Monolithically Integrated Photo-Diodes in Standard CMOS Technology for high speed optical communication: General Consideration and Analysis

Saša Radovanović, AnneJohan Annema and Bram Nauta
MESA+ Research Institute, IC-Design Group, University of Twente,
Enschede, The Netherlands.
email: s.radovanovic@el.utwente.nl
URL: icd.el.utwente.nl
Phone +31 53 489 1061
Fax +31 53 489 1034

Abstract: **Photodiodes designed in standard CMOS technology which can be monolithically integrated in high-speed optical receivers are analyzed and the advantages and drawbacks concerning bandwidth with respect to the diode geometry and structure, are discussed. Studied photodiode structures that can be realized in standard CMOS technology are: 1) N-well/P-substrate, 2) lateral N-well/P-substrate (exploiting only depletion region in between) 3) N^+/P-substrate and 4) P^+/N-well/P-substrate diodes. The maximal operating frequency as well as the impulse response of the diodes are calculated using 2-D model semiconductor device analysis.**

KEYWORDS: **CMOS photodiode, high-speed optical receiver, quantum efficiency, monolithic integration.**

I. Introduction

The cost of the optical transmission equipment for data is a critical issue in many systems. Because of this cost aspect integrating optical receivers monolithically, in CMOS, is very interesting. A key component for such systems is a high-speed monolithically integrated CMOS photodiode.

Further evolution of CMOS technology will lead to downscaling of junctions as well as increased doping levels. As a result of increased doping levels, both the depletion layer width, at a fixed voltage, is decreased and the maximum allowed voltage across junctions is decreased. The overall result is that with CMOS technology evolution, the detector-sensitivity for the same speed of response decreases significantly for long wavelength light ($\lambda = 850$ nm).

The bandwidth (speed) of the photodiode is determined inside the diode by slow diffusion carriers transit time, which can be characterized as a *physical limitation*. Outside the diode, its capacitance increases the overall input capacitance of the transimpedance amplifier which introduces low-frequency pole and limiting thus its overall bandwidth - *electrical limitation*. The maximum photodiode frequency is determined by a minimal of the former and later limitations.

The calculation of the maximal operating frequency of the diodes in this paper assume constant gain of the transimpedance amplifier(R) and constant area that photodiodes occupy on the chip i.e. the diode capacitances are taken to be the same, so the maximal operating frequency is limited only by the the excess carrier diffusion.

The widely used figure of merit for the photodiodes is their photocurrent and bandwidth product. The photocurrent is directly proportional to a quantum efficiency [2]. In the calculation presented in [3], the area of the photodiode was included, and it was shown that QE strongly depend on photodiode dimensions, while the effect of the photodiode structure is less significant. That's because it is mainly determined by absorption coefficient $\alpha(\lambda)$ of photons in silicon, the dimension of the absorbing layers and the diffusion length of the minority carriers [3].

However, the speed of the photodiode *does* depend on the structure. Light generated carriers inside N-well or N^+ and in the substrate, will diffuse from the place they are generated to a place where they are detected as a photocurrent - the depletion region edge(s). They will need time much longer (microseconds) than for the carriers generated inside the depletion layer that swiftly reach the corresponding edge of the depletion region.

Since the overall photodiode areas will be taken as constant, the quantum efficiency for all diodes will be the same. Comparing their internal bandwidth will determine which photodiode structure and geometry will give the best performance for high-speed optical communication.

In our calculation the same photodiodes chip area (same

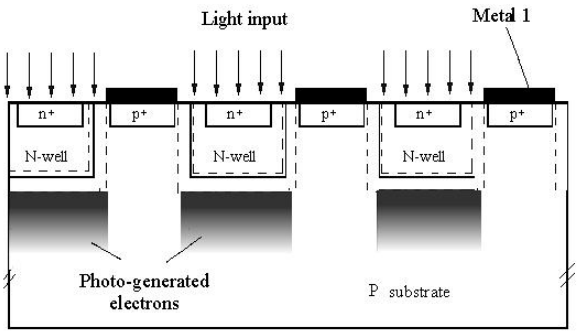

Fig. 1. Fingerprint structure of N-well/P-substrate diode in CMOS technology

QE) with a different geometries is used. We will consider fingerprint structures where the light can penetrate only through N regions, while at the P-region there will be metal contact over the whole area, behaving also as a light shield. This is standard photodiode connection within the ciruits when the distance between the N-wells is less than it's width.

II. N-WELL/P-SUBSTRATE DIODE

The photodiode structure is shown in Figure 1. Since the light signal is shielded in P region, carriers will be generated only bellow N-well and start diffusing towards junctions. The overall photocurrent density will be a sum of three contributing photocurrents: diffusion current inside N-wells, diffusion current inside P-substrate, and drift current inside depletion (space-charge) region.

Photogenerated carriers inside depletion region will be swept out much faster than the diffusion carriers and in the calculation for the overall frequency current density response, its influence will be taken as a frequency independent.

A. Two-dimensional theoretical solution of the carrier profile inside P-substrate

Light generated carriers inside substrate can diffuse up to the junctions with N-wells but also laterally towards next N-wells as well as P-contacts. The carrier profile can be calculated in a similar manner as in [1], since the structure is periodic in y-direction, with the periodicity l. By applying the Laplace transform of the time variable t we have:

$$0 = D_n \frac{\partial^2 n_p}{\partial x^2} + D_n \frac{\partial^2 n_p}{\partial y^2} - n_p(\frac{1}{\tau_n} + s) + g(x, y, s) \quad (1)$$

where τ_n is excess carrier lifetimes, D_n, diffusion coefficient of the electrons outside the depletion region, $g(x, y, s)$ is the volume generation rate due to a Dirac light input, and can be expressed as $g = g_0 e^{-\alpha l_x} e^{-\alpha x}(H(y) - H(y - l_N)$, with l_x the distance between the surface and the lower part of the depletion region, l_N is the width of the N-well and H represents a step function.

We can expand both n_p and $g(y)$ in a diffusion equation (1) as a Fourier series with the periodicity of the pitch. The diffusion equation is solved for every harmonic component of the Fourier series of both n_p and $g(y)$. We will assume the infinite depth of the substrate, so the initial conditions for the carrier profile at the boundaries are taken to be zero. Calculated carrier profile in the P-substrate is further used to calculate the photocurrent density at the lower border of the depletion region.

If the size of the both N-well and P regions are taken to be the same ($ln = l/2$), the higher order solution are all uneven and the sum of contributions corresponds to equation (9) in [1].

There are two main features of the electrons which diffuse towards the area between N-wells. First, they will further diffuse towards regions with the lowest electron profile, and that is either left or right side of the N-wells or towards P-metal contacts since the electron profile in this sections is minimal i.e. zero. The second, they will choose a minimal time-path towards those regions (the highest gradient of the electron profile).

As long as the depth of the N-well and the depth of depletion region are together larger than the half of the p-region length ($l - l_N < 2l_x$) most of the electrons will diffuse towards N-well sides and will contribute to the overall photocurrent. Only small number of carriers diffuse towards P-metal contacts at the surface where they are recombined, and do not contribute to an overall photocurrent. The smaller is the N-well depth in comparison with the distance between N-wells, more and more electrons will diffuse toward "nearer" P-metal contact and will not contribute to a photocurrent. The diffusion path of detected carriers is minimal and equal to a N-well depth, so increasing the distance between N-wells will have no influence on the speed of the diffusive electrons, but the number of detected electrons is decreased.

The electron profile in this region can be calculated as follows. The number of carriers, reaching top and side boundaries of the area between N-wells in time, is equal to the number of carriers crossing the area between the bottoms of the space-charge regions. In order to simplify complex calculation of the exact carrier profile, we can consider the ratio between the area occupied by electrons that reach the N-wells sides and contributes thus to

a photocurrent, and area occupied by electrons that reach P-metal contact and does not contribute to an overall photocurrent.

Depending on the width of the P-well region in comparison with the depth of the N-well plus depletion region, the percentage of electrons contributing to a photocurrent is

$$\xi = \begin{cases} \frac{l_x}{l-l_N} & \text{if} \quad l - l_N > 2l_x, \\ \frac{l_x - (l-l_N)/4}{l_x} & \text{if} \quad l - l_N < 2l_x. \end{cases} \tag{2}$$

Since the current density profile is found in the plain of the lower border of space-charge region, it is integrated over the distance between N-wells, and multiplied with the percentage of electrons contributing to a photocurrent (ξ). The overall photocurrent is then calculated as:

$$\frac{J_{sub(j\omega)}}{\Phi_o(j\omega)} = \frac{\alpha q l e^{-\alpha l_x}}{\pi^2 n^2} \sum_{n=1}^{\infty} \frac{(1-\cos(\frac{n\pi l_n}{l})^2)(1-\xi)}{\sqrt{\left(\frac{2n\pi L_n}{l}\right)^2 + s\tau_n + 1} + \alpha L_n} \tag{3}$$

Zero order solution can be expressed as:

$$\frac{J_0(j\omega)}{\Phi_0(j\omega)} = \frac{e\ ln\ \alpha L_n e^{-\alpha l_x}}{l^2} \frac{(ln + \xi(l-ln))}{\sqrt{1+s\tau_n} + \alpha L_n} \tag{4}$$

If we take that $ln = l/2$ and $\xi = 1$ the derived equations are equal to the equations (8) and (9) in Ref. [1]. For the photodiodes using minimal size N-well and the minimal distance in between, the high order current density solution can be neglected since ξ 1.

B. Two-dimensional theoretical solution of the carrier profile inside N-well

The minority carrier profile is inside N-well region calculated in Ref [1], using the boundary conditions that photo-generated hole concentration is zero at the three junction boundaries and the top side behave as a reflective surface since there is no metal contact above (the normal component of the gradient of the carrier density is zero [4]). The carrier distribution function p_n and the carrier generation function $g(t)$ are rewritten as the product of two Fourier series satisfying above boundary conditions. The total current is presented as:

$$\frac{J_{N-well(j\omega)}}{\Phi_0(j\omega)} = 32\frac{qL_p^2\alpha(1-e^{-\alpha l_{x1}})}{l\pi^2 l_x} \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} \frac{\frac{2l_{x1}}{l_y}\left(\frac{1}{2n-1}\right)^2 + \frac{l_y}{2l_{x1}}\left(\frac{1}{2m-1}\right)^2}{\left(\frac{(2n-1)\pi L_p}{2l_{x1}}\right)^2 + \left(\frac{(2m-1)\pi L_p}{l_y}\right)^2 + 1 + j\omega\tau_p} \tag{5}$$

and 3dB frequency is found as:

$$f_{3dB} = \frac{\pi D_p}{2}\left(\left(\frac{1}{2l_{x1}}\right)^2 + \left(\frac{1}{l_y}\right)^2 + \left(\frac{1}{L_p}\right)^2\right) \tag{6}$$

where l_{x1} is the depth of the N-well region. The above equation shows that generated excess hole carriers choose minimal time paths towards minimal (zero) hole profiles. If the depth of the N-well region is larger than half of the N-well width ($2l_{x1} > l_N$), the speed of the diffusion carriers will be determined by the width of the N-well. On the other side, if the depth of the N-well is smaller, it will be the one determining the hole diffusion speed. The wider width of the N-well, will not influence the detector maximum operating frequency.

C. Current density response of the depletion region

If the diffusion inside depletion region is neglected, the photocurrent density response will be proportional to the electron and hole transit time. This time is a linear function between the carriers drift velocities and the depletion region width. In general, these velocities depend on the electric field. Since the photodiode is reversely biased, there is strong electric field inside depletion region so drift velocities are maintained at their fixed saturation values. A frequency response of the photocurrent inside depletion region will be much higher than a slow response of the diffusion carriers outside this region. For a simplicity reasons it can be presented as:

$$\frac{J_{DR}}{\Phi_0(j\omega)} = q\left[(e^{-\alpha l_{x1}} - e^{-\alpha l_x})\frac{A_{\text{eff}}}{A_{\text{total}}} + 2(1-e^{-\alpha l_x})\frac{A_{\text{eff1}}}{A_{\text{total}}}\right. \tag{7}$$

where $l_x - l_{x1} = W_{\text{DR}}$ is depletion region length, A_{eff} is N-well effective area, A_{eff1} is side depletion effective area and A_{total} is overall photodetector area.

The overall current response of the photodiode is the sum of the three current responses (Figure 2.).

In Figure 2., the current density amplitude response for two detectors having the pitch of 3μm and 10μm and well depth of 1μm, are presented. On the same figures the responses of the hole and electron currents are also presented.

The detector with the smaller pitch size has larger depletion region current contribution and lower N-well contribution, while the substrate electron current is maximal since there are almost no electrons reaching a P-metal contact. If larger pitch is used inside detector, the N-well current contribution will be larger, but the depletion region area decrease, resulting in the lower maximum operating frequency.

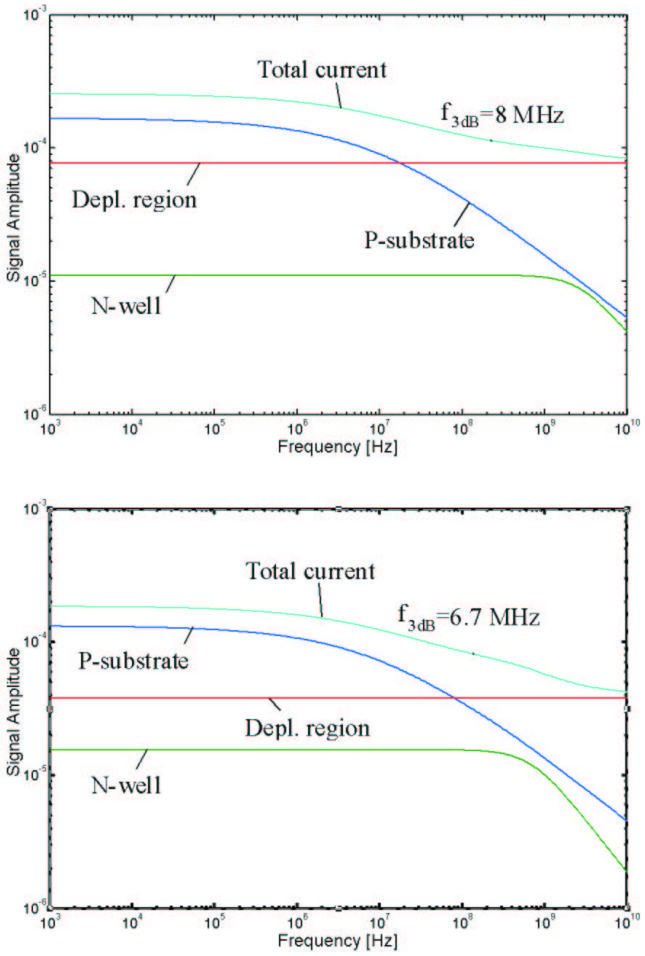

Fig. 2. Amplitude current density response of the N-well/P-substrate photodiodes in the same technology; the width of the pitch=3μm and 10μm, respectively

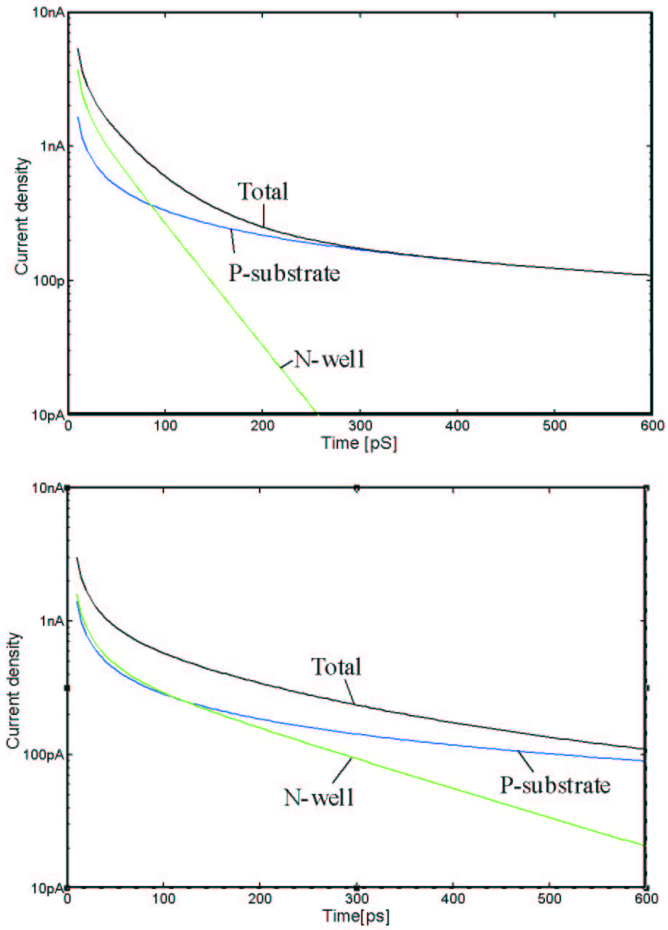

Fig. 3. Current density impulse response for the excess carriers inside the photodiodes; the width of the pitch=3μm and 10μm, respectively

The electron and hole current responses have been calculated by taking the inverse Laplace transform of (3) and (4). Since we took frequency invariant response of the depletion regions, they have no influence on the time-domain current response. The results are shown on Figure 3. The overall current density responses are compared on the Figure 4.

The faster impulse response is achieved using minimal pitch photodiode.

III. Two-dimensional response of the N+/P-substrate diode

The fingerprint structure of the N^+/P-substrate diode is presented in Figure 2. The current density response of this diode will be similar with the one calculated in the previous subsection. Since the doping concentration of the shallow N^+ is much larger than in the N-well region, the corresponding diffusion length L_{p1} will be much smaller.

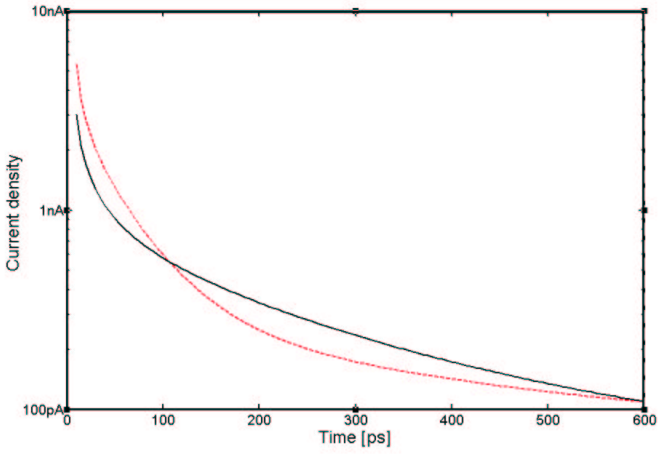

Fig. 4. Comparison of the overall current density impulse response for the the photodiodes with the width of the pitch=3μm (dashed-lines) and 10μm (full-line), respectively

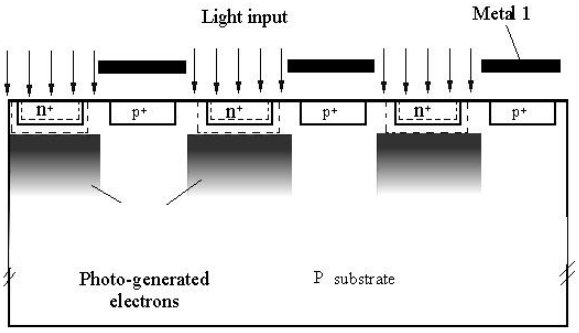

Fig. 5. N+/P-substrate photodiode in CMOS technology

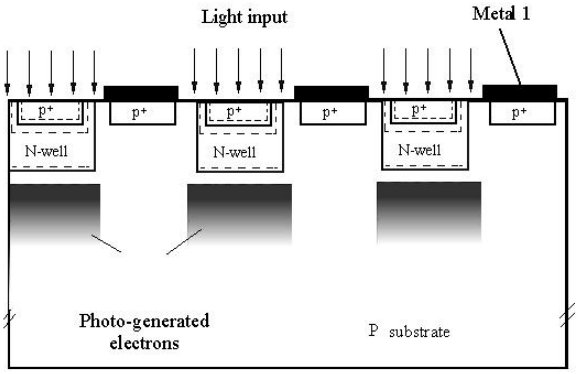

Fig. 6. P+/N-well/P-substrate photodiode in CMOS technology

The size of the N^+ diffusion layer towards the substrate l_{x1} is also lower. The maximum frequency response is determined mainly by this depth. In comparison with the N-well region, the speed of the diffusion holes will be increased while the contribution to the overall current response is decreased.

IV. Two-dimensional response of the P+/N-well/P-substrate diode (Double photo-diode)

The maximum frequency response of the excess holes inside N-well can be increased by using shallow P^+ diffusion region inside N-well. The diffusion path of the holes is than half of depth between lower side of the depletion region towards P^+ and the upper side of the depletion region towards P-substrate. This distance is very small and the hole response current will be very fast (several GHz range). The double-photodiode is shown in Figure 6.

A. Two-dimensional theoretical solution of the carrier profile inside N-well

The carrier density distribution function is calculated as a product of two Fourier series in x and y-directions. The boundary conditions for all four sides inside N-well region is zero since it is framed with depletion regions. The total contributed current is the integral of the current through the two side walls and the top and bottom layer and is calculated as:

$$\frac{J_{N-well_1}}{\Phi_0(j\omega)} = 64\frac{qL_p^2\alpha(e^{-\alpha l_{px}}-e^{-\alpha l_{x2}})}{l\pi^2 l_{x2}} \sum_{n=1}^{\infty}\sum_{m=1}^{\infty}\frac{\frac{l_{x2}}{l_y}\left(\frac{1}{2n-1}\right)^2+\frac{l_y}{l_{x2}}\left(\frac{1}{2m-1}\right)^2}{\left(\frac{(2n-1)\pi L_p}{l_{x2}}\right)^2+\left(\frac{(2m-1)\pi L_p}{l_y}\right)^2+1+j\omega\tau_p} \tag{8}$$

where l_{px} is the distance between the surface and the upper side of the first depletion region (depth of the P+ region), l_{x2} is the distance between the lower side of the first depletion region to the upper side of the second depletion region (depth of the rest of the N-well region).

In comparison with the N-well/P-substrate diode, double photodiode has *lowest diffusion paths* in the N-well region. There is the same overall photocurrent contribution, but with a larger maximum operating frequency.

Inside P^+ region, diffusion electrons choose minimal path towards upper side of the junction with N-well. Since this depth is very small, the electron current response in this region will be very fast too. The expression for the P+ diffusion current response component can be obtained by just putting the electron diffusion length and electron diffusion constant L_n and D_n instead of L_p and D_p in the equation (5).

Inside double-photodiode there is one more depletion region, in which light generated carriers can be swept out and contribute to the very fast depletion region current component.

The value of the depletion region currents can be presented as:

$$J_{DR} = q\alpha(W_{DR1}+W_{DR2})\frac{A_{\text{eff}}}{A_{\text{total}}} \tag{9}$$

where W_{DR1} and W_{DR2} are depletion region lengths of the first and second depletion region, respectively.

The overall current response is the sum of the two minority carrier currents and two depletion region currents.

V. Lateral N-well/P-substrate (exploiting only depletion region in between))

On the lateral side between N-wells and P-substrate, there is a depletion region with a thickness depending on

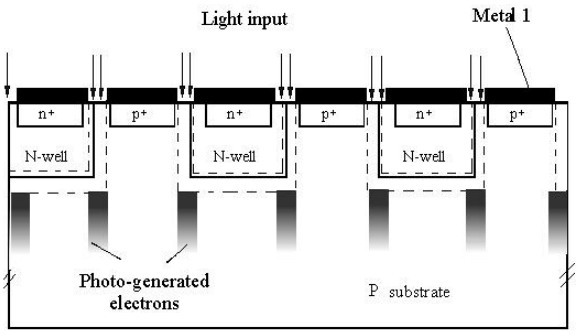

Fig. 7. Lateral N-well/P-substrate photodiode in CMOS technology

the well's doping concentrations. If light pulses incidently come into this region only, the slow diffusion inside N-wells is eliminated, since the light is shined in depletion region only. The current component contribution of the region above P-substrate have thus, flat frequency response.

One would immediately notice that a depletion region area is small compared to the total pitch area which as a consequence has low external quantum efficiency component i.e. efficiency of light transmission to the detector (fraction of incident photons that reach the silicon surface). However, the layout of the photodetector should be designed in a way to increase the effective (light sensitive) area of the detector.

Lateral N-well/P-substrate photodetector is shown in figure 4.

The width of the depletion region inside N-well will be much smaller compared to the width in the substrate so it will be neglected. The depletion region with is taken to be $l_N/h, \quad h = 2, 3, ...$ Inside P-region between the N-wells there are two of these depletion regions from the both sides of the N-well.

A. Current density response of the depletion region

The current response of this region is directly proportional to a size of the region and can be presented as:

$$J_{DR3} = q\alpha l_x \frac{A_{\text{eff}}}{A_{\text{total}}} \tag{10}$$

where A_{eff} and A_{eff} are effective and total area of the photodetector.

B. Two-dimensional theoretical solution of the carrier profile inside P-substrate

From Figure 4. it is clear that the photogenerated carrier profile $g(t, y)$ is a periodic function of l, as well as a carrier distribution function n_p. Thus, we can again expand them both as a Fourier series with the periodicity of the pitch. The zero order solution corresponds to $2l_N/(lh)$ which is due to a partial metal coverage of the detector.

$$\frac{J_0(j\omega)}{\Phi_0(j\omega)} = q\alpha L_n e^{-\alpha l_x} \frac{2l_N}{lh} \frac{1}{\sqrt{1 + j\omega\tau_n} + \alpha L_n} \tag{11}$$

For the higher order solutions the substrate current density is a sum of the currents bellow space-charge region plane:

$$\frac{J_{sub_1}}{\Phi_0(j\omega)} = q\alpha L_n(1-\xi)e^{-\alpha l_x} \sum_{n=1}^{\infty} \frac{1+\cos\left(\frac{2n\pi(h-1)l_N}{lh}\right)-\cos\left(\frac{2n\pi l_N}{lh}\right)-\cos\left(\frac{2n\pi l_N}{l}\right)}{\pi^2 n^2 \sqrt{\left(\frac{2n\pi L_n}{l}\right)^2+1+j*\omega\tau_n}+\alpha L_n} \tag{12}$$

If we take $\xi = 1$ the sum of the first two current components will be zero since the higher order solutions are all even and do not contribute to this sum [1]. The slowest and also the dominant contribution of the response is for $n = 1$. The amplitude of the other contribution decreases quadratic with n.

VI. Conclusion

An analytic current density profile for the CMOS photodiodes that can be monolithically integrated inside optical receivers is presented. The performed calculation helps to better understand influence of the structure and geometry on the speed of the photodiode. It presents the way how the layout of the photodiode should be design in order to achieve best bandwidth performances for the same occupied chip area.

The width of the N-wells and the distance between them should be less than two times the depth of the wells, in order to achieve maximum photodiode speed. Thus, the sizes should be kept as minimal as the technology limitation. Increasing the size of the pitch will decrease "intrinsic" bandwidth of the photodiode, since the speed of the hole diffusion inside N-wells remain the same but depletion region area is decreased in comparison with the overall pitch area and its contribution to the overall photocurrent decrease. The speed of the electrons inside P-substrate remains the same no matter the pitch area, but the longer the pitch, less electrons will contribute to the photocurrent (reaching P-metal contact). As far as the structure of

the photodiode is concerned the best speed performance was achieved using a P+/N-well/P-substrate photodiode (double-photodiode).

Finally, the slow substrate diffusion current has largest contribution in the overall photocurrent in all types of diodes thus, mainly determining the ”internal” speed of the detector. For the overall photodiode speed the capacitance should also be taken into account, since it determines the ”electrical” photodiode speed. With the further gate-length scaling in CMOS technology, the N-well depth will decrease, and the influence of the substrate current will dominate even more.

However, there is more and more effort put by scientist in finding ”smart solutions” for excluding the effect of the carriers inside substrate which will make feasible CMOS photodiode applications far in GHz range.

VII. References

[1] J. Genoe, D. Coppée, J. H. Stiens, R. A. Vounckx and M. Kuijk: ' *Calculation of the current response of the spatially modulated light CMOS detector*', IEEE Trans. Electr. Devices, vol. 48, No. 9, pp. 1892-1902, 2001.

[2] S. M. Sze: '*Physics of semiconductor devices*', New York: Wiley-Interscience, 2-nd edition, p. 81, 1981.

[3] I. Brouk, and Y. Neimirowsky,:'Dimensional effects in CMOS photodiodes', Solid-State Electronics, January 2002, 46, pp. 19-28

[4] L. D. Edmonds:'A Time-Dependent Charge- Collection Efficiency for Diffusion', IEEE Trans. on Nuclear Science, October 2001, vol. 48, pp.1609-1622