Available online at www.sciencedirect.com
ScienceDirect
Nuclear Instruments and Methods in Physics Research A 568 (2006) 224–232
NUCLEAR INSTRUMENTS & METHODS IN PHYSICS RESEARCH Section A
www.elsevier.com/locate/nima

A new Silicon Photomultiplier structure for blue light detection

Claudio Piemonte*

ITC-irst, Divisione Microsistemi, 38050 Povo di Trento, Via Sommarive, 18, 38050 Povo di Trento, Italy

Available online 7 August 2006

Abstract

Silicon Photomultipliers are extremely promising devices for those applications requiring the detection of very low-intensity light (down to single photon detection). The major drawback of the existing prototypes is the poor detection efficiency, especially at short wavelengths (below 10% in the blue region). In this paper, a new structure aimed at improving this parameter at wavelengths ranging from 400–450 nm is presented. With respect to a conventional structure it allows a maximization of the breakdown initiation probability for a given bias voltage and a reduction of the dead area. The analysis is supported by TCAD simulations.

PACS: 29.40.Wk; 85.60.Bt; 85.60.Ha

Keywords: Silicon photomultiplier; Device modeling

1. Introduction

Most of the systems used for the detection of very low-intensity light are based on photomultiplier tubes (PMTs). Only recently, new types of detectors, based on silicon diodes working in avalanche regime, have been developed and proved to be extremely interesting candidates to replace the existing vacuum-based systems [1]. Some of the advantages offered by the solid-state solution are: insensitivity to magnetic fields, ruggedness, compactness, low operating voltage and long lifespan. In addition, this technology facilitates the interconnection between the detector and the read-out electronics.

A diode working in a region near the breakdown voltage can be operated in 2 different ways depending on whether the bias voltage is below or above the breakdown point. In the first case the device is called avalanche photodiode (APD). Each absorbed photon creates in average a finite number M of electron–hole pairs exploiting the impact ionization process. This mode of operation is called linear because the number of collected carriers is proportional (by a factor M) to the number of absorbed photons.

In the second case the device is referred to as Geiger-mode APD (GM-APD). In this bias condition, the electric field is so high that a single carrier injected into the depletion region can trigger a self-sustaining avalanche. The carrier initiating the discharge can be either thermally generated (noise source of the device) or photogenerated (useful signal). The main limitation of a single diode working in GM is that the output signal is the same regardless of the number of interacting photons. In order to partially overcome this limitation, the diode can be segmented in tiny micro-cells (each working in GM) connected in parallel to a single output. Each element, when activated by a photon, gives the same current response, so that the output signal is proportional to the number of cells hit by a photon. The dynamic range is limited by the number of elements composing the device, and the probability that 2 or more photons hit the same micro-cell depends on the size of the micro-cell itself. This structure is called Silicon PhotoMultiplier (SiPM) (for example see [1]).

An interesting application of SiPMs is the detection of the light emitted by scintillators. Among the various types of scintillators, particular attention has recently been given to lutetium oxyorthosilicate (LSO) for its high light yield, short decay time and relatively good mechanical properties. These features are extremely useful, for example, in

*Corresponding author. Tel.: +39 0461 314 428; fax: +39 0461 314 591.
E-mail address: piemonte@itc.it.

0168-9002/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.nima.2006.07.018

positron emission tomography (PET). The spectrum of the emitted light is peaked at 420 nm. Unfortunately, the existing SiPMs have a detection efficiency of only few percent at these wavelengths, so they are not suitable for this use [1].

The first part of the paper reviews the main factors governing the functioning of a SiPM and the parameters determining the detection efficiency. In the second part, an analysis of the detection efficiency of conventional SiPMs (based on shallow n+/p or p+/n junctions) at short wavelengths is given, highlighting their limits. Finally, a new SiPM structure aimed at optimizing the detection efficiency in that part of the spectrum is presented.

2. Operation principle of a SiPM

As mentioned in the previous paragraph the SiPM is a matrix of GM-APDs connected in parallel. A schematic representation of the device is shown in Fig. 1(a). The connection between the cells is made on one side by the low-resistivity substrate and on the other side by a metal layer. The diodes (labelled as D) are asymmetric p–n junctions with a suitable edge structure (guard ring) in order to lower the electric field at the borders.

Each GM-APD has in series a quenching resistor (R_Q) which is needed to stop the avalanche current and, then, to restore the initial bias condition enabling the detection of a new incoming photon. A circuit model, which emulates the evolution of the signal of a GM-APD was developed in the 1960s to describe the behaviour of micro-plasma instability in silicon [2,3]. According to this model, the pre-breakdown state can be represented as a capacitance (junction capacitance, C_D) in series with the quenching resistor. Referring to Fig. 1(b) this state corresponds to the switch in the OFF condition. In steady state, the capacitance is charged at $V_{BIAS} > V_{BR}$ where V_{BR} is the breakdown voltage and V_{BIAS} is the operating voltage.

When a carrier traverses the high-field region, there is a certain probability, known as turn-on probability P_{01}, to initiate an avalanche discharge. If this happens, the new state of the system can be modelled adding to the circuit a voltage source V_{BR} with a series resistor R_S in parallel to the diode capacitance (switch closed in Fig. 1(b)). R_S includes both the resistance of the neutral regions inside the silicon as well as the space charge resistance. C_D, originally charged at $V_{BIAS} > V_{BR}$, discharges through the series resistance down to the breakdown voltage with a time constant τ_D given by the product $R_S C_D$. It should be noted that the discharge current is initially limited by the build up of the avalanche process which can take some hundreds of ps. Since R_s is in the order of 1 kΩ, this time can be similar to τ_D for small diodes.

As the voltage on C_D decreases, the current flowing through the quenching resistance, and as a consequence through the diode, tends to the asymptotic value of $(V_{BIAS}-V_{BR})/(R_Q+R_S)$. In this final phase, if R_Q is high enough, the diode current is so low that a statistical fluctuation brings the instantaneous number of carriers flowing through the high-field region to zero, quenching the avalanche. The probability of such a fluctuation (turn-off probability P_{10}) becomes significant when the diode current is below 10–20 μA (defined as latching current, I_L). The average time needed to stop the avalanche, when this condition is satisfied, is in the order of 1 ns. The latching current poses a strict limit on the lower value of R_Q to some hundreds of kΩ.

As the avalanche process is terminated, the switch is again open and the circuit is in its initial configuration. The capacitance charged at V_{BR}, starts recharging to the bias voltage with a time constant $C_D R_Q$, and the device becomes ready to detect the arrival of a new photon. The number of carriers created during an avalanche discharge is given by $1/q(V_{BIAS}-V_{BR})C_D$ where q is the electron charge.

Each diode composing the SiPM reacts independently in the above-described way. Thus, if n cells are activated at

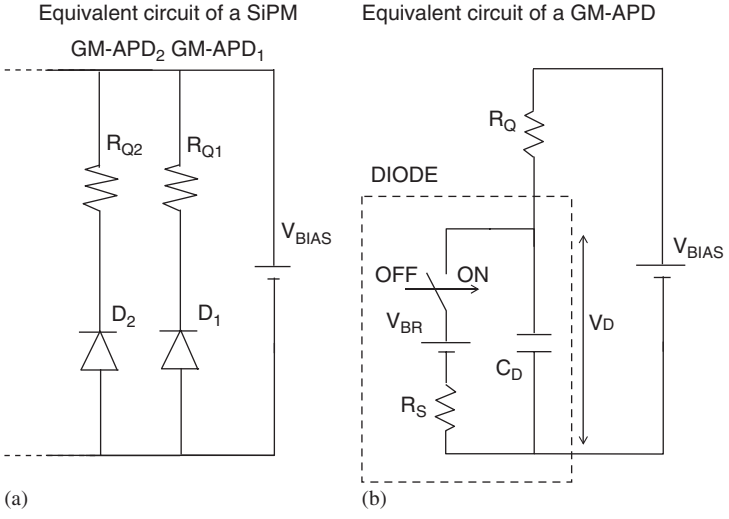

Fig. 1. Equivalent circuits of a SiPM (a) and of a single GM-APD (b).

the same time, the charge measured at the SiPM output is n times the charge developed by a single GM-APD, giving information on the light intensity.

As for every detector, the performance of a SiPM is determined by 2 features: noise and detection efficiency. The latter, which is the main topic of this paper, will be discussed in the next session. Concerning the noise, its origin is the fluctuation of non-photogenerated carriers triggering a discharge. Secondary sources are carriers trapped during a discharge and released after the recovery time (afterpulsing) and carriers generated by photons emitted during the discharge of neighbouring GM-APDs (optical cross-talk). The latter source can be suppressed by optically isolating each cell, for example by etching a trench in the border region and covering it with an aluminium layer.

3. Detection efficiency

The detection efficiency of a SiPM is the product of 3 factors: the quantum efficiency, the turn-on probability (from now on referred to as triggering probability) and the geometrical efficiency. All these factors have a relevant impact on the overall efficiency and must be carefully optimized.

3.1. Quantum efficiency

The quantum efficiency (QE) represents the probability for a photon to generate an e–h pair in the active thickness of the device. It is given by the product of 2 factors: the transmittance of the dielectric layer on top of the silicon surface and the internal QE. Both are wavelength dependent. The former can be maximized, by implementing an anti-reflective coating (ARC) [4]. The second term represents the probability for a photon that has passed the dielectric layer to generate an e–h pair in the active thickness. In a conventional n+/p/p+ diode, the active layer is roughly limited on top by the undepleted n+ layer, whereas on the bottom by the p+ layer used for the ohmic contact or by the highly doped substrate in case of epitaxial substrates. Indeed, when a pair is generated in those regions, there is a high probability for the electron and hole to recombine due to Auger or Shockley–Read–Hall (SRH) processes [4]. For short wavelengths, the problem is focused in the top layer. As an example, a 420 nm light is almost totally absorbed in the first 500 nm of silicon, which, for non-optimized fabrication processes, is usually well inside the undepleted layer.

3.2. Triggering probability

As mentioned in the previous paragraph, there is a finite probability for a carrier to initiate an avalanche when passing through a high-field region. In case of a photogeneration event, 2 carriers are created travelling in opposite directions. Both contribute to the triggering probability that can be evaluated from the following:

$$P_t = P_e + P_h - P_e P_h \quad (1)$$

where P_e and P_h are the electron and hole breakdown initiation probabilities [5]. These terms can be calculated as a function of the generation position by solving 2 differential equations involving the carrier ionization rates. In order to understand the physical impact of Eq. (1), an n+/p junction having a constant high-field region, extending beyond the n+ layer, is considered (see dashed line in Fig. 2). When a pair is generated in the left side of the high-field region, the electron is directly collected at the n+ terminal; thus, it does not contribute to the triggering

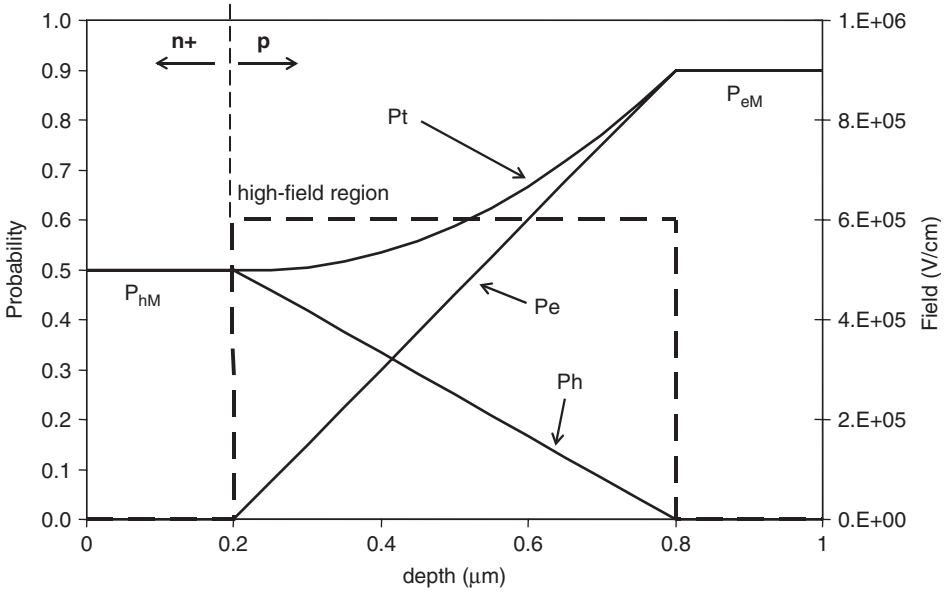

Fig. 2. Avalanche triggering probability as a function of the photogeneration position.

probability. The hole is forced to pass the whole high-field region and so its triggering probability is maximized and $P_t = P_{hM}$. In case of photogeneration on the right side, the situation is symmetrical and only electrons contribute to the triggering probability, thus, $P_t = P_{eM}$. In the central region, both carriers contribute to a different extent as a function of the interaction position and the P_t value is between P_{eM} and P_{hM}.

As mentioned above, P_e and P_h depend on the impact ionization rates of electrons (α_n) and holes (α_p), respectively. These parameters are not well determined yet, and large discrepancies exist among the values extracted from the various models (as an example see [6,7]). Anyway, despite the differences in absolute values, some features are well established: (i) both coefficients increase with the electric field, (ii) the electron has an ionization rate higher than the hole (e.g., at 5×10^5 V/cm, α_n is about twice α_p [6]), and (iii) their difference decreases with increasing fields. This behaviour is reflected in the probabilities P_e and P_h. Thus, to maximize the triggering probability: (i) the photogeneration should happen in the p side of the junction in order for the electrons to pass the whole high-field zone, and (ii) the bias voltage (V_{BIAS}) should be as high as possible. It must be noted that V_{BIAS} cannot, in any case, exceed the value for which the current flowing through the diode is higher than I_L.

3.3. Geometrical efficiency

The ratio between the active area and the total device area is a critical issue in SiPMs. As mentioned in the previous paragraph, each GM-APD cell is surrounded by a dead region determined by the guard ring and the structure preventing optical cross-talk. Considering that the area of a cell can be very small (in the order of $30 \times 30\ \mu m^2$) even few microns of dead region around the cell have a very detrimental effect on the geometrical efficiency.

4. Short-wavelength light detection efficiency in a conventional structure

A conventional structure (to our knowledge all reported SiPMs have this configuration) is built on a 3–5 μm thick lowly doped p-type epitaxial layer (π) which was grown on a highly doped p-type substrate (p+). The abrupt junction is obtained by creating an n+ zone on the superficial region of the epitaxial layer. Usually, a second p-type region is created underneath the n+ to fix the breakdown voltage to the desired value. Thus, the final structure, from top to bottom, is n+/p/π/p+ (see Fig. 3). In order to optimize the detection efficiency in the short wavelength region, the following 3 points have to be satisfied:

(1) the n+ layer has to be as shallow as possible (for optimum QE); this can be accomplished by using arsenic as n-type dopant. With standard equipment for detector fabrication, layers with a junction depth of 100 nm can be obtained. In reality, the number of collected carriers can be increased by minimizing the recombination probability both in the n+ region and in the silicon oxide/silicon interface. Concerning the first term, a "low" doping concentration (e.g. peak value of 5×10^{18}–$1 \times 10^{19}\ cm^{-3}$) layer can be implemented in order to suppress Auger processes and minimize enhanced concentration-dependent SRH recombination [8]. The second point is strictly related to the technology depending on the quality of the silicon interface;

(2) the high-field region should be as thin as possible in order to photogenerate as much as possible beyond it, maximizing Pt. This can be accomplished by increasing the doping concentration of the p-type implant, which leads also to a lowering of the breakdown voltage. An upper limit on the p concentration is posed by the increase of the tunnelling probability, which becomes significant for dopant levels in the order of few $10^{17}\ cm^{-3}$ (breakdown voltages around 10 V) [4].

Using the above-described criteria, a fabrication process has been defined and simulated with the process simulator ATHENA [9]. Successively, the breakdown voltage and the electric field have been determined by means of device simulations (device simulator ATLAS [9]). The former value has been extracted from the IV curve and found to be 25 V. It should be noted that, once an impact ionization model has been selected, the generation rates calculated from the electric field are normally over-estimated (especially for very peaked fields) [4], so the real breakdown voltage is expected to be higher.

The electric field in the active area as a function of the depth is shown in Fig. 4. On the same graph, the light absorption curve for 3 different wavelengths is shown as well. It is clearly visible that, even for such an

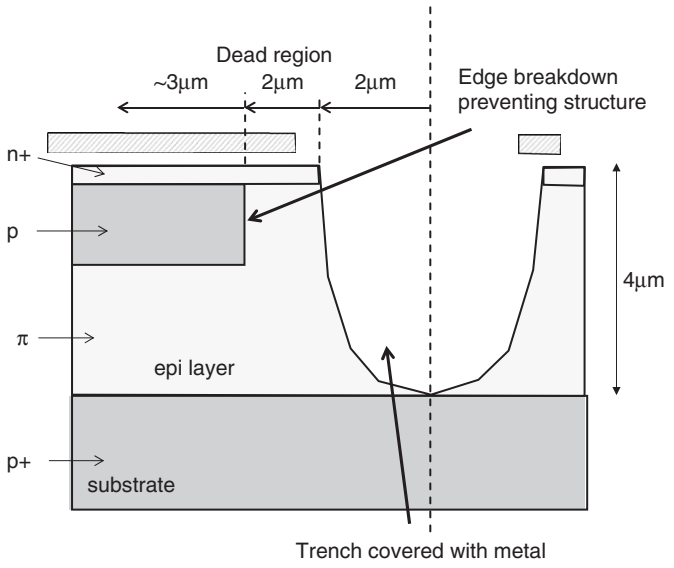

Fig. 3. Sketch of the border region of a conventional structure.

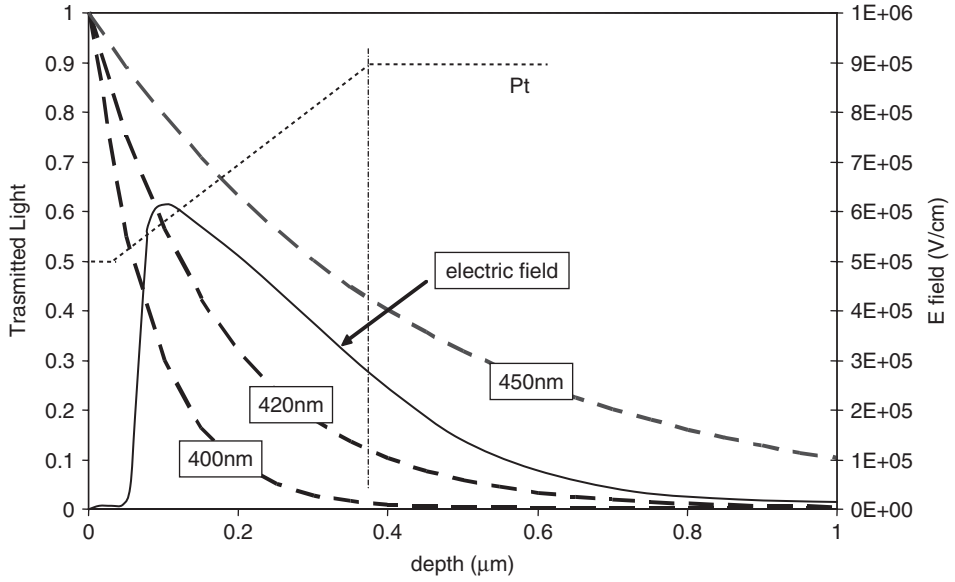

Fig. 4. Simulated electric field distribution for an n+/p diode with optimized doping profiles for short-wavelength detection. On the same graph the absorption curves for 3 wavelengths and a representation of the triggering probability are shown.

optimized structure, at 420 nm, practically 90% of the photons are absorbed before the maximum-triggering probability region. On the other hand, with a suitable ARC, the QE can reach a value as high as 0.95 at 420 nm.

(3) the dead region should be as narrow as possible (this statement is valid for every wavelength). An estimation of this width is presented in Fig. 3. In this case, the electric field at the cell edge is shaped by a "virtual" guard ring structure obtained by defining a p region smaller than the n+ one. Its contribution to the dead region can be estimated to be about 4 μm, of which 2 μm are only partially non-active and correspond to the transition from the low- to the high-field region. Concerning the structure for optical cross-talk prevention, a reasonable width of the trench, obtained with a deep-reactive ion-etching (DRIE) machine, can be estimated to be in the order of 4 μm. Considering a GM-APD cell having a size of $30 \times 30\,\mu m^2$, the area efficiency is only 36%. This value does not take into account the presence of both, the contact regions (from the silicon to R_Q and from R_Q to the metal line connecting all the cells together) and the quenching resistor. The contribution of these regions to the area efficiency cannot be estimated since it depends on the layout configuration. To minimize their impact, these structures should overlap the dead border region as much as possible.

Triggering probability can be improved by maintaining the same doping profile configuration but reversing the types, i.e. having a p+/n/n−/n+ structure, and making the junction deeper (>0.4 μm). In this case the triggering probability curve sketched in Fig. 4 is flipped and most of the carriers are absorbed in the high-Pt region. The main drawback is that the QE is lowered by the larger extent of the undepleted p+ region.

5. New approach for optimized blue detection efficiency

The structure proposed in this paper is referred to as Buried-Junction SiPM (BJ-SiPM), the main characteristic being that the junction is located deep in the bulk. In principle, this structure could be easily fabricated growing an epitaxial layer opposite in type with respect to the substrate, i.e. a p-type layer on an n+ substrate. If, at the bias operational condition, the epitaxial layer is fully depleted, an electron photogenerated near the surface drifts through the whole layer traversing always at the end of its path the high-field region. In this way, the triggering probability is always maximized.

Using a sufficiently thin epitaxial layer the high-field region can be formed by a deep Boron implant (as sketched in Fig. 5). The main concern related to this approach is the uniformity of the breakdown voltage. Two reasons can affect this aspect: (i) the poor uniformity of the substrate dopant concentration, and as a consequence of the outdiffused dopant tail in the epitaxial layer causes fluctuations in the compensation of the deep Boron implant; (ii) the non-uniformities of the epitaxial layer thickness cause variations of the depth of the implanted p-type layer.

To overcome this problem, every doped layer of the proposed structure can be created by ion implantation. Thus, the starting silicon is an n-type epitaxial layer on an n+-type silicon substrate. The thickness of the epi layer can be in the order of 3 μm. A high-energy (e.g. 1 MeV), medium-dose Phosphorous implantation is used to form the n+ side of the junction. Then, a 300 keV low-dose Boron implant forms the p-side of the junction and fixes

the breakdown voltage to the desired value. Finally, a shallow Boron layer is created to prevent the depletion region from reaching the silicon surface. The fabrication process, including implantations and annealing cycles, has been simulated and the resulting doping profiles are shown in Fig. 6.

The 3 implanted layers are clearly visible, and the junction is located at 0.8 μm from the surface. A detailed analysis of the doping profiles reveals that the shallow Boron implant has a lower concentration at the Si/SiO_2 interface with respect to the peak value, despite the very low implantation energy used. This is due to the fact that the Boron ions tend to escape from the silicon and segregate in the oxide during the annealing cycles. This creates a small retarding field in the first 50 nm that could slightly reduce the QE. It is worth pointing out that the implantation parameters of the shallow layer can be reasonably varied and tuned without affecting the main characteristics of the device.

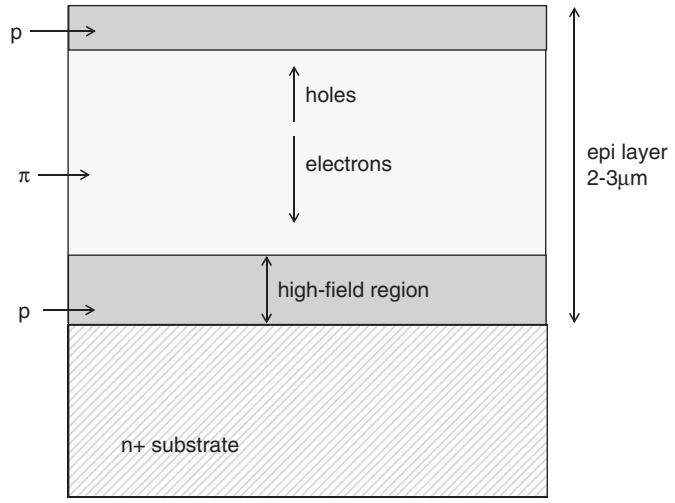

Fig. 5. Sketch of a Buried-Junction SiPM implemented with an epi layer opposite in type with respect to the substrate.

The breakdown voltage corresponding to these doping profiles is around 20 V. The electric field calculated at this voltage is shown in Fig. 6.

The depletion layer is less than 1 μm thick, so, for a cell area of $30 \times 30\,\mu m^2$ the capacitance is about 100 fF. As mentioned in the introduction, the maximum avalanche current must be below the latching value for any reasonable value of the excess bias voltage (e.g. up to 5 V), so the quenching resistor should be of at least 200 kΩ. In such conditions the recovery time constant is about 20 ns.

Plotting the absorption curves along with the electric field profile (Fig. 7), it is evident that, at 420 nm, almost every absorbed photon creates an e–h pair in a region preceding the high-field zone, maximizing the triggering probability.

The position of the high-field zone can be slightly adjusted to optimize the triggering probability (as a function of the wavelength to be detected) by varying the implant energies of the buried implants. Of course, this structure is suitable only for the detection of short-wavelength light: photons having a wavelength longer than 450 nm would require very high implantation energies, which are difficult to implement.

Besides maximizing the triggering probability, the BJ-SiPM structure allows an optimization of the area efficiency. Indeed, the fact that the electric field grows towards the surface rather than into the bulk can be exploited to create a guard ring structure with the inclined wall of the trench needed for the optical isolation of the cells. This technique is widely used in APDs for the same purpose (bevelled-edge APDs) [4]. In order to create clean and smooth walls, the trench can be etched with tetramethyl ammonium hydroxide (TMAH) [10], which is an anisotropic etchant (i.e. the etching rate depends on the

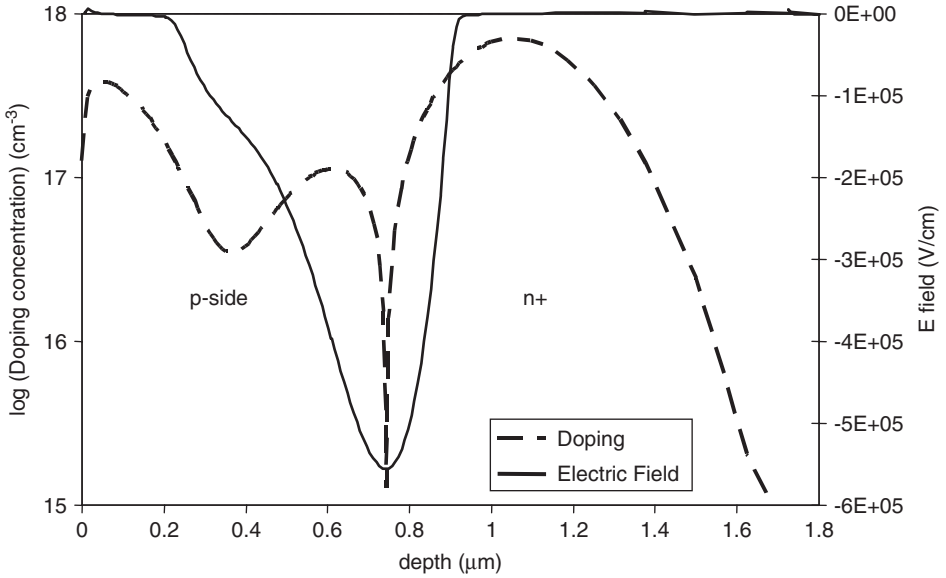

Fig. 6. Simulated dopant profile and corresponding electric field at the breakdown voltage of a BJ-SiPM.

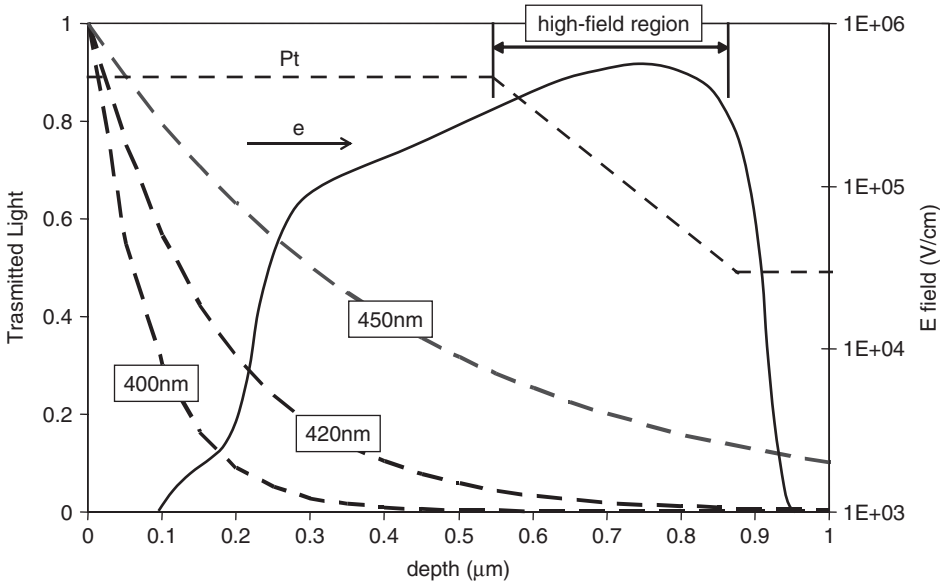

Fig. 7. Simulated electric field (represented in log scale) along with the absorption curves at 3 different wavelengths of a BJ-SiPM.

crystal orientation) that has been proved to be CMOS compatible. In this way, using a $\langle 100 \rangle$-oriented substrate, one can obtain reproducible walls with an angle of about 57° with respect to a line perpendicular to the surface. The trench depth and width are related by the tangent of this angle. As an example, a trench 4 μm wide gives a hole about 3 μm deep, which, in case of the BJ-SiPM, is enough to avoid optical-cross talk.

Both process and device simulations of a device with this border region have been performed. The simulated structure, partially shown in Fig. 8, has a total length of about 7 μm extending from the centre of the trench (on the right) to about 5 μm inside the active area (on the left). Such geometry assures that the boundary regions do not affect the solution in the zone of interest, i.e. the diode border. The device has a thick SiO_2 layer on top of a double layer composed by silicon nitride and silicon dioxide (barely visible in the picture). An aluminium layer covers the oxide in the trench. Note that, in the real device, the thicker oxide will be removed in the active area region in order to have an optimum ARC with the 2 underlying thin layers. As for the previous simulations, the junction is located at a depth of about 0.8 μm (y-axis).

Figs. 8 and 9 show the equipotential lines and the electric field, respectively. In particular, the first picture evidences the spreading of the potential lines in proximity of the inclined wall that leads to a reduction of the electric field. Notably, without using any mask for the implants, the field is much lower at the border than in the active area: from about 5×10^5 V/cm it goes down to roughly 3×10^5 V/cm. Furthermore, due to the reduced depletion layer thickness the transition region from low to high field is less than 1 μm wide. In both pictures the drift path in seven different positions is shown as well. The ionization integral along each line at the breakdown voltage is calculated and reported in the same figures. Even if the absolute value of these numbers can change according to the physical model, their behaviour is maintained and confirms that the active area extends almost up to the trench edge. Therefore, the area efficiency is consistently higher compared to a conventional structure, the dead region being 2 μm (half trench width) plus roughly 1 μm (transition region).

It is worthwhile to note that the silicon surface of the inclined wall is depleted starting from the wafer surface down to the peak of the buried n+ implant (about 0.8 μm). Even if the surface quality of the wall is optimized by using TMAH etching, this could lead to an increased dark count rate due to the surface generated carriers that are able to reach the high-field region. The electric field configuration at the border region prevents the carriers from travelling towards the active area, but, instead, they are forced to drift close to the interface. The prediction of the noise behaviour is, in any case, very difficult because the SRH generation is strongly related to the fabrication process.

6. Conclusion

In this paper, a new Silicon Photomultiplier structure for short-wavelength (up to 450 nm) light detection is presented. This structure (named BJ-SiPM) has a junction located deep in the silicon bulk and a depletion region that grows toward the surface. Such a configuration has 2 main advantages in terms of detection efficiency with respect to a conventional abrupt junction: it provides a higher avalanche triggering probability because the process is always initiated by the electrons and it gives the possibility to consistently reduce the dead area at the border region of the micro GM-APDs composing the SiPM. A third advantage derives from the fact that the diodes are completely formed and delimited by implanted layers. This will allow the fabrication of junctions featuring an extremely uniform and reproducible electric field in the

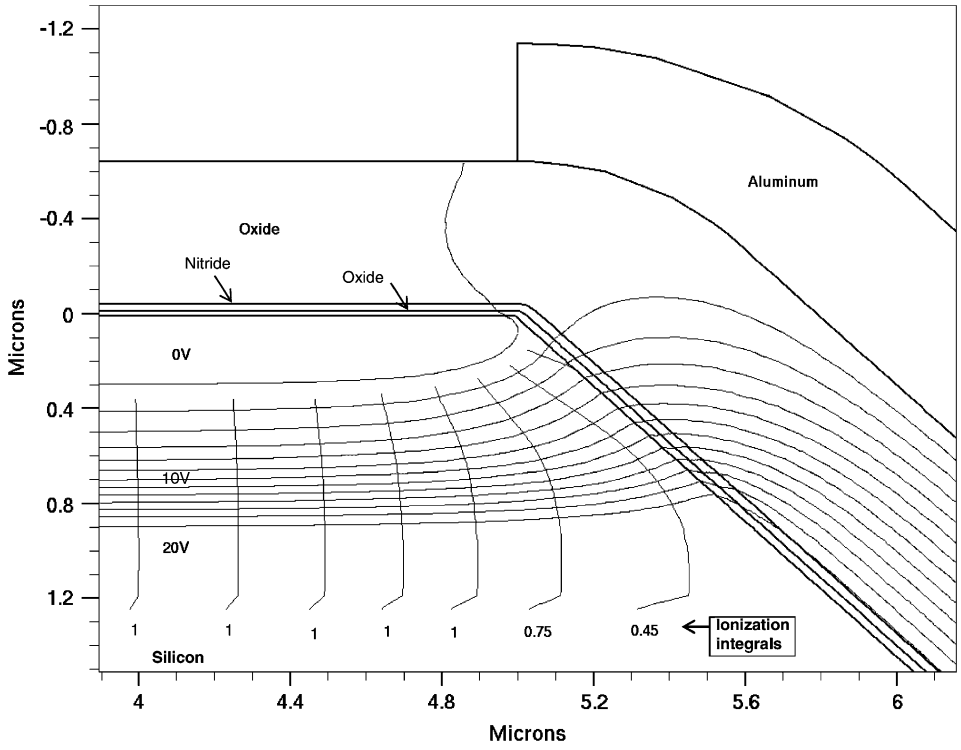

Fig. 8. Equi-potential lines and drift trajectories at the border region of a GM-APD cell composing the BJ-SiPM.

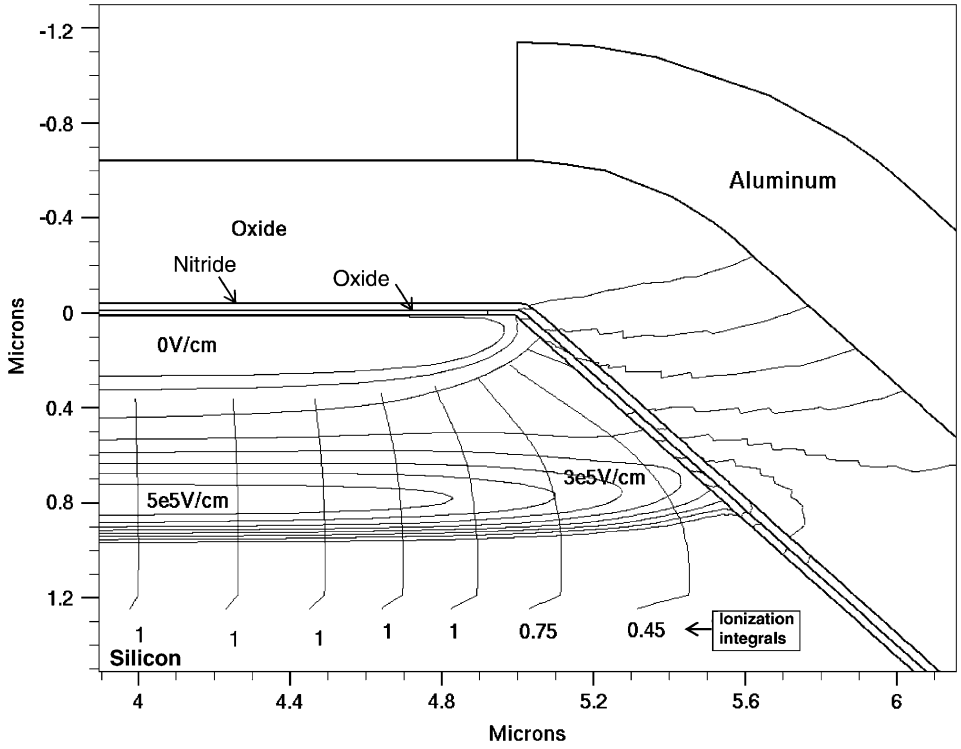

Fig. 9. Map of the electric field strength and drift trajectories at the border region of a GM-APD cell composing the BJ-SiPM.

active region regardless of substrate characteristics such as the doping level and the epitaxial layer thickness.

A disadvantage of a completely implanted structure could be a higher SRH generation due to residual defects introduced by the high-energy implantation step. This drawback should not be so important because the implantation doses are relatively low ($<10^{14}\,\mathrm{cm}^{-3}$).

A mask set implementing these devices has already been designed, and the production of the first prototypes is starting at ITC-irst (Trento).

Acknowledgments

The author would like to thank Dr. M. Boscardin (ITC-irst, Trento) and Prof. Gian-Franco Dalla Betta (University of Trento) for the useful discussions.

References

[1] P. Buzhan, et al., ICFA Instrum. Bull. 23 (2001) 28.
[2] R.J. McIntrye, J. Appl. Phys. 32 (6) (1961) 983.
[3] R.H. Haitz, J. Appl. Phys. 35 (5) (1964) 1370.
[4] M.S. Tyagi, Introduction to Semiconductor Materials and Devices, Wiley, New York, 1991.
[5] W.G. Oldham, et al., IEEE Trans. Electron. Dev. ED-19 (9) (1972) 1056.
[6] W.N. Grant, Solid-state Electron. 16 (1973) 1189.
[7] R. van Overstraeten, H. de Man, Solid-state Electron. 13 (1970) 583.
[8] D.J. Roulston, Bipolar Semiconductor Devices, McGraw-Hill, New York, 1990.
[9] ATHENA/ATLAS user's manual, SILVACO International, Santa Clara, USA.
[10] S. Ronchin, et al., Nucl. Instr. and Meth. A 530 (2004) 131.