Review Vol. 31, No. 21 / 9 Oct 2023 / Optics Express 33963
Optics EXPRESS

Timing measurements with silicon single photon avalanche diodes: principles and perspectives [Invited]

GIULIA ACCONCIA,[1] FRANCESCO CECCARELLI,[2] ANGELO GULINATTI,[1] AND IVAN RECH[1,*]

[1]Dipartimento di Elettronica, Informazione e Bioingegneria - Politecnico di Milano, Piazza Leonardo da Vinci 32, Milano, 20133, Italy
[2]Istituto di Fotonica e Nanotecnologie - Consiglio Nazionale delle Ricerche (IFN-CNR), Piazza Leonardo da Vinci 32, Milano, 20133, Italy
*ivan.rech@polimi.it

Abstract: Picosecond timing of single photons has laid the foundation of a great variety of applications, from life sciences to quantum communication, thanks to the combination of ultimate sensitivity with a bandwidth that cannot be reached by analog recording techniques. Nowadays, more and more applications could still be enabled or advanced by progress in the available instrumentation, resulting in a steadily increasing research interest in this field. In this scenario, single-photon avalanche diodes (SPADs) have gained a key position, thanks to the remarkable precision they are able to provide, along with other key advantages like ruggedness, compactness, large signal amplitude, and room temperature operation, which neatly distinguish them from other solutions like superconducting nanowire single-photon detectors and silicon photomultipliers. With this work, we aim at filling a gap in the literature by providing a thorough discussion of the main design rules and tradeoffs for silicon SPADs and the electronics employed along them to achieve high timing precision. In the end, we conclude with our outlook on the future by summarizing new routes that could benefit from present and prospective timing features of silicon SPADs.

1. Introduction

Time-resolved single-photon measurements are growing today as enabling tools in fields as diverse as quantum communication [1], near-infrared spectroscopy [2] and computational imaging [3]. Such applications pose great technological challenges such as meeting the highly demanding requirements set on the single-photon detection system. Indeed, high sensitivity, picosecond timing resolution and, often, the possibility to be integrated in multichannel modules by exploiting large and/or dense single-photon detector arrays are only some of the features that directly affect the performance of a time-resolved measurement. Such an ambitious combination of parameters forces the designer to a thorough optimization of the entire detection system.

In the large landscape of single-photon technologies, superconducting nanowire single-photon detectors (SNSPDs) [4,5] are attracting tremendous attention due to the exquisite photon detection efficiency (PDE), even higher than 90%, the very low dark count rate (DCR), commonly a few counts per second (cps), and the remarkable timing jitter, typically between 20 and 100 ps full width half maximum (FWHM). Recently, Korzh et al. [6] have also demonstrated a record-breaking jitter lower than 3 ps, even though the adopted geometry makes this specific device difficult for the exploitation in a real-life experiment. More generally, SNSPDs require a complex setup based on a cryostat for their operation (typical operating temperature is 4 K) and, in addition, the manufacturing technology is not yet as mature as needed to implement detector arrays with performance similar to the one demonstrated for a single pixel [7].

#491400 https://doi.org/10.1364/OE.491400
Journal © 2023 Received 22 Mar 2023; revised 7 Sep 2023; accepted 10 Sep 2023; published 26 Sep 2023

Optics EXPRESS

An alternative option, operating at room temperature (or slightly below), is the use of single-photon avalanche diodes (SPADs) [8,9]. These are semiconductor pn junctions biased over the breakdown voltage that are able to produce a macroscopic current pulse when triggered by a single photon. With respect to other room temperature solutions like photomultiplier tubes (PMTs) [10] and hybrid single photon detectors [11], silicon SPADs are purely solid state devices able to provide higher PDE (especially for red and near-infrared wavelengths [12]), lower operating voltages and a more compact design. In addition, silicon SPADs can be integrated in megapixel arrays [13,14] while preserving the integrity of the photon timing signal, different from what happens with large-area silicon photomultipliers (SiPMs) [15]. Last but not least, SPADs can achieve a DCR and a maximum count rate comparable to commercial SNSPDs [16–18]. However, attaining a suitable combination of all these parameters is typically a complex task, requiring not only a deep understanding of the device physics, but also a broader knowledge that involves how to design the front end electronics in order to extract precise photon timing information. SPADs and front end electronics are indeed profoundly interrelated and each design choice considered on one side will affect the solutions that will be adopted on the other side.

In this review, we aim at providing a complete guide on how to achieve high-precision time-resolved (i.e. timing) single-photon measurements with silicon SPADs in a wavelength range between 400 and 1000 nm. Hereafter, we will not specify the material anymore, unless for a few examples of SPADs and other avalanche detectors designed for operation beyond 1000 nm, which are typically based on different materials (e.g. InP/InGaAs, Ge, HgCdTe) and structures (e.g. separate absorption, charge and multiplication regions). For the sake of compactness, we will not provide details regarding these devices. The reader interested in this topic can refer to [8,19–22].

This paper is oriented both toward scientists interested in understanding the best SPAD-based detection solution for a given application and toward engineers that are investigating the main tradeoffs and guidelines for the design of a SPAD-based timing system. Therefore, we start by reviewing the physics behind the avalanche and the main models developed to link the temporal response to the main SPAD design parameters and we continue by discussing the highest performance SPADs demonstrated and reported in the literature for timing applications. Throughout the review we also address the strong link between device physics and front end electronics and, in the end, we conclude this work by providing our perspective on likely future work and technological challenges that will come in this fascinating research field.

2. Single-photon avalanche diode

In this section we start from the SPAD avalanche physics and modeling, moving through the state of the art toward the highest performance SPADs reported so far. However, we will focus only on the features that make the SPAD an ideal solution for a single-photon timing measurement, therefore we assume the readers are familiar with some background definitions. Readers interested in a more comprehensive review can refer to [8].

2.1. SPAD fundamentals and basic models

A SPAD is basically a pn junction, which is reverse biased above its breakdown voltage. In this condition, if no free carriers are present in the space charge region, the device remains in a quiescent state with no current flowing through it. However, if a photon is absorbed, a free electron-hole pair is photo-generated. The high electric field of the space charge region then accelerates any free carrier in this area to an energy high enough to produce other pairs by means of impact ionization. Since the applied voltage is greater than the breakdown value, the carrier multiplication process results in a self-sustained macroscopic current that is limited only by the parasitic resistance in series with the detector. Such a process is traditionally modeled with the equivalent circuit proposed by Cova et al. [23] (Figure 1(a)), in which the SPAD is considered

Optics EXPRESS

a voltage source (value equal to the breakdown voltage V_{bd}) surrounded by the resistive and capacitive parasitics of the main pn junction and of other junctions, which are typically employed for electrical isolation from the substrate and/or other devices integrated in the same die, and the photon arrival is mimicked through an external pulse (*Photon*) of proper duration.

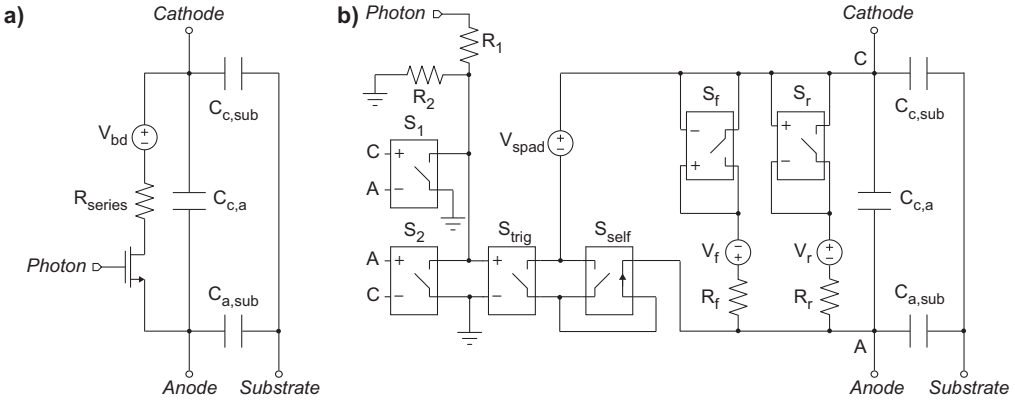

Fig. 1. Basic modeling of the SPAD relying on equivalent circuits. a) Traditional model proposed by Cova et al. [23]. b) Traditional model [23] improved and complemented by Dalla Mora et al. [26,27].

During the detection process, the external electronics (i.e. the front end) plays a key role. Indeed, once the avalanche is triggered, the detector is blind and cannot detect further photons until it is reset. In order to quench the current and reset the device to the initial condition, a proper circuit is needed. Such a circuit can be simply a high-value resistor, but this typically leads to a slow reset transient, during which parameters like timing jitter, PDE or DCR are not constant. All these drawbacks are not acceptable for the majority of the photon timing experiments and so more complex circuital solutions, typically based on active components (e.g. transistors [24]) and/or novel passive devices (e.g. adaptive resistive switches [25]), are considered. The design and role of these circuits in a photon timing experiment will be considered throughout Section 3. Here, the main point is that as the complexity of the front end increases, so does the complexity of the SPAD simulation model. Therefore, throughout the years the traditional model reported in Figure 1(a) has been complemented with additional parts, with the final aim of adding features and peculiarities of real detectors. As an example, Dalla Mora et al. [26] introduced triggering, self-sustainability and self-quenching of the avalanche by means of current/voltage controlled switches. In addition, they decided to replace the linear model adopted in Figure 1(a) for the I-V characteristic curve of the SPAD with a piece-wise linear function, implemented through the non-constant voltage source V_{spad}. Such a voltage source provides a correct modeling only in the breakdown region and was recently complemented with two additional branches in which a current flows when either forward (subscript f) or reverse (subscript r) biased above the second breakdown (due to, for example, edge effects) [27]. The circuit model, fully compatible with commercial simulators like OrCAD PSpice and Virtuoso Spectre, is reported in Figure 1(b). Models like this are ideal for designing the quenching electronics: current waveform, avalanche charge, power dissipation and quench/reset times are typically the main parameters that are accessible by employing this modeling. In addition, more complex simulations are also possible by complementing the equivalent circuits with stochastic models able to generate triggering events through the input *Photon*. A relevant example is [28], in which Giustolisi et al. developed a behavioral model including photon arrival, dark counts due to thermally-generated carriers and aftepulses. Later, Cheng et al. [29] improved this model by adding also the band-to-band tunneling contribution to the DCR and the temporal dependence of the aftepulsing counts. In

Optics EXPRESS

both cases, Verilog-A is the language employed to integrate the statistical behaviour of the detector into Virtuoso Spectre. These models find employment when system-level metrics need to be simulated, as an example in applications like light detection and ranging (LiDAR) [30–32] and positron emission tomography (PET) [33,34]. Finally, it is also worth noting that this approach to the device modeling was employed successfully also for SiPMs [35,36], InP/InGaAs SPADs operated either in gated [37,38] or free-running mode [38–40] and even for novel device structures like perimeter-gated CMOS SPADs [41] (i.e. SPADs with an additional gate terminal that surrounds the junction and prevents premature edge breakdown).

Although these models are perfect for the simulation of quenching and even system electronics, it is necessary to note that they have an important limitation. Indeed, the detection process is not modeled at a level that is sufficient to accurately reproduce the temporal response and, therefore, relying on these models for the design of a readout circuit able to extract the timing information by preserving its precision is not possible. The physical description of the temporal response and its mathematical modeling are the subjects of the next section.

2.2. Temporal response

The temporal response of the SPAD is typically quantified through the instrument response function (IRF) and an example of this curve is reported in Figure 2(a). Besides a normalization factor, the IRF represents the probability density of the photon detection time T_{det} recorded by the readout electronics. As reported in Figure 2(a), it is usually characterized by a strong Gaussian peak, whose width is referred to as timing jitter and quantified through its FWHM, and by a lower exponential tail, typically quantified by its lifetime. Often, in order to include also the effect of the exponential tail, the IRF distribution is evaluated by considering not only its FWHM, but also by means of the full width at 1/10th maximum (FWTM) or the full width at 1/100th maximum (FW1/100M).

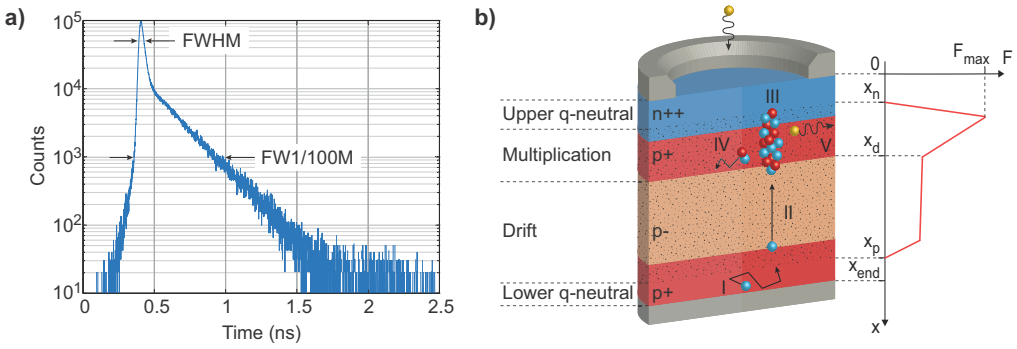

Fig. 2. Temporal response of a SPAD device. a) Example of IRF for a custom-technology SPAD [8]. Both FWHM and FW1/100M are quoted on the graph. b) Electric field and schematic structure of a SPAD, in which all the processes possibly contributing to the statistical dispersion of the photon detection time are depicted: I) carrier diffusion, II) carrier drift, III) avalanche build-up, IV) avalanche lateral propagation assisted by secondary carrier diffusion and V) avalanche lateral propagation assisted by secondary photons.

The detection time T_{det} can be seen as the sum of different contributions, each of them corresponding to the random time spent on a different part of the detection process:

$$T_{det} = T_{diff} + T_{drift} + T_{build} + T_{prop}, \tag{1}$$

where:

Optics EXPRESS

- T_{diff} is the time elapsed between the generation of a carrier pair in a quasi-neutral region and the actual collection of a carrier by the electric field present in the depletion region. It is worth noting that, in case the photon is absorbed within the depletion region, this contribution is zero. This process is depicted in Figure 2(b) (I).
- T_{drift} is the time elapsed between the appearance of the carrier in the depletion region and its arrival in the multiplication region, where the electric field is high enough to produce impact ionization events. This process is depicted in Figure 2(b) (II).
- T_{build} is the time interval in which the avalanche starts growing, but remains confined in the narrow filament that includes the point of appearance of the carrier in the depletion region. This process is depicted in Figure 2(b) (III).
- T_{prop} is the time during which the avalanche continues its growth until the current reaches the detection threshold set by the readout electronics. In this phase the avalanche propagates laterally throughout all the active area of the SPAD assisted either by secondary carrier diffusion or by secondary photons. These two processes are depicted in Figure 2(b) (IV) and (V), respectively.

All these variables, which are intrinsically related to the SPAD physics, contribute to the statistical distribution of the avalanche current waveforms thus setting the ultimate limit to the dispersion (from photon to photon) of the IRF. Hereafter, we proceed by analyzing each of these contributions. If not otherwise stated, the following discussion will always take as a reference the standard n-over-p SPAD structure reported in Figure 2(b), featuring an electric field profile encompassing both a high-field multiplication and a low-field drift region. While the first region is responsible for the avalanche ignition and growth, the second one is typically employed in SPADs to promptly collect carriers generated by photons absorbed deeper into the detector. Other structures that have demonstrated remarkable single-photon timing capabilities (e.g. the p-i-n diode) will be taken into account as well.

2.2.1. Carrier diffusion

The first contribution that we analyze is T_{diff}, which is the random delay at the origin of the SPAD exponential tail. Such a non-ideality cannot be ignored since it may produce distortion on timing measurements and, in particular, it is under the spotlight in applications like high clock-rate quantum key distribution (QKD), where it may limit the quantum bit error rate (QBER) [42,43], or time-resolved near-infrared spectroscopy, when time gating is employed to improve acquisition speed and dynamic range [44,45]. As we will discuss in this section, the exponential tail is strictly connected to the diffusion phenomenon and, therefore, many refer to the tail as "diffusion tail" and to T_{diff} as "diffusion time" even though this would not be strictly correct since here we are considering the random motion of a single carrier and not the motion of an ensemble of carriers.

The physical origin of the diffusion time T_{diff} is related to the photons absorbed in the quasi-neutral regions of the device, which generate electron-hole pairs that are not promptly accelerated by the electric field, as opposed to what happens when photons are absorbed in the depletion region. Considering the representation of Figure 2(b) (I), a minority electron generated in the lower quasi-neutral region moves as a random walker until either it recombines, it escapes laterally or it gets collected by the electric field and directed toward the multiplication region. The first two possibilities do not result in an avalanche, however they are typically negligible if the device is fabricated through a reasonably clean process (recombination in silicon is mediated by defects) and if the active area of the device is many times larger than the thickness of the quasi-neutral region. On the contrary, in the third case there is a possibility for the avalanche to be ignited. Since the initial motion of the electrons is not subject to a strong driving force, this is a process that can introduce a non-negligible random delay between the photon absorption and

Optics EXPRESS

the actual detection, which corresponds to T_{diff} and whose value can be even in the nanosecond range. Obviously, similar considerations hold also for a minority hole generated in the upper quasi-neutral region and, as a result, the exponential tail measured experimentally on a SPAD is the result of the superposition of both of these contributions. Tail lifetime and ratio of the number of counts with respect to the gaussian peak depend on the distribution of the photogenerated carriers as a function of depth. As the photon wavelength increases, also the light penetration increases and, as a result, the tail becomes slower and with higher intensity relative to the gaussian peak [46].

At first sight, the most natural way to model the exponential tail is through a Monte Carlo simulation. In a few words, after a photon is absorbed in one of the quasi-neutral regions, the random walk of the generated carrier is computed to establish if and when it will reach the depletion region. In principle, this means that the carrier motion is simulated at each step with the extraction of a random direction and a random distance that is travelled with no collision. By iterating this process, the whole motion is computed and, by repeating it for a statistically significant number of photons, it is possible to generate the whole exponential tail. However, this approach is typically time consuming and other solutions are actually implemented in practice. The first modeling was proposed by Ripamonti et al. [46] by preserving the single-carrier approach, but, on the other hand, introducing the concept of macro-collision, which accounts for a large number of real collisions. The result of a macro-collision is computed on the basis of the Gaussian solution of the time-dependent diffusion equation for a point source initial condition in an unlimited 3D domain. This solution provides the spatial probability density distribution of the carrier for any assigned delay t_d; however, the choice of this parameter for each macro-collision may be tricky especially if we consider the proximity to the boundaries, where the unlimited domain hypothesis is at stake.

Recently, Gulinatti et al. [47] proposed a different paradigm able to overcome the problems of the Monte Carlo simulation. Indeed, nothing changes if, instead of just one carrier, many non-interacting carriers are present at the same time in the quasi-neutral region. This means that the same phenomenon can be described macroscopically as the collective motion of a large number of carriers subject to the drift-diffusion equations and the result would be the same as produced by an equally large number of Monte Carlo iterations. It is worth noting that, although the diffusion plays the most important role, such a model allows one to include also the effects of the faint electric field present in the quasi-neutral regions due to doping gradients. More specifically, if we refer again to Figure 2(b) (I), the contribution to the exponential tail given by the electrons can be calculated in this framework as the electron flow $F_n(x_p, t)$, namely the number of electrons per unit time that enters the depletion region from the lower quasi-neutral region, assuming an absorbing boundary condition at $x = x_p$ (i.e. electron concentration $n(x_p) = 0$). Gulinatti and coworkers verified the validity of their assumptions by implementing a 1D version of this model and by showing good prediction on thin SPADs fabricated through the same custom technology [48], yet with different doping profiles. An example of simulation [47] is depicted in Figure 3(a), where both the upper and the lower quasi-neutral region contributions are reported. Gulinatti et al. were recently followed in this approach also by Sun et al. [49] who validated this model with p+/n-well SPADs fabricated in a 0.18 µm CMOS technology. A similar model was also exploited more recently by Helleboid and coworkers [50], who decided to start by computing the 3D electric field profile and to continue by integrating the drift-diffusion equation along the field lines calculated at the former step. In this way, it is possible to apply similar considerations also to more complex structures by preserving the 1D nature of the involved equations, but repeating the calculation for field lines developing on a 3D domain. The model was validated on different n+/p-well SPAD architectures that differ in the size of the n+ implant.

Since the exponential tail is strongly connected with the diffusion of minority carriers throughout the quasi-neutral regions, the lifetime of the exponential tail is also strongly connected

Optics EXPRESS

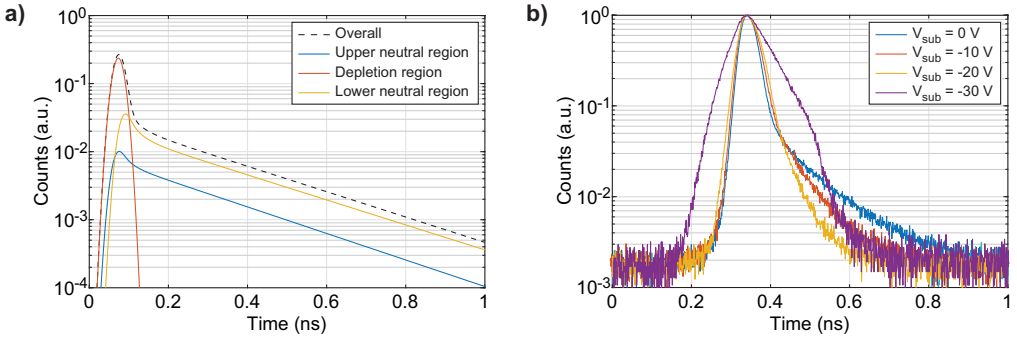

Fig. 3. Simulation and peculiar features of the exponential tail. a) Temporal response of a custom-technology thin SPAD at $\lambda = 800$ nm as simulated by Gulinatti and coworkers [47]. All the contributions coming from the different regions of the devices are reported separately. b) Temporal response of a 100 µm CMOS SPAD at $\lambda = 850$ nm as measured by Villa and coworkers [55]. The different substrate voltages V_{sub} allow the reader to appreciate the effect of reducing the lower quasi-neutral region on both the tail and the FWHM. Data courtesy of Prof. Federica Villa (Politecnico di Milano).

with the thickness of these regions, which, in order to sharpen the SPAD temporal response, have to be always designed as thin as possible. Following this idea, SPAD temporal responses featuring a lifetime lower than 100 ps [16,17,51–53] or even free of slow components [54,55] were proposed and successfully demonstrated. For example, Spinelli and coworkers [54] modified the structure of the custom-technology thin SPAD previously reported by Lacaita et al. [56] by removing the p+ buried layer in the region immediately below the active area of the device. As a result, the drift region is further extended toward the depletion region of the substrate junction that is typically employed to isolate the SPAD from the rest of the die. In this condition, if voltages are high enough, the two depletion regions merge into a single one and the quasi-neutral region in-between is completely suppressed. However, as pointed out later in [47], this solution is not well suited to large-area SPADs (diameter $d > 10$ µm). Indeed, Villa et al. [55] demonstrated experimentally in a 0.35 µm CMOS-technology SPAD featuring a diameter as high as $d = 100$ µm that a complete suppression of the quasi-neutral region, obtained by progressively decreasing the substrate voltage V_{sub} from 0 to −30 V, makes the tail progressively disappear, but, on the other hand, it leads also to a FWHM increasing from 92 to 177 ps (Figure 3(b)). This detrimental effect is connected to the high series resistance seen by the carriers generated upon the absorption of a photon close to the center of the SPAD active area and will be thoroughly addressed in Section 2.2.4.

Deep-junction SPADs As an interesting case study, let us now consider the case of deep-junction SPADs [16,17,57–64]: these devices are characterized by a buried multiplication region that results in an electric field profile typically mirrored with respect to the one reported in Figure 2(b). These SPADs are flourishing thanks to the introduction of deep n-wells in CMOS imaging and high-voltage processes, typically realized through high-energy ion implantation and thus featuring a retrograde doping profile (i.e. the doping concentration increases as one goes deeper), which is also exploited to avoid premature edge breakdown with no need of guard rings [65]. In order to study the properties of the exponential tail in deep-junction SPADs let us consider two different situations: firstly, the illumination with photons having short wavelengths ($\lambda < 500$ nm), which results in a strong absorption in the first hundreds of nanometers of the device and thus in a non-negligible probability to generate free carriers in the upper quasi-neutral region; secondly, the illumination with photons having long wavelengths ($\lambda > 600$ nm), which

Optics EXPRESS

results in broadly distributed absorption throughout the device and thus in a non-negligible probability to generate free carriers in the lower quasi-neutral region.

Let us start by considering the first scenario (short wavelengths) and the case of a deep-junction SPAD with a depletion region featuring only a thin multiplication layer (i.e. with no drift region). This situation can result in a large broadening of the IRF with two different emerging peaks [47]: one that is faster and sharper, related to the carriers absorbed in the depleted region, and another one that is slower and broader, related to the diffusing carriers absorbed in the thick upper quasi-neutral region. This phenomenon is evident if we consider the SPAD proposed in the literature by Webster et al. [58] fabricated in a 130 nm CMOS technology, whose structure and electric field profile are both reported in Figure 4(a). This device is implemented through a deep n-well on a p-type substrate, a design that results in one of the best PDE curve provided by a CMOS SPAD (Figure 4(c)), with a maximum value of 72% at $\lambda = 560$ nm ($\approx 33\%$ at $\lambda = 400$ nm). However, due to the thick upper quasi-neutral region, the temporal response at the short wavelength is not suitable for picosecond timing applications, since the FWHM = 1.55 ns at $\lambda = 443$ nm is limited by a tail lifetime as high as 1.4 ns (Figure 4(e)). A solution to this issue is already pointed out by the same authors in a different work [59], in which they report a SPAD fabricated in a 90 nm CMOS technology where they exploit a similar structure, yet with an additional shallow p-well introduced to prevent the slowest holes from reaching the multiplication region. The final result is a sharp temporal response also in the short wavelength range, with a tail lifetime of only 210.7 ps and a corresponding FWHM of 84 ps at $\lambda = 443$ nm (Figure 4(e)). However, it is worth mentioning that in this case the authors have traded off a sharper temporal response with a lower PDE, which moves from about 33% to only 5% at $\lambda = 400$ nm (Figure 4(c)). On the contrary, such a tradeoff between temporal response and PDE is not present when the upper layer is exploited to accommodate the drift region of the device. This was demonstrated lately by Sanzaro et al. [17] and Gramuglia et al. [16] in a 160 nm BCD and in a 180 nm CMOS technology, respectively. They both exploited a deep-junction SPAD as Webster et al. [58,59], however with a depletion region extending for a large part of the upper p layer and, thus, resulting in a quasi-neutral region that is concentrated only at the surface of the device. In particular, Figure 4(b) reports the SPAD structure designed by Gramuglia et al. [16], which is based on a p-well (i.e. the anode) isolated from the p-type substrate by a buried n-well layer (i.e. the cathode). The presence of a low-doped p epitaxial layer allows the implementation of a p-i-n structure, in which the electric field is approximately flat and extends over the upper p region (Figure 4(b)). The final result is a sharp temporal response along with no detrimental effect on the PDE at the short wavelengths. Such an engineered device allowed Gramuglia et al. [16] to experimentally demonstrate a tail lifetime as low as 31.5 ps at $\lambda = 515$ nm (Figure 4(e)), along with a PDE as high as 30% at $\lambda = 400$ nm (Figure 4(d)). Moreover, such an outstanding result has been recently improved to achieve lifetime values even lower than 30 ps [66].

Let us move now on the long wavelengths ($\lambda > 600$ nm) and let us consider again as a reference the SPAD proposed by Webster and coworkers in a 130 nm CMOS technology [58]. Even in this spectral range the performance in terms of PDE is remarkable, with about 60% at $\lambda = 654$ nm. Moreover, at this operating wavelength the temporal response is not subject to the same effect observed for the shorter wavelengths and, as reported in Figure 4(f), the resulting FWHM is as low as 77 ps. However, even if the sharp FWHM is preserved, the same cannot be said for the tail, which is characterized by a lifetime as high as 1.7 ns. The origin of such a slow tail can be found in the fact that the SPAD is not fabricated in an isolated well, but, on the contrary, the anode of the device is electrically connected to the p-type substrate and, thus, all the electrons generated in the substrate have a chance to reach the multiplication region and trigger an avalanche. A possible solution consists again in preventing the slower carriers from reaching the depletion region and this can be achieved by resorting to one of the SPAD structures proposed, as an example, by Sanzaro et al. [17] or by Gramuglia et al. [16], which both feature dedicated doping

Optics EXPRESS

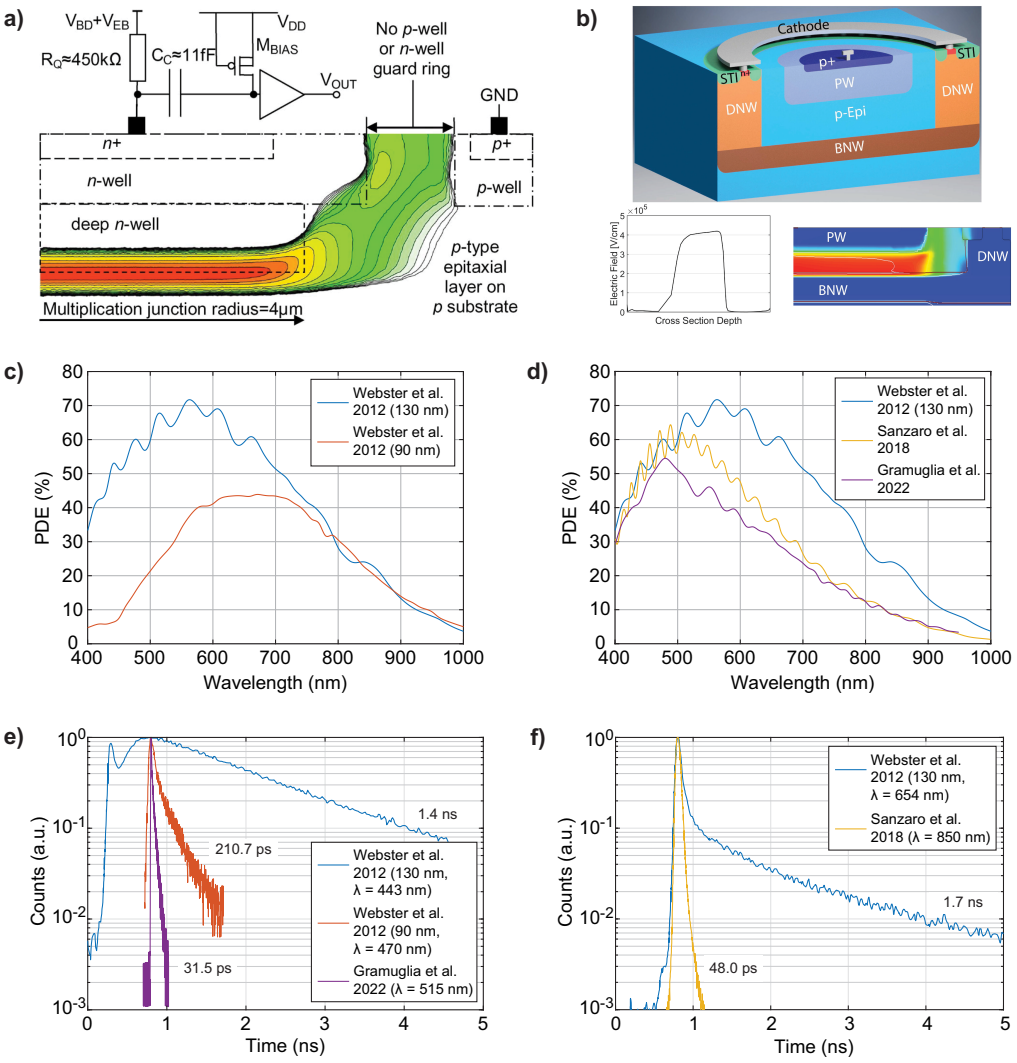

Fig. 4. Deep-junction SPADs. a) Device structure and electric field (color map) of the SPAD developed by Webster et al. [58] in a 130 nm CMOS technology. Reproduced with permission. Copyright 2012, IEEE. b) Device structure and electric field (reported through a plot and a color map) of the SPAD developed by Gramuglia et al. [16] in a 180 nm CMOS technology. Reproduced under terms of the CC-BY license. Copyright 2022, The Authors, published by IEEE. c) PDE as a function of the wavelength reported for the non-isolated SPADs proposed by Webster et al. in [58] (130 nm) and [59] (90 nm). d) PDE as a function of the wavelength reported for the isolated SPADs proposed by Sanzaro et al. [17] and Gramuglia et al. [16]. The figure reports also Webster et al. (130 nm) [58] for comparison purposes. e) IRF measured at short wavelengths for Webster et al. (130 nm) [58] compared to Webster et al. (90 nm) [59] and Gramuglia et al. [16]. Wavelengths are λ = 443 nm, λ = 470 nm and λ = 515 nm, respectively. Tail lifetime is reported for each curve. f) IRF measured at long wavelengths for Webster et al. (130 nm) [58] compared to Sanzaro et al. [17]. Wavelengths are λ = 654 nm and λ = 850 nm, respectively. Tail lifetime is reported for each curve.

Optics EXPRESS

wells that host the SPAD and that are insulated from the substrate (Figure 4(b)). Thanks to this design, Sanzaro et al. [17] managed to demonstrate a fast tail as low as 48.0 ps at $\lambda = 850$ nm (Figure 4(f)). However, once again, preventing the collection of slow carriers means also having a lower PDE, which is less than 10% at this wavelength, more than a factor of 2 lower with respect to what was demonstrated by Webster et al. with a non-isolated substrate (Figure 4(d)). The tradeoff can be broken also in this case, but in order to do that it is necessary to expand in depth both the isolated doping well and the depletion region of the SPAD, preferably by introducing a drift region in order to limit operating voltages and power dissipation. A case study will be thoroughly treated in the next section.

2.2.2. Carrier drift

After the exponential tail and its modeling, we now focus our attention on the photons absorbed within the depletion region. As depicted in Figure 3(a), these are typically the majority of all the detection events and they contribute to the overall IRF with a fast Gaussian peak, whose FWHM is connected to the statistical dispersion of different factors. The first one is the transit time T_{drift} through the SPAD depletion region. Indeed, if we consider again the graphical representation of Figure 2(b) (II), we promptly realize that, even considering only the depletion region, the photons that are absorbed at different depths will produce carriers that will reach the multiplication region with different delays.

In principle, the modeling of this phenomenon can be carried out with the same approach proposed for the exponential tail and based on the analytical/numerical solution of the drift-diffusion equation [47,49,50]. Even though this approach is theoretically faultless, it is also unnecessarily complex for most of the SPAD structures. Indeed, the SPAD electric field is typically designed to accelerate the carriers up to the saturation velocity (i.e. $v_{sat} \approx 1{\times}10^7$ cm s^{-1} in silicon, corresponding to about 10 ps µm^{-1}) and, therefore, it is higher than $F_{sat} \approx$ 2-3$\times10^5$ V m^{-1} in the entire depletion region, even in the drift part. This means that (a) the carrier can be treated as a purely drifting particle and (b) its velocity can be considered constant and equal to the saturation velocity no matter the local value of the electric field [47]. For these reasons, photons absorbed in the drift region typically contribute to the IRF peak and not to its tail. However, similarly to what happens with the diffusion time T_{diff}, the statistical dispersion of the drift time T_{drift} increases as both the thickness of the depletion region and the absorption length of the photons increases, therefore this contribution is more evident in SPADs with high PDE at long wavelengths. Given a wavelength that results in a quasi-uniform absorption probability throughout the depletion region, this contribution can be roughly estimated as the drift time $v_{sat}(x_n - x_p)$, which is about 100 ps for a SPAD having a 10 µm deep depletion region.

Thin vs. red-enhanced SPADs A useful case study to better understand the contribution given by the drift time T_{drift} is the comparison between thin SPAD and red-enhanced SPAD (RE-SPAD), both developed at Politecnico di Milano by exploiting fully-custom silicon fabrication processes. The structure of a thin SPAD is reported in Figure 5(a) and it is thoroughly described in [8,48], while the structure of the RE-SPAD is shown in Figure 5(b) and comprehensively illustrated in [8,12]. Besides some specific differences between the two structures, which allow them to have comparable performance in terms of DCR, afterpulsing and a relatively low power dissipation in both cases, we here focus on the thickness of the quasi-intrinsic layer that is of about 2 µm for the thin SPAD and about 10 µm for the RE-SPAD. Such a difference extends the drift region in the RE-SPADs, thus allowing an effective collection of the carriers generated deep into the detector with remarkable results in terms of PDE at long wavelenghts, which increases from about 15% to about 45% at 800 nm as reported in Figure 5(d). At the same time, the thicker depletion region is also responsible for the IRF peak broadening, as shown in Figure 5(e), due to the increment of the drift time in this part of the device. For the sake of completeness, the electric field profile of

Optics EXPRESS

the RE-SPAD is reported in Figure 5(c). Also in this case, electrons move at saturated velocity in the whole depletion region.

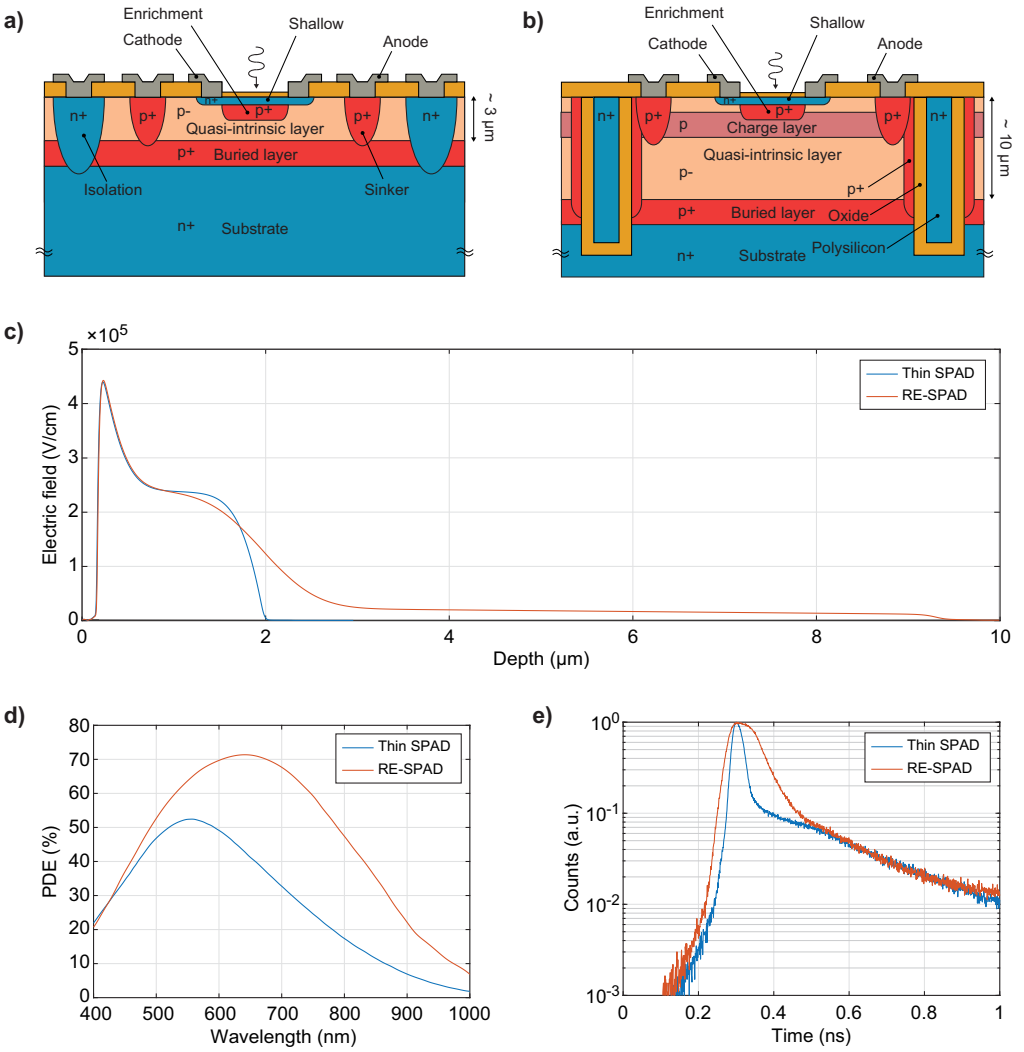

Fig. 5. Custom-technology thin and RE-SPADs. a) Device structure of the thin SPAD developed at Politecnico di Milano [48]. b) Device structure of the RE-SPAD developed at Politecnico di Milano [12] on the basis of the thin device. c) Electric field profile for a thin and a RE-SPAD. While the peak in the multiplication region is similar, the RE-SPAD features an extended drift region for the prompt collection of electrons generated by long wavelength photons. d) PDE as a function of the wavelength reported for the thin and the RE-SPAD. The PDE at λ = 800 nm increases from 17% (thin SPAD) to 47% (RE-SPAD). e) IRF measured at λ = 820 nm for the thin and the RE-SPAD. The FWHM is 34 and 94 ps, respectively.

This example demonstrates that the tradeoff between PDE and timing response described in Section 2.2.1 can be only mitigated by the extension of the depletion region. Indeed, while having an extended depletion region is effective to avoid a long tail in thick devices, this in turn broadens the IRF peak. The tradeoff between jitter and PDE is actually an open problem and,

Optics EXPRESS

in order to finally break this tradeoff, novel structures are needed. Some examples are briefly discussed in Section 4.

2.2.3. Avalanche build-up

In addition to the contribution given by the collection process (either by diffusion or drift), another major contribution to the jitter typically arises from the avalanche growth dynamics. Once the first carrier reaches the multiplication region, it starts generating other carrier pairs by impact ionization. At this stage, the current grows exponentially, but there is still a relatively small number of free carriers in the depletion region, which are confined in a small filament that encloses the seed point and extends in the same direction as the electric field [67] (see Figure 2(b) (III)). As the number of carriers increases, their charge becomes large enough to weaken the electric field, slowing down the multiplication rate. Eventually, the local electric field is depressed to approximately the breakdown level, causing the current density to saturate at a value that allows the process to be self-sustained. This phenomenon is typically modeled by introducing a parasitic resistance in series to the junction, i.e. the space-charge resistance [68]. Although the space-charge resistance limits the maximum current density in the filament, it is worth noting that during this phase the voltage drop across the series bulk resistance of the SPAD is still negligible. This stage of the avalanche growth is typically referred to as build-up, while the time T_{build} between the beginning of the multiplication process and the saturation of the current density within the first filament is referred to as the build-up time. This is a random contribution to the total detection time T_{det} due to the noise of the avalanche multiplication process, which results in random fluctuations in the leading edge of the SPAD current. The study of the build-up contribution to the timing jitter is key because it represents the ultimate limit to the timing resolution, achievable only by sensing the faint avalanche current before it starts propagating through the whole active area. However, even by employing ultra-low threshold front end circuits, attaining this limit is a really challenging task due to the presence of all the intrinsic SPAD parasitics (e.g. the SPAD junction capacitance) and to the filtering action they have on the avalanche current. These concepts will be thoroughly addressed in Section 3.2

Spinelli et al. [69] were the first to model this process. They implemented a 1D Montecarlo simulator of the avalanche dynamics, in which the carriers drift at saturated velocity in a given non-uniform electric field profile and the random length x of the ionization path is chosen step by step through the probability density function $\alpha \exp(-\alpha x)$, where α is the ionization coefficient corresponding to the electron (β for the case of a hole) and its dependence on the electric field is taken into account through the self-scattering method [70]. The total current is calculated by summing up each elementary contribution estimated by using Ramo's theorem [71]. In literature, this simulation approach is usually referred to as random path length (RPL) Montecarlo method. It is worth noting that the RPL model includes also the possibility of having a carrier that crosses the multiplication region without creating other pairs. In principle, due to this possibility, it is not certain that a self-sustained avalanche will be triggered. Indeed, Spinelli and coworkers point out the strong connection between the simulation of the build-up process and the breakdown probability (also called triggering efficiency), which is the probability of having a self-sustained avalanche given a successful photon absorption and collection [72–74]. The probability of having a carrier escape the depletion region without an ionization event is not negligible and, as reported in Figure 6(a), this is particularly evident in the first part of the avalanche build-up process (current $< 1\,\mu A$). This part of the avalanche build-up gives the major contribution to the statistical dispersion of the build-up time T_{build}. Indeed, the avalanche noise is averaged out as soon as the current increases (avalanche current $> 1\,\mu A$) and the true exponential nature of the process emerges (see again Figure 6(a)), making the build-up jitter asymptotically independent of the detection threshold [75]. Spinelli et al. applied their model to the simulation of an old generation of custom-technology thin SPADs, showing that, operating these devices at reasonably

Optics EXPRESS

high excess bias, the dispersion of the build-up time is less than 10 ps FWHM, a value which can be considered negligible with respect to the typical overall jitter (about 35 ps FWHM, note that jitter contributions add quadratically).

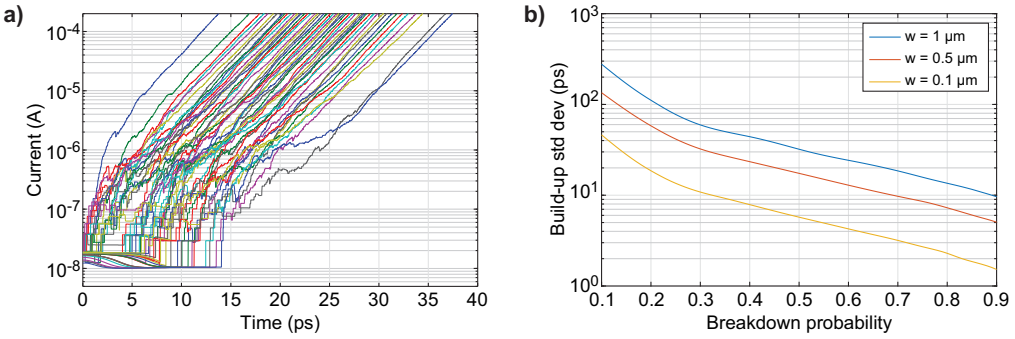

Fig. 6. Simulation of the avalanche current build-up. a) Avalanche current growth as simulated with the RPL Montecarlo model reported by Ingargiola et al. [82]. The curves correspond to 50 different carrier injection events in custom-technology thin SPADs. b) Build-up standard deviation simulated as a function of the electron breakdown probability (Tan et al. [79]). Values are reported for three silicon SPADs having uniform electric field and different depletion region width w.

Spinelli et al. have also considered at first the possibility to introduce ionization coefficients into their model that follow the non-local (also called microscopic) approach [76] originally proposed by Okuto et al. [77,78]. Indeed, in the local approximation these coefficients are considered a function only of the local electric field. Conversely, the non-local model takes into account also the dead space, namely the distance that a carrier must travel after each ionizing event in order to gain enough energy to impact ionize again. In this scenario, the ionization probability distribution is zero immediately after each ionization event and, after a distance equal to the dead space, it promptly rises to a constant value determined by the local electric field. This effect is especially marked when the SPAD under investigation features a thin region in which all the ionization events are concentrated, as in the case considered by Spinelli et al., in which the SPAD was characterized by a depletion region thinner than 1 μm. However, they evaluated this effect and, after having reported a shift in the simulated breakdown voltage, they realized that the local approximation works accurately in a context in which comparisons are made at the same relative excess bias. Later, Tan et al. [79] questioned this modus operandi, as they argued that a fair comparison should be performed at constant breakdown probability, since this quantity plays an important role also in determining other figures of merit like the PDE of the SPAD. Following this approach, they implemented another RPL Montecarlo simulator, which they used to theoretically study the statistical distribution of the avalanche current waveform and the breakdown time, defined as the time interval between the first carrier injection and the detection threshold being reached. The standard deviation of the breakdown time, which quantifies the build-up time dispersion, is studied as a function of different parameters of interest. Firstly, Tan et al. showed that the build-up time jitter decreases as the ratio between the ionization coefficients $k = \beta/\alpha$ increases (i.e. it approaches 1), which is consistent with an enhanced carrier feedback during the multiplication process. Secondly, the build-up dispersion was also predicted to decrease as the dead space decreases, demonstrating the non-negligible detrimental effect that the non-local ionization coefficients have on the avalanche dynamics, no matter the value of k. Lastly, they studied SPADs in which the multiplication regions are made either of silicon or InP. They reported the simulated build-up time dispersion for SPAD structures featuring a different width w for the multiplication region, demonstrating the superior timing performance of

Optics EXPRESS

thin detectors, regardless of the adopted material. The results obtained by Tan et al. [79] for silicon are reported in Figure 6(b). This was not actually a trivial result to predict, since thin junctions typically have higher electric fields and ionization coefficient ratios, which are expected to favor the build-up jitter, but, on the other hand, they are also affected by a more significant weight of the dead space, which is expected to worsen the build-up jitter. Similar conclusions were reported in parallel also by another research group [80,81], which was studying the SPAD avalanche dynamics in GaAs with a similar RPL model.

Even though the simulator implemented by Tan et al. has proven to be crucial to understand many aspects of the avalanche dynamics and its effect on the build-up time dispersion, this model cannot be straightforwardly applied to cases of practical interest due to the assumption of having a constant electric field over the whole multiplication region. A more realistic RPL Montecarlo model was later proposed by Ingargiola et al. [82] and applied to the current generation of custom-technology thin SPADs. Different from the devices considered by Spinelli et al., these SPADs feature an electric field profile suitably engineered in order to increase the PDE and lower the DCR. Ingargiola et al. developed the model by taking into account all the theoretical details considered by Tan et al. and by considering also the actual SPAD electric field profile as calculated from doping profile measurements. Thanks to this model, they managed to demonstrate that the build-up time dispersion in engineered custom-technology thin SPADs is not negligible with respect to the overall jitter, being about 20 ps FWHM for an excess bias of 5 V. The only flaw of the model is the need for a preliminary calibration procedure used by the authors to fit the breakdown voltage of the SPADs by suitably tuning some parameters related to the ionization coefficients. This procedure was necessary for both of the parameter sets they adopted from the existing literature [77,83].

In conclusion, none of the presented RPL models provide sufficient evidence of accurately predicting the avalanche build-up dynamics of a SPAD. For this reason, a full-band (FB) Montecarlo method was also proposed and applied to SPAD structures by Dolgos et al. [84,85]. Differently from RPL simulators, FB algorithms aim at solving the Boltzmann's transport equation in the semiclassical approximation with the most comprehensive models for the band structure of the material and the scattering phenomena in which the carriers are involved (impact ionization included). This approach also allows the second-order effects to be added, such as the average velocity overshoot that a carrier witnesses in case of an ionization event right after the end of the dead space [86]. Due to the high computational burden that the FB models typically involve, Petticrew et al. [87] proposed a simplified approach, with far less details about the band structure, but still including the main scattering mechanisms evaluated on a femtosecond timescale. Both FB and simpler simulators have extensively been benchmarked with experimental data either on impact ionization [84,85,88] or by considering typical figures of merit of avalanche photodiodes (APDs) operating in the linear regime (i.e. biased at voltages lower than the breakdown voltage) [89–91]. However, a convincing validation of these models for the case of SPADs is still missing in the literature.

2.2.4. Avalanche lateral propagation

As we have seen, carrier collection and avalanche build-up play an important role in determining both the shape and the statistical dispersion of the SPAD temporal response. However, the contributions coming from these processes are not able to explain by themselves neither the values of the IRF FWHM nor its dependence on the detection threshold in different SPAD structures. A final contribution is currently missing from our picture, which is the one due to the propagation of the avalanche in the transversal direction until the current is evenly distributed across the whole active area. As we will see, this process originates from two different physical mechanisms. The first cause is related to the multiplication-assisted diffusion of hot carriers (see Figure 2(b) (IV)). As the current density increases, the hot carriers diffuse laterally and

Optics EXPRESS

trigger the avalanche in new regions of the SPAD. In the newly activated regions the avalanche build-up takes place and, due to the statistical fluctuations of the impact ionization process, one region of the propagation front can grow faster than another one, resulting in a dispersion of the propagation time T_{prop}. In addition, even neglecting the fluctuations of the diffusion process, an avalanche triggered in the center of the active area grows faster than an avalanche triggered near the border due to the different lengths to be covered by the avalanche spread. The second cause of the avalanche propagation is instead related to the photons emitted by hot carrier relaxation and reabsorbed in other regions of the active volume which are still quiescent (see Figure 2(b)(V)). In this condition the dependence of avalanche current rise time on the seed position is typically less relevant, but the statistical fluctuations in the number of emitted/reabsorbed photons is at the origin of additional dispersion for the propagation time T_{prop}. The avalanche lateral propagation, either by hot carriers diffusion or assisted by secondary photons, is the main cause of the dependence of the timing jitter on the detection threshold shown by certain SPADs.

Lacaita et al. [92] were the first to study the lateral propagation of the avalanche current in SPAD detectors. They studied the avalanche propagation by considering the old generation of custom-technology thin SPADs and, in particular, they chose to consider a highly rectangular geometry (active area 140×14 μm^2) and measure the dependence of the current leading edge on the point where the multiplication process is triggered. The measurements were performed with the laser spot focused in different points of the active area and with variable excess bias. The result was that the time between the beginning of the rise of the avalanche current and its saturation was proportional to the distance between the seed point and the farthest end of the rectangular area. This experiment suggested that the avalanche propagation in these devices is mainly assisted by diffusion of hot carriers and that the statistical dispersion of T_{prop} is limited by the fact that photons can be absorbed at random positions, resulting in a tradeoff between timing jitter and dimension of the active area. Subsequently, Lacaita and coworkers confirmed this conclusion also by considering circular SPADs belonging to the same family, with a thorough study of their behaviour as a function of active area diameter (8 - 22 μm), operating temperature and overbias, in which they found that the FWHM of the temporal response of larger devices can be lowered to the values obtained for smaller devices by focusing the light in a spot of comparable dimension [93]. Of course, these results are highly dependent on the SPAD structure and, indeed, when the same authors considered reach-through SPADs (RT-SPADs) [8], which feature large diameter (up to 500 μm) and deep depletion region (up to tens of micrometers), they had to face completely different physics [94]. In RT-SPADs they observed that the current leading edge does not change significantly when moving the laser spot from the center to the edge of the active area and, as a result, also when employing diffused light. What is instead observed is that, when focusing the light on the edges, the avalanches occur later than the case in which light is focused in the center of the SPAD. The authors explained this phenomenon by resorting to the photon-assisted propagation [95]: since photons having energy comparable to the silicon band gap can be absorbed even hundreds of micrometers away from the seed point, they could be particularly effective in triggering secondary avalanches in RT-SPADs. As both qualitatively and quantitatively demonstrated by Lacaita et al., the statistical fluctuations in the number of emitted photons are the cause of the main contribution to the randomness in the avalanche current onset and, as a result, in the RT-SPAD timing jitter.

From a modeling point of view, Lacaita et al. decided to focus on thin SPADs. For these devices, they proposed a model based on two continuity equations, one for the electrons and one for the holes, in which they considered for the longitudinal direction (i.e. the x direction) only carrier drift at saturated velocity, while they considered for the transverse directions (i.e. the yz plane) only carrier diffusion [96]. Mathematically speaking, the electron equation was formulated as follows:

$$\frac{dn}{dt} = -v_{sat}\frac{dn}{dx} + D_n\Delta_t n + \frac{n}{\tau_n} + \frac{p}{\tau_p}, \tag{2}$$

Optics EXPRESS

where D_n is the transverse diffusion coefficient, Δ_t is the transverse Laplace operator, τ_n is the build-up time constant for the electrons and τ_p is the build-up time constant for the holes (both computable through the methods already shown in the previous paragraph). The continuity equations were coupled to the Poisson equation in order to correctly include also the space charge effects. However, the authors never managed to apply this model to realistic devices due to the high computational complexity required by a system of three fully coupled equations in three dimensions. As a result, the authors eventually applied to their SPADs only a simpler model based on an ensemble of elementary diodes connected in parallel and coupled only through diffusion terms, which they used to simulate the avalanche current waveform and to confirm that the timing jitter in these SPADs is set by the randomness of photon absorption position. In a later work, they also evaluated the effect of the avalanches generated by secondary photons, but they concluded that this phenomenon could be neglected for predicting both the current leading edge of their SPADs and the FWHM of their IRF [69].

As already mentioned in Section 2.2.3, the electric field profile of the custom-technology thin SPADs has most recently been engineered such as to decrease the DCR, enhance the PDE and allow the practical exploitation of larger devices (up to a diameter as high as 200 µm) [48]. However, the new devices are characterized by a strong dependence of the achievable timing jitter on the detection threshold, that was not present in the former generation (Figure 7(a)) and that could be particularly troublesome when a SPAD-based multichannel timing system is designed, as it will be discussed in Section 3. Therefore, Assanelli et al. [97] started to investigate whether the avalanche triggering position was still the main cause of jitter and whether this phenomenon was able to explain also the strong dependence on the detection threshold. With this goal, they developed an experimental setup in which it was possible to focus a laser beam on the active area of the SPAD with a spot diameter of about 1 µm (negligible with respect to the SPAD diameter, which was 50 µm) at an arbitrary position. The final result is showed in Figure 7(a) and it can be summarized focusing on two main aspects: firstly, regardless of the spot position, the timing jitter showed the same strong dependence on the detection threshold; secondly, at high thresholds, moving the spot toward the border of the SPAD causes a detrimental effect on the timing jitter. Both these results suggested that the photon absorption position could not be the only mechanism behind the statistical uncertainty of the lateral propagation process of the new SPADs. Based on the fact that, at high threshold, the timing jitter improves as the laser spot is moved toward the center of the SPAD and on the observation that an avalanche growing faster also has a lower jitter, they defined a figure of merit (FoM) proportional to the time derivative of the current waveform, showing that this figure was strongly linked to the avalanche build-up time constant τ and to the specific resistive parasitic R_{spec} (i.e. measured in $\Omega\,\mu m^2$) of the SPAD. Mathematically speaking:

$$FoM = \frac{1}{R_{spec}\sqrt{\tau}}. \tag{3}$$

They demonstrated that the considered FoM was able to predict the dependence of the timing jitter on the detection threshold and, therefore, suggested that, in order to produce higher performing devices, an avalanche growing faster and the optimization of the specific resistance are key factors. In particular, thanks to non-invasive electroluminescence measurements [98], they also confirmed the importance of the space charge resistance in determining the steepness of the current leading edge right after the avalanche ignition and, in the later moments, the importance of the bulk resistance due to the quasi-neutral layers adjacent to the depletion region. These results pushed the same authors toward the development of a complete simulator able to integrate Eq. (2) in order to simulate the multiplication-assisted diffusion of the avalanche current for any triggering position within the SPAD active area. The simulator was integrated with a complete 3D model for taking into account also the photon-assisted avalanche spread [82]. The quenching effect of the internal distributed bulk resistance and the filtering action of the

Optics EXPRESS

external readout circuit were considered as well. The most important result of this work was that, contrary to what happens for past devices, taking into account the photon-assisted propagation is fundamental to reproduce the dependence of the timing jitter on the detection threshold of engineered custom-technology SPADs. This result was ascribed to the increased active volume, which is from 10 to 100 times bigger than in the previous detector generation. Last, but not least, the simulator highlighted that having a reduced space charge resistance results in a strong beneficial effect also on the jitter coming from the photon-assisted avalanche propagation.

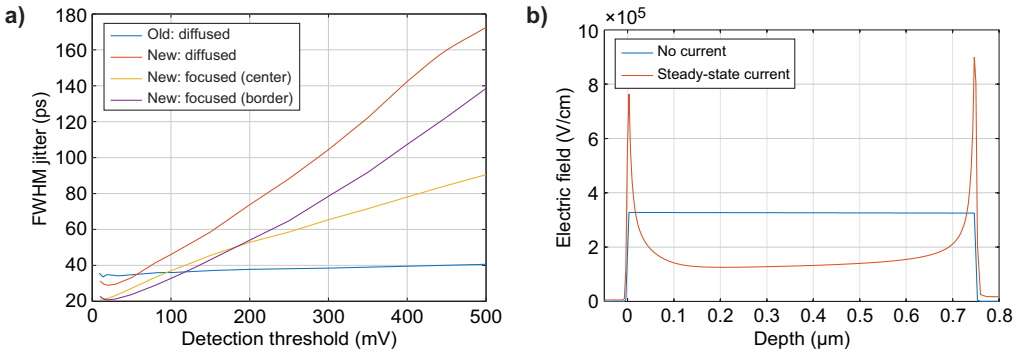

Fig. 7. Avalanche lateral propagation in custom thin SPADs. a) FWHM timing jitter as a function of the detection threshold: old generation and current generation are compared and, in addition, the latter are reported in different illumination conditions [97]. b) Electric field profile for a p-i-n SPAD [99]. The simulation is reported for the conditions in which no current is flowing and in which a steady-state current is present [100].

SPADs based on p-i-n structure From Eq. (3) it is clear that the timing jitter associated to the lateral propagation process is affected by the series parasitic resistance (either related to space charge effects or to the quasi-neutral regions) and the current build-up time constant. The latter contributes to the FoM only with its square root and therefore, as a first step, it can be neglected. The bulk resistance associated to the quasi-neutral regions comes into play only when the avalanche current value is high and, therefore, has a minor role with respect to the space charge resistance. All these hints pushed Assanelli et al. to focus their attention on the space charge resistance and on the possible ways to reduce it. To this aim, they proposed the exploitation of a p-i-n structure to develop a SPAD with a superior timing performance [99]. Indeed, such structure exhibits a negative space charge resistance [100] that, as the avalanche current grows, results in a progressive increase of the electric field at the sides of the depletion region and in a decrease of the electric field in the central part (Figure 7(b)). The final result is a positive feedback allowing the avalanche current to grow dramatically faster with respect to a simple pn junction. Simulations by Assanelli and coworkers showed that a p-i-n SPAD can exhibit an extremely favorable ratio between the PDE and the peak electric field, thus opening the way to timing-enhanced structures that do not sacrifice other crucial performance as PDE and DCR [99]. The intuition of Assanelli et al. has been confirmed experimentally by Gramuglia and coworkers, who independently designed a CMOS SPAD (Figure 4(b)) able to combine a peak PDE as high as 55% at 480 nm (and still 12% at 800 nm) with DCR down to 100 cps at room temperature and a timing jitter as low as 12.1 ps FWHM [16] (later even reduced to less than 10 ps with a slightly higher overvoltage and an optimized readout electronics [66]).

3. Front end electronics: quenching and readout

The thorough review of SPAD fundamental physics and processes involved in the generation and propagation of the avalanche current reported in Section 2. highlighted the parameters and

Optics EXPRESS

structures that set the ultimate limit to the timing performance that can be obtained with this kind of single-photon detector. In this section we will discuss what is needed to operate and exploit the SPAD detector at its best for timing purposes. We will see that the timing performance that can be obtained from a SPAD not only depends on both the detector and the front end electronics, but that a key role is also played by their specific interaction. For this reason, we will see that timing optimization can easily require a front end design deeply-rooted on the knowledge of the selected SPAD features.

3.1. Quenching circuit

The proper operation of a SPAD requires a circuit that is promptly activated by the avalanche current (*sense* phase), stops such current (*quenching* phase), and finally restores the initial bias conditions that make the SPAD ready to detect another photon (*reset* phase). Sometimes, such circuit is also able to keep the detector off for a certain amount of time: this additional phase is usually referred to as *hold-off*. A circuit able to carry out all these tasks is called quenching circuit and its characteristics strongly affect the performance that can be obtained with a SPAD. Particularly relevant is the *dead time*, that is the overall time needed to perform all the phases and whose inverse sets the maximum SPAD count rate.

Quenching strategies have been classified by Cova et al. in 1996 [23] into three main classes: (i) passive, (ii) active, and (iii) a mixed one. The simplest solution consists in placing a high-value resistor (on the order of 100 kΩ) in series to the SPAD, resulting in a passive quenching circuit (PQC). Along with its simplicity, this approach provides immediate activation of the quenching phase with beneficial effects on both power consumption and afterpulsing probability. On the contrary, the high resistor value results into a slow reset phase, causing a variation of all parameters that depend on the overvoltage [8] during all this interval as already mentioned in Section 2.1. Concerning timing, avalanche build-up and propagation processes are slower than in nominal conditions because of the lower overvoltage, thus resulting in a slower rising edge of the avalanche current. In this case, a time-shifted crossing of the current threshold may occur, causing an altered recording of the photon time of arrival. If the number of these events is statistically relevant in the experiment, the timing measurement is distorted. Moreover, passive quenching intrinsically features zero hold-off time, meaning that a SPAD terminal immediately starts its recharging as soon as the current goes to zero. In this case, the dead time theoretically coincides with the quenching time. However, PQC-based systems may suffer from loss of events occurring during the reset phase as their lower amplitude with respect to photons impinging in nominal bias conditions may not be able to cross the system detection threshold. In this scenario, some events are not registered but they trigger another dead time of the detector, resulting into a variable duration of the effective dead time that impairs dead time correction techniques in single photon counting applications [101], and completely prevents the exploitation of a recently-proposed method to achieve high-speed without distortion in time correlated single photon counting (TCSPC) that is based on the matching of the detector dead time with the period of the laser (the interested reader is referred to [102] for a detailed explanation of the method). For all these reasons, passive quenching is not well-suited to achieve a controlled and short dead time, which is paramount to achieve high count rate in timing applications as will be discussed in Section 4.

Conversely, the active quenching strategy as classified in [23] consists in a fully-active behavior: a fast comparator is used to sense the avalanche current and react on the SPAD bias voltage bringing it to the breakdown level or below. In this case, the dead time is constant and well defined, but a large current flows in the detector for all the time needed by the feedback loop to sense the current and activate the quenching circuitry, with unnecessary and detrimental effects on afterpulsing probability and power consumption. For this reason, the fully-active approach is intrinsically outperformed by the mixed passive-active quenching circuit, that includes an

Optics EXPRESS

equivalent resistor for prompt avalanche quenching already in the sense phase. Due to its active behavior throughout all the phases except for the very beginning of the avalanche management, the mixed active-quenching circuit has been historically simply referred to as active quenching circuit (AQC) without any risk of misunderstanding as the fully-active approach is rarely used.

Since the AQC is activated by the arrival of a photon and produces a synchronous output pulse marking this event, in principle the timing information could be extracted from its sense stage. However, the requirements of these two tasks (sense and timing) are quite different. On one hand, as discussed above, the sense stage should initiate the quenching process, which means that it should provide a relatively high impedance ($> 1\,\mathrm{k}\Omega$) to cause a significant voltage drop across the SPAD terminals; on the other hand, the timing circuit is used to read the avalanche current signal so its equivalent input impedance should be as low as possible to minimize the loss of current on alternative parasitic paths. Moreover, the sense stage is part of a circuit that must be able to provide quenching-reset pulses ranging from a few volts to tens of volts depending on the optimal SPAD excess bias (see [8] for more details), while the other one should feature a large bandwidth to preserve the avalanche current edge. However, high-voltage (HV) transistors are typically much slower than their low-voltage (LV) counterparts; in turn, LV transistors are quite fragile making the coexistence with HV circuits anything but trivial.

Timing measurements have been carried out with AQCs featuring a relatively high impedance of the sense stage obtaining a precision ranging from several tens of picoseconds [18,103,104] with thin SPADs (both CMOS and custom ones) to hundreds of picoseconds with thick RT-SPADs [105] and RE-SPADs [106]. High timing jitter obtained with AQCs is due neither to the detector structure itself nor to the circuit performance (e.g. its electronic-noise jitter), but in the mutual SPAD-AQC interaction that determines the avalanche current readout, as will be thoroughly explained in the next section.

3.2. Current pick-up circuit

The combination of various factors in SPADs produce a variability in the avalanche current waveforms as thoroughly explained in Section 2.

An easy and interesting way to directly visualize the temporal behavior of the SPAD current consists in the simple setup of Figure 8(a): in this case, a custom-technology thin SPAD [8] has been chosen to be operated with a simple PQC ($R_q = 100\,\mathrm{k}\Omega$) on the anode side and a $50\,\Omega$ resistor (R_s) on the cathode side. The avalanche current converted into a voltage signal by the low-value resistor is readout by means of a 4 GHz-bandwidth oscilloscope (Tektronix TDS7000 series), thus minimizing the filtering effect on the fast SPAD signal. It is worth saying that in the real setup the $50\,\Omega$ resistor is actually substituted with a coaxial cable directly connected to the $50\,\Omega$ input of the oscilloscope. A pulsed laser completes the setup: the light is attenuated to the single-photon level and it is directed toward the SPAD, while an electrical signal synchronous with the optical one is used as trigger input of the oscilloscope. In this way, only the avalanches triggered by the laser pulse are recorded while spurious dark counts are discarded. The graph reported in Figure 8(b) reproduces the oscilloscope display on which many curves developing across the $50\,\Omega$ resistor are superimposed. This experiment confirms that the SPAD avalanche current signal is affected by a spread, and, even more, it provides the key guideline to minimize jitter in timing measurements. Indeed, the current variability increases with the signal amplitude making a low-threshold sensing the solution of choice to extract the best timing performance from a SPAD. To this aim, Cova et al. in 2002 [107] patented a current pick up circuit consisting of a CR network followed by a fast comparator as the one shown in Figure 9(a). Using this kind of circuit, Gulinatti et al. demonstrated a jitter as low as 35 ps with a 100 µm-diameter thin SPAD (breakdown voltage: 24 V) for the first time in 2005 [51]. Moreover, this circuit has been proven effective to improve the timing performance of SPADs in several cases. For example, Rech et al. optimized the timing of a commercial module [108], while the timing jitter difference (about 60

Optics EXPRESS

ps vs less than 30 ps) between the work of Severini et al. [18] and the one of Sanzaro et al. [38] with the same SPAD (fabricated with a 0.16 µm BCD technology) is completely attributable to the exploitation of such a circuit in the second case.

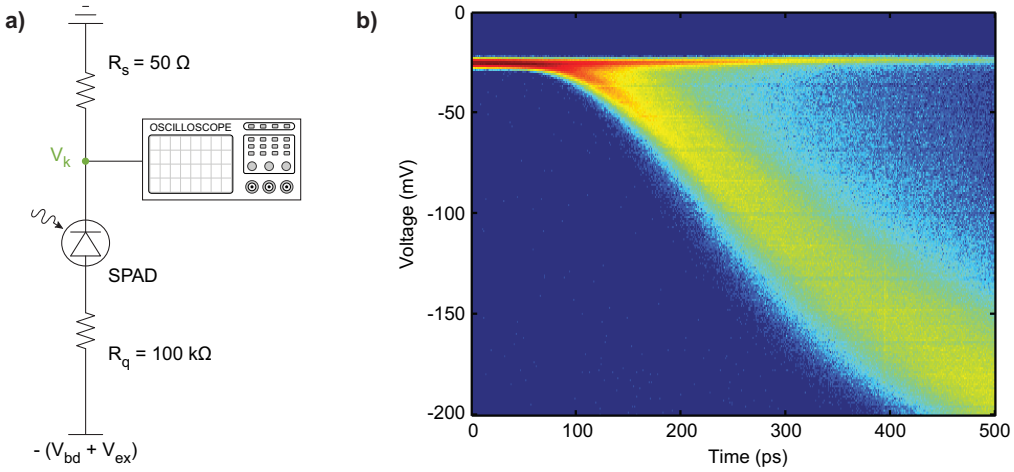

Fig. 8. Avalanche current direct measurement. a) Custom-technology thin SPAD operated with a PQC on the anode side and a 50 Ω resistor on the anode side. A 4 GHz oscilloscope reads out the cathode voltage V_k. b) Superposition of SPAD cathode waveforms showing the avalanche current growth transduced by the 50 Ω resistor.

From a front end designer perspective, it is worth highlighting that while low-threshold current sensing (at about 100 µA) is always effective, it is not always necessary. Two main aspects should be considered when designing a timing system architecture based on SPADs.

On one hand, the output of the pick-up circuit must be fed to a time measurement circuit to transform the photon time-of-arrival information carried by a leading edge into digital data suitable to be stored and/or processed by an elaboration unit (e.g. a field programmable gate array, FPGA). Such circuits can sometimes set the bottleneck to the timing performance of the system. In these cases, the readout of the avalanche current is typically simply made by the quenching circuit, having the advantage of minimizing the amount of front end circuitry to be placed around each detector.

On the other hand, the structure and the operating conditions of the selected detector directly determine the amount of avalanche current spread (as well as other crucial parameters such as PDE and DCR [8]). For example, SPADs with a high electric field peak and a thin depleted region, thus characterized by a low breakdown voltage (e.g. < 12 V), feature a low spread of the avalanche current waveforms thus providing good timing performance even at a relatively high sensing threshold [109]. However, this is not the case for all SPAD structures, as already explained in Section 2.2.4. A work by Ghioni et al. [110] highlights these aspects by comparing the timing performance of thin custom-technology SPADs differing in diameter, operating excess bias voltage and electric field shape all readout by the same pick-up circuit of Figure 9(a). Results are reported in Figure 9(b)-d. In the graph of Figure 9(b) it can be noted, for example, that a timing precision of 50 ps FWHM could be achieved with a voltage threshold higher than 200 mV with a 20 µm-diameter SPAD, while to achieve the same performance with a 100 µm-diameter SPAD a threshold lower than 50 mV was necessary. Analogously, Figure 9(c) shows that a thin SPAD operated at the most suitable excess bias voltage $V_{ex} = 5$V could achieve a timing precision better than 50 ps FWHM provided a threshold lower than 30 mV is used, while increasing the overvoltage up to 10 V (causing an increment of DCR) makes it possible to achieve the same timing performance with a remarkably higher threshold (about 180 mV). Finally, Figure 9(c)

Optics EXPRESS

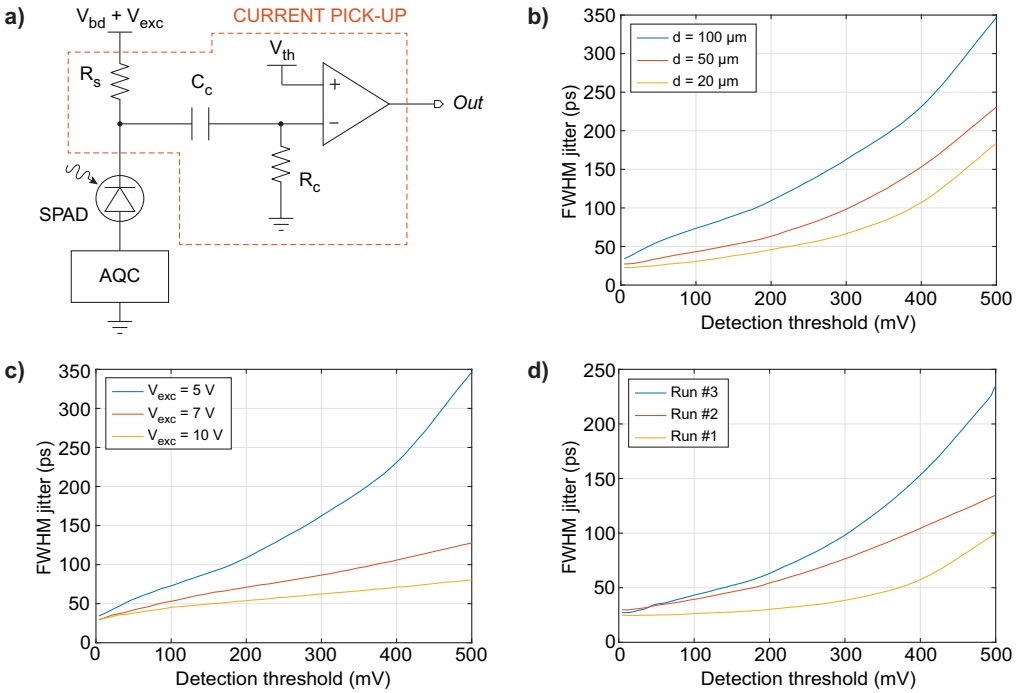

Fig. 9. Custom thin SPAD timing jitter measurement as a function of detection threshold. a) Schematic of the SPAD current pick-up circuit patented by Cova et al. [107] and exploited in this measurement. b) FWHM timing jitter of three SPADs with diameter from 20 µm to 100 µm. c) FWHM timing jitter of a 50 µm SPADs operated at V_{ex} ranging from 5 V to 10 V. d) FWHM timing jitter of SPADs with sharp (RUN#1), mild (RUN#2) and smooth (RUN #3) electric field peak.

compares three fabrication runs differing in the electric field shape. Going from RUN#1 to RUN#3 the electric field has been engineered to reduce the DCR, resulting also into a stronger dependence of the timing performance from the avalanche-current sensing threshold. As can be seen, with a sharp electric field profile (RUN#1), a timing precision better than 40 ps can be achieved with a threshold as high as 300 mV, while the same result with a smoothed electric field profile (RUN#3) can be achieved only with a threshold lower than 60 mV. For these reasons, an accurate knowledge of the selected SPAD features should be gained before opting for a low-threshold sensing strategy, because it adds complexity to the electronics and system design, but it may be ineffective or not necessary. When pursuing low-threshold sensing, parasitics due to the detector, the electronics and their interconnection can play a prominent role in reducing the amount of current that actually flows into the pick-up circuit, thus increasing the equivalent current threshold. To analyze this aspect we can refer to the simplified electrical model of Figure 10(a) where $C_{s,det}$ represents the parasitic capacitance at the detector terminal, L_b is the inductive load of the interconnection (e.g. wire bonding) in case of detector and electronics being fabricated on separate dies, and $C_{s,ele}$ and $R_{in,ele}$ are the input capacitance and resistance of the front end, respectively. The scheme has been reported for a pick-up circuit placed at the anode terminal of the SPAD side with the quenching circuit connected to the cathode side, but the same considerations hold for the opposite case. From the scheme it is evident that the three unwanted elements, i.e. $C_{s,det}$, L_b and $C_{s,ele}$, must be minimized to maximize the current flowing into the pick-up circuit. At the same time, the input resistance of this circuit should be made low with respect to the parasitics. We will now discuss solutions and design guidelines for the detector

Optics EXPRESS

and the pick-up circuit, but before going into those details, it is worth making a comment on L_b. Whenever SPADs and front end electronics are integrated in the same silicon die L_b is obviously negligible, while in case of two separate chips (typically in case of different technologies for detector and front end optimization) they should be placed as close as possible to allow a short and direct interconnection between the two.

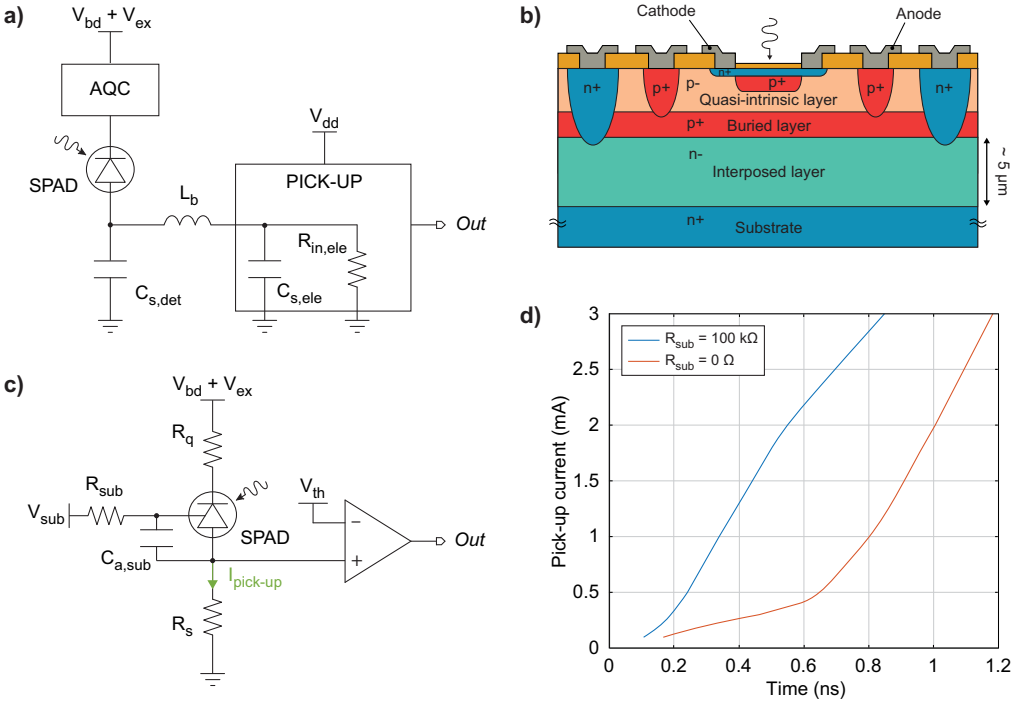

Fig. 10. The interconnection between SPADs and electronics involves parasitics that may affect the timing measurement. a) Simplified schematic of a SPAD connected to an external pick-up circuit with main parasitics explicitly reported. b) Structure of a custom technology thin SPAD with an extra epitaxial layer (interposed layer) introduced to reduce the anode-substrate capacitance. c) Substrate bias scheme (semi-floating configuration) to minimize the impact of the anode-substrate capacitance for timing measurements with a single SPAD. d) Waveform of the avalanche current flowing into the pick-up resistor indirectly measured with the setup shown in c). The orange curve has been obtained with R_{sub} = 0Ω, i.e. the substrate voltage is fixed, while the blue one has been measured using the semi-floating-substrate configuration resulting from R_{sub} = 100kΩ. Both curves have been obtained in two steps: (i) the switching time of the comparator at different threshold has been measured, (ii) the voltage threshold has been divided by the R_s-value to derive the pick-up current $I_{pick-up}$.

SPAD stray capacitance The structure of the SPAD intrinsically features various parasitic capacitances. Above all, the SPAD is a pn junction thus a junction capacitance between its two main terminals is always present. However, this is typically in the order of few hundreds of femtofarads or lower and it is equally experienced by both the anode and cathode. More critical is the substrate of the device. Silicon SPADs fabricated with a planar process typically lay on top of a lightly-doped silicon substrate. As an example, we can refer to the structure of Figure 4(b), where a p-over-n SPAD is fabricated on top of a p substrate. In this case, on the cathode side there is a stray capacitance due to the cathode-substrate reverse biased junction whose diameter is larger than that of the SPAD. On the contrary, at the anode side the stray capacitance is smaller

Optics EXPRESS

because it is only due to the metal-oxide-semiconductor stack associated with the metal path that is used to contact the anode terminal. As a result, the capacitive load at the cathode side is higher than the one at the anode side.

An effective solution to reduce the stray SPAD-substrate capacitance has been proposed by Labanca et al. in 2018 [111]. Starting from a custom technology thin SPAD structure as the one reported in Figure 5(a), featuring a double epitaxial structure as reported in [48], an extra epitaxial layer, called the interposed layer, has been added to the substrate leading to the structure reported in Figure 10(b). This solution increases the space charge region of the anode-substrate junction with a consequent reduction of the associated stray capacitance. Incidentally, the additional epitaxial layer is also effective for increasing the breakdown voltage of the substrate junction, which is in some cases required to operate the SPAD at higher excess bias [12]. Obviously, the exploitation of this approach requires some control on the fabrication steps of the device, so it cannot be exploited in standard CMOS processes.

All these aspects make the SPAD terminals easily unbalanced in terms of their capacitive load. Consequently, an optimized front end design must start from a deep understanding of the structure of the SPAD of choice. Ideally, the avalanche current pick-up circuit should be placed at the less capacitive terminal, but this may not be always possible. For example, speed maximization would also require placing the AQC at the less capacitive SPAD terminal, thus posing a speed vs precision tradeoff whenever AQC and pick-up circuit are two separated circuits to be placed at the two opposite terminals of the SPAD. A possible workaround to the SPAD stray capacitance issue consists in biasing the detector substrate by means of a high-value series resistor, i.e by using R_{sub} in the order of 100 kΩ in a circuit configuration like the one shown in Figure 10(c). In this way, the impedance of the substrate path can be increased by the resistor value and its impact on the avalanche current divider can be minimized. Figure 10(d) compares the experimental measurement of the avalanche current of a custom thin SPAD read out by a low-value resistor (R_s = 100Ω in Figure 10(c)) followed by a comparator, when the substrate of the chip is connected to the bias voltage V_{sub} by means of two different resistors. The orange curve corresponds to the exploitation of a null substrate resistor (R_{sub} = 0Ω), meaning that the substrate is fixed at V_{sub}, while the blue curve has been obtained with R_{sub} = 100kΩ corresponding to a semi-floating substrate configuration. As anticipated above, here it can be observed how the exploitation of a high-value resistor in series to the substrate remarkably increases the amount of current flowing into the pick-up circuit with respect to the fixed substrate configuration. Unfortunately, having the SPAD substrate almost electrically floating is a viable solution only for single detectors. In SPAD arrays the substrate is shared among all the pixels and leaving it floating would create a path between adjacent detectors easily leading to electrical crosstalk especially when low-threshold operation is pursued.

Pick-up circuit input impedance The ideal current readout circuit features an input impedance that is much lower than the one of any other available path in order to maximize the current collection. With SPADs, this requirement is further complicated by the speed of the avalanche current rising edge, typically lower than a few hundreds of picoseconds, demanding an extremely high input bandwidth as well. The simplest solution consists of a resistor that directly translates the current into a voltage. In the measurement setup of Figure 8(a), for example, a 50 Ω resistor was used to maximize the current collection. However, when it comes to timing measurements it has to be taken into account that a threshold is necessary to detect the avalanche event and thus activate the time-measurement electronics. In this scenario, the exploitation of a low-value resistance implicates a low voltage threshold to achieve an equivalent low-current measurement. This is the case of the work reported by Gulinatti et al. [51], for example, where the lowest jitter with a 50 µm-diameter SPAD has been achieved with a voltage threshold as low as 8 mV. As previously discussed, in that work Gulinatti and coworkers exploited the pick-up circuit patented by Cova et al. [107]. The patent guidelines recommend a pick-up resistor of few hundreds of

Optics EXPRESS

ohms (R_s<500Ω), strongly depending on the SPAD current profile, surrounding parasitics and their bias conditions as discussed throughout the previous paragraph.

One possibility to break the low-impedance vs low-voltage-threshold tradeoff is the transimpedance amplifier (TIA). The classic configuration based on an operational amplifier [112] may not be suitable for SPAD front end due to the need of an extremely large bandwidth and the necessity of keeping the size and power consumption of the front end circuit as low as possible for scalability toward SPAD arrays. For all these reasons, Crotti et al. presented a current pick-up circuit based on a TIA structure consisting of a few transistors and passive elements, first made of discrete components [113] and later in a fully-integrated version [114] suitable to be the building block of a multichannel structure. The transimpedance structure provides the necessary degrees of freedom to obtain at the same time a low input impedance and a voltage signal of several hundreds of millivolts. A negative feedback loop is paramount to minimize the input impedance, while a fast comparator with a tunable threshold allows the end user to select the value that optimizes the timing performance of the final system. Indeed, this standalone circuit has been conceived to be suitable for any external SPAD. Following Crotti's preliminary results, Acconcia et al. in 2017 [115] designed a TIA pick-up circuit able to achieve a timing jitter as low as 32 ps with the same SPAD in the same operating conditions of Gulinatti et al. [51], but with a voltage threshold as high as 200 mV, that is 25 times higher than the result obtained with a low-value resistor. This result opened the way to design a multichannel system not only because this solution is realized by a compact fully-integrated circuit (area occupation: 0.011 mm^2), but, even more important, because relatively high-voltage-threshold operation makes any front end much more robust against electrical crosstalk that may occur among pixels of an array.

Impedance of the opposite SPAD terminal with respect to the current pick-up node In this section we have focused so far on the SPAD terminal that is chosen to extract the avalanche current signal and on the guidelines and solutions that maximize the fraction of current that is actually used for timing purposes. However, before closing this subject, it is worth adding one more comment on the impedance that affects the opposite SPAD terminal, e.g. the cathode whenever the pick-up circuit is connected to the anode side, and viceversa. Indeed, the overall amount of current that can come out of the detector depends on the maximum of the impedances that are present at its two terminals. In principle, in order to maximize the current exiting the SPAD, the opposite terminal should feature a zero-impedance path toward the bias voltage [23]. However, it must be considered that either the anode or the cathode must be connected to some sort of quenching circuit to ensure the proper functionality of the sensor. In the next paragraph we will focus on circuits that combine readout and quenching capabilities in a single circuit. Here, we will discuss the case of separated quenching and pick-up circuit connected to the opposite terminals of the SPAD. In this scenario, the quenching requirements (high impedance) seem to be at odds with timing ones (low impedance) on the quenching terminal. A possible solution to break this tradeoff consists in the exploitation of a capacitor in parallel to the circuit on the quenching side. The capacitor can set a low impedance during the fast rising transient of the avalanche current, i.e. when the timing information is extracted. Then, the quenching circuit comes into play and stops the current. To this aim, Acerbi et al. [116], for example, added a metal ring on the cathode side of their SPAD thus providing a "quenching capacitance" that will be in parallel to the external quenching circuit. By doing so, they improved the extraction of the avalanche current, obtaining a better timing jitter with respect to the same structure without the metal ring (24 ps vs 41 ps in the same configuration). Nonetheless, it is worth saying that while timing requirements push toward the maximization of the current flowing in the device, the afterpulsing probability increases with the amount of charge flowing in the device given a set dead time. As a result, a tradeoff between timing precision and afterpulsing may exist especially when the device dead time is pushed down to a few nanoseconds (more details can be found in [8]).

Optics EXPRESS

3.3. AQC with low-current sensing capabilities

Quenching and timing solutions discussed so far require that these two functions are attributed to separate circuits connected to the two opposite terminals of the SPAD. Indeed, high-voltage quenching pulses can easily damage the pick-up circuit if directly applied to fast low-voltage comparators or transistors, as briefly mentioned already in Section 3.1.

Having both quenching and timing functionalities within the same circuit would have undeniable advantages in terms of system design. To better understand this point, it is worth recalling first that the two main SPAD terminals (anode and cathode) intrinsically belong to two voltage domains that differ by the breakdown voltage, ranging from 10 V to 15 V for very sharp-electric-field profile detectors [59,109], from 30 V to 50 V for engineered-electric-field profile ones [17,111,117], and up to more than 250 V for RT-SPADs [105,118]. At the system level this means that if two separate circuits are used, the co-existence of two very different voltage domains on the same circuit board must be ensured. On the contrary, having a single circuit featuring both functionalities would permit having just a single fixed bias voltage on the other terminal. Moreover, the availability of an AQC with low-current sensing capabilities would halve the number of required interconnections between the SPAD and its front end. This aspect is particularly crucial for bi-dimensional SPAD arrays fabricated with either custom technologies that prevent the integration of advanced electronics on the same chip, or a CMOS process but using a dedicated chip only for detectors to achieve a high fill factor. Both of these solutions require external electronics to be connected to the SPAD array, e.g. by means of wire or bump bonding, making a single-terminal front end practically necessary to scale up to hundreds or thousands of channels.

The first attempt to combine some of the quenching and timing requirements on the same chip was made by Tisa et al. in 2008 [119] with the so-called variable load quenching circuit (VLQC). The simplified scheme of this solution is sketched in Figure 11(a): one voltage-controlled transistor (M_a) is used both as the sense and quenching resistor. To this aim, the gate of M_a is initially biased at the positive power supply of the logic block, thus providing a relatively low impedance in the sense phase (about 800 Ω). Then, as soon as the avalanche is triggered, the inverter senses the voltage rising across the transistor and it activates the control logic that lowers the gate voltage of M_a. By doing so, the resistance of M_a is increased thus providing a prompt passive quenching. The potential success of this solution relies in two main factors: (i) the exploitation of a feedback loop to change the configuration of the sense/quenching transistor on the basis of the status of the SPAD, and (ii) the fast response of the loop that can be achieved by using only a few control elements (in principle, just an inverter) and that is crucial to minimize power consumption and afterpulsing probability. However, this circuit also presents two main limitations. On one hand, the exploitation of an inverter to detect the avalanche event is beneficial in terms of bandwidth and area occupation, but it offers no degrees of freedom to set the readout threshold, that may indeed be a key timing parameter as widely discussed in Section 3.2. The reader may think that, at least when the SPAD and the electronics are integrated on the same chip exploiting a CMOS process, the best possible inverter threshold could be set in the design phase and the transistors size could be accordingly tailored. However, we highlighted in Section 2.1 that a SPAD circuit model allowing the accurate prediction of timing performance by simulation has never been reported to date, thus making the sensing threshold a key parameter that should be tuned at system level to minimize jitter. Even worse, in the VLQC architecture the sense transistor is always connected to the SPAD, thus limiting the maximum overvoltage to the highest drain-source voltage that the transistor can tolerate. Early VLQC designs have been made with low voltage CMOS technologies resulting in a maximum overvoltage that could be applied to the SPAD of a few volts. Such operating voltage limitation has been mitigated in the VLQC architecture presented by Lindner et al. [103] (which they called passive quenching active recharge, PQAR) thanks to the exploitation of a cascoded transistor in series to the sense MOSFET. In this way, only a fraction of the overvoltage is applied to the drain-source terminals

Optics EXPRESS

of the sense MOSFET, thus allowing to apply an overvoltage that exceeds the single MOSFET voltage limitations. With this solution Lindner and coworkers were able to apply an overvoltage of 4.4 V to their SPAD using transistors that can reliably tolerate up to 2.75 V. Simulations showed that all nodes comply with safe operating conditions. Starting from this work, Gramuglia et al. exploited two cascoded transistors in series to M_a, resulting in the circuit architecture sketched in Figure 11(b), to enable an excess bias up to 11 V [16]. With this front end integrated along with the SPAD, Gramuglia and coworkers applied an overvoltage of 6.5 V obtaining a timing jitter as low as 7.5 ps FWHM [66], that is the best timing result ever reported with a SPAD so far, combined with a dead time as low as 3 ns and an afterpulsing probability as low as 0.1% [16]. However, VLQC and PQAR are both limited in timing applications by their voltage-based readout of the SPAD terminal: with a low impedance at the sensing node, this solution requires a low voltage threshold whenever a low equivalent current threshold is necessary. The state-of-art result obtained by Gramuglia and coworkers is primarily due to the sharp avalanche current edge of their p-i-n SPAD structure, which probably allowed them to use a relatively high readout voltage threshold (values are not provided in the paper as the structure is fully integrated and the threshold is fixed by design) and still obtain a very low timing jitter. Nonetheless, using a VLQC or a PQAR with other SPADs that require a low-current threshold detection could result in poor timing precision performance (if the threshold is high) or into a very-low voltage threshold that would make the pixel exposed to electrical crosstalk in a SPAD array as discussed in 3.2.

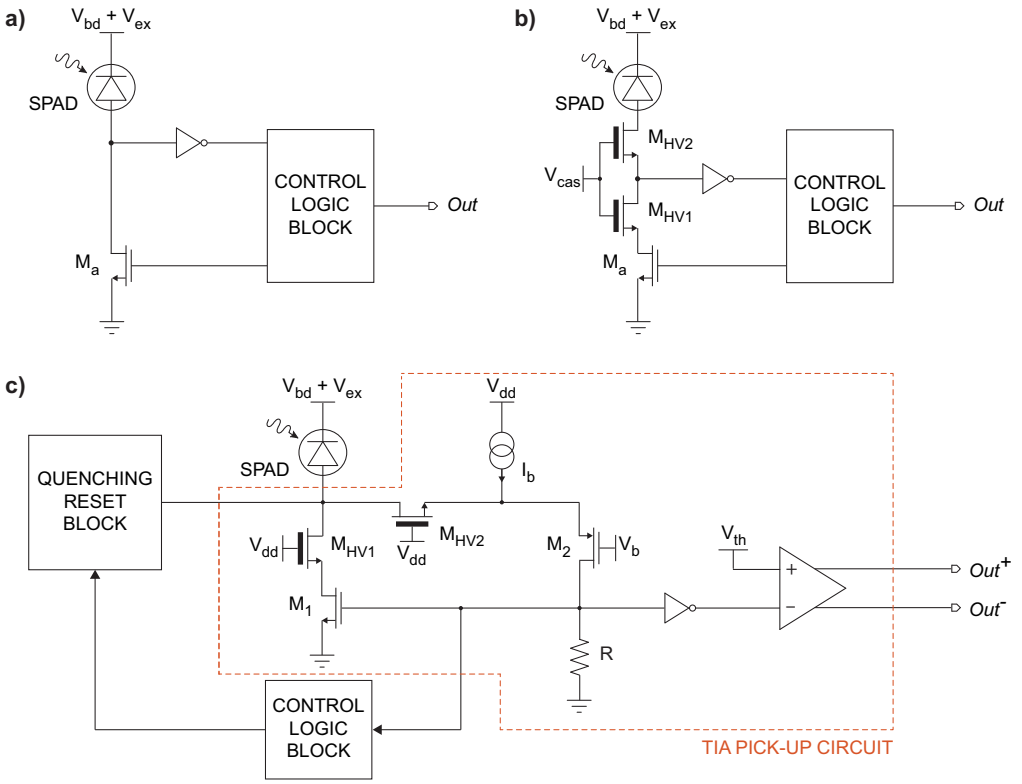

Fig. 11. AQCs with low-input impedance. a) Simplified schematic of the VLQC architecture. b) Simplified schematic of the PQAR architecture with two cascoded transistors in series to the low voltage transistor M_a to increase the AQC voltage capabilities. c) Simplified schematic of the TR-AQC. V_{dd} is 1.8 V, V_b is about 100 mV.

Optics EXPRESS

An alternative versatile solution has been presented by Acconcia et al. in 2017 [120] by combining the features of a high-voltage AQC [104] and a TIA [115] in the so-called time-resolving AQC (TR-AQC). The simplified schematic of the TR-AQC is shown in Figure 11(c). It consists of three main blocks: (i) the quenching/reset block, which is responsible for applying quenching/reset pulses, (ii) the TIA pick-up circuit, that senses the avalanche current, and (iii) the control logic block which regulates the proper behavior of the whole circuit. In this case, the AQC architecture of [104] has been substantially preserved thus providing HV quenching/reset capabilities up to 50 V. To make the coexistence with the AQC possible, the TIA has been modified with respect to [115] by inserting two HV transistors (M_{HV1} and M_{HV2} in Figure 11(c)) that protect the LV part of the circuit. The maximum voltage at the source terminal of M_{HV1} and M_{HV2} is always lower than V_{dd} by some hundreds of millivolts, that is the threshold voltage of these transistors. Above this value, the two HV MOSFETs are off and the LV part of the circuit is no more connected to the SPAD anode. Additional circuitry that is not shown in the figure ensures that no internal node is ever left floating. It is worth highlighting that the two HV MOSFETs have been placed within the TIA feedback loop to minimize their impact on the impedance and bandwidth of the system. With this circuit, Acconcia and coworkers achieved a timing jitter as low as 37 ps with a 80 µm-diameter custom technology thin SPAD with a threshold as high as 100 mV. The higher complexity of this standalone TR-AQC with respect to the VLQC architecture provides a higher degree of flexibility, making it suitable to be used with different types of SPADs requiring an overvoltage up to 50 V, but at the expense of higher power consumption and area occupation.

4. Timing measurement: methods and outlook

The thorough discussion of Sections 2. and 3. pointed out how the ultimate limit to the timing performance that can be obtained with a single SPAD is both due to the device itself and to the ability of sensing the avalanche current when the growth process is still confined to a limited region. Whilst the variety of currently-available SPADs and front end circuits are exploited in a large number of applications, there still is a strong research activity in this field. In particular, two main directions can likely convey most of the future interest and effort: extremely high timing precision and high speed.

Concerning the precision, Sections 2.2.1 and 2.2.2 have extensively described the tradeoff between PDE and timing tail/jitter that is limiting the performance of current SPAD structures. A possible solution is based on the separation of the direction along which the light propagates, which determines the PDE, and the direction along which the electric field is orientated, which affects the timing performance. A device already proposed in the literature exploiting this idea is the waveguide SPAD [8], which has gained particular interest from researchers working on the integration of single-photon detectors within photonic circuits [121,122]. Experimental realizations of such devices are now starting to come up in the literature, although at the moment all of them are based on Ge-on-Si technology for detection of telecom wavelengths such as 1310 [123,124] and 1550 nm [125]. Another promising path is represented by the use of resonant-cavity-enhanced SPADs. These devices are already an experimental reality in silicon [126] and were recently proposed also for the detection of telecom photons [127,128]. However, a resonant detector with DCR and afterpulsing similar to the best SPADs is yet to be demonstrated [129]. Similar considerations about PDE, timing jitter and DCR can be made also for light-trapping SPADs [130,131], i.e. mesa SPADs based on nanograting layers able to scatter the photons and trap them within the active volume of the detector.

A second problem, highlighted in Sections 2.2.3 and 2.2.4, is the lack of a complete understanding of the physics underpinning the avalanche build-up and propagation phenomena. In this regard, modelling will play a crucial role in the future. The works reported by Gramuglia et al. [16,66] have recently demonstrated that the timing performance of silicon SPADs can be

Optics EXPRESS

further improved by suitably engineering the electric field profile. Therefore, it is reasonable to foresee that better simulation tools will allow researchers in this field to set the bar even higher. The lack of accurate models for avalanche build-up and propagation makes it difficult to access what might be the ultimate timing jitter achievable with silicon SPADs. However, it is clear that this value is set by two phenomena: the limited carrier velocity in drift regions and the randomness of impact ionization. Therefore we may speculate that researchers will try to overcome this limit by looking into materials with higher carriers mobility and better ionization coefficients. Limiting the impact of such materials on DCR and afterpulsing probability, as well as their integration in the silicon platform will pose formidable challenges.

Concerning high speed, a short introduction is necessary. Picoseconds timing capabilities of SPADs are of particular interest in measurements based on the TCSPC technique [101]. However, while TCSPC would directly benefit from any improvement in SPADs' timing precision, the same cannot be immediately said for the detector speed. Indeed, the inverse of the SPAD timing precision sets the ultimate limit to the equivalent bandwidth of TCSPC in the acquisition of an optical waveform. On the contrary, the root principles of this technique have historically set an upper limitation to the detector count rate. Referring to the frequency of the laser that is used to produce the signal, traditional TCSPC states that the single photon detector count rate must be limited to a few percent (typically 5%) of the laser rate in order to prevent the so-called pile-up distortion [101]. Considering for example a laser rate of 80 MHz, which is typical in many applications especially in the biological field, the above-mentioned constraint would result in a detector count rate of 4 Mcps. Such value is much lower than the maximum count rate that can be nowadays achieved with a single SPAD [16,18,132]. Given the limitation set by the technique itself, high-speed acquisition in TCSPC has been pursued in an alternative way, i.e. by developing multichannel systems. Following this approach, an overall high count rate can in principle be achieved by using multiple channels in parallel, each one operated within the TCSPC speed constraint. In this scenario, the highest level of integration in terms of number of channels has been achieved by exploiting standard CMOS technologies, allowing the integration of both SPADs and electronics on the same silicon die [133,134]. Nonetheless, the development of densely integrated multichannel systems has highlighted several implementation challenges. Concerning the detector, high yield has been obtained only with planar processes (both custom and standard ones) so far. Similarly, guidelines and solutions to avoid both optical and electrical crosstalk have been extensively studied and reported in the literature (the reader interested in a thorough discussion about these aspects can refer to [8]). Concerning the electronics, the necessity of a pick-up circuit whose complexity depends on the SPAD current growth profile has been extensively discussed in Section 3. Moreover, an additional circuit is often required to measure the time of occurrence of each single photon detection event with respect to a reference signal (e.g. the laser pulse). This kind of circuit is the time measurement circuit, which is typically implemented by either exploiting a fully digital approach by means of a time-to-digital converter (TDC) [135] or with a time to amplitude converter (TAC) [136] followed by an analog to digital converter (ADC). Since the converter can require a significant amount of area and power, alternative solutions have been conceived to avoid it in some specific applications. This is the case of time of flight (ToF) measurements, for example, where a valid alternative to the direct measurement of the timing interval is the time-gating approach [13,137–139]. It consists in photon counting within multiple short time windows whose position is finely shifted over time. In the applications where this approach can be exploited, time-gating can allow the design of very compact circuitry (e.g. <7 transistors per pixel in [13]). In all other cases, the need of having both a front end and a TDC/TAC to code the time of arrival of each photon poses a significant challenge in multichannel systems because of fill factor. If we consider, for example, a multichannel system made of smart pixels like the one sketched in Fig. 12(a)), consisting of a front-side illuminated (FSI) SPAD integrated along with quenching front end, TDC and data

Optics EXPRESS

register, it is evident that the active area that can be used to collect the light is only a fraction of the overall chip area. There are basically three options to mitigate fill-factor issues: (i) circuit area minimization, (ii) resource sharing and (iii) chip stacking. Since high timing performance requires complex circuitry, circuit area minimization can easily result in limited performance and/or functionalities. Timing resolution with a compact TDC integrated in each pixel is typically on the order of several tens of picoseconds [133,140]. Alternatively, resource sharing is a well-founded option to reduce the area occupied by the time measurement circuit. Nonetheless, the number of pixels that can share one converter is limited to a few ones (typically four or eight) due to potential conflicts among pixels that could result into signal loss and/or distortion. In all cases, a quenching/reset circuit for each pixel is necessary to guarantee its independence, thus posing the ultimate limit to the achievable fill factor. For all these reasons, 3D stacking has been the real breakthrough in this field [138,139,141–144], allowing the combination of a detector-only chip and a circuit-only chip by either through silicon vias (TSVs) [145] or using backside-illuminated (BSI) SPAD architectures [103,146,147]. 3D stacking provides a twofold advantage: it minimizes the impact of the electronics area on the fill factor and, at the same time, it allows the exploitation of different technologies for detector and electronics thus allowing their independent optimization. In all cases, a further improvement to fill factor can be obtained by the exploitation of microlenses [148–150].

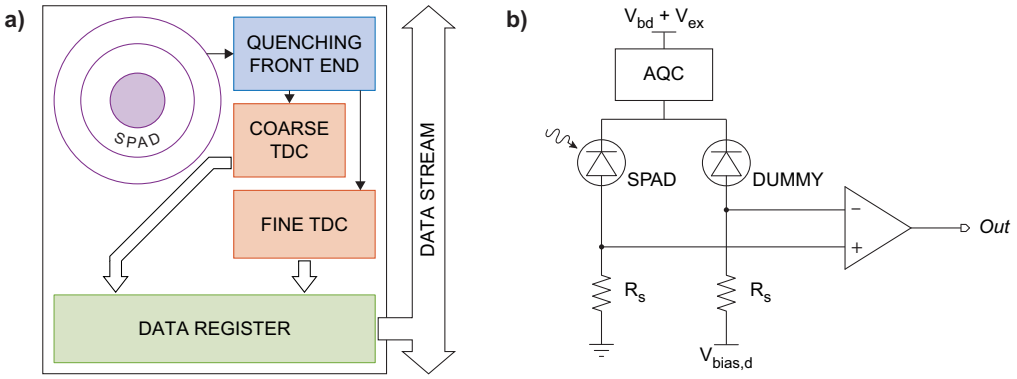

Fig. 12. High speed in timing measurements can be pursued with a multichannel structure and/or by pushing the speed of the single channel. a) Simplified scheme of a smart pixel including a SPAD with its quenching front end, a TDC and a data register. Multichannel systems can be built by replicating this building block. b) Differential architecture for high-speed operation of SPADs. An AQC drives both an active and a dummy SPAD and a differential readout filters out the spurious disturbance induced by the AQC during the fast reset transition. $V_{bias,d}$ provides a degree of freedom to set an equivalent threshold on the active-SPAD signal readout.

It is worth mentioning that the remarkable level of integration achieved in SPAD-based multichannel systems has set new challenges in data extraction. The architectures that have been presented in the literature can be classified by their readout mechanism into clock-driven, event-driven and router-based architectures. While in principle data extraction should not affect the timing performance, its impact on the overall system speed affects the design strategy, with potential effects also on timing. The router-based architecture, for example, requires a delay line on the signal path before it reaches the time measurement circuit, thus adding a stage that can directly contribute to the overall timing jitter. The reader interested in a comparison among data extraction architectures can refer to [151].

Overall, the best timing performance achieved with a single channel have not been extended to a multichannel structure so far, leaving an open door for near-future work in this direction.

Optics EXPRESS

In recent years there has been also a significant research interest in increasing the single channel count rate in timing applications. Some correction algorithms have been proposed to correct the pile-up distortion occurring in TCSPC when using a high detector count rate [152–154]. However, these approaches typically require some previous knowledge of the sample under test, thus limiting their range of applicability. In 2017, Cominelli et al. reported theoretical evidence that pile up can be actually avoided in TCSPC at high rates, provided that the detector dead time is matched to an integer number of excitation periods [102]. With this approach, a count rate as high as 32 Mcps with a single channel in a 80 MHz-laser setup has been experimentally demonstrated [155], increasing by a factor of eight the speed of a TCSPC acquisition chain with respect to the classic approach. Concerning timing, the remarkable increment of the detector count rate with respect to classic TCSPC experiments sets new challenging requirements on AQC and pick-up circuit. On one hand, the AQC should be able to reset the SPAD as quickly as possible to minimize the number of photons that can be detected during the reset phase. Indeed, those photons must be discarded to ensure negligible distortion at high rates, and so minimizing their number is crucial for system efficiency. On the other hand, a fast reset transition at one of the SPAD terminals is easily coupled to the other terminal through the anode-cathode capacitance, potentially causing the activation of the pick-up circuit. This scenario has been already faced in fast-gating systems, i.e. whenever the selective activation of the SPAD is used to filter out noise and/or interfering signals. Dalla Mora et al. in [156] showed that, with the pick-up circuit of Figure 9(a), the unwanted disturbance at the comparator input due to a fast reset transition (200 ps rising edge, 5.5 V pulse) can be higher than the signal itself (3 times higher in their case). To avoid this issue, both in fast-gating and fast TCSPC with matched dead time setups, the solution of choice is based on the exploitation of a dummy cell besides the active SPAD [157] to minimize the impact of the fast reset transition on the timing measurement.

The basic simplified schematic of this solution is sketched in Figure 12(b). An AQC, in this case connected to the cathode side, is used to drive both the active SPAD and a dummy cell, while two identical pick-up resistors are connected at the anode side, followed by a differential comparator. When the AQC applies its fast reset transition, the high spurious signal is generated on both branches thanks to the symmetry of the stage, thus generating a common mode signal that is ideally rejected. On the contrary, the absorption of a photon only generates a current on the active SPAD branch, resulting into a differential signal that triggers the comparator. By using two different bias voltages below the pick-up resistors, it is possible to set the equivalent threshold of the avalanche current detection, a parameter that can be crucial in timing measurements as widely discussed in Section 3. In this kind of setup, the exploitation of another SPAD as dummy cell is beneficial to maximize the symmetry of the two branches. However, it must be also guaranteed that this auxiliary detector always remains off. To this aim, one possibility is to use a dummy SPAD featuring a higher breakdown voltage with respect to the active one. In this way, the dummy SPAD can be kept below breakdown for any voltage applied by the shared AQC. However, this solution requires SPADs that have quite the same structure (to have the same parasitics), but feature different electric field profiles, which adds complexity, especially to integrate them both on the same silicon die, and it could be hardly feasible with standard technologies. In case the dummy and the active SPAD feature the same breakdown voltage, an independent bias network can be used on the dummy branch, and the AQC can be connected by means of an AC-coupled network. This solution has been successfully exploited by Acconcia et al. in [158]. An alternative solution could consist in the exploitation of a physically-blinded SPAD as dummy cell (e.g. covered by a metal layer), whose PDE is null also when biased above the breakdown voltage. However, in this case the DCR of the dummy cell would contribute to the overall noise of the system.

Overall, the matched dead time approach provided a new option to achieve high speed in timing measurements. We expect that this result will set new technological challenges, such as

Optics EXPRESS

integrating a dummy cell and a differential readout into a compact structure that can extend this solution to a multichannel, integrated and fast system. Furthermore, this result put into question the long-standing belief that TCSPC was an intrinsically slow technique, potentially contributing to increase the research interest and effort toward fast timing measurements.

5. Conclusion

One of the most peculiar features of SPADs is the ability to provide the information about the time of arrival of single photons with picoseconds precision. In this review we focused on silicon SPADs, suitable to detect light in the visible range and up to the near infrared region of the spectrum, and on the circuit solutions that have allowed to extract the best timing performance from these sensors. The performance of the most suitable SPADs for timing applications are summarized in Table 1. This table provides a reference summary for SPAD and front end designers and it can also be useful to system designers that have to select the detector and the front end electronics that best fit the specific requirements of the final application. Starting from the table, it is possible to summarize this work as follows.

Table 1. Summary of SPAD and front end circuit performance.

Type of SPAD	Diameter [µm]	Breakdown [V]	Overvoltage [V]	Maximum PDE	PDE @800nm	DCR[a] [cps]	Front end circuit	Min. dead time [ns]	Timing jitter [ps]	Tail time constant [ps]
CMOS p-i-n [16]	25	21.5	6	55% @480nm	12%	100	PQAR	3	**12.1**[b]	31.5 @515 nm
CMOS 130 nm [58]	8	20	12	**72% @560 nm**	28%	1.4k	PQC + inverter	30	52	>1000 @654nm
CMOS BCD 130 nm [18]	10	26.5	5	48% @490 nm	10%	~60	VLQC	**0.93**	75	80 @850 nm
Custom reach through [105]	180[c]	364.6	22	70% [c] @650nm	**62%**[c]	25-1.5k[c]	AQC + low-R sense	8.8	235	500 @780 nm
Custom red enhanced [117,159]	50	41.5	20	60% @650nm	40%	5k	AQC + TIA	15	**83**	300[d] @780 nm
Custom thin [120]	80	34.2	5	52[e] @525nm	15%[e]	~2.5k[e]	**TR_AQC**	12.5	37	350 @780 nm

[a]Data are reported at room temperature.
[b]7.5 ps @6.5 V of overvoltage [66].
[c]Data from commercial datasheet [118].
[d]Typical data reported in [12].
[e]Derived from [48].

CMOS technologies have been playing a key role especially thanks to the possibility of integrating both detector and complex electronics on the same chip. Indeed, the intrinsic parasitic minimization of this approach can have beneficial effects in terms of current collection and overall speed. Dead times of few ns [16] and even below 1 ns [18] have been demonstrated with CMOS SPADs. On the other hand, the exploitation of a single standard technology for both detector and electronics can result into SPADs featuring a relatively low PDE, especially in the near infrared. This is typically due to thin layer structures and the relatively low overvoltage required to comply with the safe operating region of the low voltage transistors within the front end circuit. With a non-isolated structure and an overvoltage exceeding 10 V, a PDE as high as 28% at 800 nm with a CMOS SPAD has been demonstrated, but at the expense of DCR (even with a diameter smaller than 10 µm) and tail time constant (> 1 ns) [58]. Custom technologies current provide the best PDE in the NIR, but combined with a timing jitter of at least 83 ps [117]. New device structures can break new grounds in the timing precision achievable with SPADs, as recently demonstrated by the first p-i-n structure achieving sub-10 ps timing jitter combined with low dark count rates, ultrashort tail time constant and low afterpulsing probability with a few ns dead time [66]. At the same time, combining best in class timing performance with high speed would be paramount to finally bring time-resolved measurements to life-changing applications.

Optics EXPRESS

Funding. Human Frontier Science Program (RGP0061/2019).

Disclosures. The authors declare no conflicts of interest.

Data availability. No data were generated or analyzed in the presented research.

References

1. F. Xu, X. Ma, Q. Zhang, H.-K. Lo, and J.-W. Pan, "Secure quantum key distribution with realistic devices," Rev. Mod. Phys. **92**(2), 025002 (2020).
2. Y. Yamada, H. Suzuki, and Y. Yamashita, "Time-domain near-infrared spectroscopy and imaging: a review," Appl. Sci. **9**(6), 1127 (2019).
3. Y. Altmann, S. McLaughlin, M. J. Padgett, V. K. Goyal, A. O. Hero, and D. Faccio, "Quantum-inspired computational imaging," Science **361**(6403), eaat2298 (2018).
4. I. Esmaeil Zadeh, J. Chang, J. W. N. Los, S. Gyger, A. Elshaari, S. Steinhauer, S. N. Dorenbos, and V. Zwiller, "Superconducting nanowire single-photon detectors: a perspective on evolution, state-of-the-art, future developments, and applications," Appl. Phys. Lett. **118**(19), 190502 (2021).
5. I. Holzman and Y. Ivry, "Superconducting nanowires for single-photon detection: progress, challenges, and opportunities," Adv. Quantum Technol. **2**(3-4), 1800058 (2019).
6. B. Korzh, Q. Y. Zhao, and J. P. Allmaras, *et al.*, "Demonstration of sub-3 ps temporal resolution with a superconducting nanowire single-photon detector," Nat. Photonics **14**(4), 250–255 (2020).
7. E. E. Wollman, V. B. Verma, A. E. Lita, W. H. Farr, M. D. Shaw, R. P. Mirin, and S. W. Nam, "Kilopixel array of superconducting nanowire single-photon detectors," Opt. Express **27**(24), 35279–35289 (2019).
8. F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, I. Rech, and R. Osellame, "Recent advances and future perspectives of single-photon avalanche diodes for quantum photonics applications," Adv. Quantum Technol. **4**(2), 2000102 (2021).
9. C. Bruschini, H. Homulle, I. M. Antolovic, S. Burri, and E. Charbon, "Single-photon avalanche diode imagers in biophotonics: review and outlook," Light: Sci. Appl. **8**(1), 87 (2019).
10. S. Donati and T. Tambosso, "Single-photon detectors: from traditional PMT to solid-state SPAD-based technology," IEEE J. Sel. Top. Quantum Electron. **20**(6), 204–211 (2014).
11. Hamamatsu Photonics K. K., "High speed compact HPD (Hybrid Photo Detector) series R11322U-40," https://www.hamamatsu.com/content/dam/hamamatsu-photonics/sites/documents/99_SALES_LIBRARY/etd/HPD_TPMH1361E.pdf (2018).
12. A. Gulinatti, F. Ceccarelli, M. Ghioni, and I. Rech, "Custom silicon technology for SPAD-arrays with red-enhanced sensitivity and low timing jitter," Opt. Express **29**(3), 4559–4581 (2021).
13. K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications," Optica **7**(4), 346–354 (2020).
14. K. Morimoto, J. Iwata, and M. Shinohara, *et al.*, "3.2 megapixel 3d-stacked charge focusing spad for low-light imaging and depth sensing," in *2021 IEEE International Electron Devices Meeting (IEDM)*, (IEEE, 2021), pp. 20–2.
15. M. Seo, H. Park, and J. S. Lee, "Evaluation of large-area silicon photomultiplier arrays for positron emission tomography systems," Electronics **10**(6), 698 (2021).
16. F. Gramuglia, M.-L. Wu, C. Bruschini, M.-J. Lee, and E. Charbon, "A low-noise CMOS SPAD pixel with 12.1 ps SPTR and 3 ns dead time," IEEE J. Sel. Top. Quantum Electron. **28**(2: Optical Detectors), 1–9 (2022).
17. M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce, and F. Zappa, "Single-photon avalanche diodes in a 0.16 μm BCD technology with sharp timing response and red-enhanced sensitivity," IEEE J. Sel. Top. Quantum Electron. **24**(2), 1–9 (2018).
18. F. Severini, I. Cusini, D. Berretta, K. Pasquinelli, A. Incoronato, and F. Villa, "SPAD pixel with sub-ns dead-time for high-count rate applications," IEEE J. Sel. Top. Quantum Electron. **28**(2: Optical Detectors), 1–8 (2022).
19. S. Dello Russo, A. Elefante, D. Dequal, D. K. Pallotti, L. Santamaria Amato, F. Sgobba, and M. Siciliani de Cumis, "Advances in mid-infrared single-photon detection," Photonics **9**(7), 470 (2022).
20. C. Liu, H.-F. Ye, and Y.-L. Shi, "Advances in near-infrared avalanche diode single-photon detectors," Chip **1**(1), 100005 (2022).
21. F. Thorburn, X. Yi, Z. M. Greener, J. Kirdoda, R. W. Millar, L. L. Huddleston, D. J. Paul, and G. S. Buller, "Ge-on-Si single-photon avalanche diode detectors for short-wave infrared wavelengths," J. Phys. Photonics **4**(1), 012001 (2022).
22. P. D. Anderson, J. D. Beck, W. Sullivan III, C. Schaake, J. McCurdy, M. Skokan, P. Mitra, and X. Sun, "Recent advancements in HgCdTe APDs for space applications," J. Electron. Mater. **51**(12), 6803–6814 (2022).
23. S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," Appl. Opt. **35**(12), 1956–1976 (1996).
24. D. P. Palubiak and M. J. Deen, "CMOS SPADs: design issues and research challenges for detectors, circuits, and arrays," IEEE J. Sel. Top. Quantum Electron. **20**(6), 409–426 (2014).
25. J. Zheng, X. Xue, C. Ji, Y. Yuan, K. Sun, D. Rosenmann, L. Wang, J. Wu, J. C. Campbell, and S. Guha, "Dynamic-quenching of a single-photon avalanche photodetector using an adaptive resistive switch," Nat. Commun. **13**(1), 1517 (2022).

Optics EXPRESS

26. A. Dalla Mora, A. Tosi, S. Tisa, and F. Zappa, "Single-photon avalanche diode model for circuit simulations," IEEE Photonics Technol. Lett. **19**(23), 1922–1924 (2007).
27. F. Zappa, A. Tosi, A. Dalla Mora, and S. Tisa, "SPICE modeling of single photon avalanche diodes," Sens. Actuators, A **153**(2), 197–204 (2009).
28. G. Giustolisi, R. Mita, and G. Palumbo, "Behavioral modeling of statistical phenomena of single-photon avalanche diodes," Int. J. Circ. Theor. Appl. **40**(7), 661–679 (2012).
29. Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen, and H. Peng, "A comprehensive and accurate analytical SPAD model for circuit simulation," IEEE Trans. Electron Devices **63**(5), 1940–1948 (2016).
30. A. Tontini, L. Gasparini, and M. Perenzoni, "Numerical model of SPAD-based direct time-of-flight flash LIDAR CMOS image sensors," Sensors **20**(18), 5203 (2020).
31. A. Incoronato, M. Locatelli, and F. Zappa, "Statistical modelling of SPADs for time-of-flight LiDAR," Sensors **21**(13), 4481 (2021).
32. A. Buchner, S. Hadrath, R. Burkard, F. M. Kolb, J. Ruskowski, M. Ligges, and A. Grabmaier, "Analytical evaluation of signal-to-noise ratios for avalanche- and single-photon avalanche diodes," Sensors **21**(8), 2887 (2021).
33. A. Corbeil Therrien, B. L. Berube, S. A. Charlebois, R. Lecomte, R. Fontaine, and J. F. Pratte, "Modeling of single photon avalanche diode array detectors for PET applications," IEEE Trans. Nucl. Sci. **61**(1), 14–22 (2014).
34. D. R. Schaart, "Physics and technology of time-of-flight PET detectors," Phys. Med. Biol. **66**(9), 09TR01 (2021).
35. F. Corsi, A. Dragone, C. Marzocca, A. Del Guerra, P. Delizia, N. Dinu, C. Piemonte, M. Boscardin, and G. F. Dalla Betta, "Modelling a silicon photomultiplier (SiPM) as a signal source for optimum front-end design," Nucl. Instrum. Methods Phys. Res., Sect. A **572**(1), 416–418 (2007).
36. A. J. Gonzalez, M. Moreno, J. Barbera, P. Conde, L. Hernandez, L. Moliner, J. M. Monzo, A. Orero, A. Peiro, R. Polo, M. J. Rodriguez-Alvarez, A. Ros, F. Sanchez, A. Soriano, L. F. Vidal, and J. M. Benlloch, "Simulation study of resistor networks applied to an array of 256 SiPMs," IEEE Trans. Nucl. Sci. **60**(2), 592–598 (2013).
37. S. Xie, J. Liu, and F. Zhang, "An accurate circuit model for the statistical behavior of InP/InGaAs SPAD," Electronics **9**(12), 2059 (2020).
38. M. Sanzaro, N. Calandri, A. Ruggeri, and A. Tosi, "InGaAs/InP SPAD with monolithically integrated zinc-diffused resistor," IEEE J. Quantum Electron. **52**(7), 1–7 (2016).
39. T. Lunghi, C. Barreiro, O. Guinnard, R. Houlmann, X. Jiang, M. A. Itzler, and H. Zbinden, "Free-running single-photon detection based on a negative feedback InGaAs APD," J. Mod. Opt. **59**(17), 1481–1488 (2012).
40. J. Liu, Y. Xu, Y. Li, Z. Liu, and X. Zhao, "Exploiting the single-photon detection performance of InGaAs negative-feedback avalanche diode with fast active quenching," Opt. Express **29**(7), 10150–10161 (2021).
41. M. Dandin, M. H. U. Habib, B. Nouri, P. Abshire, and N. McFarlane, "Characterization of single-photon avalanche diodes in a 0.5-µm standard CMOS process — Part 2: equivalent circuit model and Geiger mode readout," IEEE Sens. J. **16**(9), 3075–3083 (2016).
42. 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. **13**(7), 075008 (2011).
43. E. Amri, G. Boso, B. Korzh, and H. Zbinden, "Temporal jitter in free-running InGaAs/InP single-photon avalanche detectors," Opt. Lett. **41**(24), 5728–5731 (2016).
44. A. Tosi, A. Dalla Mora, F. Zappa, A. Gulinatti, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli, and R. Cubeddu, "Fast-gated single-photon counting technique widens dynamic range and speeds up acquisition time in time-resolved measurements," Opt. Express **19**(11), 10735–10746 (2011).
45. D. Contini, A. Dalla Mora, L. Spinelli, A. Farina, A. Torricelli, R. Cubeddu, F. Martelli, G. Zaccanti, A. Tosi, G. Boso, F. Zappa, and A. Pifferi, "Effects of time-gated detection in diffuse optical imaging at short source-detector separation," J. Phys. D: Appl. Phys. **48**(4), 045401 (2015).
46. G. Ripamonti and S. Cova, "Carrier diffusion effects in the time-response of a fast photodiode," Solid-State Electron. **28**(9), 925–931 (1985).
47. A. Gulinatti, I. Rech, M. Assanelli, M. Ghioni, and S. Cova, "A physically based model for evaluating the photon detection efficiency and the temporal response of SPAD detectors," J. Mod. Opt. **58**(3-4), 210–224 (2011).
48. M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, and S. Cova, "Progress in silicon single-photon avalanche diodes," IEEE J. Sel. Top. Quantum Electron. **13**(4), 852–862 (2007).
49. F. Sun, Y. Xu, Z. Wu, and J. Zhang, "A simple analytic modeling method for SPAD timing jitter prediction," IEEE J. Electron Devices Soc. **7**, 261–267 (2019).
50. R. Helleboid, D. Rideau, J. Grebot, I. Nicholson, N. Moussy, O. Saxod, J. Saint-Martin, M. Pala, and P. Dollfus, "Modeling of SPAD avalanche breakdown probability and jitter tail with field lines," Solid-State Electron. **194**, 108376 (2022).
51. 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. **41**(5), 272 (2005).
52. M. J. Hsu, S. C. Esener, and H. Finkelstein, "A CMOS STI-bound single-photon avalanche diode with 27-ps timing resolution and a reduced diffusion tail," IEEE Electron Device Lett. **30**(6), 641–643 (2009).
53. A. Tosi, F. Acerbi, M. Anti, and F. Zappa, "InGaAs/InP single-photon avalanche diode with reduced afterpulsing and sharp timing response with 30 ps tail," IEEE J. Quantum Electron. **48**(9), 1227–1232 (2012).

Optics EXPRESS

54. A. Spinelli, M. Ghioni, S. Cova, and L. M. Davis, "Avalanche detector with ultraclean response for time-resolved photon counting," IEEE J. Quantum Electron. **34**(5), 817–821 (1998).
55. F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm," J. Mod. Opt. **61**(2), 102–115 (2014).
56. A. Lacaita, M. Ghioni, and S. Cova, "Double epitaxy improves single-photon avalanche diode performance," Electron. Lett. **25**(13), 841 (1989).
57. 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. **21**(14), 1020–1022 (2009).
58. E. A. G. Webster, L. Grant, and R. K. Henderson, "A high-performance single-photon avalanche diode in 130-nm CMOS imaging technology," IEEE Electron Device Lett. **33**(11), 1589–1591 (2012).
59. E. A. G. Webster, J. Richardson, L. 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. **33**(5), 694–696 (2012).
60. C. Veerappan and E. Charbon, "A low dark count p-i-n diode based SPAD in CMOS technology," IEEE Trans. Electron Devices **63**(1), 65–71 (2016).
61. F. Gramuglia, P. Keshavarzian, E. Kizilkan, C. Bruschini, S. S. Tan, M. Tng, E. Quek, M.-J. Lee, and E. Charbon, "Engineering breakdown probability profile for PDP and DCR optimization in a SPAD fabricated in a standard 55 nm BCD process," IEEE J. Sel. Top. Quantum Electron. **28**(2: Optical Detectors), 1–10 (2022).
62. H. Lee, H. Choi, and I. Yun, "Junction engineering-based modeling and optimization of deep junction silicon single-photon avalanche diodes for device scaling," IEEE Trans. Electron Devices **69**(9), 4970–4975 (2022).
63. D. Issartel, S. Gao, P. Pittet, R. Cellier, D. Golanski, A. Cathelin, and F. Calmon, "Architecture optimization of SPAD integrated in 28 nm FD-SOI CMOS technology to reduce the DCR," Solid-State Electron. **191**, 108297 (2022).
64. V. Gautam, R. Casanova, S. Terzo, and S. Grinstein, "Development of single photon avalanche detectors for NIR light detection," J. Instrum. **17**(12), C12019 (2022).
65. D. Shin, B. Park, Y. Chae, and I. Yun, "The effect of a deep virtual guard ring on the device characteristics of silicon single photon avalanche diodes," IEEE Trans. Electron Devices **66**(7), 2986–2991 (2019).
66. F. Gramuglia, E. Ripiccini, C. A. Fenoglio, M.-L. Wu, L. Paolozzi, C. Bruschini, and E. Charbon, "Sub-10 ps minimum ionizing particle detection with geiger-mode APDs," Frontiers in Physics p. 370 (2022).
67. S. Cova, M. Ghioni, A. Lotito, I. Rech, and F. Zappa, "Evolution and prospects for single-photon avalanche diodes and quenching circuits," J. Mod. Opt. **51**(9-10), 1267–1288 (2004).
68. S. M. Sze and W. Shockley, "Unit-cube expression for space-charge resistance," Bell Syst. Tech. J. **46**(5), 837–842 (1967).
69. A. Spinelli and A. Lacaita, "Physics and numerical simulation of single photon avalanche diodes," IEEE Trans. Electron Devices **44**(11), 1931–1943 (1997).
70. C. Jacoboni and P. Lugli, *he Monte Carlo method for semiconductor device simulation* (Springer-Verlag, 1989).
71. S. Ramo, "Currents induced by electron motion," Proc. IRE **27**(9), 584–585 (1939).
72. W. G. Oldham, R. R. Samuelson, and P. Antognetti, "Triggering phenomena in avalanche diodes," IEEE Trans. Electron Devices **19**(9), 1056–1060 (1972).
73. R. J. McIntyre, "On the avalanche initiation probability of avalanche diodes above the breakdown voltage," IEEE Trans. Electron Devices **20**(7), 637–641 (1973).
74. R. J. McIntyre, "A new look at impact ionization-Part I: A theory of gain, noise, breakdown probability, and frequency response," IEEE Trans. Electron Devices **46**(8), 1623–1631 (1999).
75. C. Groves, C. H. Tan, J. P. R. David, G. J. Rees, and M. M. Hayat, "Exponential time response in analogue and geiger mode avalanche photodiodes," IEEE Trans. Electron Devices **52**(7), 1527–1534 (2005).
76. A. Spinelli, A. Pacelli, and A. L. Lacaita, "Dead space approximation for impact ionization in silicon," Appl. Phys. Lett. **69**(24), 3707–3709 (1996).
77. Y. Okuto and C. R. Crowell, "Ionization coefficients in semiconductors: A nonlocalized property," Phys. Rev. B **10**(10), 4284–4296 (1974).
78. Y. Okuto and C. R. Crowell, "Threshold energy effect on avalanche breakdown voltage in semiconductor junctions," Solid-State Electron. **18**(2), 161–168 (1975).
79. C. H. Tan, J. S. Ng, G. J. Rees, and J. P. R. David, "Statistics of avalanche current buildup time in single-photon avalanche diodes," IEEE J. Sel. Top. Quantum Electron. **13**(4), 906–910 (2007).
80. S. L. Tan, D. S. Ong, and H. K. Yow, "Advantages of thin single-photon avalanche diodes," Phys. Status Solidi A **204**(7), 2495–2499 (2007).
81. S. L. Tan, D. S. Ong, and H. K. Yow, "Theoretical analysis of breakdown probabilities and jitter in single-photon avalanche diodes," J. Appl. Phys. **102**(4), 044506 (2007).
82. A. Ingargiola, M. Assanelli, A. Gallivanoni, I. Rech, M. Ghioni, and S. Cova, "Avalanche buildup and propagation effects on photon-timing jitter in Si-SPAD with non-uniform electric field," in *SPIE Defense, Security + Sensing*, vol. 7320 M. A. Itzler and J. C. Campbell, eds. (2009), pp. 73200K–73200K-11.

Optics EXPRESS

83. S. Reggiani, N. Jensen, G. Groos, M. Stecher, E. Gnani, M. Rudan, G. Baccarani, C. Corvasce, D. Barlini, M. Ciappa, W. Fichtner, M. Denison, N. Jensen, G. Groos, and M. Stecher, "Measurement and modeling of the electron impact-ionization coefficient in silicon up to very high temperatures," IEEE Trans. Electron Devices **52**(10), 2290–2299 (2005).
84. D. Dolgos, H. Meier, A. Schenk, and B. Witzigmann, "Full-band Monte Carlo simulation of high-energy carrier transport in single photon avalanche diodes: Computation of breakdown probability, time to avalanche breakdown, and jitter," J. Appl. Phys. **110**(8), 084507 (2011).
85. D. Dolgos, H. Meier, A. Schenk, and B. Witzigmann, "Full-band Monte Carlo simulation of high-energy carrier transport in single photon avalanche diodes with multiplication layers made of InP, InAlAs, and GaAs," J. Appl. Phys. **111**(10), 104508 (2012).
86. P. J. Hambleton, J. P. R. David, and G. J. Rees, "Enhanced carrier velocity to early impact ionization," J. Appl. Phys. **95**(7), 3561–3564 (2004).
87. J. D. Petticrew, S. J. Dimler, X. Zhou, A. P. Morrison, C. H. Tan, and J. S. Ng, "Avalanche breakdown timing statistics for silicon single photon avalanche diodes," IEEE J. Sel. Top. Quantum Electron. **24**(2), 1–6 (2018).
88. S. Jallepalli, M. Rashed, W.-K. Shih, C. M. Maziar, and A. F. Tasch, "A full-band Monte Carlo model for hole transport in silicon," J. Appl. Phys. **81**(5), 2250–2255 (1997).
89. S. A. Plimmer, J. P. David, D. S. Ong, and K. F. Li, "A simple model for avalanche multiplication including deadspace effects," IEEE Trans. Electron Devices **46**(4), 769–775 (1999).
90. X. Zhou, J. S. Ng, and C. H. Tan, "A simple Monte Carlo model for prediction of avalanche multiplication process in Silicon," J. Instrum. **7**(08), P08006 (2012).
91. J. Zheng, S. Z. Ahmed, Y. Yuan, A. Jones, Y. Tan, A. K. Rockwell, S. D. March, S. R. Bank, A. W. Ghosh, and J. C. Campbell, "Full band Monte Carlo simulation of AlInAsSb digital alloys," InfoMat **2**(6), 1236–1240 (2020).
92. A. Lacaita, M. Mastrapasqua, M. Ghioni, and S. Vanoli, "Observation of avalanche propagation by multiplication assisted diffusion in p-n junctions," Appl. Phys. Lett. **57**(5), 489–491 (1990).
93. A. Lacaita and M. Mastrapasqua, "Strong dependence of time resolution on detector diameter in single photon avalanche diodes," Electron. Lett. **26**(24), 2053 (1990).
94. M. Ghioni and G. Ripamonti, "Improving the performance of commercially available Geiger-mode avalanche photodiodes," Rev. Sci. Instrum. **62**(1), 163–167 (1991).
95. A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes," Appl. Phys. Lett. **62**(6), 606–608 (1993).
96. A. Lacaita, A. Spinelli, and S. Longhi, "Avalanche transients in shallow p-n junctions biased above breakdown," Appl. Phys. Lett. **67**(18), 2627–2629 (1995).
97. 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. **47**(2), 151–159 (2011).
98. A. Ingargiola, M. Assanelli, I. Rech, A. Gulinatti, and M. Ghioni, "Avalanche current measurements in SPADs by means of hot-carrier luminescence," IEEE Photonics Technol. Lett. **23**(18), 1319–1321 (2011).
99. M. Assanelli, A. Gulinatti, I. Rech, and M. Ghioni, "Timing enhanced silicon SPAD design," in *2011 Numerical Simulation of Optoelectronic Devices*, (IEEE, 2011), pp. 197–198.
100. H. C. Bowers, "Space-charge-induced negative resistance in avalanche diodes," IEEE Trans. Electron Devices **15**(6), 343–350 (1968).
101. W. Becker, *Advanced time-correlated single photon counting techniques*, vol. 81 (Springer Science & Business Media, 2005).
102. A. Cominelli, G. Acconcia, P. Peronio, M. Ghioni, and I. Rech, "High-speed and low-distortion solution for time-correlated single photon counting measurements: A theoretical analysis," Rev. Sci. Instrum. **88**(12), 123701 (2017).
103. S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-PDE, backside-illuminated SPAD in 65/40-nm 3D IC CMOS pixel with cascoded passive quenching and active recharge," IEEE Electron Device Lett. **38**(11), 1547–1550 (2017).
104. G. Acconcia, I. Rech, A. Gulinatti, and M. Ghioni, "High-voltage integrated active quenching circuit for single photon count rate up to 80 Mcounts/s," Opt. Express **24**(16), 17819–17831 (2016).
105. S. Farina, I. Labanca, G. Acconcia, M. Ghioni, and I. Rech, "10-nanosecond dead time and low afterpulsing with a free-running reach-through single-photon avalanche diode," Rev. Sci. Instrum. **93**(5), 053102 (2022).
106. G. Acconcia, I. Labanca, I. Rech, A. Gulinatti, and M. Ghioni, "Note: Fully integrated active quenching circuit achieving 100 MHz count rate with custom technology single photon avalanche diodes," Rev. Sci. Instrum. **88**(2), 026103 (2017).
107. 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," (2002). US Patent 6,384,663.
108. I. Rech, I. Labanca, M. Ghioni, and S. Cova, "Modified single photon counting modules for optimal timing performance," Rev. Sci. Instrum. **77**(3), 033104 (2006).
109. F. Nolet, S. Parent, N. Roy, M.-O. Mercier, S. A. Charlebois, R. Fontaine, and J.-F. Pratte, "Quenching circuit and SPAD integrated in CMOS 65 nm with 7.8 ps FWHM single photon timing resolution," Instruments **2**(4), 19 (2018).
110. M. Ghioni, A. Gulinatti, I. Rech, P. Maccagnani, and S. Cova, "Large-area low-jitter silicon single photon avalanche diodes," in *Quantum Sensing and Nanophotonic Devices V*, vol. 6900 (SPIE, 2008), pp. 267–279.

Optics EXPRESS

111. I. Labanca, F. Ceccarelli, A. Gulinatti, M. Ghioni, and I. Rech, "Triple epitaxial single-photon avalanche diode for multichannel timing applications," Electron. Lett. **54**(10), 644–645 (2018).
112. B. Razavi, "The transimpedance amplifier [a circuit for all seasons]," IEEE Solid-State Circuits Mag. **11**(1), 10–97 (2019).
113. M. Crotti, I. Rech, A. Gulinatti, and M. Ghioni, "Avalanche current read-out circuit for low-jitter parallel photon timing," Electron. Lett. **49**(16), 1017–1018 (2013).
114. M. Crotti, I. Rech, G. Acconcia, A. Gulinatti, and M. Ghioni, "A 2-GHz bandwidth, integrated transimpedance amplifier for single-photon timing applications," IEEE Trans. VLSI Syst. **23**(12), 2819–2828 (2015).
115. G. Acconcia, I. Rech, I. Labanca, and M. Ghioni, "32ps timing jitter with a fully integrated front end circuit and single photon avalanche diodes," Electron. Lett. **53**(5), 328–329 (2017).
116. F. Acerbi, M. Cazzanelli, A. Ferri, A. Gola, L. Pavesi, N. Zorzi, and C. Piemonte, "High detection efficiency and time resolution integrated-passive-quenched single-photon avalanche diodes," IEEE J. Sel. Top. Quantum Electron. **20**(6), 268–275 (2014).
117. F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, and I. Rech, "83-ps timing jitter with a red-enhanced SPAD and a fully integrated front end circuit," IEEE Photonics Technol. Lett. **30**(19), 1727–1730 (2018).
118. Excelitas Technologies Corp., "Single Photon Counting Modules," https://www.excelitas.com/product/spcm-aqrh (2019).
119. S. Tisa, F. Guerrieri, and F. Zappa, "Variable-load quenching circuit for single-photon avalanche diodes," Opt. Express **16**(3), 2232–2244 (2008).
120. G. Acconcia, M. Ghioni, and I. Rech, "37ps-precision time-resolving active quenching circuit for high-performance single photon avalanche diodes," IEEE Photonics J. **10**(6), 1–13 (2018).
121. S. Yanikgonul, V. Leong, J. R. Ong, C. E. Png, and L. Krivitsky, "2D Monte Carlo simulation of a silicon waveguide-based single-photon avalanche diode for visible wavelengths," Opt. Express **26**(12), 15232–15246 (2018).
122. S. Yanikgonul, V. Leong, J. R. Ong, C. E. Png, and L. Krivitsky, "Simulation of silicon waveguide single-photon avalanche detectors for integrated quantum photonics," IEEE J. Sel. Top. Quantum Electron. **26**(2), 1–8 (2020).
123. N. J. D. Martinez, M. Gehl, C. T. Derose, A. L. Starbuck, A. T. Pomerene, A. L. Lentine, D. C. Trotter, and P. S. Davids, "Single photon detection in a waveguide-coupled Ge-on-Si lateral avalanche photodiode," Opt. Express **25**(14), 16130–16139 (2017).
124. H. Wang, Y. Shi, Y. Zuo, Y. Yu, L. Lei, X. Zhang, and Z. Qian, "High-performance waveguide coupled Germanium-on-silicon single-photon avalanche diode with independently controllable absorption and multiplication," Nanophotonics **12**(4), 705–714 (2023).
125. Y. Li, X. Liu, X. Li, L. Zhang, B. Chen, Z. Zhi, X. Li, G. Zhang, P. Ye, G. Huang, D. He, W. Chen, F. Gao, P. Guo, X. Luo, G. Lo, and J. Song, "Germanium-on-silicon avalanche photodiode for 1550 nm weak light signal detection at room temperature," Chin. Opt. Lett. **20**(6), 062501 (2022).
126. 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. **20**(6), 413–415 (2008).
127. M. Zavvari, K. Abedi, and M. Karimi, "Design of resonant cavity structure for efficient high-temperature operation of single-photon avalanche photodiodes," Appl. Opt. **53**(15), 3311–3317 (2014).
128. Q. Chen, S. Wu, L. Zhang, W. Fan, and C. S. Tan, "Simulation of high-efficiency resonant-cavity-enhanced GeSn single-photon avalanche photodiodes for sensing and optical quantum applications," IEEE Sens. J. **21**(13), 14789–14798 (2021).
129. M. Ghioni, G. Armellini, P. Maccagnani, I. Rech, M. K. Emsley, and M. S. Ünlü, "Resonant-cavity-enhanced single photon avalanche diodes on double silicon-on-insulator substrates," J. Mod. Opt. **56**(2-3), 309–316 (2009).
130. J. Ma, M. Zhou, Z. Yu, X. Jiang, Y. Huo, K. Zang, J. Zhang, J. S. Harris, G. Jin, Q. Zhang, and J.-W. Pan, "Simulation of a high-efficiency and low-jitter nanostructured silicon single-photon avalanche diode," Optica **2**(11), 974–979 (2015).
131. K. Zang, X. Jiang, Y. Huo, X. Ding, M. Morea, X. Chen, C.-Y. Lu, J. Ma, M. Zhou, Z. Xia, Z. Yu, T. I. Kamins, Q. Zhang, and J. S. Harris, "Silicon single-photon avalanche diodes with nano-structured light trapping," Nat. Commun. **8**(1), 628 (2017).
132. A. Giudici, G. Acconcia, I. Labanca, M. Ghioni, and I. Rech, "4 ns dead time with a fully integrated active quenching circuit driving a custom single photon avalanche diode," Rev. Sci. Instrum. **93**(4), 043103 (2022).
133. R. K. Henderson, N. Johnston, F. M. Della Rocca, H. Chen, D. D.-U. Li, G. Hungerford, R. Hirsch, D. Mcloskey, P. Yip, and D. J. Birch, "A 192×128 time correlated SPAD image sensor in 40-nm CMOS technology," IEEE J. Solid-State Circuits **54**(7), 1907–1916 (2019).
134. C. Veerappan, J. Richardson, and R. Walker, *et al.*, "A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *2011 IEEE International Solid-State Circuits Conference*, (IEEE, 2011), pp. 312–314.
135. S. Henzler, *Time-to-digital converter basics* (Springer, 2010).
136. M. Crotti, I. Rech, and M. Ghioni, "Monolithic time-to-amplitude converter for TCSPC applications with 45 ps time resolution," in *2011 7th Conference on Ph. D. Research in Microelectronics and Electronics*, (IEEE, 2011), pp. 21–24.

Optics EXPRESS

137. H. Ruokamo, L. Hallman, H. Rapakko, and J. Kostamovaara, "An 80×25 pixel CMOS single-photon range image sensor with a flexible on-chip time gating topology for solid state 3D scanning," in *ESSCIRC 2017-43rd IEEE European Solid State Circuits Conference*, (IEEE, 2017), pp. 59–62.
138. T. Al Abbas, O. Almer, S. W. Hutchings, A. T. Erdogan, I. Gyongy, N. A. Dutton, and R. K. Henderson, "A 128×120 5-wire 1.96mm^2 40nm/90nm 3D stacked SPAD time resolved image sensor SoC for microendoscopy," in *2019 Symposium on VLSI Circuits*, (IEEE, 2019), pp. C260–C261.
139. P. Padmanabhan, C. Zhang, M. Cazzaniga, B. Efe, A. R. Ximenes, M.-J. Lee, and E. Charbon, "7.4 A 256×128 3D-stacked (45nm) SPAD FLASH LiDAR with 7-level coincidence detection and progressive gating for 100m range and 10klux background light," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64 (IEEE, 2021), pp. 111–113.
140. M. Zarghami, L. Gasparini, L. Parmesan, M. Moreno-Garcia, A. Stefanov, B. Bessire, M. Unternährer, and M. Perenzoni, "A 32×32-pixel CMOS imager for quantum optics with per-SPAD TDC, 19.48% fill-factor in a 44.64-µm pitch reaching 1-MHz observation rate," IEEE J. Solid-State Circuits **55**(10), 2819–2830 (2020).
141. E. Charbon, C. Bruschini, and M.-J. Lee, "3D-stacked CMOS SPAD image sensors: Technology and applications," in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, (IEEE, 2018), pp. 1–4.
142. S. W. Hutchings, N. Johnston, I. Gyongy, T. Al Abbas, N. A. Dutton, M. Tyler, S. Chan, J. Leach, and R. K. Henderson, "A reconfigurable 3-D-stacked SPAD imager with in-pixel histogramming for flash LIDAR or high-speed time-of-flight imaging," IEEE J. Solid-State Circuits **54**(11), 2947–2956 (2019).
143. A. R. Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D.-N. Yaung, and E. Charbon, "A modular, direct time-of-flight depth sensor in 45/65-nm 3-D-stacked CMOS technology," IEEE J. Solid-State Circuits **54**(11), 3203–3214 (2019).
144. A. T. Erdogan, T. Al Abbas, N. Finlayson, C. Hopkinson, I. Gyongy, O. Almer, N. A. Dutton, and R. K. Henderson, "A high dynamic range 128×120 3-D stacked CMOS SPAD image sensor SoC for fluorescence microendoscopy," IEEE J. Solid-State Circuits **57**(6), 1649–1660 (2022).
145. B.-L. Bérubé, V.-P. Rhéaume, A. C. Therrien, S. Parent, L. Maurais, A. Boisvert, G. Carini, S. A. Charlebois, R. Fontaine, and J.-F. Pratte, "Development of a single photon avalanche diode (SPAD) array in high voltage CMOS 0.8 µm dedicated to a 3D integrated circuit (3DIC)," in *2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC)*, (IEEE, 2012), pp. 1835–1839.
146. J. M. Pavia, M. Scandini, S. Lindner, M. Wolf, and E. Charbon, "A 1×400 backside-illuminated SPAD sensor with 49.7 ps resolution, 30 pJ/sample TDCs fabricated in 3D CMOS technology for near-infrared optical tomography," IEEE J. Solid-State Circuits **50**(10), 2406–2418 (2015).
147. O. Kumagai, J. Ohmachi, and M. Matsumura, *et al.*, "7.3 A 189×600 back-illuminated stacked SPAD direct time-of-flight depth sensor for automotive LiDAR systems," in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, vol. 64 (IEEE, 2021), pp. 110–112.
148. J. M. Pavia, M. Wolf, and E. Charbon, "Measurement and modeling of microlenses fabricated on single-photon avalanche diode arrays for fill factor recovery," Opt. Express **22**(4), 4202–4213 (2014).
149. P. W. Connolly, X. Ren, and A. McCarthy, *et al.*, "High concentration factor diffractive microlenses integrated with CMOS single-photon avalanche diode detector arrays for fill-factor improvement," Appl. Opt. **59**(14), 4488–4498 (2020).
150. C. Bruschini, I. M. Antolovic, F. Zanella, A. C. Ulku, S. Lindner, A. Kalyanov, T. Milanese, E. Bernasconi, V. Pešić, and E. Charbon, "Challenges and prospects for multi-chip microlens imprints on front-side illuminated SPAD imagers," Opt. Express **31**(13), 21935–21953 (2023).
151. A. Cominelli, G. Acconcia, P. Peronio, I. Rech, and M. Ghioni, "Readout architectures for high efficiency in time-correlated single photon counting experiments—Analysis and review," IEEE Photonics J. **9**(3), 1–15 (2017).
152. S. Isbaner, N. Karedla, D. Ruhlandt, S. C. Stein, A. Chizhik, I. Gregor, and J. Enderlein, "Dead-time correction of fluorescence lifetime measurements and fluorescence lifetime imaging," Opt. Express **24**(9), 9429–9445 (2016).
153. J. Rapp, Y. Ma, R. M. Dawson, and V. K. Goyal, "Dead time compensation for high-flux ranging," IEEE Trans. Signal Process. **67**(13), 3471–3486 (2019).
154. A. K. Pediredla, A. C. Sankaranarayanan, M. Buttafava, A. Tosi, and A. Veeraraghavan, "Signal processing based pile-up compensation for gated single-photon avalanche diodes," arXiv, arXiv:1806.07437 (2018).
155. S. Farina, G. Acconcia, I. Labanca, M. Ghioni, and I. Rech, "Toward ultra-fast time-correlated single-photon counting: A compact module to surpass the pile-up limit," Rev. Sci. Instrum. **92**(6), 063702 (2021).
156. 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. **16**(4), 1023–1030 (2010).
157. A. Tomita and K. Nakamura, "Balanced, gated-mode photon detector for quantum-bit discrimination at 1550 nm," Opt. Lett. **27**(20), 1827–1829 (2002).
158. G. Acconcia, A. Cominelli, M. Ghioni, and I. Rech, "Fast fully-integrated front-end circuit to overcome pile-up limits in time-correlated single photon counting with single photon avalanche diodes," Opt. Express **26**(12), 15398–15410 (2018).
159. F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, and I. Rech, "Fully integrated active quenching circuit driving custom-technology SPADs with 6.2 ns dead time," IEEE Photonics Technol. Lett. **31**(1), 102–105 (2019).