Chapter 4: Semiconductor-based detectors

Sergio Cova, Massimo Ghioni
Politecnico di Milano
Dipartimento di Elettronica, Informazione e Bioingegneria
Piazza Leonardo da Vinci, 32
20133 Milano - Italy
sergio.cova@polimi.it

Mark A. Itzler
Princeton Lightwave, Inc.
2555 US Route 130 S.
Cranbury, NJ 08512
mitzler@princetonlightwave.com

Joshua C. Bienfang, Alessandro Restelli
National Institute of Standards and Technology, Joint Quantum Institute
100 Bureau Dr.
Gaithersburg, MD 20899
bienfang@nist.gov

Section 4.1 – Photon Counting: when and why

There is nowadays a widespread and growing interest in low-level light detection and imaging. This interest is driven by the need for high sensitivity in various scientific and industrial applications such as fluorescence spectroscopy in life and material sciences, quantum computing and cryptography, profiling of remote objects with optical radar techniques, particle sizing, and more. In particular, the use of fluorescence-lifetime spectroscopy as both an analytical and research tool has increased markedly in recent years finding remarkable applications in chemistry, biochemistry, and biology.

Photon counting has long been recognized as the technique of choice for attaining the ultimate sensitivity in measurements of optical signals. However, advanced analog detectors (such as back-illuminated charge-coupled devices (CCDs)) with ultra-low dark current can also be used in some instances to measure very weak photon fluxes. A basic issue must therefore be clearly addressed: when and why are photon-counting detectors advantageous? For applications where the measurement time is very short or the arrival time of the optical signal must be known with high precision (e.g. high-frame-rate imaging, fluorescence correlation spectroscopy (FCS), or fast optical coincidences), photon-counting detectors have an advantage over analog detectors, which have electronic readout noise in addition to dark-current noise. For short measurement times the readout noise exceeds the dark-current noise and sets the sensitivity limit of analog detectors, whereas readout noise simply does not exist in photon-counting detectors. These photon-counting detectors exploit an internal amplification mechanism, which, in response to single photons, generates macroscopic electrical signals that are much larger than any electronic circuit noise.

Section 4.2 – Why semiconductor detectors for photon counting?

Photon counting and time-correlated single-photon counting (TCSPC) techniques were developed using photomultiplier tubes (PMT), that is, vacuum-tube detectors with high internal gain (see Chapter 3), and high-performance PMTs have been produced industrially since the 1940s. Commercially available devices can provide remarkable performance, even up to rates of millions of counts per second, and compact and rugged PMT devices have been developed to address the typical drawbacks of vacuum-tube devices. Amongst their advantages, the most significant and distinct is the PMT's large sensitive area ($\approx$ cm^2), which can greatly simplifies the design of the optical system. Micro-channel plate (MCP) PMTs also offer picosecond timing jitter. However, PMTs suffer from low detection efficiency (DE). In the visible, the DE of conventional bialkali and multialkali photocathodes reach 20-25 % between 400 nm and 500 nm, whereas a DE up to ~40 % can be achieved between 450 and 650 nm using a GaAsP photocathode [1][2]. In the infrared, PMTs have much lower DEs.

Semiconductor-based detectors are a valuable alternative to PMTs. Besides the well-known advantages of solid state versus vacuum tube devices (small size, ruggedness, low power dissipation, low supply voltage, high reliability, low cost, etc.), semiconductor detectors provide inherently higher detection efficiency, particularly in the red and near-infrared spectral regions.

The development of semiconductor-based detectors of single photons has been slower than that of PMTs. Avalanche multiplication of carriers in reverse-biased p-n junctions is used in ordinary avalanche photodiodes (APDs) to obtain internal amplification in the detector similar to that in PMTs. However, in an ordinary APD, the multiplication of both holes and electrons causes an inherent positive feedback that produces strong fluctuations in the avalanche gain. Such fluctuations increase more steeply with the applied voltage than the average gain. In the best case (that is, in silicon diodes made with special structure and technology), the useful gain is limited to $\approx 5 \times 10^2$, as opposed to the $\approx 10^6$ gain easily reached by PMTs. Therefore, even in such best APDs the current

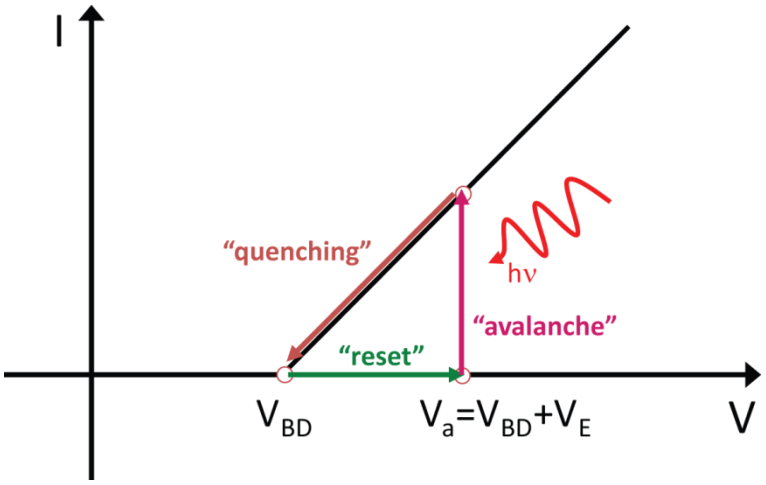


Fig. 4.1 SPAD operation in the reverse I-V characteristics of a p-n junction.

pulses due to single-photon absorption have very small and wildly fluctuating amplitudes, hence it is only marginally possible to detect single-photon pulses and their timing is highly uncertain.

Section 4.3 – Principle of operation of Single-Photon Avalanche Diodes

The positive feedback in the avalanche makes it possible to exploit a reverse-biased p-n junction in a different way for detecting single-photons. In this operation mode, the p-n junction cannot be considered a detector with an amplifier inside as is the case for APDs, but rather a detector with a digital flip-flop inside. As seen in Fig. 4.1, in the quiescent state the device is biased at voltage V_a above the breakdown voltage V_{BD}, and no current flows (the "OFF" state): in the junction depletion layer the electric field is very high, but no free carrier is present. When even a single charge carrier is injected in the high-field region it is strongly accelerated and can impact ionize and generate a secondary electron-hole pair, starting a self-sustaining avalanche multiplication process. The current then grows exponentially until the space-charge effect limits it to a constant level. This level is proportional to the excess bias voltage $V_E = V_a - V_{BD}$, hence an avalanche resistance can be defined. The p-n junction is thus switched to the "ON" state, where a constant macroscopic avalanche current flows (≈1 ma). The fast onset of the current marks the time of arrival of the photon that generated the initial charge carrier. The device remains in this ON state until the avalanche is quenched by an external circuit (quenching circuit), which drives the applied voltage down to V_{BD} (or lower). The quenching circuit then concludes the operation cycle by resetting the voltage to the original level above breakdown. The detector is insensitive to any subsequent photon arriving in the time interval from the avalanche onset to the voltage reset, which is the detector dead time.

This type of operation is called Geiger mode because of the analogy with the gas counters of ionizing radiation. The device operates in a way radically different from an ordinary APD and to avoid misunderstanding and confusion it was given a different name and acronym: single-photon avalanche diode (SPAD).

To be operated as a SPAD, a p-n junction must have a uniform breakdown over the entire active junction area in order to produce macroscopic current pulses with constant amplitude. That is, causes of localized breakdown must be avoided, such as edge effects and microplasmas within the active junction area. Besides this requirement, more stringent conditions must be fulfilled to attain acceptable SPAD performance, as discussed in Section 4.4.

Figure 4.2 outlines the structure of the silicon devices in which Geiger-mode operation was first observed. R.H. Haitz *et al.* [3][4][5] developed these devices in a planar technology: a deeply diffused guard ring was used to avoid edge breakdown effects and to define a small sensitive area of diameter < 10 µm.

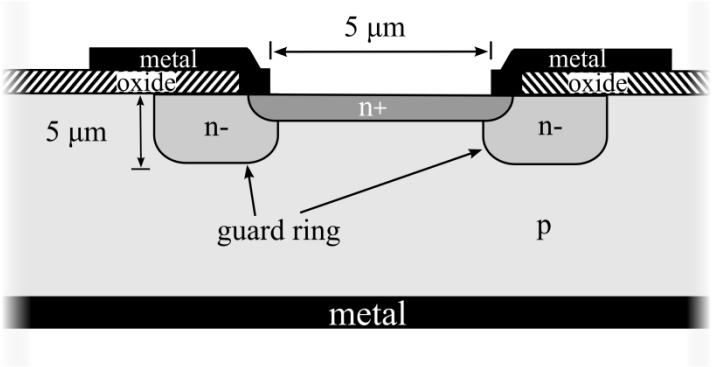


Fig. 4.2 Cross section of the p-n diode devised by Haitz *et al.* [5] for investigating the physics of avalanche above breakdown.

Silicon SPADs have been extensively investigated and are now well developed. Considerable progress has been achieved in SPAD design and fabrication techniques, and devices with good characteristics are commercially available for applications over the visible spectral range, up to 1 μm. A thorough review of the available silicon SPADs and their state-of-the-art performance is given in Section 4.6. Recent applications, such as quantum cryptography (quantum key distribution, QKD), LADAR, and VLSI circuit characterization based on light emission from hot carriers in MOSFETs, require single-photon detectors with high efficiency, low noise, and picosecond timing jitter in the wavelength range above 1 μm. The DE in the near-infrared (NIR) range is extremely low for typical PMTs, and reaches only ≈1 % for PMTs with photocathodes specifically designed for IR efficiency, but at the cost of high photon-timing jitter and high dark count rates [1]. The evolution of SPADs for extending the spectral range of photon counting beyond 1 μm started in the mid-1980s with studies carried out on commercially available APDs designed for fiber communications. Photon counting in the NIR range was first performed with Ge SPADs [6], and then extended with InGaAs/InP SPADs [7], which are now the workhorse for most experiments. A review of the state-of-the-art in InGaAs/InP SPADs is given in Section 4.8.

Section 4.4 – Performance parameters and features of SPAD devices

SPAD operation in Geiger mode is characterized by a number of basic performance parameters (see Chapter 2). The photon-detection efficiency is the probability that an incident photon triggers an avalanche (true detection event). The uncertainty in the photon arrival time is called timing jitter. The dark count rate (DCR) is the number of avalanches per unit time that occur in the absence of incident photons (false detection event, or "dark count"). Furthermore, physical phenomena specific to photon-counting devices can generate additional dark counts correlated to the occurrence of previous avalanche events, called afterpulses.

4.4.1 Photon detection efficiency

Besides the physical phenomena that determine the performance of semiconductor photodiodes in general (optical coupling, reflection, absorption, etc.), there are other physical effects that are important for SPAD operation. For a photon to be detected, not only it must be absorbed in the detector's active volume and generate a primary electron-hole pair. It is also necessary that the primary electron-hole pair succeeds in triggering an avalanche. The avalanche-triggering probability increases with the excess bias voltage V_E, since it is enhanced by a higher electric field. Theoretical and experimental studies [8][9] have shown that this probability increases linearly at low V_E, and tends to saturate at high V_E. The detection efficiency behaves accordingly, as illustrated in Fig. 4.3.

4.4.2 Dark count rate (DCR)

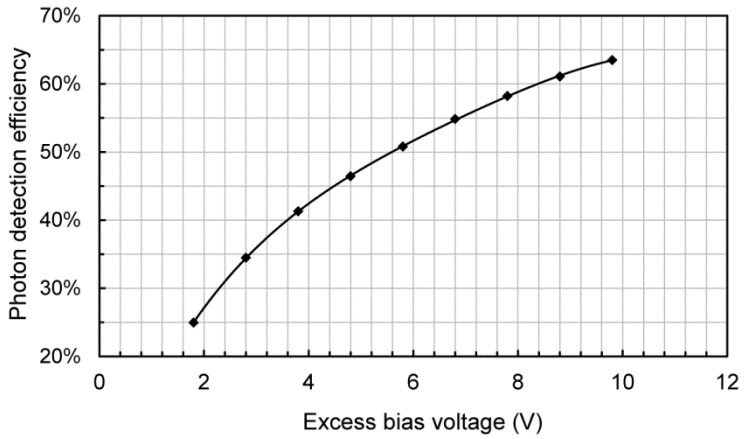


Fig. 4.3 Photon detection efficiency versus excess bias voltage for SPAD devices reported in [16].

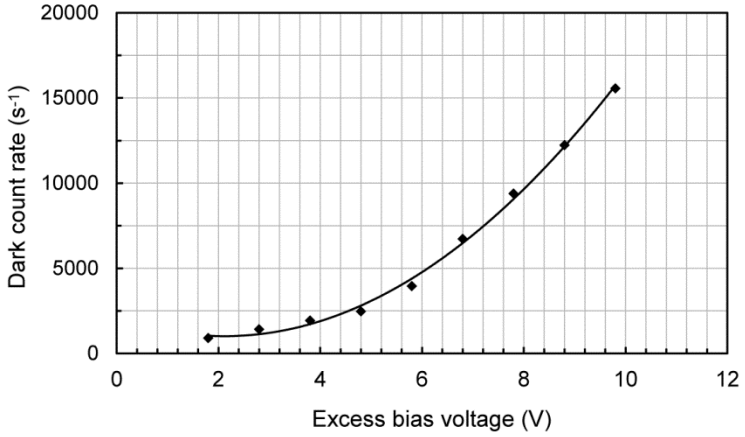
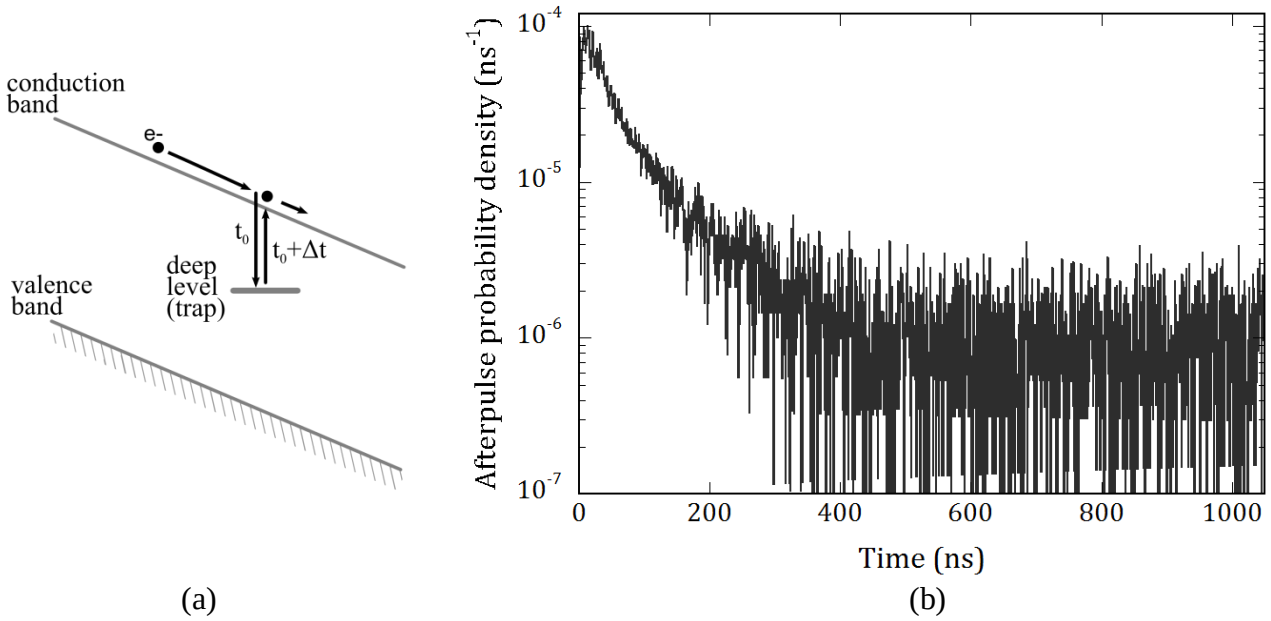


Fig. 4.4 Dark count rate versus excess bias voltage for a 50 µm diameter SPAD reported in **Error! Reference source not found.**, operated at room temperature.

Dark counts are due to carriers thermally generated within the SPAD junction, so that the dark count rate increases with the temperature. The DCR has the same role as the dark current in ordinary photodiodes, that is, its Poissonian fluctuations are the internal noise source of the detector. As shown in Fig. 4.4, the DCR of SPADs increases with the excess bias voltage V_E. The rise is due not only to the avalanche triggering probability, which also increases the DE, but also to the field enhancement of the carrier generation rate. It is well known that in silicon and other semiconductors thermal generation of carriers occurs through local energy levels located deep within the bandgap (levels closer to the midgap are the most efficient generation centers) [10]. Both the quality of the starting material and the technological processes used in the device fabrication have a strong impact on the density of deep energy levels and therefore on the generation rate. Transition metal impurities are the most common source of deep levels. Metal contamination may occur during silicon handling, high-temperature heat treatments or ion implantations. Unintentional contaminants, Fe, Cu, Ti, Ni are usually found in silicon in concentrations of $\sim 10^{11}$ - 10^{12} cm^{-3} [11].

Poole-Frenkel and trap-assisted tunneling effects that occur at high electric fields (> 10^5 V/cm) can greatly enhance the emission rate of deep energy levels (field-enhanced generation) [12][13]. At even higher field intensities (>$7 \cdot 10^5$ V/cm), direct band-to-band tunneling may take place, that is, strong generation of free carriers in the junction without the assistance of deep energy levels [14][15]. Tunnel-assisted generation is not reduced by lowering the temperature and therefore sets a limit to the reduction of the dark count rate obtained by cooling the detector. An important



(a) (b)

Fig. 4.5 (a) Trapping and release of an electron by a deep level (CB is the conduction band; VB is the valence band); (b) Probability density of afterpulse generation after an avalanche in a 100 µm diameter silicon SPAD operating at room temperature.

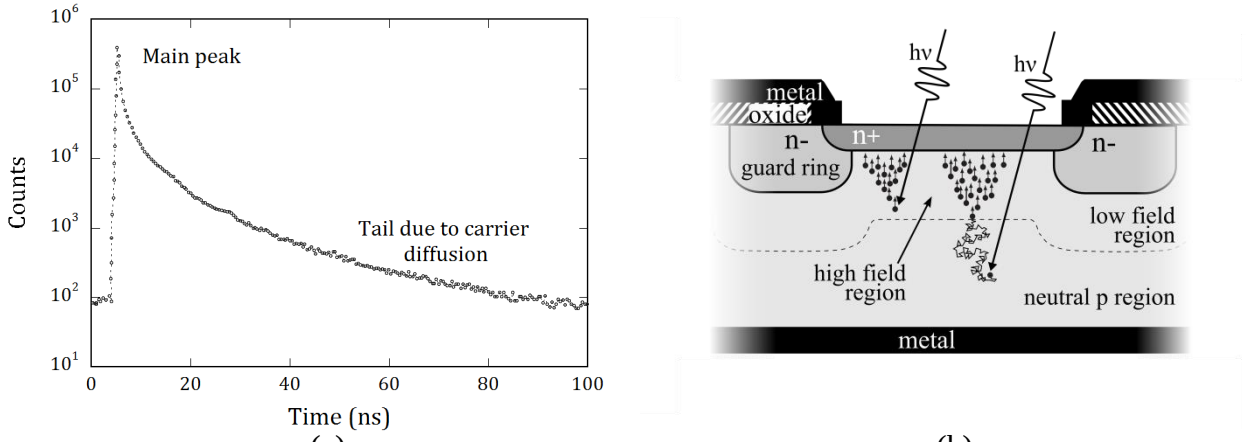
conclusion can thus be drawn for the design of SPADs: the electric field profile within the SPAD junction must be suitably tailored to avoid band-to-band tunneling and field-enhanced generation of carriers. This reduces the detector noise at room temperature and makes cooling more effective in reducing the DCR [16].

4.4.3 Afterpulsing

The noise in SPADs is further increased by an effect that does not play any role in ordinary APDs. Deep levels located at intermediate energies between mid-gap and band edge can act as minority carrier traps. During each avalanche pulse, a few carriers may be trapped in these levels and subsequently released, as outlined in Fig. 4.5a. The released carrier can re-trigger the avalanche, thereby generating "afterpulses" correlated in time to the original avalanche triggered by the photon [5][17][18]. This release is statistical; the emission probability per unit time has a characteristic value for each type of level involved, and the reciprocal of this emission probability is the exponential time constant (trap lifetime) for that level. Afterpulsing effects can be evaluated by using the time-correlated carrier counting technique described in ref.[17]. Fig. 4.5b shows the probability density as a function of time for the occurrence of an afterpulse after an initial avalanche pulse for a 100 µm Si SPAD operated at room temperature.

Afterpulsing introduces a positive feedback loop that can significantly increase the effective dark count rate [18]. Since traps are far from being saturated [18][19], their population increases linearly with the charge that flows during the avalanche pulse. Therefore, to reduce afterpulsing, the total avalanche charge should be reduced as far as possible.

If the trapped charge cannot be reduced to a sufficiently low level, a quenching procedure can be exploited to reduce the afterpulsing rate to a negligible, or at least acceptable level. After an avalanche, by deliberately maintaining the voltage at the quenching level (see Section 4.5) for an extended "hold-off" time, trapped carriers are given time to be released and thus will not retrigger the device when its voltage is returned above breakdown. While operating at lower temperatures improves noise performance of photodetectors, lower operating temperatures exacerbate the afterpulsing problem. This is because the trap-release process becomes much slower at lower



(a) (b)

Fig. 4.6 (a) Photon arrival time distribution measured in a standard TCSPC set up with laser diode pulses of about 20 ps duration at 830 nm and a planar Si-SPAD device as in Fig. 4.2; (b) Outline of avalanche generation by photons absorbed in the depletion layer and in the neutral region.

temperatures [17], requiring a much longer hold-off time to achieve the same level of depopulation and seriously limiting the dynamic range in photon-counting measurements.

4.4.4 Timing jitter

The onset of the avalanche pulse is correlated with the arrival time of the photon that generates the primary charge pair. Due to various physical effects, however, the delay of the instant at which the onset is sensed with respect to the true arrival time of the photon is not constant, but subject to statistical fluctuations. The timing jitter is usually quoted as the full-width at half maximum (FWHM) of the photon arrival time distribution [20]. A typical photon-timing distribution is shown in Fig. 4.6a for the planar device structure developed by Haitz [3][4][5]. Two main components are evident. The narrow peak has a FWHM of about 60 ps and is due to carriers photogenerated in the junction depletion layer, which are immediately accelerated by the electric field (see Fig. 4.6b). The slow tail is due to carriers photogenerated in neutral regions near the depletion layer that migrate by diffusion, eventually reaching the edge of the depletion layer where they are accelerated by the electric field. The tail limits the photon-timing resolution. In QKD applications, diffusion tails can lead to inter-symbol error when the timing jitter exceeds the clock period, resulting in photon events being recorded in subsequent bit periods. In such applications the FW10%M and FW1%M are important parameters as well as the FWHM. Furthermore, the amplitude and duration of the tail are wavelength dependent due to the dependence of the optical absorption coefficient (i.e., of the penetration length) [21]. This is a significant drawback in applications where the light source is not monochromatic. It is clear that SPAD devices intended for photon-timing applications should have a structure designed to minimize any diffusion-tail effects.

The FWHM of the main peak exceeds the contribution due to the noise in the pulse-timing circuits, and can therefore be ascribed to statistical fluctuations in the build-up of the avalanche, from the first seed (the photogenerated electron-hole pair) to the macroscopic current level of the sensing threshold of the timing circuit. Various research groups have endeavored to explain and compute it in terms of a local statistical build-up with essentially one-dimensional models of the current rise. This approach, however, predicted FWHM values shorter than 10 ps. As a possible source of further fluctuations, the lateral propagation of the avalanche from the first seed to the whole detector area was then investigated. Two lateral diffusion mechanisms were highlighted in

the literature: multiplication-assisted lateral diffusion of carriers [22], and photon emission from hot carriers in the avalanche [23]. The former mechanism is dominant in SPAD devices with thin depletion layers (≈ 1 μm), whereas the latter is dominant in SPAD devices having a thick depletion layer (≈ 20 μm), see Section 4.6. The efficiency of both processes is enhanced by a higher electric field; by increasing the excess bias voltage V_E, the photon-timing resolution is improved in all cases [24]. It has been recently demonstrated that remarkable timing performance is achievable by sensing the avalanche current at very low level (about a hundred μA), when the multiplication process is still confined within a small area around the photon absorption point [16][25]. To perform a true low-level sensing of the avalanche current it is critical to preserve the shape of the first part of the leading edge by minimizing any filtering action. To this end, a carefully designed current pick-up circuit must be used [25], as illustrated in Section 4.5.

In the context of SPAD arrays, a further set of performance parameters should be introduced besides the aforementioned ones. The most significant are crosstalk and fill factor.

4.4.5 Crosstalk

Ideally, photons absorbed within the active volume of a pixel are expected to contribute only to the signal of that pixel. In practice, however, such an event can cause detections in neighboring pixels, resulting in crosstalk. Crosstalk reduces the effective spatial resolution of an image sensor, leading to blurring. Two crosstalk mechanisms affect the performance of a SPAD array: optical crosstalk, and electrical crosstalk.

4.4.5.1 Optical crosstalk

Silicon p-n junctions emit photons when operated in avalanche regime. The emission probability is very low: on the average, about one photon is emitted every 10^5 carriers crossing the junction [26]. In monolithic SPAD arrays, photons emitted from a SPAD can trigger an avalanche in another detector, thus causing optical crosstalk between the pixels of the array. The outcome is an incorrect evaluation of the optical signal detected by each pixel of the array. The crosstalk probability increases as the distance between pixels is reduced, and therefore sets a limit to the array density. Optical barriers placed between adjacent pixels (such as deep trenches coated with metals or heavily doped diffusions) cannot completely prevent the optical crosstalk [27] because photons can be reflected at the bottom surface of the chip, thus bypassing the optical barriers and contributing substantially to the crosstalk. A good strategy to minimize this contribution is to adopt thick and highly doped substrates to increase the free-carrier absorption of avalanche-generated photons [28].

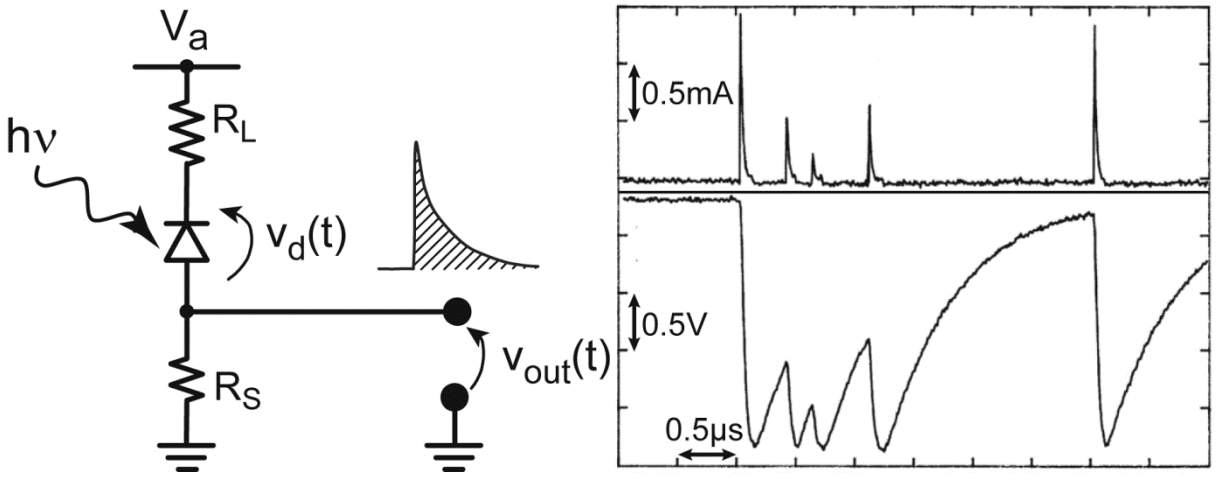
4.4.5.2 Electrical crosstalk

Carriers photogenerated in the quasi-neutral region below the p-n junction can diffuse laterally and trigger an avalanche in a neighboring pixel. Since the mean penetration depth strongly increases with wavelength, photons in the red and near-infrared ranges tend to induce more electrical crosstalk than short-wavelength photons. Electrical crosstalk may be effectively addressed by exploiting dielectric and/or junction isolation techniques, as discussed in Section 4.7.

4.4.6 Fill factor

Fill-factor is defined as the active-to-total area ratio of a single pixel. It is a key figure of merit, especially for applications involving diffuse illumination of the SPAD array (such as 3D imaging and profiling of remote objects). In general, the fill-factor is limited by the sizes of the guard ring structure, the isolation structure, and the quenching and counting/timing circuitry associated with each pixel; it is usually of the order of a few percent. Lenslet arrays can be used to improve the fill-factor [29], but at the cost of greater complexity of the system and lower flexibility in applications due to the reduced range of acceptance angles.

4.4.7 Microelectronic structure of a SPAD: outline and basic features



(a) (b)

Fig. 4.7 (a) Schematic circuit diagram of a SPAD in the passive-quenching arrangement (typical values R_L = 500 kΩ and R_S = 50 Ω); (b) Waveform of the avalanche current (upper trace) and of the voltage across the SPAD (lower trace).

Taking into account the overview given above of SPAD performance and parameters, the essential elements of the structure of a SPAD device can now be highlighted. The device core is the ***avalanche region***, where a high electric field provides carrier multiplication by impact ionization. For obtaining uniform DE, the material properties and the electric field intensity in this region must be as uniform as possible over the entire detector area. The thickness of the avalanche region has to be thin in order to limit the zone where the high electric field enhances the thermal generation of carriers, i.e. the DCR. To obtain high DE, carriers photo-generated in the contiguous ***absorption and drift region*** of the depletion layer must also be driven to the avalanche region, but by a lower electric field. In fact, moderate field intensity in this region is sufficient to ensure a saturated drift velocity (i.e. fast collection) and avoids enhancing thermal carrier generation. At the edge of the avalanche zone, deeply diffused ***guard-rings*** or other equivalent structures avoid local concentration of the electric field intensity. The depletion layer is sandwiched between neutral semiconductor layers at the top and bottom, which should be either transparent at the wavelength of interest, or at least very thin. Carriers photo-generated in these layers either recombine and do not contribute signals, or are collected at the avalanche region after diffusing in the neutral region, thus generating a diffusion tail in the photon timing distribution (c.f. Section 4.6.1).

Section 4.5 – Circuit principles for SPAD operation

The circuit that quenches the avalanche and resets the bias voltage plays an integral role in the performance of SPADs [18]. In early studies on silicon avalanche diodes in Geiger mode the simple passive quenching circuit (PQC) outlined in Fig. 4.7a was used. The bias voltage is applied through a large ballast resistor R_L; a small resistor R_S is connected to the other terminal for observing the current pulse. The avalanche current discharges the total capacitance C_T at the diode terminal, which is the sum of the junction capacitance C_d and of the stray capacitance C_s. The voltage V_d across the diode decreases towards V_{BD} and the avalanche current decreases correspondingly. As the voltage V_d approaches V_{BD} the rate of decrease slows down. All the avalanche current flows through R_L and is reduced to the value $(V_a - V_{BD})/R_L$. If R_L is high enough to reduce the current below a few tens of μA, the number of avalanche carriers is small and the probability of interruption of the multiplication chain is high, and the avalanche can be finally quenched [5]. The voltage V_d then starts to recover slowly towards the bias voltage V_a (reset transition), as the small current in R_L recharges C_T with a long time constant $R_L C_T$. During the reset transition, the diode voltage is higher

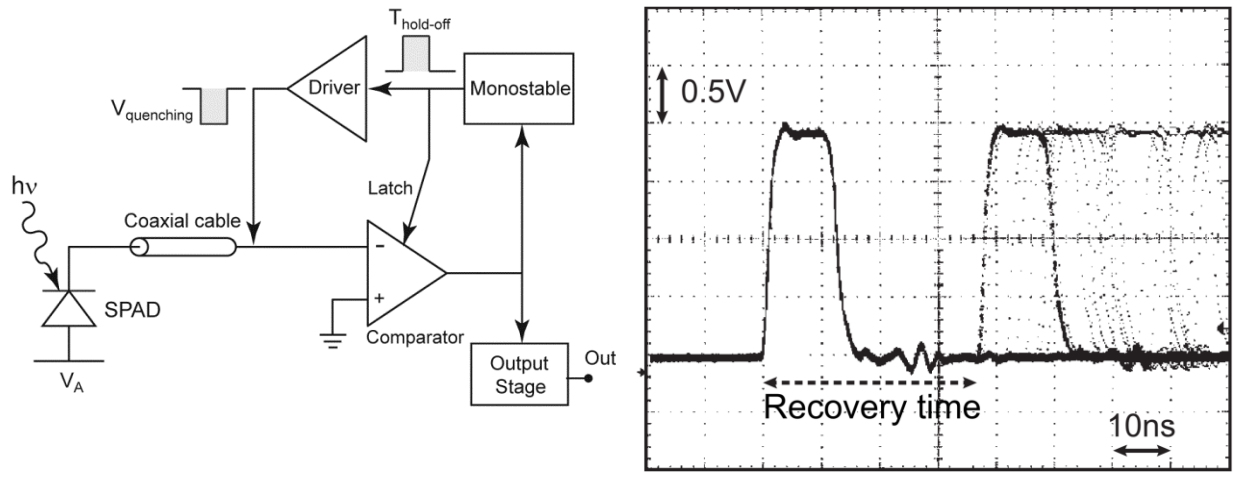


Fig. 4.8 (a) Schematic diagram of a SPAD in an active-quenching circuit (AQC); (b) output pulses from the AQC at a counting rate of ≈ 2 MHz (horizontal scale 10 ns/div; vertical scale 0.5 V/div).

than V_{BD} and an avalanche can be triggered, but the diode fires at a voltage V_d lower than V_a (see Fig. 4.7b). It then operates with lower photon detection efficiency and impaired photon-timing resolution. In a passively quenched SPAD, after each avalanche the triggering probability has a continuous evolution, starting from practically nil and finally reaching the steady-state value determined by V_a. The behavior of the detector is thus peculiar: it is paralyzable, but with a time-dependent sensitivity to triggering events [18]. Furthermore, the voltage and current pulses produced during the reset transition have smaller amplitudes, as shown in Fig. 4.7b. Since a comparator is employed for sensing avalanches, pulses smaller than the threshold level are discarded, producing a dead time that is neither well known nor stable. The result is a loss of linearity at high counting rates that can be measured empirically, but is difficult to model and correct for accurately [18]. In summary, photon-counting measurements can be accurately performed only if the total count rate is low enough to make count losses negligible and correction unnecessary.

The drawbacks of the passive quenching can be reduced, though not eliminated, with modern circuit technologies that significantly reduce the stray capacitance C_s and thus shorten the duration of the reset transition. With surface mounting techniques and miniature components C_s can be reduced to a few picofarads. A monolithic integration of detector and ballast resistor, nowadays possible at least for silicon SPAD's, can reduce C_s well below 1 pF and the reset transition time well below 1 μs.

The solution that completely avoids the drawbacks of the passive quenching is the active quenching circuit (AQC), first devised in 1975 [30]. The principle is simple: to sense the rise of the avalanche and react back at the SPAD, forcing short quench and reset transitions with a controlled bias-voltage source. As outlined in Fig. 4.8a, the sensing comparator triggers a voltage driver to switch the bias voltage down to the breakdown voltage V_{BD} or below. After a controlled hold-off time, the bias voltage is then switched back to the operating level V_a. A standard pulse synchronous to the avalanche rise is derived from the comparator output, and can be used for photon counting and timing (Fig. 4.8b).

General principles and practical features of quenching circuits have been extensively dealt with in a tutorial paper [18]; this section concentrates on recent results and prospects for further development.

The matter may be better analyzed by focusing on the main requirements of the circuit, which concern the avalanche quenching transition, the subsequent hold-off time, and the reset transition.

Quenching should be as fast as possible to reduce the avalanche charge and related effects (power dissipation, carrier trapping, and light emission from hot carriers). To this end it is necessary to minimize both the delay T_d with which the quenching transition starts and the transition time T_q. With passive quenching circuits the transition is started immediately by the avalanche, but the transition time is determined by the time constant $R_D (C_d + C_s)$ (where R_D is the diode resistance during the avalanche). This time constant is fairly long in cases where R_D is not small and/or the capacitance $(C_d + C_s)$ is not minimal. Active-quenching circuits can ensure a very fast transition, but their feedback loop may produce a significant delay T_d, particularly in cases where the SPAD is not located near the quenching circuit [31][32]. A mixed passive-active quenching approach was developed [18], in which quenching is started with a passive transition and then sped up and completed by an active loop. A remarkable reduction of the avalanche charge can be obtained with discrete-component circuits [33], but further reductions can be achieved by integrating the load resistor R_L into the detector chip, and by integrating the complete quenching circuit on a chip [34]. A variant passive-active quenching circuit that was specifically devised for monolithic integration in CMOS technology cuts the avalanche current path instead of driving down the SPAD voltage [35].

The hold-off time, that is, the duration of the low-voltage fully-quenched state, must be minimized to attain the highest counting rate. On the other hand, an adjustable hold-off time may be very useful for reducing the afterpulsing effect by allowing time for the release of the carriers trapped in deep levels [18]. This suggests developing circuits that have a negligible hold-off time and the capability of enforcing longer adjustable duration when needed.

The reset transition is a critical phase in all measurements and should therefore be very fast, at most a few nanoseconds. In photon counting it is highly desirable that the detector have a well-defined dead time. That is, it is acceptable that the detector be totally insensitive for a given time (preferably a short time, of course), provided that it is then reset abruptly to its full efficiency. For this case, accurate equations for the correction of the count losses at high rate are available, and are based on well-known concepts of statistics. In reality, however, the situation is remarkably different: the recovery from zero efficiency is gradual and follows the evolution of the voltage with a non-linear dependence. Equations for accurately correcting the count losses due to such a variation have not been reported in the literature, and carrying out a quantitative statistical analysis of such a complex situation looks problematic. A gradual reset transition is also a significant drawback in photon-timing measurements, because the temporal resolution of SPADs depends strongly on the excess bias voltage. Therefore, a gradual reset transition causes a progressive degradation of the resolution as the count rate is increased. For both photon counting and photon timing, high-quality detector performance can be achieved at high count rates only if the duration of the reset transition, and thus the probability of detecting photons during the voltage recovery, is minimized. For this reason, active reset is explicitly advantageous. Passive quenching circuits have an inherent exponential reset transition, with time constant $R_L(C_d + C_s)$. In all cases where the circuit is not integrated in the detector chip, this time constant is at least a few hundred nanoseconds. It can be reduced to a few tens of nanoseconds in cases where the total capacitance $(C_d + C_s)$ is in the range of 100 fF, as in fully integrated chips with small detector diameter (less than 20μm). But even in such cases the reset transition is gradual over tens of nanoseconds and the drawbacks are reduced, but are not negligible.

The reset must not only be fast, but also very accurate. That is, for ensuring accurate photon counting and timing the bias voltage of the detector must be cleanly restored back to the final level. Therefore, in actively driven reset transitions care must be taken to avoid perturbations such as overshoots and ringing; a final part of the recovery that slowly approaches the steady-state voltage level must be avoided as well. To this end, it may be useful to enforce a reset action that pulls the

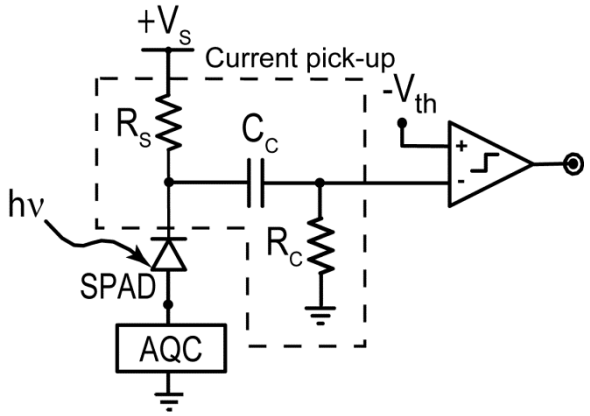


Fig. 4.9 Simplified block diagram of the current pick-up circuit that can be added to any of the existing AQCs for improving the photon timing jitter.

voltage to the target level and keeps it there for a short time (typically 10 ns), that is, to introduce a final hold-on time. However, this approach can result in incorrect timing of photons that arrive during the hold-on time.

Obtaining accurate photon timing sets further requirements that may conflict with the requirements for obtaining fast quenching with minimization of the avalanche charge. In particular, a marked dependence of the timing resolution on the circuit employed for timing the pulses is observed for SPAD devices with active area wider than 10 µm. Such dependence can be understood by taking into account the physical processes that determine the rise of the avalanche current [22] [36][37][38]. During the initial stage of the avalanche, when the multiplication of carriers is localized within a small area around the seed point (the point of incidence of the photon), the avalanche current rises with relatively small statistical fluctuations. The subsequent rise to the avalanche signal's peak value corresponds to the progressive spreading of the multiplication over the sensitive area of the detector, and has stronger statistical fluctuations. Consequently, if the triggering threshold of the circuit that senses the avalanche is reached not in the very first part of the rise, but later during the spreading of the avalanche over the area, the timing jitter is remarkably larger than if the threshold is reached early in the avalanche growth. Any low-pass filtering that slows down the rise of the signal sensed by the timing circuit will have the effect of shifting the triggering instant to later times in the avalanche process, and will therefore degrade the timing resolution. The voltage waveform developed by the avalanche on the SPAD is inherently subject to a low-pass filtering action due to the charging of the diode and stray capacitances ($C_d + C_s$). Therefore, circuits that use this voltage waveform to sense the avalanche are not suitable for high-resolution timing with SPADs that have larger active areas. In fact, they provide good timing performance only in cases where the capacitance ($C_d + C_s$) is reduced to very low level, as in SPADs with less than 10 µm diameter and a low-capacitance quenching circuit. In larger area SPADs the lateral-propagation effect is stronger, and the larger intrinsic capacitance of the diode significantly contributes to the parasitic low-pass filtering action.

Research has demonstrated that the trade-off between active area diameter and time resolution may be overcome by detecting the avalanche current during the initial part of the rise, about at the 100 µA level [25]. By employing a separate current pick-up circuit (see Fig. 4.9), an unprecedented time resolution of 35 ps was obtained with a 100 µm diameter SPAD. This patented technique [39] employs AC coupling with a very fast time constant for extracting a short signal that reflects the rise of the avalanche current, which makes it possible to maintain excellent timing performance up to very high count rates. This technique enables the use of large-area SPADs in high-performance TCSPC measurements, as illustrated by results recently obtained with 200 µm diameter SPADs

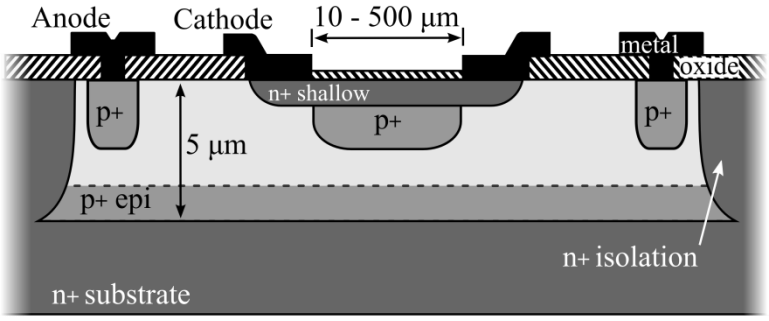


Fig. 4.10 Schematic cross section of the double-epitaxial SPAD device structure reported in [43].

[16]. A monolithic integrated circuit that includes both the active quenching circuit and the current pick-up and timing circuit in a chip has recently been developed [40].

Section 4.6 – Silicon SPAD devices

The silicon SPAD devices reported to date can be grouped in four categories, according to their fabrication technology:

- planar SPADs fabricated in a custom technology;
- non-planar SPADs fabricated in a custom technology;
- SPADs fabricated in a high-voltage complementary metal-oxide-semiconductor (HV-CMOS) technology;
- SPADs fabricated in a standard deep-submicron CMOS technology.

Their features and performance are reviewed in the following sections.

4.6.1. Planar SPAD devices fabricated in a custom technology

4.6.1.1. Early planar SPAD devices

The precursor of modern planar SPADs was introduced at the Shockley laboratory in the early 1960s, during studies of the physics of avalanche multiplication with high electric field intensities [3][4][5]. It was necessary to carry out experiments on avalanching junctions that had reasonably uniform electric field and were absolutely free from the so-called microplasmas (extended defects such as metal precipitates, dislocations, etc.). The approach was to fabricate many n+p junctions with very small diameter (a few microns) surrounded by a deeply diffused guard ring, and then select the few devices that did not contain a microplasma. The n+p junctions were fabricated by diffusing a shallow (< 0.5 µm) *n* + layer in a p-bulk substrate (Fig. 4.2). This simple structure has two key features: it operates at low voltage (about 30 V), resulting in limited power dissipation during the avalanche (a few hundred milliwatts), and it is fabricated in an ordinary silicon wafer with a planar technology, and thus is amenable to monolithic integration with other detectors and circuits. However, a thorough analysis reveals some weaknesses of the early planar structure as a SPAD device. The deep guard-ring diffusion causes the photon detection efficiency to be non-uniform in the active zone, giving it a dome-shaped distribution. This effect arises from the fact that the n-diffusion acts laterally from the edge of the device towards the center of the active junction over a distance almost equal to the diffusion depth, thereby decreasing the net p-doping and increasing the breakdown voltage from the center to the edge. Therefore the excess bias voltage progressively decreases from center to periphery, causing a decrease in the photon detection efficiency. This effect sets a strong limit to the minimum diameter of the active n+p junction, thus

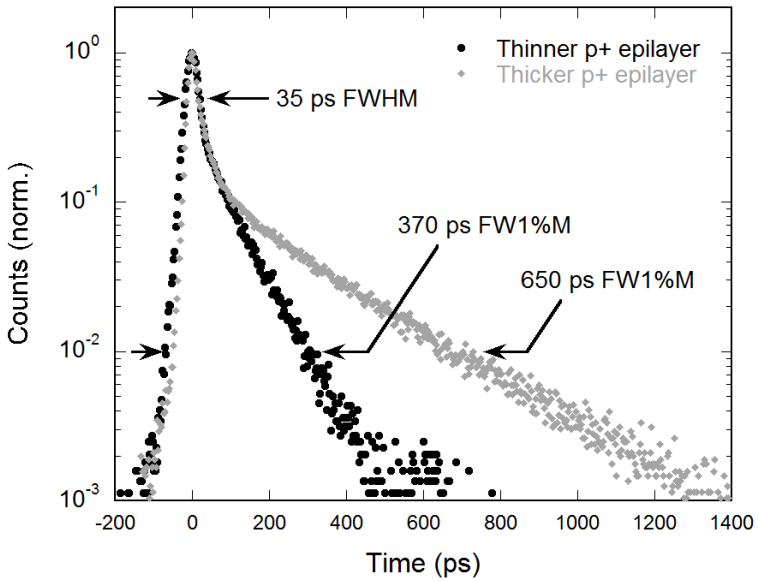


Fig. 4.11 Photon-arrival-time distributions measured with two different 200 µm SPADs and a picosecond laser at 820 nm. The curves shows a prompt peak with a FWHM of 35 ps and a clean exponential diffusion tail with a time constant of 280 ps (diamonds), and 110 ps (dots), corresponding to a thickness of the neutral p+ buried layer of 2.4 µm, and 1.5 µm, respectively.

limiting the scalability of the structure for implementing arrays. Furthermore, due to the deep guard ring, the diffusion tail of the photon-timing distribution has a multi-exponential, wavelength-dependent shape. This makes the study of fast fluorescent decays more complex, especially when the spectral distribution of the incident light is not accurately known [21]**Error! Reference source not found.**[41].

4.6.1.2. Planar SPAD devices on epitaxial silicon substrates

In the past three decades, a number of custom planar technologies have been developed for fabricating SPAD devices with optimized performance [42][43][44][45][46][47]. The planar epitaxial devices outlined in Fig. 4.10 were first introduced in 1988 [43][48] to overcome the drawbacks of the early planar structures. These devices have undergone continuous improvement and are now exploited in commercially available photon detection modules from, for example, Micro Photon Devices [49].

The SPAD fabrication starts from an n-type substrate on top of which a p+/p- double-epitaxial layer is grown. The p-n junction formed between the epitaxial layer and the substrate limits the neutral region from which carriers are collected, thereby shortening the diffusion tail. The active n+p junction is built in the upper (2.5 µm thick), low-doped p-epilayer (10 Ω-cm). The buried p+ epilayer (2 µm-thick) establishes a low-resistivity path (0.3 Ω-cm) to the side ohmic contact. The extrinsic guard ring used in the early planar devices is replaced by a 'virtual' guard ring structure. The concept of this virtual structure is straightforward: instead of using a lightly doped n-diffused guard ring to reduce the field in the peripheral region, the electric field in the central region of the shallow n+-p junction is locally enhanced by means of a higher p-doping (*enrichment*). Implantation of boron followed by a drive-in diffusion is used to produce a ~ 1 µm thick enrichment region, which defines the active area of the device. A highly-doped, p-type diffusion (*sinker*) provides a low-resistivity path for the avalanche current flowing from the buried epilayer to the anode contact. Finally, a highly doped, n-type diffusion (*isolation*) region completely surrounds the detector. As a result the SPAD is enclosed in a p-well delimited by the isolation and by the

substrate. This makes it possible to electrically isolate the detector from other SPADs or electronic devices fabricated on the same chip, thus allowing the fabrication of arrays (see Section 4.7) and monolithic integration of detectors and circuits in a chip.

Various design and fabrication parameters such as the boron-implanted dose, the conditions of the drive-in diffusion, the thickness and doping of the lightly doped epitaxial layer and of the buried layer can be easily customized for achieving a desired performance [16]. Continued improvement of the planar epitaxial technology, as described in [43], makes it possible to reliably fabricate SPAD devices with large active-area diameters (up to 500 µm) and exhibiting an excellent compromise between breakdown voltage (typically around 30 V), DE (50 % peak at 550 nm, decreasing to 25 % at 730 nm, and 12 % at 850 nm), DCR (from 10 s^{-1} to 10^3 s^{-1} at -15 °C for SPAD diameters ranging from 50 µm to 500 µm), total afterpulsing probability (about 1 % at -15 °C), and timing jitter (better than 40 ps FWHM). Figure 4.11 (diamonds) shows the photon-timing distribution of a 200 µm SPAD detector illuminated with 10 ps FWHM optical pulses at 820 nm. The curve shows a prompt peak with a FWHM of 35 ps, and a clean exponential diffusion tail whose time constant is 280 ps [43].

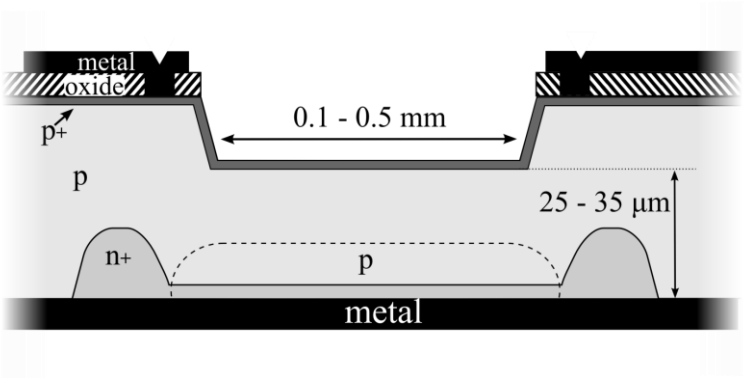
Figure 4.11 (dots) demonstrates the advantage provided by fully-custom SPAD technologies. By simply reducing the thickness of the p+ epitaxial layer from 2.4 µm to 1.5 µm (in a different device), a reduction of the tail lifetime from 280 ps to 110 ps can be obtained.

Since the lifetime of the exponential tail does not depend on the photon wavelength, the timing resolution of the double-epitaxial SPAD is almost completely wavelength independent, which provides a remarkable advantage in reconvolution analysis of fast fluorescent decays [41][21]. To attain even better timing resolution completely free from diffusion tails, a fairly complex modification of the double-epitaxial device has been devised, fabricated, and tested [50]. The basic idea is to eliminate the neutral region beneath the active junction by exploiting a patterned buried layer. It has been verified that the diffusion tail can be completely eliminated in SPADs with diameters of up to about 10 µm. However, the performance of wider detectors remains less satisfactory, and the fabrication process is clearly more difficult than that of the previous double-epitaxial devices.

New developments in planar-epitaxial SPAD technology are mainly concerned with improving the DE in the red wavelength range (600 nm to 900 nm), either by incorporating a resonant cavity in the device structure or by increasing the thickness of the absorption region.

A resonant-cavity-enhanced (RCE) SPAD fabricated on a reflecting silicon-on-insulator (SOI) substrate has been reported by Ghioni *et al.* [51] and successfully exploited in three-dimensional imaging and QKD applications [52][53]. The RCE SPAD detectors have peak detection efficiencies ranging from 42 % at 780 nm to 34 % at 850 nm, and timing jitter of 35 ps FWHM. Typical dark count rates of 450 s^{-1}, 3500 s^{-1}, and 10^5 s^{-1} were measured at room temperature with RCE SPADs having 8 µm, 20 µm , and 50 µm diameters, respectively.

More recently, a red-enhanced (RE) SPAD device was devised, fabricated and characterized [54]. The key feature of RE-SPADs is a separate absorption and multiplication structure that provides a thick (≈10 µm) absorption region with low electric field (hence no multiplication and negligible field-enhanced carrier generation), and a shallow multiplication region with a peaked electric field profile (designed to enhance the avalanche triggering probability and to reduce the photon timing jitter). The electric field profile in the two regions was designed to achieve the optimal trade-off between operating voltage, avalanche triggering probability (thus DE), DCR, and timing jitter. Experimental measurements on 50 µm RE-SPAD devices showed a significantly improved DE in the red region, reaching 40 % at 800 nm wavelength (i.e. a factor 2.5 higher than the DE of standard planar SPADs) and 60 % at 550 nm wavelength. The devices exhibit a



remarkably low DCR, less than 10^3 s^{-1} at room temperature, decreasing to a ≈50 s^{-1} at 5 °C. Although the active volume of RE-SPADs is considerably larger than that of standard planar

Fig. 4.12 SLiK™ device developed at the former RCA ElectroOptics, now Excelitas Technologies Corp. [55][56].

SPADs due to the increased thickness of the absorption layer, the DCR of the two detectors is comparable at room temperature. The thicker absorption region of the RE-SPAD does not significantly contribute to the DCR because the low electric field practically rules out field-enhanced generation of carriers. The dominant contribution to the DCR comes from the high-field multiplication region, whose design remained substantially unchanged. A timing jitter of 93 ps FWHM was obtained at room temperature, which is higher than the typical figure of standard planar SPADs (30-50 ps). This is due to the increased thickness of the absorption and drift region, resulting in increased transit times for photo-generated electrons of about 10 ps/µm at the saturated speed of 10^7 cm/s. Since photons are absorbed randomly in the drift region, timing jitter of ≈100 ps FWHM can be attributed to the 10 µm thick absorption region. Total afterpulsing probability lower than 1.5 % was measured over the entire temperature range of operation.

4.6.2. Non-planar SPAD devices fabricated in a custom technology

The SLiK™ device sketched in Fig. 4.12 was devised by R.J. McIntyre and P. Webb at the former RCA Optoelectronics (now Excelitas Technologies Corp.), and employed to produce highly successful single-photon counting modules (SPCM) [55][56]. SLiK™ stands for 'Super-low k,' where k denotes the ratio of the ionization coefficient of holes to that of electrons. The device represents a remarkable evolution of the previous reach-through avalanche diode structure pioneered by the same team [57][58]. It is built in special ultra-pure high-resistivity silicon wafers with a dedicated technological process; various device features and processing steps are proprietary and covered by patents [59][60]. The active area of the detector is fairly wide (≈180 µm). It is defined by a p+ implant and deep diffusion in the central region of the bottom silicon surface and by a localized reduction of the wafer thickness to ≈30 µm, obtained by accurately etching the back of the wafer. A lightly diffused n guard ring around the shallow n+ layer avoids edge breakdown. The device is illuminated from the back. Since the lightly doped p region (from 20 µm to 30 µm thick) is fully depleted, diffusion of photo-generated carriers takes place only in the surface p+ layer (a few microns thick). The decrease of the electric field from its maximum at the n-p junction is gradual, hence the avalanche region is fairly extended. The breakdown voltage is high and strongly varies from sample to sample over a wide range from 250 V to 500 V. Thanks to the thick depletion layer, the DE is very high in the visible region and fairly good in the NIR up to about 1 µm. The typical value is significantly higher than 50 % over the entire range from 540 nm to 800 nm (the peak DE is 65 % at 650 nm), and is still about 3 % at 1064 nm [56]. Notwithstanding the remarkably large volume of the depletion layer, the DCR is very low, ranging from ≈100 s^{-1} to ≈10^3 s^{-1} at -10 °C. The afterpulsing probability is also strongly reduced, typically well below 1 %. The

timing performance is moderate: with a broad illumination on the active area the timing distribution has a relatively broad peak with ≈450 ps FWHM, and an exponential diffusion tail that is one decade lower and has ≈160 ps lifetime. However, significant improvement in the timing

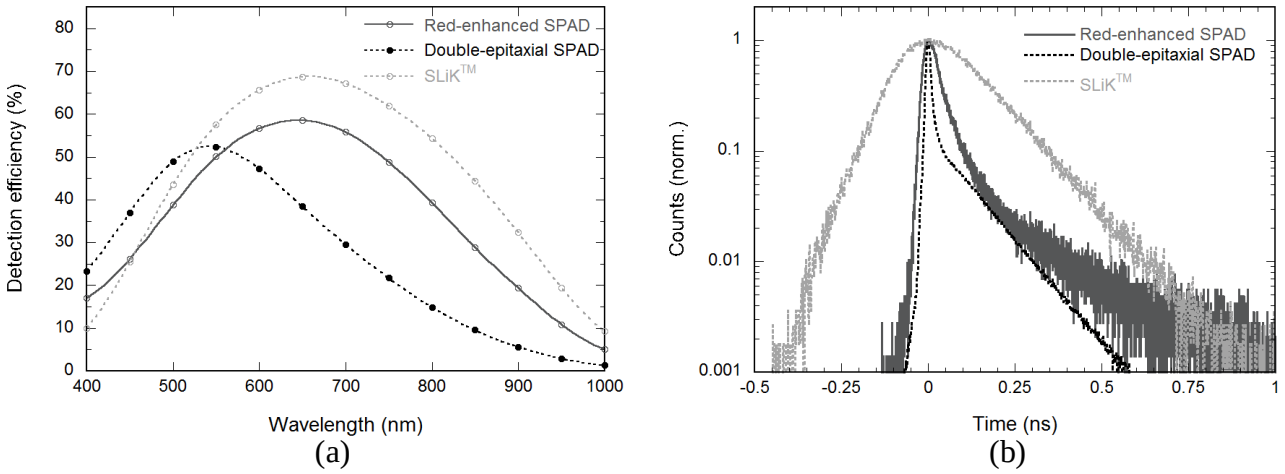


Fig. 4.13 Detection efficiency (a) and photon timing distribution (b) of three different types of SPADs: SLiK™, planar epitaxial SPAD, and red-enhanced SPAD.

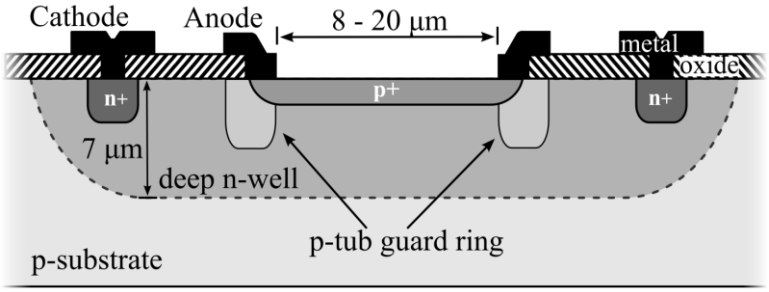
performance can be gained by focusing the light at center of the active area and by using the current pick-up circuit discussed in Section 4.5 [61]. Devices similar to the SLiK™ have been recently developed by Laser Components, Inc., [62], having a larger active area (500 µm diameter) and a timing jitter ≤ 200 ps FWHM [63].

The SLiK™ devices have very good performance, but also a number of practical drawbacks. Due to the high breakdown voltage, the power dissipation during the avalanche is high, from 5 W to 10 W, and very effective cooling of the detector under normal operating conditions is mandatory (with Peltier stages, or other means) [55]. The special fabrication technology is inherently complex, and the production yield of good devices is low and the cost is high. The devices are delicate and degradable, and integrating multiple detectors, or associated circuitry, is not possible. For reference, Fig. 4.13 compares the DE and timing jitter of three different devices, a SLiK™, a double epitaxial SPAD [43] and a RE-SPAD [54].

4.6.3 High-voltage, complementary metal-oxide semiconductor (HV-CMOS) SPADs

Foundry services for fabricating CMOS circuits have been available since the early 1990s, but for some years the available technologies have been far from meeting the requirements outlined above for fabricating SPADs, and all attempts gave poor results. The quality has been then steadily improving, but the progress in IC technologies is driven by the demands of circuits for consumer electronics, which are usually different from those of SPADs. Nevertheless, the monolithic integration of SPAD devices and CMOS circuits offers manifest advantages, from the availability of a fully supported, mature and reliable technology at reasonably low cost, to the possibility of developing complete systems on chip with a high degree of complexity.

The requirement for SPAD integration is that a suitable subset of fabrication steps can be specified within a complex CMOS process flow and used to build a planar p-n junction free from edge effects. However, some stringent technological requirements for SPAD fabrication must be carefully fulfilled. First of all, the quality of both the starting material and the fabrication process must guarantee very low concentrations of impurities (particularly transition metal contaminants) that create deep energy levels within the silicon gap acting as generation-recombination centers and afterpulsing centers. Another key requirement is the ability to keep the electric field strength within



the depletion region low enough to avoid band-to-band tunneling effects and to reduce the field-enhanced carrier generation as much as possible.

In recent years, the development of high-voltage HV-CMOS technologies has been fueled by the rising demand for integrated circuits for automotive and control electronics (that is, circuits for

Fig. 4.14 Schematic cross section of a typical HV-CMOS SPAD device, consisting of a shallow p+/deep n-well junction surrounded by a p-well acting as an extrinsic guard ring to prevent edge breakdown.

the operation of motors, actuators, sensors, etc.). These circuits must fulfill more severe specifications than ordinary CMOS circuits; in particular they must operate with much higher voltage, and this imposes various requirements that are fairly consistent with those of SPAD devices. The main advantage of HV-CMOS technologies is the availability of a relatively low-doped deep n-well that provides up to 50 V isolation from the substrate [64]. HV-CMOS SPAD devices typically consist of a shallow p+/deep n-well junction surrounded by a p-well acting as an extrinsic guard ring that prevents edge breakdown (Fig. 4.14). By relying on the p-well guard ring, a breakdown voltage in the active area above 20 V can be obtained [64]. The deep n-well/p-substrate junction (a few microns deep) limits the depth of the neutral region from which minority carriers can be collected, thus reducing the length of the diffusion tail to less than 10 ns [64]. A number of small-area SPADs with diameter ≤ 20 µm and fairly low dark count rates were obtained using a 0.8 µm HV-CMOS technology by independent research groups [64][65][66][67]. Chips with more detectors and associated circuitry were implemented [65], as well as complete photon-counting modules (detectors and active quenching circuits) [66].

Due to the limitations of the 2-metal 0.8 µm HV-CMOS technology and its eventual obsolescence, there has been a push to migrate to more advanced technologies, and this has resulted in the first successful implementation of SPAD devices in a 0.35 µm HV-CMOS technology [35] [68][69][70]. These 0.35 µm HV-CMOS SPADs exhibit a moderate DCR ($\approx 10^3$ s^{-1} for a 20 µm detector) and a maximum DE of 35 % at 450 nm when biased 4 V above breakdown (Fig. 4.15). The DE however drops to ≈20 % at 600 nm and it is < 5 % at 800 nm [35]. The reduction of the concentration of deep levels is not satisfactory in standard CMOS technologies, but the afterpulsing probability can nevertheless be reduced to fairly low levels by integrating the quenching circuit in the detector chip. In fact, the intrinsic capacitance of a small-area SPAD can be less than 100 fF. By integrating the quenching circuit, the stray capacitance can be brought to comparable level, thereby reducing the total capacitance by more than one order of magnitude with respect to an off-chip quenching circuit. The avalanche charge required for discharging a capacitance of about 100 fF is reduced to $\approx 6 \times 10^5$ electrons per Volt of excess bias. This minimization of the avalanche charge is also useful for reducing the optical crosstalk between adjacent SPAD detectors in a monolithic array, since the light emission from the hot avalanche carriers is proportional to the number of carriers that flow through the junction [26].

A significant shortcoming of HV-CMOS SPADs arises from the p+n polarity of their active junction. Most of the depletion layer of this junction is accommodated in the n-well, so that the avalanche current is mainly triggered by holes, which are the minority carriers generated in this zone. In silicon, holes have a lower probability of avalanche initiation than electrons [8][9], and as a consequence the DE of a p+n SPAD is inherently lower than that of a complementary device

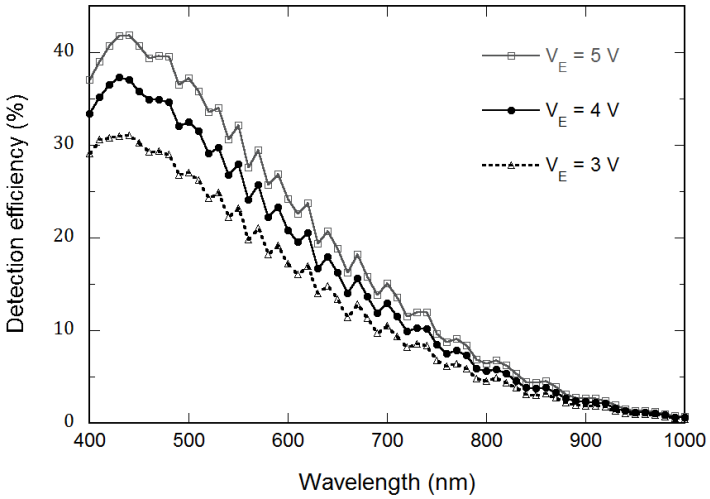


Fig. 4.15 Detection efficiency of a 0.35 µm HV-CMOS SPAD device as a function of wavelength [35]. Measurements were performed at different excess bias voltages.

structure with reversed n+p polarity. Another drawback of the currently available CMOS SPADs is the deeply diffused guard ring that causes non-uniform DE over the device active area, as discussed in Section 6.1.1.

4.6.4. Standard deep submicron CMOS SPADs

CMOS technology with deep submicron (DSM) resolution is mandatory for the fabrication of dense SPAD arrays with large numbers of pixels, adequate fill factor, and smart pixels that include integrated electronics. However, a challenging basic issue must be faced: the inherent features of DSM CMOS technologies, namely the relentless trend toward higher doping levels, lower thermal budget, and thinner p- and n-well layers, conflict with SPAD detector performance. The smaller depth of carrier-collection layers limits the DE, and the high electric fields arising from higher doping result in strongly enhanced DCRs due to band-to-band and trap-assisted tunneling effects. In addition, the reduced thermal budget and the lack of external gettering processes [71] also have adverse effects on afterpulsing.

Thus, the first major challenge for SPAD fabrication in DSM CMOS technologies is the choice of a suitable breakdown region that avoids trap-assisted and band-to-band tunneling. A second challenge that must be faced when scaling the SPAD size to a few micrometers is the design of a suitable guard-ring structure that is effective in preventing premature edge breakdown. To address these issues, designers have to cope with a number of design layers, models, and rules, without any flexibility in changing or adapting a process parameter to better match the requirements of a SPAD.

A number of SPAD structures in DSM CMOS have been proposed in recent years [72] [73][74][75][76][77]. Most of them employ an explicit guard ring of low-doped p-well material around the central p+/n-well breakdown region similar to that shown in Fig. 4.14. Two SPADs with this structure have been reported at the 180 nm process node [72][73]. However, as discussed in [72], this device cannot be scaled much below 5 µm diameter because the p-well regions become so close that the active area of the SPAD is almost fully depleted (see also Section 4.6.1.1). A SPAD fabricated with 180 nm CMOS technology [74] proposes the use of shallow-trench isolation (STI) as the guard region, this approach being free from obvious limitations on minimum detector size

and capable of fitting SPADs together compactly with other electronics. Unfortunately, an unacceptably high DCR of 10^6 s^{-1} was observed, which is likely due to the high density of deep-level carrier generation centers at the Si-SiO_2 interface [78]. If the active region of the SPAD is in direct contact with the STI walls, as in [74], the injection of free carriers into the avalanche region of the detector results in greatly increased DCR.

With 130 nm CMOS technology, various devices employing a p-well guard ring have been reported, invariably with high DCR (40×10^3 s^{-1} to 100×10^3 s^{-1}) [76]. Passivation of STI-interface traps failed to reduce DCR significantly, leading to the assumption that tunneling is the dominant mechanism [75].

Some improvements in compactness and scalability may be obtained by adopting the virtual guard ring structure shown in Fig. 4.10. Richardson *et al.* [79] recently introduced three SPAD structures based on a novel retrograde buried n-well guard ring, capable of scaling from 32 μm to 2 μm in diameter. One of these structures is compatible with a standard 130 nm, triple-well CMOS technology. A remarkable sub-100 s^{-1} DCR for an 8 μm diameter SPAD was achieved at room temperature with 0.8 V excess bias, a maximum DE of 25 % at 560 nm, and negligible afterpulsing.

SPAD devices have been demonstrated in 90 nm CMOS technologies, but with significantly lower performance [80]. A notable exception is the 90 nm SPAD device reported by Webster *et al.* [81], where the deep n-well/p-epi junction is used as the active junction, achieving a peak DE of 44 % at 690 nm and better than 20 % at 850 nm. The 6.4 μm diameter SPAD also achieves a low DCR of 100 s^{-1} along with a low afterpulsing probability of 0.375 % at 0.4 V excess bias voltage. Timing jitter as low as 50 ps FWHM was demonstrated, although the timing distribution had a relatively long diffusion tail. The key performance parameters reported for a variety of individual SPADs fabricated in standard deep-submicron technologies are summarized in Table 4.1.

Section 4.7 – Silicon SPAD Array detectors

Depending on the kind of application envisaged, two distinct directions are emerging for the fabrication of SPAD arrays. The first one is motivated by fast-growing applications like 3D imaging based on direct and indirect time-of-flight (TOF) techniques and low-light-level 2D imaging at high frame-rate, both of which require SPAD arrays with high pixel number and small pixel size, monolithically integrated in systems with electronics for information processing, that is, arrays with in-pixel electronics.

The integration of SPAD devices and associated electronics in submicron and deep-submicron CMOS technologies paved the way for the fabrication of large SPAD arrays that offer distinct advantages over charge-coupled device (CCD) and CMOS active pixel sensor (APS) imagers in these applications. Niclass *et al.*[82] first reported a large SPAD array implemented with 0.8 μm HV-CMOS technology. The array comprised 32 x 32 pixels, each with an independent SPAD device and a five-transistor digital circuit that provided quenching, pulse-shaping and column-access functions. The digital circuit occupies a square area of 54 x 54 μm^2, while the active area of the SPAD is 38 μm^2, resulting in a fill-factor of about 11 %. Photon-timing operations were performed off-chip using a CMOS time-to-digital converter (TDC), and overall timing jitter of 115 ps FWHM was measured on a pixel. The main drawback of this design is that a sequential access is used, meaning that only one pixel can be processed at a time. While optical scanning is eliminated, frame rates remain relatively low (e.g., 250 frames per second for a pixel exposure time of 4 μs).

In-column array architectures were then introduced by the same research group, whereby processing is shared among clusters of pixels (for example, columns). To address the readout bottleneck an event-driven approach was devised, consisting of using the column as a bus that is addressed every time a photon is detected. The address of the relevant row is sent to the bottom of

the column, where the photon time of arrival is evaluated, either off-chip [69][83][84], or on-chip [85]. The drawback of this approach is that multiple photon events cannot be detected simultaneously on the same column, which restricts the use of the event driven readout to applications where the expected photon flux hitting the sensor is low.

Table 4.1

Reference	Technology node	SPAD Diameter (µm)	V_E (V)	DCR (s^{-1}) @ room Temp.	DE max	Total afterpulsing probability (%)	Dead Time (ns)	Timing Jitter FWHM (ps)
Faramarzpour 2008 [72]	0.18 µm	10 - 20	2	70 k - 300 k	5.5 % @ 450nm	-	30	-
Marwick 2008 [73]	0.18 µm	10	0.5	100	-	-	-	-
Pancheri 2011 [77]	0.15 µm	10	5	230	32 % @ 470nm	2.1	30	170
Niclass 2007 [76]	0.13 µm	10	1.7	100 k	34 % @ 450nm	-	450	144
Gersbach 2009 [75]	0.13 µm	9	5	11 k	36 % @ 480nm	-	-	125
Richardson 2011 [79]	0.13 µm	8	1.2	40 @ V_E=0.8V	25 % @ 560nm	-	-	180
Karami 2010 [80]	90 nm	8	0.13	8.1k	9 % @ 480nm	32	1200	398
Webster 2012 [81]	90 nm	6.4	0.4	100	37 % @ 680nm	0.375	15	82

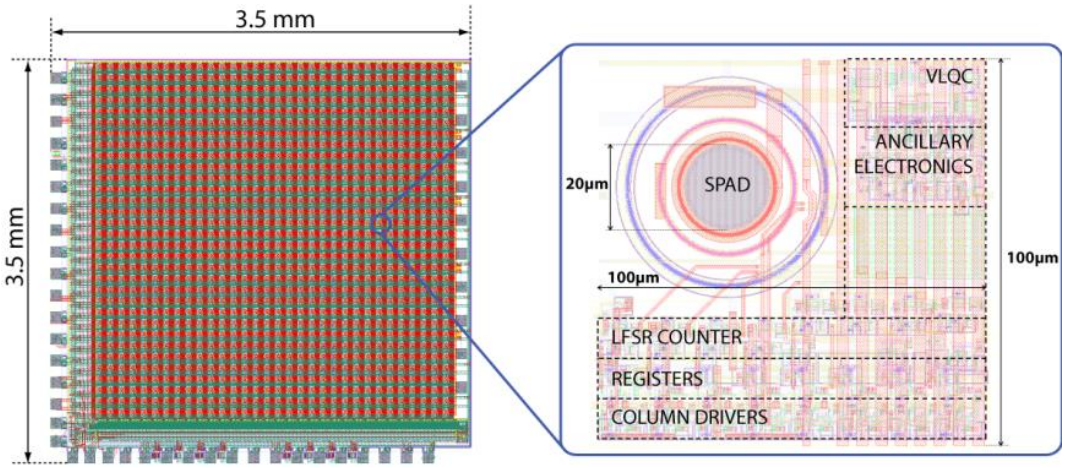


Fig. 4.16 Layout of a monolithic array of 32 × 32 smart pixels fabricated in a 0.35 µm HV-CMOS technology [89]. The zoom shows the layout of a single pixel, including the SPAD detector, the analog sensing and quenching front-end (VLQC), and the digital counter and latch register.

An alternative readout approach that allows the simultaneous detection of photons over an entire column is the latchless pipeline scheme [86]. In this approach, the absorption of a photon causes the SPAD to inject a digital signal into a delay line, which is then read externally. The timing of all injected pulses is evaluated to recover the time of arrival of the photon and to determine the pixel of origin.

The pixel access problem can also be overcome if time discrimination, photon counting, and any additional functionality (including local storage) is performed on-pixel. The advantage of this approach is the massive parallelism that can be achieved, potentially improving the number of photons that can be detected and processed at the same time at reasonable power consumption. Many embodiments of this approach exist, depending on the level of complexity implemented at each pixel. The simplest one uses in-pixel information storage capability as realized by a counter [87][88][89]. Guerrieri *et al.* [89] designed and fabricated a high-speed single-photon camera based on a monolithic array of 32 x 32 smart pixels fabricated in a 0.35 µm HV-CMOS technology (Fig. 4.16). Each pixel (100 µm x 100 µm) is a completely independent photon-counting channel that includes a 20 µm diameter SPAD, analog sensing and avalanche quenching electronics, digital processing for counting the incoming photons, and memory and buffer stages for global shutter readout with no dead-time. Better than 35 % peak DE is attained at 450 nm, decreasing to 8 % at 800 nm, with DCR in the range of 10^3 s^{-1} for more than 75 % of the SPAD devices. The 32 x 32 2D-imager can operate up to 10^5 frames per second with a dynamic range of 8 bits for counting. Noteworthy results have been obtained in challenging experiments [90], and work is in progress toward the in-pixel integration of a TDC [91]**Error! Reference source not found.**.

With the implementation of the first SPADs in 130 nm CMOS technologies [75][76, p. 130] it has been possible to integrate more functionality on a pixel, and remarkable results have been obtained [92]. A number of SPAD arrays were developed in which each pixel contains a multibit counter and a picosecond resolution TDC [93][94] or time-to-amplitude converter (TAC) [95]. Recently, a sensor was reported based on this concept, capable of detecting single photons over an array of 32 x 32 pixels, simultaneously evaluating their time-of-arrival with a time-bin width of 119 ps and a 10 bit range [96]. The array exploits the 8 µm low-noise SPAD devices described in [97], having a median DCR of 10^2 s^{-1} and a peak DE of 25 % at 460 nm when biased at 1 V above breakdown. Each channel operates independently, and contributes to an overall data rate from the

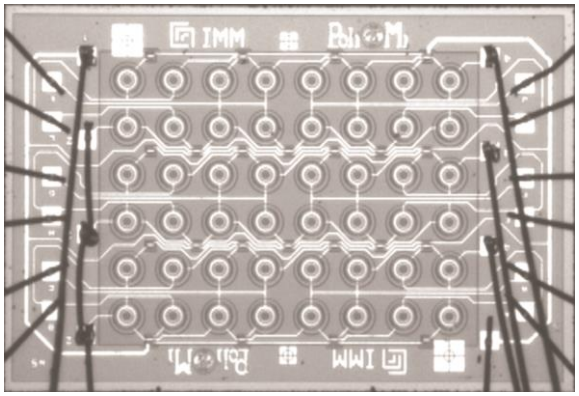


Fig. 4.17 Microphotograph of a 6 × 8 SPAD array detector fabricated using a double-epitaxial silicon technology [106].

chip of up to 10 Gb/s in time-correlated operation mode. The detector active area is 50 μm^2 and total pixel area 2500 μm^2; hence the fill factor does not exceed 2 %. To mitigate this limitation, an array of microlenses based on a design described in [98] was used. The resulting concentration factor was characterized over all the pixels and showed strong variation across the array, with a median value of ≈5, corresponding to an effective fill factor of approximately 10 %. This SPAD array was successfully used in wide-field fluorescence-lifetime imaging (FLIM) [99] experiments in the blue/green wavelength range [96]. More recently, an additional step was reported towards higher spatial and timing resolution with a new sensor of 160 x 128 pixels and 140 ps FWHM instrument response function [100]. The design includes a phase-locked-loop-stabilized 10 bit TDC array with 55 ps time bins.

The second direction in the fabrication of SPAD arrays is driven by applications in life sciences, such as fluorescence correlation spectroscopy (FCS) [101], multi-photon multifocal microscopy [102], spectrally-resolved fluorescence lifetime imaging (SFLIM) [1][99], luminescence/chemiluminescence detection in protein microarrays [103][104], fluorescence resonant energy transfer (FRET) [99]. In all of these applications the basic goal is to increase both throughput and miniaturization of the measurement system. These applications require large pixel sizes (50 μm to 100 μm diameter), high DE, and arrays of small or moderate pixel number (<100). An optimization of DE in the green/red region allows rapid and efficient detection of fluorescent emission from minimal quantities of biological material, i.e., from extremely small samples (down to single molecules of DNA and proteins). Large-area SPAD pixels are preferred to facilitate alignment of the detector array and to achieve good optical collection efficiency. Detectors used in multi-spot experiments (i.e., parallel excitation and detection of multiple spots in a sample) must be able to collect light from each individual spot with minimum contamination by emission from other spots. Although one could devise multiplexing schemes using a single detector to collect and disentangle signals originating from different locations, it is simpler and more effective to use multiple-element detectors with a distinct element for each individual spot. It is also worth noting that a low (<<1) fill factor is required in multi-spot detectors that must avoid optical cross-talk [105]. This requirement distinguishes these applications from the usual imaging applications, where a fill factor as close as possible to 100 % is generally preferred.

To meet these requirements, a research effort using the dedicated SPAD technology described in Section 4.6 was used to fabricate SPAD arrays with large-area elements. As an example, Fig. 4.17 shows a 6x8 SPAD array developed for chemiluminescent array detection and parallel FCS [106][107]. The pixels have 50 μm diameter and 240 μm pitch. A low DCR was obtained at 5 V excess bias voltage at moderately low temperature (-15°C with Peltier element cooling); the

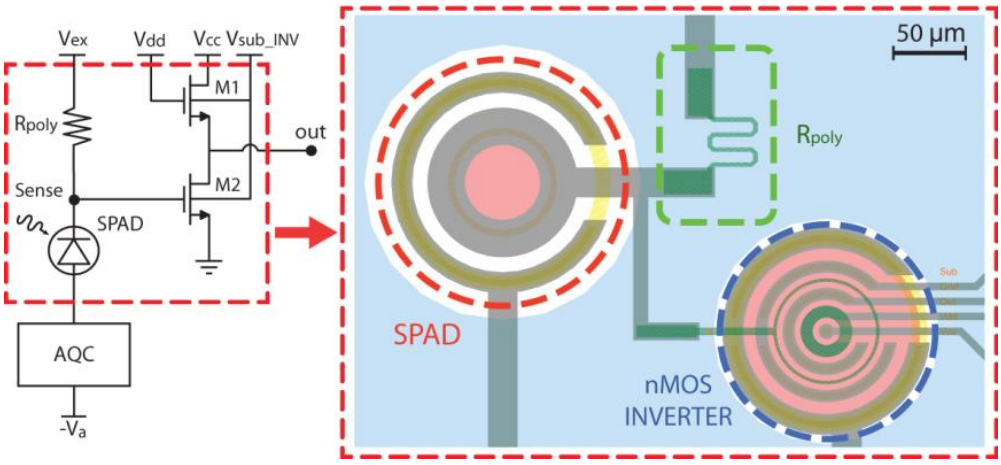


Fig. 4.18 Schematic circuit (left) and layout (right) of a monolithic pixel including the SPAD, n-MOS inverter, and polysilicon resistance [111], fabricated in a full custom technology.

individual pixel DCR is 60 s^{-1} for about 40 % of the elements and is below 5700 s^{-1} in the rest of the array. It was verified that the probability of optical crosstalk between elements is lower than 0.2 %. Each pixel of the SPAD array shown above is connected to an integrated active-quenching circuit (i-AQC) that provides active-quenching and active reset pulses with a dead-time of 65 ns, enabling a saturated photon counting rate of about 15 x 10^6 s^{-1}.

SPAD arrays with large-area elements and various geometries were successfully used in single-molecule FCS [108] and FRET [105] measurements and for wavefront sensing at high frame rate (> 10^4 frames per second) in adaptive optics systems [109][110].

Due to the electrical coupling between adjacent pixels, the incorporation of high-performance photon timing capabilities into custom SPAD arrays is a challenging task even with small numbers of pixels. Fast and large voltage transients (≈1 V/ns) generated by the AQCs cause electrical disturbances on nearby pixels, which preclude low avalanche-sensing thresholds. As explained in Section 4.5, this increases timing jitter. An effective solution is monolithic integration of the avalanche pick-up circuit in close proximity to the SPAD. Thanks to the reduction of parasitic capacitances resulting from integration, it is possible to attain low timing jitter even with higher thresholds, reducing the issues due to electrical crosstalk. The lower parasitic capacitance also reduces the number of charge carriers flowing through the device during the avalanche, reducing both the afterpulsing probability and optical crosstalk.

To enable the integration of MOS transistors in a custom SPAD technology, the dedicated SPAD process flow must be modified, and particular care must be taken to safeguard the structure and the performance of the detector. To this aim, only a few process steps necessary for the fabrication of a basic n-MOS transistor need to be added. In Ref. [111], a simple current pick-up circuit, including an n-MOS inverter and a polysilicon load resistance, were monolithically integrated near to the photodiode (see Fig. 4.18). The pixel was completed by an external standard-CMOS active quenching circuit, which provides stable timing performance up to high count rates (>10^6 s^{-1}).

In summary, research has demonstrated that SPAD arrays can attain performance comparable to that of state-of-the-art single-pixel detectors implemented in the same technology. CMOS-based SPAD arrays offer a significant functionality along with single-photon detection capability, whereas SPAD arrays manufactured in custom technologies represent a valuable tool for parallel high-throughput measurements of very low light level signals. There is a huge potential for improvement of SPAD arrays in terms of pixel number, detection efficiency and time resolution. As discussed in

Section 4.10, the progress will be mainly driven by user demands for detector performance in new and more diverse applications.

Section 4.8 – SPADs for the infrared spectral range: single detectors and arrays

4.8.0 Infrared SPADs

While silicon provides excellent performance for the detection of photons at visible and near-infrared wavelengths, the rolloff in its optical absorption beyond ≈1 μm makes this material unsuitable for longer wavelength detection. To serve applications in the wavelength range of 0.95 μm to 1.65 μm—particularly at the immensely important fiber-based telecommunications windows at 1.3 μm and 1.55 μm—photodetectors based on the InGaAsP compound semiconductor material system have been widely adopted. Avalanche diodes employing $In_{0.53}Ga_{0.47}As$ absorbers (which are lattice matched to InP substrates) and InP multipliers have been technologically significant since they were first introduced in the late 1970s [112], and the enormous growth in fiber optic communications during the late 1990s instigated dramatic improvements in the performance of InGaAs/InP APDs for use in fiber optic receivers.

However, progress related to the telecom receiver-based "linear mode" operation of these devices, for which output photocurrent is proportional to input optical power, had essentially no impact on the performance and availability of SPADs based on a similar device design and material platform. In fact, up until the mid-2000s, there had been no InGaAs/InP devices designed specifically for photon-counting operation in Geiger mode, and system developers who had sought devices with good Geiger-mode performance at fiber-optic telecommunications wavelengths had been relegated to sampling the various commercially available telecom APDs to characterize their photon-counting performance [112][111][113][114][115][116][117].

Within the past decade, this situation has improved significantly. Researchers have 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 [117][116][118][119][120][121][122], primarily because the most critical performance attributes for linear-mode APDs (such as excess noise and gain-bandwidth product) are irrelevant for SPADs. This realization, and subsequent efforts expended on advancing the performance of InGaAs/InP SPADs, has led to significant progress for many of their properties [123][123][122]. For instance, there has been notable improvement in the fundamental tradeoff between single-photon detection efficiency and dark count rate, and high precision timing resolution has been demonstrated for these detectors. There has also been impressive scaling of these detectors to large format arrays [124][125] for emerging applications requiring single-photon imaging at short-wave infrared (SWIR) wavelengths.

4.8.1 Basic InGaAs/InP SPAD design concepts

All InP-based avalanche diodes deployed today are based on the separate absorption and multiplication (SAM) regions structure [112]. **Error! Reference source not found.**9 presents a schematic representation of a widely-used InGaAs/InP device design platform [123]. This design entails a narrow bandgap $In_{0.53}Ga_{0.47}As$ layer (with bandgap $E_g \approx 0.75$ eV at 295 K), lattice-matched to InP, that provides efficient absorption of photons within a wavelength range between ≈0.9 μm and ≈1.6 μm. Adjacent to this absorption region is a wider bandgap InP region ($E_g \approx 1.35$ 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 [126]) 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

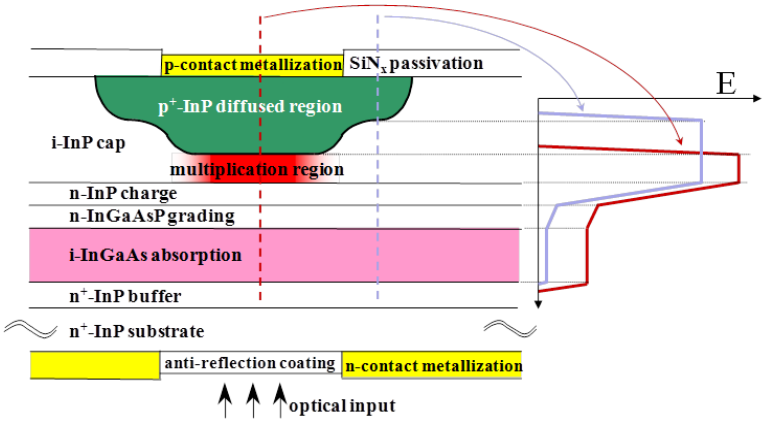


Fig. 4.19 Schematic of a typical InGaAs/InP SPAD device structure and associated internal electric-field profiles (at right) [121].

layers in the structure is important to reduce hole trapping effects that result from the valence band offset that arises in an abrupt InGaAs/InP heterojunction [127].

As in any avalanche diode, the design of the lateral structure of the InGaAs/InP SPAD should achieve uniform gain in the high-field region of the device and suppress field enhancement due to p-n junction curvature at the device periphery that can lead to edge breakdown effects. The structure presented in **Error! Reference source not found.** 4.19 illustrates one commonly used scheme in which the device periphery is shaped using two concentric diffusions of different diameters [128][120], but other approaches are possible and have been demonstrated in the APD literature.

4.8.2 DE and DCR modelling and performance

One of the most fundamental design considerations for InGaAs/InP SPADs is managing the tradeoff between the single-photon DE and the DCR. As in Si SPADs, optical coupling issues are generally set aside and the DE is taken as the product of three probabilities: DE = $\eta_{QE}P_cP_a$, where η_{QE} is the quantum efficiency for carrier creation by absorption of an incident photon in the InGaAs absorber; P_c is the probability that a photo-excited carrier is collected by injection into the InP multiplication region; and the avalanche probability P_a is the probability that a carrier injected into the multiplication region actually gives rise to a detectable avalanche. Although InGaAs/InP SPADs are often fiber pigtailed, the impact on the DE of optical coupling from the fiber to the SPAD active area is generally considered negligible in the context of the typical DE achieved with InGaAs/InP SPADs. In a well-designed device, the two dominant contributions to the DCR are thermal carrier generation in the narrow bandgap absorber and trap-assisted tunnelling in the multiplier. The relative importance of these two mechanisms is determined by operating conditions; thermal generation will dominate at high temperature and low bias, while tunnelling effects will dominate at low temperature and high bias.

DE and DCR calculations must include several dynamic processes, some of which are highly dependent on the local electric field intensity. Modelling of the avalanche probability P_a relies on a description of the avalanche process, and the adoption of appropriate expressions for the impact ionization coefficients in InP—particularly their temperature dependence [129] —is critical to the accuracy of the model. Dark-carrier creation can occur through field-dependent tunnelling processes as well as thermally driven Shockley-Read-Hall processes. The first comprehensive description of a DE vs. DCR model for InP-based SPADs was developed by Donnelly *et al.* [118], and this formalism has been employed in additional work to treat both InGaAs/InP SPADs for 1.5 μm photon counting [121] as well as InGaAsP/InP SPADs for use at 1.06 μm [130]. One salient

output of this model is that a wider multiplication region is highly beneficial for achieving a lower DCR at a given value of DE. Because a wider multiplication region reaches the avalanche-

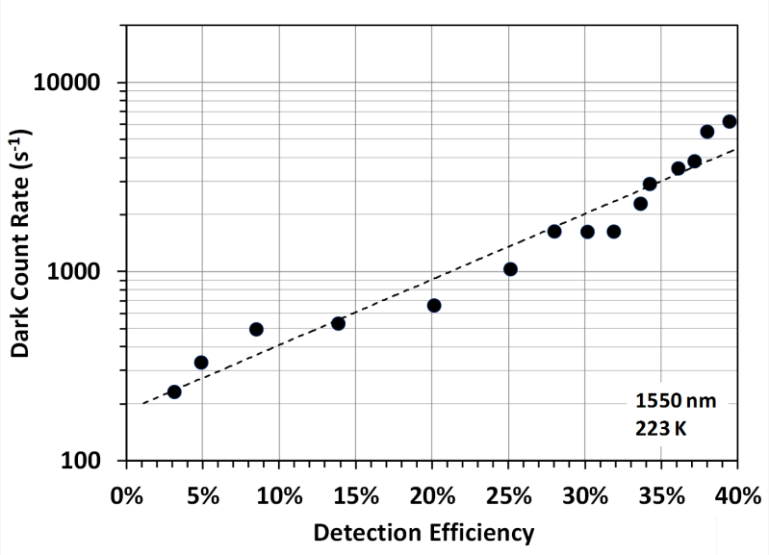


Fig. 4.20 Dark count rate versus detection efficiency performance of an InGaAs/InP SPAD operating at 223 K in response to 1550 nm photons. The dashed line indicates the general trend in performance.

breakdown condition at lower electric field intensity than a narrower one, the former structure allows for Geiger-mode operation with reduced tunnelling effects. (An alternative theoretical treatment of SPAD performance optimization with respect to multiplication-region width employs generalized breakdown probabilities [122] calculated using the recursive dead-space multiplication theory [131].) On the other hand, thermally generated dark counts originating in the InGaAs absorber are fairly independent of the width of the multiplication region, and are instead very sensitive to operating temperature. Therefore, reduction in operating temperature will improve DCR performance until temperatures are sufficiently low that tunnelling effects dominate.

With the insight of the modelling described above, as well as continuous improvements related to the fabrication of these devices, the fundamental trade-off between DCR and DE has advanced to a performance level that is sufficient to serve many applications of photon counting at wavelengths near 1.5 µm. As illustrated in **Error! Reference source not found.**0, devices with a size typical of fiber-coupled modules (e.g., 25 µm active area diameter) can provide DCR below 10^3 s^{-1} at 20 % DE, and DCR below 10^4 s^{-1} for DE values of at least 40 %, while operated with modest cooling provided by thermoelectric coolers (e.g., 223 K)

4.8.3 Timing jitter

A number of physical mechanisms within any SPAD structure can contribute to timing jitter, i.e., uncertainty in the correlation between photon arrival time at the detector and the time of avalanche detection. In InGaAs/InP SPADs, these mechanisms include differences in the transit times of photoexcited carriers resulting from differences in the location of photon absorption, carrier propagation delay caused by the temporary trapping of carriers at heterojunctions formed by dissimilar semiconductor layers, and variations in the avalanche build-up time induced by the stochastic nature of the impact ionization process. Avalanche build-up time variation also includes effects related to the randomness of the spreading of the avalanche from an initially localized filament to a saturated avalanche process that fills the entire high-field active area of the device [24]. Aside from these stochastic processes, there is also the important consideration of local excess-bias non-uniformities resulting from non-uniform breakdown voltage across a given device's active area. If the excess bias exhibits considerable variation as a function of position in the device, the associated distribution of mean times to reach threshold further broadens the timing distribution

and may increase the effective timing jitter significantly above that which would be found for a device with an ideally uniform excess bias.

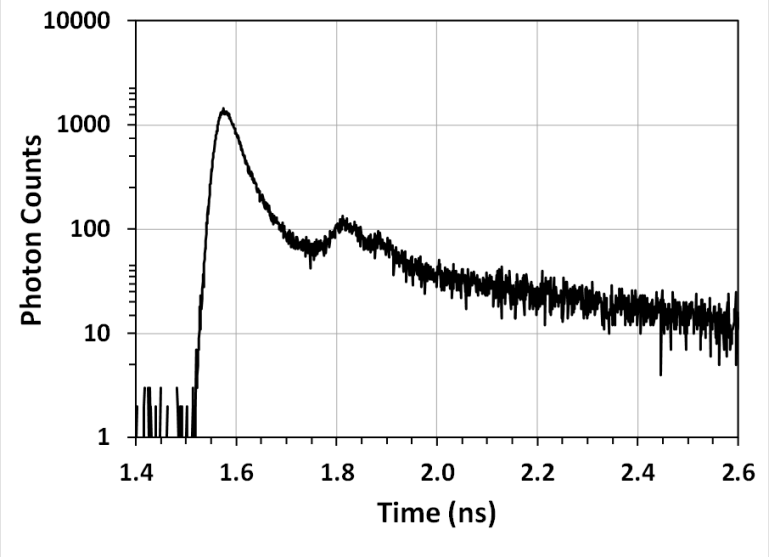


Fig. 4.21 Photon arrival time distribution for an InGaAs/InP SPAD with 25 μm diameter operated with 7 V excess bias. The FWHM timing jitter is 46 ps. The influence of the distribution tail increases the root-mean-square (RMS) timing variation to 59 ps. (The minor peak at 1.8 ns is an artefact of the measurement apparatus.)

For a well-designed InGaAs/InP SPAD, timing jitter of less than 50 ps FWHM has been demonstrated for an excess bias of ≈7 V [120]. This performance is comparable to the best results achieved with Si SPADs [25] and represents very good timing performance relative to other single-photon detector technologies. However, it should be noted that jitter is higher at lower excess bias (e.g., 100 ps at 3.5 V); obtaining very low jitter by using larger excess bias operation poses a trade-off with DCR and afterpulsing. Additionally, as discussed above in Section 4.4.4, low timing jitter is only possible with high performance circuits that can accurately detect the onset of the avalanche response at very low signal levels without spurious detections caused by circuit transient response characteristics. The critical importance of the circuit in determining the jitter performance—as well as other SPAD performance parameters— are discussed further in the next section of this chapter.

4.8.4 Afterpulsing

Among the possible strategies for the mitigation of afterpulsing in InGaAs/InP SPADs, the most direct are (i) decreasing the density of material defects that act as potential charge traps and (ii) reducing the number of charges that are trapped in the first place by reducing the amount of charge that flows during each avalanche event. A dramatic improvement in the quality of the InGaAsP material system was driven by the enormous economic opportunities of the telecommunications boom beginning in the late 1990s. However, following the collapse of this market in 2002 and its relative maturation and commoditization over the past decade, there has been little indication of further improvement in InP material quality as relates to the density of defects in the multiplication region, defects that are understood to cause afterpulsing. A key problem in this area is the paucity of knowledge concerning what types of material defects could be acting as charge traps, and what causes their formation.

Therefore, essentially all recent efforts to mitigate afterpulsing have invoked the strategy of reducing the number of charges that are trapped by restricting avalanche events to having less charge flow. There have been experimental measurements to confirm that the amount of trapped charge and the consequent afterpulsing scale linearly with the charge flow per avalanche

[132][133]. For situations in which gated mode operation with very short (sub-ns scale) gates is appropriate, avalanche charge flow can be reduced dramatically because the falling edge of the gate

Table 4.2 Comparison of state-of-the-art performance for Si and InGaAs/InP SPADs

	Si [1]	InGaAs/InP
Operating temperature	20 °C	– 70 °C
Active region diameter	50 μm	
Wavelength	550 nm	1550 nm
DCR and DE[2]	10^4 s^{-1} at 60 % $2x10^3$ s^{-1} at 40 % $0.5x10^3$ s^{-1} at 20 % -	- $6x10^3$ s^{-1}at 40 % 10^3 s^{-1} at 20 % $0.5x10^3$ s^{-1} at 10 %
Jitter (FWHM)	30 – 50 ps	50 – 100 ps
Minimum hold-off for 1% afterpulsing[3]	≈ 10 ns	≈ 100 ns

[1] Si SPAD performance corresponds to thin Si SPAD structures as in [16].
[2] Si DE values are cited for 550 nm, for which the highest Si DE is obtained.
[3] Assumes 20 % DE and free-running operation with fast active quenching of a few ns.

acts to rapidly quench the avalanche. This basic concept has been implemented at high (GHz) gating frequencies, and are discussed in detail in Section 4.9. More general non-periodic solutions have focused on sensing the avalanche with as low a detection threshold as possible and then rapidly quenching it to minimize the charge flow [134]. Finally, there has been recent work on self-quenching InGaAs/InP SPADs [135][136][137][138][139] in which the monolithic integration of passive quenching elements can lead to reduced charge flow if the quench elements can be integrated with strictly minimized parasitic capacitance. (As discussed above in the context of Si SPADs, parasitic capacitive elements must be discharged and recharged with each avalanche event, and minimizing these capacitances can reduce the overall charge flow per avalanche, c.f. Section 4.5.) However, afterpulsing continues to pose the primary challenge to free-running and high-rate photon counting using InP-based SPADs.

4.8.5 Comparison of InGaAs/InP SPADs and Si SPADs

A comparison of InGaAs/InP SPAD performance with those of state-of-the-art Si SPADs [16] is useful as an indication of how far InP-based SPADs might progress if InGaAsP materials engineering can be brought to the level of Si materials engineering. The longer-wavelength InGaAs/InP detectors will always be at a performance disadvantage relative to Si detectors given the necessarily smaller bandgap of the InGaAs absorbers. However, the primary impact of the absorber bandgap on SPAD performance is its role in determining the contribution to the DCR of carriers generated thermally by generation-recombination via mid-gap states. This suggests that we can remove the bandgap disparity by comparing Si and InGaAs/InP device performance at different temperatures that compensate for the difference in bandgaps.

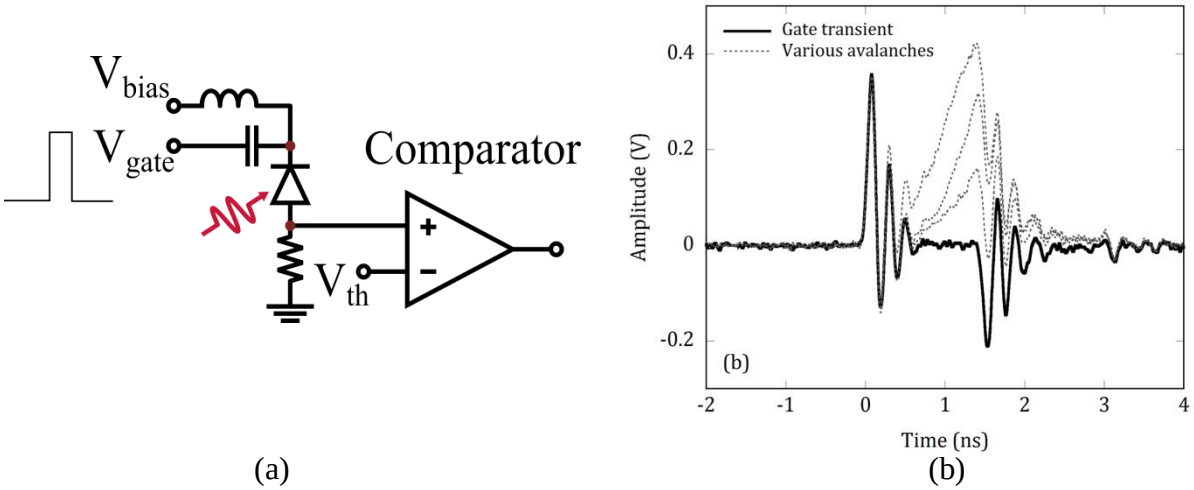
We first consider that dark-carrier thermal generation by Shockley-Read-Hall processes is proportional to $\exp(E_g/2k_BT)$ where E_g is the material bandgap, k_B is Boltzmann's constant, and T is temperature. Given that $E_g(\text{Si}) = 1.12$ eV at 20 °C, the exponent $E_g/2k_BT \approx 21.5$ for silicon. We then proceed to find the temperature for InGaAs at which $E_g(\text{InGaAs})/2k_BT$ gives the same value, which occurs at approximately –70 °C. By comparing Si SPAD and InGaAs/InP SPAD performance at these two respective temperatures, we remove the role of the material bandgap in thermal dark carrier generation to allow a direct comparison of underlying material properties. This comparison is summarized in Table 4.2, assuming devices with a 50 μm active region diameter.

Based on the rationale just described, Si SPADs exhibit superior material quality resulting in lower DCR, but only by a factor of 2 or 3 for a given value of DE. Thin Si SPADs have demonstrated somewhat lower timing jitter [25] than InGaAs/InP SPADs when operated with comparable electronic circuitry. For this parameter, the Si devices tend to operate over a range of timing jitter values that are about one-half the range exhibited by InP-based SPADs. Finally, a comparison of afterpulsing performance is complicated by the fact that it is highly circuit-dependent, so we rely on characterization in free-running operation with fast (i.e., a few ns) active quenching using the same backend electronics [140]. Si SPADs have the potential for an order of magnitude shorter hold-off times at 1 % afterpulsing levels, but while this suggests lower trap densities in Si multiplication regions, at least some of this afterpulsing performance advantage is related to the much higher temperature operation of Si SPADs allowed by their larger bandgap absorber.

Section 4.9 - Active gating techniques for InGaAs SPADs

4.9.0 Introduction

As discussed in the preceding sections of this chapter, the efficiency, noise, resolution, and maximum count rate of any SPAD detection system derive from the co-operative performance of the SPAD and the circuitry used to control it. This connection is particularly strong for actively gated detection systems, in which the SPAD is biased in the linear-multiplication regime and raised above breakdown only during a short detection gate. Active bias gates are a useful means to



(a) (b)

Fig. 4.22 (a) A prototypical active gating circuit. (b) The gate transients at the SPAD anode, along with some avalanches, when a 5 V, 1.5 ns, square gate with 100 ps edges is applied to the cathode of an InGaAs/InP SPAD.

improve the signal-to-noise (SNR) in applications with pulsed sources (e.g. ranging, fluorescence, communications), particularly (but not exclusively [141]) for devices with relatively high dark-count rates such as InGaAs or Ge SPADs [6][7], and can be simple to implement.

Over the past two decades, the demands of quantum information systems for low-noise high-efficiency single-photon detectors in the telecommunications windows [142][117][115] have motivated significant advances in active-gating techniques, resulting in dramatic improvements in detection efficiency and maximum count rate. The performance of these InGaAs detection systems can be so strongly influenced by the biasing scheme that it is common to identify them not by the type of SPAD, but by the type of gating and avalanche discrimination system used to control it. This section presents a comparative survey of these techniques and emphasizes the performance benefits various approaches can provide. While it is worthwhile to note the ongoing development of systems that combine gating with active and passive quenching circuits [115][18][143][144], this discussion is restricted to systems in which the avalanche current is terminated by the end of the gate itself, that is, quenching (as discussed above) is implemented by gate termination rather than some feedback mechanism [18].

Figure 4.22a shows one prototypical active-gating circuit. The bias on the cathode is the sum of the DC voltage, held at some value lower than the SPAD breakdown voltage, and the active gating signal AC-coupled to the SPAD. The avalanche current can be sensed as voltage across a resistive load, often chosen to be 50 Ω to keep the SPAD anode fast. As discussed in Section 4.8, InGaAs SPADs are generally devices that evolved from telecommunications applications [6], and as such tend to have low junction capacitance (c.f. Section 4.8), typically of the order of 0.1 pF when biased close to breakdown. This capacitance acts as a high-pass filter on the spectral components of the applied gate pulse. Provided that the anode supports a wide bandwidth, the signal at the output is similar to that shown in Fig. 4.22b when a 5 V, 1.5 ns square gate pulse with ≈100 ps edges is applied to the cathode. These are the so-called gate transients. Avalanches due to single-photon absorption occur between these transient signals, as shown, and the discrimination of the avalanche signal from these gate transients, particularly when the gate duration is short (≤ 1 ns), is the main subject of this section.

In the simple circuit shown in Fig. 4.22a, the SPAD presents an impedance discontinuity that may reflect spectral components of the gate signal back along any transmission line between the driving source and the SPAD. This effect can be used to increase the AC voltage experienced by the SPAD. On the other hand the transmission line may host multiple reflections that can complicate

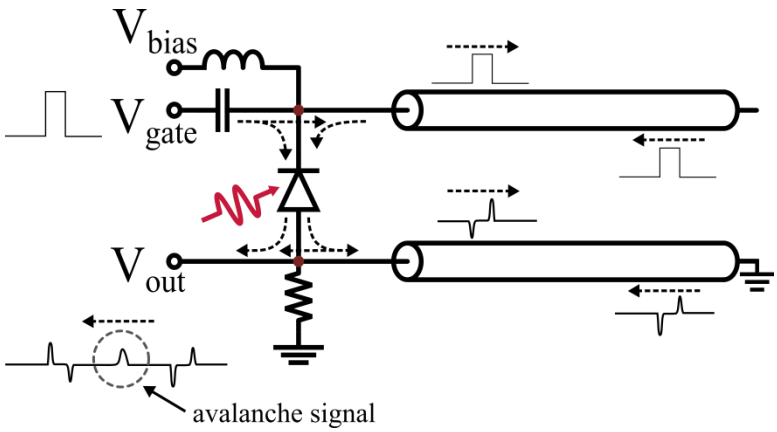


Fig. 4.23 Gate-transient cancellation based on inverting and non-inverting reflection from matched transmission lines. The avalanche signal at the output must still be discriminated from the preceding and following uncancelled transients, as illustrated.

the gating waveform. A simple way to avoid this process is to provide an AC termination, as in [145].

If nothing more than a threshold comparator is used to detect avalanches, the discrimination threshold must be set at a level higher than the rising gate transient, and it was recognized early on that this has significant consequences for the performance of the detection system. Specifically, larger avalanche signals correspond to larger amounts of charge, which populates more traps within the SPAD and increases the afterpulsing [146][114]. Better performance can be achieved with techniques that discriminate small avalanches from the gate transients, and a wide variety of techniques have been developed for this purpose, all representing different forms of filters. For classification we can categorize them as sampling schemes and cancellation schemes.

4.9.1 Sampling

Although not the first approach reported in the literature, sampling schemes are perhaps the most straightforward because they make use of the fact that the avalanche signal and the gate transients may be temporally distinguishable. Sampling can therefore be implemented with modifications only to the discriminator electronics, for example, with an AND gate that samples the voltage between the gate transients [117][147], or with an analog-to-digital converter (ADC) that records the voltage between the gate transients for further discrimination in digital format [148]. Systems of this type have been used with 800 ps gates at repetition rates as high as 14 MHz with low afterpulsing at a detection efficiency of about 14 %. Sampling becomes more challenging as the gate duration and avalanche amplitude become smaller, requiring higher bandwidth sampling systems. However, one alternative is to use the transient signal itself to sample the diode: if an avalanche occurs between the rising and falling edges of the gate, then the transient due to the falling edge of the gate becomes distorted, or even disappears, effectively reporting anything that may have occurred during the gate [149]. A comparator monitoring the falling-edge transient can then be used to identify when an avalanche occurred. This alleviates the need for high-speed sampling, and has been implemented with 1 ns gates at rates as high as 20 MHz [150].

4.9.2 Cancellation

There are a variety of approaches that all share the same basic idea: the SPAD biasing circuit is designed to generate matching replicas of the gate transients and subtract them to reveal an avalanche signal that may be obscured in one of the gate transients. One of the earliest techniques, developed at IBM in the late 1990s, is illustrated in Fig. 4.23 [151][152]. This circuit uses inverting and non-inverting reflections from the ends of coaxial cables to generate an opposing pair of gate transients that cancel in a passive network. Both signals pass through the SPAD, so the quality of

the cancellation is determined primarily by the degree to which the properties of the coaxial lines (delay, attenuation, dispersion, reflection) match, which can be quite good. An additional means still must be applied to ignore the remaining gate transients, as illustrated in Fig. 4.23, and this limits the minimum interval between gates to at least twice the round-trip time of the transmission lines. It is also worthwhile to point out that the transmission line, particularly on the gate-driver side, can host multiple reflections that can affect the minimum threshold at short gate intervals. Nonetheless, this scheme is robust and has been applied at gate rates well above 1 MHz. When properly designed and implemented, this approach strongly suppresses the gate transient and allows the discrimination of small avalanche signals, on the order of 200 fC, reducing afterpulsing [153]. Another configuration of basically the same approach uses balun transformers, rather than open and shorted transmission lines, to generate the opposing gate transients, and was shown to support gate intervals as low as 5 ns [154].

An alternative scheme implements cancellation by applying the gate pulse to two separate SPADs, as in [155]. The resulting pair of gate transients can then be subtracted with active or passive circuit elements, for example, a 180^{o} hybrid junction or a transformer. In this case the quality of the cancellation will be determined by how well the electrical response of the two diodes match, for which it is difficult, but not impossible, to control; Lu *et al.* used SPADs from adjacent locations on a wafer in a common-mode cancellation scheme with a sinusoidal gate signal (c.f. Section 4.9.4) and achieved excellent transient suppression with an 80 MHz gate frequency [156][157]. They demonstrated detection efficiency as high as 43 % at 1310 nm.

As configured in ref. [155], both SPADs can be used as detectors, and are distinguished by the orientation of the output avalanche signal (positive going or negative going), though simultaneous avalanches will result in distorted or completely missed detection events. Alternative configurations of this approach replace one of the SPADs with a ‘dummy’ element, such as a diode [158], or a capacitor [159][143], whose electrical response is similar to that of a SPAD. These approaches are generally simple to implement and are effective in supressing the gate transient to improve (reduce) the avalanche discrimination threshold. However, the quality of the match between the SPAD and the reference element critically determines the minimum discrimination threshold, and hence the afterpulse performance and usable gate rate. Systems of this type have been demonstrated at rates up to 25 MHz [160].

Rather than generating a transient for cancellation, it is also possible to choose the gate waveform to facilitate the discrimination of avalanche signals. Zhang *et al.* [161] use a Gaussian gate waveform, which produces an anti-symmetric gate transient dominated by the first derivative of the Gaussian. Avalanches within this anti-symmetric structure are revealed by summing the transient with a delayed portion of the driving pulse, thereby creating a symmetric transient structure from which avalanche signals can be more effectively detected.

One of the most advantageous features of all these cancellation schemes is that a single gate pulse generates the reference signal used to supress the gate transient. This allows for essentially arbitrary gate waveforms, and more importantly, for asynchronous gating (up to the minimum supported repetition rate of the scheme). Asynchronous operation is particularly useful in conditional measurements, in which the detector is activated by an external event coincident with a signal of interest, as in correlated or heralded photon experiments.

4.9.3 Introduction to high-speed periodic gating

The benefits demonstrated by reducing the avalanche charge motivated the investigation of methods that achieve even stronger gate-transient suppression and lower avalanche-discrimination thresholds. In 2006, Namekata *et al.* [162] demonstrated that by gating with a radio-frequency (RF) sine wave and using strong narrowband RF filters to supress the resulting gate transient, thresholds at unprecedented low levels (estimated to be 0.5 mV at the SPAD) could be used. The low threshold, in conjunction with short (sub-nanosecond) gate durations, efficiently detects avalanches with greatly reduced total charge; further developments have achieved average charge levels more

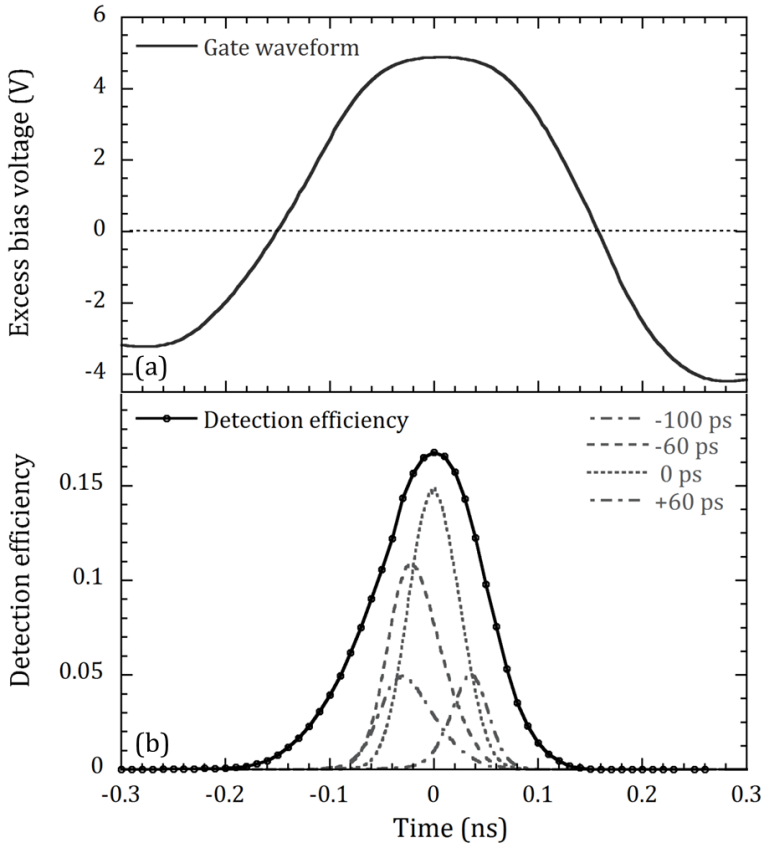


Fig. 4.24 (a) A bias gate from a high-speed (1.25 GHz) periodically gated InGaAs detection system. The gate exceeds the breakdown voltage for roughly 320 ps. (b) Detection efficiency versus time measured as a short (<30 ps) optical stimulus is stepped through the gate. Also shown are TCSPC histograms (30 second acquisition, each) when the optical stimulus arrives at four different times relative to the peak of the detection efficiency. The position of peak detection efficiency relative to the gate is unknown and arbitrarily aligned in this figure.

than an order of magnitude lower than in sampling or cancellation systems, resulting in low afterpulsing even with gate frequencies of the order of 1 GHz. This section provides a survey of a variety of techniques that implement high-speed periodic gating. First it is worthwhile to highlight some of their common characteristics.

The extremely strong transient suppression achieved in high-speed gating schemes is facilitated by both the periodicity and the high frequency of the gate. While this results in good sensitivity to minute avalanches, as is necessary for efficient high-speed operation, the periodicity also means that the device will be active every gate period, making asynchronous or triggered activation impossible unless further efforts are made. Therefore it is often the case that systems of this type employ a logical "hold-off," implemented after the output stage, that simply ignores outputs unlikely to be of interest. For example, such a hold-off is commonly applied immediately after a detection event to ignore the output from some number of subsequent gates that have high afterpulse probability (some typical hold-off times are included in Table 4.3).

Short bias-gate duration is also necessary for good performance in high-speed systems. For a fixed excess bias voltage, the total charge in an avalanche grows roughly exponentially with the gate duration [154], which means that the afterpulse performance can be significantly improved with even moderate reductions in the gate duration. However, the short bias gates, in some cases less than 200 ps, are on the order of the characteristic time scales for both the growth and the temporal jitter of the avalanche signal, and this has a significant impact on the temporal response of

Table 4.3

Table 4.3: Survey of a variety of high-speed (GHz) periodic gating techniques, along with the various refinements applied, as discussed in this section. The laser illumination rate used during the integrated afterpulse probability (AP) measurement, along with the type of counter and hold-off, are specified where available.

Technique	Temp.	Gate Frequency	Detection Efficiency (λ)	Dark Count Probability (per gate)	Integrated AP	Laser Rate during AP Meas.	AP Meas. Method, Hold-off
Self-differencing, tunable difference [170]	243 K	1 GHz 2 GHz	27.8 % 23.5 % (1550 nm)	$2.9 \cdot 10^{-5}$ $1.32 \cdot 10^{-5}$	8.8 % 4.84 %	15.6 MHz 31.25 MHz	TCSPC, --
Self-differencing, with sine gate [173]	243 K	921 MHz	9.3 % (1550 nm)	$4.3 \cdot 10^{-7}$	3.4 %	77 MHz	Gated counter, 10 ns hold-off
Sine-wave, notch filters [166]	223 K	1.244 GHz	11.6 % (1550 nm)	$5.8 \cdot 10^{-7}$	0.69 %	9.7 MHz	TIA, 23 ns hold-off
Sine-wave, notch & low-pass filter [164]	223 K	1 GHz 2 GHz	10.4 % 10.5 % (1550 nm)	$6.4 \cdot 10^{-7}$ $6.1 \cdot 10^{-7}$	1.6 % 3.4 %	10 MHz	TCSPC, 50 ns hold-off
Sine-wave, cancel & low-pass filter [168]	240 K	1 GHz	10.4 % (1550 nm)	$6.1 \cdot 10^{-6}$	3.0 %	10 MHz	-- 10 ns hold-off
Sine-wave, low-pass filter [165]	273 K	1.25 GHz	10 % (1550 nm)	$7 \cdot 10^{-7}$	1.6 %	-	QKD performance, 8 ns hold- off
Harmonic subtraction [175]	251 K	1.25 GHz	25% (1310 nm)	$2.4 \cdot 10^{-5}$	0.77 %	19.5 MHz	Gated counter, 10 ns hold-off

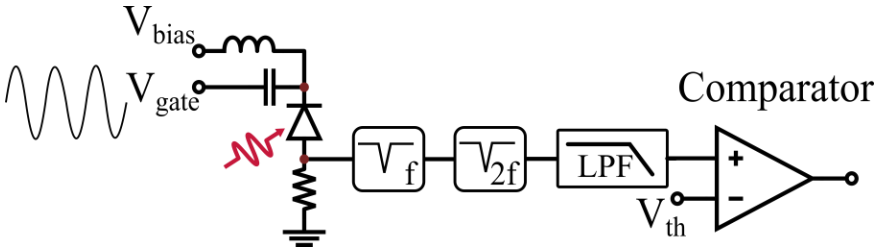


Fig. 4.25 A sine-wave gating circuit. Notch filters at the gate frequency f, and 2f, and a low-pass filter (LPF) are used to suppress the harmonics of the gate.

the detection system. One consequence is that the detection efficiency is strongly dependant on *when* in the gate a single-photon is absorbed. For example, Fig. 4.24a shows a bias gate from a high-speed gating system operating at 1.25 GHz, while Fig. 4.24b shows the detection efficiency of this gate as a function of the photon arrival time, as measured with a counter and an attenuated < 30 ps optical stimulus whose temporal position is stepped through the gate. The bias gate exceeds breakdown for roughly 320 ps, but the detection system has a region of sensitivity with a FWHM of 140 ps. Obviously, the system is most efficiently used with ultra-short optical signals aligned to the peak of the distribution. Figure 4.24b also shows histograms of detection events for four different temporal positions of the optical stimulus with respect to the gate, as measured with a traditional start-stop TCSPC system. Although the FWHM of the detection-event histograms are narrower than the full region of sensitivity, they do not accurately represent the arrival time of the photon. For example, the stimulus that arrived at -100 ps (100 ps before the peak sensitivity) produced a histogram whose peak is located at -40 ps. Moreover, all the histograms overlap significantly, making it nearly impossible to even distinguish a photon that arrived early in the gate from one that arrived later. In general, high-speed periodically gated systems do not provide timing resolution other than that defined by the gate's temporal region of sensitivity.

Given the strong relationship between gate duration and afterpulsing, it is tempting to envision ever shorter electrical bias gates to improve performance. However, with a shorter bias gate, ensuring that initiated avalanches grow to a detectable level before the end of the gate requires increasing the excess bias voltage. The challenge is therefore to produce large amplitude GHz-rate gates, and maintain good suppression of the resulting gate transient. To date, good performance has been demonstrated with gate amplitudes up to 20 V using commercially available GHz amplifiers, though improvements in both components and techniques continue to be made [163].

4.9.4 Sine-wave gating

There are a variety of configurations that expand on the idea of sine-wave gating as introduced above: a narrow-band RF signal as the gate, and passive filters to suppress the gate transient; one setup is shown in Fig. 4.25. Systems of this type can be implemented with commercially available connectorized RF components and can achieve low-noise high-speed single-photon detection at gate frequencies in the GHz range.

Although the spectrum of the gate has a single RF component, the voltage dependence of the SPAD junction capacitance generates higher-order harmonics of the gate signal. Therefore, to achieve a low discrimination threshold the filters used to supress the gate transient must address not only the gate frequency, but its harmonics as well, as illustrated in Fig. 4.25. The avalanche signal has a broad spectrum that extends to roughly the inverse of the gate duration, and that necessarily spans the gate frequency. In sine-wave gating there is a fundamental trade-off between supressing the gate signal and preserving the avalanche signal. Fortunately the gate harmonics lie at discrete frequencies, and a combination of narrow RF band-elimination (notch) filters and low-pass filters can be used to suppress them while efficiently passing the components of the avalanche signal that lie outside the filter bandwidths [164]. It has also been shown that solely low-pass filters with a

corner below the gate frequency can be used to reject all the harmonics of the gate [165]. This approach affords more tuning of the gate frequency than is allowed by narrowband notch filters, but obviously low-pass filters the avalanche signal as well.

Regardless of which configuration is used, using passive filters to supress the gate transient distorts the avalanche signal. Low-pass filters limit the steepness of the rising edge of the avalanche, and multipole filters with sharp profiles induce strong dispersion, both of which change the shape of the avalanche waveform. The effect of the filters often appears as additional jitter in the distribution of detection events in a TCSPC measurement, as reported in [165][166][167]. However, it should be noted that the region of sensitivity (c.f. Fig. 4.24b) is not affected by the filter-induced distortions, and for this reason the importance of this additional jitter depends on the application. Alternatively, Liang *et al.*[168] demonstrated a variant of sine-wave gating in which the first harmonic is suppressed by cancellation with a reference sine wave, rather than with dispersive notch filters, and only low-pass filters are used to suppress the higher-order harmonics. This approach better preserves the avalanche spectrum and reduces the observed jitter in a TCSPC measurement, and is related to the harmonic subtraction technique discussed below. A consequence of using passive filters may be more significant than jitter is that strong filtering may inhibit some avalanche signals from reaching the discrimination threshold before the end of the gate, either due to low-pass filtering or signal distortion. This is a particular concern for those avalanches that are initiated late in the gate, and may affect the detection efficiency of the system.

One of the main advantages of the sine-wave scheme is that the gate suppression with multiple filters can be strong (as high as -100 dB). These systems can therefore support extremely low discrimination thresholds, reducing avalanche charge and afterpulsing. The ultimate limit to the discrimination threshold is determined by thermally induced voltage fluctuations, or Johnson noise, at the output of the SPAD. The RMS voltage fluctuation across a resistor R, at temperature T, is given by $V_{\mathrm{Th}} = (4k_{\mathrm{B}}TRf)^{1/2}$, where f the measurement bandwidth. For a room temperature 50 Ω load in a 2 GHz bandwidth, the RMS thermal noise is 40 μV. A figure of merit for the actual usable threshold in the presence of such a Gaussian noise source is the 5σ level, or 0.2 mV, at which the probability that thermal noise would trigger an ideal comparator is in the 10^{-7} range, and thus below typical per-gate dark-count probabilities. Assuming a 1 V gate transient at the output of the SPAD, 74 dB of attenuation is required to suppress it below this thermal noise floor, an amount of attenuation that can be achieved fairly easily with RF filters. In practice, however, it is often the case that the noise floor is dominated by amplifiers in the output stage. Nonetheless, discrimination thresholds that correspond to total avalanche charges of the order of 10^4 electrons have been reported with sine-wave gating [164].

Sine-wave gating systems have an inherent inflexibility in the gate duration, as it is inextricably linked to the gate frequency and the excess bias voltage. For a given excess bias, the gate duration can be reduced by increasing the AC amplitude of the gate, making the sine wave more sharply peaked above the breakdown voltage. Following this approach Nambu *et al.* [166] used a 16 V gate signal at 1.244 GHz and report extremely low afterpulse probability. Sine-wave gating systems have been demonstrated at gate frequencies up to 2.23 GHz, and tend to operate most effectively at frequencies above 1 GHz given the link between the gate duration and frequency. Detection efficiencies up to 25 % have been reported [166]. Perhaps most distinguishing with respect to the schemes discussed earlier in this section, single-photon detection rates in the range of 10 MHz to 100 MHz can be achieved. This is the major advance enabled by high-speed periodically gated detection systems.

4.9.5 Self-differencing

Self-differencing [169] is high-speed periodic-gating scheme that, in contrast to sine-wave gating, supports arbitrary gate waveforms. A typical schematic is shown in Fig 4.26; in these systems the SPAD output is split evenly into two delay lines whose difference in propagation delay equals exactly one gate period, and the outputs of the delay lines are subtracted from each other. With a

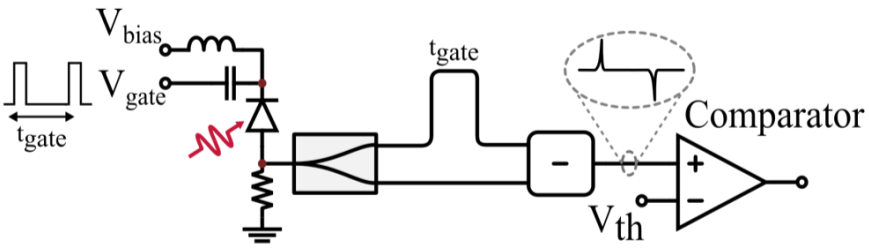


Fig. 4.26 A self-differencing circuit. The SPAD output is split between two delay lines whose difference in propagation delay is equal to the period of the gate signal, tgate. Taking the difference (-) of the outputs of the two delay lines subtracts the gate transients from successive gates, revealing avalanches as the anti-symmetric signal illustrated before the comparator.

strictly periodic gate waveform, the gate transient from one period eliminates the transient produced in the previous period, revealing any avalanche that occurred in one of the gates. Along with the gate transient, the avalanche signal is also distributed to each delay line, resulting in the characteristic anti-symmetric signal shown in Fig. 4.26. While this is of little consequence at low count rates, it does mean that avalanches in adjacent gates will interfere, which can complicate measurements of afterpulsing [154].

The ability to use arbitrary periodic waveforms offers the flexibility to optimize the detection system to a given application, often for reducing the gate duration and minimizing the avalanche charge. Also, although the delay difference must equal one gate period for good cancellation, moderate changes in the gate frequency can be accommodated with coaxial line stretchers in each delay path. In addition, self-differencing does not require strongly dispersive RF filters, and can accurately report the avalanche waveform without undue distortion. Overall, self-differencing is a highly versatile approach to high-speed periodic gating.

Self-differencing has many similarities to the cancellation schemes discussed earlier [152]. However, in self-differencing the opposing transients that cancel with each other travel through delay lines that differ by a length equivalent to one gate period. The quality of the match of the attenuation and dispersion of these two different paths determines the gate-transient suppression, and therefor has a major impact on the overall performance of the detection system. Matching the attenuation and dispersion between the different delay lines is therefore critical, and becomes increasingly challenging as the signal bandwidth is increased, as with short square-wave bias gates with sharp edges (wide bandwidth). A variety of refinements to the self-differencing circuit have been developed to improve performance.

Yuan *et al.* [170] showed that by providing fine adjustable tuning to both the splitting ratio and the delay difference they were able to greatly improve the cancellation, and hence the overall performance. They demonstrated detection efficiency up to 23.5 % at 1550 nm, operating at 2 GHz with moderate afterpulse probability. Notably, they were able to show that the average charge per avalanche was as low as 35 fC. Restelli *et al.* [171] showed that the frequency-dependent losses in the coaxial delay lines could be well matched over the detection bandwidth by designing the long and short delay lines with different types of coaxial cable, thus requiring adjustment only to the delay difference (or the gate frequency) to optimize the cancellation.

Chen *et al.* [172] noted that imperfect transient suppression in the self-differencing output stage left a systematic periodic background signal. Being periodic, they demonstrated that it could be further suppressed by a second self-differencing stage. As described above, the self-differencing output stage distributes the avalanche signal between two gate periods, and thus reduces the SNR ratio. Interestingly, the double-self-differencing scheme presented by Chen *et al.* does not further reduce the amount of avalanche signal in the detection gate because two avalanche signals produced

in the first stage combine to yield the same avalanche signal strength as with a single self-differencing stage. Unfortunately component losses cannot be ignored, and each power splitter or combiner imposes some loss to the avalanche signal. Nonetheless, Chen *et al.* were able to improve the transient suppression and better discriminate small avalanches from the transients by improving their discrimination threshold by 2 mV, and were able to demonstrate detection efficiency up to 30.5 % at 1550 nm with moderate afterpulsing.

One approach to improve transient cancellation in self-differencing systems is to relinquish some flexibility by using a narrowband sinusoidal gate in a self-differencing system [173]. As discussed above, the SPAD's voltage-dependant capacitance generates higher harmonics that can be removed with notch filters, in the same manner as in sine-wave schemes, and the remnant of the fundamental gate frequency can then be eliminated with the self-differencing circuit. In this case the broadband response of the delay lines is irrelevant, and care must only be given to match the attenuation of the two delay lines, which greatly simplifies the system and enhances the quality of the cancellation.

Although high-speed periodic gating schemes tend to operate more effectively at gate frequencies in the few-GHz range, the self-differencing scheme can be applied at lower gate frequencies by converting the output of the SPAD to an optical signal and using fiber-optic delay lines and balanced photodiode detection in the self-differencing output stage [174]. The low dispersion, low loss, and tunable power splitting available in fiber-optical components allow the gate repetition rates well into the MHz range with detection efficiency as high as 22 % with moderate afterpulsing. It is worthwhile to note that while the fundamental signal to noise ratio may be exacerbated by the electrical-to-optical-to-electrical conversion, along with attendant amplification stages, some benefit is regained by the essentially lossless signal splitting.

4.9.6 Harmonic subtraction

An alternative to sine-wave gating and self-differencing is shown in Fig. 4.27 [163][175]. Here, the gate waveform is synthesized from a discrete number of harmonics of the gate frequency, and the gate-transient suppression is achieved by destructive interference with reference signals at each harmonic that are generated directly at the RF source. The transient suppression in this scheme can approach what can be achieved with sine-wave gating, but without filtering or distorting the avalanche signal. This approach can support large-amplitude gate waveforms and excellent sensitivity to avalanche signals, both of which enhance the detection efficiency; Restelli *et al.* were able to reach detection efficiencies of 50 % at 1310 nm. The use of multiple harmonics in the gate signal allows the gate duration to be reduced well below that of a sine-wave gate of the same frequency. The number of harmonics needed for transient suppression is determined solely by the detection bandwidth.

Harmonic subtraction has significant merits in the quality of the transient suppression, in the preservation of the undistorted avalanche signal, and in the ability to reduce the gate duration. It is also worthwhile to point out that the gate-synthesis in this approach allows the use of narrowband low-noise RF amplifiers. However, these benefits come at the expense of significant circuit complexity. Moreover, the quality of the transient suppression is determined by the ability to maintain nearly perfect destructive interference of multiple RF sinusoids; for a practical detection system, active stabilization of this interference is necessary. Nonetheless, this is a promising approach and has achieved the highest detection efficiency of any high-speed periodically gated system and afterpulsing comparable to the lowest levels reported to date.

4.9.7 Summary

A comparative survey of some of the high-speed (GHz) periodic gating schemes presented in this section is given in Table 4.3. Given the singular importance of afterpulsing in such schemes, and the difficulty in characterizing afterpulsing is a system that is gated on every nanosecond, the illumination rate, and the afterpulse-probability measurement technique and hold-off that were used to characterize the system are specified (when available). The period of integration of the

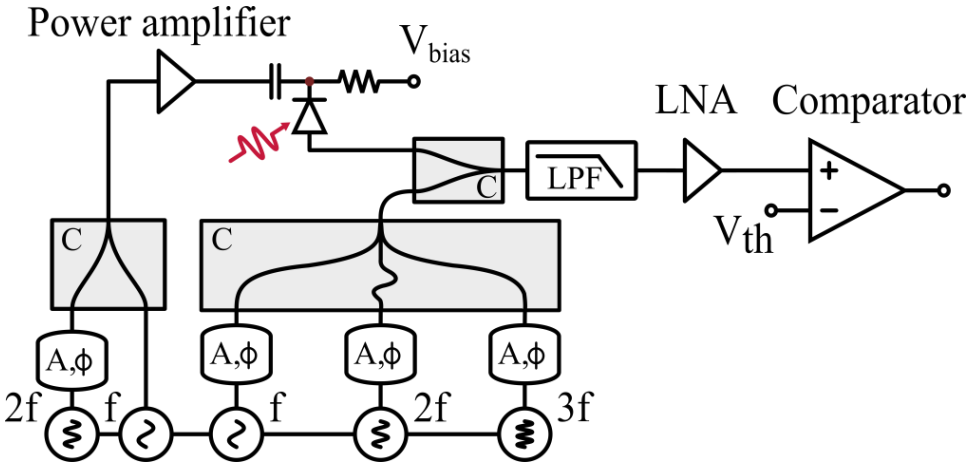


Fig. 4.27 A harmonic-subtraction setup. The gate waveform is synthesized from harmonics of the gate frequency f, and the resulting harmonics in the gate transient are eliminated by destructive interference with reference signals from the source, with fine amplitude (A) and phase (□) control, within the bandwidth of a low-pass filter (LPF). A low-loss combiner (C) and low-noise amplifier (LNA) efficiently preserves the avalanche signal.

afterpulse-probability measurement is the inverse of the laser illumination rate. It should be noted that afterpulse events that occur during the hold-off that immediately follows a detection event are not counted. By extension, afterpulses that are measured within one hold-off time of an illuminated gate indicate that a detection event did *not* occur in that illuminated gate, meaning that the counted afterpulse was necessarily due to some earlier detection event. There is roughly an order of magnitude variation in the afterpulse probabilities, which reflects the strong (exponential) relationship between the gate duration and total charge.

Active gating remains an active field of research, and continues to result in significant contributions to single-photon detection technology. The benefits gating techniques have provided, particularly in improved count rates and detection efficiencies, come with tradeoffs, the most obvious being the reduction in detection duty cycle, or the time for which a detector is active. Both the benefits and tradeoffs underscore the importance of the bias and control circuitry on the performance of the detection system as a whole. The wide variety of gating techniques is a testament to the ingenuity of experimenters in extracting ever-better performance from imperfect devices.

Section 4.10 - Future prospects for silicon SPADs

The future progress of silicon SPADs will be mainly driven by user demands for detector performance, which naturally set requirements for the device design and fabrication technology.

A first basic request arising in many applications is to improve the detection efficiency. Enhanced DE is strongly desired (particularly in the red and near-infrared spectral ranges) in several life-science applications based on in-vivo molecular imaging [176]. In order to achieve this, devices with thicker depletion layer must be designed and fabricated, which implies tailoring some steps in the fabrication process, or introducing new processing steps. In a dedicated fabrication technology this is quite natural. In standard CMOS technologies such flexibility is normally not offered; however, other applications may lead to develop new standard CMOS technologies with features suitable for producing SPADs with enhanced DE.

A second common request is a larger active area. In several techniques relying on confocal or near field microscopy (FCS, FLIM, combined FRET-FLIM), the illuminated spot size at the microscope image plane is small enough (a few tens of microns or less) to be easily covered by the

SPAD active area, provided it has sufficiently large diameter (~100 μm). Fiber pigtailing of the detector, often employed for making the optical system more flexible, also benefits from a wider detector area because greater coupling efficiency can be achieved, and fibers with larger core diameters can be more easily accommodated. An increase of the detector diameter, however, sets stringent requirements on the quality of the starting material and the fabrication process. A dedicated technology can exploit specific gettering steps performed as close as possible to the device active area. Such gettering is important not only because it is very difficult to obtain the required purity in the starting material, but also because contamination may be introduced later in the fabrication by unwanted marginal effects. Typical examples are faint residual contamination from furnaces previously employed in another fabrication, and side effects during the ion implantation. A further advantage of a dedicated technology is the capability of suitably shaping the electric field profile to minimize both band-to-band tunneling and field-assisted generation. Again, standard CMOS technologies do not offer this flexibility. Furthermore, the current trend of standard CMOS technologies toward low-thermal-budget processes and the lack of specific external gettering processes [71] that can be performed close to the SPAD raise some concerns about the evolution of these technologies towards fabrication of large-area devices.

A third quite specific requirement, arising from photon-timing applications, is to reduce the diffusion tail in the temporal response. For instance, a sub-nanosecond diffusion tail can benefit high-rate quantum key distribution (QKD) applications [53]. The diffusion tail can be reduced by keeping the neutral region very thin. Standard CMOS technologies do not allow any modification of the processing steps, whereas dedicated technologies are inherently flexible and fully customizable.

It is therefore likely that high-end applications requiring a combination of high DE, low DCR and low timing jitter will continue to rely on SPAD devices fabricated with dedicated technologies. Future developments in these technologies will likely focus on improving the DE in the red region of the spectrum. For example, a combination of red-enhanced [54] and resonant-cavity-enhanced [51] technologies might be used to develop frequency-selective SPAD devices with unprecedented DE at a desired wavelength.

There are no reasons to expect that the enhancement of sensitivity obtained with single SPAD detectors will not be extendable to array configurations used for multi-spot detection. The question to ascertain is how many SPADs such custom-technology arrays may eventually be able to contain without becoming cost-prohibitive, overly complex and, in the end, of little use to experimenters. A substantial increase in the pixel number can be achieved by resorting to more complex and sophisticated technologies, such as advanced multi-wafers and three-dimensional technologies that make possible the integration of custom SPAD arrays with high performance CMOS electronics for quenching/timing. Recently, Aull *et al.*[177] reported a fully parallel laser radar imager based on a 64 x 64 SPAD array coupled to with high-speed SOI CMOS circuits by using 3-D integration techniques.

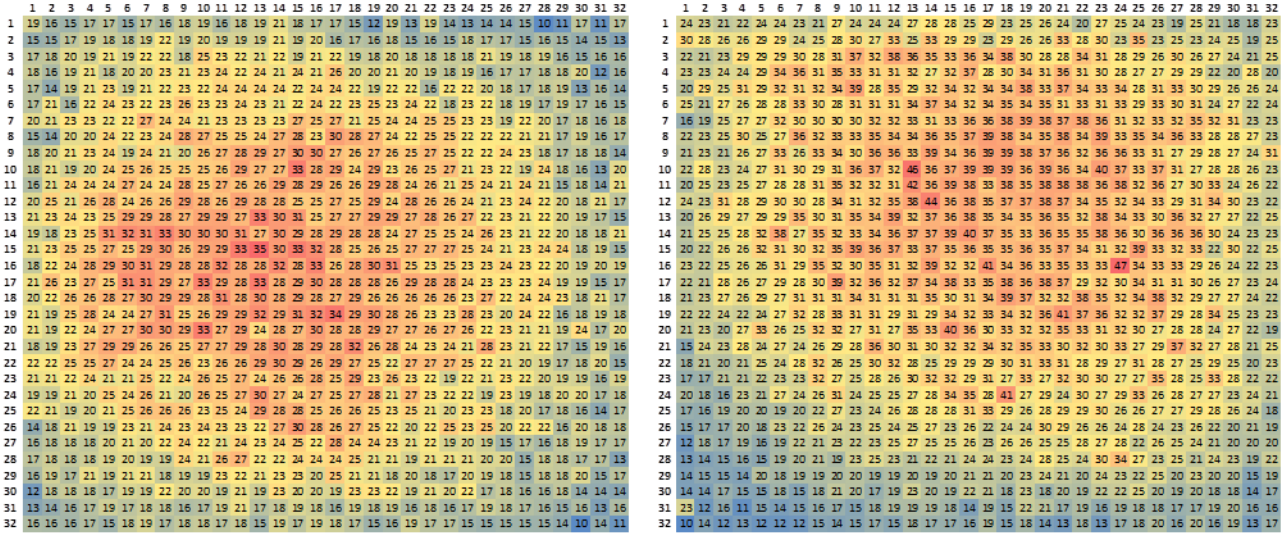
On the other hand, CMOS integration has enabled progressively smaller feature sizes, to the point where it is now possible to envision extremely large imaging systems based on SPADs. Standard CMOS technologies will definitely outperform custom SPAD technology in all those applications where a high number (> 1000) of pixels with small size, adequate fill factor (> 10 %), and integrated electronics, are mandatory requirements. Future research activity in this area will be aimed at the development of denser arrays with larger formats (>10^6 pixels) by exploiting sub-100 nm CMOS technologies [80]. Significant development efforts will be necessary for achieving a satisfactory tradeoff between detector performance (e.g. active area, fill-factor, timing jitter, noise) and system complexity (in-pixel photon counting and timing circuitry, external readout electronics).

Section 4.11 - Future prospects for InGaAs SPADs

As in the case of silicon SPADs, future improvements in the capabilities of InGaAs SPADs will be driven by the most pressing needs of applications that rely on these detectors. The tradeoff between DCR and DE described elsewhere in this chapter will continue to be a target for further progress, but this fundamental limitation poses significant challenges. DCR performance is intimately tied to materials properties, particularly with respect to bulk materials defects that lead to the thermal generation of dark carriers through Shockley-Read-Hall processes in the narrow-bandgap InGaAs absorption layer, as well as defects in the InP multiplication layer that lead to trap-assisted tunneling. Defects in the InP multiplier are also responsible for carrier trapping and detrapping that gives rise to afterpulsing effects. Dramatic improvements in epitaxial growth quality for the InGaAsP quaternary system were realized 10 to 15 years ago with the explosive growth in fiber-optic telecommunications applications that employed diode lasers and photodetectors based on this material system. Further progress on this front is likely to proceed much more modestly, especially in the absence of a similarly large new commercial market for devices employing these devices. Moreover, relative to the silicon material system, the InGaAsP material system serves vastly smaller markets and has considerably less technological maturity. Consequently, much less is known about the nature of InGaAsP materials limitations, and there is no comprehensive roadmap for materials improvement as there is in the silicon industry. The use of different III-V semiconductor materials with potentially favourable properties for SPAD devices may present interesting opportunities, but to the extent that these new materials will be even less technologically mature, they are likely to suffer from worse material quality. In light of these challenges to fundamental materials improvements, there is likely to be more rapid progress related to novel design approaches and implementation strategies, especially with regard to the electronic circuitry used to control SPAD functionality.

Beyond the fundamental DCR vs. DE tradeoff, the greatest recent focus for improvement of InGaAsP SPAD performance has been the effective photon counting rate of these devices. This need has been driven by the desire for GHz-scale bit rates for single-photon communications applications, especially in the context of quantum communications and quantum information processing. A similar requirement for much higher rate counting has also emerged in the context of applications of single-photon imaging such as 3-D laser radar and low-light level imaging. While the inherent carrier dynamics of these devices can readily support response times well below 1 ns, afterpulsing effects pose a much more difficult challenge to high-rate counting. Because the elimination of defects that give rise to afterpulsing does not appear achievable as a near-term strategy, the reduction of afterpulsing effects is an example of the derivation of more viable improvements from clever circuit-based solutions. In particular, the dominant effective strategy among practitioners in the field has been to reduce the current flow per avalanche event to limit the amount of trapped charge that can potentially give rise to afterpulses, as discussed in Section 4.9.

Despite inherent materials challenges, the gradual maturing of the InGaAsP SPAD device platform has enabled the realization of imaging arrays with an evolution to successively larger formats. **Error! Reference source not found.**4.28 illustrates pixel maps for the DCR and DE for a first-generation 32 x 32 array of InGaAs/InP SPADs on a 100 μm pitch designed for laser radar 3-D imaging at 1.5 μm. Pixel yield is 100 % with well-behaved, fairly narrow distributions of pixel-level performance parameters [178]. Similar arrays with InGaAsP quaternary absorbers optimized for detection at 1.06 μm have also been commercially realized in 32 x 32 formats, as well as in larger 128 x 32 arrays with a 50 μm pitch [125]. The largest InGaAsP SPAD array demonstrated to date has a format of 256 x 64 pixels [179], and the progression to significantly larger formats seems inevitable.



(a) (b)

Fig. 4.28 Performance maps of a 32 x 32 InGaAs/InP (1.55 μm) SPAD FPA operating with an excess bias of 3.25 V with modest cooling to 253 K. (a) The DCR, $<10^3$ s^{-1} for *all* pixels, is $< 50 \times 10^3$ s^{-1}. (b) Detection efficiency (in %) for all pixels, where the average pixel DE of 22 % includes all optical losses, including the microlens array for maintaining high fill factor.

References

[1] W. Becker, *Advanced time-correlated single-photon counting techniques*, vol. 81. Springer, 2005.

[2] Hamamatsu Photonics, MCP-PMT R3809U-50 Series Datasheet. Available online at: http://www.hamamatsu.com/resources/pdf/etd/R3809U-50_TPMH1067E09.pdf, 2011.

[3] A. Goetzberger, R. Scarlett, R. Haitz, and B. Mcdonald, "Avalanche Effects in Silicon P-N Junctions II. Structurally Perfect Junctions," *J. Appl. Phys.*, vol. 34, no. 6, pp. 1591–1600, 1963.

[4] R. H. Haitz, "Model for the Electrical Behavior of a Microplasma," *J. Appl. Phys.*, vol. 35, no. 5, pp. 1370–1376, May 1964.

[5] R. H. Haitz, "Mechanisms Contributing to the Noise Pulse Rate of Avalanche Diodes," *J. Appl. Phys.*, vol. 36, no. 10, pp. 3123–3131, Oct. 1965.

[6] B. Levine and C. Bethea, "Single Photon Detection at 1.3-Mu-M Using a Gated Avalanche Photodiode," *Appl. Phys. Lett.*, vol. 44, no. 5, pp. 553–555, 1984.

[7] B. Levine, C. Bethea, and J. Campbell, "Near Room-Temperature 1.3-Mu-M Single Photon-Counting with a Ingaas Avalanche Photodiode," *Electron. Lett.*, vol. 20, no. 14, pp. 596–598, 1984.

[8] R. J. McIntyre, "On the avalanche initiation probability of avalanche-diodes above the breakdown voltage," *IEEE Trans. Electron Devices*, vol. ED20, no. 7, pp. 637–641, 1973.

[9] W. G. Oldham, R. R. Samuelson, and P. Antognet, "Triggering Phenomena in Avalanche-Diodes," *IEEE Trans. Electron Devices*, vol. ED-19, no. 9, p. 1056, 1972.

[10] S. M. Sze, *Physics of Semiconductor Devices*. pp. 520-527, New York: Wiley, 1981.

[11] K. Graff, *Metal Impurities in Silicon-device Fabrication*, 2nd ed. Berlin: Springer-Verlag, 1995.

[12] G. Vincent, A. Chantre, and D. Bois, "Electric field effect on the thermal emission of traps in semiconductor junctions," *J. Appl. Phys.*, vol. 50, no. 8, pp. 5484–5487, Aug. 1979.

[13] P. A. Martin, B. G. Streetman, and K. Hess, "Electric field enhanced emission from non-Coulombic traps in semiconductors," *J. Appl. Phys.*, vol. 52, no. 12, pp. 7409–7415, Dec. 1981.

[14] G. A. M. Hurkx, H. C. Degraaff, W. J. Kloosterman, and M. P. G. Knuvers, "A New Analytical Diode Model Including Tunneling and Avalanche Breakdown," *IEEE Trans. Electron Devices*, vol. 39, no. 9, pp. 2090–2098, Sep. 1992.

[15] G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Knuvers, "A New Recombination Model for Device Simulation Including Tunneling," *IEEE Trans. Electron Devices*, vol. 39, no. 2, pp. 331–338, Feb. 1992.

[16] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, and S. Cova, "Progress in silicon single-photon avalanche diodes," *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 852–862, Aug. 2007.

[17] S. Cova, A. Lacaita, and G. Ripamonti, "Trapping Phenomena in Avalanche Photodiodes on Nanosecond Scale," *IEEE Electron Device Lett.*, vol. 12, no. 12, pp. 685–687, Dec. 1991.

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

[19] M. Ghioni, A. Giuduce, S. Cova, and F. Zappa, "High-rate quantum key distribution at short wavelength: performance analysis and evaluation of silicon single photon avalanche diodes," *J. Mod. Opt.*, vol. 50, no. 14, pp. 2251–2269, Sep. 2003.

[20] S. Cova, A. Longoni, and A. Andreoni, "Towards picosecond resolution with single-photon avalanche diodes," *Rev. Sci. Instrum.*, vol. 52, no. 3, pp. 408–412, Mar. 1981.

[21] G. Ripamonti and S. Cova, "Carrier Diffusion Effects in the Time-Response of a Fast Photodiode," *Solid-State Electron.*, vol. 28, no. 9, pp. 925–931, 1985.

[22] A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions," *Appl. Phys. Lett.*, vol. 57, no. 5, pp. 489–491, Jul. 1990.
[23] P. P. Webb and R. J. McIntyre, "Recent developments in silicon avalanche photodiodes," *RCS Eng.*, vol. 27, pp. 96–102, 1982.
[24] A. Spinelli and A. L. Lacaita, "Physics and numerical simulation of single photon avalanche diodes," *IEEE Trans. Electron Devices*, vol. 44, no. 11, pp. 1931–1943, Nov. 1997.
[25] A. Gulinatti, P. Maccagnani, I. Rech, M. Ghioni, and S. Cova, "35 ps time resolution at room temperature with large area single photon avalanche diodes," *Electron. Lett.*, vol. 41, no. 5, pp. 272–274, Mar. 2005.
[26] A. Lacaita, F. Zappa, S. Bigliardi, and M. Manfredi, "On the Bremsstrahlung Origin of Hot-Carrier-Induced Photons in Silicon Devices," *IEEE Trans. Electron Devices*, vol. 40, no. 3, pp. 577–582, Mar. 1993.
[27] I. Rech, A. Ingargiola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, and S. Cova, "A new approach to optical crosstalk modeling in single-photon avalanche diodes," *IEEE Photonics Technol. Lett.*, vol. 20, no. 5–8, pp. 330–332, Apr. 2008.
[28] I. Rech, A. Ingargiola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, and S. Cova, "Optical crosstalk in single photon avalanche diode arrays: a new complete model," *Opt. Express*, vol. 16, no. 12, pp. 8381–8394, Jun. 2008.
[29] E. Charbon, "Towards large scale CMOS single-photon detector arrays for lab-on-chip applications," *J. Phys. -Appl. Phys.*, vol. 41, no. 9, May 2008.
[30] P. Antognetti, S. Cova, and A. Longoni, "Study of the Operation and Performances of an Avalanche Diode as a Single Photon Detector," in *Proceedings of the 2nd Ispra Nuclear Electronics Symposium*, Stresa, May 20-23, 1975, pp. 453–456.
[31] S. Cova, "Active quenching circuit for avalanche photodiodes," US Patent No. 4,963,727; 16-Oct-1990 (priority date Oct. 20, 1988, Brevetto Italia 22367 A/88).
[32] A. Lacaita, S. Cova, C. Samori, and M. Ghioni, "Performance optimization of active quenching circuits for picosecond timing with single photon avalanche diodes," *Rev. Sci. Instrum.*, vol. 66, no. 8, pp. 4289–4295, Aug. 1995.
[33] M. Ghioni, S. Cova, F. Zappa, and C. Samori, "Compact active quenching circuit for fast photon counting with avalanche photodiodes," *Rev. Sci. Instrum.*, vol. 67, no. 10, pp. 3440–3448, Oct. 1996.
[34] F. Zappa, A. Lotito, A. C. Giudice, S. Cova, and M. Ghioni, "Monolithic active-quenching and active-reset circuit for single-photon avalanche detectors," *IEEE J. Solid-State Circuits*, vol. 38, no. 7, pp. 1298–1301, Jul. 2003.
[35] S. Tisa, F. Guerrieri, and F. Zappa, "Variable-load quenching circuit for single-photon avalanche diodes," *Opt. Express*, vol. 16, no. 3, pp. 2232–2244, Feb. 2008.
[36] A. Lacaita and M. Mastrapasqua, "Strong Dependence of Time Resolution on Detector Diameter in Single Photon Avalanche-Diodes," *Electron. Lett.*, vol. 26, no. 24, pp. 2053–2054, Nov. 1990.
[37] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes," *Appl. Phys. Lett.*, vol. 62, no. 6, pp. 606–608, Feb. 1993.
[38] M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti, and M. Ghioni, "Photon-Timing Jitter Dependence on Injection Position in Single-Photon Avalanche Diodes," *IEEE J. Quantum Electron.*, vol. 47, no. 2, pp. 151–159, Feb. 2011.
[39] S. Cova, M. Ghioni, and F. Zappa, "Circuit for high precision detection of the time of arrival of photons falling on single photon avalanche diodes," US Patent No. 6,384,663 B2; 07-May-2002. (priority date March 9, 2000).

[40] A. Gallivanoni, I. Rech, D. Resnati, M. Ghioni, and S. Cova, "Monolithic active quenching and picosecond timing circuit suitable for large-area single-photon avalanche diodes," *Opt. Express*, vol. 14, no. 12, pp. 5021–5030, Jun. 2006.
[41] T. A. Louis, G. Ripamonti, and A. Lacaita, "Photoluminescence lifetime microscope spectrometer based on time-correlated single-photon counting with an avalanche diode detector," *Rev. Sci. Instrum.*, vol. 61, no. 1, pp. 11–22, Jan. 1990.
[42] M. Ghioni, S. Cova, A. Lacaita, and G. Ripamonti, "New Silicon Epitaxial Avalanche-Diode for Single-Photon Timing at Room-Temperature," *Electron. Lett.*, vol. 24, no. 24, pp. 1476–1477, Nov. 1988.
[43] A. Lacaita, M. Ghioni, and S. Cova, "Double Epitaxy Improves Single-Photon Avalanche-Diode Performance," *Electron. Lett.*, vol. 25, no. 13, pp. 841–843, Jun. 1989.
[44] W. J. Kindt and H. W. van Zeijl, "Modeling and fabrication of Geiger mode avalanche photodiodes," *IEEE Trans. Nucl. Sci.*, vol. 45, no. 3, pp. 715–719, Jun. 1998.
[45] J. C. Jackson, A. P. Morrison, D. Phelan, and A. Mathewson, "A novel silicon geiger-mode avalanche photodiode," in *Digest of the International Electron Devices Meeting*, New York, 2002.
[46] B. F. Aull, A. H. Loomis, J. A. Gregory, and D. . Young, "Geiger-mode avalanche photodiode arrays integrated with CMOS timing circuits," in *Digest of the 56th Annual Device Research Conference*, 1998, pp. 58–59.
[47] E. Sciacca, A. C. Giudice, D. Sanfilippo, F. Zappa, S. Lombardo, R. Consentino, C. Di Franco, M. Ghioni, G. Fallica, G. Bonanno, S. Cova, and E. Rimini, "Silicon planar technology for single-photon optical detectors," *IEEE Trans. Electron Devices*, vol. 50, no. 4, pp. 918–925, Apr. 2003.
[48] A. Lacaita, M. Ghioni, and S. Cova, "Ultrafast single-photon detector with double epitaxial structure for minimum carrier diffusion effects," in *Le Journal de Physique Colloques*, Montpellier, France, Sept. 1988, 1988, vol. 49–C4, pp. 633–636.
[49] Micro-Photon-Devices, PDM Series datasheet, Available online at: http://www.micro-photon-devices.com/Docs/Datasheet/PDM.pdf, 2013.
[50] A. Spinelli, M. A. Ghioni, S. D. Cova, and L. M. Davis, "Avalanche detector with ultraclean response for time-resolved photon counting," *IEEE J. Quantum Electron.*, vol. 34, no. 5, pp. 817–821, May 1998.
[51] M. Ghioni, G. Armellini, P. Maccagnani, I. Rech, M. K. Emsley, and M. S. Unlu, "Resonant-cavity-enhanced single-photon avalanche diodes on reflecting silicon substrates," *IEEE Photonics Technol. Lett.*, vol. 20, no. 5–8, pp. 413–415, Apr. 2008.
[52] N. J. Krichel, A. McCarthy, I. Rech, M. Ghioni, A. Gulinatti, and G. S. Buller, "Cumulative data acquisition in comparative photon-counting three-dimensional imaging," *J. Mod. Opt.*, vol. 58, no. 3–4, pp. 244–256, 2011.
[53] P. J. Clarke, R. J. Collins, P. A. Hiskett, M.-J. García-Martínez, N. J. Krichel, A. McCarthy, M. G. Tanner, J. A. O'Connor, C. M. Natarajan, S. Miki, M. Sasaki, Z. Wang, M. Fujiwara, I. Rech, M. Ghioni, A. Gulinatti, R. H. Hadfield, P. D. Townsend, and G. S. Buller, "Analysis of detector performance in a gigahertz clock rate quantum key distribution system," *New J. Phys.*, vol. 13, no. 7, p. 075008, Jul. 2011.
[54] A. Gulinatti, I. Rech, F. Panzeri, C. Cammi, P. Maccagnani, M. Ghioni, and S. Cova, "New silicon SPAD technology for enhanced red-sensitivity, high-resolution timing and system integration," *J. Mod. Opt.*, vol. 59, no. 17, pp. 1489–1499, 2012.
[55] H. Dautet, P. Deschamps, B. Dion, A. D. MacGregor, D. MacSween, R. J. McIntyre, C. Trottier, and P. P. Webb, "Photon-Counting Techniques with Silicon Avalanche Photodiodes," *Appl. Opt.*, vol. 32, no. 21, pp. 3894–3900, Jul. 1993.

[56] Excelitas Technologies Corp., Single Photon Counting Module SPCM-AQRH Series Datasheet. Available online at: http://www.excelitas.com/Downloads/DTS_SPCM-AQRH.pdf, 2013.
[57] R. J. McIntyre, "Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes - Theory," *IEEE Trans. Electron Devices*, vol. ED19, no. 6, p. 703–&, 1972.
[58] P. P. Webb, R. J. McIntyre, and J. Conradi, "Properties of Avalanche Photodiodes," *RCA Rev.*, vol. 35, no. 2, pp. 234–278, 1974.
[59] R. J. McIntyre, "Silicon avalanche photodiode with low multiplication noise," US Patent No. 4,972,242; 20-Nov-1990.
[60] R. J. McIntyre and P. P. Webb, "Low-noise, reach-through, avalanche photodiodes," US Patent No. 5,583,352; 10-Dec-1996.
[61] I. Rech, I. Labanca, M. Ghioni, and S. Cova, "Modified single photon counting modules for optimal timing performance," *Rev. Sci. Instrum.*, vol. 77, no. 3, pp. 033104–033104–5, Mar. 2006.
[62] Laser Components USA, Inc., SAP500-Series Datasheet, Available online at: http://www.lasercomponents.com/fileadmin/user_upload/home/Datasheets/lcd/sap-series.pdf, 2013.
[63] M. Stipcevic, H. Skenderovic, and D. Gracin, "Characterization of a novel avalanche photodiode for single photon detection in VIS-NIR range," *Opt. Express*, vol. 18, no. 16, pp. 17448–17459, Aug. 2010.
[64] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, G. Ribordy, and N. Gisin, "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology," *Rev. Sci. Instrum.*, vol. 74, no. 7, pp. 3263–3270, Jul. 2003.
[65] A. Rochas, M. Gosch, A. Serov, P. A. Besse, R. S. Popovic, T. Lasser, and R. Rigler, "First fully integrated 2-D array of single-photon detectors in standard CMOS technology," *IEEE Photonics Technol. Lett.*, vol. 15, no. 7, pp. 963–965, Jul. 2003.
[66] F. Zappa, S. Tisa, A. Gulinatti, A. Gallivanoni, and S. Cova, "Complete single-photon counting and timing module in a microchip," *Opt. Lett.*, vol. 30, no. 11, pp. 1327–1329, Jun. 2005.
[67] L. Pancheri and D. Stoppa, "Low-noise CMOS single-photon avalanche diodes with 32 ns dead time," in *ESSDERC 2007: Proceedings of the 37th European Solid-State Device Research Conference*, D. SchmittLandsiedel and R. Thewes, Eds. New York: IEEE, 2007, pp. 362–365.
[68] A. Rochas, A. Pauchard, L. Monat, A. Matteo, P. Trinkler, R. Thew, and R. Ribordy, "Ultra-compact CMOS single photon detector," in *Advanced Photon Counting Techniques*, vol. 6372, W. Becker, Ed. Bellingham: SPIE-Int. Soc. Optical Engineering, 2006, pp. U169–U176.
[69] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in 0.35 mu m CMOS and based on an event-driven readout for TCSPC experiments," in *Advanced Photon Counting Techniques*, vol. 6372, W. Becker, Ed. Bellingham: SPIE-Int. Soc. Optical Engineering, 2006, pp. U216–U227.
[70] D. Mosconi, D. Stoppa, L. Pancheri, L. Gonzo, and A. Simoni, "CMOS single-photon avalanche diode array for time-resolved fluorescence detection," in *ESSCIRC 2006: Proceedings of the 32nd European Solid-State Circuits Conference*, Montreaux, France, September 2006, 2006, pp. 564–567.
[71] M. J. Binns, S. Bertolini, R. Wise, D. J. Myers, and T. A. McKenna, "Effective intrinsic gettering for 200mm and 300mm P/P- wafers in a low thermal budget 0.13 mu m advanced CMOS logic process," in *Semiconductor Silicon 2002*, H. R. Huff, L. Fabry, and S. Kishino, Eds. Pennington, NJ: Electrochemical Society Inc, 2002.

[72] N. Faramarzpour, M. J. Deen, S. Shirani, and Q. Fang, “Fully integrated single photon avalanche diode detector in standard CMOS 0.18-mu m technology,” *IEEE Trans. Electron Devices*, vol. 55, no. 3, pp. 760–767, Mar. 2008.
[73] M. A. Marwick and A. G. Andreou, “Single photon avalanche photodetector with integrated quenching fabricated in TSMC 0.18 mu m 1.8 VCMOS process,” *Electron. Lett.*, vol. 44, no. 10, pp. 643–U42, May 2008.
[74] H. Finkelstein, M. J. Hsu, and S. C. Esener, “STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology,” *IEEE Electron Device Lett.*, vol. 27, no. 11, pp. 887–889, Nov. 2006.
[75] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, “A low-noise single-photon detector implemented in a 130 nm CMOS imaging process,” *Solid-State Electron.*, vol. 53, no. 7, pp. 803–808, Jul. 2009.
[76] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, “A single photon avalanche diode implemented in 130-nm CMOS technology,” *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 863–869, Aug. 2007.
[77] L. Pancheri and D. Stoppa, “Low-noise single Photon Avalanche Diodes in 0.15 um CMOS technology,” in *Solid-State Device Research Conference (ESSDERC), 2011 Proceedings of the European*, Helsinki, Finland, 2011, pp. 179–182.
[78] T. Hamamoto, “Sidewall Damage in a Silicon Substrate Caused by Trench Etching,” *Appl. Phys. Lett.*, vol. 58, no. 25, pp. 2942–2944, Jun. 1991.
[79] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, “Scaleable Single-Photon Avalanche Diode Structures in Nanometer CMOS Technology,” *IEEE Trans. Electron Devices*, vol. 58, no. 7, pp. 2028–2035, Jul. 2011.
[80] M. A. Karami, M. Gersbach, H.-J. Yoon, and E. Charbon, “A new single-photon avalanche diode in 90nm standard CMOS technology,” *Opt. Express*, vol. 18, no. 21, pp. 22158–22166, Oct. 2010.
[81] E. A. G. Webster, J. A. Richardson, L. A. Grant, D. Renshaw, and R. K. Henderson, “A Single-Photon Avalanche Diode in 90-nm CMOS Imaging Technology With 44% Photon Detection Efficiency at 690 nm,” *IEEE Electron Device Lett.*, vol. 33, no. 5, pp. 694–696, May 2012.
[82] C. Niclass, A. Rochas, P. A. Besse, and E. Charbon, “Design and characterization of a CMOS 3-D image sensor based on single photon avalanche diodes,” *IEEE J. Solid-State Circuits*, vol. 40, no. 9, pp. 1847–1854, Sep. 2005.
[83] C. Niclass, M. Sergio, and E. Charbon, “A CMOS 64x48 single photon avalanche diode array with event-driven readout,” in *ESSCIRC 2006: Proceedings of the 32nd European Solid-State Circuits Conference*, C. Enz, M. Declercq, and Y. Leblebici, Eds. New York: IEEE, 2006, pp. 556–559.
[84] C. Niclass, M. Sergio, and E. Charbon, “A single photon avalanche diode array fabricated in deep-submicron CMOS technology,” in *2006 Design Automation and Test in Europe, Vols 1-3, Proceedings*, New York: IEEE, 2006, pp. 79–84.
[85] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, “A 128 x 128 Single-Photon Image Sensor With Column-Level 10-Bit Time-to-Digital Converter Array,” *IEEE J. Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, Dec. 2008.
[86] M. Sergio, C. Niclass, and E. Charbon, “A 128 x 2 CMOS Single-Photon Streak Camera with Timing-Preserving Latchless Pipeline Readout,” in *Solid-State Circuits Conference, 2007. ISSCC 2007. Digest of Technical Papers. IEEE International*, 2007, pp. 120–121.
[87] L. Carrara, C. Niclass, N. Scheidegger, H. Shea, and E. Charbon, “A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications,” in *Solid-State Circuits Conference - Digest of Technical Papers, ISSCC 2009. IEEE International*, 2009, pp. 40–41.

[88] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, “Single-Photon Synchronous Detection,” *IEEE J. Solid-State Circuits*, vol. 44, no. 7, pp. 1977–1989, Jul. 2009.
[89] F. Guerrieri, S. Tisa, A. Tosi, and F. Zappa, “Two-Dimensional SPAD Imaging Camera for Photon Counting,” *IEEE Photonics J.*, vol. 2, no. 5, pp. 759–774, Oct. 2010.
[90] F. Guerrieri, L. Maccone, F. N. C. Wong, J. H. Shapiro, S. Tisa, and F. Zappa, “Sub-Rayleigh Imaging via N-Photon Detection,” *Phys. Rev. Lett.*, vol. 105, no. 16, p. 163602, Oct. 2010.
[91] B. Markovic, S. Tisa, A. Tosi, and F. Zappa, “Monolithic Single-Photon detectors and Time-to-Digital Converters for picoseconds Time-of-Flight ranging,” in *Sensors, Cameras, and Systems for Industrial, Scientific, and Consumer Applications XII*, vol. 7875, R. Widenhorn and V. Nguyen, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2011, p. 78750P.
[92] European Community Sixth Framework Programme, IST-FET Open, project MEGAFRAME Million Frame Per Second Time-Correlated Single Photon Camera. Available Online: http://www.megaframe.eu.
[93] M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, R. Henderson, F. Borghetti, D. Stoppa, and E. Charbon, “A Parallel 32x32 Time-To-Digital Converter Array Fabricated in a 130 nm Imaging CMOS Technology,” in *2009 Proceedings of ESSCIRC*, New York: IEEE, 2009, pp. 197–200.
[94] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, “A 32x32 50ps Resolution 10 bit Time to Digital Converter Array in 130nm CMOS for Time Correlated Imaging,” presented at the IEEE Custom Integrated Circuit Conference, New York, 2009.
[95] D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R. K. Henderson, M. Gersbach, and E. Charbon, “A 32x32-Pixel Array with In-Pixel Photon Counting and Arrival Time Measurement in the Analog Domain,” in *2009 Proceedings of ESSCIRC*, New York: IEEE, 2009, pp. 205–208.
[96] M. Gersbach, Y. Maruyama, R. Trimananda, M. W. Fishburn, D. Stoppa, J. A. Richardson, R. Walker, R. Henderson, and E. Charbon, “A Time-Resolved, Low-Noise Single-Photon Image Sensor Fabricated in Deep-Submicron CMOS Technology,” *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1394–1407, Jun. 2012.
[97] J. A. Richardson, L. A. Grant, and R. K. Henderson, “Low Dark Count Single-Photon Avalanche Diode Structure Compatible With Standard Nanometer Scale CMOS Technology,” *IEEE Photonics Technol. Lett.*, vol. 21, no. 14, pp. 1020–1022, Jul. 2009.
[98] S. Donati, G. Martini, and M. Norgia, “Microconcentrators to recover fill-factor inimage photodetectors with pixel on-boardprocessing circuits,” *Opt. Express*, vol. 15, no. 26, pp. 18066–18075, Dec. 2007.
[99] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*, 3rd Edition. Springer, 2006.
[100] C. Veerappan, J. Richardson, R. Walker, D.-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, “A 160 x 128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter,” in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 312–314.
[101] R. Rigler and E. S. Elson, *Fluorescence Correlation Spectroscopy: Theory and Applications*. Berlin: Springer-Verlag, 2001.
[102] J. Bewersdorf, R. Pick, and S. W. Hell, “Multifocal multiphoton microscopy,” *Opt. Lett.*, vol. 23, no. 9, pp. 655–657, May 1998.
[103] G. MacBeath, “Protein microarrays and proteomics,” *Nat. Genet.*, vol. 32, pp. 526–532, 2002.
[104] T. Kodadek, “Protein microarrays: prospects and problems,” *Chem. Biol.*, vol. 8, no. 2, pp. 105–115, Feb. 2001.

[105] X. Michalet, R. A. Colyer, G. Scalia, A. Ingargiola, R. Lin, J. E. Millaud, S. Weiss, O. H. W. Siegmund, A. S. Tremsin, J. V. Vallerga, A. Cheng, M. Levi, D. Aharoni, K. Arisaka, F. Villa, F. Guerrieri, F. Panzeri, I. Rech, A. Gulinatti, F. Zappa, M. Ghioni, and S. Cova, "Development of new photon-counting detectors for single-molecule fluorescence microscopy," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 368, no. 1611, Feb. 2013.
[106] A. Restelli, I. Rech, P. Maccagnani, M. Ghioni, and S. Cova, "Monolithic silicon matrix detector with 50 μm photon counting pixels," *J. Mod. Opt.*, vol. 54, no. 2–3, pp. 213–223, 2007.
[107] S. Marangoni, I. Rech, M. Ghioni, P. Maccagnani, M. Chiari, M. Cretich, F. Damin, G. Di Carlo, and S. Cova, "A 6 × 8 photon-counting array detector system for fast and sensitive analysis of protein microarrays," *Sensors Actuators B Chem.*, vol. 149, no. 2, pp. 420–426, Aug. 2010.
[108] R. A. Colyer, G. Scalia, I. Rech, A. Gulinatti, M. Ghioni, S. Cova, S. Weiss, and X. Michalet, "High-throughput FCS using an LCOS spatial light modulator and an 8 × 1 SPAD array," *Biomed. Opt. Express*, vol. 1, no. 5, pp. 1408–1431, 2010.
[109] F. Zappa, A. Gulinatti, P. Maccagnani, S. Tisa, and S. Cova, "SPADA: single-photon avalanche diode arrays," *IEEE Photonics Technol. Lett.*, vol. 17, no. 3, pp. 657–659, 2005.
[110] F. Zappa, S. Tisa, S. Cova, P. Maccagnani, D. B. Calia, R. Saletti, R. Roncella, G. Bonanno, and M. Belluso, "Single-photon avalanche diode arrays for fast transients and adaptive optics," *IEEE Trans. Instrum. Meas.*, vol. 55, no. 1, pp. 365–374, 2006.
[111] C. Cammi, F. Panzeri, A. Gulinatti, I. Rech, and M. Ghioni, "Custom single-photon avalanche diode with integrated front-end for parallel photon timing applications," *Rev. Sci. Instrum.*, vol. 83, no. 3, p. 033104, 2012.
[112] K. Nishida, K. Taguchi, and Y. Matsumoto, "InGaAsP heterostructure avalanche photodiodes with high avalanche gain," *Appl. Phys. Lett.*, vol. 35, no. 3, pp. 251–253, Aug. 1979.
[113] A. Lacaita, F. Zappa, S. Cova, and P. Lovati, "Single-photon detection beyond 1 um: performance of commercially available InGaAs/lnP detectors," *Appl. Opt.*, vol. 35, no. 16, pp. 2986–2996, 1996.
[114] G. Ribordy, J. D. Gautier, H. Zbinden, and N. Gisin, "Performance of InGaAs/InP avalanche photodiodes as gated-mode photon counters," *Appl. Opt.*, vol. 37, no. 12, pp. 2272–2277, Apr. 1998.
[115] P. A. Hiskett, G. S. Buller, A. Y. Loudon, J. M. Smith, I. Gontijo, A. C. Walker, P. D. Townsend, and M. J. Robertson, "Performance and design of InGaAs/InP photodiodes for single-photon counting at 1.55 um," *Appl. Opt.*, vol. 39, no. 36, pp. 6818–6829, Dec. 2000.
[116] P. A. Hiskett, J. M. Smith, G. S. Buller, and P. D. Townsend, "Low-noise single-photon detection at wavelength 1.55 mu m," *Electron. Lett.*, vol. 37, no. 17, pp. 1081–1083, 2001.
[117] D. Stucki, G. Ribordy, A. Stefanov, H. Zbinden, J. G. Rarity, and T. Wall, "Photon counting for quantum key distribution with Peltier cooled InGaAs/InP APDs," *J. Mod. Opt.*, vol. 48, no. 13, pp. 1967–1981, Nov. 2001.
[118] J. P. Donnelly, E. K. Duerr, K. A. McIntosh, E. A. Dauler, D. C. Oakley, S. H. Groves, C. J. Vineis, L. J. Mahoney, K. M. Molvar, P. I. Hopman, K. E. Jensen, G. M. Smith, S. Verghese, and D. C. Shaver, "Design Considerations for 1.06 um InGaAsP-InP Geiger-Mode Avalanche Photodiodes," *IEEE J. Quantum Electron.*, vol. 42, no. 8, pp. 797–809, 2006.
[119] S. Pellegrini, R. E. Warburton, L. J. J. Tan, J. S. Ng, A. B. Krysa, K. Groom, J. P. R. David, S. Cova, M. J. Robertson, and G. S. Buller, "Design and performance of an InGaAs-InP single-photon avalanche diode detector," *IEEE J. Quantum Electron.*, vol. 42, no. 4, pp. 397–403, 2006.
[120] M. A. Itzler, R. Ben-Michael, C.-F. Hsu, K. Slomkowski, A. Tosi, S. Cova, F. Zappa, and R. Ispasoiu, "Single photon avalanche diodes (SPADs) for 1.5 μm photon counting applications," *J. Mod. Opt.*, vol. 54, no. 2–3, pp. 283–304, 2007.

[121] X. Jiang, M. A. Itzler, R. Ben-Michael, and K. Slomkowski, "InGaAsP-InP Avalanche Photodiodes for Single Photon Detection," *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 895–905, 2007.
[122] D. A. Ramirez, M. M. Hayat, and M. A. Itzler, "Dependence of the Performance of Single Photon Avalanche Diodes on the Multiplication Region Width," *IEEE J. Quantum Electron.*, vol. 44, no. 12, pp. 1188–1195, 2008.
[123] M. A. Itzler, X. Jiang, M. Entwistle, K. Slomkowski, A. Tosi, F. Acerbi, F. Zappa, and S. Cova, "Advances in InGaAsP-based avalanche diode single photon detectors," *J. Mod. Opt.*, vol. 58, no. 3–4, pp. 174–200, 2011.
[124] S. Verghese, J. P. Donnelly, E. K. Duerr, K. A. McIntosh, D. C. Chapman, C. J. Vineis, G. M. Smith, J. E. Funk, K. E. Jensen, P. I. Hopman, D. C. Shaver, B. F. Aull, J. C. Aversa, J. P. Frechette, J. B. Glettler, Z.-L. Liau, J. M. Mahan, L. J. Mahoney, K. M. Molvar, F. J. O'Donnell, D. C. Oakley, E. J. Ouellette, M. J. Renzi, and B. M. Tyrrell, "Arrays of InP-based Avalanche Photodiodes for Photon Counting," *IEEE J. Sel. Top. Quantum Electron.*, vol. 13, no. 4, pp. 870–886, 2007.
[125] M. A. Itzler, M. Entwistle, M. Owens, K. Patel, X. Jiang, K. Slomkowski, S. Rangwala, P. F. Zalud, T. Senko, J. Tower, and J. Ferraro, "Comparison of 32 x 128 and 32 x 32 Geiger-mode APD FPAs for single photon 3D LADAR imaging," in *Advanced Photon Counting Techniques V*, vol. 8033, M. A. Itzler and J. C. Campbell, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2011, p. 80330G.
[126] J. C. Campbell, A. G. Dentai, W. S. Holden, and B. L. Kasper, "High-performance avalanche photodiode with separate absorption grading and multiplication regions," *Electron. Lett.*, vol. 19, no. 20, pp. 818–820, 1983.
[127] S. R. Forrest, O. K. Kim, and R. G. Smith, "Optical response time of In0.53Ga0.47As/InP avalanche photodiodes," *Appl. Phys. Lett.*, vol. 41, no. 1, pp. 95–98, Jul. 1982.
[128] Y. Liu, S. R. Forrest, J. Hladky, M. J. Lange, G. H. Olsen, and D. E. Ackley, "A planar InP/InGaAs avalanche photodiode with floating guard ring and double diffused junction," *J. Light. Technol.*, vol. 10, no. 2, pp. 182–193, 1992.
[129] F. Zappa, P. Lovati, and A. Lacaita, "Temperature dependence of electron and hole ionization coefficients in InP," in *Eighth International Conference on Indium Phosphide and Related Materials, 1996. IPRM '96*, 1996, pp. 628–631.
[130] X. Jiang, M. A. Itzler, R. Ben-Michael, K. Slomkowski, M. A. Krainak, S. Wu, and X. Sun, "Afterpulsing Effects in Free-Running InGaAsP Single-Photon Avalanche Diodes," *IEEE J. Quantum Electron.*, vol. 44, no. 1, pp. 3–11, 2008.
[131] D. A. Ramirez, M. M. Hayat, G. Karve, J. C. Campbell, S. N. Torres, B. E. A. Saleh, and M. C. Teich, "Detection efficiencies and generalized breakdown probabilities for nanosecond-gated near infrared single-photon avalanche photodiodes," *IEEE J. Quantum Electron.*, vol. 42, no. 2, pp. 137–145, 2006.
[132] K. E. Jensen, P. I. Hopman, E. K. Duerr, E. A. Dauler, J. P. Donnelly, S. H. Groves, L. J. Mahoney, K. A. McIntosh, K. M. Molvar, A. Napoleone, D. C. Oakley, S. Verghese, C. J. Vineis, and R. D. Younger, "Afterpulsing in Geiger-mode avalanche photodiodes for 1.06 μm wavelength," *Appl. Phys. Lett.*, vol. 88, no. 13, pp. 133503–133503–3, Mar. 2006.
[133] M. Liu, C. Hu, J. C. Campbell, Z. Pan, and M. M. Tashima, "Reduce Afterpulsing of Single Photon Avalanche Diodes Using Passive Quenching With Active Reset," *IEEE J. Quantum Electron.*, vol. 44, no. 5, pp. 430–434, 2008.
[134] F. Zappa, A. Tosi, and S. Cova, "InGaAs SPAD and electronics for low time jitter and low noise," in *Photon Counting Applications, Quantum Optics, and Quantum Cryptography*, vol. 6583, I. Prochazka and A. L. Migdall, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2007, p. 65830E.

[135] K. Zhao, A. Zhang, Y. Lo, and W. Farr, "InGaAs single photon avalanche detector with ultralow excess noise," *Appl. Phys. Lett.*, vol. 91, no. 8, pp. 081107–081107–3, Aug. 2007.
[136] K. Zhao, S. You, J. Cheng, and Y. Lo, "Self-quenching and self-recovering InGaAs/InAlAs single photon avalanche detector," *Appl. Phys. Lett.*, vol. 93, no. 15, pp. 153504–153504–3, Oct. 2008.
[137] M. A. Itzler, X. Jiang, B. Nyman, and K. Slomkowski, "InP-based negative feedback avalanche diodes," in *Quantum Sensing and Nanophotonic Devices VI*, vol. 7222, M. Razeghi and R. Sudharsanan, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2009, p. 72221K.
[138] X. Jiang, M. A. Itzler, B. Nyman, and K. Slomkowski, "Negative feedback avalanche diodes for near-infrared single-photon detection," in *Advanced Photon Counting Techniques III*, vol. 7320, M. A. Itzler and J. C. Campbell, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2009, p. 732011.
[139] M. A. Itzler, X. Jiang, B. M. Onat, and K. Slomkowski, "Progress in self-quenching InP-based single photon detectors," in *Quantum Sensing and Nanophotonic Devices VII*, vol. 7608, M. Razeghi and R. Sudharsanan, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2010, p. 760829.
[140] A. Tosi, A. D. Mora, F. Zappa, and S. Cova, "Single-photon avalanche diodes for the near-infrared range: detector and circuit issues," *J. Mod. Opt.*, vol. 56, no. 2–3, pp. 299–308, 2009.
[141] S. Cova, A. Longoni, and G. Ripamonti, "Active-Quenching and Gating Circuits for Single-Photon Avalanche-Diodes (SPADs)," *IEEE Trans. Nucl. Sci.*, vol. 29, no. 1, pp. 599–601, 1982.
[142] C. Marand and P. Townsend, "Quantum Key Distribution Over Distances as Long as 30 Km," *Opt. Lett.*, vol. 20, no. 16, pp. 1695–1697, Aug. 1995.
[143] A. Dalla Mora, A. Tosi, F. Zappa, S. Cova, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli, and R. Cubeddu, "Fast-Gated Single-Photon Avalanche Diode for Wide Dynamic Range Near Infrared Spectroscopy," *IEEE J. Sel. Top. Quantum Electron.*, vol. 16, no. 4, pp. 1023–1030, Aug. 2010.
[144] C. Hu, X. Zheng, J. C. Campbell, B. M. Onat, X. Jiang, and M. A. Itzler, "Characterization of an InGaAs/InP-based single-photon avalanche diode with gated-passive quenching with active reset circuit," *J. Mod. Opt.*, vol. 58, no. 3–4, pp. 201–209, 2011.
[145] A. Yoshizawa and H. Tsuchida, "A 1550 nm single-photon detector using a thermoelectrically cooled InGaAs avalanche photodiode," *Jpn. J. Appl. Phys. Part 1-Regul. Pap. Short Notes Rev. Pap.*, vol. 40, no. 1, pp. 200–201, Jan. 2001.
[146] F. Zappa, A. Lacaita, S. Cova, and P. Webb, "Nanosecond Single-Photon Timing with Ingaas/Inp Photodiodes," *Opt. Lett.*, vol. 19, no. 11, pp. 846–848, Jun. 1994.
[147] G. Ribordy, N. Gisin, O. Guinnard, D. Stucki, M. Wegmuller, and H. Zbinden, "Photon counting at telecom wavelengths with commercial InGaAs/InP avalanche photodiodes: current performance," *J. Mod. Opt.*, vol. 51, no. 9–10, pp. 1381–1398, Jul. 2004.
[148] P. L. Voss, K. G. Koprulu, S. K. Choi, S. Dugan, and P. Kumar, "14 MHz rate photon counting with room temperature InGaAs/InP avalanche photodiodes," *J. Mod. Opt.*, vol. 51, no. 9–10, pp. 1369–1379, Jul. 2004.
[149] A. Yoshizawa, R. Kaji, and H. Tsuchida, "Gated-mode single-photon detection at 1550 nm by discharge pulse counting," *Appl. Phys. Lett.*, vol. 84, no. 18, pp. 3606–3608, May 2004.
[150] A. Yoshizawa, S. Odate, and H. Tsuchida, "Discharge pulse counting for low-noise single-photon detection at 1550 nm using InGaAs avalanche photodiode cooled to 130 K," *Jpn. J. Appl. Phys. Part 1-Regul. Pap. Brief Commun. Rev. Pap.*, vol. 46, no. 1, pp. 220–222, Jan. 2007.

[151] D. S. Bethune, R. G. Devoe, C. Kurtsiefer, C. T. Rettner, and W. P. Risk, "System for gated detection of optical pulses containing a small number of photons using an avalanche photodiode," US6218657 B117-Apr-2001.
[152] D. S. Bethune and W. P. Risk, "An autocompensating fiber-optic quantum cryptography system based on polarization splitting of light," *IEEE J. Quantum Electron.*, vol. 36, no. 3, pp. 340–347, Mar. 2000.
[153] D. S. Bethune, W. P. Risk, and G. W. Pabst, "A high-performance integrated single-photon detector for telecom wavelengths," *J. Mod. Opt.*, vol. 51, no. 9–10, pp. 1359–1368, Jul. 2004.
[154] A. Restelli, J. C. Bienfang, and A. L. Migdall, "Time-domain measurements of afterpulsing in InGaAs/InP SPAD gated with sub-nanosecond pulses," *J. Mod. Opt.*, vol. 59, no. 17, pp. 1465–1471, 2012.
[155] A. Tomita and K. Nakamura, "Balanced, gated-mode photon detector for quantum-bit discrimination at 1550 nm," *Opt. Lett.*, vol. 27, no. 20, pp. 1827–1829, Oct. 2002.
[156] Z. Lu, W. Sun, J. C. Campbell, X. Jiang, and M. A. Itzler, "Corrections to 'Common-Mode Cancellation in Sinusoidal Gating With Balanced InGaAs/InP Single Photon Avalanche Diodes'; [Dec 12 1505-1511]," *IEEE J. Quantum Electron.*, vol. 49, no. 1, pp. 59–59, 2013.
[157] J. C. Campbell, W. Sun, Z. Lu, M. A. Itzler, and X. Jiang, "Common-Mode Cancellation in Sinusoidal Gating With Balanced InGaAs/InP Single Photon Avalanche Diodes," *IEEE J. Quantum Electron.*, vol. 48, no. 12, pp. 1505–1511, Dec. 2012.
[158] G. Wu, C. Zhou, X. Chen, and H. Zeng, "High performance of gated-mode single-photon detector at 1.55 mu m," *Opt. Commun.*, vol. 265, no. 1, pp. 126–131, Sep. 2006.
[159] C. Y. Zhou, G. Wu, and H. P. Zeng, "Multigate single-photon detection and timing discrimination with an InGaAs/InP avalanche photodiode," *Appl. Opt.*, vol. 45, no. 8, pp. 1773–1776, Mar. 2006.
[160] Y. Liang, Y. Jian, X. Chen, G. Wu, E. Wu, and H. Zeng, "Room-Temperature Single-Photon Detector Based on InGaAs/InP Avalanche Photodiode With Multichannel Counting Ability," *IEEE Photonics Technol. Lett.*, vol. 23, no. 2, pp. 115–117, 2011.
[161] Y. Zhang, X. Zhang, and S. Wang, "Gaussian pulse gated InGaAs/InP avalanche photodiode for single photon detection," *Opt. Lett.*, vol. 38, no. 5, pp. 606–608, Mar. 2013.
[162] N. Namekata, S. Sasamori, and S. Inoue, "800 MHz single-photon detection at 1550-nm using an InGaAs/InP avalanche photodiode operated with a sine wave gating," *Opt. Express*, vol. 14, no. 21, pp. 10043–10049, Oct. 2006.
[163] A. Restelli, J. C. Bienfang, and A. L. Migdall, "Single-photon detection efficiency up to 50 % at 1310 nm with an InGaAs/InP avalanche diode gated at 1.25 GHz," *Appl. Phys. Lett.*, vol. 102, no. 14, pp. 141104–141104–4, Apr. 2013.
[164] N. Namekata, S. Adachi, and S. Inoue, "Ultra-Low-Noise Sinusoidally Gated Avalanche Photodiode for High-Speed Single-Photon Detection at Telecommunication Wavelengths," *IEEE Photonics Technol. Lett.*, vol. 22, no. 8, pp. 529–531, Apr. 2010.
[165] N. Walenta, T. Lunghi, O. Guinnard, R. Houlmann, H. Zbinden, and N. Gisin, "Sine gating detector with simple filtering for low-noise infra-red single photon detection at room temperature," *J. Appl. Phys.*, vol. 112, no. 6, Sep. 2012.
[166] Y. Nambu, S. Takahashi, K. Yoshino, A. Tanaka, M. Fujiwara, M. Sasaki, A. Tajima, S. Yorozu, and A. Tomita, "Efficient and low-noise single-photon avalanche photodiode for 1.244-GHz clocked quantum key distribution," *Opt. Express*, vol. 19, no. 21, pp. 20531–20541, Oct. 2011.
[167] N. Namekata, S. Adachi, and S. Inoue, "1.5 GHz single-photon detection at telecommunication wavelengths using sinusoidally gated InGaAs/InP avalanche photodiode," *Opt. Express*, vol. 17, no. 8, pp. 6275–6282, Apr. 2009.

[168] Y. Liang, E. Wu, X. Chen, M. Ren, Y. Jian, G. Wu, and H. Zeng, “Low-Timing-Jitter Single-Photon Detection Using 1-GHz Sinusoidally Gated InGaAs/InP Avalanche Photodiode,” *IEEE Photonics Technol. Lett.*, vol. 23, no. 13, pp. 887–889, Jul. 2011.
[169] Z. L. Yuan, B. E. Kardynal, A. W. Sharpe, and A. J. Shields, “High speed single photon detection in the near infrared,” *Appl. Phys. Lett.*, vol. 91, no. 4, Jul. 2007.
[170] Z. L. Yuan, A. W. Sharpe, J. F. Dynes, A. R. Dixon, and A. J. Shields, “Multi-gigahertz operation of photon counting InGaAs avalanche photodiodes,” *Appl. Phys. Lett.*, vol. 96, no. 7, Feb. 2010.
[171] A. Restelli and J. C. Bienfang, “Avalanche discrimination and high-speed counting in periodically gated single-photon avalanche diodes,” in *Advanced Photon Counting Techniques VI*, vol. 8375, Bellingham: SPIE-Int. Soc. Optical Engineering, 2012, p. 83750Z.
[172] X. Chen, E. Wu, G. Wu, and H. Zeng, “Low-noise high-speed InGaAs/InP-based single-photon detector,” *Opt. Express*, vol. 18, no. 7, pp. 7010–7018, Mar. 2010.
[173] J. Zhang, R. Thew, C. Barreiro, and H. Zbinden, “Practical fast gate rate InGaAs/InP single-photon avalanche photodiodes,” *Appl. Phys. Lett.*, vol. 95, no. 9, Aug. 2009.
[174] Y. Jian, E. Wu, G. Wu, and H. Zeng, “Optically Self-Balanced InGaAs-InP Avalanche Photodiode for Infrared Single-Photon Detection,” *IEEE Photonics Technol. Lett.*, vol. 22, no. 3, pp. 173–175, Feb. 2010.
[175] A. Restelli, J. C. Bienfang, and A. L. Migdall, “Gigahertz-gated InGaAs SPAD system with avalanche charge sensitivity approaching the fundamental limit,” in *Advanced Photon Counting Techniques VII*, vol. 8727, M. A. Itzler and J. C. Campbell, Eds. Bell: SPIE-Int. Soc. Optical Engineering, 2013, p. 87270F.
[176] V. Ntziachristos, J. Ripoll, L. V. Wang, and R. Weissleder, “Looking and listening to light: the evolution of whole-body photonic imaging,” *Nat. Biotechnol.*, vol. 23, no. 3, pp. 313–320, 2005.
[177] B. Aull, J. Burns, C. Chen, B. Felton, H. Hanson, C. Keast, J. Knecht, A. Loomis, M. Renzi, A. Soares, V. Suntharalingam, K. Warner, D. Wolfson, D. Yost, and D. Young, “Laser Radar Imager Based on 3D Integration of Geiger-Mode Avalanche Photodiodes with Two SOI Timing Circuit Layers,” in *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, 2006, pp. 1179–1188.
[178] M. A. Itzler, M. Entwistle, M. Owens, K. Patel, X. Jiang, K. Slomkowski, S. Rangwala, P. F. Zalud, T. Senko, J. Tower, and J. Ferraro, “Geiger-mode avalanche photodiode focal plane arrays for three-dimensional imaging LADAR,” in *Infrared Remote Sensing and Instrumentation XVIII*, vol. 7808, M. Strojnik and G. Paez, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2010, p. 78080C.
[179] R. D. Younger, K. A. McIntosh, J. W. Chludzinski, D. C. Oakley, L. J. Mahoney, J. E. Funk, J. P. Donnelly, and S. Verghese, “Crosstalk analysis of integrated Geiger-mode avalanche photodiode focal plane arrays,” in *Advanced Photon Counting Techniques III*, vol. 7320, M. A. Itzler and J. C. Campbell, Eds. Bellingham: SPIE-Int. Soc. Optical Engineering, 2009, p. 73200Q.