Journal of Modern Optics
Vol. 54, Nos. 2–3, 20 January–15 February 2007, 283–304

Taylor & Francis
Taylor & Francis Group

Single photon avalanche diodes (SPADs) for 1.5 μm photon counting applications

M. A. ITZLER*†, R. BEN-MICHAEL†, C. -F. HSU†, K. SLOMKOWSKI†, A. TOSI‡, S. COVA‡, F. ZAPPA‡ and R. ISPASOIU§

†Princeton Lightwave Inc., 2555 US Route 130 S, Cranbury, NJ 08512, USA
‡Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan I-20133, Italy
§Credence Systems Corp. – DCG, 1299 Orleans Dr., Sunnyvale, CA 94089, USA

(*Received 10 February 2006; in final form 20 April 2006*)

The paper reports on the design and characterization of InGaAs/InP single photon avalanche diodes (SPADs) for photon counting applications at wavelengths near 1.5 μm. It is shown how lower internal electric field amplitudes can lead to reduced dark count rates, but at the expense of degraded afterpulsing behaviour and larger timing jitter. Dark count rate behaviour provides evidence of thermally assisted tunnelling with an average thermal activation energy of ~0.14 eV between 150 K and 220 K. Afterpulsing behaviour exhibits a structure-dependent afterpulsing activation energy, which quantifies how carrier de-trapping varies with temperature. SPAD performance simultaneously exhibits a dark count rate of 10 kHz at a detection efficiency of 20% with timing jitter of 100 ps at 200 K, and with appropriate performance tradeoffs, we demonstrate a 200 K dark count rate as low as 3 kHz, a detection efficiency as high as 45%, and a timing jitter as low as 30 ps.

1. Introduction

The need for higher performance fibre optic telecommunications receivers spurred rapid advancement during the last decade in the understanding and performance of InP-based avalanche photodiodes (APDs) for the near-infrared wavelength range from 1.0 to 1.7 μm. However, progress related to the telecom receiver-based 'linear mode' operation of these devices, for which output photocurrent is proportional to input optical power, has only very recently begun to impact the performance and availability of single photon avalanche diodes (SPADs) based on similar device design and materials platforms. In contrast to linear mode APDs, SPADs operate in the so-called 'Geiger mode', in which a single photon can trigger a macroscopic current pulse, providing the ability to accurately sense the arrival at the detector of a single photon. The state-of-the-art in commercially available near-infrared SPADs leaves substantial room for improvement because the overwhelming majority of

*Corresponding author. Email: mitzler@princetonlightwave.com

Journal of Modern Optics
ISSN 0950–0340 print/ISSN 1362–3044 online © 2007 Taylor & Francis
http://www.tandf.co.uk/journals
DOI: 10.1080/09500340600792291

avalanche photodetectors developed for this wavelength range have been designed for linear mode performance, and system developers who have sought devices with good Geiger mode performance have been forced to sample the various commercially available telecom APDs and characterize their photon counting performance [1–5].

Researchers have recently found that the optimization of InP-based SPADs for detecting single photons requires design approaches that are quite distinct from those shown to be effective in optimizing APD linear mode performance, primarily because the most critical performance attributes for linear mode APDs (such as excess noise and gain–bandwidth product) are irrelevant for SPADs. In this paper, we present recent work we have done to design and characterize InGaAs/InP SPADs. In particular, we show the substantial performance differences introduced by sometimes subtle structural variations introduced by growth and processing variations in different regions of a SPAD wafer. In section 2, we give an overview of avalanche diode design concepts, review key SPAD performance parameters, and describe some of the salient differences between the design of SPADs and linear mode APDs. In section 3, we present data obtained for our devices pertaining to detection efficiency and dark count rate in the absence of afterpulsing. Detailed data and analysis on afterpulsing effects are presented in section 4, followed by timing jitter performance in section 5. In section 6, we summarize our results.

2. Properties and design concepts of InGaAs/InP SPADs

By virtue of the photoelectric effect, positive-intrinsic-negative (PIN) InGaAs/InP photodetectors can efficiently detect infrared light in the fibre optic communications wavelength range from 1300–1600 nm. However, their responsivity is quantum-limited to, at best, one electron–hole pair generated per input photon. Avalanche photodetectors benefit from internal gain due to a process known as impact ionization that leads to multiple electron–hole pairs per input photon. Applying a larger reverse voltage to the avalanche detector will result in a larger gain, until the breakdown voltage V_{br} is reached. For bias voltages larger than V_{br}, the electron–hole generation process can become self-sustaining and result in a runaway avalanche.

Traditional optical communications receivers make use of linear mode avalanche photodiodes (APDs), for which the output photocurrent is linearly proportional to the intensity of the optical input. Internal APD gain can provide significant improvement in high-bandwidth optical receiver sensitivity. In contrast to linear mode operation below V_{br}, if an avalanche photodetector is biased above V_{br}, then a single photoexcited carrier can induce a runaway avalanche that gives rise to an easily detectable macroscopic current. In this case, the detector is sensitive to a single photon input and is referred to as a single photon avalanche diode (SPAD). This mode of operation is often referred to as ‘Geiger mode’ because of its similarity to Geiger–Muller detectors, in which particle emission from radioactive materials gives rise to an avalanche of carriers from ionized gas atoms.

Linear mode APD performance has a noise floor determined by the shot noise associated with leakage current, or dark current, that exists in the absence of input photons. In contrast, SPAD performance is degraded by false counts, or dark counts, that arise when carriers are created by processes other than photoexcitation. Both thermal excitation and field-mediated creation of free carriers (i.e. tunnelling processes) contribute to the dark count rate (DCR). To improve SPAD performance in the presence of significant DCR, InGaAs/InP SPADs are usually operated in gated mode. The detector is biased at a baseline voltage just below the breakdown voltage, and to 'arm' the detector, a gate pulse is applied to bring the detector bias above breakdown for a short period of time, generally between 1 and 100 ns. For applications in which the photon arrival time is known, a shorter gate pulse can be used to reduce the likelihood of a dark count being generated within the gate. Once an avalanche is initiated, it must be quenched. Gated quenching allows the avalanche to persist until the bias is reduced below V_{br} according to the fixed gate duration. Passive quenching employs a resistor in series with the APD so that the avalanche current induces a voltage across the resistor and drops the APD bias below V_{br}. Active quenching uses circuitry to rapidly detect the presence of an avalanche and actively force the APD bias below V_{br}.

2.1 *Fundamental InGaAs/InP avalanche photodetector design concepts*

All InP-based avalanche photodetectors deployed today are based on the separate absorption and multiplication (SAM) regions structure [6]. Figure 1 shows a schematic representation of our basic device design platform. This design entails a narrow bandgap $In_{0.53}Ga_{0.47}As$ layer ($E_g \sim 0.75$ eV at 295 K), lattice-matched to InP,

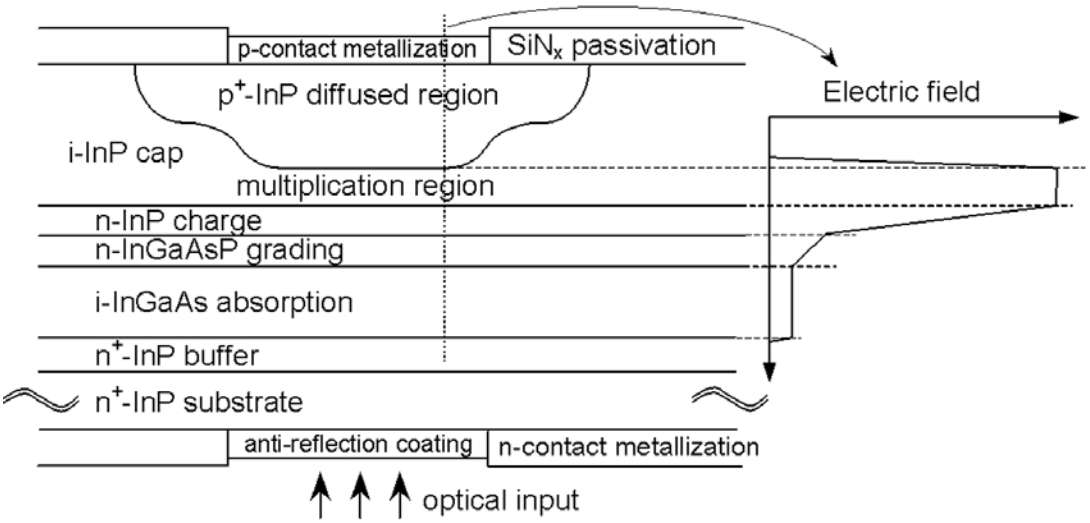

Figure 1. Schematic cross-section of SPAD device structure with a separate absorption, charge and multiplication (SACM) region structure. The charge layer maintains high field to generate avalanche gain in the multiplication region while keeping sufficiently low field in the absorption region to minimize field-induced leakage currents.

used to absorb photons with the wavelength of interest (e.g. 1550 nm) adjacent to a wider bandgap InP region ($E_g \sim 1.35$ eV) in which avalanche multiplication occurs. A primary goal of the design is to maintain low electric field in the narrow bandgap absorber (to avoid dark carriers due to tunnelling) while maintaining sufficiently high electric field in the multiplication region (so that impact ionization effects lead to significant avalanche multiplication). The inclusion of a charged layer between the absorption and multiplication regions (the SACM structure [7]) allows for more flexible tailoring of the internal electric field profile, along with the associated avalanche process, and is common to many InP-based avalanche diodes used today. The addition of grading layers between the InGaAs and InP layers in the structure is important to reduce carrier (specifically hole) trapping effects that result from the valence band offset that arises in an abrupt InGaAs/InP heterojunction [8].

The lateral structure of our design employs a buried p–n junction to guarantee edge breakdown suppression, low perimeter leakage, and high reliability. The device active area geometry is determined by the patterning of the SiN dielectric passivation layer to create a diffusion mask. A subsequent diffusion of Zn dopant atoms creates a p^+-InP region within the i-InP cap layer (see figure 1). The quasi-cylindrical junction that is formed by a single diffusion exhibits electric field enhancement where the junction curvature is maximum and leads to premature avalanche breakdown at the edges of the device. To suppress this edge breakdown phenomenon, we use two diffusions to tailor the junction profile [9] so that the junction is deeper in the central part of the active area than it is in the junction periphery. This design ensures that the gain profile across the central part of the active region is uniform and that the gain is reduced in the peripheral region of the device. With this buried junction design, the formation of a high-quality SiN passivation layer can guarantee low perimeter leakage and extremely stable long-life performance (e.g. over 25 years operation for telecom use). The use of this platform for the fabrication of linear mode APDs is described further in [10–12].

2.2 *Overview of SPAD performance parameters*

SPADs have a number of fundamental performance parameters that differ from linear mode APD parameters, and their relative importance may depend on the end application. In all applications, it is desirable to minimize the occurrence of false signals, or dark counts, that result when avalanches are seeded by carriers originating by a mechanism other than photo-excitation. Likewise, high detection efficiency is always desirable and is determined by two factors: (i) the probability that an input photon photoexcites a carrier which then reaches the multiplication region (i.e. the detector's quantum efficiency), and (ii) the probability that a carrier which enters the multiplication region successfully initiates a self-sustaining avalanche (i.e. the avalanche breakdown probability).

The importance of other photon counting performance parameters tends to be more application dependent. A specific form of dark counts known as 'afterpulsing' is caused by the trapping and subsequent release of carriers induced during a previous avalanche. The additional dark counts caused by this afterpulsing effect

can be avoided if the quench period (with APD bias below breakdown) between successive overbias gating windows is long enough to allow trapped carriers to detrap. The reduction of afterpulsing by long quench periods requires a sufficiently slow repetition rate, i.e. the rate at which single photons can be counted. For some applications, this tradeoff between afterpulsing and repetition rate forces significant compromises in overall performance. Afterpulsing also imposes another tradeoff with regard to improving dark count rate: lower operating temperatures can be used to reduce the intrinsic dark count rate, but carrier trapping and the consequent afterpulsing effects are typically exacerbated at lower temperatures due to longer trap lifetimes.

Another photon counting parameter of significance is the temporal variation, or timing jitter, of the current pulses corresponding to single photon detection. SPAD timing jitter results from the fluctuations in temporal correspondence between the arrival of a photon and the detection of a resulting avalanche. There are a number of potential mechanisms that contribute to device-level jitter, including statistical variations in the absorption location of photons, which results in a variation in the distance photoexcited carriers must travel, and the amount of time required for each avalanche pulse to reach a pre-determined threshold for detection. In particular, fluctuations in the way a runaway avalanche spreads laterally across the active area of a SPAD have been shown to dominate the timing jitter for single photon detection in larger area devices [13].

2.3 *Differences between SPAD and linear mode APD design*

Despite structural similarities, SPAD and linear mode APD design optimizations differ significantly. Linear mode InGaAs/InP APDs should achieve an optimal electric field profile below breakdown at modest gains of 10–20, since it is within this range of gain that the combination of high-bandwidth (e.g. >1 GHz) APDs and amplifiers generally achieves its maximum signal-to-noise ratio. On the other hand, with SPADs intended for operation above breakdown, their electric field profile should be optimized for operation at the target overbias. In particular, the total field control charge (see figure 1) should be larger in a SPAD than in a linear mode APD to maintain low field amplitude in the absorption region while supporting higher field amplitude in the multiplication region to realize Geiger mode operation.

Thinner multiplication regions are desirable to reduce the noise of linear mode APDs because they result in more deterministic linear mode avalanche processes. This is a manifestation of the so-called 'dead space effect' [14], in which an impact ionization event causes a carrier to lose a large fraction of its energy and not regain a significant probability of impact ionizing again until it has travelled far enough to acquire sufficient kinetic energy. Since APD gain–bandwidth product is inversely proportional to the multiplication region width, linear mode devices also require sufficiently thin multiplication regions to achieve their desired bandwidth (~0.5 μm and ~0.2 μm for 2.5 and 10 Gb/s, respectively) at typical operating gains of 10. In contrast, SPAD performance benefits from much thicker multiplication regions (~1.0 μm) due to the fact that their breakdown probability increases significantly

faster with increasing relative overbias voltage than for the thin multiplication regions typical of linear mode APDs [15].

In linear mode devices, it is generally required that the p–n junction depletion 'punches' through the absorption region at low voltage so that high bandwidth can be achieved at low gains. This early punchthrough is important to assure high dynamic range in fibre optic receivers. However, since SPADs are intended to operate solely above V_{br}, the constraints of early punchthrough are absent. The considerable difference in operating temperature, with SPADs generally cooled to $-60^{\circ}C$ or lower, must also be carefully accounted for. Finally, there is also a crucial distinction between the detection processes for the two types of devices: linear mode APDs are employed as analogue devices, while the detection of SPAD avalanche pulses is inherently digital.

3. Overview of SPAD performance characteristics

3.1 *Linear mode behaviour as an indicator of SPAD performance*

Although we have pointed out many differences between linear mode and Geiger mode performance, it is still of interest to consider whether certain aspects of linear mode behaviour can be correlated with good SPAD performance. In figure 2, we illustrate the linear mode current–voltage (I–V) characteristics of one of our typical 25 μm active area diameter SPADs, where the active area diameter is determined by the more deeply diffused region that defines the multiplication region (see figure 1). We operationally define the breakdown voltage V_{br} as the voltage for which the linear mode dark current I_d reaches 10 μA. The onset of the photocurrent response occurs when the p–n junction depletion first extends into

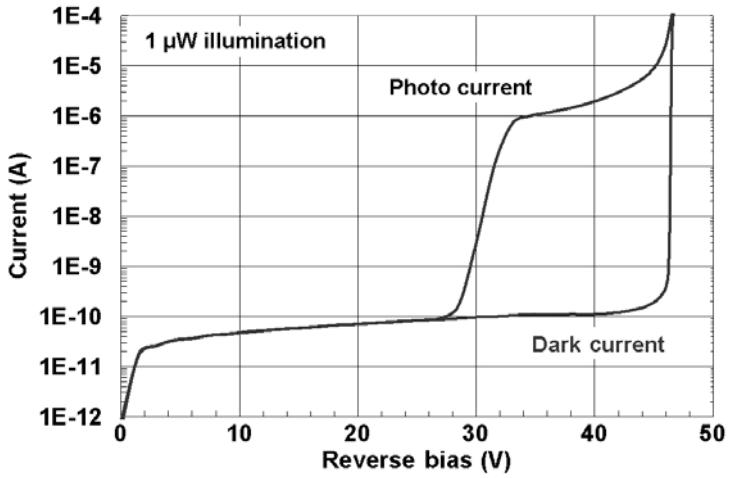

Figure 2. Typical dark and illuminated (1 μW) I–V characteristics for a 25 μm diameter SPAD.

the InGaAs absorption region. The smooth increase in photocurrent, to gains in excess of 100 before V_{br} is reached, is consistent with a uniform gain profile and the absence of edge breakdown effects.

The dark *I–V* characteristic exhibits a number of significant features. Most importantly, I_d between 30 V and a few volts below V_{br} is nearly independent of voltage. If I_d were dominated by bulk leakage from the narrow bandgap InGaAs or other layers (or interfaces) within the structure, this leakage current would be multiplied. In this case, we would expect the shape of I_d to closely match the shape of the photocurrent *I–V*, which is determined almost exclusively by the multiplication of photoexcited carriers from the InGaAs. The absence of any apparent dark current multiplication until just before breakdown suggests that perimeter leakage currents dominate the dark current behaviour in linear mode. We have confirmed for devices ranging from 25 μm to 1 mm taken from the same wafer that for sufficiently large diameters (i.e. 1 mm), bulk leakage dominates and dark *I–V* curves exhibit the same gain characteristic as photocurrent *I–V* curves.

The dark count rate (DCR) of a SPAD depends solely on bulk leakage since a dark carrier must pass through the multiplication region if it is to seed a run-away avalanche that will be detected as a dark count. I_d may be large due to significant perimeter leakage current (e.g. due to mediocre SiN passivation properties), but since only bulk leakage affects the DCR, the device might still be a good SPAD. Alternatively, if I_d exhibits the same gain behaviour as the photocurrent, then it is dominated by bulk leakage that will manifest itself in the DCR in Geiger mode. To give a quantitative example, if I_d originates entirely from bulk leakage and has a value of 0.1 nA at a gain of 10, the primary dark current (i.e. the dark current that will be multiplied when it enters the multiplication region) is 0.01 nA. If we assume that the dark carrier generation is comparable immediately above breakdown, this would correspond to about 6×10^7 dark counts per second (given that 1 electron per second is 1.6×10^{-19} A) at room temperature.

For a dark *I–V* characteristic which remains voltage-independent, its value just below breakdown provides only a worst-case bound on SPAD DCR performance. Since bulk leakage contributes only a small fraction to the total dark current just below V_{br} in this case, the DCR will be much lower than would be calculated for an estimate based on the linear mode I_d.

The data in figure 2 suggest that at room temperature, this device is dominated by perimeter leakage, with a noticeable contribution from bulk leakage starting to contribute for bias voltages only within ~5 V of V_{br}. On a larger area device (125 μm diameter) from this same wafer, we found that dark *I–V* curves taken at 150 K showed almost complete independence of voltage between 20 V and $V_{br} \sim 37$ V, with an extremely sharp onset of breakdown and $I_d \sim 20$–30 fA just below V_{br}. Based on these results, we estimate that the bulk leakage at V_{br} is no greater than 1 fA, which would correspond to a bulk carrier generation of about 10^4 electrons per second. At room temperature, the carrier generation rate (before multiplication) of this 125 μm diameter SPAD is about 10^5 times larger than the 150 K rate. This factor is consistent with thermally activated bulk generation current through mid-gap states in the InGaAs absorption region.

3.2 *Wafer-level performance variations*

The SPADs used in this study were obtained from wafers grown by metallorganic chemical vapour deposition (MOCVD). Although the MOCVD growth technology for the InGaAsP quaternary system on InP substrates has seen enormous advances during the past decade, one still finds wafer-level growth variations that can be significant with respect to device performance. In particular, variation in epitaxial layer doping concentrations and thicknesses leads to variations in the internal electric field profile of fabricated devices, and this field profile determines most of the significant SPAD performance characteristics. Regions of the wafer with the greatest deviation in layer doping and thickness from 'typical' values tend to yield devices with the greatest difference in performance from typical performance characteristics. Often these regions are near the edge of the wafer.

SPAD breakdown voltage V_{br} is particularly sensitive to internal layer structure. In figure 3(*a*), we present V_{br} data taken along a 25 mm long column of detectors indicated by the arrow in figure 3(*b*). In the central portion of the 50 mm diameter wafer, the variation in V_{br} is approximately ± 0.2 V. Towards the wafer edge, V_{br} values drift by at least 1 V. We have taken advantage of this variation in device characteristics to compare geometrically identical devices taken from different parts of the wafer. In particular, much of the data presented in this paper have been obtained from the four devices whose positions on the wafer are indicated schematically in figure 3(*b*). Devices 40E and 40C have 40 μm diameter active areas and were taken from the edge region and central region of the wafer, respectively. Similarly, 25E and 25C have 25 μm diameter active areas and were taken from edge and central regions, respectively. By focusing on performance comparisons for these two pairs of devices, we will show the extent to which fairly modest epitaxial structure variations can affect SPAD performance characteristics.

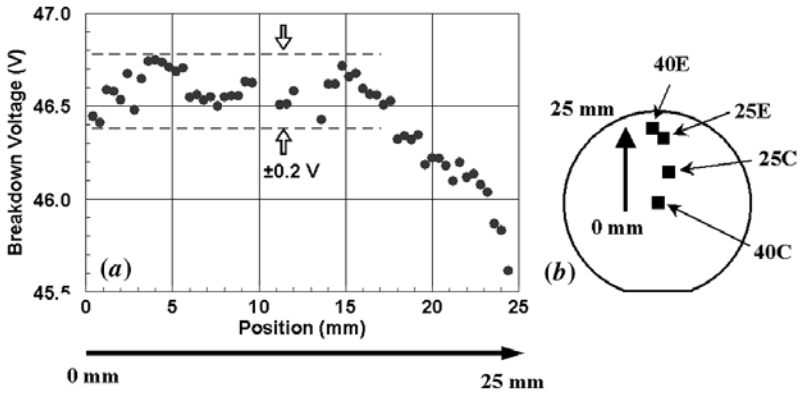

Figure 3. (*a*) Wafer-level variation of breakdown voltage along 25 μm extent showing systematic deviation near wafer edge. (*b*) Schematic illustration of positions on wafer of SPADs used in this study. 25E and 40E are from edge region; 25C and 40C are from centre region.

3.3 *Tradeoff between dark count rate and detection efficiency*

For a given SPAD, detection efficiency can be increased by operating at a higher overbias voltage V_{ov}, where $V_{ov} \equiv V - V_b$ is the bias in excess of the breakdown voltage V_b. For low overbias, the breakdown probability increases linearly with V_{ov}; as overbias increases, the breakdown probability becomes sub-linear in V_{ov} and asymptotically approaches unity [16]. However, the DCR also increases with higher V_{ov} and usually does so exponentially, consistent with a leakage mechanism involving tunnelling at the low operating temperatures (e.g. 200 K) typical for InGaAs/InP SPADs.

A plot of DCR versus detection efficiency is therefore linear on a semi-log plot, as illustrated in figure 4, in which we present data for the two 25 μm diameter devices described above. These data were obtained with a device temperature of 200 K, 200 ns gating with active quenching, and a 10 kHz gate repetition rate. Device 25E, taken from close to the edge of the wafer, has a DCR lower by a factor of 4 to 5 at any given detection efficiency compared to device 25C. The performance of device 25E, which exhibited a DCR ~ 4000 Hz at 25% detection efficiency, is comparable to the best results reported for InGaAs/InP SPADs operated under similar conditions. For clarity's sake, we point out that the units used for DCR in this paper (counts per second, or Hz) refer to the number of dark counts that would be detected per second if the device were free-running (i.e. continuously biased above breakdown) with instantaneous quenching and reset following each avalanche and assuming no afterpulsing. For DCR versus detection efficiency characterization, we used a setup that was developed for measurements in which the photon arrival was not synchronous, and a 200 ns gate optimized the overall system performance given the anticipated photon arrival rate and detector DCR. The time response of the active quenching circuit is 5–10 ns. Detection efficiency measurements were made using a continuous wave 1310 nm laser diode attenuated to optical powers

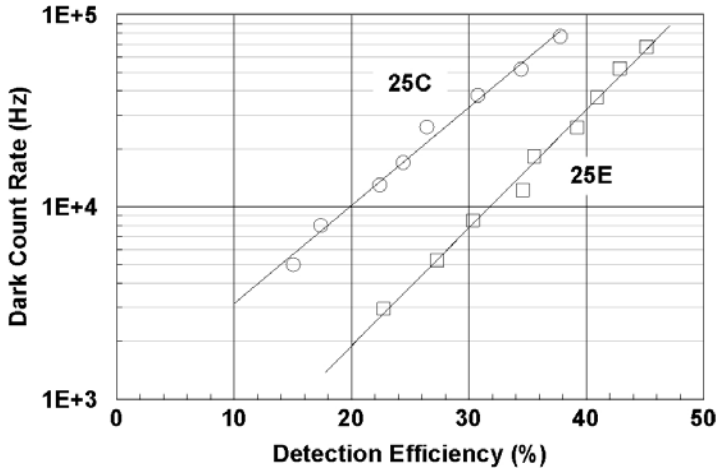

Figure 4. Dark count rate versus detection efficiency for two 25 μm diameter devices SPADs operated at 200 K using 200 ns gating with active quenching and a 10 kHz gate repetition rate.

of ~0.1 pW. Initial unattenuated output power and the attenuation used both had NIST-traceable calibrations.

Device simulations show that the structural differences in epitaxial structure, particularly a larger integrated field control charge and increased layer thicknesses, lead to reduced electric fields in the multiplication and absorption regions for 25E (relative to 25C) with a consequently lower DCR. However, the improvement in DCR comes at the cost of lower performance with respect to carrier dynamics, as seen in afterpulsing and jitter measurements presented in later sections.

3.4 *Dark count rate activation energy*

To demonstrate the dependence of DCR on both temperature T and overbias voltage V_{ov}, in figure 5 we plot DCR data obtained for device 40C as a function of $1/kT$, where k is Boltzman's constant. Data were taken using different values of V_{ov} between 2 and 6 V in increments of 1 V. Using a linear fit to $\ln(DCR)$ versus T, we extract a thermal activation energy E_a at each value of V_{ov}. As seen in the figure, a single fit to the data at the four temperatures used (150 K, 175 K, 200 K, and 220 K) yields an activation energy $E_a \sim 0.13$–0.14 eV for all values of overbias. Despite its differences in other performance characteristics, device 40E yielded similar values for E_a over the same range of V_{ov} values.

At any given operating temperature, the internal electric field in a particular layer of the device structure is linearly related to the applied bias voltage. Therefore, the DCR data in figure 5 also indicate an exponential dependence on the applied electric field. Taken in combination, the exponential dependence of DCR on both temperature and overbias is consistent with thermally assisted tunnelling. The thermal E_a of ~0.14 eV suggests thermal activation to a moderately deep

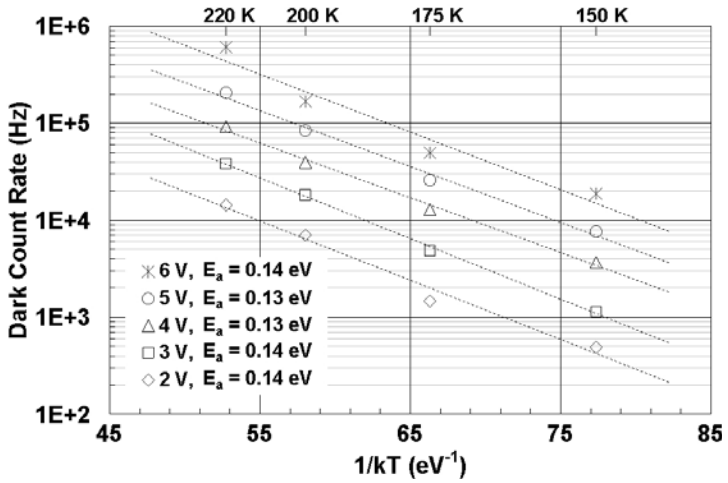

Figure 5. Dark count rate plotted versus $1/kT$, for temperature T in the range 150 K to 220 K, measured at five different overbias values from 2 to 6 V in 1 V increments for device 40C. Exponential fits for data points at each bias voltage all yield a thermal activation energy in the range 0.13–0.14 eV.

energy defect in the bandgap, with consequent tunnelling allowing carriers to cross the bandgap.

Because electric fields in the wide bandgap ($E_g \sim 1.4$ eV) InP multiplication region are calculated to be too low to lead to significant tunnelling, we believe that the thermally assisted tunnelling occurs in the narrow bandgap ($E_g \sim 0.8$ eV) InGaAs absorption region. For room temperature operation of similar devices, with InGaAs electric fields designed to be well below 150 kV/cm, one finds that bulk dark current is dominated by generation current, which arises from thermal excitation of carriers through mid-gap defect states. The corresponding activation energy for this process is ~0.35–0.40 eV. As long as the InGaAs electric field just above V_{br} is well below the tunnelling limit, we would expect that DCR at room temperature is also dominated by generation current with a similarly large activation energy. Upon closer inspection, the data in figure 5 confirms a trend towards higher activation energy at higher temperature. Considering data from all five V_{ov} values, the average slope between 175 K and 150 K gives $E_a \sim 0.11$ eV, while between 220K and 200 K, the average slope gives $E_a \sim 0.17$ eV. We suspect that the apparent E_a found at moderate temperatures around 200 K is actually a convolution of values corresponding to thermally activated tunnelling and generation current, with the thermally activated tunnelling component likely to be closer to 0.10 eV. In future work, a more detailed analysis of dark current behaviour will help to quantify these different contributions to the DCR.

The magnitude of the thermal activation energy is of interest as an indication of the primary defect type leading to the measured DCR as well as for making quantitative assessments of the change of DCR with temperature. For instance, given $E_a \sim 0.17$ eV for temperatures near 200 K, one finds that the DCR changes by a factor of about 1.5 for every 10 K increment in temperature. Note that this factor will depend on both E_a and the temperature of interest.

It is worth mentioning another mechanism shown to give rise a rather small DCR thermal activation energy in III–V SPADs. Karve *et al*. [17] have shown that large electric fields in the InAlAs multiplication layer of an InGaAs/InAlAs SPAD lead to tunnelling in this layer as the dominant DCR mechanism. The temperature dependence of the InAlAs bandgap then manifests itself as an effective activation energy of ~0.11 eV. However, these devices had thin (0.4 μm) multiplication regions, requiring excessively high fields to reach Geiger mode operation. A similar analysis accounting for the electric field in our InP multiplication region and the temperature dependence of the InP bandgap leads to an effective activation energy <0.02 eV. It is also interesting to note that the data in [17] indicate a strong tendency toward larger activation energy at higher temperatures.

4. Dark count rate and afterpulsing properties

As explained above, afterpulsing is defined to be a dark count induced by the release of a carrier trapped by a defect in the multiplication region during an earlier avalanche event. If the period between successive overbias gate pulses is sufficiently

long, carriers trapped during one avalanche are able to de-trap before the following gate is applied, and afterpulsing does not occur. If the 'hold-off' time between successive gates is made sufficiently short, the measured DCR increases sharply due to afterpulsing effects. The time scale for this critical hold-off time is related to the characteristic de-trap time τ_d, and to the extent that detrapping is thermally induced, τ_d becomes larger at lower temperature.

To study afterpulsing, we measured the DCR as a function of the hold-off time T_{off} between successive gates. In this measurement setup, a closed-loop helium-gas cryostat cools the SPAD and the front-end electronics. A pulse generator feeds the SPAD cathode with the gate pulse, while the DC bias voltage is supplied to the anode. The avalanche signal is read from the anode side by means of AC coupling. A discriminator detects the rising edge of the avalanche current pulse and feeds the counting/timing electronics. We present data for devices 40E and 40C in figure 6(*a*) and 6(*b*), respectively. The measurement conditions include an overbias $V_{ov} \sim 6\,V$ applied for 20 ns gate pulses (i.e. $T_{on} = 20\,ns$) with gated quenching (i.e. the gate is

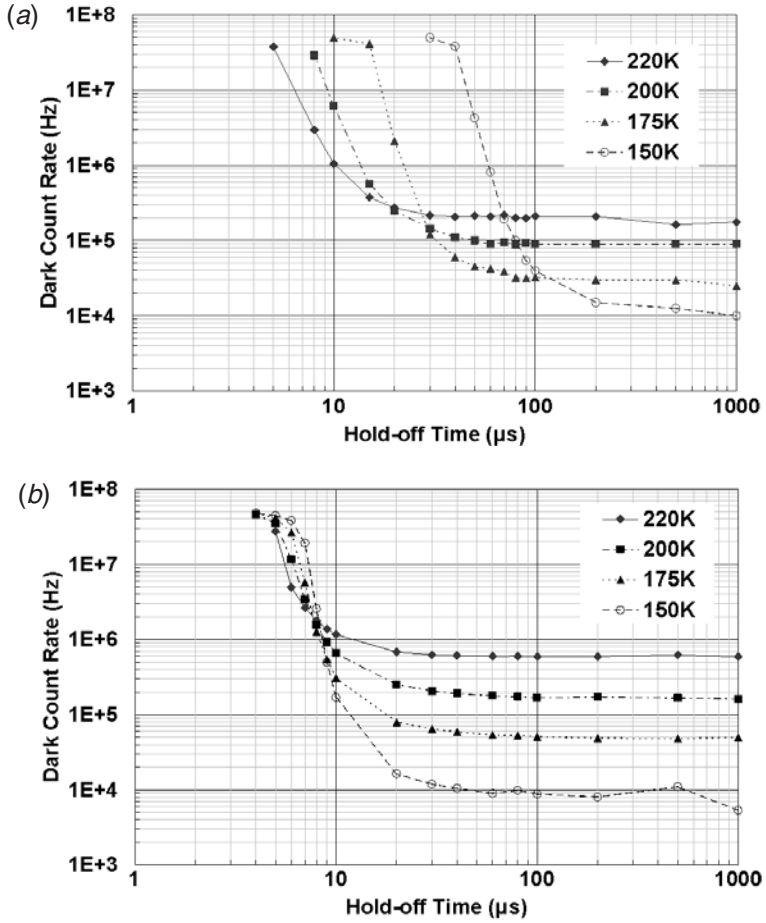

Figure 6. Dark count rate versus hold-off time at 6 V overbias for four temperatures between 150 K and 220 K illustrates influence of afterpulsing effects at short hold-off times for (*a*) device 40E and (*b*) device 40C.

applied for a full 20 ns period regardless of when an avalanche might occur within the gate). Since $T_{on} \ll T_{off}$, the gate frequency is $(T_{on} + T_{off})^{-1} \approx 1/T_{off}$. The same absolute value of $V_{ov} \sim 6$ V was used at all temperatures, and the bias below breakdown between gate pulses was held at 0.5 V below V_b in each case (at 220 K, $V_b \sim 42$V; at 150 K, $V_b \sim 38$ V). For device 40E, the 220 K data in figure 6(*a*) show a sharp increase in DCR for hold-off times shorter than about 20 μs. For lower temperatures, the hold-off time for the increase of DCR becomes substantially longer, e.g. ~200 μs at 150 K. In contrast, as seen in figure 6(*b*), device 40C shows a sharp DCR increase for hold-off times less than about 10 μs at 220 K, and even at 150 K, DCR is fairly uniform for hold-off times as short as 20 μs.

The data from either of the devices illustrate the tradeoff between DCR and afterpulsing mentioned earlier: lower temperatures provide improved DCR but worse afterpulsing. However, a comparison of the data for these two devices illustrates another tradeoff: although 40E exhibits lower DCR than 40C at a given temperature, the afterpulsing for 40E is substantially worse, especially at the lower temperatures. We will present analysis and additional data in the following sections to describe how this difference in performance might be related to the structural differences between the devices. Another feature in the data for both devices – the saturation of the DCR at 5×10^7 counts per second at sufficiently short hold-off times – is just an artefact of the gate pulse duration: only a single dark count event can be detected during each 20 ns gate period, so the measured DCR cannot exceed $(20\text{ ns})^{-1} = 5 \times 10^7$ counts per second, which corresponds to a dark count measured in every gate pulse.

4.1 *Dark carrier generation, trapping, and afterpulsing effects*

To describe quantitatively the dark count probability, Kang *et al*. recently developed a model [18] that calculates total DCR based on contributions from several dark count generation mechanism. If we assume that N_d is the mean number of dark carriers in the multiplication region and P_a is the probability that a carrier triggers a detectable avalanche effect, then $N_d P_a$ is the mean number of carriers that trigger a detectable avalanche. To find the dark count probability P_d that at least one dark carrier triggers an avalanche, one assumes Poissonian statistics to obtain

$$P_d = 1 - \exp(-N_d P_a), \tag{1}$$

where the exponential term is just the probability of finding no carriers in the multiplication region.

To solve for P_d using equation (1), we must evaluate the mean number of dark carriers N_d. *Primary dark carriers* generated during a gate pulse can generate a dark count, and the number of primary dark carriers $N_{pdc} = I_{d,m}\ \tau/q$, where $I_{d,m}$ is the multiplied (or primary) dark current, τ is the gate length, and q is the electron charge. A second dark count mechanism is the afterpulsing effect stemming from the de-trapping of previously trapped carriers during a gate pulse, giving rise to *afterpulse dark carriers*. If one considers the possible contribution of trapped carriers

from all previous gate pulses, the summation over previous pulses yields the number of afterpulse dark carriers N_{adc} given by the expression

$$N_{adc} = P_d N_{tr,0}[\exp(\tau/\tau_d) - 1]/[\exp(\Delta T/\tau_d) - 1]. \quad (2)$$

where $N_{tr,0}$ is the number of initially filled traps at the end of a gate pulse, τ is the gate pulse width, τ_d is the characteristic de-trapping time (assuming an exponential de-trapping behaviour), and ΔT is the time between successive gate pulses (i.e. the sum of τ and the hold-off time, where generally $\tau \ll$ hold-off time). Kang *et al.* [18] also consider two additional mechanisms related to dark carriers (primary or afterpulse) generated just before the gate pulse that persist long enough to be present when the gate pulse is turned on. To first order, consideration of these two additional mechanisms is roughly equivalent to assuming a slightly longer gate pulse width, and for the sake of simplicity, we have ignored them. By taking $N_d \sim N_{pdc} + N_{adc}$, we can use equation (1) to fit our data and extract the characteristic de-trapping time τ_d. For the various parameters in the model, we assumed the following: $I_{d,m}$ is fitted to data at long hold-off times; gate width $\tau = 20$ ns; $\sim 10^8$ carriers per pulse, of which $\sim 1\%$ are trapped, lead to $N_{tr,0} \sim 10^6$ (as used by Kang *et al.*); $\Delta T = (10\,\text{kHz})^{-1}$; and P_a is taken to be 0.2 (as assumed by Kang *et al.*).

We present the results of this analysis applied to the data presented in figure 6(*a*) for device 40E in figure 7, where open symbols and dashed lines indicate measured data, and filled symbols and solid lines indicate model results. The model predicts a much steeper increase in DCR with reduced hold-off time, but fitting to the onset of this increase can be used to extract de-trapping times τ_d for each curve. Modelled detrapping times are indicated in the figure. Note that the model accurately mimics the saturation of the DCR seen experimentally for sufficiently short hold-off times.

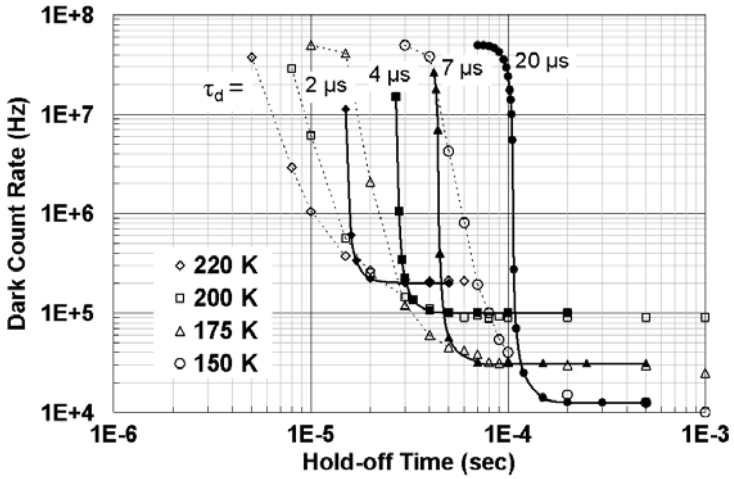

Figure 7. Simulation of dark count rate versus hold-off time using model of Kang *et al.* [18] to fit data from figure 6(*a*) for device 40E. Filled symbols and solid lines are model results; open symbols and dashed lines are measured data. De-trapping times τ_d providing best model fit to onset of dark count rate increase are indicated for each model curve.

4.2 *Afterpulsing scaling and de-trapping activation energy*

Although the model of Kang *et al*. is quantitative and provides a very physical basis for describing dark count phenomena (and photon counting behaviour as well), we do not find very accurate agreement with the rate of increase of the DCR at short hold-off times when the afterpulsing effect is strong. Additionally, the model has a relatively large number of free parameters that complicate the fitting of our data. We have therefore employed an alternative analysis below that allows us to describe the change in τ_d with variations in temperature using an effective thermal activation energy that can be deduced without having to first estimate values of τ_d.

We first consider the data in figure 6(*a*) taken for DCR versus hold-off time using device 40E with a 6 V bias. The first step in the analysis is to normalize the DCR versus hold-off time curve to the hold-off-time-independent background DCR found for each temperature. Specifically, this background DCR value at 150 K, 175 K, 200 K, and 220 K was $\sim 1 \times 10^4$, 3×10^4, 9×10^4, and 2×10^5 Hz, respectively, as can be seen in figure 6(*a*) for hold-off times ~500–1000 μs. After normalizing each curve by its respective background DCR value, we obtain the normalized data shown in figure 8(*a*). Given these normalized curves, we can then consider the hold-off time corresponding to a specific increase in dark current above the long hold-off time (time-independent) DCR, e.g. 10 times the background DCR. We find that such an attribute of the normalized DCR scales quite accurately with the de-trapping times found as modelled in the previous section (e.g. 10 times the background DCR corresponds to $\sim 4\,\tau_d$). Moreover, the scaling of hold-off times for the different temperature curves was fairly independent of the specific criterion for DCR increase. For instance, the same scaling was found for hold-off times giving 100 times the background DCR (which corresponded to $\sim 3\,\tau_d$). This suggests that the *entire curve* for each temperature can be scaled so that all curves collapse to a single universal curve.

Such a collapse of all the curves in figure 8(*a*) by appropriate re-scaling of the hold-off time axis is illustrated in figure 8(*b*), where the scaling factors for the hold-off time axis were chosen to collapse all curves to the 150 K curve; i.e. the 150 K data has a scaling factor of 1, and for the 175 K, 200 K, and 220 K data, the scaling factors are 3.0, 5.7, and 9.6, respectively. (For the specific criterion of 10 times the background DCR, the scaling factors would be 3.0, 5.7, and 9.6; for 100 times the background DCR, these factors would be 3.0, 6.0, and 10.2.) The collapsed curves overlap quite accurately until the DCR plateaus at a normalized value corresponding to a dark count being detected in every 20 ns gate. Achieving a good collapse shows that the functional form of the increase in DCR with reduced hold-off time is independent of temperature and that the de-trapping phenomenon has the same behaviour at all temperatures up to a scale factor in the de-trapping time. To derive a thermal activation energy for this de-trapping, we plot the inverse of the hold-off time scale factors versus $1/kT$ in figure 9 and find an 'afterpulsing activation energy' $E_{\mathrm{ap}} = 91$ meV.

An identical analysis was carried out for the DCR versus hold-off time data from device 40C shown in figure 6(*b*). After DCR normalization and hold-off time rescaling as described above, we again found an excellent collapse of all four curves,

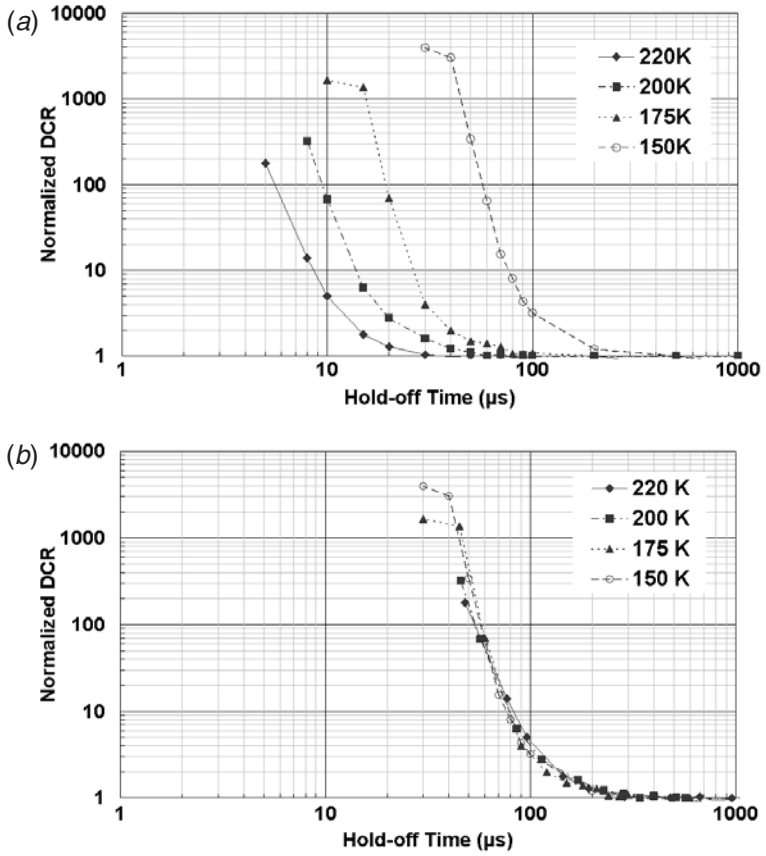

Figure 8. (*a*) Normalized dark count rate versus hold-off time obtained using data from figure 6(*a*) for device 40E. Each curve is normalized to long hold-off time (time-independent) dark count rate values. (*b*) Collapse of all four dark count rate versus hold-off time curves in (*a*) after re-scaling the hold-off time axis for the 150 K, 175 K, 200 K, and 220 K data by scale factors of 1.0, 3.0, 5.7 and 9.6, respectively.

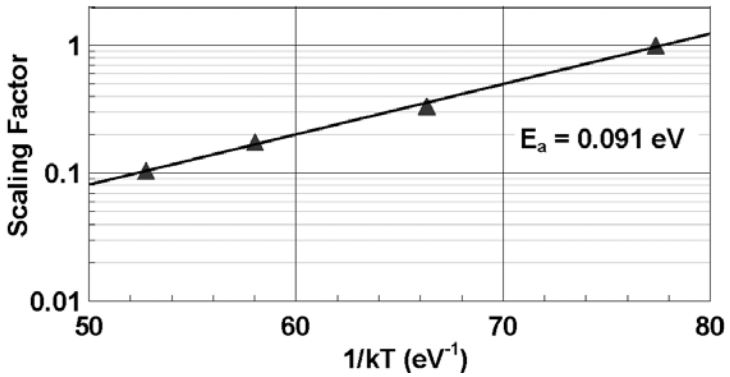

Figure 9. Dependence on $1/kT$ of the inverse of the scaling factors used to obtain the collapse of normalized dark count rate versus hold-off time curves for device 40E shown in figure 8(*b*). Exponential fit yields an afterpulsing activation energy of 0.091 eV that characterizes temperature dependence of afterpulsing effects illustrated by figure 6(*a*).

as found for device 40E. However, the much smaller scaling factors used to achieve this collapse (1, 1.25, 1.45 and 1.9, for the 150 K, 175 K, 200 K, and 220 K curves, respectively) lead to a much lower afterpulsing activation energy of $E_{ap} = 24$ meV, as shown in figure 10.

Inasmuch as carrier detrapping is a thermally driven process, the effective thermal activation energy E_{ap} describes the degree to which carrier de-trapping times change with temperature. However, to explain the substantially different values obtained for E_{ap} for devices 40C and 40E, we consider the conditions under which de-trapping occurs while the SPAD bias is held just below V_{br} (by ~0.5 V) between adjacent gate pulses. For device 40C, the electric field amplitude in the multiplication region under this condition is about 4.5×10^5 V/cm, while structural differences between devices 40E and 40C, resulting from their different locations on the wafer, lead to a corresponding field for 40E that is estimated to be ~5–10% lower. This suggests that both thermal and tunnelling mechanisms may be important in determining carrier de-trapping times and that the afterpulsing activation energy E_{ap} may depend quite sensitively on the internal electric field, consistent with a field-assisted thermionic emission de-trapping process.

It is also possible that intrinsic differences in the material properties towards the edge of the wafer could lead to different types of trap defects in 40E, with consequently difference E_{ap} values. Therefore, we intend to more definitively demonstrate the role of internal electric field magnitude by performing similar measurements in the future with a baseline voltage held further below V_{br} during the hold-off period between gates.

4.3 *Afterpulsing as a function of overbias voltage*

To further our understanding of afterpulsing behaviour, we measured DCR versus hold-off time for device 40C at lower overbias voltages V_{ov}. In figure 6(*b*), we present data for 40C taken with $V_{ov} = 6$ V. In figures 11(*a*) and 11(*b*), we show complementary data taken with V_{ov} at 4 V and 2 V, respectively.

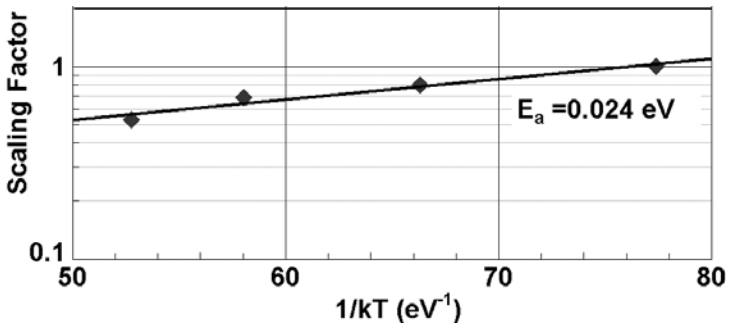

Figure 10. Dependence on $1/kT$ of the inverse of the scaling factors (1, 1.25, 1.45, and 1.9) used to obtain the collapse of normalized dark count rate versus hold-off time curves for the raw data presented in figure 6(*b*) for device 40C. Exponential fit yields an afterpulsing activation energy of 0.024 eV that characterizes temperature dependence of afterpulsing effects illustrated by figure 6(*b*).

A first observation is that afterpulsing is progressively less pronounced as the overbias is reduced. In fact, with $V_{ov} = 2\,V$, at a hold-off time of 4 μs, the DCR increases by no more than a factor of 2, even at 150 K, relative to the long hold-off time DCR values.

To further analyse these data, we applied the same normalization of DCR curves and re-scaling of the hold-off time axis to collapse all curves to a single universal curve, as described in the previous section. We find an excellent collapse for the $V_{ov} = 4\,V$ data. Although the very small variation in DCR for $V_{ov} = 2\,V$ makes the analysis of these data less accurate, we can still extract a value for E_{ap} with a somewhat larger uncertainty estimated to be ±5 meV. In summary, for V_{ov} values of 6 V, 4 V, and 2 V, we find afterpulse activation energy E_{ap} of 24 meV, 26 meV, and 29 meV, respectively. To within the uncertainty of the data and the analysis (especially of the 2 V data), $E_{ap} \sim 25 \pm 5$ meV seems to be independent of the overbias voltage. This overbias-independent result for E_{ap} is consistent with our hypothesis

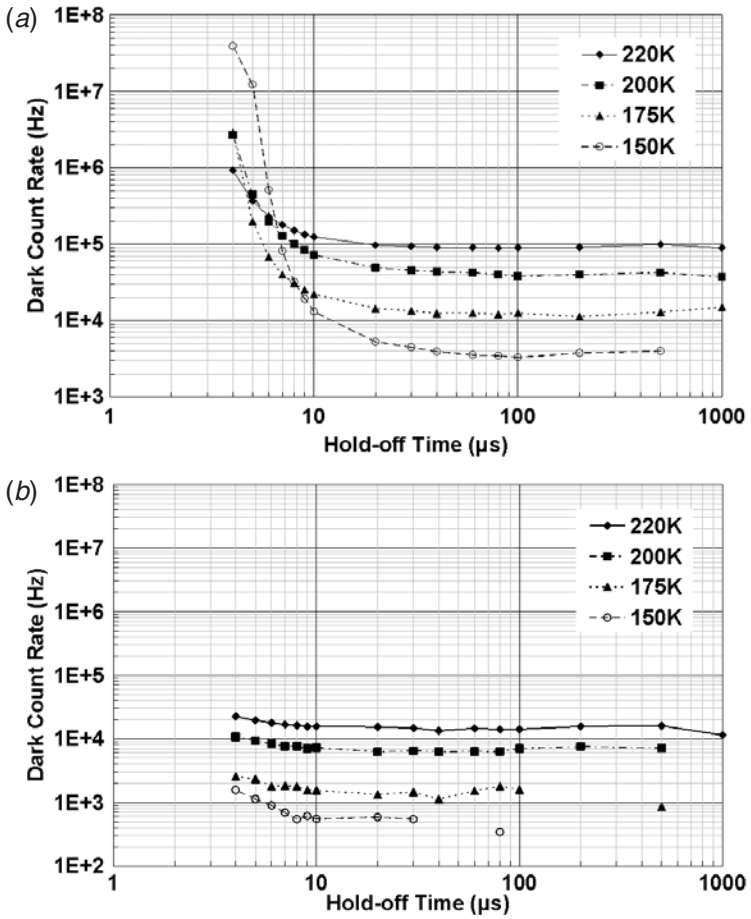

Figure 11. Dark count rate versus hold-off time for device 40C at four temperatures between 150 K and 220 K with an overbias of (*a*) 4 V and (*b*) 2 V. Taken in conjunction with complementary data at 6 V overbias in figure 6(*b*), afterpulsing effects show drastic reduction at lower overbias.

that the de-trapping time depends on the internal field amplitude in the multiplication region dictated by the baseline voltage between gate pulses, since this baseline voltage was 0.5 V below V_b for all values of V_{ov}.

The dramatic reduction in afterpulsing with lower overbias warrants further comment. It is generally accepted that the severity of afterpulsing following a given avalanche event depends on how much charge flowed through the device during the avalanche [2]. This argument presumes that the number of traps filled during the avalanche is roughly proportional to how much charge flowed past the traps. Therefore, afterpulsing behaviour is governed not only by fundamental device properties, but also by the way in which the device is operated. In particular, a low avalanche detection threshold followed by a rapid active quench can very effectively reduce total charge flow per avalanche event. In obtaining the data presented above, we used a 'gated quenching' in which a fixed gate length of 20 ns provided the cited overbias, and therefore avalanching, for the entire 20 ns period. Although this gate length is sufficiently short to provide good performance, it is also apparent that a lower overbias (e.g. 2 V) applied for 20 ns results in much less charge flow than for a higher overbias (e.g. 6 V). The reduced charge flow at lower overbias results in fewer trapped carriers and significantly reduced afterpulsing. Further information concerning afterpulsing and associated performance optimization can be obtained in the future by using different gate lengths and implementing active quenching.

5. Timing jitter properties

Variation in the temporal correspondence between the arrival of a photon and the detection of a resulting avalanche, referred to as timing jitter, originates in the stochastic nature of the carrier dynamics involved with avalanche breakdown. Variation in the 'vertical' build-up of the regenerative avalanche, intrinsic to the impact ionization process, is one contribution to the total jitter. There are also two other important mechanisms related to the way in which the avalanche laterally spreads throughout the active region [13]. Multiplication-assisted diffusion is the result of the lateral spreading of an initial cylindrical avalanche filament, enhanced by multiplication at the edge of the filament; and photon-assisted avalanche is the process by which infrequent radiative recombination of an electron–hole pair in the avalanche region radiates a photon that can be re-absorbed elsewhere in the multiplication region to initiate a new avalanche filament. Although these lateral mechanisms can dominate the timing jitter of larger area SPADs (e.g. Si-based devices), they may be less significant in the behaviour of the small area (25 and 40 μm diameter) devices studied for this paper. The design of the timing pick-up circuit for avalanche threshold detection is also a critical factor in determining overall jitter performance [19].

In figure 12, we illustrate the dependence of the timing jitter on the overbias V_{ov}. The jitter improves by over an order of magnitude as the overbias is increased from 2 V to 6 V. At any given overbias below 5 V, device 25C has notably lower jitter than device 25E. In addition to the mechanisms described above, we believe that an

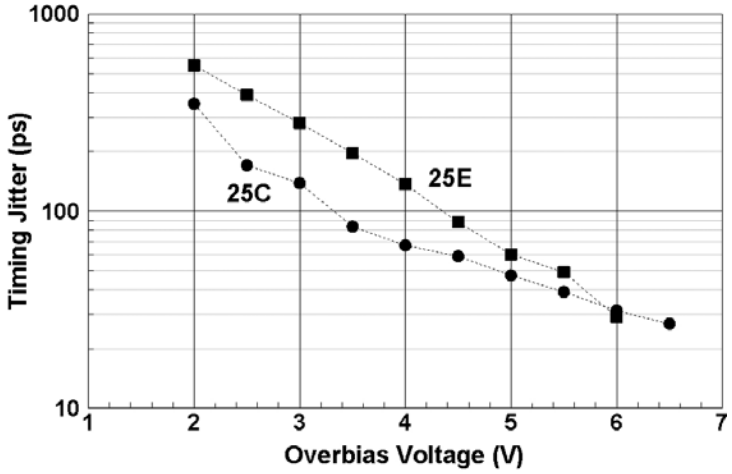

Figure 12. Timing jitter versus overbias voltage for devices 25C and 25E at 200 K shows strong reduction in jitter with increasing overbias. Performance of 25E is degraded relative to 25C, likely due to interface trapping effects.

important limiting factor for the timing jitter at low bias can be the residual trapping of a photoexcited hole at the InGaAs/InGaAsP (absorption/grading) interface (refer to figure 1) due to an imperfect grading structure. Larger electric fields at this interface will minimize carrier trap times, and the larger interface field for device 25C, relative to that of device 25E, explains the factor of two improvement in timing jitter for 25C for low V_{ov}. For overbias values beyond 5 V, the interface field for 25E is sufficiently large that both devices exhibit comparable jitter values. It should also be noted that the 30 ps timing jitter values obtained at 6 V are among the lowest jitter values ever reported for an InGaAs/InP SPAD. The jitter measurement setup used to obtain these data is based on an Ortec 9308 picosecond timing analyser (pTA) from which timing histograms are built up for the time between the measurement trigger and the 'stop' signal corresponding to the SPAD avalanche current crossing a preset threshold. The pTA has a minimum time bin of 1.2 ps and enables the measurement of timing jitter as short as 30 ps for the FWHM of the timing histogram.

To demonstrate the effect of temperature on the timing jitter, we present the data in figure 13 for devices 40C and 40E. These data were obtained for a relatively large overbias of 6 V using a 1300 nm pulsed laser. The optical waveform was reconstructed by a timing board (Becker&Hickl SPC630) with 7 ps intrinsic timing jitter, and the full-width at half-maximum (FWHM) of the laser pulse was about 60 ps. Due to non-optimal electronics (i.e. long coaxial cables inside the cryostat and the shielding of the SPAD fixture), the setup itself imposed a minimum jitter level of ~100 ps. Nevertheless, the difference between the two devices is clear. The jitter for device 40C was apparatus-limited for temperatures of 175 K and above, whereas device 40E shows a rapid degradation in jitter performance, by a factor of 5, as temperature is decreased below 220 K. The larger InGaAs/InGaAsP interface field present in device 40C makes the jitter performance of this device much less sensitive to temperature (at least at the level of ~100 ps) until a value

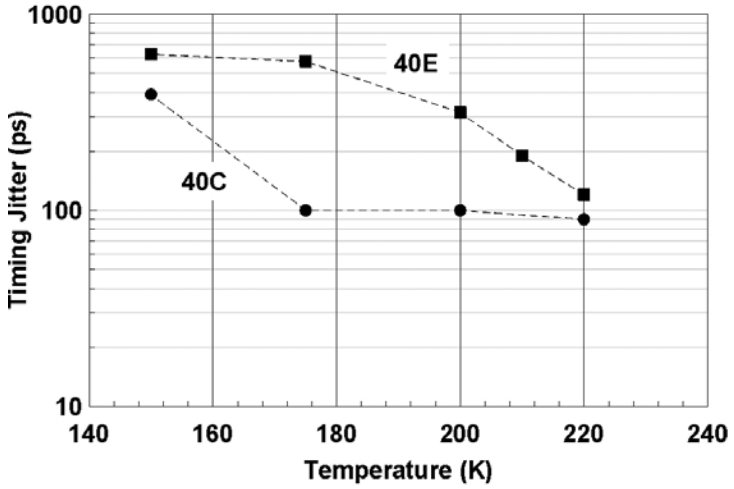

Figure 13. Timing jitter versus temperature for devices 40C and 40E at 6 V overbias exhibits progressively worse jitter with lower temperature, likely due to interface trapping effects. 40C shows much lower jitter consistent with higher interface electric fields. Experimental setup limitations imposed jitter floor of ~100 ps.

as low as 150 K is reached, whereas the lower interface field in device 40E makes this device more susceptible to the effects of interface trapping even at relatively high temperatures of ~220 K.

6. Conclusions

With the results obtained from this study, we have shown various ways in which photon counting performance characteristics of InGaAs/InP SPADs are sensitive to the internal electric field profile arising from the detailed structure of the device. In particular, by comparing typical devices from the centre of a wafer with edge devices that contain significant structural differences relative to the typical devices, we have shown how lower electric field magnitudes can lead to reduced dark count rates, but at the expense of degraded afterpulsing behaviour and larger timing jitter.

For all devices studied, the dark count rate exhibited a consistent average thermal activation energy of ~0.14 eV between 150 K and 220 K. Its exponential dependence on both temperature and overbias suggests that thermally assisted tunnelling through moderately deep defects in the InGaAs absorption region is responsible for dark carrier generation in this temperature range. However, there is also evidence that the effective activation energy is actually temperature-dependent, with a trend towards higher activation energy at higher temperatures, consistent with a crossover to mid-gap generation current generally found at room temperature for these structures

We have also presented an analysis of afterpulsing, as exhibited in dark count rate measured as a function of hold-off time, which provides an afterpulsing

activation energy quantifying how carrier de-trapping varies with temperature. We believe that larger electric fields in the multiplication region enhance de-trapping, leading to lower thermal activation energy. Future measurements in which electric fields are varied more directly should allow us to confirm this interpretation of our afterpulsing data.

Finally, we have demonstrated an InGaAs/InP SPAD design that simultaneously exhibits a dark count rate of ~10 kHz at a detection efficiency of ~20% with timing jitter of ~100 ps at 200 K operating temperature accessible with thermoelectric coolers. With appropriate tradeoffs in certain performance parameters, we have demonstrated a 200 K dark count rate as low as 3 kHz, a detection efficiency as high as 45%, and a timing jitter as low as 30 ps.

Acknowledgements

We thank Dr Mike Krainak for low-temperature *I–V* characteristic measurements and insightful discussions.

References

[1] A. Lacaita, P.A. Francese, F. Zappa and S. Cova, Appl. Opt. **35** 2986 (1996).
[2] G. Ribordy, J.D. Gautier, H. Zbinden and N. Gisin, Appl. Opt. **37** 2272 (1998).
[3] P.A. Hiskett, G.S. Buller, A.Y. Loudon, *et al.*, Appl. Opt. **39** 6818 (2000).
[4] P.A. Hiskett, J.M. Smith, G.S. Buller and P.D. Townsend, Electron. Lett. **37** 1081 (2001).
[5] D. Stucki, G. Ribordy, A. Stefanov, *et al.*, J. Mod. Opt. **48** 1967 (2001).
[6] K. Nishida, K. Taguchi and Y. Matsumoto, Appl. Phys. Lett. **35** 251 (1979).
[7] J.C. Campbell, A.G. Dentai, W.S. Holden and B.L. Kasper, Electron. Lett. **19** 818 (1983).
[8] S.R. Forrest, O.K. Kim and R.G. Smith, Appl. Phys. Lett. **41** 95 (1982).
[9] Y. Liu, S.R. Forrest, J. Hladky, *et al.*, J. Lightwave Technol. **10** 182 (1992).
[10] M.A. Itzler, C.S. Wang, S. McCoy, *et al.*, in *Proceedings of 24th European Conference on Optical Communications (ECOC '98)*, **1**, 59–60 (1998).
[11] M.A. Itzler, K.K. Loi, S. McCoy, *et al.*, in *Proceedings of 12th Annual Meeting of Lasers and Electro-Optics Society (LEOS '99)*, **2**, 748–749 (1999).
[12] M.A. Itzler, K.K. Loi, S. McCoy, *et al.*, in *Proceedings of 25th Optical Fiber Communication Conference (OFC 2000)*, **4**, 126–128 (2000).
[13] A. Spinelli and A.L. Lacaita, IEEE Trans. Electron Devices **44** 1931 (1997).
[14] M.M. Hayat, B.E.A. Saleh and M.C. Teich, IEEE Trans. Electron Devices **39** 546 (1992).
[15] S. Wang, F. Ma, X. Li, *et al.*, Appl. Phys. Lett. **82** 1971 (2003).
[16] W.G. Oldham, R.R. Samuelson and P. Antognetti, IEEE Trans. Electron Devices **ED-19** 1056 (1972).
[17] G. Karve, S. Wang, F. Ma, *et al.*, Appl. Phys. Lett. **86** 063505 (2005).
[18] Y. Kang, H.X. Lu, Y.-H. Lo, *et al.*, Appl. Phys. Lett. **83** 2955 (2003).
[19] A. Gulinatti, P. Maccagnani, I. Rech, *et al.*, Electron. Lett. **41** 272 (2005).