7-5-2012

Modeling and engineering impact ionization in avalanche photodiodes for near and mid infrared applications

David Ramirez

Recommended Citation

Ramirez, David. "Modeling and engineering impact ionization in avalanche photodiodes for near and mid infrared applications." (2012). https://digitalrepository.unm.edu/ece_etds/215

David Alejandro Ramirez

Candidate

Electrical and Computer Engineering

Department

This dissertation is approved, and it is acceptable in quality and form for publication:

Approved by the Dissertation Committee:

Majeed M. Hayat, Chairperson

Sanjay Krishna

Steven R. J. Brueck

Sudhakar Prasad

Modeling and Engineering Impact Ionization in Avalanche Photodiodes for Near and Mid Infrared Applications

by

David A. Ramirez

M.S., Engineering, University of Concepcion, 2005
B.S., Engineering, University of Concepcion, 2002

DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
in Engineering

The University of New Mexico

Albuquerque, New Mexico

May, 2012

Dedication

Dedicated to the memory of my mother.

Acknowledgments

I would like to thank my advisors Professor Majeed M. Hayat and Professor Sanjay Krishna for their continued support, guidance, and encouragement in my research work. I greatly appreciate the opportunity I have had to work both in theory and experiments as a result of having worked with them. That diversity of research has given me invaluable tools in my professional life. Furthermore, I would like to thank them for giving me the opportunity to have presented my work in several conferences. I would also like to thank Professor Majeed M. Hayat for sending me to the Department of Electronic and Electrical Engineering, at the University of Sheffield, to gain experimental knowledge involving experimental characterization of avalanche photodiodes.

I also want to thank Professor Steven R. J. Brueck and Professor Sudhakar Prasad for participating in my dissertation committee and for providing insightful comments that contributed in improving this work.

I am especially grateful to Professor John P. R. David for all his help during my visit to the Department of Electronic and Electrical Engineering at the University of Sheffield, UK. I appreciate the time he spent with me, sharing his expertise in the area of avalanche photodiodes. I would also like to thank all the people in the impact ionization group (IIG) at the University of Sheffield for their help during my visit. In particular, I want to thank Lionel Tan for his help on the characterization of avalanche photodiodes and for the time and effort he extended in helping me during my time in Sheffield.

I greatly appreciate the insightful comments about avalanche photodiodes provided by Professor Joe C. Campbell from the University of Virginia and Professor Chee Hing Tan from the University of Sheffield during discussions held at LEOS

conferences. I would like to thank Professor Graham Rees from the University of Sheffield for sharing his profound knowledge of avalanche photodiodes and for providing insightful comments, which greatly contributed to this work. I am also grateful to Dr. Mark Itzler from Princeton Lightwave Inc. for spending great lengths of time providing valuable comments and thoughts about my research and for providing data that contributed immensely to this work.

I would like to thank my research group, which is composed by a splendid group of people with whom I have had the privilege of sharing all these years. In particular, I want to thank Woo-Yong Jang for his friendship and all the cups of coffee we have shared during these years.

I would like to thank my friends in Albuquerque: Jorge and Angely, Sebastian, Tim, Alejandro, Diego, Ulises, and all my pals from my soccer team Universitarios. I want to thank my extraordinary Sardinian friends Stefano and Paolo (Mascagna and Crestino) with whom I have shared great moments. I am especially grateful to Sophia for her continued support, care, and patience during all these years. I also want to thank Beowulf, my cat, because according to Sophia, petting him decreases my stress level. I do not know if that is true because Beowulf and I spend our time just fighting and playing.

Most importantly, I would like to acknowledge my family. I want to thank my parents, brothers, and sisters for their unconditional support and great sense of humor. I want to especially thank my beloved mother whose immeasurable love has been the pillar of our family.

This work was supported by: The National Science Foundation (NSF), under Grants Nos. ECS-0601645, ECS-0428756, and EEC-0812056 (Smart Lighting Engineering Research Center, ERC); Comision Nacional de Investigacion Cientifica y Tecnologica (CONICYT) de Chile; AFRL, under Grant No. FA9453-07-C-0171; and Princeton Lightwave Inc. (through a NASA funded SBIR program).

Modeling and Engineering Impact Ionization in Avalanche Photodiodes for Near and Mid Infrared Applications

by

David A. Ramirez

ABSTRACT OF DISSERTATION

Submitted in Partial Fulfillment of the
Requirements for the Degree of

Doctor of Philosophy
in Engineering

The University of New Mexico

Albuquerque, New Mexico

May, 2012

Modeling and Engineering Impact Ionization in Avalanche Photodiodes for Near and Mid Infrared Applications

by

David A. Ramirez

M.S., Engineering, University of Concepcion, 2005
B.S., Engineering, University of Concepcion, 2002

Ph.D., Engineering, University of New Mexico, 2012

Abstract

Avalanche photodiodes (APDs) are the preferred photodetector in many applications in which low light levels need to be detected. The reason why APDs are important in such applications is due to their internal gain, which improves the APD's sensitivity. Compared to receivers based on PIN photodiodes, which do not present internal gain, APD-based receivers achieve 5–10 dB improved sensitivity. The origin of the APD's internal gain is the impact ionization process. However, due to the stochastic nature of the impact ionization process the multiplication gain comes at the expense of extra noise. This multiplication noise is called the excess noise, and it is a measure of the gain uncertainty. In addition, as the multiplication gain increases the buildup time, which is the time required for all the impact ionizations to complete, also increases. Thus, for a given multiplication gain the buildup time limits the bandwidth of the APD.

The main challenge for state-of-the-art APDs, operating in linear and Geiger modes, is to achieve higher operating speeds. For application in which the APD is operated in linear mode the limited speed of APD-based receivers have limited their use in systems that operate at 2.5 and 10 Gbps. However, to meet the demand of the exponential growth in data transfer, the telecommunication industry has been moving toward 40-Gbps and 100-Gbps protocols for their core fiber-optic backbone networks alongside the existing 10-Gbps infrastructure operating at the low-loss wavelength of 1.55 μm. Moreover, the fast progress on quantum communications requires Geiger-mode APDs to operate at higher repetition rates. Currently, Geiger-mode APDs are limited to operate at detection rates of about 20 MHz. In addition, there has been relatively little work on infrared APDs, although there are many applications in remote sensing, medical imaging, and environmental monitoring. In particular, there is no GaAs-based APD operating in Geiger mode beyond 2 μm.

This dissertation provides theoretical analysis and experimental exploration of APDs working in linear and Geiger modes in the near infrared (NIR) and mid-infrared (MIR) ranges of wavelength. This research effort is geared to address the aforementioned current challenges of the state-of-the-art APD technology. In the theoretical part of this work the focus is on the development of new theoretical methods that allow us to model, understand, and characterize avalanche photodiodes working in linear and Geiger modes. The objective is that the developed methods help the design and optimization of high performance, high speed APDs. The experimental part of this research effort consists of the design, fabrication and characterization of a novel mid-infrared sensor, based on GaAs technology, called the quantum-dot avalanche photodiode (QDAP). The main motivation for the QDAP is to exploit its potential of working in Geiger mode regime, which can be utilized for single-photon detection. In addition, the QDAP represents the first GaAs-based APD operating in the mid infrared range of wavelength.

Contents

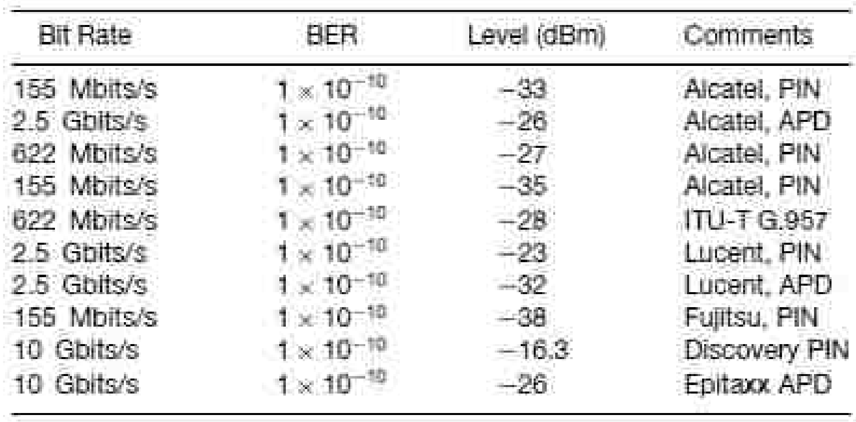
List of Figures

Chapter 1

Introduction

1.1 Overview

Avalanche photodiodes (APDs) are very important devices in a wide range of commercial, military, and research applications. These applications include optical communications, satellite laser ranging [1], deep-space laser communications [2], time-resolved photon counting [3], quantum key distribution [4, 5, 6] and quantum imaging [7]. In recent years, the primary driving force for research and development of APDs has been optical communications systems, especially at high bit rates. The popularity of APDs in optical communications is due to their internal gain, which increases the output signal of the device above the thermal-noise level of the receiver. Thus, in thermal-noise-limited operation, APD-based receivers achieve a higher sensitivity than those based on PIN photodiodes, which do not exhibit internal gain. Compared to receivers with PIN photodiodes, APD-based receivers achieve 5–10 dB improved sensitivity [8, 9]. Figure 1.1, extracted from [9], compares the sensitivity, measured as the average received power level (in dBm), of commercial APDs and PIN photodiodes at a bit error rate (BER) of 1×10^{-10}, for some common bit rates. In practice, the lower sensitivity of PIN photodiodes can be overcome by using erbium doped fiber amplifiers (EDFAs) to pre-amplify the signals optically before the are detected by the PIN photodiode. The use EDFA-PIN receivers is well established; they offer fast response time and good signal-to-noise ratio characteristics. However, the optical amplifier is an expensive and bulky component. An EDFA requires the use of meters of fiber, and it requires the use of a pump laser to provide the optical amplification. On the other hand, in the APD-based receiver there is no need for the optical pre-amplification stage, since the amplification is performed by the APD. As a result, APD-based receivers benefit from small form-factor packaging, and offer a more cost-effective solution compared to the EDFA–PIN receiver.



Bit Rate	BER	Level (dBm)	Comments
155 Mbits/s	1×10^{-10}	−33	Alcatel, PIN
2.5 Gbits/s	1×10^{-10}	−26	Alcatel, APD
622 Mbits/s	1×10^{-10}	−27	Alcatel, PIN
155 Mbits/s	1×10^{-10}	−35	Alcatel, PIN
622 Mbits/s	1×10^{-10}	−28	ITU-T G.957
2.5 Gbits/s	1×10^{-10}	−23	Lucent, PIN
2.5 Gbits/s	1×10^{-10}	−32	Lucent, APD
155 Mbits/s	1×10^{-10}	−38	Fujitsu, PIN
10 Gbits/s	1×10^{-10}	−16.3	Discovery PIN
10 Gbits/s	1×10^{-10}	−26	Epitaxx APD

Figure 1.1: Sensitivity, measured as the average received power level (in dBm), of commercial APDs and PIN photodiodes at a bit error rate (BER) of 1×10^{-10}, for some common bit rates. Data extracted from [9].

APDs operate on the basis of highly energetic (hot) carriers that exploit the impact ionization mechanism to achieve avalanche multiplication gain. However, due to the stochastic nature of the impact ionization process the multiplication gain comes at the expense of extra noise. This multiplication noise is called the excess noise, and it is a measure of the gain uncertainty. In addition, as the multiplication gain increases the buildup time, which is the time required for all the impact ionizations to complete, also increases. This gives rise to intersymbol interference (ISI), which limits the receiver performance in high-speed systems. McIntyre showed that the excess noise factor, F, associated with the mean gain, M, of the APD can be expressed in the case of uniform electric field and pure electron injection as [10]

$$F(M) = \frac{\langle M^2 \rangle}{\langle M \rangle^2} = kM + (1-k)\left(2 - \frac{1}{M}\right) \tag{1.1}$$

where $k = \beta/\alpha$ is the ratio of the ionization coefficients for electrons, α, and holes, β, of the semiconductor material. In the case of pure hole injection, k in (1.1) is given by $k = \alpha/\beta$. The value of α and β, which represent the probability of impact ionization per unit length (cm^{-1}), depend on the band structure of the semiconductor, the scattering processes (mainly phonon scattering), and the electric field [11]. Equation (1.1) has been derived under the condition that the ionization coefficients at a specific position are determined solely by the electric field at that position, the so-called *local* approximation. It is well known that the impact ionization is non-local, in the sense that carriers injected in the multiplication region require a minimum distance before acquiring sufficient energy to impact ionize. The distance in which no impact ionizations occur is called the *dead space*. However, for thick multiplication regions (> 1 μm) the dead space can be neglected, and the local approximation provides an accurate prediction of the excess noise factor [11]. From (1.1) it is clear that the lowest excess noise is obtained when k is small. Therefore, the more disparate the ionization coefficients α and β are in a semiconductor material, the lower the excess noise is. Examples of semiconductor materials that exhibit small values of k are Si and HgCdTe.

Essentially, an APD is a *pin* junction that operates under reversed bias, as shown in Fig. 1.2a. As a result of the reverse applied voltage a depletion region is formed across the i region, which establishes an intense electric field. When an incoming

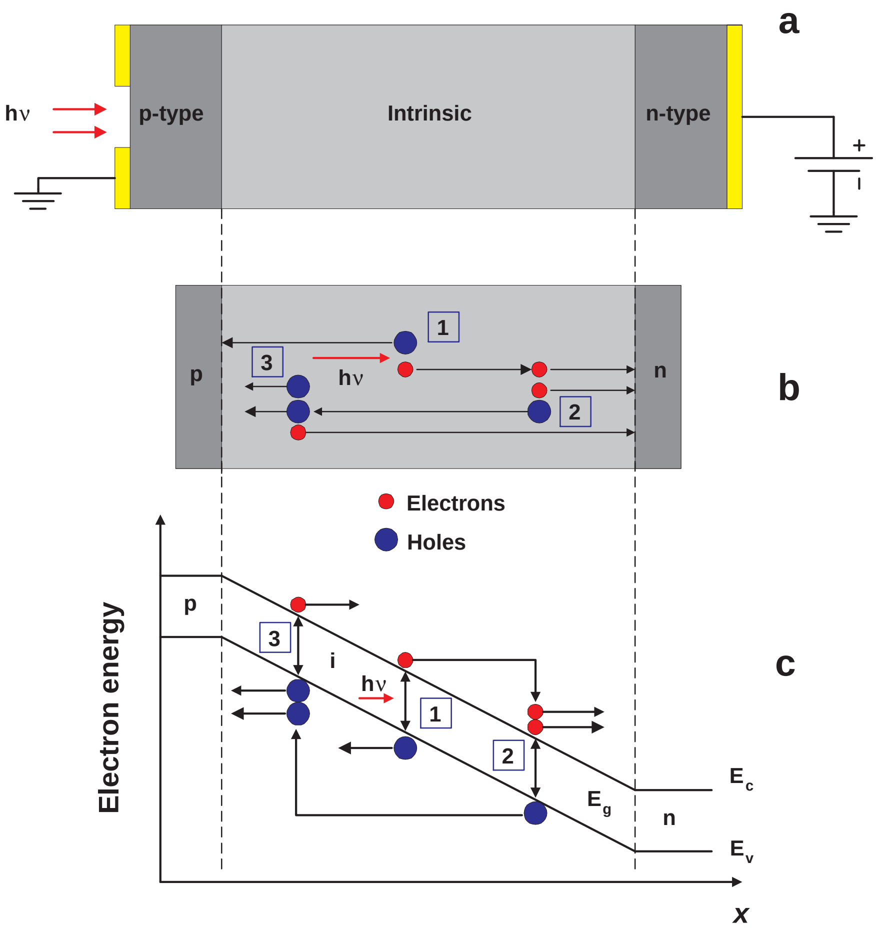


Figure 1.2: (a) Reverse biased *pin* junction. (b) Illustration of three ionization events in the depletion region from a spatial point of view. (c) Illustration of the energy band transitions of the three ionization events shown in (b).

photon with sufficient energy $h\nu$, where h is the Planck's constant and ν is the photon frequency, is absorbed an electron-hole pair is generated. This event is identified with the number 1 in Figs. 1.2 b and c. Under the influence of the electric field the photogenerated electron and hole are forced to travel in opposite directions. As the electric field increases these carriers can gain kinetic energy from the electric field at a faster rate than they lose it to the various scattering processes. If the electron or the hole acquire sufficient energy, i.e., higher than the ionization threshold energy, a random collision with an atom of the material results in an impact ionization event, which generates a new electron-hole pair. This is event 2 in Figs. 1.2 b and c. The electric fields required to observe impact ionization depend on the band gap of the material and may range at room temperature from $\approx 10^4$ V/cm^{-1} in low-gap semiconductors, such as InAs ($E_g = 0.33$ eV), to values well in excess of 10^5 V/cm^{-1} in wide-gap materials, such as GaP ($E_g = 2.24$ eV) [12]. During the impact ionization the carrier responsible for the ionization (parent carrier) looses part or all of its energy to create the new electron-hole pair. As a result, right after an impact ionization the parent carrier and the newly-generated electron and hole have almost no kinetic energy. However, at this point the electric field accelerates the carriers and they start to acquire kinetic energy. If any of these carriers acquire sufficient energy they can impact ionize once again generating another electron-hole pair, shown as event 3 in Figs. 1.2 b and c. The repetition of this process can yield a cascade of impact ionization events. This process is called avalanche multiplication. When an APD is properly biased it operates by converting each electron hole pair, resulting from the absorption of a photon, to a large number of electron hole pairs via a cascade of impact ionizations. The three ionization events described above are illustrated in Figs 1.2b from a spatial points of view. On the other hand, Fig. 1.2c illustrates the schematic of the energy band transitions associated with the same ionizations events.

The impact ionization rate is not only affected by the ionization threshold energy but also by the phonon scattering rate [12]. At nonzero temperature, the atoms in the crystal lattice vibrate around their fixed equilibrium. These vibrations are quantized and the quantum of lattice vibrations is called a phonon [13]. In III–V semiconductor materials phonon scattering is the dominant scattering mechanism. In particular, the impact ionization rate is strongly affected by phonon scattering [12,

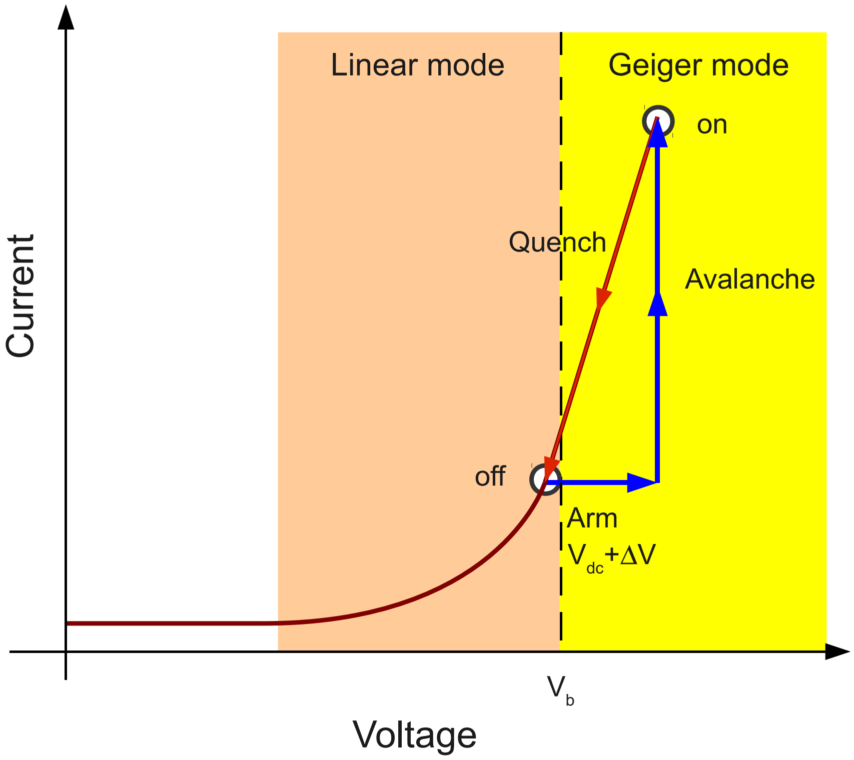


Figure 1.3: Current-voltage characteristics of an APD under reverse bias. As the voltage across the APD increases the APD goes from linear mode operation to Geiger mode operation. The figure also shows the transitions between the *on* and *off* states in Geiger mode operation.

14]. Collisions with phonons control the energy and momentum losses of the carriers and thus influence the average distance required to create an electron-hole pair by impact ionization [12]. In the absence of phonon collisions, this distance would be E_{th}/qE where E is the electric field. Phonon scattering greatly increases this distance. Scattering against the direction of the field is very effective in this respect, since after suffering these collisions, carriers are slowed down by the electric field and lose a considerable portion of their energy. This increases considerably the distance

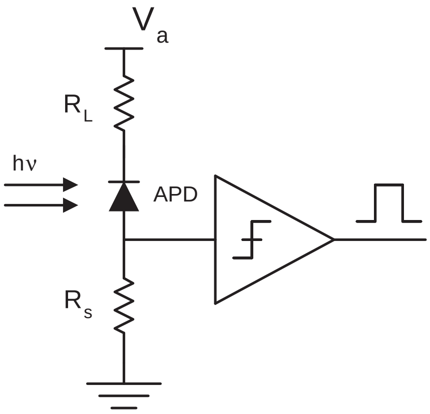


Figure 1.4: Basic quenching circuit.

required to gain the ionization energy [12].

1.1.1 Modes of operation

APDs can be operated in two modes: linear mode and Geiger mode. In the linear mode operation, the APD is biased below its breakdown voltage, V_b[1]. Thus, the cascade of impact ionizations resulting from each injected carrier pair ends within a finite, stochastic time, which is the aforementioned avalanche buildup time. The total number of carrier pairs, injected or generated via impact ionization, constitutes the multiplication factor by which the photocurrent is amplified. On the other hand, in Geiger mode operation, the APD is biased above breakdown. As a result, the number of impact ionizations may increase indefinitely, yielding, in principle, an infinite multiplication factor. In this mode of operation the APD functions as a switch alternating between the *on* and *off* states. Figure 1.3 shows a typical current-voltage characteristics of an APD under reverse bias. It can be seen from the figure that

---

[1] The breakdown voltage is the voltage after which the multiplication gain diverges.

as the voltage across the APD increases the device goes from linear mode operation (zone marked in orange), in which the multiplication gain is proportional to the incident light, to Geiger mode operation (yellow zone), in which the multiplication factor diverges.

To prevent the runaway of the avalanche current, also know as the persistent current, in Geiger mode operation, a ballast resistor, R_L, is sometimes introduced in series with the APD to provide negative feedback, as shown in Fig. 1.4. In this configuration the applied voltage, V_a, is split between the APD and the ballast resistor. As a result, the voltage across the APD decreases as the avalanche current increases. This reduction in the voltage across the APD causes, in turn, the avalanche current through the device to decrease. Depending upon the value of the applied bias, the load resistor and the breakdown voltage, the avalanche current may terminate due to stochastic fluctuations in the carrier production at a stochastic time, known as the quenching time, after which the diode behaves once again as an open circuit [15]. After a recovery period the voltage across the APD once again reaches the value of the voltage supply and the APD is ready for another avalanche trigger. This mode of operation is referred to as the passive quenching mode [15], since the persistent current is allowed to terminate spontaneously. Figure 1.3 shows the transitions between the on and off states in an APD under Geiger mode operation.

The use of passive quenching circuits is the simplest way to terminate the avalanche current of an APD working in Geiger mode [15]. The alternative way to terminate the avalanche current is by employing active quenching circuits [15], which are much more complex. The operation of an active quenching circuit includes the early detection of the avalanche current after an avalanche is triggered and the ultra fast reduction in the voltage across the APD to terminate avalanche pulse. The repetition rate that can be achieved by using active quenching circuits is much higher compared to that of the passive quenching approach. However, the simplicity of the latter makes them very attractive in applications such as single photon imaging and quantum key distribution [16].

1.1.2 APD versus PIN photodiode

Next, the performance of APDs operating in linear mode and PIN photodiodes is compared by calculating the signal to noise ratio (SNR) of both photodetectors.

The most common way to measure the performance of an optical receiver is by calculating the SNR [17, 18, 19, 20]. The SNR of an electrical signal is defined as [18]

$$\mathrm{SNR} = \frac{\text{average signal power}}{\text{noise power}} = \frac{I_p^2}{\sigma_s^2 + \sigma_T^2}, \tag{1.2}$$

where the two fundamental noise mechanisms σ_s^2 and σ_T^2 are the shot noise and the thermal (Johnson) noise, respectively. Additional noise is generated if the incident optical power, P_{in}, is itself fluctuating because of noise produced by optical amplifiers [19]. However, this section focuses only on shot and thermal noises. The total shot noise in an APD is given by [18, 19]

$$\sigma_{s,APD}^2 = 2qM^2F(\Re P_{in} + I_d)\Delta f, \tag{1.3}$$

where M is the multiplication gain, F is the excess noise factor, $\Re$ is the responsivity, q is the charge of the electron, I_d is the dark current, and Δf is the effective noise bandwidth of the receiver. For a PIN photodiode, in which $M = 1$ and $F = 1$, (1.3) reduces to

$$\sigma_{s,pin}^2 = 2q(\Re P_{in} + I_d)\Delta f. \tag{1.4}$$

The thermal noise term, σ_T^2, in an APD is given by

$$\sigma_{T,APD}^2 = (4k_bT/R_L)F_n\Delta f, \tag{1.5}$$

where k_b is the Boltzmann constant, R_L is the load resistor in the front end of the optical receiver, T is the temperature, and F_n, which accounts for the amplifier noise, is called the amplifier noise figure. The thermal noise in (1.5) remains the same in PIN receivers since it originates in the electrical components that are not part of the device.

Using equations from (1.2) to (1.5), and having in mind that the photocurrent in an APD is given by $I_p = M\Re P_{in}$, we can calculate the SNR of an APD as [18, 19]

$$\text{SNR}_{APD} = \frac{I_p^2}{\sigma_s^2 + \sigma_T^2} = \frac{(M\Re P_{in})^2}{2qM^2F(\Re P_{in} + I_d)\Delta f + (4k_bT/R_L)F_n\Delta f}, \tag{1.6}$$

similarly, for a PIN photodiode ($M = 1$ and $F = 1$) the SNR is calculated as

$$\text{SNR}_{pin} = \frac{(\Re P_{in})^2}{2q(\Re P_{in} + I_d)\Delta f + (4k_bT/R_L)F_n\Delta f}. \tag{1.7}$$

Figure 1.5 shows the calculated SNR as a function of the optical incident power, P_{in}, for a PIN photodiode and an APD. The multiplication gain of the APD was assumed to be $M = 5$. The rest of the parameters used in the SNR calculation are typical parameter values for a 1.55 μm InGaAs receiver: $R_L = 1$ kΩ, $F_n = 1$, $\Re = 1$ A/W, $I_d = 1$ nA, $\Delta f = 10$ GHz. It can be seen from the figure that the SNR of an APD receiver is worse than that of a PIN when input powers are relatively large. The reason behind this behavior is related to the enhancement of shot noise in APD receivers [18, 19, 20]. At low power levels, thermal noise dominates over shot noise, and the APD gain is beneficial. However, as the incident optical power increases, shot noise begins to dominate over thermal noise, and APD performs worse than a PIN photodiode under the same operating conditions [18, 19, 20]. To illustrate this point, the thermal-noise limited SNR and the shot-noise limited SNR are considered separately.

In the thermal noise limit, in which $\sigma_T \gg \sigma_s$, the ratio between the APD's SNR

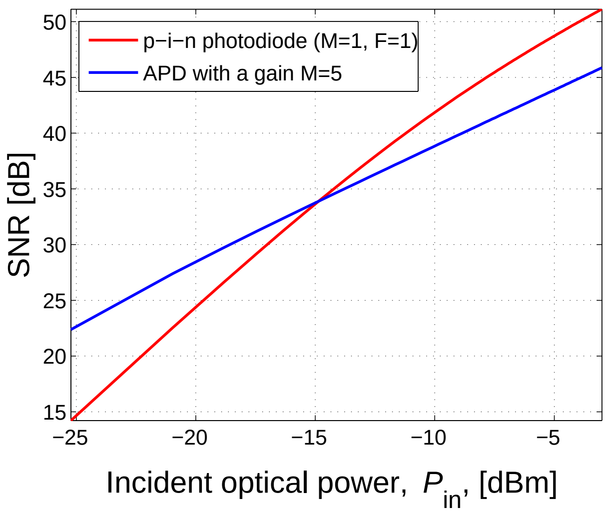


Figure 1.5: Calculated SNR as a function of the optical incident power, P_{in}, for a PIN photodiode and an APD. The multiplication gain of the APD was assumed to be $M = 5$. The rest of the parameters used in the SNR calculation are typical parameter values for a 1.55 μm InGaAs receiver: $R_L = 1$ kΩ, $F_n = 1$, $\Re = 1$ A/W, $I_d = 1$ nA, $\Delta f = 10$ GHz.

and the SNR of the PIN photodiode, given by equations (1.6) and (1.7), is

$$\frac{\mathrm{SNR}_{APD}}{\mathrm{SNR}_{pin}} = M^2 \tag{1.8}$$

As expected, the SNR of the APD is improvement by a factor of M^2 compared to that of the PIN photodiode. On the other hand, in the shot-noise limit, in which $\sigma_s \gg \sigma_T$, the ratio between the APD's SNR and the SNR of the PIN photodiode is

$$\frac{\mathrm{SNR}_{APD}}{\mathrm{SNR}_{pin}} = \frac{1}{F} \tag{1.9}$$

In this case the SNR of the APD is reduced by the excess noise factor, F, compared to that of the PIN photodiode. This analysis illustrates the detrimental effect of the excess noise factor on the APD's performance.

1.2 Single photon avalanche photodiodes (SPADs)

APDs operating in Geiger mode are also know as *single-photon avalanche diodes* (SPADs). These photodetectors are very important in sensing very weak optical signals in applications which span a wide range of the electromagnetic spectrum, from the ultraviolet (10–400 nm) to the long-wave infrared (8–12 μm). As described earlier, APDs operating in Geiger mode employ a nonlinear detection scheme, in which the absorption of a single photon results in a large, saturated current which can easily be detected without ambiguity by electronic circuitry. This detection scheme is usually implemented to measure the arrival time of the incoming photon or in photon counting applications.

1.2.1 Performance characterization of SPADs

Contrary to the case of linear-mode APDs, the concept of gain is less important in SPADs since the operation of the SPAD devices moves between the *off* state and the *on* state. Thus, performance metrics like the excess noise factor, which is very important to characterize APDs working in linear mode is not relevant when characterizing SPADs. This reveals important differences between the linear and the Geiger-mode that need to be considered at the moment of the design and characterization of one or the other type of device.

The performance of a SPAD is primarily measured by the *photon detection efficiency* (PDE), the *dark count rate* (DCR), and the *afterpulsing probability* (AP). The PDE is the product of the detector quantum efficiency and the avalanche breakdown probability. The DCR constitutes false counts and it is a measure of how noisy the detector is. Dark counts originate from dark carriers generated in the absence of illumination; the larger the number of dark carriers, the larger the dark-count probability is [21]. There are several mechanisms that contribute to the concentration of dark carriers. At high electric fields, the dark-carrier concentration is strongly affected by band-to-band tunneling, which depends exponentially on the electric field [22], and it constitutes a limiting factor in APDs that have thin multiplication regions. Another important mechanism that contributes to the number of dark carriers is tunneling through defects. It has been reported that in some materials the tunneling current due to defects is higher than that of band-to-band tunneling [23]. The AP is a measure of the likelihood of afterpulses events, which are false counts originated by detrapped carriers. When charge flows during an avalanche event, some fraction of the carriers are trapped at defect sites in the avalanche region of the SPAD. If these carriers are detrapped at a later time when the device is ready for another avalanche trigger, they can lead to dark counts, referred to as "afterpulses." Larger avalanches involve the flow of more charge, which results in a greater number of trapped charges and consequently larger AP [24, 25]. One way of reducing afterpulsing, in Geiger mode operation, is to keep the voltage across the SPAD below breakdown (hold-off time) for a sufficiently long time interval, longer than the lifetime of the trap. Thus, the next time the SPAD is ready for another avalanche trigger the trap levels are empty. However, since the typical detrapping time is in the μs range, this approach limits the photon counting rate to a few MHz [24, 25].

1.2.2 Breakdown probability

The breakdown probability is a key parameter in the operation of a SPAD; it is a measure how likely it is to trigger an avalanche event. Besides the strength of the electric field and the properties of the material, the probability of a carrier triggering an avalanche breakdown is determined by the place where it is born [26]. A carrier created in the start of the multiplication region has a greater probability of triggering an avalanche event compared to that created close to the end of the multiplication region. This is because a primary carrier created early on in the multiplication region has a larger distance to travel compared to those created close to the end of the multiplication region. For example, for a separate absorption separate multiplication (SAM)[2] APD, a carrier created in the absorption layer is more likely to cause an avalanche compared to that created in the multiplication region. More generally, the dependence of the breakdown probability on the birth location of a carrier is crucial in determining the SPAD's performance when the number of dark carriers inside the absorption and multiplication layers is taken into account.

1.3 Midwave infrared detection

Mid-infrared (MIR) sensors in the 3–25 μm range are very important devices in applications such as medical imaging, fire fighting equipments, and defense and security applications [27, 28]. Among the different technologies available the most important midwave infrared detectors are HgCdTe [29], InSb [30], type II InAs/GaSb strained layer superlattice (SLS) [31], and bolometers. In addition, the quantum dot infrared photodetector (QDIPs) and the quantum dots-in-a-well (DWELL) photodetector are among the most promising alternatives for high background threshold applications due to their beneficial characteristics, which include normal absorption of the incident radiation and low dark current. In addition, QDIP and DWELL detectors benefit from a mature growth and processing technology of III–V semiconductors, which makes it possible to produce devices with good spatial uniformity over a large

---

[2] Separate absorption multiplication APDs are avalanche photodiodes in which the photo-generation of carriers and the avalanche multiplication take place in different layers. The goal of this structure is to provide sufficiently high electric field in the multiplication region to achieve avalanche gain while maintaining sufficiently low electric field in the absorber region to suppress tunneling effects in this layer.

area. This characteristic is essential for fabricating large area focal plane arrays (FPAs).

The structure of a DWELL detector is a hybrid between a conventional quantum well infrared photodetector (QWIP) and the QDIP and benefits from the advantages of both of them [32]. Apart from sensitivity to normal incidence radiation and dark current, the DWELL detector has demonstrated bias tunability and multicolor operation in the mid-wave infrared (MWIR, 3–5μm), long wave infrared (LWIR, 8–12μm) and very long-wave infrared (VLWIR, >14μm). In the DWELL design the quantum dots are placed inside the quantum well and the intersubband transitions are from the dot to the well and from the dot to the quasi-bound state [27].

1.3.1 Avalanche multiplication in midwave infrared detection

It it known that DWELL detectors suffer from low quantum efficiency (QE), which translates into low responsivity and detectivity [27]. Several ways have been suggested to improve the conversion efficiency of DWELL detectors. These ideas include the use of a resonant cavity and the addition of a photonic crystal cavity. One other proposed design involves incorporating gain in the device, the addition of an avalanche photodiode (APD) in conjunction with the DWELL produces a novel sensor called quantum dot avalanche photodiode (QDAP) [33]. In the QDAP, an intersubband quantum dot (QD) detector is coupled with an avalanche photodiode (APD) through a tunnel barrier. The tunnel barrier reduces the dark current while the avalanche section supplies the photocurrent with internal gain. In this three-terminal device, the applied bias of the QD-detector and the APD section of the QDAP are controlled separately. This feature permits the control of the responsivity and dark current of the QD detector independently of the operating avalanche gain.

When operated in Geiger mode the QDAP has the potential for use as a single-photon detector.

1.4 Motivation for this dissertation

1.4.1 State of the art

The type of detector technology used for SPADs is determined by the range of wavelengths of the specific application. In the wavelength range from near infrared to mid infrared (0.7 – 5 μm) the most important materials for single-photon detection are: Silicon and the compounds InGaAs-InP, InGaAsP-InP, HgCdTe. In the following, a brief overview of the most relevant detectors technologies in the wavelengths 0.7 – 5 μm will be presented. Additionally, their state-of-art performance will also be discussed.

Near infrared single photon detectors

For wavelengths shorter than 1 μm Silicon SPADs are the detectors that exhibit the best performance [34]. The PDE for a Si SPAD with an active region diameter of 200 μm under an excess bias of 5 V has a peak of 52 % at 550 nm and it is about 15 % at 820 nm [34]. The DCR decreases almost exponentially with temperature: at a temperature of -25 °C the typical DCR is 5, 50, and 1500 counts/s for SPADs with an active region diameter of 50, 100, and 200 μm, respectively [34]. However, for wavelengths beyond 1 μm the PDE of Si SPADs do not exceed a few percent, due to the low absorption of Si at these wavelengths, and single photon detection using Si SPADs is not longer possible [24]. For the wavelengths beyond 1 μm devices that use the alloy $In_{0.53}Ga_{0.47}As$, from now on referred to as InGaAs, as the absorber layer and

InP as the multiplication region are used. These detectors are the separate absorption multiplication InGaAs/InP SPADs. They cover the 1.1–1.65 μm wavelength interval, which covers the telecommunication range of wavelengths. The typical value of PDE of commercially available InGaAs/InP SPADs is 20 % and the DCR is about 50 kHz measured at 218 K [35]. However, the relatively narrow bandgap of the InGaAs absorber, used to achieve a wavelength cutoff of 1.65 μm, leads to relatively high dark count rates. This performance tradeoff is unnecessary if detection of these longer wavelengths is not desired [36]. Therefore, for application at 1.06 μm the InGaAs absorber layer is changed for the wider bandgap InGaAsP [22, 36]. For InGaAsP/InP SPADs typical values of the photon detection efficiency of is about 30 % with a DCR of 1000 Hz measured at 237 K.

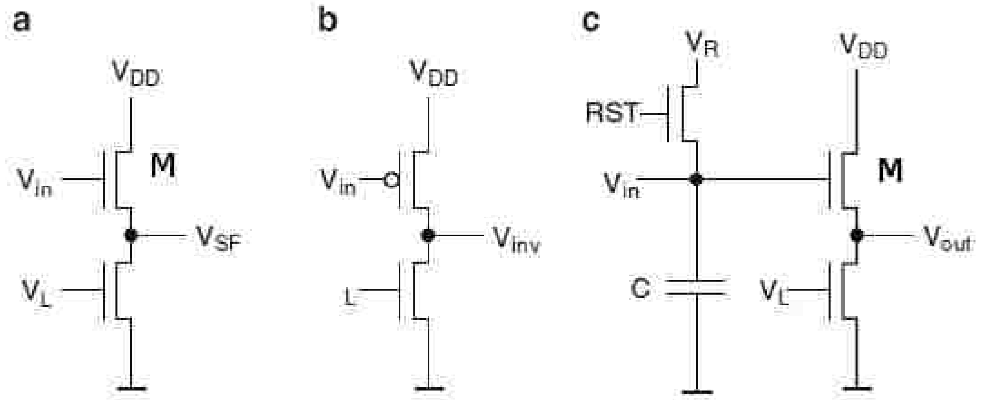


Figure 1.6: Basic circuits for the electronic detection of charge [37]. (a) Source follower with load transistor, (b) Current-sink inverter, (c) Source follower with reset transistor RST and effective input capacitance C.

Mid-Infrared detectors

HgCdTe electron APDs working in linear mode with cutoff wavelengths from 2 μm to 11 μm exhibit single-carrier electron ($k = \beta/\alpha \approx 0$) ionization properties that are a consequence of the band structure and the scattering processes characteristics of HgCdTe [38]. This results in a extremely low, gain-independent, excess noise factor close to unity at 77 K for gains up to greater than 1000. To date, the highest sensitivity that has been demonstrated is 7.5 photon input at a gain of 964. The low excess noise property of HgCdTe, which provides an almost deterministic gain, has motivated the interest of many research groups [39, 40, 41] to develop HgCdTe-based mid infrared single photon imaging systems.

According to the definition proposed by Seitz and Theuwissen [37] "Single-photon imaging is the detection of two-dimensional patterns of low-intensity light, i.e. mean photon numbers in the pixels of less than 10, where the electronic photocharge detection process contributes such little noise that the probability of erroneously reporting a photon where there is none is appreciably smaller than the probability of having at least one photon in a pixel [37]." To achieve such low levels of noise the read-out integrated circuit (ROIC) needs to be carefully designed. Figure 1.6 shows three basic ROICs for the electronic detection of charge. Among these circuits, the one shown in Fig. 1.6c is the more complete pixel circuit [37]. It is based on the source follower scheme, shown in Fig. 1.6a, for the detection of photogenerated charge Q on the gate of the measurement transistor, M. This scheme includes a reset transistor (with reset signal RST) and the effective capacitance C at the gate of the measurement MOSFET. The root-mean-square noise, σ_Q, of this ROIC is [37]

$$\sigma_Q = C\sqrt{\frac{4k_bTB\alpha_R}{g_m}} \tag{1.10}$$

where k_b is the Boltzmann constant, T is the temperature, B is the measurement bandwidth, g_m is the transconductance of the MOSFET, and α_R is a parameter that depends on the operation regime of the MOSFET. In saturation $\alpha_R = 2/3$. Using this source-follower based detection approach, and assuming the following parameter values: $C = 50$ fF, $1/g_m = 1$ kΩ, $T = 300$ K, and $B = 20$ MHz; the thermal-noise limited charge measurement resolution is $\sigma_Q = 4.6$ electrons [37].

The most effective way to reduce the ROIC noise given by (1.10) is to lower the effective capacitance C [37]. One way to do this is by employing very small transistors. This approach has led to capacitances C of only a few fF. Another possibility is to reduce the operating temperature. However, this is not very effective because the absolute temperature appears under the square root of (1.10). The real benefit of lowering the temperature is the reduction of the dark current of the photodetector [37]. The reduction of the measurement bandwidth B is a practical and successful approach to single-electron photocharge detection. However, reducing the output bandwidth of the image sensor will necessarily decrease the system's frame-rate [37]. An alternative to circumvent this problem is to provide the image sensor with multiple output channels, each of them operating at a reduced bandwidth.

One way to relax the noise requirement on the ROIC is to employ physical amplification mechanisms to produce more than one charge per incoming photon. In particular, single photon avalanche photodiodes have received a lot of attention by the single photon imaging community due to their ability to provide internal gain. Today, HgCdTe avalanche photodiodes are the photodetector of choice to be used in mid infrared single photon imaging systems [39, 40, 41]. Beck *et al.* [39] reported a gated-mode infrared imaging system based on a 128x128 FPA that uses HgCdTe APDs and a custom designed ROIC, which shows an rms noise of 100 electrons. The authors reported median gains as high as 946 at 11 V bias with noise equivalent photon inputs as low as 0.4 photon at 80 K and 1 μs gate times. The gated-mode operation of this system was demonstrated at ranges out to 9 km. As a future work the authors aim to reduce the ROIC/system noise, which translates into lower gain requirements on the APDs. Another example is the FPA of the James Webb Space telescope, called the NIRSpec, which has two 2048x2048 HgCdTe arrays based on the Teledyne HAWAII-2RG ROIC, which shows a readout noise of 6 electrons rms per 1008 second exposure. The median readout noise of the HAWAII-2RG at 100 kHz pixel readout rate at a wavelength of 2.5 μm is ≤ 18 electrons.

1.4.2 Prior work

Modeling methods: Metrics of performance

The development of models to describe the performance of SPADs has been investigated by many authors. Kang *et al.* [21] developed a model to calculate the dark count probability and the single-photon quantum efficiency (SPQE) of SPADs. The SPQE is calculates as SPQE $= (P_{on} - P_d)/P_{ph}$, where P_d is the dark count probability, which can be determined as $P_d = 1 - exp(-N_d P_a)$ where P_a is the probability of a carrier to cause an avalanche and N_d is the total number of dark carriers in the multiplication region. The authors identify four sources of dark carriers that contribute to P_d and relate them with the operating condition and the physical parameters of the device such as transit time, gain bandwidth product, detrap time constant, etc. The model assumes the probability of a carrier to cause an avalanche, P_a, as a parameter independent of the structure and electric field profile. In a more material specific work Karve *et al.* investigated the origin of dark counts in $\mathrm{In_{0.53}Ga_{0.47}As/In_{0.52}As}$ SPADs with a multiplication region of 400 μm. This thin multiplication region gives a very good timing performance. However, as a result of the thin multiplication region the dark count rate is dominated by band-to-band tunneling due to the high electric field. Based on experimental data, Donnelly *et al.* [23] developed a model that can predict the DCR and the PDE of a 1.06 μm InGaAsP–InP SPAD as a function of the overbias and temperature. By fitting the parameters of the model to the experimental data they predicted that tunneling through defects in the avalanche region is an important contributor to DCR even at room temperature and dominates at lower temperatures. The model also predicts that in general Geiger-mode performance is better in devices with thicker avalanche regions. The reason for this enhancement is that in wider avalanche regions, for a given PDE, a substantially lower DCR can be achieved compared to that obtained by a thinner avalanche region.

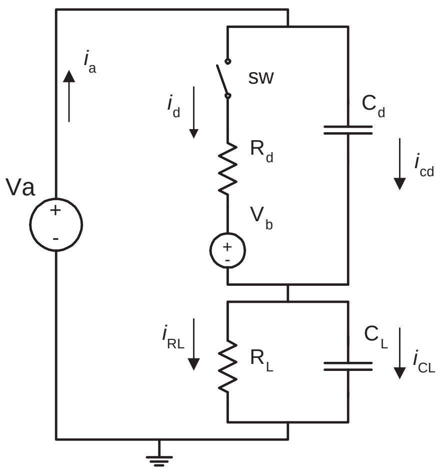


Figure 1.7: Traditional model for a passively quenching SPAD circuit. i_d represents the self sustaining current through the multiplication region of the SPAD; R_d is its equivalent dynamic resistance; C_d is its junction capacitance; R_L is the load resistor and C_L is its parasitic capacitance. The traditional model neglects the effect of feedback on the impact ionization process; it assumes that after the trigger of an avalanche, the electric field remains constant at the breakdown threshold, so that the core of the device is represented by a voltage generator, V_b.

Modeling methods: Traditional model for passively quenched SPADs

The most accepted model that describes the quenching characteristics of passively quenched SPADs was proposed by Haitz in 1964 [42]. In this model, the SPAD is represented by its depletion capacitance, C_d, in parallel with a series combination of a switch, *sw*, a dynamic resistance, R_d, and a DC bias source, V_b, representing the breakdown voltage of the SPAD. This model is shown in Fig. 1.7. In the absence of an avalanche trigger the switch is open and the bias across the diode is V_a, which is set slightly above breakdown. When an avalanche is triggered the model assumes that the switch is instantly closed and the capacitance C_d discharges through the diode's dynamic resistance R_d, which reduces the voltage across the SPAD to a value that depends on the ratio of R_d and R_L. In steady state, the voltage across the SPAD is given by $V_{SPAD} = V_a - I_{R_L} R_L \approx V_b$, the breakdown voltage of the device.

The presence of the DC source, V_b, in the traditional model implicitly assumes that after an avalanche event is triggered the electric field in the avalanche region (responsible for impact ionizations to persist) remains precisely at the breakdown level until the persistent current is quenched due to the stochastic fluctuations inherent in the impact ionization process, that is when all carriers chance to exit the multiplication region without ionizing.

A major concern about the traditional model, which arises from the constant field assumption, is that it predicts that the quenching time should have memory. In fact, a quenching process with memory is not observed in self quenched SPAD circuits. Figure 1.8a shows measurements of the probability density function (pdf) of the quenching time showing exponential decay, which implies that the decay process is memoryless. (The data was provided by Princeton Lightwave Inc.) In addition, The traditional model fails to predict the oscillatory behavior in persistent current also observed by Itzler *et al.* [45, 43] and shown in Fig. 1.8b. Therefore, there is

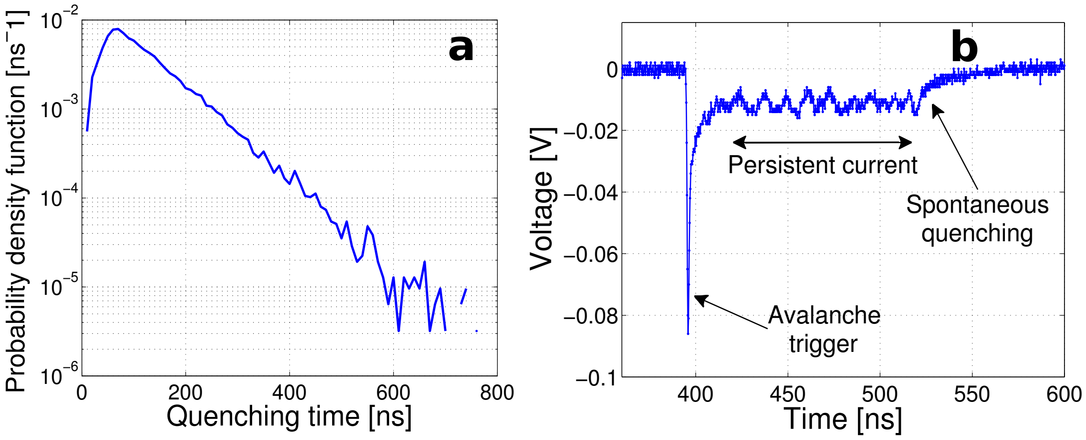


Figure 1.8: (a) Measured pdf of the quenching time [43]. The exponential decay of the pdf implies that the quenching time is memoryless. (b) Measured voltage across the SPAD for an excess bias of $V_{ex} \approx 1.7$ V [43]. The current shows oscillatory behavior about the steady state before it quenches spontaneously. The complete structure of the device can be found elsewhere [44].

the need for a more extensive model beyond the traditional model, which is able to explain the behavior observed in the new generation of SPADs. This dissertation introduces such a model.

New generation of SPAD structures: self-quenched SPADs

The reduction of the quenching time in passively quenched SPADs offers several desirable features. Foremost, short quenching times reduce the total charge flow during an avalanche event, and hence reduce the fraction of carriers trapped at defect sites in the SPAD. This, in turn, leads to reduced afterpulsing, which is one of the most severe drawbacks of SPADs as it limits the recovery time needed to allow the trapped carriers to be released so that the SPAD is reset afresh to detect a new photon.

Recently, new SPAD structures capable of achieving very short quenching times have been developed. These devices are the negative feedback avalanche diode (NFAD) [45] and the self-quenching and self-recovery avalanche detector [46]. The operation of self-quenching SPADs and in particular NFADs and self-quenching and self-recovery avalanche detectors rely heavily on the introduction of negative feedback, which rapidly lowers the internal electric field of the avalanche diode following buildup of the avalanche current and forces the stochastic avalanche to terminate quickly. Presently, there is no model that is capable of predicting the statistics of the stochastic quenching time in passively quenched SPADs in general and the new generation of SPAD structures that heavily exploit the negative feedback effect in particular. In this dissertation, a new fundamental model is developed, which helps us to understand how NFADs work, and assists us in their optimization.

1.5 Contributions of this dissertation

1.5.1 Theoretical modeling of SPADs

The main focus of this dissertation on the theoretical analysis of APDs is to develop new modeling methods that assist the design and characterization of high performance APDs working in linear and Geiger modes. The first contribution is a model that sheds light on the dependence of the performance of SAM SPADs on the width of the multiplication region by comparing the effects of field-assisted tunneling with temperature-assisted dark carriers as the width is varied. This model captures the effect of the dead space and heterojunction multiplication regions. An aspect of importance that had not been explored before, namely, the random locations where carriers are born in each layer, will be thoroughly analyzed and studied. In particular, it is assumed that photogenerated carriers are generated in the absorber at random locations according to an exponential pdf. On the other hand, dark carriers are assumed to be generated randomly in the multiplication region and the absorber according to a uniform pdf in each layer. In addition, along with the DCR and PDE, the single-photon quantum efficiency (SPQE) will be also used as a figure of merit to assess the SPAD's performance [21]. The ability of the SPQE to admit an optimal operating overbias makes it a very useful metric [47].

The second contribution is the development of a stochastically self-regulating avalanche model for passively quenched SPAD, which is the first significant expansion beyond the model presented by Haitz in 1964 [42]. Specifically, the stochastically self-regulating avalanche model reported in this dissertation addresses three important phenomena that are entirely beyond the scope of the traditional modeling methods. First, it predicts the existence of an oscillatory behavior of a persistent avalanche current. Second, it predicts that the probability density function of the stochastic quenching time of the persistent avalanche current has an exponential decay. Third, under device and operational conditions that lead to strong feedback, the stochastic avalanche current can collapse before persistent avalanche current can be realized. All three of these behaviors are in agreement with recent experimental demonstrations employing negative-feedback SPADs (NFADs) that had until now not been theoretically explained. The model specifically captures the effect of the load's feedback on the stochastic avalanche multiplication, an effect believed to be key in breaking today's counting rate barrier in the $1.55-\mu$m detection window.

In addition, it will be shown that the traditional model for passively quenched SPADs fails to determine the quenching characteristics of passively quenched SPAD. In particular, it will be shown that the *constant-field assumption*, namely the electric field in the avalanche region remaining at the breakdown level while the persistent current is ongoing, implies the unrealistic consequence that the quenching time, T_q, has memory. This unrealistic consequence is a result of the simplistic, deterministic approach adopted in the traditional model, which neglects the effect of the feedback on the stochastic impact ionization process.

1.5.2 Mid-infrared: Quantum dot avalanche photodiode

The experimental part of this dissertation focuses on the development of a GaAs-based midwave infrared photodetector. The device exploits the impact ionization mechanism to increase the conversion efficiency of III–V based semiconductor detectors. The goal is to pave the path to realize single photon detectors (SPADs) with III–V based semiconductors operating beyond 2 μm. Presently, there are no III–V based semiconductor SPADs available beyond 2 μm. However, there are many applications ranging from biomedical imaging, astronomy to laser detection and ranging that require SPADs in the MWIR (3–5 μm) range. In this dissertation, it is demonstrated the linear mode operation of a GaAs-based avalanche photodiode (APD) at 5 μm. The device, called the QDAP, exploits quantum confined transitions to obtain MWIR absorption and couples the photogenerated carriers into an APD to provide multiplication. A conversion efficiency of 12% is obtained. This is also the longest wavelength APD reported with III–V semiconductors and opens up the exciting possibility of realizing SPADs for these wavelengths.

1.6 Outline of this dissertation

This dissertation is divided into two parts, Parts I and II. The first part, which includes from Chapter 2 until Chapter 4, is devoted to the modeling of avalanche photodiodes operating in linear and Geiger modes. On the other hand, Part II, which includes Chapter 5, deals with the implementation of a new midinfrared photodetector called the quantum dot avalanche photodiode.

In Chapter 2, new modeling methods are developed to design and characterize the performance of *separate absorption multiplication* (SAM) SPADs. The model uses the recursive dead-space multiplication theory (DSMT) according to a field-dependent spatial distribution of carriers to calculate the generalized breakdown probability for all the carriers generated in the SPAD. The characterization of the devices is made by calculating the performance metrics: *photon detection efficiency*, the *dark count rate* and also the *single-photon quantum efficiency* as a function of the width of the multiplication region, the applied voltage and the temperature of operation. Moreover, An aspect of importance that had not been explored before, namely, the random locations where carriers are born in each layer, is thoroughly analyzed and studied. In particular, it assumed that photogenerated carriers are generated in the absorber at random locations according to an exponential pdf. On the other hand, dark carriers are assumed to be generated randomly in the multiplication region and the absorber according to a uniform pdf in each layer.

Chapter 3 explores the characteristics of the impact ionization process under the influence of an time-varying electric field. It is shown rigorously that a sinusoidal biasing scheme that is synchronous with the optical pulse stream and has a properly selected DC level, peak-to-peak value and phase, can offer an increase in the effective gain-bandwidth product of the APD. Presently, to meet the demand of the exponential growth in data transfer, the telecommunication industry has been moving toward 40-Gbps and 100-Gbps protocols for their core fiber-optic backbone networks alongside the existing 10-Gbps infrastructure operating at the low-loss wavelength of 1.55 μm. However, the limited speed of APD-based receivers have limited their use in systems that operate at 2.5 and 10 Gbps. The proposed biasing scheme represents a promising effort to enable the current InP-based APDs to meet the expectations of 40 Gbps systems.

Chapter 4 deals with the models for passively quenched SPADs. Three models are analyzed: (1) The first model is the traditional model, which assumes that after an avalanche trigger the voltage across the SPAD remains at a fixed value. It will be shown that this assumption leads to unrealistic consequences. (2) The second model to be examined is a deterministic self-regulating model. This model captures the effect of the feedback from the load in the current-voltage characteristics of a passively quenched SPAD. However, the stochastic nature of the impact ionization process is neglected. (3) The third model is a stochastically self-regulating avalanche model, which captures the dynamic effect of the feedback from the load in the stochastic nature of the impact ionization process. The stochastically self-regulating avalanche model represents the first significant expansion beyond the Haitz model [42], for passively quenched SPADs, since it was proposed more than 45 years ago. It will be shown that the proposed model predicts, the important phenomena observed in NFADs that traditional models are unable to predict.

In Chapter 5, a new midwave infrared photodetector is presented, the Quantum-dot avalanche photodiode (QDAP). In the QDAP, an intersubband quantum dots-in-a-well (DWELL) detector is coupled with an APD through a tunnel barrier. The idea behind the QDAP is to increase the conversion efficiency of GaAs-based midinfrared detectors by providing avalanche multiplication gain. In the operation of the QDAP the photon absorption and generation of carriers take place in the DWELL section while the avalanche section of the device provides internal gain. It is shown that the conversion efficiency of the DWELL detector is increased by a factor of 14 due to the gain introduced by the avalanche multiplication stage of the QDAP.

1.7 List of publications

1.7.1 Refereed Journal publications

1. **D. A. Ramirez**, M. M. Hayat, Graham J. Rees, Xudong Jiang, Mark A. Itzler, "New perspective on passively quenched single photon avalanche diodes: effect of feedback on impact ionization," submitted to Optics Express.

2. M. M. Hayat, **D. A. Ramirez**, "Theory of dynamically biased avalanche photodiodes: new limits for gain bandwidth product," (under preparation).

3. **D. A. Ramirez**, J. Shao, M. M. Hayat, and S. Krishna, "Midwave infrared quantum dot avalanche photodiode," Applied Physics Letters, vol. 97, p. 221106, Dec. 2010.

4. **D. A. Ramirez**, M. M. Hayat, and M. A. Itzler, "Dependence of the Performance of Single Photon Avalanche Diodes on the Multiplication Region Width," to IEEE J. Quantum Electronics, vol. 44, pp. 1188–1195, 2008.

5. T. E. Vandervelde, M. Lenz, E. Varley, A. Barve, J. Shao, R. V. Shenoi, **D. A. Ramirez**, W. Jang, Y. Sharma, and Sanjay Krishna, "Quantum Dots-in-a-Well Focal Plane Arrays," IEEE Journal of Selected Topics in Quantum Electronics, Volume 14, pp. 1150–1161, 2008.

6. E. Varley, M. Lenz, S. J. Lee, J. S. Brown, **D. A. Ramirez**, A. Stintz, and S. Krishna, Axel Reisinger and Mani Sundaram, "Single bump,two-color quantum dot camera," Applied Physics Letters 91, 081120, 2007.

7. **D. A. Ramirez**, M. M. Hayat, S. N. Torres, B. E. A. Saleh, and M. C. Teich, "Information-Theoretic Criterion for the Performance of Single-Photon Avalanche Photodiodes," IEEE Photonic Technology Lett., vol. 17, pp. 2164–2166, 2005.

8. **D. A. Ramirez**, M. M. Hayat, G. Karve, J. C. Campbell, S. N. Torres, B. E. A. Saleh, and M. C. Teich, "Detection Efficiencies and Generalized Breakdown Probabilities for Nanosecond-Gated Near Infrared Single-Photon Avalanche Photodiodes," IEEE J. Quantum Electronics, vol. 42, pp. 137–145, 2006.

1.7.2 Conferences, proceedings and abstracts

1. **D. A. Ramirez**, J. Shao, M. M. Hayat, and S. Krishna, "Linear-mode operation of the quantum dot avalanche photodiode (QDAP)," Phys Status Solidi C, 7, No. 10, pp. 2548–2551, 2010.

2. M. M. Hayat, **D. A. Ramirez** "Sensitivity of high-speed receivers using InAlAs avalanche photodiodes," 23rd Annual Meeting of the IEEE Photonics Society (formerly LEOS), pp. 126–127, 2010.

3. M. M. Hayat, **D. A. Ramirez** "Impact ionization under dynamic field," 23rd Annual Meeting of the IEEE Photonics Society (formerly LEOS), pp. 128–129, 2010.

4. M. M. Hayat, **D. A. Ramirez**, G. J. Rees, and M. A. Itzler, "Modeling Negative Feedback in Single-Photon Avalanche diodes," SPIE Proceedings Vol. 7681, p. 76810W, 2010.

5. M. M. Hayat, M. A. Itzler, **D. A. Ramirez**, and G. J. Rees, "Model for Passive Quenching of SPADs," SPIE Proceedings Vol. 7608, p. 76082B, 2010.

6. **D. A. Ramirez**, J. Shao, M. M. Hayat, and S. Krishna, "Recent Progress on the Quantum-dot Avalanche Photodiode (QDAP)," 22nd Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), pp 295–296, 2009.

7. **D. A. Ramirez**, J. Shao, M. M. Hayat, S. Krishna, "Demonstration of the

quantum dot avalanche photodiode (QDAP)," Proceedings of SPIE Vol. 7320 (SPIE, Bellingham, WA 2009) 73200L.

8. S. Krishna, A. V. Barve, R. V. Shenoi, W-Y. Jang, S. Myers, H. S. Kim, **D. A. Ramirez**, J. Montoya, M. N. Kutty, A. Khoshakhlagh, N. Gautam, J. Shao, Y. Sharma, E. Plis, S. J. Lee and S. K. Noh, Infrared Retina using Quantum Dots in a Well and Type II InAs/GaSb Strained Layer Superlattice Detectors, the 34th International Conference on Infrared, Millimeter, and Terahertz Waves (IRMMW-THz), September 21-25th, 2009, Busan, Korea (Keynote).
9. M. M. Hayat, G. J. Rees, **D. A. Ramirez**, and M. A. Itzler, "Statistics of Self-quenching Time in Single Photon Avalanche Diodes," 21st Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), pp. 230–231, 2008.
10. Thomas E. Vandervelde, Michael C. Lenz II, Eric Varley, Ajit Barve, Jiayi Shao, Rajeev Shenoi, **David A. Ramirez**, Wooyong Jang, Yagya Sharma, and Sanjay Krishna, "Multicolor quantum dots-in-a-well focal plane arrays" Infrared Technology and Applications XXXIV Proc. SPIE Vol.(6940), PP694003, 2008.
11. **D. A. Ramirez**, M. M. Hayat, and M. A. Itzler, "Dependence of the Performance of Single Photon Avalanche Diodes on the Multiplication Region Width: Temperature and Field Effects," 20th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), pp. 511–512, 2007.
12. R. V. Shenoi, **D. A. Ramirez**, Y. Sharma, R. S. Attaluri, J. Rosenberg, O. J. Painter, and S. Krishna, "Plasmon Assisted Photonic Crystal Quantum Dot Sensors" in Nanophotonics and Macrophotonics for Space Environments, edited by Edward W. Taylor, David A. Cardimona, Proceedings of SPIE Vol. 6713, p. 67130P (2007).
13. Eric S. Varley, **David A. Ramirez**, Jay S. Brown, Sang Jun Lee, Andreas Stintz, Michael Lenz, Sanjay Krishna, Axel Reisinger and Mani Sundaram,

"Demonstration of a two-color 320x256 quantum dots-in-a-well focal plane array" in Infrared Spaceborne Remote Sensing and Instrumentation XV, Proceedings of SPIE Vol. 6678, p. 66780T, 2007.

14. **D. A. Ramirez**, M. M. Hayat, Gauri Karve, J. C. Campbell, Sergio N. Torres, B. E. A. Saleh and M. C. Teich, "Detection Efficiencies for InP, InAlAs, and InAlAs-InP Single-Photon Avalanche Photodiodes," 18th Annual Meeting of the IEEE Lasers and Electro-Optics Society (LEOS), pp. 387–388, 2005.

Part I

Modeling of avalanche photodiodes

Chapter 2

Optimization of the width of multiplication region

2.1 Introduction

Recently, there has been a growing interest in optimizing the width of the multiplication layer for the best *photon detection efficiency* (PDE) and *dark count rate* (DCR) performance [23]. As mentioned in the introduction of this dissertation, the PDE and DCR are two of the most important performance metrics for SPADs. The PDE is the product of the detector quantum efficiency and the avalanche breakdown probability. The DCR constitutes false counts and it is a measure of how noisy the detector is. Dark counts originate from dark carriers generated in the absence of illumination; the larger the number of dark carriers, the larger the dark-count rate is [21]. There are several mechanisms that contribute to the concentration of dark carriers. At high electric fields, the dark-carrier concentration is strongly affected by band-to-band tunneling, which depends exponentially on the electric field [22]. Another important mechanism that contributes to the number of dark carriers is tunneling through defects. It has been reported that in some materials the tunneling currents due to defects concentration is higher than that of band-to-band tunneling [23].

Besides the strength of the electric field and the properties of the material, the probability of a carrier triggering an avalanche breakdown is determined by the place where it is born [26]. A carrier created in the start of the multiplication region has a greater probability of triggering an avalanche event compared to that created close to the end of the multiplication region. This is because a primary carrier created early on in the multiplication region has a larger distance to travel compared to those created close to the end of the multiplication region. For example, for a separate absorption separate multiplication (SAM) APD, a carrier created in the absorption layer is more likely to cause an avalanche compared to that created in the multiplication region. More generally, the dependence of the breakdown probability on the birth location of a carrier is crucial in determining the SPAD's performance when the number of dark carriers inside the absorption and multiplication layers is taken into account. Since the width of the multiplication layer significantly affects the electric field (and hence tunneling current), it is important to have a model that can predict the DCR and PDE required for Geiger-mode operation for various SPAD structures and geometries while taking into account the types of dark carriers and the randomness in the location where they are generated in the absorption and multiplication layers.

The main focus of this chapter is to present theoretical results based on new modeling tools that shed light on the dependence of the performance of SAM SPADs on the width of the multiplication region by comparing the effects of field-assisted tunneling with temperature-assisted dark carriers as the width is varied. This study also reveals the characteristic difference in the performance between low-temperature operation and room-temperature operation while identifying and quantitatively examining the main factors that govern the performance of the SPAD. Moreover, an aspect of importance that had not been explored before, namely, the random locations where carriers are born in each layer, is thoroughly analyzed and studied. In particular, it is assumed that photogenerated carriers are generated in the absorber at random locations according to an exponential probability density function (pdf). On the other hand, dark carriers are assumed to be generated randomly in the multiplication region and the absorber according to a uniform pdf in each layer. To calculate the generalized breakdown probability for all the carriers generated in the SPAD the recursive dead-space multiplication theory (DSMT) according to a field-dependent spatial distribution of carriers is used [26, 47]. In addition, along with the DCR and PDE, the single-photon quantum efficiency (SPQE) is also used as a figure of merit to assess the SPAD's performance [21]. The ability of the SPQE to admit an optimal operating overbias makes it a very useful metric [47]. The theory developed is applied to SPADs that operate in a short-pulse gated-mode regime, in the 1.3–1.55 μm range, with InP multiplication regions and either InGaAs or InGaAsP as absorbers.

2.2 Model

In this section I draw upon existing models for dark current [48, 23], breakdown probability [47, 26] and SPAD-performance metrics [21, 47] to develop new expressions for distributed breakdown probability for SAM SPADs and their performance.

2.2.1 Review of dark current model

The dominant mechanism of dark-carrier generation in a specific SPAD will depend upon its physical structure and operating conditions such as the bias voltage, repetition rate in gated operation, and temperature. In this study dark-carrier generation in both of the absorption and multiplication regions are considered. In both regions the mechanisms to be considered are GR, band-to-band tunneling and tunneling through defect states. Accordingly, the number of dark carriers generated per second in the absorber is $N_{d,abs} = N_{gen,abs} + N_{def,abs} + N_{tun,abs}$. In the same way, the number of dark carriers generated per second in the multiplication region is given by $N_{d,mul} = N_{tun,mul} + N_{def,mul} + N_{gen,mul}$.

The GR current density, which is the dominant mechanism of dark-carrier generation at low voltage, is given by the expression [48]

$$J_{gen} = \frac{qn_iW}{\tau_{\text{eff}}}\left(1 - \exp(qV/2kT)\right) \tag{2.1}$$

where, W is the width of the depletion region, n_i is the intrinsic carrier concentration, V is the applied voltage, and τ_{eff} is the effective carrier lifetime. (The units of J_{gen} are Amperes per square meter.) Thus, the number of dark carriers due to GR is $N_{gen} = J_{gen}A/q$, where A is the SPAD's cross-sectional area, and q is the charge of the electron.

At high electric fields, the dominant mechanism of dark-carrier generation is tunneling [48]. Consequently, tunneling currents become very important for thin multiplication layers [49]. In InGaAs dark carrier generation due to tunneling becomes importance at electric field > 200 kV/cm [50]. Generally, tunneling current increases exponentially as the electric field increases [22]; more precisely [48],

$$J_{tun} = \frac{\sqrt{2m^*}q^3E_mV}{4\pi^2\hbar^2E_g^{1/2}}\exp\left(-\frac{\theta\sqrt{m^*}E_g^{3/2}}{qE_m\hbar}\right) \tag{2.2}$$

where E_m is the electric field, V is the voltage across the avalanche region, m^* is the electron effective mass, and θ is a parameter that depends on the shape of the tunneling barrier. As in the case of N_{gen}, the number of dark carriers in the avalanche region due to band-to-band tunneling is $N_{tun} = J_{tun}A/q$.

Defects in the material also contribute to increase the dark-carrier generation [25, 23, 22]. The tunneling current density due to defect states is given by the expression [23]

$$J_{def} = \frac{A_d E_m V N_T \exp\left(\frac{-(B_1 E_{B1}^{3/2} + B_2 E_{B2}^{3/2})}{E_m}\right)}{N_v \exp\left(\frac{-B_1 E_{B1}^{3/2}}{E_m}\right) + N_c \exp\left(\frac{-B_2 E_{B2}^{3/2}}{E_m}\right)} \tag{2.3}$$

where $A_d = q^3 \sqrt{\frac{2m_r}{E_g}}/(4\pi^3\hbar^2)$, $m_r = \frac{2(m_c m_{lh})}{m_c + m_{lh}}$ is the reduced effective mass, m_c being the conduction band effective mass and m_{lh} being the light hole effective mass, $B_1 = \pi(m_{lh}/2)^{1/2}/(2q\hbar)$, and $B_2 = \pi(m_c/2)^{1/2}/(2q\hbar)$. In the above expression, E_{B1} is the barrier height of tunneling from valence band to trap and is equal to aE_g $(a < 1)$, and E_{B2} is the barrier height of tunneling from trap to the conduction and is equal to $(1 - a)E_g$. The quantities N_v and N_c are the light hole valence and conduction band density of states and N_T represents the number of defects per unit volume [23]. The number of dark carriers in the avalanche region due to defects states is $N_{def} = J_{tun}A/q$. The average number of dark carriers generated in the SPAD is given by

$$N_d = N_{d,mul} + N_{d,abs} \tag{2.4}$$

2.2.2 Calculation of breakdown probability

In order to apply the DSMT to calculate the generalized breakdown probabilities for all the carriers generated in the SPAD illustrated in Fig. 2.1 [47, 26], it has been used (i) the nonlocalized ionization coefficients, also called enabled ionization coefficients (the ionization coefficient assumed once the carrier travels the dead-space distance), and the threshold energies for each material [51], and (ii) the electric-field profile through the device. The non-localized electron and hole ionization coefficients and threshold energies for InP, InGaAs and InGaAsP are readily available [51, 52, 53].

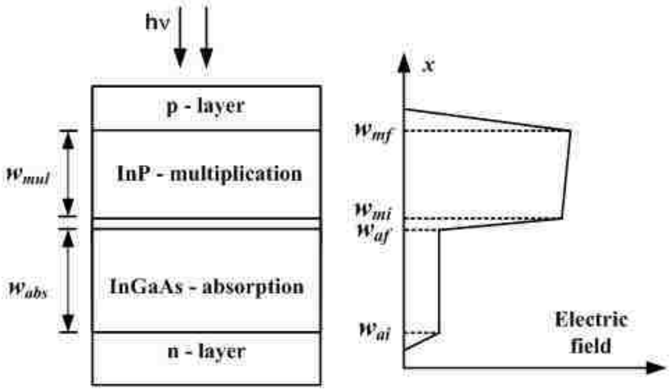


Figure 2.1: Device structure and electric-field profile of a SAM SPAD with InP multiplication region and InGaAs absorber.

Probability density function of the free path

The following shifted-exponential model for the probability densities of the distance to ionization, y, measure from the location, x, where a carrier is born is adopted [26]. The mathematical expressions presented in this section are further explained in the Appendix A. For an electron born at location x, with α being the enabled ionization coefficient, the probability that it impact ionizes at location y and assuming that electrons move to the direction of increasing x is [54]

$$h_e(y|x) = \alpha(y)\exp\left(-\int_{x+d_e(x)}^{y}\alpha(u)\,du\right), \quad y \geq x + d_e(x) \tag{2.5}$$

where $d_e(x)$ is the dead space of an electron born at location x, and $h_e(y|x) = 0$ when $y < x + d_e(x)$. The dead space is the distance a carrier must travel within the SPAD before acquiring the energy threshold needed for effecting an impact ionization; $d_e(x)$ satisfies the equation [54]

$$E_{th,e}\big(x + d_e(x)\big) = q\int_{x}^{x+d_e(x)} E(u)\,du, \tag{2.6}$$

where $E_{th,e}(x)$ is the ionization threshold energy for electrons at location x in the APD (this energy varies from layer to layer). Equation (2.6) neglects the effect of scattering, which, as described in Chapter 1, increases the dead space. However, (2.6) gives a good approximation of the dead space for high electric fields ($> 1 \times 10^5$ V/cm) where the carriers in the depleted multiplication region can gain energy from the electric field at a faster rate than they lose it to the various scattering processes. There are similar expressions for holes. For an hole born at location x, with β being the enabled ionization coefficient, the probability that it impact ionizes at location y and assuming that holes move to the direction of decreasing x is

$$h_h(y|x) = \beta(y) \exp\left(-\int_y^{x-d_h(x)} \beta(u) du \right), \quad y \leq x - d_h(x) \tag{2.7}$$

where $d_h(x)$ is the dead space of a hole born at location x, and $h_h(y|x) = 0$ when $y > x - d_h(x)$. The hole dead space satisfies the following equation

$$E_{th,h}\big(x - d_h(x)\big) = q \int_{x-d_h}^{x} E(u)\, du, \tag{2.8}$$

In the case where the field is constant, and scattering effects are neglected, the position-independent dead space is calculated using $d = E_{th}/qE$ [51]. The equations from (2.5) to (2.8) of the DSMT are generalized equations; they constitute a powerful tool that allows us to model APDs with any electric-field profile and any structure, like multilayer devices with heterostructure multiplication regions to be reviewed next.

Breakdown probability

Suppose that we know the total electron and hole population, $Z(x)$, resulting from a parent electron born at x, and the total electron and hole population, $Y(x)$, resulting from a parent hole born at x, where $0 \leq x \leq w$, and w is the width of the SPAD. Let us define $P_Z(x)$ as the probability that $Z(x)$ is finite, and similarly,

$P_Y(x)$ as the probability that $Y(x)$ is finite [26]. These quantities reflect the non-breakdown probabilities for carriers generated at location x anywhere in the SPAD. Thus, for example $1 - P_Z(x)$ is the probability that $Z(x)$ is infinite, which is precisely the case when avalanche breakdown occurs. On the other hand, the probability that an electron-hole pair born at x collectively triggers an avalanche breakdown is $P_b(x) = 1 - P_Z(x)P_Y(x)$. Recursive integral equations describing $P_Z(x)$ and $P_Y(x)$ are developed elsewhere [26] and are repeated here for completeness:

$$P_Z(x) = \int_{w-x}^{\infty} h_e(\xi|x)\, d\xi + \int_0^{w-x} \Big(P_Z^2(x+\xi) P_Y(x+\xi) \Big) h_e(\xi|x)\, d\xi \tag{2.9}$$

$$P_Y(x) = \int_{x}^{\infty} h_h(\xi|x)\, d\xi + \int_0^{x} \Big(P_Y^2(x-\xi) P_Z(x-\xi) \Big) h_h(\xi|x)\, d\xi \tag{2.10}$$

These integral equations can be solved using a straightforward numerical iterative approach similar to that described in [26].

Let us assume that the electron-hole pairs are created at random locations in the absorption and multiplication regions extending from $x = w_{ai}$ to $x = w_{mf}$, as shown in Fig. 2.1. It is also assumed that holes (electrons) are transported in the positive (negative) x direction. Moreover, let $f(x)$ denote the pdf of the birthplace of the parent electron-hole pair. Thus, the average probability that an electron-hole pair, randomly generated in the interval $[w_{ai}, w_{mf}]$ according to the pdf $f(x)$, triggering an avalanche breakdown is given by

$$Q_f = \int_{w_{ai}}^{w_{mf}} f(x)\Big(1 - P_Z(x)P_Y(x)\Big)\, dx. \tag{2.11}$$

The expression for Q_f represents the general form of the breakdown probability for any random distribution of carriers and it accounts for avalanche breakdown occurring either in the absorption or multiplication regions. We can further specialize this expression for two distinct forms of f representing the following physical scenarios: (a) the scenario for which the avalanche breakdown is triggered by electron-hole pairs photo-generated inside the absorption region, in which case f is denoted by f_{ph}; and (b) the scenario for which the avalanche breakdown is initiated by dark carriers randomly generated in either the absorption or the multiplication region, in which case f is denoted by f_d. The use of f_{ph} and f_d in (2.11) will lead to the *injected-carrier breakdown probability*, Q_{ph}, and the *distributed-carrier breakdown probability*, Q_d, respectively. The former represents the breakdown probability caused by a carrier pair photogenerated in the absorber; on the other hand, the latter represents the breakdown probability caused by a dark carrier that is randomly generated inside the SPAD, taking into account the dark carriers generated in the multiplication and the absorption regions.

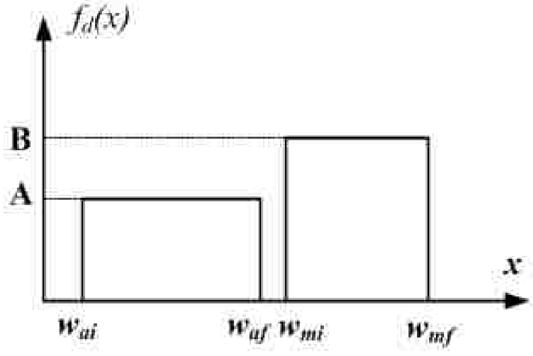


Figure 2.2: Probability density function f_d for the case where the avalanche breakdown is triggered by dark carriers randomly generated in the SPAD.

Let us consider first the case where the avalanche breakdown is triggered by dark carriers randomly generated in the SPAD. In this case the pdf f is given by, as depicted in Fig. 2.2,

$$\begin{aligned} f_d = & \frac{A}{w_{af}-w_{ai}}\Bigg(u(x-w_{ai})-u(x-w_{af})\Bigg) \\ & +\frac{B}{w_{mf}-w_{mi}}\Bigg(u(x-w_{mi})-u(x-w_{mf})\Bigg) \end{aligned} \tag{2.12}$$

where $A = \frac{N_{d,abs}}{N_{d,abs}+N_{d,mul}}$, $B = \frac{N_{d,mul}}{N_{d,abs}+N_{d,mul}}$, and $u(x)$ is the unit step function. Note that A (resp. B) is the probabilities that an arbitrary dark-carrier pair already generated in the SPAD was actually generated in the absorption (resp. multiplication) region. The quantities w_{mi} and w_{mf} respectively represent the start and end of the multiplication region, and w_{ai} and w_{af} respectively represent the start and end of the absorption region, where it has been assumed $(w_{mf}-w_{mi})+(w_{af}-w_{ai}) \approx w$ (see Fig. 2.1). (In all the calculations the effect of the charge layer, which is between the absorber and the multiplication region, has been neglected leading to the approximation $w_{af} \approx w_{mi}$.) Consequently, the probability Q_d simplifies to

$$\begin{aligned} Q_d &= \frac{N_{d,abs}}{N_{d,abs}+N_{d,mul}}\frac{1}{w_{af}-w_{ai}}\int_{w_{ai}}^{w_{af}}\big(1-P_Z(x)P_Y(x)\big)\,dx \\ &+ \frac{N_{d,mul}}{N_{d,abs}+N_{d,mul}}\frac{1}{w_{mf}-w_{mi}}\int_{w_{mi}}^{w_{mf}}\big(1-P_Z(x)P_Y(x)\big)\,dx. \end{aligned} \tag{2.13}$$

In the case of the injected-carrier breakdown probability, the absorption of photons in the absorption region obeys an exponential behavior. Hence, f_{ph} will be of the form $f_{ph}(x) = C_1 e^{-C_2 x}$, for $w_{ai} \leq x \leq w_{af}$, and $f_{ph} = 0$ elsewhere. For simplicity, $x = w_{ai} = 0$ and therefore $w_{af} = w_{abs}$, which is the width of the absorber (see Fig. 2.1). The constant C_2 can be determined by equating $\int_0^{w_{abs}} C_2 e^{-C_2 x}\,dx$ to the SPAD's quantum efficiency, η. This yields $C_2 = -\ln(1-\eta)/w_{abs}$. The constant C_1, on the other hand, is chosen so that f_{ph} has unit area, as we would expect from a valid pdf; this yields $C_1 = C_2/\eta$. In summary, the injected-carrier breakdown probability is given by

$$Q_{ph} = \frac{-\ln(1-\eta)}{\eta w_{abs}} \int_0^{w_{abs}} \exp\left\{\frac{\ln(1-\eta)}{w_{abs}} x\right\} \times \Big(1 - P_Z(x)P_Y(x)\Big)\, dx. \tag{2.14}$$

2.2.3 SPAD performance

The traditional performance metrics, photon-detection efficiency and the dark-count rate are respectively defined as $PDE \equiv \eta Q_{ph}$ and $DCR \equiv N_d Q_d$. Additionally, the single-photon quantum efficiency is another useful metric to assess the SPAD performance [21, 47]. The latter is defined as the probability that a photon triggers an avalanche breakdown, given that an optical pulse is present and at least one photon impinges on the SPAD, and provided that no dark carrier triggers a breakdown. Mathematically, it is given by

$$SPQE = \frac{(1-P_d)P_{opt}}{p_o}, \tag{2.15}$$

where P_d is the dark count probability, calculated throughout the absorption and the multiplication regions altogether. The quantity P_{opt} is the probability that at least one photogenerated carrier in the absorber triggers the avalanche, and p_o is the probability that one photon impinges on the SPAD during the detection time. The dark count probability is given by

$$P_d = 1 - e^{-N_d Q_d}, \tag{2.16}$$

where N_d is the average number of dark carriers generated in the SPAD (calculated in (2.4)). Note that in Kang *et al.* [21], the breakdown probability Q_{ph} is used in place of Q_d; however, the use of Q_d, as done here, accounts for dark-carrier generation at random locations across the entire device. The quantity P_{opt} is calculated using the following expression

$$P_{opt} = 1 - e^{-\eta Q_{ph} N_o}, \tag{2.17}$$

where η is the detector quantum efficiency and N_o is the average number of photons per pulse.

2.3 Results

The theory described in the previous section is applied to SAM SPADs with InP homojunction multiplication regions and InAlAs-InP heterojunction multiplication regions. In both cases the absorber layer is InGaAs. In the case where GR dark carriers are included the operating temperature is 300 K. Figure 2.1 illustrates the structure and the electric-field profile of the SPAD with InP homojunction multiplication region. A schematic of the electric-field profile across the device is also shown.

To see the role played by the width of the multiplication region on the performance of the SPAD, the PDE, DCR and SPQE curves have been calculated, as the width of the multiplication region is varied, considering two scenarios: (i) when the dominant dark-carrier-generation mechanism is field-assisted and (ii) when the dominant mechanism of dark-carrier generation is temperature assisted. The comparison of the performance of the SPAD under these scenarios will illustrate the characteristic difference in the performance between low-temperature operation and room-temperature operation and how this attribute varies as the multiplication-region width is changed.

2.3.1 InGaAs/InP homojunction SAM photodiode

The DCR, PDE and SPQE for a SAM SPAD with InP homojunction multiplication region of width in the range 500–2000 nm and absorber of 1μm were calculated. It is expected that the effect on the number of dark carriers, and hence on the DCR, of the temperature-assisted generation of dark carriers is more relevant at lower bias voltages since as we increase the bias voltage the dark-carrier generation will be dominated by field-assisted mechanisms. Figure 2.3 shows the calculated DCR as a function of the normalized excess applied voltage for three different widths of the multiplication region. The normalized excess applied voltage is defined as $\Delta V/V_{BR}$, where $\Delta V = (V - V_{BR})$, V_{BR} is the breakdown voltage and V is the voltage across the device. The solid lines correspond to the case when field-assisted and temperature-assisted generation of dark carriers are both present in the model. For clarity, the case when only field-assisted generation is taking place (dashed lines) is also shown. The figure shows that at higher normalized excess bias voltages the DCR curve is almost completely dictated by tunneling effects for all the widths of the multiplication region. It is also noticed that the effect of temperature-assisted dark carrier generation on the DCR is more important in devices with thick multiplication regions, e.g., > 800 nm. On the other hand, for devices with thin multiplication regions the DCR curve is dominated, over almost the whole range of normalized excess voltages, by field-assisted mechanisms.

Figure 2.4 shows the calculated PDE versus DCR for InP multiplication regions of 700, 900, 1200 and 2000 nm, and an InGaAs absorption layer of 1 μm. There are two groups of curves generated according to the different mechanisms for dark-carrier generation; in the figure these two groups are labeled by their respective ellipses. The lower group of curves corresponds to the cases for which both field-assisted and temperature-assisted generation of dark carriers are included. On the other hand, in the upper group of curves only field-assisted generation of dark carriers is considered. It can be seen that the PDE versus DCR behavior varies as temperature-assisted dark carriers along with field-assisted dark carriers are included. In the case when only field-assisted dark-carrier generation is considered (upper group), the calculated PDE, for a given DCR, is higher as the multiplication region becomes wider. On the other hand, in the case for which both mechanisms of dark-carrier generation are considered (lower group) two distinct behaviors are observed as the width of the

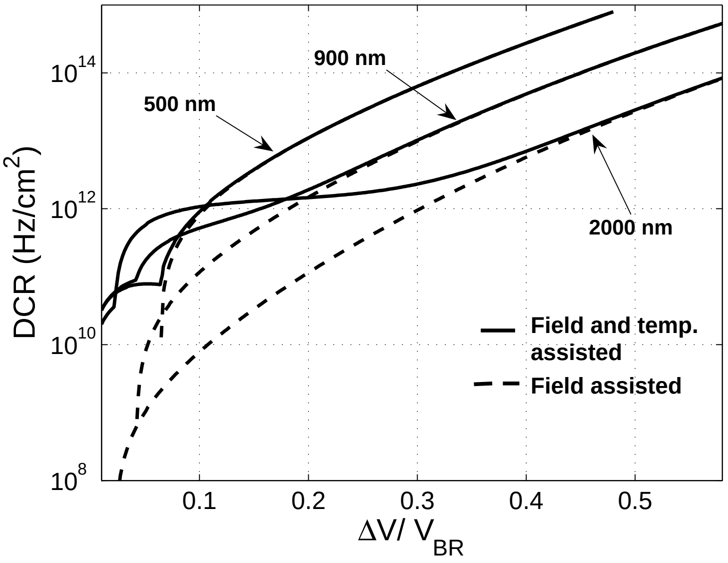


Figure 2.3: DCR versus normalized excess voltage for 500 nm, 900 nm and 2000 nm multiplication region widths. Dashed lines correspond to the case when GR dark carriers are absent and solid lines correspond to the case when both field-assisted and GR dark carriers are present.

multiplication region increases. First, for the low values of the DCR ($< 10^{11}$Hz/cm^2), it can be seen an improvement in PDE as the width of the multiplication region is increases. However, for larger DCR values, the PDE degrades as the multiplication region becomes wider. Hence, the calculated results illustrated in Fig. 2.4 suggest that in cooled devices, the performance will improve as the width of the multiplication region is increased. However, for devices working at room temperature the increment in PDE, due to a wider multiplication region, is counteracted by an increment in DCR and the performance will be degraded as the multiplication region becomes wider. The improvement in the PDE versus DCR characteristics at low temperatures is attributable to fact that as the width of the multiplication region increases the

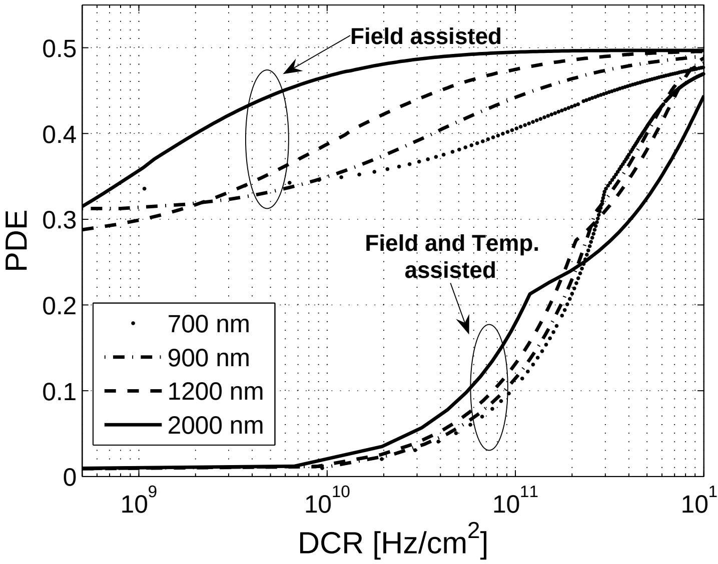


Figure 2.4: PDE versus DCR for InP multiplication regions of 700, 900, 1200 and 2000 nm. The absorber is a 1 μm layer of InGaAs. The maximum value of the PDE versus DCR curve is determined by the quantum efficiency η, which in this case is 0.5.

tunneling current decreases due to the lower electric field. It should be pointed out that the maximum value of the PDE versus DCR curve is determined by the quantum efficiency η, which in this case is 0.5.

Our calculations of the SPQE, as a function of the applied bias, indicate a similar trend to that suggested by the PDE versus DCR curves. Moreover, the SPQE curves provide further insight by suggesting an optimal thickness of the multiplication region that achieves the highest SPQE at the appropriate applied voltage. Figure 2.5 (solid lines) shows that the peak value of each SPQE curve increases as the width of the multiplication region increases, reaching a maximum value (between 1200–1400 nm) beyond which it starts to decrease. Nonetheless, for a scenario dominated by field-assisted dark-current generation both the peak SPQE and the FWHM (full-width-at-half-maximum) of each curve increase as the width of the multiplication region increases (dashed lines). The existence of an optimal peak SPQE at room temperature is a result of the competing effects of the field- and temperature-assisted generation of dark carriers.

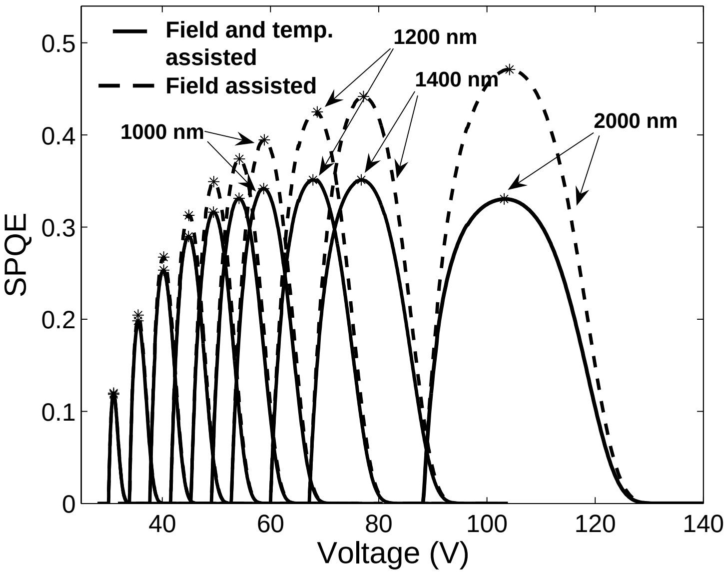


Figure 2.5: SPQE versus applied voltage for several widths of the multiplication region. The maximum achievable value of the SPQE curve is determined by the quantum efficiency η, which in this case is 0.5.

2.3.2 InAlAs-InP heterojunction multiplication regions

Thin heterojunction multiplication regions have proven to be beneficial in reducing the excess noise factor due to the strong effect of the dead space in devices with thin multiplication region [55, 56, 57]. However, their desirable characteristics decreases in devices with thick multiplication regions due to the reduced importance of dead space in these devices. Additionally, in an earlier theoretical work [47] it has been shown that the fractional width of the $\mathrm{In_{0.52}Al_{0.48}As}$ layer in an $\mathrm{In_{0.52}Al_{0.48}As}$-InP heterojunction multiplication region can be optimized to attain a maximum SPQE that is greater than that offered by a homojunction InP multiplication region. (The *fractional width* of the $\mathrm{In_{0.52}Al_{0.48}As}$ layer in an $\mathrm{In_{0.52}Al_{0.48}As}$-InP heterojunction multiplication region is defined as the ratio between the width of the $\mathrm{In_{0.52}Al_{0.48}As}$ energy buildup layer to the total width of the heterojunction multiplication region comprising the $\mathrm{In_{0.52}Al_{0.48}As}$ and InP layers.) It was also shown that this effect became more pronounced in thin multiplication regions as a result of the increased significance of dead space. Therefore, it would be of interest to further investigate the performance of SPADs with heterojunction multiplication regions.

Figure 2.6 shows the DCR as a function of the normalized excess voltage for four different widths of the multiplication region. By comparing Fig. 2.6 with Fig. 2.3, it is observed that the curves show a similar trend in the DCR as the width of the multiplication region is varied. Similarly to the case of a homojunction multiplication region and in accord with our understanding of the dominance of field-assisted effects over GR effects in high-fields, the change in the DCR, as the role of GR is varied (for a certain width of the multiplication region), is only noticeable in thicker multiplication regions.

The SPQE curves, on the other hand, give as a slightly more informative account of things. As a function of the applied voltage, the SPQE exhibits a different behavior in the cases of a homojunction and heterojunction multiplication regions. Figure 2.7 shows the calculated SPQE versus the applied voltage for the homojunction and heterojunction multiplication regions for several widths of the multiplication region. For a given width of the multiplication region, the calculated SPQE of the heterojunction multiplication region is higher than that for the homojunction case.

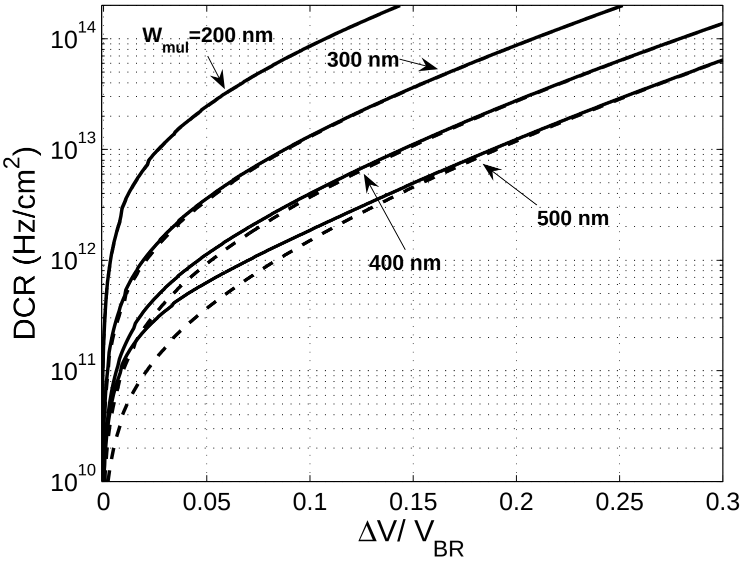


Figure 2.6: DCR versus normalized excess voltage for 200 nm, 300 nm , 400 nm and 500 nm multiplication region widths. Dashed lines correspond to the case when GR dark carriers are absent and solid lines correspond to the case when both field-assisted and GR dark carriers are present.

Moreover, this enhancement in the SQPE, as we move from a homojunction to a heterojunction, becomes more pronounced as the width of the multiplication region is reduced. This is attributed to the fact that for a given width of the multiplication region, the electric field required to achieve a certain breakdown probability is smaller in the heterojunction multiplication-region case than that in a homojunction multiplication-region case [47], which, in turn, results in a reduction in the number of dark carriers generated though field-assisted mechanisms. The improvement in breakdown characteristics in properly designed heterojunction multiplication layers is a result of the so-called initial-energy effect, which takes advantage of injecting “hot” carriers from a high bandgap layer (InAlAs in our case) of the multiplication region to the lower bandgap layer (InP) [54, 47]. It is to be noted, however, that this conclusion does not take into account the possibility of an increase in hole trapping in a heterojunction multiplication region, which may aggravate after-pulsing.

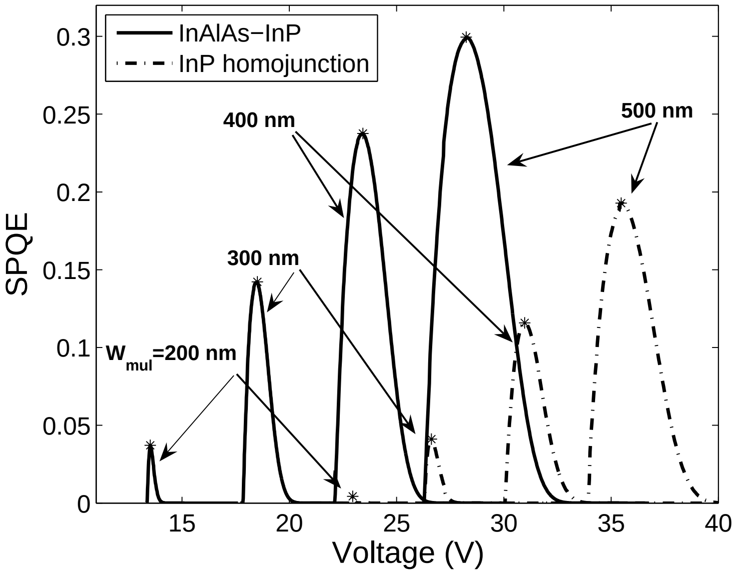


Figure 2.7: SPQE versus applied voltage for InP homojunction and InAlAs-InP heterojunction multiplication region.

2.4 Conclusions

This theoretical study shows that the thickness of the multiplication region plays a different role in the performance of a SPAD depending upon what mechanism of dark-carrier generation is dominant. At low temperatures, for which field-assisted mechanisms are dominant, an increment in the thickness of the multiplication region will result in an improved PDE vs. DCR characteristics. The same behavior is seen in the SPQE curve at low temperatures. At room temperatures, on the other hand, the PDE vs. DCR characteristics show a weaker performance as the width of the multiplication region is increased. However, the SPQE curves show a maximum achievable peak SPQE at an optimal overbias and an optimal multiplication-region width. It is important to note that the behavior of an APD as a function of the multiplication-region width in the linear mode, where excess noise factor decreases as the multiplication-region width is decreased, is characteristically different from that of a SPAD.

Chapter 3

Impact ionization under dynamic electric field

3.1 Introduction

As mentioned in the introduction of this dissertation, the exponential growth in video, voice, data and mobile-device traffic over the Internet, has motivated the telecommunication industry to move toward 40-Gbps and 100-Gbps protocols for their core fiber-optic backbone networks alongside the existing 10-Gbps infrastructure operating at the low-loss wavelength of 1.55 μm. Clearly, operation at such high speeds requires high sensitivity detectors. Separate absorption and multiplication (SAM) InP-InGaAs APDs are normally the most preferred photodetectors for direct-detection high-data rate systems. However, due to the stochastic nature of the impact-ionization process, the buildup time of an APD is also stochastic and can degrade its performance. In particular, the APD's finite and stochastic buildup time, the time needed for all the impact ionizations to settle, gives rise to intersymbol interference (ISI) and limits the bandwidth of the communication system [58]. In fact, the long avalanche buildup time in InP has limited the speed of InP-based APDs and stopped them from meeting the expectations of 40-Gbps systems. While the buildup time of an APD is dependent on the material (e.g., on k, the hole-to-electron ionization coefficient ratio) it also has a dependence on the applied electric field and its profile in time and space [59] as it governs the cascade of impact ionizations. Several approaches have been explored to model the buildup time in order to increase the gain-bandwidth product (GBP) of APDs, including GBPs for heterojunction multiplication regions [60, 59]. However, all existing models focused, mainly, on optimizing the APD's structure. In particular, the effect of modulating the electric field on the impact ionization characteristics of APDs remains unexplored, and no analytical model for avalanche multiplication exists for APDs that are driven by a time-varying bias voltages. The modulation of the applied electric field to control the impact ionization process could be beneficial in communication systems since it opens up the possibility of increasing the GBP in a simple and efficient way. The optimization problem becomes that of finding the optimal electric-field profile, for a fixed mean gain, that maximizes the GBP. In this chapter it is outlined the potential benefits of modulating the applied electric field on the performance of APDs. This approach enables the calculation of the impulse response, gain and and excess noise factor, breakdown probability, as well as pulse duration time all under conditions of a dynamic field in the multiplication region.

The theory of impact ionization process under dynamic electric fields was developed by Hayat *et al.* [61], and the most important findings are repeated here for completeness.

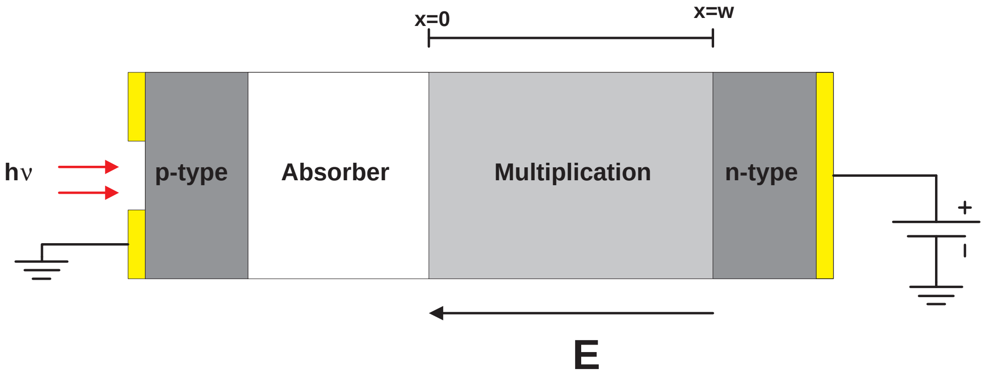


Figure 3.1: Separate-absorption multiplication APD.

3.2 Concept of impact ionization under dynamic electric field

Suppose that a time-varying bias, $V_B(t)$, $t \geq 0$, is applied to an APD. Consider a charge-depleted multiplication region of an APD extending from $x = 0$ to $x = w$, as shown in Fig. 3.1, with the convention that the electric field is pointing in the negative x direction. Let $E(x, t)$ denote the dynamic electric field in the multiplication region at position x and at time t. If the field is uniform, then $E(x, t) \equiv E(t) = V_B(t)/w$. Suppose that a parent hole (electron) is created at an arbitrary location x in the multiplication region of the APD, and assume that the field is sufficiently high so as conduction-band electrons and valence-band holes travel at their material-specific saturation velocities, v_e and v_h, respectively. As the hole travels the multiplication region, it can impact ionize at a stochastic location, say ξ, and at time $\tau = (\xi x)/v_h$. It will be assumed that the change in the electric field is slow with respect to the successive ionizations that a carrier undergoes so that the ionization coefficient of a carrier depends adiabatically on the instantaneous electric field $E(\xi, \tau)$ at the point where and when it impact ionizes. The ability of a carrier to impact ionize also depends on the carrier's history, consistent with the dead-space phenomenon. In particular, a newly born carrier cannot impact ionize before traveling a dead space [56], which is the minimum distance a carrier must travel before it acquires sufficient energy to effect an impact ionization. Upon impact ionization, the parent hole is replaced with two offspring holes and an offspring electron. Each offspring carrier moves on to further impact ionize, and so on. This process continues and it may or may not terminate, depending on the field and device and material properties. The stochastic multiplication factor is the total number of electron-hole pairs generated as a result of a parent carrier; it can be either finite or infinite. On the other hand, the buildup time is the time measured from the creation of the parent carrier to the time when all carriers have exited the multiplication region. In Section 3.3 the equations that enable us to calculate the statistics of the multiplication factor and the impulse response function, as well as the breakdown probability under the dynamic field $E(x,t)$ will be derived. However, before doing it is important to extend the notions of the ionization coefficient, dead space, and the probability density function of the carrier's free path (prior to ionization) to a dynamic-field setting.

Under the adiabatic field assumption, $\alpha(x,t)$ and $\beta(x,t)$ are defined as the electron and hole time-varying non-localized ionization coefficients associated with carriers at location x in the multiplication region and at time t. These are the ionization coefficients for those carriers that have already traveled the dead space. Following the model for non-localized ionization coefficients under a static electric field [62] and by replacing the static field with its dynamic counterpart, the dynamic ionization coefficients are given by

$$\begin{aligned} \alpha(x,t) &= A_e \exp\left[-\left(\frac{\mathcal{E}_e}{E(x;t)}\right)^{m_e}\right] \\ &\text{and} \\ \beta(x,t) &= A_h \exp\left[-\left(\frac{\mathcal{E}_h}{E(x;t)}\right)^{m_h}\right], \end{aligned} \tag{3.1}$$

where A, $\mathcal{E}$, and m are parameters chosen by fitting measured excess-noise-factor data [63, 64]. The value of the constants A, $\mathcal{E}$, and m, are known for various III–V materials [62].

3.2.1 Probability density function of the carrier's free path under dynamic fields

Consider an electron and hole created at location x and of age s relative to the launch instant of the dynamic electric field (at $t = 0$), and let X_e and X_h be their stochastic free-path distances to their first impact ionization. As it turns out, the age of a carrier will play a key role in the formulation of the theory for avalanche multiplication under dynamic fields, as described in Section 3.3. As an extension to shifted exponential model for the free path [54], the time-varying probability density function of X_e and X_h as $h_e(\xi; x, s)$ and $h_h(\xi; x, s)$, respectively, are given by

$$h_e(\xi; x, s) = \begin{cases} \alpha(\xi, s + \frac{\xi - x}{v_e}) \exp\left(\int_{x+d_e(x,s)}^{\xi} \alpha(\sigma, s + \frac{\sigma - x}{v_e})\, d\sigma\right) & \xi \geq x + d_e(x,s) \\ 0 & \text{Otherwise} \end{cases} \tag{3.2}$$

$$h_h(\xi; x, s) = \begin{cases} \beta(\xi, s + \frac{x - \xi}{v_h}) \exp\left(\int_{\xi}^{x-d_h(x,s)} \beta(\sigma, s + \frac{x - \sigma}{v_h})\, d\sigma\right) & \xi \leq x - d_h(x,s) \\ 0 & \text{Otherwise} \end{cases}$$

(3.3)

where $d_e(x, s)$ and $d_h(x, s)$ represent the time-varying dead spaces for electron and hole, respectively, that were created at position x and of age s. The electron's time-varying dead space is computed as the minimum d value that satisfies the equation

$$q \int_{x}^{x+d_e(x)} E(y, s + \frac{y - x}{v_e})\, dy = E_{th,e}(x + d_e(x)), \tag{3.4}$$

where for any $0 \leq x \leq w$, $E_{th,e}(x)$ is the ionization threshold energy for electrons for the material at position x in the multiplication region. Similarly, the hole's dynamic dead space is computed as the minimum d value that satisfies the equation

$$q \int_{x}^{x+d_e(x)} E(y, s + \frac{x - y}{v_h})\, dy = E_{th,h}(x - d_h(x)), \tag{3.5}$$

where $Eth, h(x)$ is the ionization threshold energy for holes for the material at position x in the multiplication region. Note that given the knowledge of the material composition in the multiplication region and the electric field profile and its evolution in time, the dynamic dead space can be calculated for all x and s.

3.3 Multiplication theory under dynamic electric fields

The recurrence theory for the avalanche multiplication, including the gain and impulse-response function, under non-uniform, static electric fields was originally formulated by Hayat et al. in [65, 56, 66]. In addition, the generalization of the recurrence theory under dynamic electric fields was introduced by Hayat *et al.* in [61] and it is repeated for completeness. As it was mentioned before, the age of the parent carrier triggering the avalanche process is key in modeling the avalanche multiplication process when the field is allowed to be time variant because carriers with different ages will experience different dynamical electric field ahead of them as they travel the multiplication region. Specifically, if we assume a causal and spatially non-uniform electric-field, $E(x,t)$, launched at time $t=0$, then an electron born at location x with age 0 will experience this time-varying field in its lifetime. In contrast, if an electron is born at location x with age s (relative to the launch time of the field at $t=0$) then it will experience in its lifetime the s-delayed version of the dynamic field, namely $E(x,t)u(t-s)$, where $u(.)$ is the unit-step function. To take the carrier's "age" element into account in a model for the avalanche multiplication process, we must formulate a model in which the ionization probability is parameterized by the time at which the parent carrier is injected in the multiplication region. In what follows, sets of age-dependent recurrence equations are derived that enable us to calculate the mean gain, the excess noise factor, the probability distribution function of the gain, the mean of the impulse-response function, as well as the breakdown probability, all under a dynamic electric field.

3.3.1 Analysis of the gain statistics

$Z(x,s)$ and $Y(x,s)$ are defined to be the totality of all electrons and holes, including the parent carrier, initiated by an electron and hole, respectively, injected at location x with age s. Note that $Z(w,s)=Y(0,s)=1$ since an electron (hole) placed at the left (right) edge of the multiplication region will exit with ionizing and result in no offspring carriers. Now consider a parent and electron at location x and of age s. The stochastic multiplication factor, $M(x,s)$, defined as the total number of electron-hole pairs generated as a result of an electron-hole pair whose initial location in the multiplication region is x and whose ages are s is simply

$$M(x,s)=0.5[Z(x,s)+Y(x,s)]. \tag{3.6}$$

Note that in the case of a SAM APD where holes are injected at the edge of the multiplication region at $x = w$ and with age $s \geq 0$, then the stochastic age-dependent gain is given by

$$G(s) = M(w, s) = 0.5[1 + Y(w, s)]. \tag{3.7}$$

Mean gain

Now the equations that allow us calculate the statistics of the quantities $Z(x, s)$ and $Y(x, s)$ are derived. Once we find the first and second moments of Z and Y we can related them to the mean gain and the excess noise factor via (3.6) and (3.7).

Suppose that the first ionization for a parent electron of age s (from the launch time of the field at $t = 0$) positioned at location x and occurs at location $X_e = \xi$, where $x \leq \xi \leq w$. Note that the time of this ionization is necessarily $s + \tau$, where $\tau = (\xi - x)/v_e$. Note that the two offspring electrons at ξ, whose ages are $s + \tau$, will independently generate $Z_1(\xi, s + \tau)$ and $Z_2(\xi, s + \tau)$ carriers, respectively. The offspring hole, on the other hand, will generate $Y(\xi, s + \tau)$ carriers, which is also statistically independent of $Z_1(\xi, s + \tau)$ and $Z_2(\xi, s + \tau)$. Thus, conditional on the event that the first impact ionization for the parent electron occurs at location ξ, the sum of $Z_1(\xi, s + \tau)$, $Z_2(\xi, s + \tau)$ and $Y(\xi, s + \tau)$ will simply amount to $Z(x, s)$. Since we can always express the mean of $Z(x, s)$ as the expectation of the conditional mean given that the first ionization of the parent electron (triggering $Z(x, s)$) occurs at location X_e, we can write

$$\begin{aligned} E[Z(x, s)] &= E[E[Z(x, s)|Xe]] \\ &= E[Z_1(X_e, s + (X_e - x)/v_e) + Z_2(X_e, s + (X_e - x)/v_e) \\ &+ Y(X_e, s + (X_e - x)/v_e)]. \end{aligned} \tag{3.8}$$

Now if we define the notation for the mean of the quantities $z(x, s) = E[Z(x, s)]$ and $y(x, s) = E[Y(x, s)]$, then the expression on the right hand side of (3.8) can recast as

$$z(x,s) = E[2z(X_e, s + (X_e - x)/v_e) + y(X_e, s + (X_e - x)/v_e)]. \tag{3.9}$$

Note that the only thing that is stochastic in (3.9) is X_e, whose pdf is given by (A.1); thus, we can write (3.9) explicitly as

$$\begin{aligned} z(x,s) &= \int_w^\infty h_e(\xi;x,s)\,d\xi + \int_x^w \left[2z(\xi, s + \tfrac{\xi - x}{v_e}) + y(\xi, s + \tfrac{\xi - x}{v_e})\right] h_e(\xi;x,s)\,d\xi \\ & \qquad\qquad 0 \le x \le w, \quad s \ge 0. \end{aligned} \tag{3.10}$$

The first term on the right side of (3.10) represents the scenario when the parent electron does not impact ionize before it exits the multiplication region.

We can repeat the above argument that led to (3.10) by starting from a parent hole at ξ and of age s instead of the parent electron. Note that the location of the first ionization of the parent hole, X_h, can be in the interval $[0, x]$ (instead of $[x, w]$ as in the case of the parent electron). The recursive equation for $y(x,s)$ is

$$\begin{aligned} y(x,s) &= \int_x^\infty h_h(\xi;x,s)\,d\xi + \int_0^x \left[2y(\xi, s + \tfrac{\xi - x}{v_e}) + z(\xi, s + \tfrac{\xi - x}{v_e})\right] h_h(\xi;x,s)\,d\xi \\ & \qquad\qquad 0 \le x \le w, \quad s \ge 0. \end{aligned} \tag{3.11}$$

The pair of linear coupled integral equation in (3.10) and (3.11) can be solved numerically, e.g., by the method of iterations.

The age-dependent mean multiplication factor can be calculated using $m(x,s) = 0.5(z(x,s)+y(x,s))$, and the age-dependent gain, $g_a(s)$, is simply $m(w,s)$ in the case of a hole injection APD. As a special case, the usual mean gain, g, of an APD under a static bias is simply $g = g_a(0)$ for the cases of hole injection.

Excess noise factor

The first equations to be derived are the recursive equations for the second moments $z_2(x,s) = E[Z(x,s)^2]$ and $y_2(x,s) = E[Z(x,s)^2]$. Note that the second moment of the multiplication factor, $m_2(x,s) = E[M(x,s)^2]$, can be related to $z_2(x,s)$ and $y_2(x,s)$ using $m_2(x,s) = 0.25[z_2(x,s) + y_2(x,s) + 2z(x,s)y(x,s)]$, and the age-dependent excess noise factor, $F_a(s) = m_2(w,s)/g_a(s)^2$, is given by

$$F(s) = \frac{y_2(w,s) + 2y(w,s) + 1}{[y(w,s) + 1]^2}. \tag{3.12}$$

Next, the renewal equations characterizing $z_2(x,s)$ and $y_2(x,s)$ are derived. Since we can always express the second moment of $Z(x,s)$ as the expectation of the conditional second moment of $Z(x,s)$ given that the first ionization of the parent electron (triggering $Z(x,s)$) occurring at location X_e, we can write

$$\begin{aligned} E[Z(x,s)^2] &= E[E[Z(x,s)^2|Xe]] \\ &= E[\{Z_1(X_e, s + (X_e - x)/v_e) + Z_2(X_e, s + (X_e - x)/v_e) \\ &\quad +Y(X_e, s + (X_e - x)/ve)\}^2] \\ &= E[Z_1(X_e, s + (X_e - x)/v_e)^2 + Z_2(X_e, s + (X_e - x)/v_e)^2 \\ &\quad +Y(X_e, s + (X_e - x)/v_e)^2 \\ &\quad +2Z_1(X_e, s + (X_e - x)/v_e)Z_2(X_e, s + (X_e - x)/v_e) \\ &\quad +2Z_1(X_e, s + (X_e - x)/v_e)Y(X_e, s + (X_e - x)/v_e) \\ &\quad +2Z_2(X_e, s + (X_e - x)/v_e)Y(X_e, s + (X_e - x)/v_e)], \end{aligned} \tag{3.13}$$

which simplifies to

$$\begin{aligned} z_2(x,s) &= E[2z_2(X_e, s + (X_e - x)/v_e) \\ &\quad +y_2(X_e, s + (X_e - x)/v_e) + 2z(X_e, s + (X_e - x)/v_e)^2 \\ &\quad +4z(X_e, s + (X_e - x)/v_e)y(X_e, s + (X_e - x)/v_e)]. \end{aligned} \tag{3.14}$$

Upon writing down the averaging explicitly using the pdf of X_e, we obtain

$$\begin{aligned} z_2(x,s) &= \int_w^\infty h_e(\xi;x,s)\,d\xi + \int_x^w \left[2z_2\left(\xi, s+\frac{\xi-x}{v_e}\right) + y_2\left(\xi, s+\frac{\xi-x}{v_e}\right)\right. \\ &\quad \left. 4z\left(\xi, s+\frac{\xi-x}{v_e}\right)y\left(\xi, s+\frac{\xi-x}{v_e}\right) + 2z\left(\xi, s+\frac{\xi-x}{v_e}\right)\right] h_e(\xi;x,s)\,d\xi \\ &\quad 0 \le x \le w, \quad s \ge 0. \end{aligned} \tag{3.15}$$

Similarly, a recursive equation for $y_2(x,s)$ can be obtained:

$$\begin{aligned} y_2(x,s) &= \int_x^\infty h_h(\xi;x,s)\,d\xi + \int_0^x \left[2y_2\left(\xi, s+\frac{x-\xi}{v_h}\right) + z_2\left(\xi, s+\frac{x-\xi}{v_h}\right)\right. \\ &\quad \left. 4z\left(\xi, s+\frac{x-\xi}{v_h}\right)y\left(\xi, s+\frac{x-\xi}{v_h}\right) + 2y\left(\xi, s+\frac{x-\xi}{v_h}\right)\right] h_h(\xi;x,s)\,d\xi \\ &\quad 0 \le x \le w, \quad s \ge 0. \end{aligned} \tag{3.16}$$

The pair of linear coupled integral equation in (3.15) and (3.16) can be solved numerically once the quantities $z(x,s)$ and $y(x,s)$ have already been computed by first solving the pair of equations in (3.10) and (3.11).

Breakdown probability

Here it is developed recursive equations that characterize the probability that a parent carrier triggers breakdown, that is, the probability that infinitely many offspring carriers are generated. Note that if an electron at position x in the multiplication region and of age s (relative to the launch instant of the dynamic electric field) impact ionizes for the first time at location ξ, then the probability that the parent electron generates a finite number of offspring carriers is precisely the product of the probabilities that each of the two offspring electrons and offspring hole created at ξ with age $s + (\xi - x)/v_e$ generates a finite number of offspring carriers. (Implicit in this statement is that each carrier acts independently of the other carries, which is a correct assumption since here no feedback effect from the created charges on the electric field are included.) Note that in the special case when the parent electron exits the multiplication region without ionizing, the conditional probability that produces a finite number of carriers is trivially equal to one.

Mathematically, $P_Z(x,s) = P_Z(x,s) < \infty$ and $P_Y(x,s) = PY(x,s) < \infty$ are defined, and they are given by:

$$P_Z(x,s) = \int_w^\infty h_e(\xi;x,s)\,d\xi + \int_x^w P_Z^2\left(\xi, s + \frac{\xi - x}{v_e}\right) P_Y\left(\xi, s + \frac{\xi - x}{v_e}\right) h_e(\xi;x,s)\,d\xi \tag{3.17}$$

A similar argument can lead to the equation

$$P_Y(x,s) = \int_x^\infty h_h(\xi;x,s)\,d\xi + \int_0^x P_Y^2\left(\xi, s + \frac{x - \xi}{v_h}\right) P_Z\left(\xi, s + \frac{x - \xi}{v_h}\right) h_h(\xi;x,s)\,d\xi \tag{3.18}$$

Once $P_Z(x,s)$ and $P_Y(x,s)$ are numerically calculated by solving the nonlinear coupled integral equations in (3.17) and (3.18) (using iterations, for example), the breakdown probability is calculated. For example, for a hole injection APD, the age-dependent breakdown probability for a photon absorbed at time s, is simply

$$P_B(s) = 1 - P_Z(w,s). \tag{3.19}$$

3.3.2 Mean impulse response function

$I_e(t,x,s)$ is defined as the stochastic impulse-response function at time t, initiated by an electron injected at location x and with age s. Similarly, $I_h(t,x,s)$ is the stochastic impulse-response function at time t, initiated by a hole injected at location x with age s. The age variable s can also be associated with the absorption time of a photon.

Mathematically, if we define $i_e(t, x, s)$ and $i_h(t, x, s)$ as the mean quantities of $I_e(t, x, s)$ and $I_h(t, x, s)$, respectively, then we can write that conditional mean of $I_e(t, x, s)$ given that the first ionization of the parent electron occurs at location X_e as

$$E[I_e(t, x, s)|X_e] = 2i_e(t, X_e, s + (X_e - x)/v_e) + i_h(t, X_e, s + (X_e - x)/v_e), \quad (3.20)$$

where $0 \leq X_e \leq w$. On the other hand, when no ionization occurs (namely, $X_e > w$), then $E[I_e(t, x, s)|Xe > w] = (qv_e/w)u(t) - u(t(w - x)/v_e)$, which is simply a square pulse of duration equal to the transit time of the original electron (born at x) as it drifts across the remainder of the multiplication region. Averaging the above conditional expectation over all possible values of the stochastic position, X_e of the first ionization, the following coupled recursive equations are obtained:

$$\begin{aligned} i_e(t, x, s) &= \frac{qv_e}{w}\left[u(t) - u\left(t - \frac{w - x}{v_e}\right)\right] \int_w^\infty h_e(\xi; x, s) \\ &+ \int_x^{min(w, x+v_e t)} 2i_e\left(t - \frac{\xi - x}{v_e}, \xi, s + \frac{\xi - x}{v_e}\right) \\ &\quad i_h\left(t - \frac{\xi - x}{v_e}, \xi, s + \frac{\xi - x}{v_e}\right) h_e(\xi; x, s)\, d\xi \end{aligned} \quad (3.21)$$

A similar equation can be obtained for $i_h(t, x, s)$:

$$\begin{aligned} i_h(t, x, s) &= \frac{qv_h}{w}\left[u(t) - u\left(t - \frac{x}{v_h}\right)\right] \int_w^\infty h_h(\xi; x, s) \\ &+ \int_{max(0, x - v_h t)}^{x} 2i_e\left(t - \frac{x - \xi}{v_h}, \xi, s + \frac{x - \xi}{v_h}\right) \\ &\quad i_h\left(t - \frac{x - \xi}{v_h}, \xi, s + \frac{x - \xi}{v_h}\right) h_h(\xi; x, s)\, d\xi \end{aligned} \quad (3.22)$$

The two coupled equations can be solved numerically using an iterative approach. The mean impulse-response function, $i(t, s)$, (in the case of a hole injection to the multiplication region at $x = 0$) for a photon absorbed at time s is then obtained using $i(t, s) = i_h(t, w, s)$.

3.3.3 Pulse response, pulse bandwidth and pulse-generated mean gain

It was mentioned earlier that under dynamic biasing, the impulse response function $i(t, s)$ is dependent on the birth time of the photogenerated parent carrier triggering the APD (or the arrival time s of the absorbed photon). Since in on-off keying optical communication photons arrive randomly within each optical pulse, the appropriate quantity to look at when assessing ISI would be the response of the APD to a pulse, rather than an impulse, which can be easily obtained through the relation

$$i_p(t) = \int_0^T i(t, s) p_{ph}(s)\, ds, \tag{3.23}$$

where $p_{ph}(s)$ is the probability density of photons within an optical pulse (bit) of duration T, and it is proportional to the intensity of the received optical pulse within the optical bit. Note that if the electric field is static, then $i(t, s)$ is simply $i(t-s)$, and $i_p(t)$ would become simply the convolution between $i(t)$ and $p_{ph}(t)$. An alternative way to view $i_p(t)$ is to regard it as a photon-arrival time averaged impulse response. Since early and late photons have long and short impulse responses, it would make sense to look at the average of the impulse response functions over all possible photon arrival times within each received optical pulse. By calculating the 3 dB bandwidth of the Fourier transform of $i_p(t)$, we can obtain the bandwidth, B_p, which combines the APD's buildup limited bandwidth with the bandwidth of the optical pulse in each bit of duration T.

Note that if we integrate $i_p(t)$ over the interval $[0, T]$, then we obtain an estimate of the charge generated by the APD due to the optical pulse received in the interval $[0, T]$. In particular, if $i(t, s)$ is very narrow compared to the pulse $p_{ph}(.)$, then the integral of $i_p(t)$ over the interval $[0, T]$ is simply $q \int_0^T g_a(s) p_{ph}(s)\, ds$ (assuming a hole injection APD), which can be approximated as $nT^{-1} q \int_0^T g_a(s)\, ds$, where $n = \int_0^T p_{ph}(s)\, ds$ is the average number of photons in the received pulse, and as defined in Section 3.3, $m(w, s)$ is the age-dependent mean multiplication factor due to a carrier-pair born at w with age s. This motivates us to define the average pulse-gain factor, $\bar{g_p}$, as

$$\bar{g_p} = T^{-1} \int_0^T g_a(s)\, ds \tag{3.24}$$

which is simply the average of the age-dependent mean gain ($g_a(s)$), as defined in Section 3.3, over all photon arrival times. Hence, in the photocurrent is integrated over one bit period, then the output of the integrator is $nq\bar{g_p}$, which is the average number of detected photons multiplied by the average age-dependent gain. As a special case when the field is static, $\bar{g_p} = g$ and the output of the integrator is the usual expression nqg.

Finally, the pulse gain-bandwidth product, GBP_p, is defined as

$$GBP_p = g_p B_p, \tag{3.25}$$

which collapses to the standard gain bandwidth product whenever the biasing is static. It is worthwhile reiterating that when the photocurrent is integrated over each bit in the receiver, the total charge is proportional to the product of the pulse-gain factor $\bar{g_p}$ and the average number n of detected photons in the optical pulse in each bit. In other words, in the dynamically biased integrate-and-dump receiver, the charge produced in each bit remains proportional to the energy in the optical pulse in each bit. Thus, while the dynamically biased APD may not be directly applicable to simple analog detection (unless gain equalization is employed), it is a perfect fit to digital communications.

3.4 Results

To exploit the dynamic biasing scheme in on-off keying direct detection communication, it is desirable to use a periodic dynamic bias so as to minimize the tail of the APD's response to an optical pulse, in either RZ or NRZ formats, over subsequent bits, thereby minimizing ISI. In principle, there are many choices for the waveform selection in each bit-duration; however, from the perspectives of practicality of implementation and minimizing waveform distortion at high speeds, a sinusoidal waveform may be the best choice.

$$V_B(t) = A + m \cdot sin(2\pi f_c t + \phi), \tag{3.26}$$

where f_c is set to be equal to the reciprocal of the bit transmission rate, $f_c = 1/R$. The parameters A, m and ϕ are free parameters that can be selected to maximize the benefit of dynamic biasing. It is implicitly assumed that the bias signal is synchronous with the bit stream. In practice, a clock recovery circuit and a phase lock loop can be employed to maintain synchronization. In addition, a spatially uniform electronic field, $E(t) = V_B(t)/w$, will be assumed.

3.4.1 Breakdown probability under linearly varying biasing

The calculation of the breakdown probability assumes that the voltage across the multiplication region has a duration of 8 electron-transit times, after which the voltage is zero, as shown in the inset of Fig. 3.3a. First, the constant-field case is examined. Figure 3.3a shows the calculated probability that the pulse, initiated by an electron injected at location x with age s, terminates by time t. The inset shows, in the blue curve, the applied constant voltage of 23 V as a function of time and the breakdown voltage level in the red curve. Two cases were simulated: (i) when the electron that initiates the avalanche is created at the start of the electric field, i.e.,

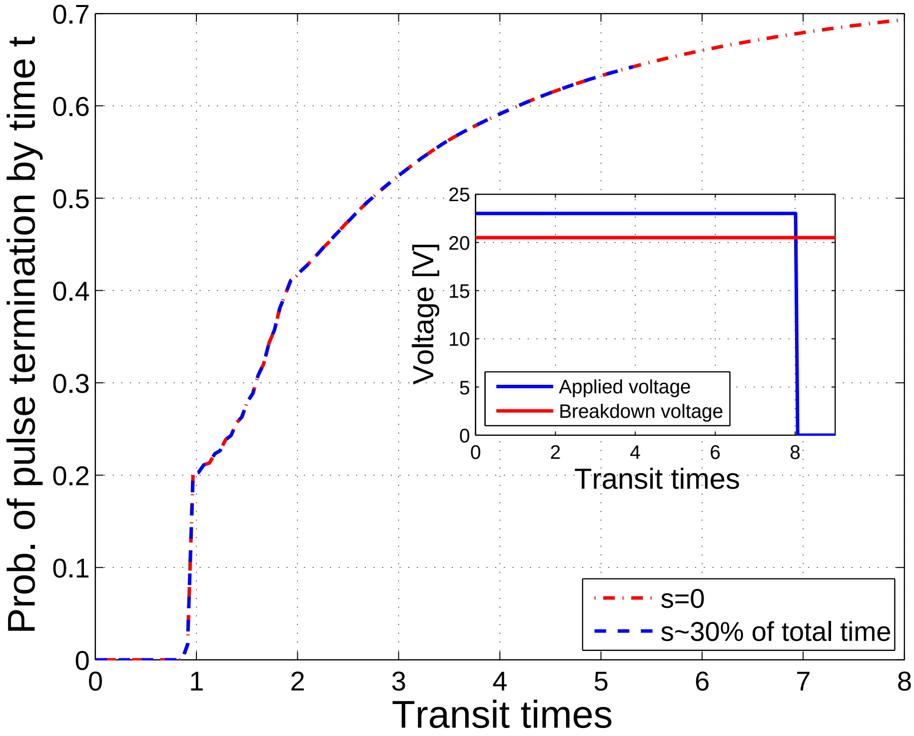


Figure 3.2: Probability that the pulse created by an electron injected at $x = 0$ with age s terminates by time t. The dash-dotted red curve shows $P_e(x, t, s)$ for $s = 0$, which is the case when the electron that initiates the pulse is injected at the start of the electric field. On the other hand, the dashed blue curve shows $P_e(x, t, s)$ when the electron that initiates the pulse has an age $s \sim 30\%$ of the total time. The inset shows the applied constant voltage (blue curve) and the breakdown voltage (red curve).

with age $s = 0$ (dash-dotted red curve), and (ii) when the electron that triggers the avalanche has an age $s \sim 30\%$ of the total duration of the applied voltage (dashed blue curve). It can be seen that both curves overlap. This is an expected result since under a constant electric field the probability of pulse termination is not affected by the age of the carrier that initiates the avalanche. On the other hand, when we

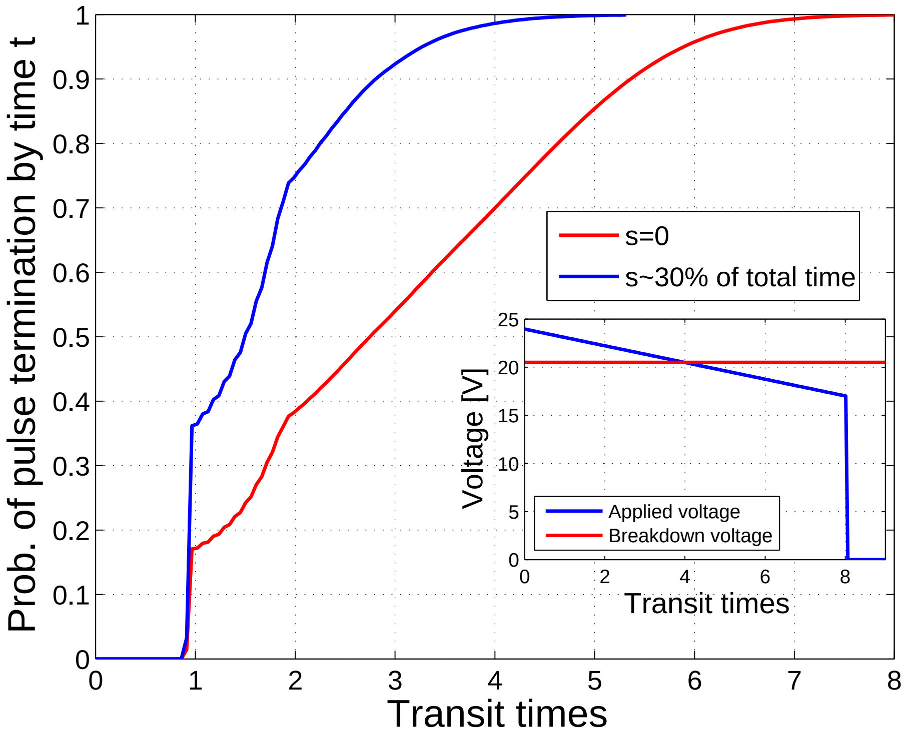


Figure 3.3: Probability that the pulse initiated by an electron injected at $x = 0$ with age s terminates by time t. The red curve shows $P_e(x, t, s)$ for $s = 0$ and the blue curve shows $P_e(x, t, s)$ when the electron that generates the pulse has an age $s \sim 30\%$ of the total time. The inset shows the applied voltage (blue curve) and the breakdown voltage (red curve).

examine the time-varying-field case we obtain more informative results. Figure 3.3b shows the calculated probability that the pulse, initiated by an electron born at location x with age s, terminates by time t. The inset shows the applied voltage in the blue curve and the breakdown voltage level in the red curve. The same two cases were simulated: (i) when the electron that initiates the avalanche has an age $s = 0$ (red curve), and (ii) when the electron that initiates the avalanche has an age $s \sim 30\%$ of the total duration of the applied voltage (blue curve). It can be clearly seen that since the voltage across the device is time dependent the probability of pulse termination will depend on the age of the carrier initiating the avalanche. As expected, Fig. 3.3b shows that the probability of pulse termination is smaller in the case when the carrier that initiates the avalanche has an age $s = 0$ (red curve). This is because, when $s = 0$, the carrier that initiates the avalanche sees a higher electric field compared to the case when the carrier that initiates the avalanche has an age $s \sim 30\%$ of the total duration of the applied voltage (blue curve). This is consistent with the fact that higher electric fields increase the impact ionization probability.

3.4.2 Impulse response under sinusoidal biasing

In the calculations of the impulse response a SAM APD (see Fig. 3.1) with an InP multiplication layer of width $w = 200$ nm is considered. The ionization parameters for InP were extracted from [67]. For reference, we begin by showing the calculated mean impulse response function, triggered by a hole injected at position $x = w = 200$ nm, under a static bias of $V_B = 14.45$ V, as shown in the red curve of Fig. 3.4. The figure also shows four mean impulse response functions triggered by four holes of different ages. As expected, the impulse response is the same regardless the age of the parent hole triggering the avalanche. The mean gain for this case is around $g = 28$. The standard gain bandwidth product for this device is found to be 238 GHz, which is the same as the pulse GBP_p in this case since the field is static.

The calculation of the age-dependent impulse response function in the case of the dynamic bias is shown in Fig. 3.6 for different values of the age variable s. The dynamic electric field profile is also shown in Fig. 3.6. The pulse-gain factor is around 27, making the static and dynamic biasing schemes equivalent in our example at from an “average-gain” perspective. In Fig. 3.6, the curve with $s = 4$ transit times (green curve), for example, corresponds to a parent hole triggering the avalanche 4 transit

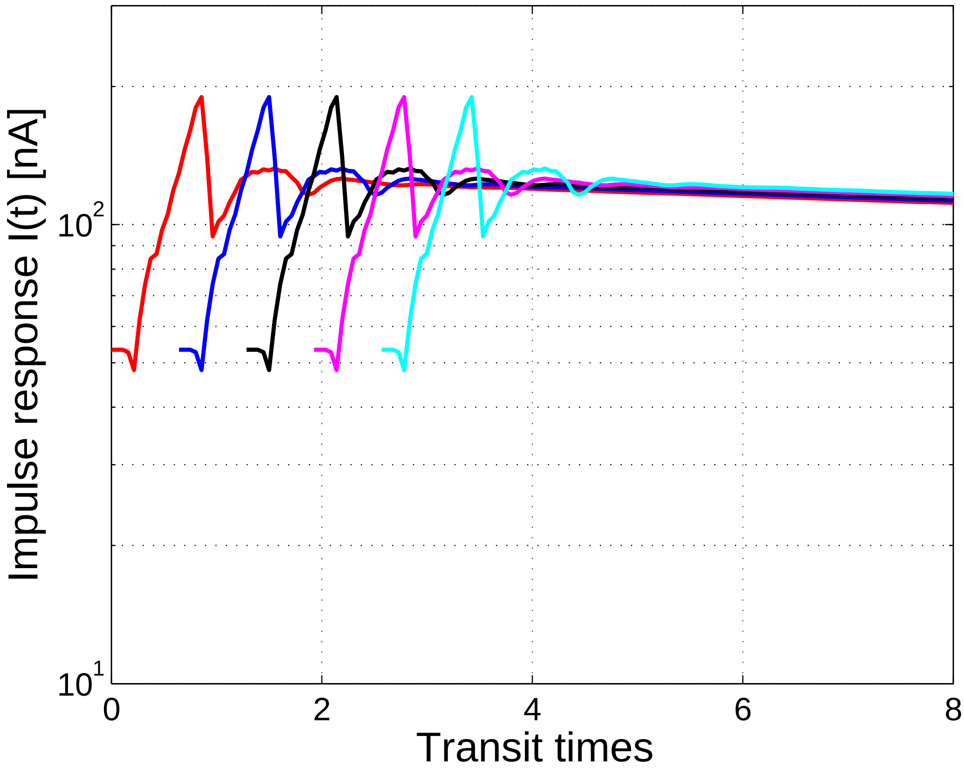


Figure 3.4: Mean impulse response under a constant electric field

times after the launch of the bias at time $t = 0$. It is interesting to note the change in the shape of the impulse response as result of dynamic biasing. In particular, unlike the static-bias case, the impulse response no longer peaks at the hole transit time w/v_h.

Figure 3.6 shows the calculated age-dependent impulse response function under dynamic biasing using the following values for the dynamic-bias in (3.26): $A = 13$ V, $m = 6$ V, and $\phi = \pi/3$. These parameters were chosen, in part, so that the pulse-gain factor (calculated using (20)) is $= 27$, making the static and dynamic biasing

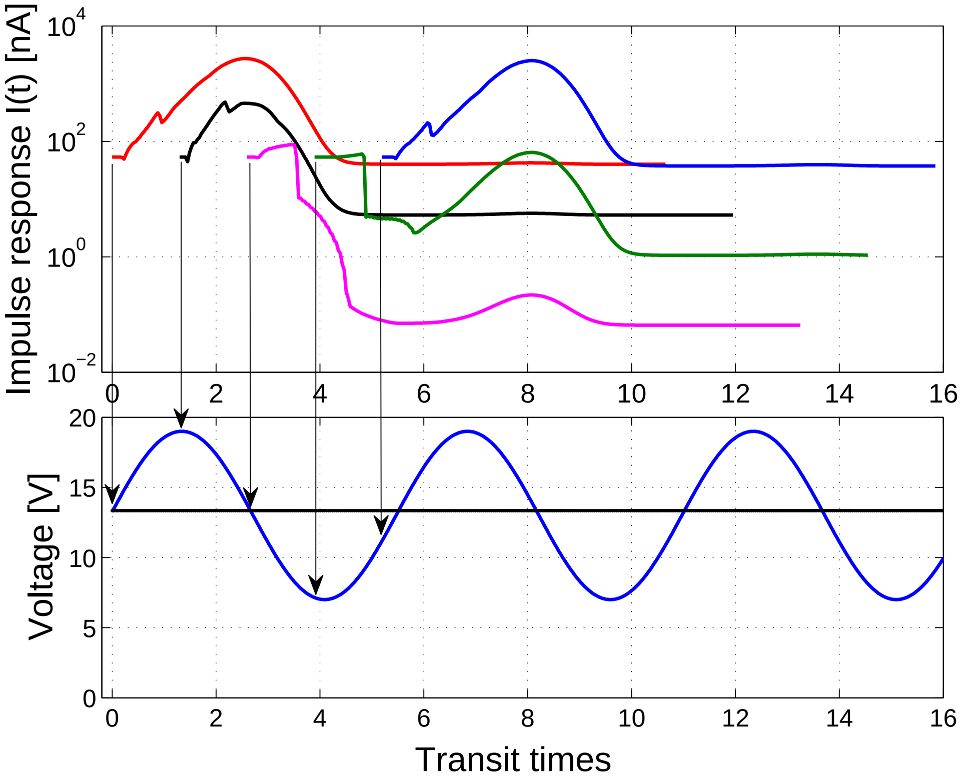


Figure 3.5: Calculated age-dependent impulse response function under a sinusoidal dynamic bias.

schemes equivalent in our example at from an average-gain perspective.It is to be noted that due to the modulated field, for small values of the age (or equivalently for early photons) the tail of the impulse response is far shorter than that for the static-bias impulse response. Moreover, the gain associated with this s value is quite high. This is due to the rise in the field initially, where a high gain is built up, followed by a drop in the field, where the tail of the impulse response is shortened due to high probability of the avalanche terminating. For example, when $s = 0$ the age-dependent bandwidth corresponding to the impulse response is 62 GHz and the

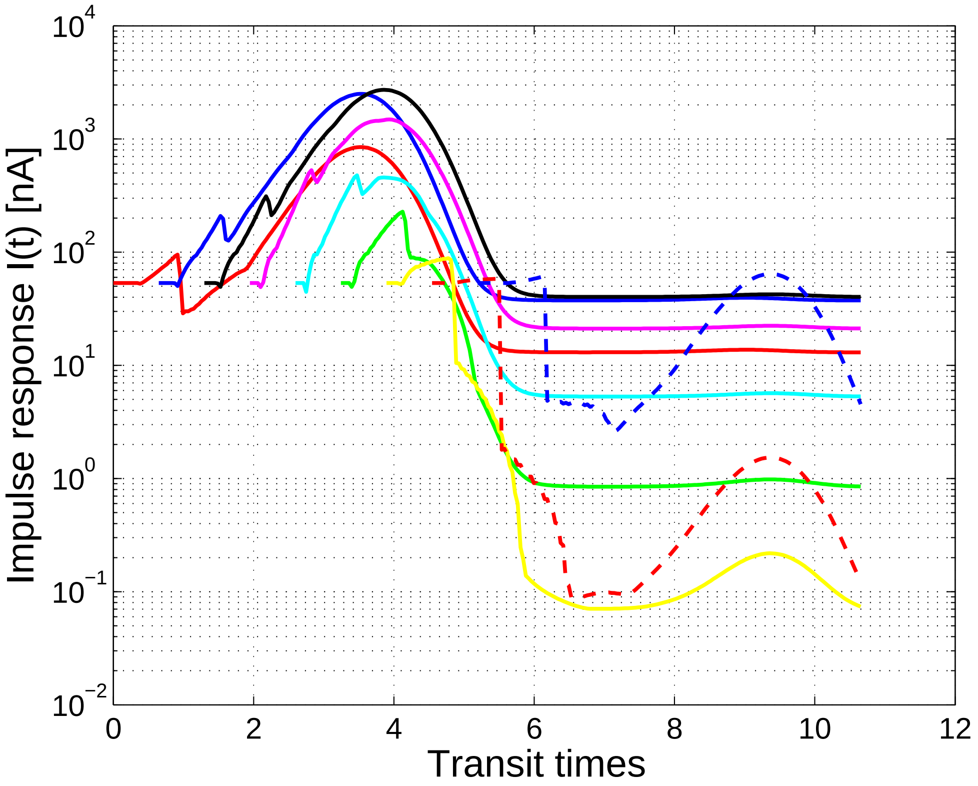


Figure 3.6: Calculated age-dependent impulse response function under dynamic biasing using the following values for the dynamic-bias in (3.26): $A = 13$ V, $m = 6$ V, and $\phi = \pi/3$.

mean gain is 82, while in the static-bias case the bandwidth is 11.4 GHz and the gain is 28. Meanwhile, if we look at larger age variables (corresponding to photons arriving late in the pulse), we will see that the gain is generally small and so is the bandwidth. For example, at $s = 5\pi/3$ the age-dependent bandwidth corresponding to the impulse response is 23 GHz and the mean gain is 3. This is because the carriers have reduced probability of impact ionizing due to the low field in the second half of the pulse.

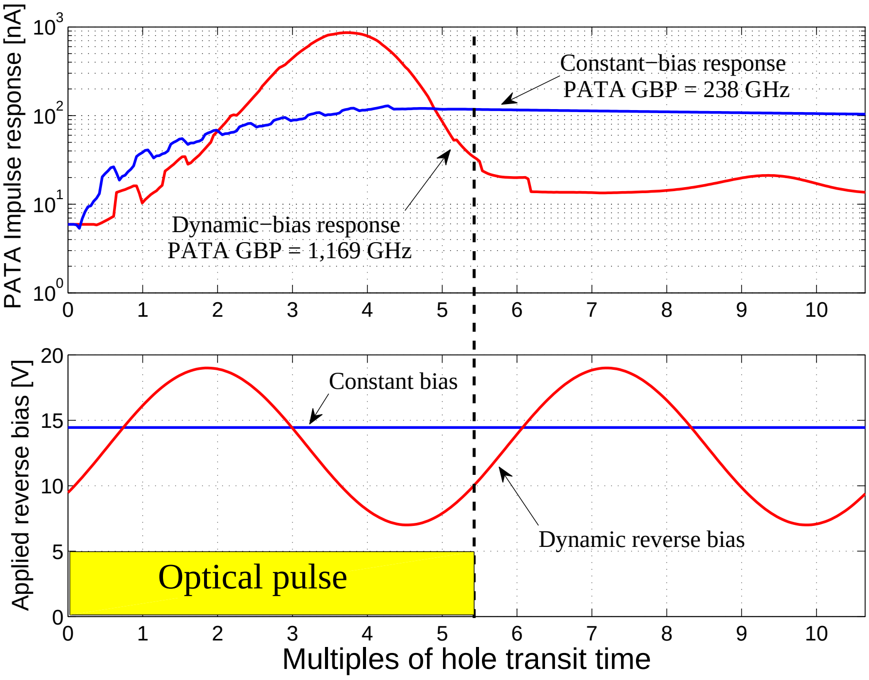


Figure 3.7: Calculated time response to a 8.3-ps rectangular optical pulse of the proposed DBE InP APD, with a sinusoidal-dynamic bias function, and a conventional InP APD. A 5X enhancement in the GBP is predicted.

Figure 3.7 shows the calculated mean pulse-response function, $i_p(t)$, for the same device, once with the sinusoidal dynamic-field profile and once with the traditional static reverse bias. In this example the width of the optical pulse is 8.3 ps (consistent with 60-Gbps NRZ bit stream). The amplitude, dc-bias and phase shift of the sinusoidal periodic function has been selected to maximize the pulse gain-bandwidth product, GBP_p. The calculations predict an enhancement in the pulse gain-bandwidth product from 238 GHz (corresponding to a mean gain of 28 and bandwidth of 8.5 GHz) in the traditional static-bias scheme to a pulse gain-bandwidth product of 1169 GHz (corresponding to a average pulse-gain factor of 27 and pulse bandwidth of 43.3 GHz). This shows that a dynamically biased APD with the bias parameters described earlier can increase the pulse gain-bandwidth product of an APD by a factor of 5 compared to the same APD operated under the conventional static biasing scheme.

3.5 Conclusion

In this chapter a theory that models the impact ionization process in APDs under dynamic biasing was presented. The model allows us to predict the breakdown probabilities, the gain, the mean impulse response, the excess noise factor, and the gain-bandwidth product of SAM APDs under an arbitrary time-varying electric field. The model predicts that by using an sinusoidal biasing scheme we are able to increase the pulse gain-bandwidth product of a SAM APD by a factor of 5 compared to the same APD operated under the conventional static biasing scheme.

Chapter 4

Models for passively quenched SPADs

4.1 Introduction

Recently, new SPAD structures capable of achieving very short quenching times have reactivated the interest in passively quenched SPADs. These devices are the negative feedback avalanche diode (NFAD) [45] and the self-quenching and self-recovery avalanche detector [46]. The operation of self-quenching SPADs and in particular NFADs and self-quenching and self-recovery avalanche detectors rely heavily on the introduction of negative feedback, which rapidly lowers the internal electric field of the avalanche diode following buildup of the avalanche current and forces the stochastic avalanche to terminate quickly. The main motivation for achieving quick quenching times is to increase the repetition rate of passively quenched SPADs. In effect, the quick termination of the avalanche pulse reduces the total charge flow during an avalanche event, which, in turn, reduces the fraction of carriers trapped at defect sites in the SPAD that can lead to afterpulsing. It is well known that the increment of the repetition rate of SPADs is limited by afterpulsing.

Recent experiments on NFAD devices have revealed new aspects on the operation of passively quenched SPADs, which are entirely beyond the scope of the traditional modeling tools. First, the experiments have shown the existence of an oscillatory behavior of a persistent avalanche current. Second, it has been observed that the probability density function of the stochastic quenching time of the persistent avalanche current has an exponential decay. Third, under certain device and operational conditions the stochastic avalanche current can collapse before persistent avalanche current can be realized. All three of these behaviors have not been theoretically explained, since there is no model that is capable of predicting the statistics of the stochastic quenching time in passively quenched SPADs in general and the new generation of SPAD structures that heavily exploit the negative feedback effect in particular.

In this chapter three models to calculate the current-voltage evolution of passively quenched SPAD after an avalanche is triggered by an injected carrier are presented. The first model is the traditional model, which assumes that after an avalanche trigger the voltage across the SPAD remains at a fixed value. It will be shown that this assumption leads to unrealistic consequences. The second model to be examined is a deterministic self-regulating model. This model captures the effect of the feedback from the load in the current-voltage characteristics of a passively quenched SPAD. However, the stochastic nature of the impact ionization process is neglected. The third model is a stochastically self-regulating avalanche model, which captures the dynamic effect of the feedback from the load in the stochastic nature of the impact ionization process. The proposed model represents the first significant expansion beyond the Haitz model [42], for passively quenched SPADs, since it was proposed more than 45 years ago. It will be shown that the proposed model predicts the aforementioned three phenomena that have been experimentally observed in passively quenched NFAD.

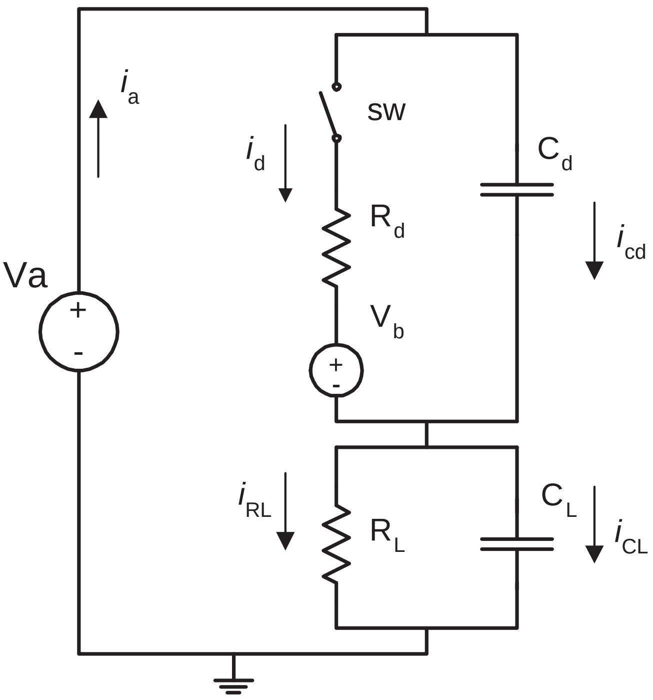


Figure 4.1: Traditional model for a passively quenching SPAD circuit. i_d represents the self sustaining current through the multiplication region of the SPAD; R_d is its equivalent dynamic resistance; C_d is its junction capacitance; R_L is the load resistor and C_L is its parasitic capacitance. The traditional model neglects the effect of feedback on the impact ionization process; it assumes that after the trigger of an avalanche, the electric field remains constant at the breakdown threshold, so that the core of the device is represented by a generator of voltage, V_b.

4.2 Traditional model

Figure 4.1 shows the traditional model of a passively quenched SPAD that was reported by Haitz in 1964 [42]. Since then, this model has become the most accepted one in describing the current-voltage characteristics of a passively quenched SPAD after an avalanche trigger, and it has been adopted by many authors [15]. In this model, the SPAD is represented by its depletion capacitance, C_d, in parallel with a series combination of a switch, *sw*, a dynamic resistance, R_d, and a DC bias source, V_b, representing the breakdown voltage of the SPAD. In the absence of an avalanche trigger the switch is open and the bias across the diode is V_a, which is set slightly above the breakdown voltage, V_b by the excess voltage, V_{ex}. When an avalanche is triggered the model assumes that the switch is instantly closed and the capacitance C_d discharges through the diode's dynamic resistance R_d, which reduces the voltage across the SPAD to a value that depends on the ratio of R_d and R_L. In steady state, the voltage across the SPAD is given by $V_{SPAD} = V_a - V_{ex}R_L/(R_L + R_d) \approx V_b$, for $R_L \gg R_d$. In addition, the steady state avalanche current is given by $I_{ss} \approx V_{ex}/R_L$, and the voltage across the resistor, R_L, is $V_{R_L} \approx V_{ex}$, for $R_L \gg R_d$.

The presence of the DC source, V_b, in the traditional model reflects the assumption that after an avalanche event is triggered the electric field in the avalanche region, responsible for the persistence of impact ionization, remains precisely at breakdown until the persistent current collapses owing to the stochastic fluctuations inherent in the impact ionization process, that is when all carriers chance to exit the multiplication region without ionizing.

4.2.1 Equations that describe the traditional model

In this section, the equations that describe the traditional model of Fig. 4.1 are derived from the instant at which the avalanche is triggered to the spontaneous quenching of the persistent current. It is assumed that an avalanche is triggered at time $t = 0$, which is represented by the closing of the switch, sw. After the system has reached steady state the switch is open once again, which represents the instant at which the avalanche current spontaneously quenches. We can summarize the different stages of the current-voltage characteristics of the circuit in Fig. 4.1 as follows:

1. At $t = 0^-$ (pre-avalanche trigger) the SPAD is reversed bias at V_a. There is no current flow, the voltage across the capacitor C_d is V_a, and the voltage across the capacitor C_L is zero.

2. At $t = 0$ (transient response) the avalanche is triggered (the switch, *sw*, closes). Simultaneously the capacitor C_d discharges through R_d and V_b until it reaches a steady-state voltage $V_{C_d} = (V_b + V_{R_d})$, where V_{R_d} is the steady state voltage across the resistor R_d. As soon as the switch closes the capacitor C_L starts charging until it reaches a voltage $V_{C_L} = (V_A - V_{C_d})$.

3. For $t \gg 0$ (steady state) the voltage across C_d is $V_{C_d} = V_b + V_{R_d}$ and the voltage across C_L is $V_{C_L} = V_A - (V_b + V_{R_d})$. No current is flowing through either of the capacitors, and the only current in the circuit is the current flowing through R_d and R_L and its value is given by $i_{ss} = (V_A - V_b)/(R_d + R_L)$.

4. After a random time the persistent current is quenched due to stochastic fluctuations inherent to the impact ionization process. In Fig. 4.1 the quenching of the persistent current is represented by the closing of the switch sw. At this point the voltage across C_d start recharging until its final voltage level, V_a.

Next, the derivation of the equations that describe the current-voltage characteristics of the circuit of Fig. 4.1 from points 1 to 3 described above is provided. By applying the KCL we have

$$i_d(t) + i_{C_d}(t) = i_{C_L}(t) + i_{R_L}(t) \tag{4.1}$$

where

$$\begin{aligned}
i_d(t) &= \frac{\big(v_{C_d}(t) - V_b\big)}{R_d} && (4.2)\\
i_{C_d}(t) &= C_d \frac{dv_{C_d}(t)}{dt} && (4.3)\\
i_{C_L}(t) &= C_L \frac{dv_{C_L}(t)}{dt} = C_L \frac{d}{dt}\big(V_A - v_{C_d}(t)\big) && (4.4)\\
i_{R_L}(t) &= \frac{v_{C_L}(t)}{R_L} = \frac{\big(V_A - v_{C_d}(t)\big)}{R_L} && (4.5)
\end{aligned}$$

Substituting from (4.2) to (4.5) into (4.1) we have

$$\frac{\big(v_{C_d}(t) - V_b\big)}{R_d} + C_d \frac{dv_{C_d}(t)}{dt} = C_L \frac{d}{dt}\big(V_A - v_{C_d}(t)\big) + \frac{\big(V_A - v_{C_d}(t)\big)}{R_L} \tag{4.6}$$

where $dV_A/dt = 0$. Now

$$v_{C_d}(t) R_L - V_b R_L + R_L R_d C_d \frac{dv_{C_d}(t)}{dt} = R_L R_d C_L \Big(-\frac{dv_{C_d}(t)}{dt}\Big) + R_d V_a - R_d v_{C_d}(t) \tag{4.7}$$

Rearranging we obtain the following first order differential equation

$$\boxed{\frac{dv_{C_d}(t)}{dt} + \frac{R_L + R_d}{R_L R_d \big(C_d + C_L\big)} v_{C_d}(t) = \frac{1}{\big(C_d + C_L\big)} \left(\frac{V_b}{R_d} + \frac{V_A}{R_L}\right) u(t)} \tag{4.8}$$

Applying the Laplace transform to (4.8)

$$\mathscr{L}\left[\frac{dv_{C_d}(t)}{dt} + \frac{R_L + R_d}{R_L R_d \big(C_d + C_L\big)} v_{C_d}(t)\right] = \mathscr{L}\left[\frac{1}{\big(C_d + C_L\big)} \left(\frac{V_b}{R_d} + \frac{V_A}{R_L}\right) u(t)\right]$$

making

$$\boxed{C_1 = \frac{R_L + R_d}{R_L R_d \big(C_d + C_L\big)}}$$

$$\Big[s V_{C_d}(s) - V_{C_d}(0)\Big] + C_1 \cdot V_{C_d}(s) = \left[\frac{1}{\big(C_d + C_L\big)} \left(\frac{V_b}{R_d} + \frac{V_A}{R_L}\right)\right] \frac{1}{s} \tag{4.9}$$

Rearranging

$$V_{C_d}(s)(s + C_1) = \left[\frac{V_b R_L + V_A R_d}{R_d R_L \big(C_d + C_L\big)}\right] \frac{1}{s} + V_{C_d}(0) \tag{4.10}$$

In this way

$$V_{C_d}(s) = \frac{\frac{V_bR_L+V_AR_d}{R_dR_L\left(C_d+C_L\right)}}{s\left[s+\frac{R_L+R_d}{R_LR_d\left(C_d+C_L\right)}\right]} + \frac{V_{C_d}(0)}{\left[s+\frac{R_L+R_d}{R_LR_d\left(C_d+C_L\right)}\right]} \tag{4.11}$$

Using partial fraction expansion for the first term on the right hand side of (4.11) we obtain the final equation for $V_{C_d}(s)$ given by

$$\begin{aligned} V_{C_d}(s) &= \left\{\frac{V_bR_L+V_AR_d}{R_dR_L(C_d+C_L)}\right\}\cdot\frac{1}{s} \\ &\quad -\left\{\frac{V_bR_L+V_AR_d}{R_dR_L(C_d+C_L)}\right\}\cdot\frac{1}{s+\frac{R_L+R_d}{R_LR_d\left(C_d+C_L\right)}} \\ &\quad +\frac{V_{C_d}(0)}{s+\frac{R_L+R_d}{R_LR_d\left(C_d+C_L\right)}} \end{aligned} \tag{4.12}$$

Taking the inverse transform of (4.12) gives

$$v_{C_d}(t) = \frac{V_bR_L+V_AR_d}{R_dR_L(C_d+C_L)}\cdot u(t) \tag{4.13}$$

$$-\frac{V_bR_L+V_AR_d}{R_dR_L(C_d+C_L)}\exp\left(-\frac{R_L+R_d}{R_LR_d(C_d+C_L)}t\right) + V_{C_d}(0)\exp\left(-\frac{R_L+R_d}{R_LR_d(C_d+C_L)}t\right) \tag{4.14}$$

Finally, rearranging (4.14) we obtain the expression for the voltage across the capacitor C_d

$$\begin{aligned} v_{C_d}(t) &= \left\{V_{C_d}(0)-\frac{V_bR_L+V_AR_d}{R_dR_L(C_d+C_L)}\right\}\exp\left(-\frac{R_L+R_d}{R_LR_d(C_d+C_L)}t\right) \\ &\quad +\frac{V_bR_L+V_AR_d}{R_dR_L(C_d+C_L)}\cdot u(t) \end{aligned} \tag{4.15}$$

The rest of the currents and voltages of the circuit can be obtained using (4.15) as follows

$$\begin{aligned} v_{C_L}(t) &= \big(V_A - v_{C_d}(t)\big) && (4.16)\\ i_d(t) &= \frac{\big(v_{C_d}(t) - V_b\big)}{R_d} && (4.17)\\ i_{C_d}(t) &= C_d\frac{d}{dt}v_{C_d}(t) && (4.18)\\ i_{R_L}(t) &= \frac{v_{C_L}(t)}{R_L} && (4.19)\\ i_{C_L}(t) &= C_L\frac{d}{dt}v_{C_L}(t) && (4.20)\\ i_a(t) &= i_{R_L}(t) + i_{C_L}(t) && (4.21) \end{aligned}$$

The voltage across the capacitor C_d during the recharge period is given by the following equation.

$$v_{C_d}(t) = V_a\left(1 - \exp\left[-\frac{t}{R_L(C_d + C_L)}\right]\right) + V_{C_d}(0)\exp\left(-\frac{t}{R_L(C_d + C_L)}\right) \quad (4.22)$$

4.2.2 Current-voltage characteristics

The equations derived above were used to calculate the response of a passively quenched SPAD reversed biased at a voltage V_a. In the simulation the avalanche is triggered at $t = 0$, and the spontaneous quenching of the persistent current is assumed to occur at $t = 15$ ns. The simulated SPAD has an InP multiplication region of 1600 nm. The values of the circuit parameters used are as follows: junction capacitance: $C_d = 0.1$ pF, dynamic resistance: $R_d = 3$ kΩ, load resistor: $R_L = 22$ kΩ and load capacitance: $C_L = 0.51$ pF. The circuit is biased by the power supply at a

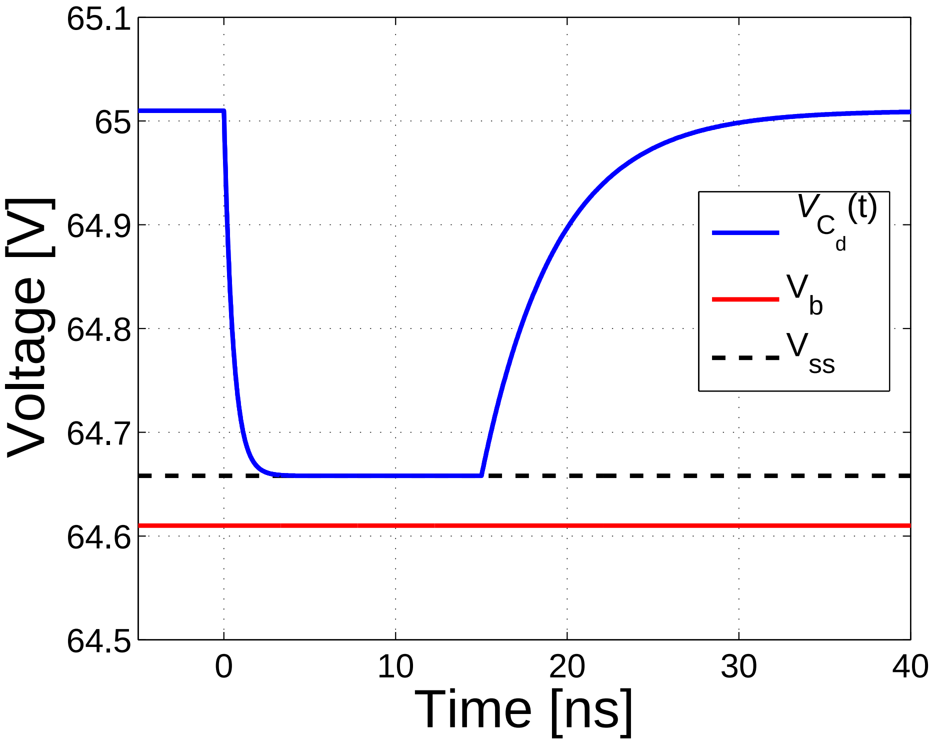


Figure 4.2: Voltage across the capacitor C_d from the instant at which the avalanche is triggered ($t = 0$) until the spontaneous quenching of the persistent current, which occurs at $t = 15$ ns.

voltage $V_a = V_b + V_{ex}$ so that the SPAD is reverse biased beyond its breakdown voltage, V_b by the excess voltage, V_{ex}. The theoretical breakdown voltage was calculated using McIntyre's multiplication expression: [10]

$$M = \frac{1-k}{\exp(-(1-k)\alpha w) - k}, \tag{4.23}$$

which yields $V_b = 64.61$ V. The electron and hole ionization coefficients for InP were obtained from the work of Tan *et al.* [67].

Figure 4.2 shows the voltage across the junction capacitor from the triggering of

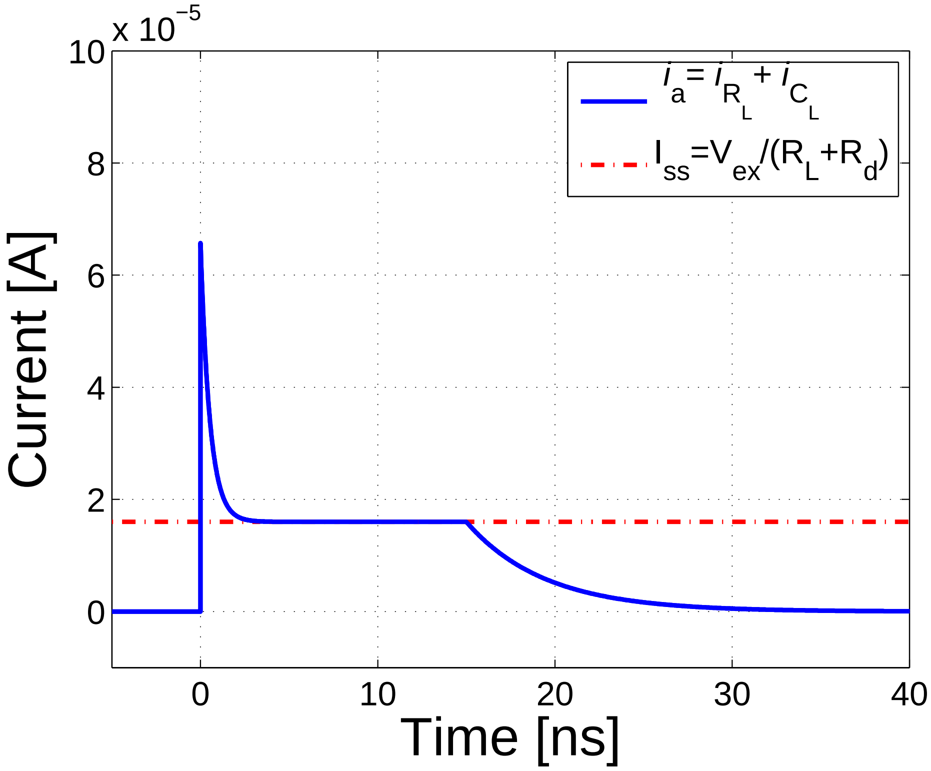


Figure 4.3: Current in the load from the instant at which the avalanche is triggered ($t = 0$) until the spontaneous quenching of the persistent current, which occurs at $t = 15$ ns.

the avalanche to the spontaneous quenching time. Figure 4.3 shows the load current.

4.2.3 Limitations of the traditional model

Next, it will be shown that the assumption that the electric field in the multiplication region remains at the breakdown level leads to unrealistic consequences, namely it implies that the quenching time has memory. In particular, it is shown that the

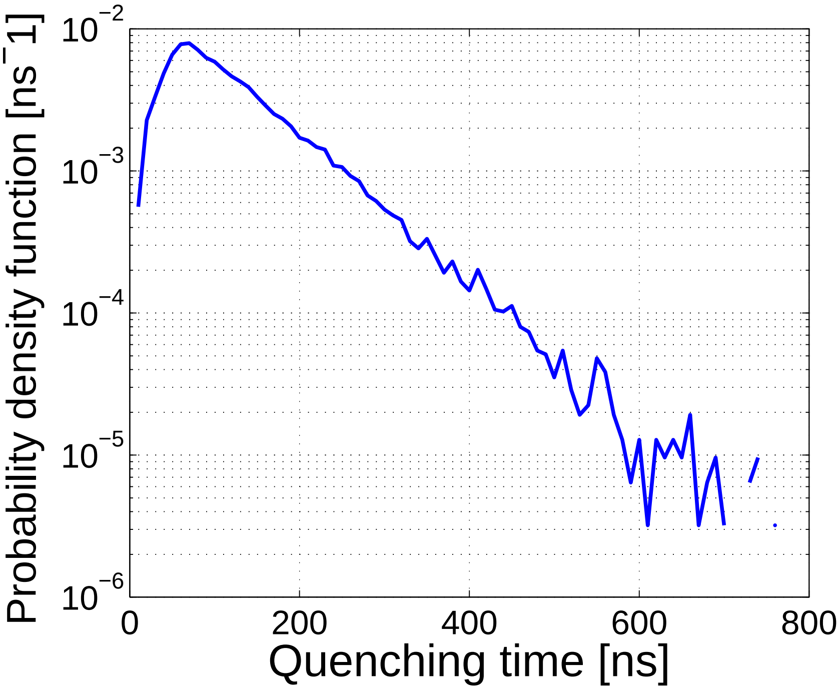


Figure 4.4: Measured pdf of the quenching time [43]. The exponential decay of the pdf implies that the quenching time is memoryless. The data was provided by Princeton Lightwave Inc.

traditional model implies a time dependent decay in the tail of the probability density function (pdf) of the quenching time, which, in turn, implies that the quenching time has memory, in contrast with measurements. More precisely, this result implies that the probability that quenching occurs in a small time interval, $(t,\ t+\Delta t)$, provided it has not occurred earlier, falls as t increases, which describes a system with memory, whereas in a memoryless system this probability is independent of t. This unrealistic consequence of the traditional model is a result of neglecting the effect of the stochastic feedback on the impact ionization process. In fact, in this section it will be shown that, as a consequence of this coupling, the voltage across

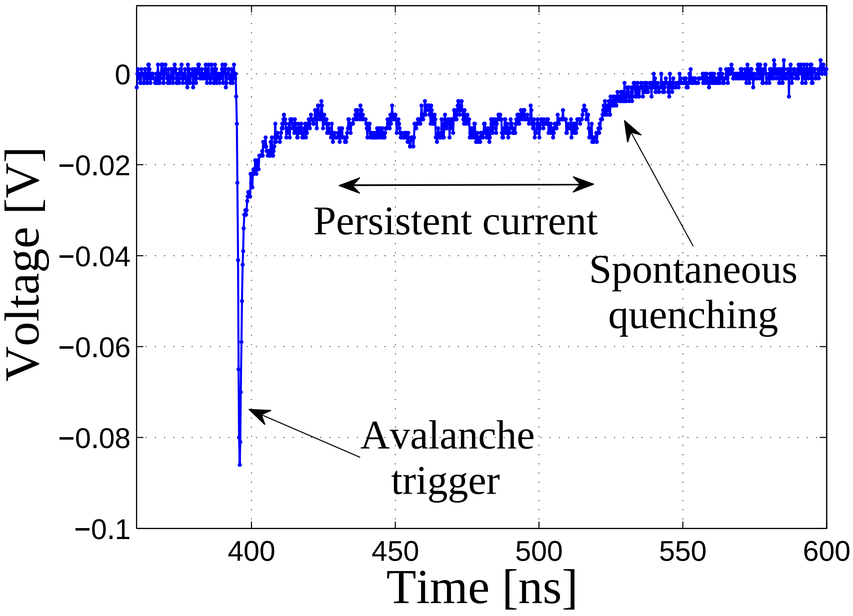


Figure 4.5: Measured voltage across the SPAD for an excess bias of $V_{ex} \approx 1.7$ V. The current shows oscillatory behavior about the steady state before it quenches spontaneously. The complete structure of the device can be found elsewhere [44].

the SPAD, and so the electric field in the multiplication region, oscillate around breakdown before the SPAD quenches spontaneously. This oscillatory behavior has been observed by Itzler *et al.* [45, 43] and is not predicted by the simplistic traditional model.

Following the assumption that after an avalanche event is triggered the electric field in the avalanche region remains precisely at breakdown and building upon the recursive technique for avalanche multiplication developed by Hayat *et al.* [66, 58], it has been shown (details can be found elsewhere [68]) that the probability that the avalanche current, I self quenches before time t has elapsed is given by

$$F_I(t) \triangleq \mathsf{P}\{T_q \leq t\} \approx \exp(-T/t), \tag{4.24}$$

where $T = (IC\tau_0^2 J)/q$, $J = 2/\ln(k)(2/\ln(k) + \frac{1+k}{1-k})$, q is the electronic charge, C is a dimensionless constant of order unity, τ_0 is the average of the electron and hole transit times across the multiplication region and $k = \beta/\alpha$ is the hole/electron ionization coefficient ratio.

A major concern about the formula in (4.24) arising from the constant field assumption, is that it predicts that the quenching time should have memory, since the form of the probability $F_I(t)$ suggests that the pdf of quenching time should diminish with time. The conditional probability that quenching should occur between times t and $t + \Delta t$ is $\mathsf{P}\{T_q < t + \Delta t \mid T_q > t\} = \Delta t \cdot (T/t^2)\exp(-T/t)$, instead of being proportional to Δt as in the memoryless case. Hence, the probability that spontaneous quenching occurs in a small time increment Δt becomes smaller as t increases. This can be understood under the constant field assumption; a delayed quenching time implies that the stochastic avalanche current has been large in the past, which reduces the probability of its quenching in the future.

However, a quenching process with memory is not observed in self quenched SPAD circuits. Figure 4.4 shows measurements of the pdf of the quenching time showing exponential decay, which implies that the decay process is memoryless. The data was provided by Princeton Lightwave Inc. The traditional model also fails to predict the oscillatory behavior in persistent current also observed by Itzler *et al.* [45, 43] and shown in Fig. 4.5.

The above discussion suggests that a more realistic approach is needed to model the current voltage evolution and quenching characteristics of a passively quenched SPAD. In the next section a first attempt to achieve a more complete description of a passively quenched SPAD is presented.

4.3 Deterministic self-regulating model

This model captures the feedback from the load resistor but ignores the stochastic element of the avalanche persistent current. Once an avalanche is triggered (while the diode is biased above breakdown), the average avalanche current grows exponentially according to the theory of average impulse response of APDs above breakdown [58, 69]. This growth tends to discharge the capacitor, and therefore, reduce the junction voltage (voltage across C_d), which, in turn, causes the avalanche current to drop. The reduction in the avalanche current continues and the junction bias drops below the breakdown voltage. After this point of time, the DC source begins to recharge the capacitor with a time constant $\tau_r \approx R_L C_d$, causing the avalanche current to increase once again. The repetition of this process can yield an oscillatory behavior, where the current inside the diode oscillates about the steady state value. Moreover, the field inside the multiplication region of the APD oscillates above and below the breakdown threshold.

4.3.1 Mathematical description

The current in the voltage-controlled current source (VCCS), shown in Fig. 4.6, is controlled by the voltage across the junction capacitor, V_{C_d}, and the breakdown voltage, V_b. The equation that describe the current in the VCCS is given by

$$i_d(t) = I_o e^{\alpha \int_0^t (v_{C_d}(t) - V_b) dt}, \tag{4.25}$$

where $\alpha > 0$ and $I_0 = qv_e/w$ is the initial current triggered by a single photo-absorption event.

From Fig. 4.6 we know that the avalanche current, $i_a(t)$, must be equal to the current through the APD and the load. More specifically, the avalanche current, $i_a(t)$, is equal to the current through the capacitor, C_d, plus the current in the

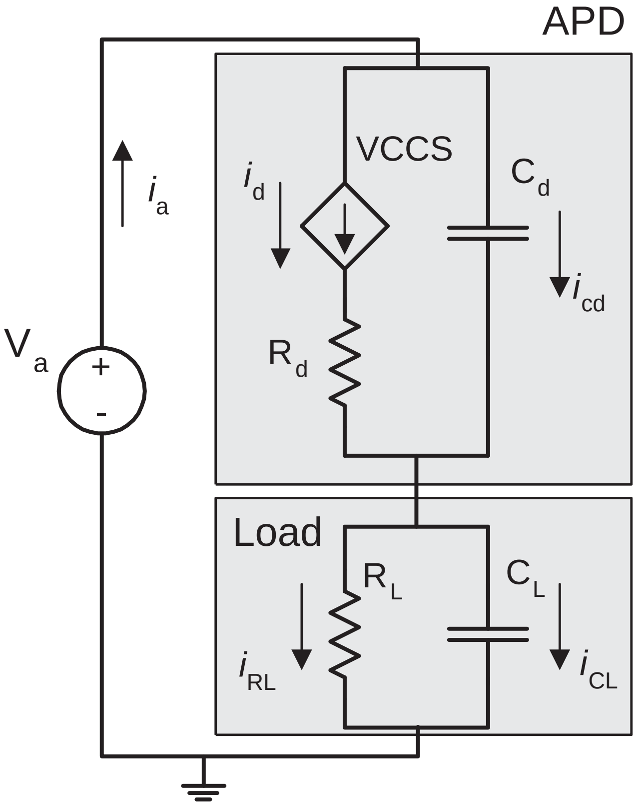


Figure 4.6: Circuit that models the diode and the quenching circuit.

VCCS, $i_d(t)$. Moreover, $i_a(t)$ is equal to the current through the load resistor R_L plus the current in the load capacitor C_L. Thus

$$i_a(t) = i_d(t) + i_{C_d}(t)$$

and,

$$i_a(t) = i_{C_L}(t) + i_{R_L}(t)$$

then

$$i_d(t) + i_{C_d}(t) = i_{C_L}(t) + i_{R_L}(t) \tag{4.26}$$

where

$$i_d(t) = I_o e^{\alpha \int_0^t (v_{C_d}(t)-V_b)dt} \tag{4.27}$$

$$i_{C_d}(t) = C_d \frac{dv_{C_d}(t)}{dt} \tag{4.28}$$

$$i_{C_L}(t) = C_L \frac{dv_{C_L}(t)}{dt} = C_L \frac{d}{dt}\big(V_A - v_{C_d}(t)\big) \tag{4.29}$$

$$i_{R_L}(t) = \frac{v_{C_L}(t)}{R_L} = \frac{\big(V_A - v_{C_d}(t)\big)}{R_L} \tag{4.30}$$

Substituting from (4.27) to (4.30) into (4.26) we have

$$I_o e^{\alpha \int_0^t (v_{C_d}(t)-V_b)dt} + C_d \frac{dv_{C_d}(t)}{dt} = C_L \frac{d}{dt}\big(V_A - v_{C_d}(t)\big) + \frac{\big(V_A - v_{C_d}(t)\big)}{R_L}, \tag{4.31}$$

where $dV_A/dt = 0$. Rearranging and solving for $dv_{C_d}(t)/dt$

$$\begin{aligned}
C_d \frac{dv_{C_d}(t)}{dt} + C_L \frac{dv_{C_d}(t)}{dt} &= \frac{\big(V_A - v_{C_d}(t)\big)}{R_L} - I_o e^{\alpha \int_0^t (v_{C_d}(t)-V_b)dt} \\
\big(C_d + C_L\big) \frac{dv_{C_d}(t)}{dt} &= \frac{\big(V_A - v_{C_d}(t)\big)}{R_L} - I_o e^{\alpha \int_0^t (v_{C_d}(t)-V_b)dt} \\
\frac{dv_{C_d}(t)}{dt} &= \frac{1}{\big(C_d + C_L\big) R_L} \Big\{ V_A - v_{C_d}(t) - R_L I_o e^{\alpha \int_0^t (v_{C_d}(t)-V_b)dt} \Big\} \qquad (4.32)
\end{aligned}$$

Using the fact that

$$\int \frac{dv_{C_d}(t)}{dt} dt = v_{C_d}(t) - V_{C_d}(0)$$

where $V_{C_d}(0) = V_b + \Delta V$ and taking the integral in both sides of (4.32) gives

$$v_{C_d}(t) = V_b + \Delta V + \frac{1}{\big(C_d + C_L\big) R_L} \int_0^t \Big\{ V_A - v_{C_d}(s) - R_L I_o e^{\alpha \int_0^u \big(v_{C_d}(u)-V_b\big)\, du} \Big\}\, ds \tag{4.33}$$

Substituting V_A by $V_b + \Delta V$ we get to the final equation

$$v_{C_d}(t) = V_b + \Delta V + \frac{1}{\big(C_d + C_L\big) R_L} \int_0^t \Big\{ V_b + \Delta V - v_{C_d}(s) - R_L I_o e^{\alpha \int_0^u \big(v_{C_d}(u)-V_b\big)\, du} \Big\}\, ds \tag{4.34}$$

The effect of the resistor R_d is captured by the VCCS. In particular, any change in the voltage across R_d is absorbed by the voltage across the VCCS so that voltage across the series combination V_{R_d} and V_{VCCS} is always equal to the voltage across the junction capacitor. As a result, the resistor R_d does not appear in (4.34).

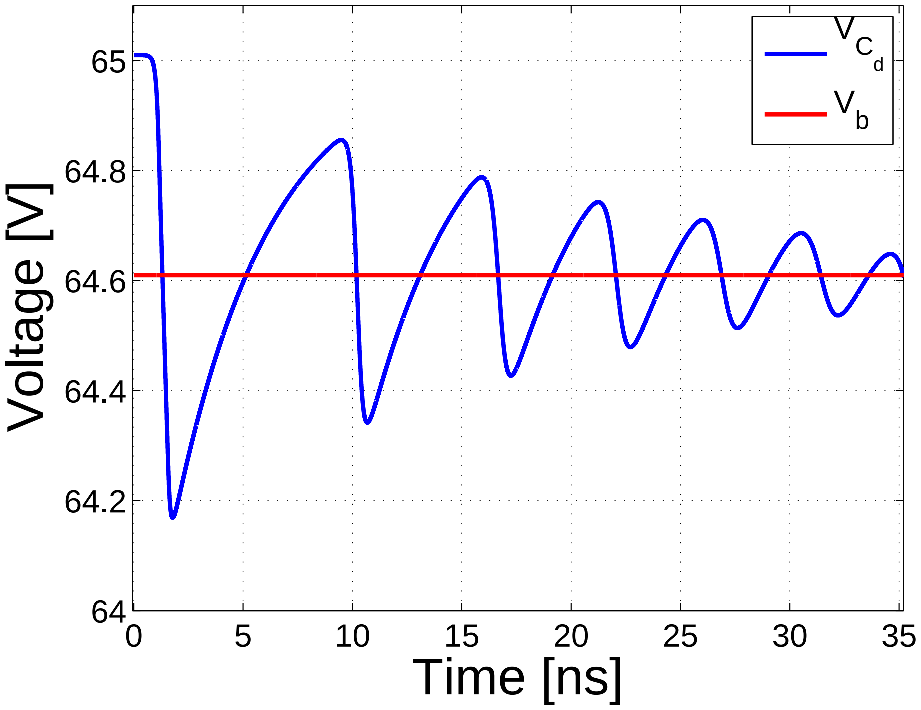


Figure 4.7: Voltage across the capacitor C_d from the instant at which the avalanche is triggered ($t = 0$) until the spontaneous quenching of the persistent current, which occurs at $t = 15$ ns.

4.3.2 Delay in the response of the VCCS

To account for the non-instantaneous response of the field as the voltage across the diode is changed due to feedback, a *delay* in the dynamics of the diode is introduced. The origin of the delay has to do with the non-instantaneous re-arrangement of the charge in the multiplication region as a result of the finite time the carries take to travel the multiplication region and effectively contribute to the negative feedback. The effect of the delay on the dynamics of the diode can be an important factor that

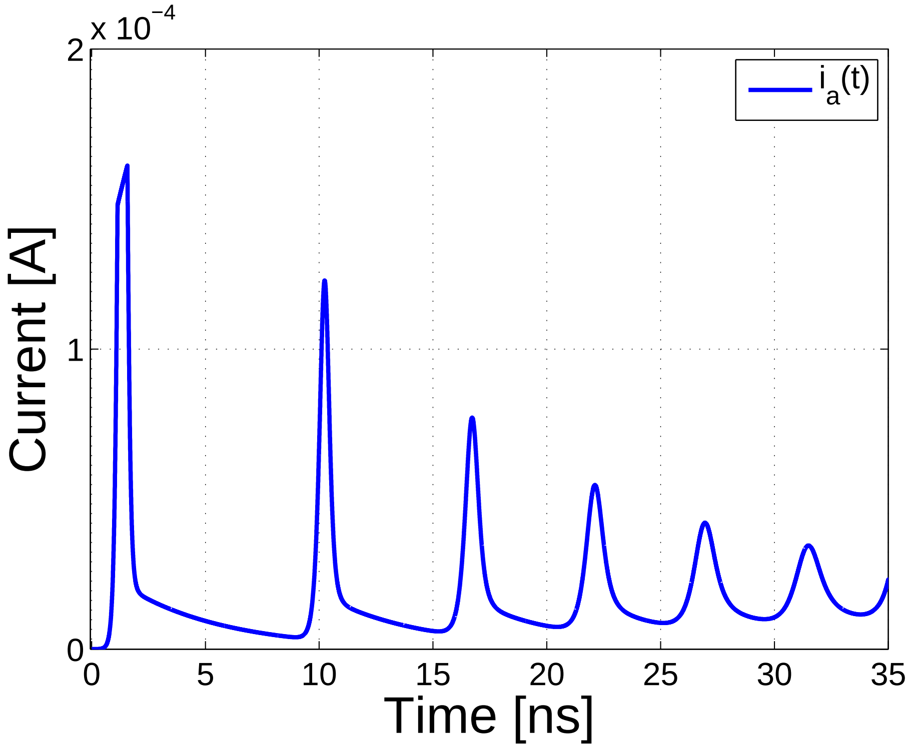


Figure 4.8: Current in the load from the instant at which the avalanche is triggered ($t = 0$) until the spontaneous quenching of the persistent current, which occurs at $t = 15$ ns.

affects the action of the feedback. The delay is introduced by retarding the response of the VCCS to the changes in the field across the junction capacitance.

4.3.3 Results

Next, results obtained from the simulation of the circuit of Fig. 4.6 by solving the equations derived above are shown. In the simulations it has been assumed that the voltage dependent source, i_d, that generates the exponential growth of the current inside the diode responses to the voltage across the capacitor C_d with a finite delay. This delay provides the circuit with feedback, which is responsible of creating an oscillatory response. The value of the delay is given by $delay = (RC \text{ time constant})/d$, where d is a constant. The simulated SPAD has the same circuit parameters as the ones used in the traditional model. Those parameters are repeated here for completeness: multiplication region width: 1600 nm (InP), junction capacitance: $C_d = 0.1$ pF, dynamic resistance: $R_d = 3$ kΩ, load resistor: $R_L = 22$ kΩ and load capacitance: $C_L = 0.51$ pF. The circuit is biased by the power supply at a voltage $V_a = V_b + V_{ex}$ so that the SPAD is reverse biased beyond its breakdown voltage, V_b by the excess voltage, V_{ex}.

Figure 4.7 shows the calculated voltage across the load resistor according to this deterministic model. It can be seen from the figure that the oscillations are indeed centered about the steady state current through the diode. Figure 4.8 shows the calculate persistent avalanche current, which also shows the oscillatory behavior seen in the load resistor.

The oscillatory behavior is governed by two factors: (1) how quickly the junction capacitor can be discharged, which, in turn, depends on the RC time constant and on the growth rate of the avalanche pulse; and (2) how fast the change in the junction voltage can alter the avalanche current in the diode. The latter effect, characterized by a delay factor, d, is akin to a delay brought about by an inductor, which induces oscillations.

4.4 Stochastically self-regulating avalanche model

In this section a stochastically self-regulating avalanche model is presented. The model combines an analytical circuit representation with a Monte Carlo simulator of the multiplication process to calculate the response of a passively quenched SPAD, reverse biased above breakdown. This model considers a passively quenched SPAD as a closed loop system, capturing the effect of the feedback introduced by the load on the stochastic nature of the avalanche multiplication. The carrier dynamics and multiplication are simulated at successive instants by a custom made Monte Carlo simulator. Notably, the self-regulating avalanche model is able to predict the stochastic current-voltage evolution and quenching characteristics of the new generation of SPAD structures that use the negative feedback effect. In addition, the stochastically self-regulating avalanche model correctly predicts that the decay of the tail of the pdf of the stochastic quenching time of the persistent avalanche current is exponential, in agreement with the observed memoryless behavior. The model can also be used to predict the conditions under which very short quenching times are achieved, similar to the behavior seen in NFADs.

4.4.1 Model

Figure 4.9 shows the proposed stochastically self-regulating avalanche model of a passively quenched SPAD [70]. The main difference between this and the traditional model of Fig. 4.1 is that the switch and voltage generator V_b in Fig. 4.1, which represented the on/off state of the SPAD, are now replaced by a stochastic voltage controlled current source (VCCS) i_d. A Monte Carlo simulator of the dynamics of the avalanche multiplication is used to produce the current in the VCCS. Moreover, as the voltage across R_L changes so does the bias on the SPAD, and hence also the stochastic avalanche current i_d, since the ionization coefficients, α and β used by

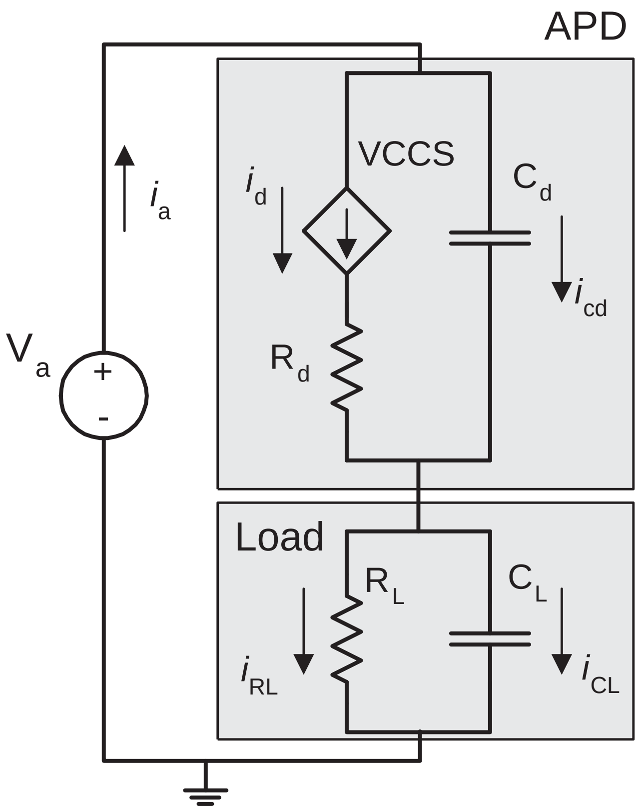


Figure 4.9: Stochastically self-regulating avalanche model for passively quenched SPADs. The circuit represents a series combination of a SPAD and a negative feedback load. The load is described as a parallel combination of a resistance, R_L and a capacitance, C_L. The SPAD is modeled as two parallel branches; one branch consists of the diode depletion capacitance, C_d, the other includes the Monte Carlo simulator, which is represented by the stochastic voltage controlled current source (VCCS) i_d. The resistor R_d, in series with the VCCS, accounts for the resistance of the bulk regions.

the Monte Carlo simulator depend on the instantaneous electric field through the junction capacitance, C_d. As the carriers multiply stochastically their resulting cur-

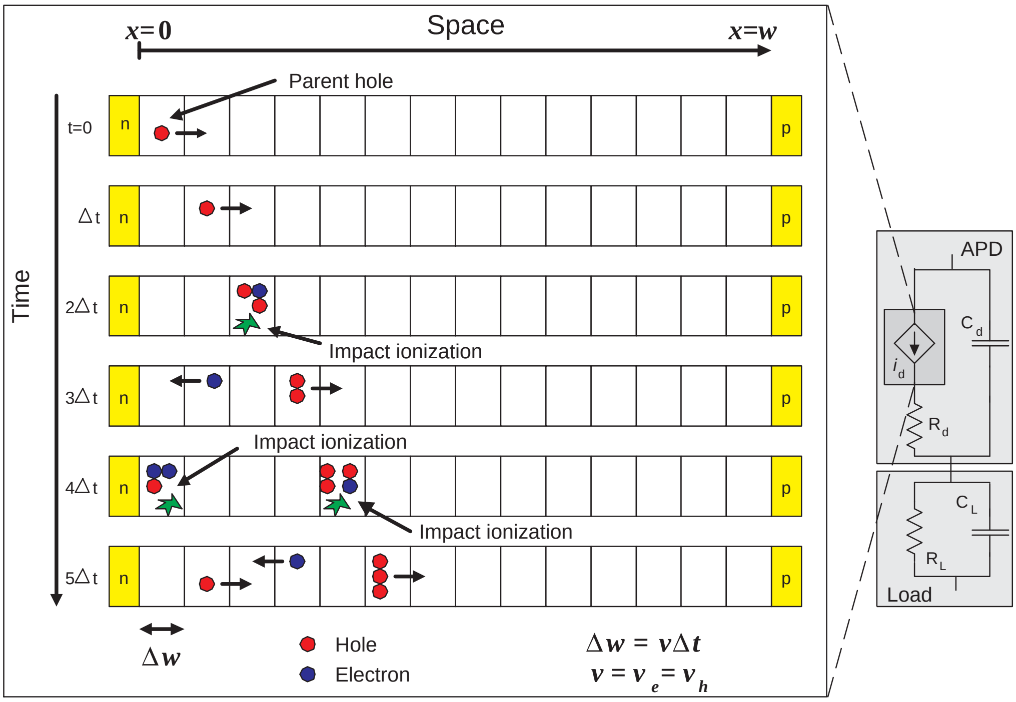


Figure 4.10: Monte-Carlo simulator for i_d. The expanded section on the left describes the simulator represented in the circuit on the right by the stochastic VCCS i_d. In the example a hole is injected at the start of the multiplication region, $x = 0$, at time $t = 0$. At time $2\Delta t$ the first impact ionization occurs and as a result one hole and one electron are created in the multiplication region. For simplicity it is assumed that electrons and holes have the same drift velocity, v, i.e., $v = v_e = v_h$.

rent is calculated using Ramo's theorem [17] from the number of carriers inside the multiplication region. Hence, by contrast with the traditional model, the stochastically self-regulating avalanche model captures the effect of feedback on the stochastic evolution of carrier multiplication associated with the persistent avalanche current.

Monte Carlo simulator

Figure. 4.10 illustrates the operation of the Monte Carlo simulator used to produce the current in the VCCS by mimicking a SPAD with a dynamic and stochastic bias. In the simulator, the multiplication region extends from $x = 0$ to $x = w$, and this region is divided into L small increments each of width Δw, representing L bins, each terminating at $x_k = k\Delta w$, where $k = 1, 2, 3, \ldots,\ L$. The total number of bins L, which determines the spatial increment Δw, is chosen so that the product $\alpha \cdot \Delta w$ $(\beta \cdot \Delta w)$ is small. The binomial model described here is a good approximation of the continuous-space ionization process provided that $\alpha \cdot \Delta w(\beta \cdot \Delta w) \ll 1$. Indeed, in the simulations with $L = 1600$ and $\Delta w = 1$ nm. The maximum value for $\beta \cdot \Delta w \approx 9 \times 10^{-4}$, which is much less than one. The total simulation time, from $t = 0$ to $t = T_{max}$, is divided into M small increments, each of duration Δt, where Δt represents the time taken for a carrier to travel a distance Δw. For simplicity it is assumed that electrons and holes have the same drift velocity, v. A particular time in the simulation is described as $t_j = j\Delta t$, where $j = 1, 2, 3, \ldots, M$. It is assumed that holes (electrons) move in the positive (negative) direction of x. In addition, it is employed the common rule that at any time interval $[t, t + \Delta t]$ the probability that an electron will impact ionize is given by $\alpha\big(E_{C_d}(t)\big)\Delta w$, where $E_{C_d}(t)$ is the instantaneous electric field through C_d. Similarly, the probability that a hole will impact ionize is given by $\beta\big(E_{C_d}(t)\big)\Delta w$.

To track the stochastic evolution of the total number of carriers at each instant $X_e(t_j, x_k)$ and $X_h(t_j, x_k)$ are defined as the number of electrons and holes, respectively, at bin location x_k and time t_j. The effect of the dead space is ignored in the carrier multiplication process since SPADs with thick multiplication regions are considered ($> 1\mu$m), which are preferred for Geiger mode operation [16]. It is well known that for thick multiplication regions the effect of the dead space does not play a relevant role in the carrier multiplication process. Therefore, in thick multiplication regions the dead space can be ignored. On the other hand, for thin multiplication regions (< 500 nm) the dead space becomes important and to accurately describe the impact ionization process the dead space must be taken into account. Considering the transport and ionization properties of the carriers, and by ignoring their dead space, we can write the following stochastic dynamical equations:

$$\begin{aligned} X_e(t_{j+1}, x_k) &= X_e(t_j, x_{k+1}) + b\Big(X_e(t_j, x_{k+1}), \alpha(E_{C_d}(t))\Delta w\Big) \\ &\quad + b\Big(X_h(t_j, x_{k-1}), \beta(E_{C_d}(t))\Delta w\Big) \end{aligned} \tag{4.35}$$

and

$$\begin{aligned} X_h(t_{j+1}, x_k) &= X_h(t_j, x_{k-1}) + b\Big(X_h(t_j, x_{k-1}), \beta(E_{C_d}(t))\Delta w\Big) \\ &\quad + b\Big(X_e(t_j, x_{k+1}), \alpha(E_{C_d}(t))\Delta w\Big). \end{aligned} \tag{4.36}$$

In the above equations the notation $b(n, p)$ stands for a binomial random variable of size n and success probability p; thus $b(n, p)$ represents the total number of successful ionization events resulting from n independent attempts, each with success probability p. The boundary conditions at $k = 1$ and L must clearly be handled separately in (4.35) and (4.36).

To trigger an avalanche the multiplication region is reverse biased above breakdown and a carrier is injected at the start of the multiplication region. The stochastic dynamical equations (4.35) and (4.36) are implemented at every time increment and samples of the required binomial random variables are generated. Figure 4.10 shows a fictitious example which illustrates the total number of carriers in the multiplication region at each time, the direction of motion of the carriers and the impact ionization events generated, during 5 intervals of time Δt, by a hole injected at time $t = 0$ and at location $x = 0$. After time t_j has elapsed the instantaneous stochastic current $i_d(t_j)$ is calculated using Ramo's theorem:

$$i_d(t_j) = \frac{qv}{w} \sum_{k=1}^{L} \Big(X_e(t_j, k) + X_h(t_j, k) \Big). \tag{4.37}$$

All other currents and voltages in Fig. 4.9 are calculated by solving the standard circuit equations. The instantaneous values of the electric field dependent ionization coefficients are recalculated at every time increment to allow for the change of voltage across the SPAD as a result of the instantaneous feedback from the load.

4.4.2 Derivation of the circuit equations

Finite difference equations

Applying Kirchhoff current and voltage laws to the circuit of Fig. 4.9 we can establish the following equations:

$$i_d + i_{C_d} = i_L + i_{CL} \tag{4.38}$$

$$V_A = v_{C_d} + v_{R_L} + v_L \tag{4.39}$$

$$V_A = v_{C_d} + v_{C_L} \tag{4.40}$$

$$v_{R_L} + v_L = v_{C_L} \tag{4.41}$$

$$i_{C_d} = C_d \frac{d}{dt} v_{C_d} = C_d v'_{C_d} \tag{4.42}$$

$$i_{C_L} = C_L \frac{d}{dt} v_{C_L} = C_L v'_{C_L} \tag{4.43}$$

$$v_L = L \frac{d}{dt} i_L = L i'_L \tag{4.44}$$

from (4.38)

$$\begin{aligned} i_L &= i_d + i_{C_d} - i_{C_L} \\ i_L &= i_d + C_d v'_{C_d} - C_L v'_{C_L} \\ i_L &= i_d + C_d v'_{C_d} + C_L v'_{C_d} \\ i_L &= i_d + \Big(C_d + C_L\Big) v'_{C_d} \end{aligned} \tag{4.45}$$

In addition

$$\begin{aligned} V_A &= v_{C_d} + v_{R_L} + v_L \\ V_A &= v_{C_d} + R_L \Big[i_d + \Big(C_d + C_L\Big) v'_{C_d} \Big] + L \frac{d}{dt} i_L \end{aligned} \tag{4.46}$$

Substituting (4.45) into (4.46) we obtain

$$V_A = v_{C_d} + R_L \Big[i_d + \Big(C_d + C_L\Big) v'_{C_d} \Big] + L \Big[i'_d + \Big(C_d + C_L\Big) v''_{C_d} \Big] \tag{4.47}$$

Rearranging this equation we arrive at

$$\begin{aligned} V_A &= v_{C_d} + R_L i_d + R_L (C_d + C_L) v'_{C_d} + L \Big[i'_d + L(C_d + C_L) v''_{C_d} \Big] \\ L(C_d + C_L) v''_{C_d} &+ R_L (C_d + C_L) v'_{C_d} + v_{C_d} = V_A - R_L i_d - L i'_d \\ v''_{C_d} &+ \frac{R_L}{L} v'_{C_d} + \frac{1}{L(C_d + C_L)} v_{C_d} = \frac{V_A - R_L i_d}{L(C_d + C_L)} - \frac{i'_d}{(C_d + C_L)} \end{aligned} \tag{4.48}$$

A more compact form of (4.48) can be obtained by grouping the constant terms into three new constants. Thus, by making

$$C_1 = \frac{R_L}{L} \tag{4.49}$$

$$C_2 = \frac{1}{L(C_d + C_L)} \tag{4.50}$$

$$C_3 = \frac{1}{C_d + C_L}, \tag{4.51}$$

and substituting from (4.49) to (4.51) into (4.48) we obtain the second order differential equation to solve.

$$v''_{C_d} + v'_{C_d} C_1 + v_{C_d} C_2 = (V_A - R_L i_d) C_2 - i'_d C_3 \tag{4.52}$$

Equation (4.52) can be solved numerically by using the finite difference method. First, we need to express the derivatives in terms of finite difference. The finite difference approximation of first order derivatives is given by

$$f'(x) = \frac{f(x) + f(x-1)}{h}, . \tag{4.53}$$

Similarly, the finite difference approximation of second order derivatives is

$$f''(x) = \frac{f(x) - 2f(x-1) + f(x-2)}{h^2} \tag{4.54}$$

Substituting (4.53) and (4.54) into (4.52) we obtain the following equation

$$\frac{v_{C_d}(x) - 2v_{C_d}(x-1) + v_{C_d}(x-2)}{\Delta t^2} + C_1 \frac{v_{C_d}(x) - v_{C_d}(x-1)}{\Delta t} + v_{C_d}(x) C_2 = \\ (V_A - i_d R_L) - i'_d C_3$$

Rearranging the previous equation to express it in terms of $v_{C_d}(x)$ we obtain

$$\frac{v_{C_d}(x) - 2v_{C_d}(x-1) + v_{C_d}(x-2)}{\Delta t^2} + C_1 \frac{v_{C_d}(x) - v_{C_d}(x-1)}{\Delta t} + v_{C_d}(x) C_2 = \\ (V_A - i_d R_L) - i'_d C_3$$

$$\left\{ \frac{1}{\Delta t^2} + \frac{C_1}{\Delta} + C_2 \right\} v_{C_{d}(x)} - \left\{ \frac{2}{\Delta t^2} + \frac{C_1}{\Delta t} \right\} v_{C_d}(x-1) + \frac{1}{\Delta t^2} v_{C_d} = \\ (V_A - i_d R_L) - i'_d C_3$$

$$\begin{aligned} v_{C_d}(x) = \left\{\frac{\frac{2}{\Delta t^2}+\frac{C_1}{\Delta t}}{\frac{1}{\Delta t^2}+\frac{C_1}{\Delta t}+C_2}\right\} v_{C_d}(x-1) - \frac{1}{\Delta t^2\left\{\frac{1}{\Delta t^2}+\frac{C_1}{\Delta t}+C_2\right\}} v_{C_d}(x-2) + \\ \frac{(V_A - i_d R_L)C_2}{\left\{\frac{1}{\Delta t^2}+\frac{C_1}{\Delta t}+C_2\right\}} - \frac{(i_d(x) - i_d(x-1))C_3}{\left\{\frac{1}{\Delta t^2}+\frac{C_1}{\Delta t}+C_2\right\}\Delta t} \end{aligned}$$

$$\begin{aligned} v_{C_d}(x) = \left\{\frac{2 + C_1\Delta t}{1 + C_1\Delta t + C_2\Delta t^2}\right\} v_{C_d}(x-1) - \\ \frac{v_{C_d}(x-2)}{1 + C_1\Delta t + C_2\Delta t^2} + \frac{(V_A - i_d R_L)C_2\Delta t^2}{1 + C_1\Delta t + C_2\Delta t^2} - \frac{(i_d(x) - i_d(x-1))C_3\Delta t}{1 + C_1\Delta t + C_2\Delta t^2} \end{aligned} \tag{4.55}$$

Equation (4.55) can be written in a more compact way by grouping the constant terms in two new constants given by

$$\begin{aligned} C_4 &= 2 + C_1\Delta t \\ C_5 &= 1 + C_1\Delta t + C_2\Delta t^2 \end{aligned} \tag{4.56}$$

Finally, substituting the constant terms defined in (4.56) into (4.55) we obtain the equation that describe the voltage across the capacitor C_d

$$\begin{aligned} v_{C_d}(x) = \frac{C_4}{C_5} \cdot v_{C_d}(x-1) - \frac{1}{C_5} \cdot v_{C_d}(x-2) + (V_A - i_d R_L) \cdot \frac{C_2\Delta t^2}{C_5} \\ +(i_d(x) - i_d(x-1)) \cdot \frac{C_3\Delta t}{C_5} \end{aligned} \tag{4.57}$$

Equation (4.57) is solved in the interval $0 < x < w$ for each time increment Δt. The solution requires an initial condition for $v_{C_d}(x-1)$ and $v_{C_d}(x-2)$. The value of i_d is obtained from the Monte-Carlo simulator.

4.4.3 Results

Next, the stochastically self-regulating avalanche model is used to simulate the unique attributes of the new generation of self quenched SPADs, which can be described by the circuit shown in Fig. 4.9. In particular, the statistics of the quenching time and the observed oscillatory behavior of the persistent current will be predicted. A passively quenched InP SPAD is simulated using the following values of the circuit parameters: junction capacitance: $C_d = 0.1$ pF, load resistor: $R_L = 22$ kΩ and load capacitance: $C_L = 0.001$ pF. The resistor R_d is not included in the equations that describe the model because its effect is absorbed by the voltage controlled current source, i_d. In the simulations it is assumed that the electric field in the multiplication region is spatially uniform, which corresponds to a multiplication region without doping. To start the simulation a hole is injected at the edge of the multiplication region of width $w = 1600$ nm. The circuit is biased by the power supply at a voltage $V_a = V_b + V_{ex}$ so that the SPAD is reverse biased beyond its breakdown voltage, V_b by the excess voltage, V_{ex}. The theoretical breakdown voltage was calculated from the divergence of McIntyre's multiplication expression [10]

$$M = \frac{1-k}{\exp(-(1-k)\alpha w) - k}. \tag{4.58}$$

By using the electric-field dependent expressions for the electron and hole ionization coefficients for InP [67], the breakdown voltage is found to be $V_b = 64.61$ V.

Circuit behavior after an avalanche trigger

Figure 4.11 shows the calculated avalanche current, i_a, the feedback voltage, V_{R_L} and the voltage across the SPAD, V_{C_d} as a function of time, displayed in terms of both the carrier transit time, w/v, and the RC time constant of the circuit, $R_L(C_L + C_d)$. It is assumed that both holes and electrons travel at the velocity $v = 6.7 \times 10^6$ cm/s. In the simulations the value of the excess voltage is 0.39 V.

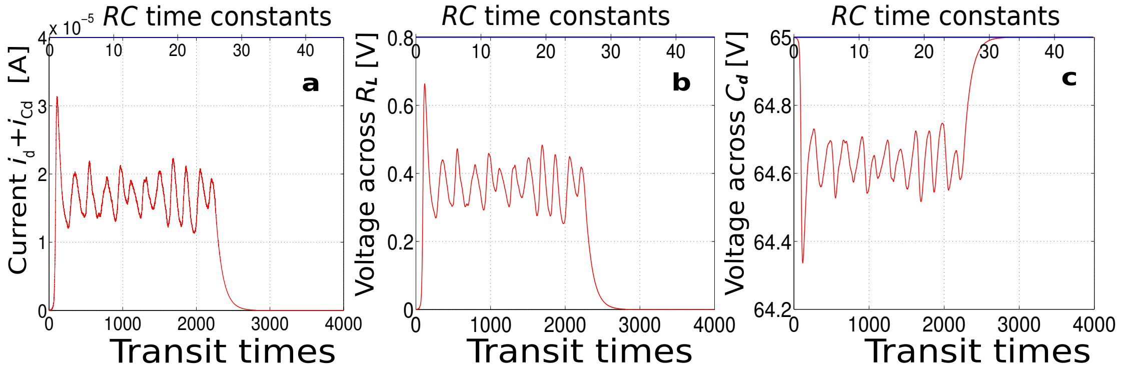


Figure 4.11: Calculated current-voltage evolution of the SPAD after an avalanche trigger. (a) Calculated avalanche current, $i_a = i_d + i_{C_d}$, (b) voltage across the feedback resistor, R_L, and (c) voltage across the SPAD, V_{C_d} as a function of time for an excess bias voltage $V_{ex} \approx 0.39$ V and a feedback resistor $R_L = 22$ kΩ. It can be seen that the oscillations are centered around their steady state values; thus, the avalanche current oscillates around $I_{ss} \approx 18$ μA, the feedback voltage oscillations are centered around $V_{R_L} = V_{ex} \approx 0.39$ V and the voltage across the SPAD fluctuates around the breakdown voltage $V_f = V_b \approx 64.61$ V. Note that quenching occurs at about 2340 transit times. In the simulations it is assumed that the electric field in the multiplication region is spatially uniform, which corresponds to a multiplication region without doping.

The current and voltages fluctuate around the steady state values predicted by the traditional model; the persistent current fluctuates around $I_{ss} \approx V_{ex}/R_L$ [42, 15], since $R_L \gg R_d$, the feedback voltage, V_{R_L}, fluctuates around the excess bias voltage V_{ex} and the voltage across the junction capacitor, V_{C_d}, fluctuates around the breakdown voltage, V_b.

Once an avalanche is triggered, then when the diode is biased above breakdown the mean avalanche current grows exponentially, after a brief transient of the order of the transit time, according to the theory of mean impulse response of APDs above breakdown [58, 69]. This growth discharges the capacitor C_d and therefore reduce the junction voltage V_{C_d}, which in turn causes the avalanche current to increase more slowly. Equivalently, from a feedback perspective the large avalanche current flowing through the junction increases the Ohmic drop across R_L, causing a drop in the junction voltage V_{C_d}. The avalanche current eventually falls until the junction bias falls below the breakdown voltage. This is a significant outcome of the stochastically self-regulating avalanche model and it is contrary to the traditional model, which dictates that the junction voltage never drops below V_b. The DC source then begins to recharge the capacitor with a time constant $\tau_r \approx R_L C_d$, causing the avalanche current to increase once again. The repetition of these discharge and recharge cycles yields the oscillatory behavior seen in Fig. 4.11, where the current through the diode oscillates about $I_{ss} \approx 18$ μA, the feedback voltage oscillates around the excess bias voltage $V_{ex} \approx 0.39$ V, and the voltage across the SPAD oscillates above and below the breakdown threshold, $V_b \approx 64.61$ V. This repetition continues until the stochastic fluctuations inherent in the impact ionization process cause the spontaneous quenching of the avalanche current. In the simulation shown in Fig. 4.11 quenching occurs after about 2340 transit times.

In Fig. 4.12 the voltage across the junction capacitor, V_{C_d} (red curve) is plotted together with the current, i_d (blue curve), calculated by the Monte Carlo simulator, to illustrate the timing relationship between the voltage and the current in the SPAD. In the figure the different stages of the voltage and current described above are identified, from the onset of the avalanche until the spontaneous quenching of the persistent current. Stage 1 indicates the onset of the avalanche, where the avalanche current starts growing. (For clarity the curve i_d was truncated and its first peak is not shown.) It can be seen from the figure that after the onset of the avalanche the junction capacitor starts to discharge, which reduces the voltage V_{C_d} and causes the avalanche current to drop, as shown in stage 2. This state of affairs continues and the

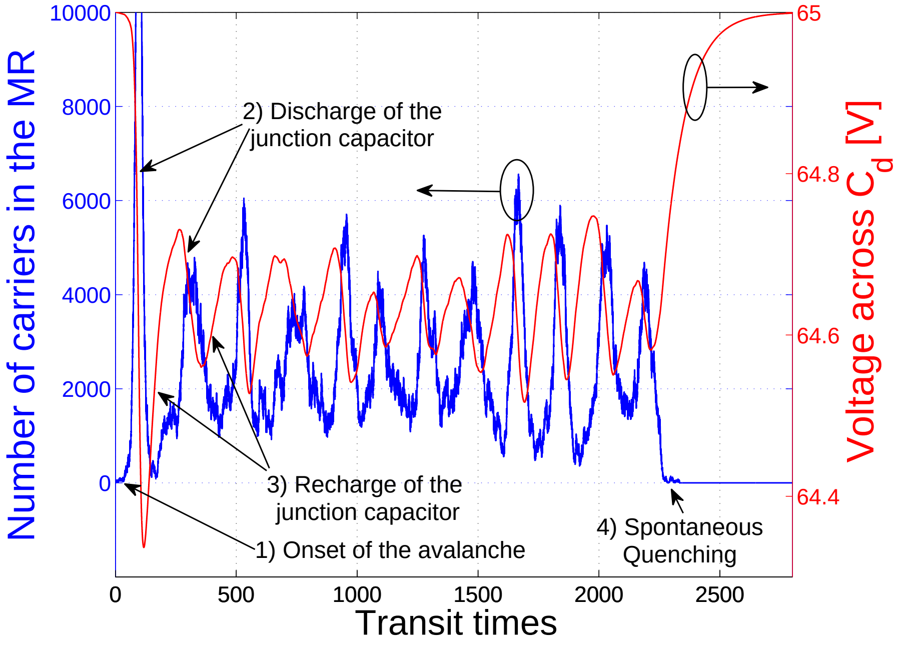


Figure 4.12: Timing relationship between the voltage across the junction capacitance and the number of carriers in the multiplication region. The red curve shows the voltage across the junction capacitor V_{C_d} and the blue curve shows the current i_d calculated by the Monte-Carlo simulator. For clarity, the current i_d was truncated and its first peak is not shown. The stages of the current-voltage evolution identified are: (1) onset of the avalanche, (2) discharge of the junction capacitor, (3) recharge of the junction capacitor and (4) spontaneous quenching.

voltage V_{C_d} drops below the breakdown voltage level. As described above, after this point the DC source begins to recharge the junction capacitor, which increases the voltage V_{C_d} and, in turn, causes the avalanche current to increase again, as shown in stage 3. The repetition of the discharge and recharge cycles of the junction capacitor yields the oscillatory behavior shown in Fig. 4.12. Eventually, the stochastic fluctuations inherent in the impact ionization process cause the spontaneous quenching of the avalanche current. A key point that was learned from the simulations is that spontaneous quenching invariably occurs during the recharge cycle of the junction capacitor, where the persistent current is at its lowest and the number of carriers is at a minimum. This is shown in stage 4. This observation is critical in understanding the exponential form of the pdf of the quenching time, which is discussed later on.

Quenching behavior

In the simulations an observation window of 8000 transit times was considered (~ 200 ns). The observation window is the interval of time during which the persistent current was observed when the quenching time was determined. The quenching time was measured within the observation window. A realization that shows a persistent current that does not spontaneously quench within the observation window is considered to be self sustaining. It was found that within the considered observation window the probability of spontaneous quenching increases as the current I_{ss} decreases. This is because when the current is reduced so is the number of ionizing carriers, increasing the chance that all carriers present in the multiplication region exit without impact ionizing.

Figure 4.13 shows representative examples of the persistent current regime without quenching (red curve), the case where spontaneous quenching occurs after a period of persistent current flow (blue curve) and the case when quenching occurs immediately following the first current peak, shown in the black curve. In this example the excess bias voltage is about 0.39 V and R_L was varied to achieve the different values of I_{ss}. It should be mentioned that the three quenching behaviors described above have been observed on NFAD devices by Princeton Lightwave, Inc., with appropriate variations in the feedback resistor R_L. Moreover, a similar fast self-collapse of the avalanche current was reported by Shushakov *et al.* [71, 72] and Zhao *et al.* [46] in devices where the feedback was provided by means of a chargeaccumulation effect due to a potential barrier outside the multiplication region. The work of Shushakov *et al.* [71, 72] also included a Monte-Carlo simulation of the stochastic avalanche process in the presence of feedback, which was used to calculate the distribution of the gain.

The quenching behavior described above was reproduced consistently for excess voltages below 0.8 V. However, for higher excess bias voltages the behavior shown in the blue curve of Fig. 4.13, in which there is a period of persistent current flow followed by spontaneous quenching, was not observed. Instead, in the higher excess voltage case the system goes from the regime of persistent current flow that does not quench to that of quenching immediately following the first peak of the current for higher R_L. This may be because with higher excess voltages the feedback is stronger and the current i_d decreases too fast, making it difficult for the system to execute even a single period of persistent oscillating current.

The stochastic nature of the avalanche current can be appreciated from Fig. 4.14, which shows one realization of the calculated current i_d, represented by the number of carriers in the multiplication region. After this oscillatory persistent current, which extends for about 2340 transit times, spontaneous quenching occurs. Figure 4.14 also shows a zoomed view of three specific times in the simulation. Figure 4.14b and 4.14c show two instants where the random fluctuations of i_d reach their minimum values. In both cases the number of carriers reaches a very small value, although spontaneous quenching does not occur. Figure 4.14d shows the instant when spontaneous quenching occurs. The smallest fluctuations of the number of carriers in the multiplication region is around one carrier.

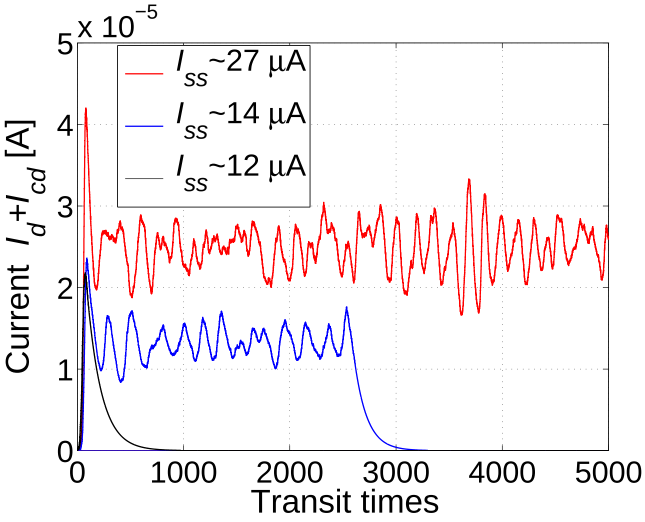


Figure 4.13: Quenching behavior of the simulated passively quenched SPAD for different values of the current I_{ss}. As the current I_{ss} decreases the avalanche current spontaneously quenches sooner, on average.

Probability density function of the quenching time

The pdf of the quenching time, T_q, for an excess bias voltage of 0.39 V was estimated by repeating the simulation of the persistent current (from trigger instant to quenching instant) 2267 times. The quenching time, T_q, is the interval of time measured from the start of the avalanche until its spontaneous quenching. The result is shown in Fig. 4.15, where the decay of the tail of the pdf is exponential, implying that the quenching time is memoryless. This observation is consistent with the measurements shown in Fig. 4.4.

The memoryless property of the quenching process can be understood from the fluctuating behavior of the voltage across the SPAD and the persistent current. Recall that these quantities oscillate about V_b and I_{ss}, respectively. It can be noticed from the simulations that quenching invariably occurs only during the recharge cycle

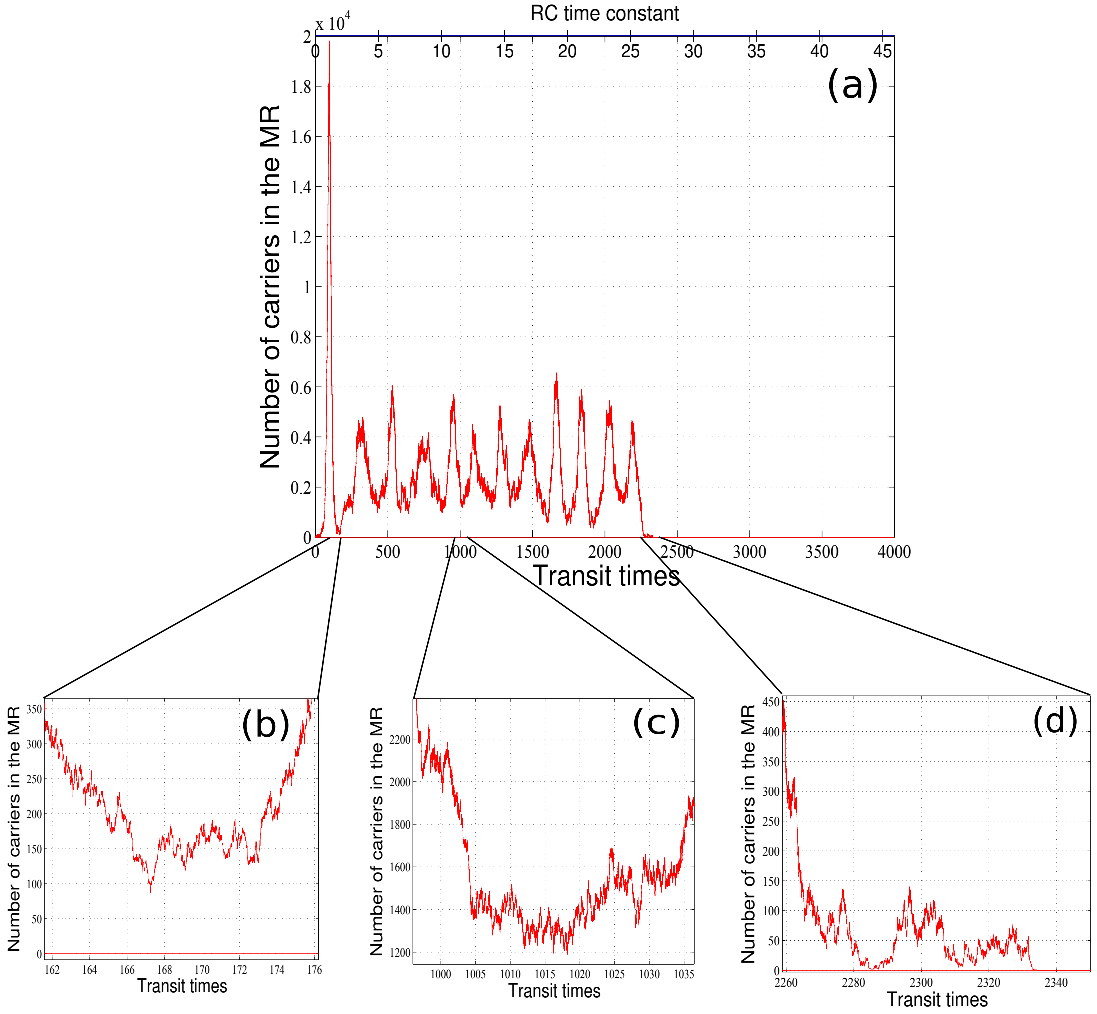


Figure 4.14: Calculated current i_d. (a) Number of carriers in the multiplication region as a function of time. The curve shows the oscillatory behavior of the persistent current before spontaneous quenching occurs. (b) and (c) show two instants where the random fluctuations of i_d reach their minimum value. In both cases only a few carriers remain but spontaneous quenching does not occur. (d) shows the instant of spontaneous quenching.

of the junction capacitance, as shown in Fig. 4.12, and that the probability that quenching occurs during the discharge stage of the capacitance is negligible. Two key observations are made. The first is that the probability that quenching occurs during the recharge cycle of the junction capacitance is the same for all recharge cycles. (The beginning of each recharge cycle of the junction capacitance starts right after the voltage across C_d reaches a local minimum.) This is due to two factors: (a) the electric field profile is almost identical in all recharge cycles. The point here is that the electric field remains above breakdown in half cycle and then remains below breakdown in the second half. (b) The number of carriers at the beginning of each recharge cycle is almost the same for all recharge cycles, owing to the periodicity of the persistent current. The number of carriers at the beginning of each recharge cycle is almost the same in a statistical sense, meaning that the probability distribution of this number is approximately the same from cycle to cycle but the actual numbers can be different. Hence, prior to quenching, both the electric field and carrier number conditions are almost reproduced periodically at the beginning of each recharge cycle. This, in turn, implies that the probability of quenching is approximately the same for all recharge cycles. It is emphasized that the probability of quenching is approximately the same on average (in a statistical sense), although the actual values may vary from cycle to cycle and from experiment to experiment. The second observation is that quenching events over different recharge cycles are statistically independent. We can assume that the quenching events are statistically independent because different recharge cycles involve different carriers, since the duration of the cycle is much greater than the carrier transit time across the multiplication region. Thus, if P represents the probability that quenching occurs in a specific recharge cycle, given that quenching has not occurred earlier, then by using the two observations made above we can write the probability that quenching occurs at the nth recharge cycle as $P(1-P)^{n-1}$. This is exponential in form and thus satisfies the memoryless property.

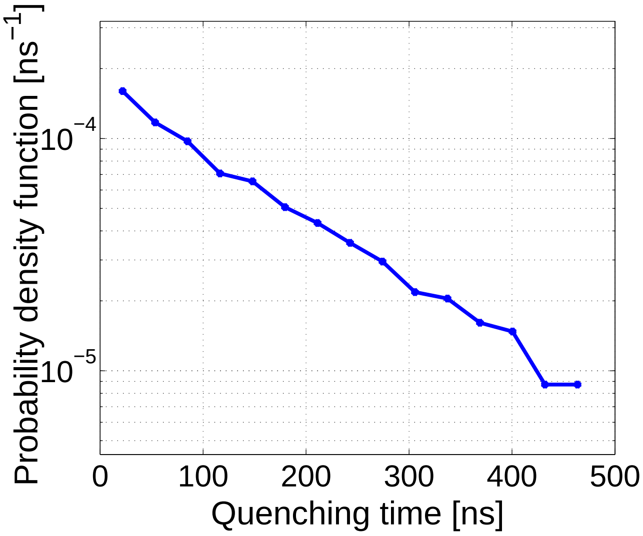


Figure 4.15: Calculated probability density function of quenching time, T_q.

4.4.4 Conclusions

A stochastic model that calculates the response of a passively quenched SPAD, reverse biased above breakdown, was presented in this section. The model considers a closed loop system, capturing the effect of the feedback introduced by the resistive load on the stochastic nature of the avalanche multiplication. This approach differs from the conventional traditional model [42, 15], which captures the deterministic feedback, maintaining the device at breakdown, but neglects the dynamic coupling between the voltage across the SPAD, the feedback from the load and the impact ionization process. As a consequence the traditional model provides no way of determining the oscillatory behavior of the persistent avalanche current and the statistics of the quenching time. Moreover, it has been shown that the traditional model leads to unrealistic predictions of the pdf of the quenching time. By contrast the stochastically self-regulating avalanche model enables us to predict the stochastic current-voltage evolution and quenching characteristics in passively quenched SPAD circuits. The model predicts key attributes of the stochastic avalanche current seen in experiments performed on the new generation of SPAD structures that rely on negative feedback. The proposed model therefore constitutes a reliable simulation framework to aid the design and optimal operation of an emerging generation of SPAD devices that rely on negative feedback.

Part II

Mid wave infrared avalanche photodiodes

Chapter 5

QDAP

5.1 Introduction

Present day GaAs-based APDs are limited to wavelengths below 2 μm. In this chapter, the first demonstration of a GaAs based APD operating in the mid wave infrared region (MWIR, 3–5 μm) is reported. The device, called Quantum Dot Avalanche Photodiode (QDAP) [33], exploits quantum confined transitions to obtain intersubband absorption in the quantum dots-in-a-well (DWELL) heterostructure for MWIR detection and couples the photogenerated carries into an APD to obtain a large conversion efficiency (CE) via avalanche multiplication. The conversion efficiency η_{conv} is the product of the internal quantum efficiency η and the photoconductive gain g and is given by $\eta_{conv} = \eta g = \Re(h\nu)/q$, where $\Re$ is the responsivity, q is the charge of the electron and $h\nu$ is the photoexitation energy. Using this approach, the photocurrent has been increased by a factor of 14 and reached a CE of 12%, which is one of the highest reported CE for any quantum dot detector.

Various approaches have been explored to realize single photon detectors. They include superconducting detectors, photomultiplier tubes (PMTs) and APDs. While

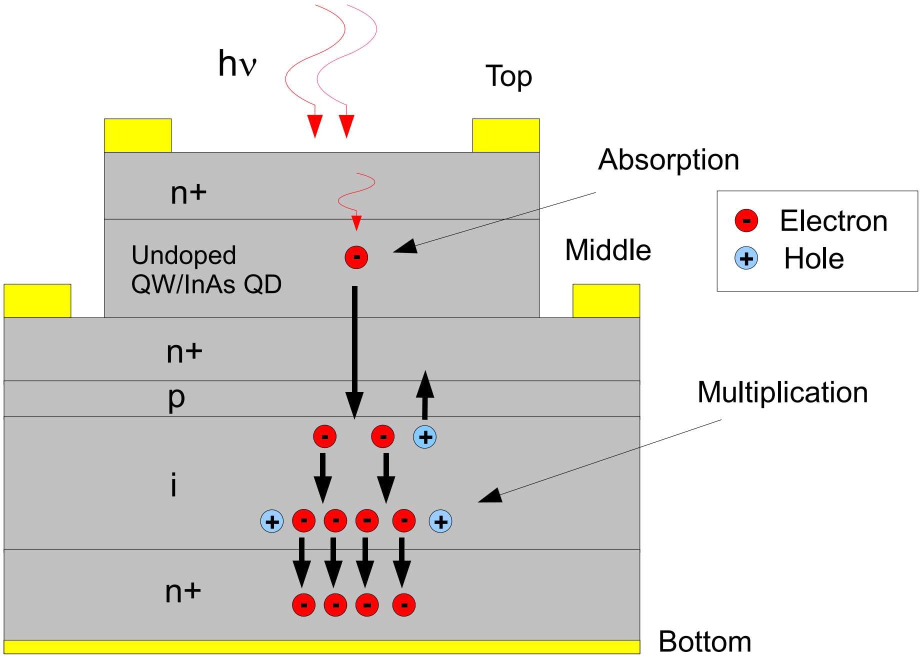


Figure 5.1: heterostructure schematic of the Quantum Dot Avalanche Photodiode (QDAP) showing the separate absorption and multiplication regions

superconducting detectors offer the highest quantum efficiency and the lowest dark count rate, they are limited to very low operating temperature (<10 K). PMTs have demonstrated very good performance in the visible region but are very bulky and are not compatible with the standard semiconductor manufacturing processes. PMTs also require very high voltages and are sensitive to magnetic fields. APDs represent the most promising scalable technology as they are based on semiconductors such as Si, GaAs and InP and can be made into large format arrays [11, 8, 29, 50]. APDs can be operated in two modes. In the linear mode, the APD is biased below breakdown and the output photocurrent is “amplified” through the avalanche of impact ionizations. In the Geiger or single-photon avalanche detector mode, the APD is biased just below to the breakdown voltage and a gated pulse is used to

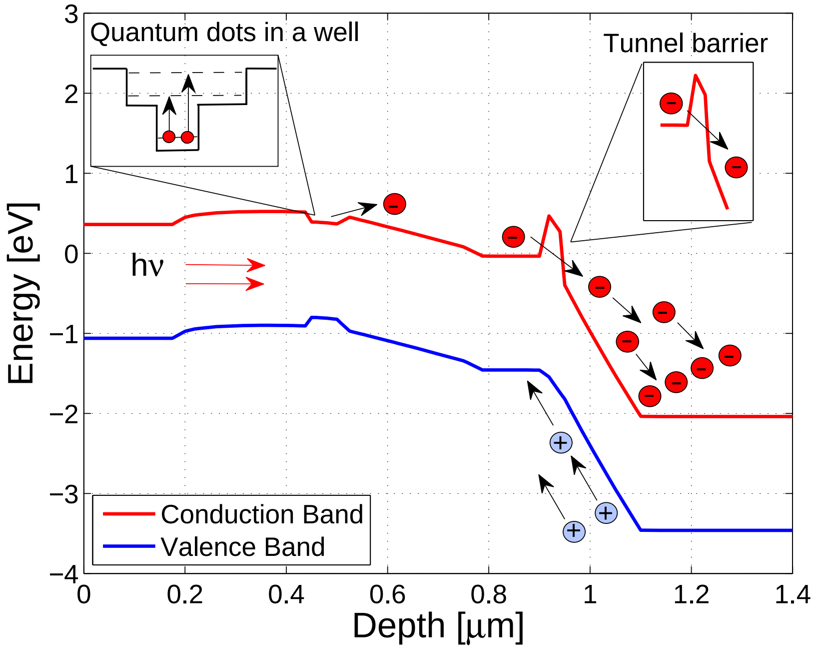


Figure 5.2: Calculated band diagram of the QDAP using Sentaurus simulation tool with the absorption and multiplication sections reverse biased at a voltage of 0.5 V and 2.0 V, respectively.

drive it above breakdown for a short duration. An incoming photon during this gated pulse triggers a cascade, or avalanche, of impact-ionization events leading to a large current. Various active and passive quenching circuits have been developed to limit the current flowing in the APD and to reset it for the next pulse [15]. However, all GaAs based SPADs available today are limited to wavelengths less than 2 μm [73]. Hamamatsu has an extended InGaAs APD which operates till 1.7 μm [73].

In the QDAP, an intersubband quantum dots-in-a-well detector is coupled with an APD through a tunnel barrier, see Fig. 5.1. The photon absorption and generation of carriers take place in the DWELL section while the avalanche section provides

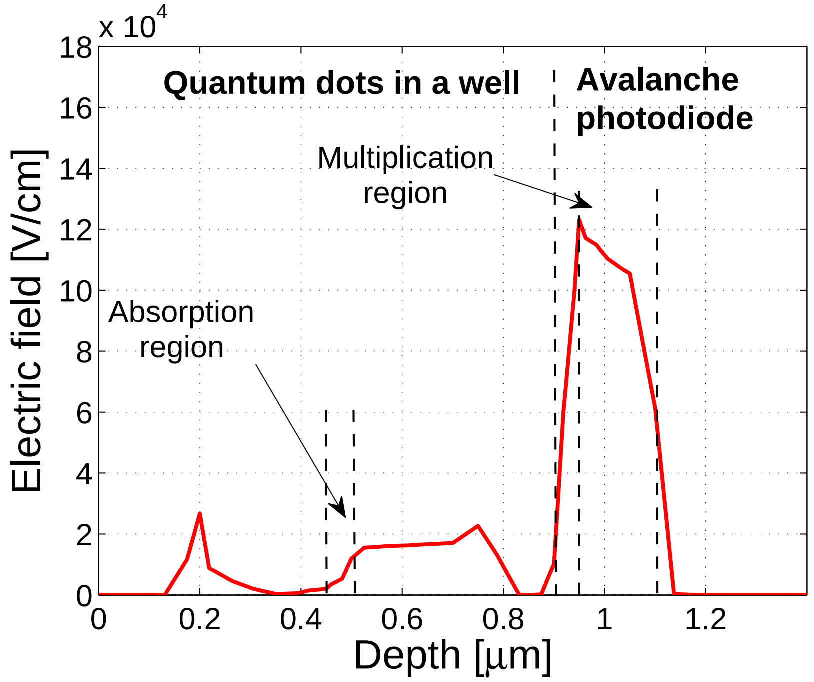
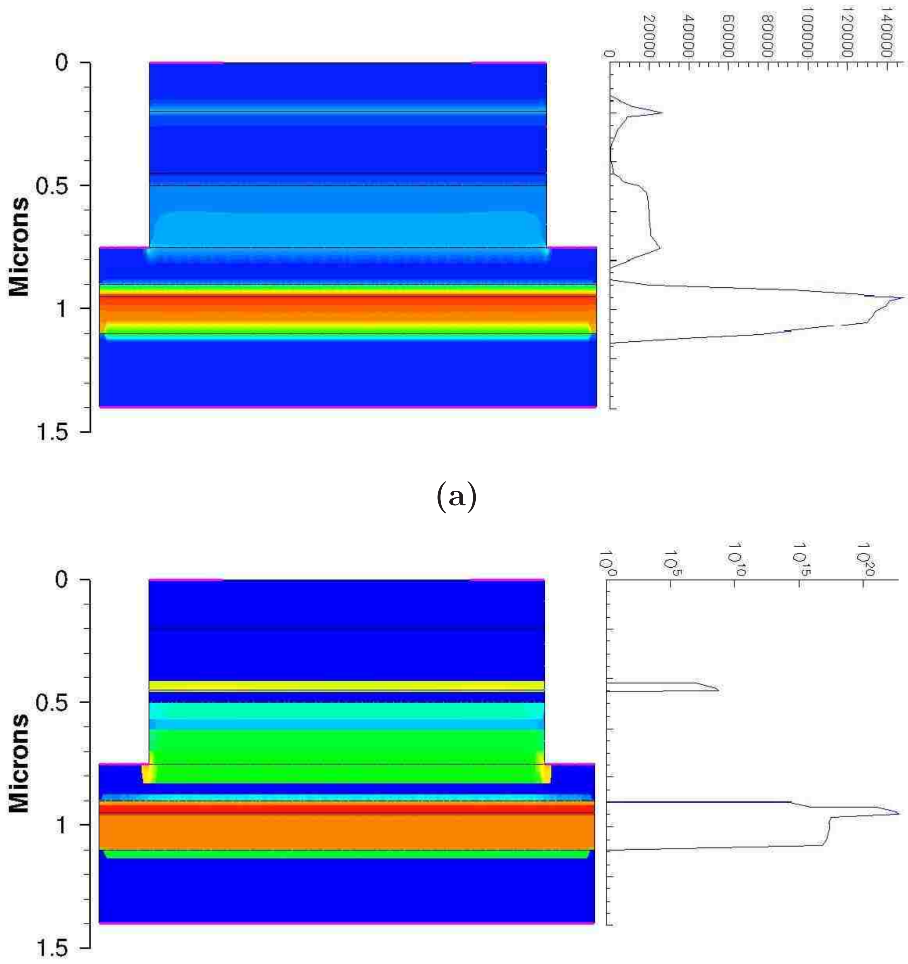


Figure 5.3: Theoretical simulation of the electric field profile in the device confirming that the maximum electric field is established across the multiplication region.

internal gain. The demonstration of the linear-mode operation of the QDAP signifies a major achievement as it opens up the possibility of obtaining SPADs in the midwave infrared range. This would have a dramatic impact on many applications that require high sensitivity MWIR detectors including astronomy and biomedical diagnostics.

5.2 Operating principle

The operating principle of the QDAP is shown in Fig. 5.1. Operationally, the QDAP can be divided into two stages: (a) the photogeneration of carriers, and (b) the avalanche multiplication of the photogenerated carriers. The absorption of photons



(a)

(b)

Figure 5.4: Theoretical modelling using Sentaurus. (a) Electric field profile (V/cm) (b) impact ionization rate ($cm^{-3}s^{-1}$).

and photogeneration of electrons is due to intersubband transitions in the quantum dots (QDs). Among the various QD heterostructures, the DWELL detector is among the most promising alternatives for terrestrial applications [74]. The DWELL heterostructure is a hybrid between a conventional quantum well infrared photodetector (QWIP) and a quantum dot infrared photodetector (QDIP) and draws from the advantages of both of these technologies. DWELL detectors operate under normal incidence conditions with low dark current like the QDIP and have good operating wavelength control like the QWIP. In addition, DWELL detectors benefit from a mature growth and processing technology of III–V semiconductors, making it possible to produce devices with good spatial uniformity over a large area. This characteristic is essential for fabricating large format FPAs. Recently, Ting *et al.* have demonstrated a 1 megapixel DWELL FPA with peak response at 8.5 μm [75].

Thus the spectral response of the QDAP is determined by the DWELL section. On the other hand, the avalanche multiplication, resulting from the injection of the photogenerated electrons into the avalanche region, takes place in the APD section. To control the individual responses of the DWELL and the APD sections, the voltage applied between the top and middle contacts and the voltage between the middle and bottom contacts are independently varied. Figure 5.1 shows the heterostructure of the QDAP and illustrates the absorption and multiplication stages. When the incident radiation of energy $h\nu$, where h is the Planck's constant and ν is the frequency of the incident radiation, is shone on the device, the incident photons are absorbed in the active region of the DWELL section. As a result, the electrons in the ground state of the quantum dots are promoted to a set of bound states within the quantum well. Once these electrons are extracted from the quantum well they drift, as a result of the applied electric field, toward the APD section of the device. This is illustrated in Fig. 5.2. To reach the avalanche region located in the APD section the electrons need to tunnel through the barrier established by the p-type layer, as shown in Fig. 5.2. In the high-field avalanche region the injected electrons experience a series of impact ionizations that multiply them. Finally, the multiplied electrons are collected in the bottom contact. For these processes to take place, the external biasing circuit used to operate the QDAP has to be designed in such a way that the flow of electrons is from the top to the bottom contact. This mode of operation is achieved when both sections of the QDAP are reverse biased, i.e.,

$V_{Top} < V_{Mid} < V_{Bot}$. A customized biasing circuit was built to enable this configuration. Figure 5.3 shows the theoretically calculated electric field in the QDAP indicating that most of the applied field drops across the APD.

5.3 Device modeling

To help the design of the QDAP structure, in particular the APD section, and to predict the device behavior theoretical modelling, using the software Sentaurus, was carried out. Figure 5.4 shows the calculated 1D and 2D profiles of the electric field and the impact ionization across the QDAP. In the simulation the voltage across the DWELL section was 0.5 V and the voltage across the APD was 2.0 V. It can be noticed that the electric field and the impact ionization profiles have their maximum values in the avalanche multiplication region of the APD section. This result supports the designed APD section as the multiplication stage in the device.

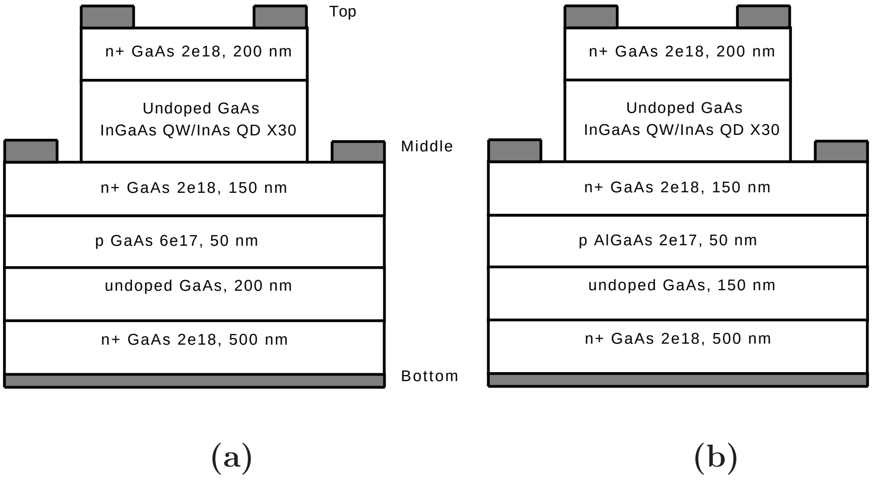


Figure 5.5: Two different QDAP designs (a) QDAP 1 (b) QDAP 2

5.4 Device structure

The n-i-n DWELL detector was grown on top of the p-i-n APD section. The DWELL consists of *n*-doped InAs dots in an $In_{0.52}Al_{0.48}As$ /GaAs well and $Al_{0.10}Ga_{0.90}As$ as the barrier. These layers are sandwiched between two highly doped *n*-GaAs layers. The APD is a standard PIN diode with a GaAs multiplication region of 0.15 μm. The multiplication region was made thin to achieve a small avalanche breakdown voltage and to enhance the dead-space effect [54]. It is known that in thin multiplication regions the dead-space reduces the multiplication noise, also called as excess noise, introduced by the avalanche multiplication process in APDs [76, 54]. The multiplication noise is a result of the stochastic nature of the impact ionization process.

Figure 5.5 shows the two different QDAP structures fabricated. The first design, referred to as QDAP 1, shown in Fig. 5.5a, includes an APD with a GaAs *p*-type layer of 50 nm, a GaAs homojunction multiplication region of 200 nm, and a GaAs *n*-type layer of 500 nm. In the second design, referred to as QDAP 2, the *p*-layer of the avalanche photodiode consists of an AlGaAs layer of 50 nm, a multiplication layer of GaAs material of 150 nm, and a *n*-type layer of 500 nm of GaAs, see Fig. 5.5b. The idea behind the design of QDAP 2 is to reduce the dark current by introducing an AlGaAs potential barrier.

5.5 Growth and fabrication

In the QDAP structure the n-i-n DWELL detector was grown on top of the p-i-n APD section in a single-step epitaxy. The DWELL section of the QDAP consists of the structure reported by Shenoi *et al.* [77]. This DWELL structure was designed to maximize the obsoption of photons by increasing the number of active region stacks to thirty, compared to previous designs [78]. This larger number of active region stacks is achieved by minimizing the strain. In the design, the DWELL consists of n-doped InAs dots in an $\mathrm{In_{0.52}Al_{0.48}As/GaAs}$ well and $\mathrm{Al_{0.10}Ga_{0.90}As}$ as the barrier. These layers are sandwiched between two highly doped n-GaAs layers. [1]

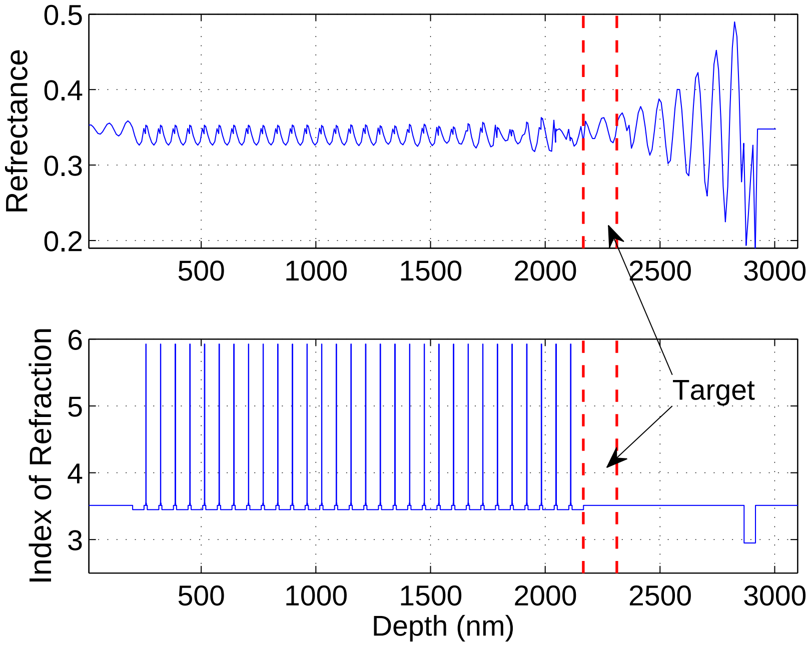


Figure 5.6: Simulated reflectance of the sample as a function of the depth (upper plot) and index of refraction profile (lower plot).

Post-growth processing was done in a class 100 clean-room environment by performing three rounds of: standard contact photolithography, plasma etching, and metal deposition using an e-beam evaporator. Finally, the processing was concluded by annealing the contacts at 400 °C using rapid thermal annealing. Due to the thin middle contact layer (middle n-type layer), the critical part of the device processing was to perform the first mesa etch. To overcome this difficulty the real-time reflectance of

---

[1] A detailed description of the DWELL structure design can be found elsewhere [77].

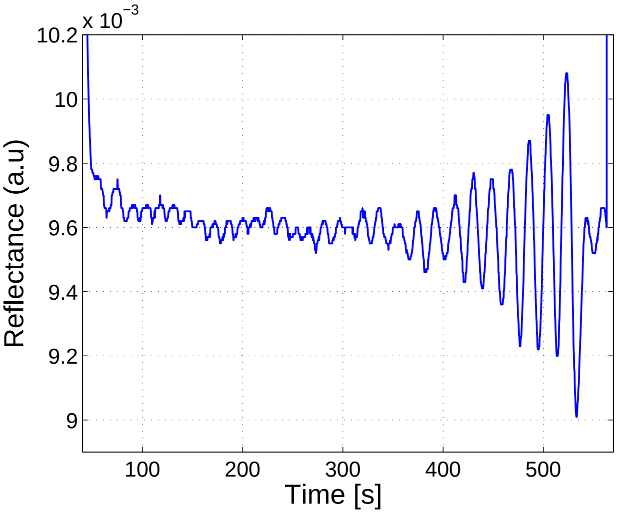


Figure 5.7: Real-time measured reflectance as a function of etching time.

the sample was measured while the etching process was carried out. Figure 5.6 shows a simulation of the theoretical reflectance profile (upper plot) as a function of the depth of the sample. The index of refraction profile of the sample and the desired target is also shown (lower plot). In Fig. 5.7 the measured reflectance (arbitrary units) as a function of time is plotted. While the etching process is carried out, the real-time measurement of the reflectance is compared with the theoretical reflectance profile. When the profiles match, the etching process is stopped when the predicted target is reached. The dark current of the APD section of the processed devices was measured for several temperatures, see figure 5.8. It can be seen from the figure that at 77 K the breakdown voltage is around -1.8 V.

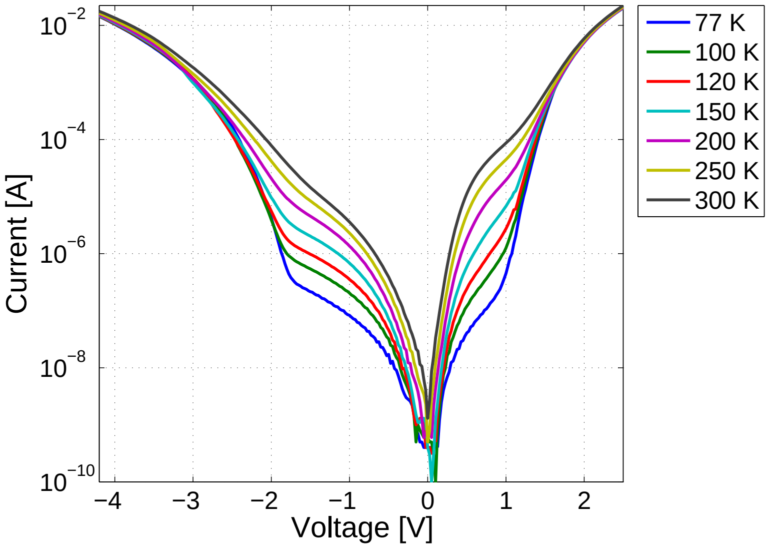


Figure 5.8: Dark current of the APD section of the QDAP as a function of the applied voltage, measured at several temperatures. The breakdown voltage at 77 K is around -1.8 V.

5.6 Results

To demonstrate the operation of the QDAP a series of radiometric measurements were carried out. First the spectral response of the QDAP was measured using a Thermo Nicolet Fourier transform infrared (FTIR) spectrometer. Figure 5.9 shows the measured conversion efficiency of the DWELL section as a function of wavelength for a reverse bias of 2 V across the DWELL. It can be seen that the conversion efficiency peaks at 5 μm. The next step was to demonstrate the ability to separately control the absorption and multiplication response of the QDAP. To this end, the photocurrent of the QDAP detectors was measured with the devices cooled down to liquid nitrogen temperature (77 K). To create carriers in the DWELL section the

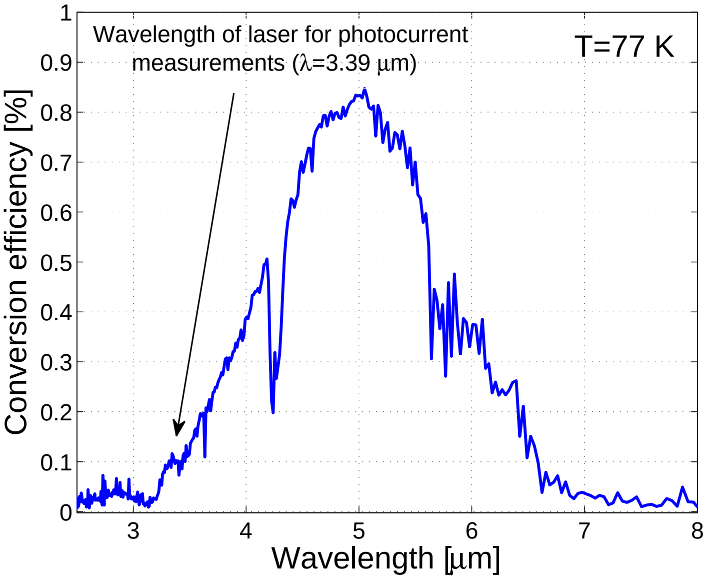


Figure 5.9: Conversion efficiency of the DWELL absorber as a function of wavelength for an applied bias of 2 V.

devices were irradiated using 3.39 μm laser beam chopped at a frequency of 400 Hz. A specially designed circuit board was used to bias the devices and amplify the photocurrent. Figure 5.10 shows the designed biasing scheme used to measure the photocurrent and the multiplication gain of the QDAP. (It can be seen from the figure that the QDAP is modelled as a series combination of a resistor, which represents the DWELL section, and an APD.) Then, the amplified signal was fed into an SR770 Fast Fourier Transform (FFT) Network Analyzer. In the first experiment, the photocurrent in the QDAP was plotted as a function of the bias across the DWELL section (shown in Fig. 5.11). The current increases linearly with the applied bias as expected. The inset of Fig. 5.11 shows the structure of the device under test. In the second experiment, the photoelectron generation in the DWELL section was fixed by keeping the DWELL section biased at a fixed value (0.5 V). The multiplication gain

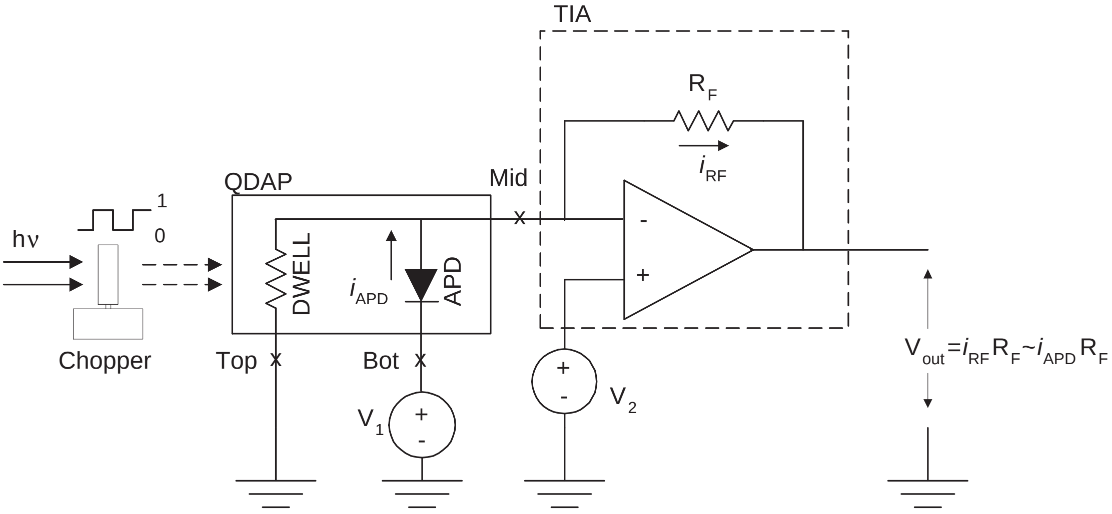


Figure 5.10: Biasing scheme to measure the photocurrent of the QDAP.

of the APD was then varied by changing the voltage across the APD. Figure 5.12 shows the photocurrent and the total multiplied noise of the heterojunction QDAP structure as a function of the reverse bias across the APD for a fixed reverse bias of 0.5 V across the DWELL section. It can be seen that for a fixed voltage across the DWELL section, as the reverse bias across the APD increases, the photocurrent of the device increases in a fashion dictated by the multiplication gain of the APD. In addition, compared to the photocurrent characteristics of the DWELL section alone the QDAP photocurrent is distinctly different. Our measurements demonstrate that the response of the DWELL and APD sections of the device can be controlled independently and the overall response of the QDAP can be enhanced. Figure 5.13a shows the calculated excess noise factor of the device (dots) and the excess noise factor predicted by the dead-space multiplication theory [65] (DSMT, solid curve) as a function of the multiplication gain. The excess noise factor was estimated by fitting the measured data to the excess noise factor of a GaAs multiplication region

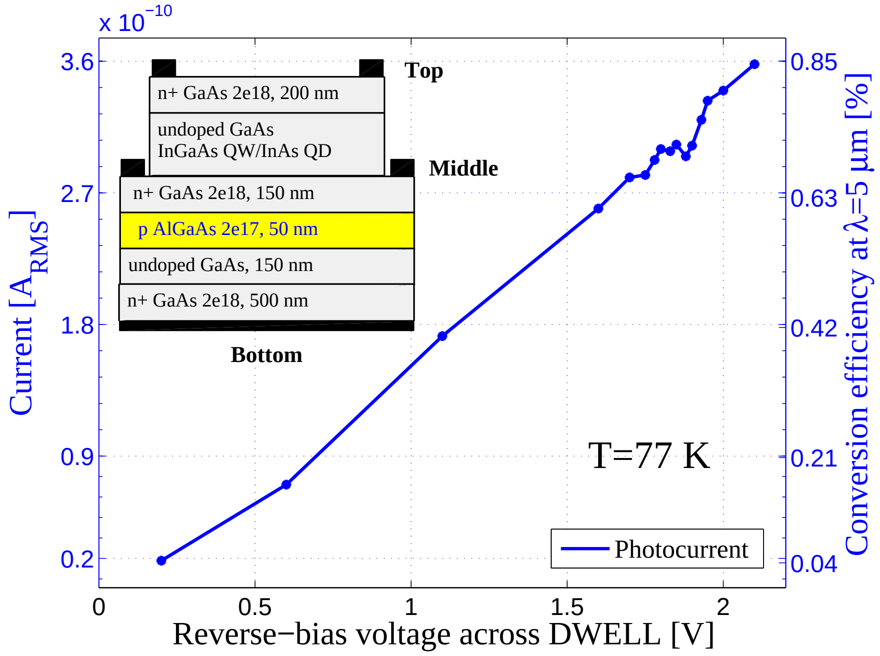


Figure 5.11: Photocurrent and conversion efficiency of the DWELL section at $\lambda = 5\ \mu$m as a function of the reverse bias.

of 150 nm predicted by the DSMT. It has been previously shown [51] that the DSMT accurately predicts the excess noise factor of thin III-V avalanche photodiodes. It can be seen that, with the exception of the excess noise factor for a multiplication gain of $\sim$ 8.2, the calculated excess noise factor closely follows the trend predicted by the DSMT.

We have found that the detectivity of the QDAP decreases as the reverse voltage across the APD increases. The detectivity, D^*, is calculated as [79]

$$D^*(\lambda) = \frac{\sqrt{A_d \Delta f}}{i_n}\Re \tag{5.1}$$

where A_d is the area of the detector, Δf is the measurement bandwidth, $\Re$ is the responsivity, and i_n is the current noise. Figure 5.13b shows the detectivity of the

QDAP at $\lambda = 3.39$ μm, as a function of the reverse voltage across the APD section for a fixed value of the DWELL section of 0.5 V. It can be seen from the figure that as the reverse voltage across the APD section increases the detectivity curve decreases. This behavior is an inevitable result of the operation of the QDAP, in which the DWELL and the APD sections contribute to the noise of the device. In the QDAP, both the photogenerated and the dark carriers injected from the DWELL section into the APD are amplified by the same factor, which is the multiplication gain of the APD. In addition, the dark carriers created in the APD are also amplified by the multiplication gain. As a result as the gain of the APD is increased the noise of the device increases faster than the signal.

To compare the performance of the QDAP with that of the DWELL, their conversion efficiencies η_{conv} were measured. The conversion efficiency in the DWELL was increased by a factor of 14 due to the gain introduced by the avalanche multiplication stage of the QDAP. From Fig. 5.9 we see that at $\lambda = 5$ μm the peak conversion efficiency of the DWELL alone is measured to be 0.84% at a reverse bias of 2.0 V. Therefore, by operating the QDAP with the DWELL section reverse biased at 2.0 V it is possible to obtain a maximum conversion efficiency of about 12%.

5.7 Conclusion

A new mid infrared detector, called the QDAP, was presented. The QDAP structures under test use an optimized DWELL structure as the absorption layer and a p-i-n avalanche photodiode as the multiplication stage. The photocurrent shows an increase as the reverse voltage across the APD section was increased. In fact, the photocurrent of the QDAP increased in a nonlinear fashion as the reverse voltage across the APD was increased. Compared to the photocurrent characteristic of the DWELL section alone, the QDAP photocurrent is very distinct and its characteristic

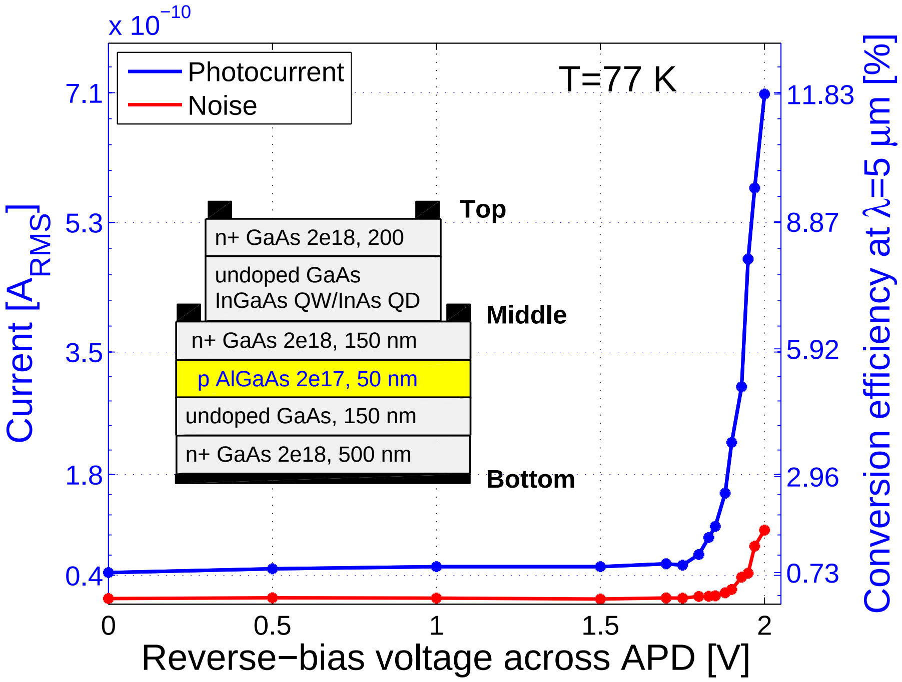
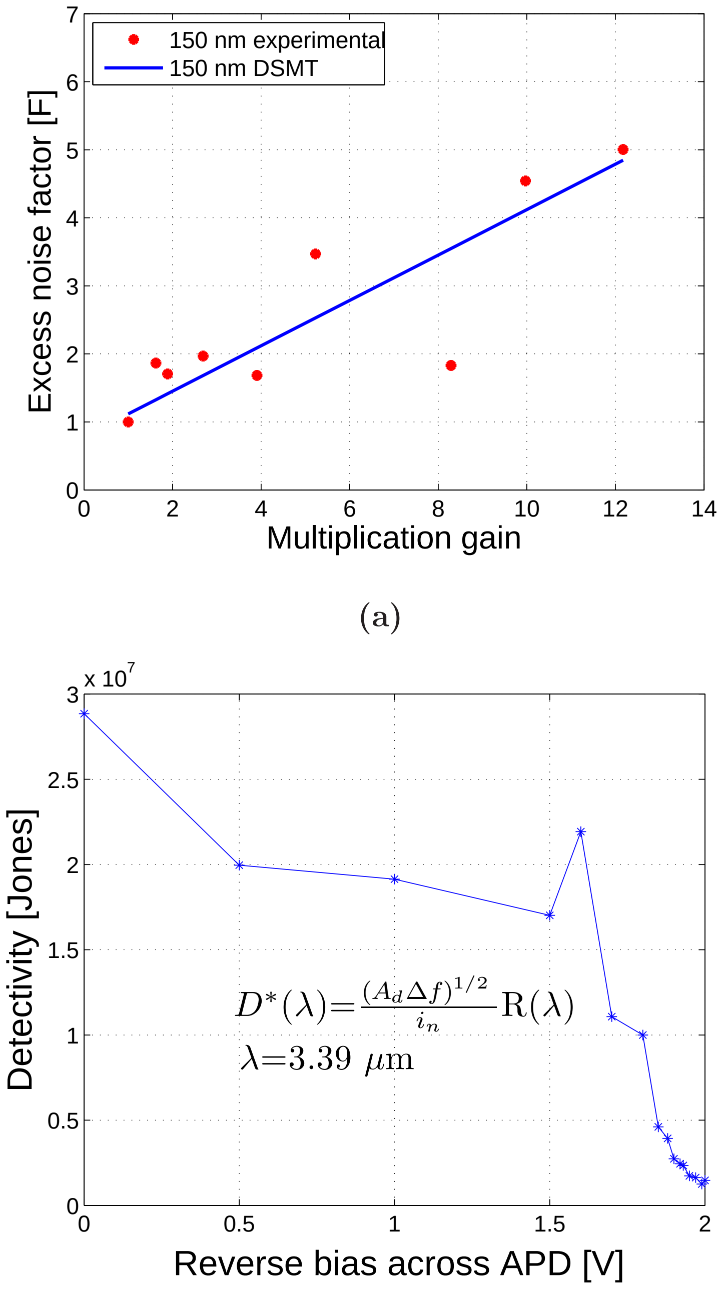


Figure 5.12: Photocurrent and conversion efficiency of the QDAP as functions of the reverse-bias voltage across the APD section for a fixed applied bias of 0.5 V across the DWELL section. The inset illustrates the structure of the device. Note that if the DWELL absorber was operated at 2 V (as in Fig. 5.9), the peak CE at $\lambda = 5$ μm would be about 12%, which is one of the highest reported CE for any QD based mid-infrared detector.

is very similar to that of the I–V curve of the APD section. These results suggest that the photocarriers generated in the DWELL section are amplified by the APD section. It is expected that by operating the QDAP with the DWELL section reverse biased at 2.0 V it is possible to obtain a maximum conversion efficiency of about 12%.



(a)

(b)

Figure 5.13: (a) Measured (dots) and predicted (blue curve) excess noise factor of the QDAP. (b) Detectivity of the QDAP, measured at 3.39 μm, as a function of the reverse voltage across the APD section for a fixed value of the DWELL section of 0.5 V.

Chapter 6

Summary and future work

In this dissertation, theoretical methods have been presented that allow us to characterize the performance of APDs working in linear and Geiger modes. In addition, it has been demonstrated the linear mode operation of a novel mid infrared photodetector that incorporates avalanche multiplication gain.

The first modeling method presented is a model that sheds light on the dependence of the performance of SAM SPADs (APDs operating in Geiger mode) on the width of the multiplication region by comparing the effects of field-assisted tunneling with temperature-assisted dark carriers as the width is varied. The characterization of the devices is made by calculating the performance metrics: *photon detection efficiency* (PDE), the *dark count rate* (DCR) and also the *single-photon quantum efficiency* as a function of the width of the multiplication region, the applied voltage and the temperature of operation. The model captures the effect of the dead space and heterojunction multiplication regions. An aspect of importance that had not been explored before, namely, the random locations where carriers are born in each layer, is thoroughly analyzed and incorporated in the model. In particular, it is assumed that photogenerated carriers are generated in the absorber at random locations according to an exponential probability density function (pdf). On the other hand, dark carriers are assumed to be generated randomly in the multiplication region and the absorber according to a uniform pdf in each layer. This theoretical study shows that the thickness of the multiplication region plays a different role in the performance of a SPAD depending upon what mechanism of dark-carrier generation is dominant. At low temperatures, for which field-assisted mechanisms are dominant, an increment in the thickness of the multiplication region will result in an improved PDE vs. DCR characteristics. The same behavior is seen in the SPQE curve at low temperatures. At room temperatures, on the other hand, the PDE vs. DCR characteristics show a weaker performance as the width of the multiplication region is increased. However, the SPQE curves show a maximum achievable peak SPQE at an optimal overbias and an optimal multiplication-region width.

The second modeling method explores the characteristics of the impact ionization process under the influence of a time-varying electric field. A theory is presented that models the impact ionization process in APDs under dynamic biasing. The model allows us to predict the breakdown probabilities, the gain, the mean impulse response, the excess noise factor, and the gain-bandwidth product of SAM APDs under an arbitrary time-varying electric field. The model predicts that by using a sinusoidal biasing scheme we should be able to increase the pulse gain-bandwidth product of a SAM APD by a factor of 5 compared to the same APD operated under the conventional static biasing scheme. This result is important because the telecommunication industry has been moving toward 40-Gbps and 100-Gbps protocols for their core fiber-optic backbone networks alongside the existing 10-Gbps infrastructure operating at the low-loss wavelength of 1.55 μm. However, the limited speed of APD-based receivers have limited their use in systems that operate at 2.5 and 10 Gbps. Consequently, the proposed biasing scheme represents a promising effort to enable the current InP-based APDs to meet the expectations of 40 Gbps systems.

Future work is needed on the study of the impact ionization process under dynamic biasing. The future work includes the theoretical optimization of the biasing scheme, the determination of the fundamental trade-offs, and the experimental feasibility of the scheme. In particular, the optimal time-varying-voltage profile that provides the highest speed at the maximum achievable mean gain needs to be determined. This task is not simple since there are several aspects to take into account. For example, in the proposed scheme the gain produced by photons arriving early in the bit is higher than the mean gain produced by photons arriving close to the end of the bit. As a result, the mean gain produced by the APD in one bit depends on the gain profile across the duration of the bit. Assuming that the frequency of the sinusoidal bias voltage is fixed at the desired bit rate, the gain profile is a function of the phase of the sinusoidal bias voltage, its amplitude, and its DC level. Some of the constrains are: (a) the minimum applied bias has to be such that its associated gain has to be at least equal to 1, so that the APD is still able to detect the incoming photons, and (b) the maximum voltage has to be such that the gain does not diverge. In addition, the optimal bias voltage profile depends on the particular I–V characteristics of the APD considered. Another aspect to investigate has to do with the noise performance of the APD under the dynamic biasing scheme. As seen in Chapter 1, the excess noise factor strongly affects the performance of an APD receiver. Therefore, it is necessary to investigate the effect of the dynamic biasing on the excess noise characteristics of the APD. In particular, since the excess noise factor increases with the mean gain, it is important to know the penalty introduced by the high multiplication gain produced by photons arriving early in the bit.

The third modeling method is a stochastically self-regulating avalanche model for passively quenched SPADs, which is the first significant expansion beyond the model presented by Haitz in 1964 [42]. The presented model considers a passively quenched SPAD as a closed loop system, capturing the effect of the feedback introduced by the resistive load on the stochastic nature of the avalanche multiplication.

This approach differs from the traditional model [42, 15], which captures the static feedback, maintaining the device at breakdown, but neglects the dynamic coupling between the voltage across the SPAD, the feedback from the load and the impact ionization process. To capture the coupling between the feedback and its effect on the stochastic nature of the avalanche current a hybrid model has been developed that uses an analytical circuit model in conjunction with a Monte Carlo simulator of the multiplication process that allows the impact ionization coefficients to change continuously in time according to the dynamic and stochastic feedback received from the load. The stochastically self-regulating avalanche model reported in this dissertation addresses three important phenomena that are entirely beyond the scope of the traditional modeling tools. First, it predicts an oscillatory behavior of a persistent avalanche current. Second, it predicts that the probability density function of the stochastic quenching time of the persistent avalanche current has an exponential decay. Third, under device and operational conditions that lead to strong feedback, the stochastic avalanche current can collapse before persistent avalanche current can be realized. All three of these behaviors are in qualitative agreement with recent experimental demonstrations employing negative-feedback SPADs (NFADs) that had until now not been theoretically explained. The model specifically captures the effect of the load's feedback on the stochastic avalanche multiplication, an effect believed to be key in breaking today's counting rate barrier in the $1.55-\mu$m detection window.

Future work is needed to investigate the dependence of the probability density function (pdf) of the quenching time on the strength of the feedback. This can be determined by calculating the pdf of the quenching time by using different values of the load resistor R_L. The goal would be to find out how the probability density function of the quenching time changes as the system goes from a self-sustaining regime (lower values of R_L) to the case in which the stochastic avalanche current collapses fast, before a persistent avalanche current can be established (higher values of R_L).

In the experimental part of this dissertation a new GaAs-based midwave infrared photodetector, called the quantum dot avalanche photodiode (QDAP), was presented. The QDAP structures under test use an optimized DWELL structure as the absorption layer and a p-i-n avalanche photodiode as the multiplication stage. The proposed device exploits the impact ionization mechanism to increase the conversion efficiency of the DWELL photodetector. The photocurrent of the QDAP structures under test shows an increase as the reverse voltage across the APD section was increased. In fact, the photocurrent of the QDAP increased in a nonlinear fashion as the reverse voltage across the APD was increased. Compared to the photocurrent characteristic of the DWELL section alone, the QDAP photocurrent is very distinct and its characteristic is very similar to that of the I–V curve of the APD section. These results suggest that the photocarriers generated in the DWELL section are amplified by the APD section. It is shown that the conversion efficiency of the DWELL detector is increased by a factor of 14 due to the gain introduced by the avalanche multiplication stage of the QDAP. It is expected that by operating the QDAP with the DWELL section reverse biased at 2.0 V it is possible to obtain a maximum conversion efficiency of about 12%.

Further work is needed to obtain lower noise QDAP structures in order to achieve Geiger mode operation. As seen in Chapter 5, the sensitivity of the QDAP worsened as the voltage across the APD section increased. This is an inevitable result of the operation of the QDAP, in which the DWELL and the APD sections contribute to the noise of the device. Thus, the real advantage of the QDAP is in Geiger mode operation. In Geiger mode operation, in which an APD can be considered as a photon-activated switch, performance metrics like the excess noise factor and the concept of gain are irrelevant since in Geiger mode what is important is the detection of the photon-arrival event rather than the linear amplification of the photogenerated carriers. However, to achieve Geiger mode operation the dark current level needs to be dramatically decreased in order to keep the dark count rate at a reasonable level.

The dark current level that makes Geiger mode operation possible is of the order of $< 1 \times 10^{-9}$ A. Currently, at 77 K the dark current of the QDAP, before breakdown occurs, is of the order of 1×10^{-7} A. Therefore, it is necessary to decrease the dark current by around two orders of magnitude. New DWELL sections with resonant tunneling structures, which have lower dark current levels, have been considered to be used in the next generation of QDAP devices. Another way to reduce the dark current of the device is to employ APD sections with thicker multiplication regions to reduce the dark current contribution due to tunneling effects.

Appendix A

Calculation of the breakdown probabilities

In this chapter I freely draw upon derivations and results reported elsewhere [26, 51, 54, 56, 62]

A.1 Preliminaries

Consider a multiplication region extending from $x = 0$ to $x = w$, and assume that the electric field therein is $E(x)$, pointing in the opposite direction of x. We will further assume, in general, that the multiplication region consists of multiple layers. The goal is to characterize the probability density function (pdf) of the distance from the birth location of a carrier to the location of its first impact ionization thereafter. If an electron (respectively, hole) is born at position x, we let $h_e(\xi, x)$ [respectively, $h_h(\xi, x)$] denote the pdf of the distance to the first ionization, measured from the carrier's birth position at x. For example, $h_e(\xi, x)\Delta$ is approximately the probability that an electron born at x first impact ionizes somewhere in the interval $[x+\xi, x+\xi+\Delta]$. We begin by identifying the key physical parameters that govern this pdf. These are: 1) the multiplication-region's ionization threshold-energy profile; 2) the carrier's dead-space profile; and 3) the profile of the ionization coefficients of enabled carriers (those that have traveled the required dead space). To accommodate the requirement that the multiplication region may consist of layers of different materials, we will allow the electron and hole ionization threshold energies, $E_{th,e}(x)$ and $E_{th,h}(x)$, respectively, to be position-dependent.

For an electron (respectively, hole) created at position x, let $d_e(x)$ [respectively, $d_h(h)$] be the dead space with which it is associated. With this convention, an electron (respectively, hole) which is newly created at x position cannot impact ionize before reaching $x+d_e(x)$ [respectively, $x-d_h(x)$]. Finally, let α and β denote the electron and hole ionization coefficients, respectively, associated with carriers that have acquired the ionization threshold energy. These coefficients are material specific and depend only on the electric field $E(x)$, independently of the multiplication-layer width. A model for the electron and hole impact ionization coefficients of enabled carriers has been developed by Saleh *et al.* [51, 62]. This model is given by

$$\begin{aligned} \alpha(x) &= A_e \exp\left[-\Big(\mathcal{E}_e/E(x)\Big)^{m_e}\right], \\ \beta(x) &= A_h \exp\left[-\Big(\mathcal{E}_h/E(x)\Big)^{m_h}\right] \end{aligned} \tag{A.1}$$

where A, $\mathcal{E}$, and m are parameters chosen by fitting measured excess-noise-factor data [63, 64]. After calculating unique pairs of electron and hole ionization coefficients for every pair of experimental gain and excess noise factor (corresponding to a specific electric field E), one can obtain the parameters A, $\mathcal{E}$, and m by fitting the ionization coefficients to the model given by (A.1).

With the availability of profiles of the dead-space and the ionization coefficients, the expression for $h_e(\xi|x)$ and $h_h(\xi|x)$ are given by [56]

$$h_e(\xi|x) = \alpha(\xi)\exp\left(-\int_{x+d_e(x)}^{\xi}\alpha(u)\,du\right), \quad \xi \geq x + d_e(x) \tag{A.2}$$

and

$$h_h(\xi|x) = \beta(\xi)\exp\left(-\int_{\xi}^{x-d_h(x)}\beta(u)du\right), \quad \xi \leq x - d_h(x) \tag{A.3}$$

To make the above pdfs suitable for multilayer multiplication regions, we must thoroughly characterize the dead space profiles in heterostructure. Figure A.1 shows a qualitative example of the probability density function of the ionization distance, $h_e(\xi, x)$, described by (A.2). In this example the electron is born at $x = 0$, and the width of the multiplication region is $w = 1600$ nm. To better visualize the shifting of the probability density function due to the dead space, $d_e(x)$, the value of the dead space was assumed to be larger (300 nm) than it is in a real material. Typical values of the dead space in GaAs are around [25 – 30] nm at electric fields of the order of 6×10^5 V/cm [80].

A.1.1 Characterization of the dead space

Under the simplifying assumption that after each impact ionization a carrier starts from zero initial energy, the minimum distance that an electron, born at position x, must travel before acquiring the ionization threshold energy is governed by the following energy relation

$$E_{th,e}\big(x + d_e(x)\big) = q\int_{x}^{x+d_e(x)} E(u)\,du, \tag{A.4}$$

The above expression is a simple extension of the dead-space definition in [56], which now captures position-dependent ionization thresholds. Equation (A.4) neglects the effect of scattering, which, as described in Chapter 1, increases the dead space.

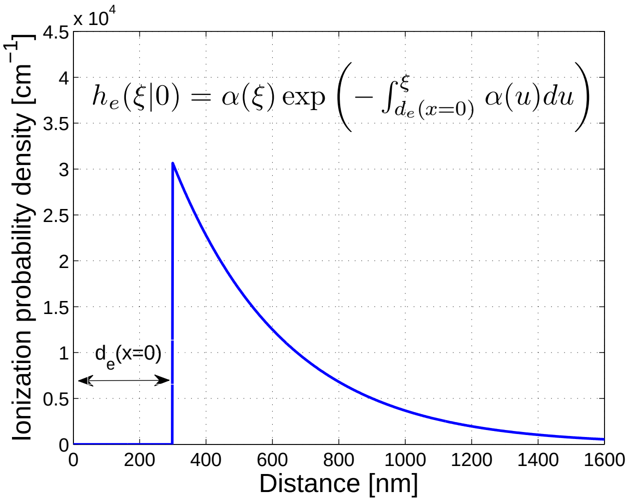


Figure A.1: Probability density function of the ionization distance. In this example the electron is born at $x = 0$. The value of the dead space was assumed to be larger (300 nm) than it is in a real material. Typical values of the dead space in GaAs are around [25 – 30] nm at electric fields of the order of 6×10^5 V/cm [80].

However, (A.4) gives a good approximation of the dead space for high electric fields ($> 1 \times 10^5$ V/cm) where the carriers in the depleted multiplication region can gain energy from the electric field at a faster rate than they lose it to the various scattering processes. Recall that the threshold energy $E_{th,e}$ may vary with x according to the type of material at x. Furthermore, observe that for each x, the relevant ionization threshold energy is the value at the point where the carrier attains the ionization threshold. Hence, for an electron born at location x, the dead space which must be traveled, is the minimum nonnegative solution δ to the following equation:

$$E_{th,e}\big(x+\delta\big) = q\int_{x}^{x+\delta} E(u)\,du, \tag{A.5}$$

Similarly, the hole dead space $d_h(x)$ is the minimum nonnegative solution δ to the following equation

$$E_{th,h}\big(x-\delta\big) = q\int_{x-\delta}^{x} E(u)\,du, \tag{A.6}$$

In our formulation of the dead-space model, we adopted the commonly-accepted assumption that the dead space is deterministic. In actuality, the dead space is a random variable since a carrier may not necessarily loose all of its kinetic energy after each impact ionization.

A.2 Breakdown probabilities

We now characterize the breakdown probability. Let $Z(x)$ denote the total electron and hole population resulting from a parent electron born at x with zero initial energy. Similarly, let $Y(x)$ denote the total electron and hole population resulting from a parent hole born at x. Thus, for the case of electron injection (at $x=0$), the APD gain G is given by $0.5(Z(0)+1)$. Let be $P_z(x)$ defined as the probability that $Z(x)$ is finite, and similarly, let $P_Y(x) \equiv \mathrm{P}\{Y(x) < \infty\}$. McIntyre invoked a recurrence argument and characterized P_Z and P_Y through the following two nonlinear integral equations [81]:

$$P_Z(x) = \int_{w-x}^{\infty} h_e(\xi|x)\,d\xi + \int_0^{w-x}\Big(P_Z^2(x+\xi)P_Y(x+\xi)\Big)h_e(\xi|x)\,d\xi \tag{A.7}$$

$$P_Y(x) = \int_{x}^{\infty} h_h(\xi|x)\,d\xi + \int_0^{x}\Big(P_Y^2(x-\xi)P_Z(x-\xi)\Big)h_h(\xi|x)\,d\xi \tag{A.8}$$

The coupled recursive equations (A.7) and (A.8) can be numerically solved using an iterative approach. Once these equations are solved the breakdown probability is calculated as $1-P_Z(0)$.

References

[1] C. P. Morath, K. Vaccaro, W. R. Clark, W. A. Teynor, M. A. Roland, and W. Bailey. Performnace characterization of an InGaAs/InP single photon avalanche diode. In *Proceedings of the 49th Annual Meeting of the SPIE: Applications of Digital Image Processing XXVII*, Denver, CO, 2004.

[2] A. Biswas and W. H. Farr. Detectors for ground-based reception of laser communications from Mars. *Lasers and Electro-Optics Society (LEOS). The 17th Annual Meeting of the IEEE*, 1:74–75, 2004.

[3] A. Spinelli, L. M. Davis, and H. Dautet. Actively quenched single-photon avalanche diode for high repetition rate time-gated photon counting. *Rev. Sci. Instrum.*, 67:55–61, 1996.

[4] D. Stucki, G. Ribordy, A. Stefanov, H. Zbinden, and J. G. Rarity. Photon counting for quantum key distribution with Peltier cooled InGaAs/InP APD's. *J. Modern Optics*, 48:1967–1981, 2001.

[5] H. Zbinden, H. Bechmann-Pasquanucci, N. Gisin, and G. Ribordy. Quantum cryptography. *Appl. Phys. B*, 67:743–748, 1998.

[6] J. M. Merolla, Y. Mazurenko, J. P. Goedgebuer, L. Duraffourg, H. Porte, and W. T. Rhodes. Quantum cryptographic device using single-photon phase modulation. *Physical Rev. A*, 60:1899–1905, 1999.

[7] M. B. Nasr, B. E. A. Saleh, A. V. Sergienko, and M. C. Teich. Dispersion-cancelled and dispersion-sensitive quantum optical coherence tomography. *Optics Express*, 12:1353–1362, 2004.

[8] Joe C. Campbell, Stephane Demiguel, Feng Ma, Ariane Beck, Xiangyi Guo, Shuling Wang, Xiaoguang Zheng, Xiaowei Li, Jeffrey D. Beck, Senior Member, Michael A. Kinch, Andrew Huntington, Larry A. Coldren, , Jean Decobert, and Nadine Tscherptner. Recent Advances in Avalanche Photodiodes. *IEEE J. of Selected Topics in Quantum Electronics*, 10:777–787, 2004.

[9] Roger L. Freeman. *Telecommunication System Engineering*. Wiley InterScience, fourth edition edition, 2004.

[10] R. J. McIntyre. Multiplication noise in uniform avalanche photodiodes. *IEEE Trans. Electron devices*, ED. 13:164–168, 1966.

[11] J. P. R. David and C. H. Tan. Material considerations for avalanche photodiodes. *IEEE Journal of Selected Topics in Quantum Electronics*, 14:998–1009, 2008.

[12] Federico Capasso. *Physics of Avalanche Photodiodes, in semiconductors and semimetals, Vol. 22, Part D*. Academic Press, Inc., 1985.

[13] Ansgar Jüngel. *Transport Equations for Semiconductors*. Springer, 2009.

[14] S. M. Sze. *Physics of semiconductor devices*. New York: Wiley, 1981.

[15] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa. Avalanche photodiodes and quenching circuits for single-photon detection. *Applied Optics*, 35:1956–1976, 1996.

[16] M. A. Itzler, X. Jiang, M. Entwistle, K. Slomkowski, A. Tosi, F. Acerbi, F. Zappa, and S. Cova. Advances in InGaAsP-based avalanche diode single photon detectors. *Journal of Modern Optics*, 58:174–200, 2011.

[17] B. E. A. Saleh and M. C. Teich. *Fundamentals of Photonics*. Wiley, New York, 1991.

[18] Govind P. Agrawal. *Fiber-Optic Communication Systems*. Wiley InterScience, third edition edition, 2002.

[19] Govind P. Agrawal. *Lightwave Technology: Telecommunication Systems*. Wiley InterScience, 2005.

[20] Giovanni Ghione. *Semiconductor Devices for High-Speed Optoelectronics*. Cambridge University Press, 2009.

[21] Y. Kang, H. X. Lu, Y. H. Lo, D. S. Bethune, and W. P. Risk. Dark count probability and quantum efficiency of avalanche photodiodes for single-photon detection. *Appl. Phys. Lett.*, 83:2955–2957, 2003.

[22] K. A. McIntosh, J. P. Donnelly, D. C. Oakley, A. Napoleone, S. D. Calawa, L. J. Mahoney, K. M. Molvar, E. K. Duerr, S. H. Groves, and D. C. Shaver. InGaAsP/InP avalanche photodiodes for photon counting at 1.06 μm. *Appl. Phys. Lett*, 81:2505–2507, 2002.

[23] J. P. Donnelly, E. K. Duerr, K. A. McIntosh, E. A. Dauler, D. C. Oakley, S. H. Groves, C. J. Vineis, L. J. Mahoney, K. M. Molvar, P. I. Hopman, K. E. Jensen, G. M. Smith, S. Verghese, and D. C. Shaver. Design Considerations for 1.06-μm InGaAsP-InP Geiger-Mode Avalanche Photodiodes. *IEEE J. Quantum Electron.*, 42(8):797–809, 2006.

[24] G. Ribordy, N. Gisin, O. Guinnard, D. Stucki, M. Wegmuller, and H. Zbinden. Photon counting at telecom wavelengths with commercial InGaAs/InP avalanche photodiodes: current performance. *J. Modern Optics*, 51:1381–1398, 2004.

[25] S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa. Evolution and prospects for single-photon avalanche diodes and quenching circuits. *J. Modern Optics*, 51:1267–1288, 2004.

[26] M. M. Hayat, Ünal Sakoğlu, O-H. Kwon, S. Wang, J. C. Campbell, B. E. A. Saleh, and M. C. Teich. Breakdown probabilities for thin heterostructure avalanche photodiodes. *IEEE Journal of Quantum Electron.*, 39:179–185, 2003.

[27] S. Krishna. Quantum dots-in-a-wells infrared photodetectors. *J. of physics D: Appl. Phys.*, 38:2142–2150, 2005.

[28] Z. Ye, J. C. Campbell, Z. Chen, E-T. Kim, and A. Madhukar. Normal-incidence InAs sefl-assembled quantum-dot infrared photodetectors with high detectivity. *J. Quantum Electronics*, 38:1234–1237, 2002.

[29] M. A. Kinch. HgCdTe: Recent Trends in the Ultimate IR Semiconductor. *Journal of Electronic Materials*, 39:1043–1052, 2010.

[30] Donald N. B. Hall, Richard S. Aikens, Richard Joyce, and Thomas W. McCurnin. Johnson Noise Limited Operation of Photovoltaic InSb Detectors. *Applied Optics*, 14:450–453, 1975.

[31] G. A. SaiHalasz, R. Tsu, and L. Esaki. A new semiconductor superlattice. *Appl. Phys. Lett.*, 30:651, 1977.

[32] T. E. Vandervelde, M. C. Lenz, E. Varley, A. Barve, J. Shao, R. V. Shenoi, D. A Ramirez, W. Jang, Y. D. Sharma, and S. Krishna. Quantum dots-in-a-well focal plane arrays. *J. Sel. Top. in Quantum Electronics*, 14:1150–1161, 2008.

[33] S. Krishna, O H. Kwon, and M. M. Hayat. Theoretical investigation of Quantum-dots avalanche photodiodes for Mid-infrared applications. *J. Quantum Electronics*, 41:1468–1473, 2005.

[34] Massimo Ghioni, Angelo Gulinatti, Ivan Rech, Franco Zappa, and Sergio Cova. Progress in silicon single-photon avalanche diodes. *IEEE Journal of Selected Topics in Quantum Electronics*, 13:852–862, 2007.

[35] http://www.princetonlightwave.com.

[36] Xudong Jiang, Mark A. Itzler, Rafael Ben-Michael, Krystyna Slomkowski, Michael A. Krainak, Stewart Wu, and Xiaoli Sun. Afterpulsing Effects in Free-Running InGaAsP Single-Photon Avalanche Diodes. *IEEE Journal of Quantum Electronics*, 44:3–11, January 2008.

[37] Peter Seitz and Albert J.P. Theuwissen. *Single-Photon Imaging*. Springer, fourth edition edition, 2011.

[38] J. Beck, C. Wan, M. Kinch, J. Robinson, P. Mitra, R. Scritchfield, F. Ma, and J. Campbell. The HgCdTe electron avalanche photodiode. *Journal of Electronic Materials*, 35:1166–1173, June 2006.

[39] J. Beck, M. Woodall, R. Scritchfield, M. Ohlson, L. Wood, P. Mitra, and J. Robinson. Gated IR Imaging with 128 x 128 HgCdTe Electron Avalanche Photodiode FPA. *J. Electron. Mater*, 37:1334–1343, 2008.

[40] J. Rothman, E. de Borniol, S. Bisotto, L. Mollard, F. Guellec, F. Pistone, S. Courtas, and X. Lefoule. HgCdTe APD-Focal Plane Array development at DEFIR for low flux and photon-counting applications. *Quantum of Quasars workshop*, 2009.

[41] John Vallerga, Jason Mcphate, Larry Dawson, and Maryn Stapelbroek. Mid-IR photon counting array using HgCdTe APDs and the Medipix2 ROIC.

[42] R. H. Haitz. Model for the electrical behavior of a microplasma. *J. Appl. Phys.*, 35:1370–1376, 1964.

[43] M. A. Itzler, Xudong Jiang, B. M. Onat, and Krystyna Slomkowski. Progress in self-quenching InP-based single photon detectors. *Proc. of SPIE*, 7608:760829, 2010.

[44] M. A. Itzler, R Ben-Michael, C. F. Hsu, K. Slomkowski, A. Tosi, S. Cova, F. Zappa, and R. Ispasoiu. Single photon avalanche diodes (SPADs) for 1.5 μm photon counting applications. *Journal of Modern Optics*, 54:283–304, 2007.

[45] M. A. Itzler, Xudong Jiang, B. Nyman, and Krystyna Slomkowski. InP-based Negative Feedback Avalanche Diodes. *Proc. of SPIE*, 7222:72221K, 2009.

[46] K. Zhao, S. You, J. Cheng, and Y. Lo. Self-quenching and self-recovering InGaAs/InAlAs single photon avalanche detector. *Appl. Phys. Lett.*, 93:153504, 2008.

[47] D. A. Ramirez, M. M. Hayat, G. Karve, J. C. Campbell, S. N. Torres, B. E. A. Saleh, and M. C. Teich. Generalized breakdown probabilities and detection efficiencies for single-photon avalanche photodiodes. *IEEE J. Quantum Electron.*, 42(2):137–145, 2006.

[48] S. R. Forrest, R. F. Leheny, R. E. Nahony, and M. A. Pollack. $In_{0.53}Ga_{0.47}As$ photodiodes with dark current limited by generation-recombination and tunneling. *Appl. Phys. Lett*, 37(3):322–325, 1980.

[49] G. Karve, X. Zheng, X. Zhang, X. Li, S. Wang, F. Ma, A. Holmes, J. C. Campbell, G. S. Kinsey, J. C. Boisvert, T. D. Isshiki, R. Sudharsanan, D. S. Bethune, and W. P. Risk. Geiger mode of an $In_{0.53}Ga_{0.47}As$-$In_{0.52}Al_{0.48}As$ avalanche photodiodes. *IEEE Journal of Quantum Electron.*, 39:1281–1286, 2003.

[50] C. H. Tan, J. S. Ng, S. Xie, and J. P. R. David. Potential Materials for Avalanche Photodiodes Operating above 10 Gb/s. *Inter. Conf. on Computer and Dev. for Communication*, 2009.

[51] M. A. Saleh, M. M. Hayat, P. Sotirelis, A. L. Holmes, J. C. Campbell, B. E. A. Saleh, and M. C. Teich. Impact-ionization and noise characteristics of thin III-V avalanche photodiodes. *IEEE Trans. Electron Devices*, 48:2722–2731, 2001.

[52] J. S. Ng, C. H. Tan, J. P. R. David, G. Hill, and G. J. Rees. Field dependence of impact ionization coefficients in $In_{0.53}Ga_{0.47}As$. *IEEE Transaction on electron devices*, 50(4):901–905, April 2003.

[53] Fukunobu Osaka, Takashi Mikawa, and Takao Kaneda. Electron and hole ionization coefficients in (100) oriented $Ga_{0.33}In_{0.67}As_{0.70}P_{0.30}$. *Appl. Phys. Lett.*, 45(3):292–293, 1984.

[54] M. M. Hayat, O-H. Kwon, S. Wang, J. C. Campbell, B. E. A. Saleh, and M. C. Teich. Boundary effects on multiplication noise in thin heterostructure avalanche photodiodes: Theory and experiment. *IEEE Trans. on Electron Devices*, 49(12):2114–2123, 2002.

[55] B. E. A. Saleh, M. M. Hayat, and M. C. Teich. Effect of dead space on the excess noise factor and time response of avalanche photodiodes. *IEEE Trans. Electron Devices*, 37:1976–1984, 1990.

[56] M. M. Hayat, W. L. Sargeant, and B. E. A. Saleh. Effect of dead space on gain and noise in Si and GaAs avalanche photodiodes. *IEEE J. Quantum Electron.*, 28:1360–1365, 1992.

[57] O-H. Kwon, M. M. Hayat, S. Wang, J. C. Campbell, A. Holmes, B. E. A. Saleh, and M. C. Teich. Optimal excess-noise reduction in thin heterojunction $Al_{0.6}Ga_{0.4}As/GaAs$ avalanche photodiodes. *Journal of Quantum Electron.*, 39:1287–1296, 2003.

[58] M. M. Hayat and G. Dong. A new approach for computing the bandwidth

statistics of avalanche photodiodes. *IEEE Trans. Electron Devices*, 47:1273–1279, June 2000.

[59] M.M. Hayat, J.C. Campbell, B.E.a. Saleh, and M.C. Teich. Gain-bandwidth product optimization of heterostructure avalanche photodiodes. *J. Lightwave Technol.*, 23:1896–1906, May 2005.

[60] M.M Hayat, O-h. Kwon, Y Pan, P Sotirelis, J C Campbell, B E A Saleh, and M C Teich. Gain-bandwidth characteristics of thin avalanche photodiodes. *IEEE Trans. Electron Devices*, 49:770–781, 2002.

[61] M. M. Hayat and D. A. Ramirez. Theory of dynamically biased avalanche photodiodes: New limits for Gain-Bandwidth product. *In preparation.*

[62] M. A. Saleh, M. M. Hayat, B. E. A. Saleh, and M. C. Teich. Dead-space-based theory correctly predicts excess noise factor for thin GaAs and AlGaAs avalanche photodiodes. *IEEE Trans. Electron Devices*, 47:625–633, 2000.

[63] G. E. Bulman, V. M. Robbins, K. F. Brennan, K. Hess, and G. E. Stillman. Experimental determination of impact ionization coefficients in (100) GaAs. *IEEE Electron Devices Lett.,*, EDL-4:181–185, 1983.

[64] V. M. Robbins, S. C. Smith, and G. E. Stillman. Impact ionization in $\mathrm{Al}_x\mathrm{Ga}_{1-x}\mathrm{As}$ for x=0.1 – 0.4. *Appl. Phys. Lett.*, 52:296, 1988.

[65] M. M. Hayat, B. E. A. Saleh, and M. C. Teich. Effect of dead space on gain and noise of double-carrier multiplication avalanche photodiodes. *IEEE Trans. Electron Devices*, 39:546–552, 1992.

[66] M. M. Hayat and B. E. A. Saleh. Statistical Properties of the Impulse Response Function of Double-Carrier Multiplication Avalanche Photodiodes Including the Effect of Dead Space. *J. Lightwave Technol.*, 10:1415–1425, 1992.

[67] L. J. J. Tan, J. S. Ng, C. H. Tan, and J. P. R. David. Avalanche Noise Characteristics in Submicron InP Diodes. *IEEE Journal of Quantum Electronics*, 44:378–382, 2008.

[68] D. A. Ramirez, M. M. Hayat, G. J. Rees, Xudong Jiang, and M. A. Itzler. New perspective on passively quenched single photon avalanche diodes: effect of feedback on impact ionization. *Submitted to Optics Express.*

[69] C. Groves, C. H. Tan, J. P. R. David, G. J. Rees, and M. M. Hayat. Exponential Time Response in Analogue and Geiger Mode Avalanche Photodiodes. *IEEE Trans. Electron Devices*, 52:1527–1534, 2005.

[70] M. M. Hayat, M. A. Itzler, D. A. Ramirez, and G. J. Rees. Model for Passive Quenching of SPADs. *Proc. of SPIE*, 7608:76082B–76082B–8, 2010.

[71] D. Shushakov and V. Shubin. New solid state photomultiplier. *Proc. of SPIE*, 2397:544–554, 1995.

[72] D. Shushakov and V. Shubin. New avalanche device with an ability of a few-photon light pulse detection in analog mode. *Proc. of SPIE*, 2699:173–183, 1996.

[73] InGaAs APD: G8931-20. *Hamamatsu datasheet*, pages 1–2, 2009.

[74] A. Barve, Sang Jun. Lee, Sam Kyu. Noh, and S. Krishna. Review of current progress in quantum dot infrared photodetectors. *Laser and Photon. Rev.*, 13:1–13, 2009.

[75] David Z.-Y. Ting, Sumith V. Bandara, Sarath D. Gunapala, Jason M. Mumolo, Sam A. Keo, Cory J. Hill, John K. Liu, Edward R. Blazejewski, Sir B. Rafol, and Yia-Chung Chang. Submonolayer quantum dot photodetector. *Appl. Phys. Lett*, 94:111107–1–3, 2009.

[76] P. Yuan, S. Wang, X. Sun, X. G. Zheng, A. L. Holmes, Jr., and J. C. Campbell. Avalanche photodiodes with an impact-ionization-engineered multiplication region. *IEEE Photon. Technol. Lett.*, 12:1370–1372, 2000.

[77] R. V. Shenoi, R. S. Attaluri, A. Siroya, J. Shao, Y. D. Sharma, A. Stintz, T. E. Vandervelde, and S. Krishna. Low-strain InAs/InGaAs/GaAs quantum dots-in-a-well infrared photodetector. *J. Vac. Sci. Technol. B*, 26:1136–1139, 2008.

[78] R. S. Attaluri, S. Annamali, K. T. Posani, A. Stintz, and S. Krishna. Influence of Si doping on the performance of quantum dots-in-well photodetectors. *J. Vac. Sci. Technol. B*, 24:1553–1555, 2006.

[79] E. L. Dereniak and G. D. Boreman. *Infrared Detectors and Systems.* Wiley InterScience, 1996.

[80] Xiaowei Li, Xiaoguang Zheng, Shuling Wang, Feng Ma, and Joe C. Campbell. Calculation of Gain and Noise With Dead Space for GaAs and $Al_xGa_{1-x}As$ Avalanche Photodiodes. *IEEE Transactions on Electron Devices*, 49:1112–1217, 2002.

[81] R. J. McIntyre. A new look at impact ionization-part I: A theory of gain, noise, breakdown probability, and frequency response. *IEEE Trans. Electron devices*, 46:1623–1631, 1999.