Nuclear Inst. and Methods in Physics Research, A 926 (2019) 16–35

Contents lists available at ScienceDirect

Nuclear Inst. and Methods in Physics Research, A

journal homepage: www.elsevier.com/locate/nima

Understanding and simulating SiPMs

Fabio Acerbi [a,*], Stefan Gundacker [b,c,**]

[a] Fondazione Bruno Kessler, via Sommarive 18, Trento, Italy
[b] UniMIB, Piazza dell'Ateneo Nuovo, 1 - 20126, Milano, Italy
[c] CERN, 1211 Geneve 23, Switzerland

ARTICLE INFO

Keywords:
SiPM review
SPAD
PDE
Noise in SiPM
Equivalent electrical model
Fast timing
TOF-PET
SPTR

ABSTRACT

The silicon-photomultiplier (SiPM) is becoming the device of choice for different applications, for example in fast timing like in time of flight positron emission tomography (TOF-PET) and in high energy physics (HEP). It is also becoming a choice in many single-photon or few-photon based applications, like for spectroscopy, quantum experiments and distance measurements (LIDAR). In order to fully benefit from the good performance of the SiPM, in particular its sensitivity, the dynamic range and its intrinsically fast timing properties it is necessary to understand, quantitatively describe and simulate the various parameters concerned. These analyses consider the structure and the electrical model of a single photon avalanche diode (SPAD), i.e. the SiPM microcell, and the integration in an array, i.e. the SiPM. Additionally, for several applications a more phenomenological and complete view on SiPMs has to be done, e.g. photon detection efficiency, single photon time resolution, SiPM signal response, gain fluctuation, dark count rate, afterpulse, prompt and delayed optical crosstalk. These quantities of SiPMs can strongly influence the time and energy resolution, for example in PET and HEP. Having a complete overview on all of these parameters allows to draw conclusions on how best performances can be achieved for the various needs of different applications.

Contents

* Corresponding author.
** Corresponding author at: CERN, 1211 Geneve 23, Switzerland.
E-mail addresses: acerbi@fbk.eu (F. Acerbi), stefan.gundacker@cern.ch (S. Gundacker).

https://doi.org/10.1016/j.nima.2018.11.118
Received 3 September 2018; Received in revised form 26 November 2018; Accepted 26 November 2018
Available online 3 December 2018

1. Introduction

The silicon photomultiplier (SiPM) (also solid-state photomultiplier, SSPM, or multi pixel photon counter, MPPC) is a solid state photodetector made of an array of hundreds or thousands of integrated single-photon avalanche diodes (SPADs), called microcells or pixels [1–6]. All cells are independent and connected to a common readout. In analog SiPMs each cell has a quenching resistor and they are connected in parallel. Each cell is typically square with an edge length between less than 10 μm [7] and 100 μm [8].

Upon the detection of a photon the SPAD generates a large electric output signal due to internal avalanche multiplication. In a SiPM it is possible to count each fired SPAD separately: (i) in a digital fashion (digital SiPM), where each SPAD is connected to its own readout electronic circuit [6,9] or (ii) by the amplitude (or charge) of the sum of the single SPAD signals like in an analog SiPM [1,2,10]. Either way, the SiPM allows to detect and count photons with good resolution and with single-photon sensitivity [1,11,12]. The internal avalanche amplification is as well fast enough to obtain very good timing information of the arrival time of the detected photons [13–15], within several tens of picoseconds.

These properties, along with advantages like low bias voltage, compactness and robustness, makes the SiPM a good device for light detection from single photon to several thousand of photons, especially when fastest timing is a requisite. Typical applications based on low light intensity are light detection and ranging (LIDAR) [16,17], functional optical spectroscopy and fluorescence light detection in biology and physics [18–21], quantum physics [22] and quantum informatics [23], etc. Coupled to organic or inorganic scintillators the SiPM sense the scintillation light and/or Cherenkov light [24,25] with highest time precision. They are used in nuclear medical imaging [26–30], for gamma spectroscopy and for time tagging of high energetic particles [5,31–33]. In these applications they exploit their higher granularity with respect to PMTs and their insensitivity to magnetic fields. For example, in oncological diagnostics time of flight (TOF) in positron emission tomography (PET) was resumed after the first studies in the ’80s by the commercial availability of high performance SiPMs around 2010. New applications like in the search for dark matter or double beta decay demand novel developments of the SiPM to extend its photon detection efficiency (PDE) towards the vacuum UV (VUV) or deep UV. On the other hand the already mentioned LIDAR market calls for a high PDE on the other side of the spectrum in the near infrared. Furthermore new challenges in TOF-PET, high energy physics, time resolved X-ray detection and spectroscopy push developments of the SiPM to achieve single photon time resolutions (SPTRs) as good as 10 ps.

The SiPM is an already established photodetector having entered many fields of basic scientific research to social and medical applications; however it is still a device with plenty of room for further developments. This paper will give an overview of the basics of a SiPM, in order to get a deep understanding of its working principles and main parameters. It starts with the description of the SPAD physics itself, the equivalent electrical model of SPADs and SiPMs followed by a definition of SiPM parameters and characterization methods with state-of-the-art results. An additional focus will as well be given on phenomenological simulations in applications like TOF-PET and new arising challenges in this field.

2. Structure and simulation models

2.1. Single photon avalanche diode

Single photon avalanche diodes (SPADs), also called Geiger-mode avalanche photodiodes (Gm-APD) are solid-state single-photon sensitive photodetectors. They exploit avalanche multiplication as internal gain mechanism. The avalanche breakdown process has been studied in the ’60s and ’70s using avalanche photodiodes operated close to breakdown voltage or above [34–37]. Instead, the first avalanche photodiodes working above breakdown, in Geiger mode, have been proposed and studied in the ’80s and ’90s [38–40,13]. Nowadays, SPADs are realized in silicon, with custom or CMOS processes, with quenching and readout circuitry in-pixel, or made in different materials, like III/V materials for near-infrared range detection.

A SPAD is essentially a p–n junction, specifically designed to be biased above the breakdown voltage [13,38]. In such conditions, the electric field is so high (in the order of few 10^5 V/cm) that a single carrier injected or generated into the depletion layer can trigger a self-sustaining avalanche process. The current increases rapidly to a macroscopic level and the leading edge of the avalanche current pulse marks with good time resolution the arrival time of the detected photon. The current theoretically would continue to flow until the avalanche is quenched by lowering the bias voltage to or below the breakdown voltage, by a so called “quenching circuit”. The bias voltage must then be restored in order to be able to detect another photon (reset phase).

The quenching circuit is a series resistor, with a relatively high resistance value: when the current in the SPAD increases, due to avalanche build-up, the voltage drop at the quenching resistor rises, thus the voltage at the SPAD consequently decreases, reaching values close to the breakdown voltage. Then the bias is restored through the same resistor, with a time constant:

$$\tau_{reset} = R_q \cdot (C_{SPAD} + C_{node}) \quad (1)$$

where R_q is the quenching resistor, C_{SPAD} is the junction capacitance of the SPAD, and C_{node} is the total capacitance at the high-impedance SPAD node, beside the SPAD one (e.g. the parasitic capacitance of connections or the capacitance given by the quenching resistor, or quenching circuit). In analog SiPMs the capacitance at the node is mostly the one given by the integrated quenching resistor, which is placed on the top of the cell junction, C_q in parallel to R_q.

Fig. 1 shows an example of the structure of a p-on-n silicon SPAD, the reverse current vs bias curve with avalanche quenching and reset phase, the typical single-SPAD readout circuitry and the combination of SPADs to form an analog SiPM.

For single SPADs (i.e. SPADs with external not-integrated quenching circuit) or typically for CMOS SPADs, the passive quenching has been generally replaced by active quenching or mixed-active–passive quenching solutions [13], where a transistor is used to force the bias to either quench or reset the SPAD. With such active solutions, the recharge is faster and the dead-time (i.e. the time when the SPAD is not sensitive) can be set and is well defined (no more exponential recharge where SPADs gradually become more and more sensitive [13]).

Generally speaking, the closer the quenching circuit is to the SPAD, the smaller is the amount of charge flowing per avalanche, since the detection of the avalanche triggering happens earlier and the parasitic capacitance C_{node} is reduced. In a CMOS “active pixel” the quenching circuit is realized in each pixel which can be very compact with the pixel

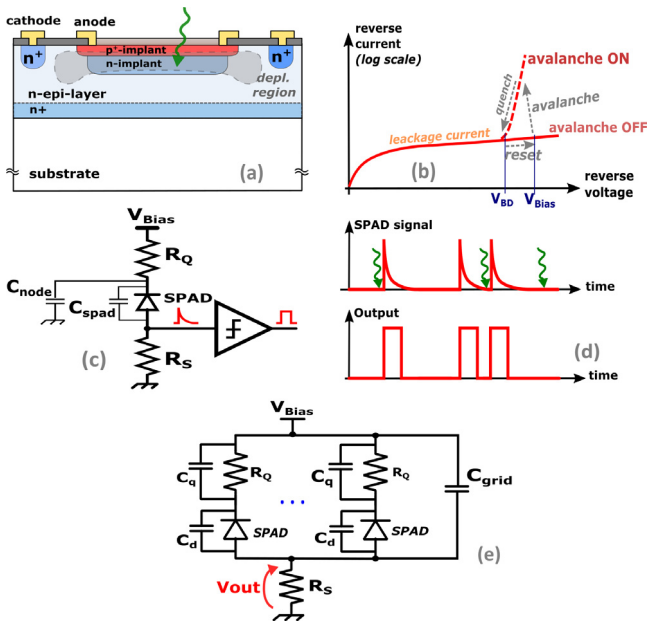

Fig. 1. (a) Example of p-on-n silicon SPAD structure, (b) reverse current vs bias curve, with avalanche, quenching and reset phase, (c) typical single-SPAD readout circuit, with discriminator to digitize the signal, (d) example of analog output and digitized signal, and (e) typical analog SiPM circuit, composed by many SPADs with integrated quenching resistors (R_q), and with sensing resistor (R_s).

pitch in the order of few tens of micrometers. Nevertheless, the active quenching circuit is area-consuming and this is particularly relevant for example when building an imager, or when the SPADs are integrated in a 2D array, creating a silicon photomultiplier (SiPM). In such cases, the fill-factor (FF) of the pixel is very important. It has to be as high as possible to increase the photon detection efficiency (which includes the geometrical FF). In these cases, passive quenching is preferable: the resistor is very compact and for example, in analog SiPMs it can be realized at the side of the SPADs active area, without any significant FF reduction [17].

2.2. Avalanche process

In a diode the reverse current–voltage curve generally shows a divergence at a certain bias voltage. This is called the breakdown voltage (V_{bd}): it is defined as the voltage where the multiplication factor (M) [41], i.e. the number of secondary carriers produced per each primary one diverges. The breakdown voltage is the edge between the linear multiplication (typical of APDs) and the diverging "avalanche" breakdown (typical of SPADs), where the photodetector works in Geiger mode. The breakdown voltage can be easily identified in externally quenched single SPADs, where the reverse current–voltage curve starts to rise very steeply. However, in recent SiPM cells, i.e. SPADs with integrated quenching resistor, the so called "gain" (the amount of charge flowing per avalanche) is made small, in the order of 10^5 electrons (compared to 10^7–10^8 in externally-quenched SPADs) [13]. This makes the current rise not as accentuated as in the externally quenched SPADs, thus the breakdown voltage identification can be less straightforward. Moreover, sometimes it can happen that an intense leakage current (see Fig. 1b), combined with a small bulk generation of the SPAD(s) prevent the correct identification of the breakdown voltage from the current–voltage curve. The bulk generation is the multiplied current or avalanche pulses from the depleted region, whereas the leakage current is due to current flowing at the periphery of the device (e.g. at the surface) and not multiplied.

The value of the breakdown voltage depends on the internal structure of the diode (particularly the doping profiles at the p–n junction) and on the temperature. The "ionization integral" is defined as in Eq. (2) and conventionally at the breakdown voltage it is considered equal to one [42].

$$\int_0^W \alpha_n \cdot \exp\left(-\int_0^x (\alpha_n - \alpha_p) dx'\right) dx = 1 \tag{2}$$

Here α_n and α_p are the ionization coefficient of electrons and holes respectively and W is the depleted region width. Different authors report different models and different ionization coefficient values (see for example the comparison in [42]). It can be divided among local models and non-local models, taking into account the history of the particle. Van Overstraeten–de Man [43] and Okuto–Crowell [44] are among the most used models in device simulators [45]. Generally, ionization coefficients show an exponential dependency on the electric field. The dependence on temperature, instead, can be qualitatively understood thinking about the physics of the avalanche multiplication: every carrier entering the high-field region is accelerated by the electric field and has a certain probability to hit another atom of the reticle or a phonon while traveling the depleted region. In case of a collision with an atom, there is a minimum energy, called threshold energy to produce a second electron–hole pair (impact ionization), which as a first approximation is $E_{th} = 3/2E_g$ [46] (when the effective masses of electron and holes are assumed equal). Every time the accelerated carrier with energy below impact ionization hits an atom or a phonon it looses energy without generating secondary carriers. The higher the temperature, the higher is the rate of collisions and, hence, the smaller is the average energy of the accelerated carriers. Thus, the critical electric field needed to generate a self-sustained avalanche process increases with the temperature and therefore the breakdown-voltage increases as well.

In addition, the width of the depleted region affects significantly the breakdown voltage. Supposing (as a first approximation) that the electric field is constant inside the depleted region, then in a wider depleted region there will be a higher number of possible collisions, thus a higher probability of impact generation. Therefore, a wider high-field region has a lower critical electric field (at breakdown), which is generally preferable to reduce the field-enhanced noise generation. However, despite the lower peak field, the breakdown voltage is higher, since the depleted region is wider. Indeed, the breakdown voltage is the integral of electric field over the whole width. This has to be considered also to understand the temperature dependence. The critical electric field (at breakdown) increases with temperature. With a wider depleted region, the increment in critical electric field will be multiplied for a higher value, thus giving a larger variation with temperature. Vice-versa, with a narrower depleted region, the temperature dependence will be smaller.

Note that the depleted region width considered here (for breakdown voltage variation) is just the one at the breakdown voltage: the depleted region width generally varies in the device with the applied bias.

What was discussed so far is related mostly to avalanche build-up and breakdown voltage. Avalanche quenching, as described above, is made with simple or more complicated avalanche circuits. Considering a simple passive-quenching, after ignition and avalanche build-up the current discharges the capacitance at the high-impedance node $C_{SPAD} + C_{node}$ and the voltage across the junction falls towards the asymptotic value V_f given by: $V_{bd} + R_d I_f$, where $I_f = V_{ov}/R_q$ is the steady state current and R_d the internal SPAD resistance. The avalanche multiplication process is stochastic and when I_f is small enough and V_f very close to the breakdown voltage, the amount of multiplied carriers is reduced and it can happen that instantaneously the avalanche process is no more self-sustaining. This leads to a progressive reduction of the avalanche current, thus the avalanche is quenched. The threshold current value for avalanche quenching is not well defined and probably it depends on the internal structure, the high-field region depth, the peak electric field and the overall active area extension. In literature it is reported that the avalanche seems to extend to a

diameter of about 10 μm before quenching [47], for different active area dimensions. As a rule of thumb it is considered that 20 μA is a correct value for the steady-state current for a prompt passive-quenching [13].

2.2.1. Electric field uniformity

So far we discussed about a one-dimensional analysis of the breakdown. As a better approximation, the p–n junction has to be analyzed at least in two dimensions, to account for the "edge effects". In the center of the p–n junction the previous one-dimensional analysis is valid, whereas at the edges there can be edge premature breakdown issues or a non-uniform electric field going from the center to the edges of the p–n junction. In the former case, the curvature effect of the p–n junction creates a higher electric field at the edges, thus a higher multiplication factor and a smaller "local" breakdown voltage at the edges. At the breakdown voltage, identified by the reverse I–V curve, only a small portion of the diode area is actually working in Geiger mode, whereas the central part is not. Different "local breakdown voltages" can be estimated on the different carrier paths [48].

To avoid edge breakdown, typically a guard-ring or a virtual guard-ring structure is used. Examples of this implementations in custom processes or CMOS processes can be found in [49,50]. Fig. 2(a) shows an example of a guard-ring (GR) in a CMOS-compatible SPAD, created with p+ implant and n-well, using p-well for GR, whereas Fig. 2(b) shows a typical custom-process SPAD, using a p-enriched implant to create a "virtual guard ring" (i.e. higher depletion at the edges). The electric field at the edges is effectively reduced, avoiding edge breakdown, but as a secondary effect this slightly reduces the effective active area. This is due to the transition region between the central high-field area to the edge, where the electric field "gradually" reduces. Moreover, carriers in the border region are not drifted vertically, as in the central region, but laterally towards a lower-field region. As a result the effective active area is smaller than the nominal (layout) one. This transition region can be in the order of 1–2 μm, thus negligible in big area SPADs (e.g. 50 μm), but important in small SPADs.

2.2.2. Photon detection efficiency

Photon detection efficiency (PDE) quantifies the ability of a single-photon detector to detect photons. This is the ratio between the number of detected photons and the photons arriving at the detector. The PDE is calculated as in Eq. (3).

$$PDE(V_{OV}, \lambda) = QE(\lambda) \cdot P_T(V_{OV}, \lambda) \cdot FF_{eff}(V_{OV}, \lambda) \tag{3}$$

Here QE is the quantum efficiency, P_T the avalanche triggering probability, V_{OV} is the overvoltage and FF_{eff} is the effective geometrical fill-factor. The geometric fill factor is typically not included in the PDE calculation when characterizing a single SPAD (in this case normally the photon detection probability PDP is quoted), but it has to be considered when it is part of an array or of a SiPM.

The quantum efficiency includes the probability of a photon to enter into the detector (i.e. without being reflected at the surface) and then to be absorbed in the "useful" part of the device, i.e. where the photo-generated carriers have some chance to reach the active region before being re-absorbed. The avalanche triggering probability depends on the electric field, thus on the overvoltage and on the position where the carriers are generated. In particular, it is significantly different for photo-generated electrons or holes, since the ionization coefficient for electrons is typically higher than the one for holes [37].

The PDE of a SPAD or SiPM based on a p-on-n junction is quite different from the one of an n-on-p junction [51]. Referring to a structure like the one in Fig. 2(b), we can consider two cases: (1) a p-type shallow implant with n-type enrichment (n-type epi/substrate) and (2) an n-type shallow with p-type enrichment (p-type epi/substrate). Supposing a junction depth of few hundreds of nanometers, the majority of photons in the green and red part of the spectrum are absorbed in depth, beneath the junction, whereas the photons in the blue part are absorbed close to the surface, above the junction. Only electrons in p-on-n junction type and only holes in n-on-p junction type, will trigger the avalanche process when photons are absorbed close to the surface [52,53]. The photo-generation happens always above the junction, thus only one type of carrier is drifted towards the high-field region and can trigger the avalanche. Instead, for green and red photons, light is absorbed above or below the junction (the high field region). Photo-generated electrons or holes can trigger the avalanche (not just one of them), from above or below the junction. For red and NIR photons mostly holes (in p-on-n junction) or mostly electrons (n-on-p junction) are generated and trigger the avalanche (assuming the collection region below the junction is bigger than above the junction). Even considering a fixed depleted region width and the same device structure for the two junction type cases (thus the same QE), the different type of carriers triggering the avalanche give a different spectral shape of the PDE [51–53].

Finally, the effective FF depends specifically on the layout and on the internal SPAD structure, in particular on the uniformity of electric fields and on the border effect. As described in the previous section, at the border the electric field is lower and there are also depletion effects which make the photo-generated carriers in those regions to drift laterally instead of vertically, thus not triggering any avalanche (see Fig. 2(c)) [7,15]. This border effect can be very important in small SPADs (i.e. microcells in SiPMs). It reduces when increasing the overvoltage, due to saturation of the avalanche triggering probability with a higher electric field. Moreover, the border region created by the border effect is different depending on the depth in the SPAD [52], thus the FF can be also considered partially dependent on the wavelength.

The PDE can be simulated in different ways, for example with the TCAD simulation software [45]. Here, instead, we report just a simple example of QE and PDE estimation based on effective values of FF and the collection depths. Fig. 3(left) shows the estimated internal quantum efficiency (QE) for different effective absorption region thicknesses (i.e. "effective" epitaxial layer thickness), whereas Fig. 3(right) shows the estimated PDE, for two different absorption region thicknesses and for n-on-p or p-on-n structures. This estimation has been done considering as a first approximation the absorption region divided in two, where electrons or holes are the dominant carriers triggering the avalanche, see Eq. (4).

$$\begin{aligned} PDE &= \left(\frac{QE_{top}}{QE(\lambda)} \cdot P_{t,h}(V_{ov}) + \frac{(1 - QE_{top})}{QE(\lambda)} \cdot P_{t,e}(V_{ov}) \right) \cdot \\ &\quad \cdot T_{ARC}(\lambda) \cdot QE(\lambda) \cdot FF \\ QE_{top} &= \left(1 - \exp\left(-\frac{x_d}{l(\lambda)}\right)\right) - \left(1 - \exp\left(-\frac{x_1}{l(\lambda)}\right)\right) \\ QE &= \left(1 - \exp\left(-\frac{T_{abs}}{l(\lambda)}\right)\right) - \left(1 - \exp\left(-\frac{x_1}{l(\lambda)}\right)\right) \end{aligned} \tag{4}$$

QE is the internal quantum efficiency, QE_{top} is the quantum efficiency considering the absorption only in the upper region of the junction, x_d is the junction depth (in this particular example ~500 nm), x_1 is the top region thickness where carriers recombine before reaching the depleted region (in this particular example 5 nm), T_{abs} is the absorption region thickness, T_{ARC} is the transmission of anti-reflective coating (calculated for each wavelength) and l is the absorption length. $P_{t,e} = 0.86$ and $P_{t,h} = 0.39$ are the triggering probability for electrons and holes, calculated at 4 V overvoltage using the plots and fitting parameters reported in [52]. FF in this case is 0.72. The reported formula is valid for p-on-n junction types. For n-on-p type $P_{t,e}$ and $P_{t,h}$ have to be exchanged.

2.2.3. Gain and amplitude

The gain of the single cell represents the number of carriers flowing per each triggered avalanche. In analog SiPMs it is generally "well defined" due to the integrated resistor, the internal capacitances and the good uniformity obtained in the fabrications processes. This is in contrast for example with single SPADs with external quenching circuit,

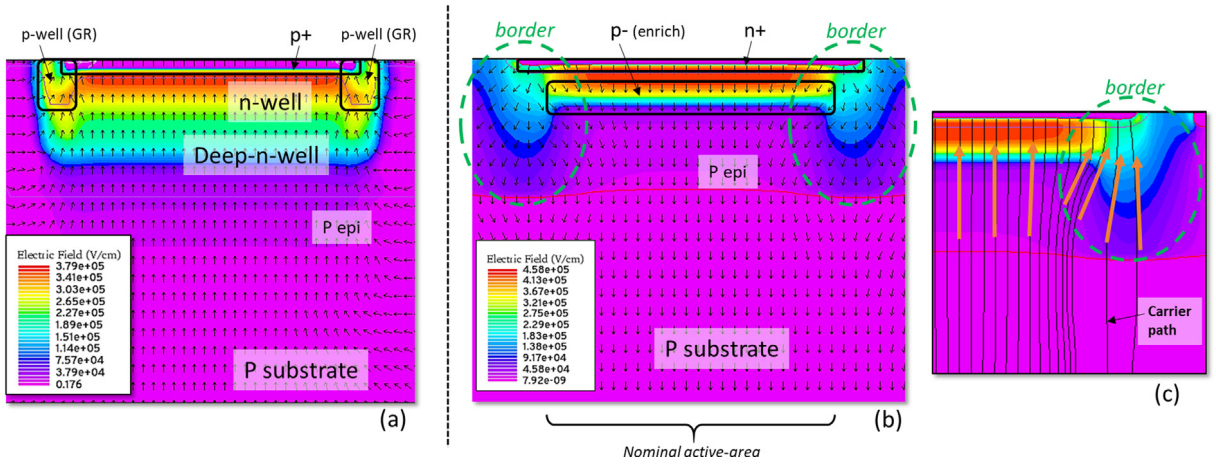

Fig. 2. TCAD simulation of electric field inside SPADs. The structures are examples of a CMOS SPAD with guard ring made with p-well (a), of a custom process SPAD with p-type enrichment implant (b) and a detail of the carrier path in this second structure at the border (c). Arrows represent the local current direction. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

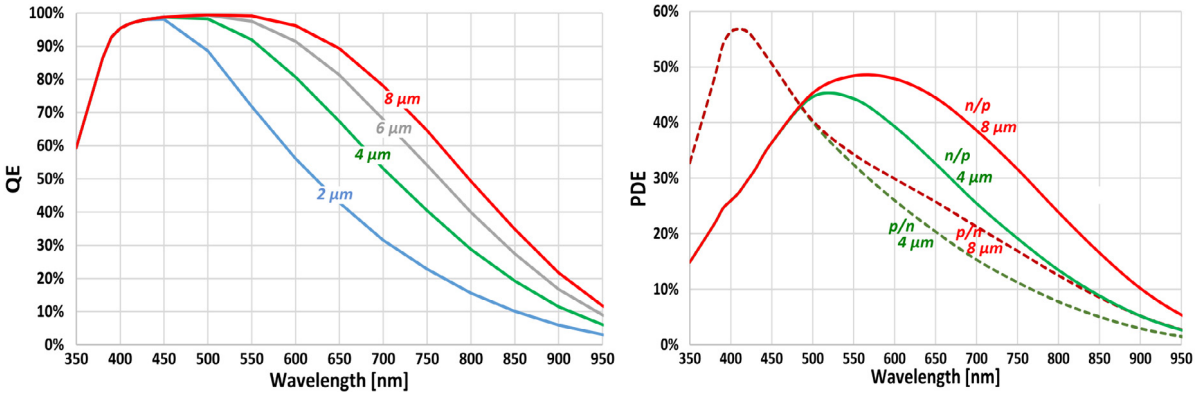

Fig. 3. Estimated internal QE and SiPM PDE for different absorption region (or epitaxial layer) thickness, and for p-on-n and n-on-p junction structure types.

where bonding wires, connections and different front-end readout can change significantly the value of C_{node}, thus the amount of charge flowing per avalanche.

The well defined gain gives the analog SiPM very good photon number resolving capabilities, which can be estimated by the amplitude spectrum or by the charge spectrum, when the signal is integrated over a certain time synchronous with the light pulses [11,12]. For an analog SiPM cell, as in Fig. 1(e), the average gain is generally expressed, as in Eq. (5).

$$Gain = \frac{avalanche\_charge}{q} = \frac{V_{ov} \cdot (C_q + C_d)}{q} \tag{5}$$

With q denoting the elementary charge $q = 1.602 \cdot 10^{-19}$ C. The gain is typically in the order of 10^5 to 10^7 [54,11] and produces a single photon signal well above the electronic noise level. Hence, constraints on the readout electronics are not as severe as in the case of APDs or PIN photodiodes. The excess noise factor component due to the variation of the gain, defined as F, or ENF is defined as $1 + \sigma_G^2/\langle G \rangle^2$, where σ_G is the standard deviation of the gain fluctuation, is almost unity ($F \approx 1$). In SiPMs, instead, other components are more important and dominate in defining the excess noise factor, in particular: the correlated noise (afterpulsing and optical crosstalk) and saturation effects. The mean charge at the output is therefore not just proportional to the input number of detected photons (plus noise) multiplied by the average gain of the SiPM cells, but due to correlated noise the mean and the variance are larger. This enlargement can be quantified by the excess charge factor (ECF) and the ENF:

$$ENF = \frac{SNR_{in}^2}{SNR_{out}^2} = \frac{\sigma_{Q,out}^2/\langle Q_{out} \rangle^2}{\sigma_{Q,primary}^2/\langle Q_{primary} \rangle^2} \tag{6}$$

$$ECF = \frac{\langle Q_{out} \rangle}{\langle Q_{primary} \rangle} \tag{7}$$

The internal capacitances, C_d and C_q depend on the SPAD dimensions (and layout). A smaller cell gives a smaller gain, thus a smaller number of carriers flowing during the avalanche. This could give a worse photon number resolution (since the peaks in the charge spectrum are less separated), but it reduces the correlated noise (as described in the next sections) and it makes the recharge of the single cell faster. The former aspect gives a better peak-to-valley ratio [11], whereas the second aspect makes it possible to integrate for a smaller time window, integrating less dark noise.

Examples of gain (average) and signal amplitude of SiPMs (NUV-HD from FBK) with 40 µm and 25 µm cell pitch are shown in Fig. 4. As in the reported example, in some devices the gain dependence on overvoltage is not linear. This is due to the progressive depletion of the epi-layer beneath the p–n junction, leading to a diode capacitance reduction with increasing bias voltage, thus a non linear gain dependence. Evidence of this can be seen in previous works [11]. TCAD simulations of the exact cell structure (not shown here) confirm that the depletion of the epi-layer is gradually increasing with the overvoltage of the SiPM.

The amplitude of the signal instead, might be not entirely related to the gain. It is true that bigger cells (i.e. SPADs) have typically higher gain and larger signal amplitude, but the single-cell signal is usually composed by a fast component and a slow component [10,55]. The fast

component is due to the capacitive coupling of the avalanche signal through the quenching capacitance (C_q) and then partitioned towards the output (due to the other capacitances and parasitics), whereas the slow component is due to the recharge current flowing through the quenching resistor (R_q), being higher at the beginning and then exponentially decreasing (exponential recharge). The front-end electronics has a significant role in signal shaping, depending on its bandwidth the fast component can be prompt and very high in amplitude or can be filtered out completely. Fig. 4(b) shows an example of measured peak amplitude of SiPMs with 40 μm and 25 μm cell pitch. In Fig. 5 the corresponding signal shape can be seen. In this measurement the SiPM was amplified by a front-end based on the AD8000 operational amplifier chip, in a trans-impedance configuration, with a second amplification stage. Total trans-impedance gain is 5000 V/A, considering the 50 Ω termination of the oscilloscope input. It can be qualitatively inferred that the fast component has a frequency content in the range of 50–250 MHz. Similar low-pass filtering effects can also be given by the total capacitance of the SiPM itself (i.e. sum of grid capacitance and the overall capacitance of all not triggered cells) and by further parasitics of the package. A big SiPM, e.g. 6 × 6 mm^2 or 10 × 10 mm^2, has a large capacitance. Without a very small input impedance of the front-end an output-current low-pass filtering effect can be observed, smoothing the fast component similar to a reduced front-end bandwidth [14].

Fig. 6 shows an example of charge spectrum and amplitude spectrum, acquired with a 1 × 1 mm^2 SiPM with 40 μm pitch. In this example we used a pulsed LED illumination, with narrow pulses and the acquisition was synchronous with the light. The amplitude spectrum acquired with lower bandwidth shows a better peak-to-valley ratio. Indeed, amplitude spectra are generally significantly affected by the electronic noise of the front-end and by peak amplitude oscillations. Both are generally reduced when the signal is low-pass filtered. The charge spectrum instead is not affected by the bandwidth (as a first approximation) but is more affected by the length of the integration window and by dark count rate and correlated noise probability. Indeed, due to delayed correlated noise there can be subsequent pulses (after the primary one), with time delays between few to hundreds of nanoseconds. Afterpulsing generates pulses with "fractional" charge (smaller than the single photo-electron charge), whereas delayed crosstalk and primary noise could generate pulses at the edge of integration window, thus being partially integrated. Both effects increase the "valley amplitude" in the charge spectrum, thus worsening the peak-to-valley ratio. There is typically an "optimum" integration time, reducing the amount of fractional-charge events but preserving the useful information about primary pulses.

2.3. SPAD equivalent electrical model

The analog SPAD with integrated quenching resistor is usually modeled as a parallel connection between the internal resistance of the diode space-charge region R_d and the inner depletion layer capacitance C_d, which itself can be the sum of the SPAD area capacitance and any kind of perimeter capacitance (see Fig. 7). When the SPAD has the quenching resistor integrated, i.e. like in analog SiPM cells, the model includes also the quenching part [10]. This is described by the quenching resistor R_q and a "parasitic" (or designed) capacitance C_q in parallel, which can be beneficial to increase fast signal extraction [14,56]. A detected photon (or noise event) triggering the avalanche in the SPAD is modeled by closing a switch in Fig. 7 [13,57]. A current starts to discharge the internal node with total capacitance equal to the sum of C_q and C_d through R_d [57,58]. The voltage drop at the internal node is almost equal to the overvoltage, i.e. the difference between cathode–anode reverse bias and the breakdown voltage, modeled as the DC power supply V_{bd}. Indeed, as reported in Eq. (5), the avalanche charge is expressed as the overvoltage times the sum of the two capacitances.

In literature, for the SiPM equivalent electrical models, the avalanche is sometimes modeled via a pulsed current source [10,59] in place of the series of the switch, the breakdown DC power supply and R_d [58]. These

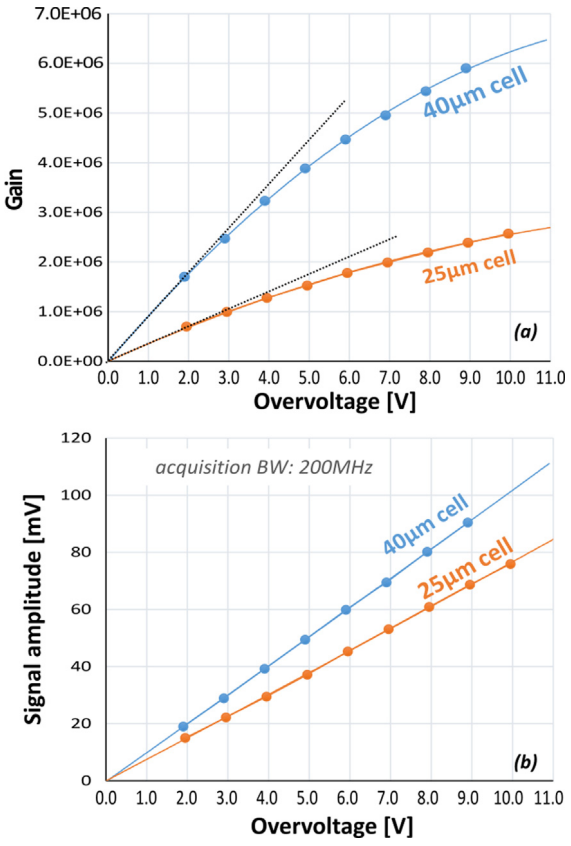

Fig. 4. Example of measured cell gain of FBK NUV-HD 2018 1x1 mm^2 SiPMs with 40 μm and 25 μm pitch, as a function of overvoltage (a). Example of single-cell signal amplitude of SiPMs with 40 μm and 25 μm pitch (acquisition bandwidth 200 MHz) (b).

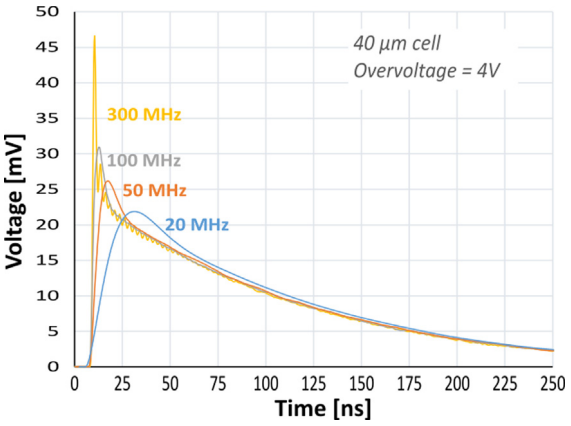

Fig. 5. Effect of acquisition bandwidth on signal shape, measured with a 1 × 1 mm^2 SiPMs (FBK NUV-HD 2018) with 40 μm pitch.

two approaches are to a certain extend equivalent, especially when the user is only interested in simulating the output signal from the SiPM. However, the approach with the switch and the DC V_{bd} source can simulate, if needed, the avalanche quenching or non-quenching.

Before photon detection (switch open) C_d is charged to the SiPM bias voltage V_{bias} applied on the anode and cathode with the positive voltage on the cathode. Upon photon detection (or in general avalanche triggering by photon or dark counts), the switch in Fig. 7 is closed, leading to a discharge of the capacitance C_d via the resistor R_d. The initially high current I_d (in the range of several milliamperes) is given by the overvoltage V_{OV} divided by R_d [13]. V_{OV} is the difference of the applied bias voltage V_{bias} and the breakdown voltage V_{bd}. The voltage

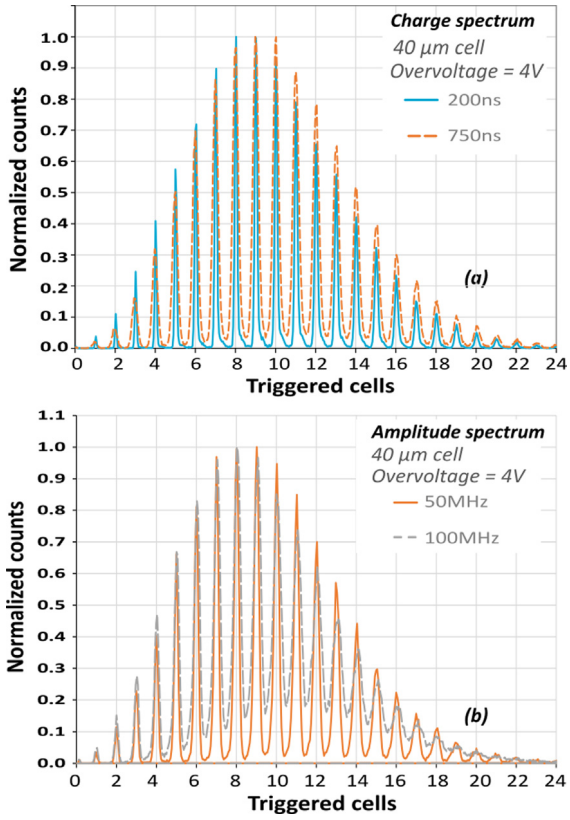

Fig. 6. Examples of charge spectrum (a) and amplitude spectrum (with fixed integration window) (b), showing the effect of different integration times (200 ns and 750 ns) and bandwidths (50 MHz and 100 MHz).

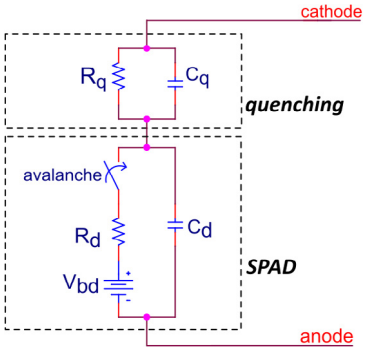

Fig. 7. Equivalent electrical circuit of the single photon avalanche diode (SPAD) with integrated quenching resistor.

drop on C_d provokes a similar voltage change on C_q producing an external current spike giving rise to a fast initial signal part in the front-end electronics. The intrinsic limit of the signal rise time (i.e. when not limited by the amplifier or front-end bandwidth or slew-rate) is given by $\tau_r = (C_d + C_q) \cdot (R_d \parallel R_q)$ [10] and can be in the range of tens of picoseconds only. The discharge of C_d and recharge of C_q is stopped when the current I_d through R_d reaches a certain value, i.e. the "threshold current", which is the minimum value of current to get a self sustainable avalanche process. This value is a bit higher then the asymptotic minimum value I_{df} that I_d would reach. This asymptotic current I_{df} is given by the overvoltage V_{OV} divided by $R_q + R_d$ [58], i.e. $I_{df} = V_{OV}/(R_q + R_d) \sim V_{OV}/R_q$. Once the avalanche is quenched the cell recovery or recharge time is given by $\tau_{recharge} = R_q \cdot (C_q + C_d)$. Hence, as described in the previous chapter, in the passively quenched

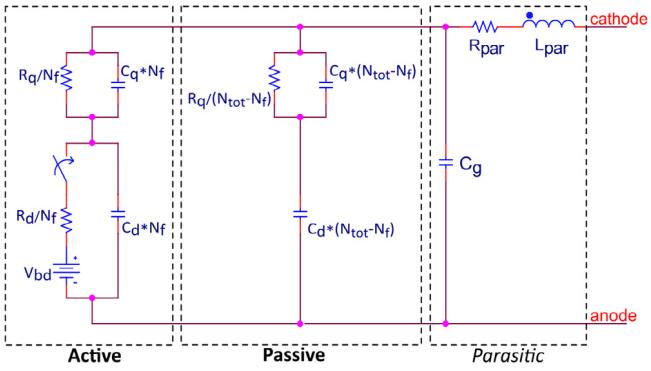

Fig. 8. Equivalent circuit of the SiPM, including the switch that mimics avalanche triggering, the diode series resistance R_d, the diode capacitance C_d, the quenching part R_q and C_q, the parasitic part from the not triggered cells and the series and parallel parasitic components.

SPAD a fast signal component followed by a slow component can be observed.

The model described here assumes that the avalanche stops when the voltage across the junction reaches V_{bd}, where the avalanche can no more self-sustain. This naturally gives the correct gain of the SPAD with the integrated quenching circuit defined as $V_{OV} \cdot (C_q + C_d)/q$. Recently it was reported in [60] that a voltage at which the avalanche stops (i.e. extrapolated from linear fit on the gain plot) is ~1 V lower than V_{bd} for a SiPM with 15 µm pitch (following the model in [61]). However, there is no general agreement on this point and the majority of measurements in literature are compatible with the assumption that the avalanche stops at V_{bd}.

2.4. SiPM equivalent electrical model

The most common representation of the SiPM equivalent electrical circuit can be seen in Fig. 8. The differences to the SPAD equivalent circuit are the additional passive components of the other not-triggered SPADs plus the grid inductances and capacitances [10,57,58]. The output current produced by the triggered cell(s), especially the fast component via C_q is divided by the parasitic capacitance C_g (i.e. mainly the capacitance of the metal grid and bonding pads) and the series connection of the passive capacitances $(N_{tot} - n_f) \cdot C_q$ and $(N_{tot} - n_f) \cdot C_d$ of the $N_{tot} - n_f$ inactive cells of the SiPM [55]. In the schematic n_f represents the number of triggered cells of a total of N_{tot} cells available in the SiPM, thus $N_{tot} - n_f$ is the number of not triggered cells. Hence, depending on the input impedance of the used front-end electronics, the grid capacitance and the inactive cells can cause a noticeable drop in the single SPAD signal amplitude. Using an amplifier with very low input impedance, e.g. few Ohms (like transimpedance amplifier) the impedance of the passive and parasitic part is generally higher than the input impedance of the amplifier and this helps in extracting the current signal [62]. The parasitic trace (and bonding) inductance L_{par} and resistance R_{par} limit the SiPM signal rise time and should be minimized in a proper printed circuit board (PCB) design in order to benefit from the intrinsically fast rise time of the SiPM signal for fastest timing.

In the frequency domain the absolute impedance of the SiPM can be obtained as in Eq. (8). This function can serve to estimate the SiPM's equivalent circuit component values by measuring $|Z|$ in the frequency range and fitting to the model [63].

$$|Z(\omega)| = \left| j\omega L_{par} + R_{par} + \left(j\omega C_g + N_{tot} Y_{cell} \right)^{-1} \right| \tag{8}$$

$$Y_{cell} = \left[\frac{1}{j\omega C_d} + \frac{1}{R_q^{-1} + j\omega C_q} \right]^{-1} \tag{9}$$

Y_{cell} (also called Y_{pix}) is the admittance of the single cell, which is then multiplied by the total number of pixels in the $|Z(\omega)|$ calculation, since

they are all supposed inactive. For low to intermediate frequencies $|Y_{cell}|$ is dominated by C_d. The capacitive component of $|Z|$ is $\sim N_{tot} \cdot C_d$. At high frequencies instead the contribution of C_q dominates over the R_q one, thus the capacitive component of $|Z|$ is $\sim N_{tot} \cdot \frac{C_d \cdot C_q}{C_d + C_q}$.

Because of epitaxial layer depletion, the impedance will depend on the absolute bias of the measurement. The smaller depletion of low bias will give a higher C_d, whereas a progressively higher depletion reaching the breakdown voltage will lower the value of C_d. In case the change of depletion region thickness is relevant above breakdown, there will be a dependence of the measured impedance on the overvoltage, even though direct impedance measurement above breakdown is generally difficult. Moreover, a change in the equivalent circuit as a function of the number of pixels firing can be seen, leading to a slight change in the output signal (due to different load and charge partitioning).

Neglecting the parasitic part in Eq. (8) the complex impedance $Z(\omega)$, or the complex admittance $Y(\omega)$, can be rewritten as in Eq. (10) [58], further assuming that no cells in the SiPM are firing.

$$Y(\omega) = \left(\frac{1}{G(\omega)} \parallel \frac{1}{j\omega C(\omega)} \right)^{-1} = G(\omega) + j\omega C(\omega)$$
$$= \left[\left(\frac{R_q}{N_{tot}} \parallel \frac{1}{j\omega N_{tot} C_q} + \frac{1}{j\omega N_{tot} C_d} \right) \parallel \frac{1}{j\omega C_g} \right]^{-1} \quad (10)$$

where $G(\omega)$ and $C(\omega)$ are the measurable parallel conductance and capacitance of the SiPM, which can be obtained with a precision LCR meter. Eq. (10) can be solved to obtain the solution for C_d and C_g from $G(\omega)$ and $C(\omega)$, as can be seen in Eqs. (11) and (12), respectively.

$$C_d = \sqrt{\frac{1 + \omega^2 (C_d + C_q) R_q^2}{\omega^2 N_{tot} R_q} G(\omega)} \quad (11)$$

$$C_g = C(\omega) - N_{tot} C_d + \frac{\omega^2 C_d^2 R_q^2 N_{tot} (C_d + C_q)}{1 + \omega^2 R_q^2 (C_d + C_q)^2} \quad (12)$$

As already mentioned the sum $C_d + C_q$ can be measured via the gain of the SiPM, $Gain = (V_{bias} - V_{bd})(C_d + C_q)/q$ and the quench resistor value R_q can be obtained from the forward current measurements, knowing the number of cells in the SiPM N_{tot}. This defines all parameters of the SiPM equivalent model except the diode series resistance R_d, which in theory could be estimated by the single SPAD signal rise time. However, other effects as bandwidth limitations of the used electronics and parasitic inductances makes the direct measurement of R_d via the signal rise time not reliable. In most of the cases, however, R_d can be assumed to be small in the range of ~1 kΩ and simulations show that changing R_d in a broad range does not impact the model output too much [58].

2.5. Noise and secondary effects in SiPMs

The noise in the SPADs and in the SiPMs can be divided in:

1. Primary noise: this identifies the avalanche pulse triggered by thermally generated carriers (possibly field-enhanced thermal generation [64]) or carriers generated due to tunneling in the high-field region [50]. Tunneling generation is important at low temperatures [65].
2. Correlated noise: this identifies all the avalanche pulses subsequent to a primary event, which are generated because of the primary ones, thus "correlated" to this one. These pulses are generated due to: (i) afterpulsing (in the same cell) or (ii) optical crosstalk (in neighboring microcells of the SPAD-array or SiPM).

Primary noise at room temperature is generally dominated by Shockley–Read–Hall (SRH) generation–recombination. The generation rate depends typically on the "quality" of the epi layer, in terms of number of deep levels and activation energy. The higher the defect concentration is the lower will the equivalent "lifetime" (τ_{SRH}) be, thus the higher is the rate of avalanche pulse generation in dark conditions, called dark count rate (DCR). It also depends on the volume of the depleted region, thus on the micro-cell fill-factor and on the depleted region thickness. NIR sensitive SiPMs, having a thicker epitaxial layer and thicker depleted region [66] have a higher DCR (if all other parameters are the same). SRH generation is also affected by the high electric field which is present in the avalanche region. This induces effects like Pole–Frenkel and thermal-enhanced trap-assisted-tunneling [64,67]: all these effects can be modeled by lowering the effective generation lifetime of the SRH process, due to the electric fields.

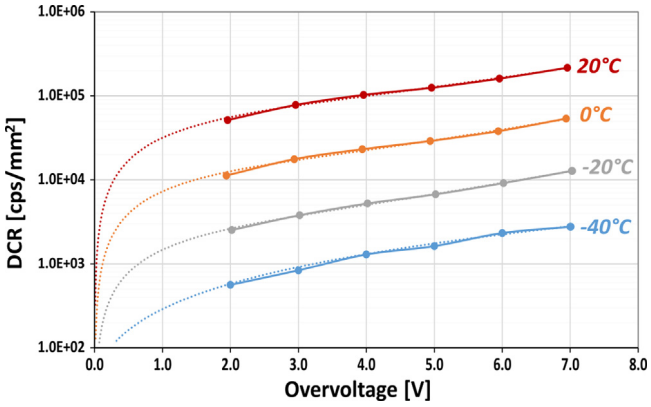

Fig. 9. Example of measured DCR of a SiPM per unit area (1 × 1 mm^2), as a function of overvoltage at different temperatures.

At higher temperatures, or in some devices already at room temperature, the diffusion current from neutral regions around the depleted one can be significant, too.

Decreasing the temperature the DCR significantly decreases. As in the example shown in Fig. 9 the DCR halves every 10 degrees. However, at even lower temperatures tunneling generation becomes dominant and the dependence on temperature reduces. This is particularly important for cryogenic temperatures based applications [65]. The "corner temperature" (between tunneling and thermal generations) depend on the microcell design, in particular on the electric field at breakdown voltage. The higher it is the higher is the tunneling generation contribution. Indeed, recent developments apply low-fields to reduce the DCR at low temperatures [65].

2.5.1. Afterpulsing

Afterpulsing is a correlated noise component of SPADs and SiPMs that is due to trapping and subsequent release of carriers in the high-field region. Deep levels up to shallow levels in the bandgap of the avalanche region can act as traps for the large amount of carriers flowing during the avalanche. Some of the carriers can be trapped and then released subsequently (typically exponential distribution of the release times) generating a secondary spurious avalanche. The afterpulsing probability depends on the number of the effective traps in the high-field region and on their release time constant compared to the recharge time constant of the microcells (or the hold-off time in an actively quenched SPAD). Indeed, to mitigate the afterpulsing probability, the recharge time constant (or hold-off) can be regulated to have the majority of traps released when the cell is not yet completely recharged.

Afterpulsing can also be "optically-induced", thus not related to traps in the high field region. During each avalanche, secondary photons are produced [68,69] and some of them can be re-absorbed in the same microcell in the neutral region beneath the active region. This can photo-generate carriers that can reach via diffusion the depleted region, where they can trigger a secondary spurious avalanche, as shown in Fig. 10. For optically induced afterpulsing the correspondent of the trap release time constant is the carrier lifetime in the neutral region. In the substrate this can be between few nanoseconds to hundreds of nanoseconds [69]. To reduce optically-induced afterpulsing, a low-lifetime substrate has to be used [69]. When the lifetime is much smaller

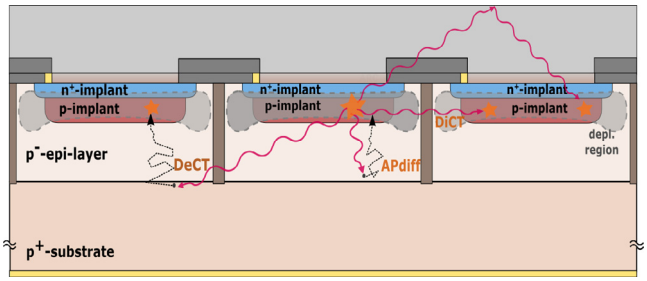

Fig. 10. Typical SiPM crosssection with the different type of correlated noise represented: direct (or prompt) crosstalk (DiCT), delayed crosstalk (DeCT), afterpulsing optically induced (or diffused) (APdiff) and external crosstalk due to reflection on the top protective layer of the SiPM (this can as well generate either direct or delayed crosstalk events). Afterpulsing due to trapping in the high-field region in not represented.

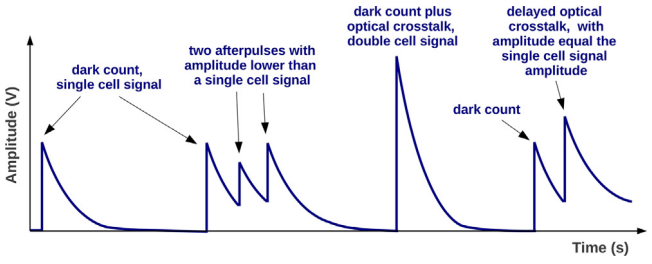

Fig. 11. Representation on the SiPM output signal of the different kinds of noise observable: primary events, prompt crosstalk, afterpulsing and delayed crosstalk events.

than the recharge time constant of the microcell, this contribution is negligible. Another possible solution is to use an inverted-doping substrate, creating a second p–n junction, blocking all carriers photo-generated in the substrate to diffuse towards the avalanche region [50].

2.5.2. Prompt and delayed cross-talk

As described above, secondary photons are produced during the avalanche. The amount has been estimated to $3 \cdot 10^{-5}$ photons per avalanche carrier [70,71]. This photon emission is isotropic and gives rise to absorption and photon generated carriers in neighboring SPADs, as shown in Fig. 10 (i.e. neighboring cells in the SiPM). This phenomena is the cause of optical crosstalk. When a photon is detected in one cell, the avalanche pulse in this cell can trigger (with a certain probability) avalanches in the neighboring cells, thus for example creating pulses with two times or three times the single-cell amplitude in a SiPM, even though the "primary" photon was only one. This type of optical crosstalk is as well called "direct" optical crosstalk, or "prompt" optical crosstalk. The output signal of the SiPM is twice as high in amplitude as can be seen in Fig. 11.

Another type of optical crosstalk can be seen on the very right in Fig. 11, the "delayed" optical crosstalk. It is caused by secondary photons generating an electron–hole pair in the bulk, or generally in the neutral regions near the depleted one (see Fig. 10). The charge carriers will diffuse and some of them can reach the active region and trigger an avalanche with a delay of several nanoseconds to microseconds. It has to be considered that due to the finite bandwidth and sampling rate of the front-end and SiPM signal acquisition, it is possible that part of the delayed crosstalk events are interpreted as prompt crosstalk. For example, if the analysis method can distinguish events down to 2 ns every delayed crosstalk happening with smaller diffusion time will be considered as a prompt crosstalk.

In Fig. 12 the frequency of dark count events and crosstalk events is shown. Measurements were performed by triggering on random dark count events by setting the trigger threshold well below the single SPAD signal amplitude and recording the corresponding charge spectrum of the resulting pulse. Hence, the crosstalk events seen in the plot

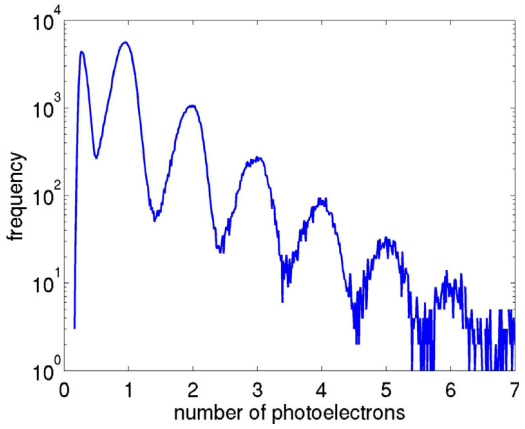

Fig. 12. Frequency of dark count events and optical crosstalk in the SiPM. Higher order optical crosstalk decreases rapidly with the number of triggered cells.

are induced by dark count events. In this example, for a Hamamatsu S10931 SiPM with 50 μm SPAD pitch, the probability of triggering N avalanches simultaneously via optical crosstalk can be described decreasing exponentially with N. Often such a simple model of the crosstalk probability is sufficient and can be used to phenomenologically include crosstalk in Monte Carlo simulations. However, there are more complex models for the optical crosstalk probabilities. Comprehensive descriptions of optical crosstalk with a description of the generalized Poisson distribution which has been shown to be a powerful tool to characterize crosstalk upon light detection can be found in [72–74]. Optical crosstalk can be mitigated by inserting optical trenches between the microcells in the SiPM [75,76,54].

3. Experimental methods and results

3.1. Signal pick-up and front-end electronics

Depending on the gain (or amount of charge per avalanche) of the SPADs or SiPM cells, the output current signal can be more or less pronounced, possibly requiring amplification. Analog SiPM signals are generally amplified in a transimpedance configuration, having a gain of 1000–10000 V/A. This gives an amplified output signal in the order of few tens of millivolts in response to a single photon. For some applications the required amplification can be smaller (about one order of magnitude lower) when it is not important to have single-photon sensitivity, for example when using the SiPM to read-out the signal from scintillators (i.e. many photons per each event).

Front-end amplifiers for analog SiPMs are generally based on transimpedance amplifiers, like for example represented in Fig. 13. The input impedance should be low to reduce the filtering effect of the grid capacitance and of parasitic components. Alternatively, a series shunt resistor to the SiPM can be used, followed by a voltage amplifier. This approach is generally not preferred since the series resistance sets compromises between amplification and bandwidth (i.e. steepness of the signal), as discussed in [62], however, it can be beneficial when used with RF amplifiers, with adapted input impedance and very high bandwidths. Recent work shows very good signal steepness with analog SiPM using baluns (transformer) to reduce the input impedance at the SiPM, followed by AC coupled high-bandwidth RF amplifiers, inspired by the patent in [77].

The signal coming from the SiPM is the superposition of many pulses. Depending on the application it is possible to measure the current or to count the avalanche pulses (photon-counting mode). When the count rate is low, the pulses are well separated but increasing the count rate they start to overlap, preventing correct photon counting. To reduce this problem some techniques can be employed:

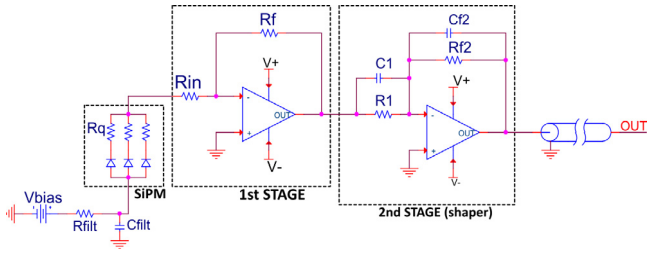

Fig. 13. Example of front-end circuit for SiPM amplification. First stage is based on a transimpedance configuration and the gain is given by R_f (expressed in volt over ampere). The second stage can further amplify and shape the signal.

1. High-pass filtering;
2. Pole-zero cancellation [78];
3. DLED (delayed leading-edge discrimination) [79];
4. Modification of the device to enhance or to extract only the fast signal component, thus minimizing or avoiding the slow recharge tail [80].

The first three methods modify the signal, "hiding" the slow recharge tail, thus making pulses more narrow and allowing photon counting up to higher frequencies. A simple high-pass filtering can leave some undershoot after the main peak, whereas pole-zero compensation, despite the need to be tuned on the specific signal, can avoid this problem.

There are other applications not based on single-photon sensitivity where SiPMs are used to read-out the signal from a scintillator (many photons per event), the readout of scintillating fibers or Cerenkov radiators, etc. The pile-up of pulses if manifesting in severe baseline shifts can be a problem in these applications, especially when doing timing measurements. The previously reported techniques can as well be applied in these applications.

3.2. Current–voltage characteristics

The current–voltage (I–V) curve of SPADs and SiPMs is an important measurement. It is used typically to get a quick overview of the functionality of the device and to estimate its breakdown voltage. Both forward bias and reverse bias curve are important in analog SiPMs.

The reverse-bias I–V curve shows the leakage current at bias lower than the breakdown voltage, whereas at higher biases the current suddenly increases. The leakage current can be almost flat or increasing with the bias, as shown in Fig. 14(a). The current I_{GM} after the breakdown voltage is proportional to the rate of dark counts (DCR) and to the gain of the cell (i.e. charge per each avalanche):

$$I_{GM} = DCR(V_{ov}) \cdot GAIN(V_{ov}) \cdot q \cdot ECF(V_{ov}) \tag{13}$$

$$DCR(V_{ov}) \propto Gen\_rate(Temp, V_{ov}) \cdot P_{trig}(V_{ov})$$

where q is the elementary charge and ECF is the excess charge factor, which quantifies the average extra charge produced per each avalanche due to correlated noise. P_{trig} is the avalanche triggering probability, which can be expressed in its simpler form as $1 - \exp(-\frac{V_{OV}}{V_c})$, where V_{OV} is the overvoltage and V_c is a "characteristic impact ionization voltage" [13], or P_{trig} can be expressed in more complex form like in [52].

Example of reverse I–V curves are reported in Fig. 14(a) and Fig. 14(b), for 1×1 mm^2 SiPMs with different cell pitches, thus different $GAIN$ and ECF, and for a 4×4 mm^2 SiPM. The forward bias curve can be used to extract the quenching resistor value, knowing the number of cells connected in parallel in the device. The slope of the curve, after the threshold voltage is indeed given by the parallel resistance of all quenching resistors in the SiPM. An example of a forward I–V curve is shown in Fig. 14(c).

The measurement of the reverse I–V curve requires typically a source and measurement unit (SMU) with several order of magnitude in dynamic range. It is important to measure from the leakage current level (typically between hundreds of picoampere to tens of nanoampere) up to the multiplied current (between hundreds of nanoampere to hundreds of microampere). In single SPADs or small SiPMs the leakage current can be one order of magnitude lower. The measurement of forward current instead can be sensitive to the series resistance of the package or to any series resistance in the setup. These can lower the measured current and prevent a correct estimation of the quenching resistors R_q of the SPADs. This issue is typically negligible in small-area SiPMs or with a small number of cells in parallel.

3.3. Time domain: SiPM signal

As described in previous sections, the current signal from the SiPM has to be properly amplified and shaped (if needed by the application). The signal is generally composed by a fast and a slow component, having a different frequency content.

In order to have a good time resolution it is important to preserve the steep rising edge of the signal (i.e. the fast component of the output current), thus a high bandwidth of the front-end is needed with values of about few hundreds of megahertz and low electronics noise [14,15,81]. For other applications like gamma-ray spectroscopy or large experiments in particle and astro-particle physics (e.g. [65]), the time resolution is not critical but the signal-to-noise ratio is important. In these cases the signal is shaped or integrated, thus the fast component is filtered out.

Moreover, it has to be considered that SiPMs with large areas provide typically a smaller signal amplitude. This is mainly due to the bigger grid capacitance and due to the parasitics of the longer interconnections between the cells and the bonding pads, acting as a low pass filter on the output current. Fig. 15 compares the average signals of a 1×1 mm^2 SiPM and a 6x6 mm^2 SiPM (with the same microcell characteristics). It can be seen that with a low bandwidth (e.g. 20 MHz) the shape of the signal is similar, although the amplitude is smaller for the bigger-area SiPM. This is likely due to the higher inductance and parasitics seen in larger SiPMs additionally to a higher SiPM capacitance which lowers the bandwidth and produces non-stable responses in the front-end electronics. It can also be seen that with larger bandwidth the fast-component of the signal is still suppressed in the bigger SiPM, whereas it is clearly visible in the 1×1 mm^2 SiPM.

3.4. Functional characterization

The functional characterization of the SiPM performance includes the measurement of the photon detection efficiency (PDE), the primary noise rate, i.e. the primary dark count rate (DCR) and the probabilities of the different correlated noise components. There are also important other quantities which can be measured or derived from the previous ones, which are sometimes very useful from the application point-of-view, like the excess noise factor (ENF), and the excess charge factor (ECF). These are both quantities related to the correlated noise.

One common method to measure the dark count rate and the correlated noise component is the analysis of the output pulses from the SiPM in dark condition, typically at a controlled temperature. When collecting trains of pulses, the first problem is how to handle all the pulses that are piled-up. The second problem is how to distinguish the primary events from the correlated noise. As an example, direct crosstalk events are easily distinguishable (they produce pulses with higher amplitudes) but afterpulsing and delayed crosstalk events are mixed within the primary ones. One efficient way is to evaluate the inter-time between the events, as for example proposed in [82] and used by others [83,53]. Among proposed implementations, the method in [83] is based on signal filtering (low-pass filtering, plus DLED to remove the long tails of pulses) and peak-detection with subsequent

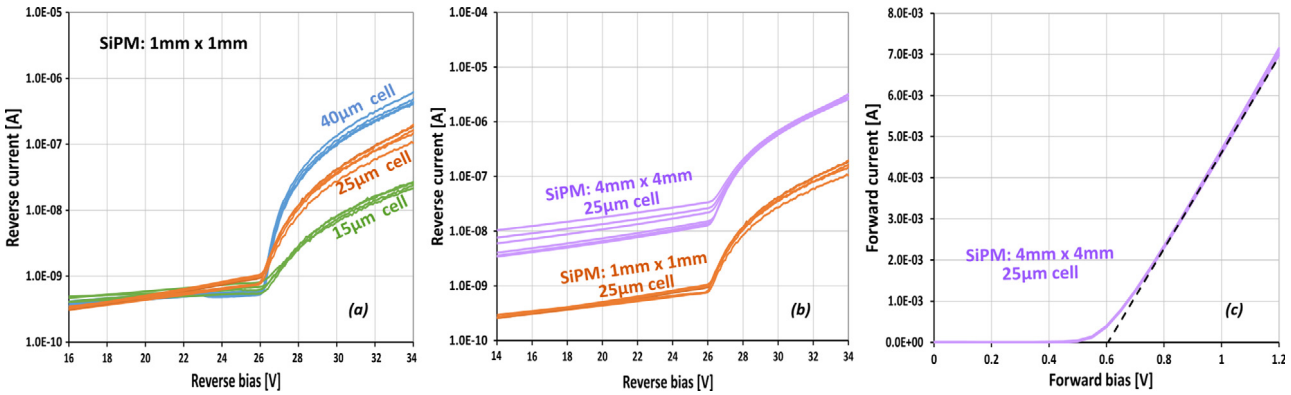

Fig. 14. Examples of I–V curves in reverse bias (from 5 different samples per type), of 1 × 1 mm^2 SiPMs with different cell pitch (a) and of 1 × 1 mm^2 SiPMs compared with 4 × 4 mm^2 SiPMs (b). Forward I–V curve of 4 × 4 mm^2 SiPM (c). Data are taken with FBK NUV-HD 2016 and 2018 production runs.

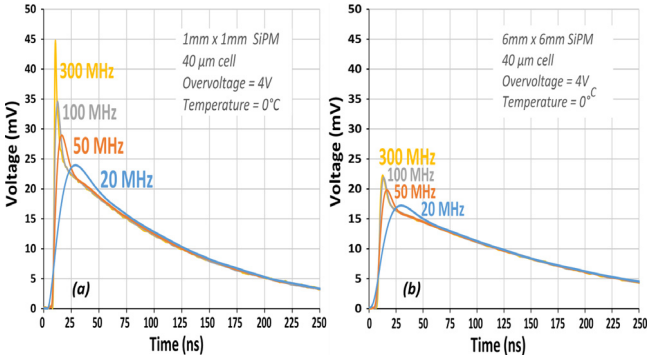

Fig. 15. Examples of average single-cell signals measured on 1×1 mm^2 SiPM (a) and 6x6 mm^2 SiPM (b), with different front-end bandwidths. Measurement done at 0 °C and with FBK NUV-HD 2018 devices.

extraction of the signal inter-times and the amplitude (normalized to the single-cell amplitude) for each event. Others use different methods for signal filtering, for example "Moving Window Difference" plus "Moving Window Average", obtaining similar results [84]. Then, by plotting the amplitude versus inter-time and the histogram of the inter-times a plot like in Fig. 16 can be obtained. The primary dark count rate follows a Poisson distribution, thus the inter-times have an exponential distribution. To extract the primary DCR it is possible to fit the inter-time histogram with an exponential function, but considering just the high inter-times part, where no afterpulsing or correlated noise is present. Then, analyzing the remaining part, i.e. the difference between the measurement and the fit, it is possible to extract the afterpulsing and crosstalk probabilities, as described below.

3.5. Correlated noise

Correlated noise probabilities can be extracted from the very same plots and with the same procedure described above [83]. From the inter-time histogram shown in Fig. 16 (middle), it is possible to evaluate the excess of events with respect to the exponential fit (relative to primary generation). This excess of occurrences, normalized to the total number of events acquired, gives an estimation of the "delayed correlated noise" probability, i.e. afterpulsing and delayed crosstalk. These two components are generally distinguishable in plots like in Fig. 16 (middle). In this case they can be calculated separately based on the inter-time. Referring to the specific case on the cited figure, it is possible to evaluate the afterpulsing probability and the delayed-crosstalk probability as in the following formulas.

$$P_{ap} = \left[\sum_{\Delta t=2.5E-8}^{\Delta t=1E-5} \left(N_{meas}(\Delta t) - N_{fit}(\Delta t)\right)\right] / \sum N_{meas} \tag{14}$$

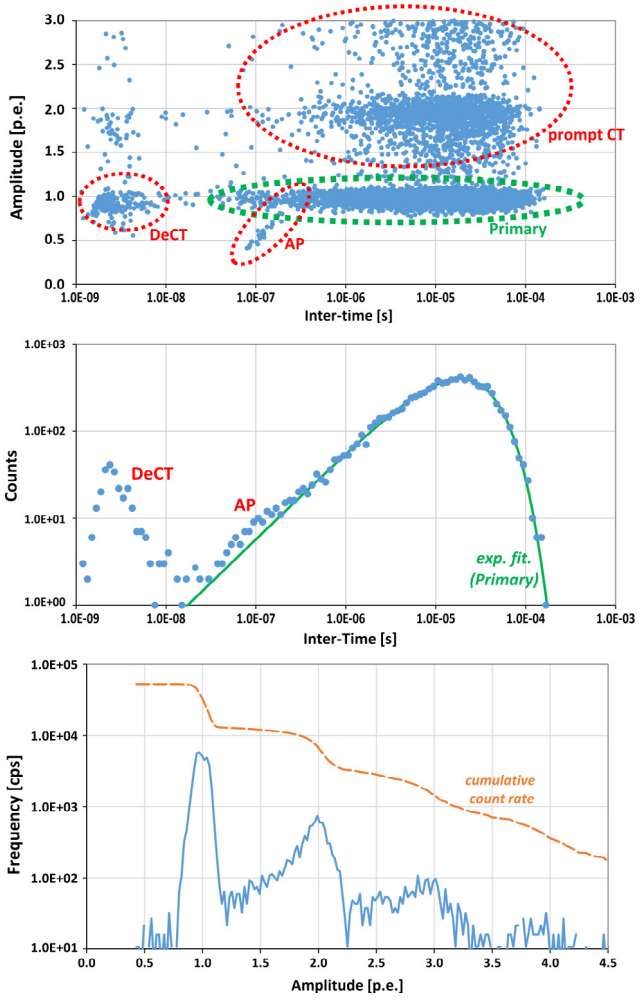

Fig. 16. On the top: examples of amplitude (expressed in photo-electrons) vs. inter-time scatter plot, from the functional characterization of pulses from the SiPM (top). In the middle: histogram of the intertimes (horizontal log scale and log bins), where the exponential fit to primary events is visible. On the bottom: histogram of the event amplitudes, normalized and expressed in photo-electrons.

$$P_{DeCT} = \left[\sum_{\Delta t=1E-9}^{\Delta t=2E-8} \left(N_{meas}(\Delta t) - N_{fit}(\Delta t)\right)\right] / \sum N_{meas} \tag{15}$$

The maximum and minimum inter-times are chosen with a certain degree of freedom, based on the scatter plots.

When instead the cell recharge is very fast, thus afterpulsing events appear with inter-times down to few nanoseconds, or when the lifetime of the photo-generated carriers in the neutral regions is in the order of hundreds of nanoseconds (thus the tail of delayed crosstalk events extends up to hundreds of nanoseconds) [11], afterpulsing and delayed crosstalk cannot be calculated separately. The two "clouds" of events in scatter plots overlap. In such cases the summed probability of the two components can be given.

This procedure is relatively simple and can be easily implemented when the DCR is moderate, i.e when it is possible to well fit the exponential distribution of the primary generation events in a region without afterpulsing. However, it has to noted that Eqs. (14) and (15) do not consider the correlated effects of dark-counts and correlated pulses, as instead it is done in [85]. For example, at short intertimes, due to the presence of both primary and correlated noise, there will be less events due to DCR, than in the absence of delayed cross-talk. Therefore, the simple extrapolation of the primary-events via an exponential fit in the delayed cross-talk region might be correct only as a first approximation (giving a small bias). If DCR is not high the effect can be considered negligible.

In the reported example (Fig. 16), the primary DCR has been estimated considering intertimes higher than 10 μs, where afterpulsing is negligible. DCR plus direct crosstalk events are counted as one event in the histogram, thus the direct crosstalk contribution is neglected. It should be noted that in other cases, when the rate of primary noise is high or the afterpulsing is more important, the "effective" time constant of AP and DCR become similar. In such case it might be difficult to completely disentangle the two contributions, introducing some errors in the DCR and P_{ap} estimation. However, in case of small and medium area SiPM, as in the example, they are well separated. The time constant is just an "equivalent" quantity for afterpulsing, since its contribution vs. the inter-time is actually given by a combination of the traps release time constants (one or more) and the cell recovery time.

Direct (prompt) crosstalk probability instead, can be extracted by the event amplitude histogram. Fig. 16 (bottom) shows an example where the vertical axis is normalized to the acquisition time, thus giving the occurrence frequency and the horizontal axis gives the event amplitude in photo-electrons (p.e.), i.e. normalized to the single-cell response amplitude. The cumulative probability is also shown, which quantifies the DCR as a function of discrimination threshold. Under certain assumptions, the direct crosstalk probability can be extracted as the probability of having 2 p.e. events, over the probability of having 1 p.e. event.

3.6. Photon detection efficiency and saturation

To measure the photon detection efficiency (PDE) of a SiPM special attention has to be given to the crosstalk, afterpulse and dark count rate. Upon the detection of one impinging photon on the SiPM the measured output signal can either be a single SPAD signal or a multiple of the single SPAD signal due to crosstalk and/or afterpulse. In order to avoid a bias of the PDE measurement by the correlated noise one method is to exploit the probability of no events detected, if using a pulsed light source [86–88,53]. One possible PDE measurement setup can be seen in Fig. 17. The light source is a pulsed light emitting diode (LED) which is diffused in an integrating sphere with a reference diode mounted on one port. Another port of the integrating sphere serves to illuminate the SiPM under test at a certain position for which the reference diode mounted in the integrating sphere was previously calibrated with a NIST calibrated avalanche photodiode. All three ports of the integrating sphere are perpendicular to each others to avoid direct light. By measuring the count spectrum of the SiPM under test the zero peak (no photons detected) N_{ped} can be identified and via Poisson statistics (knowing the total event number N_{tot}) the number of photons detected n_{pe} calculated,

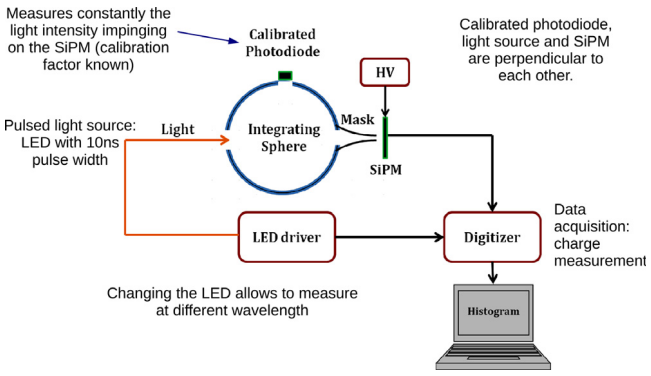

Fig. 17. Setup to measure the photon detection efficiency of a SiPM.

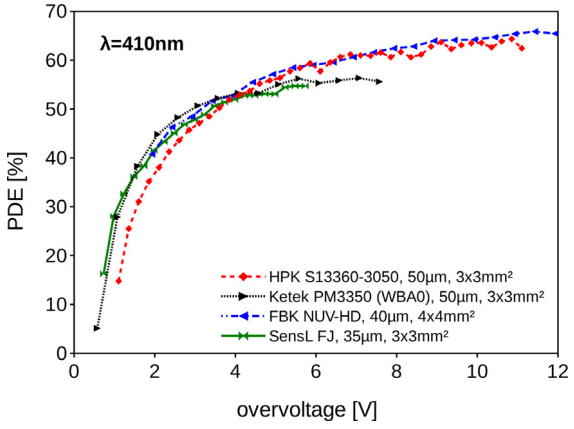

Fig. 18. PDE as a function of SiPM bias overvoltage at 410 nm for different SiPM producers (HPK, Ketek, FBK 2016 and SensL). The measurement error is in the range of 5%, not shown in the plot for clarity.

correcting for randomly detected dark count events (N_{ped}^{dark} and N_{tot}^{dark}), as given in Eq. (16) [86].

$$n_{pe} = -\ln\left(\frac{N_{ped}}{N_{tot}}\right) + \ln\left(\frac{N_{ped}^{dark}}{N_{tot}^{dark}}\right) \tag{16}$$

Knowing the number of photoelectrons detected the PDE can be calculated via the measured reference power, as reported in [86] and in Eq. (17) or from the reference diode current, knowing the afore determined calibration factor.

$$PDE = \frac{n_{pe}}{N_{ph,pulse}} = \frac{n_{pe}}{(P_{opt,DUT} \cdot T_{led})/E_{ph}} \tag{17}$$

$N_{ph,pulse}$ is the average number of photons per LED pulse, $P_{opt,DUT}$ is the optical power on the device under test active-area, T_{led} is the repetition period of the light pulse and E_{ph} is the energy of photons.

In Figs. 18 and 19 state-of-the-art PDE values measured with p-on-n SiPMs for different producers at 410 nm and 525 nm can be seen, respectively. It can be observed that for a wavelength of 410 nm all producers obtain very similar and good PDE values, whereas for 525 nm HPK SiPMs are better performant. This can be due to a thicker depletion width as well seen in their higher breakdown voltage. Furthermore it can be seen that for 410 nm the PDE saturates faster as for 525 nm. In these p-on-n devices at 410 nm the avalanche is triggered mostly by electrons whereas at 525 nm holes start to play a role (both electron and holes can trigger, depending on the absorption position). Holes have a lower and slowly increasing impact ionization coefficient. Therefore, the avalanche triggering probability saturates at higher overvoltages.

If the number of impinging photons ($N_{photons}$) times the PDE is small compared to the total number of microcells (N_{total}) the SiPM output

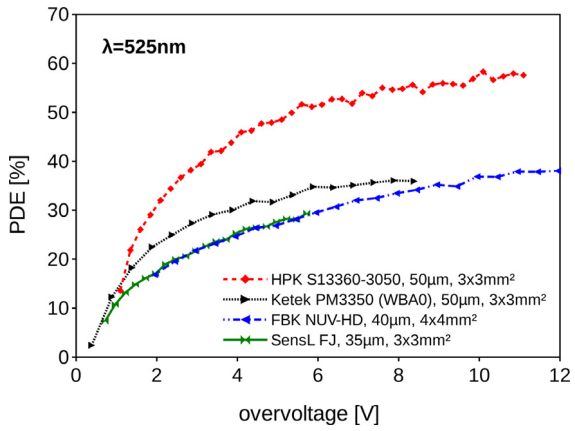

Fig. 19. PDE as a function of SiPM bias overvoltage at 525 nm for different SiPM producers (HPK, Ketek, FBK 2016 and SensL). The measurement error is in the range of 5%, not shown in the plot for clarity.

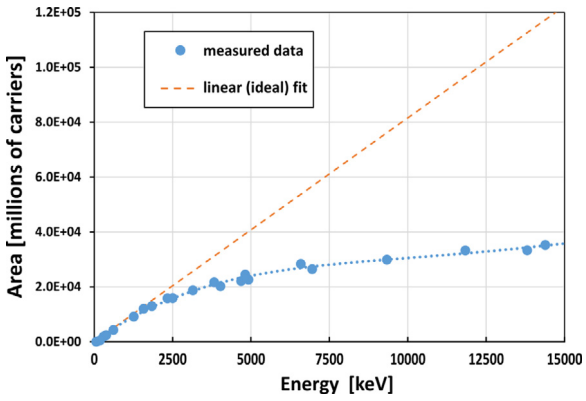

Fig. 20. Example of measured charge ("Area" of integrated signal) as a function of energy of gamma source (either real or emulated, as described in [89]). SiPM is a FBK RGB-HD 4×4 mm^2 with 25 μm cell pitch at 5 V overvoltage and illuminated by a 420 nm LED. The real behavior deviates from the ideal linear one, extrapolated from the first points.

signal ($N_{fired\ cells}$) is proportional to the input photon signal (N_{photon}). If the input photon flux increases, SiPMs show saturation effects leading to non-linear behavior (i.e. "non-linearity" effect) [89] . This is inherently given by their limited number of microcells. An approximation of the input–output transfer function can be seen in Eq. (18) [2].

$$N_{fired\ cells} = N_{total} \cdot \left(1 - \exp\left[-\frac{N_{photon} \cdot PDE}{N_{total}}\right]\right) \quad (18)$$

Respective plots and measurements of various SiPMs can be found in [90]. In this reference as well "over-saturation" effects of some devices are described. In this case the simple Eq. (18) is not further valid and more photoelectrons as SPADs available in the SiPM can be seen.

To illustrate the effect of saturation, Fig. 20 shows an example of measured charge as a function of energy of gamma source (either real or emulated, as described in [89]). The real behavior deviates from the ideal linear one, extrapolated from the first points, at low energies. This translates into a compression of the reconstructed energy spectrum. As an example, Fig. 21 shows the collected charge spectrum (integrating the SiPM signal) of a LSO:Ce scintillator coupled to a Hamamatsu MPPC S10931-025P SiPM with different gamma sources, i.e. ^{22}Na, ^{57}Co, ^{60}Co, ^{137}Cs. An energy resolution of ~10% for an energy of 511 keV can be deduced. For gamma energies above 511 keV the number of produced scintillation photons become too high and saturation effects can be observed. In a phenomenological simulation of the SiPM signal such saturation effects have to be taken into account. However, it should be mentioned that the correct description of light yield measurements with crystals is rather complex, as several effects have to be considered, e.g optical crosstalk, afterpulsing, recharging of the SPADs and so on.

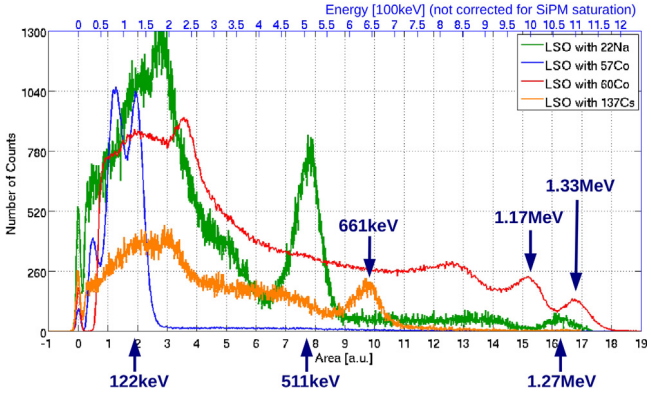

Fig. 21. Typical energy spectra of a LSO:Ce scintillator measured with different gamma energies. The scintillator light output was measured with a Hamamatsu MPPC S10931-025P SiPM. At higher gamma energies saturation effects due to the limited number of microcells can be noticed.

4. Simulation framework

4.1. Electrical SPICE simulations

It is useful to have an equivalent electrical model (e.g. SPICE model) for the SiPM, in combination with the front-end model, to estimate the front-end behavior and to simulate the signal shape and amplitude [57,58]).

Here we report an electrical model which takes into account the parasitics of the interconnections, the multiple cell ignitions and which further simulates the physical avalanche ignition, self-sustaining and quenching (or not quenching) through two switches, one for ignition and one for self-sustaining. The threshold value is based on the value of the threshold current for the avalanche process [13,58] (which here we set to 20 μA). The schematic is reported in Fig. 22. The front-end amplifier (AD8000 in trans-impedance configuration, with gain 1000 V/A and input impedance 20 Ω, not shown in the figure) was also included in the simulation.

The SiPM parameters were extrapolated from measurements and for the simulation here reported we set R_q to 700 kΩ, C_q is 40 fF, C_d is 90 fF, R_d is 700 Ω, the breakdown voltage is 28 V and the bias voltage 33 V. Fig. 23 shows some simulation comparisons. In Fig. 23(a) we varied the number of cells in the SiPM, thus simulating the behavior of one single cell, a 1×1 mm^2 SiPM and a 3×3 mm^2 SiPM (as measured in [14]). In Fig. 23(b) we compare the behavior of a 3×3 mm^2 SiPM with different number of triggered cells. In Fig. 24(a) we varied the quenching resistor (on a 3×3 mm^2 SiPM), from a very low value to a very high one. It can be noticed that with the lowest value, the resistor value is not high enough and the avalanche process is not quenched, whereas with higher resistor values the avalanche is correctly quenched. By increasing the R_q value, the fast component changes slightly, however the slow component of the signal is significantly different. With the highest value of the resistor the signal rapidly decreases but slowly reaches the baseline after a very long time. The signal offset in all plots is due to the amplifier in the front-end. Finally in Fig. 24(b) we varied the amplifier input impedance of the front-end amplifier. A high input impedance does not allow to extract the fast component of the signal, whereas a very small input impedance is beneficial for extracting the fast component, however, in this case can create oscillations in the signal.

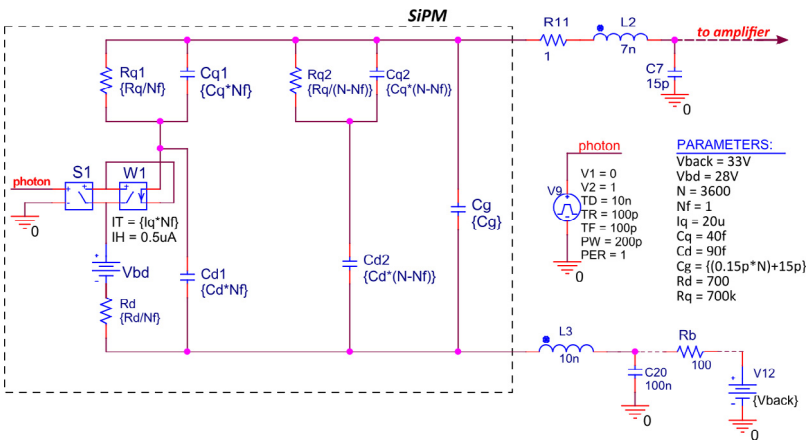

Fig. 22. Electrical model used for the SPICE electrical simulations. It takes into account the parasitics of the interconnections, the multiple cell triggering and it emulates the physical steps of avalanche triggering, self-sustaining and quenching (or not quenching), through two switches — one for ignition and one for self-sustaining.

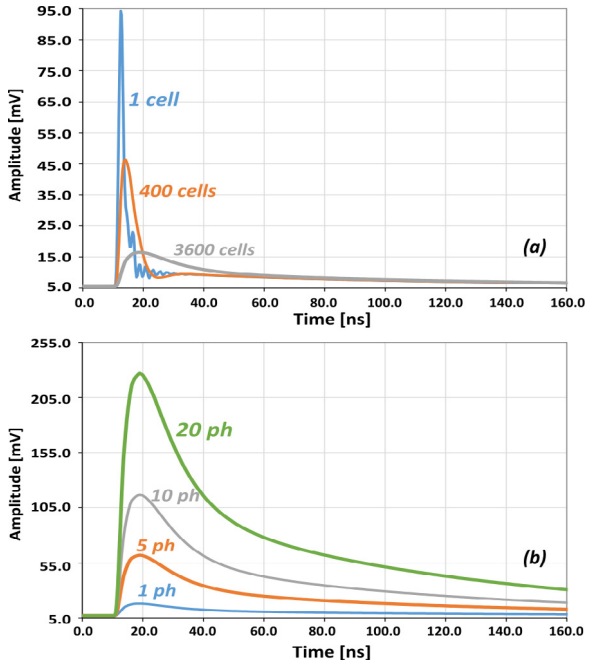

Fig. 23. SPICE electrical simulation results. (a) Plot that compares signals with 1 triggered cell, varying the total number of cells, thus the SiPM dimensions. (b) Signals from a 3×3 mm^2 SiPM with different number of triggered cells.

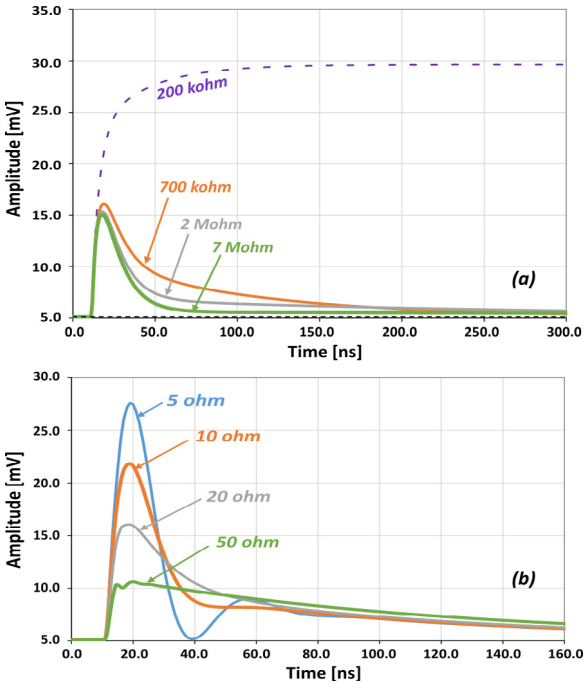

Fig. 24. SPICE electrical simulation results. (a) Comparison of signals from 3×3 mm^2 SiPM with different quenching resistor values. (b) Signals from a 3×3 mm^2 SiPM with different input impedance of the front-end amplifier.

4.2. Phenomenological simulations

The applications of SiPMs in view of light detection can be divided in three main sections, (i) single photon detection, (ii) low light intensity (multi-photon) detection and (iii) high photon rate (high photon time density) detection. In the single photon detection regime, the SPAD signal (SiPM behavior) can be described by the equivalent circuit model seen above together with the used readout electronics and parameters like the single photon time resolution (SPTR), photon detection efficiency (PDE), direct and delayed crosstalk, afterpulse probability and dark count rate. These parameters and ways to measure them have already been discussed in the preceding chapters. In single-photon detection the SiPM is working in photon counting mode, which means that the rate is low enough, hence, the single SPAD signals do not overlap and the system is adequately described by these mentioned parameters and models.

In the range of multiple photon detection up to very high photon rates the single SPAD signals start to overlap and secondary correlation effects appear, hence, analytical statistical models or sound Monte-Carlo simulation have to be applied. The essential ingredients of such statistical models and simulations are the same phenomenological parameters as given for single photon detection. Additionally non-linearity or saturation effects have to be taken into account, e.g. saturation of the SiPM signal due to the limited amount of SPADs available. Furthermore, in such analysis the recharge-time of the SPAD starts to play an important role as well.

A classical application of SiPMs is to sense the emitted light of a pair of scintillators in time resolved spectroscopy of high energetic particles, as in time of flight positron emission tomography (TOF-PET). Additionally, in high energy physics the time tagging of minimum ionizing particles becomes more and more important. It was shown that the best timing in such systems can be achieved by a leading edge discrimination of the SiPM signal [91]. Because of the determination of the time information within the first photoelectrons detected, Monte-Carlo simulations of such scintillator based detectors, readout by SiPMs,

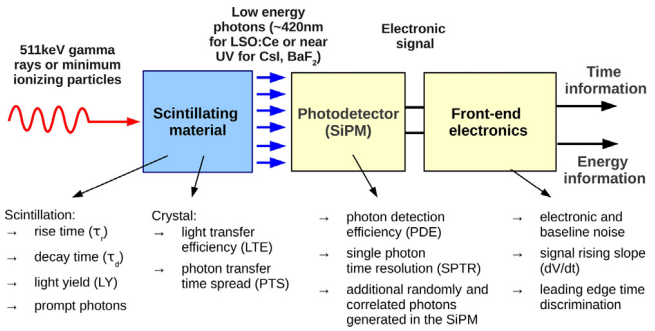

Fig. 25. Scintillator based detectors typically used in TOF-PET consists of three main building blocks, the scintillator, the photodetector and the readout front-end electronics.

usually neglect effects like saturation, afterpulsing or the recharge-time of the SPAD.

4.2.1. *Example of time resolution simulations in TOF-PET*

In Fig. 25 the detector system in TOF-PET is shown, which typically comprises an inorganic scintillator which detects the 511 keV gamma converting the energy into optical photons in the visible range ~420 nm. These photons are emitted with a certain time structure, mostly modeled by a bi-exponential function with ~70 ps rise time and ~40 ns decay time constant [92]. In the crystal the optical transport imposes additional time smearing and losses, which can be modeled e.g. in SLitrani [93] or Geant4 [94]. The sensing of the optical photons in state-of-the-art systems is done with SiPMs, which gives the best timing performance of all commercially available detectors.

If one of the microcells or SPADs is fired upon sensing a scintillation photon, they give rise to a single cell signal defined by the SiPM's equivalent circuit and the circuitry around the SiPM [95]. As discussed in Section 2.4, it arises from RC-filters and thus should only be a combination of exponential functions. The time constants are dependent on the overvoltage, i.e. the cell capacitance changes with the overvoltage as the depletion zone will change in dimension. However, these changes are assumed to be negligible in this kind of application, and to good approximation, it can further be assumed that the time constants determining the single cell signal are independent of the overvoltage. In this case, the single cell signal can be modeled with a bi-exponential function (Eq. (19)) with τ_{Mr} and τ_{Md} being the rise time and fall time of the signal, respectively. For simplicity the second long recharge tail of the signal can be neglected if considering timing discrimination on the leading edge of the SiPM signal and if the dark count rate is low enough in order that the signal tails do not pile-up causing baseline fluctuations. If the dark count rate is higher a second long recharge tail can easily be modeled with an additional exponential term in Eq. (19). The parameter A denotes the amplitude of the single SPAD signal, and ϵ the moment when each microcell fires. Typically the rise time τ_{Mr} is in the order of tens to hundreds of picoseconds and the fall time τ_{Md} in the order of tens of nanoseconds. A single cell amplitude jitter caused by gain fluctuations can be modeled by a Gaussian distribution with probability $p(A) = \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left[-(\bar{A}-A)^2/(2\sigma_A^2)\right]$, with $\bar{A}$ the mean value and σ_A the standard deviation of the fluctuation. In this way mathematically the single SPAD signal $s_\epsilon(t)$ can be described according to Eq. (19). The maximum of the signal in Eq. (19) is normalized to the amplitude A. Here the operator $\Theta(t)$ denotes the Heaviside function which is 0 for $t < 0$, 0.5 for $t = 0$ and 1 for $t > 0$.

$$
\begin{aligned}
s_\epsilon(t) &= A'\left(\exp\left[-\frac{t-\epsilon}{\tau_{Md}}\right] - \exp\left[-\frac{t-\epsilon}{\tau_{Mr}}\right]\right)\Theta(t-\epsilon) \\
A' &= A\left(b^{\frac{1}{1-b}} - b^{\frac{1}{1/b-1}}\right)^{-1} \\
b &= \frac{\tau_{Md}}{\tau_{Mr}}
\end{aligned}
\tag{19}
$$

Coupling the SiPM to a scintillator with an intrinsic light yield LY and a light transfer efficiency LTE it will give rise to a mean value of $LY \cdot PDE \cdot LTE$ avalanches. This is due to the fact that not all photons generated by the scintillation n with $\langle n \rangle = LY$ will actually produce an avalanche due to absorption in the crystal and limited photodetection efficiency of the SiPM. Thus the probability of one photon emitted by the scintillation to produce an avalanche is $PDE \cdot LTE$. This process of loosing photons from the scintillation to the production of an avalanche can be denoted as random deletion, underlining its stochastic nature. The output signal of the device S_{SiPM} is described as the sum of the single cell signals (Eq. (20)). Because of the long decay time of a single cell signal this process is often referred to as signal pile-up.

$$
S_{SiPM} = \sum_{k=1}^{n} s_{\epsilon(k)}(t) \cdot \Theta\left[PDE \cdot LTE(\chi) - rand(1)\right] \tag{20}
$$

The function "rand(1)" generates a random number between 0 and 1 for each photon emitted k by the scintillation process. Thus the Heaviside function $\Theta\left[PDE \cdot LTE(\chi) - rand(1)\right]$ represents the random deletion due to the limited detection efficiency of the SiPM (PDE) and the light transfer efficiency (LTE) of the crystal. The light transfer efficiency $LTE(\chi)$ depends on the gamma absorption point in the crystal χ, which in a first approximation is mainly dependent on the depth-of-interaction (DOI), and thus changes with each gamma interaction.

In the Monte Carlo simulation $\epsilon(k)$, the detection time of each photon k, is a crucial parameter as it models the entire timing behavior of the system. Four contributions to $\epsilon(k)$ can be identified, i.e. the time Δt from the emission of the 511 keV gamma until its absorption in the crystal (at χ with a certain DOI), the scintillation statistics (given by the light yield, scintillation emission rise- and decay time), the light transfer time spread (LTTS) and the single photon time resolution (SPTR) of the SiPM. Supposing that the $T_k(p)$ operator generates a single random time stamp with the underlying probability density function p, then $\epsilon(k)$ can be expressed as in (21).

$$
\epsilon(k) = \Delta t + T_k(f) + T_k(LTTS(\chi)) + T_k(g) \tag{21}
$$

The function $T_k(f)$ gives one time stamp per k-th photon emitted of the scintillation process with a certain light yield, rise time and decay time as given by the time probability density function f. This intrinsic scintillation rate $f(t)$ is defined in Eq. (22) as a bi-exponential function with τ_r the rise time and τ_d the decay time. In full generality it can consist of one or more N bi-exponential terms with relative weights ρ_i. The percental weights in terms of the area or relative light abundance are then calculated according to Eq. (23).

$$
f(t) = \sum_{i=1}^{N} \frac{\exp\left(-\frac{t}{\tau_{d,i}}\right) - \exp\left(-\frac{t}{\tau_{r,i}}\right)}{\tau_{d,i} - \tau_{r,i}} \cdot \rho_i \cdot \Theta(t) \tag{22}
$$

$$
R_i = \frac{\rho_i}{\sum_{i=1}^{N} \rho_i} \tag{23}
$$

The time $T_k(LTTS(\chi))$ accounts for the random time spread caused by the light transfer in the crystal for the k-th photon, simulated by a light ray tracing software, e.g. SLitrani or Geant4. The light transfer time spread (LTTS) is as well a function of the gamma interaction point in the crystal χ and, hence, the DOI. The combined influence of the gamma delay time (first term in Eq. (21)) and scintillation light transfer time spread in the crystal (third term in Eq. (21)) is referred to as photon transfer time spread or photon travel spread (PTS) [96]. $T_k(g)$ models the photodetector's transit time spread and gives a random time stamp generated by a Gaussian distribution. The sigma of this Gaussian is the single photon time resolution (SPTR) of the SiPM, as can be seen in Eq. (24). Here the parameter Δ_M denotes a possible electronic delay and σ_{SPTR} is the single photon time resolution of the SiPM.

$$
g(t) = \frac{1}{\sqrt{2\pi}\sigma_{SPTR}} \exp\left[-\frac{(t-\Delta_M)^2}{2\sigma_{SPTR}^2}\right] \tag{24}
$$

The Monte Carlo simulation according to these formulas is summarized in Fig. 26 [97]. After generating the SiPM signal, i.e. the pile-up of

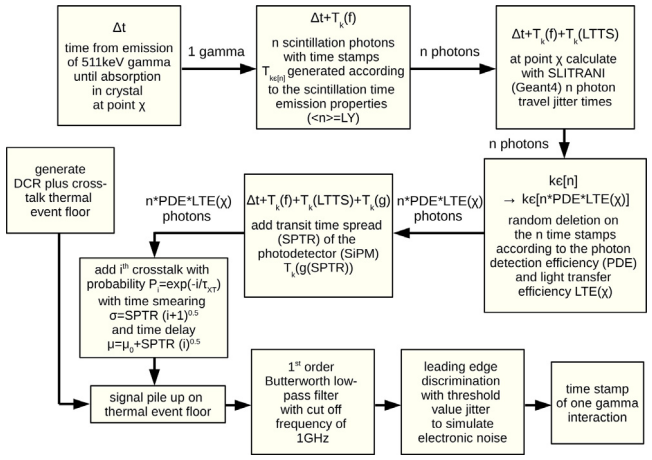

Fig. 26. Flow chart of the Monte Carlo program to simulate the time resolution in a TOF-PET system.

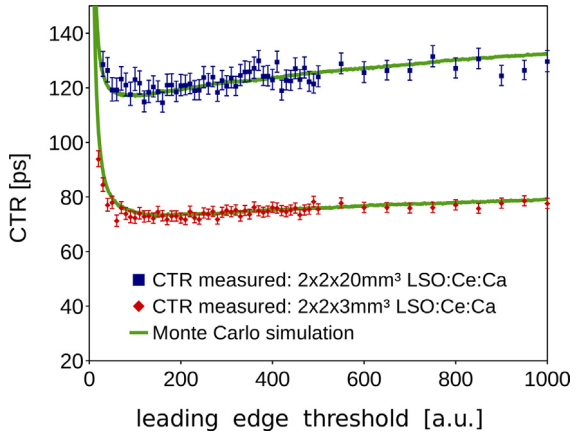

Fig. 27. CTR measurements with 4×4 mm^2 NUV-HD SiPMs from FBK 25 μm SPAD size operated at 12.5 V overvoltage, coupled to 2×2×3 mm^3 and 2x2x20 mm^3 LSO:Ce:Ca crystal from Agile. Green solid lines represent the outcome of the phenomenological Monte Carlo simulations. Data taken from [29].

individual SPAD signals firing at the proper times, a low pass filtering is done in order to mimic the limited electronic bandwidth typically in the order of 1 GHz. A leading edge discrimination is used to produce the time stamp of the simulated output pulse, like it is done in most of the systems applied in TOF-PET. Electronic noise can be added by randomly changing the leading edge threshold around the given mean value with a Gaussian distribution, where the used sigma is equal to the rms noise level.

A comparison of the results of such a Monte Carlo simulation model with measurements in a TOF-PET like setup are shown in Fig. 27. Shown measurements were performed in a coincidence setup coupling LSO:Ce codoped with Ca crystals of 2×2×3 mm^3 and 2x2x20 mm^3 from the producer Agile with 4×4 mm^2 NUV-HD SiPMs from FBK with 25 μm SPAD size. The SiPM signal was readout with the NINO ASIC [98] to obtain the time information and an analog amplifier for the energy discrimination [97]. For a general information on the coincidence time resolution measurement setup we refer for example to [28,79,99–101] and for details on the discussed results here we refer to [29]. The Monte Carlo simulations shown in Fig. 27 and as described in this paper further implement the crosstalk probability of the SiPM modeled via an increased time smearing of the i-th crosstalk event according to $\sigma = SPTR \cdot \sqrt{i+1}$ and an additional time delay $\mu = \mu_0 + SPTR \cdot \sqrt{i}$, where the ith crosstalk probability is equal to $P_i = \exp(-i/\tau_{XT})$ with $\tau_{XT} = 2.3$ crosstalk events [29]. As can be seen in

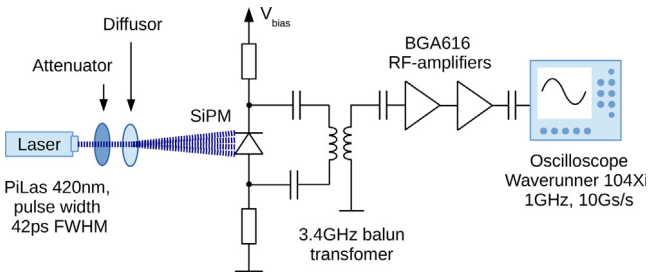

Fig. 28. Setup to measure the SPTR with balun transformer electronics.

Fig. 27 the phenomenological Monte Carlo simulation is able to predict the measurements with a coincidence time resolution (CTR) of 73 ± 3 ps FWHM for 2×2×3 mm^3 and 117 ± 3 ps FWHM for 2x2x20 mm^3 crystals very well. It is important to mention that the "correct" SPTR has to be used in the simulations (~50 ps sigma [29]) which is the SPTR without the influence of electronic noise, in order to obtain matching results.

5. Discusssion

5.1. Impact of front-end electronics on the SPTR

For precise timing measurements the correct design of the front-end electronics is important. Especially when measuring the single photon time resolution with large area SiPMs the obtained values can be deteriorated significantly with respect to small area ones. One reason is the signal transfer time skew of the avalanche signal, generated in one cell and propagating to the readout node, which increases with SiPM size due to cells that can be far away from the common bonding pad. As well the higher capacitance of large area SiPMs creates a low pass filtering effect and leads to a smaller and slower single SPAD signal [14,15] (smaller signal slew rate dV/dt) which together with a possible higher noise level (as well given by additional baseline fluctuation due to a higher DCR) deteriorates the SPTR. Recently it was shown that with large area SiPMs (4×4 mm^2 NUV-HD from FBK with 40 μm SPAD size) SPTR values below 100 ps FWHM can be achieved. This was done with a special front-end electronics, based on an RF-transformer (Balun) design, which reduces the impact of the parasitic and passive capacitances of the SiPM facilitating the extraction of the fast single-SPAD signal components [102]. It is indeed important to lower as much as possible the front-end impedance seen by the SiPM, without creating instabilities in the amplifier decreasing its gain. A schematics of the SPTR measurement setup can be seen in Fig. 28. In Fig. 29 the state-of-the-art SPTRs for different SiPMs (from various producers) are summarized, measured with a pulsed laser at 420 nm illuminating the whole device uniformly. These values include all effects, i.e. the laser pulse width of 42 ps FWHM, the electronic noise and the setup jitter. In Fig. 30 we report the estimated electronic noise components, for the different tested SiPMs. This quantity can significantly depend on many parameters of the device, e.g. the cell size, the SiPM area and the DCR. For some of the SiPMs tested it is almost negligible. It is important to notice that the "real" SPTR without the influence of electronic noise is a very important quantity, not only to get insight on the real timing capabilities of the device, but also as an input parameter for timing simulations. Using an SPTR spoiled by electronic noise will lead to wrong conclusions. Fig. 30 should emphasize this important point.

5.2. Considerations for fast timing and future outlook

Besides improving the electronic noise or the speed of the readout amplifier in order to improve the SPTR, at the single SPAD level the electric field uniformity is important for fastest single photon timing (with single SPADs or SiPMs). In the SiPM cell, the SPTR is a function of

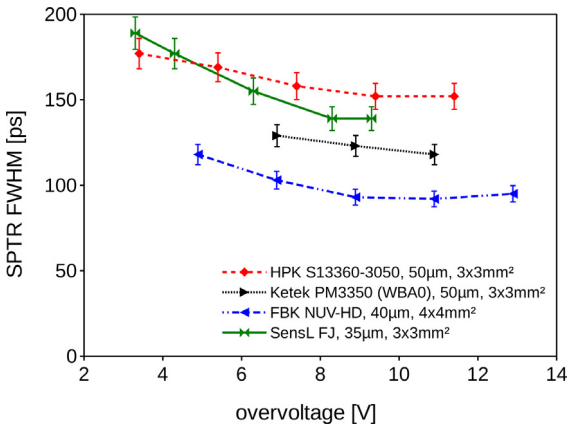

Fig. 29. Measured SPTR with a 420 nm PiLas picosecond laser (42 ps FWHM intrinsic pulse width). The SPTR in nowadays devices is around 100 ps to 140 ps FWHM for a SiPM size around ~3×3 mm^2.

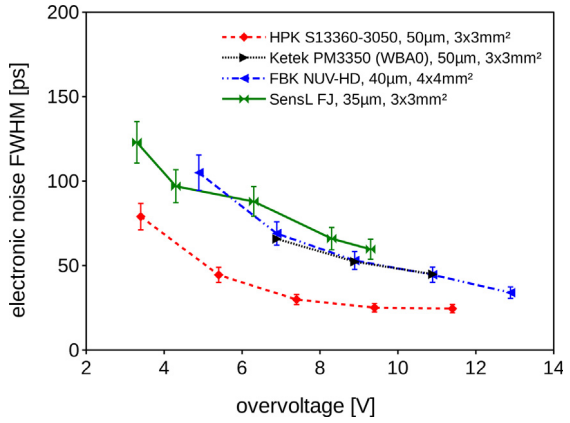

Fig. 30. Measured electronic noise contribution to the SPTR for different SiPM types. Even with optimized electronics readout this contribution is noticeable for large area devices as the 4×4 mm^2 NUV-HD SiPM.

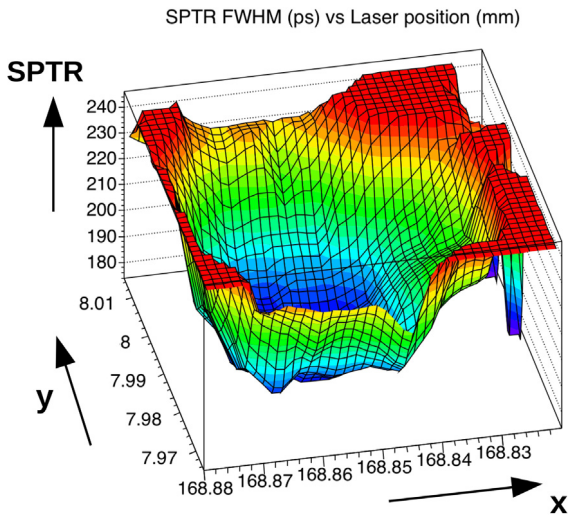

Fig. 31. SPTR laser illumination scan in 5 µm steps of a single SPAD within a Hamamatsu SiPM [15]. Axis units: x in mm, y in mm and SPTR in ps.

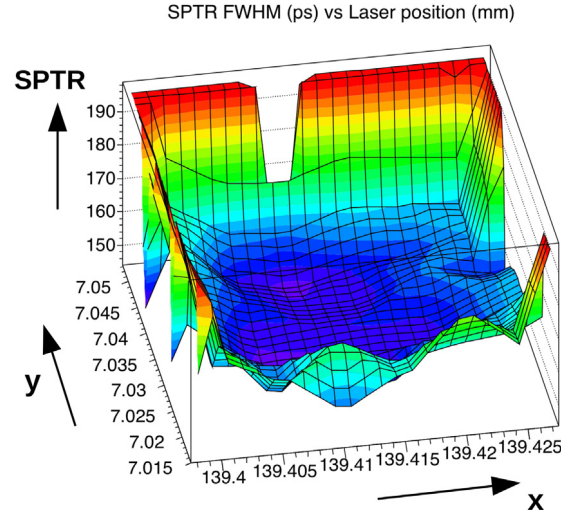

Fig. 32. SPTR laser illumination scan in 5 µm steps of a single SPAD within a FBK SiPM [15]. Axis units: x in mm, y in mm and SPTR in ps.

the photon impact point with observed worse values at the edges of the SPAD, where the electric field is lower, as discussed in previous sections. This can be seen in Fig. 31 for a single SPAD within a Hamamatsu SiPM. On the other hand, taking care of these edge effects (i.e. reducing the border region width of the cell or in general reducing the border effects) a much sharper transition can be achieved, seen in Fig. 32 for FBK SPADs [15].

Fast timing becomes a requirement in many domains of molecular imaging applications and high energy physics. In time of flight positron emission tomography the prompt photon emission in scintillators has the potential to significantly improve the coincidence time resolution [92]. Prompt photon emission can be e.g. Cherenkov radiation caused by the hot-recoil electron upon photoelectric absorption of a 511 keV gamma or as well possible, but less probable, by Compton scattering with a large energy transfer to the quasi-free electron [25,103]. Additionally hot intraband luminescence can be a possible source of prompt emission in crystals [104,105]. A CTR around 20 ps FWHM can change several limiting aspects of PET with new possibilities in PET diagnostics, pharmaceutical research or dose monitoring for proton therapy. The Cramér–Rao lower bound [106] can be used to describe the theoretical CTR improvement due to prompt photons and improvements in the SPTR. Results are shown in Fig. 33, where solid lines show the necessary SPTR and photon detection efficiency (PDE) in order to reach a CTR of 20 ps FWHM for the case of 10, 20, 30, 100 and 500 prompt photons produced at the onset of the standard LYSO:Ce scintillation emission. Indeed already in standard LYSO:Ce crystals about 10–15 Cherenkov photons are produced by the hot recoil electron upon photoelectric absorption of the 511 keV gamma and in addition 5–10 prompt photons via hot intraband luminescence [107]. Parameters for LYSO:Ce used in these calculations are; $\tau_{r1} = 9$ ps with 78 % abundance, $\tau_{r2} = 306$ ps with 22 % abundance, $\tau_d = 41$ ns and an intrinsic light yield of 40 kph/MeV [108]. These lower bound calculations include the additional photon transport time spread (PTS) and light transfer efficiency (LTE) or loss in a 2×2×3 mm^3 crystal. Theoretically 30 prompt photons produced, could be already enough to reach a CTR of 20 ps FWHM, if the SPTR of the photodetector reaches values of close to 10 ps FWHM and a PDE higher than 50%. SiPMs are indeed able to achieve such performance with manageable amount of future research effort.

Harvesting these prompt photons with highest SPTR in an analog system, however, will impose additional challenges on the lowest achievable leading edge threshold regarding the readout electronics, as can be seen in Fig. 34. The figure shows Monte Carlo simulations of a 2×2×3 mm^3 LYSO:Ce crystal assuming that 30 prompt photons are produced at the very onset of the scintillation process instantaneously (Dirac delta), readout with a state-of-the-art SiPM (PDE ~ 55%) and a SPTR of 118 ps FWHM. It can be seen that in the case of an SPTR

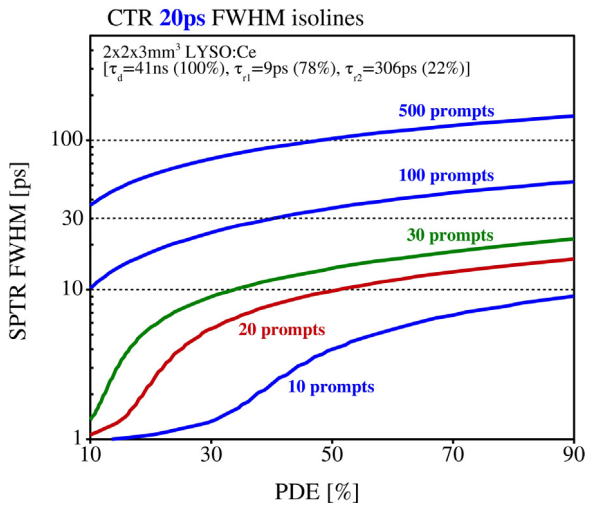

Fig. 33. CTR isolines showing SPTR and PDE values on the axis in order to reach 20 ps FWHM for the case of 10 to 500 prompt photons produced [108].

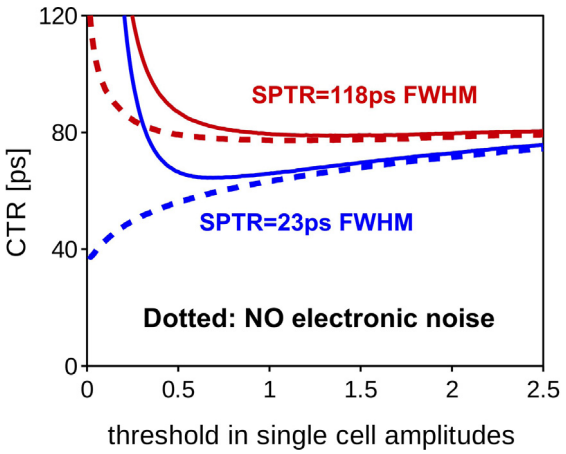

Fig. 34. Full CTR Monte-Carlo simulation of a standard 2×2×3 mm³ LYSO:Ce with 30 prompt photons produced in the crystal. Improving the SPTR improves the CTR noticeably if the electronic leading edge threshold can be lowered due to an improvement in electronic noise or SiPM output signal.

of 118 ps FWHM, lowering the electronic noise to zero has almost no influence on the highest CTR achievable, which is confirmed by several groups achieving very similar CTR benchmarks in their laboratory. However, if the SiPM would provide an SPTR of 23 ps FWHM, electronic noise becomes the dominant factor if the CTR should be improved by a small amount of prompt photons. This is a direct implication that in this case the highest timing information lies on the first photons detected. Considering only scintillation statistics this statement can be found by calculating the variance of the *n*-th photon arrival time [92,109–111] which gives the lowest CTR value for the first photon detected and increases for the subsequent ones. If the scintillation rate is convolved with a Gaussian like time smearing, e.g. in the case of relative high SPTR values (similar or higher as the achieved CTR), the first photon emitted does not deliver the best timing anymore [91,106,112–114]. Qualitatively this can be understood by the long tail the Gaussian introduces at the beginning of the scintillation emission, leaving the first photon detected with a large timing jitter, as it in theory can be detected much earlier as compared to the start of the scintillation pulse. The second photon detected, per definition after the first photon, already experiences a more "limited" range and, hence, shows better timing. In this sense, a low SPTR sets the best timing on the very first photons which is emphasized if prompt photons (like Cherenkov) are being detected in addition [92]. However, because of the limited electronics bandwidth and signal rise time, in this case, the leading edge threshold has to be lowered to very low levels in order to sense these very first photoelectrons, which at some point is prohibited by electronic noise. In this situation future work directed in more powerful high-frequency readout, pixelating the SiPM or pursuing a multi-digital SiPM approach might be indispensable.

6. Conclusion

The silicon photomultiplier has found its way in many applications from industry to fundamental physics experiments. It represent an interesting and versatile photodetector, being both a scintillation-light detector, useful in nuclear and high energy physics and a single-photon detector being used in low-light applications, like time-resolved optical spectroscopy and quantum physics experiments. Being a solid state device it is as well compact and robust.

The increasing number of applications using SiPMs encouraged large developments in the last years, making this device also interesting for large markets. Nowadays the detection efficiency reaches very high values, especially thanks to the very high fill-factor and optimized internal structure. However there is still room for improvements. As highlighted in this paper, the parameters describing the SiPM electrical behavior and its functional performance are many, and they are all interdependent to each other (at a certain degree).

We described the internal microcell (i.e SPAD) structure, along with its design issues and tradeoffs, e.g. triggering probability optimization for the wavelength of interest and border effect minimization using TCAD simulations. This is a relevant topic nowadays, especially for the ongoing optimization towards higher efficiency in the near infrared.

We discussed the amplitude and gain dependence and their fluctuations. The cell dimension is an important parameter, closely related to the gain, which in turn affects many of the SiPM parameters, like the correlated noise probabilities, the peak separation in the charge and amplitude spectra thus the signal-to-noise ratio, and the signal duration.

We introduced the equivalent electrical models and showed simulations of the output signals: the proper signal extraction is a crucial aspect in timing based applications, where the prompt leading edge of the signal, marking the photon arrival time, has to be preserved as much as possible.

Finally, with this complete picture we discussed some phenomenological aspects for SiPMs in timing applications. Here all the parameters, signal extraction and noise aspects have to be taken into account and Monte Carlo simulations are mandatory. The lower bound for energy resolution or time resolution can be estimated in such ways, but then many other practical parameters can limit the real performance. As an example there have been some important developments in the front-end for SiPM readout in single photon timing. Typical circuitry showed limited performances with big area devices, whereas a refined circuit design using RF specific components can perform significantly better.

Acknowledgment

Some of the results presented in this work have been supported in part by the Crystal Clear Collaboration and the European advanced ERC grant TICAL 338953.

References

[1] D. Renker, E. Lorenz, Geiger-mode avalanche photodiodes, history, properties and problems, Nucl. Instrum. Methods Phys. Res. A 567 (1) (2006) 48–56.
[2] D. Renker, E. Lorenz, Advances in solid state photon detectors, J. Instrum. 4 (2009) P04004.
[3] P. Buzhan, B. Dolgoshein, L. Filatov, A. Ilyin, V. Kantzerov, V. Kaplin, A. Karakash, F. Kayumov, S. Klemin, E. Popova, S. Smirnov, Silicon photomultiplier and its possible applications, Nucl. Instrum. Methods Phys. Res. A 504 (1–3) (2003) 48–52.

[4] V. Golovin, V. Saveliev, Novel type of avalanche photodetector with Geiger mode operation, Nucl. Instrum. Methods Phys. Res. A 518 (1–2) (2004) 560–564.

[5] D.J. Herbert, V. Saveliev, N. Belcari, N.D. Ascenzo, A.D. Guerra, A. Golovin, First results of scintillator readout with silicon photomultiplier, IEEE Trans. Nucl. Sci. 53 (1) (2006) 389–394.

[6] T. Frach, G. Prescher, C. Degenhardt, R. Gruyter, A. Schmitz, R. Ballizany, The digital silicon photomultiplier principle of operation and intrinsic detector performance, in: IEEE Nuclear Sci. Symp. Conf., 2009, pp. 1959–1965.

[7] F. Acerbi, A. Gola, V. Regazzoni, G. Paternoster, G. Borghi, N. Zorzi, C. Piemonte, High Efficiency, Ultra-High-Density Silicon Photomultipliers, IEEE J. Sel. Top. Quantum Electron. 24 (2) (2018) 3800608.

[8] A.A. Wagadarikar, S. Katz, A. Ivan, S. Dolinsky, Performance of low afterpulsing probability multi-pixel photon counters for time-of-flight positron emission tomography, in: Proc. of 2013 IEEE Nuclear Science Symposium and Medical Imaging Conference, 2013, pp.1–5.

[9] Z. Liu, S. Gundacker, M. Pizzichemi, A. Ghezzi, E. Auffray, P. Lecoq, M. Paganoni, In-depth study of single photon time resolution for the Philips digital silicon photomultiplier, J. Instrum. 11 (2016) P06006.

[10] F. Corsi, C. Marzocca, A. Perrotta, A. Dragone, M. Foresta, A.D. Guerra, S. Marcatili, G. Llosa, G. Collazzuol, G.F.D. Betta, N. Dinu, C. Piemonte, G.U. Pignatel, G. Levi, Electrical characterisation of silicon photo-multiplier detectors for optimal front-end design, in: IEEE nuclear scinece symposium conference record, N30-222, 2006, pp. 1276–1280.

[11] F. Acerbi, A. Ferri, G. Zappala, G. Paternoster, A. Picciotto, A. Gola, N. Zorzi, C. Piemonte, NUV silicon photomultipliers with high detection efficiency and reduced delayed correlated-noise, IEEE Trans. Nucl. Sci. 62 (3) (2015) 1318–1325.

[12] P. Buzhan, B. Dolgoshein, A. Ilyin, V. Kaplin, S. Klemin, R. Mirzoyan, E. Popova, M. Teshima, The cross-talk problem in SiPMs and their use as light sensors for imaging atmospheric Cherenkov telescopes, Nucl. Instrum. Methods Phys. Res. A 610 (1) (2009) 131–134.

[13] S. Cova, M. Ghioni, A.L. Lacaita, C. Samori, F. Zappa, Avalanche photodiodes and quenching circuits for single photon-detection, Appl. Opt. 32 (2) (1996) 1956–1976.

[14] F. Acerbi, A. Ferri, A. Gola, M. Cazzanelli, L. Pavesi, N. Zorzi, C. Piemonte, Characterization of single-photon time resolution: From single SPAD to silicon photomultiplier, IEEE Trans. Nucl. Sci. 61 (5) (2014) 2678–2686.

[15] M.V. Nemallapudi, S. Gundacker, P. Lecoq, E. Auffray, Single photon time resolution of state of the art SiPMs, J. Instrum. 11 (2016) P10016.

[16] R. Agishev, A. Comeron, J. Bach, A. Rodriguez, M. Sicard, J. Riu, S. Royo, Lidar with SiPM: Some capabilities and limitations in real environment, Opt. Laser Technol. 49 (2013) 86–90.

[17] F. Acerbi, G. Paternoster, A. Gola, V. Regazzoni, N. Zorzi, C. Piemonte, High-density silicon photomultipliers: Performance and linearity evaluation for high efficiency and dynamic-range applications, IEEE J. Quantum Electron. 54 (2) (2018) 4700107.

[18] A.D. Mora, E. Martineghi, D. Contini, A. Tosi, G. Boso, T. Durduran, S. Arridge, F. Martelli, A. Farina, A. Torricelli, A. Pifferi, Fast silicon photomultiplier improves signal harvesting and reduces complexity in timedomain diffuse optics, Opt. Express 23 (11) (2015) 13937.

[19] R. Zimmermann, F. Braun, T. Achtnich, O. Lambercy, R. Gassert, M. Wolf, Silicon photomultipliers for improved detection of low light levels in miniature near-infrared spectroscopy instruments, Biomed. Opt. Express 4 (5) (2013) 659–666.

[20] R. Re, E. Martinenghi, A.D. Mora, D. Contini, A. Pifferi, A. Torricelli, Probe-hosted silicon photomultipliers for time-domain functional near-infrared spectroscopy: Phantom and in vivo tests, Neurophotonics 3 (4) (2016) 045004.

[21] L. Mik, W. Kucewicz, J. Barszcz, M. Sapor, S. Glab, Silicon photomultiplier as fluorescence light detector, in: Proc. of the 18th International Conference Mixed Design of Integrated Circuits and Systems-MIXDES 2011, 2011, pp.663–666.

[22] D. Kalashnikov, L. Krivitsky, Measurement of photon correlations with multipixel photon counters, J. Opt. Soc. Amer. B 31 (10) (2014) B25–B33.

[23] K.A. Balygin, V.I. Zaitsev, A.N. Klimov, S.P. Kulik, S.N. Molotkov, A quantum random number generator based on the 100-Mbit/s poisson photocount statistics, J. Exp. Theor. Phys. 126 (6) (2018) 728–740.

[24] S. Korpar, R. Dolenec, P. Krizan, R. Pestotnik, A. Stanovnik, Study of TOF PET using Cherenkov light, Nucl. Instrum. Methods Phys. Res. A 654 (2011) 532–538.

[25] S.E. Brunner, L. Gruber, J. Marton, K. Suzuki, A. Hirtl, Studies on the Cherenkov effect for improved time resolution of TOF-PET, IEEE Trans. Nucl. Sci. 61 (1) (2014) 443–447.

[26] E. Roncali, S.R. Cherry, Application of Silicon Photomultipliers to Positron Emission Tomography, Ann. Biomed. Eng. 39 (4) (2011) 13581377.

[27] R. Lecomte, Novel detector technology for clinical PET, Eur. J. Nucl. Med. Mol. Imaging 36 (1) (2009) 69–85.

[28] R. Vinke, H. Löhner, D.R. Schaart, H.T. van Dam, S. Seifert, F.J. Beekman, P. Dendooven, Optimizing the timing resolution of SiPM sensors for use in TOF-PET detectors, Nucl. Instrum. Methods Phys. Res. A 610 (2009) 188–191.

[29] S. Gundacker, F. Acerbi, E. Auffray, A. Gola, M.V. Nemallapudi, G. Paternoster, C. Piemonte, P. Lecoq, State of the art timing in TOF-PET detectors with LuAG, GAGG and L(Y)SO scintillators of various sizes coupled to FBK-SiPMs, J. Instrum. 11 (2016) P08008.

[30] J.W. Cates, R. Vinke, C.S. Levin, Analytical calculation of the lower bound on timing resolution for PET scintillation detectors comprising high-aspect-ratio crystal elements, Phys. Med. Biol. 60 (2015) 5141–5161.

[31] A. Benaglia, S. Gundacker, P. Lecoq, M. Lucchini, A. Para, K. Pauwels, E. Auffray, Detection of high energy muons with sub-20 ps timing resolution using L(Y)SO crystals and SiPM readout, Nucl. Instrum. Methods Phys. Res. A 830 (2016) 30–35.

[32] E. Garutti, Silicon photomultipliers for high energy physics detectors, J. Instrum. 6 (C10003) (2011) C10003.

[33] K. Pauwels, C. Dujardin, S. Gundacker, K. Lebbou, P. Lecoq, M. Lucchini, F. Moretti, A.G. Petrosyan, X. Xub, E. Auffray, Single crystalline LuAG fibers for homogeneous dual-readout calorimeters, J. Instrum. 8 (2013) P09019.

[34] R.H. Haitz, A. Goetzberger, R.M. Scarletta, W. Shockley, Avalanche effects in silicon p-n junctions. I. localized photomultiplication studies on microplasmas, J. Appl. Phys. 34 (6) (1963) 1581–1590.

[35] R.H. Haitz, Model for the electrical behavior of a microplasma, J. Appl. Phys. 35 (5) (1964) 1380.

[36] R.J. McIntyre, Theory of microplasma instability in silicon, J. Appl. Phys. 32 (6) (1961) 983–995.

[37] W.G. Oldham, R.R. Samuelson, P. Antognetti, Triggering phenomena in avalanche diodes, IEEE Trans. Electron Devices 19 (9) (1972) 1056–1060.

[38] R.J. McIntyre, Recent developments in silicon avalanche photodiodes, Measurement 3 (4) (1985) 146–152.

[39] S. Cova, A. Lacaita, M. Ghioni, G. Ripamonti, T.A. Louis, 20-ps timing resolution with single-photon avalanche diodes, Rev. Sci. Instrum. 60 (6) (1989) 1104–1110.

[40] H. Dautet, P. Deschamps, B. Dion, A.D. MacGregor, D. MacSween, R.J. McIntyre, C. Trottier, P.P. Webb, Photon counting techniques with silicon avalanche photodiodes, Appl. Opt. 32 (21) (1993) 3894–3900.

[41] W.J. Kindt, H.W.V. Zeijl, Modelling and fabrication of Geiger mode avalanche photodiodes, IEEE Trans. Nucl. Sci. 45 (3) (1998) 715.

[42] W. Maes, K.D. Meyer, R.V. Overstraeten, Impact ionization in silicon: A review and update, Soid State Electron. 33 (6) (1990) 705–718.

[43] R. van Overstraeten, H. de Man, Measurement of the ionization rates in diffused silicon p-n junctions, Solid State Electron. 13 (1) (1970) 583–608.

[44] Y. Okuto, C.R. Crowell, Threshold energy effect on avalanche breakdown voltage in semiconductor junctions, Solid State Electron. 18 (2) (1975) 161–168.

[45] Synopsys Inc., Sentaurus Device User Guide, Version D-2010.03, 2010.

[46] C.L. Anderson, C.R. Crowell, Threshold energies for electron-hole pair production by impact ionization in semiconductor, Phys. Rev. 5 (1972) 2267–2272.

[47] M.L. Knoetig, J. Hose, R. Mirzoyan, SiPM Avalanche size and crosstalk measurements with light emission microscopy, IEEE Trans. Nucl. Sci. 61 (3) (2014) 1488–1492.

[48] M. Anti, F. Acerbi, A. Tosi, F. Zappa, 2D simulation for the impact of edge effects on the performance of planar InGaAs/InP SPADs, in: Optical Systems Design, Proc. SPIE 8550, 2012, 855025.

[49] M. Perenzoni, L. Pancheri, D. Stoppa, Compact SPAD-based pixel architectures for time-resolved image sensors, Sensors 16 (745) (2016) 1–12.

[50] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, S. Cova, Progress in silicon single-photon avalanche diodes, IEEE J. Sel. Top. Quantum Electron. 13 (4) (2007) 852–862.

[51] T. Pro, A. Ferri, A. Gola, N. Serra, A. Tarolli, N. Zorzi, C. Piemonte, New developments of Near-UV SiPMs at FBK, IEEE Trans. Nucl. Sci. 60 (3) (2013) 2247–2253.

[52] G. Zappala, F. Acerbi, F. Ferri, G. Paternoster, V. Regazzoni, N. Zorzi, C. Piemonte, Study of the photo-detection efficiency of FBK High-Density silicon photomultipliers, J. Intrumentation 11 (P11010) (2016) P11010.

[53] A.N. Otte, D. Garcia, T. Nguyen, D. Purushotham, Characterization of three high efficiency and blue sensitive silicon photomultipliers, Nucl. Instrum. Methods Phys. Res. A 846 (2017) 106–125.

[54] C. Piemonte, F. Acerbi, A. Ferri, A. Gola, G. Paternoster, V. Regazzoni, G. Zappala, N. Zorzi, Performance of NUV-HD silicon photomultiplier technology, IEEE Trans. Electron Devices 63 (3) (2016) 1111–1116.

[55] G. Collazuol, Review of Silicon Photo-Multiplier Physics and Applications, Including a Study at Low Temperature, 2008, http://www.bo.infn.it/sm/sm08/presentations/10-04/collazuol.pdf.

[56] D. Marano, G. Bonanno, M. Belluso, S. Billotta, A. Grillo, S. Garozzo, G. Romeo, O. Catalano, G.L. Rosa, G. Sottile, D. Impiombato, S. Giarrusso, Improved SPICE electrical model of silicon photomultipliers, Nucl. Instrum. Methods Phys. Res. A 726 (2013) 1–7.

[57] D. Marano, M. Belluso, G. Bonanno, S. Billotta, A. Grillo, S. Garozzo, G. Romeo, O. Catalano, G.L. Rosa, G. Sottile, D. Impiombato, S. Giarrusso, Silicon photomultipliers electrical model etensive analytical analysis, IEEE Trans. Nucl. Sci. 61 (1) (2014) 23–34.

[58] S. Seifert, H.T. van Dam, J. Huizenga, R. Vinke, P. Dendooven, H. Löhner, D.R. Schaart, Simulation of silicon photomultiplier signals, IEEE Trans. Nucl. Sci. 56 (6) (2009) 3726–3733.

[59] F. Corsi, A. Dragone, C. Marzocca, A.D. Guerra, P. Delizia, N. Dinu, C. Piemonte, M. Boscardin, G.F.D. Betta, Modelling a silicon photomultiplier (SiPM) as a signal source for optimum front-end design, Nucl. Instrum. Methods Phys. Res. A 572 (1) (2007) 416–418.

[60] V. Chmill, E. Garutti, R. Klanner, M. Nitschke, J. Schwandt, Study of the breakdown voltage of SiPMs, Nucl. Instrum. Methods Phys. Res. A (845) (2017) 56–59.

[61] O. Marinov, J. Dean, J.A.J. Tejada, Theory of microplasma fluctuations and noise in silicon diode in avalanche breakdown, J. Appl. Phys. 101 (2007) 064515–1 – 064515–21.

[62] J. Huizenga, S. Seifert, F. Schreuder, H.T. VanDam, P. Dendooven, H. Lohner, R. Vinke, D.R. Schaart, A fast preamplifier concept for SiPM-based time-of-flight PET detectors, Nucl. Instrum. Methods Phys. Res. A 695 (2012) 379–384.

[63] F. Scheuch, D. Führen, T. Hebbeker, C. Heidemann, M. Merschmeyer, Electrical characterization and simulation of SiPMs, Nucl. Instrum. Methods Phys. Res. A 787 (2015) 340–343.

[64] M. Ghioni, A. Gulinatti, I. Rech, P. Maccagnani, S. Cova, Large-area low-jitter silicon single photon avalanche diodes, in: Proc of SPIE Integrated Optoelectronic Devices, Quantum Sensing and Nanophotonic Devices V, vol. 6900, 2008, p. 69001D.

[65] F. Acerbi, S. Davini, A. Ferri, C. Galbiati, G. Giovanetti, A. Gola, G. Korga, A. Mandarano, M. Marcante, G. Paternoster, C. Piemonte, A. Razeto, V. Regazzoni, D. Sablone, C. Savarese, G. Zappala, N. Zorzi, Cryogenic characterization of FBK HD Near-UV sensitive SiPMs, IEEE Trans. Electron Devices 645 (2) (2017) 521–526.

[66] F. Acerbi, G. Paternoster, A. Gola, N. Zorzi, C. Piemonte, Silicon photomultipliers and single-photon avalanche diodes with enhanced NIR detection efficiency at FBK, Nucl. Instrum. Methods Phys. Res. A (2018) in press. https://doi.org/10.1016/j.nima.2017.11.098.

[67] G. Vincent, A. Chantre, D. Bois, Electric field effect on the thermal emission of traps in semiconductor junctions, J. Appl. Phys. 50 (1979) 5484–5487.

[68] D.K. Gautam, W.S. Khokle, K.B. Garg, Photon emission from reverse-biased silicon p-n junction, Solid State Electron. 32 (2) (1988) 219–222.

[69] F. Acerbi, A. Ferri, G. Zappala, G. Paternoster, A. Picciotto, A. Gola, N. Zorzi, C. Piemonte, NUV silicon photomultipliers with high detection efficiency andreduced delayed correlated-noise, IEEE Trans. Nucl. Sci. 62 (3) (2015) 1318–1325.

[70] A.L. Lacaita, F. Zappa, S. Bigliardi, M. Manfredi, On the Bremsstrahlung origin of hot-carrier-induced photons in silicon devices, IEEE Trans. Electron Devices 40 (3) (1993) 577–582.

[71] A.N. Otte, On the efficiency of photon emission during electrical breakdown in silicon, Nucl. Instrum. Methods Phys. Res. A 610 (2009) 105–109.

[72] S. Vinogradov, Analytical models of probability distribution and excess noise factor of solid state photomultiplier signals with crosstalk, Nucl. Instrum. Methods Phys. Res. A 695 (2012) 247–251.

[73] L. Gallego, J. Rosado, F. Blanco, F. Arqueros, Modeling crosstalk in silicon photomultipliers, J. Instrum. 9 (2013) P05010.

[74] V. Chmill, E. Garutti, R. Klanner, M. Nitschke, J. Schwandt, On the characterisation of SiPMs from pulse-height spectra, Nucl. Instrum. Mehtods Phys. Res. A (854) (2017) 70–81.

[75] P. Buzhan, B. Dolgoshein, L. Filatov, A. Ilyin, V. Kaplin, A. Karakash, S. Klemin, R. Mirzoyan, A.N. Otte, E. Popova, V. Sosnovtsev, M. Teshim, Large area silicon photomultipliers: Performance and applications, Nucl. Instrum. Methods Phys. Res. A 567 (1) (2006) 78–82.

[76] W.S. Sul, C.H. Lee, G.S. Cho, Influence of guard-ring structure on the dark count rates of silicon photomultipliers, IEEE Electron Device Lett. 34 (3) (2013) 336–338.

[77] US patent US 2016/0327657 A1, 2016.

[78] A. Gola, C. Piemonte, A. Tarolli, Analog circuit for timing measurements with large area SiPMs coupledto LYSO crystals, IEEE Trans. Nucl. Sci. 60 (2) (2013) 1296–1302.

[79] A. Gola, C. Piemonte, A. Tarolli, The DLED algorithm for timing measurements on large area SiPMs coupled to scintillators, IEEE Trans. Nucl. Sci. 59 (2) (2013) 358–365.

[80] J.Y. Yeom, R. Vinke, N. Pavlov, S. Bellis, L. Wall, K. O'Neill, C. Jackson, C.S. Levin, Fast timing silicon photomultipliers for scintillation detectors, IEEE Photonics Technol. Lett. 25 (15) (2013) 1309–1312.

[81] S. Gundacker, E. Auffray, N.D. Vara, B. Frisch, H. Hillemanns, P. Jarron, B. Lang, T. Meyer, S. Mosquera-Vazquez, E. Vauthey, P. Lecoq, SiPM time resolution: From single photon to saturation, Nucl. Instrum. Methods Phys. Res. A 718 (2013) 569–572.

[82] Y. Du, F. Retiere, After-pulsing and cross-talk in multi-pixel photon counters, Nucl. Instrum. Methods Phys. Res. A 596 (3) (2008) 396–401.

[83] C. Piemonte, A. Feri, A. Gola, A. Picciotto, T. Pro, N. Serra, A. Tarolli, N. Zorzi, Development of an automatic procedure for the characterization of silicon photomultipliers, in: Proc. of IEEE Nuclear Science Symposium and Medical Imaging Conference, 2012, pp. 428–432.

[84] E. Engelmann, SiPM Noise Measurement with Waveform Analysis (ICASIPM 2018 Schwetzingen - Working group: SiPM Nuisance Parameters), 2018,.

[85] E. Garutti, M. Gensch, R. Klanner, M. Ramilli, C. Xu, Afterpulse effect in SiPM and neutron irradiation studies, in: Proc. of 2014 IEEE Nuclear Science Symposium and Medical Imaging Conference, NSS/MIC, 2014, pp. 1–7.

[86] P. Eckert, H.C. Schultz-Coulon, W. Shen, R. Stamen, A. Tadday, Characterisation studies of silicon photomultipliers, Nucl. Instrum. Methods Phys. Res. A 620 (2010) 217–226.

[87] G. Zappala, F. Acerbi, A. Ferri, A. Gola, G. Paternoster, N. Zorzi, C. Piemonte, Set-up and methods for SiPM photo-detection efficiency measurements, J. Intrumentation 11 (P08014) (2016) P08014.

[88] A.N. Otte, J. Hose, R. Mirzoyan, A. Romaszkiewicz, M. Teshima, A. Thea, A measurement of the photon detection efficiency of silicon photomultipliers, Nucl. Instrum. Methods Phys. Res. A 567 (1) (2006) 360–363.

[89] V. Regazzoni, F. Acerbi, G. Cozzi, A. Ferri, C. Fiorini, G. Paternoster, C. Piemonte, D. Rucatti, G. Zappala, N. Zorzi, A. Gola, Characterization of high density SiPM non-linearity and energy resolution for prompt gamma imaging applications, J. Instrum. 12 (P07001) (2017) P07001.

[90] L. Gruber, S.E. Brunner, J. Marton, K. Suzuki, Over saturation behavior of SiPMs at high photon exposure, Nucl. Instrum. Methods Phys. Res. A (737) (2014) 11–18.

[91] S. Gundacker, E. Auffray, P. Jarron, T. Meyer, P. Lecoq, On the comparison of analog and digital SiPM readout in terms of expected timing performance, Nucl. Instrum. Methods Phys. Res. A 787 (2015) 6–11.

[92] S. Gundacker, E. Auffray, K. Pauwels, P. Lecoq, Measurement of intrinsic rise times for various L(Y)SO and LuAG scintillators with a general study of prompt photons to achieve 10 ps in TOF-PET, Phys. Med. Biol. 61 (2016) 2802–2837.

[93] F.X. Gentit, Litrani: A general purpose Monte-Carlo program simulating light propagation in isotropic or anisotropic media, CMS-NOTE-2001-044, 2001.

[94] S. Agostinelli, et al., Geant4—a simulation toolkit, Nucl. Instrum. Mehtods Phys. Res. A 506 (2003) 250.

[95] P. Avella, A. Santo, A. Lohstroha, M.T. Sajjada, P.J. Sellina, A study of timing properties of silicon photomultipliers, Nucl. Instrum. Methods Phys. Res. A 695 (2011) 257.

[96] S. Gundacker, A. Knapitsch, E. Auffray, P. Jarron, T. Meyer, P. Lecoq, Time resolution deterioration with increasing crystal length in a TOF-PET system, Nucl. Instrum. Methods Phys. Res. A 737 (2014) 92–100.

[97] S. Gundacker, E. Auffray, B. Frisch, P. Jarron, A. Knapitsch, T. Meyer, M. Pizzichemi, P. Lecoq, Time of flight positron emission tomography towards 100 ps resolution with L(Y)SO: An experimental and theoretical analysis, J. Instrum. 8 (2013) P07014.

[98] F. Anghinolfi, P. Jarron, F. Krummenacher, E. Usenko, M.C.S. Williams, NINO: An ultrafast low-power front-end amplifier discriminator for the time-of-flight detector in the ALICE experiment, IEEE Trans. Nucl. Sci. 51 (5) (2004) 1974–1978.

[99] M.V. Nemallapudi, S. Gundacker, P. Lecoq, E. Auffray, A. Ferri, A. Gola, C. Piemonte, Sub-100ps coincidence time resolution for positron emission tomography with LSO:Ce codoped with Ca, Phys. Med. Biol. 60 (2015) 4635–4649.

[100] S. Gundacker, E. Auffray, B. Frisch, H. Hillemanns, P. Jarron, T. Meyer, K. Pauwels, P. Lecoq, A systematic study to optimize SiPM photodetectors for highest time resolution in PET, IEEE Trans. Nucl. Sci. 59 (5) (2012) 1798–1804.

[101] J.Y. Yeom, R. Vinke, C. S.L.evin, Optimizing timing performance of silicon photomultiplier-based scintillation detectors, Phys. Med. Biol. 58 (2013) 1207–1220.

[102] J.W. Cates, S. Gundacker, E. Auffray, P. Lecoq, C.S. Levin, Improved single photon time resolution for analog sipms with front end readout that reduces influence of electronic noise, Phys. Med. Biol. (63) (2018) 11p.

[103] P. Lecoq, E. Auffray, S. Brunner, H. Hillemanns, P. Jarron, A. Knapitsch, T. Meyer, F. Powolny, Factors Influencing Time Resolution of Scintillators and Ways to Improve Them, IEEE Trans. Nucl. Sci. 57 (5) (2010) 2411–2416.

[104] S. Omelkov, V. Nagirnyi, A. Vasilev, M. Kirm, New features of hot intraband luminescence for fast timing, J. Lumin. 176 (2016) 309–317.

[105] P. Lecoq, M. Korzhik, A. Vasiliev, Can transient phenomena help improving time resolution in scintillators? IEEE Trans. Nucl. Sci. 61 (1) (2014) 229–234.

[106] S. Seifert, H.T. van Dam, D.R. Schaart, The lower bound on the timing resolution of scintillation detectors, Phys. Med. Biol. 57 (2012) 1797–1814.

[107] S. Omelkov, V. Nagirnyi, S. Gundacker, D.A. Spassky, E. Auffray, P. Lecoq, M. Kirm, Scintillation yield of hot intraband luminescence, J. Lumin. 198 (2018) 260–271.

[108] S. Gundacker, R.M. Turtos, E. Auffray, P. Lecoq, Precise rise and decay time measurments of inorganic scintillators by means of X-ray and 511 keV excitation, Nucl. Instrum. Methods Phys. Res. A 891 (2018) 42–52.

[109] R.F. Post, L.I. Schiff, Statistical limitations on the resolving time of a scintillation counter, Phys. Rev. 80 (6) (1950) 1113.

[110] S. Gundacker, Time Resolution in Scintillator Based Detectors for Positron Emission Tomography (Ph.D. thesis), Vienna University of Technology, 2014, 206 pages.

[111] H. Wieczorek, A. Thon, T. Dey, V. Khanin, P. Rodnyi, Analytical model of coincidence resolving time in TOF-PET, Phys. Med. Biol. (61) (2016) 4699–4710.

[112] M.W. Fishburn, E. Charbon, System tradeoffs in gamma-ray detection utilizing SPAD arrays and scintillator, IEEE Trans. Nucl. Sci. 57 (5) (2010) 2549–2557.

[113] S. Seifert, H.T. van Dam, R. Vinke, P. Dendooven, H. Löhner, F.J. Beekman, D.R. Schaart, A comprehensive model to predict the timing resolution of SiPM-Based scintillation detectors: Theory and experimental validation, IEEE Trans. Nucl. Sci. 59 (1) (2012) 190–204.

[114] S. Vinogradov, Approximations of coincidence time resolution models of scintillator detectors with leading edge discrimination, Nucl. Instrum. Methods Phys. Res. A (2017).