We are IntechOpen, the world’s leading publisher of Open Access books
Built by scientists, for scientists

5,000
Open access books available

125,000
International authors and editors

140M
Downloads

154
Countries delivered to

TOP 1%
Our authors are among the most cited scientists

12.2%
Contributors from top 500 universities

WEB OF SCIENCE™
Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com
High Sensitivity Photodetector for Photon-Counting Applications

Fabio Acerbi and Matteo Perenzoni

Abstract

In the last years, there has been a large development of low-light applications, and many of them are based on photon counting using single-photon detectors (SPDs). These are very sensitive detectors typically with an internal gain. The first candidate SPD was the photomultiplier tube (PMT), reaching a very high gain (~10^6), but there have been a large development of many other solutions, like solid-state solutions. Among them, single-photon avalanche diodes (SPADs) have been used in spectroscopy, florescence imaging, etc., particularly for their good detection efficiency and time resolution (tens of picoseconds). SPADs have been developed in silicon and III–V materials, for the NIR wavelength range. SPADs can be used as single high-performance pixels, or in arrays. SPAD arrays have imaging capabilities, with high sensitivity. Another kind of array is the silicon photomultiplier (SiPM), where all the pixels are connected to a common anode and a common cathode. SiPMs are used in nuclear medicine, physics experiments, quantum-physics experiments, light detection and ranging (LIDAR), etc., due to their high detection efficiency combined with large sensitive areas, and high dynamic range. SiPMs with many small cells present several advantages and nowadays the SPAD pitch can be reduced down to 5 μm.

Keywords: photon counting, single-photon, dynamic range, photodetector, high sensitivity, imaging

1. Introduction

In the last years, there has been a large development and an increasing interest in photodetectors for low-light applications. Single-pixel detectors, like photomultiplier tube (PMT) [1, 2], and single-photon avalanche diode (SPAD) [3], can be extended to the more complex
single-photon imagers [4]. These technologies enabled several new applications, physical experiment, and techniques. They also significantly increased the image quality and the sensitivity in biomedical and medical imaging techniques [5, 6].

Photon counting is a statistically accurate technique to measure faint light signals, based on the measurement of a random sequence of pulses generated by detection of single photons. In analog “linear” detection technique, the incoming light is composed by many photons and the output of the photodetector is typically continuous and proportional to the intensity of the light signal. In photon counting, the intensity of the light is so low that the output signal from the photodetector is composed by separated pulses (see Figure 1). Photon counting can give a better signal-to-noise ratio [1]. The baseline fluctuation and the electronic noise in the front-end are no more relevant, since the detection is based on a thresholding on the photo-generated pulses, resulting in a more “digital” approach. Photodetectors are single-photon detectors (SPDs); they have particular characteristics like a high internal gain. They produce a prompt and strong signal in response to a small photogenerated charge like the one from a single photon. This gives them the potentiality for a very good time resolution, which is exploited, for example, in time-correlated single photon counting (TCSPC) technique.

1.1. Time-correlated single photon counting

TCSPC is a technique to reconstruct the temporal shape of faint light signal. It is based on the detection of single photons of periodical light signal and on the measurements of their

Figure 1. Representation of photon flux and detector output in case of high light intensity (a) and of low light intensity (b) and representation of the time-correlated single photon counting (TCSPC) technique (c) for the reconstruction of low light level periodic light signals.
detection times. For each photon, the detection time is measured, building a histogram. After many detections, the histogram will represent the waveforms of the detected optical signal. The TCSPC principle is represented in Figure 1, showing example of detection cycles and the resulting histogram. The assumption is that the light signal intensity is low enough so that the probability of having one photon at the detector in each cycle is much less than one; thus, the probability to have more than one photon is negligible. This is to avoid distortion because the TCSPC system can detect only one event per cycle. Each measurement starts with the arrival of a pulse from the reference signal (START) and stops with the arrival of the signal related to the photon detection (STOP).

2. Single-photon detectors

Photon counting requires a very-high-sensitivity detector (down to single photon level), thus single-photon detector (SPD). Moreover, the signal produced from the detector has to have an amplitude high enough (i.e., higher than the electronic noise) to be “sensed” by the front-end electronics. In a semiconductor-based detector, for example, the absorption of a photon produces a single electron-hole pair, which is typically too low compared to the sensitivity of a front-end circuit. An internal charge multiplication mechanism is needed inside the detector. This can be obtained, for example, by secondary electron emission or by avalanche multiplication. Single-photon detectors can be divided into the following groups:

- Vacuum based: photomultiplier tubes (PMTs), micro-channel plates (MCPs), etc.
- Solid-state: electron-multiplying CCD (EMCCDs), single photon avalanche diodes (SPADs), SPAD array, silicon photomultipliers (SiPMs), quanta-imagers (QIS), etc.
- Cryogenic-temperature based: superconducting nanowire single-photon detectors (SNSPDs), transition-edge sensors (TESs), etc.

Vacuum-based detectors are mature technologies with big active areas and low noise, but they can be bulky, they require high biases, and are sensitive to magnetic fields [7]. Cryogenic-based detectors have good performance: low noise and high detection efficiency; thus, they are often used for quantum-physics experiments [8, 9]. They require typically a multi-stage cryostat, which can be bulky. Solid-state solution are compact and requires low biases (<100 V). They have a good detection efficiency and time resolution, but usually a higher noise. In the following, some of the main SPDs are better described.

The photomultiplier tubes are one of the first photodetector used for photon counting [1]. It can reach gains up to $10^6$ or $10^7$. The PMT has more than 50 years of history and has been used in a variety of applications due to its great versatility. A PMT is a vacuum tube that consists of an input window, a photocathode, focusing electrodes, an electron multiplier section (dynodes), and an anode. Incoming photons can be absorbed in the photocathode.
material and an electron-hole pair is generated. The electron can escape the material due to photoelectric effect, being then focused on the primary dynode and accelerated by the electric field. The impact creates several other secondary electrons, which are then all accelerated toward the successive dynode, and so on, until all the generated electrons are collected by the anode. Different version have been developed during the years: PMT-MCP (microchannel plate) exploits microchannels to obtain the electron multiplication instead of the dynodes. Electron, extracted from the photocathode “bounces” in this microchannel and produces secondary electrons. MCP-PMTs are among the fastest photon counting detectors [2]. Moreover, MCP can be used to build position-sensitive detectors, when coupled with a multianode structure or it can be used to create image intensifiers, typically used in front of a CCD imager (charge-coupled device) for night-vision or to build very sensitive imagers.

Among cryogenic-based SPD, superconducting nanowire single-photon detectors (SnSPDs) [10, 11] and transition edge sensors (TESs) [12] are the most used. SNSPDs are nanostructured devices based on long stripes of an ultrathin superconducting film, operated well below the critical temperature (TC), and biased with a subcritical current. The absorption of a photon produces a hot-spot region, in which the superconductivity is suppressed. The hot spot grows in size until electrons diffuse out of the spot. The current locally exceeds the critical limit, thus generating secondary hot spots. The superconductivity is destroyed and a resistive barrier is formed; thus, a voltage pulse can be detected. After a certain delay, superconductivity is restored. Recently, the performance of SNSPD improved significantly, also due to better cryo-stat solutions. Cryogenic-based single-photon detectors demonstrate high quantum efficiency (QE) at visible and near-infrared (NIR) wavelengths [10], low dark count rate, picosecond pulse-to-pulse timing jitter, and a gigahertz counting rate. SnSPDs are also recently being used with “imaging” capabilities [13].

Solid-state detectors are typically the preferable choice in applications like consumer electronics, in portable instrumentation or to build imagers. Most of solid-state SPDs are based on avalanche multiplication process, like in single-photon avalanche diodes (SPADs). The single pixel of an array of SPAD or, in general one single SPAD, can have a sensitive area between tens and hundreds of micrometers. This is significantly low compared to PMTs, but comparable with cryogenic-based detectors. To extend the sensitive area, an array of pixels can be realized, creating a bigger detector. All these pixels can be connected together, like in silicon photomultipliers (SiPMs) or each pixel can be read-out separately, creating an imager.

3. Single-photon avalanche diode

SPADs are photodetectors essentially based on a p-n junction, designed to be biased above the breakdown voltage [3]. In such conditions, the electric field is so high (typically >10⁹ V/cm) that a single carrier injected or generated into the depletion layer can trigger a self-sustaining avalanche multiplication process. As represented in Figure 2, when the
avalanche is triggered, the current rises swiftly (nanoseconds or sub-nanosecond rise time) to a macroscopic level (milliampere range). If the primary carrier is photo-generated, the leading edge of the avalanche pulse marks (with good time resolution) the arrival time of the detected photon. The current continues to flow until the avalanche is quenched by lowering the bias voltage down to or below the breakdown voltage (“quenching”) [3]: this lower electric field is no more able to accelerate the carriers at a sufficient energy. After the avalanche quenching, the bias voltage must be restored in order to be able to detect another photon (reset phase). All these operations require a suitable electronics: this circuit is usually referred to as quenching circuit.

The simpler quenching circuit is just a series resistor, with a relatively high resistance value. With this resistor, when the current in the SPAD increases, due to avalanche buildup, the voltage drop at the quenching resistor rises, and thus the voltage at the SPAD consequently decreases, reaching values close to $V_{BD}$, eventually quenching the avalanche. Then the SPAD is reset through the same resistor, restoring the bias to $V_{BIAS}$ value, with a time constant:

$$\tau_{reset} = R_Q \cdot C_{SPAD} = R_Q \cdot (C_D + C_{InCathode})$$  \hfill (1)

On the other SPAD terminal, there is the avalanche-sensing part. The front-end circuit has to sense the avalanche and to provide an output pulse per each detection. This can be done with a sensing resistor, a transistor or a trans-impedance amplifier, followed by a voltage discriminator, to obtain a digital pulse.

Figure 2. Schematic representation of a SPAD cross-section, in custom technology (a) and in CMOS technology, exploiting well and deep-well implants (b). The typical reverse I-V curve of a SPAD, with the metastable behavior is represented (c) along with the typical SPAD biasing, quenching and front-end circuit (d) and the resulting output digitalized signal (e).
Alternatively to passive quenching, active quenching or mixed active/passive quenching solutions can be used [3]. With active solutions, a transistor is used to force the bias of one of the SPAD node to either quench or reset the bias at the SPAD terminals [14]. In such way, the recharge is faster and the dead-time (i.e., the time when the SPAD is not sensitive) can be set and it is well-defined.

3.1. SPAD parameters

The active area of silicon SPAD is generally circular, to have uniform electric fields, with a diameter between 10 μm and 500 μm [15]. With SPADs and with other photon-counting detectors, due to the “digital” operation mode (different from “linear” mode), there are specific parameters identifying the performance:

- **Photon detection efficiency** (PDE), i.e., the ability to detect photons. This is the ratio between the number of detected photons and the photons arriving at the detector. PDE is calculated as the product of: (i) the quantum efficiency (QE) and (ii) the avalanche triggering probability ($P_T$). The latter is the probability that photo-generated carriers can reach the high-field region and trigger a self-sustaining detectable avalanche.

- The noise is typically divided into “primary” noise and correlated noise. The primary noise represents all the avalanche pulses due to thermally generated carriers (or generated by tunneling or field-assisted thermal generation). The dark count rate (DCR) is typically in the order of 10–1000 counts per second.

- The correlated noise for a SPAD is represented by the *afterpulsing*. During the avalanche, a large amount of carriers flows through the depleted region and some of them can be trapped in deep-levels (traps), being subsequently released with a delay, causing retriggering of another spurious avalanche, not related to photon absorption but to a previous avalanche, thus “correlated noise.” The time distribution of carrier release follows normally an exponential (or multiexponential) distribution. Its time constant depends on temperature, being slower at low temperatures. The afterpulsing probability depends on the SPAD itself, on the quenching circuit, and on the dead-time.

- The **dead-time** is the time interval after an avalanche, where the SPAD is not sensitive to another photon. This interval is necessary to recharge the SPAD and to let the traps to release the carriers without triggering a spurious avalanche. This is typically in the order of tens of nanosecond. Differently from active quenching, with passive quenching, the recharge is exponential, and thus it is not easy to identify a precise dead-time. The recharge time-constant can be used as a parameter.

- The time resolution of the SPAD, i.e., the ability of precise time-tag the photon arrival time, is another important parameter. The “*timing jitter*” or “*single-photon time resolution*” (SPTR) quantifies the time spread between the photon arrival and the pulse detection by the front-end electronics. This spread is due to the different absorption position and the statistical avalanche buildup time. It is in the order of a few tens of picoseconds [15]. To measure the timing jitter, it can be used a pulsed laser, attenuated to single-photon level. Using TCSPC
technique, the resulting time-resolution histogram shape is generally Gaussian, with an exponential tail, as shown in Figure 3. The tail is due to photons absorbed in the neutral region: once photo-generated the carriers diffuse randomly and can reach the depleted region but with a certain delay. The tail in the timing jitter histogram is particularly detrimental in some applications [16].

3.2. SPAD for the near-infrared range

Some applications require detecting single photons with a wavelength in the near-infrared (NIR) range, above 1000 nm, for example when using laser emitting in the telecom wavelengths, at 1310 nm and 1550 nm. There are SPADs made by semiconductor materials different from silicon. In particular, there has been recently a development of SPADs made with III/V materials like InGaAs/InP [17, 18] or InGaAs/InAlAs [19] or InGaAsP/InP [20]. InGaAs/InP SPADs are used to detect photons at 1550 nm. They are based on a separate absorption, grading, charge and multiplication (SAGCM) heterostructure as shown in Figure 4. The absorption layer is made of InGaAs and the multiplication layer is of InP. Between them, there

![Figure 3](image1.png)

**Figure 3.** Representation of two photon absorption cases in the depleted region and in the neutral region beneath it (a) and examples of the relative timing histograms (b). At \( \lambda = 400 \) nm, all the photons are absorbed in the depleted region, whereas at 850 nm, they are absorbed mostly in the neutral region, creating the tail.

![Figure 4](image2.png)

**Figure 4.** Schematic cross-section of an InGaAs/InP SPAD, with different layers (a). Comparison of photon detection efficiency (PDE) of some silicon SPADs and InGaAs SPADs (b): “thin” silicon SPAD [23], “thick” silicon SPAD [23], the FBK RGB technology (SiPM PDE divided by the FF) [25] and the typical state-of-the-art InGaAs/InP SPADs [24].
is one or more layers of InGaAsP, called “grading layer.” This is needed to smooth the hetero-

barrier in the valence band between the InGaAs and the InP layers. On the top of it, there is a

“charge layer” that shapes the electric field profile.

InGaAs has an energy gap of about 0.75 eV [17]. The PDE of InGaAs/InP SPAD is typically

around 40% between 1000 nm and 1550 nm (see Figure 4(b)), being limited by the thickness

of the absorption layer. This cannot be too high in order to limit the noise. The primary noise

is typically higher in InGaAs/InP SPADs compared to silicon ones. The technology of the

III/V material is commonly less mature. The main noise source is thermal generation in the

InGaAs layer down to about 225 K, whereas at lower temperatures, it is trap-assisted-tun-

deling in the multiplication layer. The afterpulsing probability is also higher in InGaAs/InP

SPADs, compared to silicon. The typical dead-time can be in the order of few microseconds.

However, despite these limitations, InGaAs/InP SPADs have been used in several applica-
tions. Advantages such as good PDE, compactness, and low power consumption make this a

competitive solution for NIR single-photon counting. They are typically used inside compact

modules and cooled at 230 K (or lower) to decrease the primary noise [24].

The high afterpulsing probability can be a limiting factor in several applications. Some solu-
tions have been proposed that aim to reduce the amount of charge flowing per each ava-

lanche. The first is based on very fast quenching circuit, reducing the bias at the SPAD quickly

once an avalanche is detected [21]. Another solution is based on fast gating the detector, for

e example, using sinusoidal signals with a frequency of few gigahertz: the bias at the SPAD is

modulated and, once an avalanche is triggered, it can last at maximum for the duration of

half a gating period [18]. This technique is called sinusoidal gating. Another approach is based

on the integration of a quenching resistor directly on the SPAD, very close to the active area.

This approach limits the overall capacitance at the node between the SPAD and the quenching

resistor, thus the avalanche charge. Some implementations of this kind of detector are called

negative-feedback avalanche diode (NFAD) [18].

4. SPAD array and low-light imagers

SPADs can be arranged in arrays, 1D or 2D, manufactured in custom process or CMOS pro-

cess. SPAD array in CMOS process has the advantage of having all the quenching, control and

read-out electronics in each pixel. Unfortunately, usually with the CMOS process, the SPAD

performance is not as good as with a custom process, due to the non-optimized implants and

electric fields. In particular, they have a higher DCR.

SPAD arrays are one of the solutions for low-light-level imaging [4, 5]. With such technol-
ygy, it is possible to have a sensitivity down to the single-photon level, working in photon-
counting mode. With the proper in-pixel circuitry, each pixel counts the number of photons

detected in the integrating period. This information is stored and then downloaded, to

build an image. Scientific imaging applications often require such low level of sensitivity,

typically with the addition of time-resolving capabilities. As few examples: fluorescence
lifetime imaging microscopy (FLIM), Raman spectroscopy, time-resolved near-infrared spectroscopy, and consumer applications like three-dimensional (3-D) imaging based on time-of-flight [26]. In particular, fluorescence microscopy is an established technique for the analysis of biological processes and relies on the measurement of the fluorescence intensity upon an excitation [22] at different wavelengths. Fluorescent light emission can occur in a variety of temporal scales, from nanoseconds upward. Its lifetime measurements add valuable information and require a specific kind of detectors, like a SPAD array with time-tagging circuitry per pixel. In the same way, 3D imaging, based on time-of-flight (TOF), is an application that is rapidly emerging in many fields [26], to have a three-dimensional image of the scene or to measure the distance from the objects. 3D imaging can be obtained by means of direct TOF or indirect TOF. The first one is based on the direct detection of the time-of-arrival of the reflected light pulse.

The first approach to have SPAD pixels with time-tagging capabilities is the addition of a time-to-digital converter (TDC) in each pixel [27]. The SPAD array will have in each pixel the photon-number information as well as the time-of-arrival information (typically of the first photon), as shown, for example, in Figure 5(a). Devices realized with this approach, showed good performance in many biological and 3D-ranging applications. However, the TDC generally occupies a significant part of the pixel area, reducing the fill factor (FF). There are also special implementations without the TDC, but performing distance measurements with indirect time-of-flight (iToF) method, by using three different counters per pixel [4].

In general, to improve the FF in a CMOS implementation, it is necessary to reduce the read-out circuitry occupation and complexity. The SPAD structure has to be placed very close to the read-out electronics, but being electrically isolated with proper implants, as shown in Figure 5(b). To minimize the pixel circuitry, it is possible to implement just the time-to-amplitude conversion

Figure 5. Example of 32 × 32 SPAD array, where each pixel contains a SPAD, the quenching circuitry, counters, and the time-to-digital converter (TDC) [27] (a). Example of possible SPAD CMOS implementation with read-out electronics isolated inside the deep p-well (b) [28]. Example of analog time-gated SPAD pixel, with a reduced electronics-complexity to obtain a higher FF (c) [5].
(TAC) in pixel [29, 30], or using an all-analog approach [5, 28], by means just of few transistors and capacitances. Figure 5(c) shows an example of analog time-gated SPAD pixel, based on analog time-gating and analog counter [5].

An alternative implementation of low-level imager is the so-called “quanta image sensor” (QIS) [31]. This is based on the concept of extending the sensitivity of a “classical” image sensor, which is not based on avalanche multiplication process. The pixel is composed of a pinned photodiode (PPD), collecting the photo-generated charge, which then is transferred to a small floating diffusion (FD). FD has a small dimension and its capacitance is very small, thus the conversion gain between charge and voltage amplitude is very high, enough to measure a single photo-electron. In a single-bit QIS, the output after each acquisition is a binary bit plane, where each bit represents the presence or absence of at least one photoelectron. A series of bit planes has to be generated, with a high-speed readout, eventually being able to create an image.

5. Silicon photomultiplier

Silicon photomultipliers (SiPMs) are arrays of many single-photon avalanche diodes (SPADs), all connected in parallel [25, 32, 43]. Each SPAD is called microcell (or cell) and has a square area with a side between 10 μm and 100 μm, whereas the overall SiPM active area can be typically between 1 × 1 mm$^2$ and 10 × 10 mm$^2$. This is one of the main advantages of SiPM over other SPDs: they can have big active area, but preserving the good performance of the single SPAD, with the additional advantage of being photon-number resolved (i.e., being able to count the number of photons arriving at the detector simultaneously, with a high dynamic range). These characteristics are becoming more and more important in a large number of applications [33–37] (Figure 6).

The silicon photomultipliers can be divided into: (i) analog SiPM [25] (see (Figure 6a), (Figure 6b), and (Figure 6c)) and (ii) digital SiPM [37, 38] (see (Figure 6d) and (Figure 6e)). In analog SiPM (aSiPM), the microcell, composing the array, is just made of a SPAD and a quenching resistor. The output current is the sum of the currents from the triggered cells in the array; thus, the output (amplitude or charge) is proportional to the number of detected photons (see (Figure 6b)). In digital SiPM (dSiPM), each microcell is typically composed by a SPAD and a more-complex quenching circuit (see (Figure 6d)). The cell provides a digital output to the internal dSiPM circuitry, which eventually digitally sum-up the signals from the microcells and can contain a time-to-digital converter (TDC) (see (Figure 6e)) to time-stamp the photons arrival times [37, 38]. In dSiPM, the signals are digital starting from the single microcell, thus less affected by the front-end noise. However, due to the more complex quenching and due to the presence of the control circuitry, the fill factor is lower as well as the photon detection efficiency.

5.1. SiPM applications

SiPMs have obtained a growing attention as alternative to the traditional photomultiplier tubes in the detection of low photon fluxes due to a number of advantages typical of solid-state...
detectors, such as compactness, ruggedness, ease of use, low operational voltage, and insensitivity to magnetic fields [32]. One of the most common applications is the detection of fast scintillation light in particle detectors that are used in nuclear medicine [32, 38] and in high-energy physics experiments [33, 34]. In these applications, SiPMs are coupled with scintillator crystals, which convert high-energy particle, X or gamma ray into visible or NUV light, being detected by the SiPM. An example of 4 × 4 mm² SiPM with some scintillator crystals (with different heights) is shown in Figure 7a. SiPMs allowed significant advancements in positron-emission tomography (PET) [35] and other medical applications, especially due to their excellent time resolution [37]. This made possible to develop PET scanners with improved performance as

Figure 6. Example of silicon photomultiplier chip (SiPM), with back contact and common top PAD (a). Example of SiPM signal, acquired with oscilloscope in persistence mode (b). Typical circuit for the readout of a SiPM, with trans-impedance amplifier (c) with the SiPM equivalent circuit, composed by quenching resistor (R_Q), quenching capacitance (C_Q), i.e., parasitic capacitance of the quenching resistor through the SPAD, and the metal grid equivalent capacitance (C_GRID). Example of digital SiPM with TDC per each subpixel [38] (d) and schematic of dSiPM with one global TDC (e).

Figure 7. Picture of a test SiPM with several scintillator crystals (to be mounted on the top of it) (a). Example of setup for the measurement of coincidence resolving time (CRT), with two SiPMs with crystals detecting two coincident gamma rays (b). Example of a SiPM TILE with 6 × 6 element of 4 × 4 mm² SiPMs (c).
regards the time-of-flight (TOF) technique, which significantly improves image quality. In these applications, SiPMs are typically grouped in TILEs, to cover areas of few centimeters squared. An example is shown in Figure 7c.

SiPMs are also becoming an interesting choice in other applications based on single-photon or few-photon detection. For example: (i) light detection and ranging (LIDAR) [39], where many cells with good detection efficiency are highly desirable, (ii) optical spectroscopy [40], where high detection efficiency and big sensitive areas are very useful, (iii) fluorescence detection, (iv) flow cytometry, (v) Cherenkov detection for physics experiments [36], etc.

5.2. From SPAD to SiPM

Moving from a single SPAD to a silicon photomultiplier, there are some additional parameters and other noise sources that have to be considered.

• The fill factor of the cell has to be included in the photon detection efficiency (PDE) calculation. The cell FF is the ratio between the sensitive area and the total area of the cell. Nowadays, typical FF for analog SiPMs are between 40 and 80%.

• Optical crosstalk between the cells is an additional source of correlated noise [43]. During the avalanche in one cell, not only the carriers can be trapped but also the secondary photons are produced [3]. Being emitted isotopically, some can reach the neighboring cells, triggering another spurious avalanche. They can be absorbed in the depleted region or in the neutral region. In the former case, there is a direct crosstalk (DiCT) (see Figure 8), giving an instantaneous triggering of the neighboring cell. In the latter case, the crosstalk event happen delayed in time (typically few nanoseconds or tens of nanoseconds), creating a delayed crosstalk (DeCT). The direct crosstalk produces a current pulse that has twice the amplitude of a single-cell event.

• Another source of correlated noise is the diffused afterpulsing, as shown in Figure 8. The secondary photon can be reabsorbed in the same cell, and the photo-generated carrier can diffuse and reach the depleted region with a certain delay, producing an afterpulsing.

• An important parameter for SiPMs is the gain of the cell, i.e., the number of carriers produced in response to a single photo-generated carrier. The presence of an integrated resistor right above each cell reduces the amount of carriers flowing per each avalanche and reduces the amount of carriers flowing per each avalanche and makes this quantity well defined. The gain is typically between few $10^5$ and $10^6$, similarly to a photomultiplier tubes.

• Finally, the single-photon time resolution (SPTR) [42] is important for SiPMs. Differently from SPAD, SPTR of analog SiPM is mainly limited by the effect of electronic noise from the front-end circuit and by the transit-time spread. The former is the effect of the baseline fluctuation due to the noise: given a limited signal slope, the baseline variation induces a threshold crossing time variation, thus worsening the time resolution. The latter is the effect of different lengths of the path connecting the cells to the common PAD.
The integrated quenching resistor also reduces the amplitude of the single-cell signal. Considering the SiPM equivalent circuit, when the avalanche switch is closed, the bias at the internal node (between SPAD and quenching resistor) is lowered, discharging $C_D$ and $C_Q$. Then, it is recharged through $R_Q$. However, the detectable signal is only the current that flows through the anode and cathode pins of the SiPM. The SiPM signal is composed by a fast peak (capacitive coupling through $C_Q$) and a slower component, due to the recharge current through $R_Q$. Both are “filtered” by the presence of $C_{GRID}$. The bigger the SiPM the larger is $C_{GRID}$, thus the higher is the filtering effect on the signal [42].

PDE is one of the most important parameter for a SiPM. It has been significantly improved over the last years. One possible improvement is having the PDE spectrally peaked in the region of interest of the specific application, for example, in the blue or in the green wavelength region. An Examples are the RGB [25] and NUV [43] technologies from FBK, made with p-on-n junction or n-on-p junction type. Another improvement is the increment of the cell FF, which can be obtained reducing the border region, i.e., the not-sensitive region between one cell and the neighboring one.

5.3. Front-end and read-out

The signal coming out from the SiPM is the superposition of many pulses, either in light or in dark. Depending on the application, it is possible to measure directly the current level produced by the SiPM, or count the avalanche pulses (photon-counting mode). If the count rate is low, the pulses are clearly distinguishable, but increasing the count rate, they start to overlap, making the counting more difficult. To avoid this situation, some techniques have been developed:

- **High-pass filtering or pole-zero cancellation.** At the front-end level, it is possible to filter the signal to remove the slow tails of the signals.
- **DLED (delayed leading-edge discrimination) method** [44]. This method consists of subtracting from the signal its delayed replica, creating some sort of high-pass filtering.
• At the device level, one producer added a third output, called fast-output [45]. In the SiPM, there is a capacitive pick-up in each microcell, connected between the internal nodes and a common output. This produces a faster signal which is used for timing purpose.

Another problem is how to distinguish the primary events from the correlated noise. 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, with a method described in Ref. [46], and used in Ref. [47]. This method is based on the collection of a train of many events, filtering and peak-detection. For each event, the inter-time and the amplitude (normalized to single-cell amplitude) are extracted. Plotting the amplitude vs. inter-time and the histogram of the inter-times, it creates a plot like in Figure 9(a). By fitting the events with high inter-times, with an exponential function, it is possible to identify and distinguish the primary events.

5.4. SiPM performance

The performance of SiPMs has been significantly improved over the last years [48–52]. Producers pushed the technology limits to obtain a higher PDE, now at levels of 50–60% (peak), with FF around 70–80%. In particular, some examples are: PDE of 34% at 400 nm, with 15 μm cell pitch [48], PDE of 49% at 420 nm, with cell size of 35 μm [51], PDE of 43% at 400 nm, with cell pitch of 25 μm [50], and PDE of 33% at 520 nm, with a cell pitch of 50 μm [52]. The primary noise has been reduced, being now at the level of 50–100 kcps/mm² [48–52] at T = 20°C (see Figure 9b). The correlated noise has also been reduced due to: (i) an improvement in the silicon materials [43] and (ii) a better cell-to-cell isolation with trenches [46, 49, 53]. In particular, some examples are: CT probability of 10%, at 8 V of excess bias, with 15 μm cell pitch [48], which increases to 35%, with 7 V of excess bias when the cell pitch is 30 μm, or CT probability of just 3% in a 50 μm cell, with 3 V of excess bias, due to better cell isolation [49]. The reduction of the correlated noise and the uniformity of gain between cells give the SiPMs...
a good photon-number resolution. Figure 9c shows a typical charge spectrum obtained with a 40-μm cell-pitch SiPM [43].

5.5. SiPM with small cells

The recent efforts to increase the PDE lead to an enhanced FF [48, 49, 54], meaning a smaller border region between cells. This gives a higher optical crosstalk. One efficient way to limit the correlated noise is to reduce the cell gain, giving a smaller amount of carriers flowing, thus smaller probability of trapping and emitting secondary photons. A lower gain can be obtained by means of small cells [54]. However, reducing the cell pitch normally means smaller FF, thus lower PDE. To have both smaller cells and a good PDE, the border structure of the cell has to be redesigned [48].

SiPM with small cells also have other advantages: a higher cell density and a shorter recovery time due to the smaller diode capacitance. Both these features increase the dynamic range of the SiPM, which for some applications is very important. Few examples are: (i) for calorimetry in high-energy physics experiments with high luminosity, (ii) in LIDAR and (iii) for prompt gamma imaging in proton therapy [55]. Short recovery time means reduced pile-up of the avalanche pulses, thus higher maximum count rate. Another interesting feature of SiPMs with small cell is their improved radiation hardness. The smaller is the cell size the lower is the performance degradation caused by the effects of radiations. Indeed, with a lower gain there is a smaller correlated noise, thus the noisy cells that have an increased primary dark count rate (DCR) due to radiation damage produce a lower number of correlated pulses, reducing the total overall DCR of the SiPM. With a lower total DCR and reduced gain, the SiPM has a smaller power consumption, even when damaged by radiation. Furthermore, due to the higher number of cells with a lower correlated noise, even after radiation damages there are a larger number of cells ready to be triggered by a light signal, thus a smaller PDE reduction.

5.5.1. High density silicon photomultipliers

One interesting example of SiPM with small cells is the so-called “high density” (HD) SiPM technology, developed in FBK [48, 54]. In HD technology, there are deep trenches (few micrometers deep), with high aspect ratio between the cells to electrically isolate them (as shown in Figure 10). The border region at the edge of each active area in the cells have been reduced to less than 2 μm, making possible to have a high FF of about 80% for a 30-μm cell-pitch SiPM and more than 50% for a 12-μm cell-pitch SiPM. For the FBK previous non-HD technology, the FF was about 60% for a 40-μm pitch SiPM [41]. The gain decreases as the cell pitch reduces: it is about $2.4 \times 10^6$ for the 30-μm pitch SiPM and $3 \times 10^5$ for the 12-μm pitch SiPM, at 5 V of excess bias. Due to the small active area and the trench isolation, the direct crosstalk probability is around 9% for 15 μm cell-pitch SiPM, which have a FF of 62%, whereas it was about 35% for non-HD SiPM with 40 μm cell-pitch (FF of 60%). The PDE, in HD technology, with n-on-p junction is peaked at 550 nm, reaching values of ~40% at 4 V of excess bias, and ~50% at 10 V of excess bias, for the 30 μm pitch SiPM-HD, and 25% for the 12 μm pitch SiPM-HD (see Figure 10).
5.5.2. Ultra-high density silicon photomultipliers

The HD technology has been further improved developing the “ultra-high density” (UHD) technology [56]. UHD SiPMs have very small cells and high cell density. All the feature size of the manufacturing process have been reduced (e.g., contact dimension, resistor width, etc.). The cells have a circular active area, to avoid corners with smaller electric field, and they are arranged in a honeycomb configuration (see Figure 11). The border region is now less than 1 μm. UHD SiPMs have been produced with cell pitch between 5 μm and 12.5 μm. With a pitch of 5 μm, the FF is about 40%, and it is higher than 70% for the 12.5-μm cell-pitch SiPM. The cell density is between 7400 cell/mm$^2$ and 46,190 cells/mm$^2$. Moreover, very small cells mean low gain, low correlated noise, and very fast recovery time.

The realization of very small cell sizes poses different challenges not only in the design and in the microfabrication process, but also due to intrinsic problems. The “border region” at the edge of the high-field region is no longer negligible but starts to play a very important role. Figure 12 shows the TCAD simulation of electric field inside a 10 μm SiPM cell. The effective region, where the electric field is high, is smaller than the nominal one (defined by layout). Moreover, the depleted region close to the trench extends laterally toward the center of the cell. This makes the carriers photo-generated in that region to drift laterally toward the peripheral region, instead of drifting vertically, thus they are not detected. These effects are collectively called “border effect.” This issue worsens the

---

**Figure 10.** Schematic representation of RGB-HD SiPM structure (a). SEM image of a 12-μm cell pitch SiPM with details of the single microcell (b). Measured performance of high density SiPMs: gain (c), direct crosstalk probability (d), and PDE vs. wavelength (e) of RGB-HD and one NUV-HD (e).
performance of small cells, and it is more and more important as the cell pitch decreases. To reduce this issue, a new version of the cell has been designed, with modified doping profile. The new version has a wider high-field region and a reduced lateral depletion underneath the junction.

UHD SiPM with new structure have a higher PDE: an UHD SiPM with 7.5 μm pitch reaches a PDE of about 30% in the peak, with 6 V of excess bias at a wavelength of 470 nm, whereas a 10-μm pitch SiPM have a PDE peak of ~40% in the same conditions. SiPM with 5 μm cell pitch reaches a PDE higher than 15%. Due to the small cell capacitance, the signals from UHD SiPMs are very fast, in the order of few nanoseconds FWHM, as shown in Figure 11. The noise of UHD SiPM is generally in the order of few kcps/mm^2, but in the new structure, it is higher. This is probably due to an increased electric field; but, this technology is relatively new and there is room for improvement. The gain of the cells, thus the crosstalk probability, is low (even without absorbing material in the trenches). In a 7.5 μm cell, the gain is about 2 × 10^5, at 6 V of excess bias, and the direct crosstalk probability is smaller than 5%.

Figure 11. SEM image of 10-μm cell ultra-high density SiPM (a), showing active areas, metal and polysilicon resistors. Nominal FF of UHD, HD and non-HD technologies from FBK (b). Typical single-cell signals of UHD SiPMs (c).

Figure 12. TCAD simulations of electric fields inside a 10-μm pitch and a 7.5-μm pitch microcells of the ultra-high density SiPMs, with depleted region highlighted.
6. Conclusions

We have reviewed some of the most interesting photodetectors technologies for photon counting. Solid-state solution, like single-photon avalanche diodes (SPADs), is able to reach high detection efficiencies and good time resolution, in the order of few tens of picoseconds. SPAD can be made in silicon or III/V materials, for the detection in the NIR wavelength range. Building a silicon SPAD in CMOS technology, it is possible to integrate some electronics into each pixel, to count the photons and to time-tag them. An array of such kind of SPAD pixels can be exploited to create low-light imagers. SPAD imagers are nowadays used in several biomedical applications (e.g., FLIM, Raman spectroscopy, etc.) and in 3D ranging. This technology is continuously evolving with a pixel density getting higher due to the CMOS technology development. Another interesting detector based on SPAD arrays is the silicon photomultiplier (SiPM). Here, all the SPADs are connected in parallel, in analog or digital way. It has single-photon sensitivity but, differently from single-SPADs, it is able to reach big active areas (few millimeters squared) and it is able to count the number of photons arriving simultaneously with good photon-number resolution. SiPM performance has been significantly improved over the last years, reaching a high FF and a high detection efficiency. This promising technology is now starting to be used not only in the typical applications (e.g., nuclear medicine and physics experiments) but also in biomedical and 3D ranging applications. SiPM technologies are evolving in the direction of smaller cells (SPADs), which is advantageous for SiPM performance, but it requires improvements in the manufacturing processes. For example, there are new technologies for SiPMs with cell-pitches smaller than 10 μm, down to 5 μm. This provides a higher cell density and a larger dynamic range. Some mixed solutions are also emerging, with imagers made by an array of many mini-SiPMs. This allows to have imaging capabilities but with a high dynamic range per pixel.

Author details

Fabio Acerbi* and Matteo Perenzoni

*Address all correspondence to: acerbi@fbk.eu

Fondazione Bruno Kessler (FBK), Trento, Italy

References


[40] Dalla Mora A et al. Fast silicon photomultiplier improves signal harvesting and reduces complexity in time domain diffuse optics. Optics Express. 2015;23(11)


[56] Acerbi F et al. High efficiency, ultra high-density silicon photomultipliers. IEEE Journal of Selected Topics in Quantum Electronics; March-April 2018;24(2)