.18). This technique is most effective for a passively quenched diode, with a transistor implemented as R_q; in this case, the bias voltage to the

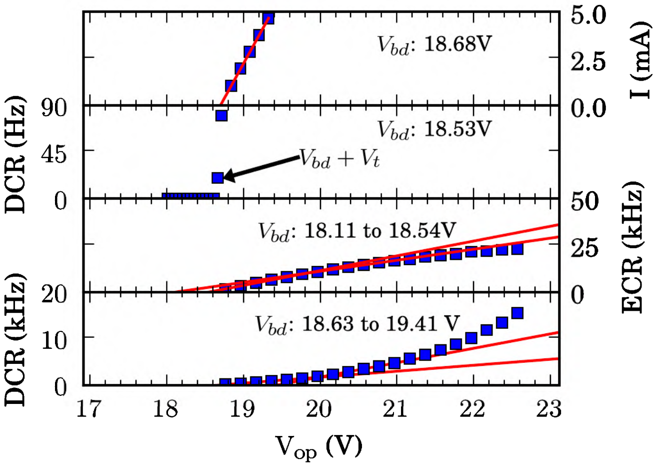


Figure 3.2: **Measurements of V$_{bd}$** — shown are results from four different methods, listed above with their strengths and weaknesses, for estimating the breakdown voltage of a single SPAD in an array

transistor can be varied, and the SPAD's dead time can be compared to the simulated value for different capacitances, allowing better estimation of C$_p$.

3.3 Noise

Measuring a SPAD's DCR seems simple; just count the number of output pulses, n, per second while the SPAD is in the dark. The output sampling is subject to shot noise variations, with RMS $\sqrt{n}$[1], though the mean noise can be measured arbitrarily close to the actual value by increasing the integration time. Afterpulsing and dead time will also distort this estimate; in practice these distortions are small.

As described in Sec. 2.4.6, if the diode spends a large fraction of the time recharging, all count rates must be multiplied by $1/(1-f_d)$, with f_d being the fraction of time the diode is inactive, to compensate for the dead time effects.

3.3.1 Random Telegraph Signal Noise

Some SPADs exhibit random telegraph signal (RTS) noise[81], also called burst noise or popcorn noise. Fig. 3.3 shows such the noise rate of a SPAD

[1] The sampling will follow a Poisson distribution.

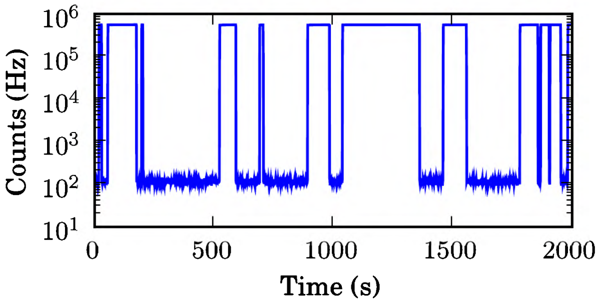


Figure 3.3: **Random Telegraph Signal Noise** — The DCR of a non-irradiated SPAD is shown to switch between two noise rates, one of ~600kHz and another of 110Hz.

with such noise. The phenomena has been previously observed in SPADs irradiated either with γ-rays or $\alpha+$ particles. Ch. 6 presents a non-irradiated device that exhibits the phenomena. For large arrays, it might not be practical to plot the waveform of every single pixel's DCR. In this case, RTS noise can be observed by repeatedly sampling the count rate, and testing whether the fraction of events that are within 2 or 3 σ from the mean count rate matches that of a Poisson or Gaussian distribution; when sampling switching, disparate count rates, such as those that RTS-containing diodes exhibit, too few samples will fall close to the mean. Quantitatively, ~95% of samples are expected to fall within 2σ of the mean μ; if fewer than some similar threshold, say 90%, fall within $\mu \pm 2\sigma$, then the waveform should probably be examined for RTS noise.

3.3.2 Afterpulsing

There are four possible methods of measuring afterpulsing:

1. The autocorrelation method
2. The inter-arrival time histogram method
3. The DCR-based method
4. Statistical methods

Method one, the autocorrelation method, is the most common method. Neglecting crosstalk, non-afterpulsing noise will be uncorrelated, i.e.

$$P[a_{\overline{ap}}(t)|a(0)]\Delta t \approx e_{\overline{ap}}\Delta t, \tag{3.1}$$

with $e_{\overline{ap}}$ being the event rate of uncorrelated avalanches, Δt being a small time interval, and $a(t)$ signifying that an avalanche occurs at time t. $a(t)$ is composed of mutually exclusive sub-components that the avalanche was not afterpulsing, $a_{\overline{ap}}(t)$, or was afterpulsing, $a_{ap}(t)$. The probability density of afterpulsing occurring at time t, $P[a_{ap}(t)]$, will be related to $P[a(t)|a(0)]$, the probability of an avalanche at time t given an avalanche at time 0, by

$$\begin{aligned} P[a(t)|a(0)]\Delta t &= (P[a_{ap}(t)|a(0)] + P[a_{\overline{ap}}(t)|a(0)])\Delta t, \quad (3.2) \\ &\approx P[a_{ap}(t)|a(0)]\Delta t + e_{\overline{ap}}\Delta t. \quad (3.3) \end{aligned}$$

$P[a_{ap}(t)]$ is not a probability density function, it is simply probability density. $P[a(t)|a(0)]$ can be acquired by tracking the time between the rising edge of avalanche pulses. Let $h[i]$ be a discrete histogram of time difference between the rising edge of a particular avalanche with the rising edges of all following avalanches, with bin i signifying the number of avalanches with a time difference $i\Delta t$ and $(i+1)\Delta t$ between a particular avalanche and subsequent avalanches. The probability of an avalanche occuring between $i\Delta t$ and $(i+1)\Delta t$ in this configuration is $P[a(i\Delta t)|a(0)]\Delta t \approx \frac{h[i]}{\sum_{j=-\infty}^{\infty}(h[j])}$. The probability density of an afterpulse occurring at time t corresponding to index i (i.e. $i = \lfloor t/\Delta t \rfloor$) will be

$$\begin{aligned} P[a_{ap}(t)|a(0)]\Delta t &\approx P[a(t)|a(0)]\Delta t - e_{\overline{ap}}\Delta t, \quad (3.4) \\ &\approx \frac{h[i]}{\sum_{j=-\infty}^{\infty}(h[j])} - e_{\overline{ap}}\Delta t. \quad (3.5) \end{aligned}$$

Because the fraction $h[i]/(\sum_{j=-\infty}^{\infty}(h[j]))$ is the autocorrelation function of the rising edges of the output comparators waveform, (3.5) can be represented as

$$P[a_{ap}(t)|a(0)]\Delta t \approx n_{\overline{ap}} \cdot (G^{(2)}(s) - 1), \quad (3.6)$$

with s being a waveform containing the rising edges of the comparator's output[1].

Equation (3.6) estimates the probability density that an afterpulse will occur at some time after an avalanche, rather than the probability that an afterpulse will occur. The two definitions are not the same, and integrating the density derived from the autocorrelation function will yield the expected number of afterpulses for an avalanche, rather than the probability of an afterpulse. As described in Sec. 2.4.4, these two quantities are not equal, so

[1] Depending on how the normalized autocorrelation is performed, (3.6) may or many not have $e_{\overline{ap}}$ in the left-hand side. Note, however, the normalization must correct for the lack of the Δt term multiplying $h[i]$ in (3.5).

the actual probability of afterpulsing follows as

$$n_a = E[\text{Afterpulses per event}] = \int_{t_d}^{\infty} P[a_{ap}(t)|a(0)]dt, \tag{3.7}$$

$$P_{ap} = \frac{n_a}{1+n_a}, \tag{3.8}$$

where t_d is the dead time. The autocorrelation method will distort the probability density of afterpulsing if the dead time varies. In a passive recharge scheme, afterpulses can occur while the diode is partially restored but still not below the counting logic's trigger threshold, causing the output pulse's duration to lengthen and delaying subsequent pulses by some amount of time. This creates problems when deciding what value to actually use for t_d, since the density could conceptually drop to a negative value if the dead time variance is very high. When sizable distortion occurs in the dead time for a particular passive restore circuit, the afterpulsing probability density is only good for that particular restore circuit. If the restore occurs due to some sort of bias level, such as a transistor which controls the dead time, the bias must be swept to find P_{ap} as a function of the operating point. Actively quenched diodes do not suffer from these problems.

The inter-arrival time histogram method relies on extracting correlations between the rising edge of the time of avalanche pulses. Because afterpulsing occurs on time scales of hundreds of nanoseconds at room temperature, the histogram of inter-avalanche arrival times will show multi-exponential behavior, with the slowest exponential resulting from the uncorrelated noise and any incident light. The afterpulsing probability at a specific dead time can be found by taking the inter-avalanche time histogram, fitting an exponential to the uncorrelated noise source, and then finding the fraction of events above the fit curve but below the experimental curve[1]. The exponential fit to the uncorrelated noise source will have a time constant of $1/e_{\overline{ap}}$.

Like the first method, the acquisition and fit procedures must be carried out for multiple dead times in a passively restored SPAD with high afterpulsing, or distortions will arise from the variable dead time.

Afterpulsing has also been experimentally measured using gating[82]. The gating scheme acquires the same information regarding avalanche triggering probability following an avalanche as the autocorrelation or the inter-avalanche time methods, with a nearly identical analysis of the probability following.

Method three relies on relating the expected number of afterpulses to the base noise rate. It is expected that the event rate, e, will increase as a

[1] Because afterpulsing is uncorrelated with the event rate, additional light can be added to low noise diodes to make it easier to gather statistically significant data for fits.

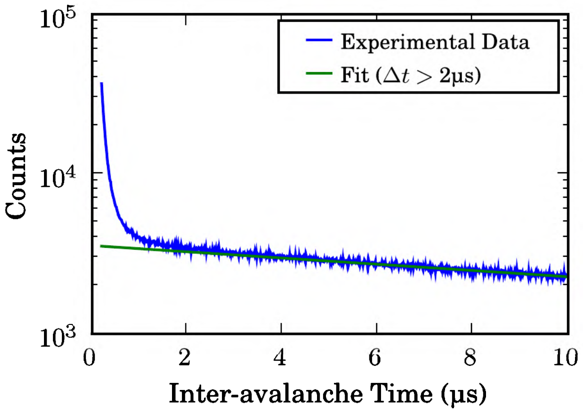


Figure 3.4: **Inter-avalanche Time Method** — shown is a histogram of inter-avalanche arrival times, along with an exponential fit to times larger than 1µs. The afterpulsing probability is the area between the two curves, divided by the area under the experimentally acquired curve.

function of P_{ap} as

$$e = e_{\overline{ap}}/(1 - P_{ap}), \tag{3.9}$$

with $e_{\overline{ap}}$ being the afterpulsing-free event rate. Because afterpulsing occurs in the first few microseconds following an avalanche, $e_{\overline{ap}}$ can be found by setting the dead time to a duration longer than, say, ∼10µs. Then, the dead time can be swept, and the afterpulsing probability as a function of the count rate e will be $1 - e/e_{\overline{ap}}$.

Method four relies on relating the distribution of DCR samples to a Poisson distribution distorted by the afterpulsing. Discussion of this method is beyond the scope of this thesis.

The four methods for measuring afterpulsing all have different strengths and weaknesses. The first and second methods are very similar, both relying on inter-avalanche time differences to estimate the afterpulsing, though method one looks at the time differences between one and many avalanches, whereas the second method looks at the difference in pair timing. However, the first method requires more complex data acquisition, while the second method requires more complex data analysis. In the first method, the time between a particular avalanche and subsequent avalanches must be acquired, compared to the second method which only requires the time difference between consecutive avalanches. Exponential fits with time constant $1/e_{\overline{ap}}$ must be made to the data in the second method, a potentially expensive operation

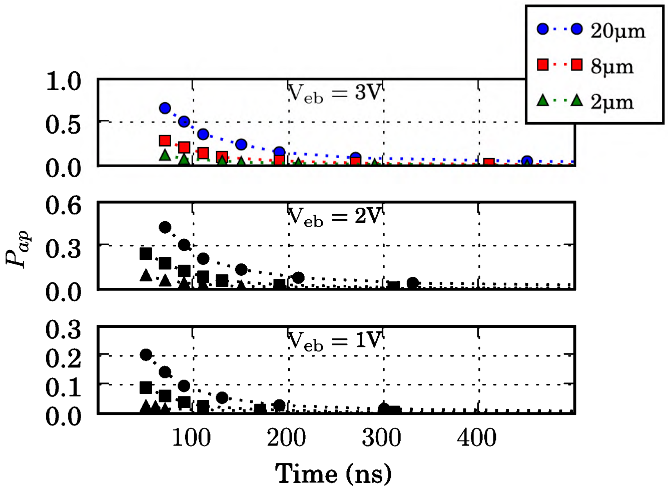


Figure 3.5: **Afterpulsing Measurements** — shown are afterpulsing measurements for three active area diameters.

depending on the size of the data. The first method does require an estimate of $e_{\overline{ap}}$, though this is easily achieved by the average number of counts per bin in the later portion of h. The third method is the simplest, requiring no timing information, only sampling, though it has never been widely used in the literature for some reason, possibly due to concerns about correctly extracting the $e_{\overline{ap}}$ parameter. The fourth method has also never been widely reported in the literature, ostensibly due to concerns that the method is too complex to be practical.

As described in Sec. 2.4.4, afterpulsing has a dependence on both the number of traps near a SPAD's active region, and the total current through the diode. Since the coupled capacitance can be measured as described in Sec. 3.2 and the total active area is known, afterpulsing can be normalized to these two quantities to see which is dominant. Specifically, if the probability of an afterpulse per unit measure is $P_{\square}$, then the probability for m units will be $P_{ap} = 1-(1-P_{\square})^m$. $P_{\square}$ can be estimated as $1-(1-P_{ap})^{1/m}$. Fig. 3.5 shows the afterpulsing probabilities measured at three excess biases for SPADs of different sizes and coupled capacitances. Fig. 3.6 shows these afterpulsing values normalized to a unit area of 100µm, or a unit charge of 1pC. The total charge was estimated using the relation $Q = \mathrm{C_p V_{eb}}$; current that flowed through the quenching resistor was assumed to be negligible. $P_{\square}$ shows a much better match to the unit charge that flows through a diode compared to the unit area. The dominating factor in the afterpulsing measurement will thus be the unit charge, not the unit area, that flows through a diode for

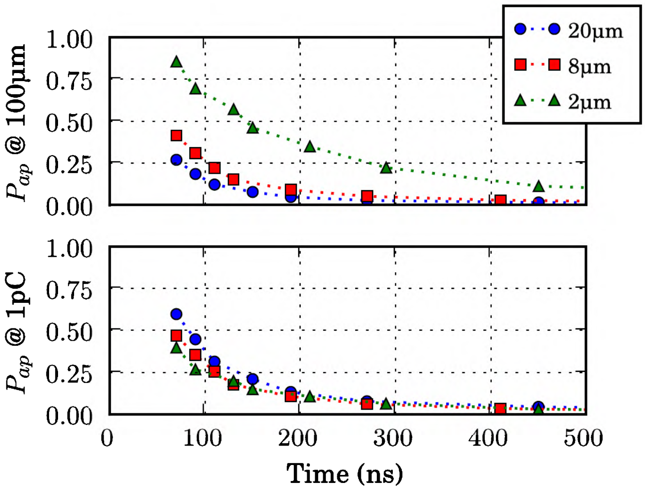


Figure 3.6: **Normalized Afterpulsing** — shown are afterpulsing measurements when normalized for area (top) or carriers (bottom).

total charges in the region of 1pC.

It is important to note that afterpulsing will not create a higher SNR in estimating the incident light on the diode, the SNR will always be at least shot noise limited, and will be DCR-limited for low light levels and saturation limited for high light levels[1].

3.4 Crosstalk

Because crosstalk is correlation between SPADs, whereas afterpulsing is correlated noise in a particular SPAD, measuring crosstalk is nearly identical to measuring afterpulsing. For the correlation method, instead of using the nor-

[1] Try to find the problem with the following analysis, which erroneously shows that the SNR will improve with afterpulsing: Even though afterpulsing causes correlated noise, the distribution of samples of count rate per unit time from a SPAD with non-trivial afterpulsing will not significantly vary from that of an afterpulsing-free SPAD, implying the SNR can be increased for free. To see why this is the case, imagine that sampling the dark counts per a unit time from a SPAD would yield $e_{\overline{ap}}$ non-afterpulsing events. The distribution of samples e from the SPAD itself will be the sum of $e_{\overline{ap}}$ and $e_{\overline{ap}}$ independent geometric distributions with rate parameter P_{ap}. Due to the Barry-Esseen theorem[83] and [84], which quantifies the rate at which the distributions will approach a normal one, the resulting distribution of afterpulsing events will vary from a normal distribution by no more than $\frac{(1-P_{ap}/2)P_{ap}^6}{(1-P_{ap})^{7/2}\sqrt{e_{\overline{ap}}}}$. For $P_{ap} = 0.5$ and $e_{\overline{ap}} = 100$, the CDF of the afterpulsing-seeded samples will vary from a normal CDF by less than 0.001, and when convolved with the distribution of $e_{\overline{ap}}$, itself nearly normal, a normal distribution will result.

malized autocorrelation, the normalized correlation should be used instead. For the inter-avalanche method, the histogram of times it takes one of the diodes to trigger following an avalanche in the other diode should be used. The count-rate-based method can also be used, but it requires that one of the two diodes can be shut off while the other operates, which may not be possible in an array.

3.5 Inactive Distance

The importance of the active area is obvious, considering its relationship to all of a SPAD's FOMs. However, the actual active area of a SPAD may not be the expected active region from a submitted mask set. Annealing of implant layers, processing variations, undocumented processing by the CMOS foundry — there are many factors that can reduce the actual active area of the diode itself. This section discusses using three techniques to estimate the active area. All techniques rely on the assumption that the breakdown voltage is identical across the entire active area.

For example, if implant annealing from the guard rings slowly shifts the doping concentration near the edge of the device, causing a gradual increase in breakdown voltage, the techniques listed below would roughly estimate the halfway distance along the spatial variation of the breakdown voltage, modeling the diode with two regions: the active region in breakdown, and the inactive region not in breakdown. The difference between the expected active area and the realized active area will be captured by the inactive distance d_i, which is the distance from expected active area's edge to the realized active area's edge. Whether the inactive distance decreases or increases with increasing excess bias is important in determining the distance's cause. If the infringement of the depletion region from the guard ring causes the inactive distance, d_i will increase with V_{op}. If the guard ring's implants' annealing causes the inactive distance, d_i will decrease as V_{op} increases.

3.5.1 Optical Emission Test

The first technique for estimating d_i, the optical emission test, relies on observation of the hot-carrier generated photons which produce crosstalk and signal avalanching[85, 61]. If the SPAD is left free-running, then a sensitive infrared camera and multi-second acquisition time will allow direct observation of emitted photons[34]. Because CMOS SPADs tend to be very sensitive to blue light, but emitted photons are red-shifted, illuminating the diode with blue light but placing a high-pass filter between the diode and the camera may aid in acquiring the breakdown uniformity. The drawback of illumination is that, in a large diode, if a specific position is triggering most of the noise, uniform illumination may make the current appear more uniform over the diode than would be the case in the dark. Illumination may more accurately reflect operating conditions — for the purposes of the optical emission test, it is best to accurately reflect the diode's sensitivity during operation, which would include light.

If the quenching resistance is lowered to the point that an avalanche will not properly quench, then the diode will emit more light. The emitted photons may even be visible to the unaided eye[1]. However, when improperly quenched, the emitted carriers will not exactly match the carriers emitted during an avalanche, due to the effects of the well's ohmic resistance. Depending on the dopings and the well contact configuration, certain areas of the SPAD may appear dim, though these areas would have the same breakdown voltage if the SPAD was properly quenched. Care must be taken so that these effects do not appreciably distort the result.

Fig. 3.7 shows the optical emission from a SPAD with a circular active area of an expected diameter of 24µm. The guard rings in this device are completely covered with metal for positional reference, and they are also highlighted in the figure. Using the guard rings as an absolute reference for a distance of 24µm, the inactive distance can be estimated by measuring the circumference of the thresholded circle, which is slightly smaller than 20µm in this case. There are large numbers of well contacts around the outside of this diode, so any ohmic resistance effects would cause the middle of the diode to appear dimmer than the outside edge, not effecting the estimate of d_i. Limited evidence implies that the hot spots in Fig. 3.7 are caused by doping variations in the well[85].

The main advantage of the optical emission test is that it can work on a single diode. The disadvantage of this method is that an optical microscope is required. Also, if the diode needs to be improperly quenched, the quenching resistor must be variable, with a range large enough to allow sizable amounts of current to flow through the diode. For the result in Fig. 3.7, current amplitudes of several mA were required for the optical emission test with a standard camera. In an array, it may not be feasible to source this amount of current to all diodes in the array.

Care should be taken to limit the maximum current through the system so that the diode does not melt, nor the current density rules are exceeded in the routing or contacts.

[1] The effect is easier to see if the diode is made to blink, e.g. if V_{op} is modulated as a square wave with a frequency of, say, 0.5Hz.

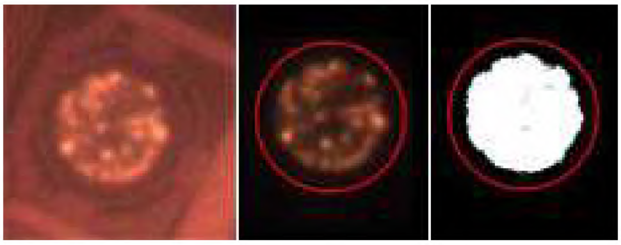


Figure 3.7: **A SPAD's Optical Emission** — Optical emission is shown from an improperly quenched diode with small amounts of incident light (left), in the dark with red overlaid on the guard ring (center), and thresholded at a quarter of the peak intensity (right).

3.5.2 Count Rate Matching

The total DCR is a function of, amongst other things, the total number of traps in the diode and the excess bias. Because the trap concentration is expected to be relatively constant within a particular chip, the DCR normalized to the unit area of diodes with different geometries can be compared to find the inactive distance. As Ch. 7 will show, the trap concentration is not uniform in high noise diodes, so these diodes should not be used for DCR matching. This method is most easily performed with large, circular diodes. The diodes should be large to avoid breakdown voltage distortions caused by the edge effects of the guard rings; if the diodes are circular, corner effects will not be a factor, and can be ignored.

For two circular diodes with radii r_1 and r_2 having DCRs DCR_1 and DCR_2, the inactive distance will satisfy the equation

$$\frac{\pi(r_1 - d_i)^2}{\pi(r_2 - d_i)^2} = \frac{\mathrm{DCR}_1}{\mathrm{DCR}_2}, \tag{3.10}$$

$$d_i = \frac{r_1 - r_2\sqrt{\mathrm{DCR}_1/\mathrm{DCR}_2}}{1 - \sqrt{\mathrm{DCR}_1/\mathrm{DCR}_2}} \tag{3.11}$$

A similar technique can be used with the ECR of two diodes, instead of the DCR, since the count rate from incident light is expected to scale with the active area. ECR matching can be used with high noise diodes, so long as the count rate does not begin to saturate, but an external light source is required. The light source's power must be static, though the absolute value is unimportant, and the source should be placed far compared to the distances between diodes themselves.

Fig. 3.8 compares the inactive distance measured using the two types of count rate matching with the same diode pictured in Fig. 3.7. After [41], the

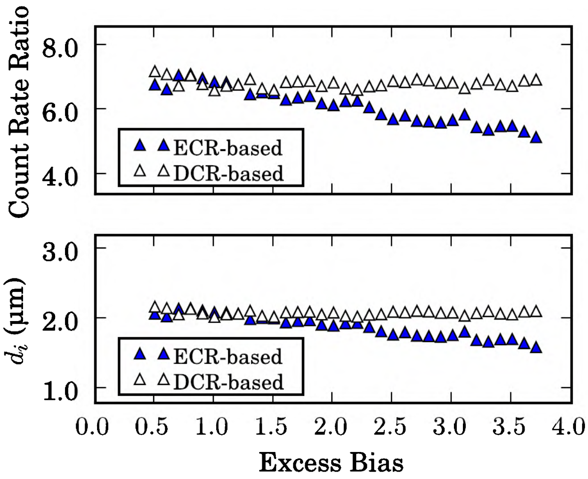


Figure 3.8: **Count Rate Density Matching**

results for the ECR-based method show a decreasing inactive distance from roughly 2.2µm to 1.8µm as V_{eb} increases from 0.5V to 3.5V. However, the DCR-based estimation remains relatively constant at 2µm. It is not clear what a constant d_i implies, though if the effects of decreasing d_i with V_{op} balance out the effect of an increasing d_i depletion width, a constant d_i is possible. The emission test was unable to be run at higher excess biases due to current density limitations of the metal routing.

3.6 Photon Detection Probability

The most common method of measuring the photon detection probability is to create an area with uniform flux of photons of a particular wavelength, and compare the responsivity of the SPAD under test to a reference diode's responsivity. A wide-spectrum light source passed to an integrating sphere through a monochromator will produce a uniform flux of wavelength-specific photons at the output ports of the integrating sphere. The output light will be diffuse, though the mean penetration depth does not need to be modified, due to the high real refractive index of silicon (>3.5 for visible light). The ECR, which is the count rate above the DCR, divided by the absolute photon flux over the diode's active area, in photons per second, gives the PDP. The ECR must be compensated for afterpulsing, and the active area must be compensated for the inactive distance, to measure the PDP.

$$\mathrm{PDP}(\lambda) = \frac{\mathrm{ECR}(\lambda)}{A_{\mathrm{SPAD}} \cdot \mathrm{Flux}(\lambda)} \left(c_{\mathrm{corr}}\right), \tag{3.12}$$

where A_{SPAD} is the SPAD's active area and c_{corr} is a correction factor for afterpulsing and crosstalk. The correction factor needs to account for the probability that afterpulsing will cause crosstalk and crosstalk will cause afterpulsing. The probability P_c that one or more correlated noise avalanches will occur is $P_c = 1-(1-P_{ap})(1-P_{ct})$, where P_{ct} is the probability of crosstalk as previously defined in Sec. 2.4.4. The correction factor will be the fraction of events seeded by non-correlated noise, which is

$$c_{\mathrm{corr}} = \frac{1}{1 + 1/(1 - P_c)} = 1 - P_c. \tag{3.13}$$

In practice it is better to set a high dead time and use a SPAD or SPADs free of crosstalk effects than to worry about these factors being measured correctly or changing with different measurement conditions. The photon flux will be

$$\mathrm{Flux}(\lambda) = \frac{I(\lambda) - I_{\mathrm{dark}}}{r(\lambda) \cdot A_{\mathrm{ref}} \cdot (hc/\lambda)}, \tag{3.14}$$

where $I(\lambda)$ is the reference photodiode's current, I_{dark} is the reference diode's dark current, $r(\lambda)$ is the reference diode's responsivity (in amperes per watt per unit area), A_{ref} is the reference diode's active area, h is Planck's constant, and c is the speed of light.

3.7 Jitter

The timing uncertainty in SPADs is normally measured using a source that produces light correlated with an electrical signal, such as a laser. The time uncertainty between the output light and the correlated electrical signal ranges from nanoseconds to femtoseconds, though visible, solid-state lasers tend to have values in the picosecond to hundreds of picoseconds range. Measuring the histogram of time differences between the electrical signal and the SPAD output yields the system time uncertainty, and removing the jitter generated by the non-SPAD portions of the measurement setup yields the SPAD's jitter. The laser jitter must be attenuated such that an avalanche is unlikely per pulse, otherwise distortions from the number of photons incident on the diode will cause problems, as Ch. 4 presents.

Finding the non-SPAD portions of the jitter has the inherent difficulty of trying to measure something without changing it. If any signals must be routed from or to off-chip, such as from the laser or to an off-chip TAC or

TDC, the routing jitter may be larger than the SPAD jitter. For example, there might be problems matching the impedance of the laser's electrical output to the TDC trigger. In these cases, the time difference between two SPADs' avalanches, rather than the avalanche vs. laser time difference can be used to aid in reducing these jitters[29]. The lowest reported jitter, which uses an optimized setup with a cooled SPAD, is 20ps, with low-threshold comparators or current pick-up methods can achieve sup-50ps jitter at room temperature[86, 65]. Care must be taken when comparing reported RMS jitter to FWHM jitter; the two are not the same.

Chapter 4

Multi-Photon Distortions

This chapter discusses three types of distortions when a SPAD[1] is triggered by multiple photons rather than single photons: distortions in the event rate, the quench time[87, 88], and the jitter[67]. A single system capable of measuring all these distortions at the same time is presented. These distortions are important for two reasons. First, correct characterization of SPADs is important, and clues in characterization data that show whether or not these distortions are occurring is relevant to any device-level characterization. Second, hacks to quantum key distribution (QKD) systems have relied on forcefully triggering the SPADs[89]. Any shifts in SPAD behavior that result from forced triggering can detect classes of hacks to QKD systems that rely on forced triggering of devices. While previous work has focused on capturing information about the intensity of the incoming light burst to a SPAD [90], this chapter shows that multi-photon measurement in a single diode are also possible.

4.1 Theory of Distortions

As Ch. 2 describes, in large diodes the avalanche propagates outward from the initial electron-hole pair's injection point. If multiple photons are simultaneously incident on the diode itself, the avalanche propagation will occur from more than one location, and the avalanche is expected to quench more quickly. Fig. 4.1 shows the basic theory behind this idea. Complete modeling

[1] When multiple photons are incident on the SPAD, technically it is no longer a single-photon avalanche diode, but is rather a photon-number-resolving, Geiger-mode avalanche photodiode. However, for reasons of consistency with the rest of the text, the diode will still be referred to as a SPAD in this chapter.

of this phenomena is beyond the scope of this text; only experimental results will be presented.

There are other distortions that will appear in the SPAD's characterization when multiple photons simultaneously seed the avalanche. As Sec. 2.4.3 describes, the timing response of a SPAD can be thought of as the convolution of a normal curve with the sum of a delta function and a single exponential function. The delta function represents carriers generated in the depletion region, whereas the exponential represents carriers that must diffuse to the depletion region. If multiple carriers are simultaneously incident on the diode, electron-hole pairs generated in the depletion region may mask an avalanche that would have been caused by a carrier that needed to diffuse. This decreases the magnitude of the exponential tail compared to the magnitude of the delta function. Fig. 4.2 shows how this can occur.

The effect is not simply limited to the diffusion tail of the timing curve; as more electron-hole pairs are generated, the Gaussian component of the timing jitter will also begin to be distorted. Because simulating multiple trigger locations is beyond the scope of this thesis, the analysis here will focus on the reduction in the diffusion tail.

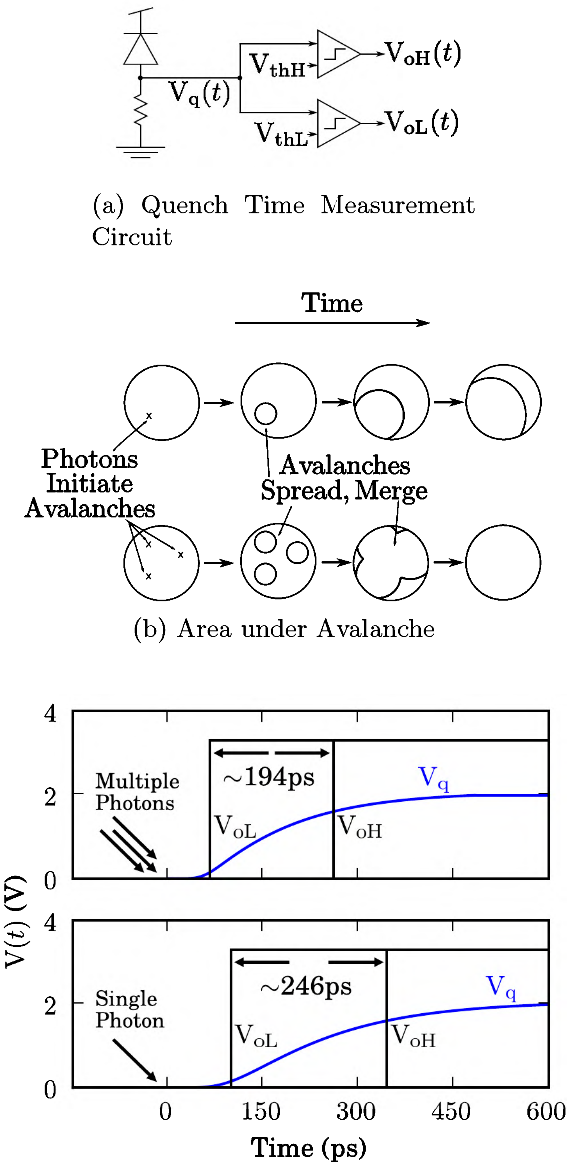
Let n be the number of simultaneously incident photons on a SPAD. When $n = 1$, there are four possibilities for a photon-generated electron-hole pair: the pair is in the p+ region, termed E_p; the pair is in the depletion region, event E_d; the pair is in the n-well region, termed E_n; and the photon doesn't generate a pair in these three region, event E_o. The four are mutually exclusive, but one must occur, so $P[E_p] + P[E_d] + P[E_n] + P[E_o] = 1$.

The probability of an avalanche for a photon of wavelength λ can be written as $\sum_{x\in\{p,d,n,o\}} (P[A|E_x]P[E_x|\lambda])$, with $P[A|E_x]$ denoting the probability of an avalanche given that the carrier is generated in the regions described above. The timing jitter, $f_\mathrm{j}(t)$ will be

$$f_\mathrm{j}(t) \propto \sum_{x\in\{p,d,n,o\}} \left(f_{\mathrm{j}|E_x}(t)P[A|E_x]P[E_x|\lambda]\right), \tag{4.1}$$

with $f_{\mathrm{j}|E_x}(t)$ being the timing jitter when the carrier is generated in region E_x.

For condition E_d, the timing jitter will be completely dependent on the ionization noise. Due to the central limit theorem, a normal distribution results; $f_{\mathrm{j}|E_d}(t)$ is $\sim \mathcal{N}[\mu,\, \sigma^2](t)$ with mean μ and variance σ^2. In condition E_o, no avalanche can occur, and this condition can be ignored. Conditions E_p and E_n may generate an avalanche, but the carrier must first diffuse to the depletion region. The diffusion time is approximately exponential, so the timing jitter will be $f_{\mathrm{j}|E_n}(t) = f_e(t,\tau) * \mathcal{N}[\mu,\, \sigma^2](t)$, with $f_e(t,\tau)$ being a



(a) Quench Time Measurement Circuit

(b) Area under Avalanche

(c) Quench Waveforms

Figure 4.1: **Quench Waveforms for Single vs. Multiple Photons** — (a) shows a circuit capable of measuring the quench time. As (b) shows, when multiple photons are simultaneously incident on the diode, the avalanching area is larger due to the multiple seed locations. Measuring this difference, as (c) displays, allows a probabilistic estimate of whether or not multiple photons triggered the avalanche.

single exponential with time constant τ. Additionally, if the wavelength value is restricted to a particular value, 637nm in this case, the fraction of photons generated for condition E_p can be ignored due to the long penetration depth into the silicon.

Thus, the total timing jitter at $\lambda = 637$nm will be

$$f_{\mathrm{j}}(t|n=1) \propto \mathcal{N}[\mu,\, \sigma^2_{\text{laser}}](t) * \mathcal{N}[\mu,\, \sigma^2](t) * (\delta(t)P[A|E_d]P[E_d|\lambda] + \\ f_e(t,\tau)P[A|E_n]P[E_n|\lambda]). \quad (4.2)$$

If the timing jitter is split into sub-components $\alpha_d(t, n=1)$ and $\alpha_n(t, n=1)$ representing the time jitter from depletion- and well-generated carriers to reach the depletion region, the equation becomes

$$f_{\mathrm{j}}(t|n=1) \propto \mathcal{N}[\mu,\, \sigma^2](t) * [\alpha_d(t, n=1) + \alpha_n(t, n=1)]\,, \quad (4.3)$$

with $*$ denoting convolution. One important thing to note in this equation is that the ratio of the exponential tail in the timing jitter to the $\delta(t)$ function will be $\frac{P[A|E_n]P[E_n|\lambda]}{P[A|E_d]P[E_d|\lambda]}$. $P[E_x|\lambda]$ is easy to estimate, given knowledge of the photon's mean penetration depth. $P[A|E_x]$ will follow from the CDP curve. Thus, the weight of the exponential tail compared to the normal component will allow confirmation of some components in the PDP model.

Assume that the first photon incident on the depletion region will dominate the timing jitter. If two or more photons are incident on the SPAD, then the weights of the component jitters will change. Because any seeding carriers in the depletion region would immediately trigger an avalanche, the weight for the E_d condition changes from a single carrier possibly generating an avalanche in this region to the probability that any region-seeded carrier generates an avalanche. With n_d pairs generated in the depletion region, the coefficient of $\delta(t)$ in $f_{\mathrm{j}|C_d}(t)$ will change from $P[A|E_d]P[E_d|\lambda]$ to

$$\alpha_d(n) = \sum_{n_d=0}^{n} (1 - (1 - P[A|E_d])^{n_d}) f_{N_d}[n_d|n] \quad (4.4)$$

with a binomial distribution governing the probability that n_d carriers (out of the n incident photons) are generated in the region,

$$f_{N_d}[n_d|n] = \binom{n}{n_i} P[E_d|\lambda]^{n_i}(1 - P[E_d|\lambda])^{n-n_i}. \quad (4.5)$$

Because any avalanche generated by carriers in the depletion region are assumed to mask avalanches generated by carriers diffusing from the well, there must be zero avalanche-generating carriers in the depletion region if an avalanche is to be triggered by a well-sourced carrier. Additionally, the timing distribution from well-sourced carriers will no longer be a single exponential with time constant τ, but will rather be the first carrier to reach the depletion region and cause an avalanche. Normally, if n photons are incident, a binomial distribution with parameter $P[E_n|\lambda]$ governs the number of carriers that would be expected in the well, n_i below. However, because the carriers generated in the depletion region have been captured by the n_d variable, the rate parameter is actually $\frac{P[E_n|\lambda]}{1-P[E_d|\lambda]}$, and n_i can range from 0 to $n - n_d$.

$$f_{N_i}[n_i|n-n_d] = \binom{n-n_d}{n_i}\left(\frac{P[E_n|\lambda]}{1-P[E_d|\lambda]}\right)^{n_i}\left(1-\left(\frac{P[E_n|\lambda]}{1-P[E_d|\lambda]}\right)\right)^{n-n_d-n_i} \tag{4.6}$$

Another binomial distribution with parameter $P[A|E_n]$ governs the number of well-sourced carriers, n_n below, that will trigger an avalanche,

$$f_{N_n}[n_n|n_i] = \binom{n_i}{n_n} P[A|E_n]^{n_n}(1-P[A|E_n])^{n_i-n_n}, \tag{4.7}$$

From the theory of order statistics, the timing distribution of the first carrier reaching the depletion region will still be exponential[1], but will now have a time constant τ/n_n. Thus, with n incident photons, the weighted time for the E_n condition will change from $f_e(t,\tau)P[A|E_n]P[E_n|\lambda]$ to

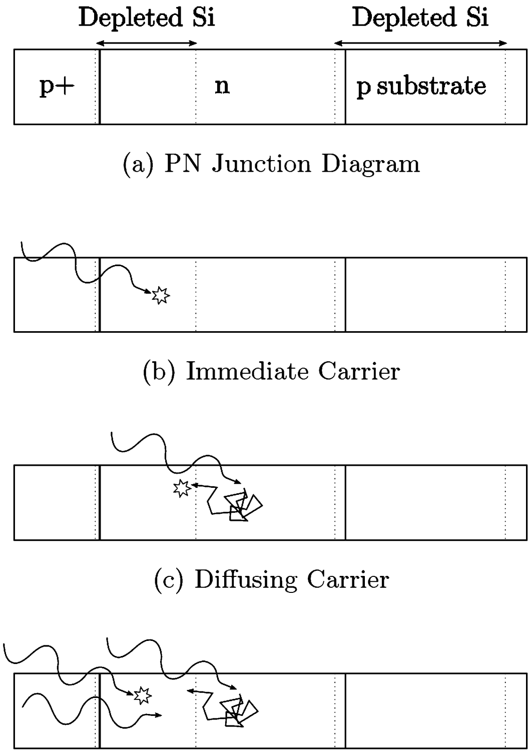
$$\begin{aligned}\alpha_n(t,n) = \sum_{n_d=0}^{n}\Bigg((1-P[C_d|\lambda])^{n_d} f_{N_d}[n_d|n]\cdot\sum_{n_i=1}^{n-n_d}\Bigg(f_{N_i}[n_i|n-n_d]\cdot \\ \sum_{n_n=1}^{n_i}\Big(f_{N_n}[n_n|n_i]f_e(t,\tau/n_n)\Big)\Bigg)\Bigg)\end{aligned} \tag{4.8}$$

Ignoring distortions to the Gaussian component from multiple seed locations, the total timing jitter with n incident carriers will be

$$f_{\mathrm{j}}(t|n) \propto \mathcal{N}[\mu,\, \sigma^2_{\mathrm{laser}}](t) * \mathcal{N}[\mu,\, \sigma^2](t) * (\alpha_d(t,n)+\alpha_n(t,n))\,. \tag{4.9}$$

If multiple seed locations are included, then the avalanche is expected to occur more quickly. Thus eq. (4.9) will be an upper-bound on the timing distribution for $n > 1$. Technically, (4.9) should be a sum of weighted Poisson components, due to the shot noise in the number of generated carriers, but to keep the task computationally feasible for large n, only the case of n incident photons will be considered.

[1] Ch. 8 presents the Rènyi representation, from which this fact is easily derived.



(a) PN Junction Diagram

(b) Immediate Carrier

(c) Diffusing Carrier

(d) Immediate Carrier Masks Diffusing Carrier

Figure 4.2: **Processes Contributing to Jitter**

Fig. 4.3 shows the expected distortion to the diffusion tail for the diode that will be measured. There are four things of note on the graph. First, the distortions to the normal component are not modeled. Second, the fall-off in the tail should be visible when several photons are incident on the diode itself. Third, the time constant in the exponential tail does not vary greatly before the tail is expected to disappear below the noise floor. Even though the tail is the sum of multiple exponentials with different time constants, the single carrier case dominates while the tail can be observed. Finally, the noise floor is not shown on the plot; however, since the signal will increase as the number of photons is increased, the SNR should increase with increasing power.

There is one other distortion that multiple photons will cause. If the diode's applied voltage can be varied and the flux of incident photons does not vary, the sampled count rate can be used to detect the expected number of simultaneously incident carriers. Specifically, if there are n simultaneously incident photons, the ratio between the count rates at bias $\mathrm{V}_{\mathrm{eb1}}$ compared to

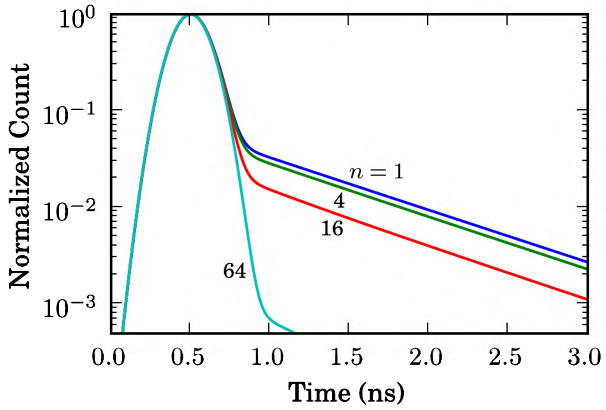


Figure 4.3: **Simulated Multi-Photon Distortions to the Diffusion Tail** — shown is the expected distortion to the exponential tail in the timing jitter curve for the SPAD presented in Ch. 2. The n variable depicts the number of photons incident on the diode.

$\mathrm{V_{eb2}}$ will be

$$r = \frac{1 - (1 - \mathrm{PDP}(\lambda, \mathrm{V_{eb1}}))^n}{1 - (1 - \mathrm{PDP}(\lambda, \mathrm{V_{eb2}}))^n}. \tag{4.10}$$

If the PDP is known as a function of the wavelength and the excess bias, the bias can be modulated between values $\mathrm{V_{eb1}}$ and $\mathrm{V_{eb2}}$, with the ratio providing a numerical estimate of n. Note that in an actual measurement, n will be subject to shot noise, and r will actually be the sum of values when $n \geq 1$, weighted by a Poisson distribution with mean and variance $E[n]$. To simplify the analysis, the closest non-negative number to $E[n]$ will be used in place of n in (4.10).

If the SPAD is operating in a passive quenching scheme with a dead time close to the laser clock frequency, distortions in the pulse duration will cause the measured event rate to appear lower than expected. Assume, for a moment, that the afterpulsing probability density when the SPAD is at its full excess bias as a function of time is known to be $p_{ap}(t)$, and the probability density of a dark count is known to be $p_d(t)$. If the carrier detection probability, CDP, as a function of the excess bias is assumed to follow the relation[91]

$$\mathrm{CDP}(\mathrm{V_{eb}}) = 1 - \exp(\mathrm{V_{eb}}/\mathrm{V_o}), \tag{4.11}$$

and the excess bias is restored by a weakly inverted transistor acting as a current source, causing the excess bias following an avalanche at $t = 0$ to be roughly

$$\mathrm{V_{eb}}(t) = \begin{cases} (\mathrm{V_{op}} - \mathrm{V_{bd}}) \cdot \frac{t}{d_i} & t \leq t_d, \\ 0 & t > t_d, \end{cases} \tag{4.12}$$

then the probability $p_{\text{longer}}(\mathrm{V_{eb}})$ that the dead time will last longer than the nominal value t_d will be

$$p_{\text{longer}}(\mathrm{V_{eb}}) = \int_0^{t_l} (\mathrm{CDP}(\mathrm{V_{eb}}(t)) \cdot (p_{ap}(t) + p_d(t)))\, dt, \tag{4.13}$$

with t_l being the laser clock frequency. When the dead time is near the laser period t_l, a first order estimate of the lost count rate given event rate e is $c_{\text{corr}}(\mathrm{V_{eb}}) = 1 - e/t_l \cdot p_{\text{longer}}(\mathrm{V_{eb}})$. Each of the count rates in the ratio above needs to be tempered with this factor, yielding

$$r = \frac{(1 - (1 - \mathrm{PDP}(\lambda, \mathrm{V_{eb1}}))^n)(c_{\text{corr}}(\mathrm{V_{eb1}}))}{(1 - (1 - \mathrm{PDP}(\lambda, \mathrm{V_{eb2}}))^n)(c_{\text{corr}}(\mathrm{V_{eb2}}))}. \tag{4.14}$$

4.2 Measurement Setup

Fig. 4.4 depicts a setup capable of measuring all three types of distortions. The previously characterized circular SPAD is coupled to a 20ps, Vernier-delay-line based TDC via two comparators, one with a threshold voltage of 0.1V and the other with a threshold voltage of 2.0V. The SPAD is the same one presented in Ch.3 and its active area, following compensation for inactive distance effects, is 20µm. The output from the comparator with the lower threshold is also run directly to the FPGA, allowing sampling of the count rate, and to an off-chip 61ps TDC whose stop signal is input from a laser. Both TDCs are coupled to the FPGA with serial peripheral interface (SPI) buses. A laser beam (photons of wavelength 637nm) with fixed position and power shines on the SPAD after being routed through one or more neutral filters with optical density between 0 and 4.5. In other words, the power transmitted through the filters ranges from $10^{-4.5} = 0.00003$ to $10^{-0} = 1$. Additionally, this laser has an electrical jitter of 230ps when run at a count rate of 2.4MHz, though the optical jitter is roughly 40ps. That is, the photons in the beam occur temporally with a FWHM of roughly 40ps, though the electrical output compared to the mean temporal time of any pulse will have a FWHM of roughly 230ps. The number of SPAD-incident photons expected to trigger an avalanche, $E[p_a]$, can be extrapolated from the count rate measured with higher optical densities. For example, if the ECR is 80kHz at an optical density of 4, and the laser frequency is 4MHz, then $E[p_a]= \frac{\text{80kHz}}{\text{4MHz}} = 0.02$. If the optical density is changed to 0 in this example,

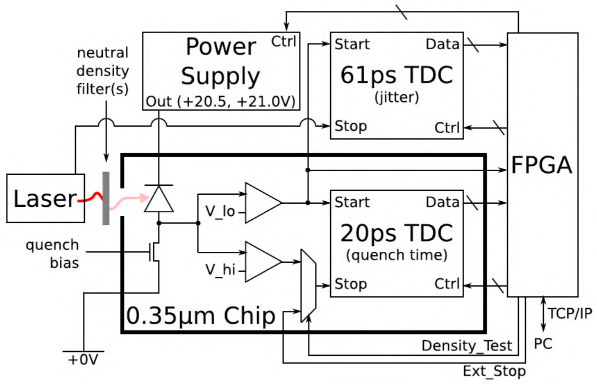


Figure 4.4: **Measurement Setup for Multi-Photon Distortions** — two TDCs are necessary in the setup, one to measure the jitter with respect to the laser, and one to measure the quench time.

$E[p_a]$ will change to $0.02 \cdot 10^4 = 200$. The number of incident photons can be derived by dividing $E[p_a]$ by the PDP at 637nm and the excess bias used. In this particular case, the PDP will be assumed to be roughly linear with the excess bias[91], being $\sim 16\%$ when $\mathrm{V_{eb}} = 2.0\mathrm{V}$ and $\sim 20\%$ at 2.5V. Thus $E[n] \approx 4E[p_a]$.

Due to hardware limitations, the laser could not be run at a frequency lower than 2.4 MHz, requiring the diode to have a dead time of 400ns. Even at this dead time, the use of passive quenching in such a large diode will imply distortions in the dead time. For high incident photon count rates, nearly 10% of all pulses will be missed because the SPAD will not have recharged, since another avalanche will occur before the sensed voltage returns below the low threshold of 0.1V. This will cause the aforementioned distortions in the ratio of the excess count rate that would be mitigated in an active quenching scheme.

In addition to being connected to the two TDCs, the readout FPGA is also connected to a computer workstation and a controllable $\mathrm{V_{op}}$ supply via TCP/IP. In the present setup, the power supply outputs either 20.5V (2.0V excess bias) or 21.0V (2.5V excess bias) at the command of the FPGA. The FPGA flips the output voltage to the other value once every second. When the output voltage is 20.5V, the FPGA also streams the values from the two TDCs to a computer workstation. The values are stored and then histogrammed and analyzed offline.

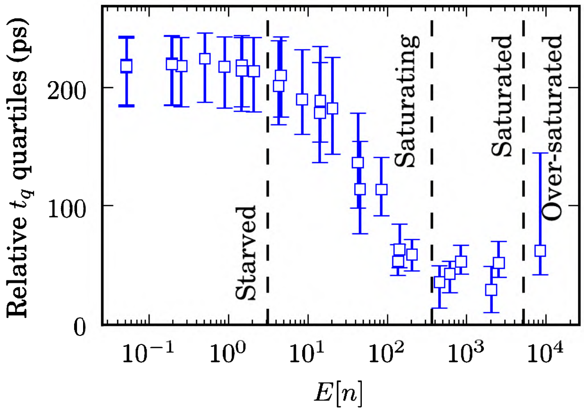


Figure 4.5: **Quench Time vs. Simultaneously Incident Photons**

4.3 Quench Time Distortions

Fig. 4.5 shows the measured quench time t_q as a function of the incident photon number. As would be expected, the quench time begins to decrease as the expected number of photons exceeds one, though the decrease is not noticeable until 10 or more photons are simultaneously incident on the diode.

A curious distortion appears when the expected number of photons exceeds 10,000. The rise time begins to increase again, though the variance appears to be much larger. It isn't clear what is occurring at this stage, but the optical power incident on the chip is quite large, with more than 30 photons incident per laser pulse per square micron. At this amount of incident optical power, it is possible that IR drop due to carriers generated in the substrate or the comparator wells, for example, begins to cause distortions. Due to concerns about damaging the chip, this test condition was not repeated. Future work should focus light solely on the diode, using an optical microscope for example, to avoid these possible complications.

4.4 Event Rate Distortions

Fig. 4.6 shows the measured distortions to the event rate ratio as a function of n. Also shown in the expected value from (4.10), with the closest non-negative value of $E[n]$ being used in place of n. The expected curve would be smooth if the shot noise variations were to be included.

If the diode had no dead time distortions, the event rate ratio would start

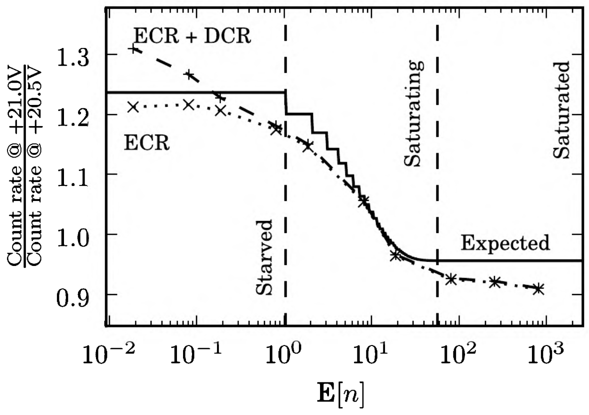


Figure 4.6: **ECR Ratios with Multi-Photon Distortions** — the n variable on the x-axis is the expected number of photons incident on the SPAD.

at the ratio of the two PDPs, simulated to be 1.25 in this case, and then decrease to a value of 1. Due to the larger afterpulsing probability at the higher excess bias, the ratio approaches a value below one. The mismatch between experimental and simulated values of $n << 1$ is due to PDP saturation — i.e. the PDP does not scale linearly with the excess bias, even at excess biases of 2V.

4.5 Timing Distortions

Fig. 4.7 shows the jitter curves for different numbers of simultaneously incident photons, along with the FW(1/N)M and selected left- and right-widths. As predicted by the theory, the diffusion tail begins to disappear as more photons are simultaneously incident on the diode. Also, as predicted, SNR decreases as more photons are simultaneously incident on the diode. The diffusion tail is visible compared to the noise floor with a decreased magnitude when $n \approx 20$, but is not observable when $n = 80$. Also of note is that the normal component to the curve shows little to no distortion when $n \leq 20$, but appears to have a smaller FWHM at $n > 80$.

The disappearance of the tail can be quantitatively observed by plotting the RW(1/100)M as a function of the expected number of simultaneously incident photons. The RW(1/100)M has a relatively high constant value of 2ns until $E[n] > 2$, at which point it begins to decrease until it is less than 200ps, several times the value of the FWHM. The LW(1/100)M is also seen to increase above the inaccuracy level defined by the TDC's LSB duration when $E[n] > 1,000$. This effect is likely due to distortions to the normal component of the jitter from multiple avalanches.

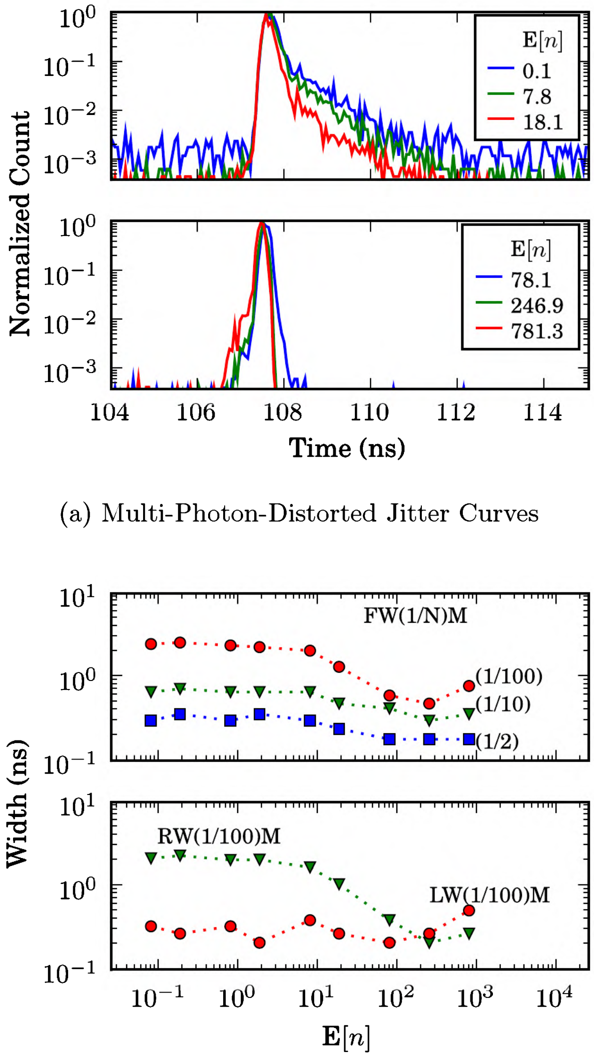
Fig. 4.8 compares the predicted decrease in the tail's magnitude to the simulated value based on the theory presented above. There is a good qualitative match between the theory and the expected result. The diffusion tail's time constant remains relatively steady until the tail cannot be observed, as predicted by the theory. However, the theory predicts a slightly larger initial magnitude in the tail. Due to the $(1 - P[C_d|\lambda])^n$ terms in (4.8) and (4.4), even small shifts in $P[C_d|\lambda]$ can create order of magnitude differences in the relative weights when $n >> 1$, so it is not surprising that the goodness of the match will change as n changes. The final subplot of Fig. 4.8 also shows that the plot does not follow a Gaussian distribution for extremely large $E[n]$, ostensibly due to multi-avalanche spreading.

4.6 Discussion

While each presented method of detecting multi-photon distortions has several strengths and weaknesses, the efficiency modulation methods shows the greatest promise for determining whether a detector is operating in a photon-starved regime or not. The diffusion tail method requires knowledge of the incident photon stream's timing jitter *a priori*; shifts in the incident photon stream's generation requires complex timing measurements. Especially for QKD, where an attacker may have precise control over the incident photon's timing, this method is not likely to be useful. Additionally, this method is likely to be sensitive to shifts in the operating conditions, such as the excess bias or temperature.

While the quench time method allows for a single-shot estimate of the incident photon number, an advantage that neither of the other two methods share, this method may also be susceptible to shifts in the operating condition of the avalanche photodiode. This method has an additional shortcoming for QKD, which is that an attacker could stream single photons separated by the quench time of the photon-starved diode. In this mode of illumination, the diode would always measure a single photon incident on its active area; however, the diode would still be triggered with high probability.

Of the three methods, the efficiency modulation method shows the most promise for use in an actual system. The method is better resistant to shifts in operating point than the other two methods, and does not require a TDC. However, this method requires a shift in diode operating point, with the associated cost of more complex voltage supplies and operating the diode for



(a) Multi-Photon-Distorted Jitter Curves

(b) Jitter Curves vs. Simultaneously Incident Photons

Figure 4.7: **Jitter vs. Simultaneously Incident Photons** — n is the expected number of simultaneously incident photons.

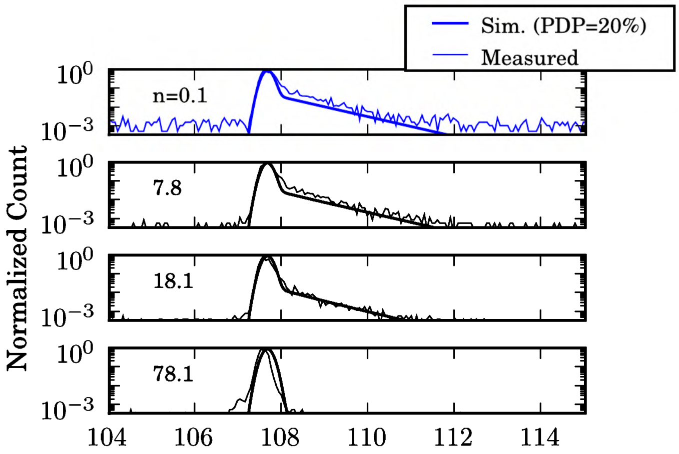


Figure 4.8: **Simulated vs. Measured Values of the Diffusion Tail for Simultaneously Incident Photons**

a period of time at a lower detection efficiency than the optimal efficiency. Additionally, a minimum integration time is required for this method, though there will be a trade-off between the integration time and the uncertainty in whether the avalanches are being caused by a single photon or multiple photons.

The three methods, which are independent of one another and use the same setup, have the potential to be used at the same time to cover one each other's weaknesses. For example, utilizing both the diffusion tail and quench time methods would eliminate the mentioned avenues of attack against QKD systems using such methods, since one attack relies on an imprecise triggering method using precisely timed single photons while the other attack relies on precisely timing a large number of photons. In addition, the methods show compatibility with methods relying on observing correlations in avalanches across multiple diodes[90]. If such methods were to be used in a QKD system, a more secure system is possible.

Chapter 5

Hostile Environments

This chapter presents data from two SPAD-based integrated circuits operating in hostile environments. The first IC, the RADHARD2 chip, was developed by Lucio Carrara in collaboration with the European Space Agency as an oxygen airglow sensor for backup navigation on satellites[73]. The second IC, the MOSAIC chip, was custom developed for the express purpose of studying SPAD device physics using electrical measurements. Following a brief discussion of the experimental setups employed, results will be shown for various SPAD characteristics when the diode is in a magnetic field, and noise levels in γ-ray flooded environments. This chapter is based on results published in [41], [92] and [93].

5.1 Magnetic Fields

5.1.1 Simulations and Expected Results

As Ch. 2 describes, in planar diodes the avalanche process is known to spread via multiplication-assisted diffusion. During multiplication, carriers travel at the saturation speed, negating a critical assumption used in the derivation of the Hall coefficient — a carrier's velocity varies in proportion to the electric field.

Thus, in an avalanche diode operating in a strong magnetic field, the analysis that leads to the Hall coefficient cannot be applied. Instead, the force from the magnetic field will be governed by the Lorentz force[94],

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}). \tag{5.1}$$

When traveling under high force at an average speed that is the saturation speed[20], the magnitude of the force from magnetic field will be fixed at roughly $|\vec{F}| = |\vec{V}||\vec{B}|$ so long as the direction of the carrier's propagation remains roughly orthogonal to the magnetic field, an assumption which will be checked later. In a 9.4T field, the magnitude of the magnetic field's force will be about $0.15 \cdot 10^{-12}$N, or 0.15pN. In an abrupt one-sided junction with a V$_{\mathrm{bd}}$ of roughly 20V, the peak electric field is $\sim 5 \cdot 10^5$ V/cm, implying the average electric field in the multiplication region, the region with significant ionization, is larger than $4 \cdot 10^5$ V/cm[20]. The magnitude of the force on the particle from the electric field is $q\vec{E} \approx 6.4 \cdot 10^{-12}$N, or 6.4pN.

The ratio of the magnitude of the force from the electric field to that from the magnetic field is 6.4pN:0.15pN$\sim$40:1. At the saturation speed, the carrier will travel in the direction set by the sum of these two components. So long as the electric field is oriented orthogonally to the magnetic field, the 40:1 ratio validates the assumption that the carrier's direction remains orthogonal to the magnetic field. If the electric field, as per previous convention, is oriented in the $\hat{z}$ direction, with the force from the magnetic field being in the $\hat{x}$ direction (the magnetic field itself would be in the $\hat{y}$ direction), then the velocity vector will be

$$\begin{aligned}
\vec{v} &= v_z\hat{z} + v_y\hat{y} + v_x\hat{x}, && (5.2)\\
&= \frac{|\vec{F}_E|}{\sqrt{|\vec{F}_E|^2 + |\vec{F}_B|^2}}|\vec{v}|\hat{z} + 0\hat{y} + \frac{|\vec{F}_B|}{\sqrt{|\vec{F}_E|^2 + |\vec{F}_B|^2}}|\vec{v}|\hat{x}, && (5.3)\\
&\approx 0.975|\vec{v}|\hat{z} + 0.025|\vec{v}|\hat{x}, && (5.4)
\end{aligned}$$

with planar component $v_x \approx 0.025 \cdot 10^5$m/s ≈ 2.5μm/ns. Given that the total force acting on any free carriers changes by a factor of $\sqrt{|\vec{F}_E|^2 + |\vec{F}_B|^2}/|\vec{F}_E| \approx 1.0003$, or a 0.03% change, no significant shift should be expected in the breakdown voltage. Because neither of the noise mechanisms is directly related to the presence of a magnetic field, and also due to the negligible shift in the ionization rate, there should also be no sizable shift in the DCR.

The planar component will cause convection, and hence when modeling the avalanche propagation, the governing equation from (2.14) must be modified to include the term from convection, $\vec{v_c} \cdot \nabla c$,

$$\frac{\partial c}{\partial t} = D_{eff}\nabla^2 c - \vec{v_c} \cdot \nabla c + \frac{c}{\tau}, \tag{5.5}$$

with $\vec{v_c}$ being the convection velocity[95]. If τ and $\vec{v_c}$ are taken to be constants and the boundary condition for c is that it is a delta function at the origin at time 0, then the substitutions[96], $u(\vec{r}, t) = \exp(\gamma \cdot t + \vec{\lambda} \cdot \vec{r})c(\vec{r}, t)$, $\vec{\lambda} = \vec{v}/2D$, and $\gamma = 1/\tau - |\vec{v}|^2/4D$ can be used to derive the analytical solution,

$$c(\vec{r}, t) = \frac{\exp{(t/\tau)}}{4Dt} \exp\left(\frac{-|\vec{r} - t\vec{v_c}|^2}{4Dt}\right). \tag{5.6}$$

The solution with inclusion of the convection term appears identical to the previously derived solution in the convection-free case, except that $\vec{r}$ is replaced with $\vec{r} - t\vec{v_c}$. Appendix B shows the exact mathematics of this derivation in detail. When $t >> 0$, the avalanche propagates with velocity $\vec{v_c} + 2\sqrt{D/\tau}$, reducing to the well-known propagation velocity [97] of $2\sqrt{D/\tau}$ when the convection term is negligible.

The same simulation method used in Ch. 2's Fig. 2.12 can be used with (5.5) replacing (2.14) to simulate avalanche propagation in a magnetic field. However, Ch. 2.12 uses a static breakdown voltage for the entire diode; as described in Ch. 3, the breakdown voltage actually varies as a function of position.

The dominant component in the spatially varying breakdown voltage is the compensation doping from the well implants. The uncompensated doping can be estimated from the breakdown voltage of diodes with sizable active regions that are "far" from the guard ring. The word "far" should be taken to mean large when compared to the inactive distance. On the chip that is to be modeled, breakdown voltages were acquired from circular structures with diameters 6µm, 12µm, and 24µm, along with a pill-shaped, 6µm by 24µm structure capped by two semi-circles of diameter 6µm (Ch. 7's Fig. 7.6 summarizes the $\mathrm{V_{bd}}$ variations). The pill-shaped and 6µm diameter diodes show marked increases in $\mathrm{V_{bd}}$, with a larger increase in the 6µm diameter structure. The ratio of the $\mathrm{V_{bd}}$ increase between the two structures, roughly a factor of 4.2 between the two structures, will be important when estimating the characteristics of the diffusion of the well's implants. Following implantation in a CMOS process, dopants diffuse during annealing steps. For a pre-annealing impulse function with value c_0, the concentration of post-annealing dopants at a particular depth can be numerically estimated as a normal distribution

$$c_d(\vec{r}, t) = \frac{c_0}{4\pi D t} \exp\left(-\frac{|\vec{r}|^2}{4Dt}\right), \tag{5.7}$$

with c_0 being the concentration of the initial impulse function and $\vec{r}$ being the position from the origination impulse function. In the case of the two SPAD structures with compensated doping, the concentration shift from compensation can be numerically approximated by integrating (5.7) over the region of interest. Specifically, for the 6µm circular diode, the concentration shift c_s in the central portion of the diode from compensation will be estimated to be

$$c_s = \int\int_R c_d(\vec{r}, t)dxdy, \tag{5.8}$$

with the region of interest R being the set of x and y points such that 3µm$< \sqrt{x^2 + y^2} < 3 + d_g$µm, where d_g is the planar thickness of the guard ring. For the pill-shaped structures, it will be approximately the set of points with 3µm$< |x| < 3 + d_g$µm. The parameter Dt must be found for which the ratio of c_s from the two structures is approximately equal to 4.1 — in this particular case, the Dt value is ∼0.55µm^2. The concentration shift in the central portion of the 6µm diameter circular diode is roughly 2.5% of the post-annealing guard ring's doping concentration.

Given that the breakdown voltages of abrupt one-sided junctions are well-known[20], the absolute value of the doping compensation can be estimated from the shift in the breakdown voltage. For diodes with $\mathrm{V_{bd}} \approx$ 18V, a compensation shift of roughly $4 \cdot 10^{15}$ per cm^3 would cause the increase in $\mathrm{V_{bd}}$ in the small circular diodes[20], leading to a guard ring doping of $\sim 1.6 \cdot 10^{17}$ per cm^3. This value is in good agreement with the value of $1.73 \cdot 10^{17}$ per cm^3 at the junction depth measured using spreading resistance profiling analysis.

The concentration shift from annealing of the guard rings can now be estimated for each point in the diode by numerically integrating (5.8) with the breakdown voltage based on the known $\mathrm{V_{bd}}$ values for an abrupt one-side junction. Fig. 5.1 shows the numerically estimated breakdown voltage for the pill-shaped diode, with Fig. 5.2 showing the cross-sectional $\mathrm{V_{bd}}$. As would be expected from Ch. 3's inactive distance of roughly 2µm for SPADs using this structure in this process, the breakdown voltage is relatively constant in the middle micron of the diode, varying less than 300mV, but rapidly increases several volts for positions that are just one micron closer to the edge of the diode. The breakdown voltage will be limited to 26V, since there will be negligible ionization in any region with a greater $\mathrm{V_{bd}}$ given that the operating point will be below 21V.

Table 5.1 lists several other measured and estimated parameters, along with notes for how the estimated parameters are derived from the literature.

Now that all of the parameters are known, the FEM model described in Ch. 2 can be combined with (5.5) to yield the quench waveforms as a function of triggering position. Fig. 5.4 shows how the avalanche quench waveforms are expected to differ when the avalanche is seeded in the center portion of the diode compared to the edge of the diode. When comparators sample a waveform such as the one in Fig. 5.4, there will be some dependence on the slope of the input signal. Cadence Spectre[80] simulations of the comparators implied that, for every 100ps shift in the rise time of the quench waveform to the level of the high comparator, the output time difference only shifts by 85ps — the measurement should be multiplied by 100/85 to compensate for this effect. Fig. 5.4 also shows the time that the comparators are expected to measure for this waveform, along with a histogram that would be output when uniform triggering across the diode is expected. The propagation velocity is not uniform, since center-seeded avalanches will have

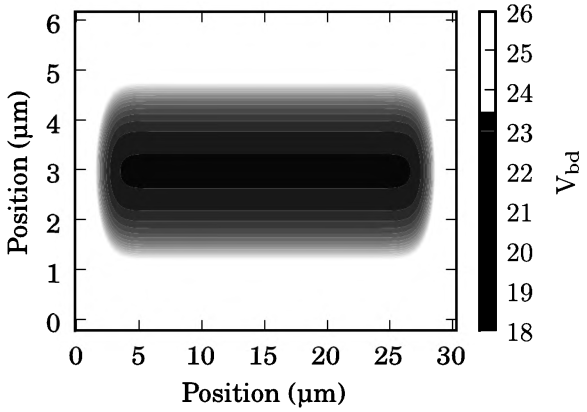


Figure 5.1: **Simulated V_{bd} vs. Position**

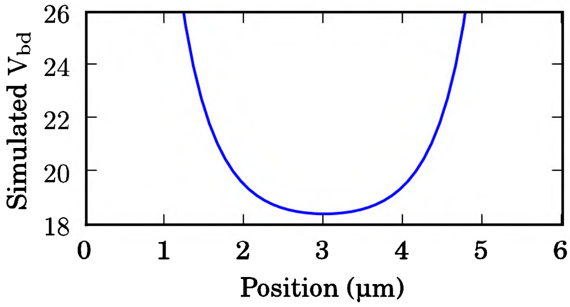


Figure 5.2: **Simulated V_{bd} vs. Position (Cross-section)**

Parameter	Value	Notes
V_{bd}	18.5-26.0V	See Fig. 5.1
V_{op}	20.9V	Free parameter
D_{eff}	$101 \cdot \exp(-.006 \cdot T)$ cm^2/s	[51, 52, 98]
$\lvert\vec{v}_s\rvert$	$\frac{1.9 \cdot 10^7}{1+.8 \cdot \exp(T/600)}$ m/s	[99, 20]
T	223.15K to 323.15K (-50°C to +50°C)	Controlled variable
Element diode size	100 by 100 nm^2	[51], Diffusion out of region during Δt is unlikely
$\vec{v}_c$	0, 2.5µm/ns $\hat{x}$, 2.5µm/ns $\hat{y}$	Based on b-fields
$c_p + c_d$	190fF	Measured
c_d	40fF	Calculated
R_o	700Ω	Measured
R_q	1MΩ	Simulated
R_{sc}	1.6MΩ	Calculated

Table 5.1: **Model Parameters** — see Fig. 2.10 for example circuit

more current flowing more quickly; the quench time changes more slowly for center-seeded avalanches than for edge seeded avalanches.

When small amounts of light are incident on the edge-open diode, there will be two peaks in the histogram; the first peak will convey the quench time of the center-seeded avalanches from the noise, and the second peak will convey the quench time of the edge-seeded avalanche. Comparing the two peaks allows a direct measurement of whether or not the avalanche propagation has changed, along with an indirect measurement of the avalanche propagation itself.

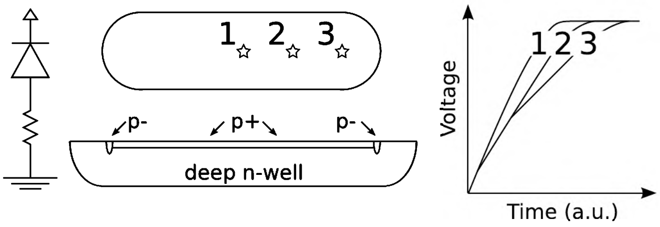


Figure 5.3: **Theory of Quench Time Shift for Differing Seed Positions** — shown are the circuit (left), device and trigger locations (center), and quench time for different seed positions (right), showing how the quench time will vary for different seed positions

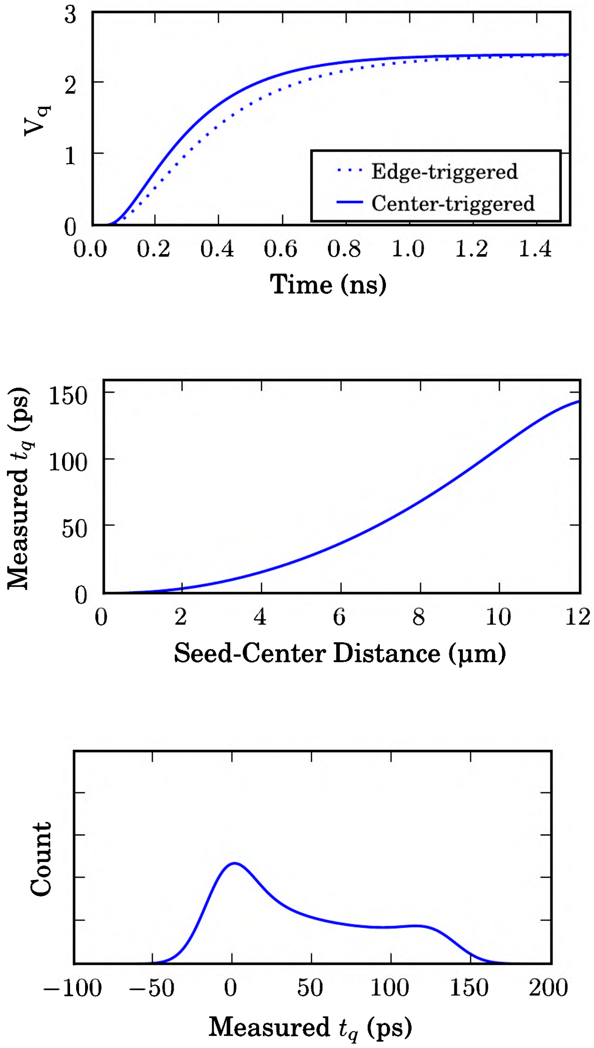


Figure 5.4: **Simulated Quench Times** — shown are the simulated quench times for center- and edge-seeded avalanches (top), the quench time as measured by the coupled comparators (middle), and the quench time histogram when seeding is uniform across the diode (bottom). The center-seeded avalanches correspond to position 1 in Fig. 5.3, while the edge-seeded avalanches correspond to 3.

5.1.2 Experimental Results

Two different setups were used to assess SPAD performance in magnetic fields up to magnitude 9.4T. In both cases an IC, assembled on a daughterboard, was placed on a runner which was inserted into a small animal MRI scanner with a static magnetic field varying between <0.1T and 9.4T as a function of the distance the chip was inserted into the scanner. The readout system, including the motherboard and computer workstation, were placed outside the field.

In the first setup, which acquired noise rates as a function of $|\vec{B}|$, the RADHARD2 chip was placed on the runner. The DCRs (not including the inactive distance) of 1,024 SPADs with a 6µm diameter were read from the chip to a computer workstation via an FPGA-based motherboard. Details of the RADHARD2 architecture, which includes a per-SPAD 1-bit memory read out in a rolling shutter readout approach, can be found in [73].

Fig. 5.5 shows the DCR quartiles of the SPADs as a function of the incident magnetic field. The DCR was integrated over a 5s period for all diodes. Less than a 2% shift was observed in the median DCR. No significant change was observed in noisy SPADs vs. quiet SPADs — SPADs with DCRs two times or more above the median remained at least two times above the median. Given that the median DCR is ~110Hz, a 5s integration period will have a count rate of ~650 counts and a std. dev. of roughly 25 counts, giving an expected std. dev. in the DCR sampling of $25/5 = 5$. The 5Hz std. dev. is about 4% of the 130Hz base count rate, implying that shifts above 5σ, which is 20% in this case, should be seen as statistically significant. No pixels were observed to have shifts above 17%.

Unfortunately it was not possible to evaluate the jitter nor the speed of avalanche propagation with the RADHARD2 chip. For this reason, an experimental setup with the MOSAIC chip was also utilized. In the second setup, the MOSAIC chip was placed on the runner. On the MOSAIC chip, multiple pill-shaped SPADs with 6x24µm^2 bodies capped by two semi-circles of diameter 6µm, which have a geometry suitable for observation of differing quench times depending on the avalanche seed position, were coupled to an 18ps, Vernier-delay-line TDC via two comparators. The pill-shaped diodes are either completely open, completely covered, covered except for a 2µm opening in the diode's middle, or covered except for a 2µm opening on the edge of the diode's major axis. Use of these coverings allows confirmation that, due to the finite propagation speed, the avalanche quench time varies as a function of the seed position. Fig. 5.6 shows this setup.

Also included on the chip are circular SPADs with a 12µm diameter. Bias signals allow compensation for TDC resolution changes from process,

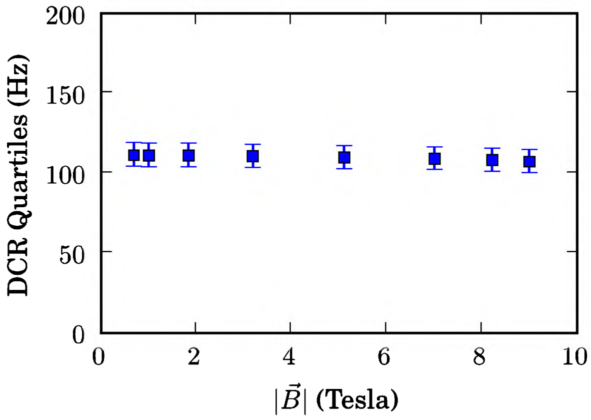


Figure 5.5: **DCR Quartiles vs.** $|\vec{B}|$ — data is from 1,024 circular diodes of diameter 6µm (without inactive distance compensations) in a magnetic field that was oriented in the planar direction.

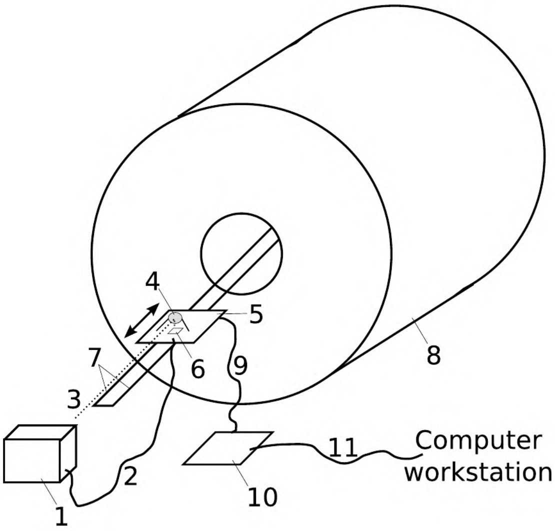


Figure 5.6: **Schmetic of Measurement Setup** — A laser (1) was electrically coupled (2) and optically coupled (3) via a reflector (4) to the custom daughterboard (5) with the ASIC (6). The daughterboard was placed on runners (7), allowing free motion into an MRI machine (8). The daughterboard is coupled via parallel cables (9) to a motherboard (10) outside the B-field, with the motherboard connected via a TCP/IP link (11) to a workstation.

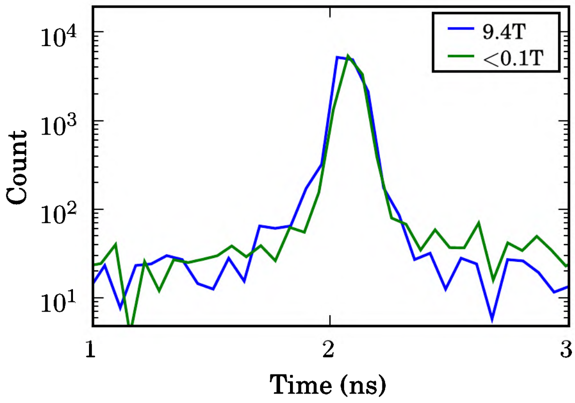


Figure 5.7: **SPAD Jitter for differing** $|\vec{B}|$ — uncompensated for TDC's INL and DNL.

voltage, and temperature variations. The CMOS chip was assembled on a daughterboard with a commercial, 60ps TDC (an Acam MessElectronic GP2). A blue laser from Advanced Laser Diode Systems GmbH, whose ouotput wavelength was 405nm with temporal FWHM 40ps, was optically coupled to the SPAD and electrically coupled to the 60ps TDC. An FPGA-based readout system, allowing acquisition of both TDC's data as well as the dark count rate of the selected SPAD, interfaced to the chip to a computer workstation.

The sweep and subtract method was used to acquire the breakdown voltage at a sub-0.1T field and at a 9.4T field — V_{bd} was found to be within 20mV of 18.6V in both cases. A density test, described in Appendix C, characterized both TDCs in and out of the magnetic fields; neither TDC exhibited a statistically significant change in behavior as a function of $|\vec{B}|$.

Fig. 5.7 shows the jitter at two magnetic field strengths measured using the 60ps TDC and the pulsed laser. The FWHM showed a small increase, from 145ps to 150ps, but the increase is not significant given the resolution of the TDC. The difference in noise floor is likely due to reflections — because the chip needed to be inserted into an MRI scanner, which was not located on an optical table in a completely dark room, it was not possible to ensure the same noise floor behavior.

Fig. 5.8 shows the quench time histograms for the center- and edge-triggered diodes for various magnetic fields, along with results from an identical setup in a temperature chamber. As previously described, the difference in the initial peak and the secondary peak in the edge-triggered histogram can be used to measure any changes in avalanche propagation. The error bars represent a 20ps uncertainty from the limiting TDC resolution. No significant change in propagation speed was noted when under the 9.4T field; however, there was a measurable shift when the temperature was shifted between -50°C and +50°C. The data looks slightly noisier when the IC is in the magnetic field because a full density test to characterize the TDC could not be run in the strong magnetic field; compensation for the TDC non-uniformities is thus not completely possible. A fit of a normal curve to the difference between the quench times in the light and the dark allows an estimation of the quench time difference; Fig. 5.9 summarizes the measured avalanche propagation difference with the expected results. The model underpredicts the quench time by roughly 25%. A number of causes are possible for the underprediction, including the following: the junction is probably not an abrupt one-sided junction, there will be some effects from localized charge differences in the diode[52], and the breakdown voltage estimate is not exact. However, the model is able to predict the relative shift in quench time differences.

5.2 Gamma-ray Flooded Environments

Another important consideration for SPAD-based sensors targeting PET, along with SPADs being used in other radiation-laden fields such as space, is the long-term viability of the sensors following exposure to radiation. In a commercial PET scanner, gamma-rays of energy 511keV are incident on the scanner over the course of many years. However, it is not practical to expose a sensor to low dose rates of radiation for years; instead, radiation damage is usually characterized by irradiating a sensor to a high dose rate, and then allowing some annealing to occur[77]. Damage from high-energy gamma-rays, which causes defects in silicon and silicon dioxide, is expected to increase the noise rate of SPAD imagers but have no effect on any of the other FOMs.

To evaluate the effect of radiation on SPAD noise, the same setup used to evaluate noise in the 9.4T magnetic field was setup next to two different Co_{60} sources. Co_{60} emits gamma-rays of roughly 1.25MeV. The first Co_{60} source created a dose rate of 40mGy/s; the other created a dose rate of 800mGy/s. The motherboard and readout system did not receive any significant dose (less than 100Gy in both cases). Different chips were used for the different dose rates.

Fig. 5.10 shows the transient waveforms for the initial 7kGy dose during

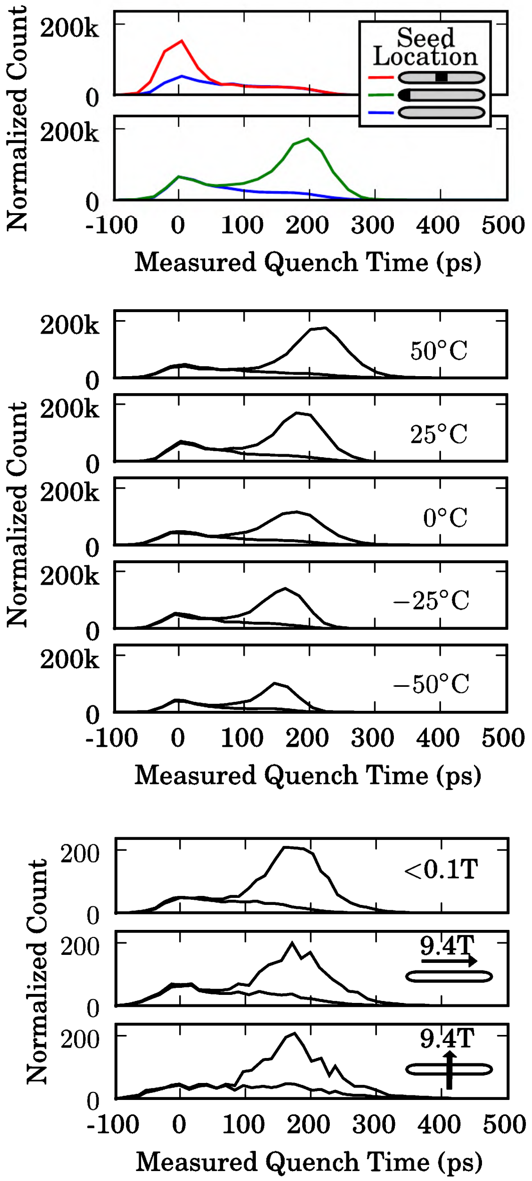


Figure 5.8: **Quench Time Histograms for Different Environmental Conditions**

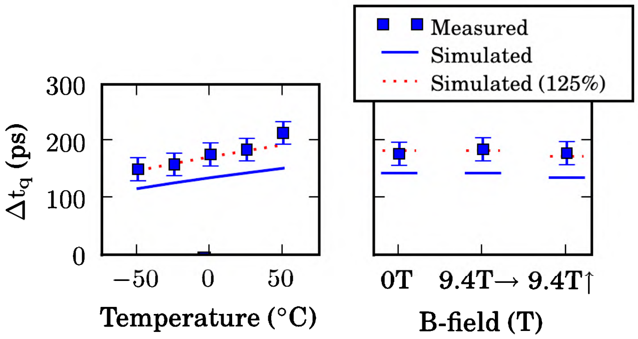


Figure 5.9: **Quench Time's Dependence on Environment** — shown is the measured dependence of the quench time on various environmental factors, along with simulated values and 125% of the simulated values.

the two exposures. There is a significant increase in DCR as a function of the dose rate; the DCR increases by roughly 1Hz for every 1mGy/s increase in dose rate. Because the increase is linear with the dose rate, it is likely due to radiation-generated carriers causing avalanches. When the dose rate was 800mGy/s, the noise began to increase roughly quadratically starting at a total dose of 1kGY, until the readout system experienced a failure at 7kGy. Though it is not apparent from the figure, no failure was ever observed when the dose rate was 40mGy/s; data was only acquired for the initial 7kGy dose at the dose rate of 40mGy/s due to a mistake during setup. The median DCR did not steadily increase at the lower dose rate, but a small decrease in median DCR occurred between dose rates of 3kGy and 6kGy, though the DCR's 75th percentile did steadily increase during the entire 40mGy/s exposure.

The chip at a dose rate of 800mGy/s was irradiated until a total dose of 300kGy, with the chip at 40mGy/s being continuously irradiated until 12kGy. Fig. 5.11 summarizes the DCR increase as a function of total dose. The fraction of SPADs which are "high-noise", defined as the fraction with twice the median count rate, steadily increases from roughly 5% with no radiation dose to almost 35% at 12kGy of total dose. The median noise rate is seen to increase from 150Hz at baseline to a value slightly less than 1kHz at a total dose of 12kGy, with a large increase above 10kHz at 300kGy. It should be noted that very little annealing occurred during these exposures; for the total doses of 12kGy or below, no annealing occurred, while the dose of 300kGy had one week of annealing at room temperature.

In order to understand the importance of the results, the total dose of a PET sensor across its lifetime must be estimated. During a PET scan, a dose of roughly 5mGy[100] is given to a human being. At a dose to the

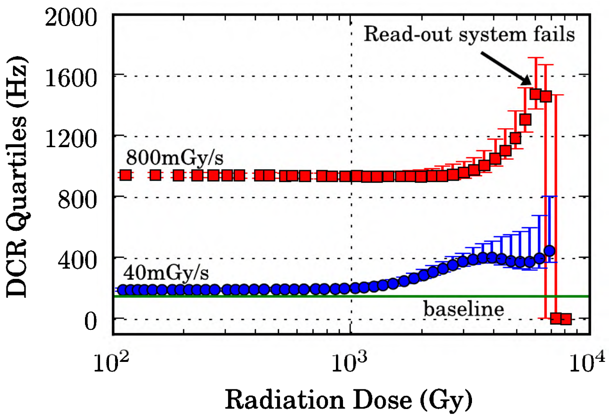


Figure 5.10: **Transient DCR Increase from 1.25MeV Gamma-rays** — shown is the increase in the dark count rate of 1,024 SPADs for doses of 1.25MeV gamma-rays at two dose rates (labeled). No annealing occurred during these experiments.

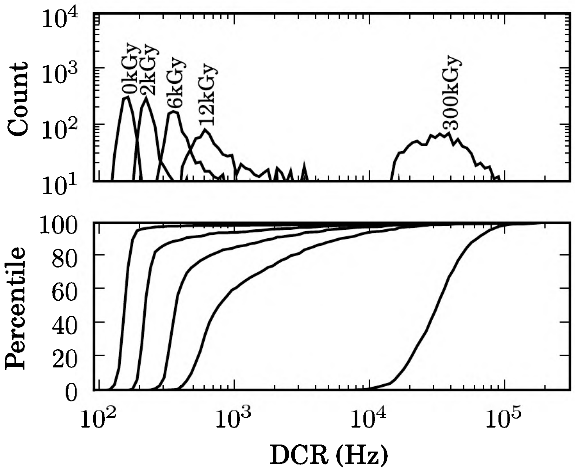


Figure 5.11: **DCR Increase from 1.25MeV Gamma-rays** — shown is the increase of 1,024 SPADs' DCRs for various total doses of 1.25MeV gamma-rays (labels). The 300kGy exposure, which was at a dose rate of 800mGy/s, is following one week of annealing at room temperature. Other data was at a dose rate of 40mGy/s with no annealing.

silicon of 1mGy per scan, probably an order of magnitude too high given the larger distance between the sensor and the radiotracer, several thousand scans are necessary to double the noise rate for these particular diodes if no annealing is allowed. Thus these particular diodes would be quite resistant to radiation damage. However, these diodes use a well-based guard ring, not one that relies on silicon dioxide, which will be much more sensitive to radiation damage. Further work will need to address whether the larger diodes commonly used in PET sensors will be able to withstand the long-term effects of radiation damage, especially architectures that are sensitive to noise[75].

Chapter 6

Electrically Controlling the Breakdown Voltage

Some TCSPC applications require spectral sensitivity. For example, time-resolved Raman spectroscopy[101] and microchip electrophoresis[102] are applications requiring spectral sensitivity in addition to precision in observing the photon's arrival time accurately. To achieve spectral sensitivity, optical filters, gratings, or prisms are often used.

This chapter will present a SPAD that shows promise for achieving spectral sensitivity in a conventional, high-voltage CMOS process without the use of external components. Rather than relying on wells of differing junction depths[103, 104], the structure attempts to implement a guard ring based on polycrystalline-silicon (poly). Modulating the bias on the poly will be shown to change the breakdown voltage at the edge of the diode, though not in the center, allowing control of the region under breakdown. If the poly can double as a color filter, a spectrally sensitive SPAD could be realized in standard CMOS. Fig. 6.1 shows the conceptual difference in PDP graphs when the diode has or does not have a poly filter.

6.1 Guard Rings Utilizing Electrical Effects

Aside from the common guard rings presented in Ch. 2, there are many other styles of guard rings. One such style uses a gate electrode placed above the device edge to avoid edge breakdown[39]. With this type of structure, displayed in Fig. 6.2, the voltage on the guard ring can be modulated to partially control the electrical field underneath the guard ring.

By itself, a gate electrode would not be sufficient to control the breakdown

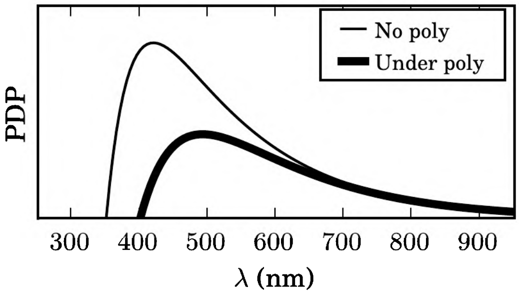


Figure 6.1: **Conceptual Shift in PDP Under a Polysilicon-Filter** — neglecting transmission effects, the polysilicon would absorb nearly all of the blue light, a bit of the green, and little of the red, allowing it to act as a filter.

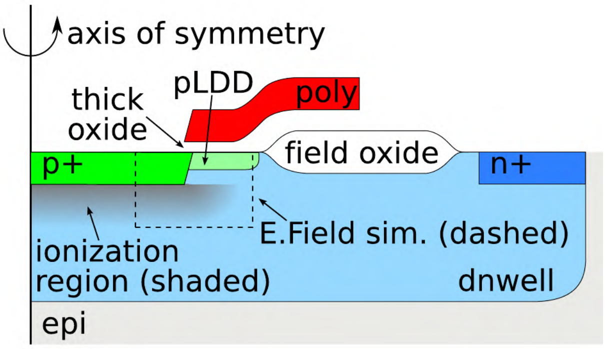


Figure 6.2: **Structure of a Poly-bound SPAD**

voltage at the edge of the guard ring. However, in a 2-poly CMOS process, the lightly-doped drain (LDD) layer will be implemented underneath the second poly if the poly is used in combination with the active implant. The addition of the LDD layer may be enough to prevent premature edge breakdown. Additionally, the extra oxide thickness between the poly and the surface will be crucial in preventing oxide breakdown, which would be problematic as the voltage difference of the p+-poly-n-well path needs to be ~20V in this particular process.

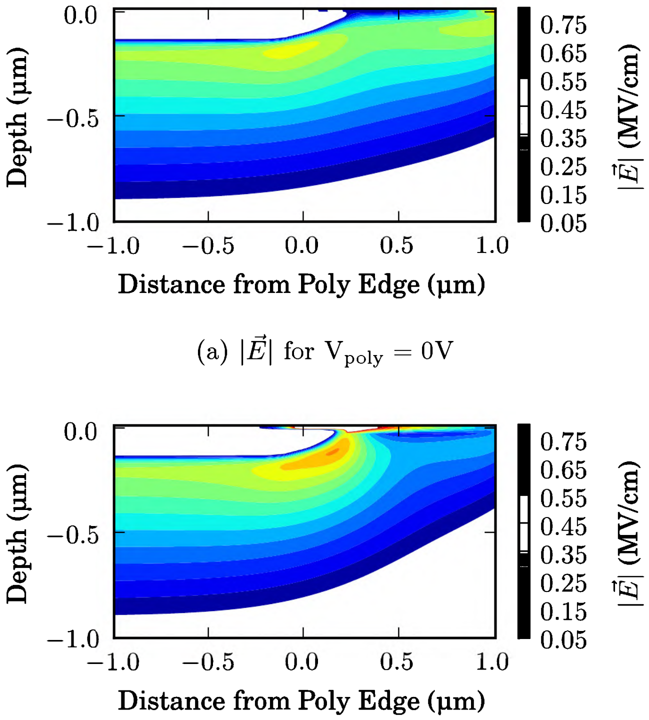
To estimate the breakdown voltage as a function of the gate voltage, a commercial program[24] was used to solve Poisson's equation for the electrical field given the electrical operating points and doping concentrations in the silicon. Fig. 6.3 shows the simulated electric field for different voltages on the poly. Using knowledge of the ionization coefficients as a function of the electric field[58], (2.2) can be solved for equality, yielding the breakdown voltage. The breakdown voltage is expected to be lower at the device edge than at the device center. Even when the poly has the same potential as the p+ implant, there is still a ~10% increase in the $|\vec{E}|$ near the structure edge. When V_{poly} is comparable to the n-well voltage instead of the p+ voltage, there is a >30% increase in $|\vec{E}|$.

Because poly is silicon, it shares roughly the same optical characteristics as silicon. Thus, the region underneath the poly is expected to have roughly the same PDP for red and near IR light, but with a reduced sensitivity to blue light. This creates the possibility of using standard CMOS devices in sensors requiring only a few colors, such those used in multi-color electrophoresis[102]. There is one additional benefit to this structure over the traditional guard ring structure: there is no inactive distance. The effects causing the inactive distance from the diffusion of the guard ring's implants[41] and the horizontally oriented depletion region[42] are both absent in this design.

Thus, if a CMOS process has a poly layer under which an LDD can be implanted, with thick oxide separating the poly layer from the substrate or the n-well, it may be possible to generate a single-junction, color SPAD.

6.2 Fabrication and Characterization

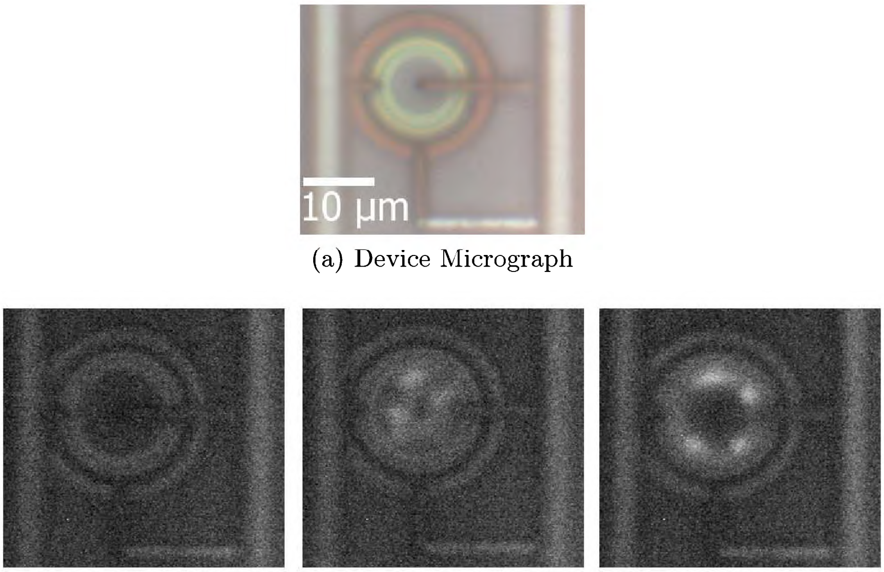
To test the idea of a SPAD with an electrically modulated breakdown voltage, structures based on the layers shown in Fig. 6.3 were generated in a high-voltage CMOS process with 4 metal and 2 poly layers. The SPADs differ in their active area diameters, horizontal guard ring size, and guard ring to n-well contact distance. All SPADs are coupled to their own sized inverter



(a) $|\vec{E}|$ for $\mathrm{V_{poly}} = 0\mathrm{V}$

(b) $|\vec{E}|$ for $\mathrm{V_{poly}} = +16\mathrm{V}$

Figure 6.3: **Simulated $|\vec{E}|$ in a Poly-bound SPAD** — shown are electric field magnitudes near the poly edge when $\mathrm{V_{poly}}= 0V$ (top), or $\mathrm{V_{poly}}= 16\mathrm{V}$ (bottom). Simulated using TCAD [24].



(a) Device Micrograph

(b) Device Emission Micrograph

Figure 6.4: **Poly-bound SPAD Micrographs** — shown are a micrograph (top), and a micrograph of the device with an applied voltage below the breakdown voltage (bottom left), with an applied voltage above the breakdown voltage with V_{poly}= 0V (bottom center), and with an applied voltage above the breakdown voltage with V_{poly}= 16V (bottom right).

with a threshold of roughly 1.2V.

Fig. 6.4 shows a micrograph of one realized diode, along with a micrograph of the diode's optical emission. As the optical emission clearly shows, modulating the voltage on the guard ring allows the breakdown voltage at the edge of the device to shift. Because the breakdown voltage in the middle of the device does not rely on electrical effects, less emitted light comes from the center when the V_{poly} is higher.

However, there is no light observed from underneath the poly, though there is probably light emitted from beneath. Unfortunately, in this particular CMOS process, it is not possible to prevent silicidation of the poly. Silicided poly is opaque to all visible light[105], and thus the spectral sensitivity of this diode could not be tested. However, the diode was still characterized to see if the lack of inactive distance will be useful in an array.

6.2.1 The Breakdown Voltage

The sweep and subtract method, described in Ch. 3 was used to estimate the diode's breakdown voltage as a function of the voltage on the poly. Fig. 6.5

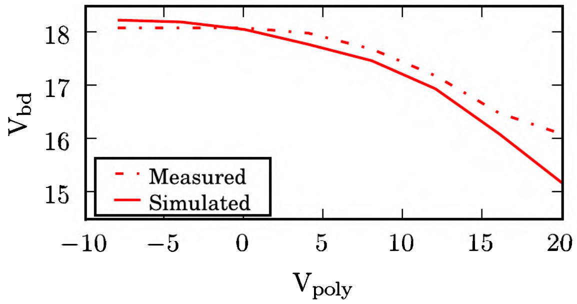


Figure 6.5: **Simulated and measured V_{bd} vs. V_{poly} for a poly-bound SPAD**

compares the simulated results with the measured values. The curves show a good match over the range of interest, which is from V_{poly}=−8V to 20V. The breakdown voltage does not continue to increase when V_{poly} is decreased below 0V. It is also important to note the difference between the V_{bd} in the center of the diode, and the V_{bd} at the diode's edge. For comparison, measurements of the p+-n junction's breakdown voltage from a different run of the same process were, on average, 18.5V[41]. Diodes identical to those previously measured also showed a breakdown voltage of roughly this value. Thus, the smaller V_{bd} for these particular diodes, which is 0.5V below the value expected for a p+-n-well junction, is likely due to the larger electric field at the edge of the structure. Though the poly-bound SPADs always have a small amount of PEB, a difference of 0.5V is not necessarily enough to render the diode useless.

6.2.2 DCR

Fig. 6.6 shows 5s integration time samples of the noise of a 9µm diameter SPAD for various operating points. Also shown in the figure is a well-bound SPAD whose active area, following compensation for the inactive distance, is roughly 9µm. The poly-bound SPADs show an order of magnitude larger noise, ∼ 2kHz compared to ∼ 400 Hz when V_{poly}=0V and V_{eb}≈2V. The poly-bound SPAD's noise increases by almost an order of magnitude for each 4V increase in the poly's potential, with a noise of rate slightly less than 1MHz when V_{poly}=20V.

If a negative voltage was placed on the poly, a drastic increase in noise was seen for low excess biases, with a noise rate of roughly 600kHz even when V_{eb}< 1.5V. However, the noise rate decreases as the excess bias increases. Samples of the count rate at a point where the slope is negative imply that

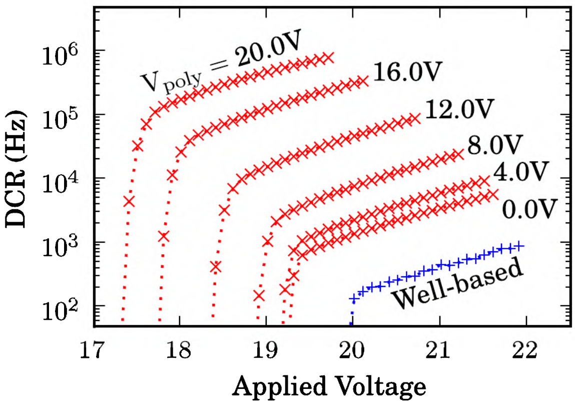
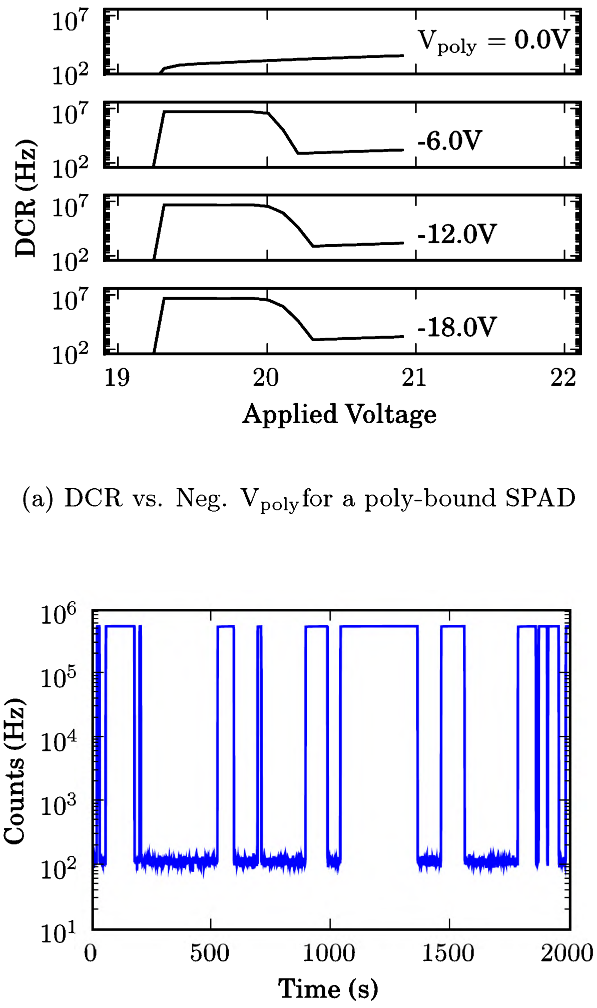


Figure 6.6: **Measured DCR for a poly-bound SPAD when $\mathbf{V}_{\text{poly}} \geq 0$V**

the noise is RTS noise[81]. Given that a negative poly voltage would force electrons, which are minority carriers in the p+ region, generated at the trap-filled surface towards the junction, the large increase in noise at low excess biases is unsurprising. Additionally, because RTS noise is usually caused by surface traps, it is also expected that this particular type of noise would appear when exposing the diode to carriers from generation processes at the surface.

Two things remain unclear about the RTS noise: why does the noise maintain a constant level at low excess biases, even as the excess bias is increased, and why does the RTS noise disappear when the excess bias is increased across the junction? One possibility for explaining the disappearance is that the low-field portion of the diode's depletion region interfaces with the noise-generating location when the excess bias is high, but not when it is low. If this is the case, with a high excess bias, carriers generated at the surface would be swept through the low-field portion of the depletion region, failing to cause ionization. However, at low excess biases, the depletion region might not extend far enough to capture the carrier, allowing it to diffuse to the high-field portion of the depletion region and cause an avalanche. Fig. 6.8 shows the simulated electric field when the diode has a negative potential on the poly, with the applied voltage near the breakdown voltage. It does not appear that the depletion region from the well will come anywhere near the surface region with an applied electric field. Thus, the switch between the high RTS noise state and the low RTS noise state is not obvious, nor is the cause of the noise remaining constant for low excess bi-



(a) DCR vs. Neg. $V_{poly}$for a poly-bound SPAD

(b) DCR when V_{eb}= 19.8V, V_{poly}= -2V

Figure 6.7: **RTS Noise in a poly-bound SPAD**

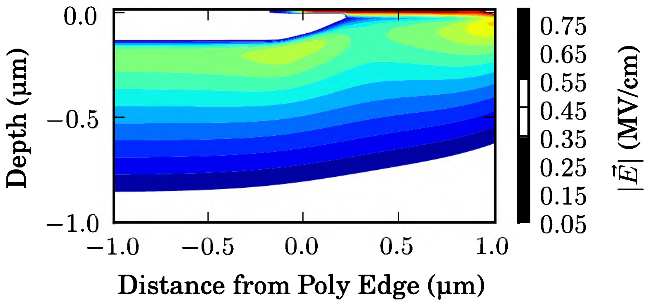


Figure 6.8: $|\vec{E}|$ **for a poly-bound SPAD when** $\mathbf{V_{poly}} = -16\mathbf{V}$

ases; future work will need to address this fact if the diode is ever to be used with a negative voltage on the V_{poly}.

6.2.3 PDP

Fig. 6.9 shows PDP measurements at V_{eb}= 3V using the monochromator and integrating sphere setup described in Ch.3. Note that the constant excess bias in this curve requires that V_{op} be changed to deal with the different breakdown voltage as a function of the gate voltage. There is no inactive distance compensation in this figure; the active diameter used in the calculation is the active area that was laid out. As V_{poly} increases, the PDP decreases, but does so more slowly than might be expected. This may be caused by the increase in the PDP at the edges of the diode; even though the excess bias is constant, the larger electric field will improve the chances of detecting a photon incident on the diode's edge.

6.3 Discussion

A SPAD guard ring allowing adjustable breakdown voltage has been theorized, simulated, fabricated, and experimentally verified. The SPAD uses both solid-state and electrical effects to form a guard ring. Using electrical effects creates the possibility of modulating the breakdown voltage at the edge of the device, but not in the center. The active area modulation has been experimentally verified using light emission techniques. Experimentally measured values are within 2V of the simulated values for the structure. The device shows a larger active area compared to SPADs with well-based guard rings, but suffers from noise that is roughly an order of magnitude larger in the best case. The structure is potentially interesting to time-correlated

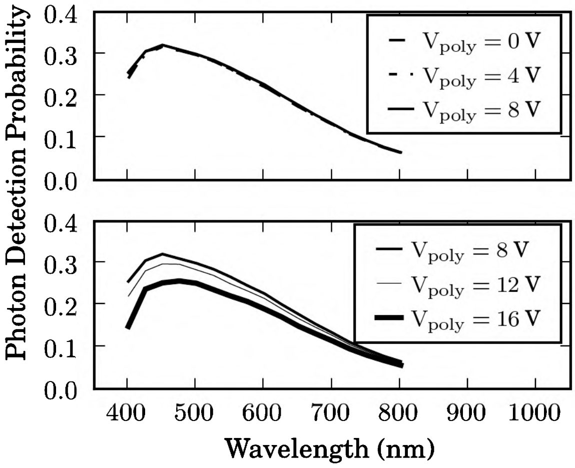


Figure 6.9: **PDP of a poly-bound SPAD**

applications requiring spectral sensitivity, such as DNA detection and time-resolved Raman spectroscopy, but at present the required salicidation of the poly prevents realization of spectrally sensitive diodes.

Unfortunately, in this particular CMOS process, all poly layers use salicided poly, with no facility to prevent salicidation. Processing poly with salicide is known to turn the poly nearly opaque to incident visible light[105]. If the un-salicided poly was available, then all available evidence implies that the poly would be available for use as a color filter, allowing an electrically modulated spectral response, both from the poly and the shift in depletion region depth.

Chapter 7

Fixed-Pattern Noise: Characterization and Mitigation

A small fraction of SPADs causes the great majority of dark counts in arrays fabricated in many different CMOS processes. Radiation damage exacerbates this effect, as presented in Ch. 5. If tunneling-assisted noise was dominant in these high-DCR SPADs, the noise should be uniformly spread across the array and, within a single SPAD, uniformly across the SPAD. Due to the uncorrelated nature of the few high-DCR SPADs' locations[78], the underlying mechanism for this noise is more likely to be trap-assisted or possibly localized tunneling effects. In both of these cases, the noise will be fixed-pattern noise. In a position-sensitive SPAD[106] with a high-noise rate, it should be possible to localize this noise, and selectively ignore avalanches from specific diode regions to increase the signal-to-noise ratio. This chapter presents noise localization and reduction in a single high-noise, position-sensitive SPAD, first published in [41].

7.1 Differences between Noisy and Quiet Diodes

As described in Sec. 5.1.1, it is possible to estimate the originating seed location of an avalanche based on the quench time of the SPAD in long, thin diodes. Avalanches triggered near the middle of such devices will have the avalanche propagate in two directions, creating more current and a faster quench time. The effect requires a high-precision TDC, and two comparators.

An experimental setup based on a chip previously described in Ch. 5 was created to measure the quench time profiles of many different devices. In this

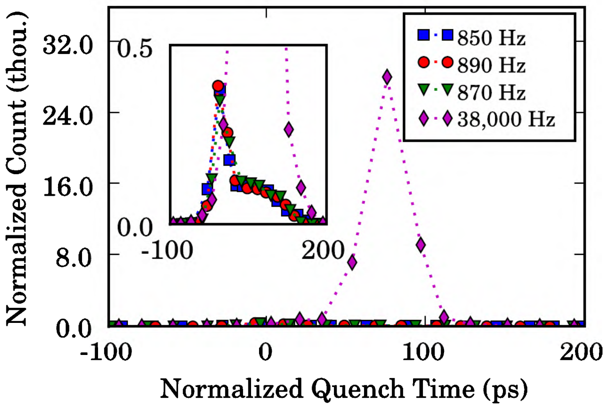


Figure 7.1: **Quench Time Differences between Noisy and Quiet SPADs** — line labels indicate the DCR of the SPAD; the same integration time (1s) was used with the two tests.

setup, the excess bias of all SPADs was set to 2.5V. The SPADs were coupled to a 20ps TDC via two comparators, one with a threshold set to 0.1V and the other with a threshold set to 2.0V. An FPGA senses the avalanche onset, waits for the TDC to sample the quench time, and then reads out the TDC. The FPGA contains two logic programs. One program sends the TDC data to a computer workstation for analysis. The other logic program outputs a real-time digital pulse following acquisition of the TDC data code based on whether or not the code is defined to be a "noisy" code in a look-up table.

From eight tested chips, only one completely open SPAD exhibits high noise. This SPAD exhibits a DCR of ~40kHz, compared to the ~850Hz output by the other diodes. Fig. 7.1 shows the quench time histogram of the high noise diode, along with the quench time histograms of three low-DCR SPADs. The low-DCR SPADs' histograms have the same shape as the simulated predictions shown in Ch. 5; however, the high-DCR SPAD's histogram is dominated by a Gaussian curve corresponding to triggering caused by a single location.

Let t_q be the measured quench time. Fig. 7.2 shows the difference in the pill-shaped, high-DCR SPAD's t_q histogram when in the dark and when under light. Under light sufficient to cause a count rate several times larger than the DCR, the histogram appears to have a similar shape to that of the low-noise SPADs, but with the addition of a Gaussian component represent-

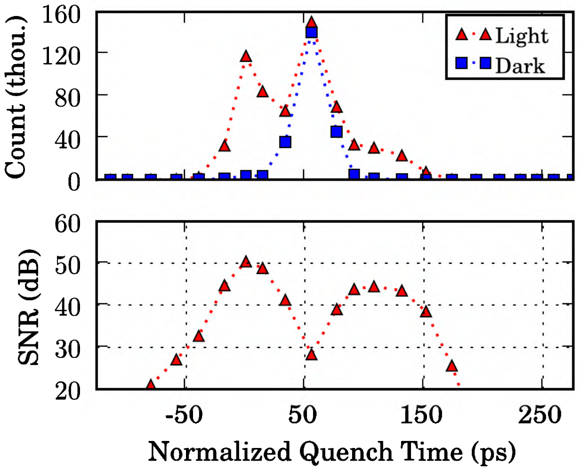


Figure 7.2: **SNR Variation vs. Quench Time (Noisy SPAD)** — integration time was 10s.

ing the fixed position noise. Also shown in the figure is the SNR[1] of each TDC bin using the definition from (2.29).

7.2 Selectively Ignoring Noise

Most of the TDC bins in Fig. 7.2 exhibit shot-noise-limited SNR, but the bin representing a quench time of ~60ps shows a surprisingly low SNR, which is due to the noise. By selectively ignoring this bin, the SNR may improve. Specifically, if the active area of the diode is reduced to fraction f of its previous value, with a corresponding reduction in noise to μ_l, then when light is uniformly seeding avalanches across the remainder of the diode, the SNR will change to

$$\mathrm{SNR} = 20\log_{10}\left(\frac{f\mu - \mu_l}{\sqrt{f\mu}}\right). \tag{7.1}$$

In an experimental setup with the position determined by the quench time, f will be constrained to discrete values governed by what fraction of the diode's

[1] It should be noted that the SNR definition used by the imaging community differs with that used by other communities. Technically the definition should be $10\log_{10}(P_{\mathrm{signal}}/P_{\mathrm{noise}})$, not $20\log_{10}(P_{\mathrm{signal}}/P_{\mathrm{noise}})$. Historically the imaging community has used $20\log_{10}$ since imagers measure voltages, and normally the square of the voltage varies with the power. However, in an ideal imager, the voltage will vary linearly with the incident optical power. A coefficient of 20 will be used here for purposes of consistency with the rest of the imaging community.

active area corresponds to particular t_q values and TDC codes. Selectively ignoring particular TDC codes will eliminate some noise, but at the cost of signal. The exact fraction f to be ignored will depend on the incident optical power and the distribution of the noise in the t_q histogram compared to the signal distribution.

It may not be immediately obvious, but once the incident optical power reaches a certain level, all of the TDC codes should be used. However, at high enough optical powers, even if the noise was perfectly localized the signal increase from including the noisy TDC bin would be larger than the noise increase from ignoring noisy TDC bins. Quantitatively, if the DCR is perfectly localized to a single TDC code that contained fraction f_0 of all events, then the active fraction will be either $1 - f_0$ with noise reduction or 1 without. Use of (7.1) with and without the removal of the particularly noisy diode region gives

$$\begin{aligned}
20\log_{10}\left(\sqrt{(1-f_0)\mu}\right) &< 20\log_{10}\left(\frac{\mu-\mu_{\text{noise}}}{\sqrt{\mu}}\right), & (7.2)\\
\sqrt{(1-f_0)\mu} &< \frac{\mu-\mu_{\text{noise}}}{\sqrt{\mu}}, & (7.3)\\
\mu\sqrt{(1-f_0)} &< \mu-\mu_{\text{noise}}, & (7.4)\\
\frac{\mu_{\text{noise}}}{1-\sqrt{(1-f_0)}} &< \mu, & (7.5)
\end{aligned}$$

yielding the event rate μ corresponding to the incident optical power for which no TDC codes will be ignored. Above this optical power, noise mitigation is pointless — the signal is shot-noise dominated. For a noise rate of 1kHz perfectly localized to 10% of the quench time histogram codes, when $\mu > 19,500$kHz no TDC bins should be deactivated.

However, the noise will not be perfectly localized because the statistical nature of the ionization process creates uncertainty in the quench time. Thus, it will be necessary to determine which regions should be ignored and which should be included for specific levels of light incident on the diode. Algorithms 1 and 2 describe a brute force method of experimentally determining the TDC bins which should be ignored for a specific light level. The algorithms work by testing every possible combination of active and inactive TDC codes, and then choosing the masked codes that yield the best SNR increase.

Algorithm 1 Calculate SNR for SPAD-coupled TDC of n codes

```
Require: mask (which TDC bins to mask), crFromBin (count rate for a
    particular TDC bin), dcrFromBin (dark count rate for a particular TDC
    bin)
    $\mu \leftarrow 0$
    $\sigma^2 \leftarrow 0$
    for $b = 0 \rightarrow n$ do
        if $2^b$ & mask then
            $\mu \leftarrow \mu+\mathrm{crFromBin}[b]-\mathrm{dcrFromBin}[b]$
            $\sigma^2 \leftarrow \sigma^2+\mathrm{crFromBin}[b]$
        end if
    end for
    return $20 \cdot \log(\mu/\sigma)$
```

Algorithm 2 Calculate which of n TDC bins to mask for the best SNR

```
Require: crFromBin (count rate for a particular TDC bin), dcrFromBin
    (dark count rate for a particular TDC bin)
    $\mathrm{SNR}_{\mathrm{max}} \leftarrow -\infty$
    bestMask$\leftarrow 0$
    for $i = 0 \rightarrow 2^n - 1$ do
        $\mathrm{SNR}_{\mathrm{masked}} \leftarrow$maskedSNR($i$, crFromBin, dcrFromBin)
        comment Alg. 1 shows how to compute the masked SNR
        if $\mathrm{SNR}_{\mathrm{masked}} \geq \mathrm{SNR}_{\mathrm{max}}$ then
            $\mathrm{SNR}_{\mathrm{max}} \leftarrow \mathrm{SNR}_{\mathrm{masked}}$
            bestMask$\leftarrow i$
        end if
    end for
    return bestMask, $\mathrm{SNR}_{\mathrm{max}}$
```

7.2.1 Results

Fig. 7.3 shows the experimentally measured SNR increase for the diode previously presented in Figs. 7.1 and 7.2. Photon flux in this figure is derived using the mean ECR increase and a previously measured PDP value of 30% for the specific operating conditions. The diode exhibits an 8dB increase in SNR in photon-starved light levels, though the SNR increase goes to nothing as the signal shot noise increasingly dominates the noise power. Also shown in the figure are the shot-noise-limited, noise-reduced, and raw SNRs. The noise-reduced SNR is still quite low compared to the shot-noise-limited SNR at low photon fluxes. This is due to the fact that the noise is not perfectly localized — i.e. the DCR in this diode is the sum of standard DCR processes, causing a DCR of ~850Hz spread uniformly across the diode, and a fixed-position defect which causes ~37,000Hz of noise. The shot-noise-limited curve also has the PDP of 30% included in the curve — the ideal image sensor would be roughly 3.3× better for the same photon flux.

The SNR increase will saturate at roughly +8dB, since the noise cannot be completely removed. As the incident optical power goes to zero, the SNR increase will approach

$$20\log_{10}\left(\frac{(1-f_0)\mu-\mu_l}{\sqrt{(1-f_0)\mu}}\right)-20\log_{10}\left(\frac{\mu-\mu_{\text{noise}}}{\sqrt{\mu}}\right)\approx 10\log_{10}\left((1-f_0)\frac{\mu_{\text{noise}}}{\mu_{\text{l}}}\right). \quad (7.6)$$

The experimentally measured increase of 8dB is in good agreement with this formula for the reduced noise level at an active fraction of 0.6, given that the measurement of μ_l is several kHz.

Fig. 7.5 shows what fraction of the area is active or inactive as the ECR increases. At low ECRs the active area is expected to be roughly 60% of the low-noise SPAD's active areas. The active area during noise reduction remains so high because the fixed-position defect is close to the edge, whereas nearly all of the center-seeded avalanches end up in the same TDC bin. Fig. 7.3 exposes a major limitation of the technique, which is that for most diodes, the SNR increase is expected to be much smaller than what this diode exhibits. Because the fixed-position defect is so close to the edge, much of the diode's area can remain active, even when noise-reduced. However, many diodes with fixed-position defects will have the defect near the middle, possibly forcing the removal of the TDC bin with the most events. This would cause a smaller increase in the SNR, since a larger fraction of the active area would need to be removed.

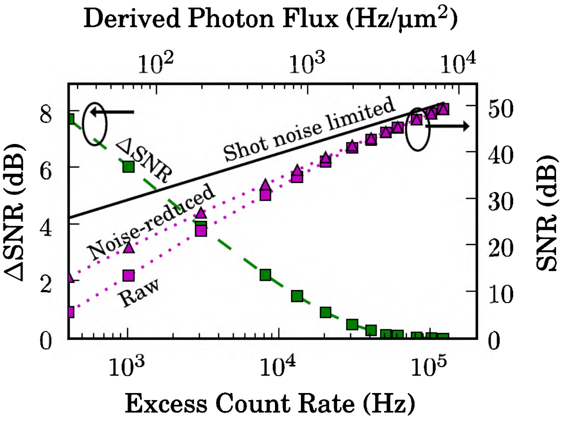


Figure 7.3: **Experimental SNR Increase vs. ECR** — integration time was 1s. The curve for shot noise limited conditions includes the PDP[1].

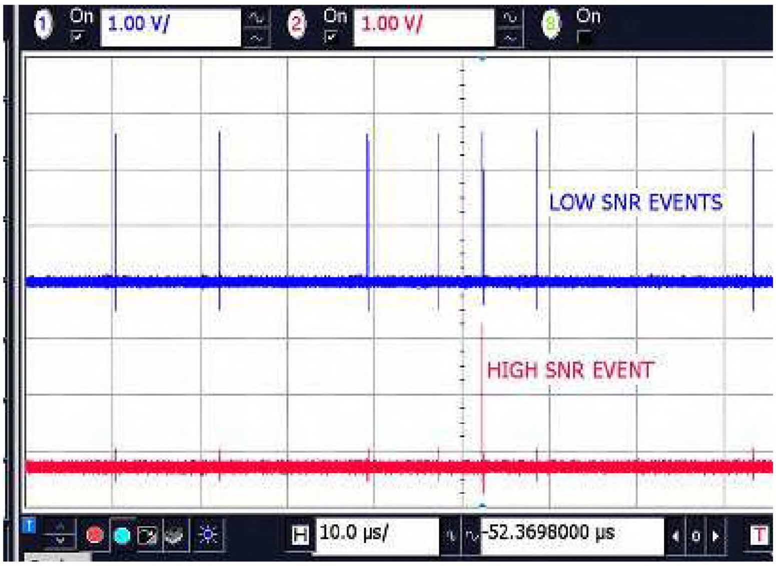


Figure 7.4: **Real-Time Noise Reduction**

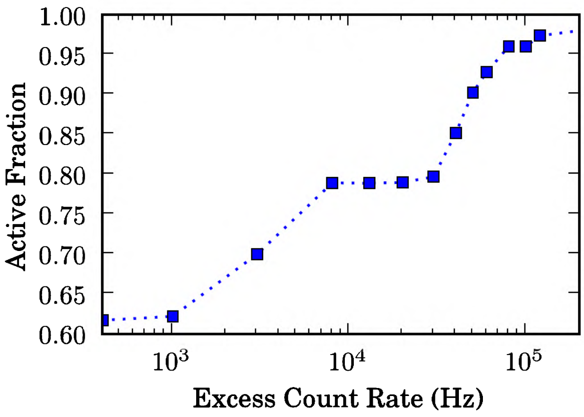


Figure 7.5: **Active Fraction of Noise-Reduced SPAD vs. Excess Count Rate**

7.3 Chip-to-chip Variation

An important question to address when assessing the feasibility of noise reduction is whether or not chip-to-chip variability will cause sizable distortions to the quench time. Whether or not uniform conditions exist across chips is the most important consideration when assessing the uniformity of avalanche propagation. To test the uniformity, the breakdown voltage of different SPADs from eight different chips[1] was acquired using the sweep and subtract method with 0.02V steps presented in Ch. 3. Fig. 7.6 presents the V_{bd} variation; Table 7.1 presents this data in a tabular format. Of interest is the smaller intra-chip variation for larger diodes compared to smaller diodes, and the smaller intra-chip variation compared to the inter-chip variation. Because the accuracy of the sweep-and-subtract method is 0.02V, the standard deviation of the large diodes may simply be a artifact of the acquisition technique; the small set size of $N < 40$ spread across eight chips does not help in this regard. However, the larger variation in the pill-shaped and small diodes, which show a distorted V_{bd}, will be an important consideration when trying to properly bias these diodes.

As Fig. 7.6 shows, chip D is the single outlier, exhibiting SPADs with

[1] The presentation of effects here is different from the order in which they were found and characterized. Attempting to explain the V_{bd} variation resulted in the idea that the compensation from the well doping caused the inactive distance and the shift in the breakdown voltage, spurring the discussion of non-uniformity.

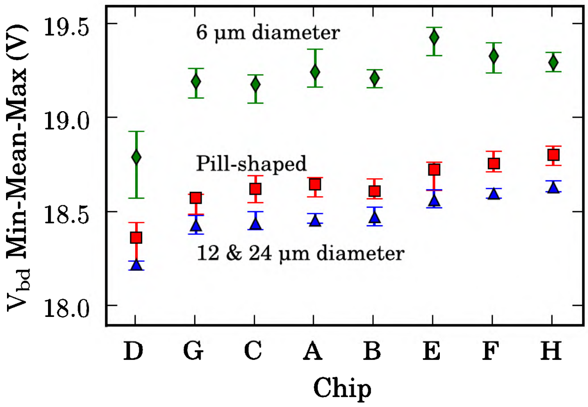


Figure 7.6: **Chip-to-chip V_{bd} Variation**

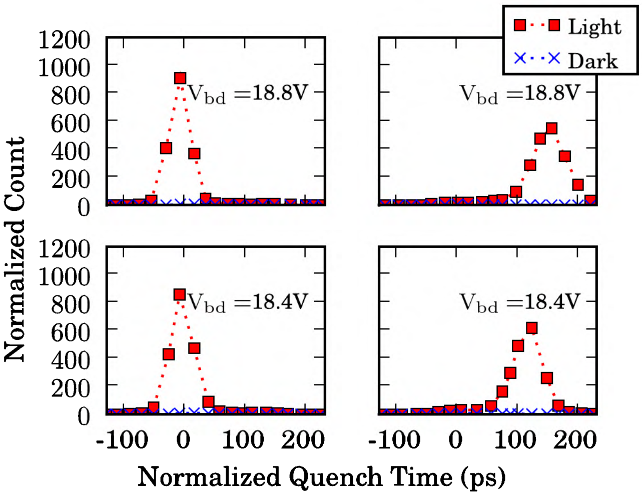


Figure 7.7: **Chip-to-chip Quench Time Variation** — shown are center-seed (left) or edge-seeded (right) avalanche quench times for a low-V_{bd} SPAD (bottom) or high-V_{bd} SPAD (top). Both SPADs operate at the same excess bias, i.e. the top operates at 18.8V+V_{eb} and the bottom 18.4V+V_{bd}.

		SPAD Geometry		
		Large Circular	Pill-shaped	Small Circular
All Chips	Mean V_{bd}	18.48 V	18.64 V	19.21 V
	Maximum V_{bd}	18.69 V	18.85 V	19.50 V
	Minimum V_{bd}	18.19 V	18.29 V	18.70 V
	V_{bd} Range	0.50 V	0.55 V	0.81 V
	V_{bd} Std. Dev.	0.12 V	0.13 V	0.19 V
	Intra-chip variation (mean of chips' V_{bd} std. dev.)	**0.02 V**	**0.04 V**	**0.07 V**
	Inter-chip variation (std. dev. of chips' mean V_{bd})	**0.12 V**	**0.13 V**	**0.18 V**
Without Outlying Chip D	Mean V_{bd}	18.52 V	18.68 V	19.27 V
	Maximum V_{bd}	18.69 V	18.85 V	19.50 V
	Minimum V_{bd}	18.39 V	18.49 V	19.00 V
	V_{bd} Range	0.30 V	0.36 V	0.50 V
	V_{bd} Std. Dev.	0.08 V	0.09 V	0.12 V
	Intra-chip variation(mean of chips' V_{bd} std. dev.)	0.02 V	0.04 V	0.07 V
	Inter-chip variation (std. dev. of chips' mean V_{bd})	0.08 V	0.08 V	0.08 V

Table 7.1: **V_{bd} Variation Statistics** of different SPAD geometries from eight chips in a temperature chamber at 25° C are summarized. Each chip contains 3 small circular SPADs (6 micron diameter), 6 large circular SPADs (12 or 24 micron diameter), and 12 pill-shaped SPADs (roughly 6 microns by 30 microns).

mean V_{bd}s several standard deviations from the other chips. The effect of chip D is largest on the inter-chip variation of the V_{bd}; without chip D, all diodes show a V_{bd} std. dev. of 0.08V, whereas with the chip the other diodes show a larger variation. To assess the performance difference, the quench times acquired from diodes on chip H were compared to the quench times acquired from SPADs on chip D. Fig. 7.7 shows that the quench time for the chip with the higher breakdown voltage is slightly longer than that of the chip with the lower breakdown voltage. From a qualitative standpoint, this makes sense; a higher breakdown voltage implies less ionization occurring at the same excess bias, with slower carrier build-up, avalanche propagation, and quenching.

The SNR increase of +8dB presented above is for a SPAD from chip D, which has a slightly worse quench time difference of ~120ps than the ~140ps exhibited by chip H. The variability in breakdown voltage from chip to chip does not appear to change the quench time differences much, implying negligible variability when utilizing position localization techniques.

7.4 Discussion

While a modest SNR increase of +8dB could be the difference in statistical significance in a light-starved setup, the technique's drawbacks are too large at this point in time to consider it useful for widespread adoption. In large arrays of small pixels, it is probably more practical to include a programmable memory that shuts off the noisy pixels, rather than including another comparator and a TDC with a precision better than 20ps. Additionally, the technique was not found to be useful in circular SPADs due to the smaller quench time differences between edge- and center-triggered avalanches. The technique may be interesting once 3D integration of CMOS circuits becomes more commonplace, and the TDC and comparators can be placed on a different chip, but until that point in time the loss in fill factor is likely to prevent wide-spread use. The effect may also be improved if different versions of position-sensitive SPADs become available, e.g. quadraphonic diodes operating in Geiger-mode, comparable to existing devices that operate in the linear-mode[107].

Despite these drawbacks, noise localization in this manner could open up interesting possibilities when studying noise from a research perspective. Ch. 5 presents noise increases from radiation — are these increases caused mostly by fixed-pattern noise, or is the noise uniform throughout the diode? Is afterpulsing dominated by a few traps with particularly poor release times, or is afterpulsing spread uniformly throughout the diode? Do all SPADs with high noise exhibit fixed-pattern noise, or do some high-noise SPADs have their noise uniformly generated across the active area? These are important questions that the noise localization technique may help answer.

Chapter 8

Sensors for Positron Emission Tomography: A Case Study

This chapter presents an analysis of SPAD trade-offs relevant to positron emission tomography, specifically focusing on whether the fill factor or timing resolution is more important for sensors targeting time-of-flight positron emission tomography (TOF PET). The work was inspired by questions as to whether or not the MEGAFRAME system[108], a SPAD-TDC array with superior timing resolution but low fill factor, would be an effective PET sensor when compared to higher fill factor devices with lower timing resolution, such as digital silicon photomultipliers[75]. This chapter is based on results published in [109].

8.1 A Short Overview of PET

Positron emission tomography is a type of functional imaging often used in cancer detection. PET works as follows: a positron-emitting substance, often the glucose analog fludeoxyglucose (FDG), is injected into a patient. The substance is designed such that it will concentrate in areas using a lot of energy, such as tumors undergoing rapid growth. When a molecule emits a positron, the positron will react with nearby electrons, usually creating two anti-parallel gamma-rays, each with 511keV of energy. Though the gamma-rays can be detected by conversion to Cherenkov photons[110], in commercial systems a scintillator-coupled photomultiplier converts the 511keV γ-rays to processable electrical signals. When multiple gamma-ray detectors are placed in specific geometries, such as a ring, gamma-rays simultaneously incident on two of the detectors, a coincidence event, imply a positron-electron annihilation between the two detectors. Use of algorithms such as filtered back-projection can recover images of positron activity from coincidence events[111]. Due to the possibility of Compton scattering, the photomultipliers must be able to estimate the energy in the arriving gamma-ray; if the energy is too low, the gamma-ray may have scattered before reaching the scintillator, and the position estimate will not be useful[3].

If the gamma-ray detectors are able to determine the gamma-ray's arrival time with an accuracy that is <500ps, timing information can aid in determining the originating positron's position. This is TOF PET, which requires lower radiation doses and generates images with less noise, especially for large-volume scanners like human scanners[112].

Recently, there has been a growing interest in using silicon-based photomultipliers rather than PMTs for TOF PET. Silicon is easy to produce, process, and requires lower bias voltages. Additionally, silicon does not distort magnetic fields, creating the possibility of a dual PET-MRI system[113].

Thus, the goal of a PET sensor is to detect a gamma-ray's energy and arrival time as accurately as possible using knowledge about the timing distribution of emitted scintillation photons, generally modeled as a single exponential[3]. The electrical waveform output by PMTs and analog SiPMs can be used to estimate both of these quantities. However, due to the MRI-incompatible materials used in PMTs, along with distortions in analog SiPMs and readout caused by the Hall effect, the following discussion will focus on so-called "digital SiPMs," which output digital time information using on-chip TDCs.

8.2 A Single-Exponential Model

Imagine, for a moment, that the "holy grail" sensor discussed in Ch. 1 is available — a silicon chip that gives the arrival time of every incident photon with no distortion — and this chip is coupled to a shot-noise-free scintillator which outputs light with a single exponential decay. What would the inaccuracy in the timing estimate be if this setup was available? When the measurements of the scintillation photons' arrival times are i.i.d. with a single exponential governing the distribution, an elegant, closed-form solution exists for estimating the gamma-ray's arrival time.

Assume that a scintillator-coupled SPAD array measures time stamps $t_1 ... t_n$ corresponding to the photons emitted by a scintillation when a gamma-ray is incident at time t_0. Each one of these time stamps, as per the assumption in the previous paragraph, has a PDF governed by a single exponential

$$f_{T_i}(t_i) = \begin{cases} 0 & \text{if } t_i < t_0, \\ \frac{1}{\tau_d} \exp\left(-\frac{t_i - t_0}{\tau_d}\right) & \text{if } t_i \geq t_0, \end{cases} \tag{8.1}$$

with τ_d being the time constant of the exponential. The time stamps can be sorted to find the order statistics, $t_{1:n}...t_{n:n}$. The governing distribution of these statistics will be the Rènyi representation[114],

$$T_{i:n} = t_0 + \tau_d \sum_{k=1}^{i} \left(\frac{Z_k}{n-k+1} \right), \tag{8.2}$$

with the Z_k variables being i.i.d. exponential distributions with time or rate constant one, i.e. $\lambda = \tau = 1$. As the Z_k variables are i.i.d., with $E[Z_k] = 1$ and $\text{var}(Z_k) = 1$, the moments of T_i follow from basic tenets of probability[69], being

$$E[T_{i:n}] = t_0 + \tau_d \sum_{k=1}^{i} \left(\frac{1}{n-k+1} \right), \tag{8.3}$$

$$\text{var}(T_{i:n}) = \tau_d^2 \sum_{k=1}^{i} \left(\frac{1}{n-k+1} \right)^2. \tag{8.4}$$

To derive the covariance between $T_{i:n}$ and $T_{j:n}$, note that the two have the $Z_1...Z_{\min(i,j)}$ variables in common, along with these variables' coefficients. Thus,

$$\text{covar}(T_{i:n}, T_{j:n}) = \text{var}(T_{\min(i,j):n}). \tag{8.5}$$

First it will be shown that, for an unbiased, linear estimator based on any pair of $(T_{i:n}, T_{j:n})$, all weight in the estimator should be given to the order statistic with lower rank. This result will then be expanded to non-pair estimators, proving that only the lowest order statistic should be used when estimating t_0.

For the pair $(T_{i:n}, T_{j:n})$, the unbiased, linear estimator of t_0 will be

$$\begin{aligned} \tilde{t_0}(T_{i:n}, T_{j:n}) &= \alpha \left(T_{i:n} - \tau_d \sum_{k=1}^{i} \left(\frac{1}{n-k+1} \right) \right) + \\ &\quad (1-\alpha) \left(T_{j:n} - \tau_d \sum_{k=1}^{j} \left(\frac{1}{n-k+1} \right) \right). \end{aligned} \tag{8.6}$$

The variance of this estimator will be

$$\begin{aligned} \mathrm{var}(\tilde{t_0}(T_{i:n}, T_{j:n})) &= \mathrm{var}\,(\alpha T_{i:n}) + \mathrm{var}\,((1-\alpha)T_{j:n}) + \\ & \quad 2 \cdot \mathrm{covar}\,(\alpha T_{i:n}, (1-\alpha)T_{j:n})\,, \end{aligned} \tag{8.7}$$

with the sums disappearing because the variance of constants is zero. Assuming WLOG that $i < j$, by (8.5), $\mathrm{covar}(T_{i:n}, T_{j:n}) = \mathrm{var}(T_{i:n})$, and the expression can be simplified to

$$\mathrm{var}(\tilde{t_0}(T_{i:n}, T_{j:n})) = \mathrm{var}(T_{i:n}) + (1-\alpha)^2\,(\mathrm{var}(T_{j:n}) - \mathrm{var}(T_{i:n}))\,. \tag{8.8}$$

Because $\mathrm{var}(T_{j:n}) > \mathrm{var}(T_{i:n})$, both $(\mathrm{var}(T_{j:n}) - \mathrm{var}(T_{i:n}))$ and $(1-\alpha)^2$ are non-negative quantities. The expression is minimized when $\alpha = 1$, implying all weight in the estimator should be placed on the order statistic with lower rank to achieve a better estimator.

The unbiased, linear expression for all order statistics will be

$$\mathrm{var}(\tilde{t_0}(T_{1:n} \ldots T_{n:n})) = \sum_{i=1}^{n} \left(\alpha_i \left(T_{i:n} - \tau_d \sum_{k=1}^{i} \left(\frac{1}{n-k+1} \right) \right) \right)\,, \tag{8.9}$$

with $\sum_{i=1}^{n}(\alpha_i) = 1$. By the previous result, for any pair consisting of $\alpha_1 T_{1:n}$ and $\alpha_j T_{j:n}$, the estimator variance will be lower if $\alpha_j = 0$ and α_1 has the highest possible value. By induction, when $\alpha_1 = 1$ and all other αs are zero, the estimator will have the least variance of any linear estimator. Thus, the minimum-variance, linear, unbiased estimator will be solely the result of the first order statistic, with the estimation error following an exponential distribution with time constant τ_d/n. The error in estimating any β^+ particle's position will be the convolution of these two distributions, yielding a FWHM of $\tau_d/n(2\ln(2))$. The FWHM yields a very good result, though the tail behavior of this distribution does not match that of a normal distribution, falling off as $\exp(1/t)$, and not a normal distribution's $\exp(1/t^2)$.

The fact that only the first order statistic is necessary is intuitive when the memorylessness of the generating process is considered. The time difference between the first order statistic and the gamma-ray's arrival time is exponentially distributed with time constant $\tau = \tau_d/n$, due to the fact that the order statistic is the first sample chosen from n exponentially distributed processes with time constant τ. By induction, and due to the fact that the processes are memoryless, the difference in time between the second order statistic and the first order statistic is an exponential process with time constant $\tau = \tau_d/(n-1)$. Similarly, the difference between the second and third will be exponential with constant $\tau = \tau_d/(n-2)$. In other words, the first order statistic poorly samples the arrival time, the second order statistic poorly samples the first, the third poorly samples the second, etc. Because the second order statistic is a sample of the first, there is no additional information that is added by the second order statistic.

The lack of new information comes from the memorylessness of the governing processes — if any information were to be present in the second order statistic that is not present in the first order statistic, then the process must have some memory of prior events[1]. This fact will be an important guiding principle when working with distributions that approach memoryless distributions over time, such as double-exponential distributions. As the distribution approaches memorylessness, higher rank order statistics become poorer estimators.

8.3 Sensor Requirements

A number of questionable assumptions were made to achieve the statistical lower bound in the previous section: shot noise was neglected; the SiPM was assumed to be ideal; and the scintillation photons were assumed to follow an exponential distribution. In reality, none of these assumptions are reasonable.

However, the initial proof provides a basis for a first-order analysis of the timing uncertainties inherent in TOF PET utilizing SiPMs. If the distribution governing the scintillation photons' arrival times is known, the best single-photon-based estimator can be derived. The parameters influencing the scintillation photons' arrival times can be swept to answer important questions when considering constructing PET sensors, especially concerning the fill factor's trade-off with timing uncertainty. The major weakness in this analysis is that only estimators based on single order statistics will be compared.

8.4 Estimating Performance Results with Order Statistics

Let $f_a(t)$ and $F_a(t)$ be the PDF and CDF, respectively, governing the measurement of the arrival time of scintillation photons to a digital SiPM. This PDF will be assumed to be a double-exponential function

$$f_a(t) = \begin{cases} 0 & \text{if } t < 0, \\ \frac{\exp\left(-\frac{t}{\tau_d}\right) - \exp\left(-\frac{t}{\tau_r}\right)}{\tau_d - \tau_r} & \text{if } t > 0, \end{cases} \tag{8.10}$$

[1] The result only follows if the definition of a memoryless process is intuition.

with τ_d and τ_r being the decay and rise times, respectively[115]. The PDF of the measured arrival times, $f(t)$, is assumed to be the convolution of $f_a(t)$ with a normal distribution representing the SPAD's jitter,

$$f(t) = f_a(t) * \mathcal{N}[\mu,\, \sigma^2](t), \tag{8.11}$$

where $\mathcal{N}[\mu,\, \sigma^2](t)$ is a normal distribution (mean μ, std. dev. σ) and σ is the std. dev. jitter of the SPAD comprising the digital SiPM array. Photons emitted by fast scintillators such as LYSO tend to be around 400nm in wavelength; thus the exponential tail will be negligible in the timing response, and the Gaussian approximation will be a good one[67]. The variable μ will be set to 1 nanosecond to account for delays introduced by the readout circuitry, though this variable will be unimportant when considering the final result as it will not appear in the distribution of coincident gamma-rays.

Any distribution independently sampled n times will yield order statistics with PDFs given by

$$f_{k:n}(t) = n\binom{n-1}{k-1} f(t)F(t)^{k-1}\left[1 - F(t)\right]^{n-k}, \tag{8.12}$$

where k denotes the k^{th} order statistic[114]. In the context of this problem, when sampling $f(t)$ the variable n will denote the number of avalanches expected to be generated in a dead-time free detector from scintillation photons. n is also known as "primary photoelectrons." Because the double-exponential process is not memoryless like the single-exponential, there is no longer any guarantee that the first order statistic will be the basis for the best estimator. Neglecting, for a moment, afterpulsing, noise, shot noise, and dead time, the FWHM of the order statistic's distributions can be compared to find the index of the minimum-FWHM estimator based on a single order statistic. The assumptions about noise, dead time, shot noise, and afterpulsing will then be revisited, to see if it is reasonable to ignore these effects.

Fig. 8.1 shows the double-exponential governing the arrival of scintillation photons from LYSO on the digital SiPM, with τ_r = 500ps and τ_d=40ns. There is still disagreement in the rise time of the generation process; two different results will be compared for LYSO, with τ_r = 80ps or τ_r = 500ps, to see what effect this variable will have on the end result[116, 115].

Fig. 8.2 shows the first five order statistics from a simulated, LYSO-coupled SiPM with n =100 or 800. When n = 100, the first order statistic is a better estimator than any other order statistic; however, when n = 800, this is no longer the case. Imagine that the SPAD jitter σ was many times larger than even τ_d — in this case, $f(t)$ would be $\sim \mathcal{N}[\mu, \sigma^2](t)$, and the median order statistic would be the best estimator. If the double-exponential

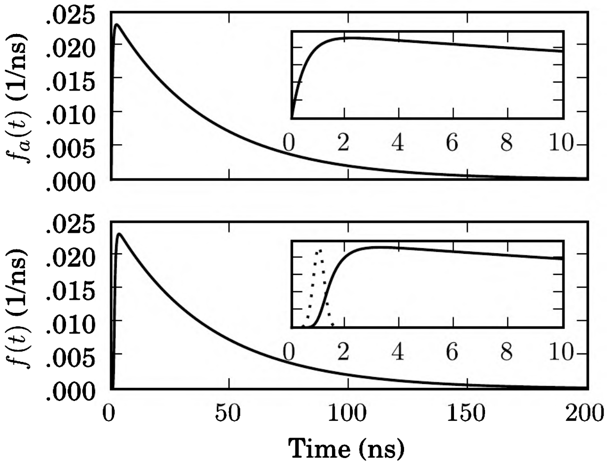


Figure 8.1: **Generation and Measurement Distributions for LYSO** — shown are the simulated generation (top) and measurement (bottom) distributions for LYSO, as defined in (8.10) and (8.11), respectively, with insets showing the first few nanoseconds. Overlaid on the bottom inset is the sensor jitter. In these graphs, $\tau_r = 500\text{ps}$, $\tau_d = 40\text{ns}$, and $\sigma = 190\text{ps}$.

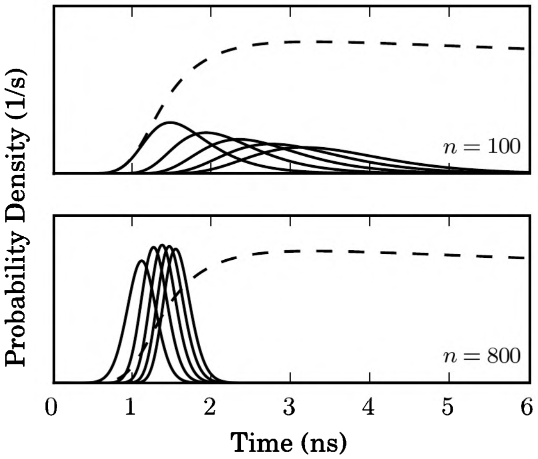


Figure 8.2: **Initial Scintillation Photon Arrival Time Distributions** — Shown are the order statistics of rank 1, 2, 3, 4, and 5, when 100 (top) or 800 (bottom) samples are taken from the $f(t)$ in (8.11). These distributions reflect the arrival times of the 1st to 5th scintillation photons. Also included is a 100x scaled version of original $f(t)$ (dashed curves), the distribution from which the samples are taken.

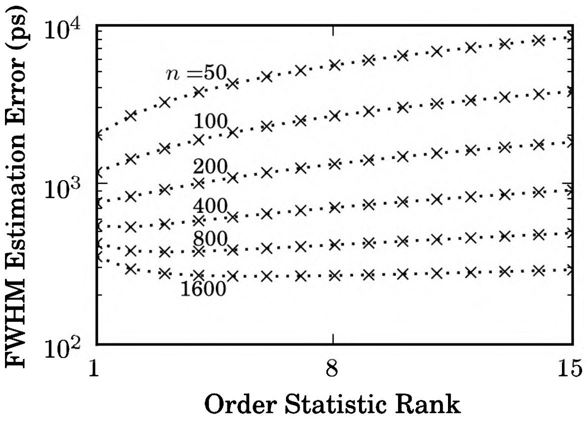


Figure 8.3: **Estimation Error vs. Order Statistic Rank for LYSO** — shown is the FWHM of the k^{th} order statistic's distribution for various detection efficiencies when $\tau_r = 500$ps, $\tau_d = 40$ns, and the SPAD jitter (σ) is 190ps. The FWHM of the distribution is identical to the FWHM of the estimation error when that particular order statistic is used as the basis of the gamma-ray's arrival time estimator.

is viewed as, approximately, a Gaussian convolved with a single-exponential, then the situation depicted in Fig. 8.2 can be intuitively viewed as a match between the best estimator for the Gaussian, which is the median order statistic, and the best estimator for the exponential, which is the initial order statistic.

Fig. 8.3 shows how the FWHM of order statistics with ranks between 1 and 15 will vary for a LYSO-coupled SiPM with a microcell jitter of 190ps. If this particular order statistic's rank is used as the basis for the estimator of the gamma-ray's arrival time, then the FWHM will also describe the error in such an estimator. As Fig. 8.3 shows, n increasing will cause the rank of the best order statistic to also increase.

8.5 Revisiting the Assumptions

Before discussing the results, it is necessary to check, based on the new information, whether or not the lack of several factors in the model is justified.

8.5.1 Shot Noise

Now that the rank of the optimal order statistic for an estimator is known, the contribution of shot noise can be quantified. Let the optimal rank be r_o. To find the distortion from shot noise, the distribution of $t_{r_o:n}$ can be compared to the sum of the $t_{r_o:n}$ distributions when weighted by a Gaussian with mean n and std. dev. $\sqrt(n)$. In other words, the distribution distorted by shot noise, $f_{\text{shot}}(t)$, will be

$$f_{\text{shot}}(t) = \frac{\sum_{k=1}^{\infty} \left(\mathcal{N}[n,n](k) \cdot f_{r_o:n}(t)\right)}{\sum_{k=1}^{\infty} \mathcal{N}[n,n](k)}, \tag{8.13}$$

with the $\sum_{k=1}^{\infty} \mathcal{N}[n,n](k)$ term in the denominator necessary for the purposes of normalization. Because the value of $\mathcal{N}[n,n](k)$ will be negligible when k is not in $n \pm 3\sqrt{n}$, (8.13) can be evaluated with a restricted support.

Fig. 8.4 shows $f_{3:770}(t)$, $f_{3:800}(t)$, $f_{3:830}(t)$, and the scaled deconvolution of $f_{3:800}(t)$ with (8.13)'s result. The deconvolution can be thought of as a graphical description of the shot noise's degradation of the estimate. If the deconvolution has a σ larger than or similar to $f_{3:800}(t)$'s std. dev., then the shot noise will cause degradation. However, Fig. 8.4 shows this is not the case for $n = 800$ — the shot noise's contribution to the timing uncertainty is trivial compared to the inherent timing uncertainty. All tested conditions had increases of less than 5% in FWHM due to the shot noise. Thus, the shot noise does not cause significant error in the arrival time estimate.

8.5.2 Noise

Three types of noise may interfere with the estimate — uncorrelated noise, afterpulsing, and crosstalk.

Because afterpulsing occurs after the avalanche diodes themselves have fired, afterpulsing will not cause any impact on the arrival times of the initial scintillation photons. Thus, afterpulsing can be safely disregarded.

Unfortunately crosstalk cannot be so easily disregarded. However, the effects of crosstalk can be ignored so long as the probability of an avalanche being caused by crosstalk before the measurement of the ideal order statistic is set to be low, say 5%. If the probability of crosstalk is 0.5%[1], then only the first ten order statistics can be considered, and in this case crosstalk can be neglected.

[1] This probability may seem quite low; however, crosstalk is highly correlated in space, and a digital silicon photomultiplier might be able to suppress nearby cells' firing to aid in the reduction of crosstalk. Future work will need to address how much of an issue crosstalk is in tightly packed arrays.

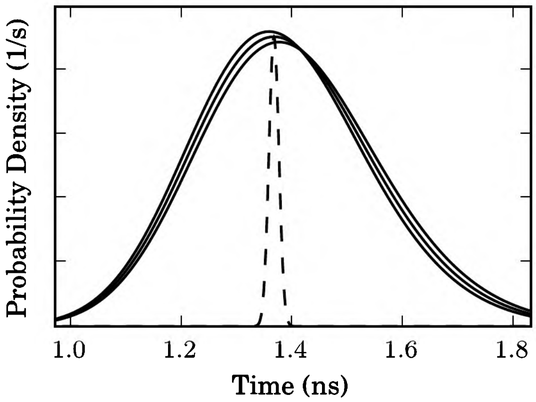


Figure 8.4: **Shot Noise Degradation of a Gamma-ray's Arrival Time Estimate** — shown are the distributions of the measured third photon arrival when n =770, 800, or 830 (solid), along with a scaled Gaussian distribution that approximates the shot noise degradation (dashed) for $n = 800$. In all simulated trials, the shot noise caused the estimation error's FWHM to vary less than 5%.

The same is not true of uncorrelated noise. To understand the effects of uncorrelated noise, it is necessary to have a trigger condition for when photons from a gamma-ray are arriving. Let the trigger condition be as follows: assume that a gamma-ray has arrived if four or more avalanches occur in the five nanoseconds following an initial avalanche, and assume that this initial avalanche is the order statistic with rank one. Two things must be checked. First, no avalanches should occur five nanoseconds before or after the initial scintillation photons. Second, the ideal order statistic for estimation must occur within five nanoseconds of the initial order statistic.

The probability that a dark count will occur in the ten nanoseconds surrounding the event is $1 - \exp(r \cdot (10 \text{ ns}))$, with r being the event rate of noise. The probability of a dark count in this window is less than 1% if the noise rate is ~1MHz, increasing to ~5MHz if the probability is relaxed to 5%. Thus, so long as the noise rate is less than 5MHz, there will be a negligible effect from uncorrelated noise.

8.5.3 Dead Time

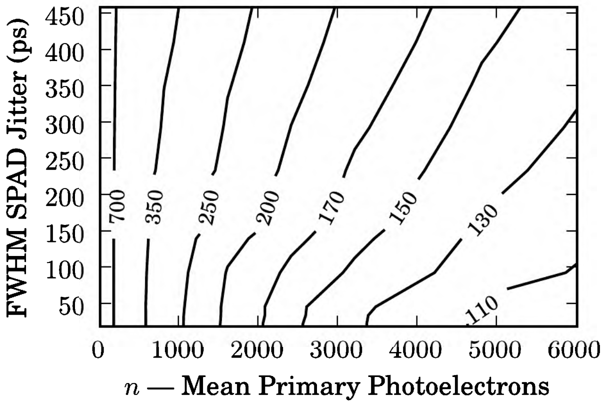
When a SPAD avalanches, it must be restored before it can avalanche again. If the ideal order statistic for the gamma-ray's arrival time has a large rank, then there will be some distortions in the order statistic's distributions due to the decrease in active area as more and more SPADs avalanche. The probability that a scintillation photon will impinge on an avalanching SPAD, that is that two SPADs will share a photon, is exactly analogous to the birthday problem, which is the probability that two people in a room share a birthday[117]. For a 1,000 cell SPAD array, fewer than 11 SPADs must be expected to fire for a >95% chance that none of the avalanches are expected to occur in the same cell. The exact same criteria used to limit the performance in the case of afterpulsing, that fewer than 10 cells fire, is also valid in this case.

8.6 Results

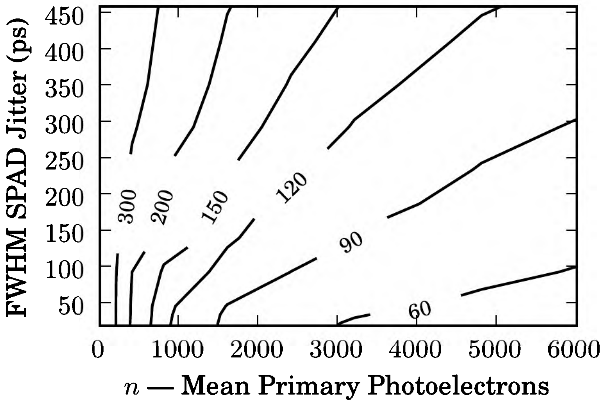
The minimum for each curve in Fig. 8.3 can be found, and then the detector efficiency can be swept to yield the FWHM error of a gamma-ray's arrival time for a particular set of conditions. As per the discussion of distortions from dead time and crosstalk, the minimum will also be considered if the order statistics are restricted to being of rank 10 or less. Figs. 8.5 and 8.6 show these sets of points as contour plots for four different (τ_d, τ_r) pairs, representing: LYSO with τ_r=500ps; LYSO with τ_r =80ps; LaBr$_3$; and an imaginary scintillator. Fig. 8.6 also contains curves produced when only order statistics of rank 10 or less are considered.

As the figures show, at low detection efficiencies the microcell jitter has little to no effect on the estimation error of the gamma-ray's arrival time. For small n, the distortion from the rise-time will be minimal. As n increases, the overlap between the first order statistic's distribution with the section of the measurement CDF and PDF will increase, with increasingly divergent behavior in the initial order statistics. For scintillators with faster decay times, the timing jitter becomes increasingly important as more than 1,000 scintillation photons are expected to create avalanches. However, even when a 200ps jitter SiPM coupled to LaBr$_3$ is expected to observe 2,000 scintillation photons, it is better to double the photon detection efficiency than it is to halve the jitter.

Figs. 8.5 and 8.7 shows the important role that the rise time plays in the estimate of the gamma-ray's arrival time. In Fig. 8.5, as τ_r changes from 500ps to 80ps, the estimation error is cut roughly in half. There is

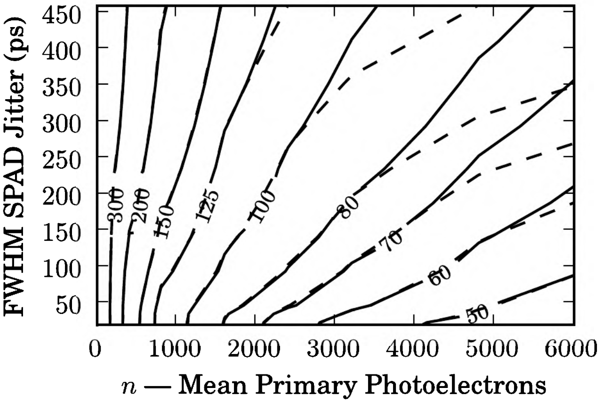


(a) LYSO with $\tau_d = 40$ns, $\tau_r = 500$ps

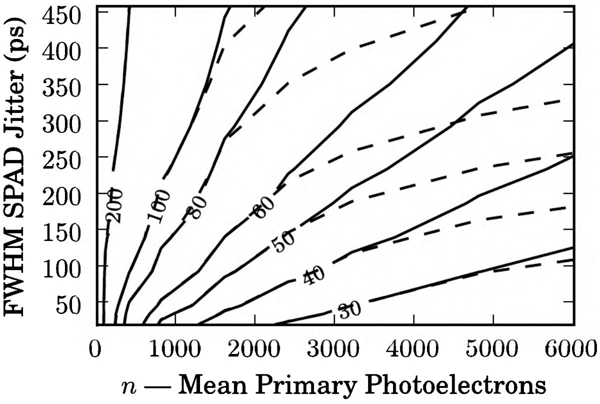


(b) LYSO with $\tau_d = 40$ns, $\tau_r = 80$ps

Figure 8.5: **Single Gamma-ray Arrival Time Estimation Errors** — shown are FWHM errors (in ps) of the gamma-ray arrival time's estimate for LYSO coupled to a digital silicon photomultiplier with various microcell jitters (ordinate) and detection efficiencies (abscissa). The rise time of the LYSO was simulated as either 500ps (top), or 80ps (bottom).

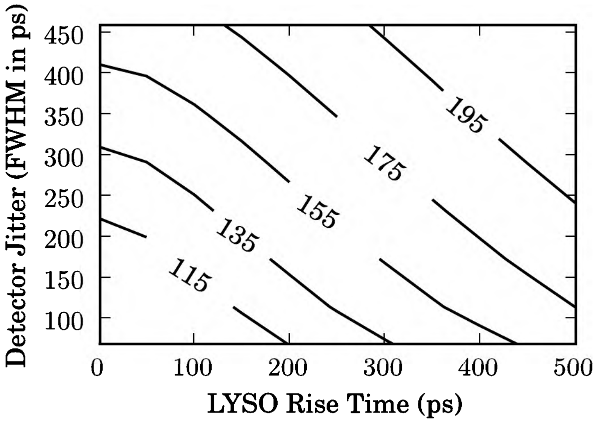


(a) LaBr$_3$ with $\tau_d = 17$ns, $\tau_r = 500$ps

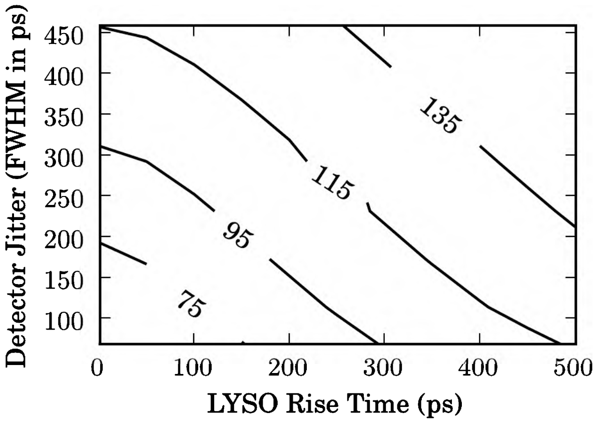


(b) Imaginary Scintillator with $\tau_d = 10$ns, $\tau_r = 500$ps

Figure 8.6: **Single Gamma-ray Arrival Time Estimation Errors (cont.)** — shown are FWHM errors (in ps) of the gamma-ray arrival time's estimate for LaBr$_3$ (top) or an imaginary scintillator (bottom) coupled to a digital silicon photomultiplier with various microcell jitters (ordinate) and detection efficiencies (abscissa). The decay and rise times of the LaBr$_3$ were 17ns and 100ps[118]; they were 10ns and 50ps for the imaginary scintillator. The solid lines show the estimate with dead time and crosstalk effects, with the dashed lines show the estimate with the compensation discussed in Sec. 8.5 for these effects.



(a) LYSO with τ_d = 40ns, 2,400 primary photoelectrons



(b) LYSO with τ_d = 40ns, 4,800 primary photoelectrons

Figure 8.7: **Single Gamma-ray Arrival Time Estimation Errors (cont.)** — shown are FWHM errors (in ps) of the gamma-ray arrival time's estimate for LYSO coupled to a digital silicon photomultiplier with various microcell jitters (ordinate) and scintillator rise times (abscissa). The number of primary photoelectrons was simulated as either 2,400 (top), or 4,800 (bottom).

still a large disagreement in the literature as to the exact value and cause that the rise time should take. Values as low as 80ps have been reported for small LYSO crystals, though larger rise times have been reported for larger LYSO crystals. Fig. 8.7 shows the nearly linear trade-off between the SPAD detector jitter and the rise time from a LYSO crystal for set detector efficiencies. To keep the same minimum bound when using a crystal with a higher rise time, the detector must have a better jitter. No matter the cause of the rise time, it is an important factor when simulating the error in estimations of a gamma-ray's arrival time.

If LYSO is assumed to give off 14,000 scintillation photons per 511keV gamma-ray, for system detection efficiencies (consisting of the PDP multiplied by the fill factor) below 10%, the SiPM's timing resolution plays almost no part in the estimation error of the gamma-ray's arrival time. If the system detection efficiency increases to 20%, the timing resolution begins to factor if τ_r is 80ps, but for LYSO with a longer rise time of 500ps, the timing resolution plays very little role until the system detection efficiency is 30% or greater. Similarly for LaBr_3 and the imaginary, 10ns decay time scintillator, the timing resolution is not critical until the number of primary photoelectrons increases past several thousand.

Finally, Fig. 8.6 shows the effect of limiting the estimators to the first 10 order statistics to mitigate the distortions from crosstalk and dead time. The effect will cause distortions for an efficient detector with a large jitter, which is reasonable given that as the jitter and detection efficiency increase, more order statistics have their distributions distorted by the rise time.

There are several weaknesses in the present model that future work will need to address when using these types of statistical analysis. First and foremost, there is no experimental data presented here, just a model. Experimental data is needed to verify the timing information, and systems have begun to be created that can perform this verification[68]. The assumed crosstalk value, 0.5%, does not accurately capture the high rates of crosstalk in tightly packed SPAD arrays, and this particular point is probably the weakest portion of this analysis. Additionally, estimators based on single order statistics, while ideal for the single exponential case, are not likely to be the best estimators for generation processes that differ from the single exponential case. Ongoing work is already examining how estimators based on multiple order statistics can improve the estimation error[119]. This work is important for understanding the tradeoffs in different chip architectures, and making the best possible PET sensor with the limitations of current technology.

Chapter 9

Conclusion

Many applications, especially biomedical ones, rely on the unique nature of light. This thesis has discussed how understanding the physics behind CMOS-integrated SPADs is important when deciding on sensors and architectures to detect photons for applications such as positron emission tomography.

9.1 Contributions

Following presentation of background information, Ch. 3 presented characterization techniques for measuring SPADs. Four different methods of measuring the breakdown voltage *in situ* were compared, with errors ranging from 0.1V to 0.5V depending on the measurement conditions. Three methods for measuring the afterpulsing were discussed, along with an estimation of the afterpulsing probability per unit charge. A comparison of techniques for measuring the inactive distance showed good agreement with one another.

Ch. 4 discussed the importance of ensuring that single photons are incident on the device during a timing jitter measurement is shown; a good match is shown between an experimentally measured decrease in the diffusion tail along with the predicted value. A decrease in quench time of roughly 200ps was observed when multiple photons were simultaneously incident on a single SPAD.

SPADs' insensitivity to magnetic fields with magnitudes of nearly 10T was hypothesized and experimentally measured in Ch. 5. The multiplication-assisted diffusion model has been extended to include the effects from a convective force acting on the carriers in the magnetic field, predicting no change and estimating a minimum field strength when the avalanche propagation would show statistically signficant distortion. Also shown is the increase in noise from ~1.25MeV γ-rays, with a discussion on the lifetime of PET sensors.

Motivated by the use of position sensitivity, Ch. 6 presents a diode with an electrically controllable breakdown voltage in a portion of the diode via modulation of the voltage on a polysilicon layer above the edge of the diode. The breakdown voltage, which can be modulated between 16V and 18V, is shown to be in good agreement with the theory. Additionally, the use of a negative voltage on the polysilicon, ostensibly exposing the high field region to surface-generated carriers, triggers RTS noise in the avalanche diode.

Ch. 7 presents a SPAD capable of localizing fixed-pattern noise and selectively ignoring this noise. Along with the underlying theory, an 8dB increase in SNR is shown for a diode.

Ch. 8 examines under what circumstances noise will begin to effect a SPAD-based sensor's performance when targeting positron emission tomography. The importance of fill factor is clear in this case study, especially when considering slow scintillators with high rise times. Fill factor is shown to be the dominant consideration in detector performance when LYSO with 40ns decay and 500ps rise time is coupled to a SPAD-based sensor; even if the rise time is decreased to 80ps, fill factor remains the dominant consideration under the SPAD-based sensor collects at least one third of the scintillator's output light.

9.2 Future Work

Like many theses, more questions have been raised here than answered. This section contains unanswered questions and their importance to future work.

9.2.1 Operation in Hostile Environments

In Ch. 5, an electrical technique was used to study the avalanche propgation. This technique relies on comparators which were not fully characterized in the magnetic field. Were the comparators unaffected by the field, like the TDC transistors? Do the models accurately predict the comparator's performance as a function of temperature? Is a statistically significant shift in avalanche propagation measureable as a function of magnetic field strength?

There were also several weaknesses in the noise increase from the irradiation. Can the noise increase be predicted? How will this noise increase change for larger or smaller diodes?

9.2.2 Reduction of Fixed-Position Noise

If stereophonic methods are used to achieve position sensitivity instead of sampling the quench time, how much more accurately can the seed position be localized? How uniform is afterpulsing across the SPAD's active region? Under what circumstances would switching the diode off be superior to localizing the noise intra-diode? How uniform is the breakdown voltage as a function of space?

9.2.3 Multi-Photon Distortions

How accurately can each of the three methods predict how many photons were incident on the diode? How resistant are these methods to environmental effects? Will these methods help prevent against forced triggering in quantum key distribution systems?

9.2.4 Control of the Breakdown Voltage

If the un-silicided polysilicon was used to control the region under breakdown, would the SPAD exhibit spectral sensitivity? Will variations in the poly's thickness cause yield issues? Why does the RTS noise go away as the excess bias across the diode is increased?

9.2.5 Positron Emission Tomography

How much of an advantage would using a multi-photon estimator for the gamma-ray's arrival time have over a single-photon estimator? What happens when the crosstalk in an array is non-trivial? What is the ideal number of initial timestamps to acquire for a detector given a specific amount of noise and detection efficiency in a detector? How well can the energy resolution be estimated with the initial time stamps?

9.3 Parting Remarks

The role that SPADs will play in PET-MRI seems clear; there is no contemporary competing detector that allows simultaneous acquisition of PET images with MRI compatible materials. SPADs also show promise in a plethora of single-photon applications including quantum key distribution, fluorescence lifetime imaging microscopy, and scintillator characterization. Understanding the device fundamentals creates the possibility of predicting performance, and improving detectors for these applications, as this thesis has shown.

Appendix A

Ionization: Parameters and Governing Formulas

A.1 Carrier Acceleration

In the diode itself, an electron of mass m_o will feel the Lorentz force,

$$\vec{F} = q(\vec{E} + \vec{v} \times \vec{B}). \tag{A.1}$$

With an electric field that is roughly $5 \cdot 10^5$ V per cm and no magnetic field, an electron will feel a force with a magnitude that is approximately $|\vec{F}| = q|\vec{E}| \approx (1.6 \cdot 10^-19\mathrm{C})(5 \cdot 10^5 \frac{\mathrm{V}}{\mathrm{cm}}) \approx 8\mathrm{pN}$. According to Newton's second law, this will accelerate the electron at a rate of roughly $|\vec{a}| = |\vec{F}|/m_0 \approx \frac{8\mathrm{pN}}{9.1 \cdot 10^{28}\mathrm{g}} \approx 8.8 \cdot 10^{18} \frac{\mathrm{m}}{\mathrm{s}^2}$.

At this acceleration rate, the electron would reach a relativistic velocity in less than 40 picoseconds! Furthermore, this happens over a distance $.5|\vec{a}|t^2 \approx$ 5mm — just a few millimeters. In reality, however, the saturation speed[20] will limit the top speed of the atom due to scattering in the lattice. The saturation speed in silicon, roughly $10^7 \frac{\mathrm{cm}}{\mathrm{s}}$, is reached in $\sim 10fs$. Even at electric field magnitudes that are 10% of the peak strength, free carriers are expected to reach the saturation speed within $100fs$. When considering interactions on the order of picoseconds, it is a reasonable assumption that the free carriers travel at the saturation velocity within the entire depletion region.

A.2 Deriving the Multiplication Region's Current

The derivation in this section is based on that from [29].

As described in the previous section, any free carriers in the depletion region will travel at the saturation velocity within the junction itself. The current densities for electrons and holes will be

$$\begin{aligned} \vec{j}_n(z,t) &= -q \cdot \vec{v}_s \cdot n(z,t), && \text{(A.2)} \\ \vec{j}_p(z,t) &= q \cdot (-\vec{v}_s) \cdot p(z,t). && \text{(A.3)} \end{aligned}$$

with the total current density being the sum of the two component densities,

$$\vec{j}(z,t) = \vec{j}_n(z,t) + \vec{j}_p(z,t). \tag{A.4}$$

Henceforth the vector portion of the saturation velocity and the current densities will be ignored, as per the assumption above.

The continuity equations,

$$\begin{aligned} \frac{\partial n}{\partial t} &= G_n - U_n + \frac{1}{q}\nabla \cdot \vec{j}_n, && \text{(A.5)} \\ \frac{\partial p}{\partial t} &= G_p - U_p - \frac{1}{q}\nabla \cdot \vec{j}_p, && \text{(A.6)} \end{aligned}$$

will govern the magnitude of the current during the build-up phase[20]. The carrier generation rates, G, will be assumed to be governed completely by impact ionization, with a rate of $\overline{\alpha} \cdot (n(z,t) + p(z,t)))$. Any carrier generation by light, for example, will need to be governed by initial conditions. The recombination rates, U, will be assumed to be zero.

Combining (A.3) and (A.6),

$$\begin{aligned} \frac{\partial p(z,t)}{\partial t} &= G_p - U_p - \frac{1}{q}\nabla \cdot J_p, \\ &= \overline{\alpha}|\vec{v}_s|(n(z,t) + p(z,t)) - \frac{1}{q}\nabla \cdot (q \cdot (-\vec{v}_s) \cdot p(z,t)), \\ &= \overline{\alpha}|\vec{v}_s|(n(z,t) + p(z,t)) + |\vec{v}_s|\nabla \cdot p(z,t). && \text{(A.7)} \end{aligned}$$

Similiarly for n,

$$\frac{\partial n(z,t)}{\partial t} = \overline{\alpha}|\vec{v}_s|(n(z,t) + p(z,t)) - |\vec{v}_s|\nabla \cdot n(z,t). \tag{A.8}$$

Summing (A.7) and (A.8), and simplifying with the carrier concentration, a sum of the electron and hole concentrations $p(z,t)+n(z,t)=c(z,t)$, gives

$$\begin{aligned}
\frac{\partial[n(z,t)+p(z,t)]}{\partial t} &= 2\overline{\alpha}|\vec{v}_s|(n(z,t)+p(z,t))+ \\
&\quad |\vec{v}_s|\nabla\cdot[p(z,t)-n(z,t)], \qquad \text{(A.9)} \\
\frac{\partial c(z,t)}{\partial t} &= 2\overline{\alpha}|\vec{v}_s|c(z,t)+|\vec{v}_s|\nabla\cdot[p(z,t)-n(z,t)]. \qquad \text{(A.10)}
\end{aligned}$$

This equation can be integrated over the multiplication region and then divided by the multiplication region's width to yield the average carrier concentration solely as a function of time. The initial boundary conditions will be that there are no electrons on the p+ side of the junction and no holes on the deep n-well side of the multiplication — that $n(z_0,t)=p(z_m,t)=0$. Additionally, by Kirchoff's current law, the hole carrier concentration at the p+ edge of the multiplication region must equal the electron concentration at the n edge of the multiplication region, and both of these quantities will be equal to the mean carrier concentration within the diode, $n(z_m,t)=p(z_0,t)=c(t)$. Thus,

$$\begin{aligned}
\int_{z_m}\frac{\partial c(z,t)}{\partial t}dz &= \int_{z_m}\left(2\overline{\alpha}|\vec{v}_s|c(z,t)+|\vec{v}_s|\nabla\cdot[p(z,t)-n(z,t)]\right)dz, \\
z_m\frac{\partial c(t)}{\partial t} &= 2z_m\overline{\alpha}|\vec{v}_s|c(t)+|\vec{v}_s|\int_{z_m}\left(\nabla\cdot[p(z,t)-n(z,t)]\right)dz, \\
z_m\frac{\partial c(t)}{\partial t} &= 2z_m\overline{\alpha}|\vec{v}_s|c(t)+|\vec{v}_s|\left([p(z,t)-n(z,t)]_{z=z_0}^{z=z_m}\right), \\
z_m\frac{\partial c(t)}{\partial t} &= 2z_m\overline{\alpha}|\vec{v}_s|c(t)+ \\
&\quad |\vec{v}_s|\left([p(z_m,t)-n(z_m,t)]-[p(z_0,t)-n(z_0,t)]\right), \\
z_m\frac{\partial c(t)}{\partial t} &= 2z_m\overline{\alpha}|\vec{v}_s|c(t)+|\vec{v}_s|\left([0-c(t)]-[c(t)-0]\right), \\
z_m\frac{\partial c(t)}{\partial t} &= 2z_m\overline{\alpha}|\vec{v}_s|c(t)-2|\vec{v}_s|c(t), \\
\frac{\partial c(t)}{\partial t} &= 2\overline{\alpha}|\vec{v}_s|c(t)-2|\vec{v}_s|c(t)/z_m. \qquad \text{(A.11)}
\end{aligned}$$

This expression can be simplified if the transit time for a carrier across the multiplication region, $t_m = z_m/v_s$, is considered as a constant,

$$\begin{aligned}
\frac{\partial c(t)}{\partial t} &= 2\overline{\alpha}|\vec{v}_s|c(t) - 2|\vec{v}_s|c(t)/z_m, && \text{(A.12)}\\
\frac{\partial c(t)}{\partial t} &= 2\overline{\alpha}z_m c(t)/t_m - 2c(t)/t_m, && \text{(A.13)}\\
\frac{\partial c(t)}{\partial t} &= c(t)\left(2\overline{\alpha}z_m/t_m - 2/t_m\right), && \text{(A.14)}\\
\frac{\partial c(t)}{\partial t} &= c(t)\left(\frac{1}{t_m/(2\overline{\alpha}z_m)} - \frac{1}{(t_m/2)}\right), && \text{(A.15)}\\
\frac{\partial c(t)}{\partial t} &= c(t)\left(\frac{1}{\tau_p} - \frac{1}{\tau_n}\right). && \text{(A.16)}
\end{aligned}$$

The coefficient 2 appears in τ_p because each ionization creates two carriers, and in τ_n because the carriers exit the diode via two positions.

A.3 Estimating Mean Ionization

$\overline{\alpha}$ scales with the electric field magnitude as:

$$\alpha(|\vec{E}|) = A\exp\left(-(a/|\vec{E}|)^m\right), \qquad \text{(A.17)}$$

where A, a, and m are constants dependent on the material, being roughly 10^7 per cm and 1, respectively[120]. Because it is impractical during a simulation with a large number of element diodes to calculate the ionization rate for a set of points in each diode and then take the average, a function will be fit to this curve, and then the average ionization rate can be easily estimated.

In order to estimate a, the magnitude of the electric field must be known. Using the relation between the peak electric field magnitude and the depletion region width [20],

$$\begin{aligned}
|\vec{E}| &= q\mathrm{N_d}W/\epsilon_s, && \text{(A.18)}\\
&\approx (1.6\cdot 10^{-19}\ \mathrm{C})(5\cdot 10^{16}\ \mathrm{cm}^{-3})\ (720\ \mathrm{nm})/(1.0\cdot 10^{-10}\ \mathrm{F/m}),\\
&\approx 4.9\cdot 10^{5}\ \mathrm{V/cm},
\end{aligned}$$

gives the peak electric field in an abrupt one-sided junction. As described in Chapter 2 the multiplication region of an abrupt one-sided junction composes roughly the third the depletion region with the highest field strength. If the peak field is roughly $5.4\cdot 10^5$ V/cm, the multiplication region will cover regions with fields varying from 3.6-5.4$\cdot 10^5$ V/cm, with an average field strength of roughly $4.5 \cdot 10^5$ V/cm. As (2.3) describes, the depletion region width will vary as the square root of the applied voltage. If the excess bias varies from -0.5V to +4.0V, the width of the depletion region will vary from about 710 nm to 800 nm, causing the average electric field of the multiplication region to range 4.5-5.0$\cdot 10^5$ V/cm.

Thus the average ionization coefficient as a function of the applied voltage must solve the equation

$$\overline{\alpha}(\mathrm{V_{op}}) = \frac{1}{z_m} \int_{z_{1/3}}^{z_0} \left(A \cdot \exp\left(-(a/|\vec{E}(z)|)^m \right) \right) dz, \tag{A.19}$$

Using (A.18), along with the knowledge that $|\vec{E}(z)|$ varies linearly from 2/3 of the peak value at z_0 to the peak value at z_1, the integral can be rewritten as,

$$\overline{\alpha}(\mathrm{V_{op}}) = \frac{1}{z_m} \int_{z_{1/3}}^{z_0} \left(A \cdot \exp\left(-(a/|\vec{E}(z)|)^m \right) \right) dz, \tag{A.20}$$

and as per (2.10),

$$\begin{aligned} \frac{1}{z_m} &= \frac{1}{z_m} \int_{z_{1/3}}^{z_0} \left(A \cdot \exp\left(-(a/|\vec{E}(z)|)^m \right) \right) dz, && \text{(A.21)} \\ 1 &= \int_{z_{1/3}}^{z_0} \left(A \cdot \exp\left(-(a/|\vec{E}(z)|)^m \right) \right) dz, && \text{(A.22)} \end{aligned}$$

Use of a binary search allows a to be estimated at this point — a is approximately 2.5 MV per cm.

Fig. A.1 shows how α varies with $|\vec{E}|$, with the value of a from above. Additionally, the mean value of this curve, averaged over a window of 2E5V/cm is also show, along with a fit to $\overline{\alpha}$ based on the $\overline{\alpha}$ value at $\mathrm{V_{eb}}$, which causes $\overline{\alpha}$'s error in the region of interest to be less than 7%. For this reason, the mean ionization coefficients within the excess bias range can be estimated by (A.19), with the a being given by the value at the breakdown voltage.

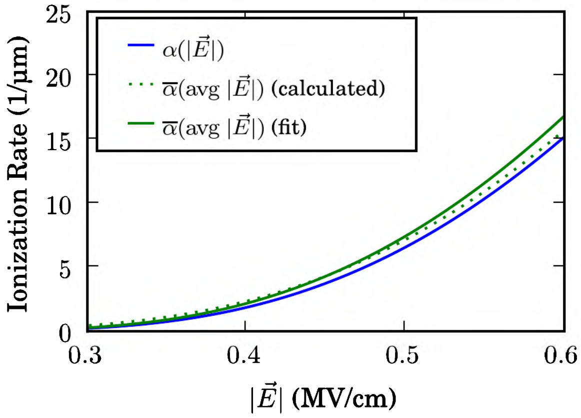


Figure A.1: **Extracted Ionization Rate Vs. Electric Field** with the average ionization rate is fit to be 1/(254 nm) at an average field magnitude of 0.5 MV/cm. As can be seen from the graph, the fit values match the average values in the region of interest, showing less than a 7% error.

Appendix B

The Convective-Diffusion Equation's Vector Solution

This chapter derives the analytical solution to the convective-diffusion equation using vector notation in cartesian coordinates, based on results for the single dimension case. The vector $\vec{r}$ signifies $x\hat{x} + y\hat{y}$, conveying cartesian, not polar, coordinates.

Important identities are[121]

$$\nabla^2(eu) = u\nabla^2(e) + e\nabla^2(u) + 2(\nabla e) \cdot (\nabla u), \tag{B.1}$$

$$\nabla(eu) = e\nabla(u) + u\nabla(e). \tag{B.2}$$

The diffusion equation's solution for $t > 0$ in two dimensions (in other dimensions the normalization term will vary), assuming boundary conditions with a delta function at the origin at time $t = 0$, is

$$u(\vec{r}, t) = \frac{1}{4\pi Dt} \exp\left(-\frac{|\vec{r}|^2}{4Dt}\right). \tag{B.3}$$

Here forward, unless noted, it is assumed that $c = 0$ when $t < 0$, $c(\vec{r}, 0) = 0$ when $|\vec{r}| > 0$ and $c(\vec{r}, 0) = 1$ when $|\vec{r}| = 0$.

Some substitutions are necessary to simplify the equation for these equations in a non-vector format ([96]):

$$\frac{\partial c}{\partial t} = D\nabla^2 c - \vec{v}\cdot\nabla c + \frac{c}{\tau}, \tag{B.4}$$

$$c(x,y,t) = \exp(\gamma t + \vec{\lambda}\cdot\vec{r})u(x,y,t), \tag{B.5}$$

$$e(x,y,t) = \exp(\gamma t + \vec{\lambda}\cdot\vec{r}), \tag{B.6}$$

$$c(x,y,t) = e(x,y,t)\cdot u(x,y,t), \tag{B.7}$$

$$\vec{\lambda} = \frac{\vec{v}}{2D}, \tag{B.8}$$

$$\gamma = 1/\tau - \frac{|\vec{v}|^2}{4D}. \tag{B.9}$$

These are the identities for simplifying e expressions:

$$\nabla e = \vec{\lambda} e, \tag{B.10}$$

$$\nabla^2 e = |\vec{\lambda}|^2 e, \tag{B.11}$$

$$\frac{\partial e}{\partial t} = \gamma e, \tag{B.12}$$

$$|\vec{\lambda}|^2 = \frac{|\vec{v}|^2}{4D^2}. \tag{B.13}$$

The full expansions, using basic identities, are

$$\frac{\partial c}{\partial t} = D\nabla^2 c - \vec{v}\cdot\nabla c + \frac{c}{\tau}, \tag{B.14}$$

$$\frac{\partial (eu)}{\partial t} = D\nabla^2 (eu) - \vec{v}\cdot\nabla (eu) + \frac{eu}{\tau}, \tag{B.15}$$

$$\begin{aligned} e\frac{\partial u}{\partial t} + u\frac{\partial e}{\partial t} &= eD\nabla^2 u + uD\nabla^2 e + \\ &2D(\nabla e)\cdot(\nabla u) - \vec{v}\cdot(e\nabla u + u\nabla e) + \frac{eu}{\tau}. \end{aligned} \tag{B.16}$$

Using (B.10), (B.11), and (B.12), the expression becomes

$$\begin{aligned} e\frac{\partial u}{\partial t} + \gamma ue =& eD\nabla^2 u + uD|\vec{\lambda}|^2 e + \\ &2D\vec{\lambda}e\cdot\nabla u - e\vec{v}\cdot\nabla u - u\vec{v}\cdot\vec{\lambda}e + \frac{eu}{\tau} \end{aligned} \tag{B.17}$$

Replacing $\vec{\lambda}$ in one term, and cancelling with another term, this simplifies to

$$e\frac{\partial u}{\partial t} + \gamma ue = eD\nabla^2 u + uD|\vec{\lambda}|^2 e + \left(2D\frac{\vec{v}}{2D}e \cdot \nabla u - e\vec{v} \cdot \nabla u\right) - u\vec{v} \cdot \vec{\lambda}e + \frac{eu}{\tau}, \tag{B.18}$$

$$e\frac{\partial u}{\partial t} + \gamma ue = eD\nabla^2 u + uD|\vec{\lambda}|^2 e - u\vec{v} \cdot \vec{\lambda}e + \frac{eu}{\tau}. \tag{B.19}$$

Shifting γue, replacing γ, expanding and cancelling a term yields

$$e\frac{\partial u}{\partial t} = -\gamma ue + eD\nabla^2 u + uD|\vec{\lambda}|^2 e - u\vec{v} \cdot \vec{\lambda}e + \frac{eu}{\tau}, \tag{B.20}$$

$$e\frac{\partial u}{\partial t} = -\left(1/\tau - \frac{|\vec{v}|^2}{4D}\right) ue + eD\nabla^2 u + uD|\vec{\lambda}|^2 e - u\vec{v} \cdot \vec{\lambda}e + \frac{eu}{\tau}, \tag{B.21}$$

$$e\frac{\partial u}{\partial t} = \left(\frac{eu}{\tau} - \frac{eu}{\tau}\right) + \frac{|\vec{v}|^2}{4D}eu + eD\nabla^2 u + uD|\vec{\lambda}|^2 e - u\vec{v} \cdot \vec{\lambda}e, \tag{B.22}$$

$$e\frac{\partial u}{\partial t} = \frac{|\vec{v}|^2}{4D}eu + eD\nabla^2 u + uD|\vec{\lambda}|^2 e - u\vec{v} \cdot \vec{\lambda}e. \tag{B.23}$$

Replacing the λ terms and cancelling three terms finally gives

$$e\frac{\partial u}{\partial t} = \frac{|\vec{v}|^2}{4D}eu + eD\nabla^2 u + uD\frac{|\vec{v}|^2}{4D^2}e - u\vec{v} \cdot \frac{\vec{v}}{2D}e, \tag{B.24}$$

$$e\frac{\partial u}{\partial t} = eD\nabla^2 u + \left(\frac{|\vec{v}|^2}{4D}eu + \frac{|\vec{v}|^2}{4D}eu - \frac{|\vec{v}|^2}{2D}eu\right), \tag{B.25}$$

$$e\frac{\partial u}{\partial t} = eD\nabla^2 u. \tag{B.26}$$

$$\tag{B.27}$$

Since $e \neq 0$,

$$\frac{\partial u}{\partial t} = D\nabla^2 u. \tag{B.28}$$

Since u is known, an expression for c follows as

$$
\begin{aligned}
c &= eu, && \text{(B.29)}\\
&= \exp(\gamma t + \vec{\lambda}\cdot\vec{r}) \cdot \frac{1}{4\pi Dt}\exp\left(-\frac{|\vec{r}|^2}{4Dt}\right), && \text{(B.30)}\\
&= \frac{1}{4\pi Dt}\exp\left(\gamma t + \vec{\lambda}\cdot\vec{r} - \frac{|\vec{r}|^2}{4Dt}\right), && \text{(B.31)}\\
&= \frac{1}{4\pi Dt}\exp\left((1/\tau - \frac{|\vec{v}|^2}{4D})t + \frac{\vec{v}}{2D}\cdot\vec{r} - \frac{|\vec{r}|^2}{4Dt}\right), && \text{(B.32)}\\
&= \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(-\frac{t|\vec{v}|^2}{4D} + \frac{2\vec{v}\cdot\vec{r}}{4D} - \frac{|\vec{r}|^2}{4Dt}\right), && \text{(B.33)}\\
&= \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(-\frac{t^2|\vec{v}|^2}{4Dt} + \frac{2t\vec{v}\cdot\vec{r}}{4Dt} - \frac{|\vec{r}|^2}{4Dt}\right), && \text{(B.34)}\\
&= \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(\frac{-1}{4Dt}\left(t^2|\vec{v}|^2 - 2t\vec{v}\cdot\vec{r} + |\vec{r}|^2\right)\right), && \text{(B.35)}\\
&= \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(\frac{-1}{4Dt}\left(|\vec{r} - t\vec{v}|^2\right)\right), && \text{(B.36)}\\
&= \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(\frac{-|\vec{r} - t\vec{v}|^2}{4Dt}\right) && \text{(B.37)}
\end{aligned}
$$

When the originating impulse function for the diffusion equation is at $\vec{r}_0$ rather than the origin, $\vec{r}$ should be replaced with $\vec{r} - \vec{r}_0$, and the full equation is

$$c = \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(\frac{-|\vec{r} - \vec{r}_0 - t\vec{v}|^2}{4Dt}\right) \qquad \text{(B.38)}$$

Similar to the derivation from [52], for a low carrier threshold, c_l, the outer edge of the avalanche at time t will be at

$$
\begin{aligned}
c_l &= \frac{\exp(t/\tau)}{4\pi Dt}\exp\left(\frac{-|\vec{r} - \vec{r}_0 - t\vec{v}|^2}{4Dt}\right), && \text{(B.39)}\\
4\pi Dtc_l &= \exp\left(t/\tau - \frac{|\vec{r} - \vec{r}_0 - t\vec{v}|^2}{4Dt}\right), && \text{(B.40)}\\
\ln(4\pi Dtc_l) &= t/\tau - \frac{|\vec{r} - \vec{r}_0 - t\vec{v}|^2}{4Dt}, && \text{(B.41)}\\
\frac{|\vec{r} - \vec{r}_0 - t\vec{v}|^2}{4Dt} &= t/\tau - \ln(4\pi Dtc_l), && \text{(B.42)}\\
|\vec{r} - \vec{r}_0 - t\vec{v}|^2 &= 4Dt\cdot(t/\tau - \ln(4\pi Dtc_l)), && \text{(B.43)}\\
|\vec{r} - \vec{r}_0 - t\vec{v}| &= \sqrt{4Dt\cdot(t/\tau - \ln(4\pi Dtc_l))} && \text{(B.44)}
\end{aligned}
$$

When $t >> 0$ and c_l is small, $(t/\tau - \ln(4\pi Dtc_l)) \approx t/\tau$, giving the spreading speed of the avalanche,

$$|\vec{r} - (\vec{r}_0 + t\vec{v})\,| = 2t\sqrt{D/\tau} \tag{B.45}$$

implying that avalanche still spreads at a speed $2\sqrt{D/\tau}$, but does so from the moving point $\vec{r}_0 + t\vec{v}$. For the more specific case $\vec{v} = 0$, c reduces to the previous solution, as does the derivation of the avalanche spread.

Appendix C

TDC Terminology and Characterization

C.1 Overview and Terminology

A time-to-digital converter (TDC) is an electronic component which measures the time between two events, and outputs a digital representation. This chapter discusses the characterization of TDCs using a density test and uniform time interval generators (UTIGs). The discussion in this section will be limited to TDCs that output a code that is linear with the input time difference.

Usually, a TDC will receive both a start signal and a stop signal, and output the time difference between the signal edges, either rising or falling. Some types of TDCs may take only a single signal, and output the width of the digital pulse. A full description of TDCs is beyond the scope of this text, but information about the characterization is included, since these devices are so critical for measuring SPAD performance.

Due to the interface similarity with ADCs, TDCs share many of the same terms for the figures of merit: input range, DNL, INL, and resolution. The terms LSB duration, mean bin duration, and LSB all refer to the step difference in output code. The single-shot jitter, sometimes called the jitter or time uncertainty, captures the error when the same input time difference is given to the TDC. The input range is the range of input time differences that can be given to the TDC. The differential non-linearity, or DNL, is the difference between an actual output code's duration and the LSB duration. The integral non-linearity, or INL, is an integration of the DNL. The TDC offset, which is the difference between the input time difference and the time difference implied by the output code, will be a combination of system delay and the INL. In this text, the term resolution will not be used, since some texts use resolution for LSB duration, whereas other texts use resolution to refer to the jitter.

Each output code has its own DNL and INL. In CMOS TDCs, DNL is often caused by variations between the transistors used to create the TDCs. For example, some transistors that propagate the start signal more quickly than expected might cause a shorter bin duration, and a negative DNL.

To give an example for these terms, if a TDC with an LSB duration of 1ps outputs code 100 when receiving an input time difference of 80ps, then code 101 should be output when receiving an input time difference of 81ps. The TDC's mean output code of 80 would imply an input time difference of 80ps, not 100ps, so the TDC's offset is 20ps at this input time difference. If this offset is constant throughout the entire input range, then this offset would be considered system-level delay, and would not be reflected in the INL. If the offset varies across the input range, then a portion of the offset would be reflected in the INL. If an input time difference of 100ps is constantly given to the TDC, but the TDC outputs code 80 with $p = 0.5$, code 79 with $p = 0.25$, and code 81 with $p = 0.25$, then the TDC's single-shot jitter's expected standard deviation is $\sqrt{1/2}$LSB, or roughly 0.7LSB RMS.

More information on various TDC architectures can be found in [122].

C.2 Characterization Using Density Tests

A density test is often used to find the DNL and INL over a TDC's entire range, or just a portion of this range[123]. In such a test, a UTIG provides the TDC with an input time difference that is equally likely to occur for any value in the input range. The UTIG is normally created by coupling a probabilistic element, such as a PMT or a SPAD, to one TDC input and a reference clock whose period is the input range to the other TDC input. Such a test will introduce probabilistic uncertainty in the resulting measurement, though this uncertainty can be made arbitrarily small by increasing the number of samples. The exact uncertainties are detailed in this section.

It is important to note that a density test cannot measure the single-shot jitter. An input time difference of fixed duration is most often used to acquire single-shot jitter at several points in the TDC's range. It should also be noted that the single-shot jitter can be seen as inherent to the TDC — i.e. impossible to remove — and any distortion to the characterization will be identical to distortions when using the TDC in an actual situation.

In order to use a density test, the input range of the TDC must be known *a priori*, or must be found from the density test data. As a crystal oscillator or other known time reference is normally used by the TDC for the start or stop signal, fixing the input range to be the clock reference's period, this knowledge is rarely a problem. In many experimental setups, the reference clock is not only used for characterization, but also in any actual measurements — if this is the case, any jitter in the reference clock will contribute to the single-shot jitter.

C.2.1 Density Test Statistics

Assume a b-bit TDC with $c = 2^b$ output codes and input range m is connected to a UTIG that generates time intervals uniformly in the range $(0, m)$. The LSB duration of this TDC will be the input range divided by the number of output codes, or m/c. Take n samples from the TDC. Let s_i be the number of samples that are output code x, with code x having a DNL of d_x measured as $\hat{d}_x$ and an INL of i_x measured as $\hat{i}_x$. In this setup, the DNL of each sample is measured to be $\hat{d}_x = \frac{s_x - E[s_x]}{E[s_x]}$ LSB. Since all of the samples have an equal expected value, $E[s_x] = n/c$, the DNL measurement can be simplified to $\hat{d}_x = s_x/(n/c) - 1$ LSB.

Even if the TDC is ideal, there will be correlated shot noise in a realization of the density test. For an ideal TDC, each sample has a $1/c$ probability of being a specific code, with the resulting single sample PDF governed by a Bernoulli with $p = 1/c$. Repeated sampling will create a binomial distribution, with mean n/c and variance $n(1/c)(1 - 1/c)$. $\hat{d}_x$'s variance will be

$$\begin{aligned}
\hat{d}_x &= s_x/(n/c) - 1 \text{ LSB}, &\text{(C.1)}\\
\operatorname{var}\left(\hat{d}_x\right) &= \operatorname{var}(s_x/(n/c) - 1) \text{ LSB}^2, &\text{(C.2)}\\
&= c^2/n^2 \cdot \operatorname{var}(s_x) \text{ LSB}^2, &\text{(C.3)}\\
&= c^2/n^2 \cdot n(1/c)(1 - 1/c) \text{ LSB}^2, &\text{(C.4)}\\
&= c^2/n^2 \cdot n(1/c)((c-1)/c) \text{ LSB}^2, &\text{(C.5)}\\
&= (c-1)/n \text{ LSB}^2, &\text{(C.6)}
\end{aligned}$$

with the standard deviation of $\hat{d}_x$ being $\sqrt{(c-1)/n}$ LSB.

The variance in the INL's measurement error will be the sum of the variances of prior DNL measurement errors, $\operatorname{var}(\hat{i}_y) = \operatorname{var}(\sum_{x=0}^{y} \hat{d}_x)$. The random variables $\hat{d}_x$ are not independent because the s_x variables are not independent, and hence their variance sums cannot be directly separated. Instead the variables' covariances, which are negative, must be considered. The negative covariance is easily observed when $n = 1$. Since one of the variables must be zero during the same trial, $E[s_x s_y] = 0$, and thus

$$\begin{aligned} \text{Cov}(s_x, s_y) &= E[s_x s_y] - E[s_x]E[s_y], && \text{(C.7)} \\ &= 0 - (1/c)^2, \text{ when } n{=}1, && \text{(C.8)} \end{aligned}$$

with (C.7) found in any probability textbook[69]. Splitting the calculation into n independent trials, each trial having covariance $1/c^2$, shows that $\text{Cov}(s_x, s_y)$ is n/c^2, with the $\hat{d}_x$'s covariance being simply $c^2/n^2 \cdot \text{Cov}(s_y, s_z) = 1/n$.

Hence,

$$\begin{aligned} \text{var}(\hat{i}_z) &= \text{var}\left(\sum_{y=1}^{z} \hat{d}_y\right) && \text{(C.9)} \\ &= \sum_{x=1}^{z} \text{var}(\hat{d}_x) + 2\sum_{y=1}^{z}\sum_{x=1}^{y-1} \text{Cov}(\hat{d}_x, \hat{d}_y) && \text{(C.10)} \\ &= z(c-1)/n - (z-1)z(1/n) \text{ LSB}^2, && \text{(C.11)} \\ &= \frac{z}{n}(c-z) \text{ LSB}^2, && \text{(C.12)} \end{aligned}$$

implying that the standard deviation of $\hat{i}_z$'s measurement error will scale as $\sqrt{z}$ times the DNL's measurement error for small z, peak at value $c^2/(4n)$ when $z = c/2$, and then decrease back to zero when $z = c$. Simplification from (C.10) to (C.11) uses the identity $1 + 2 + ... + z = (z-1)(z/2)$ by way of

$$\begin{aligned} \left(\sum_{y=1}^{z}\sum_{x=1}^{y}(1/n)\right) &= (1 + 2 + ... + z)(1/n), && \text{(C.13)} \\ &= (z-1)(z/2)(1/n). && \text{(C.14)} \end{aligned}$$

It may be somewhat surprising that $\hat{i}_c$ is zero, but the result logically follows when it is noted that the UTIG's range has been assumed to match the TDC's range exactly. Because the INL will exhibit symmetry — the last code could just as easily be seen as first code when the input range is matched — $i_c = 0$ and the measurement error reflects this. Non-ideals UTIGs will be considered later.

For a non-ideal TDC with non-zero DNL values, the analysis is similar, but the PDFs for $\hat{d}_x$ will no longer be uniform, and hence the simplification used for changing (C.10) into (C.11) is no longer valid. Additionally the covariances will now be different. An upper bound is easily placed on the INL measurement error if the worst DNL values are known. Take the largest

DNL value to be $\overline{w} - 1$ and the smallest to be $\underline{w} - 1$, implying a probability of $\overline{w}/c$ and $\underline{w}/c$ for a per-trial probability of sampling these worst codes. The worst single-case variance in the "largest" code's sample count will be $n(\overline{w}/c)(1 - \overline{w}/c)$, which increases from 0 at $\overline{w} = 0$ to a maximum of 0.25 at a value of $\overline{w} = 0.5c$. So long as the code with the worst DNL is expected to have fewer than one half of the total samples (a reasonable assumption), its error variance will be an upper bound on the variances of all errors. The variance for measurement of a DNL value will now have an upper bound of $(c\overline{w} - \overline{w}^2)/n$. The per-trial s covariances will be at least the covariance from two "smallest" codes, or $(\underline{w}/c)^2$, instead of $(1/c)^2$, with the n trial DNL covariances being at least $\underline{w}^2/n$. Thus the variance in $\hat{i}_z$ will be bound by

$$\begin{aligned}
\mathrm{var}(\hat{i}_z) &= \mathrm{var}\left(\sum_{y=1}^{z} \hat{d}_y\right) && \text{(C.15)}\\
&= \sum_{y=1}^{z} \mathrm{var}(\hat{d}_y) + 2\sum_{y=1}^{z}\sum_{x=1}^{y-1} \mathrm{Cov}(\hat{d}_x, \hat{d}_y) && \text{(C.16)}\\
&< z(c\overline{w} - \overline{w}^2)/n - 2(z-1)(z/2)(\underline{w}^2/n)\ \mathrm{LSB}^2, && \text{(C.17)}\\
&< \frac{z}{n}\left((c\overline{w} - \overline{w}^2) - \underline{w}^2(z-1)\right)\ \mathrm{LSB}^2, && \text{(C.18)}
\end{aligned}$$

being roughly $\sqrt{\overline{w}}$ times the case for the ideal TDC when i is small and $\overline{w} << c$ — i.e. when the worst case DNL is much smaller than the number of TDC bins. Because the UTIG's range is matched to the TDC range, $\hat{i}_c$ should be 0 to reflect $i_c = 0$, as in (C.12), but is not in (C.18) because the underestimation of the covariance causes too little to be subtracted as z increases. However, symmetry can be exploited by noting that, when taking code c's INL value to start at zero and summing with an index going from c to z, $\mathrm{var}(\hat{i}_z)$ will increase from $\hat{i}_c = 0$ to a maximum at $\hat{i}_1$. Thus, only when the UTIG matches the TDC input range, the error can be constrained to be

$$\begin{aligned}
\mathrm{var}(\hat{i}_z) &< \frac{z'}{n}\left((c\overline{w} - \overline{w}^2) - \underline{w}^2(z'-1)\right)\ \mathrm{LSB}^2,\\
z' &= \min(z, c+1-z) && \text{(C.19)}
\end{aligned}$$

with an identical from to (C.18) except that z has been replaced by z', which is the smaller of the index z's difference with the two edges. Like the case for the ideal TDC, the variance in the INL measurement still peaks when $z = c/2$, but the value is now $\frac{c}{2n}\left((c\overline{w} - \overline{w}^2) - \underline{w}^2(c-2)/2\right)$. Note that, as per their definition or other constraints, $1 \leq \overline{w} \leq c/2$ and $0 \leq \underline{w} \leq 1$. An ideal TDC should have the same behavior when run through the analysis for a non-ideal TDC. For an ideal TDC, all DNL values are 0, $\overline{w}-1=0 \Rightarrow \overline{w}=1$ and $\underline{w}-1=0 \Rightarrow \underline{w}=1$, with (C.18) reducing to (C.12). So the $\hat{i}_z$ variances for the ideal TDC are identical with the two analyses, as expected.

If the upper bound provided by (C.19) is still too coarse, a full expansion of (C.16) is possible. Let the actual DNL value of code x be w_x-1. In this case, the per-trial probability of sampling code x is w_x/c. Because the per-trial probabilities must add to one, $c=\sum_{x=1}^{c}(w_x)$. The variance in $\hat{d}_x$ will be $(cw_x-w_x^2)/n$. The covariance between s_x and s_y is $-w_xw_y/c^2$, with $\mathrm{Cov}(\hat{d}_x,\hat{d}_y)=-w_xw_y/n$. The full expression for the variance of $\hat{i}_z$ is

$$\begin{aligned}
\mathrm{var}(\hat{i}_z) &= \mathrm{var}\left(\sum_{y=1}^{z}\hat{d}_y\right), && \text{(C.20)}\\
&= \sum_{y=1}^{z}\mathrm{var}(\hat{d}_y)+2\sum_{y=1}^{z}\sum_{x=1}^{y-1}\mathrm{Cov}(\hat{d}_x,\hat{d}_y), && \text{(C.21)}\\
&= \sum_{y=1}^{z}\left(\frac{cw_y-w_y^2}{n}\right)-2\sum_{y=1}^{z}\sum_{x=1}^{y-1}\left(\frac{w_xw_y}{n}\right), && \text{(C.22)}\\
&= \frac{1}{n}\sum_{y=1}^{z}\left(cw_y-w_y^2-2w_y\sum_{x=1}^{y-1}(w_x)\right). && \text{(C.23)}
\end{aligned}$$

By symmetry, when $z=c$ the error should be zero. This can be checked by noting,

$$\begin{aligned}
\mathrm{var}(\hat{i}_c) &= \frac{1}{n}\sum_{y=1}^{c}\left(cw_y-w_y^2-2w_y\sum_{x=1}^{y-1}(w_x)\right), && \text{(C.24)}\\
&= \frac{1}{n}\sum_{y=1}^{c}(cw_y)-\frac{1}{n}\sum_{y=1}^{c}\left(w_y^2+2w_y\sum_{x=1}^{y-1}(w_x)\right), && \text{(C.25)}\\
&= \frac{1}{n}c^2-\frac{1}{n}\sum_{y=1}^{c}\left(w_y^2+\sum_{x=1}^{y-1}(w_yw_x)+\sum_{x=1}^{y-1}(w_yw_x)\right), && \text{(C.26)}\\
&= \frac{1}{n}c^2-\frac{1}{n}\sum_{y=1}^{c}(w_y^2)-\frac{1}{n}\sum_{y=1}^{c}\sum_{x=1}^{y-1}(w_yw_x)-\frac{1}{n}\sum_{y=1}^{c}\sum_{x=1}^{y-1}(w_yw_x), && \text{(C.27)}\\
&= \frac{1}{n}c^2-\frac{1}{n}\sum_{y=1}^{c}(w_y^2)-\frac{1}{n}\sum_{y=1}^{c}\sum_{x=1}^{y-1}(w_yw_x)-\frac{1}{n}\sum_{x=1}^{c}\sum_{y=1}^{x-1}(w_xw_y), && \text{(C.28)}
\end{aligned}$$

$$= \frac{1}{n}c^2 - \frac{1}{n}\sum_{y=1}^{c}\sum_{x=1}^{c}(w_x w_y), \tag{C.29}$$

$$= \frac{c^2}{n} - \frac{c^2}{n}, \tag{C.30}$$

$$= 0, \tag{C.31}$$

which gives the expected result that the final variance is zero. The simplification from (C.28) to (C.29) can be visually imagined as summing $w_x w_y$ over the set of points in the square $1 \leq x \leq c, 1 \leq y \leq c$, with points on diagonal $x = y$ being the single sum, the points above the diagonal being in the first double sum, and the points below the diagonal being in the other double sum.

It should be noted that if the w values are normally distributed about 1, the worst measurement of the INL value, i.e. the measurement with highest variance, should occur in the middle of the TDC's range. Studying (C.23), this value will occur roughly when the covariance removes more from the measurement error than the variances add. Quantitatively, this is

$$\operatorname{var}(\hat{d}_z) \quad < \quad -2\sum_{y=1}^{z-1}\operatorname{Cov}(\hat{d}_y, \hat{d}_z), \tag{C.32}$$

$$c w_y - w_y^2 \quad < \quad 2 w_y \sum_{y=1}^{z-1}(w_y), \tag{C.33}$$

$$c - w_y \quad < \quad 2\sum_{y=1}^{z-1}(w_y), \tag{C.34}$$

$$c/2 - w_y/2 \quad < \quad \sum_{y=1}^{z-1}(w_y), \tag{C.35}$$

which, since $c = \sum_{x=1}^{c}(w_x)$, first occurs for roughly $z = c/2$ as expected.

C.2.2 Reference Clock Jitter

There is one case that has not been considered — what occurs if the TDC is being characterized by a UTIG created by a probabilistic source and a reference clock that has a lot of jitter, but any measurement will occur with inputs having lower jitter? Such a setup might occur, for example, if an on-chip TDC will measure time waveforms generated on-chip, but the reference clock suffers large amounts of jitter when injected from off-chip. In this case, the measured DNL values of the final codes will show distortions, and the density test might be considered valid only for the codes that are smaller than the reference clock period minus several times the jitter. A full treatment of this case is beyond the scope of this text.

C.2.3 UTIG Non-Uniformity

Oftentimes a probabilistic exponential source, such as a SPAD or PMT with a low event rate is used as a UTIG. These elements create electrical pulses whose rising edges are exponentially distributed in time, with the pulse originated created by some random phenomena such as quantum tunneling in a semiconductor. With a low event rate, there is a probability that no event is produced in the time interval, though proper reset behavior will remove any negative effects from the lack of an event. Given a probabilistic exponential source with an event rate λ, the expected time until the next electrical pulse at any point in time for such devices follows an exponential distribution, whose PDF[69] is

$$f(t) = \lambda \exp\left(-\lambda t\right), \tag{C.36}$$

if $t > 0$ and $f(t) = 0$ otherwise. If the uniformity criteria for this generator is that the probability should vary by less than ϵ across the entire time range, starting at time 0, then the event rate must meet the criteria that

$$\begin{aligned} 1 - \epsilon \cdot f(0) &< f(m), & \text{(C.37)}\\ 1 - \epsilon\lambda &< \lambda \exp\left(-\lambda m\right), & \text{(C.38)}\\ \frac{-\ln(1-\epsilon)}{m} &> \lambda, & \text{(C.39)}\\ \sim \frac{\epsilon}{m} &> \lambda, & \text{(C.40)} \end{aligned}$$

with the second line being simplified using the Taylor expansion of the natural logarithm.

The DNL distortions will sum in the INL. If the shift in the measurement of code x's DNL value is linearly approximated as $\epsilon/2 - x\epsilon/c$, which is a shift of $\epsilon/2$ for the initial DNL values and $-\epsilon/2$ in the final values, then the distortion to the INL of code z will approximately be

$$\begin{aligned} \sum_{x=1}^{z}(\epsilon/2 - x\epsilon/c) &= z\epsilon/2 - \epsilon/c\sum_{x=1}^{z}(x), & \text{(C.41)}\\ &= \epsilon\left(\frac{z}{2} - \frac{z(z-1)}{2c}\right), & \text{(C.42)}\\ & & \text{(C.43)} \end{aligned}$$

which has a maximum at $z = c/2$ of roughly $c\epsilon/8$. Should the maximum error for any INL value be ϵ', then $\epsilon = 8\epsilon'/c$, with the maximum event rate being $\sim 8\epsilon'/(mc)$.

For example, when characterizing a 16-bit, 100ns input range TDC with a density test, if a SPAD being used as a UTIG should cause DNL distortions less than 5%, then the event rate must be less than $\sim 0.05/100\text{ns} = 500\text{kHz}$. If the distortion to the INL should be less than 5%, then $2^{16}\epsilon/8 \approx 0.05 \Rightarrow \epsilon \approx 6 \cdot 10^{-6}$, with an event rate lower than $6 \cdot 10^{-6}/100\text{ns} \approx 60\text{Hz}$.

C.3 Summary

The measurement of code x's INL will have std. dev. $\sqrt{\frac{x}{n}(c-x)}$ LSB RMS given an ideal TDC with c codes measured with an n sample density test, with the error roughly scaling in quadrature for small x with the measurement error in any code's DNL, which is $\sqrt{(c-1)/n}$ LSB RMS. The worst error in any INL value, occurring for code $c/2$, will have a standard deviation of $\sim \sqrt{c^2/(4n)}$.

A non-ideal TDC with highest DNL value $(\overline{w} - 1)$ LSB and lowest DNL value $(\underline{w} - 1)$ will have a measurement error in code i's INL value of at most $\frac{c}{2n}\left((c\overline{w} - \overline{w}^2) - \underline{w}^2(c-2)/2\right)$ LSB RMS.

If an exponential source, such as a SPAD or a PMT, is to be used with a fixed reference clock as a uniform time interval generator to characterize a TDC with c codes over an input range of m, but should introduce an error no more than ϵ' to the INL, the event rate of this exponential source should be smaller than approximately $8\epsilon'/(mc)$.

The INL distortions from the density test and from the UTIG do not scale in quadrature. The density test distortion is a statistical variation effect from shot noise, whereas the UTIG distortion is an expected result from non-uniformities in the time interval generator.

Bibliography

[1] A. Einstein, "Über einen die erzeugung und verwandlung des lichtes betreffenden heuristischen gesichtspunkt," *Annalen der Physik*, vol. 322, no. 6, pp. 132–148, 1905.

[2] A. Wang, P. Gill, and A. Molnar, "An angle-sensitive CMOS imager for single-sensor 3D photography," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, Feb. 2011, pp. 412–414.

[3] G. F. Knoll, *Radiation Detection and Measurement*, 3rd ed. Wiley, 2000.

[4] F. Stuker, C. Baltes, K. Dikaiou, D. Vats, L. Carrara, E. Charbon, J. Ripoll, and M. Rudin, "Hybrid small animal imaging system combining magnetic resonance imaging with fluorescence tomography using single photon avalanche diode detectors," *Medical Imaging, IEEE Transactions on*, vol. 30, no. 6, pp. 1265–1273, June 2011.

[5] ID Quantique SA, "Random number generation using quantum physics," Tech. Rep., Apr 2010.

[6] J. Vallerga, J. McPhate, A. Tremsin, and O. Siegmund, "High-resolution UV, alpha and neutron imaging with the timepix CMOS readout," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 591, no. 1, pp. 151–154, 2008.

[7] "The Medipix Home Page." [Online]. Available: http://medipix.web.cern.ch/medipix/

[8] D. R. Schuette, R. C. Westhoff, A. H. Loomis, D. J. Young, J. S. Ciampi, B. F. Aull, and R. K. Reich, "Hybridization process for back-illuminated silicon Geiger-mode avalanche photodiode arrays," M. A. Itzler and J. C. Campbell, Eds., vol. 7681, no. 1. SPIE, 2010, p. 76810P. [Online]. Available: http://link.aip.org/link/?PSI/7681/76810P/1

[9] S. Hecht, S. Shlaer, and M. H. Pirenne, "Energy, quanta, and vision," *The Journal of General Physiology*, vol. 25, no. 6, pp. 819–840, 1942.

[10] H. Iams and B. Salzberg, "The secondary emission phototube," *Proceedings of the Institute of Radio Engineers*, vol. 23, no. 1, pp. 55–64, Jan. 1935.

[11] H. Kume, K. Koyama, K. Nakatsugawa, S. Suzuki, and D. Fatlowitz, "Ultrafast microchannel plate photomultipliers," *Appl. Opt.*, vol. 27, no. 6, pp. 1170–1178, Mar 1988.

[12] 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*, vol. 567, no. 1, pp. 230–232, 2006.

[13] A. J. Theuwissen, *Solid-State Imaging with Charge-Coupled Devices*, 1st ed. Springer, 1995.

[14] J. Hynecek, "Impactron-a new solid state image intensifier," *Electron Devices, IEEE Transactions on*, vol. 48, no. 10, pp. 2238–2241, Oct 2001.

[15] R. Dyck and G. Weckler, "Integrated arrays of silicon photodetectors for image sensing," *Electron Devices, IEEE Transactions on*, vol. 15, no. 4, pp. 196–201, Apr 1968.

[16] Y. Chen, "Low-noise cmos image sensors for radio-molecular imaging," Ph.D. dissertation, Delft University of Technology, 2012,.

[17] H.-J. Yoon, S. Itoh, and S. Kawahito, "A CMOS image sensor with in-pixel two-stage charge transfer for fluorescence lifetime imaging," *Electron Devices, IEEE Transactions on*, vol. 56, no. 2, pp. 214–221, Feb. 2009.

[18] H. W. Li, B. E. Kardynal, P. See, A. J. Shields, P. Simmonds, H. E. Beere, and D. A. Ritchie, "Quantum dot resonant tunneling diode

for telecommunication wavelength single photon detection," *Applied Physics Letters*, vol. 91, no. 7, p. 073516, 2007.

[19] G. N. Gol'tsman, O. Okunev, G. Chulkova, A. Lipatov, A. Semenov, K. Smirnov, B. Voronov, A. Dzardanov, C. Williams, and R. Sobolewski, "Picosecond superconducting single-photon optical detector," *Applied Physics Letters*, vol. 79, no. 6, pp. 705–707, 2001.

[20] S. M. Sze, *Physics of Semiconductor Devices*, 2nd ed. Wiley-Interscience, 1981.

[21] P. Seitz and A. J. P. Theuwissen, Eds., *Single-Photon Imaging*. Springer, 2011.

[22] H. Bowers, "Space-charge-induced negative resistance in avalanche diodes," *Electron Devices, IEEE Transactions on*, vol. 15, no. 6, pp. 343–350, Jun 1968.

[23] S. Sze and G. Gibbons, "Effect of junction curvature on breakdown voltage in semiconductors," *Solid-State Electronics*, vol. 9, no. 9, pp. 831–845, 1966.

[24] Synopsys, "Medici two-dimensional device simulation program user manual," 2003, version 2002.4.

[25] J. Richardson, L. Grant, and R. Henderson, "Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology," *Photonics Technology Letters, IEEE*, vol. 21, no. 14, pp. 1020–1022, July 2009.

[26] A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche diode performance," *Electronics Letters*, vol. 25, no. 13, pp. 841–843, June 1989.

[27] P. Webb and A. R. Jones, "Large area reach-through avalanche diodes for radiation monitoring," *Nuclear Science, IEEE Transactions on*, vol. 21, no. 1, pp. 151–158, Feb 1974.

[28] S. Cova, A. Longoni, and A. Andreoni, "Towards picosecond resolution with single-photon avalanche diodes," *Review of Scientific Instruments*, vol. 52, no. 3, pp. 408–412, 1981.

[29] A. Rochas, "Single photon avalanche diodes in CMOS technology," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, 2003.

[30] C. Niclass, A. Rochas, P.-A. Besse, and E. Charbon, “Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes,” *Solid-State Circuits, IEEE Journal of*, vol. 40, no. 9, pp. 1847–1854, Sep. 2005.

[31] M. A. Karami, M. Gersbach, H.-J. Yoon, and E. Charbon, “A new single-photon avalanche diode in 90nm standard cmos technology,” *Opt. Express*, vol. 18, no. 21, pp. 22 158–22 166, Oct 2010.

[32] Y. Maruyama (personal correspondence), June 2011.

[33] H. Finkelstein, M. Hsu, and S. Esener, “STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology,” *Electron Device Letters, IEEE*, vol. 27, no. 11, pp. 887–889, Nov. 2006.

[34] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, “A low-noise single-photon detector implemented in a 130 nm CMOS imaging process,” *Solid-State Electronics*, vol. 53, no. 7, pp. 803–808, 2009.

[35] R. Henderson, E. Webster, R. Walker, J. Richardson, and L. Grant, “A 3× 3, 5µm pitch, 3-transistor single photon avalanche diode array with integrated 11V bias generation in 90nm CMOS technology,” in *Electron Devices Meeting (IEDM), 2010 IEEE International*, Dec. 2010, pp. 14.2.1–14.2.4.

[36] M. Adler, V. Temple, A. Ferro, and R. Rustay, “Theory and breakdown voltage for planar devices with a single field limiting ring,” *Electron Devices, IEEE Transactions on*, vol. 24, no. 2, pp. 107–113, Feb 1977.

[37] K.-P. Brieger, W. Gerlach, and J. Pelka, “Blocking capability of planar devices with field limiting rings,” *Solid-State Electronics*, vol. 26, no. 8, pp. 739–745, 1983.

[38] W. J. Kindt, “Gieger-mode avalanche photodiode arrays,” Ph.D. dissertation, Delft University of Technology, Delft, the Netherlands, 1999.

[39] P. A. Tove, “Methods of avoiding edge effects on semiconductor diodes,” *Journal of Physics D: Applied Physics*, vol. 15, no. 4, p. 517, 1982.

[40] A. Rochas, A. Pauchard, P.-A. Besse, D. Pantic, Z. Prijic, and R. Popovic, “Low-noise silicon avalanche photodiodes fabricated in conventional cmos technologies,” *Electron Devices, IEEE Transactions on*, vol. 49, no. 3, pp. 387–394, Mar 2002.

[41] M. Fishburn, Y. Maruyama, and E. Charbon, "Reduction of fixed-position noise in position-sensitive single-photon avalanche diodes," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2354–2361, Aug. 2011.

[42] N. Faramarzpour, M. Deen, S. Shirani, and Q. Fang, "Fully integrated single photon avalanche diode detector in standard CMOS 0.18-μm technology," *Electron Devices, IEEE Transactions on*, vol. 55, no. 3, pp. 760–767, Mar. 2008.

[43] D. A. Ramirez, M. M. Hayat, G. J. Rees, X. Jiang, and M. A. Itzler, "New perspective on passively quenched single photon avalanche diodes: effect of feedback on impact ionization," *Opt. Express*, vol. 20, no. 2, pp. 1512–1529, Jan 2012.

[44] S. Cova, A. Lacaita, and G. Ripamonti, "Trapping phenomena in avalanche photodiodes on nanosecond scale," *Electron Device Letters, IEEE*, vol. 12, no. 12, pp. 685–687, Dec 1991.

[45] R. G. W. Brown, K. D. Ridley, and J. G. Rarity, "Characterization of silicon avalanche photodiodes for photon correlation measurements. 1: Passive quenching," *Appl. Opt.*, vol. 25, no. 22, pp. 4122–4126, Nov 1986.

[46] R. G. W. Brown, R. Jones, J. G. Rarity, and K. D. Ridley, "Characterization of silicon avalanche photodiodes for photon correlation measurements. 2: Active quenching," *Appl. Opt.*, vol. 26, no. 12, pp. 2383–2389, Jun 1987.

[47] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *Nuclear Science, IEEE Transactions on*, vol. 57, no. 6, pp. 3815–3826, Dec. 2010.

[48] F. Zappa, A. Lotito, A. Giudice, S. Cova, and M. Ghioni, "Monolithic active-quenching and active-reset circuit for single-photon avalanche detectors," *Solid-State Circuits, IEEE Journal of*, vol. 38, no. 7, pp. 1298–1301, July 2003.

[49] C. Niclass and M. Soga, "A miniature actively recharged single-photon detector free of afterpulsing effects with 6ns dead time in a 0.18 µm CMOS technology," in *Electron Devices Meeting (IEDM), 2010 IEEE International*, Dec. 2010, pp. 14.3.1–14.3.4.

[50] J. Richardson, R. K. Henderson, and D. Renshaw, “Dynamic quenching for single photon avalanche diode arrays,” in *Proceedings of the 2007 International Image Sensors Workshop*, June 2007.

[51] A. Spinelli and A. Lacaita, “Physics and numerical simulation of single photon avalanche diodes,” *Electron Devices, IEEE Transactions on*, vol. 44, no. 11, pp. 1931–1943, Nov. 1997.

[52] A. Spinelli, “Limits to the timing performance of single-photon avalanche diodes,” Ph.D. dissertation, Politecnico di Milano, Milano, Italy, 1995.

[53] K. Dekker and J. G. Verwer, *Stability of runge-kutta methods for stiff nonlinear differential equations.* Elsevier Science Ltd, 1984.

[54] W. Oldham, R. Samuelson, and P. Antognetti, “Triggering phenomena in avalanche diodes,” *Electron Devices, IEEE Transactions on*, vol. 19, no. 9, pp. 1056–1060, Sep. 1972.

[55] R. Wolffenbuttel, “Integrated silicon colour sensors,” Ph.D. dissertation, Delft University of Technology, Delft, Netherlands, 1988.

[56] M. E. Hoenk, P. J. Grunthaner, F. J. Grunthaner, R. W. Terhune, M. Fattahi, and H. Tseng, “Growth of a delta-doped silicon layer by molecular beam epitaxy on a charge-coupled device for reflection-limited ultraviolet quantum efficiency,” *Applied Physics Letters*, vol. 61, pp. 1084–1086, June 1992.

[57] J. A. del Alamo and R. M. Swanson, “Modelling of minority-carrier transport in heavily doped silicon emitters,” *Solid-State Electronics*, vol. 30, no. 11, pp. 1127–1136, 1987.

[58] W. Grant, “Electron and hole ionization rates in epitaxial silicon at high electric fields,” *Solid-State Electronics*, vol. 16, no. 10, pp. 1189–1203, 1973.

[59] B. N. Brockhouse, “Lattice vibrations in silicon and germanium,” *Phys. Rev. Lett.*, vol. 2, pp. 256–258, Mar 1959.

[60] R. Hull, *Properties of Crystalline Silicon.* Institution of Engineering and Technology, 1999.

[61] A. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, “On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices,” *Electron Devices, IEEE Transactions on*, vol. 40, no. 3, pp. 577–582, Mar 1993.

[62] R. H. Haitz, “Studies on optical coupling between silicon pn junctions,” *Solid-State Electronics*, vol. 8, no. 4, pp. 417–425, 1965.

[63] M. Sergio and E. Charbon, “An intra-chip electro-optical channel based on CMOS single photon detectors,” in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, Dec. 2005, pp. 822–826.

[64] C. Kurtsiefer, P. Zarda, S. Mayer, and H. Weinfurter, “The breakdown flash of silicon avalanche photodiodes-back door for eavesdropper attacks?” *Journal of Modern Optics*, vol. 48, no. 13, pp. 2039–2047, 2001.

[65] A. Gulinatti, P. Maccagnani, I. Rech, M. Ghioni, and S. Cova, “35 ps time resolution at room temperature with large area single photon avalanche diodes,” *Electronics Letters*, vol. 41, no. 5, pp. 272–274, Mar. 2005.

[66] M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti, and M. Ghioni, “Photon-timing jitter dependence on injection position in single-photon avalanche diodes,” *Quantum Electronics, IEEE Journal of*, vol. 47, no. 2, pp. 151–159, Feb. 2011.

[67] G. Ripamonti and S. Cova, “Carrier diffusion effects in the time-response of a fast photodiode,” *Solid-State Electronics*, vol. 28, no. 9, pp. 925–931, 1985.

[68] J. R. Meijlink, C. Veerappan, S. Seifert, D. Stoppa, R. Henderson, E. Charbon, and D. R. Schaart, “First measurement of scintillation photon arrival statistics using a high-granularity solid-state photosensor enabling time-stamping of up to 20,480 single photons,” in *2011 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, Oct 2011, pp. 2254–2257.

[69] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*, 1st ed. Athena Scientific, 2002.

[70] D. A. Neamen, *Semiconductor Physics and Devices*, 3rd ed. McGraw-Hill, 2003.

[71] G. Hurkx, D. Klaassen, and M. Knuvers, "A new recombination model for device simulation including tunneling," *Electron Devices, IEEE Transactions on*, vol. 39, no. 2, pp. 331–338, Feb 1992.

[72] W. Shockley and W. T. Read, "Statistics of the recombinations of holes and electrons," *Phys. Rev.*, vol. 87, pp. 835–842, Sep 1952.

[73] L. Carrara, C. Niclass, N. Scheidegger, H. Shea, and E. Charbon, "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications," in *Solid-State Circuits Conference - Digest of Technical Papers, 2009. ISSCC 2009. IEEE International*, Feb. 2009, pp. 40–41,41a.

[74] C. Veerappan, J. Richardson, R. Walker, D. Li, M. Fishburn, D. Stoppa, F. Borghetti, Y. Maruyama, M. Gersbach, R. Henderson, C. Bruschini, and E. Charbon, "Characterization of large-scale non-uniformities in a 20k tdc/spad array integrated in a 130nm cmos process," in *Solid-State Device Research Conference (ESSDERC), 2011 Proceedings of the European*, Sept. 2011, pp. 331–334.

[75] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier — principle of operation and intrinsic detector performance," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2009 IEEE*, Nov. 2009, pp. 1959–1965.

[76] I. Rech, A. Ingargiola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, and S. Cova, "Optical crosstalk in single photon avalanche diode arrays: a new complete model," *Opt. Express*, vol. 16, no. 12, pp. 8381–8394, Jun 2008.

[77] K. Iniewski, Ed., *Radiation Effects in Semiconductors*, 1st ed. CRC Press, 2011.

[78] C. L. Niclass, "Single-photon image sensors in CMOS: Picosecond resolution for three-dimensional imaging," Ph.D. dissertation, École Polytechnique Fédérale de Lausanne, 2008.

[79] D. Stoppa, D. Mosconi, L. Pancheri, and L. Gonzo, "Single-photon avalanche diode CMOS sensor for time-resolved fluorescence measurements," *Sensors Journal, IEEE*, vol. 9, no. 9, pp. 1084–1090, Sep. 2009.

[80] Cadence, "Spectre circuit simulator reference," 2003.

[81] M. Karami, L. Carrara, C. Niclass, M. Fishburn, and E. Charbon, "RTS noise characterization in single-photon avalanche diodes," *Electron Device Letters, IEEE*, vol. 31, no. 7, pp. 692–694, July 2010.

[82] H. T. Yen, S. D. Lin, and C. M. Tsai, "A simple method to characterize the afterpulsing effect in single photon avalanche photodiode," *Journal of Applied Physics*, vol. 104, no. 5, p. 054504, 2008.

[83] A. C. Berry, "The accuracy of the gaussian approximation to the sum of independent variates," *Transactions of the American Mathematical Society*, vol. 49, no. 1, pp. 122–136, 1941.

[84] I. S. Tyurin, "Refinement of the upper bounds of the constants in Lyapunov's theorem," *Russian Mathematical Surveys*, vol. 65, no. 3, p. 586, 2010.

[85] A. Zanchi, F. Zappa, and M. Ghioni, "A probe detector for defectivity assessment in p-n junctions," *Electron Devices, IEEE Transactions on*, vol. 47, no. 3, pp. 609–616, Mar 2000.

[86] F. Zappa, S. Tisa, A. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 103–112, 2007.

[87] J. Blazej, "Photon number resolving in geiger mode avalanche photodiode photon counters," *Journal of Modern Optics*, vol. 51, no. 9-10, pp. 1491–1497, 2004.

[88] B. E. Kardynal, Z. L. Yuan, and A. J. Shields, "An avalanche-photodiode-based photon-number-resolving detector," *Nature Photonics*, vol. 2, no. 7, pp. 425–428, 2008.

[89] L. Lyderson, C. Wiechers, C. Wittmann, D. Elser, J. Skaar, and V. Makarov, "Hacking commercial quantum cryptography systems by tailored bright illumination," *Nat. Photonics*, vol. 4, no. 10, pp. 686–689, Aug. 2010.

[90] P. Eraerds, M. Legré, A. Rochas, H. Zbinden, and N. Gisin, "SiPM for fast photon-counting and multiphoton detection," *Opt. Express*, vol. 15, no. 22, pp. 14 539–14 549, Oct 2007.

[91] M. Ghioni, S. Cova, F. Zappa, and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Review of Scientific Instruments*, vol. 67, no. 10, pp. 3440–3448, 1996.

[92] M. Fishburn and E. Charbon, "Environmental effects on photomultiplication propagation in silicon," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2011 IEEE*, Oct. 2011, pp. 572–574.

[93] E. Charbon and M. Fishburn, "Radiation-hardened and radiation-sensitive single-photon imagers," in *2011 International Workshop on Radiation Imaging Detectors Book of Abstracts*, Jul. 2011, p. 44.

[94] O. Heaviside, "On the electromagnetic effects due to the motion of electrification through a dielectric," *Philosophical Magazine*, vol. 5, pp. 324–339, 1889.

[95] R. F. Probstein, *Physicochemical Hydrodynamics*, 1st ed. Butterworths, 1989.

[96] A. D. Polyanin, *Handbook of Linear Partial Differential Equations for Engineers and Scientists.* Chapman and Hall, 2001.

[97] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions," *Applied Physics Letters*, vol. 57, no. 5, pp. 489–491, Jul. 1990.

[98] R. Brunetti, C. Jacoboni, F. Nava, L. Reggiani, G. Bosman, and R. J. J. Zijlstra, "Diffusion coefficient of electrons in silicon," *Journal of Applied Physics*, vol. 52, no. 11, pp. 6713–6722, 1981.

[99] C. Jacoboni, C. Canali, G. Ottaviani, and A. A. Quaranta, "A review of some charge transport properties of silicon," *Solid-State Electronics*, vol. 20, no. 2, pp. 77–89, 1977.

[100] G. Brix, U. Lechel, G. Glatting, S. I. Ziegler, W. Mnzing, S. P. Müller, and T. Beyer, "Radiation exposure of patients undergoing whole-body dual-modality ^{18}F-FDG PET/CT examinations," *Journal of Nuclear Medicine*, vol. 46, no. 4, pp. 608–613, 2005.

[101] J. Blacksberg, Y. Maruyama, E. Charbon, and G. R. Rossman, "Fast single-photon avalanche diode arrays for laser Raman spectroscopy," *Opt. Lett.*, vol. 36, no. 18, pp. 3672–3674, Sep. 2011.

[102] S. Stenirri, M. Cretich, I. Rech, A. Restelli, M. Ghioni, S. Cova, M. Ferrari, L. Cremonesi, and M. Chiari, "Dual-color microchip electrophoresis with single-photon avalanche diodes: application to mutation detection," *Electrophoresis*, vol. 29, pp. 4972–4975, Dec 2008.

[103] G. Lu, M. Chouikha, G. Sou, and M. Sedjil, "Colour detection using a buried double p-n junction structure implemented in the CMOS process," *Electronics Letters*, vol. 32, no. 6, pp. 594–596, Mar 1996.

[104] H. Finkelstein, M. Hsu, and S. Esener, "Dual-junction single-photon avalanche diode," *Electronics Letters*, vol. 43, no. 22, 25 2007.

[105] V. Nayar, J. Russell, R. T. Carline, A. J. Pidduck, C. Quinn, A. Nevin, and S. Blackstone, "Optical properties of bonded silicon silicide on insulator (S2OI): a new substrate for electronic and optical devices," *Thin Solid Films*, vol. 313–314, pp. 276–280, 1998.

[106] G. Ripamonti, S. Cova, M. Ghioni, M. Mastrapasqua, and S. Vanoli, "$(PS)^2$: a new semiconductor device for positron-sensitive picosecond detection of single optical photons," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 310, no. 1-2, pp. 184–188, 1991.

[107] K. Shah, R. Farrell, R. Grazioso, E. Harmon, and E. Karplus, "Position-sensitive avalanche photodiodes for gamma-ray imaging," *Nuclear Science, IEEE Transactions on*, vol. 49, no. 4, pp. 1687–1692, Aug 2002.

[108] "MEGAFRAME - million frame per second, time-correlated single photon camera [IST FP6 FET open]." [Online]. Available: http://www.megaframe.eu

[109] M. Fishburn and E. Charbon, "System tradeoffs in gamma-ray detection utilizing SPAD arrays and scintillators," *Nuclear Science, IEEE Transactions on*, vol. 57, no. 5, pp. 2549–2557, Oct. 2010.

[110] R. Pestotnik, R. Dolenec, S. Korpar, P. Krizan, and A. Stanovnik, "Time-of-flight PET with Cherenkov photons," in *2011 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, Oct 2011.

[111] M. E. Phelps, Ed., *PET: Physics, Instrumentation, and Scanners*, 1st ed. Springer, 2006.

[112] J. S. Karp, S. Surti, M. E. Daube-Witherspoon, and G. Muehllehner, "Benefit of time-of-flight in PET: Experimental and clinical results," *Journal of Nuclear Medicine*, vol. 49, no. 3, pp. 462–470, 2008.

[113] M. S. Judenhofer *et al.*, "Simultaneous PET-MRI: a new approach for functional and morphological imaging," *Nature Medicine*, vol. 14, pp. 459–465, 2008.

[114] B. Arnold, *A First Course in Order Statistics.* Wiley-Interscience, 1992.

[115] S. Seifert, J. Steenbergen, H. van Dam, R. Vinke, P. Dendooven, H. Lohner, F. Beekman, P. Dorenbos, E. van der Kolk, and D. Schaart, "Accurate measurements of the rise and decay times of fast scintillators with solid state photon counters," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2010 IEEE*, Nov. 2010, pp. 1736–1739.

[116] S. Derenzo, J. Weber, W. Moses, and C. Dujardin, "Measurements of the intrinsic rise times of common inorganic scintillators," in *Nuclear Science Symposium, 1999. Conference Record. 1999 IEEE*, vol. 1, 1999, pp. 152–156.

[117] E. H. Mckinney, "Generalized birthday problem," *The American Mathematical Monthly*, vol. 73, no. 4, pp. 385–387, 1966.

[118] J. Glodo, W. Moses, W. Higgins, E. van Loef, P. Wong, S. Derenzo, M. Weber, and K. Shah, "Effects of Ce concentration on scintillation properties of LaBr3:Ce," *Nuclear Science, IEEE Transactions on*, vol. 52, no. 5, pp. 1805–1808, Oct. 2005.

[119] S. Seifert *et al.*, "The lower bound on the timing resolution of photon counting scintillation detectors," *Physics in Medicine and Biology*, (forthcoming).

[120] C. A. Lee, R. A. Logan, R. L. Batdorf, J. J. Kleimack, and W. Wiegmann, "Ionization rates of holes and electrons in silicon," *Phys. Rev.*, vol. 134, pp. A761–A773, May 1964.

[121] H. M. Schey, *DIV, Grad, Curl, And All That: An Informal Text on Vector Calculus*, 3rd ed. W. W. Norton and Company, 1997.

[122] S. Henzler, *Time-to-Digital Converters*, 1st ed. Springer, 2010.

[123] B. Swann, B. Blalock, L. Clonts, D. Binkley, J. Rochelle, E. Breeding, and K. Baldwin, "A 100-ps time-resolution CMOS time-to-digital converter for positron emission tomography imaging applications," *Solid-State Circuits, IEEE Journal of*, vol. 39, no. 11, pp. 1839–1852, Nov. 2004.

Summary

Motivated by the demand for time-correlated imaging and single-photon detectors in biomedical and research applications, this thesis covers how single-photon avalanche diode (SPAD) performance relies on the underlying physics. Special attention is focused on operation of SPADs in hostile environments, including radioactive environments, locations with strong magnetic fields, and distortions from forced triggering.

Many applications, especially biomedical ones, use the unique nature of light and single photons. This thesis discusses how understanding the physics behind CMOS-integrated SPADs is important when detect single photons for applications such as positron emission tomography. Figures of merit, state-of-the-art detectors, and current understanding of the physical processes involved with an avalanche are presented in Ch. 2. Ch. 3 presented characterization techniques for measuring SPADs. Four different methods of measuring the breakdown voltage *in situ* are compared, with errors ranging from 0.1V to 0.5V depending on the measurement conditions. Three methods for measuring the afterpulsing are discussed, along with an estimation of the afterpulsing probability per unit charge. A comparison of techniques for measuring the inactive distance showed good agreement with one another.

Ch. 4 discusses the importance of ensuring that single photons are incident on the device during a timing jitter measurement; a good match is shown between an experimentally measured decrease in the diffusion tail along with the predicted value. A decrease in the measured quench time of roughly 200ps is presented when multiple photons were simultaneously incident on a single SPAD. The chapter also discusses distortions to the triggering probability when multiple photons are incident during light pulses on a single SPAD.

SPADs' insensitivity to magnetic fields with magnitudes of nearly 10T is hypothesized and experimentally measured in Ch. 5. The multiplication-assisted diffusion model is extended to include the effects from a convec-

tive force acting on the carriers in the magnetic field, predicting the lack of changes. Also shown is the increase in noise from $\sim$1.25MeV γ-rays, with a discussion on the lifetime of PET sensors.

Ch. 6 presents a diode with an electrically controllable breakdown voltage in a portion of the diode, which shares a larger noise rate with the radiation damaged diodes. Control occurs via modulation of the voltage on a polysilicon layer above the edge of the diode. The breakdown voltage, which can be modulated between 16V and 18V, is shown to be in good agreement with the theory. Additionally, the use of a negative voltage on the polysilicon, ostensibly exposing the high field region to surface-generated carriers, triggers RTS noise in the avalanche diode.

In an attempt to mitigate the greater noise observed from radiation damage or from fabrication issues, Ch. 7 presents a position-sensitive diode capable of localizing fixed-pattern noise and selectively ignoring this noise. Along with the underlying theory, an 8dB increase in SNR is shown for a diode.

Ch. 8 examines under what circumstances noise will begin to effect a SPAD-based sensor's performance when targeting positron emission tomography. The importance of fill factor is clear in this case study, especially when considering slow scintillators with high rise times. Fill factor is shown to be the dominant consideration in detector performance when LYSO with 40ns decay and 500ps rise time is coupled to a SPAD-based sensor; even if the rise time is decreased to 80ps, the fill factor remains the prime consideration until the SPAD-based sensor collects at least one third of the scintillator's output light.

The role that SPADs will play in PET-MRI seems clear; there is no contemporary competing detector that allows simultaneous acquisition of PET images with MRI compatible materials. SPADs also show promise in a plethora of other single-photon applications, including quantum key distribution, flouresence lifetime imaging microscopy, and scintillator characterization. Understanding the device fundamentals creates the possibility of predicting performance, and improving detectors for these applications.

Samenvatting

Gemotiveerd door de vraag naar tijd-gecorreleerde beeldvorming en enkelfoton detectors in biomedische en onderzoeks-toepassingen, behandelt dit proefschrift de vraag hoe de prestatie van single-photon avalanche diodes (SPADs) afhangen van de onderliggende fysica.

Veel toepassingen, in het bijzonder biomedische, gebruiken de unieke eigenschappen van licht. Dit proefschrift beschrijft hoe het juiste begrip van de fysica achter CMOS-geintegreerde SPADs van belang is als we enkele fotonen willen detecteren voor toepassingen zoals positron emission tomography (PET). Prestatie-indicators, de momenteel beste detectors, en het huidige begrip van de fysische processen die een rol spelen in het lawine-effect worden gepresenteerd in Hfst. 2. Hfst. 3 presenteert manieren om SPADs via metingen te karakteriseren. Vier verschillende methoden om de afkapspanning *in situ* te meten worden vergeleken, met fouten in de orde van 0.1V tot 0.5V, afhankelijk van de meetomstandigheden. Drie methoden om het napulsen te meten worden behandeld, evenals een schatting van de waarschijnlijkheid van napulsen per eenheid lading. Een vergelijking van technieken voor het meten van de inactieve afstand liet zien dat die goed met elkaar overeenkomen.

Hfst. 4 beschrijft het belang van ervoor te zorgen dat een enkel foton op de SPAD valt tijdens een meting van de tijdvariatie; een goede overeenkomst wordt aangetoond tussen de experimenteel bepaalde afname in de diffusiestaart en de voorspelde waarde. Een afname van de gemeten afkaptijd van ongeveer 200ps wordt getoond voor het geval waar meerdere fotonen tegelijkertijd op een enkele SPAD vallen.

De ongevoeligheid van SPADs voor magnetische velden met sterkes tot bijna 10T wordt gesteld en experimenteel gemeten in Hfst. 5. Het vermenigvuldiging-ondersteund diffusie-model wordt uitgebreid om de effecten van een convectieve kracht op de dragers in het magnetisch veld mee te nemen; dit voorspelt dat er geen verandering is. Ook wordt de toename in ruis van ~1.25MeV γ-stralen getoond, met een behandeling van de levensduur van PET sensoren.

Hfst. 6 presenteert een diode met een elektrisch regelbare afkapspanning in een deel van de diode, deze deelt een hogere ruisfrequentie met de diodes die door straling zijn aangetast. De sturing vindt plaats door de modulatie van de spanning op een poly-silicon laag boven de rand van de diode. Voor de afkapspanning, die gemoduleerd kan worden tussen 16V en 18V, wordt getoond dat deze goed overeenkomt met de theorie. Verder geeft het gebruik van een negatieve spanning op het poly-silicon RTS ruis op de lawinediode, naar aangenomen wordt door de hoge veldzone bloot te stellen aan oppervlakte-gegenereerde dragers.

In een poging om de toegenomen ruis vanwege schade door straling of fabricagefouten te verminderen, presenteert Hfst. 7 een positie-gevoelige diode die in staat is om de positie te bepalen van ruis op een vaste plaats, en deze ruis te onderdrukken. Samen met de onderliggende theorie wordt voor een diode een 8dB toename in SNR getoond.

Hfst. 8 bestudeert onder welke omstandigheden ruis de prestatie van een SPAD-gebaseerde sensor begint de beinvloeden, voor toepassingen rond positron emission tomography. Het belang van de opvulfactor is in dit geval duidelijk, in het bijzonder voor langzame scintillators met snelle stijgtijden. Het wordt getoond dat de opvulfactor de dominante factor in de prestatie van de detector is als LYSO met 40ns afval- en 500ps stijgtijd wordt gecombineerd met een SPAD-gebaseerde sensor; zelfs als de stijgtijd wordt verminderd tot 80ps is de opvulfactor de belangrijkste factor voor de SPAD-gebaseerde sensor welke tenminste een derde van het vrijkomende scintillator-licht verzamelt.

De rol die SPADs zullen spelen in PET-MRI lijkt duidelijk; er is momenteel geen alternatieve detector die gelijktijdige opname van PET beelden met MRI-compatibele materialen combineert. SPADs zijn ook veelbelovend in een reeks van andere enkel-foton toepassingen, waaronder kwantum-sleutel distributie, fluorescent levensduur microscopen, en scintillator karakterisatie. Het begrijpen van de fundamenten geeft mogelijkheden om de prestatie te voorspellen, en om detectors voor deze toepassingen te verbeteren.

Acknowledgments

I have spent a great deal of time thinking about how blue photons interact with silicon over the past four years. Focus so sharp requires support from a plethora of people, many but not all of whom I have the honor to thank hereafter.

First and foremost, I thank my adviser, Professor Edoardo Charbon. This dissertation was made possible by his wisdom and sage advice on many issues, ranging from providing insight on my philosophosical quandries to basic knowledge of CMOS circuits. Finally, no matter how stupid or seemingly meaningless my question, and I think there were some questionable questions, Edoardo was always able to provide some insight on the issue.

I am also deeply indebted to Dr. Yuki Maruyama. His ability to take a seemingly insourmountable problem with an experimental setup and turn said problem into a workable issue has saved me many times. The processing knowledge he shared with me was crucial in many simulations I have performed.

The members of the MEGAFRAME consortium, including Claudio Bruschini, Dr. David Stoppa, Dr. Robert Henderson, Dr. Fausto Borghetti, Dr. Justin Richardson, Dr. Marek Gersbach, Richer Walker[1], and Steve East, taught me a great deal about circuit design and project planning. I am also indebted to their advice on long-term research plans, as I am to Prof. Dennis Schaart.

The present and former staff of the Circuits and Systems group, including Prof. Alle-Jan van der Veen, Antoon Frehe, Rosario Salazar, Laura Bruns, and Minaksie Ramsoekh, have helped me in all sorts of matters that, while small individually, would be no small feat in sum. For their help with my research, I also thank my labmates: Dr. Mohammad Azim Karami, Chockalingam Veerappan, Lucio Carrara, Dr. Claudio Favi, Dr. HyungJune Yoon,

[1] Richard will probably have a title of Dr. by the time this work is published.

L. Harmen Menninga, Joop Meijlink, and Rahmadi Trimananda. I am especially grateful to Shingo Mandai for his circuit help. Many other people at or formerly at TUDelft, including Sten Vollebregt, Aslihan Arslan, Senad Hiseni, Duan Zhao, Ruimin Yang, Yu Bi, Dr. Claude Simon, Dr. Yiyin Wang, Prof. Nick van der Meijs, Qin Tang, Prof. Geert Leus, Prof. René van Leuken, and Alexander de Graaf all deserve my thanks. I also would like to thank Joe Neesan.

I used many open and freely available tools during my thesis. I am grateful to the communities of and contributors to the python programming language, python's matplotlib library, python's scipy library, LaTeX, subversion, Ubuntu, and Wikipedia. I am also grateful to Xilinx, Inc., Acam Messelectronic GmbH, and Enterpoint, Ltd. for the generous donations of hardware, firmware, and software to my research endeavours. I thank Dr. Florian Stuker and Dr. Katerina Dikaiou of ETH Zürich for their help with the experiments utilizing strong magnetic fields.

My publications were possible because of the anonymous work of the people who reviewed my submissions. Thank you to these individuals. Even when I respectfully disagree with your opinions, the current form of science is only possible by your efforts.

Though an ocean separates America from Europe, I owe a debt of gratitude to many people on the other side of the Atlantic. For listening and her helpful, peer-oriented advice, Dr. Jennifer Gagner has my heartful gratitude. It is quite possible I would have given up in year one of my PhD without her support. Finally, my appreciation also goes to Gireeja Ranade, Jehan deFonseka, Clare Davis, Jackie Hitchingham, Elizabeth Power, L. Brown Westrick, Eric Jonas, my uncle Dr. John Fishburn, and my cousin Laura Fishburn. I must also thank my teachers, to whom I owe much of my identity and achievements, including Mr. McAndrew, Mr. Flater, Mr. Brandt, and Mrs. Schmidt. My gratitude especially goes to Mrs. Monroe, Mrs. Flater, and Mr. Anderson.

My whole-hearted appreciation goes to Alison Cichowlas, whose patience has no end that I have encountered, at least not yet, and who has helped me thrive during circumstances most frustrating.

Finally, my love and gratitude goes to my brother Aaron, my sister Mary, and especially to my parents. Their constant support, patience, and unquestioning love continue to amaze me.

Matthew William Fishburn
February 2012
Delft, the Netherlands

About the Author

Matthew W. Fishburn was born in Cedar Rapids, Iowa, the United States of America in 1984. In 2001, he qualified for the United States of America's Mathematical Olympiad. He graduated from Linn-Mar High School in Marion, Iowa in 2003 after being named a National Merit Scholarship Finalist. Matt received the Dandekar Excellence Scholarship in 2004 and in 2005 was a Carl P. and was a Marie G. Dennett Memorial Scholar. He received a Bachelor's of Science in Electrical Engineering and Computer Science from the Massachusetts Institute of Technology in 2007. He worked at Bridgewater Associates, LP, before starting work on his Ph.D. in the research group of Professor Edoardo Charbon at Delft University of Technology in 2008. Matt's research interests include biomedical applications of embedded devices, especially applications involving time-correlated imaging.

Publications

Journal Papers

M. Fishburn and E. Charbon, "System tradeoffs in gamma-ray detection utilizing SPAD arrays and scintillators," *Nuclear Science, IEEE Transactions on*, vol. 57, no. 5, pp. 2549-2557, Oct. 2010.

M. Fishburn, Y. Maruyama, and E. Charbon, "Reduction of fixed-position noise in position-sensitive single-photon avalanche diodes," *Electron Devices, IEEE Transactions on*, vol. 58, no. 8, pp. 2354-2361, Aug. 2011.

M. Karami, L. Carrara, C. Niclass, **M. Fishburn**, and E. Charbon, "RTS noise characterization in single-photon avalanche diodes," *Electron Device Letters, IEEE*, vol. 31, no. 7, pp. 692-694, July 2010.

S. Mandai, **M. Fishburn**, Y. Maruyama, and E. Charbon, "Wide spectral range single-photon avalanche diode fabricated in advanced 180nm CMOS technology," *Optics Express*, vol. 20, no. 6, pp. 5849–5857, Mar 2012.

M. Gersbach, Y. Maruyama, R. Trimananda, **M. Fishburn**, D. Stoppa, J. Richardson, R. Walker, R. Henderson, and E. Charbon, "A time-resolved, low-noise single-photon image sensor fabricated in deep-submicron cmos technology," *Solid-State Circuits, IEEE Journal of*, 2012 (accepted).

M. Fishburn and E. Charbon, "A preliminary study on the environmental dependencies of avalanche propagation in silicon," *Electron Devices, IEEE Transactions on*, (submitted).

M. Fishburn, L. H. Menninga, C. Favi, and E. Charbon, "An FPGA-based TDC with multiple channels for open source applications," *Nuclear Science, IEEE Transactions on*, (submitted).[1]

Book Chapters

E. Charbon and **M. Fishburn**, "Chapter 7: Monolithic Single-Photon Avalanche Diodes: SPADs," from Single-Photon Imaging, P. Seitz and A. J. P. Theuwissen, Eds. Springer, 2011.

Conference Papers

M. Fishburn and E. Charbon, "Environmental effects on photomultiplication propagation in silicon," in *Nuclear Science Symposium Conference Record (NSS/MIC), 2011 IEEE*, Oct. 2011, pp. 572-574.[2]

M. Fishburn and E. Charbon, "Distortions from multiphoton triggering in a single CMOS SPAD," in *SPIE Defense Security+Sensing*, Apr. 2012.[2]

C. Veerappan, J. Richardson, R. Walker, D. Li, **M. Fishburn**, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. Henderson, and E. Charbon, "A 160x128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, Feb. 2011, pp. 312-314.[1][2]

E. Charbon and **M. Fishburn**, "Radiation-hardened and radiation-sensitive single-photon imagers," in *2011 International Workshop on Radiation Imaging Detectors Book of Abstracts*, Jul. 2011, p. 44.[2][3]

[1] First authorship of this work is shared among two or more people, myself included

[2] I either gave or will give the oral presentation for this conference publication.

[3] Due to a printing error, my name does not appear in the book of abstracts, but it is shown on the iWoRiD web site.

L. H. Menninga, C. Favi, **M. Fishburn**, and E. Charbon, "A multi- channel, 10ps resolution, FPGA-based TDC with 300MS/s throughput for open-source PET applications," in *2011 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, Oct 2011, pp. 1515–1517.

C. Veerappan, J. Richardson, R. Walker, D. Li, **M. Fishburn**, D. Stoppa, F. Borghetti, Y. Maruyama, M. Gersbach, R. Henderson, C. Bruschini, and E. Charbon, "Characterization of large-scale non-uniformities in a 20k TDC/SPAD array integrated in a 130nm CMOS process," in *Solid-State Device Research Conference (ESSDERC), 2011 Proceedings of the European*, Sept. 2011, pp. 331-334.

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 using a novel SPAD-based image sensor," in *Detectors and Imaging Devices: Infrared, Focal Plane, Single Photon*, E. L. Dereniak, J. P. Hartke, P. D. LeVan, A. K. Sood, R. E. Longshore, and M. Razeghi, Eds., vol. 7780, no. 1. SPIE, 2010, p. 77801H.

Delft Institute of
Microsystems and
Nanoelectronics
dimes

TUDelft Delft
University of
Technology

9 789491 030291