Journal of Modern Optics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tmop20

Advances in InGaAsP-based avalanche diode single photon detectors

Mark A. Itzler [a], Xudong Jiang [a], Mark Entwistle [a], Krystyna Slomkowski [a], Alberto Tosi [b], Fabio Acerbi [b], Franco Zappa [b] & Sergio Cova [b]

[a] Princeton Lightwave Inc., 2555 US Route 130 S., Cranbury, NJ 08512, USA

[b] Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan I-20133, Italy

Available online: 31 Jan 2011

To cite this article: Mark A. Itzler, Xudong Jiang, Mark Entwistle, Krystyna Slomkowski, Alberto Tosi, Fabio Acerbi, Franco Zappa & Sergio Cova (2011): Advances in InGaAsP-based avalanche diode single photon detectors, Journal of Modern Optics, 58:3-4, 174-200

To link to this article: http://dx.doi.org/10.1080/09500340.2010.547262

Journal of Modern Optics
Vol. 58, Nos. 3–4, 10–20 February 2011, 174–200

Taylor & Francis
Taylor & Francis Group

TOPICAL REVIEW

Advances in InGaAsP-based avalanche diode single photon detectors

Mark A. Itzler[a]*, Xudong Jiang[a], Mark Entwistle[a], Krystyna Slomkowski[a], Alberto Tosi[b], Fabio Acerbi[b], Franco Zappa[b] and Sergio Cova[b]

[a]*Princeton Lightwave Inc., 2555 US Route 130 S., Cranbury, NJ 08512, USA;* [b]*Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan I-20133, Italy*

(*Received 19 August 2010; final version received 30 November 2010*)

In this Topical Review, we survey the state-of-the-art of single photon detectors based on avalanche diodes fabricated in the InGaAsP materials system for photon counting at near infrared wavelengths in the range from 0.9–1.6 μm. The fundamental trade-off between photon detection efficiency and dark count rate can now be managed with performance that adequately serves many applications, with low dark count rates of ~1 kHz having been demonstrated at photon detection efficiencies of 20% for 25 μm diameter fiber-coupled devices with thermoelectric cooling. Timing jitter of less than 50 ps has been achieved, although device uniformity is shown to be essential in obtaining good jitter performance. Progress is also reported towards resolving the limitations imposed on photon counting rate by afterpulsing, with at least 50 MHz repetition frequencies demonstrated for 1 ns gated operation with afterpulsing limited to the range of 1–5%. We also present a discussion of future trends and challenges related to these devices organized according to the hierarchy of materials properties, device design concepts, signal processing and electronic circuitry, and multiplexing concepts. Whereas the materials properties of these devices may pose significant challenges for the foreseeable future, there has been considerable progress in device concepts and circuit solutions towards the present imperatives for higher counting rates and simpler device operation.

Keywords: single photon avalanche diode; SPAD; photon counting; Geiger mode; InGaAs/InP; near infrared

1. Introduction

Single photon detectors based on avalanche diode structures are frequently the best choice for applications requiring not only high performance but also high reliability, ease of implementation, and scalability. Over the past decade, significant progress has been achieved for many properties of InP/InGaAs single photon avalanche diodes (SPADs). For instance, there has been notable improvement in the fundamental trade-off between photon detection efficiency (PDE) and dark count rate (DCR), and high precision timing jitter has been demonstrated for these detectors. There has also been impressive scaling of these detectors to large format arrays.

However, important limitations remain. Many photon counting applications now demand higher counting rates, and SPADs have not kept pace with these rapidly evolving requirements. Complex back-end electronics have been necessary to achieve high performance SPAD operation, and this need for sophisticated circuitry presents significant challenges to low-cost scaling, especially in the deployment of large-area SPAD-based sensors. Current trends in the field have emphasized two strategies for circumventing at least some aspects of these present limitations: specifically, the extraction of enhanced device performance using improved hybrid back-end electronic circuitry, and new monolithic chip-level concepts for obtaining improved performance through avalanche self-quenching. In this topical review, we survey the state-of-the-art in present InGaAsP-based SPAD capabilities as well as the trends that have emerged from many groups working towards demonstrating next-generation performance for these detectors.

Most fundamentally, progress in the DCR versus PDE trade-off for 1.55 μm photon detection [1] has been considerable. For PDE ~20%, fiber-coupled 25 μm diameter devices routinely exhibit DCR values of a few kHz, and 'hero' devices demonstrate that it is possible to achieve sub-kHz DCR performance at temperatures readily accessible using thermoelectric coolers [2]. Higher PDE values in the range of 40–50% can be obtained with acceptable DCR. High precision timing jitter has also been demonstrated, with 100 ps jitter found for typical operating conditions, and less than 50 ps obtained [1,3] for sufficiently high excess bias.

*Corresponding author. Email: mitzler@princetonlightwave.com

ISSN 0950–0340 print/ISSN 1362–3044 online

DOI: 10.1080/09500340.2010.547262
http://www.informaworld.com

Low jitter performance requires very good uniformity of response across the detector area, and we illustrate the impact of non-uniform response on timing properties. The gradual maturing of this detector platform has also made possible an evolution to successively larger formats of InP/InGaAs SPAD arrays [4,5].

More constraining performance limitations have generally been found with respect to maximum count rate. Although intrinsic SPAD response is reasonably high bandwidth, with avalanche build-up occurring in well under 1 ns, afterpulsing effects have often limited counting rates to the range of 1 to 10 MHz. Given the materials challenges inherent in reducing the density of defects that give rise to the carrier trapping and detrapping events that cause afterpulsing, a more prevalent strategy has been to reduce the potential number of carriers that can be trapped by limiting the charge flow per avalanche event. We show for the first time that the use of 1 ns gates and a matched delay line transient cancellation technique can be extended to gate repetition rates of at least 50 MHz with acceptable afterpulsing. Additionally, low-parasitic hybrid integration approaches for passive quenching/active reset circuits [6,7] can achieve 65 MHz counting with afterpulsing limited to ~1% using ~2 ns gates. Very intriguing results have also been shown using novel electronics circuitry such as a self-differencing architecture [8,9] and sinusoidal gating [10–12] for which much higher 1–2 GHz gating frequencies have been attained with 1–5% afterpulsing. The fact that this variety of higher frequency counting results has been achieved using SPADs of the same pedigree fabricated by the authors indicates the significant differentiation that can be offered by specific circuit implementations for SPAD operation.

As another path towards reducing avalanche charge flow, there have been efforts to implement rapid self-quenching using monolithic chip-level solutions. These designs are essentially based on passive quenching but with very low parasitic effects. We have pursued self-quenching solutions with integrated resistive feedback [13–15], and other approaches have been reported using epitaxially embedded hetero-barriers [16,17] and distributed feedback mechanisms [18].

As a means of organizing these and other current trends in InP-based SPAD development, we describe a hierarchy consisting of four levels of technology that influence SPAD performance: (i) materials properties, (ii) device design concepts, (iii) signal processing and circuits, and (iv) higher-level multiplexing approaches. This hierarchy is useful to the extent that it helps to define the constraints imposed on researchers working at one level by limitations at lower levels. For instance, device designers are limited by the state-of-the-art in materials quality; and designers of circuit-based solutions are limited by the state-of-the-art in devices. A survey of recent trends in this field is presented in the context of this hierarchy.

The remainder of this review is organized as follows. In Section 2, we present a brief survey of the applications of SPADs operating at near infrared wavelengths near 1.5 μm and the role of these applications in driving recent trends in SPAD development. We summarize InP-based SPAD design concepts in Section 3, and in Section 4, we describe the state-of-the-art in fundamental SPAD performance metrics, including the trade-off between dark count rate and photon detection efficiency, timing jitter performance, and afterpulsing properties. Section 5 is organized according to the hierarchy of SPAD technologies (i.e. materials, devices, circuits and signals, and multiplexing), and we describe recent advances and promising prospects – as well as challenges that may remain for the foreseeable future – in the context of this hierarchy. Finally, in Section 6, we present a concluding discussion of this work.

2. Dominant applications and consequent trends in NIR SPAD development

2.1. Basics of Geiger-mode operation

Avalanche photodiodes (APDs) based on the InGaAsP materials system are a mature device technology for detecting radiation in the near infrared (NIR) wavelength range from 0.9–1.6 μm. In this device, an incident photon creates an initial electron–hole pair by photoexcitation in a semiconductor absorber with an appropriate bandgap. Transport of at least one of these photoexcited carriers to a region of strong internal electric field leads to impact ionization, through which an energetic carrier ionizes an additional electron–hole pair by promoting an electron from the valence band to the conduction band. These secondary carriers can in turn ionize further pairs, and repeated ionization events lead to an avalanche of charge triggered by each photoexcited carrier. For optical receivers in which the photodetector is followed by an amplifier, this internal gain process offers great utility for increasing receiver sensitivity in cases where the amplifier noise would be the limiting noise source in the absence of internal detector gain.

The gain of an APD increases with stronger internal electric field but remains finite up to a threshold breakdown electric field E_b established by a corresponding externally applied breakdown voltage V_b. For applied voltages less than V_b, the output photocurrent of the APD is proportional to the input optical intensity. For this reason, device operation below the

breakdown voltage is referred to as 'linear mode'. In contrast, for materials in which both electrons and holes exhibit impact ionization, applied voltages greater than V_b lead to a finite probability that a single carrier will trigger a self-sustaining avalanche characterized by divergent (i.e. infinite) gain. This rapid generation of easily measured avalanche currents triggered by just a single photoexcited carrier allows for the efficient detection of single photons. Devices designed for operation above breakdown are therefore referred to as single photon avalanche diodes (SPADs) and are quite distinct from the more prevalent linear mode APDs. By virtue of the similarity of SPAD carrier avalanching to the behavior of Geiger–Muller detectors used in the detection of radioactive particles, biasing above breakdown is referred to as 'Geiger-mode' operation, and SPADs are also often called Geiger-mode APDs (GmAPDs).

Because avalanche breakdown can be self-sustaining in the Geiger mode, the concept of gain is not well defined. It is more appropriate to consider the SPAD to be a photon-activated switch. A photon arrival corresponds to closing the switch, which provides a macroscopic avalanche current to be detected by a suitable threshold circuit. Once detected, the avalanche must be quenched by some means of reducing the bias voltage to a value for which the avalanche is no longer self-sustaining. Quenching of the avalanche corresponds to opening the switch. In practice, Geiger-mode operation is a sequence consisting of (i) arming the device by biasing above V_b, (ii) triggering an avalanche with an incident photon, (iii) quenching the avalanche by lowering the bias, and (iv) re-arming the device by again biasing above V_b. As we will discuss, there are several approaches to avalanche quenching, and the specific implementation of this functionality can have a dramatic effect on SPAD photon counting performance.

SPAD operation in Geiger mode involves a number of fundamental performance parameters. The probability with which an incident photon triggers a detection event is the photon detection efficiency (PDE). The probability that a detection event occurs in the absence of an incident photon is the dark count rate (DCR). The accuracy with which the photon arrival time can be determined is the timing jitter. There is also a phenomenon specific to photon counting devices in which additional dark counts called afterpulses are induced at high counting rates and are correlated to the occurrence of previous detection events. We will discuss these four principle SPAD parameters in detail in the following sections.

A key feature of SPAD operation is that the Geiger mode process involving threshold detection of avalanche events is digital and, as such, is effectively noiseless. The origin of noise in these devices is simply the shot noise of the dark counts. Sensitivity metrics such as signal-to-noise ratio and noise equivalent power can be defined for SPADs using an analysis similar to that used for analog detectors for which the shot noise of electron flow determines the detector's noise properties [19]. By substituting the fluctuations of electron flow according to Poisson statistics in analog detectors with the fluctuations of counts in SPADs, one can show that the noise equivalent power (NEP) of a SPAD is given by $h\nu(2 \cdot \mathrm{DCR})^{1/2}/\mathrm{PDE}$ where h is Planck's constant and ν is the optical frequency. Assuming a DCR of 1 kHz, a PDE of 40%, and a wavelength of 1.5 μm, one finds $\mathrm{NEP} \sim 1 \times 10^{-17}\,\mathrm{W\,Hz^{-1/2}}$, which corresponds to a sensitivity several orders of magnitude better than any analog detector suitable for this wavelength. It is also interesting to note that the proportionality of NEP to $(\mathrm{DCR})^{1/2}/\mathrm{PDE}$ provides a performance metric that is not unique. Some applications may dictate different metrics, such as the scaling of quantum bit error rate with DCR/PDE in quantum key distribution [20].

2.2. Current dominant applications for NIR SPADs

In any measurement of optical radiation, the ability to detect a single photon is the ultimate level of sensitivity, and many optical techniques have evolved to the point where single photon detection is a critical enabling capability. In numerous cases, applications require single photon detectors because they involve physical processes in which only a very small number of photons – often just one – are available for detection. Examples of such applications at NIR wavelengths include fiber-based optical time domain reflectometry [21], semiconductor circuit diagnostics [22], laser-based remote sensing and range finding [23], free space optical communications in photon-starved environments [24], singlet oxygen detection [25], and single-photon three-dimensional LADAR imaging [26,27]. In other instances, it is the quantum properties of a single photon that are exploited, and the broad field of quantum optics, particularly quantum information processing, is critically dependent on the means for sensing individual photons. Notable examples of quantum optics applications of single photon detectors include quantum cryptography [20], quantum computing [28], and fundamental studies of quantum physics [29].

All of these applications benefit from high PDE and low DCR, and for many of them, the performance of the SPADs described in this paper meet at least their minimum requirements for PDE and DCR.

However, in some of the most active areas of research and development requiring single photon detection, achieving higher counting rates has become an imperative. In particular, applications involving communications-related techniques, such as quantum cryptography and free space optical communications, have benchmarks set by the common deployment of data transmission rates in the range of 100 MHz to 1 GHz. The development of photon counting techniques capable of providing comparable single photon counting rates has been a dominant theme in recent work on SPADs.

Another area of recent focus for NIR SPADs has been the push to employ single photon sensitivity in the pixels of large-format imaging arrays. Beyond the adequacy of single pixel performance, imaging applications requires high yield and tight uniformity, and they have been a driver for these attributes that are of much less concern for applications that make use of discrete detectors.

Finally, it is notable that, at present, NIR SPADs serve very few applications in the field of biomedicine. This field has been a source of enormous opportunity for silicon-based SPADs because numerous biomedical flourescence techniques involve the emission of photons at wavelengths in the range of 0.6 to 0.9 μm, for which detection by Si SPADs is excellent. Single-photon sensitive photomultiplier tubes (PMTs) are the legacy detector technology serving this sizable existing market, and Si SPADs are poised to replace PMTs in many of these applications. For NIR SPADs, there is an interesting opportunity in the detection of singlet oxygen based on weak emission of 1.27 μm radiation as a means of establishing more precise and effective photodynamic therapy for the treatment of some forms of cancer [25]. However, this is one of the few medical applications of NIR SPADs to emerge to date, and it is still in the early stages of development.

3. Fundamental SPAD device design and fabrication considerations

3.1. Overview of APD device design considerations

The SPADs described in this paper evolved from avalanche photodiodes (APDs) originally designed for use in high-bandwidth fiber-optic telecommunications receivers [30]. They have fundamental design elements that are common to virtually all InP-based APDs in use today, including the separate absorption and multiplication (SAM) region structure first introduced over 30 years ago [31]. The goal of this structure is to provide sufficiently high electric field in the InP multiplier to achieve avalanche gain by impact ionization while maintaining sufficiently low electric field in the $In_{0.53}Ga_{0.47}As$ absorber so that tunneling effects are suppressed in this layer (see Figure 1). The use of a charge – or field control – layer [32] between the absorber and multiplier regions of the structure provides flexibility in controlling the internal electric field profile of the device. Additional InGaAsP grading layers are often employed to minimize hole trapping effects that arise from the valence band discontinuity that exists at abrupt InGaAs/InP heterointerfaces [33].

The long-wavelength spectral response of the device is governed by the InGaAs absorber, which is lattice-matched to InP, and has a room-temperature bandgap of $E_g \sim 0.75$ eV that corresponds to a cut-off wavelength of ~1.67 μm. The InP bandgap of $E_g \sim 1.35$ eV dictates the short-wavelength response since photons

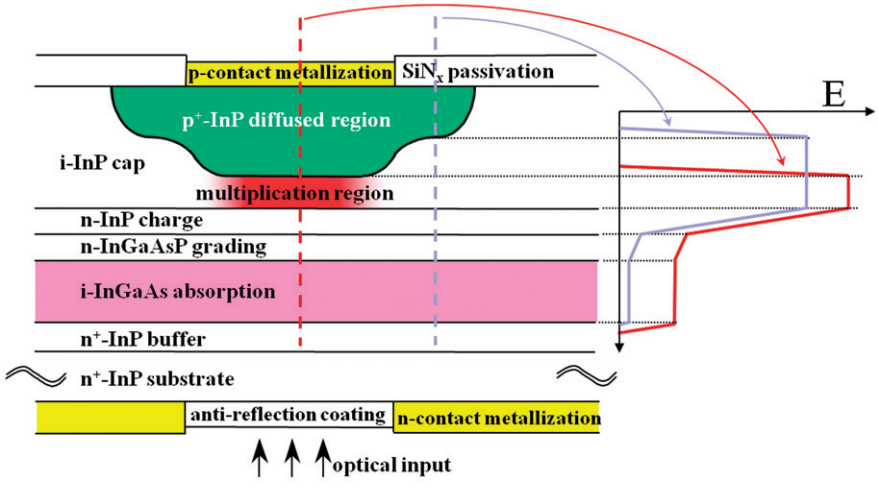

Figure 1. Schematic illustration of a diffused-junction planar-geometry avalanche diode structure. The electric field profiles at right show that the peak field intensity is lower in the peripheral region of the diffused p-n junction than it is in the center of the device. (The color version of this figure is included in the online version of the journal.)

with a wavelength of less than ~0.92 μm will be absorbed by InP before they can reach the InGaAs absorber. A back-illuminated configuration is used to achieve the desired optical active area with minimal overall device area, thereby minimizing device capacitance and dark count rate. Back illumination also provides improved detection efficiency since photons that are not absorbed during their initial transit through the InGaAs absorber experience a partial reflection from the front-side anode contact metallization for a second pass through the InGaAs. Finally, the back-illuminated structure is compatible with high fill factor arrays since it allows for flip-chip bonding of the anode contacts to readout integrated circuits.

The lateral configuration of the avalanche diode is determined by creating a buried p-n junction using the diffusion of Zn dopant atoms through a SiN dielectric passivation layer patterned with diffusion windows. The two-dimensional electric field profile resulting from a single diffusion tends to exhibit field peaking and associated edge breakdown effects near the quasi-cylindrical junction edges. To suppress edge breakdown, we employ two concentric diffusions [34] so that the p-n junction is deeper in the central region of the diode than it is in the peripheral region. The wider multiplication region in the periphery of the device has a lower peak field intensity (see right side of Figure 1) and higher breakdown voltage compared to the center of the device, so avalanche breakdown is confined to the central region of the structure. This buried-junction planar geometry provides low perimeter leakage and stable long-lifetime operation, and this device platform has been shown to have excellent reliability in the context of telecom receiver qualification [35].

It is important to stress that although the InGaAs/InP SPAD shares a common design platform with linear mode APDs, the optimization of these two different types of devices is quite distinct [1]. Linear mode APDs are operated below breakdown, and their design goals generally emphasize high gain-bandwidth product and low excess noise factor. Both of these performance attributes are aided by the use of narrow multiplication layer widths of $\leq$0.5 μm to achieve more rapid carrier transport and to also derive the benefits of so-called 'dead space' effects that result in more deterministic avalanche dynamics with consequently lower excess noise [36]. However, in Geiger mode APDs, design priorities include the minimizing of dark count rate and maximizing of the probability for detectable avalanche events (to achieve high photon detection efficiency). For Geiger mode operation, gain-bandwidth product and excess noise factor are irrelevant, leading to rather different design strategies than those used historically in developing linear mode APDs. In the next sub-section, we illustrate an example of this divergence of Geiger mode design from linear mode design in modeling the dependence of photon counting performance on SPAD multiplication layer width.

3.2. PDE versus DCR modeling, DCR mechanisms, and impact on device design

The most fundamental consideration in the design of SPADs is managing the trade-off between photon detection efficiency (PDE) and dark count rate (DCR). The PDE is the product of three probabilities: $\mathrm{PDE} = \eta_q P_c P_a$, where η_q is the quantum efficiency for carrier creation by absorption of an incident photon in the InGaAs absorber; P_c is the probability that a photoexcited carrier is collected by injection into the InP multiplication region; and the avalanche probability P_a is the probability that a carrier injected into the multiplication region actually gives rise to a detectable avalanche. The DCR is dictated by the probability that an electrical carrier is created by any mechanism other than photoexcitation, and it is also proportional to the avalanche probability P_a. (There is an additional subtlety with DCR in that carrier creation can occur in either the absorber or the multiplier, and for the latter case, P_a will depend on the position within the InP multiplier at which the dark carrier is created.)

The modeling of PDE and DCR requires the calculation of several dynamic processes within the device structure. Several of these processes are highly dependent on the local electric field intensity; therefore, calculation of the internal electric field profile is essential and depends on the doping charge concentration profile within the structure. Calculation of the avalanche probability P_a relies on a description of the avalanche process arising from impact ionization, and the adoption of appropriate expressions for the impact ionization coefficients – particularly their temperature dependence [37] – is critical to the accuracy of the model. Dark carrier creation can occur through field-dependent tunneling processes, which can be band-to-band or trap-assisted, as well as thermally driven Shockley–Read–Hall processes. The first comprehensive exposition of a PDE versus DCR model for InP-based SPADs was developed by Donnelly et al. [38], and we have adopted this formalism in previous work on both InGaAs/InP SPADs for 1.5 μm photon counting [39] as well as InGaAsP/InP SPADs for use at 1.06 μm [40].

As an example of the output of this model, we show in Figure 2 the calculated dependence of DCR per unit area on the PDE as a function of the SPAD multiplication layer width. A wider multiplication region

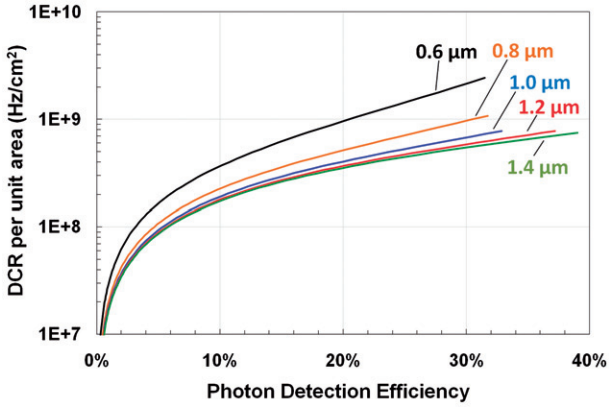

Figure 2. Modeling of dark count rate (DCR) per unit area as a function of the photon detection efficiency at 213 K and 1.5 µm wavelength. Results are shown for five different values of multiplication width W_m in the range 0.6–1.4 µm. The modeled structure includes a 1.0 µm InGaAs absorption layer and an integrated field control sheet charge of $2.0 \times 10^{12}\,cm^{-2}$. (The color version of this figure is included in the online version of the journal.)

is clearly beneficial for achieving a lower DCR at a given value of PDE. The model has additional utility in that it provides us with quantitative information concerning the contributions to the DCR of different leakage mechanisms. For the same structure, the modeling output in Figure 3 shows that increasing the multiplication layer width from 0.6 µm to 1.0 µm provides a significant reduction in trap-assisted tunneling in this layer since the wider multiplier reaches the breakdown condition at significantly lower electric field intensity, and the lower field operation reduces tunneling effects. On the other hand, thermally generated dark counts originating in the InGaAs absorber remain fairly independent of multiplication width and are instead very sensitive to operating temperature. This modeling platform is very useful for optimizing device structures to achieve specific targets for DCR or PDE subject to constraints on operating parameters such as excess bias and temperature. The optimization of SPAD performance with respect to multiplication region width has also received a theoretical treatment [41] invoking generalized breakdown probabilities calculated using the recursive dead-space multiplication theory [42].

4. Survey of state-of-the-art Geiger mode performance in IR SPADs

Using the modeling tool described in the previous section, we have made significant progress in improving the fundamental DCR versus PDE trade-off in InP-based SPAD performance relative to device characteristics reported in earlier work [1]. We present recent results for these two performance metrics, and in this section we also summarize results for two other critical performance parameters: timing jitter and afterpulsing.

4.1. PDE versus DCR trade-off

To demonstrate the state-of-the-art in the PDE versus DCR trade-off in InGaAs/InP SPADs, we present in Figure 4 experimental measurements for 20 devices fabricated in the same process lot. Measurements were taken at a temperature of 218 K with illumination at 1.55 µm using an attenuated source calibrated to supply 170 ps pulses with a mean photon number of $\mu = 0.1$. The SPADs were operated with gated biasing using short 1 ns gate pulses at a repetition frequency of 500 kHz [43,44]. The SPADs were assembled into a butterfly-style package that allowed for highly accurate and stable fiber-coupling to the 25 µm optical active area of the detectors [45]. Given that these devices had a multiplication width of ~1.5 µm, the modeling from Figure 2 at a PDE of 20% predicts a DCR of ~2 kHz, which is close to the median of the measured distribution for DCR versus PDE.

The data in Figure 4 establish that sub-kHz DCR performance can be achieved at PDE values as high as ~25% for the best of our devices. On the other hand, the data also show a rather wide distribution of performance, with DCR varying by more than an order of magnitude for any given PDE value. This performance variation does not exhibit a strong correlation to device position on the wafer, and based on significantly narrower performance distributions found for large-format (e.g. 32×32) SPAD arrays described below, we believe that significant contribution to this performance variation comes from factors unrelated to as-processed material quality and device parameter uniformity.

For higher operating temperatures, DCR increases by approximately a factor of two for every 10 K increase in temperature based on the dominance of thermally generated dark carriers in the absorber; see Figure 3(*b*). Further reduction in operating temperature to below ~200 K can be used to reduce DCR, although at sufficiently low temperature (e.g. 150–175 K), DCR reduction saturates as it becomes dominated by trap-assisted tunneling [46], which has only a weak temperature dependence.

The limitation in the maximum PDE values reported in Figure 4 to the range of 30% to 35% is dictated by the constraint that our 1 ns gating circuit can apply only ~4.0 V excess bias. Higher PDE can be achieved with larger excess bias, although

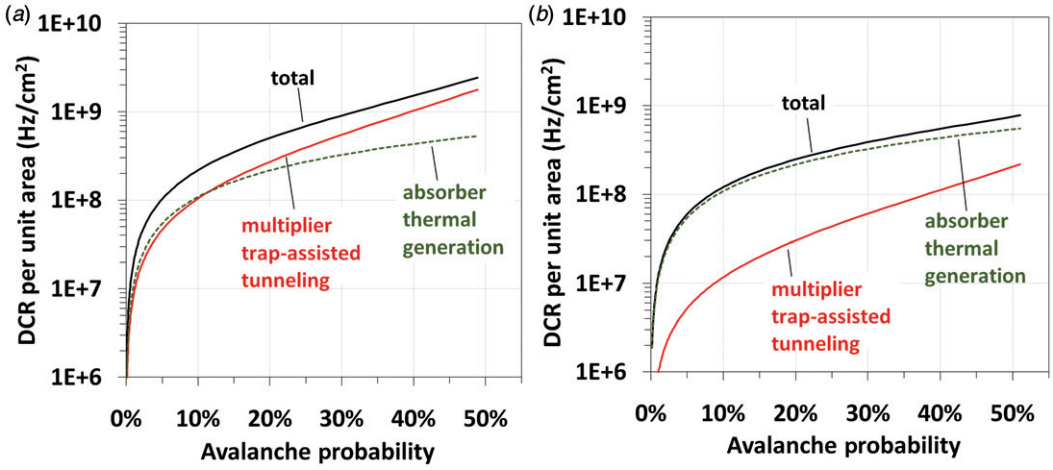

Figure 3. Simulation results for dark count rate as a function of avalanche breakdown probability for an InGaAs/InP SPAD structure with a multiplier width of (*a*) 0.6 μm and (*b*) 1.0 μm. The wider multiplier provides significant reduction of trap-assisted tunneling in this device layer. Calculations are for the same structure described in Figure 2. (The color version of this figure is included in the online version of the journal.)

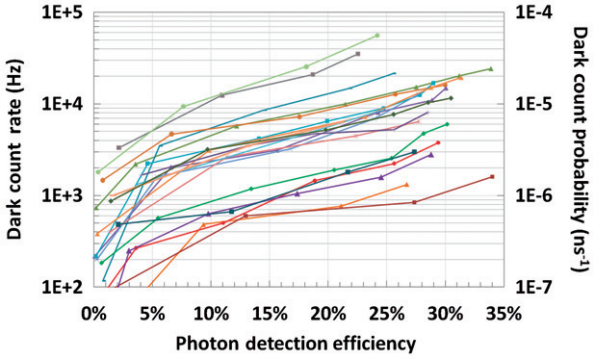

Figure 4. Dark count rate as a function of photon detection efficiency for twenty 25 μm diameter InGaAs/InP SPADs from a single process lot. Data was obtained with 1 ns gated operation at 500 kHz repetition rate and 218 K using 1.55 μm illumination. (The color version of this figure is included in the online version of the journal.)

with a consequent increase in DCR that will follow an extrapolation of the DCR versus PDE behavior seen in the figure.

4.2. *Timing jitter*

Another SPAD performance attribute that is important in many applications is the accuracy with which the precise arrival time of a photon can be determined. There is an average latency between the time a photon impinges on the SPAD and the time at which an avalanche event is detected by electronic circuitry connected to the SPAD. However, what is more critical than the average value of this latency is its variation, commonly referred to as the timing jitter. With repeated high-precision measurements of the time of avalanche detection relative to a fixed reference for the photon arrival time, one can build up an experimental distribution of detection times. The timing jitter is then determined from some measure of the width of this distribution. The distribution full-width at half-maximum (FWHM) is a frequently cited criterion for the jitter and is most relevant when the timing distribution is close to Gaussian. This is sometimes not the case, such as when non-Gaussian tails are present in measured timing histograms. We cite FWHM jitter values below, but we note that the numerically computed root-mean-square (rms) standard deviation is sometimes a preferred measure of the timing jitter since it captures non-Gaussian elements of the timing distribution.

4.2.1. *Factors contributing to device-level timing jitter*

There are a number of physical mechanisms within any SPAD structure that can contribute to the timing jitter. These mechanisms include (i) differences in the transit times of photoexcited carriers resulting from differences in the location of photon absorption; (ii) carrier propagation delay caused by the temporary trapping of carriers at heterojunctions formed by dissimilar semiconductor layers; and (iii) variations in the avalanche build-up time induced by the stochasticity of the impact ionization process that produces avalanches. Avalanche build-up time variation also includes effects related to the randomness inherent in the spreading of the avalanche from an initially localized filament to a saturated carrier multiplication process that fills the entire high-field active area of the device [47].

4.2.2. Simple Monte Carlo model of dominant build-up time contribution

Among the various timing jitter mechanisms, the principal contribution – particularly at low excess bias voltage – is the fundamental build-up time required for the avalanche amplitude to reach a predetermined threshold level. Although studies involving jitter modeling are not extensive in the literature on SPADs, the avalanche build-up time has been modeled using Monte Carlo techniques [48]. Using this approach, we begin by assuming that the random ionization path length of an electron x_e is described by the probability density function $h_e(x_e)$:

$$h_e(x_e) = \begin{cases} 0, & x_e \le d_e, \\ \alpha^* \exp[-\alpha^*(x_e - d_e)], & x_e > d_e, \end{cases}$$

where d_e is the electron dead space (within which no impact ionization event can occur) and α^* is the 'enabled' electron ionization coefficient. The electron survival probability is then

$$S_e(x_e) = \int_0^{x_e} h_e(x)\,\mathrm{d}x = \begin{cases} 1, & x_e \le d_e, \\ \exp[-\alpha^*(x_e - d_e)], & x_e > d_e. \end{cases}$$

By substituting a uniformly distributed random number r between 0 and 1 for $S_e(x_e)$, the electron random ionization path length can be obtained as

$$d_e = x_e - \frac{\ln(r)}{\alpha^*}.$$

Similar expressions can be defined for the hole ionization path length, and at any given time t, the current $I(t)$ is

$$I(t) = \frac{q}{W}[N(t)v_{se} + P(t)v_{sh}],$$

where q is the electron charge and W is the multiplication width of the avalanche diode structure. $N(t)$ and $P(t)$ are the number of electrons and holes inside the multiplication region at time t, respectively. v_{se} and v_{sh} are the electron and hole saturation velocity, respectively.

We apply this model to our canonical SPAD structure (see Figure 1) for two different multiplication widths of 1.0 and 1.7 μm. v_{se} and v_{sh} are taken to be 1×10^7 and $7 \times 10^6\,\mathrm{cm\,s^{-1}}$, respectively [49]. The threshold current was chosen to be 0.5 μA, and the operating temperature assumed for the simulations was 235 K. In Figure 5, we show simulated results for (*a*) the mean delay time between photon arrival and avalanche detection at the modeled threshold current, and (*b*) the variation in this mean delay time, i.e. the timing jitter. Both the mean time and the timing jitter are very sensitive to avalanche probability (as determined by the excess bias voltage) for low values of avalanche probability (e.g. <30%), whereas the dependence on avalanche probability becomes weaker at higher values (e.g. >60%). The results show the predicted advantage of using thinner multiplication regions for improving the jitter performance. In particular, the model predicts that the narrower 1.0 μm multiplier reduces timing jitter due to avalanche

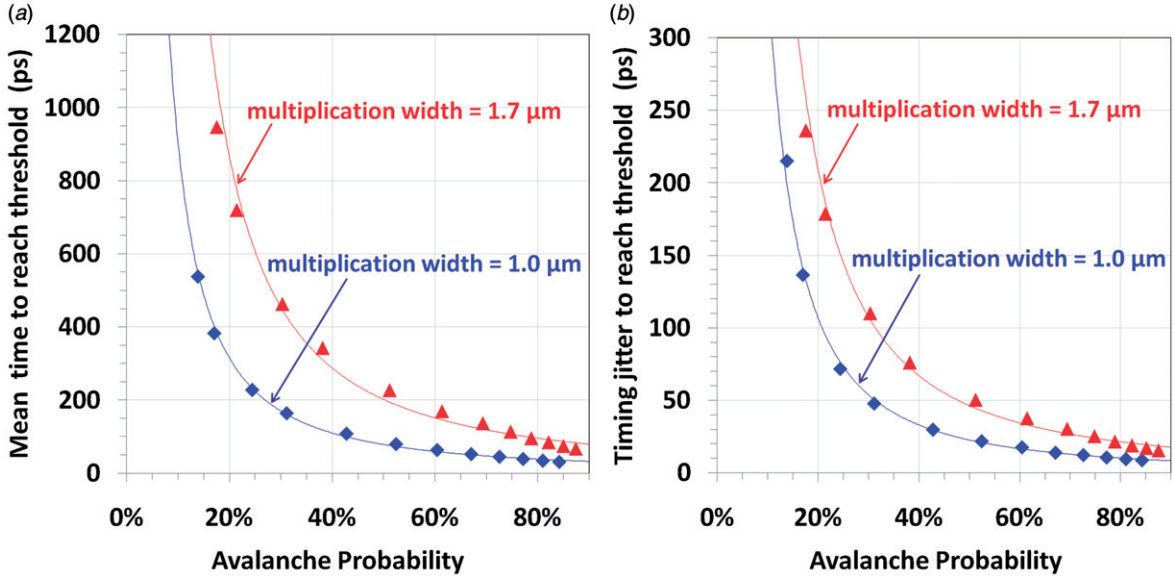

Figure 5. Results for Monte Carlo model calculations for (*a*) mean delay time between photon arrival and avalanche detection and (*b*) variation in the mean delay time, i.e. timing jitter, as a function of the avalanche breakdown probability. Power law fits match the model output well but are not yet theoretically motivated. (The color version of this figure is included in the online version of the journal.)

build-up effects by ~50% relative to the 1.7 μm multiplier for avalanche probability >30%.

It should be noted that the predicted timing jitter in Figure 5(*b*) corresponds only to the contribution of the avalanche build-up process. Aside from the additional stochastic dynamic processes mentioned above (i.e. transit time, interface carrier trapping, and lateral avalanche build-up effects), there is also the important consideration of local excess bias non-uniformities resulting from non-uniform breakdown voltage across the device active area. If the excess bias exhibits considerable variation as a function of position in the device – leading to correlated variation in the avalanche probability – then the associated distribution of mean times to reach threshold (see Figure 5(*a*)) further broadens the timing distribution and may increase the effective timing jitter significantly above that which would be found for a device with an ideally uniform excess bias, as assumed for the modeling in Figure 5. We demonstrate the experimental consequences of this effect below, but first we show results for a typical InGaAs/InP SPAD exhibiting good jitter performance in the next sub-section.

4.2.3. *InP SPAD timing jitter performance*

The data in Figure 6 show the timing response distribution for a low-jitter InP/InGaAs SPAD operated at 7 V excess bias and 175 K. The absolute timescale on the abscissa is arbitrary, but relative measures of the distribution width reflect the uncertainty in the timing of avalanche detections, or timing jitter. Based on a full-width half-maximum (FWHM) criterion, the jitter for this distribution is 46 ps, which is reasonably close to the best results achieved with Si SPADs [50] and represents excellent timing performance relative to other single photon detector technologies. Such good jitter performance reflects the timing capability of the SPAD but is only possible with very high performance circuits that can accurately detect the onset of the avalanche response at very low signal levels without spurious detections caused by circuit transient response characteristics. The critical importance of circuit capability is illustrated in [51], where the same device is shown to have a much larger FWHM timing jitter of 95 ps when a simpler threshold detection circuit without transient cancellation is employed.

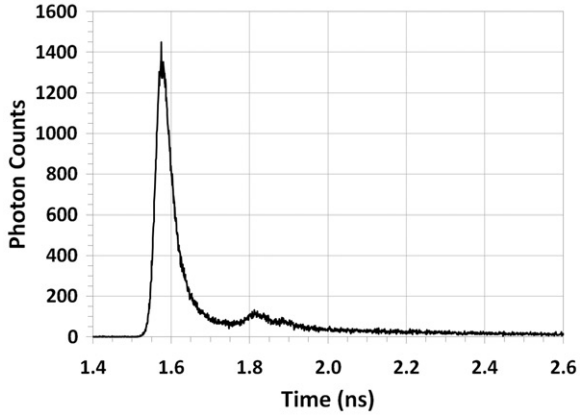

Figure 6. Timing response distribution for a 25 μm diameter InP/InGaAs SPAD operated at 7 V excess bias, 175 K, and 1.55 μm illustrating a 46 ps timing jitter based on a full-width half-maximum criterion. Although the computed rms deviation of the main peak is only 24 ps, the tail of the distribution increases the rms deviation to 59 ps.

4.2.4. *Jitter degradation caused by spatial non-uniformity in excess bias*

If the spatial dependence of the breakdown voltage of a SPAD exhibits non-uniformity within the device active area, then the effective excess bias voltage – and any parameter that depends on it – will demonstrate corresponding non-uniformity. The most readily measurable effect of non-uniform excess bias is the spatial dependence of the PDE. In Figure 7 we illustrate two-dimensional scans of the PDE for a SPAD with non-uniform response. Data were obtained for 2 μm step sizes using a focused 1.55 μm laser with a 5 μm spot size in the plane of the SPAD active area. The operating temperature was 225 K, and scans were taken at several values of the excess bias.

The non-uniformity is exacerbated for low excess bias values, as in Figure 7(*a*) corresponding to $V_{ex} = 1.5$ V, where strong peaking of PDE is evident around the perimeter of the 25 μm diameter active region defined by the portion of the device structure where the p^+-dopant diffusion is deepest (see Figure 1). As discussed with reference to Figure 1, the control of edge breakdown requires the central portion of the diffused p-n junction to be slightly deeper than the junction in the peripheral region. However, if this inner diffused region protrudes too far beyond the diffusion in the peripheral region, edge breakdown control is no longer effective, and enhanced electric field amplitudes at the perimeter of this inner diffusion induce the peaking in PDE shown in Figure 7(*a*). In the fabrication of this device, the diffusion of the central region was deeper than established design targets, and similar scans of the linear mode gain below breakdown showed qualitatively similar edge peaking for this device, as expected.

For higher values of V_{ex}, the dependence of PDE on V_{ex} is reduced due to saturation of the avalanche probability P_a, and so the variation in PDE across the area of this device is also reduced. The scans in Figure 7(*b*) and (*c*) for $V_{ex} = 3$ V and $V_{ex} = 5$ V clearly

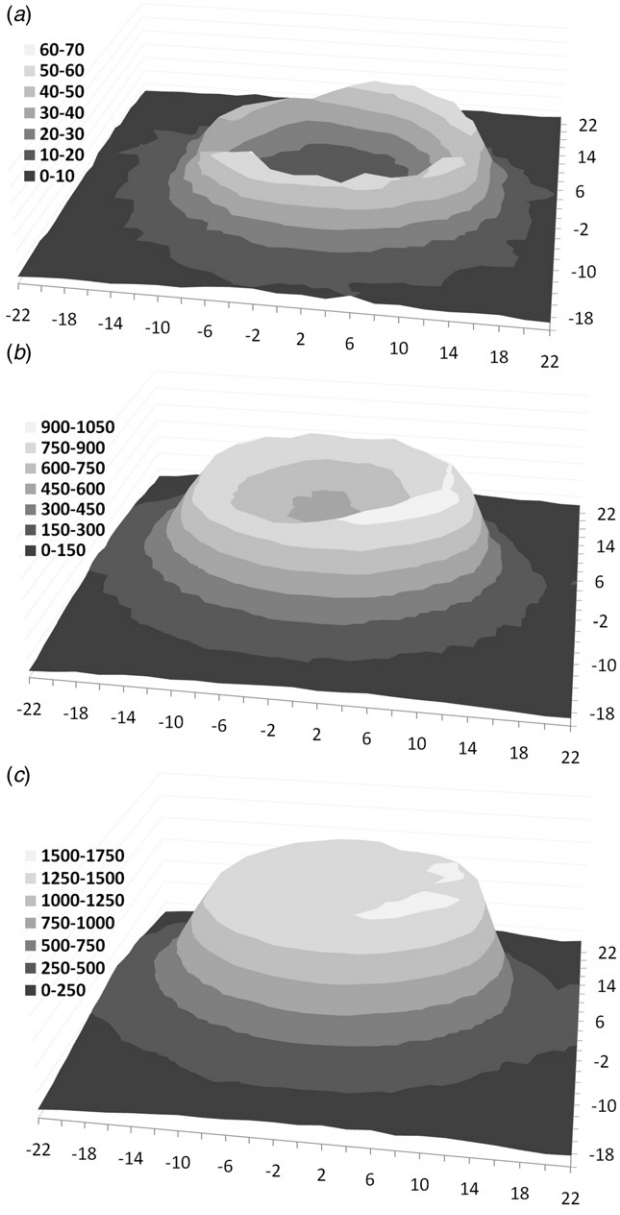

Figure 7. Two-dimensional scans of the spatial dependence of the PDE for a SPAD with non-uniform response across the active area. Data were obtained with 2 μm step sizes using a 5 μm spot size at a wavelength of 1.55 μm and an operating temperature of 225 K. Scans are shown for excess bias V_{ex} of (*a*) 1.5 V, (*b*) 3 V, and (*c*) 5 V. Axis labels are in μm; amplitudes are in arbitrary units.

show that PDE becomes more uniform for larger values of V_{ex}.

In addition to PDE, the mean delay time between photon arrival and avalanche detection also exhibits a strong dependence on V_{ex}, especially for low values of V_{ex}, as seen in the modeling of avalanche timing presented in Figure 5. Given the non-uniform V_{ex} of the device which exhibited the PDE edge peaking seen in Figure 7, one would expect to find a corresponding

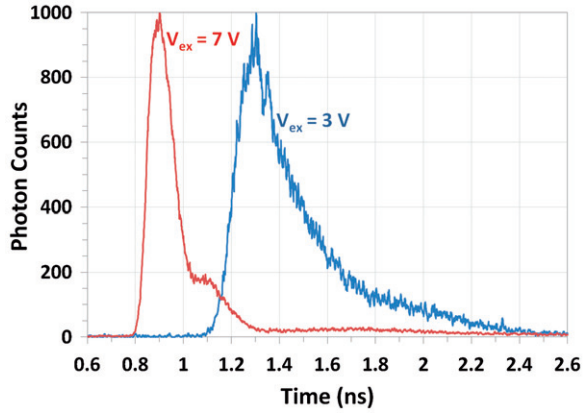

Figure 8. Timing distributions at 3 and 7 V excess bias for the SPAD with non-uniform PDE illustrated in Figure 7. A collimated illumination source samples the variation in timing response across the non-uniform active region and leads to broad timing distributions characterized by large timing jitter. (The color version of this figure is included in the online version of the journal.)

variation in the mean delay time across the SPAD active area, leading to substantially higher timing jitter. Jitter measurements on this same device are illustrated in Figure 8 for excess bias voltages of 3 and 7 V. The device was illuminated with a collimated beam approximately filling the 25 μm diameter active region so that avalanche responses effectively sample the timing characteristics of all locations in the active region. At a fairly low excess bias of $V_{ex} = 3$ V, the jitter is quite large because of the greater variation in mean delay time with V_{ex}, and the FWHM metric gives a value of 239 ps. Even for a much higher excess bias of $V_{ex} = 7$ V, the FWHM timing jitter is still 124 ps, which is considerably larger than the 46 ps FWHM value found at the same excess bias for the device that yielded the jitter performance illustrated in Figure 6. (We note that scans of the PDE for the lower jitter device confirmed good uniformity across the active region.) These results emphasize the importance of ensuring uniform excess bias across the active region to achieve excellent timing jitter performance.

4.3. *Afterpulsing and impact on photon counting rate*

Avalanche events in SPADs typically induce the flow of a large number of charge carriers through the device multiplication region (e.g. 10^6–10^8 carriers, depending on details of the avalanche quenching circuitry), and some fraction of these carriers can become trapped at atomic-level defects in the multiplication region. Over time, trapped carriers are detrapped by thermionic emission, and their population decays exponentially. If carrier detrapping occurs while the SPAD is disarmed – i.e. while the bias voltage is below

breakdown – then the detrapped carriers drift out of the multiplication region without consequence. However, if the SPAD is re-armed while traps are still populated, then there is a finite probability that subsequently detrapped carriers can trigger additional (dark) avalanche events referred to as afterpulses. It is possible to reduce afterpulsing to arbitrarily low levels by using a hold-off time that is sufficiently long to allow nearly all trapped carriers to be detrapped before the SPAD is re-armed. However, this strategy limits SPAD operation to low counting rates on the order of the inverse of the hold-off time. As counting rates are increased by using shorter hold-off times, the probability of afterpulsing increases. The net effect is an increase in dark count rate for higher counting rates, with the additional afterpulsing dark counts being correlated to the occurrence of prior avalanches. Given that one of the recent trends in photon counting applications is the need for higher counting rates, afterpulsing has emerged as a central limitation of SPAD performance.

In this section, we outline strategies for afterpulsing mitigation and describe the dominant role of reducing avalanche charge flow. To describe considerations related to avalanche charge flow, we summarize the principle equivalent circuit model in use today to simulate SPAD behavior. We then present results for three types of measurements used to characterizing afterpulsing effects. With two of these measurements, we demonstrate a significant increase in repetition rate to 50 MHz or more for gated operation with 1–2 ns gates.

4.3.1. *Afterpulsing mitigation strategies and the role of total avalanche charge flow*

There are several plausible strategies for the mitigation of afterpulsing at high counting rates. Since this effect arises from the trapping of avalanche charges, one approach is to (i) decrease the density of material defects that act as potential charge traps. A second strategy is to (ii) induce a rapid intentional detrapping of carriers by some appropriate applied stimulus. Finally, a third option is to (iii) reduce the number of charges that are trapped in the first place by reducing the amount of charge that flows during each avalanche event.

In the last decade of research on InP-based SPAD design, fabrication, and characterization, there has been no indication of improvement in material quality or concepts for achieving such improvement as relates to the density of defects that cause afterpulsing. A key problem in this area is the paucity of knowledge concerning what types of material defects could be acting as charge traps leading to afterpulsing and what causes their formation. More will be said about this topic in Section 5.1.2 below.

There have been a few attempts to implement the second strategy of intentional charge detrapping. Most notable among these is the idea of using longer wavelength radiation (beyond the long-wavelength cutoff for absorption in InGaAs) to photoexcite trapped carriers out of their traps immediately after the quenching of each avalanche. Although there has been a report [52] relating some initial promise for this approach, further study indicated that the long-wavelength radiation reduced afterpulsing by faster thermal excitation simply because it was heating the SPAD [53]. Attempts by other research groups at using photoexcitation of trapped charges to reduce afterpulsing have also been unsuccessful to date. (We also note that the operation of SPADs at higher temperatures to achieve faster detrapping for reduced afterpulsing could also be categorized as a variant of this induced detrapping strategy, but this approach is penalized by higher DCR.)

Therefore, essentially all recent efforts to mitigate afterpulsing have invoked the third strategy of reducing the number of charges that are trapped by restricting avalanche events to having less charge flow. There have been experimental measurements to confirm that the amount of trapped charge and the consequent afterpulsing do in fact scale linearly with the charge flow per avalanche [6,54]. For situations in which gated mode operation with very short (~sub-ns scale) gates is appropriate, avalanche charge flow can be reduced dramatically because the falling edge of the very short gate acts to rapidly quench the avalanche. This basic concept has been implemented to achieve very high (~GHz) gating frequencies using new schemes such as self-differencing [8] and sine-wave gating [10] that will be discussed further in Section 5.3. More general solutions that can accommodate non-periodic gating scenarios have focused on sensing the avalanche with as low a detection threshold as possible and then rapidly quenching it to minimize the charge flow [55]. Finally, there has been recent work on self-quenching SPADs [13–18] in which the monolithic integration of passive quenching elements can lead to reduced charge flow if the quench elements can be integrated with negligible parasitic capacitance. (Since parasitic capacitive elements must be discharged and recharged with each avalanche event, their elimination can reduce the overall charge flow per avalanche.)

4.3.2. *Description of equivalent circuit for modeling SPAD operation (a la Haitz)*

To gain more perspective on controlling the charge flow within each SPAD avalanche event, it is instructive to consider a simple equivalent circuit first introduced by Haitz in 1964 for an avalanche diode

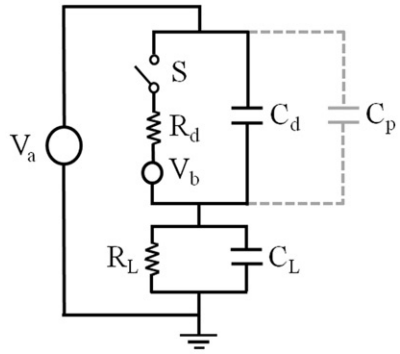

Figure 9. Equivalent circuit for an avalanche diode operating in Geiger mode. The SPAD consists of two parallel branches with diode capacitance C_d and dynamic resistance R_d. The resistive branch supports the breakdown voltage V_b without avalanche current flow, and the closing of switch S emulates the onset of avalanche current flow through the device. Avalanche current can flow when the applied voltage $V_a = V_b + V_{ex}$ exceeds V_b by the excess bias voltage V_{ex}. Parasitic capacitance C_p in parallel with C_d will increase the effective capacitance of the device. A generalized load is represented by parallel resistive and capacitive branches R_L and C_L.

operating in Geiger mode [56]. As illustrated in Figure 9, an applied voltage $V_a = V_b + V_{ex}$ is imposed across the entire circuit such that the SPAD is reverse biased beyond its breakdown voltage V_b by the excess bias V_{ex}. We also include an arbitrary load between the diode and ground. The diode itself is modeled as two parallel branches. One branch consists of the reverse-biased diode capacitance C_d determined by the area of the p-n junction and the width of its depletion region. The second branch includes the diode dynamic resistance R_d above breakdown in series with a voltage source equivalent to the breakdown voltage V_b as well as switch S. Inclusion of V_b in this branch reflects the fact that no avalanche current will flow unless $V_a > V_b$, and it is only $V_{ex} = V_a - V_b$ that determines avalanche current in the circuit. (Current leakage mechanisms below breakdown are ignored since they will not cause detectable avalanche events.) The switch S is used to represent the presence or absence of avalanche current: when no avalanche current flows, the switch is open; the onset of an avalanche is captured by closing the switch.

The simplest description of SPAD operation using this equivalent circuit is provided by assuming that the load is a passive load resistance R_L, in which case the associated operation of the SPAD is referred to as 'passive quenching'. (The inclusion of the load capacitance C_L in Figure 9 represents a more general case, including capacitive parasitics of the load.) If we assume the SPAD is initially in its armed state, then the voltage across the SPAD exceeds V_b by the excess bias V_{ex}. With the switch S open, no current flows. The onset of an avalanche is modeled by the closing of switch S, at which point capacitance C_d discharges through the diode dynamic resistance R_d with a time constant on the order of $\tau_{dis} \sim R_d C_d$. The removal of charge from C_d reduces the voltage across the SPAD structure, although the precise amount of voltage removed depends on the ratio of R_d and R_L. In fact, in steady state, the excess voltage V_{ex} is split between the SPAD structure and the load according to the voltage divider presented by these two resistances in series. The amount of voltage removed from the SPAD structure is precisely the amount of voltage $I \cdot R_L$ developed across the load by the introduction of a current I from the voltage source. When the total current through the switch S drops to a value smaller than a characteristic quench value I_q, the avalanche will spontaneously quench, represented by the re-opening of switch S. With S open, C_d is re-charged through load resistance R_L, with a recharging time constant $\tau_r \sim R_L C_d$ dictating how long it takes to re-arm the SPAD.

The presence of any parasitic capacitance C_p in parallel with C_d increases the effective capacitance dictating device behavior. Therefore, to minimize the current flow associated with avalanche events, as well as to decrease the time constants governing circuit response times, it is desirable to eliminate C_p. More general loads – including inductance or non-linear elements – are easily incorporated into the model. Moreover, the simple passive quenching operation just described can be enhanced by the incorporation of additional circuitry to provide active quenching in which the applied bias is actively lowered below V_b. Many possible variations on passive and active quenching have been comprehensively reviewed in [57].

The analytical implementation of this model has been combined with detailed statistical descriptions of carrier dynamics in the semiconductor device structure to extract useful insights concerning device behavior. In particular, with canonical passive quenching, the internal electric field is found to oscillate along with the consequent avalanche charge flow [58], and the spontaneous quenching of the avalanche is seen to correspond to the stochastic decrease of charge flow to zero during the low-field portions of the electric field oscillation [59]. This modeling effort was carried out to explain recent experimental results for self-quenching SPADs described in Section 5.2.

On an even more practical level, the SPAD equivalent circuit has recently been incorporated into a SPICE simulation environment [60,61]. The ability to monitor currents and voltages at arbitrary points in the circuit simulation provides significant new detail pertaining to device behavior. For instance, Dalla Mora et al. [60] point out that the monitoring of the apparent current flow at an accessible probe point external to

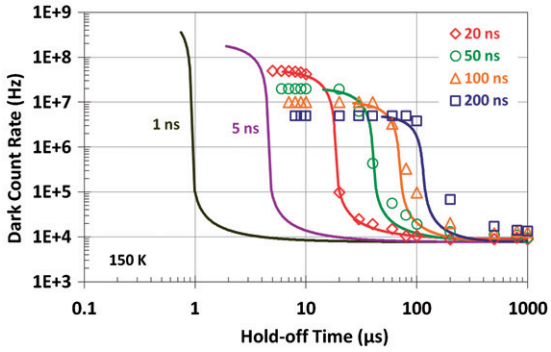

Figure 10. DCR versus hold-off time for different excess bias gate durations. Symbols are experimental data obtained from an InGaAs/InP SPAD operated with a 5 V excess bias at 150 K. Solid lines indicate modeling results; see text for details. (The color version of this figure is included in the online version of the journal.)

the device may not accurately reflect the magnitude of the internal avalanche current actually flowing through the multiplication region. To effectively mitigate afterpulsing by reducing avalanche current flow through the SPAD multiplier, an accurate description of the current distribution in the overall circuit is needed from simulations such as these.

4.3.3. *Impact of afterpulsing on DCR as a function of hold-off time*

As an example of the effects of afterpulsing on SPAD performance, we present in Figure 10 experimental data as well as modeling results for the DCR as a function of the hold-off time. These results have been obtained for a wide range of gate lengths, covering 20 to 200 ns for experiment and model, with additional modeling results included for shorter gate lengths of 1 and 5 ns. For very long hold-off times, the DCR is independent of hold-off time and defines the intrinsic background DCR.

For a timescale between gates that is shorter than a characteristic timescale τ_{AP}, afterpulsing effects become large. Presumably, this phenomenological timescale τ_{AP} is related to the physical detrapping time τ_d describing the exponentially decaying release of trapped carriers. The qualitative behavior of the measured DCR versus hold-off time is consistent with this concept given the very sharp increase in DCR for hold-off times shorter than a critical hold-off time that defines τ_{AP}. This sharp rise occurs when the probability of afterpulsing becomes sufficiently large that one afterpulse is likely to induce a subsequent afterpulse, resulting in long cascades of afterpulses.

The behavior for different gate lengths illustrates another important factor. Although avalanches were passively quenched within a few ns, the longer gate lengths present a higher overall probability for dark counts resulting in more carriers trapped per unit time. The experimental data in Figure 10, shown as symbols, clearly illustrate that more trapped carriers per unit time for longer gate lengths lead to worse afterpulsing manifested as a dramatic rise in DCR occurring at longer hold-off times. Similar effects have also been demonstrated by the use of gated quenching in which charge is allowed to flow from the moment of avalanche initiation until the end of the gate. In this scenario, longer gates will also lead to greatly enhanced carrier trapping. In either case, there is a gate-length-dependent saturation of the DCR for sufficiently short hold-off times at which afterpulses induce the maximum possible DCR of one count in every gate.

To better understand the afterpulsing effects found experimentally, we have employed a model based on work by Kang et al. [62] with which we can calculate the total DCR from several dark count generation mechanisms. We have described the details of our implementation of this model elsewhere [1,40], but what is most significant is that we assume (i) a single type of trap with a single characteristic detrapping time τ_d, and (ii) the number of trapped carriers is proportional to the total charge flow per avalanche event.

The output of the model is illustrated by the solid curves in Figure 10. The agreement with the experimental data obtained for gate durations of 20, 50, 100, and 200 ns is reasonably good. Perhaps most notable is that the characteristic hold-off time at which the DCR rises steeply varies almost linearly with gate duration, even though the same value for the characteristic detrapping time τ_d has been assumed in all simulations. Therefore, the timescale for the onset of the rapid rise in DCR is very sensitive to not only τ_d, but also the total number of filled traps. Taken together, these two factors determine the phenomenological afterpulsing timescale τ_{AP}, below which strong afterpulsing effects are exhibited. This is emphatically demonstrated by the simulated results with a much shorter gate duration of 1 ns, for which the sharp rise in DCR occurs for a hold-off time of $\sim$1 µs, even though the detrapping time in the model is $\tau_d \sim 20$ µs.

4.3.4. *Afterpulsing characterization using the double-pulse method*

As another example of afterpulsing behavior in SPADs, we describe results from the most frequently employed experimental technique used to characterize afterpulsing. The time-correlated carrier counting method [63], often referred to as the double pulse method, is a type of 'pump–probe' measurement in which traps are filled by intentionally triggering an avalanche using incident photons during a first gating

pulse, followed by a second gating pulse used to probe for secondary avalanches (i.e. afterpulses) caused by detrapped carriers. By varying the temporal separation (i.e. the hold-off time) between the two pulses, it is possible to characterize the decay of the trapped carrier population and to quantify the afterpulse probability as a function of the hold-off time. Successive pairs of pump–probe pulses are separated by a sufficiently long time that there is no afterpulsing due to previous pulse pairs. The contribution of non-afterpulse dark counts can be substracted by measuring the dark count probability in the absense of optical triggering of the first gate pulse.

The circuit used in making these measurements was a passive-quenching/active-reset (PQAR) circuit implemented by Hu et al. [7] in which passive avalanche quenching is accomplished using the very large off-state impedance of a GaAs field-effect transistor (FET) followed by rapid recharging of the SPAD with the FET in its low impedance on-state. This implementation of the PQAR circuit provided significant improvement over earlier versions [6] through the use of low-parasitic hybrid assembly of the SPAD chip, the quench/reset FET chip, and other circuit elements. As seen in Figure 11, the afterpulse probability decreases roughly as the inverse of the hold-off time (the dashed line is a power law fit with a slope of approximately −1 on a log–log plot). These data were obtained at PDE of ~10% and ~30% with an effective gate width of ~2 ns. As is usual for the double pulse method, the afterpulse probabilities plotted in the figure are for detecting an afterpulse in a *single* gate

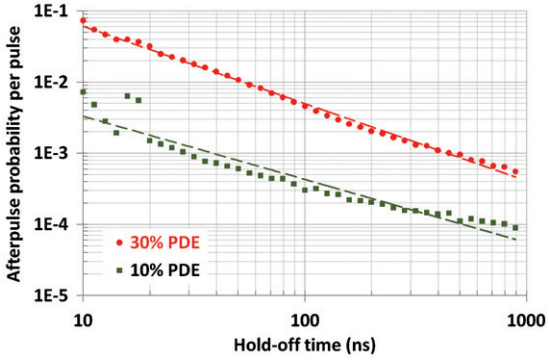

Figure 11. Afterpulse probability as a function of hold-off time measured using a PQAR circuit by the double pulse method at 230 K for PDE of 10% and 30% with ~2 ns effective gate window durations [7]. Afterpulse probabilities are for a single gate pulse occurring at a specified hold-off time. Cumulative afterpulsing for gated mode operation at a given repetition rate can be computed by summing contributions of all subsequent gates. For 10% PDE, cumulative afterpulsing is <1% for gate separations as short as ~15 ns. (The color version of this figure is included in the online version of the journal.)

pulse (i.e. the second gate pulse) after the specified hold-off time. Therefore, to extract an operationally relevant figure of merit, the 'cumulative' afterpulse probability should be computed by considering the probability of an afterpulse at *any* gate occurring after the initial trigger avalanche. For periodic gating, this amounts to summing the afterpulse probabilities for gates at all multiples of the minimum hold-off time. Given the power law dependence of the afterpulsing on hold-off time, the 'cumulative' afterpulse probability can be computed to be about a factor of 4 larger than the measured 'single gate' afterpulsing. As an example, the single gate afterpulse probability is ~0.002 at 10% PDE for hold-off times as short as ~15 ns, and the related 'cumulative' afterpulse probability is computed to be ~0.008, or 0.8%. This result indicates that afterpulsing can be limited to <1% for gated mode operation at 10% PDE at 230 K for gate repetition rates of ~65 MHz. For 30% PDE operation with this PQAR circuit, afterpulsing is approximately an order of magnitude larger.

4.3.5. *Afterpulsing with periodic short-gate operation*

For photon counting applications in which photon arrival times are known a priori, considerable performance enhancement can be realized by arming the SPAD for only short periods of time coinciding with the expected photon arrival times. In particular, the detection probabilities for dark counts and afterpulses can be reduced proportionally with the reduction in SPAD arming duration. This realization has led to the widespread use of short-gating techniques for which the SPAD is operated with gate pulse durations on the scale of 1 ns or less. Short-gate operation with periodic gates is prevalent for many communications protocols requiring single photon detection, and the focus of recent work in this area has been to increase the gate repetition frequency to rates that are compelling for communications signaling (e.g. 100s of MHz to GHz).

We have employed a short-gate technique developed by Bethune et al. [43] that is ideally suited for use with periodic communications protocols such as quantum key distribution [64]. The SPAD bias control circuitry applies periodic excess bias gates consisting of a fixed voltage swing $\Delta V \sim 4$ V with rise and fall times of ~0.1 ns and a gate plateau duration of ~0.9 ns. The voltage swing ΔV is added to a dc bias level $V_{dc} < V_b$, and the excess bias is set by adjusting the dc bias level so that $V_{ex} = V_{dc} + \Delta V - V_b$. To enable afterpulsing measurements, we use a scheme [43,65] in which 'lit' and 'dark' gates are interleaved. A pulsed diode laser source is synchronized so that single photons are temporally coincident only with the 'lit' gate pulses; for clarity, we define all odd gates as 'lit' gates and all even

gates as 'dark' gates. A laser source of the appropriate wavelength (i.e. 1.55 μm) is attenuated to generate a mean photon number of $\mu = 0.1$ per 'lit' gate. (In this case, only approximately 1 in 10 'lit' gates will actually have a photon incident on it.)

The DCR is obtained by measuring the dark count probability per gate in the absence of input photons. The PDE is determined by monitoring the total number of counts occurring in the odd 'lit' gates when the single photon source is activated. (DCR versus PDE data shown in Figure 4 were obtained in this way.) During these lit measurements, an increase in the count rate found for the even 'dark' gates (which are interleaved between the odd 'lit' gates) above the measured intrinsic DCR indicates the presence of afterpulsing and can be used to quantify the afterpulse probability. In Figure 12, we show the afterpulse probability as a function of PDE measured over a wide range of gate repetition frequency. This afterpulse probability is per detected photon and given by the ratio of the afterpulse counts to photon counts.

Through recent improvements to this short-gating circuit, we can operate at gate repetition rates as high as 50 MHz with afterpulse probabilities limited to 2.5% for 10.8% PDE and 5.2% afterpulsing for 17.5% PDE. Measurements taken at four frequencies ranging from 1 to 50 MHz show that afterpulsing probability is roughly proportional to the gate repetition frequency. This scaling is simply related to the increase in the relative duty cycle with which the SPAD is armed as the repetition frequency of the 1 ns gates increases. These results represent a five- to ten-fold increase in gate repetition rate achieved with acceptable afterpulsing compared with previous implementations of this 1 ns gating technique [45].

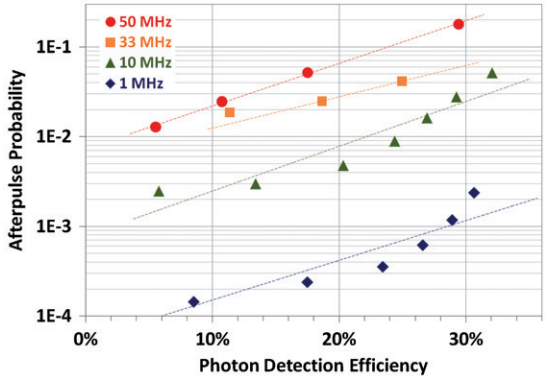

Figure 12. Measured afterpulse probability per detected photon as a function of photon detection efficiency for periodic gating with 1 ns gate pulses at four different frequencies between 1 and 50 MHz. Dashed lines indicate trends in the data. (The color version of this figure is included in the online version of the journal.)

5. Technology hierarchy for future NIR SPAD development

To provide a framework for assessing the current status of NIR SPAD performance as well as the dominant trends that have emerged in recent work on these devices, we propose a hierarchy consisting of four levels of relevant technology. At the first level consisting of materials properties, we discuss the impact of materials on DCR and afterpulsing performance and also present a comparison of InGaAs/InP SPADs to Si SPADs. At the second level, SPAD device design strategies emphasize making the best use of existing materials. In this context we describe recent interest in self-quenching SPADs, in part for reducing avalanche current flow without external circuitry, but more so for reducing the complexity of SPAD device operation. The third level of SPAD technology is the signal processing and electronic circuitry devised to extract the best possible performance from existing devices, and the most important trend at this level is recent demonstrations of much higher frequency SPAD operation. Finally, the multiplexing of large numbers of SPADs presents an opportunity to make further strides towards current priorities such as higher counting rates and also offers new capabilities such as NIR imaging with single photon sensitivity.

5.1. InP-based SPAD materials

The first level in the SPAD technology hierarchy is the underlying properties of the epitaxial material. SPAD performance is very dependent on defect types and concentrations in both the InP multiplier and the InGaAs absorber layers. Dramatic advances were achieved in the quality of the InGaAsP materials system throughout the 1990s driven by the enormous market growth for photodetectors and lasers for fiber optic telecommunications at 1.5 μm. Without a comparably large existing market opportunity to drive continued materials technology investment, further fundamental materials advances are likely to be more modest for the foreseeable future.

5.1.1. Materials impact on DCR

Materials quality is a key factor determining the SPAD DCR. As described in Section 3.2, modeling shows that the primary DCR contributions are trap-assisted tunneling in the InP multiplier and generation-recombination carrier leakage in the InGaAs absorber. Although dark counts in SPADs are in some ways analogous to dark current in linear mode APDs, it is only bulk carrier creation that impacts the SPAD DCR. Since perimeter leakage dominates the measured

dark current below the breakdown voltage, the linear mode dark current–voltage characteristics generally provide little predictive information concerning Geiger-mode DCR performance [1,66,67].

By fitting measured DCR data with our device model in which we assume one dominant deep level trap in the InP multiplier, we find this dominant trap to have an energy equal to $E_v + 0.75 \times E_g(\text{InP})$, where E_v is the valence band energy and $E_g(\text{InP})$ is the InP bandgap [39]. Similar investigations reported by researchers at MIT Lincoln Laboratory provided a very similar result [68], and these authors cite the possible correlation of this defect level with phosphorus vacancies in the InP lattice. However, there is not yet a definitive identification of the trap (or traps) responsible for the trap-assisted tunneling contribution to the DCR.

The second main contribution to SPAD DCR is dark carrier creation by generation–recombination (G-R) mechanisms in the InGaAs absorber, and these processes are common to all InGaAs/InP photodiode structures. As with all G-R leakage, dark carrier creation is thermally driven via mid-gap states in the InGaAs bandgap, and this DCR component can be depressed considerably by reduction of the operating temperature. Moderate cooling to temperatures on the order of 210 K (e.g. using thermoelectric coolers) is often sufficient to reduce the G-R contribution well below the InP trap-assisted tunneling contribution as seen in studies of the DCR activation energy [69] and other direct measurements showing tunneling-limited performance [46]. However, to achieve low DCR performance with room-temperature operation will require very significant reduction of the mid-gap state density in InGaAs, and this is a challenge that is unlikely to be met with rapid progress.

5.1.2. *Materials impact on afterpulsing*

In addition to its impact on DCR, material quality has a critical impact on afterpulsing effects in SPADs. As described earlier, this phenomenon – which involves carrier trapping and detrapping at defects in the multiplication region – is presently the main inhibitor preventing photon counting at higher repetition rates in InGaAsP-based SPADs. Techniques such as the double pulse method (see Section 4.3.4) have been used to characterize the temporal behavior of carrier detrapping by measuring the decay of the afterpulsing probability with increasing hold-off time, as illustrated in Figure 11. It has then been common practice to try to fit this measured decay by an exponential decay process with a characteristic time constant τ_d. The physical interpretation of τ_d is that it describes the exponential detrapping of carriers from defects in the multiplier. If a sufficiently narrow range of hold-off times is used (e.g. one order of magnitude or less), a single exponential of the form $R_{AP}(t) = C + A\exp(-t/\tau_d)$ will provide a reasonable fit to the afterpulsing rate $R_{AP}(t)$, where t is the hold-off time, and A and C are constants. For instance, Jensen et al. apply this analysis to afterpulsing measurements with hold-off times in the range of 1 to 10 μs, and they find a τ_d of 0.9 μs at 250 K [54]. However, when wider ranges of hold-off time have been used, a single value of τ_d does not provide accurate fits. It has been assumed that additional defects are involved with different time constants, and the model has been generalized to

$$R_{AP}(t) = A_0 + A_1\exp(-t/\tau_{d,1}) + A_2\exp(-t/\tau_{d,2}) + A_3\exp(-t/\tau_{d,3}) + \cdots, \quad (1)$$

where A_0 is the background dark count rate, and A_i and $\tau_{d,i}$ are the exponential pre-factor and detrapping time constant, respectively, associated with the ith defect type. Using such a procedure, Trifonov et al. [70] fit afterpulsing data obtained at 193 K with hold-off times ranging from 1.25 to 100 μs by using three time constants and found $\tau_{d,1} \sim 0.5$ μs, $\tau_{d,2} \sim 6.1$ μs, and $\tau_{d,3} \sim 99$ μs. Similarly, Liu et al. [67] used hold-off times ranging from 0.02 to 50 μs, and fitting of their data obtained at 240 K required four detrapping time constants $\tau_{d,1} \sim 0.07$ μs, $\tau_{d,2} \sim 0.9$ μs, $\tau_{d,3} \sim 4.2$ μs, and $\tau_{d,4} \sim 33$ μs.

The goal of detrapping studies such as these is to better understand the behavior of carrier detrapping and related afterpulsing effects as well as to ultimately identify the origin of these traps and eliminate them. However, it is sobering to review the literature on deep level traps in InP (e.g. the excellent survey by Anderson and Jiao [71]), in which a rather dense spectrum of levels has been identified in studies conducted up to 1992. The procedure of fitting double pulse method afterpulsing decay data with a sum of exponentials tends to yield an additional detrapping time constant for every additional factor of 5 to 10 in hold-off time that is used in the experiment, and the number of detrapping time constants extracted in this way is just a minimum number required for a reasonable fit to the data. It seems quite possible that this procedure may be just a mathematical exercise that does not adequately reflect the more complex reality of the materials defects in InP. We have also tried to use other measurements in addition to the double pulse method – such as dark count rate versus hold-off time data [39] similar to that presented in Figure 10, as well as free-running measurements [40] – to extract detrapping time constants, but have found comparably arbitrary results.

Furthermore, analytical techniques such as deep level transient spectroscopy (DLTS) and other

capacitive spectroscopic methods are probably not sufficiently sensitive to provide definitive information about the InP defects relevant to afterpulsing. In fact, Giudice et al. [72] turned the problem around and demonstrated the use of afterpulsing in silicon SPADs as potentially the most sensitive probe of material quality related to sensing the defects involved in afterpulsing. The use of these spectroscopic techniques, which are inherently low field techniques, is further complicated [63] by the fact that afterpulsing occurs in the presence of very high electric fields ($\sim 5 \times 10^5\,\mathrm{V\,cm^{-1}}$), for which Poole–Frenkel effects can significantly change the detrapping behavior at multiplication region defects by enhancing the carrier emission probability [73].

Despite these complexities in the afterpulsing phenomenon, there have been some valuable insights provided by the various afterpulsing measurements and models reported in the literature. For one, it appears that the number of carriers trapped per avalanche in typical small-area devices (e.g. 25–50 μm diameter) may be surprisingly small. Results in [54] indicate that the number of trapped carriers associated with afterpulsing is on the order of a few hundred, and more detailed analysis [74] of afterpulsing data in [7] points to no more than tens of trapped carriers per avalanche. Nevertheless, the physical significance of detrapping time constants mathematically extracted from the double pulse method and other measurements seems unclear, and the nature of the defects giving rise to afterpulsing in InGaAsP-based SPADs is still unknown.

5.1.3. *Comparison of InP SPADs with Si SPADs as benchmark for long-term targets*

A comparison of InP SPAD results with those of state-of-the-art Si avalanche diodes [75] is instructive because it suggests what may be possible if InGaAsP materials engineering can be brought to the level of Si materials engineering. InP-based NIR detectors will always be at a performance disadvantage relative to Si detectors, which are used at shorter (visible) wavelengths, given the necessarily smaller bandgap of the InGaAs absorbers. However, the primary impact of the absorber bandgap on SPAD performance is its role in determining the contribution to the DCR of carriers generated thermally by generation–recombination via mid-gap states. This suggests that we can remove the bandgap disparity by comparing Si and InGaAs/InP device performance at different temperatures that compensate for the difference in bandgaps.

We first consider that dark carrier thermal generation by Shockley–Read–Hall processes is proportional to $\exp(E_g/2kT)$, where E_g is the material bandgap, k is Boltzmann's constant, T is temperature, and the factor of 2 comes from thermal excitation via mid-gap states near energy levels of $E_g/2$. Given that $E_g(\mathrm{Si}) = 1.12\,\mathrm{eV}$ at 20°C, the exponent $E_g/2kT \sim 21.5$ for silicon. We then proceed to find the temperature for InGaAs at which $E_g(\mathrm{InGaAs})/2kT$ gives the same value of 21.5, noting that $E_g(\mathrm{InGaAs})$ will also be temperature-dependent. We find this equivalence at approximately −70°C, where $E_g(\mathrm{InGaAs}) \sim 0.775\,\mathrm{eV}$ [76]. Therefore, by comparing Si SPAD and InGaAs/InP SPAD performance at temperatures of 20°C and −70°C, respectively, we remove the role of the material bandgap in thermal dark carrier generation to allow a direct comparison of underlying material properties. This comparison is summarized in Table 1 assuming devices with a 50 μm active region diameter.

Table 1. Comparison of state-of-the-art performance for Si and InGaAs/InP SPADs.

	Si[a]	InGaAs/InP
Temperature	20°C	−70°C
Active region diameter	50 μm	
Wavelength	400–800 nm	1000–1600 nm
DCR and PDE[b]	10 kHz at 60%	–
	2 kHz at 40%	10 kHz at 40%
	0.5 kHz at 20%	2 kHz at 20%
	–	1 kHz at 10%
Min hold-off for 1% afterpulsing[c]	~10 ns	~100 ns
Jitter (FWHM)	30–50 ps	50–100 ps

[a]Si SPAD performance corresponds to thin Si SPAD structures as in [75]. [b]Si PDE values are cited for 550 nm, for which the highest Si PDE is obtained. [c]Assumes 20% PDE and free-running operation with fast active quenching of a few ns.

Based on the rationale just described, Si SPADs exhibit superior material quality resulting in lower DCR by about a factor of 5 for a given value of PDE. A more detailed analysis would be required to identify the contributions of the two dominant DCR mechanisms – i.e. trap-assisted tunneling in the multiplier and thermal generation in the absorber – but the present comparison gives at least an approximate measure of the improvement in DCR that could be expected if deep-level and mid-gap defect densities in InGaAs/InP SPADs could be reduced to defect densities that exist in Si SPADs. A comparison of afterpulsing performance is complicated by the fact that it is so highly circuit-dependent, so we have relied on characterization in free-running operation with fast (i.e. a few ns) active quenching using the same backend electronics [51]. Si SPADs have the potential for an order of magnitude shorter hold-off

times at 1% afterpulsing levels. While this suggests lower trap densities for Si in the multiplication region, at least some of this afterpulsing performance advantage is related to the operation of Si SPADs at considerably higher temperature allowed by its larger bandgap absorber. (We also note that it is not known whether the deep-level traps involved in DCR trap-assisted tunneling are related to the traps that give rise to afterpulsing.) Finally, although thin multiplier Si SPADs have demonstrated somewhat lower timing jitter [50] than InGaAs/InP SPADs when operated with comparable electronic circuitry, for this parameter, the two types of devices are close to parity.

5.2. *InGaAsP-based SPAD device design*

5.2.1. *Electric field profile and avalanche breakdown probability optimization*

In Section 3 we briefly summarized some of the considerations in designing InGaAs/InP SPADs. In particular, the dependence on electric field amplitude of dark carrier generation mechanisms as well as avalanche breakdown probability dictates that the optimization of DCR and PDE performance relies critically on the SPAD internal electric field profile. The lower breakdown fields of wider multiplication regions are consistent with reduced trap-assisted tunneling in the presence of lower fields, and so wide multiplication regions (e.g. $>1\,\mu m$) are beneficial for DCR without compromising PDE. On the other hand, increasing the absorber thickness to achieve higher detection efficiency also increases the thermal component of the DCR, so this trade-off needs to be balanced based on particular performance goals for DCR and PDE. Device models pertaining to InGaAsP-based SPADs described by a number of groups [38,39,41,77] allow for further optimization tailored to specific operating regimes – e.g. accommodating constraints on operating temperature, upper limits for excess bias voltage, or constraints on absolute bias voltage – but the key design advances related to the fundamental structure of InGaAs/InP avalanche diodes for Geiger-mode operation seem to have been realized.

There is also the potential for improvement to NIR SPADs through the use of higher performance materials, but promising candidates are not obvious at present. InGaAs lattice-matched to InP substrates remains the most practical high-performance material for absorption at $1.5\,\mu m$. There are perhaps more intriguing prospects for improving on InP as the multiplication region. The use of impact ionization engineered structures [78] may provide benefits, as suggested by the theoretical study of InAlAs/InP heterojunction multiplication regions [41], for which improved PDE versus DCR performance is predicted relative to InP for thinner multipliers allowing operation at lower applied voltages. However, the experimental realization of more complex multiplication regions may pose considerable challenges with respect to defect densities that exacerbate both trap-assisted tunneling and afterpulsing. From this perspective, the simplicity and maturity of InP as a comparatively high-quality SPAD multiplier material remains attractive.

5.2.2. *Avalanche charge flow reduction and self-quenching SPADs*

In Section 4.3.2, we described passively quenched SPAD operation in the context of the Haitz equivalent circuit model. Although passive quenching circuits have been investigated extensively and have significant limitations [57], they can still be compelling in situations where their simplicity is beneficial. The drawbacks to using a simple resistive load for SPAD quenching include (i) the inability to impose a hold-off time before re-charging and (ii) a typically long recharge time imposed by the time constant $R_L C_d$; refer to Figure 9 and accompanying discussion. However, there has been recent renewed interest in passive quenching for the purpose of designing self-quenching SPADs with reduced parasitics that can be multiplexed to provide higher counting rates with very simple operation. The concept for such a detector has been equated to a 'solid-state photomultiplier'. This focus for InGaAsP-based SPADs follows earlier demonstrations in silicon-based detectors [79,80] and the introduction of commercial products [81,82] based on self-quenching silicon SPADs in multiplexed configurations.

One of the key design choices for self-quenching SPADs is the magnitude of the quenching load resistance R_L. Larger R_L provides the benefit of faster quenching with less avalanche current flow but carries with it the drawback of longer recharge times and consequently slower counting rates. Since the incentive for reduced avalanche current flow is the reduction of afterpulsing to allow for high counting rates, in principle there is an optimal R_L that balances afterpulsing rate limitations and recharge rate limitations. Another consequence of long recharge times is that the instantaneous PDE of the SPAD is reduced while the device is recharging, and so the average PDE tends to be lower for higher photon arrival rates.

Recent studies of self-quenching have included the use of discretely integrated resistive loads with InGaAs/InP SPADs operating at room temperature without gating electronics [83]. Improved devices provided encouraging results with noise equivalent power on the order of $10^{-15}\,W\,Hz^{-1/2}$, although PDE

was limited to <5% and afterpulsing caused a ten-fold increase in the apparent DCR at counting rates on the order of 1 MHz relative to the background DCR without illumination [84].

To provide scalability for the eventual multiplexing of large numbers of devices, several groups have pursued the monolithic integration of passive quenching elements to demonstrate self-quenching NIR SPADs. In one approach, an InGaAsP/InAlAs heterobarrier with a large valence band offset is epitaxially integrated to act as a barrier to hole transport in an otherwise standard SPAD structure [16]. The accumulation of holes at this heterobarrier 'quenching layer' during an avalanche serves to shield the internal electric field, thereby quenching the avalanche process. A PDE of 11.5% has been achieved, but at high DCR levels of ~3 MHz at 160 K [17]. Another example of self-quenching InGaAs/InP detectors has been reported based on a discrete amplification mechanism involving avalanche multiplication in InP, although details of the device concept have not been disclosed [18].

We have pursued a design concept for self-quenching that entails the monolithic integration of thin film quenching resistors with our canonical SPAD structure [13–15]. This approach has the benefit of using the same epitaxial design that provides excellent performance for our existing SPADs while providing design flexibility through the realization of quenching elements with thin film processing on the surface of our wafers (in contrast to designs that require different epitaxy to vary quenching properties). Because these self-quenching elements provide negative feedback that counteracts the inherent positive feedback of the impact ionization process in avalanche diodes, we refer to these devices as 'negative feedback avalanche diodes', or NFADs, to distinguish them from conventional SPADs.

One of the most attractive features of the NFAD (as with all self-quenching avalanche diodes) is the simplicity of its operation: it will execute the entire Geiger mode operating cycle of avalanching, quenching, and recharging with just a dc bias voltage. Because the underlying device structure for our NFAD is our existing SPAD structure, we can achieve the same PDE and DCR as a function of excess bias for both devices. The more challenging parameter related to NFAD performance is afterpulsing probability since this device does not allow for an imposed hold-off time between quenching and recharging. Nevertheless, we have demonstrated NFAD performance with PDE as high as 10% with afterpulsing that – though considerably higher than SPAD afterpulsing with controllable hold-off times – is already acceptable for certain applications such as free space optical communications. The data in Figure 13 were obtained for two NFAD designs E3G3 and E2G6 with quenching resistances (active region diameters) of 741 kΩ (32 μm) and 992 kΩ (22 μm), respectively.

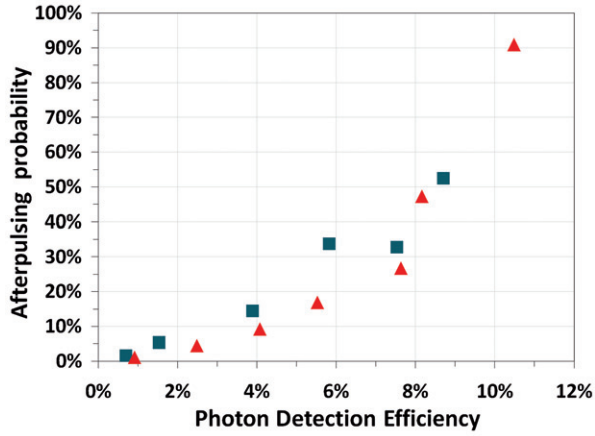

Figure 13. Afterpulsing probability as a function of PDE for two self-quenching SPADs with monolithically integrated passive quenching resistances of 740 kΩ (squares) and 990 kΩ (triangles). Data were obtained at 236 K. (The color version of this figure is included in the online version of the journal.)

For these two devices, integration over measured avalanche current pulses indicated an avalanche charge flow in the range of $(3-6) \times 10^5$ carriers for PDE in the range of 1–10%. We have also confirmed that the negative feedback of the monolithic passive quench resistance provides fairly reproducible avalanches. We have measured the statistical distribution of the avalanche charge Q with average integrated charge $\langle Q \rangle$ and standard deviation σ, and we find $\sigma/\langle Q \rangle \sim 0.3$. The significance of these avalanche statistics is that they determine whether avalanche pulses from more than one detector can be superimposed on a single output with the number of pulses still being accurately resolvable. We note that many authors have defined a 'charge excess noise' $F(Q) = 1 + \sigma^2/\langle Q \rangle^2$ as an alternate expression of the avalanche charge statistics. (It is important to realize that this charge 'excess noise' for self-quenching SPADs has nothing to do with the more established avalanche multiplication excess noise $F(M)$ used to describe fluctuations in the impact ionization avalanche process.) Our results yield an $F(Q)$ in the range of 1.08–1.09.

Finally, we have described a technique for extracting recovery times for the recharging of these devices after avalanching based on the correlation between interarrival times of consecutive avalanche pulses and the amplitude of the second pulse in this consecutive pair [69]. If the interarrival time between pulses is very short, the device will not have time to fully recharge before the arrival of the second pulse. The amplitude of

the second pulse will then be smaller, in proportion to the instantaneous value of the recharging excess bias. The saturation of the second peak amplitude plotted as a function of the interarrival time will therefore mimic the saturation in the exponential recharging of the excess bias. By analyzing the interarrival time and peak amplitude data for strings of thousands of avalanche pulses, we determined that these devices had exponential recharging time constants τ_r of 100 ns and 70 ns for E3G3 and E2G6, respectively.

The fact that self-quenching SPADs independently carry out the Geiger mode operation cycle with just a dc bias suggests that a large number of these devices can be connected in parallel with just a single common anode and cathode connection. In this configuration, they act as a single detector but with independent active regions all supplying avalanche pulses to the same output in response to single photon detections. We will return to this topic of multiplexed self-quenching devices in Section 5.4.

5.3. InP-based SPAD control circuitry and signal processing

With greater interest in applying single photon detection to communications-related applications such as quantum cryptography and free space laser communications, the limitation on counting rate posed by afterpulsing has become the primary concern for many end users. As described in the previous two sections, mitigating afterpulsing through materials improvement does not pose near-term promise, and solutions at the device design level are still nascent. Therefore, the most substantial progress in achieving higher counting repetition rates has been enabled by advances in SPAD control circuitry and signal processing. The two primary objectives of the various approaches to higher counting rates have been (i) the compensation of transient parasitics due to fast gating and (ii) the reduction of avalanche charge flow to reduce afterpulsing.

5.3.1. Compensation of parasitics related to fast gating and counting

A fundamental requirement for very high counting rates approaching GHz frequencies is the use of high-speed switching of the excess bias on sub-ns time scales. Even when such high counting rates are not required, there is still a substantial benefit to the use of very short (~1 ns) excess bias gates when photon arrival times are deterministic (as in many communications protocols) since dark count and afterpulsing probabilities are proportional to the gate duration. The sub-nanosecond rise and fall times associated with such short gates generate large capacitive transients that can couple to the output line when they are imposed on the SPAD. (For the SPADs discussed in this paper, C_d is generally in the range of 0.1–0.3 pF.) Therefore, the first requirement for short-gate circuits is the suppression of these capacitive transients to allow for the accurate detection of the often much smaller signal due to a SPAD avalanche event.

Bethune and Risk developed a transient cancellation scheme [43,44] based on the use of two matched delay lines – one inverting and one non-inverting – to linearly cancel the parasitic capacitive transients resulting from 1 ns gating, leaving any induced avalanche signal to be detected on a flat baseline. It is this type of circuit that we used to obtain the data reported in Section 4.3.5 showing 50 MHz repetition rates with 2.5% afterpulsing at 10.8% PDE. Tomita and Nakamura introduced a related concept [85] in which they used two nominally identical SPADs biased with identical gate pulses to obtain balanced outputs that could be subtracted using a 180° hybrid junction to eliminate common mode transients and facilitate measurement of the smaller SPAD avalanche signal. Finally, Zappa et al. [3] have demonstrated a monolithic circuit solution for transient cancellation by generating a 'dummy' signal on an integrated active quenching circuit that emulates the capacitive transients produced by the SPAD and subtracts the transients on-chip. This solution possesses the elegance of CMOS integration, although the generated dummy signal must be correctly matched for the particular SPAD in use.

5.3.2. Avalanche charge flow reduction by external circuitry

To achieve gate repetition rates on the order of ~1 GHz with acceptable afterpulsing using existing InGaAs/InP SPADs, appropriate transient cancellation must be augmented with adequate reduction of avalanche charge flow. By necessity, such high repetition rates require gates of sub-nanosecond duration, which are consistent with reduced charge flow. There have been two techniques exploited in the past several years that have been used to demonstrate ~2 GHz gating, and afterpulsing has been maintained at tolerable levels through the use of extremely short (~0.1–0.2 ns) effective gate widths.

The first of these schemes described by Yuan et al. [8] consists of a self-differencing circuit in which the signal response from each gate is applied to a 50 : 50 splitter so that one-half of the signal can be delayed

by exactly one gate period. This delayed signal is then subtracted from the non-delayed half of the signal from the next gate period. In this fashion, identical transients reproduced during each gate period are subtracted, leaving the less frequent net avalanche signal to be detected when it occurs. (This solution bears some conceptual resemblance – albeit at much higher frequencies – to the scheme of Bethune and Risk in that it uses the subtraction of delayed transients from the same detector.) The most recent results [9] using the self-differencing circuit have shown about 1.4% afterpulsing at 11.8% PDE with a gate repetition rate of 2 GHz.

The second new scheme to show multi-GHz repetition rates is the sine-wave gating method described by Namekata et al. [10]. The innovation of this method is to avoid the generation of capacitive transients in the first place by gating the SPAD with a purely sinusoidal gate signal. The elimination of the gate signal is then facilitated by narrow notch filtering at the gating frequency, leaving any avalanche signal (which will have broad spectral content outside the narrow filtered band) to be detected. The most recent results [11] of the sine-wave gating technique have demonstrated 2 GHz repetition frequencies with 3.4% afterpulsing at 10.5% PDE. Sine-wave gating circuit results at ~2 GHz have also been demonstrated by Zhang et al [12].

In Table 2, we have summarized various measurements discussed in this paper for which the focus was higher frequency operation. The repetition rates achieved for matched delay lines and the PQAR circuit represent a ten-fold improvement over earlier results with these circuit types, but their present limitation to ≥1 ns gate durations and lower frequency circuitry restricts them to rates far below those achieved with sine-wave gating and the self-difference circuit. Finally, we want to stress the fact that the range of performance seen for the different circuits in Table 2 was achieved with SPADs of the same pedigree fabricated by the authors. Although good SPAD device quality is a prerequisite for high performance, the operating circuit design and signal processing implementation are of paramount importance in governing the end performance realized with very high repetition rate SPAD operation.

5.4. *Multiplexing of InP-based SPADs*

Subject to whatever constraints exist on SPAD device and circuit performance, the multiplexing of large numbers of detectors is arguably the highest level solution available today for improving photon counting performance. Following a detection event in a single SPAD, the detector is inactive, either during active circuit hold-off times or during the early part of a passive circuit recharging period. This delay in the resetting of the detector to sense additional photons limits counting rate and causes a so-called 'blocking' problem at high photon flux. However, if a collection of N detectors are used in parallel, the reset time of one detector results in a blocking problem of only order $1/N$ assuming that incident photons are uniformly spread among all N detectors.

One scheme for exploiting the benefits of detector multiplexing is the use of intelligent switching among a collection of discrete detectors [86]. As soon as the first detector is triggered, subsequent input photons are switched to a second detector. When the second detector is triggered, input is switched to a third detector, and so on. Ideally, by the time the last detector is triggered, the first detector has recovered, and the cycle repeats. A first experimental proof-of-principle of the concept was presented in [87], and principle challenges include the availability of sufficiently fast (~1 ns) optical switches and the impact of optical losses on PDE.

The more prevalent approach to SPAD multiplexing has been fabricating multiple detectors in array or matrix formats so that input photons can be optically spread among all the detectors. In addition to higher

Table 2. Repetition frequency, PDE, and afterpulsing probability for SPADs with various types of operating circuitry.

SPAD Circuit type	Repetition frequency	PDE (%)	Afterpulse probability (%)	Ref.
Matched delay lines	50 MHz	10.8	2.5	this work
		17.5	5.2	
Passive-quench/active-reset	65 MHz	10	0.8	[7]
Sine-wave gating	2 GHz	10.5	3.4	[11]
	2.23 GHz	10	8.3	[12]
Self-differencing	2 GHz	11.8	1.4	[9]
		23.5	4.8	

count rates, this concept also possesses the potential for some degree of photon number resolution even though the individual SPAD detectors are insensitive to photon number.

5.4.1. *Multiplexing with large arrays of individually addressable devices*

The most familiar example of large-scale multiplexing of detectors is in the context of imaging arrays. Pioneering work has been done by researchers at MIT Lincoln Laboratory for the development of arrays of InGaAsP-based SPADs optimized for operation at either 1.06 or 1.55 μm [68] that are intended to serve as sensor engines for three-dimensional (3D) laser ranging and detection (LADAR) imaging systems [88,89]. In these sensors, every array pixel provides an independent time-of-flight distance measurement using laser ranging techniques to provide the third spatial dimension complimenting the usual x–y pixel coordinate data to create 3D image point clouds that can be rendered as 3D images. The use of SPADs allows for pixels with single photon sensitivity, and arrays as large as 256×64 elements have been demonstrated.

Pixel arrays provide images when all pixel data is captured and read out as a single frame. For the purpose of higher frequency photon counting, identical detector arrays can be used if the backend electronics instead read pixel data asynchronously only when a specific pixel has been triggered. While triggered pixels have an imposed hold-off time to avoid afterpulsing, untriggered pixels are left active. This greatly increases the photon flux that can be accommodated (relative to discrete detectors) before blocking problems start to saturate the counting rate. Lincoln Lab SPAD arrays in an 8×8 configuration were used in this fashion to demonstrate high speed communications links with single photon sensitivity [68,90].

There are several challenges specific to the development of arrays of SPADs that are not present for arrays of other detector technologies. To establish individually addressable pixels for either image capture or high counting rates, SPAD arrays must be hybridized to appropriate CMOS readout integrated circuits (ROICs) to put a SPAD in series with ROIC circuitry at every pixel. If a SPAD pixel is defective and can not support the typical Geiger-mode bias voltages on the order 40 to 80 V without generating excessive leakage current, the connected CMOS circuitry will be damaged and the entire array may be rendered unusable. This situation requires either extremely good SPAD pixel yield or highly specialized ROIC protection circuitry.

Another challenge unique to SPAD arrays is the mitigation of optical crosstalk. Every photon detection event involves avalanche charge flow in which the acceleration of charge in the high-field avalanche region can give rise to photon emission by hot carrier luminescence at the rate of one photon per $\sim 10^5$ carriers that flow through the avalanche region. Because all array pixels are sensitive to single photons, the coupling of emitted photons to neighboring pixels can cause correlated dark counts at the neighbors that are defined as crosstalk events. Strategies for the reduction of this optical crosstalk include reducing the avalanche charge flow and introducing structures that the provide optical isolation or absorption outside the pixel active regions [91–93].

To illustrate the state-of-the-art in the fabrication of arrays of diffused-junction planar-geometry InGaAs/InP SPADs described in this review, we show data in Figure 14 for DCR and PDE obtained from every pixel of a 32×32 focal plane array (FPA) with 100 μm pixel pitch. These FPAs consist of the SPAD photodiode array hybridized by indium flip-chip bonding to a CMOS ROIC, with a GaP microlens array attached to the backside of the SPAD array to maintain at least 70% optical fill factor. The FPA is hermetically packaged, and temperature control is provided by an integrated thermoelectric cooler.

As Figure 14 illustrates, array-level maps of DCR and PDE indicate excellent performance with 100% pixel yield. The DCR map in Figure 14(*a*) shows that *all* 1024 pixels have DCR less than 50 kHz, with an average DCR of 28 kHz and standard deviation of 6.5 kHz using relatively modest cooling to a temperature of -20°C. For the PDE performance map in Figure 14(*b*), the mean PDE is 22.2%, with a standard deviation of 4.6%. Somewhat lower values of DCR and PDE near the edges of the FPA are explained by process-related variations in V_b, as opposed to wafer-level variability in V_b caused by epi-growth parameter gradients that lead to performance variability over longer length scales.

In Figure 15, we have plotted the dependence of DCR on PDE for a random selection of pixels from the FPA maps in Figure 14. Although there is some variability in DCR versus PDE across the array, we note that the spread in performance is much smaller than the variability seen in Figure 4 for discrete devices. We have confirmed that a sampling of the performance of discrete devices over an area comparable to our array dimension (i.e. 3.2 mm × 3.2 mm) has a distribution similar to that seen in Figure 4. This finding serves as evidence that factors other than intrinsic device reproducibility are responsible for the degree of performance variation found for discrete SPADs. We will address this topic again briefly in our concluding discussion.

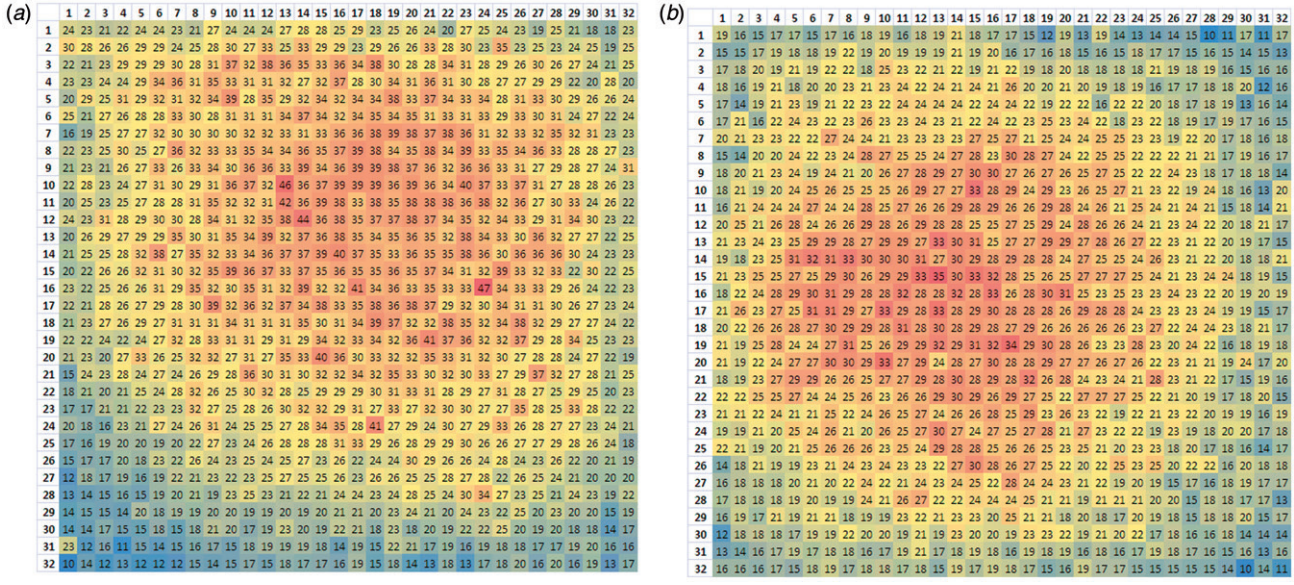

Figure 14. Performance maps of all 1024 pixels of a 32×32 InGaAs/InP (1.55 μm) SPAD FPA operating with an excess bias of 3.25 V at 253 K. (*a*) Dark count rate (DCR) in kHz for all pixels. *All* pixels are <50 kHz. (*b*) Photon detection efficiency (PDE) in % for all pixels, where the average pixel PDE of 22% includes all optical losses related to the microlens array and other sources of insertion loss. (The color version of this figure is included in the online version of the journal.)

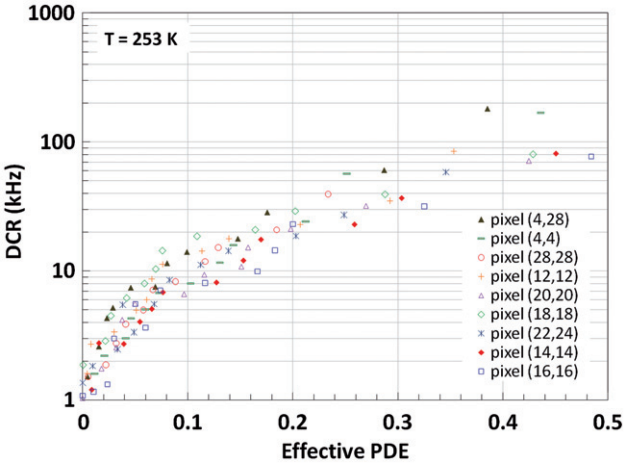

Figure 15. Dependence of DCR on the effective PDE at −20°C for a random sample of InGaAs/InP SPAD pixels from the performance maps presented in Figure 14. Regardless of position on the FPA, pixels show consistent DCR versus PDE behavior. (The color version of this figure is included in the online version of the journal.)

5.4.2. *Prospects for multiplexing with self-quenching diodes*

Although SPAD arrays with individually addressable pixels – as in the focal plane array just described – can provide high performance and functionality, there is substantial complexity and cost associated with this solution. One reason that overhead is unavoidable is that each SPAD pixel must have its own control circuitry for biasing and quenching. A simpler alternative exists in the use of self-quenching SPADs such the NFADs described in Section 5.2.2 since these devices require no control circuitry other than input–output connections for applying a dc bias and collecting the pulse responses to avalanches.

Moreover, the ability of self-quenching SPADs to independently execute the avalanche, quench, and recharge cycle allows multiple active regions to be connected in parallel to act as a single detector with a single pair of I/O connections. Pulses from all active regions are superimposed, and as with arrays of separate pixels, even if one self-quenching region has just avalanched and is temporarily blocked from sensing additional photons, the availability of the remaining regions can greatly reduce the blocking problem and therefore the limitation to the counting rate. The availability of commercial products employing this concept using silicon-based detectors [81,82] at least serves as a proof-of-feasibility for a comparable solution using InGaAsP-based self-quenching SPADs.

As mentioned in Section 5.2.2, matrices of parallel self-quenching SPADs have been touted as a solid state analog to photomultiplier tubes given their potential for single photon sensitivity over a large area with each individual active region of the detector acting independently of the other regions. In fact, a matrix of self-quenching SPADs is very similar to microchannel plate photomultiplier tubes

(MCP-PMTs), in which a matrix of parallel microcapillaries, or microchannels, with a large electric field applied along their length provides charge multiplication of a single injected photoelectron by repeated secondary electron generation resulting from electron collisions with an electron-emissive material coating the microchannel walls. Secondary electron emission in a particular microchannel induces charging of its wall and a consequent strong reduction in the local electric field, with that microchannel unable to provide further amplification until its initial charge state is restored [94]. The microchannel recharging phenomenon, which takes a time in the microsecond range, is very similar to the recharging required for triggered active regions in a matrix of self-quenching SPADs.

At present, there are still considerable challenges to realizing high performance devices based on multiplexed self-quenching SPADs. For one, the DCR of such a device will be the sum of the DCRs for all of the individual active regions connected together. To obtain substantial increases in counting rate by using at least ~100 active regions in parallel, the user will have to tolerate DCRs that are 100 times larger than they are for a discrete SPAD. There is also more progress needed to further reduce afterpulsing effects, as indicated by the NFAD results for afterpulsing versus PDE in Figure 13. Optical crosstalk due to hot carrier luminescence must also be managed, as in the case of SPAD FPAs, although our work on these FPAs suggests that cumulative crosstalk probabilities per avalanche can be limited to ~10% or less at target operating conditions for PDE and DCR [95].

6. Final discussion and summary

In this review, we have aimed to present an overview of the state-of-the-art in InGaAs/InP SPAD performance for single photon detection at NIR wavelengths spanning 0.9–1.6 μm. We have described that while PDE, DCR, and timing jitter have reached levels that are suitable for many applications, the recent emphasis on higher counting rates has prompted concerted efforts to improve the mitigation of afterpulsing effects in these devices.

With regard to current device performance, it is interesting to note that the fairly wide distribution of DCR versus PDE results seen for discrete devices is probably not inherent to the underlying device structure given the much higher degree of performance consistency realized for array pixels in our 32×32 format FPAs. Device design, epitaxial growth, and wafer processing details are essentially identical for the discrete devices and the arrays; if anything, the array processing is more complex due to the introduction of optical isolation trenches for crosstalk reduction and additional features to provide cathode contacts on the front side of the wafer. These results strongly suggest that factors other than intrinsic materials quality and wafer-level device fabrication contribute signficantly to the performance variation seen in our discrete devices. Such factors may include die singulation, die mounting to ceramic carriers, wire bonding, and other chip-level packaging processes, as well as back-end electronics. Narrowing this distribution towards the best demonstrated performance for DCR versus PDE is a challenge to be met going forward, and further study is required to isolate specific root causes for performance variation.

In our discussion of the hierarchy of technologies involved in determining what performance can ultimately be obtained from SPADs, we summarized the current situation pertaining to the materials properties of InGaAs/InP SPADs. We also explained why we believe that achieving improvements in material quality that would be sufficient to bring significant device performance improvements is likely to be a slow, challenging process. The use of different III–V semiconductor materials within the framework of the existing device structure may present interesting opportunities, but newer materials with potentially favorable properties are likely to suffer – at least initially – from worse material quality. A comparison of the properties of InGaAs/InP SPADs with those of silicon SPADs was presented to show what level of performance improvement might be expected with further maturing of the InGaAsP materials system. Performance gains from this level of improved materials quality alone would be respectable but are probably limited to five- to ten-fold improvements for parameters like DCR and afterpulsing.

Instead, we believe that there are much better near-term prospects for improved SPAD performance brought about by innovations in device design and circuit implementation. We presented a brief discussion of recent work on self-quenching SPADs as an example of device-level design concepts that may provide photon counting solutions with significant reduction in operational complexity relative to canonical SPADs. At least for the present, however, it appears that continued interest in self-quenching devices will primarily be driven by ease-of-use considerations since the challenges of device operation without control signals are likely to always force performance trade-offs relative to discrete devices operated with well-designed external circuitry.

There have also been efforts to achieve single photon detection using avalanche diode structures operated in linear mode. Linear mode devices historically have been operated at fairly modest gains no greater than $\sim 10^2$,

and higher gains must be achieved to detect single photons. In contrast, canonical Geiger mode operation has involved charge flow on the order of 10^7 to 10^8 carriers per detection event, and the recent emphasis of the work on these devices has been to reduce the charge generated per avalanche to minimize undesirable charge trapping that leads to afterpulsing. In many respects, these two distinct approaches (i.e. linear mode and Geiger mode) seem to be converging to an operating regime in which they may behave very similarly. The ideal carrier generation per detection event will provide minimal carrier trapping consistent with being large enough to ensure accurate measurement. Whether achieved with linear mode or Geiger mode operation, the residual afterpulsing and related counting rate limitations will be the same.

For the realization of higher photon counting rates, the biggest strides have clearly been made with the implementation of improved circuit designs and signal processing. Most notably, the ability to count photons at rates on the scale of 1 GHz has been demonstrated for periodic gating using very short gates with the self-differencing and sine-wave gating techniques. Similar improvements to provide broad operating capability for free-running detection pose a compelling challenge for continued work of this nature.

Acknowledgements

We wish to acknowledge the many collaborators who have provided valuable feedback on our NIR SPAD device performance and insightful discussions related to this device technology.

References

[1] Itzler, M.A.; Ben-Michael, R.; Hsu, C.-F.; Slomkowski, K.; Tosi, A.; Cova, S.; Zappa, F.; Ispasoiu, R. *J. Mod. Opt*. **2007,** *54,* 283–304.

[2] Itzler, M.A.; Jiang, X.; Nyman, B.; Ben-Michael, R.; Slomkowski, K. InP-based single photon avalanche diodes. *IEEE LEOS Annual Meeting Conference Proceedings (LEOS '08)*, Newport Beach, CA, November 9–13, 2008.

[3] Zappa, F.; Tosi, A.; Cova, S. *Proc. SPIE* **2007,** *6583,* 65830E.

[4] Verghese, S.; Donnelly, J.P.; Duerr, E.K.; McIntosh, K.A.; Chapman, D.C.; Vineis, C.J.; Smith, G.M.; Funk, J.E.; Jensen, K.E.; Hopman, P.I.; Shaver, D.C.; Aull, B.F.; Aversa, J.C.; Frechette, J.P.; Glettler, J.B.; Liau, Z.L.; Mahan, J.M.; Mahoney, L.J.; Molvar, K.M.; O'Donnell, F.J.; Oakley, D.C.; Ouellette, E.J.; Renzi, M.J.; Tyrrell, B.M. *IEEE J. Sel. Top. Quantum Electron*. **2007,** *13,* 870–886.

[5] Itzler, M.A.; Entwistle, M.; Owens, M.; Patel, K.; Jiang, X.; Slomkowski, K.; Rangwala, S.; Zalud, P.F.; Senko, T.; Tower, J.; Ferraro, J. *Proc. SPIE* **2010,** *7808,* 78080C.

[6] Liu, M.; Hu, C.; Campbell, J.C.; Pan, Z.; Tashima, M.M. *IEEE J. Quantum Electron*. **2008,** *44,* 430–434.

[7] Hu, C.; Zheng, X.; Campbell, J.C.; Onat, B.M.; Jiang, X.; Itzler, M.A. *Proc. SPIE* **2010,** *7681,* 76810S.

[8] Yuan, Z.L.; Kardynal, B.E.; Sharpe, A.W.; Shields, A.J. *Appl. Phys. Lett*. **2007,** *91,* 041114.

[9] Yuan, Z.L.; Sharpe, A.W.; Dynes, J.F.; Dixon, A.R.; Shields, A.J. *Appl. Phys. Lett*. **2010,** *96,* 071101.

[10] Namekata, N.; Sasamori, S.; Inoue, S. *Opt. Express* **2006,** *14,* 10043–10049.

[11] Namekata, N.; Adachi, S.; Inoue, S. *IEEE Photonics Technol. Lett*. **2010,** *22,* 529–531.

[12] Zhang, J.; Eraerds, P.; Walenta, N.; Barreiro, C.; Thew, R.; Zbinden, H. *Proc. SPIE* **2010,** *7681,* 76810Z.

[13] Itzler, M.A.; Jiang, X.; Nyman, B.; Slomkowski, K. *Proc. SPIE* **2009,** *7222,* 72221K.

[14] Jiang, X.; Itzler, M.A.; Nyman, B.; Slomkowski, K. *Proc. SPIE* **2009,** *7320,* 732011.

[15] Itzler, M.A.; Jiang, X.; Onat, B.M.; Slomkowski, K. *Proc. SPIE* **2010,** *7608,* 760829.

[16] Zhao, K.; Zhang, A.; Lo, Y.-H.; Farr, W. *Appl. Phys. Lett*. **2007,** *91,* 081107.

[17] Zhao, K.; You, S.; Cheng, J.; Lo, Y.-H. *Appl. Phys. Lett*. **2008,** *93,* 153504.

[18] Linga, K.; Yevtukhov, Y.; Liang, B. *Proc. SPIE* **2009,** *7320,* 73200Z.

[19] Kingston, R.H. *Detection of Optical and Infrared Radiation*; Springer-Verlag: Berlin, 1978; Chapter 2.

[20] Gisin, N.; Ribordy, G.; Tittel, W.; Zbinden, H. *Rev. Mod. Phys*. **2002,** *74,* 145–195.

[21] Levine, B.F.; Bethea, C.G.; Campbell, J.C. *Appl. Phys. Lett*. **1985,** *46,* 333–335.

[22] Lacaita, A.L.; Zappa, F.; Bigliardi, S.; Manfredi, M. *IEEE Trans. Electron Devices* **1993,** *40,* 57–82.

[23] Measures, R.M. *Laser Remote Sensing: Fundamentals and Applications*; Wiley: New York, 1988.

[24] Special issue on 'Free-space communication techniques for optical networks', *IEEE LEOS Newsl*. **2005**, *19*, 6–39.

[25] Niedre, M.J.; Patterson, M.S.; Giles, A.; Wilson, B.C. *Photochem. Photobiol*. **2005,** *81,* 941–943.

[26] Albota, M.A.; Aull, B.F.; Fouche, D.G.; Heinrichs, R.M.; Kocher, D.G.; Marino, R.M.; Mooney, J.G.; Newbury, N.R.; O'Brien, M.E.: Player, B.E.; Willard, B.C.; Zayhowski, J.J. *MIT Lincoln Lab. J*. **2002,** *13,* 351–370.

[27] Aull, B.F.; Loomis, A.H.; Young, D.J., Stern, A.; Felton, B.J.; Daniels, P.J.; Landers, D.J.; Retherford, L.; Rathman, D.D.; Heinrichs, R.M.; Marino, R.M.; Fouche, D.G.; Albota, M.A.; Hatch, R.E.; Rowe, G.S.; Kocher, D.G.; Mooney, J.G.; O'Brien, M.E.; Player, B.E.; Willard, B.C.; Liau, Z.-L.; Zayhowski, J.J. *Proc. SPIE* **2004,** *5353,* 105–116.

[28] Knill, E.; Laflamme, R.; Milburn, G.J. *Nature* **2001,** *409,* 46–52.

[29] Rarity, J.G.; Tapster, P.R. *Phys. Rev. Lett*. **1990,** *64,* 2495–2498.

[30] Itzler, M.A.; Loi, K.K.; McCoy, S.; Codd, N.; Komaba, N. High-performance, manufacturable avalanche photodiodes for 10 Gb/s optical receivers. *Proceedings of 25th Optical Fiber Communication Conference (OFC 2000)*, Baltimore, MD, March 7–10, 2000.

[31] Nishida, K.; Taguchi, K.; Matsumoto, Y. *Appl. Phys. Lett*. **1979,** *35,* 251–253.

[32] Campbell, J.C.; Dentai, A.G.; Holden, W.S.; Kasper, B.L. *Electron. Lett*. **1983,** *19,* 818–820.

[33] Forrest, S.R.; Kim, O.K.; Smith, R.G. *Appl. Phys. Lett*. **1982,** *41,* 95–98.

[34] Liu, Y.; Forrest, S.R.; Hladky, J.; Lange, M.J.; Olsen, G.H.; Ackley, D.E. *J. Lightwave Technol*. **1992,** *10,* 182–193.

[35] Itzler, M.A.; Loi, K.K.; McCoy, S.; Codd, N.; Komaba, N. Manufacturable planar bulk-InP avalanche photodiodes for 10 Gb/s applications. *Proceedings of 12th Annual Meeting of Lasers and Electro-Optics Society (LEOS '99)*, San Francisco, CA, November 8–11, 1999.

[36] Hayat, M.M.; Saleh, B.E.A.; Teich, M.C. *IEEE Trans. Electron Devices* **1992,** *39,* 546–552.

[37] Zappa, F.; Lovati, P.; Lacaita, A. Temperature dependence of electron and hole ionization coefficients in InP. *Proceedings of the 8th International Conference on Indium Phosphide and Related Materials (IPRM)*, Schwabisch-Gmund, Germany, April 21–25, 1996.

[38] Donnelly, J.P.; Duerr, E.K.; McIntosh, K.A.; Dauler, E.A.; Oakley, D.C.; Groves, S.H.; Vineis, C.J.; Mahoney, L.J.; Molvar, K.M.; Hopman, P.I.; Jensen, K.E.; Smith, G.M.; Verghese, S. *IEEE J. Quantum Electron*. **2006,** *42,* 797–809.

[39] Jiang, X.; Itzler, M.A.; Ben-Michael, R.; Slomkowski, K. *IEEE J. Sel. Top. Quantum Electron*. **2007,** *13,* 895–905.

[40] Jiang, X.; Itzler, M.A.; Ben-Michael, R.; Slomkowski, K.; Krainak, M.A.; Wu, S.; Sun, X. *IEEE J. Quantum Electron*. **2008,** *44,* 3–11.

[41] Ramirez, D.A.; Hayat, M.M.; Itzler, M.A. *IEEE J. Quantum Electron*. **2008,** *44,* 1188–1195.

[42] Ramirez, D.A.; Hayat, M.M.; Karve, G.; Campbell, J.C.; Torres, S.N.; Saleh, B.E.A.; Teich, M.C. *IEEE J. Quantum Electron*. **2006,** *42,* 137–145.

[43] Bethune, D.S.; Risk, W.P. *IEEE J. Quantum Electron*. **2000,** *36,* 340–347.

[44] Bethune, D.S.; Risk, W.P.; Pabst, G.W. *J. Mod. Opt*. **2004,** *51,* 1359–1368.

[45] Itzler, M.A.; Jiang, X.; Ben-Michael, R.; Nyman, B.; Slomkowski, K. *Proc. SPIE* **2008,** *6900,* 69001E.

[46] Tosi, A.; Dalla Mora, A.; Zappa, F.; Cova, S.; Itzler, M.A.; Jiang, X. *Proc. SPIE* **2009,** *7222,* 72221G.

[47] Spinelli, A.; Lacaita, A.L. *IEEE Trans. Electron Devices* **1997,** *44,* 1931–1943.

[48] Tan, C.H.; Ng, J.S.; Rees, G.J.; David, J.P.R. *IEEE J. Sel. Top. Quantum Electron*. **2007,** *13,* 906–910.

[49] Adachi, S. *Handbook on Physical Properties of Semiconductors*; Kluwer Academic: Dordrecht, 2004.

[50] Gulinatti, A.; Maccagnani, P.; Rech, I.; Ghioni, M.; Cova, S. *Electron. Lett*. **2005,** *41,* 272–274.

[51] Tosi, A.; Dalla Mora, A.; Zappa, F.; Cova, S. *J. Mod. Opt*. **2009,** *56,* 299–308.

[52] Krainak, M.A. Photoionization of trapped carriers in avalanche photodiodes to reduce afterpulsing during Geiger-mode photon counting. *Conference on Lasers and Electro-Optics (CLEO 2005)*, Baltimore, MD, May 22–27, 2005.

[53] Krainak, M.A. NASA Goddard Space Flight Center, Greenbelt, MD. Private communication, 2007.

[54] Jensen, K.E.; Hopman, P.I.; Duerr, E.K.; Dauler, E.A.; Donnelly, J.P.; Groves, S.H.; Mahoney, L.J.; McIntosh, K.A.; Molvar, K.M.; Napoleone, A.; Oakley, D.C.; Verghese, A.; Vineis, C.J.; Younger, R.D. *Appl. Phys. Lett*. **2006,** *88,* 133503.

[55] Zappa, F.; Tosi, A.; Cova, S. *Proc. SPIE* **2007,** *6583,* 65830E.

[56] Haitz, R.H. *J. Appl. Phys*. **1964,** *35,* 1370–1376.

[57] Cova, S.; Ghioni, M.; Lacaita, A.; Samori, C.; Zappa, F. *Appl. Opt*. **1996,** *35,* 1956–1976.

[58] Hayat, M.M.; Itzler, M.A.; Ramirez, D.A.; Rees, G.J. *Proc. SPIE* **2010,** *7608,* 76082B.

[59] Hayat, M.M.; Ramirez, D.A.; Rees, G.J.; Itzler, M.A. *Proc. SPIE* **2010,** *7681,* 76810W.

[60] Dalla Mora, A.; Tosi, A.; Tisa, S.; Zappa, F. *IEEE Photonics. Technol. Lett*. **2007,** *19,* 1922–1924.

[61] Zappa, F.; Tosi, A.; Dalla Mora, A.; Tisa, S. *Sens. Actuators, A* **2009,** *153,* 197–204.

[62] Kang, Y.; Lu, H.X.; Lo, Y.-H.; Bethune, D.S.; Risk, W.P. *Appl. Phys. Lett*. **2003,** *83,* 2955–2957.

[63] Cova, S.; Lacaita, A.; Ripamonti, G. *IEEE Electron Device Lett*. **1991,** *12,* 685–687.

[64] Bethune, D.S.; Risk, W.P. *J. Quantum Electron*. **2000,** *36,* 340–347.

[65] Ben-Michael, R.; Itzler, M.A.; Nyman, B.; Entwistle, M. *2006 Digest of the LEOS Summer Topical Meetings*, IEEE: Piscataway, NJ, 2006; pp 15–16.

[66] Itzler, M.A.; Jiang, X.; Ben-Michael, R.; Nyman, B.; Slomkowski, K. *Proc. SPIE* **2008,** *6900,* 69001E.

[67] Liu, M.; Hu, C.; Bai, X.; Guo, X.; Campbell, J.C.; Pan, Z.; Tashima, M.M. *IEEE J. Sel. Top. Quantum Electron*. **2007,** *13,* 887–894.

[68] Verghese, S.; Donnelly, J.P.; Duerr, E.K.; McIntosh, K.A.; Chapman, D.C.; Vineis, C.J.; Smith, G.M.; Funk, J.E.; Jensen, K.E.; Hopman, P.I.; Shaver, D.C.; Aull, B.F.; Aversa, J.C.; Frechette, J.P.; Glettler, J.B.; Liau, Z.L.; Mahan, J.M.; Mahoney, L.J.; Molvar, K.M.; O'Donnell, F.J.; Oakley, D.C.; Ouellette, E.J.; Renzi, M.J.; Tyrrell, B.M. *IEEE J. Sel. Top. Quantum Electron*. **2007,** *13,* 870–886.

[69] Itzler, M.A.; Jiang, X.; Entwistle, M.; Onat, B.M.; Slomkowski, K. *Proc. SPIE* **2010,** *7681,* 76810V.

[70] Trifonov, A.; Subacius, D.; Berzanskis, A.; Zavriyev, A. *J. Mod. Opt*. **2004,** *51,* 1399–1415.

[71] Anderson, W.A.; Jiao, K.L. Deep Levels in InP and Related Materials. In *Indium Phosphide and Related Materials: Processing, Technology, and Devices*: Katz, A., Ed.; Artech House: Boston, 1992; Chapter 3.

[72] Giudice, A.C.; Ghioni, M.; Cova, S.; Zappa, F. A process and deep level evaluation tool: afterpulsing in avalanche junctions. *33rd Conference on European Solid-State Device Research (ESSDERC)*, Estoril, Portugal, September 16–18, 2003.
[73] Ghioni, M.; Gulinatti, A.; Rech, I.; Maccagnani, P.; Cova, S. *Proc. SPIE* **2008,** *6900,* 69001D.
[74] Hu, C. Advanced Devices and Circuits for Single Photon Avalanche Diodes. Ph.D. Dissertation, University of Virginia, Charlottesville, VA, 2009.
[75] Ghioni, M.; Gulinatti, A.; Rech, I.; Zappa, F.; Cova, S. *J. Sel. Top. Quantum Electron.* **2007,** *13,* 852–862.
[76] Paul, S.; Roy, J.B.; Basu, P.K. *J. Appl. Phys.* **1991,** *69,* 827–829.
[77] Pellegrini, S.; Warburton, R.E.; Tan, L.J.J.; Ng, J.S.; Krysa, A.B.; Groom, K.; David, J.P.R.; Cova, S.; Robertson, M.J.; Buller, G.S. *IEEE J. Quantum Electron.* **2006,** *42,* 397–403.
[78] Yuan, P.; Wang, S.; Sun, X.; Zheng, X.G.; Holmes, A.L. Jr; Campbell, J.C. *IEEE Photonics Technol. Lett.* **2000,** *12,* 1370–1372.
[79] Saveliev, V. *Nucl. Instrum. Methods A* **2004,** *535,* 528–532.
[80] Mazzillo, M.; Condorellli, G.; Piazza, A.; Sanfilippo, D.; Valvo, G.; Carbone, B.; Fallica, G.; Billotta, S.; Belluso, M.; Bonanno, G.; Pappalardo, A.; Cosentino, L.; Finocchiaro, P. *Nucl. Instrum. Methods A* **2008,** *591,* 367–373.
[81] Stewart, A.G.; Greene-O'Sullivan, E.; Herbert, D.J.; Saveliev, V.; Quinlan, F.; Wall, L.; Hughes, P.J.; Mathewson, A.; Jackson, J.C. *Proc. SPIE* **2006,** *6119,* 61190A.
[82] Yamamoto, K.; Yamamura, K.; Sato, K.; Kamakura, S.; Ota, T.; Suzuki, H.; Ohsuka, S. Newly developed semiconductor detectors by Hamamatsu. *Proceedings of International Workshop on New Photon-Detectors (PD 07)*, Kobe, Japan, June 27–29, 2007; PoS(PD07)004.
[83] Warburton, R.E.; Itzler, M.; Buller, G.S. *Appl. Phys. Lett.* **2009,** *94,* 071116.
[84] Warburton, R.E.; Itzler, M.A.; Buller, G.S. *Electron. Lett.* **2009,** *45,* 996–997.
[85] Tomita, A.; Nakamura, K. *Opt. Lett.* **2002,** *27,* 1827–1829.
[86] Castelletto, S.A.; Degiovanni, I.P.; Schettini, V.; Migdall, A.L. *J. Mod. Opt.* **2007,** *54,* 337–352.
[87] Schettini, V.; Polyakov, S.V.; Degiovanni, I.P.; Brida, G.; Castelletto, S.; Migdall, A.L. *J. Sel. Top. Quantum Electron.* **2007,** *13,* 978–983.
[88] Albota, M.A.; Aull, B.F.; Fouche, D.G., Heinrichs, R.M.; Kocher, D.G.; Marino, R.M.; Mooney, J.G.; Newbury, N.R.; O'Brien, M.E.; Player, B.E.; Willard, B.C.; Zayhowski, J.J. *MIT Lincoln Lab. J.* **2002,** *13,* 351–370.
[89] Aull, B.F.; Loomis, A.H.; Young, D.J.; Stern, A.; Felton, B.J.; Daniels, P.J.; Landers, D.J.; Retherford, L.; Rathman, D.D.; Heinrichs, R.M.; Marino, R.M.; Fouche, D.G.; Albota, M.A.; Hatch, R.E.; Rowe, G.S.; Kocher, D.G.; Mooney, J.G.; O'Brien, M.E.; Player, B.E.; Willard, B.C.; Liau, Z.-L.; Zayhowski, J.J. *Proc. SPIE* **2004,** *5353,* 105–116.
[90] Verghese, S.; Cohen, D.M.; Dauler, E.A.; Donnelly, J.P.; Duerr, E.K.; Groves, S.H.; Hopman, P.I.; Jensen, K.E.; Liau, Z.-L.; Mahoney, L.J.; McIntosh, K.A.; Oakley, D.C.; Smith, G.M. *IEEE LEOS Newsl.* **2005,** *19,* 24–25.
[91] Younger, R.D.; McIntosh, K.A.; Chludzinski, J.W.; Oakley, D.C.; Mahoney, L.J.; Funk, J.E.; Donnelly, J.P.; Verghese, S. *Proc. SPIE* **2009,** *7320,* 73200Q.
[92] Itzler, M.A.; Entwistle, M.; Owens, M.; Jiang, X.; Patel, K.; Slomkowski, K.; Koch, T.; Rangwala, S.; Zalud, P.F.; Yu, Y.; Tower, J.; Ferraro, J. *Proc. SPIE* **2009,** *7320,* 73200O.
[93] Rech, I.; Ingargiola, A.; Spinelli, R.; Labanca, I.; Marangoni, S.; Ghioni, M.; Cova, S. *Proc. SPIE* **2007,** *6771,* 677111.
[94] Gatti, E.; Oba, K.; Rehak, P. *IEEE Trans. Nucl. Sci.* **1983,** *30,* 461–468.
[95] Itzler, M.A.; Entwistle, M.; Owens, M.; Patel, K.; Jiang, X.; Slomkowski, K.; Rangwala, S.; Zalud, P.F.; Senko, T.; Tower, J.; Ferraro, J. *Proc. SPIE* **2010,** *7780,* 77801M.