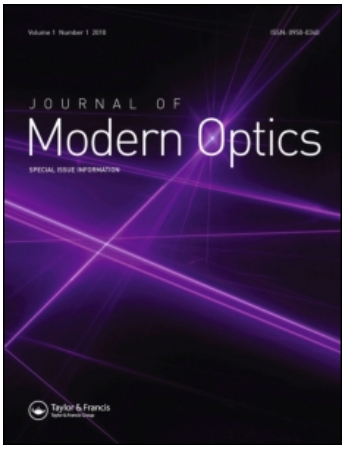

Journal of Modern Optics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713191304

A physically based model for evaluating the photon detection efficiency and the temporal response of SPAD detectors

A. Gulinatti[a]; I. Rech[a]; M. Assanelli[a]; M. Ghioni[a]; S. Cova[a]
[a] Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy

First published on: 08 December 2010

To cite this Article Gulinatti, A. , Rech, I. , Assanelli, M. , Ghioni, M. and Cova, S.(2011) 'A physically based model for evaluating the photon detection efficiency and the temporal response of SPAD detectors', Journal of Modern Optics, 58: 3, 210 — 224, First published on: 08 December 2010 (iFirst)
To link to this Article: DOI: 10.1080/09500340.2010.536590
URL: http://dx.doi.org/10.1080/09500340.2010.536590

Journal of Modern Optics
Vol. 58, Nos. 3–4, 10–20 February 2011, 210–224

Taylor & Francis
Taylor & Francis Group

A physically based model for evaluating the photon detection efficiency and the temporal response of SPAD detectors

A. Gulinatti*, I. Rech, M. Assanelli, M. Ghioni and S. Cova

Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milano, Italy

(*Received 20 June 2010; final version received 25 October 2010*)

After a brief review of the physics of photon detection in single photon avalanche diode (SPAD) devices, in this paper we will outline the principle of operation of a model we developed with the aim of calculating both photon detection efficiency (PDE) and temporal response (TR) of these detectors. Then we will apply the model to the devices currently available in order to critically analyze some experimental results. We will show in particular how the use of the model allows us to gain a better understanding of the influence of each device parameter in determining both the PDE and the TR. Finally we will discuss some modifications that can be applied to the device structure in order to overcome such limitations. Their effectiveness in improving both the PDE and the TR will be investigated by means of the aforementioned model. The aim is to provide the reader with an insight of which performances can be expected in the next few years if a strong development of the SPAD structure is pursued.

Keywords: single photon avalanche diode (SPAD); photon counting; photon timing; photon detection efficiency; TCSPC; quantum efficiency

1. Introduction

In the last few years, silicon single photon avalanche diodes (SPADs) have gained wide acceptance as an alternative to photomultiplier tubes (PMTs) in many photon counting and photon timing applications thanks to their remarkable performance [1–6]. In particular, low dark count rate devices are available with diameters ranging from 50 to 200 μm, and a very good timing resolution of about 35 ps FWHM can be achieved even with the larger detectors, and photon detection efficiency (PDE) is higher than the one usually achievable with PMTs [7,8].

However, current available devices still suffer some limitations. The main issue is related to the PDE at long wavelengths ($800\,\text{nm} < \lambda < 1000\,\text{nm}$); many applications would greatly benefit if it could be further improved. On the other hand, other applications, such as label-free analysis of proteins, would equally benefit from an improvement in the PDE at short wavelengths ($400\,\text{nm} < \lambda < 500\,\text{nm}$).

Another limitation is related to the presence of a slow component in the temporal response of the detector. The latter is commonly known as diffusion tail since it is due to the diffusion of the carriers photo-generated into the neutral regions of the device. While some applications are almost unaffected by this aspect, others are strongly influenced even by a small amplitude tail. For example in quantum key distribution (QKD), the diffusion tail reduces the attainable quantum bit error rate (QBER) while in fluorescence lifetime imaging (FLIM) it can make it difficult to assess the time constants of the multi-exponential fluorescence decays. Both the issues can be addressed by suitable modifications of the detector structure; however, the complexity of the problem requires a deep understanding of the phenomena involved and the availability of tools that can help device designers in the optimization task. To this aim we developed a physically based model aimed at calculating the PDE and the temporal response of a SPAD with a given structure. Validation of the model has been carried out by comparing simulations and experimental results of a few generations of detectors previously fabricated in our laboratory.

After a brief review of the physics of photon detection in SPAD devices, we will outline the principle of operation of the model developed along with the main assumptions made. Then we will apply the model to the devices currently available in order critically to analyse some experimental results. Based on the results

*Corresponding author. Email: angelo.gulinatti@polimi.it

ISSN 0950–0340 print/ISSN 1362–3044 online

DOI: 10.1080/09500340.2010.536590
http://www.informaworld.com

obtained we will discuss some modifications that can be applied to the device structure in order to overcome such limitations. Their effectiveness in improving both the PDE and the temporal response will be investigated by means of the aforementioned model.

2. Principle of operation and real structures

In principle, a SPAD is simply a semiconductor p–n junction reverse biased above the breakdown voltage, in a metastable state in which no current flows through the device. When a photon is absorbed into the space charge region, it generates an electron-hole pair. The high electric field present in that region accelerates the carriers that can in turn gain enough energy to create other pairs through impact ionization. Since the applied voltage is above the breakdown value, the multiplication process is self-sustained and a macroscopic current is generated allowing for the detection of the photon.

In practice, real SPADs require a more sophisticated structure than a simple pn junction [9]. As an example, Figure 1 sketches the typical structure of a SPAD designed in our laboratory (thin SPAD). The device is fabricated starting from an n-type wafer, on which has been grown a p-type double epitaxial layer. The cathode region is obtained by diffusing an n-type layer (*shallow n*) into the p^- epitaxy. In order to prevent edge breakdown, so-called *virtual guard rings* are implemented: instead of increasing edge breakdown value through an n-type guard ring, the breakdown voltage is reduced only in central region of the device by performing a local p^+ implantation (*enrichment*). A heavily doped p-type region (*sinker*) is built under the anode metallization in order to make the contact ohmic. The sinker, in combination with the p^+ epitaxy (*buried layer*), helps also to reduce the resistance seen by the current flowing from the cathode to the anode. Another heavily doped region is built all around the device (*isolation*). Since that region is n-type and since it extends vertically to the substrate, by reverse biasing the anode-substrate junction it is possible electrically to insulate the SPAD from any other device on the chip. Actually, the n-type substrate has also a very important role in improving the device temporal behavior as will be clarified at the end of the next section. A more detailed description of the structure as long as of its advantages can be found in [10].

While the structure depicted in Figure 1 is quite common, other configurations are certainly possible. For example, commercial SPAD made available from Perkin Elmer Optoelectronics [11] are based on a back-illuminated reach-through structure [12] (thick SPAD). Anyhow, all the work described in the remaining part of this article is specifically targeted to the structure of Figure 1, despite the fact that a large part of it can be easily extended to other structures.

3. Photon detection physics

The physical processes that lead to the detection of a photon in a SPAD are well understood and fully reported in the literature. They will be briefly reviewed in this section with a special focus on the structure described in the previous section. Given a generic single-photon detector, it is possible to define PDE as the probability that a photon

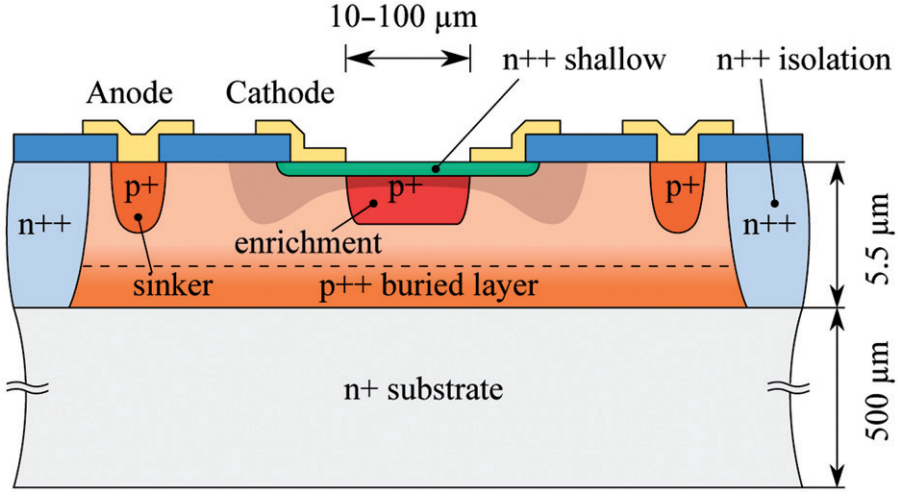

Figure 1. Vertical cross-section of a typical thin double epitaxial SPAD. The main regions of the detector are clearly visible in the figure. (The color version of this figure is included in the online version of the journal.)

impinging onto the device active area is effectively detected. In order to be detected, a photon must be absorbed into the active region. Furthermore, either the photogenerated electron or the hole must reach the high field region and must trigger the avalanche. Therefore, the PDE is a combination of the efficiencies of photons absorption and avalanche initiation.

To gain a better insight into the detection process, it is possible to consider the simplified one-dimensional case shown in Figure 2. In particular, Figure 2(*a*) represents the central region of the structure depicted in Figure 1, while Figure 2(*b*) is a simplified sketch of the corresponding electric field. For the sake of photon detection efficiency calculation, four different regions can be identified: the substrate, the multiplying space charge region and the two neutral regions at the two sides of the space charge layer itself.

Only a fraction T of the photons impinging onto the detector area enters into the device, while the fraction $R = 1-T$ is back-reflected due to the discontinuity into the refraction index. Photons that succeeded in entering into the device are absorbed in one of the four layers of Figure 2, with a probability that depends on the absorption coefficient.

If the photon is absorbed into the space charge region, a pair of photo-generated carriers are promptly separated and accelerated by the electric field. However, owing to the randomness of impact ionization process, this does not guarantee that a self-sustained avalanche is triggered even if the applied voltage is above the breakdown level. This phenomenon can be described through a triggering efficiency $\eta_{trig}(x)$ defined as the probability that an electron-hole pair generated in x, triggers a self-sustained avalanche. η_{trig} depends not only on the position x, but also on the electric field profile and consequently on the applied voltage.

In case a photon is absorbed into the lower neutral region (p-type), the photo-generated hole thermalizes with the other majority carriers, while the electron diffuses randomly through the region owing to the absence of a strong electric field. During its random walk it can recombine with a hole or it can reach the substrate. Since the electron is lost, in both the cases the avalanche is not triggered. Conversely, if the electron reaches the multiplying space charge region, it is accelerated toward the high field zone and it can trigger an avalanche. Therefore, the probability that the photon herein absorbed is detected, is the product of the probability that the electron reaches the space charge region, the so-called electron collection efficiency $\eta_{n\text{-}collect}(x)$, times the triggering efficiency $\eta_{trig}(x_p)$. A similar situation occurs if a photon is absorbed into the upper neutral region (n-type). The only difference is that now the minority carrier is a hole instead of an electron and that it will be lost if it recombines with an electron into the neutral region volume or at the interface with the silicon dioxide.

Finally, if the photon is absorbed into the substrate (n-type) it will be not detected since there is no way for the minority hole to initiate an avalanche in the active region.

In such application as the time correlated single photon counting (TCSPC), it is mandatory not only to

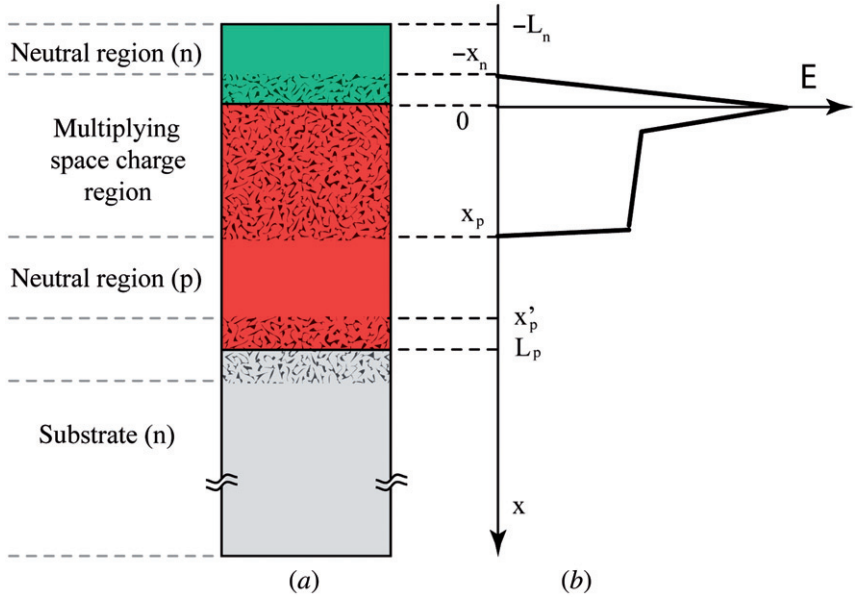

Figure 2. Cross-section of the active region of a thin SPAD (*a*) with a qualitative electric field profile for the multiplying space charge region (*b*). (The color version of this figure is included in the online version of the journal.)

detect photons but also to detect their arrival time with high accuracy. Since the triggering of the avalanche is synchronous with the photon absorption, SPADs are particularly suited for this kind of measurements.

Figure 3 reports the statistical distribution of the time of avalanche detection as measured in a conventional setup [13]; although the photon hits the detector always at the same instant, nevertheless the detection time is affected by a certain jitter that limits the temporal resolution of the detector when employed in TCSPC measurement. The shape of that curve is once again determined by the phenomena discussed in the first part of this paragraph.

Actually if a photon is absorbed into the depletion region, the avalanche process starts almost immediately. Photons absorbed in space charge region are therefore responsible of the part of the curve referred as *peak* in Figure 3. Note that even in this case a certain amount of jitter is present, due mainly to the following phenomena [14,15]:

- the time needed for the carrier to reach the high field region depends on the absorption position (transit time dispersion);
- every time that the avalanche is triggered, the current grows in a different way due to the randomness of the impact ionization process.

Provided that a suitable current pick-up circuit is used [16,17], this contribution can be as low as a few tens of picoseconds with thin SPADs represented in Figure 1, independently of the active area diameter [8].

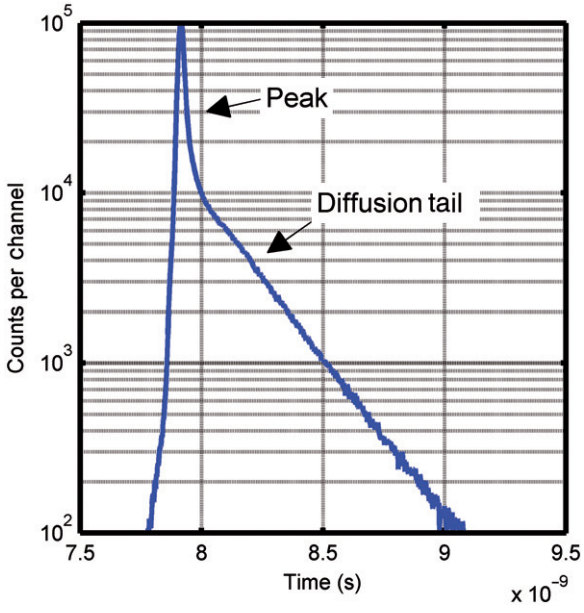

Figure 3. Typical temporal response of a double epitaxial SPAD. The main peak and the diffusion tail are clearly visible. (The color version of this figure is included in the online version of the journal.)

A photon absorbed into one of the two neutral regions gives rise to a completely different behavior. The minority carrier can randomly move through the neutral region for a relatively long time before it finally triggers the avalanche. These photons are therefore responsible of the part of the curve referred as *diffusion tail* in Figure 3 and the time of avalanche detection is spread on a time interval of few nanoseconds.

Let us consider for example the case of a photon absorbed into the lower neutral region. Defined $\mathrm{F}(x;t)\,\mathrm{d}t$ as the probability that an electron generated in x reaches the upper space charge region at the time interval between t and $t+\mathrm{d}t$, then the tail component given by the photon absorbed into the lower neutral region is given by:

$$p_{\text{lower\_neutral\_region}}(t) = T\left[\int_{x_p}^{x'_p} e^{-\alpha x}\cdot\alpha\cdot F(x;t)\cdot\eta_{\text{trig}}(x_p)\cdot\mathrm{d}x\right]. \tag{1}$$

Since the integral in time of $F(x,t)$ is exactly $\eta_{\text{n-collect}}(x)$, by integrating (1) in time, one obtains the part of the PDE associated to the lower neutral region. In this regards the PDE calculation can be considered as a special case of the calculation of the detector temporal response.

4. Model description and validation

In order to design devices with better performances, we introduced a model capable of calculating the PDE and the TR of a SPAD. In particular the purpose of the model is two fold. On one hand, it must be able to forecast the detector performances at design time, thus avoiding the need of fabrication test structure for optimization. On the other hand, it should make it possible to investigate the behavior of the detector with the aim of understanding what are the limitations to its performance and how to operate to overcome them. These requirements led us to develop a physically based model, i.e. a model capable of describing the physical phenomena that are involved in photon detection process in a way that is physically correct. In particular three tasks are sequentially performed: as a first step, the electron-hole generation profile along the device is calculated according to the silicon absorption coefficient at the considered wavelength; successively, the temporal evolution of the carriers' distribution along the device is calculated by solving drift-diffusion equations; finally, the avalanche triggering probability is calculated as a function of the photon absorption point.

In this paragraph we will not go into the details of the model; we will simply outline its principle of operation and we will discuss some of the assumptions made. Then we will report some comparison between simulations and experimental results in order to show how the developed model is really capable of reproducing PDE and TR of actual devices. A detailed description of the model along with a complete validation against experimental results can be found in [18].

The starting point for the calculation of both the PDE and the TR is the evaluation of the photon absorption distribution along the device. In particular, the fraction of the photons that enters the device at the entrance interface is calculated by using standard theory of transmission/reflection at multi-dielectric interfaces [19]; similarly the absorption profile inside the device is evaluated, as function of the wavelength, by using values of the silicon absorption coefficients reported in literature [20]. This allows us to obtain the concentration of photo-generated carriers as a function of the position, both in the depleted and in neutral regions.

In the neutral regions, the attained position-dependent concentration is used as initial condition for the solution of drift diffusion equations [21]; in effect they provide the temporal evolution of minority carriers' concentration, from which it is straightforward to obtain the carrier flux F at the edges between neutral and depleted regions.

Our modeling of transport in neutral regions deserves some considerations. First of all, we neglected volume recombination in both upper and lower neutral regions. This because minority carriers lifetime is orders of magnitude larger than the time that carriers spend in those regions before exiting owing to diffusion [18]. For the same reasons, in the upper neutral region, we neglected the recombination at the oxide interface as well. Based on these assumptions, a minority carrier generated inside the upper neutral region, sooner or later will reach the space charge region; consequently the carriers collection efficiency for this region is always unitary. However, the solution of the drift diffusion equations is still needed, since it allows us to calculate the distribution of the delays with which the carriers reach the depleted region. The situation is different for the carriers generated in the lower neutral region; although they do not recombine, nevertheless they can get lost at the interface with the substrate, leading to a less than 100% collection efficiency.

Concerning the last step for the evaluation of PDE and TR, triggering efficiency can be calculate by using well known equations introduced by Oldham et al. [22]. By solving these equations it is possible to obtain $P_{\mathrm{be}}(x)$ and $P_{\mathrm{bh}}(x)$, respectively, the probability that an electron or a hole generated in x, triggers a self-sustained avalanche. The overall avalanche probability, also know as triggering efficiency, can therefore be calculated as:

$$\eta_{\mathrm{Trig}}(x) = P_{\mathrm{be}}(x) + P_{\mathrm{bh}}(x) - P_{\mathrm{be}}(x) \cdot P_{bh}(x) \qquad (2)$$

where the last term is needed in order to prevents to accounts twice for the avalanches triggered both by an electron and by a hole. A thorough knowledge of the impact ionization coefficients for electrons and holes, namely α_{e} and α_{h}, is of the utmost importance for the solution of Oldham's equations. Since their value strongly depends on the electric field, Poisson equation is preliminary solved by using a commercial device simulator.

Actually, the presence in our devices of thin multiplication regions requires to explicitly account for the effects of dead space, i.e. the distance that a carrier must travel in order to acquire an energy high enough to ionize. As a consequence, the ionization coefficient becomes a non-local property [23]; in particular we used the model introduced by Okuto and Crowell [24]. However, in this case, Oldham's equations are not applicable anymore and must be replaced by more general relations introduced by McIntyre [25].

The modeling of the statistical phenomena involved in avalanche build-up and propagation processes deserves a few considerations as well. The complexity of these phenomena as well as the fact that they are not fully understood yet [26,27], has prevented their physically correct modeling. However, their effect has been included in the model based on the following approach. As suggested by measurements, we supposed that the combination of build-up and avalanche propagation affects avalanche detection with a random delay, characterized by a Gaussian distribution having a width of about 30 ps FWHM. Therefore, delay distributions previously calculated have been convoluted with such a distribution in order to simulate the effect of the statistics of avalanche growth. This approach makes it possible to reproduce accurately the device TR, but of course does not allow to study the effects of the structure on these phenomena. As opposite, transit time dispersion has been correctly modelled by calculating for each carrier the delay it takes to reach the multiplication region as a function of the starting position.

In order to validate our model, we applied it on a few families of devices already available in our laboratory and compared simulations with experimental results. For example Figure 4 reports the PDE as function of the wavelength for a device belonging to production lot S62. Both measurement and simulation

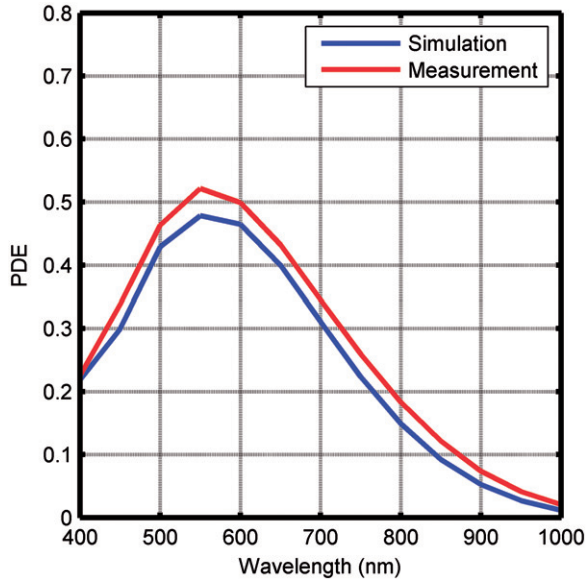

Figure 4. Comparison between calculated and measured photon detection efficiency. Both simulation and measurement have been carried out at an overvoltage of 5 V. (The color version of this figure is included in the online version of the journal.)

have been carried out at an overvoltage of 5 V. The agreement between simulations and experimental results along the entire spectrum is remarkable. Similar results have been obtained also at other overvoltages (ranging from 2 to 8 V) and with devices belonging to other families. As another example, Figure 5 reports a comparison between calculated and measured temporal response, evaluated at a wavelength of 520 nm, for a device belonging to lot S44. Once again a good agreement between the two curves has been obtained.

5. Current available devices: a critical analysis of the performances

The aim of this section is critically to analyze currently available devices in order to identify the phenomena that limit their PDE and TR performance. In this respect, the model described in the previous section constitutes a valuable tool since it makes it possible to separate the effect of every single phenomena involved in the photon detection process and to investigate how each of them affects the overall result. Conversely, it is more difficult to achieve the same kind of information without using a physical model since PDE and TR are determined by the superposition of many effects, and individual contributions such avalanche probability or carriers collection efficiency cannot be easily measured.

As an example, in this section we will refer to devices belonging to production lot S62 [28]. They are typical double epitaxial devices with the structure represented in Figure 1; they are characterized by a breakdown voltage V_{BD} of about 35 V, and by a dark count rate of about 1000 count/s for a device with 50 μm diameter active area, operated at room temperature. The thicknesses of the upper neutral region, of the space charge region and of the lower neutral region are approximately $x_{SUP} = 0.8\,\mu m$, $x_{ZCS} = 1.8\,\mu m$, $x_{INF} = 2.4\,\mu m$. If not otherwise stated, the devices are operated at an overvoltage $V_{OV} = 5\,V$. Both simulations and measurements have been carried out at room temperature; however, modifications to PDE and TR are negligible when the device is moderately cooled (e.g. down to $-20^{\circ}C$) [29].

5.1. Photon detection efficiency

Figure 4 reports the PDE measured as a function of the wavelength for a device belonging to lot S62. The PDE rapidly decreases as long as the wavelength increases above 700–800 nm. This behavior is certainly expected and is due to the strong reduction in silicon absorption coefficient at long wavelengths. However, the PDE reduces considerably also at short wavelengths where the silicon is a good absorber; moreover the peak value of the PDE, attained at a wavelength of about 550 nm, is considerably lower than in other silicon detectors such as APDs or CCDs.

This behavior can be understood on the basis of the four physical phenomena underlying the photon detection process, namely: transmission of the light at the entrance interface, absorption into the active layers, collection of the photo-generated carriers and avalanche triggering. Figures 6, 7 and 8 report the behavior of some relevant quantities useful to interpret the data of Figure 4. In particular, Figure 6 shows the fraction T of the photons transmitted at the device entrance interface as a function of the wavelength. Figure 7 reports once again as a function of the wavelength, the probability that a photon entered into the device is absorbed in one of its layers. In particular blue, green and red curves represent the probability that the photon is absorbed respectively in the upper neutral region, in the space charge region or in the lower neutral region, while the black curve is the sum of the three terms. Finally, Figure 8 reports the hole and the electron avalanche probability, namely P_{bh} and P_{be}, as a function of the depth.

At short wavelengths (400–450 nm), the absorption length in silicon is short (100–500 nm) if compared with the device thickness and therefore the absorption efficiency is 100%, as confirmed by black

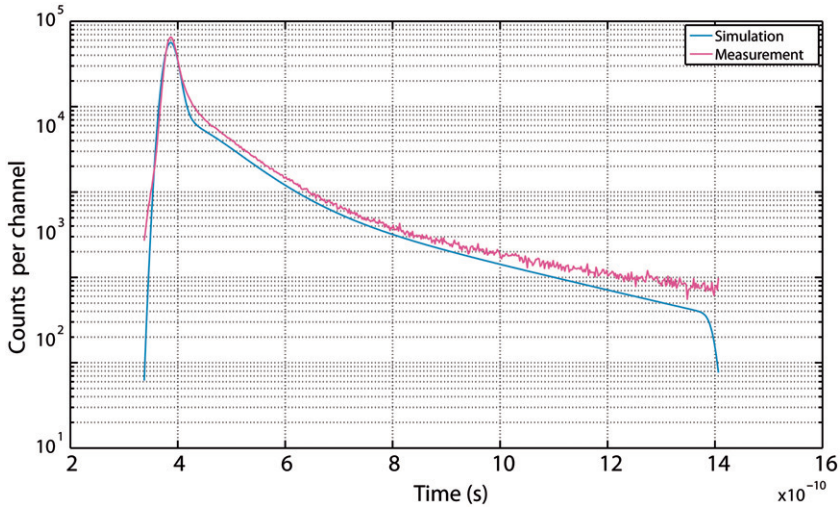

Figure 5. Comparison between calculated and measured temporal response. Both simulation and measurement have been carried out at a wavelength of 520 nm. (The color version of this figure is included in the online version of the journal.)

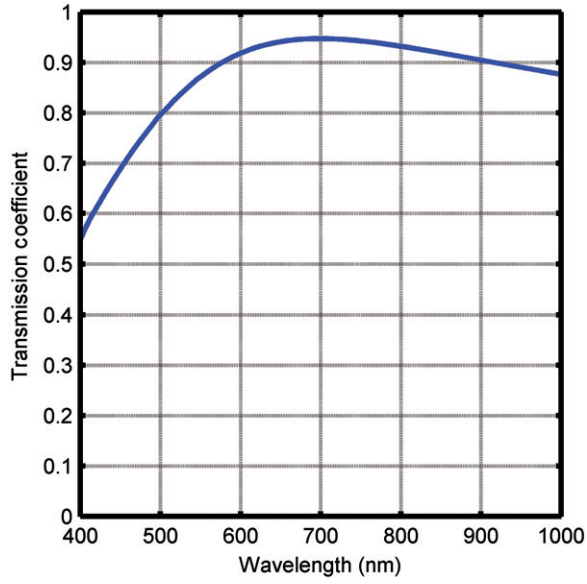

Figure 6. Optical transmission coefficient at the detector entrance interface. Values have been calculated using standard theory of reflection at multi-dielectric interfaces. (The color version of this figure is included in the online version of the journal.)

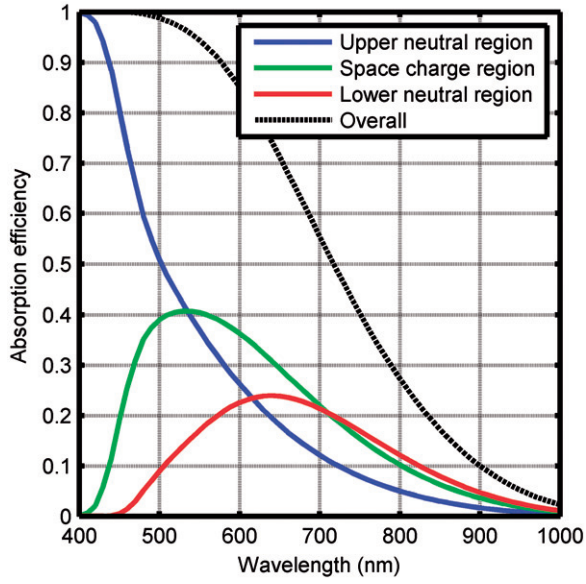

Figure 7. Absorption efficiency as a function of the wavelength. The figure reports the probability that a photon entered into the device is absorbed in one of its layers. (The color version of this figure is included in the online version of the journal.)

curve of Figure 7. Moreover, almost all the photons are absorbed into the upper neutral region (see blue curve) and the correspondingly photo-generated holes have to reach the space charge region by diffusion in order to trigger the avalanche. Following the considerations of Section 4, it is possible to assume that this transport process takes place with almost 100% efficiency. Therefore, limitations to PDE come neither from photon absorption nor from carrier collection. What really limits PDE is the avalanche triggering process. In this case the avalanche is initiated by the holes coming from the upper neutral region that are known [30] to be considerably less effective than the electrons. This behavior is confirmed from data of Figure 8. Note that a remarkable role is played also by transmission coefficient T; the PDE between 400 and 500 nm would considerably benefit of an improvement in the performances of antireflection coating.

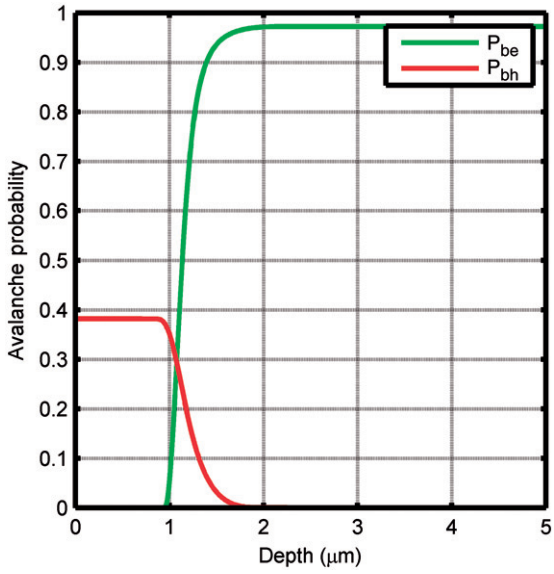

Figure 8. Electron and hole avalanche probability as a function of the depth. (The color version of this figure is included in the online version of the journal.)

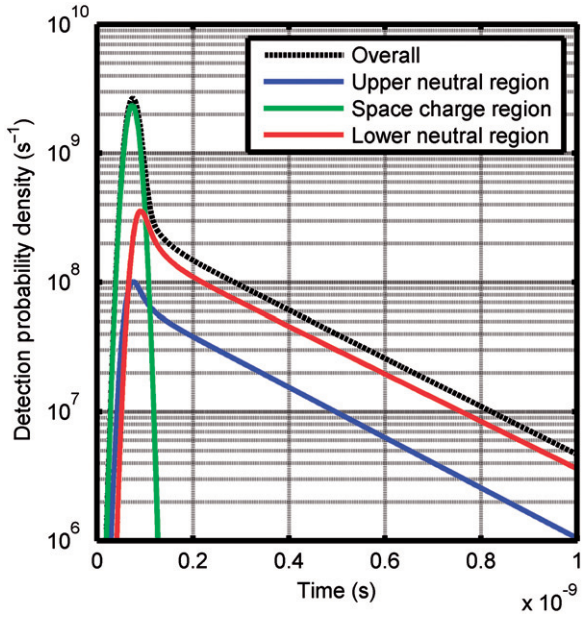

Figure 9. Temporal response of a standard SPAD calculated at a wavelength of 800 nm. The figure reports the overall response as well as the individual contributions owing to neutral regions and to the space charge region. (The color version of this figure is included in the online version of the journal.)

At intermediate wavelengths (450–550 nm) the absorption efficiency remains almost 100% (black curve). However, as the wavelength is increased, the fraction of the photons absorbed into the space charge region increases with a correspondingly increase of the effective triggering efficiency. Notably, at 550 nm, where the PDE peaks, a significant fraction ($\approx 40\%$) of the photons is still absorbed into the upper neutral region and therefore experiences low triggering efficiency. Moreover, most of the photons that reach the space charge region are absorbed in its first part where the triggering efficiency is still low. The combination of these two phenomena set the limitation to the maximum PDE attainable with current devices.

When the wavelength is increased further, the PDE starts decreasing. This is due to a combination of many phenomena. On the one hand, the fraction of photons absorbed into the upper neutral region decreases further with a benefit in term of effective triggering efficiency; on the other, the overall fraction of photon absorbed into the device decreases owing to the increase in silicon absorption length; moreover, the minority electrons generated by the photons absorbed into the lower neutral region experience a collection efficiency well below 100%.

5.2. *Temporal response*

Figure 9 represents the temporal response of a device belonging to lot S62, calculated at a wavelength of 800 nm, along with the individual contributions coming from each layer. First of all, let us consider the contribution due to the photons absorbed into the space charge region. Since the latter is quite thin ($<2\,\mu$m) and the field herein present is high enough to guarantee the saturation of the electrons velocity, then the dispersion of the transit times is almost negligible. Therefore, the corresponding contribution assumes a nearly Gaussian shape determined by the statistical phenomena associated with the avalanche build-up and propagation.

Let us consider now the two contributions due to the neutral regions. At a first sight, one would expect a lifetime considerably shorter for the contribution coming from the upper neutral region. Actually it is well known that, to a first instance, the lifetime is quadratically related to the thickness of the region [1]; on the other hand the upper neutral region is 3 times thinner than the lower one. However, this point does not find a confirmation in Figure 9 which shows that the two lifetimes are almost identical. This behavior is due to the combination of other phenomena that contribute to determine the value of the lifetime. First of all in the upper neutral region the diffusing minority carriers are holes and therefore are intrinsically 2.5 to 3 times slower than the electrons. Moreover, the really high doping level present in that region (in the order of

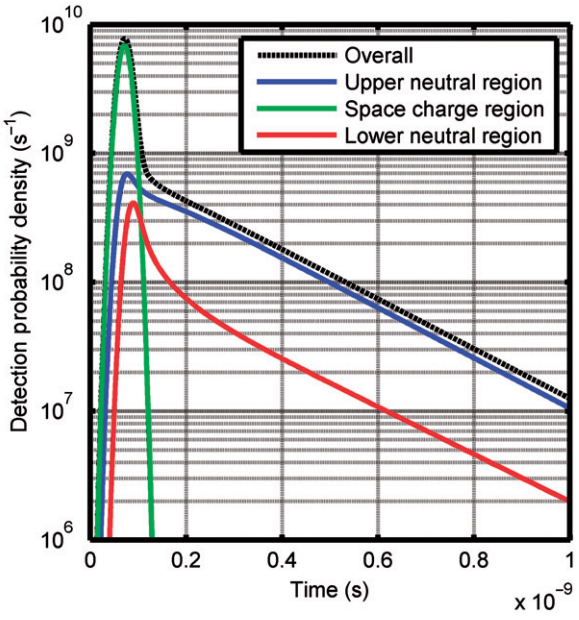

Figure 10. Temporal response of a standard SPAD calculated at a wavelength of 500 nm. The figure reports the overall response as well as the individual contributions owing to neutral regions and to the space charge region. (The color version of this figure is included in the online version of the journal.)

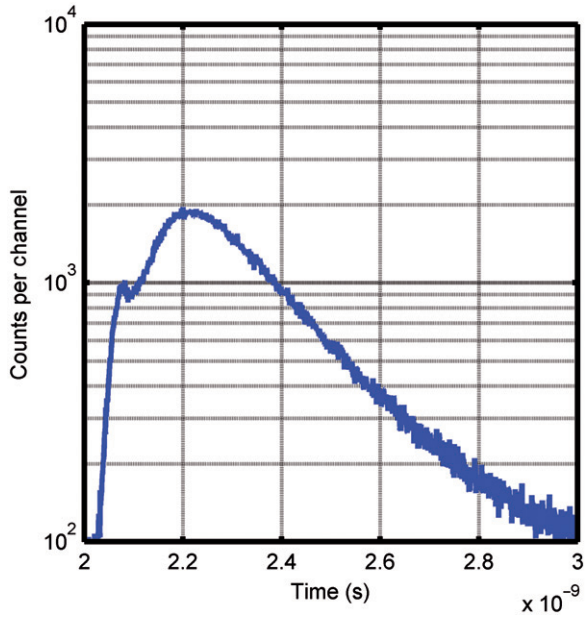

Figure 11. Temporal response of a SPAD belonging to lot S62 measured at a wavelength of 420 nm. (The color version of this figure is included in the online version of the journal.)

10^{19} cm^{-3}) remarkably contribute in further reducing the diffusivity of these carriers. Simulation accounts also for the small electric field owing to the gradient of the doping concentration. Although it tends to improve the motion of the holes, its contribution is largely insufficient to compensate the effect of the high doping level.

The area subtended by each contribution reflects the percentage of the photons absorbed into the corresponding region, apart from the effect of the collection efficiency and of the avalanche probability. Therefore, as the wavelength is reduced, the amplitude of the contribution coming from the lower region reduces while the one coming from upper region increases. This trend can be seen for example in Figure 10 that reports temporal response contributions calculated at a wavelength of 500 nm.

Figure 11 represents the temporal response of the same kind of device measured at a wavelength of 420 nm. In this case neither a sharp peak nor a quasi exponential tail can be distinguished anymore. The same behaviour is obtained also from the simulation results reported in Figure 12. However, the decomposition of the overall curve in its elementary components allows us to understand such a behaviour. The first peak in the temporal response can be attributed to the photons absorbed into the depleted region. Since, at this wavelength, they are only a small fraction of

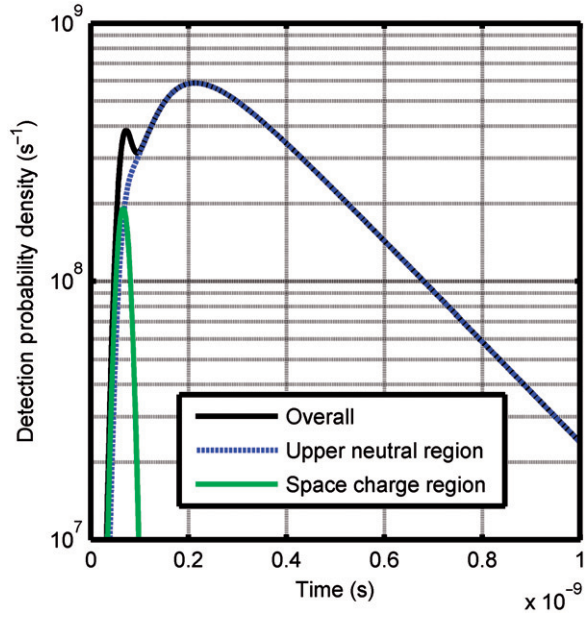

Figure 12. Temporal response of a standard SPAD calculated at a wavelength of 420 nm. The figure reports the overall response as well as the individual contributions owing to the upper neutral region and to the space charge region. The contribution from the lower neutral region is negligible and therefore has not been represented here. (The color version of this figure is included in the online version of the journal.)

the overall photons absorbed into the device, the amplitude of the peak is so small as to make it barely visible in the overall curve. The second part of the temporal response is entirely attributable to the

photon absorbed into the upper neutral region. However, unlike in the previous cases, its behavior is far from being exponential. A thorough analysis of the simulation results provided a simple explanation to this behavior. Owing to the short absorption length, photons are not uniformly absorbed into the neutral region; actually most of them are absorbed very close to the device surface. Therefore, at the very first beginning, the contribution to the tail comes only from the small amount of carriers photo-generated close to the space charge region; as time goes by, the tail amplitude increases since carriers generated close to the surface reaches the depleted region; finally the tail amplitude decreases as the neutral region depletes from photo-generated carriers.

6. Evaluation of the performances attainable with future devices

In the previous section main phenomena limiting both the PDE and the TR of currently available devices have been investigated as a function of the wavelength. In this section we will discuss some approaches that can be adopted in order to overcome these limitations. The effectiveness of these solutions will be evaluated with the help of the model described in Section 4. The aim is to provide the reader with an insight of the performance that can be expected in the next years if a strong development of the SPAD structure is pursued.

6.1. Photon detection efficiency at long wavelengths

As pointed out in the previous section, the basic limitation to the PDE at long wavelengths is the low absorption efficiency. An obvious approach to overcome this problem is to increase the overall thickness of the device active layer. In particular, it is possible to increase the thickness either of the space charge region or of the lower neutral region. The latter is certainly the easier solution to be implemented from a technological point of view since it does not influences the field profile inside the device and it only requires to start the processing from a thicker epitaxial layer [10]. However, this approach presents some remarkable drawbacks. First of all, the increase of the thickness of the neutral region results in a longer lifetime of the diffusion tail and therefore in a worsening of the temporal response. On the other hand, this solution is not optimal also for improving the PDE since the additional photon absorption takes place into the neutral region where the carrier's collection efficiency is well below 100%.

A more effective approach requires therefore to modify the space charge region. However, stretching out this region is much more complicated than increasing the thickness of the neutral layer. In effect the thickness of the depleted region cannot be increased simply by extending the low-doped region of the epitaxial wafer on which the device is fabricated. In this case, we would obtain a device with an electric field profile remarkably different from that of current devices. This would adversely affect all those device characteristics that are strongly influenced by the electric field profile, i.e. the dark count rate, the temporal resolution and the avalanche triggering probability.

We showed [31] that a suitable modification to the doping profile of the epitaxial wafer makes it possible to expand the space charge region without modifying the electric field profile across the junction. In particular in the structure we proposed it is possible to distinguish between a multiplication region and a drift region. The former is a high field region where the avalanche takes place, while the latter has the sole purpose of collecting the photo-generated carriers and accelerate them towards the multiplication region. The only requirement for the drift region is to have a field at least of 10^4V/cm to guarantee the saturation of the carrier's velocities and therefore the minimization of transit time dispersion; on the contrary, the electric field profile in the multiplication region must be carefully designed in order to guarantee the best device performances. In particular we proposed a modification of the starting epitaxial layer that makes it possible to obtain a multiplication field nearly identical to the one of current devices. Since the dark count rate, avalanche probability and temporal resolution mainly depend on the field in the multiplication region, this approach should allow to keep them unvaried while increasing the thickness of the space charge region. As an example of the results attainable with this approach, Figure 13 reports the calculated PDE of a standard device as compared to that of an extended device having a space charge region 10 μm-thick; a remarkable improvement in detection efficiency at long wavelengths is achieved, with a PDE of about 40% at a wavelength of 800 nm.

6.2. Photon detection efficiency at short wavelengths

An increase of the thickness of the depleted region has a limited effect at the intermediate wavelengths (500–550 nm). In fact in Section 5 we showed that in this range the PDE is not limited by absorption efficiency but by the avalanche probability. Therefore, a different approach is needed to improve the PDE at short wavelengths.

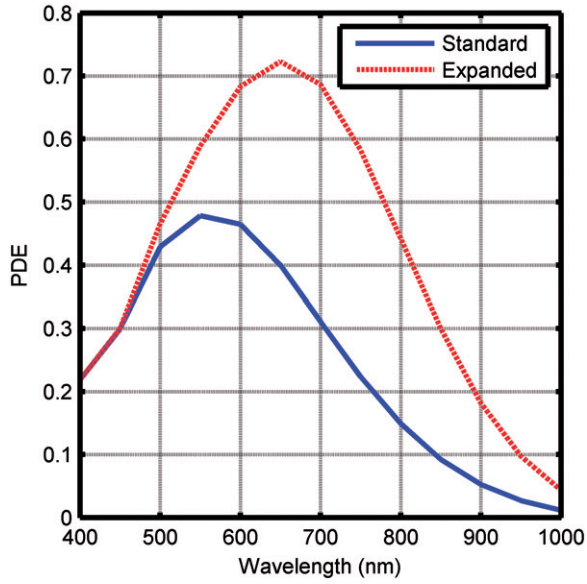

Figure 13. Comparison between the PDE of two different devices. Blue curve represents a standard device while red curve corresponds to a detector with an extended space charge region (thickness of 10 μm). (The color version of this figure is included in the online version of this journal.)

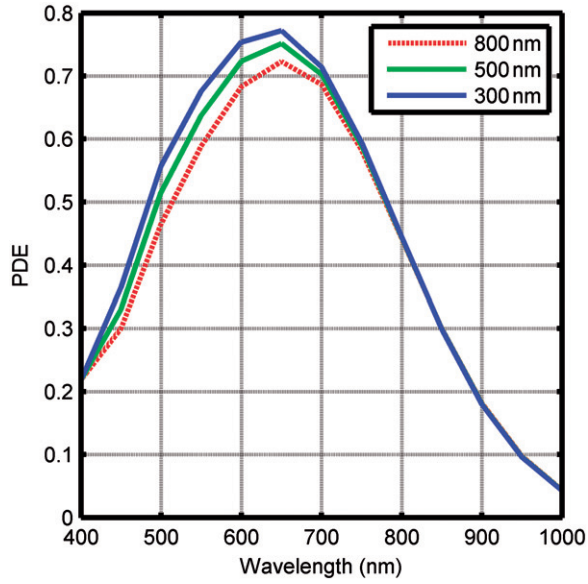

Figure 14. Comparison between the PDE of three devices characterized by different values for the thicknesses of the upper neutral region, respectively of 800, 500 and 300 nm. All the devices have an extended space charge region of 10 μm. (The color version of this figure is included in the online version of the journal.)

A first possibility is to design a device with a complementary doping type for each region. This would lead to a device with a p-type upper neutral region in which, at short wavelengths, the avalanche would be initiated by the electrons. On the one hand, this greatly improves the PDE at short wavelengths, but at same time it results in a strong reduction of the PDE at longer ones where the avalanche is initiated by the holes. Therefore, this solution can be considered viable only for devices specifically targeted to operate in the blue region of the spectrum.

A more general approach relies on the reduction of the thickness of the shallow n region in order to limit the fraction of the photons herein absorbed. In order to evaluate the effectiveness of this solution, in Figure 14 we compared the PDE of three devices characterized by the same structure ($x_{ZCS} = 10\,\mu m$) but with a different thickness of the upper neutral region. Red, green and blue curves correspond to a thickness of 800, 500 and 300 nm, respectively. Even the latter can be easily obtained in real devices thanks to modern planar processes; however, this requires a partial redesign of the device in order to keep unvaried the electric field profile inside the depleted region.

The improvement attainable by thinning the upper neutral layer is certainly not negligible; however, the resulting PDE is still far from being ideal, especially at the wavelengths below 500 nm. The reason is once again related to the poor triggering efficiency that characterizes the first part of the space charge region. Therefore, if a larger improvement of the PDE at the short wavelengths is needed, a different approach must be adopted.

A first option is to engineer the electric field profile inside the depleted region in order to make the growth of the triggering efficiency across the junction steeper. However, a discussion involving the influence of the electric field profile on the behavior of P_{be} and P_{bh}, and the trade off that must be faced in the design of such device is well beyond the scope of this paper. Some consideration on these topics can be found for example in [31] and [8].

However, a more promising solution to improve the PDE at the short wavelengths is to completely reverse the device structure in order to obtain an electric field profile like the one depicted in Figure 15. This configuration is optimal since the avalanche is mainly initiated by the electrons regardless of the wavelength. Moreover, at the short to medium wavelengths the effective triggering efficiency would be particularly high, since the carriers are generated far from the junction where P_{be} is close to unity. Unfortunately, the implementation of such a structure in a planar silicon technology is quite challenging.

6.3. Temporal response at short wavelengths

As discussed in the previous section, at short wavelengths, the SPAD temporal response is strongly

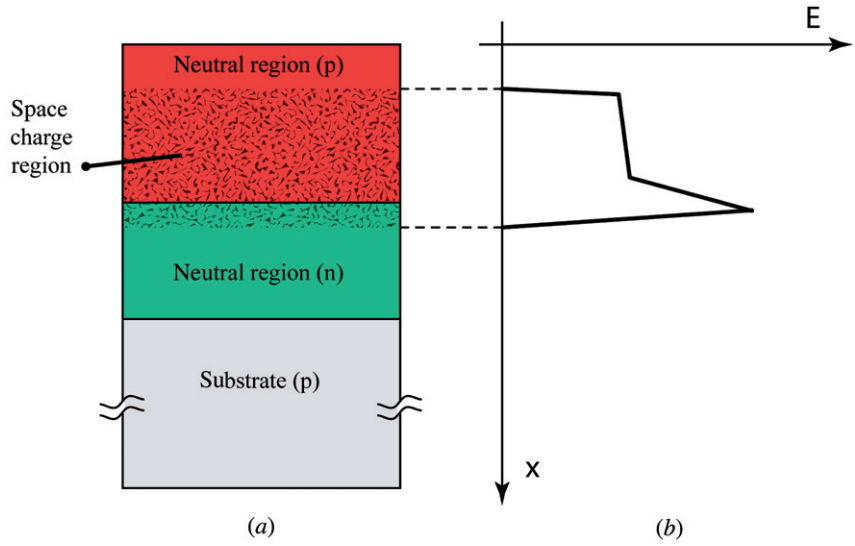

Figure 15. Cross-section of the active region of a thin SPAD with an inverted structure (*a*) with a qualitative electric field profile for the multiplying space charge region (*b*). (The color version of this figure is included in the online version of the journal.)

influenced by the diffusion of minority holes in the upper neutral region. In order to evaluate the improvement that can be obtained by thinning the shallow n region, we compared the temporal responses of two devices; the first is a standard one, while the second has an upper neutral region only 300 nm thick. Figure 16 illustrates the results of the simulation carried out at a wavelength of 420 nm. The attainable improvement is significant; in particular the temporal response is again dominated by a sharp peak while the tail lifetime strongly reduces.

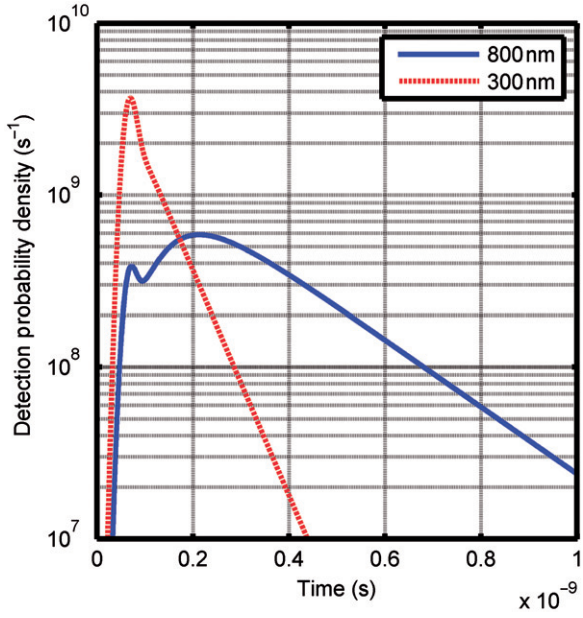

Figure 16. Temporal response calculated at a wavelength of 420 nm. Both detectors have a standard depleted region, but are characterized by different thicknesses of the upper neutral region (800 and 300 nm, respectively). (The color version of this figure is included in the online version of the journal.)

6.4. Temporal response at long wavelengths

Let's consider once again the devices characterized by an extended space charge region. Our aim here is to investigate the temporal response that is expected from these detectors. First of all one would expect that the tail lifetime is unvaried since it depends mainly on the thicknesses of the neutral regions that are unchanged. On the other hand, one would expect also a reduction of the tail amplitude; in effect the extension of the space charge region reduces the amount of photons that reaches the lower neutral region and that are herein absorbed. In order to verify this hypothesis we calculated the temporal response of these devices. Figure 17 reports a comparison between the temporal responses of two devices calculated at a wavelength of 800 nm; one is a standard SPAD, while the other has a 10 μm thick space charge region. A few features can be observed. First of all, the peak of the extended device is considerably broadened. The reason is the dispersion of transit times; in effect, since the carriers travel at a saturated velocity of about 1 μm/10 ps, the maximum transit time across a region of 10 μm is about 100 ps. The corresponding width obtained from simulation is about 90 ps FWHM. Similarly, the area subtended by the peak is considerably increased due to the growth

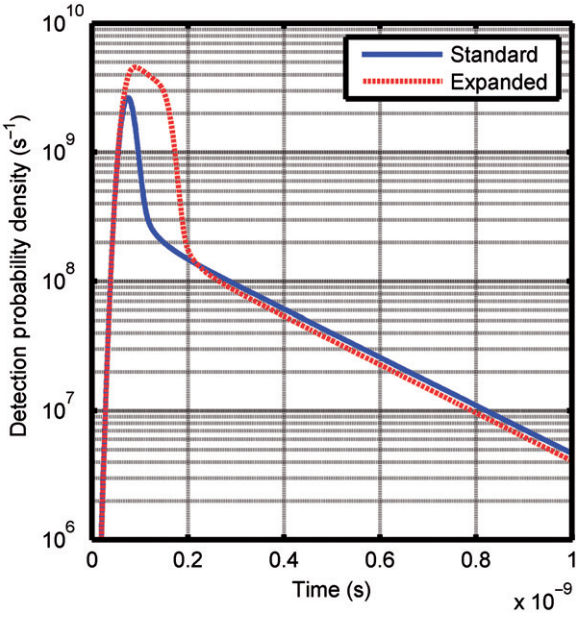

Figure 17. Temporal response calculated at a wavelength of 800 nm. The blue curve corresponds to a standard device, while the red curve corresponds to an extended device, characterized by a 10 μm thick depleted region. (The color version of this figure is included in the online version of this journal.)

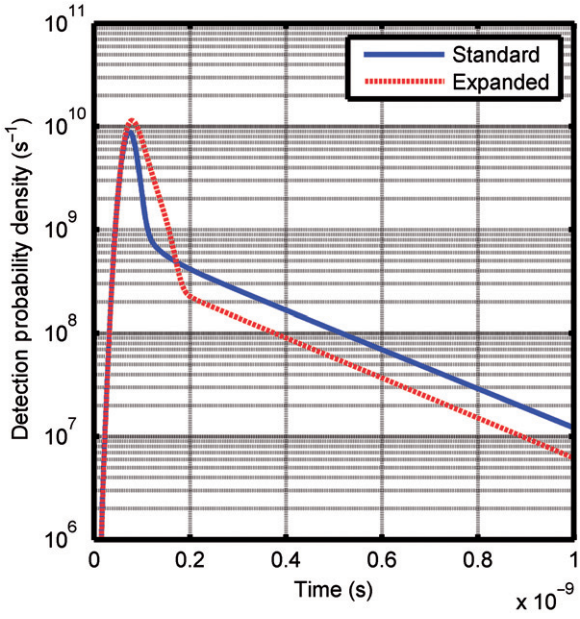

Figure 18. Temporal response calculated at a wavelength of 600 nm. The blue curve corresponds to a standard device, while the red curve corresponds to an extended device, characterized by a 10 μm thick depleted region. (The color version of this figure is included in the online version of this journal.)

of the number of photons absorbed into the depleted region. Unexpectedly, the amplitude of the diffusion tail has undergone only a minor decrease. The cause is a combination of two concurrent phenomena. On the one hand, the number of the photons absorbed into the lower neutral region decreases of a factor

$$\Gamma = \exp\left(\Delta L / L_{\text{ass}}\right) \tag{3}$$

where ΔL is the increase in the thickness of the depleted region and L_{ass} is the silicon absorption length at the considered wavelength. At 800 nm this factor amount only to 1.6. On the other hand, the increase of the thickness causes a delay of about 80 ps of the tail due to the transit time across the space charge region. In this specific case, the two phenomena compensate almost perfectly leading to a negligible reduction in the tail amplitude.

Figure 18 shows the same kind of comparison for a wavelength of 600 nm. Although more evident, also in this case the amplitude reduction is quite limited. In effect, on the one hand, the smaller value of L_{ass} allows for a stronger reduction in the number of photons absorbed into the lower neutral region ($\Gamma \approx 20$); however, the overall reduction is limited by the strong contribution due to the upper neutral region.

We can therefore conclude that while the extension of the space charge region remarkably improves the PDE at longer wavelengths, at the same time it has a limited effect on the slow part of the temporal response both at longer and short wavelengths.

In order to improve the TR at short to medium wavelengths, it is mandatory to reduce the thickness of the upper neutral region. As an example of the attainable results, Figure 19 shows a comparison between the TRs, calculate at a wavelength of 600 nm, for two different devices. Both the detectors have a space charge region of 10 μm, but a different thickness of the upper neutral region, 800 and 300 nm, respectively. In the TR of the latter device the two contributions of the tail are well distinguishable, being characterized by quite different lifetimes. The corresponding improvement is remarkable.

In order to improve the temporal response at longer wavelengths, it would be also required to reduce the thickness of the lower neutral region. However, this conflicts with other design requirements; for example the lower neutral region have to provide a low resistivity path for the current to be collected at the anode; this in turn gives some constraints on the thickness and on the doping of that region. A few solutions have already been proposed in the past; for example Spinelli et al. [32] developed a device with a patterned buried layer such that there is no neutral layer below the active region. Although this solution

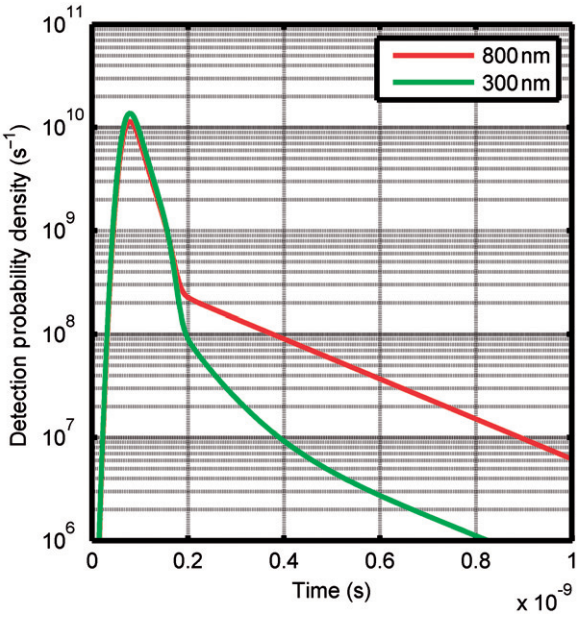

Figure 19. Temporal response calculated at a wavelength of 600 nm. Both the detectors have an extended space charge region (10 μm), but are characterized by different thicknesses for the upper neutral region (800 and 300 nm, respectively). (The color version of this figure is included in the online version of the journal.)

proved to be effective in reducing the slow component of the tail, it is really critical from a technological point of view; moreover this approach is suitable only for devices with small active area diameter (8–10 μm).

In conclusion, the reduction of the diffusion tail at longer wavelengths is still an open issue; new solutions have to be devised and thoroughly investigated.

7. Conclusions

In this paper we reported a physically based model that proved to be useful for the investigation of the phenomena that limits both the photon detection efficiency and the temporal response of currently available SPADs. In particular, we outlined how the limitations to the PDE at short wavelengths are basically due to poor triggering efficiency while at longer wavelengths can be attributed to limited absorption efficiency.

Subsequently, we discussed some modification that can be applied to the device structure in order to overcome these limitations and we verified their effectiveness by using the propose model. In particular we found that a suitable increase of the thickness of the depleted region can greatly improve the PDE at longer wavelengths (i.e. greater than 600 nm) while it has a limited effect at the short ones. We demonstrated that a significant improvement at short wavelengths can be obtained only through a complete modification to the device structure since the advantages that can be obtained by thinning the upper neutral region are limited. On the contrary, the latter modification proved to be very effective in improving the device temporal response at short to medium wavelength if combined with the extension of the depleted region; unfortunately the same conclusion does not apply to higher wavelengths. Therefore, the problem of reducing the slow component of the temporal response at wavelengths higher than 600–700 nm is still an open issue.

Acknowledgements

This work was supported by the EC grant agreement no. 232359 (PARAFLUO) FP7–SME–2008–1 and by EC grant agreement no. 248095 (Q-ESSENCE) FP7–ICT–2009–4.

References

[1] Yang, H.; Luo, G.; Karnchanaphanurach, P.; Louie, T.M.; Rech, I.; Cova, S.; Xun, L.; Xie, X.S. *Science* **2003,** *302,* 262–266.
[2] Michalet, X.; Siegmund, O.H.W.; Vallerga, J.V.; Jelinsky, P.; Millaud, J.; Weiss, S. *J. Mod. Opt.* **2007,** *54,* 239–281.
[3] Michalet, X.; Colyer, R.A.; Scalia, G.; Kim, T.; Levi, M.; Aharoni, D.; Cheng, A.; Guerrieri, F.; Arisaka, K.; Millaud, J.; Rech, I.; Resnati, D.; Marangoni, S.; Gulinatti, A.; Ghioni, M.; Tisa, S.; Zappa, F.; Cova, S.; Weiss, S. *Proc. SPIE* **2010,** *7608,* 76082D.
[4] Buller, G.S.; Wallace, A. *IEEE J. Sel. Top. Quantum Electron.* **2007,** *13,* 1006–1015.
[5] Gordon, K.J.; Fernandez, V.; Buller, G.S.; Rech, I.; Cova, S.; Townsend, P.D. *Opt. Express* **2005,** *13,* 3015–3020.
[6] Barbieri, C.; Naletto, G.; Occhipinti, T.; Facchinetti, C.; Verroi, E.; Giro, E.; Di Paola, A.; Billotta, S.; Zoccarato, P.; Bolli, P.; Tamburini, F.; Bonanno, G.; D'Onofrio, M.; Marchi, S.; Anzolin, G.; Capraro, I.; Messina, F.; Belluso, M.; Pernechele, C.; Zaccariotto, M.; Zampieri, L.; Da Deppo, V.; Fornasier, S.; Pedichini, F. *J. Mod. Opt.* **2009,** *56,* 261–272.
[7] Gulinatti, A.; Rech, I.; Maccagnani, P.; Ghioni, M.; Cova, S. Large-area Avalanche Diodes for Picosecond Time-correlated Photon Counting. Proceedings of the 35th European Solid-State Device Research Conference (ESSDERC 2005), Grenoble, France, September 12–16, 2005.
[8] Ghioni, M.; Gulinatti, A.; Maccagnani, P.; Rech, I.; Cova, S. *Proc. SPIE* **2006,** *6372,* 63720R.
[9] Lacaita, A.L.; Ghioni, M.; Cova, S. *Electron. Lett.* **1989,** *25,* 841–843.
[10] Ghioni, M.; Gulinatti, A.; Rech, I.; Zappa, F.; Cova, S. *IEEE J. Sel. Top. Quantum Electron.* **2007,** *13,* 852–862.

[11] PerkinElemer Optoelectronics. SPCM-AQ Datasheet. http://www.perkinelmer.com (accessed May 27, 2010).
[12] Dautet, H.; Deschamps, P.; Dion, B.; MacGregor, A.D.; MacSween, D.; McIntyre, R.J.; Trottier, C.; Webb, P.P. *Appl. Opt*. **1993,** *32,* 3894–3900.
[13] Becker, W. *Advanced Time-correlated Single Photon Counting Techniques*; Springer: Berlin, 2005.
[14] Lacaita, A.L.; Spinelli, A.; Longhi, S. *Appl. Phys Lett*. **1995,** *67,* 2627–2629.
[15] Spinelli, A.; Lacaita, A.L. *IEEE Trans. Electron Devices* **1997,** *44,* 1931–1943.
[16] Gulinatti, A.; Maccagnani, P.; Rech, I.; Ghioni, M.; Cova, S. *Electron. Lett*. **2005,** *41,* 272–274.
[17] Rech, I.; Resnati, D.; Gulinatti, A.; Ghioni, M.; Cova, S. *Rev. Sci. Instrum*. **2007,** *78,* 086112.
[18] Gulinatti, A.; Rech, I.; Fumagalli, S.; Assanelli, M.; Ghioni, M.; Cova, S. *Proc. SPIE* **2009,** *7355,* 73550X.
[19] Born, M.; Wolf, E. *Principles of Optics*; Cambridge University Press: Cambridge, 1999.
[20] Hull, R., Ed. *Properties of Crystalline Silicon (Emis Series)*; The Institution of Engineering and Technology: London, 1999.
[21] Sze, S.M.; Ng, K.K. *Physics of Semiconductor Devices*; Wiley-Interscience: New York, 2006.
[22] Oldham, W.G.; Samuelson, R.R.; Antognetti, P. *IEEE Trans. Electron Devices* **1972,** *19,* 1056–1060.
[23] Okuto, Y.; Crowell, C.R. *Phys. Rev. B* **1974,** *10,* 4284–4296.
[24] Okuto, Y.; Crowell, C.R. *Phys. Rev. B* **1972,** *6,* 3076–3081.
[25] McIntyre, R.J. *IEEE Trans. Electron Devices* **1999,** *46,* 1623–1631.
[26] Ingargiola, A.; Assanelli, M.; Rech, I.; Gallivanoni, A.; Ghioni, M.; Cova, S. *Proc. SPIE* **2009,** *7320,* 73200K.
[27] Assanelli, M.; Ingargiola, A.; Rech, I.; Gulinatti, A.; Ghioni, M. *Proc. SPIE* **2010,** *7681,* 76810L.
[28] Ghioni, M.; Gulinatti, A.; Rech, I.; Maccagnani, P.; Cova, S. *Proc. SPIE* **2008,** *6900,* 69001D.
[29] Rech, I.; Labanca, I; Armellini, G.; Gulinatti, A.; Ghioni, M.; Cova S. *Rev. Sci. Instrum*. **2007,** *78,* 063105.
[30] McIntyre, R.J. *IEEE Trans. Electron Devices* **1973,** *20,* 637–641.
[31] Gulinatti, A.; Panzeri, F.; Rech, I.; Maccagnani, P.; Ghioni, M.; Cova, S. *Proc. SPIE* **2010,** *7681,* 76810M.
[32] Spinelli, A.; Ghioni, M.; Cova, S.; Davis, M.L. *IEEE J. Quantum Electron*. **1998,** *34,* 817–821.