InP-Based Single-Photon Detectors and Geiger-Mode APD Arrays for Quantum Communications Applications

Xudong Jiang, Mark Itzler, Fellow, IEEE, Kevin O’Donnell, Mark Entwistle, Mark Owens, Krystyna Slomkowski, and Sabbir Rangwala

(Invited Paper)

Abstract—To meet the increasing demand from quantum communications and other photon starved applications, we have developed various InP-based single-photon detectors, including discrete single-photon avalanche diodes (SPADs), negative feedback avalanche diodes (NFADs), and Geiger-mode avalanche photodiode (GmAPD) arrays. A large quantity of InP SPADs have been fabricated. Out of ~1000 devices with a 25-μm active area diameter, operated under gated mode at temperature of 233 K, with a pulse repetition rate of 1 MHz and pulse width of 1 ns, the average dark count rate and afterpulsing probability are 30 kHz and 8 × 10−5, respectively. Smaller (16-μm active area diameter) and larger (40-μm active area diameter) discrete devices have been fabricated as well, and their performances are presented along with the 25-μm diameter devices. NFAD devices can operate in free running mode and photon detection efficiency of ~10–15% can be achieved without applying any hold-off time externally. When the temperature decreases from 240 to 160 K, the noise equivalent power (NEP) decreases from 1.9 × 10−16 to 1.8 × 10−18 W/Hz1/2, with the activation energy being 0.2 eV. The very low NEP at ~160 K makes NFAD devices an ideal choice for long distance, entanglement-based quantum key distributions. GmAPD arrays provide an enabling technology for many active optical applications, such as 3-D laser detection and ranging (LADAR) and photon starved optical communications. Both 32 × 32 and 128 × 32 GmAPD arrays have been fabricated with high performance and good uniformity. GmAPD focal plane arrays (FPAs) with framed readout mode have enabled very high-performance flash LADAR systems. GmAPD FPAs with asynchronous readout mode will enable high rate quantum key distributions and other quantum communications applications.

Index Terms—Single-photon avalanche photodiode (SPAD), Geiger mode, negative feedback, APD array, short-wave infrared (SWIR), quantum key distribution (QKD), quantum communications, laser detection and ranging (LADAR).

I. INTRODUCTION

Detectors and detector arrays with single photon sensitivity in the short-wave infrared (SWIR) spectral band are critical for many applications, including optical time domain reflectometry [1], [2], quantum communication [3], laser detection and ranging (LADAR) [4], deep space acquisition and tracking [5], and semiconductor device and material characterization [6]. Historically, photomultiplier tubes (PMTs) have been widely used in many areas that require low light level detection. However, although the PMT still plays an important role in some applications today, as with many vacuum tube-based devices, this 80-year-old technology [7] is gradually being replaced by newer solid-state devices. PMTs are fragile, bulky, sensitive to magnetic fields, require very high operating voltages, and are not conducive to making large format detector arrays. Moreover, their sensitivity in the SWIR spectral band is poor. Due to these shortcomings of PMTs and the increasing role that single photon detectors are playing in many fields, there has been a growing effort to develop a variety of solid-state photon detectors [8]–[10] such as single-photon avalanche diodes (SPADs), superconducting transition-edge sensors, superconducting nanowire single-photon detectors (SNSPDs), frequency up-conversion single-photon detectors, and visible-light photon detectors (VLPCs).

Among the various solid-state detectors, some of them, such as silicon-based SPADs and VLPCs do not have adequate sensitivity in the SWIR band. Others have excellent performance in SWIR but require rather onerous operating conditions: for instance, the SNSPD has excellent single-photon detection performance over a wide spectral range, including the SWIR spectral band, but requires cryogenic cooling, which limits its practical use for many applications. In contrast, InP-based SPADs have high performance in the SWIR band, and they also provide high reliability, compactness, and low cost. For quantum communications and other applications where single photon detection in the SWIR band is critical, InP-based SPADs provide an excellent solution that is robust and scalable.

Since its initial proposal [11], [12] and experimental demonstration [13], quantum cryptography [14], and in particular quantum key distributions (QKDs) have been developing rapidly in recent years. For QKD and other quantum communications applications, single-photon detectors with SWIR sensitivity are a key enabling technology. Very long distance (≥250 km) QKD over fiber links has been achieved using SNSPDs [15], [16]. Free-space QKD has been successfully demonstrated over long distance [17], and direct and full-scale experimental
verifications towards ground-satellite QKD have been performed \cite{18}. Thanks to complementary advances in device technology and operating circuitry, InP-based SPADs have enabled rapid progress in QKD implementation in both fiber links and free space optical channels. Mb/s secure key rates have been achieved in QKD networks using InGaAs/InP SPADs with GHz system clock rates and over 50-km of fiber links \cite{19}. For large scale deployment of QKD and other quantum communications systems, at present InP-based SPADs provide the only viable practical solution.

To meet the consequent demand for high-performance SWIR single photon detectors, we have focused on developing high-performance InP-based single photon detectors and arrays in the past several years, and have made significant advancements in the device technology \cite{20}. We summarize this progress in the remainder of the paper, which is organized as follows: In Section II we present the performance of InGaAs/InP SPADs, focusing on the photon detection efficiency (PDE) and dark count rate (DCR) performance metrics. In Section III we discuss the performance of InP-based negative feedback avalanche diodes (NFADs), InP-based Geiger-mode APD (GmAPD) arrays and their applications for quantum communications are discussed in Section IV. A discussion of our results and final conclusions are provided in Section V.

II. INP-BASED SPADS

Our InP-based SPAD device design platform employs a separate absorption and multiplication structure that introduces a high electric field in the multiplication region to induce avalancheing and maintains a low electric field in the absorption region to suppress tunneling. The p-n junction is fabricated using zinc diffusion, and the resulting planar buried junction structure ensures low perimeter leakage, stable performance, and high reliability. Further details of the design and operation of these SPADs have been presented elsewhere \cite{20}, \cite{21}.

Key performance parameters for SPADs include PDE, DCR, afterpulsing probability (APP) and timing jitter (TJ). These parameters have important implications in QKD applications. For example, TJ and the ratio of DCR to PDE (DCR/PDE) affect the quantum bit error rate and therefore limit the maximum transmission distance of a QKD system, and APP limits the maximum key generation rate. Over the past few years, we have carried out extensive experimental and theoretical investigations of InP-SPAD performance and its dependence on various material and device parameters and operating conditions. We have also fabricated a large quantity of InP-SPAD devices. Sub-100 ps RMS TJ has been demonstrated before \cite{22}, and here we focus on PDE, DCR and APP, particularly the distribution of these performance parameters over a large quantity of devices.

SPADs were characterized with gated mode operation using circuitry that employs two matched RF delay lines to cancel the capacitive transient response \cite{23} of the SPAD impedance to the very short gate bias pulses. The bias pulse width was 1 ns, the repetition rate of the gate pulses was 1 MHz, and the input photo-signal provided by a laser diode had a mean photon number of 0.1 photon/pulse. To enable afterpulsing measurements we used a scheme \cite{23}, \cite{24} in which “lit” and “dark” gates are interleaved for each pair of neighboring pulses. 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. With the laser source attenuated to a mean photon number of $\mu = 0.1$ per “lit” gate, only approximately one in ten “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. 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 is used to quantify the APP.

A histogram summarizing the distribution of DCR at 20% PDE for 990 SPADs operated at 1550 nm is presented in Fig. 1. These devices have a 25 $\mu$m active area diameter, and the testing temperature was 233 K. The cumulative percentage of the distribution is also shown. These devices exhibit a fairly broad distribution in DCR, from a minimum of 1.8 kHz to a maximum of 99.1 kHz, with a mean of 29.7 kHz and a standard deviation of 20.4 kHz. As can be seen from this figure, 90% of the devices have DCR $\leq 60$ kHz.

Distribution of the APP for the same population of devices tested under the same conditions is shown in Fig. 2. The minimum and the maximum of the APP are 2.66 $\times 10^{-7}$ and 9.76 $\times 10^{-4}$, respectively; the mean and standard deviation for the distribution are 8.04 $\times 10^{-5}$ and 9.08 $\times 10^{-5}$, respectively. A large majority of the population (i.e., 90%) have APP $\leq 1.5 \times 10^{-4}$.

Similar analysis has been done for devices with active region diameters of 16 $\mu$m and 40 $\mu$m, although with much smaller populations of devices (i.e., ~35 devices for each device type). In Fig. 3 we summarize DCR and APP for three types of devices that have active region diameters of 16,25 and 40 $\mu$m. The solid squares in Fig. 3(a) and solid circles in Fig. 3(b) indicate the average DCR and APP of each device type, respectively. The
Fig. 2. Distribution of APP at 20% PDE for 990 SPADs tested at 1550 nm and $T = 233$ K. Cumulative percentage for the distribution (red solid curve, right ordinate) is also shown. The active region diameter of all devices is 25 $\mu$m.

Fig. 3. (a) DCR at 20% PDE and (b) APP at 20% PDE for three different device types. Solid squares (■) in (a) and solid circles (•) in (b) indicate the average DCRs and APPs, respectively, and the error bars indicate the standard deviation ($\pm \sigma$) of the DCR and APP distributions. The gate pulse repetition rate is 1 MHz with a gate width of 1 ns, and the measurement temperature is $T = 233$ K.

ear bars indicate the standard deviation ($\pm \sigma$) of the DCR and APP distributions for each device type. When device active region diameter decreases from 40 to 16 $\mu$m, DCR shows a clear decrease, but APP only decreases slightly and exhibits a very mild dependence on device size.

To better understand the dependence of DCR and APP on device sizes, statistical analysis results are tabulated in Table I. Even though the population of 16 and 40 $\mu$m diameter devices are much smaller than that of 25 $\mu$m diameter devices, i.e., ~35 versus ~1000, the results presented in Table I show some very interesting and definite trends. We will discuss some of the relevant issues.

Devices of all three types show significant variation in both DCR and APP. The causes of these variations can be grouped in three categories: (i) epitaxial wafer non-uniformities in material properties such as layer thickness, doping concentration and defect/trap densities; (ii) device processing non-uniformities, such as zinc diffusion depth and the resulting multiplication layer width, and the introduction of local surface defects; and (iii) device packaging processes such as die singulation, die mounting, wire-bonding, and device burn-in, which can apply electrical, thermal, and mechanical stresses to devices that may cause variations in device performance. The relative importance of these three potential causes of device performance variation will become more clear in the discussion of GmAPD array performance in Section IV. In particular, the 1024 devices comprising each $32 \times 32$ array show much less variation than the discrete devices discussed in this section. Given that epitaxial material and device processing are extremely similar for both discrete devices and arrays, the device packaging processes appear to be a strong contributor to the wider distribution of discrete device performance.

All three device sizes show remarkably similar ratio of $\sigma$/mean for DCR. A possible explanation is that at the characterization temperature of 233 K and the over-bias voltage corresponding to 20% PDE, DCR is dominated by thermal generation processes in the InGaAs absorber [21]. In this case, the fluctuation of DCR is dominated by the defect distribution in the absorber and does not depend much on other device properties such as multiplication layer width and defect distribution in the multiplier. The defect distribution in the absorber is expected to be similar for all three sizes of devices since they were fabricated on the same wafers, and similar defect distributions should dictate similar DCR $\sigma$/mean ratios for different devices.

Another interesting phenomenon is that $\sigma$/mean ratio for APP increases monotonically with device size. Afterpulsing is

<table>
<thead>
<tr>
<th>Device diameter</th>
<th>16 $\mu$m</th>
<th>25 $\mu$m</th>
<th>40 $\mu$m</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCR (kHz)</td>
<td>Maximum</td>
<td>99.1</td>
<td>97.5</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.8</td>
<td>4.7</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>29.7</td>
<td>40.4</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>20.4</td>
<td>27.3</td>
</tr>
<tr>
<td></td>
<td>$\sigma$/mean</td>
<td>0.69</td>
<td>0.67</td>
</tr>
<tr>
<td>APP</td>
<td>Maximum</td>
<td>$9.76 \times 10^{-4}$</td>
<td>5.26 $\times 10^{-4}$</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>$9.45 \times 10^{-5}$</td>
<td>9.16 $\times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>$8.04 \times 10^{-5}$</td>
<td>8.95 $\times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>$9.08 \times 10^{-5}$</td>
<td>5.14 $\times 10^{-5}$</td>
</tr>
<tr>
<td></td>
<td>$\sigma$/mean</td>
<td>1.13</td>
<td>1.30</td>
</tr>
</tbody>
</table>
caused by the trapping and subsequent de-trapping of carriers generated during previous avalanches, and it increases with avalanche charge size. Under the same temperature and at the same PDE, larger area devices will have larger avalanche sizes, and more carriers will be trapped. Due to the complex nature of trapping/de-trapping, it is not surprising that larger area devices exhibit more stochastic variation in the trapping/de-trapping process, and therefore larger \( \sigma/\text{mean} \) ratio in APP. We have not seen a clear correlation between DCR and APP. Under the characterization conditions presented in this paper, DCR and APP are originated in different parts of the device and have different mechanisms as well. Deep generation-recombination centers in the absorber are mainly responsible for the dark count events, and shallow level traps in the multiplier are responsible for the afterpulsing behavior.

As device diameter decreases, DCR decreases significantly. On the other hand, APP decreases only slightly. A dominant factor in the very weak dependence of APP on active region size is that the bonding pad for all device sizes is the same size, as dictated by wire-bonding requirements for device packaging. The bonding pad introduces parasitic capacitance that must be charged and discharged during the arm, avalanche, and quench process of Geiger-mode operation. The charge flow through the device can be roughly estimated to be \( (C_{\text{bi}} + C_{\text{p}}) \times V_{\text{ex}} \), where \( C_{\text{bi}} \), \( C_{\text{p}} \) and \( V_{\text{ex}} \) are the intrinsic diode capacitance, parasitic capacitance introduced by the bonding pad and the excess bias voltage, respectively. When the device active area diameter is reduced, the bonding pad will be partially located on non-diffused area, and this will introduce certain parasitic capacitance \( C_{\text{p}} \). The bonding pad parasitic capacitance for the smaller area devices will increase the avalanche charge flow size and will reduce the charge flow difference between large and small area devices. As a result of this, the APP does not show as sensitive a dependence on device size as DCR does.

The major challenge for high rate QKD is afterpulsing, and currently the main approach to reducing afterpulsing is to try to reduce the avalanche charge size. With various clever circuit designs, such as self-differencing [25], sine-wave gating [26], [27], and harmonic subtraction [28], very narrow overbias gate widths of \( \sim 100-200 \) ps have been achieved at GHz repetition rates. These narrow gate widths significantly reduce the avalanche charge size and minimize afterpulsing effects, which makes GHz QKD possible.

### III. InP-Based NFADs

To date, the SPADs presented in Section II have been most often used for gated mode operation in applications where expected photon arrival time is known. In applications where photon arrival time is not precisely known, free-running operation is needed. The simplest implementation of free-running operation relies on passively quenched SPADs, in which a sufficiently large quench resistor is hybridly integrated with a SPAD. However, the hybrid approach generally introduces a large parasitic capacitance that increases the charge flow during the avalanche process and limits the maximum counting rate by exacerbating afterpulsing effects. To overcome some of the limitations of SPADs operated in either gated mode or free-running mode with hybridly integrated quench resistors, we have developed NFADs by monolithically integrating thin film resistors with InP SPADs. Our platform of fabricating NFADs is very versatile and allows to fabricate NFAD devices with different SPAD active area sizes and various resistors. The concept, design, and performance of NFADs have been discussed previously [29], [30].

Avalanches in properly designed NFADs are quenched very rapidly \((<1 \text{ ns})\), which significantly reduces the avalanche charge flow and associated afterpulsing effects. Once the avalanche is quenched, the device is re-charged through the integrated resistor to re-apply the over-bias voltage in preparation for the next detection. The recovery time, which depends on the resistance of the integrated resistor and the diode capacitance, determines the device’s maximum counting rate. In addition to the common Geiger-mode performance parameters shared by both SPADs and NFADs, such as PDE, DCR, APP, and TJ, minimum recovery time is another important parameter related to NFAD performance.

An effective way to determine the NFAD recovery time is through the dependence of the pulse height of the second peak \( P_2(t_2) \) of each pair of consecutive dark avalanche pulses \( [P_1(t_1), P_2(t_2)] \) on the inter-arrival time \( \Delta t = t_2 - t_1 \) between these two pulses [30]. Fig. 4 presents the \( P_2 \) versus \( \Delta t \) relationship for a device referred to as E2G6 tested at \( T = 236 \) K and...
off time after each detection event, the PDE can be extended beyond 20%. Several groups have reported results for device E2G6 characterized at different temperatures [30]–[33], and the noise equivalent power \( \text{NEP} = \frac{\hbar \nu}{\sqrt{2} \text{DCR}} / \text{PDE} \) (where \( h \) is Planck’s constant and \( \nu \) is the incident photon frequency) obtained from these different studies is summarized in Fig. 6. The red solid circles are the measured dependence of NEP on \( 1/(kT) \), where \( k \) and \( T \) are the Boltzmann constant and temperature, respectively; and the blue dashed line is an exponential fit to the measured data. From the exponential fit, we find an activation energy of \( \sim 0.2 \text{ eV} \) in the temperature range between 163 and 240 K.

At 163 K, the NEP of less than \( 2 \times 10^{-18} \text{ WHz}^{-1/2} \) approaches that of an SNSPD, and this NFAD operating temperature can be achieved with relatively modest resources compared to those required for cryogenic operation of an SNSPD at \( \sim 2 \text{ K} \). As mentioned in Section I, very long haul QKD demonstrations have utilized SNSPDs due to their high performance; however, with nearly comparable NEP performance from an NFAD, which is much simpler to operate and does not have stringent requirement on operating temperature, NFADs may be an ideal choice for more practical long haul (>150 km) QKD links, as demonstrated in [33]. Although reset times may be a limitation of NFADs, earlier counting rate limitations encountered with gated operation (e.g., using SPADs) may be relaxed by using free-running operation. Beyond the application to QKD experiments, NFADs have also been used in other quantum optics experiments to characterize entangled quantum particles [34], [35].

### IV. GmAPD Arrays

In parallel with the push for higher performance discrete SPADs and NFADs, there have been significant drivers for the integration of these devices into large-format arrays [36]. To date, the most crucial of these has been deployment of GmAPD focal plane arrays (FPAs) in three-dimensional (3-D) imaging LADAR systems [37]. Three-dimensional LADAR provides...
many advantages compared to passive (2-D) imaging techniques, including the measurement of explicit depth information, which is impossible to obtain using 2-D imaging. Relative to traditional RADAR techniques that employ RF wavelengths, 3-D LADAR provides much higher resolution. By using optical wavelengths along with sub-ns laser pulse widths, 3-D LADAR techniques routinely provide centimeter scale range resolution over extremely long distances on the order of tens of kilometers. By providing high-performance single-photon sensitivity, LADAR systems based on GmAPD FPAs achieve an order of magnitude improvement in mapping rate over other competing LADAR technologies.

To meet the requirements of 3-D LADAR imaging applications, we have developed GmAPD arrays with formats of $32 \times 32$ and $128 \times 32$, and we have integrated these arrays into FPAs and turn-key camera systems [38], [39]. To provide LADAR FPA functionality, GmAPD arrays are hybridized to CMOS readout integrated circuits (ROICs) that provide independent time-of-flight measurements for each pixel in the array. The FPAs operate in a framed read-out mode, and frame rates are as high as $\sim 185$ kHz for $32 \times 32$ sensors and $\sim 115$ kHz for $128 \times 32$ sensors. We have developed two families of GmAPD arrays optimized for use with source wavelengths near 1.06 and 1.5 $\mu$m.

To facilitate the comparison of GmAPD arrays with discrete SPADs described earlier in this paper, we focus our discussion of GmAPD FPAs designed for 1.5 $\mu$m operation. The spatial map of PDE for 1024 pixels of a 1.5 $\mu$m $32 \times 32$ GmAPD FPA is presented in Fig. 7(a), and the corresponding distribution of PDE values is described by the histogram in Fig. 7(b). Pixel-level PDE values are obtained by illuminating the array with short laser pulses with a pulse width of $\sim 150$ ps and a mean photon number of 0.1 photon per pixel area per pulse. 25 000 frames of data are collected to extract the pixel PDE. The data presented in the figure are obtained at an excess bias corresponding to a mean DCR of 17.9 kHz and an FPA temperature of 248 K. As can be seen in Fig. 7(a), there is some non-uniformity in the distribution of PDE: pixels from the central region have somewhat higher PDE values relative to pixels from near the edge of the array. The mean and standard deviation $\sigma$ of the PDE distribution in Fig. 7(b) are 19.7% and 3.7%, respectively, and the corresponding ratio of these values is $\sigma$/mean = 19%.

Under the same operating conditions used to obtain the PDE performance map in Fig. 7, we demonstrate the DCR performance for the same FPA in Fig. 8. The spatial map of DCR values for all 1024 pixels of the FPA is shown in Fig. 8(a), and the corresponding distribution is summarized by the histogram in Fig. 8(b). These pixel DCR values are obtained from 25 000 frames of data with the camera operating at an average PDE of 19.7% and an FPA temperature of 248 K. A pattern similar to the PDE distribution is found for the DCR spatial distribution (i.e., central pixels have higher DCR values relative to the edge pixels). This similarity in spatial variations in pixel-level PDE and DCR across the FPA suggests that the root cause is systematic variations in the breakdown voltage $V_b$ of the GmA/DPds in each pixel, which is caused by the non-uniform Zn diffusion depth across the array, with shallower depths occurring in the edge region.

The minimum, maximum, mean, and standard deviation of DCR from the 1024 pixels are 7.8, 27.3, 17.9 and 3.49 kHz, respectively. The ratio of the standard deviation to mean is $\sigma$/mean = 0.2, which is significantly smaller than the corresponding ratio from $\sim 1000$ discrete devices presented in Section II. Certainly, the 1000 discrete devices shown in Section II come from areas larger than the area covered by a $32 \times 32$ GmAPD array ($\sim 3.2$ mm $\times$ 3.2 mm). However, we have confirmed that even a more limited set of discrete devices obtained from an area comparable to that of the array exhibit a distribution similar to that shown in Fig. 1. This strongly suggests that DCR performance is related to manufacturing processes beyond the epitaxial growth and wafer fabrication. As pointed out in Section II, each discrete device goes through packaging processes such as die singulation, die mounting, wire-bonding, and device
burn-in. These packaging processes apply various mechanical, thermal and electrical stresses to individual devices which affect their final performance. On the other hand, GmAPD array pixels are connected to the mating pixels on a ROIC die via indium-bump flip-chip bonding, which involves relatively low temperature and low mechanical stresses.

In addition to serving as the sensor engine for 3-D LADAR imagers, GmAPD arrays offer tremendous benefits for other photon-starved optical applications such as free-space optical communications (FSOCs). High data rate classical FSOC requires gigaphoton/second flux rate [40], and high-rate quantum FSOC (e.g., QKD) demonstrations can demand clock rates on the order of GHz, although the incident photon rate is generally much lower (e.g., 1/16 or 1/64 of the clock rate, in the tens of MHz range). For these high-rate FSOC applications, counting rate limitations of discrete SPADs due to afterpulsing effects can be overcome by multiplexing many devices. It is in this context that GmAPD FPAs provide a practical solution for FSOC.

The GmAPD FPAs described in Section IV have framed operation, for which all pixels in the array are simultaneously armed at the start of the frame. Photon detections are time-stamped by the ROIC during a range gate which generally has a period of several $\mu$s. Once a pixel fires, it remains disarmed for the remainder of the range gate. At the end of each range gate, the timestamps for all photon detections are read out, and this readout time limits the maximum frame rate. During the readout period, the array is not able to detect any incoming photons. The framed operation provides an excellent solution for LADAR imaging in which frame rates in the range of 10–100 kHz are often ideal. However, for high data rate photon counting applications such as FSOC, the framed operation will result in significant blocking loss. Asynchronous ROICs provide a solution to overcome this difficulty [41].

In asynchronous mode, each individual GmAPD pixel executes the cycle of firing, quenching, hold-off, and re-arming independent of the other pixels. When one pixel fires and becomes unavailable for detecting photons, other pixels are still armed and able to detect incoming photons. In asynchronous operation, every pixel experiences free-running operation, analogous to the NFAD operation described in the previous section. Automatic pixel reset and continuous readout minimize blocking loss and enable continuous high-rate photon detection.

To describe the performance of asynchronous GmAPD arrays, we have simulated the effective PDE and its limitation due to blocking losses. The relevant parameters are: (i) intrinsic pixel device PDE$_0$; (ii) intrinsic pixel device DCR; (iii) hold-off time $T_{ho}$ dictated by afterpulsing effects; and (iv) array format (i.e., number of pixels in the array $N_{pix}$). The bit rate is denoted as B, and the mean number of photons contained in an optical pulse is $N_{ph}$.

Fig. 9 shows the dependence of average DCR on average PDE at $T = 248$ K for the same GmAPD FPA used to obtain...
the data presented in Figs. 7 and 8. At 10% and 20% average PDE, the average DCR is 9 and 18 kHz, respectively. There are three relevant time scales: (i) mean time interval between photon detection events \( T_{ph} \), (ii) mean time interval between dark count events \( T_d \), and (iii) hold-off time \( T_{ho} \). \( T_d \) is simply the reciprocal of the DCR, and \( T_{ph} \) can be expressed as

\[
T_{ph} = \frac{1}{\left( \frac{B \times N_{ph}}{N_{pix}} \right) \times \text{PDE}_0}.
\]  

(1)

The blocking probability is

\[
P_{\text{blocking}} = \frac{T_{ho}}{T_{ho} + \left( \frac{1}{T_{ph}} + \frac{1}{T_d} \right)}.
\]  

(2)

and the effective PDE \( \text{PDE}_{\text{eff}} \) is

\[
\text{PDE}_{\text{eff}} = \text{PDE}_0 \times \left( 1 - P_{\text{blocking}} \right).
\]  

(3)

The relative loss in PDE \( L \) is

\[
L = \frac{\text{PDE}_0 - \text{PDE}_{\text{eff}}}{\text{PDE}_0}.
\]  

(4)

The calculated effective PDE of a GmAPD array as a function of array size (number of pixels) for different hold-off times is shown in Fig. 10. We consider two cases: (a) intrinsic average pixel PDE = 10% and the corresponding average DCR = 9 kHz; and (b) intrinsic average pixel PDE = 20% and the corresponding average DCR = 18 kHz. Other parameters used in the calculation include bit rate \( B = 1 \) Gb/s and number of photons per pulse \( N_{ph} = 0.1 \) ph/pulse. We have assumed that the incident photon pulse is spread uniformly across the array, and the optical coupling is 100% [i.e., neglecting coupling losses such as microlens arrays (MLAs) fill factor loss].

As can be seen from Fig. 10, the effective PDE decreases with increasing hold-off time and increases with the size of array. However, the effective PDE saturates beyond a certain characteristic array size, where this size depends on the hold-off time. Hold-off time is the parameter that effective PDE is most sensitive to. When the required hold-off time is less than 100 ns, \( \text{PDE}_{\text{eff}} \) of an array is comparable to the intrinsic pixel \( \text{PDE}_0 \) with a rather small format (e.g., \( 15 \times 15 \)). When the hold-off time is 1 \( \mu \)s, the relative loss in PDE \( L \) for \( 32 \times 32 \) GmAPD array is 3.6% at \( \text{PDE}_0 = 20% \) and 1.8% at \( \text{PDE}_0 = 10% \), respectively.

When the hold-off time is further increased to 10 \( \mu \)s, the relative loss \( L \) is 27.3% at 20% \( \text{PDE}_0 \) and 15.8% at 10% \( \text{PDE}_0 \), respectively, for a \( 32 \times 32 \) array. By increasing the array size to \( 128 \times 32 \), relative loss \( L \) for 10 \( \mu \)s hold-off time can be reduced to 18.6% at 20% \( \text{PDE}_0 \) and 10.3% at 10% \( \text{PDE}_0 \), respectively. The results presented in Fig. 10 indicate that GmAPD arrays with formats of \( 32 \times 32 \) and \( 128 \times 32 \) can be very useful for high counting rate communications applications. On the other hand, if using discrete devices, the achievable effective PDE will be very low. At \( \text{PDE}_0 = 10% \) and \( \text{PDE}_0 = 20% \), the effective \( \text{PDE}_{\text{eff}} \) is \( \sim 1% \) for both cases using a discrete device with a 1 \( \mu \)s hold-off time.

It is worth pointing out that an array device does have a much larger effective active area compared to a discrete device, and naturally the DCR becomes a concern for array devices. For certain devices such as silicon photomultipliers (SiPMs) and matrix NFAD devices, all the pixels share the same cathode and anode connections, and the aggregate DCR is a big concern since a dark count event from any of the pixels can impact the behavior of the whole device. In the GmAPD array device discussed here, each pixel has its own anode connection and will fire and quench independent of other pixels. What matters here is still the DCR of each individual pixel device. As a result of this, the DCR poses no bigger a challenge to a GmAPD array device than to a discrete SPAD device with an active area size comparable to that of the GmAPD pixel device.

V. DISCUSSION AND CONCLUSION

The performance of a large quantity of discrete devices presented in Section I represents current state-of-the-art InP-based single-photon avalanche diodes. Out of about one thousand 1550 nm discrete devices with a 25 \( \mu \)m diameter, 90% of them...
have a DCR below 60 kHz at 20% PDE and 233 K, and 90% of the devices have a corresponding afterpulse probability of $<1.5 \times 10^{-4}$ at a 1 MHz pulse repetition rate and 1 ns gate width. We have also presented performance data for 16 $\mu$m diameter and 40 $\mu$m diameter devices. Interestingly, all three types of devices show a similar spread in DCR performance in terms of the ratio of the standard deviation to mean value, i.e., $\sigma_{DCR}/\mu_{DCR} \approx 0.7$ for all cases. On the other hand, APP shows a spread that increases with device size: $\sigma_{APP}/\mu_{APP} \approx 0.76, 1.13,$ and 1.30 for 16, 25, and 40 $\mu$m diameter devices, respectively.

In addition to epitaxial growth and device processing, packaging is a potentially important source of variation in DCR and APP performance. In section IV we showed that the performance of 1024 pixels from a 1550 nm 32 $\times$ 32 GmAPD array (solid red circles). The DCR data for the GmAPD array pixels have been scaled to compensate for differences in device size and operating temperature.

Comparing 16, 25, and 40 $\mu$m diameter devices, DCR reduces significantly with device size, while APP does not. In Section II, we described how requirements on the bonding pad dimensions dictated by assembly processes result in bonding pads of similar size for all devices, and the parasitic capacitance of these bonding pads is the likely cause of similar APP performance for all device sizes. The reduction of this parasitic capacitance is a priority for future design improvements focused on realizing the benefits of smaller active area devices—namely, reductions in DCR and APP. Smaller device areas pose more difficult challenges in obtaining high efficiency optical coupling from optical fibers, and improved optical designs will be implemented as we continue to decrease device size.

Relative to SPADs, the NFAD has the advantage of very simple operation. Only a dc bias across the two terminals of the device is needed, and the device will execute the avalanching, quenching and recovering cycle without additional electronic control. This behavior inherently provides free-running operation, as desired for applications where photon arrival time is not known such as fluorescence measurements, spontaneous down conversion processes, and entanglement-based long distance quantum communications. Currently, discrete NFAD PDE is limited to $\leq 15\%$ without any externally applied hold-off time. However, for very long distance quantum communications, photon arrival rates are likely to remain low due to significant fiber loss. Thus, a hold-off time can be implemented to allow operation at higher PDE. In this case, NFAD devices provide a very suitable solution for long distance entanglement-based QKD. By reducing the device size and properly taking care of parasitic capacitance in the device design, we expect to reduce the APP and reach higher intrinsic PDE for NFADs in free running mode.

With a recovery time of $\sim 100$ ns, the NFAD will be limited to a maximum counting rate on the order of 10 MHz. However, by aggregating many individual NFAD devices into a matrix format and biasing them all in parallel by sharing the same P and N contacts, the counting rate can be significantly increased. Such a device is analogous to a microchannel plate (MCP) PMT, in which each NFAD in the matrix configuration plays the role of a single pore of the MCP. This "solid-state photomultiplier" concept has been implemented fairly extensively with silicon SPADs to form SiPMs.

An additional advantage of matrixed NFAD devices is the possibility of photon number resolution [30], which provides a solution for applications where photon number resolution capability in the SWIR spectral band is critical. The fill factor of NFAD matrix devices that we have fabricated to date (e.g., $2 \times 2$, $4 \times 4$ and $8 \times 8$) is low and results in a rather low effective PDE.
(i.e., \(\sim 3\%\) at 1550 nm). However, by designing suitable MLAs to increase the fill factor—as we have done in GmAPD FPIAs—combined with the implementation of smaller pixel sizes, we can significantly increase the effective PDE of NFAD matrix devices. Even with fairly low single-photon PDEs, the ability of the matrixed NFADs to provide a quasi-analog output can have significant value—e.g., in detecting LADAR returns over extended distances (i.e., for “soft” targets).

The development of GmAPD FPIAs is now enabling critical active optical applications such as 3-D LADAR and high-rate FSOCs. Formats of \(32 \times 32\) and \(128 \times 32\) are rapidly achieving maturity, including the ability to more routinely achieve 100\% pixel operability with very good performance uniformity. By shrinking the active region sizes, even better GmAPD array performance can be achieved including reduced DCR and APP, as well as reduced crosstalk and increased radiation tolerance. Active region size reduction will become even more important as we migrate to smaller pitch and larger format arrays. However, with reduced active area size, there are increasing challenges of optical coupling, including the necessity for sub-micron alignment precision of MLAs.

A dominant trend in photon counting applications is the need for greatly increased counting rates, with many applications demanding rates on the scale of GHz. To circumvent the present limitations on discrete SPAD counting rates, we have considered the effectiveness of using arrays for high-rate photon counting in the simulations presented in Fig. 10. Based on the afterpulsing performance of our present SPADs, a \(1 \mu s\) hold-off time is acceptable, in which case a \(32 \times 32\) array can provide an effective PDE essentially equivalent to an intrinsic SPAD PDE of 20\%. For a much more conservative hold-off time of \(10 \mu s\), APP will be very low, but the effective PDE of the array will be reduced to about 15\%. One caveat in these simulations is that the loss from MLA coupling has been neglected. Given that we achieve an MLA fill factor of about 75\%, the effective PDE will be reduced by a similar factor. Nonetheless, even with this penalty for optical coupling, GHz counting rates are still achievable using a \(32 \times 32\) GmAPD array.

All three device technologies discussed in this paper—discrete SPADs, NFADs, and GmAPD arrays—will have improved performance with smaller active area sizes, but they all share the common challenge of optical coupling when the device size is reduced. At present, afterpulsing remains a major limiting factor for achieving higher counting rates. We expect that innovative concepts for operating circuits will continue to dominate the near-term solutions for high counting rates with InP SPADs. From the perspective of intrinsic device behavior, a better long-term understanding of defects and traps in this material system is needed, particularly if the afterpulsing issue is to be tackled at a fundamental materials level.

Currently, InP-based SPADs appear to provide the only practical solution for large-scale deployment of quantum cryptography systems. Their viability in this context is bolstered by the past success of related InGaAs/InP linear mode APDs as an essential detector technology for supporting the enormous growth in fiber-optic networks over the last two decades to satisfy the ever-increasing demand for telecommunications bandwidth. With information security emerging as one of the most important current issues facing internet-based technology and commerce, quantum cryptography may become critical to future internet-based communications. Its large-scale adoption will drive further improvements in SPAD technology, which will enable even wider adoption of quantum communications systems by providing better devices at reduced cost. This virtuous cycle of system-level and component-level performance improvement suggests a bright future for both quantum communications and InP-based single photon detectors.

REFERENCES


Mark Itzler (M’96–SM’07–F’11) received the B.S. degree in physics from Brown University, Providence, RI, USA, in 1986, and the Ph.D. degree in physics from the University of Pennsylvania, Philadelphia, PA, USA.

He did his Postdoctoral Research from Harvard University from 1992 to 1995. He joined Epitaxx Optoelectronics, Inc., in 1996 and was promoted to the Director of R&D in 1999, and following the acquisition of Epitaxx by JDS Uniphase in Nov., 1999, he became the Chief Technical Officer (CTO) and the Vice President of Device Engineering for the Epitaxx Division of JDSU in 2000. He joined Princeton Lightwave Inc., (PLI), Cranbury, NJ, USA, where he is currently a Principal Scientist and leads programs focused on the design and characterization of high-performance photodetectors and detector-based products. He is the author/coauthor of more than 60 technical publications on solid-state physics and optoelectronic devices.

Mark Itzler (M’96–SM’07–F’11) received the B.S. degree in physics from Brown University, Providence, RI, USA, in 1986, and the Ph.D. degree in physics from the University of Pennsylvania, Philadelphia, PA, USA.

He did his Postdoctoral Research from Harvard University from 1992 to 1995. He joined Epitaxx Optoelectronics, Inc., in 1996 and was promoted to the Director of R&D in 1999, and following the acquisition of Epitaxx by JDS Uniphase in Nov., 1999, he became the Chief Technical Officer (CTO) and the Vice President of Device Engineering for the Epitaxx Division of JDSU in 2000. He joined Princeton Lightwave Inc., (PLI), Cranbury, NJ, USA, where he is currently a Principal Scientist and leads programs focused on the design and characterization of high-performance photodetectors and detector-based products. He is the author/coauthor of more than 60 technical publications on solid-state physics and optoelectronic devices.
Mark Entwistle received the B.S. degree in electrical engineering from the DeVry Institute of Technology, Woodbridge, NJ, USA, in 1986.

He joined Tram Electronics, Inc., as an Electronic Engineer, while graduating, and was later promoted to the Manager of Engineering, where his main responsibilities were high-reliability design and test engineering of military electronic assemblies. In 1996, he joined Epitaxx Optoelectronics, Inc., (which became a division of JDS Uniphase, after its acquisition of Epitaxx in 1999) as a Senior Manufacturing Engineer and was promoted to the Manager of Test Engineering. His main responsibilities were in the development of high-speed optoelectronic test platforms in support of 10-Gb/s receiver and laser production. He was also involved in the development of 40-Gb/s receiver technologies. He joined Princeton Lightwave, Inc., Cranbury, NJ, USA, in 2003, as the Manager of Software and Test Engineering, where he is responsible for the development of test platforms in support of new product development as well as support of existing platforms on established product lines, he is currently the Director of Software and Test Engineering. Also, he is responsible for electronic hardware and firmware development on PLI’s module-level product lines, he has substantial experience in camera electronics design and sensor testing, including board design and layout, FPGA and microcontroller programming, and module testing.

Mark Owens received the B. S. degree in physics from the State University of New York at Stony Brook, Stony Brook, NY, USA, in 1992.

He joined Gage Laboratory Corp. as a Metrology Engineer with responsibility for practical thermodynamic standards in 1992. He joined Epitaxx (which later was acquired by JDS Uniphase) as a Senior Engineer in 1999. There, he worked on the development, qualification, and production of high-reliability underwater receivers, and he fulfilled similar roles for high bit rate receiver and source laser product lines. He joined Princeton Lightwave Inc. (PLI), Cranbury, NJ, USA, as a Product Engineer to support quality, process, and reliability engineering for the company’s aerospace-qualified 820-nm SLD module, in 2005. Since joining PLI, he has functioned in multiple roles such as the Process Engineer and the Program Manager for space-qualified chip development and packaging efforts. He is currently a Principal Engineer at PLI. He has considerable experience in package assembly processes and failure analysis pertaining to photodiode sensors in both discrete and array-based formats. He is a U.S. Air Force Veteran and a Member of the International Microelectronics and Packaging Society and the American Society for Quality.

Krystyna Slomkowski received the B.A. degree in microbiology from Rutgers University, Douglass College, Piscataway, NJ, USA, in 1976.

In 1978, she joined Laser Diode Laboratories as a Senior Process Technician and was responsible for all processing of GaAs Burus LEDs. In 1985, she joined Epitaxx Optoelectronics Inc. as a Senior Process Engineer. At Epitaxx, she was responsible for the process development of planar InGaAs-InP p-i-n photodiodes and establishing a manufacturing line for this product. In 1989, she joined the R&D development group at Epitaxx. During this time, she was responsible for developing new processes and product platforms for fiber-optic detectors and receiver packages, and supporting and improving existing wafer fabrication processes. In 1999, she was promoted to the Technical Manager leading the Process Development group at the Epitaxx Division of JDS Uniphase. She managed the R&D Process Development Staff in the development of new processes and platforms for photodiodes with successful programs for 10-Gb/s APDs and 40-Gb/s PIN detectors. In January 2004, she joined Princeton Lightwave Inc., Cranbury, NJ, USA, where she is responsible for all photodetector and laser diode process development and execution, including photomask design and process flow definition. She is currently a Senior Staff Engineer at Princeton Lightwave Inc.

Sabbir Rangwala received the B.Eng. degree from Bombay University, Mumbai, India, the M.S. degree from the University of Texas, Austin, TX, USA, and the Ph.D. degree from the University of California at Berkeley, Berkeley, CA, USA, all in mechanical engineering. He joined AT&T Bell Laboratories after graduation, followed by other positions within AT&T’s Network System’s Division, working on deployment of fiber optics technology in underwater and terrestrial environments. This was followed by experiences in management consulting with Deloitte Consulting in the Telecommunications Practice. Thereafter, he joined Epitaxx, and following its acquisition by JDSU was in charge of development for all active components. In 2003, he joined Princeton Lightwave Inc., Cranbury, NJ, USA, where he has focused on product development, operations and sales, and in leading business development efforts for LIDAR in the commercial sector. He is the President and the Chief Operating Officer of Princeton Lightwave Inc.