New perspective on passively quenched single photon avalanche diodes: effect of feedback on impact ionization

David A. Ramirez,[1,*] **Majeed M. Hayat,**[1] **Graham J. Rees,**[2] **Xudong Jiang,**[3] **and Mark A. Itzler**[3]

[1]*Department of Electrical and Computer Engineering and the Center for High Technology Materials, University of New Mexico, 1313 Goddard St. SE, Albuquerque, USA*

[2]*Department of Electronic and Electrical Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK*

[3]*Princeton Lightwave Inc., 2555 US Route 130 South, Cranbury, NJ 08512, USA*

[*]*davramir@unm.edu*

Abstract: Single-photon avalanche diodes (SPADs) are primary devices in photon counting systems used in quantum cryptography, time resolved spectroscopy and photon counting optical communication. SPADs convert each photo-generated electron hole pair to a measurable current via an avalanche of impact ionizations. In this paper, a stochastically self-regulating avalanche model for passively quenched SPADs is presented. The model predicts, in qualitative agreement with experiments, three important phenomena that traditional models are unable to predict. These are: (1) an oscillatory behavior of the persistent avalanche current; (2) an exponential (memoryless) decay of the probability density function of the stochastic quenching time of the persistent avalanche current; and (3) a fast collapse of the avalanche current, under strong feedback conditions, preventing the development of a persistent avalanche current. 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.

OCIS codes: (250.1345) Avalanche photodiodes (APDs); (230.5160) Photodetectors; (250.0250) Optoelectronics.

References and links

1. W. P. Risk and D. S. Bethune, "Quantum cryptography," Opt. Photonics News **13**, 26–32 (2002).
2. D. M. Boroson, R. S. Bondurant, and D. V. Murphy, "LDORA: A novel laser communications receiver array architecture," Proc. of SPIE **5338**, 56–64 (2004).

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

3. B. F. Levine, C. G. Bethea, and J. C. Campbell, "1.52 μm room temperature photon counting optical time domain reflectometer," Electron. Lett. **21**, 194–196 (1985).
4. B. F. Aull, A. H. Loomis, D. J. Young, R. M. Heinrichs, B. J. Felton, P. J. Daniels, and D. J. Landers, "Geiger-mode avalanche photodiodes for three-dimensional imaging," Lincoln Lab. J. **13**, 335–350 (2002).
5. M. A. Albota, B. A. Aull, D. G. Fouche, R. M. Heinrichs, D. G. Kocher, R. M. Marino, J. G. Mooney, N. R. Newbury, M. E. OBrien, B. E. Player, B. C. Willard, and J. J. Zayhowski, "Three-dimensional imaging laser radars with Geiger-mode avalanche photodiode arrays," Lincoln Lab. J. **13**, 351–367 (2002).
6. R. H. Haitz, "Model for the electrical behavior of a microplasma," J. Appl. Phys. **35**, 1370–1376 (1964).
7. S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," Appl. Opt. **35**, 1956–1976 (1996).
8. M. A. Itzler, X. Jiang, B. Nyman, and K. Slomkowski, "InP-based Negative Feedback Avalanche Diodes," Proc. of SPIE **7222**, 72221K (2009).
9. 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).
10. 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).
11. M. M. Hayat and G. Dong, "A new approach for computing the bandwidth statistics of avalanche photodiodes," IEEE Trans. Electron Devices **47**, 1273–1279 (2000).
12. M. M. Hayat, G. J. Rees, D. A. Ramirez, and M. A. Itzler, "Statistics of self-quenching time in single photon avalanche diodes," The 21st Annual Meeting of The IEEE Lasers and Electro-Optics Society pp. 203–231 (2008).
13. M. A. Itzler, X. Jiang, B. M. Onat, and K. Slomkowski, "Progress in self-quenching InP-based single photon detectors," Proc. of SPIE **7608**, 760829 (2010).
14. 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," J. Mod. Opt. **54**, 283–304 (2007).
15. 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).
16. B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics* (Wiley, New York, 1991).
17. 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," J. Mod. Opt. **58**, 174–200 (2011).
18. R. J. McIntyre, "Multiplication noise in uniform avalanche photodiodes," IEEE Trans. Electron devices **ED. 13**, 164–168 (1966).
19. L. J. J. Tan, J. S. Ng, C. H. Tan, and J. P. R. David, "Avalanche noise characteristics in submicron InP diodes," IEEE J. Quantum Electron. **44**, 378–382 (2008).
20. 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).
21. D. Shushakov and V. Shubin, "New solid state photomultiplier," Proc. of SPIE **2397**, 544–554 (1995).
22. 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).
23. J. S. Ng, C. H. Tan, J. P. R. David, and G. J. Rees, "A general method for estimating the duration of avalanche multiplication," J. Lightwave Technol. **10**, 1067–1071 (2002).
24. E. T. Whittaker and G. N. Watson, *A course on Modern Analysis* (Cambridge Mathematical Library, 1996).
25. K. B. Athreya and P. Ney, *Branching Processes* (Berlin-Germany: Springer-Verlag, 1972).
26. R. B. Emmons, "Avalanche-photodiode frequency response," J. Appl. Phys. **38**, 3705–3714 (1967).

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

1. Introduction

Recent growth of applications such as quantum cryptography [1], photon-counting optical communication [2], time resolved spectroscopy, time resolved reflectometry [3], quantum imaging [4] and three dimensional laser radar [5] has fueled considerable interest in single photon avalanche diodes (SPADs) [7]. SPADs operate by converting each photogenerated electron hole pair to a large number of carriers via an avalanche of impact ionizations. When the applied bias is above breakdown the number of impact ionizations may increase indefinitely, yielding, in principle, an infinite multiplication factor. However, a ballast resistor may be introduced in series with the SPAD to provide negative feedback and prevent runaway avalanche. The avalanche current then saturates at a level governed by the power supply and the resistance [6, 7]. This current is referred to as the self sustaining or persistent avalanche current. The persistent current may terminate owing to fluctuations in the carrier production at a stochastic time, known as the quenching time [7], after which a new incoming photon may be detected. This mode of operation is referred to as the passive quenching mode and it offers a considerable simplification over the active-quenching mode [7], which requires a complex bias circuitry.

Presently, many photon-counting applications at $1.55-\mu$m are constrained by afterpulsing. Afterpulses are dark counts triggered by the release of trapped carriers from earlier avalanches. These prolong the SPAD's recovery time as they prohibit error-free counting of subsequent incoming photons. Many practitioners [13] in the field believe that a key to breaking the present-day counting rate barrier (10 MHz) for free-running or non-periodic gating at the $1.55-\mu$m wavelength rests in exploiting negative feedback in SPADs to minimize the charge flow after an avalanche trigger and hence reduce afterpulsing. Increasing the achievable free-running photon counting rate to 100 MHz or 1GHz in a simple and cost effective way is essential in applications, e.g., the maximum data transmission rate in photon-counting pulse-position modulation is directly proportional the maximum achievable counting rate. Another example would be photon-counting imaging, in which the flux arrival time at each pixel is random and can involve very short interarrival times requiring high free-running counting rates. To elevate the photon-counting rate, a new generation of fast-quenching SPADs, the negative feedback avalanche diode (NFAD) [8] and the self-quenching and self-recovery avalanche detector [9], have been reported that can potentially transform SPAD technology in providing a solid-state equivalent of the ideal vacuum-tube microchannel plate multipliers. Interestingly, the physical structure of NFADs with a monolithically integrated quench resistor, has revealed a new phenomenon neither observed nor predicted before: the persistent avalanche current in NFADs has an oscillatory behavior. This phenomenon, which is key to understanding the observed memoryless property of the quenching time, cannot be explained by conventional theory, which fails to account for the effect of feedback on the stochastic impact ionization process.

Here we present a stochastically self-regulating avalanche model that fully captures the effect of negative feedback on the stochastic nature of the impact ionization process. It is the first significant expansion beyond the original SPAD device description

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

by Haitz in 1964 [6] and predicts three important behaviors that cannot be addressed by the traditional modeling methods. First, it predicts the oscillatory behavior of the persistent avalanche current. Second, it predicts that the probability density function of the stochastic quenching time of the persistent avalanche current has an exponential (memoryless) decay. Third, under conditions that lead to strong feedback, the stochastic avalanche current can collapse before a persistent avalanche current can be established. All three behaviors are in qualitative agreement with recent experimental studies of NFADs that have not yet been theoretically explained.

2. Limitations of the traditional model for passively quenched SPADs

We have used a stochastic approach to calculate the probability density function (pdf) of the quenching time under a key assumption implicit in the traditional model. As shown in the following the results are unrealistic, demonstrating the inadequacy of this model to capture the effect of feedback on impact ionization. Figure 1a shows the traditional model of a passively quenched SPAD [6, 7]. 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. We shall show that this *constant field assumption* implies the unrealistic consequence that the quenching time, T_q, has memory.

Following the constant field assumption and building upon the recursive technique for avalanche multiplication developed by Hayat *et al.* [10, 11], we have shown (see the Appendix) that the probability that the avalanche current, I self quenches before time t has elapsed is given by

$$F_I(t) \triangleq \mathsf{P}\{T_q \leq t\} \approx \exp(-T/t), \tag{1}$$

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. Equation (1) was first pointed out by the authors in [12] without proof

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

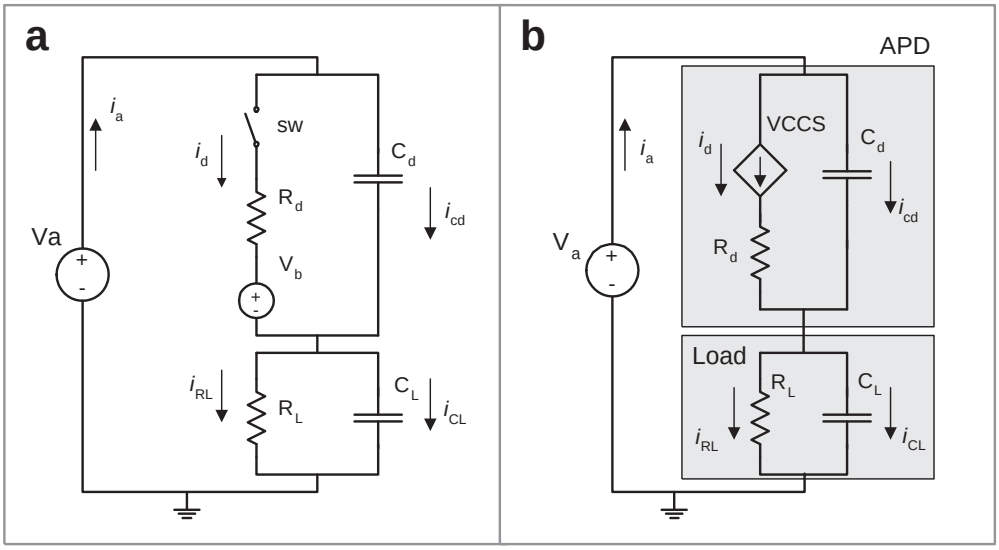

Fig. 1. Models for passively quenched SPADs. (a) Traditional model for a passively quenching SPAD circuit. i_d represents the self sustaining current through the multiplication region of the SPAD; R_d is its equivalent dynamic resistance; C_d is its junction capacitance; R_L is the load resistor and 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. (b) 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.

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

and here we provide its derivation in the Appendix.

The problem with this formula, Eq. (1), arising from the constant field assumption, is that it predicts that the quenching time should have memory, since the form of $F_I(t)$ (not being exponential in t) implies that the probability rate of quenching diminishes in time. More precisely, if we assume that quenching has not occurred by time t, then the probability that quenching should occur between times t and $t+\Delta t$ is $\mathrm{P}\{t \le T_q \le t+\Delta t \mid T_q > t\} = \Delta t \cdot (T/t^2)\exp(-T/t)$, instead of being simply proportional to Δt as in the memoryless case.

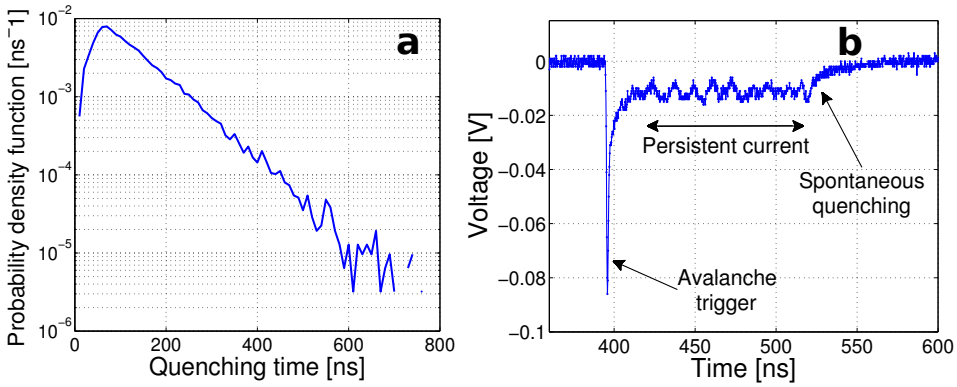

Fig. 2. Experimental results. (a) Measured pdf of the quenching time [13]. The exponential decay of the pdf implies that the quenching time is memoryless. (b) Measured voltage across the SPAD for an excess bias of $V_{ex} \approx 1.7$ V [13]. The current shows oscillatory behavior about the steady state before it quenches spontaneously. The complete structure of the device can be found elsewhere [14].

Indeed, memory is not observed in the quenching process of self quenched SPADs. Figure 2a shows measurements of the pdf of the quenching time for an NFAD SPAD, showing exponential decay, implying 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.* [8, 13] and shown in Fig. 2b.

3. Stochastically self-regulating avalanche model

Figure 1b shows our proposed stochastically self-regulating avalanche model of a passively quenched SPAD [15]. The main difference between this and the traditional model of Fig. 1a is that the switch and voltage generator V_b in Fig. 1a, 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 the Monte Carlo simulator depend

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

on the instantaneous electric field through the junction capacitance, C_d. As the carriers multiply stochastically their resulting current is calculated using Ramo's theorem [16] from the number of carriers inside the multiplication region. Hence, by contrast with the traditional model, our stochastically self-regulating avalanche model captures the effect of feedback on the stochastic evolution of carrier multiplication associated with the persistent avalanche current.

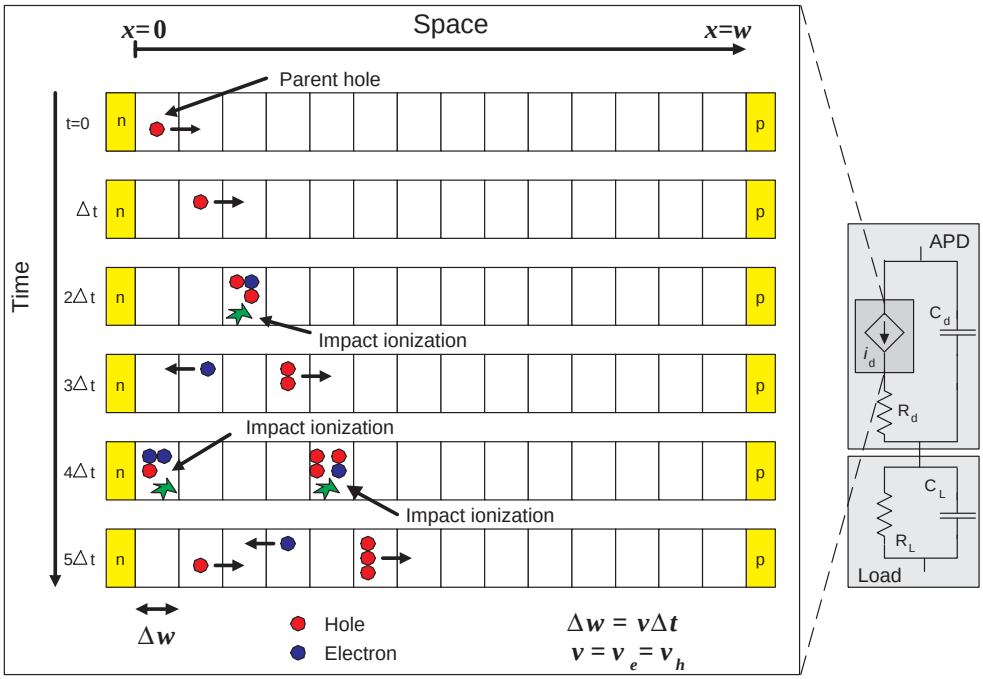

Fig. 3. Monte-Carlo simulator for i_d. The expanded section on the left describes the simulator represented in the circuit on the right by the stochastic VCCS i_d. In the example a hole is injected at the start of the multiplication region, $x = 0$, at time $t = 0$. At time $2\Delta t$ the first impact ionization occurs and as a result one hole and one electron are created in the multiplication region. For simplicity we have assumed that electrons and holes have the same drift velocity, v, i.e., $v = v_e = v_h$.

3.1. Monte Carlo simulator of avalanche multiplication under dynamic electric field

In Fig. 3 we illustrate 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 our simulations

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

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 we have 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 $j = 1,2,3,\ldots,M$. We assume that holes (electrons) move in the positive (negative) direction of x. In addition, we employ 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 we define $X_e(t_j, x_k)$ and $X_h(t_j, x_k)$ 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 we consider SPADs with thick multiplication regions ($> 1\mu$m), which are preferred for Geiger mode operation [17]. 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:

$$\begin{aligned} X_e(t_{j+1}, x_k) &= X_e(t_j, x_{k+1}) + b\Big(X_e(t_j, x_{k+1}), \alpha(E_{C_d}(t))\Delta w\Big) \\ &\quad + b\Big(X_h(t_j, x_{k-1}), \beta(E_{C_d}(t))\Delta w\Big) \end{aligned} \tag{2}$$

and

$$\begin{aligned} X_h(t_{j+1}, x_k) &= X_h(t_j, x_{k-1}) + b\Big(X_h(t_j, x_{k-1}), \beta(E_{C_d}(t))\Delta w\Big) \\ &\quad + b\Big(X_e(t_j, x_{k+1}), \alpha(E_{C_d}(t))\Delta w\Big). \end{aligned} \tag{3}$$

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 Eqs. (2) and (3).

To trigger an avalanche the multiplication region is reverse biased above breakdown and a carrier is injected at the start of the multiplication region. Equations (2) and (3) are implemented at every time increment and samples of the required binomial random variables are generated. Figure 3 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

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

instantaneous stochastic current $i_d(t_j)$ is calculated using Ramo's theorem:

$$i_d(t_j) = \frac{qv}{w} \sum_{k=1}^{L} \Big(X_e(t_j,k) + X_h(t_j,k) \Big). \tag{4}$$

All other currents and voltages in Fig. 1 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.

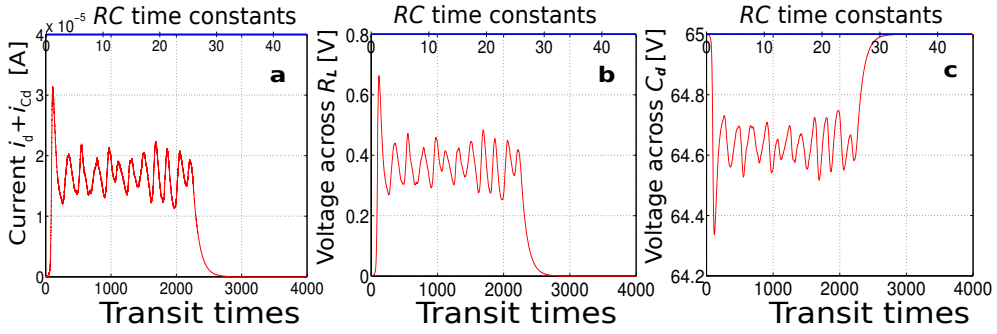

Fig. 4. Calculated current-voltage evolution of SPAD after an avalanche trigger. (a) Calculated avalanche current, $i_a = i_d + i_{C_d}$, (b) voltage across the feedback resistor, R_L, and (c) voltage across the SPAD, V_{C_d} as a function of time for an excess bias voltage $V_{ex} \approx 0.39$ V and 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 $I_{ss} \approx 18\ \mu$A, 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.

4. Results

We next apply our stochastically self-regulating avalanche model to simulate the unique attributes of the new generation of self quenched SPADs. In particular, we are interested in predicting the statistics of the quenching time and the observed oscillatory behavior of the persistent current. We have simulated a passively quenched InP SPAD 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

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

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 [18]

$$M = \frac{1-k}{\exp(-(1-k)\alpha w) - k}. \tag{5}$$

By using the electric-field dependent expressions for the electron and hole ionization coefficients for InP [19], the breakdown voltage is found to be $V_b = 64.61$ V.

4.1. Circuit behavior after an avalanche trigger

Figure 4 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, $R_L(C_L + C_d)$. We have 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.

The current and voltages fluctuate around the steady state values predicted by the traditional model; the persistent current fluctuates around $I_{ss} \approx V_{ex}/R_L$ [6, 7], 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 [11, 20]. 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, where the current through the diode oscillates about $I_{ss} \approx 18\ \mu$A, 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. 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 quenching occurs after about 2340 transit times.

To better illustrate the cycles in the oscillatory behavior described above we plot in Fig. 5 the calculated voltage across the junction capacitor, V_{C_d} (red curve) together with the current, i_d (blue curve). In particular, we mark four successive stages of behavior, from the onset of the avalanche until the spontaneous quenching of the persistent current. (For clarity the curve i_d was truncated and its first peak is not shown.) A key point

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

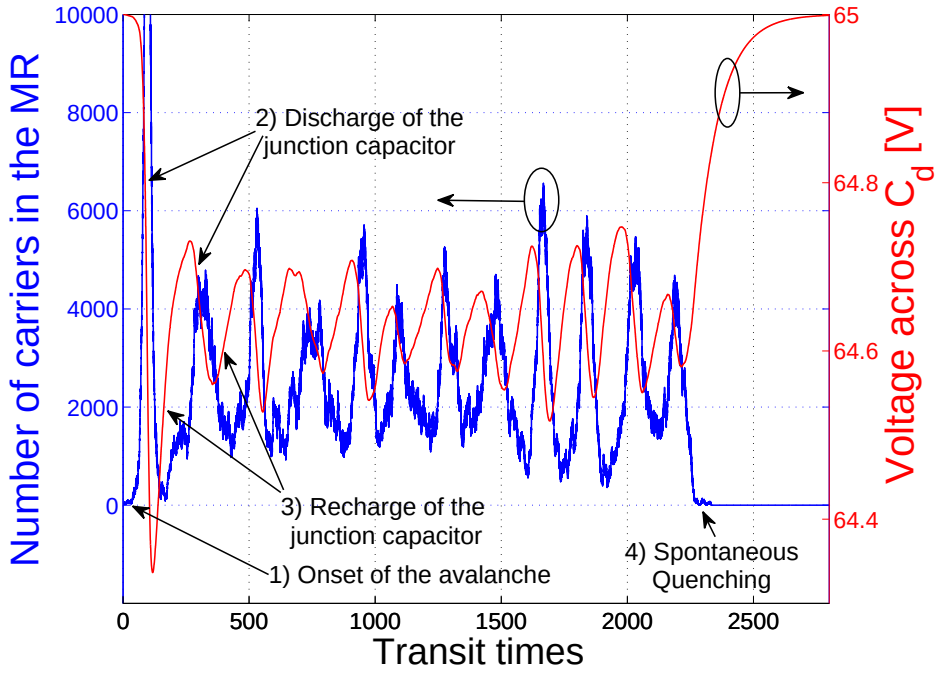

Fig. 5. Timing relationship between the voltage across the junction capacitance and the number of carriers in the multiplication region. The red curve shows the voltage across the junction capacitor V_{C_d} and the blue curve shows the current i_d calculated by the Monte-Carlo simulator. For clarity, 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.

that we have noted in our 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 ionizing carriers is at a minimum. This observation is critical in understanding the exponential decay of the pdf of the quenching time, which is discussed later.

4.2. Quenching behavior

In our simulations we have considered an observation window of 8000 transit times ($\sim$ 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. We have found that within the considered observation window the probability of spontaneous quenching increases as the current I_{ss} decreases. This

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

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 6a 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.* [21, 22] and Zhao *et al.* [9] in devices where the feedback was provided by means of a charge-accumulation effect due to a potential barrier outside the multiplication region. The work of Shushakov *et al.* [21, 22] 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.

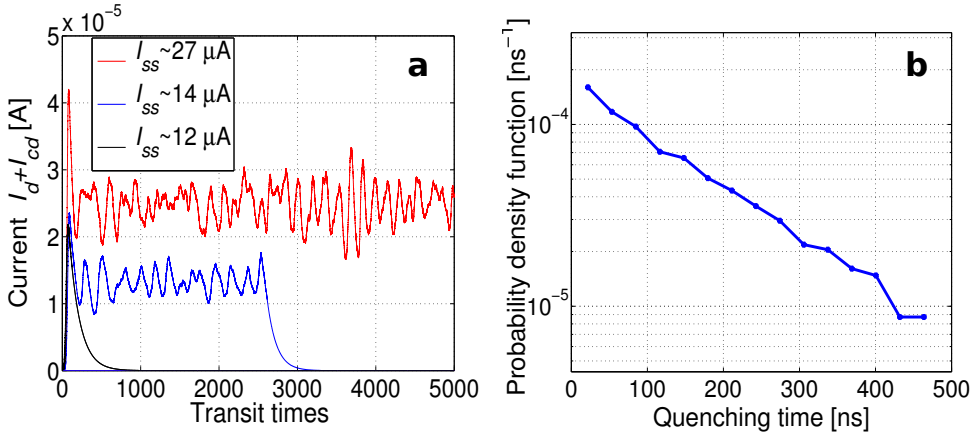

Fig. 6. Quenching characteristics of the simulated passively quenched SPAD. (a) Quenching behavior of the simulated passively quenched SPAD for different values of the current I_{ss}. As the current I_{ss} decreases the avalanche current spontaneously quenches sooner, on average. (b) Calculated probability density function of quenching time, T_q.

The quenching behavior described above was reproduced consistently for excess voltages below 0.8 V. However, for higher excess bias voltages we did not observe the behavior shown in the blue curve of Fig. 6a, in which we have a period of persistent current flow followed by spontaneous quenching. 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 resulting stronger feedback causes the current i_d to overshoot as it falls, making it difficult for the system to execute even

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

a single period of persistent oscillating current.

4.3. 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. 6b, revealing that the decay of the tail of the pdf is exponential, implying, in turn, that the quenching process is memoryless. This observation is consistent with the measurements performed on NFAD devices shown in Fig. 2a.

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. We have learned from our simulations that quenching invariably occurs only during the recharge cycle of the junction capacitance, as shown in Fig. 5, and that the probability that quenching occurs during the discharge stage of the capacitance is negligible. We now make two key observations. 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.

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

5. Conclusions

In conclusion, we have developed a model to calculate the response of a passively quenched SPAD, reverse biased above breakdown. Our 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 [6, 7], 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, we have 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. Our 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. To the best of our knowledge, the proposed stochastically self-regulating avalanche model is the only one capable of predicting the oscillatory behavior of the persistent current and the statistics of the quenching time. Our 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.

6. Appendix

Consider an electron (hole) generated at position z in a multiplication region with uniform electric field, which spans $0 < z < w$, and assume that it will ionize for the first time at a distance ζ downstream from its generation point and at a time t later with probability density function $h_{e(h)}(\zeta,t)$. A carrier injected at z in a multiplication region will then give rise to an avalanche current which terminates before time t has elapsed with probability $F_e(z,t)$ for an injected electron, and $F_h(z,t)$ for an injected hole. Along the lines of the recursive multiplication theory [11, 20, 23], these probabilities are given by

$$F_e(z,t) = R_e(z,t)Q_e(z,t) + \int_0^{w-z}\int_0^t h_e\big(\zeta,\tau\big)F_e^2(z+\zeta,t-\tau)F_h(z+\zeta,t-\tau\big)d\tau\, d\zeta \quad (6)$$

and

$$F_h(z,t) = R_h(z,t)Q_h(z,t) + \int_0^{z}\int_0^t h_h\big(\zeta,\tau\big)F_h^2(z-\zeta,t-\tau)F_e(z-\zeta,t-\tau\big)d\tau\, d\zeta. \quad (7)$$

Here $R_e(z,t) = 1 - \int_0^{w-z}\int_0^t h_e(\zeta,\tau)d\tau\, d\zeta$ and $R_h(z,t) = 1 - \int_0^{z}\int_0^t h_h(\zeta,\tau)d\tau\, d\zeta$ are the probabilities that the injected electron and hole avoid ionizing within the multiplication region before time t and $Q_{e,h}(z,t)$ are the probabilities that these carriers drift and diffuse out of the multiplication region before this time.

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

Equations (6) and (7) hold above, below and precisely at breakdown and may be solved numerically to follow the temporal evolution of $F_{e(h)}(z,t)$. The behavior below breakdown has been studied numerically by Hayat and Dong [11] and by Ng *et al.* [23] and the Malthusian behavior was pointed out by Hayat *et al.* [10] and examined further by Groves *et al.* [20]. A brief exposition of the behavior at breakdown was first reported in [12] without proof and we provide it here.

To obtain the result in Eq. (1) it is necessary to determine the asymptotic behavior of the avalanche duration distribution function at breakdown. To do this we write the avalanche duration probabilities in terms of the $f_{e,h}(z,t) = 1 - F_{e,h}(z,t)$, since at breakdown we expect $f_{e,h}(z,t) \to 0$ at long times. Without loss of generality we can write Eqs. (6) and (7) in terms of $f_{e,h}(z,t)$ as

$$f_e(z,t) = R_e(z,t)\big(1 - Q_e(z,t)\big) + \int_0^{w-z}\int_0^t h_e(\zeta,\tau)\big(2f_e + f_h - f_e^2 - 2f_e f_h + f_e^2 f_h\big)\,d\tau\,d\zeta \tag{8}$$

and

$$f_h(z,t) = R_h(z,t)\big(1 - Q_h(z,t)\big) + \int_0^{z}\int_0^t h_h(\zeta,\tau)\big(2f_h + f_e - f_h^2 - 2f_h f_e + f_h^2 f_e\big)\,d\tau\,d\zeta. \tag{9}$$

The arguments of f_e and f_h on the right-hand side of Eq. (8) are all $(z+\zeta, t-\tau)$ and in Eq. (9) are all $(z-\zeta, t-\tau)$.

At times t earlier than a transit time the inhomogeneous "source" terms, $R_{e,h}(z,t)\big(1 - Q_{e,h}(z,t)\big)$ in Eqs. (8) and (9) are non-zero. The effects of carrier diffusion and the gradual collapse of the electric field at the edges of the multiplication region in a real device will ensure that these terms fall to zero with time in a continuous and smooth manner, so that the only discontinuity is at time $t = 0$, when the system is switched on. The simple model used here, where diffusion is absent and the field falls abruptly to zero at the edges of the multiplication region, may be considered a limiting case of the real system. Since the integral terms on the right-hand side of Eqs. (8) and (9) are well behaved and we expect $f_{e,h}(z,t)$ to approach zero at long times and to be analytic away from $t = 0$, we attempt a Laurent expansion [24] of the solutions in powers of inverse time

$$f_{e,h}(z,t) = \sum_{n=1}^{\infty} f_{e,h;n}(z) \times (t)^{-n}. \tag{10}$$

Terms of only negative power in t are included because Malthusian behavior [25, 20], of the form $\exp(-\gamma t)$, and requiring positive powers, would lead to an inconsistently finite and constant breakdown probability, since γ approaches zero at breakdown [10]. In fact numerical simulations of Eqs. (6) and (7) (not shown here) confirm that the coefficient of the Malthusian exponential approaches zero at breakdown, consistent with $n > 0$.

Inserting the expansion in Eq. (10) into Eqs. (8) and (9) and equating coefficients of equal powers of t we obtain a coupled series of equations for the coefficients, $f_{e,h;n}(z)$. At sufficiently long times the lowest order terms, in $1/t$ will dominate. The equations for the coefficients, $f_{e,h;n}(z)$ are homogeneous (since the inhomogeneous source terms

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

involving the R and Q vanish at times longer than a transit time) and so these quantities are given only to within constant factors (although the factors for $f_{e;1}(z)$ and $f_{h;1}(z)$ are related). The values of these factors are determined by the initial conditions contained in the source terms, which do not appear in this asymptotic analysis. However, Eqs. (8) and (9) are nonlinear and the equations for the higher order coefficients, $f_{e,h;n}(z)$, with $n = 2,3,\ldots$, involve the lower order coefficients so that these can be determined successively after the $f_{e,h;1}(z)$ are known.

We now determine the coefficients $f_{e,h;1}(z)$. For simplicity, to derive expressions for the coefficients, $f_{e,h;1}(z)$ of the leading terms of the Laurent expansion in Eq. (10) for $f_{e,h}(z)$, we use the local, constant velocity model. The ionization event pdfs then become

$$h_e(\zeta,\tau) = \alpha\exp(-\alpha\zeta)\delta(\tau-\zeta/v_e) \text{ and } h_h(\zeta,\tau) = \beta\exp(-\beta\zeta)\delta(\tau-\zeta/v_h). \quad (11)$$

Using this model it is convenient to write Eqs. (8) and (9) in terms of dimensionless length variables, $s = z/w$ and $p = \zeta/w$ and dimensionless time variable, $r = t/\tau_0$, where $\tau_0 = (\tau_e+\tau_h)/2$ and $\tau_{e,h} = w/v_{e,h}$ are the electron and hole transit times across the multiplication region. Writing $\varphi(s,r) \equiv f_{e,h}(z,t)$, Eqs. (8) and (9) become

$$\begin{aligned}\varphi_e(s,r) &= \exp(a(s-1))\theta(1-s-r/\rho_e)\\ &+ a\int_0^{1-s}\exp(-ap)\big(2\varphi_e+\varphi_h-\varphi_e^2-2\varphi_e\varphi_h+\varphi_e^2\varphi_h\big)dp \end{aligned} \quad (12)$$

and

$$\begin{aligned}\varphi_h(s,r) &= \exp(b(s-1))\theta(1-s-r/\rho_b)+b\int_0^{1-s}\exp(-bp)\big(2\varphi_h\\ &+ \varphi_e-\varphi_h^2-2\varphi_h\varphi_e+\varphi_h^2\varphi_e\big)dp. \end{aligned} \quad (13)$$

Here $\theta(x)$ is the unit step function, $\rho_{e,h} = \tau_{e,h}/\tau_0$, the $\varphi_{e,h}$ on the right-hand side (RHS) of Eq. (12) are all to be understood as $\varphi_{e,h}(s+p, r-\rho_e p)$ and on the RHS of Eq. (13) as $\varphi_{e,h}(s-p, r-\rho_h p)$. The solutions $\varphi_{e,h}$ are evidently determined by the values of the dimensionless parameters, $a = \alpha w$, $b = \beta w$ and $\rho_{e,h}$, which are all of order unity (indeed, when $\alpha = \beta$ and $v_e = v_h$ then they are precisely unity, since we are considering a device biased at breakdown). The $\varphi_{e,h}(s,r)$ themselves are dimensionless since they represent probabilities.

Equations for the coefficients $f_{e,h;1}(z) \equiv \varphi_{e,h}(s)$ can now be found by discarding all terms nonlinear in the $\varphi_{e,h}$ on the RHS of Eq. (12) (since they do not contribute to the leading terms in the Laurent expansion) and writing $\varphi_{e,h}(s,r) \sim \varphi_{e,h}(s)/r$ on the left and $\varphi_{e,h}(s\pm p, r\pm\rho_{e,h}p) \sim \varphi_{e,h}(s\pm p)/r$ on the right, since at long times $r \gg \rho_{e,h}p$. The inhomogeneous terms also disappear at long times and we find

$$\varphi_e(s) = a\int_0^{1-s}\exp(-ap)\big(2\varphi_e(s+p)+\varphi_h(s+p)\big)dp \quad (14)$$

and

$$\varphi_h(s) = b\int_0^{s}\exp(-bp)\big(2\varphi_h(s-p)+\varphi_h(s-p)\big)dp. \quad (15)$$

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

Changing the integration variable in Eq. (14) to $u = s + p$ and in Eq. (15) to $u = s - p$ these equations become

$$\exp(-as)\varphi_e(s) = a\int_0^1 \exp(-au)\big(2\varphi_e(u) + \varphi_h(u)\big)du \tag{16}$$

and

$$\exp(bs)\varphi_h(s) = b\int_0^s \exp(bu)\big(2\varphi_h(u) + \varphi_e(u)\big)du\,. \tag{17}$$

By differentiating these equations with respect to s we can find differential equations for the $\varphi_{e,h}(s)$. With the boundary conditions, $\varphi_e(1) = 0 = \varphi_h(0)$ these yield solutions which we write as

$$\varphi_e(s) = \frac{C}{d}\big(\exp(d(1-s)) - 1\big) \text{ and } \varphi_h(s) = \frac{C}{d}\big(1 - \exp(-ds)\big). \tag{18}$$

Here $d = a - b$ and C is a dimensionless constant, determined by the values of a, b and via Eq. (12). Numerical solutions of Eqs. (6) and (7) confirm the $1/t$ behavior of the $f_{e,h}(z,t)$ and show that C is of the order of unity. We are indebted to C. H. Tan for these results. The d in the denominator of Eq. (18) is included to preserve good behavior as $d \to 0$. The analysis also yields the breakdown condition, $b\exp(a) = a\exp(b)$, confirming that these arguments are valid only at breakdown threshold and not above or below.

Finally, writing $\delta = \alpha - \beta$ we can deduce the form of the leading terms of Laurent expansions. Thus, since $f_{e;1}(z)/t = j_e(s)/r$, and $f_{h;1}(z)/t = j_h(s)/r$ it follows that

$$f_{e;1}(z) = \frac{C\tau_0}{\delta w}\big(\exp(\delta(w-z)) - 1\big) \text{ and } f_{h;1}(z) = \frac{C\tau_0}{\delta w}\big(1 - \exp(-\delta z)\big). \tag{19}$$

The dynamical equations for the avalanche carrier densities in a uniform multiplication region are given, e.g., by Emmons [26]. In terms of the electron and hole concentrations per unit length, $n(z)$ and $p(z)$. They are given by

$$n(z) = \frac{I\big(1 - \exp(\delta z)\big)}{qv_e\big(\exp(\delta z) - 1\big)} \text{ and } p(z) = \frac{I\big(\exp(\delta z) - \exp(\delta w)\big)}{qv_h\big(\exp(\delta w) - 1\big)}, \tag{20}$$

where I is the current carried by these distributions.

We now determine the probability distribution function of the quenching time in a passively quenched SPAD under constant electric field at breakdown. To calculate the statistics of the duration of such avalanche pulses we observe that the mean avalanche current is generated by electrons and holes distributed throughout the multiplication region. In the local model we can regard these as primary carriers, each generating its own individual avalanche current, all of which flow in parallel to generate a total mean current, I. For the avalanche pulse to quench each of these individual avalanche currents must terminate independently. The probability that this happens before time t elapses is given by $F_I(t) = \prod_i F_{e,h}(z_i,t)$, where the product is over all electrons and holes, situated at z_i in the multiplication region, and $F_{e,h}(z_i,t)$ is the probability that

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012

an electron (hole) injected at z_i will give rise to an avalanche pulse which terminates before time t has elapsed. When the SPAD is biased precisely at breakdown (i.e., prior to the avalanche current collapsing) $F_{e,h}(z_i,t) = 1 - f_{e,h}(z_i)/t$. Interestingly, we note that this asymptotic behavior is different from those corresponding to below or above breakdown, for which the asymptotic behavior is exponential [11, 20]. Thus, at times long compared with the carrier transit times we find

$$\ln\big(F_I(t)\big) = \sum_i \ln\big(1 - f_{e,h}(z_i)/t\big) \approx -\frac{1}{t}\sum_i f_{e,h}(z_i). \tag{21}$$

If the electron and hole distributions per unit length in the multiplication region are $n(z)$ and $p(z)$ then Eq. (21) becomes

$$\ln\big(F_I(t)\big) \approx -\frac{1}{t}\left(\int_0^w n(z)\, f_{e;1}(z)dz + \int_0^w f_{h;1}\, p(z)dz\right). \tag{22}$$

By using the expressions for the $f_{e,h;1}(z)$ given by Eq. (19), and the expressions for the $n(z)$ and $p(z)$ given by Eq. (20), respectively, we arrive at Eq. (1).

Acknowledgments

We thank CH Tan for numerical solutions of Eqs. (6) and (7). This work was supported in part by NASA under Grant NNG06LA04C.

#156323 - $15.00 USD Received 11 Oct 2011; revised 6 Dec 2011; accepted 19 Dec 2011; published 10 Jan 2012