1. **Introduction**

The silicon photomultiplier (SiPM), also known as a multi-pixel photon counter (MPPC), is a light sensor consisting of a matrix of thousands of single-photon avalanche diodes (SPADs) called microcells, or pixels [1]. All the pixels are connected in parallel to the same feed through quenching resistors. When a photon is absorbed by a pixel, a significant electronic pulse is generated upon output; tens of micro-amps from a single pixel, as a result of the avalanche that occurs at the SPAD junction. The electrical charge generated as a result of absorbing the photon depends only on the junction voltage and capacitance, and does not depend on the energy of the striking photon. This state is called Geiger mode. Thus, a semi-analog output is obtained, and its size depends on the number of photons that interacted with the SiPM sensor. The SiPM is unique since thanks to the high intrinsic (internal( amplification, it is the first solid-state device that has been able to identify a single photon at room temperature [2]. The SiPM was developed in the early 2000s in a collaboration between the Moscow Engineering Physics Institute (MEPhI) and Pulsar. From around 2005, many companies recognized the potential of the new technology and began to develop and manufacture poly-silicon-based SiPM components. Among the leading companies in the field are Hamamatsu, SensL, FBK, STMicroelectronics, KETEK, and more. The SiPM replaced older light sensors in many applications with low luminosity (signal/information). Among other things, SiPM components are used as light sensors in ~~range detection,~~ light detection and ranging (LIDAR) systems, which are used in the navigational systems in autonomous cars, quantum informatics [3], for the purposes of spectroscopy in physical experiments, and more. In addition, SiPM components are used as light sensors in scintillation-based radiation sensors. They are used in medical imaging devices such as positron emission tomography (PET) and single photon emission tomography (SPECT), nuclear safety products (homeland security), radiation detectors, and detectors for space exploration. Its advantages over other light sensors, for example, the photomultiplier tube (PMT), are resistance to electromagnetic interference (an especially important advantage in the context of medical imaging devices), low operating voltage (tens of volts), compactness, and cost. Together with the above, because an SiPM is an array of solid-state devices that are powered by high back voltage, the SiPM suffers from significant dark currents, some of which are the result of tunneling. The tunneling phenomenon becomes more dominant for ~~low temperatures,~~ cryogenic temperatures. The function of the SiPM at such temperatures is especially interesting for SiPM-based detectors located on satellites, and use of SiPM-based detectors in particle exploration. In both cases, detectors are exposed to especially low temperatures, in which the process of tunneling is dominant. In this work, we’ll review the SiPM and its features, we’ll review tunneling and its effect on the dark current of the component, and we will present the use of tunneling in the configuration of the SiPM, to expand the dynamic range of the device.

1. **Introduction to SiPMs**

The SiPM is assembled like a matrix of thousands to tens of thousands of pixels, single-photon avalanche diodes (SPADs), each one connected through an RQ resistor to common-mode bias voltage. The analog SiPM output is a junction that connects all the SPAD cell anodes. Figure 2 shows a schematic diagram of the SiPM structure. The load resistor, RL, is not part of the device, rather it is externally connected.

The SPAD cells are also called Geiger-mode avalanche photodiodes (Gm-APD). They are a solid-state device that exploits the avalanche that occurs in the depletion layer to create an internal amplification mechanism. Avalanche photodiodes (APD) were investigated at voltages close to breakdown voltages as early as the 1960s. In the 1990s, APDs were presented for the first time operating in Geiger mode above the breakdown voltage; in this state any discharge produced was “digital” with an identical charge independent of the energy of the striking photon. These components are called GM-APD or SPADs. Figure 3 shows the division between the working range of the SiPM (as a component based on SPADs) and the working range of the APD.

Besides the physical operating voltages, SPADs differ from APDs in their internal structure, in the internal amplification (hundreds for APDs, in comparison with 106 for SPADs) and in the signal-to-noise ratio. In addition, with APDs the avalanche process fades on its own, without the need for restraint.

Today, SPADs are made out of silicon, mainly using CMOS technology. Each SPAD pixel is in fact a p-n junction that was designed to be powered above the breakdown voltage of the junction. The field generated in the depletion area is on the order of 105 V/cm. An electron generated in the depletion area creates an avalanche of electrons thanks to the high electrical field. This is a self-sustaining avalanche. The current generated causes a voltage drop on the RQ resistor which causes the voltage on the cell to fall to the breakdown voltage, which leads to fading of the avalanche process. Figure 4 describes the avalanche process in a single pixel.

In the first stage, the SPAD is powered at VB voltage, beyond the breakdown voltage VBR. At this stage, no current flows through the pixel. During the formation of a free charge carrier in the depletion layer, if as a result of a photon strike and electron release as a part of the photoelectric effect, or as a result of thermal excitation of an electron, the charge carrier is swept away in the high electrical field. The high speed of the charge carrier causes the release of additional charge carriers that are also swept away, and in turn release additional charge carriers, and as such intrinsic charge amplification is generated. The current flows through an RQ resister (with a typical value of several hundred kilo Ohm), which causes a junction voltage drop close to the breakdown voltage and stoppage of the avalanche process. The cell voltage rises back to the bias voltage, a reset process, with a time constant of the pixel capacitance multiplier, and the parasitic capacitance of the RQ resister at the output load of the SiPM, as described in Equation (1).

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

where

CSPAD is the capacitance of the p-n junction.

CQ is the parasitic capacitance of the RQ resistor.

Figure 5A shows the structure of a single SPAD unit. Figure 5B shows an X-ray image of a SPAD unit. The implementation of the RQ resistor can be identified around the luminosity-sensitive area.

Implementation of an RQ resistor in an integral configuration shown in Figure 5B enables obtaining the Fill Factor (FF), an optimal ratio between the sensitive area of the sensor and the total area.

1. **The breakdown voltage**

The breakdown voltage, VBD, is the voltage from which the doubling coefficient started of the charge carriers formed as a result of interaction with the initial charge carrier increases sharply. VBD voltage is the boundary between the linear doubling range that characterizes the APDs and the avalanche ranges that is typical of SPADs. For SiPMs in which an RQ resister is implemented outside the device, a very sharp increase in current and breakdown voltage can be identified relatively easily. In SiPMs in which the RQ resistor is implemented integrally, the internal amplification is smaller by an order of about three (from 108 to about 105; in such a situation, it is necessary to perform a more complex analysis for accurate identification of the breakdown voltage. The exact identification of this voltage is necessary for obtaining uniform amplification when working with an array of SiPMs. Breakdown voltage, VBR, depends mainly on the level of doping of the junction and the temperature. An additional definition of the breakdown voltage is possible by “ionization integral”, where the breakdown voltage is the voltage at which the integral across the depletion layer is equal to 1, as displayed in Equation 2,

|  |  |  |
| --- | --- | --- |
|   |  | 2 |

Where

αp , αn, are the ionization constants of electrons and holes.

W is the width of the depletion layer.

In general, values of the ionization constants vary between different models [4], and present an exponential relation to the electric field at the junction. Figure 6 shows ionization constants of various materials as a function of the electrical field. It can be seen that for silicon, the ionization constant for electrons, αn, (solid line) is greater than the ionization constant for holes, αh, (dotted line) in every range.

The breakdown voltage VBR, rises as the temperature rises. This increase can be explained as follows: that when a charge carrier is swept away in the high electric field, collisions with the atoms in the lattice are formed. In the first approximation, the threshold energy for which that charge carrier created another electron-hole pair is ETH = 3/2Eg , under the assumption of an identical mass for electron and hole. Every time the initial charge carrier is accelerated and strikes a phonon or atom with less energy than the threshold energy, ETH, it loses some of its energy without the creation of secondary charge carriers. The rise in temperature causes a rise in the probability of collisions, and therefore the average energy for collisions decreases. This causes the need to raise the field (and therefore the voltage) at the junction to obtain a self-sustaining avalanche, as required in Geiger mode. Also, the width of the depletion layer affects the breakdown voltage. It would appear, that a large depletion layer causes more collisions at the junction, and therefore causes a lower required critical field for switching to Geiger mode; but since the breakdown voltage is integral on the field across the entire width of the junction, increasing the depletion layer actually causes an increase in the breakdown voltage. When the depletion layer increases, so does the effect of the temperature on the breakdown voltage. With the occurrence of an avalanche, a peak current Lf is formed as described in Equation 3.

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

Voltage VOV is the device voltage beyond the breakdown voltage. Equation 4 presents the voltage on the junction VF, at the peak of the avalanche process,

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

Where RD is the internal resistance of the pixel. When the voltage on the junction falls to voltage Vf, the avalanche ceases to sustain itself.

A rise in the electric field near the device terminals causes a rise in the multiplying coefficient and lowering of the breakdown voltage near the ends [5]. To reduce this phenomenon, which affects the uniformity of the amplification, use is made of a structure called a guard ring. The structure is made using P-well technology, and its goal is to reduce the electrical field near the ends, and thus prevent the beginning of an earlier avalanche in the area. An example of a structure and simulation of the electrical field with guard ring is shown in Figure 7. The disadvantage in this is the reduction of the active area by 1-2 μm and obtaining a lower FF. This is rare in SiPMs based on large pixels (for example, 50x50 μm2), but is significant in components based on smaller pixels (10x10 μm2).

1. **Internal Amplification**

Internal amplification is defined as the area under the output pulse of the SiPM, i.e. the total of the entire charge at the output for a single avalanche. The amplification is given according to

|  |  |  |
| --- | --- | --- |
|  |  | (5) |

Where qe is the electron charge. In general, the amplification is linear in the first approximation. Factors affecting the linearity of the SiPM are lack of uniformity of the cells, both in terms of uniformity of the doping and in terms of CQ values. Similarly, the dynamic range of the SiPM is affected by the density of the cells for a field unit and the time of the recovery process of the cell (reset). The larger the number of cells per field unit and the smaller the “dead time”, in which the cell is disabled from detecting new events, the smaller the dynamic range of the amplification of the SiPM in relation to the luminosity. The amplification is on the order of 105-107 amplification that generates a detectablesignal, beyond the noise level, for a single photon. This has consequences for the power requirements and noise level of electronics fed by the SiPM, in comparison with other light sensors. The noise added as a result of the amplification variance in the various pixels, excess noise factor (ENF) is defined according to Equation 6.

|  |  |  |
| --- | --- | --- |
|  |  | (6) |

Where

σG - amplification variance between SPAD cells.

<G> - Average amplification in the various cells.

The internal capacitances CSPAD and CQ depend on the size of the SPAD. Therefore, a smaller cell provides a smaller amplification, and therefore reduces the signal to noise ratio (SNR). At the same time, a small cell reduces the noise (as described in Chapter 3), and also reduces the reset time. Reducing the reset time also contributes to the reduction of the SNR by enabling the reduction of the signal integration time, and thus reducing the noise integration. Figure 8a shows the pixel amplification and signal amplitude at the output, 8b, obtained from an FBK SiPM measuring 1x1 mm2. As can be seen for an increase in VOV voltages, the linear amplification is impaired [6]. This phenomenon is a result of penetration of the depletion layer into the device substrate, the epitaxial layer. Penetration of the depletion layer leads to a decrease in capacitance of the diode, CSPAD, which leads to reduction in amplification in accordance with Equation 5.

1. **Signal Form**

The signal amplitude is not dependent only on the amplification. The reason for this is that it is possible to separate the output signal into two separate components, a fast component and a slow component [7]. The fast component is a result of the discharge path through the parasitic capacitance CQ, the slow path is the result of the charge current of the reset process through resistor RQ. Figure 9 shows an electrical diagram of an SiPM with N pixels. The left rectangle is the schematic diagram of the SPAD in which the avalanche occurs, the middle rectangle is the schema of the rest of the passive cells, and the capacitance from the right describes the parasitic capacitance of the grid which connects the pixels. The red path is the rapid current supply path at the beginning of the avalanche. The blue path is the slow path, the reset path of the cell. The time constants of the output pulse are given according to equations7-9.

|  |  |  |
| --- | --- | --- |
|  |  | (7) |
|  |  | (8) |
|  |  | 9)) |

Where

RL is the load at the SiPM output.

RD is the internal resistance of a single cell.

Ctotal is the total output capacitance of the pixels.

The general form of the signal is given according to equation 10,

|  |  |  |
| --- | --- | --- |
|  |  | 10)) |

Where Qtotal is the total charge released from all the pixels in the component.

Similarly, the output signal depends on the load and bandwidth of the electronics being fed. Figure 10 shows different output signals, as a function of/ depending on the bandwidth of the front-end amplifier (FEA) which is fed by them. For the FBK device, the bandwidth of the high-speed component is in the range of 50-250 MHz. Therefore, for an FEA having bandwidth of 20 MHz, only the slow component of the signal remains.

**4. Detection Efficiency**

The detection efficiency of the SiPM, called photon detection efficiency (PDE), is the probability that a photon striking an SiPM cell will indeed produce an avalanche, or in other words, it is the ratio of the number of striking photons to the number of photons detected. The PDE depends on both the voltage supplied to the SiPM beyond the breakdown voltage, VOV, and the wavelength of the striking photon, λ. The PDE value is given according to Equation 11.

|  |  |  |
| --- | --- | --- |
|  |  | (11) |

Where,

QE is quantum efficiency

AIP is avalanche initiation probability.

FF is fill factor --geometric efficiency

The quantum efficiency, QE, is the probability of a photon striking the detector to enter the detector (and not be returned, for example) and be absorbed in the active region. The probability of an avalanche being initiated, AIP, depends on the intensity of the electric field and the location of the formation of charge carriers in the cell. The FF is the probability that the photon will strike the active area in the array and not between two pixels or in the area of the guard ring. Since the low field width around the guard ring depends on the voltage, the FF also depends on the VOV voltage. In addition, since the shape of the field near the ends depends on the location of photon interaction in the depth of the SPAD, the FF also depends on the wavelength of the striking photon.

The PDE obtained for components based on a p-on-n junction is different from the PDE obtained for components based on an n-on-p junction, since the absorption location of the photon in the cell depends on the wavelength of the striking photon [8], as can be seen in Figure 11.

While red-green light (500-780 nm) is absorbed within the high field area (assuming a junction width of hundreds of nanometers), blue light (450-490 nm) is absorbed close to the detector before the junction. As shown in Figure 6, electrons have a higher ionization constant than holes. Since for a p-on-n and n-on-p structure the field directions are reversed, therefore the p-on-n structure has higher PDE values for wavelengths of blue light, while for green red light the n-on-p structure provides a higher detection probability. Figure 12A shows PDE curves as a function of wavelength for SiPM components in the configuration of p-on-n and n-on-p for epitaxial layers with width of 4 μm and 8 μm. Figure 12B shows quantum efficiency curves as a function of striking light wavelength for various substrate widths.

From Figure 12a, it can be seen that the p-on-n-based device gives a better response for the blue light range. This is because for a p-on-n junction interaction before the junction results in electron drift to the junction for which the ionization constant is higher in comparison with light at higher wavelengths for which the interaction was done after the junction and therefore formed by drift of holes.

**5.Noise Factors**

The noise in SiPM components can be divided into two main groups:

1. Main noise, dark current, the source of this noise is charge avalanches created as a result of charges created as a result of thermal effects or charge carriers generated as a result of tunneling in the high electric field [9].
2. Secondary noise, excess noise, secondary avalanches caused as a result of a main event that happened in the component. The secondary noise is also classified into two types, Afterpulsing, an avalanche caused in the same SPAD in which the main event occurred, and Crosstalk, an avalanche caused in cells near the pixel in which the main event occurred.

The SiPM output is also proportional to the charge generated as a result of secondary noise factors mentioned above. This value is defined by the ratio of the total charge to the secondary charge, excess charge factor (ECF), as defined in Equation 12,

|  |  |  |
| --- | --- | --- |
|  |  | (12) |

Where

QOut is the average total charge at the SiPM output.

QPrimary – Total charge due to noise, afterpulsing, dark noise, and crosstalk events.

Dark Current

The dark current is current composed of avalanches created not as a result of interaction with a photon in the pixel in which the avalanche was made or in a neighboring pixel. The dark current is a result of the charge carriers created as a result of thermal transfer or tunneling in the substrate layer of the cell, or at the end of the depletion layer adjacent to the junction. The dark current is the main source of noise in SiPM components. It increases linearly together with VOV [11] (assuming that the added effect of the crosstalk and afterpulsing is negligible). The rate of dark current events, the Dark Count Rate (DECR) can be calculated for a given work point (work voltage and temperature), and when the measurement is performed in conditions of darkness, according to

|  |  |  |
| --- | --- | --- |
|   |  | (13) |

Where

ID is the SiPM current at the given work point.

G is the amplification of the SiPM at the measured work point.

Figure 13 shows the dependence of voltage of the DCR for SiPM components of various manufacturers with pixels of various sizes. It can be seen that in some of the components for high voltage, the function becomes parabolic. This is because the afterpulsing and crosstalk that are created as a result of the initial dark current event. In addition, it can be seen that the larger the pixel area, the larger the DCR. This can be explained by the fact that for larger pixels, the fill factor, the effective area for detection increases, while the dead area, which was created as a result of the RQ restraint resistor and the separation between the cells, decreased.

At room temperature, the dark current is generated mainly in Shockley-Read-Hall (SRH) recombination. A situation in which electron-hole pairs pass from the conduction level to the valence level by “local energy states” generated in the forbidden strip by defects in the lattice, is called a trap. The transition between states is based on passing of thermal energy through a phonon. The frequency of noise, the DCR, is dependent on the quality of the epitaxial layer lattice. In addition, the DCR depends on the volume of the depletion layer, i.e. the size of a single pixel and the width of the depletion layer. In addition, creation of SRH pairs is also affected by the high electric field, and is expressed in processes such as the Pole-Frenkel effect, in which electron-hole pairs are created in a non-conductive material, under the influence of a high electric field, or a thermal-enhanced trap-assisted-tunneling which will be explained at length in the next chapter. All these phenomena can be moderated by lowering the electric field – i.e. by working at a lower VOV voltage. For higher temperatures, the diffusion current from neutral areas adjacent to the depletion level can also cause an increase in DCR. For low temperatures, the DCR rate decreases significantly by a factor of 2 for every 10⁰C. At the same time, at lower temperatures, production of pairs as a result of tunneling becomes dominant, and the dependence on temperature becomes negligible. This is especially significant for applications designed for low temperature conditions such as particle detectors, and detectors carried on satellites or research balloons.

The knee temperature, in which the dominant process for creating the dark current passes from thermal production to a process of variable tunneling from one SiPM device to another dependent on the device structure, but is dependent mainly on the size of the electrical field at the junction at breakdown voltage. For a larger electric field, the tunneling is more dominant. Figure XXX shows DCR curves for mm2 as a function of VOV voltage for various temperatures. A decrease of the DCR rate can be seen with a decrease in temperature. In addition, it can be discerned that starting with a VOV of 2V, the DCR has a linear rise together with VOV.

Afterpulsing

The afterpulsing noise is a secondary noise resulting from trapping and delayed release of charge carriers in the high field. The release time of the trapped charge carrier distributes exponentially and therefore the probability of obtaining an afterpulsing event is smaller the further away from the main event. When the charge carrier is released, it causes a separate avalanche. The probability of obtaining afterpulsing depends on the number of traps, on the area of high field, and also on the ratio of the time constant for release from the traps to the reset time of the avalanche. It can be understood from this, that it is possible to decrease the frequency of the afterpulsing by increasing the reset time, but this is at the expense of the dynamic range of the SiPM, as it becomes saturated at a lower rate of striking photons. In addition, a longer pulse requires a longer integration time, which increases the electronic noise collected by the electrons fed by the SiPM. Another configuration of afterpulse generation is the result of optical production. During the avalanche, secondary photons are formed as a result of charge carriers hitting lattice atoms. These photons can be absorbed in the neutral area outside the depletion layer. Absorbing the photon causes production of a charge carrier that can diffuse to the depletion layer and initiate a secondary avalanche in the same SPAD where the original avalanche was formed. Figure 15 shows an afterpulsing event created as a result of optical production. Corresponding to the lifetime of the charge carrier release from the trap, for optical production, is the lifetime of the charge carrier in the epi-layer. This time varies from several nanoseconds to hundreds of nanoseconds. In order to reduce the afterpulse that results from optical production, a substrate with a shorter lifetime than the reset time of the SPAD should be used. In such a situation, the contribution of the charge carrier to the avalanche would be negligible. Another possible solution to reduce the afterpulse that results from optical production is to create another p-n junction that blocks the charge carriers from diffusing to the depletion layer.

Crosstalk

As described above, during the avalanche process, secondary photons are formed. The quantity of photons was estimated to be about 3×10-5 for every charge carrier in the avalanche [7]. The emission of photons as a result of the avalanche is an isotropic process, i.e. the photons distribute in various directions at equal probability. These photons can cause an avalanche in pixels neighboring the pixel in which the initial interaction took place. Such a response is called direct optical crosstalk, and it is manifest in obtaining an increased output signal relative to the signal that emanates from the optical signal. Another form of optical crosstalk is delayed optical crosstalk. It results from the formation of electron-hole pairs in the neutral area outside the depletion layer. Again, as in the case of afterpulsing, the charge carriers are formed as a result of photons being absorbed in the substrate. The charges diffuse and some reach the depletion layer, and can form there a secondary avalanche with a delay of a few nanoseconds to microseconds after the initial pulse. Figure 15 illustrates the path of the photons and charge carriers.