Stochastic Quenching Mechanisms and a Scaling Law for Single Photon Avalanche Diodes

Akito Inoue

***Abstract*— A comprehensive scaling law for single photon avalanche diodes (SPADs) is presented through stochastic analyses of quenching mechanisms using a Monte Carlo method. By simulating random impact ionization events for individual carriers, two distinct quenching mechanisms are identified: successful quenching (SQ) and unsuccessful quenching (UQ). SQ occurs when quenching is achieved after the initial pulse of avalanche multiplication (AM), mainly attributed to the minimum average carrier number within a multiplication region (MR). In contrast, UQ involves prolonged and repetitive pulses, caused by stochastic fluctuations around the equilibrium carrier number. This study has derived an analytical expression for the probability of quenching failure ($1–P_Q$) as functions of the quenching resistance and the capacitance of the MR. This analytical expression exhibits a good agreement with the simulation results. Moreover, analytical formulas for the threshold quenching resistance and the dead time have been derived as a function of the desired P_Q value. Notably, the tradeoff relationship between the dead time and the standard deviation of the voltage swing is elucidated, leading to the scaling limitation. Additionally, avalanche triggering probability (ATP), breakdown voltage, and the average voltage swing are revealed to be scale-invariant. Based on these aforementioned observations, the comprehensive scaling law is established.**

***Index Terms*— Avalanche breakdown, avalanche photodiodes, CMOS image sensors (CISs), Monte Carlo simulation, quenching, quenching probability, scaling law, single photon avalanche diodes (SPADs).**

I. Introduction

IN RECENT years, significant progress has been made in the development of single photon avalanche diode (SPAD)-based CMOS image sensors (CISs) [1], [2], [3]. These sensors have been extensively utilized in various applications, such as time-of-flight ranging sensors for autonomous driving [4] and mobile phones [5], wide dynamic range imagers for surveillance cameras [6], and high sensitivity imagers [7]. The driving force behind these advancements is the miniaturization of pixels, leading to pixel sizes below 10 μm and resolutions reaching to several megapixels [6], [8], [9]. One crucial challenge in pixel miniaturization is quenching, which is the mechanism that stops avalanche multiplication (AM) immediately after its occurrence. Through the appropriate design of the external quenching resistance, large voltage swing, low power consumption, and short dead time can be achieved [1], [10].

To date, the two types of quenching models based on the carrier dynamics inside the multiplication region (MR) have been reported: 1) the deterministic models that numerically calculate the average behavior of voltage, current, and carrier number based on the carrier continuity equations [11], [12], [13] and 2) the stochastic models that simulate the coordinates of individual carriers as outcomes of successive random impact ionization events [14], [15], [16]. Despite these individual efforts, a unified explanation is still lacking. Additionally, the previous study on SPAD scaling mainly focuses on the size of the MR to enhance photon detection probability and reduce dark count rate [17], without considering stochastic carrier dynamics during quenching. Therefore, it is important for successful design of SPAD-CISs with sub-10-μm pixels to understand the interplay among stochastic carrier dynamics, quenching, and SPAD scaling. This understanding enables low quenching errors and short dead time within the limited pixel area.

This article employs Monte Carlo simulations to expand the deterministic analyses conducted in our previous studies [12], [13] into the stochastic framework. By investigating the time evolutions of voltage across the MR and the carrier number within it, two distinct quenching mechanisms have been identified: one attributed to the minimum average carrier number and the other arising from the fluctuations around the equilibrium carrier number within the MR. Probabilities associated with the occurrence of these two mechanisms are analyzed in detail as functions of the MR capacitance and the quenching resistance. This analysis results in the derivation of analytical formulas for the probability of quenching failure, the threshold value of quenching resistance, and the dead time. The dead time is subject to the tradeoff relationship with the standard deviation of the voltage swing, which imposes the scaling limitation. Avalanche triggering probability (ATP), breakdown voltage (V_{BD}), and average voltage swing are shown to be scale-invariant. These findings are summarized to propose the comprehensive scaling law for SPADs.

Manuscript received 15 August 2023; revised 4 November 2023; accepted 28 November 2023. Date of publication 12 December 2023; date of current version 2 January 2024. The review of this article was arranged by Editor Y. Xu.

The author is with Panasonic Industry Company Ltd., Osaka 571-8501, Japan (e-mail: inoue.akito@jp.panasonic.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TED.2023.3338596.

Digital Object Identifier 10.1109/TED.2023.3338596

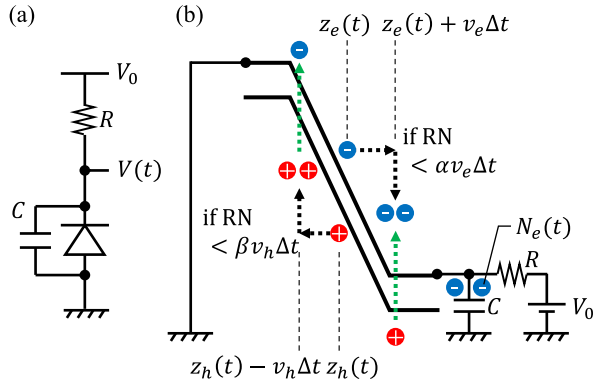

Fig. 1. (a) Circuit diagram of a SPAD. (b) Equivalent band diagram of a SPAD MR with associated circuit elements. During the simulation, the positions of all electrons [$z_e(t)$] and holes [$z_h(t)$] are computed at each time step. The impact ionizations are simulated by comparing the probabilities of impact ionization events with RN.

II. Formulation of a Monte Carlo Quenching Model

A 1-D SPAD with the p-i-n structure is considered, where the i-region is the MR. In the simulations, electric field within the MR is uniform and the width of the MR is constant. The circuit model is depicted in Fig. 1(a), where the quenching resistor, R, is connected to the cathode of the SPAD. The total capacitance, including the capacitance of the MR and stray capacitances of the cathode node, is illustrated by the parallel capacitor C. If the parasitic capacitances are negligible, C is proportional to the MR's area S as

$$C = \varepsilon \frac{S}{W} \tag{1}$$

where ε represents the dielectric constant of the SPAD. The bias voltage $V_0 = V_{\mathrm{BD}} + V_{\mathrm{ex}}$ is a variable above the breakdown voltage (V_{BD}). The excess voltage (V_{ex}) is much smaller than V_{BD} ($V_{\mathrm{ex}} \ll V_{\mathrm{BD}}$). The anode of the SPAD is connected to the ground. In Fig. 1(b), the band diagram of the MR with the circuit elements is depicted. The carriers transport with their saturation velocities and the positions of all carriers are tracked for each time step Δt. Within Δt, impact ionizations are simulated by comparing random numbers (RN) with the probabilities of impact ionization events: $\alpha v_e \Delta t$ for electrons and $\beta v_h \Delta t$ for holes. Once the electrons drift out of the MR, they are accumulated at the cathode node of the SPAD. The number of stored electrons, denoted as $N_e(t)$, increases as the number of electrons drifting out of the MR within Δt [$\Delta n_e(t)$]. Simultaneously, it diminishes as electrons discharge toward the voltage source through R. Consequently, $N_e(t)$ is calculated by the following recurrence equation:

$$N_e(t+\Delta t) = N_e(t) + \Delta n_e(t) - N_e(t)\Delta t/RC. \tag{2}$$

The voltage across the SPAD [$V(t)$] is reduced by $N_e(t)$ as

$$V(t) = V_0 - \frac{q}{C} N_e(t). \tag{3}$$

The electric field is then calculated as $E(t) = V(t)/W$. It is noted that (3) ignores the space charge effect, since the primary focus of this study is to elucidate a scaling law for SPADs.

TABLE I
Definitions and Values of the Physical Quantities

Symbol	Meanings	Values
W	Width of the MR	0.80 μm
v_e	Saturation velocity of electron [18]	1.02×10^7 cm/s
v_h	Saturation velocity of hole [18]	8.31×10^6 cm/s
$\alpha(E)$	Impact ionization rates	$\alpha_0 \exp(-a/E)$
$\beta(E)$		$\beta_0 \exp(-b/E)$
α_0	Coefficients of impact ionization ratio [19]	3.80×10^6 cm^{-1}
β_0		2.25×10^7 cm^{-1}
a		1.75×10^6 V/cm
b		3.26×10^6 V/cm

This iterative process continues until all electrons and holes disappear from the MR.

III. Results of Monte Carlo Simulations

A. Stochastic Quenching Mechanisms

Time evolutions of voltage across the MR [$V(t)$] and the number of electrons within the MR [$n_e(t)$] are presented in Fig. 2(a)–(c) and (d)–(f), respectively. The initial condition is set as the scenario, in which an electron–hole pair is generated at the anode edge of the MR assuming single photon detection. The parameters employed for this analysis pertain to a silicon SPAD and are summarized in Table I. Through simulations utilizing these parameters, V_{BD} is determined to be 29.55 V. The initial voltage V_0 is set to 30 V ($V_{\mathrm{ex}} = 0.45$ V) and the capacitance values are $C = 30$ fF [see Fig. 2(a) and (d)], 6 fF [see Fig. 2(b) and (e)], and 0.1 fF [see Fig. 2(c) and (f)]. It should be noted that the typical capacitance values for the SPAD-CISs fall within the range of tens of fF when employing a pixel pitch of 10 μm, while they diminish to less than 1 fF with a 1 μm pitch. At their lower extreme, the capacitance value reaches approximately 0.1 fF. To align with (8) in Section IV-A, the values of R are determined as a function of C. The simulations are conducted for 20 000 steps, each with a time increment Δt of 0.2 ps. Results are classified based on whether AM is triggered or not, and within the AM-triggered cases, whether quenching is successful or unsuccessful. In Fig. 2, the blue curves represent successful quenching (SQ) cases, while the red and green curves depict unsuccessful quenching (UQ) cases. Here, "unsuccessful" implies a longer quenching duration compared with that of SQ.

In the case of SQ, a sharp voltage drop followed by a recharge is observed in Fig. 2(a)–(c). The voltage swing, $\Delta V_{\mathrm{Q}} = V_0 - \min[V(t)]$, is approximately $2V_{\mathrm{ex}}$. Concurrently, the number of electrons exponentially increases and then exponentially decreases, as shown in Fig. 2(d)–(f). With the aid of carrier number fluctuations, the electron number eventually reaches 0 (highlighted by the blue hatches). As indicated by vertical dashed lines, the electron number reaches its maximum at t_{BD}, coinciding with $V(t_{\mathrm{BD}}) = V_{\mathrm{BD}}$. These observations align with the findings of the deterministic models [11], [12], [13]. Thus, the randomness has an insignificant impact on the time waveforms in the case of SQ.

In the case of UQ, the time evolutions of the voltage and the electron number exhibit the similar pattern to SQ

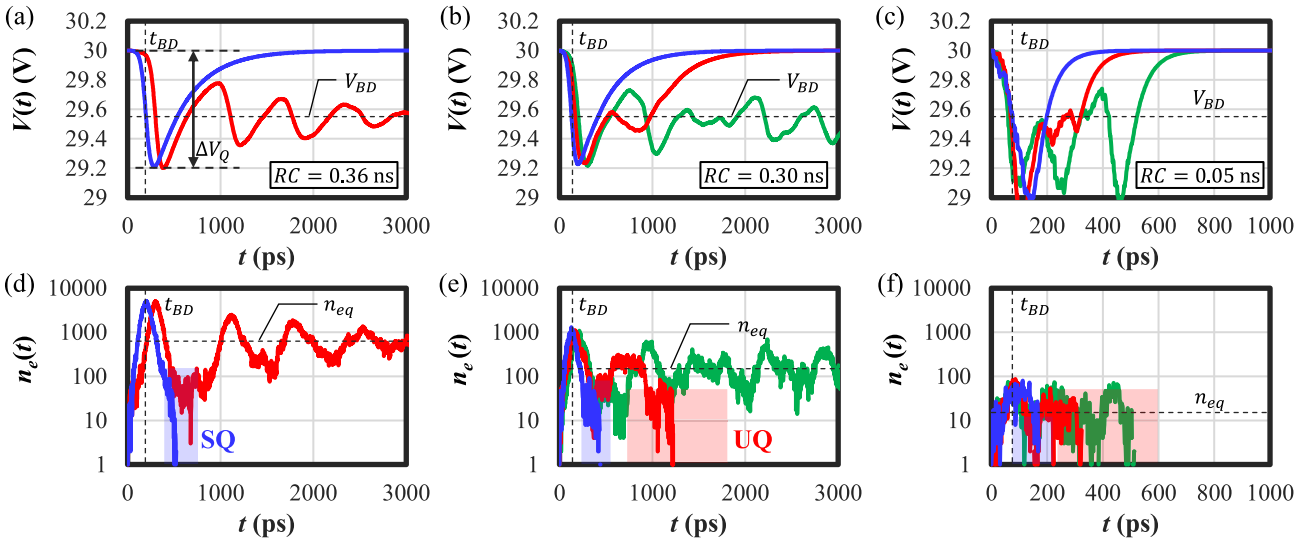

Fig. 2. Time evolutions of (a)–(c) voltage across the MR and (d)–(f) number of electrons inside the MR. The MR capacitance and quenching resistance values are (a) and (d) $C = 30$ fF and $R = 12$ kΩ ($RC = 0.36$ ns), (b) and (e) $C = 6$ fF and $R = 50$ kΩ ($RC = 0.30$ ns), and (c) and (f) $C = 0.1$ fF and $R = 500$ kΩ ($RC = 0.05$ ns). The blue curves represent the results for SQ, while the red and green curves illustrate the results for UQ. The arrow in (a) indicates the voltage swing ΔV_Q. The horizontal dashed lines in (a)–(c) and (d)–(f) correspond to V_{BD} and n_{eq}, respectively. The vertical dashed lines indicate t_{BD}. The blue and red hatches present the time intervals, where SQ and UQ occur. Note that the horizontal axes in (c) and (f) are limited up to 1 ns.

until the middle of the recharging phase. However, subsequent oscillations or fluctuations emerge. In the case where $C = 30$ fF and $R = 12$ kΩ ($RC = 0.36$ ns), as shown in Fig. 2(a) and (d), $V(t)$ and $n_e(t)$ oscillate around $V_{BD} = 29.55$ V and the equilibrium carrier number n_{eq}, respectively, with decreasing envelopes. These oscillations correspond to the results of the deterministic model shown in [12, Fig. 3]. When $C = 6$ fF and $R = 50$ kΩ ($RC = 0.30$ ns), as depicted in Fig. 2(b) and (e), n_{eq} decreases and both $V(t)$ and $n_e(t)$ exhibit irregular fluctuations. Once $n_e(t)$ reaches 0 because of these fluctuations, no impact ionizations occur afterward, resulting in quenching (highlighted by the red hatches). This type of quenching is UQ, characterized by a longer duration than SQ. In this manner, SQ and UQ are characterized by different quenching times as discussed in detail in Sections III-B and IV-A. For $C = 0.1$ fF and $R = 500$ kΩ ($RC = 0.05$ ns), as illustrated in Fig. 2(c) and (f), n_{eq} becomes even smaller, and the amplitudes of the initial AM pulse and the subsequent fluctuations become comparable, and, as a result, UQ occurs earlier than the $C = 6$ fF case. In this manner, the fluctuations due to the randomness give rise to UQ, leading to the differences in time waveforms when compared to the results of the deterministic models [11], [12], [13]. The fluctuations are enhanced as C decreases or as the SPAD area diminishes. Even in the presence of large fluctuations, $V(t)$ drops below V_{BD}, indicating that V_{BD} still acts as an equilibrium point.

B. Probability Distributions of Quenching Time

Fig. 3 illustrates the cumulative probability distributions of T_{quench}, denoted as $P(T < T_{quench})$. T_{quench} represents the quenching time at which both electrons and holes vanish from the MR. It should be noted that the horizontal axes of Fig. 3 are inverted, with the right end of the figures at 0 ns and the left end at 2 ns.

For $C = 30$ fF [see Fig. 3(a)], $P(T < T_{quench})$ exhibits distinct slopes for UQ (the red hatches) and SQ (the blue hatches), indicating different quenching mechanisms. The plateau-like UQ slopes indicate that the time required for UQ is longer than 2000 ps because of the weak fluctuations compared with large n_{eq}. As R increases, the probability of UQ decreases, while the total of the probabilities of UQ and SQ, i.e., ATP remains constant. In the case of infinite R (the purple line), all trial results are classified as SQ. Nonavalanche (NA) (green hatch) slopes appearing near $T = 0$ is clearly separated from the SQ slopes by the plateau regimes. In the case of $C = 6$ fF [see Fig. 3(b)], the NA, SQ, and UQ slopes are still distinguishable, but the UQ slopes become nonnegligible. As R increases, the UQ slopes become steeper, indicating that the time required for UQ occurrence is reduced. When $C = 1$ fF [see Fig. 3(c)], the UQ slopes get even steeper and the inflection points between the SQ and UQ slopes become unclear. For $C = 0.1$ fF [see Fig. 3(d)], these slopes cannot be distinguished. This can be attributed to the fact that the magnitude of fluctuations is comparable to that of AM, as observed in Fig. 2(f). It is noteworthy that the distributions of NA remain almost unchanged regardless of R and C. As the influence of R and C comes through the voltage across the MR, the carrier behaviors are not affected by the voltage change until AM is triggered.

The quenching probability is defined as the ratio of the probability of SQ (P_{SQ}) to ATP as

$$P_Q = \frac{P_{SQ}}{\text{ATP}} = \frac{P_{SQ}}{P_{SQ} + P_{UQ}}. \tag{4}$$

In Fig. 4(a), the failure probability of quenching $(1 - P_Q)$ is plotted as a function of RC. The results for $C = 1$, 6, and 30 fF are plotted by red, green, and blue dots with corresponding lines, respectively. It can be observed that $(1 - P_Q)$ decreases

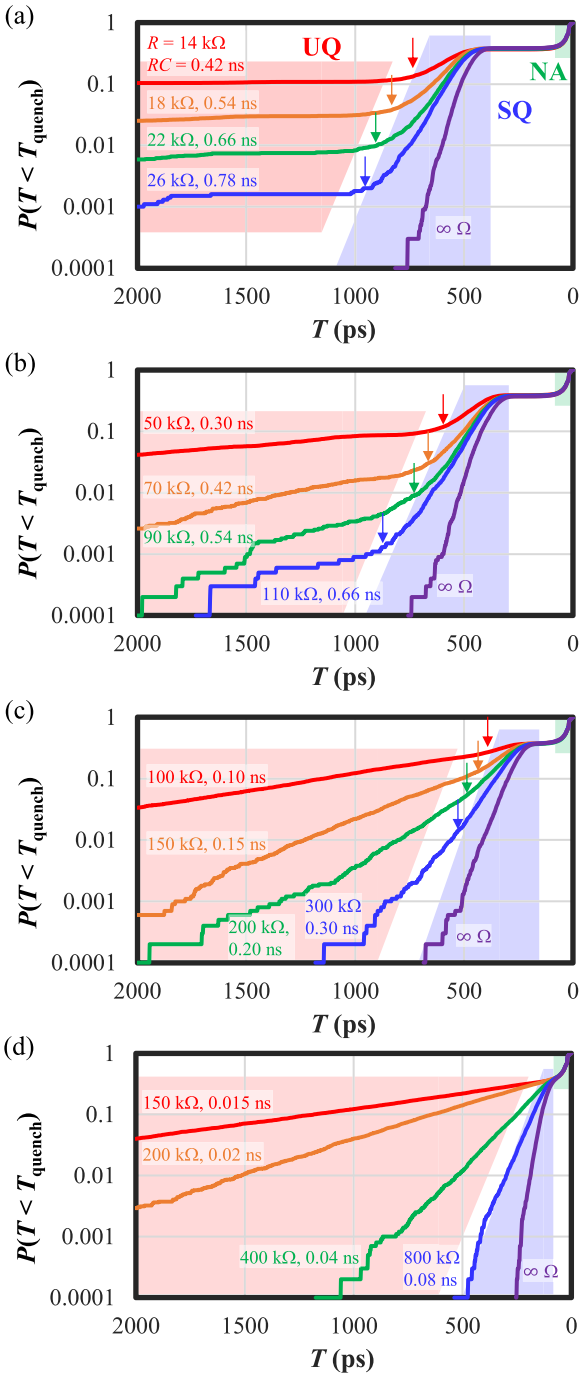

Fig. 3. Cumulative probability distributions $P(T < T_{\text{quench}})$ for (a) $C = 30$ fF, (b) $C = 6$ fF, (c) $C = 1$ fF, and (d) $C = 0.1$ fF. The red, blue, and green hatches represent the slopes for UQ, SQ, and NA components, respectively. The arrows indicate the inflection points between the SQ and UQ slopes except for (d) where the SQ and UQ slopes cannot be distinguished.

exponentially with RC, and for the same RC value, $(1 - P_Q)$ is reduced further by decreasing C. Notably, as depicted in Fig. 4(b), all data points of $(1 - P_Q)/C$ align on the same fitting line calculated by (6) (explained in Section IV-A in detail). It is important to note that [14] and [20] contain the experimental results regarding the time waveforms of SQ and UQ, as well as the probability functions for quenching time. This study distinguishes itself from [14] and [20] by focusing

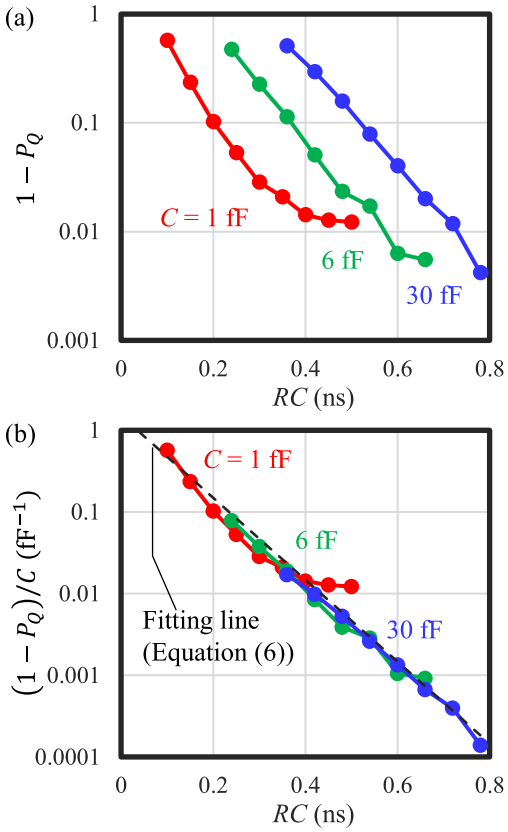

Fig. 4. (a) $1 - P_Q$ and (b) $(1 - P_Q)/C$ as a function of RC. Red, green, and blue dots with corresponding lines represent the results for $C = 1$, 6, and 30 fF, respectively. The dashed line in (b) indicates the fitting line obtained by (6). The all data points align along this line.

on elucidating the distinct mechanisms for SQ and UQ and introducing the scaling law for SPADs.

C. Probability Distributions of Voltage Swings

Fig. 5 represents the probability distributions of ΔV_Q, for initial voltages of $V_{ex} = 0.45$, 0.95, and 1.45 V ($V_0 = 30.0$, 30.5, and 31.0 V). When a voltage above V_{BD} is applied, the distributions of ΔV_Q are binarized into two components: one near $\Delta V_Q = 0$ (the NA component) and the other centered around $\Delta V_Q = 2V_{ex}$ (the AM-triggered component). The peak positions of the AM-triggered component remain constant regardless of C. In contrast, the distribution width of ΔV_Q increases as C decreases. For $C = 30$ and 6 fF [see Fig. 5(a) and (b)], the two distributions are clearly separated for all V_{ex} values. However, the overlaps between the NA and AM-triggered components are observed up to $V_{ex} = 0.45$ V for $C = 1$ fF and $V_{ex} = 0.95$ V for $C = 0.1$ fF, as shown in Fig. 5(c) and (d), respectively.

Fig. 6(a)–(c) presents ATP, the average of ΔV_Q ($E[\Delta V_Q]$), and the standard deviation of ΔV_Q ($\sigma[\Delta V_Q]$), respectively. To distinguish between the NA component and the AM-triggered component, the threshold voltage swing is set to $\Delta V_{Q,th} = 0.5$ V, which is close to the minimum for $P(\Delta V_Q)$ with $V_{ex} = 0.45$ V. Both ATP and $E[\Delta V_Q]$ are independent of the MR capacitance. Moreover, $E[\Delta V_Q]$ is almost $2V_{ex}$ for all V_{ex} and C values, consistent with [12], [13], and [21]. As $V_0 = V_{BD} + V_{ex}$, both V_{BD} and V_{ex} are

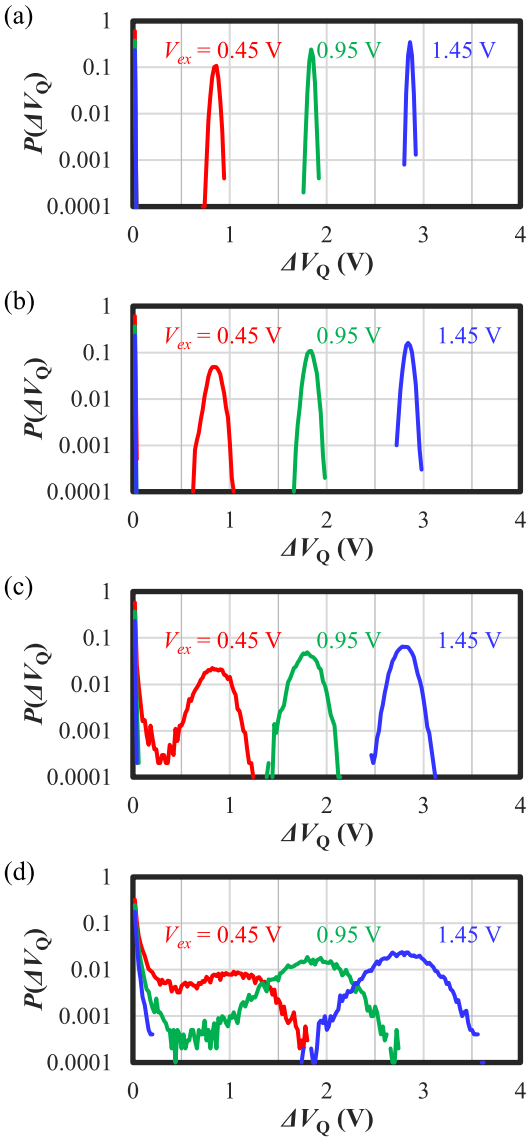

Fig. 5. Probability distributions of ΔV_Q for (a) $C = 30$ fF and $R = 24$ kΩ, (b) $C = 6$ fF and $R = 100$ kΩ, (c) $C = 1$ fF and $R = 400$ kΩ, and (d) $C = 0.1$ fF and $R = 2.8$ MΩ. To minimize the influence of UQ, sufficiently large resistance values are chosen. The initial voltages are $V_{ex} = 0.45$ V (red), 0.95 V (green), and 1.45 V (blue).

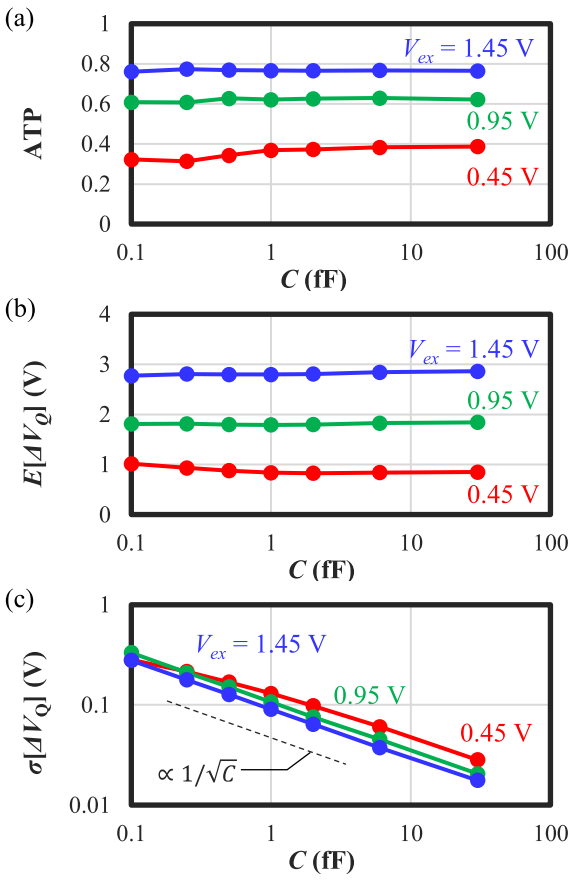

Fig. 6. (a) ATP, (b) average voltage swing $E[\Delta V_Q]$, and (c) standard deviation of the voltage swing $\sigma[\Delta V_Q]$ as a function of C. Red, green, and blue dots with corresponding lines are results for $V_{ex} = 0.45$, 0.95, and 1.45 V, respectively. The dashed line in (c) represents a slope proportional to $1/\sqrt{C}$, indicating the observed trend.

also independent of C. In contrast, as indicated by the dashed line in Fig. 6(c), $\sigma[\Delta V_Q]$ decreases proportional to $1/\sqrt{C}$, due to the fluctuations enhanced with decreasing C. In addition, the decrease in $\sigma[\Delta V_Q]$ with increasing V_{ex} aligns with the experimental results demonstrated in [21, Fig. 7].

IV. Discussion

A. Quenching Conditions for SQ and UQ

In both the cases of SQ and UQ, the following two conditions must be achieved for quenching: 1) the "average" carrier number in the MR, as determined by the deterministic model, becomes insufficient and 2) carriers within the MR vanish ($n_e(t) = 0$) due to stochastic fluctuations. As the amplitude of the fluctuations is determined by the impact ionization rates and the average carrier number within the MR, the variation in the average carrier numbers between SQ and UQ gives rise to the difference in the probability distributions of the quenching time. To fulfill the condition 1), the upper thresholds for the average carrier numbers must be defined for both UQ ($n_{th,UQ}$) and SQ ($n_{th,SQ}$).

For the UQ case, the corresponding average carrier number is n_{eq}, which can be calculated using the following equation:

$$n_{eq} = \frac{V_{ex}}{qR} \cdot \frac{W}{\eta v_e} \tag{5}$$

where η represents a parameter that characterizes the bias in electron generation locations within the MR toward the cathode side. By fitting n_{eq} to the time waveforms shown in Fig. 2(d)–(f), $\eta = 3$ is obtained. Equation (5) indicates that the probability of UQ is solely dependent on R. To enhance the probability of UQ events, R should exceed 50 kΩ, as illustrated in Fig. 3, where nonzero slopes become noticeable. By using this resistance value, the threshold is calculated as $n_{th,UQ} = 300$ below which UQ can occur.

On the other hand, SQ occurs immediately after the first AM pulse, when the average electron number reaches its minimum value ($\min[n_e(t)]$), according to [12, eq. (32)]. The quenching condition is met when $\min[n_e(t)]$ falls below $n_{th,SQ}$, and the failure probability of SQ is reduced as $\min(n_e(t))$ decreases. Assuming that $(1 - P_Q)$ can be expressed as the ratio of

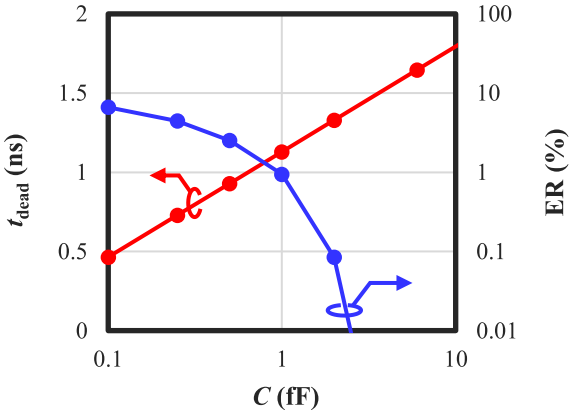

Fig. 7. Relationship between the MR capacitance and two key parameters: t_{dead} represented by the red dots on the left axis, and ER represented by the blue dots on the right axis. The data indicate a tradeoff correlation between these parameters.

$\min(n_e(t))$ to $n_{th,SQ}$, it can be formulated as

$$1 - P_Q = \frac{\min(n_e(t))}{n_{th,SQ}} \sim \frac{1}{n_{th,SQ}} \frac{CV_{ex}}{q\tau_Q} \frac{W}{\eta v_e} \exp\left(-\frac{\ln(2)}{2\tau_Q} RC\right) \tag{6}$$

where τ_Q is the time constant of the voltage swing, defined by [12, eq. (13)] as

$$\tau_Q = \left(\alpha'(E_{BD}) + \beta'(E_{BD})\right) \cdot v_e \, V_{ex}/W \tag{7}$$

where $\alpha'(E_{BD}) = d\alpha(E_{BD})/dE$ and $\beta'(E_{BD}) = d\beta(E_{BD})/dE$. Equation (6) reproduces the results shown in Fig. 4(b) well. This supports the validity of the aforementioned assumption regarding to (6). In addition, $n_{th,SQ}$ equals to 170, which is close proximity to, and slightly smaller than $n_{th,UQ}$. This discrepancy arises from the shorter time allowance for SQ compared to UQ. SQ is required to take place immediately after the initial AM pulse; otherwise, the carrier number increases again.

In light of (6), two essential parameters can be derived: the threshold quenching resistance and the dead time. These parameters are interconnected with the desired quenching probability, $P_{Q,des}$, as follows:

$$R_Q = \frac{2\tau_Q}{C\ln(2)} \ln\left(\frac{CV_{ex}}{q\tau_Q} \frac{W}{\eta v_e} \frac{1}{n_{th}(1 - P_{Q,des})}\right) \tag{8}$$

$$t_{dead} = R_Q C = \frac{2\tau_Q}{\ln(2)} \ln\left(\frac{CV_{ex}}{q\tau_Q} \frac{W}{\eta v_e} \frac{1}{n_{th}(1 - P_{Q,des})}\right). \tag{9}$$

Equations (6), (8), and (9) extend the previously derived formulas presented in [12].

B. A Scaling Law for SPADs

A scaling law for SPADs can be established through the aforementioned equations and discussions. Table II presents the scaling law based on the aforementioned stochastic analyses. The capacitance scales proportional to the pixel area according to (1), indicating that when S is multiplied by a factor of k, C is also multiplied by k. V_{BD}, $E[\Delta V_Q]$, and ATP remain unchanged under scaling, with $E[\Delta V_Q]$ approximately equal to $2V_{ex}$. $\sigma[\Delta V_Q]$ and t_{dead} are proportional to $1/\sqrt{k}$ and $\ln(k)$, respectively, exhibiting a tradeoff relationship. $\sigma[\Delta V_Q]$ is associated with the detection error rate (ER). Fig. 7 illustrates t_{dead} and ER as a function of C, highlighting the above tradeoff. In the calculation of t_{dead}, a threshold value of $(1 - P_{Q,des}) = 0.01$ is used. By referring (9), t_{dead} changes logarithmically with C and k as $\ln(k\gamma)/\ln(\gamma)$, where $\gamma = (CV_{ex}W)/(q\tau_Q\eta v_e n_{th}(1 - P_{Q,des}))$. The capacitance value must be within a few tens of femtofarads to achieve t_{dead} with a few nanoseconds. The ER is obtained by calculating the probability that ΔV_Q becomes less than $\Delta V_{Q,th}$ even when the AM is triggered. It is obtained by fitting the AM-triggered components depicted in Fig. 5 using the Gaussian distribution with the mean $E[\Delta V_Q]$ and the standard deviation $\sigma[\Delta V_Q]$ as follows:

$$\mathrm{ER} = \frac{1}{\sqrt{2\pi}\sigma[\Delta V_Q]} \int_{-\infty}^{\Delta V_{Q,th}} \exp\left(-\frac{(\Delta V_Q - E[\Delta V_Q])^2}{2\sigma[\Delta V_Q]^2}\right) d\Delta V_Q. \tag{10}$$

The ER is highly sensitive to C, varying by orders of magnitude. When considering the application to SPAD-CISs, the allowable ER is at most 1%, corresponding to a lower threshold of 1 fF. These results give the scaling limitation of the pixel area, as calculated by (1), to be within the range of 10–100 μm^2 for an about 1 μm width of MR.

TABLE II
SCALING LAW FOR SPADs

Parameters	Symbols	Scaling factor
SPAD pixel area	S	k
SPAD capacitance	C	k
Breakdown voltage	V_{BD}	1
Voltage swing	ΔV_Q	1
Avalanche triggering probability	ATP	1
Dead time	t_{dead}	$\frac{\ln(k\gamma)}{\ln(\gamma)}$
Standard deviation of voltage drop	$\sigma[\Delta V_Q]$	$\frac{1}{\sqrt{k}}$

V. CONCLUSION

The stochastic carrier dynamics in SPADs have been investigated using Monte Carlo simulations. Time evolutions of voltage and carrier number have revealed two distinct quenching mechanisms: SQ and UQ (see Fig. 2). SQ is attributed to the minimum value of the average carrier number within the MR after the initial AM pulse, while UQ arises from fluctuations around the equilibrium carrier number (see Section IV-A). The cumulative probability distributions

of T_{quench} further confirm the different mechanisms of SQ and UQ, as indicated by their different slopes (see Fig. 3). The magnitude of fluctuations is enhanced by increasing the quenching resistance and decreasing the MR capacitance, making the distinctions of SQ and UQ unclear. The quenching failure probability, $(1 - P_{\text{Q}})$, exhibits a good agreement between the analytical expression and the simulation results [see (6) and Fig. 4]. This analytical expression leads to the derivations of fundamental formulas for a threshold quenching resistance (8) and a dead time (9). The analysis of voltage swing has identified scale-invariant parameters, including ATP, V_{BD}, and the average of ΔV_{Q} (see Figs. 5 and 6). The scaling limitation is derived from the tradeoff relationship between the dead time and the standard deviation of ΔV_{Q} (see Fig. 7). Based on these findings, the comprehensive scaling law for SPADs has been established (see Table II), providing valuable design guidelines to optimize pixel capacitance and pixel area. This study significantly contributes to the advancement of high-resolution SPAD-CISs with fine pixels, expanding the possibilities for implementing cutting-edge imaging applications.

Acknowledgment

The author would like to thank Dr. Yutaka Hirose and Dr. Shinzo Koyama for valuable discussions.

References

[1] E. Charbon and M. W. Fishburn, "Monolithic single-photon avalanche diodes: SPADs," in *Single-Photon Imaging*. Heidelberg, Germany: Springer, 2011, pp. 123–158, doi: 10.1007/978-3-642-18443-7.

[2] E. Charbon, C. Bruschini, and M.-J. Lee, "3D-stacked CMOS SPAD image sensors: Technology and applications," in *Proc. 25th IEEE Int. Conf. Electron., Circuits Syst. (ICECS)*, Bordeaux, France, Dec. 2018, pp. 1–4.

[3] R. K. Henderson et al., "A 256 × 256 40 nm/90 nm CMOS 3D-stacked 120 dB dynamic-range reconfigurable time-resolved SPAD imager," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2019, pp. 106–108.

[4] O. Kumagai et al., "A 189 × 600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2021, pp. 96–97.

[5] K. Ito et al., "A back illuminated 10 μm SPAD pixel array comprising full trench isolation and cu-cu bonding with over 14% PDE at 940 nm," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2020, p. 16.

[6] K. Morimoto et al., "3.2 megapixel 3D-stacked charge focusing SPAD for low-light imaging and depth sensing," in *IEDM Tech. Dig.*, San Francisco, CA, USA, Dec. 2021, pp. 450–453.

[7] J. Ogi et al., "A 124-dB dynamic-range SPAD photon-counting image sensor using subframe sampling and extrapolating photon count," *IEEE J. Solid-State Circuits*, vol. 56, no. 11, pp. 3220–3227, Nov. 2021.

[8] K. Morimoto et al., "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications," *Optica*, vol. 7, no. 4, p. 346, Apr. 2020, doi: 10.1364/optica.386574.

[9] T. Okino et al., "A 1200 × 900 6 μm 450 fps geiger-mode vertical avalanche photodiodes CMOS image sensor for a 250 m time-of-flight ranging system using direct-indirect-mixed frame synthesis with configurable-depth-resolution down to 10 cm," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2020, pp. 96–97.

[10] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, pp. 1956–1976, 1996.

[11] E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Transient single-photon avalanche diode operation, minority carrier effects, and bipolar latch up," *IEEE Trans. Electron Devices*, vol. 60, no. 3, pp. 1188–1194, Mar. 2013.

[12] A. Inoue and Y. Hirose, "Nonlinear carrier dynamics in a single photon avalanche diode: Stability, bifurcation, and quenching condition," *IEEE Trans. Electron Devices*, vol. 68, no. 12, pp. 6222–6227, Dec. 2021.

[13] A. Inoue, S. Koyama, and Y. Hirose, "A dynamic nonlinear impedance model of a single photon avalanche diode," *IEEE Trans. Electron Devices*, vol. 70, no. 6, pp. 3160–3165, May 2023.

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

[15] P. Windischhofer and W. Riegler, "Passive quenching, signal shapes, and space charge effects in SPADs and SiPMs," *Nucl. Instrum. Methods Phys. Res. A, Accel. Spectrom. Detect. Assoc. Equip.*, vol. 1045, Jan. 2023, Art. no. 167627.

[16] T. Cazimajou et al., "Quenching statistics of silicon single photon avalanche diodes," *IEEE J. Electron Devices Soc.*, vol. 9, pp. 1098–1102, 2021.

[17] K. Morimoto and E. Charbon, "A scaling law for SPAD pixel miniaturization," *Sensors*, vol. 21, no. 10, p. 3447, May 2021.

[18] R. S. Müller and T. I. Kamins, *Device Electronics for Integrated Circuits*, 3rd ed. Hoboken, NJ, USA: Wiley, 2003, p. 33.

[19] S. M. Sze and G. Gibbons, "Avalanche breakdown voltages of abrupt and linearly graded p-n junctions in Ge, Si, GaAs, and GaP," *Appl. Phys. Lett.*, vol. 8, no. 5, pp. 111–113, Mar. 1966.

[20] M. A. Itzler, X. Jiang, B. Nyman, and K. Slomkowski, "InP-based negative feedback avalanche diodes," *Proc. SPIE*, vol. 7222, Jan. 2009, Art. no. 72221K.

[21] A. Inoue, T. Okino, S. Koyama, and Y. Hirose, "Modeling and analysis of capacitive relaxation quenching in a single photon avalanche diode (SPAD) applied to a CMOS image sensor," *Sensors*, vol. 20, no. 10, p. 3007, May 2020.