Improved breakdown model for estimating dark count rate in avalanche photodiodes with InP and InAlAs multiplication layers

Andrew S. Huntington*, Madison A. Compton, George M. Williams
Voxtel Inc., 12725 SW Millikan Way, Suite 230, Beaverton, OR, USA 97005-1782

ABSTRACT

We present an improved method for estimating the dark count rate of single-photon-sensitive avalanche photodiodes (SPADs) with either InP or InAlAs multiplication layers. Our simulation of junction breakdown probability can easily accommodate arbitrary electric field profiles and APD bias conditions. In combination with local models of dark carrier generation, our technique can provide more realistic estimates of dark count rate than are obtained by multiplying the primary dark current by a single junction breakdown probability, or by assuming constant electric fields in the multiplication layer. Our method can assist in the design of SPADs for demanding laser radar applications.

Keywords: Avalanche photodiodes, Geiger mode, photon counting, InGaAs, InAlAs, InP, dark count rate, breakdown

1. INTRODUCTION

1.1. InGaAsP and AlGaInAs SACM APDs for near infrared applications

Two semiconductor alloys are commonly used to fabricate photodetectors sensitive in the near infrared (NIR) between 1000 – 1600 nm. Germanium has a room temperature indirect bandgap of 0.66 eV (1880 nm) and an absorption coefficient of 350 cm\(^{-1}\) at 1550 nm; In\(_{0.53}\)Ga\(_{0.47}\)As – the composition lattice-matched to InP – has a room temperature direct bandgap of 0.75 eV (1650 nm) and an absorption coefficient of 6100 cm\(^{-1}\) at 1550 nm. In addition to being a stronger NIR absorber, InGaAs is preferred for the manufacture of APDs because heterostructures in which it is combined with wider gap alloys suitable for avalanche multiplication can be easily grown.

InGaAs is the narrow gap endpoint of two similar quaternary alloy systems: it is blended with InP in the InGaAsP alloy system, and with In\(_{0.52}\)Al\(_{0.48}\)As – the composition lattice-matched to InP – in the AlGaInAs alloy system (Figure 1). APDs and other optoelectronic devices such as lasers are commercially manufactured in both the InGaAsP and AlGaInAs alloy systems. Although both alloy systems span a similar bandgap range (InP has a room temperature bandgap of 1.34 eV; for InAlAs it is 1.46 eV), the band offset ratio is higher in AlGaInAs than in InGaAsP (0.7 versus 0.4), and the impact ionization rate for electrons in InAlAs is higher than that for holes, but vice versa for InP.

NIR APDs usually employ the separate absorption, charge, and multiplication (SACM) design in which generation of primary photocarriers takes place in a different layer than avalanche multiplication (Figure 2). Narrow bandgap alloys are susceptible to serious tunnel leakage when a strong electric field is applied, so a weakly doped space charge layer inserted between the absorber and multiplication layer is used to keep the field in the absorber low when the field in the multiplication layer is high.

* andrewh@voxtel-inc.com

Figure 1: Plot of band gap versus lattice constant of several semiconductor alloys. Alloy systems lattice-matched to InP substrates are aligned vertically on the dashed line.
1.2. Geiger mode operation of an APD for photon counting

APDs can be operated in two different modes: Linear mode and Geiger mode (Figure 3). In Linear mode, the reverse bias applied to the APD is held constant, and the primary photocurrent generated in the APD’s absorber is amplified by a proportional multiplication factor that is independent of signal strength (below saturation). The output of a linear APD is proportional to the level of illumination it receives. Linear mode APDs are typically used in optical receivers to boost weak signals above the noise floor of the receiver’s amplifier. In Geiger mode the reverse bias is modulated, and the APD’s response is binary. Geiger mode operation of an APD involves momentarily biasing the diode above its avalanche breakdown voltage \( V_{br} \). The excess voltage applied is called the overbias. In this active state, avalanche breakdown of the diode junction can be triggered by as little as a single primary carrier. The current that flows during avalanche breakdown is determined by the characteristics of the external circuit rather than the number of primary carriers that initiated the breakdown, so the breakdown current of a Geiger APD is the same for all signal strengths. Thus, the ability to read signal amplitude is sacrificed for maximum sensitivity in a Geiger APD. The extreme sensitivity of Geiger mode operation makes Geiger APDs suitable for photon counting tasks. However, Geiger SPADs are correspondingly susceptible to upset by dark counts and must be operated under conditions where generation of primary carriers of all types is sparse.

Geiger APDs require a quench circuit to bring them out of breakdown and reset them for the next detection event. The simplest quenching circuit is just a high-value load resistor in series with the diode. In the passive quenching scheme, the large current flow that accompanies avalanche breakdown of the Geiger APD causes a corresponding voltage drop across the load resistor. This, in turn, acts to reduce the reverse bias applied to the APD below \( V_{br} \). When the avalanche process quenches, the current flow through the load resistor ceases, and the full overbias is restored. Passive quenching is relatively rapid, but restoration of bias afterwards can be slow. More seriously, there is no way to adjust the timing of a passive quenching circuit. Quenching circuits that employ active components to modulate the bias supplied to a Geiger APD are called active quenching circuits (AQCs). AQCs enable the user to gate the overbias, controlling both the duration of the gate-on periods during which the overbias is applied, and also the hold-off time between quench and reset.

1.3. Dark counts in Geiger SPADs

Geiger SPADs suffer two types of spurious detection events: ‘dark’ counts from leakage in the diode junction, and non-signal counts from stray illumination. Both types of spurious detection event are Poisson-distributed and uncorrelated; they can be characterized by their mean rates of occurrence, and those rates are additive. Since the amount of stray light that reaches a Geiger SPAD is a feature of its application rather than its internal dynamics, this paper is only concerned with spurious detection events from junction leakage.

The dark count rate (DCR) of a Geiger SPAD is one of its most important parameters. There is no way to distinguish an avalanche breakdown event caused by a dark carrier from one caused by a photocarrier, so weak optical signals will only be detectible if they boost the rate at which the Geiger SPAD fires above the DCR. Dark counts also seriously limit the duty cycle of a Geiger SPAD. Once a Geiger SPAD fires, it must be quenched below \( V_{br} \) and held off long enough for any traps Figure 2: The separate absorption, charge, multiplication (SACM) structure with layer sequence ordered for electron-initiated avalanche multiplication. Layer design parameters are indicated.

Figure 3: APD operation illustrated by \( I-V \) characteristics. Current rises with increasing illumination. In Linear mode, avalanche gain is finite, and the primary current is amplified proportionally; in Geiger mode, the junction breaks down in response to any level of primary current.
that were filled during the discharge to release their carriers. If the overbias is restored before the traps have emptied, the carriers emitted by the traps will increase the likelihood of a new spurious count—an echo of the detection event that filled the traps in the first place. This is called ‘afterpulsing’ (Figure 4). A valid signal that arrives during the period of time that the detector is held off will be missed. Thus, higher DCR means lower detector duty cycle.

1.4. Models of DCR

Dark counts in Geiger SPADs have been studied by many researchers because of the DCR’s dominant role in detector performance. The simplest model of DCR is obtained by estimating the number of dark carriers present in the multiplication layer of the APD during a gate-on period \( N_d \), and a uniform breakdown probability \( P_a \) such that the probability of a dark count during the gate-on period \( P_d \) is:

\[
P_d = 1 - \exp\left[-N_d P_a\right].
\]

Ramirez et al. have pointed out that representing the junction breakdown probability as a single parameter is misleading. The junction breakdown probability is a function of where carriers are injected, and it is different for electrons and holes. Since generation of a Geiger SPAD’s primary dark current is distributed throughout the diode, it is necessary to use a local model of breakdown probability. Drawing upon the work of their co-authors, Ramirez et al. applied the recursive dead-space multiplication theory (DSMT) to compute carrier-type-dependent breakdown probabilities as a function of location and estimated the generation of primary dark current based upon the band-to-band tunneling mechanism. However, to simplify their calculations, they assume uniform electric field profiles within the multiplication region and uniformly-distributed dark current generation.

Computational band edge modelers that employ the finite difference method to solve the coupled drift-diffusion and Poisson equations are a cornerstone of modern semiconductor device design. Accurate electric field profiles as a function of bias can be obtained with a minimum of mathematical sophistication on the part of the operator. Here, we report on a convenient way to use the output of such a device simulator to model dark counts in a Geiger SPAD.

2. DARK COUNT MODEL

2.1. Overview

This model applies to axially-symmetric vertical-junction APDs that can adequately be described using a one-dimensional analysis (e.g., circular APDs with epitaxially-grown junctions). We seek expressions for the generation rate of primary dark carriers \( G_{dark} \) and for carrier-type-dependent breakdown probability \( P_{be} \) and \( P_{bh} \) as functions of location. To include the effects of afterpulsing, we must also consider the time-dependency of dark carrier generation. The instantaneous DCR is then:

\[
DCR(t) = A_{SPAD} \int dx \ G_{dark} (x,t) \times [P_{be}(x) + P_{bh}(x)],
\]

where time \( t \) is measured relative to the end of the previous Geiger pulse.

The breakdown probabilities and some of the dark current components in our model depend upon the local electric field strength \( F(x) \), which can be obtained from a computational band edge modeler such as SimWindows. The finite difference method used to solve for the electric field requires that the model environment be discretized. Accordingly, our model is described in terms of a 1-D diode structure that is broken down into \( n \) discrete elements, and the integral...
above is replaced by an equivalent summation. As our starting point, we assume that a discrete 1-D electric field profile has been obtained for the desired bias, expressed as $F_i$ for the $i$'th element of the structure.

### 2.2. Breakdown probability

In analyzing the breakdown probability, we adapt the methodology of Hayat et al. to a discretized version compatible with electric field profiles obtained from band edge modelers. The breakdown problem is stated in terms of two functions – $P_{e}(\text{start}, \text{finish})$ and $P_{h}(\text{start}, \text{finish})$ – which are the respective probabilities that an electron (or hole) injected in element start travels to element finish without triggering an ionization event. To find $P_{e}(\text{start}, \text{finish})$ and $P_{h}(\text{start}, \text{finish})$, one must have the probability of electron- and hole-imitated ionization as a function of location.

In the simplest model of impact-ionization, the probability that a carrier will impact-ionize while traveling through an element of width $w_i$ is:

\[
P_i^{\text{ionize}, e} = \alpha_i \times w_i, \\
P_i^{\text{ionize}, h} = \beta_i \times w_i,
\]

where the ionization rates $\alpha_i$ and $\beta_i$ are material parameters that depend upon $F_i$.

Room-temperature fit parameters published for InP multiplication layers are $A_e = 3.01 \times 10^6$ cm$^{-1}$, $A_h = 4.29 \times 10^6$ cm$^{-1}$, $F_e = 2.45 \times 10^6$ V cm$^{-1}$, $F_h = 2.08 \times 10^6$ V cm$^{-1}$, $ce = 1.08$, and $ch = 1.12$; for InAlAs, $A_e = 4.17 \times 10^6$ cm$^{-1}$, $A_h = 2.65 \times 10^6$ cm$^{-1}$, $F_e = 2.09 \times 10^6$ V cm$^{-1}$, $F_h = 2.79 \times 10^6$ V cm$^{-1}$, $ce = 1.20$, and $ch = 1.07$. It is worth pointing out that ionization rate coefficients increase under cooled operation – as would be the case for a Geiger SPAD – because of reduced scattering with optical phonons. However, accurate fits of ionization rates are often not available except at room temperature.

This simple picture of impact-ionization that depends only on the local field strength is complicated by the fact that carriers generated at rest cannot trigger impact-ionization until they have acquired the necessary kinetic energy. A rigorous treatment of impact-ionization requires Monte Carlo modeling of the complete high-field transport problem in order to track carrier energy and scattering on an individual basis. Plainly, this is prohibitively complicated for the purposes of practical device design. The simplest approximation that represents this physics is to set the ionization rate to zero within the ‘dead space’ of a newly generated carrier – the minimum distance over which the carrier must drift in the applied field to accumulate the ionization threshold energy. Once a carrier has traveled through its dead space, the local field-dependent ionization rate represented by $\alpha_i$ (or $\beta_i$, as appropriate) can be applied. This approximation is the DSMT of Saleh et al. referenced earlier, which has been successful in predicting the statistics of the Linear mode amplification of APDs with thin multiplication layers.

In our model, dead space is represented by $d_e(\text{start})$ or $d_h(\text{start})$ – the index of the element closest to element start in which the electron (or hole) is active. A simple algorithm can be used to find these addresses if the thresholds for electron- and hole-initiated ionization ($E_{i}^{\text{th}, e}$ and $E_{i}^{\text{th}, h}$) are known. In InP, the threshold for electron-initiated impact ionization is 2.05 eV and the hole ionization threshold is 2.20 eV; for InAlAs the electron ionization threshold is 2.15 eV and that for holes is 2.30 eV. If the model is oriented so that the $p$ contact is at $i=0$ and the $n$ contact is at $i=n+1$, then electrons drift in the direction of increasing index, and:

\[
d_e(\text{start}) = \begin{array}{ll}
\text{LET } j = \text{start}; & \text{LET accumulation} = 0; \\
\text{WHILE } j \leq n & \\
& \text{IF accumulation} \geq E_{j}^{\text{th}, e}, \text{THEN RETURN } j \\
& \text{ELSE accumulation} = \text{accumulation} + F_j \times w_j; \quad j = j + 1 \\
\text{RETURN } j 
\end{array}
\]

An equivalent algorithm for the counter-propagating holes can be used to find $d_h(\text{start})$.
IF accumulation ≥ $E_{j}^{th, h}$, THEN RETURN $j$
ELSE accumulation = accumulation + $F_j \times w_j$; \quad j = j - 1
RETURN $j$

In either case, the possibility that a carrier generated in element $start$ cannot ionize before leaving the diode is handled by dead space indices that coincide with the contacts (i.e., 0 or $n+1$). If desired, a conditional expression may be added that allows carriers to lose energy if the local electric field is too low for too long. This provision is useful in modeling structures in which the electric field is modulated by large values over short distances. Monte Carlo simulations of high field carrier transport suggest that energy relaxation lengths in related III-V semiconductor alloys are on the order of 500 Å. However, such structures are not commonly employed in Geiger SPADs.

Returning to $P_{se}(start, finish)$ and $P_{sh}(start, finish)$, the probability that a carrier injected at element $start$ manages to reach element $finish$ without causing an ionization is simply the compound probability that the carrier fails to ionize in every element on its path from the first element in which it is active − $d_e(start)$ or $d_h(start)$:

$$P_{se}(start, finish) = \prod_{j=d_e(start)}^{finish-1} (1 - P_{ionize,e}^{j});$$
$$P_{sh}(start, finish) = \prod_{j=d_h(start)}^{finish-1} (1 - P_{ionize,h}^{j}).$$

A conditional should be used to assign 100% survival probability in such case that $finish$ falls within the dead space of $start$.

Finally, the probability that a carrier generated in element $i$ does not cause breakdown can be expressed as a system of $2n$ coupled equations:

$$P_{nbe}(i) = P_{se}(i, n + 1) + \sum_{j=d_e(i)}^{n} P_{se}(i, j) \times P_{ionize,e}^{j} \times P_{nbe}(j)^2 \times P_{nbe}(j);$$
$$P_{nbb}(i) = P_{sh}(i, 0) + \sum_{j=1}^{d_h(i)} P_{sh}(i, j) \times P_{ionize,h}^{j} \times P_{nbb}(j)^2 \times P_{nbe}(j).$$

In English, each equation simply says that the probability an electron (or hole) injected at element $i$ does not cause breakdown is equal to the probability that either it makes it to the appropriate contact without causing an ionization event, or if it does cause ionization, that all of its progeny eventually escape without causing breakdown. The breakdown probability for injection at any given location is just one minus these unknowns:

$$P_{be}(i) = 1 - P_{nbe}(i);$$
$$P_{bh}(i) = 1 - P_{nbb}(i).$$

This system of equations can be solved numerically using standard techniques, such as are implemented in software packages like Mathematica. The solutions derived by this method assume that carriers of both types complete their journey across the diode without recombining or being trapped. This is a good approximation in the depleted layers of the APD, but a prefactor of $\exp(-\Delta x/L_d)$ should be included for points in undepleted material $\Delta x$ away from the edge of the depletion region, where $L_d$ is the diffusion length of the carrier involved. Alternatively, later in the calculation of DCR, one can simply discard dark count generation more than $L_d$ away from the depletion region, with the assumption that carriers generated in those areas will not be collected by the junction.

This procedure is demonstrated for a SACM APD with a 2.0 µm InGaAs absorber, a 0.1 µm InAlAs charge layer doped $3 \times 10^{17}$ cm$^{-3}$ p-type, and a 3.0 µm InAlAs multiplication layer. An unintentional p-type background doping level of $10^{15}$ cm$^{-3}$ was assumed for both the absorber and multiplication layer. The structure was discretized on a 250-element grid.
mesh, and simulated at 134.4 V reverse bias. The electric field profile obtained from a band edge simulator is plotted in Figure 5, and the carrier-type-dependent breakdown probabilities are shown in Figure 6.

2.3. Generation of primary dark carriers

The main dark carrier generation processes active in a Geiger APD are thermal (also known as G-R), trap release (afterpulsing), trap-assisted tunneling, and band-to-band tunneling. There is also a perimeter leakage current associated with the edge of the device, but to the extent that it bypasses the multiplying junction, it does not participate in the generation of dark counts. The dark carrier generation rate of each of these mechanisms must be found as a function of location and time in order to calculate the instantaneous DCR.

In general, both the carriers generated inside the depletion region and those carriers that are generated within one diffusion length of the junction have a reasonable chance of being collected and contributing to the dark current. Carriers drift across the depletion region according to their charge and the direction of the electric field. The direction of transport by drift is always from minority side to majority side (e.g., electrons drift from the p side to the n side, and holes drift in the opposite direction). Written in discretized form, the DCR is:

\[
DCR(i, t) = A_{SPAD} \sum_{\text{collection}} G_{\text{dark}}(i, t) \times [P_{be}(i) + P_{bh}(i)],
\]

where the range of the summation is all elements within one diffusion length of the depletion region, and:

\[
G_{\text{dark}}(i, t) = G_{\text{G-R}}(i) + G_{\text{trap}}(i) + G_{\text{band-to-band}}(i).
\]

2.3.1. G-R leakage

The G-R current is tied to the carrier recombination rate in the vicinity of the junction because the rate of thermal generation balances the rate of recombination. Recombination rates can be modeled in varying degrees of detail: a third-order polynomial in carrier concentration is often used to combine Shockley-Read-Hall (SRH) recombination at traps (linear term), band-to-band recombination (quadratic term), and Auger recombination (cubic term) into a single expression. The various recombination rates as a function of location are typically calculated by band edge modelers because they are required for continuity of the carrier transport equations. As with the electric field profile, we will assume that \(G_{\text{G-R}}(i)\) is available as output from commonly available device modeling software.

Under normal operating conditions, the G-R leakage will be only weakly dependent upon the applied bias because the width of the depletion region – and hence the volume of collection – will be essentially constant. The biggest
influence on $G_{G,R}$ is the intrinsic carrier concentration, which is a function of alloy composition and temperature. Generally speaking, dark counts attributable to $G_{G,R}$ originate in the narrow-gap InGaAs absorber where the intrinsic carrier concentration is highest, and can be reduced by cooling.

2.3.2. Trap release and trap-assisted tunneling

Traps contribute to the DCR in three ways. The SRH portion of G-R leakage discussed above is calculated by the band edge modeler assuming steady state conditions. However, immediately following Geiger breakdown, the population of filled trap states is larger than its steady state value. Assuming an excess concentration $\Delta N_t$ of filled trap states and a characteristic lifetime $\tau_{trap}$, the generation rate by trap release at time $t$ after breakdown is:

$$G_{trap}(i, t) = \frac{\Delta N_t(i)}{\tau_{trap}} \exp\left(-\frac{t}{\tau_{trap}}\right).$$

The initial excess concentration of filled traps $\Delta N_t$ can be found by application of the Fermi distribution, provided the physical concentration of trap states $N_t$ and their energy $E_t$ are known:

$$\Delta N_t(i) = N_t(i) \left[1 - \left(1 + \exp\left(\frac{E_t - E_F}{k_B T}\right)\right)^{-1}\right].$$

Trap emission is normally modeled as a thermally-excited process that can be characterized by an activation energy, which is found experimentally. Afterpulsing certainly exhibits such temperature-dependent behavior, but at high electric fields, more than one trap emission mechanism is active. In particular, quantum mechanical tunneling out of the trap state is an important field-dependent process that can affect trap lifetime. Ultimately, a net trap emission rate that reflects both thermal and tunnel processes running in parallel must be obtained to properly characterize $G_{G,R}$ and $G_{gap}$.

Generation by trap-assisted tunneling is a two-step process. In the first step, an electron is promoted out of the valence band and into a mid-gap trap state. This generates a mobile hole in the valence band and an occupied trap. The second step occurs if the trapped electron manages to tunnel from the trap state into the conduction band before it recombines with a hole in the valence band. Tunneling completes the generation of a new electron-hole pair, whereas recombination resets the system to its original state. The generation rate depends upon the density of trap states ($N_t$), their occupancy (determined by their energy $E_t$), and the tunnel lifetime of the trap ($\tau_{tunnel}$):

$$G_{trap}(i) = \frac{N_t(i) \times \left[1 + \exp\left(\frac{E_t - E_F}{k_B T}\right)\right]}{\tau_{tunnel}(i)}.$$

The trap has a tunnel lifetime because it is technically not a stationary state; rather, it is a resonance state with a complex energy (the imaginary component means the wave function evolves with time and ‘leaks’ out of the trap potential). Numerical techniques are required for the tunnel lifetime, which is highly dependent upon the trap depth and applied electric field. This is because these two factors determine the height and width of the potential barrier that separates the trap state from the conduction band. Generally speaking, tunnel currents have a strong exponential dependence on electric field (bias). Cooling can reduce tunnel leakage in two ways: by reducing the equilibrium occupancy of the traps, and through widening of the band gap.

Trap-assisted tunneling complicates the DCR calculation considerably, as the tunnel rate influences both $G_{G,R}$ and $G_{gap}$ through its impact on trap lifetime and steady state occupancy. Accordingly, a model for $\tau_{tunnel}$ must be implemented in the band edge modeler in order to obtain accurate simulations. Unfortunately, we cannot provide a closed-form expression for $\tau_{tunnel}$ here. However, the technique of Choe et al. suggests a relatively simple and rapid computational method. The trap can be modeled as a delta-function potential with a (field-free) bound state energy below the conduction band. To either side of the trap, the conduction band states vary in energy according to the local electric
field strength ($F_i$), establishing a triangle barrier between the trap state and conduction band states of equal and lower energy to one side of the trap. The transfer matrix method outlined by Choe et al. can be used to solve for the tunnel lifetime through the resulting triangle barrier.

2.3.3. Band-to-band tunneling

Generation by band-to-band tunneling is very similar to the second step of trap-assisted tunneling. Two important differences are the density of states that is involved (the density of states at the valence band edge greatly exceeds the possible density of trap states) and the height of the potential barrier (the full band gap). Consequently, band-to-band tunneling rates are only significant in narrow-gap semiconductor alloys, which is why the SACM structure is used to protect narrow-gap NIR absorbers. An analytic expression for the band-to-band tunneling rate in direct gap semiconductors is commonly used that depends upon the electric field strength ($F_i$), the reduced carrier effective mass ($m_r^*$), and the temperature-dependent band gap: \[ G_{\text{band-to-band}}(i) = C_{1,i}(T) F_i^2 \exp \left[ -\frac{C_{2,i}(T)}{F_i} \right] \]

where
\[
C_1(T) = \frac{q^2}{\pi \hbar^2} \sqrt{\frac{2m_r^*}{E_g(T)}} \]
\[
C_2(T) = \frac{\pi^2}{q \hbar} \sqrt{\frac{m_r^*}{2}} E_g(T)^{\frac{3}{2}}
\]

3. CONCLUSION

We have presented a discretized implementation of the DSMT of Saleh et al. that makes it easy to simulate position- and carrier-type-dependent breakdown probabilities for arbitrary bias conditions and electric field profiles, using standard 1-D band edge modelers. Our method makes realistic simulation of dark count rate from spatially distributed sources of primary dark carrier generation feasible, provided that the contribution from trap states can be accurately characterized.

REFERENCES


9. G. Karve, S. Wang, F. Ma, X. Li, and J. C. Campbell, “Origin of dark counts in In$_{0.25}$Ga$_{0.75}$As/In$_{0.75}$Al$_{0.25}$As avalanche photodiodes operated in Geiger mode,” *Applied Physics Letters*, vol. 86, pp. 63 505-1-3 (2005).

10. *SimWindows* is a freeware program by David W. Winston.


