Master degree thesis
Nanotechnologies for ICTs

3D multiphysics transient modeling of vertical Ge-on-Si *pin* waveguide photodetectors

Paolo Franco
* * * * **

Supervisors
Prof. Michele Goano, Supervisor
Mr. Matteo G. C. Alasio, Co-supervisor
Dr. Alberto Tibaldi, Co-supervisor
Prof. Francesco Bertazzi, Co-supervisor
Prof. Giovanni Ghione, Co-supervisor
Politecnico di Torino
October 17, 2022

Abstract

We present a methodological contribution to the modeling of waveguide photodetectors for silicon photonics. This technology specializes in making silicon-compatible photonic devices and thus takes advantage from state-of-the-art MOS technology. After a general introduction to photodetectors, the thesis presents simulation results obtained with an approach not conventionally applied to these devices, validated against experimental measurements provided by the group's research partner Cisco System. We use commercial 3D multiphysics simulation tools to solve first the optical problem (with FDTD) and then the electrical problem (with drift-diffusion). In order to assess the dynamic response of the devices, rather than using a conventional small-signal approach, we have performed transient simulations taking into account the temporal variations of the optical signal. From these simulations it is possible to directly extract all the figures of merit of the device and other relevant information that can be exploited for designing the next generation of Ge-on-Si photodetectors for silicon photonics.

Summary

Since the foundation of modern communication systems, there has been a dramatic increase in data exchange, which continues to grow along with the complexity and quality of the services provided to users around the world. This is made possible by the means of a paradigm whereby the calculations required to operate these services have been centralized in data centers. Here, information is elaborated and then delivered to the final user. A crucial part of this infrastructure is the realization of low-power high-speed short-range communications. In this context, Silicon Photonics (SiPh) appears as the ideal platform, bridging the bit rate of optical communications and providing synergistic, low-cost, and fully CMOS-compatible integration of optical and electronic systems. Aiming to keep up with such an unstoppable internet traffic growth requires optimizing each block of the optical transmission chain. The core of this thesis is on the receiver block, which is one of the current bottlenecks for increasing the performance of optical links. In particular, it presents the study and the simulations of photodetectors with a CAD (computer aided design) approach, trying to develop a model capable of understanding and predicting the operation of the device. The simulative approach lowers both production costs, by reducing prototyping, and specific critical issues found by the simulations themselves. The difficulty lies in the modeling of the device, which inherently needs a multiphysics model where both the optical and the electrical problem are taken into account. This method involves an initial solution to the optical problem, providing the spatial distribution of the charges photogenerated by the device illumination, and then the use of this term as a source into the transport (electrical) problem. Here, through the use of the Poisson equation coupled in a self-consistent way with the continuity equations of the carriers and with the constitutive relations of drift-diffusion for the current densities of electrons and holes, the transport problem is solved. It can be then studied either under stationary conditions, small signal conditions, or with a transient simulation that also considers temporal variations of the applied signal. The latter solution is a novel one and it is not present in the state-of-the-art for the simulation of devices of this type. It allows us to extract all the quantities and figures of merit of interest with a single simulation, as well as to understand the temporal behavior of the device,

although it has a higher computational cost, and thus further improve and optimize the design of future generations of devices.

The thesis is organized as follows:

- The first chapter is an introduction on Silicon Photonics and on its growing importance in the perspective of optical telecommunications. A fundamental building block for the simulations is the theory of semiconductors. This part run trough equilibrium and out-of-equilibrium semiconductors to understand the description of the drift-diffusion model. There is then a brief description of the generation-recombination events to better appreciate the dark current and how those events influence the response of the devices. The last part of the introduction is the methodology of the simulations. The focus is on two main points: FDTD solution of the optical problem and Sentaurus simulation tools. A particular focus is given to the transient analysis: from the solution of the optical problem to the transient loop for the transport equations.

- The second chapter concentrates on photodetectors theory, explaining the different relations that hold and the different quantities which give a quality analysis of the devices. The discussion continues with a general description of the dynamic pin model and the main limitations of the detector cutoff frequency: the transit time and the capacitance effects. Finally, it ends with the description and characteristics of the materials adopted, the configuration, and the explanation of how the light coming from the waveguide enters the detector.

- The third chapter is the one describing the results. It introduces the structure of the detectors under study and their two-dimensional counterparts. After a brief description of the figures of merit, the core part of the results is presented: starting from the 2D transient simulations, we have gone through the study of the tool and of the break criteria to finally reach a satisfying transient simulation performed on the 3D devices. The results are compared to both small-signal analysis and measurements given by an industrial partner (Cisco Systems). The discussion continues on high optical input power cases, in which we can exploit the transient analysis. In conclusion, it is presented the exploration of some parameters, whose influence on device operation can be studied more in detail with a transient analysis.

Contents

1 Introduction **5**
- 1.1 Research context . . . 5
- 1.2 Theory of semiconductors . . . 6
  - 1.2.1 Out of equilibrium . . . 9
  - 1.2.2 Drift-Diffusion . . . 11
  - 1.2.3 Generation-Recombination . . . 12
- 1.3 Multiphysics modeling . . . 16
  - 1.3.1 FDTD . . . 16
  - 1.3.2 Synopsys TCAD Sentaurus . . . 18
  - 1.3.3 Quasistationary . . . 18
  - 1.3.4 ACCoupled . . . 19
- 1.4 Transient Analysis . . . 20
  - 1.4.1 Methodology . . . 20
  - 1.4.2 Flow Chart . . . 20
  - 1.4.3 Solution of the optical problem: FDTD . . . 21
  - 1.4.4 Properties of the optical input pulse . . . 21
  - 1.4.5 Generation term in the drift-diffusion model and explanation of the transient loop . . . 23
  - 1.4.6 Discretization method . . . 24

2 Photodetectors **27**
- 2.1 Parameters and relations . . . 32
  - 2.1.1 Responsivity . . . 33
  - 2.1.2 Electrical bandwidth . . . 37
  - 2.1.3 Frequency response in pin . . . 39
- 2.2 Dynamic pin model . . . 40
  - 2.2.1 Transit time limitation . . . 42
  - 2.2.2 Capacitance effect . . . 46
- 2.3 Bandwidth-efficiency trade-off . . . 46
- 2.4 Si-Ge photodiode, the choice of the materials and the heterostructure . . . 48

3 Results **53**
3.1 Structure 53
3.2 Figures of merit and preliminary analysis 55
3.3 Transient analysis 62
3.3.1 2D 63
3.3.2 Break criteria 65
3.3.3 3D results 66
3.3.4 Fourier Transform and small-signal comparison 68
3.4 High optical input power 72
3.5 Optimization and exploration 77
3.5.1 Velocity saturation 77
3.5.2 Ge doping profile 82
3.5.3 Mobility 85

4 Conclusions **92**

Chapter 1

Introduction

1.1 Research context

Silicon photonics is one of the most attractive and interesting branches of optoelectronics and it is arousing interest in the communication world [1]. Nowadays, the challenging aim is to increase the speed of the transmission of data, especially in close-distance systems, because the speed of interconnections is crucial for interconnected and complex system on which new technologies relays on. In many different cases this condition is the uprising reason of bottleneck problems. In this work, what we are trying to achieve is a higher and higher bandwidth of one of the components present in the interconnection chain, the *photodetector*. To make devices as fast as possible and to do this, it is important to study those devices from a physical point of view, in such a way that it is possible to optimise and solve device level issues that affects the overall performance of the system. This approach based on physics and simulations is preferred to a direct approach where measurements and fabricated device are involved. The simulation part is an important part of industry workflow and it is needed to save both money and time. Even more importantly it can be used to fully understand the the behaviour of the devices to upgrade them to the next generation and to exploit it in our favour [2]. We do not have to confuse simulation as a theoretical approach, but as a predictive approach. In fact, being able to simulate the device means being able to interpret the results, and at the same time design and predict the capability of not-yet-fabricated devices.

At the system level the optical interconnects have to be both low-cost and efficient, and germanium photodetectors are engaging devices for this applications due to the heteroepitaxy of germanium on silicon, making them CMOS-compatible. As we are going to see, germanium is an almost direct-gap semiconductor and it suits the use at standard telecom wavelengths such as 1310 nm and 1550 nm. The epitaxy consists of a low-temperature procedure to achieve a Ge buffer layer of few nanometers over Silicon and

then a high-temperature one, which let us complete the structure with a high-quality layer [3]. The simulations of such devices are complex, due to the hetero-interface modeling challenge present between the Ge absorber and the silicon substrate [4]. Another complexity level is derivated by the coupling of the light with the detector. In fact, this technology typically takes advantages of an evanescent coupled waveguide, requiring a solution of the complete Maxwell's equations.

In this thesis, first we start by describing the device and the modeling challenges, with an introduction to the simulation software to be able to understand next what we are going to do. Then, we report the methodology used and the type of simulation we choose for the description of the device. Finally, the results that were obtained, with remarks on the state-of-the-art devices and comparisons with measurements provided by the group's partner Cisco System.

1.2 Theory of semiconductors

Once we have in mind the description of the bands, we can define the carriers as the electrons promoted in the conduction band and the holes as the ones present in the valence one due to the voids left from the promoted electrons. The representation of the holes as an electron counterpart (consequent positive charge due to the lack of a negative one and its compensation to have an overall equilibrium) can be extremely useful and can make the description of what happens in a device quite easier. Both the carriers interacts with the external electric field, photons, phonons and other particles or defects. The two populations depend on the number of states per unit volume in both bands and on the energy dependence of how these states are populated. In an equilibrium situation, we can say that the carriers follow a Fermi-Dirac distribution, while we can adjust it in case of applied bias with a quasi-Fermi distribution. The Fermi-Dirac distribution describes both carriers equilibrium occupation statistics in relation to the Fermi level E_F, which is the energy value at which the electrons fill the bands:

$$f_n(E) = \frac{1}{1 + \exp\left(\frac{E - E_F}{k_B T}\right)}, \quad f_h(E) = \frac{1}{1 + \exp\left(\frac{E_F - E}{k_B T}\right)} \tag{1.1}$$

This value is in general in a semiconductor constant in the whole system as far as we assume an equilibrium case and it is in the energy gap between the valence and conduction bands if we are dealing with a nondegenerate semiconductor. A nondegenerate semiconductor has a Fermi level in the band gap, between the valence and the conduction bands. In case of a degenerate case, this energy level take place in one of the two bands. In this case, the condition for the approximation of the Fermi-Dirac statistic in the Boltzmann one is not valid anymore. We have the necessity to use the full

Fermi-Dirac statistics. In this specific case the Fermi-Dirac distribution can be approximated with the Boltzmann distribution:

$$f_n(E) \underset{E \gg E_F}{\approx} \exp\left(\frac{E_F - E}{k_B T}\right), \quad f_h(E) \underset{E \ll E_F}{\approx} \exp\left(\frac{E - E_F}{k_B T}\right). \tag{1.2}$$

In the other case, the degenerate one, the Fermi level falls in one of the two bands and the condition for the approximation is not anymore met: in those specific cases we have to deal with the full Fermi-Dirac statistic. While we assume the effective mass approximation, the density of states (DOS) in a bulk or 3D material can be written as:

$$\begin{aligned} N_c(E) \equiv g_c(E) &= \frac{4\pi}{h^3}\left(2m^*_{n,D}\right)^{3/2}\sqrt{E - E_c} \\ N_v(E) \equiv g_v(E) &= \frac{4\pi}{h^3}\left(2m^*_{h,D}\right)^{3/2}\sqrt{E_v - E} \end{aligned} \tag{1.3}$$

The valence band DOS is larger than the conduction one due to the contribution of heavy holes to its value. Now we have everything in order to calculate the number of carriers by integrating the multiplication of the density of states and the statistical distribution (which is the Fermi-Dirac one in general or we can use the Boltzmann one in the nondegenerate case) over all the energies of the conduction band for what concern the electrons and of the valence one for the holes.

$$\begin{aligned} n &= \int_{E_c}^{\infty} N_c(E) f_n(E) \mathrm{d}E = N_c \exp\left(\frac{E_F - E_c}{k_B T}\right) \\ p &= \int_{-\infty}^{E_v} N_v(E) f_h(E) \mathrm{d}E = N_v \exp\left(\frac{E_v - E_F}{k_B T}\right), \end{aligned} \tag{1.4}$$

where we still have to explicitly write the effective densities which shows up in the formula:

$$N_c = 2\frac{\left(2\pi m^*_{n,D} k_B T\right)^{3/2}}{h^3}, \quad N_v = 2\frac{\left(2\pi m^*_{h,D} k_B T\right)^{3/2}}{h^3}. \tag{1.5}$$

In the most basic case, the intrinsic one, we are not implanting or diffusing any kind of doping and the concentration of the intrinsic number of carriers is equal to both the electron and hole population and it can be calculated by:

$$n_i = N_c \exp\left(\frac{E_{Fi} - E_c}{k_B T}\right) = p_i = N_v \exp\left(\frac{E_v - E_{Fi}}{k_B T}\right), \tag{1.6}$$

From there, thanks to the fact that we are calculating the number of carriers on the intrinsic Fermi level, we can derive it by equating the intrinsic formulas passing through an electron or an hole point of view.

$$E_{Fi} = k_B T \log\sqrt{\frac{N_c}{N_v}} + \frac{E_c + E_v}{2} \tag{1.7}$$

On the other hand, if we do not equate these two formulas, but we multiply them, the multiplication of the exponential means the addition of the exponent with the respective signs (negative for the division) and this lead to the well-known formula which relates the number of intrinsic carriers to the energy gap of the specific semiconductor.

$$n_i p_i = n_i^2 = N_c N_v \exp\left(-\frac{E_g}{k_B T}\right) \tag{1.8}$$

As we can see from that, the intrinsic concentration of carriers is hugely dependent on the temperature, which explains why at very high T doping becomes no more important. This compromises a lot of working of the device, altering for example the ohmic contacts. Other problems that may show up are for example the channels that need to be almost intrinsic in order to have good conduction or in the case of the main topic of this thesis the photodetectors: the germanium detector. Its intrinsicity is necessary to have a large absorption region and as much as possible a low number of trap states due to the doping, so we have less recombination as possible and the absorption of most photons for the creation of electron-hole pairs. We can also obtain the mass action law, which tells us that in an equilibrium condition, the product of the electrons and the holes is, by definition, equal to the square of the intrinsic number of carriers.

$$np = n_i^2 \tag{1.9}$$

We can now move on to the case of doped semiconductors, the mass action law is valid also in this case, but the electron and hole number are not anymore equal to the square of the intrinsic one, but this has to deal now with the number of doping we are using. The doping can be done in two ways: through a donor with density N_D which provides an additional electron to the conduction band when it substitutes the original semiconductor or through an acceptor element with density N_A which attracts an electron from the valence band leaving a hole. Some examples of those two types are As and P as donors and B as acceptors. The ionization energy in both cases is on the order of 10meV and they can be ionized leading to an additional electron or hole, respectively, quite easily, increasing the carriers that participate to conduction. If we assume the complete ionization of each doping atom, so we are in a full ionization regime and we remember that the mass action law is still valid, we can simply derive the new expressions and numbers for the electrons and holes populations.

$$\begin{aligned} n \approx N_D^+ \approx N_D, \quad p \approx n_i^2/N_D \quad & n\text{-type semiconductor} \\ p \approx N_A^- \approx N_A, \quad n \approx n_i^2/N_A \quad & p\text{-type semiconductor.} \end{aligned} \tag{1.10}$$

The influence of the temperature on the number of carriers is now not anymore related simply to the number of intrinsic carriers, which depends exponentially from it, but we have to take into account also the ionization energy

necessary for the doping to be full ionized. At very low temperatures, the thermal energy is not enough for a good ionization process and the carrier population decreases with T well below the N_D value (freeze-out range). The next range is the one which starts at $(3/2)k_BT \approx 20$meV with a temperature of 150 K and it ends when the number of intrinsic carriers equals the number of dopants we have in our semiconductor. This is called saturation range and in that we have something really close to the full ionization: electrons equal the dopants number in case of N_D or hole equal the dopants number in case of N_A type of doping. The last range is the one with temperatures even higher than the last mentioned. The intrinsic concentration of carriers flood the semiconductor with carriers not anymore due to the ionization of the doping agents, being their number negligible. In the case of a donor doping we can call the semiconductor of n-type, because it will have more electrons than an intrinsic one: its Fermi level (which as we have already said, tells us in part the filling of the electrons in the bands) will be higher in energy than the intrinsic one, so closer to the conduction band edge. The exact opposite is going to happen in case of an acceptors doping and consequently a p-type semiconductor, with the Fermi level closer to the valence band maximum. The full ionization condition can be satisfied only in case of relatively low levels of doping: in case of 10^{19} cm^{-3} this assumption is not anymore valid and the ionization cannot be assumed as 100%, but the number is related to the very position of the Fermi level and it decreases when the Fermi level is larger than the donor energy level or smaller than the acceptor's one. It is even more extreme the same discussion about degenerate semiconductors, but we have to remember than it is almost impossible than the degeneration is caused merely by the doping, being it a condition that can be met only in particular cases (for example in a direct-bias condition under high carrier injection).

1.2.1 Out of equilibrium

We can now move on to a non-equilibrium description of the semiconductor. In equilibrium the electron and hole populations follow the mass-action law and any difference is compensated by some kind of generation or recombination process. The difference between the number of the carriers and their number at equilibrium follows an exponential decay law:

$$n'(t) = n'(0) \exp\left(-t/\tau_n\right) \tag{1.11}$$

with a characteristic lifetime that can go from milliseconds to nanoseconds. In the case of an external cause or by an external electric field, the condition is characterized by a different form of the probability distribution which is not anymore the Fermi-Dirac one. In the second case, there is an increase in the average velocity and a non symmetrical velocity distribution concerning the origin. In all our discussions, the electric field will never be such dramatic

as to make the quasi-Fermi approach not consistent. The first thing to do in order to describe the disequilibrium of carriers with respect to equilibrium is to introduce the quasi-Fermi levels E_{Fn} and E_{Fh} modifying the Fermi-Dirac ones. These also change the Boltzmann approximation when it is valid:

$$\begin{aligned} f_n\left(E, E_{Fn}\right) &= \frac{1}{1+\exp\left(\frac{E-E_{Fn}}{k_BT}\right)} \underset{E\gg E_{Fn}}{\approx} \exp\left(\frac{E_{Fn}-E}{k_BT}\right) \\ f_h\left(E, E_{Fh}\right) &= \frac{1}{1+\exp\left(\frac{E_{Fh}-E}{k_BT}\right)} \underset{E\ll E_{Fh}}{\approx} \exp\left(\frac{E-E_{Fh}}{k_BT}\right), \end{aligned} \tag{1.12}$$

While the Boltzmann equations holds, the carrier density expressions become:

$$n = N_c \exp\left(\frac{E_{Fn}-E_c}{k_BT}\right), \quad p = N_v \exp\left(\frac{E_v-E_{Fh}}{k_BT}\right) \tag{1.13}$$

Thereafter we can derive another time the mass action law and, as we can see, there is an extra term which includes the difference between the two quasi-Fermi levels:

$$np = n_i^2 \exp\left(\frac{E_{Fn}-E_{Fh}}{k_BT}\right) \tag{1.14}$$

This is an even more general formula: if we assume back the equilibrium condition in which the Fermi level is unique, we simply restore the first formulation we have given of the mass action law. In this case we can distinguish two different conditions: the first one is when E_{Fn} is higher than E_{Fh} and the exponential has an overall plus sign, the result is that $np > n_i^2$ and we are in a carrier injection condition. On the other hand, when E_{Fn} is lower than E_{Fh}, we have a negative sign as exponent and the exponential will dramatically reduce the np product, making it smaller than n_i^2 $np < n_i^2$: this condition is called carrier depletion. In this case, if we are talking about degeneration we have to be a little more careful, by using the Fermi-Dirac integrals to write down the charge density.

$$n = \frac{2}{\sqrt{\pi}} N_c \mathcal{F}_{1/2}\left(\frac{E_{Fn}-E_c}{k_BT}\right), \quad p = \frac{2}{\sqrt{\pi}} N_v \mathcal{F}_{1/2}\left(\frac{E_v-E_{Fh}}{k_BT}\right). \tag{1.15}$$

There are various approximations such as the Joyce-Dixon inverse formula:

$$\begin{aligned} E_{Fn} &\approx E_c + k_BT\left[\log\frac{n}{N_c} + \frac{1}{\sqrt{8}}\frac{n}{N_c}\right] \\ E_{Fh} &\approx E_v - k_BT\left[\log\frac{p}{N_v} + \frac{1}{\sqrt{8}}\frac{p}{N_v}\right] \end{aligned} \tag{1.16}$$

1.2.2 Drift-Diffusion

We can now describe more in detail the case in which there is the application of an external electric field that we can expect will increase the average velocity of the carriers. We cannot underestimate the velocity in equilibrium condition, because the average kinetic energy is 39meV at 300 K. This is quite high, on the order of $v_{\text{ave}} \approx 1 \times 10^7 \,\text{cm}\,\text{s}^{-1}$ even if the distribution of the velocities is symmetric respect to the origin (the average velocity is equal to 0). In the case of an applied electric field, we can explicit the proportionality of the velocity with the electric field, which is called mobility, and these are measured in $\text{cm}^2\ \text{V}^{-1}\ \text{s}^{-1}$:

$$\underline{v}_{n,\text{ ave}} = -\mu_n \underline{\mathcal{E}}, \quad \underline{v}_{h,\text{ ave}} = \mu_h \underline{\mathcal{E}}, \tag{1.17}$$

These mobilities are a good approximation of the linear proportionality between velocity and electric field when we are in a low-field regime, as we are exploiting the linear part of the curve. This mobility depends on many objects that can influence the movement of the particles due to many kinds of scattering. The carriers can interact with lattice vibrations, which are called phonons, impurities (due to a not perfect semiconductor or to the doping), lattice mismatch for the heterostructures between two different semiconductors, etc. The mobility will also decrease with increasing doping because of the increased number of scattering centers. The same happens with an increasing temperature: the higher energy that the whole lattice has, there is increase in the creation of phonons. We can simplify their description as the vibration of the lattice: as we know an increased temperature causes an increase of its vibration. All what we have said by now is true in a low field condition, but when we increase it to values higher than 10kV/cm, we can see a saturation of the average velocity $v_{n,\text{ ave}} \to v_{n,\text{ sat}}$, $v_{h,\text{ ave}} \to v_{h,\text{ sat}}$. These saturated values are around 10^7 cm/s. The movement of carriers due to the electric field is called drift motion and this is called the drift component of the current density if we express it in the form of a current:

$$\begin{aligned} \underline{J}_{n,\text{dr}} &= -qn\underline{v}_{n,\text{ ave}} = qn\mu_n\underline{\mathcal{E}} \\ \underline{J}_{h,\text{dr}} &= qp\underline{v}_{h,\text{ ave}} = qp\mu_h\underline{\mathcal{E}}. \end{aligned} \tag{1.18}$$

The electron velocity-field curve is non-monotonic for every different semiconductor. It increases for low electric fields because electrons are mainly in the Gamma minimum of the conduction band in case of direct gap semiconductors, but when we are dealing with higher field values, electrons are scattered in the indirect-bandgap minima, with consequent lower velocity, which reduce the overall velocity of electrons. As an example we can mention the GaAs that have the maximum and the consequent beginning of the decrease in the correspondence of 3kV/cm which has a difference of 300meV between the two different gaps, while the InP has the maximum close to

higher fields accordingly to the higher energy different between direct and indirect gaps. In general the mobility of the holes is never better than the electron's one, so n-type transistors are in general more used in high speed applications. But carriers show up another cause of motion aside than the electric field which is present in the depletion region or by the application of an external one, which is the concentration gradient difference between different regions. This gives rise to the other component of the motion that corresponds to the diffusion current density:

$$\underline{J}_{n,\,\mathrm{d}} = qD_n\nabla n, \quad \underline{J}_{h,\,\mathrm{d}} = -qD_h\nabla p \tag{1.19}$$

The D_n and D_h are the diffusivity coefficients and those are related to the mobilities present in the drift part of the current through the Einsten's relation (it holds for both electrons and holes):

$$D_\alpha = (k_B T/q)\,\mu_\alpha \alpha = n, h. \tag{1.20}$$

Even if we have talked about velocity saturation as a consequence of scattering with some kind of particle, this is not the absolute rule and this is not what always happens. In very particular conditions, such as when the required movement of the carrier is on the order of picoseconds long or for very short distances way lower than micrometers, we are in a condition of ballistic transport. This means that we can assume there is no collision at all and consequently the average carrier velocity can reach values quite higher than the saturation one in the presence of a high electric field. This is what we call a velocity overshoot and this lead to an increase in speed of some kind of devices such as nanometer-gate FET. However, we have to say that in our transient analysis this could compromise the simulation and its convergence. The conditions themselves of the structure we analyse and the voltage we apply are far away enough to meet these conditions, which are meant to be used in other kind of devices where the saturation velocity is considered not good enough for their purpose.

1.2.3 Generation-Recombination

The net recombination rates take into account both the generation processes (electrons and holes generated per unit time and volume) and the recombination ones (carriers recombined per unit time and volume):

$$U_n = R_n - G_n, \quad U_h = R_h - G_h. \tag{1.21}$$

In some conditions the total rates of electrons and holes are equal: for example in DC stationary conditions or when the process is band-to-band with a direct transition between valence and conduction due to the creation of both an electron and a hole. The mechanisms can happen also through intermediate traps or recombination centers inside the band gap, working

only for one kind of carrier, making the instantaneous net recombination rates different in time-varying conditions. We know that the net rate should be zero at equilibrium, so we can write:

$$U_n = r_n \left(pn - n_i^2 \right) \tag{1.22}$$

where we can distinguish the generation associated to the n_i^2 term and the recombination associated to the pn one. The pn term is the recombination one because we are considering both carrier's populations densities in a sort of collision between the two. This term can be also expressed in lifetime terms: tau. If we define the electron density n, the excess lifetime can be derived by the rate equation:

$$\frac{\mathrm{d}n'}{\mathrm{d}t} \approx -\frac{n'}{\tau_n} \tag{1.23}$$

The lifetime of minority carriers is constant and the definition tells us that the population exponentially decreases with time:

$$n'(t) = n'(0) \exp\left(-t/\tau_n\right) \tag{1.24}$$

and can also be defined as the ratio between the average time of the carrier pair creation and their annihilation with a recombination event:

$$\langle t \rangle = \frac{\int_0^\infty t n'(t) \mathrm{d}t}{\int_0^\infty n'(t) \mathrm{d}t} = \frac{\int_0^\infty t \exp\left(-t/\tau_n\right) \mathrm{d}t}{\int_0^\infty \exp\left(-t/\tau_n\right) \mathrm{d}t} = \frac{\tau_n^2}{\tau_n} = \tau_n. \tag{1.25}$$

The GR mechanisms can be both direct, so interband transitions or indirect, assisted by trap levels present in the gap. In direct gap semiconductors, the main event is due to optical (radiative) mechanisms, while in the indirect ones, the trap-assisted mechanism is the main one. As we have already described in the drift-diffusion formula, if we are in a case of 0 electric field, then the spatial evolution of the excess carrier densities is dominated by diffusion movement and by generation-recombination events. The solution of the continuity equation in this case can be described through the carriers' lifetime with an exponential kind of solution:

$$n'(x) = A \exp\left(-x/L_n\right) + B \exp\left(x/L_n\right) \tag{1.26}$$

where the constants are derived by boundary condition and the terms at the denominator of the exponent are the diffusion lengths of the carriers, that can be written as:

$$L_\alpha = \sqrt{D_\alpha \tau_\alpha} \quad \alpha = n, h_\mathrm{a} \tag{1.27}$$

We can talk about some of these processes, for example, SRH recombination. We are assuming a trap state inside the band gap, where in an ideal case there should not be any kind of possible state. The assumption of states at

different energies from one of the valence and conduction bands complicates our discussion. We can consider a certain number of trap levels N_t at certain energy included in the band-gap E_t. In general, transitions (for example thermal ones) are quite easier, since two steps are required with a lower amount of energy each time. In the case of a not sufficient thermal energy for carrier pair creation between the two bands, it could be enough for the intermediate step of $\Delta E \approx E_g/2$ if the state is exactly in the middle of the gap. We are not writing and demonstrating the passages, but the net trap-assisted recombination rate is equal to:

$$U^{SRH} = \frac{np - n_i^2}{\tau_{h0}^{SRH}\,(n + n_1) + \tau_{n0}^{SRH}\,(p + p_1)} \tag{1.28}$$

where the lifetimes are written as:

$$\tau_{h0}^{SRH} = \frac{1}{r_{ch}^{SRH} N_t}, \quad \tau_{n0}^{SRH} = \frac{1}{r_{cn}^{SRH} N_t} \tag{1.29}$$

The r parameters are the trap capture coefficients respectively for electrons and holes and $n1$ and $p1$ are:

$$p_1 = n_i g \exp\left(\frac{E_{Fi} - E_t}{k_B T_0}\right), \quad n_1 = n_i \frac{1}{g} \exp\left(-\frac{E_{Fi} - E_t}{k_B T_0}\right) \tag{1.30}$$

g is the trap degeneracy factor and it has no dimensions, while $E_f i$ is another time the intrinsic Fermi level. We can consider a n-type doped semiconductor in low-injection condition: excess carriers are negligible respect to equilibrium ones $n' \ll N_D$ $p' \gg n_i^2/N_D$ The complete formula can be approximated as:

$$U^{SRH} \approx \frac{p'}{\tau_0^{SRH}} \left[1 + \frac{2n_i}{n} \cosh\left(\frac{E_{Fi} - E_t}{k_B T_0}\right)\right]^{-1}, \tag{1.31}$$

where

$$\tau_0^{SRH} = \frac{1}{r_c^{SRH} N_t}. \tag{1.32}$$

Starting from $n_i \ll n$, the minimum lifetime is obtained for a trap energy close to the middle of the gap. The lifetime is completely independent from the doping, while as we see in the formula, it depends on the number of traps. The derivation is the same in case of a p-doped semiconductor and the formula is identical, but referred to holes when we have something relative to electrons and vice versa. The name of these traps in the middle of the gap are called recombination centers and they drop the thermal lifetime respect to the case of an intrinsic material. We can derive a different formulation in case of a high-injection condition. In this case this is case a direct competitor even to the radiative recombination. With some simple calculations we can find

$r_c^{SRH} \approx 10^5$ m/s $\cdot\, 10^{-17}$ m^2 $= 10^{-12}$ m^3/s and $\tau_0^{SRH} = \frac{1}{10^{-12}.10^{20}} = 10$ ns, which is a value quite close to the one we obtained when we calculated the radiative lifetime. If these calculation were performed in silicon, we can find even higher numbers respect to radiative transitions in case of direct-gap semiconductors. We assume both quasi-neutrality $n \approx p$ and equal trap capture coefficients for electrons and holes $r_{ch}^{SRH} \approx r_{cn}^{SRH} = r_c^{SRH}$:

$$U^{SRH} = \frac{p}{2\tau_0^{SRH}} \left[1 + \frac{n_i}{n} \cosh\left(\frac{E_{Fi} - E_t}{k_B T_0} \right) \right]^{-1}. \tag{1.33}$$

The lifetime in high-injection is equal to:

$$\tau_{0hi}^{SRH} = 2\tau_0^{SRH} = \frac{2}{r_c^{SRH} N_t} \tag{1.34}$$

r parameters can be written as the product between the thermal velocity and the cross section:

$$r_{cn}^{SRH} = v_{th}\sigma_n, \quad r_{ch}^{SRH} = v_{th}\sigma_h. \tag{1.35}$$

The order of magnitude of the cross sections are 10^{-15} cm^2 for Silicon for both electrons and holes to $\sigma_n \approx 10^{-14}$ cm$^2, \sigma_h \approx 10^{-13}$ cm^2 for III-V materials and the thermal velocity on the order of $v_{th} \approx 10^7$ cm/s. In case the quasi-neutrality assumption is not valid, but at least $np \gg n_i^2$ and $E_t \approx E_{Fi}$ we have:

$$U^{SRH} \approx \frac{np}{\tau_0^{SRH}(n+p)} \tag{1.36}$$

Another recombination process is the Auger one, which uses one additional electron or one additional hole and the rate is proportional to $p^2 n$ or pn^2, meaning not only the colliding populations are involved, but also the type of carrier which supply the energy necessary for the recombination. Logically, this process, which needs more carriers and the right amount of energy to happen, is more significant in high-injection devices. The reflected process, from a generation point of view, is the impact ionization, which is important when we consider very high fields (on the order of 100kV/cm). Carriers in this case receive enough energy from the field to be able to ionize another carrier (for example another electron in the conduction band) after a first scattering event. In the time between two different scattering the field should be high enough to give to the carrier sufficient energy from the drift motion in order to impact another carrier, transferring to it enough energy for the ionization. This process happens for each carrier, which, in their motion range, can create a certain number of pairs. If the field is too high, this will diverge since each generated carrier will generate a high amount of carriers itself. This could cause many problems and it is called avalanche breakdown. This may be avoided by making the field always lower than the breakdown value, which value has an exponential dependence with the semiconductor

gap. This is why for high power applications or in harsh conditions, for example, it is better to use materials with a big gap such as SiC or GaN to avoid any kind of instant breakdown of our device. The description of this latter process can be done with the following formulas:

$$\nabla \cdot \underline{J}_n = -qG_n - qG_h, \quad \nabla \cdot \underline{J}_h = qG_n + qG_h, \tag{1.37}$$

where:

$$G_\alpha = \frac{1}{q}\alpha_\alpha(\mathcal{E})J_\alpha, \quad \alpha = n, h. \tag{1.38}$$

Impact coefficients are strongly dependent with the electric field and the values for electrons and holes are equal in some semiconductor such as GaAs and Ge, while the electron's value is higher in silicon for example.

1.3 Multiphysics modeling

In this section we are going to talk about the tools we use in the simulation. These are performed in order to obtain:

- the optical simulation of the evanescent coupling in the device;
- information for the preliminary analysis to obtain the main quantities of the device before the transient analysis.

1.3.1 FDTD

In Senturus we can use different optical solvers to obtain, as we are going to see, the evanescent coupling of our devices. The correct solver for the structure and the type of coupling is the finite-difference time domain. The Finite-Difference Time-Domain (FDTD) method is a rigorous and powerful tool for modeling nano-scale optical devices. FDTD solves Maxwell's equations directly without any physical approximation, and the maximum problem size is limited only by the computing power available. It is possible to use it in Synopsys TCAD Sentaurus (RSoft FullWave tool). What it can do is evaluate the optical field, and it then converts the field distribution into a generation rate of photons inside the different materials. But, how does this method work? We can start with the Maxwell equations in an isotropic medium:

$$\begin{aligned} \frac{\partial B}{\partial t} + \nabla \times E &= 0 \\ \frac{\partial D}{\partial t} - \nabla \times H &= J \\ B &= \mu H \\ D &= \epsilon E, \end{aligned} \tag{1.39}$$

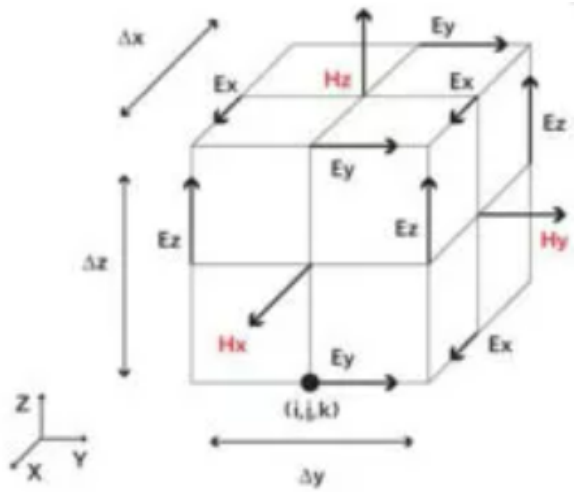


Figure 1.1: FDTD Yee Cell[5].

with J, μ, and ϵ functions of both space and time. The FDTD method solves them on a certain mesh and computes both E and H at grid points spaced δx, δy, and δz, with E and H interlaced in all three spatial dimensions as they are in the Maxwell equations. We can see in Fig. 1.1 the Yee cell of the dimensions we have written, with the presence of the field components: the E-components in the middle of the edges and the H-components in the center of the faces. The space grid size must be such that over one single increment, the electromagnetic field has no significant changes. To have relevant results, the linear dimension of the grid has to be a certain fraction of the wavelength. If we want to know the stability of the method, there is the requirement of satisfaction of a relationship between the space increment and the time increment δt. This is very difficult to obtain in case of the variable value of μ and ϵ, while in case these are constant, the computational stability requires:

$$\sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} > c\Delta t = \sqrt{\frac{1}{\epsilon\mu}}\Delta t \tag{1.40}$$

with c velocity of the light in the simulated medium. We can further enlarge the inequality by using c_{max} maximum light velocity in the analysed region:

$$\sqrt{(\Delta x)^2 + (\Delta y)^2 + (\Delta z)^2} > c_{\max}\Delta t \tag{1.41}$$

The choice of the points in which the method evaluates the field components is crucial, making the set of finite difference equations possible to solve and the solution satisfies the boundary conditions. FDTD includes many effects such as scattering, transmission, reflection, and absorption. This method follows a leapfrog scheme for going forward in the time direction:

the electric field updates are done in a time step between two consecutive time-steps in which the magnetic field is updated and vice versa. This avoids the simulator's need to solve more equations simultaneously, and it allows the propagation of a dissipation-free wave. The main problem of this scheme is the requirement of an upper bound on the time step to avoid numerical instability.

1.3.2 Synopsys TCAD Sentaurus

Technology Computer-Aided Design (TCAD Sentaurus) is an advanced device simulator developed by Synopsis which solves the electric transport of the device and it is used coupled with Synopsis RSoft FullaWAVE, which instead solves the electromagnetic (EM) problem.

Synopsis Sentaurus Workbench (SWB) tool is the graphical interface we can see with all the simulation in one single environment, or better, a table. It is possible to plan and make different simulations run consequentially. It is possible to add to the simulation a great number of tools in order to perform bot only TCAD simulations, but also RSoft, codes written in tcl, MATLAB script and others third-party tools, with the possibility to parameterize different values that are going to be used in the various simulations.

Therefore, the graphical used interface (GUI) simplifies the management of complex projects and let you parameterize input files or values to run more than one simulation automatically.

1.3.3 Quasistationary

We can briefly explain this kind of simulation because it is used in many part of the simulations for the analysis of the device before the transient one. This command ramps a device from one to another solution changing the boundary conditions or some kind of parameter we want to analyse. At each iteration the device is solved with a different parameter value and this is done until we reach the maximum value we have imposed for that particular parameter (ex. voltage applied to the device). Internally, the *Quasistationary* ramps a variable t from 0 to 1. The voltage at the contacts changes with the variable t through the formula $V = V_0 + t\,(V_1 - V_0)$, being V_0 the initial value of voltage and V_1 the final one we have imposed as the Goal of the simulation. All the control parameters are made on the t variable and the step control parameters are:

- MaxStep
- MinStep
- InitialStep
- Increment

- Decrement

The first two limit the step size and gives a range the simulation can use to solve the device. The InitialStep defines only the first one and the next step sizes are automatically increased or decreased depending on the rate of success of the inner solve command. This rate of increase is controlled solely by the number of performed Newton iterations and by the factor Increment, while the step size is reduced by Decrement when the inner solve fails. All the process aborts if the step becomes even smaller than the MinStep. In case of a simulation over the voltages with no illumination, we are able to obtain the dark current of the device by the IV characteristic, while if the goal is the OpticalGeneration rate defined as a constant or as a function of the power, the curve we are able to obtain is the photocurrent-power curve. These two simulations are functional to obtain the values of dark and photocurrent for the use of the break criteria for the optimization of the transient analysis.

1.3.4 ACCoupled

This command allows us to perform a small-signal AC analysis. This computes the frequency-dependent admittance matrix Y between circuit nodes of the specified electric system. For a given frequency v, it describes the small signal model in this way:

$$\delta I = Y \delta V \tag{1.42}$$

where δV and δI are the vectors of voltage and current excitations at selected nodes. The admittance matrix is defined as:

$$Y = A + i2\pi v\mathrm{C} \tag{1.43}$$

with A real conductance matrix and C capacitance matrix. As in the *Quasistationary*, we need to specify many parameters of the simulation:

- StartFrequency
- EndFrequency
- NumberOfPoints
- Linear
- Decade

With those, one is able to to select the frequencies of the simulation. Node is used to specifying the list of AC nodes considered in the admittance matrix. Exclude specify a list of system instances that should not be part of the AC system. Optical tells the simulation to consider the optical part of the

problem. In case of dark conditions, without the Optical command specified, we obtain the electric response of the device. What we have done is both this simulation and the one in which we do not consider the resistive load. In case we specify the Optical part, we obtain the electro-optic response of the device to our optical input and consequently its cutoff frequency.

1.4 Transient Analysis

1.4.1 Methodology

We have now all the concepts for what concern both the theory of the photodetectors (our devices in particular) and the working flow and method of Sentaurus [6]. We can move on to describe the main analysis of this thesis. We have to remember another time that everything was done firstly on a two-dimensional version of the reference project in order to have the possibility to perform more simulations in a reasonable time: hundreds of simulations on the three-dimensional would take too much time if we think that the time order of one simulation is between 30 and 40 hours. In the next part we are going to describe the flow of the transient, the basic equations and what it is based on.

1.4.2 Flow Chart

The transient simulation starts from the solution to the optical problem, through a finite-difference time-domain (FDTD) method. Its solution is then used as a pulsed generation term in the drift-diffusion equation and then solved in an iterative way for each time step until the final time you chose at the beginning of the simulation is reached or the break criteria is encountered (see Fig. 1.2). Thus, Sentaurus solves the transient problem by solving the device first, and then by increasing the time step in a loop and resolving it. The equations used for the transient can be written as a set of ordinary differential equations which can be divided into both DC and transient parts of the partial differential equations:

$$\frac{d}{dt}q(z(t)) + f(t, z(t)) = 0 \tag{1.44}$$

Composite trapezoidal rule / backward differentiation formula is the implicit method which is used for the discretization of the problem, which requires as input:

- the time interval
- the initial condition

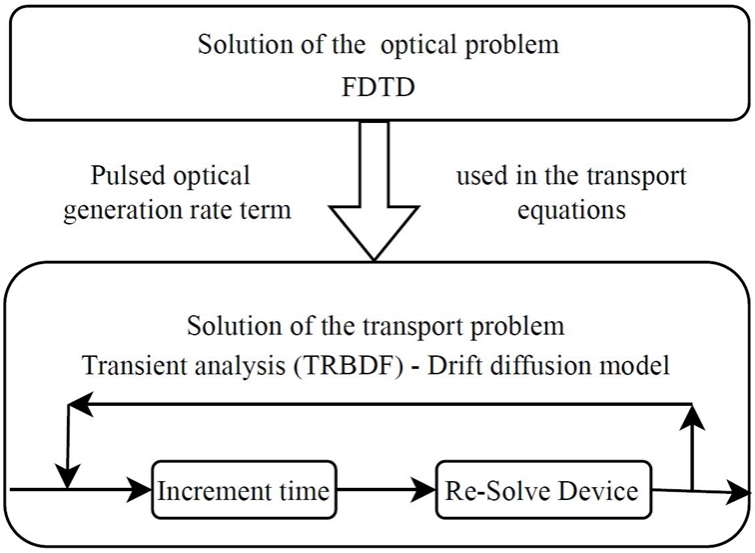


Figure 1.2: Flow chart of the transient simulation.

1.4.3 Solution of the optical problem: FDTD

We can start to say that in the two-dimensional case, we are not going to use Synopsis RSoft FullaWAVE using Yee method [5], but we are going to use a constant Optical Generation Rate. The value we are using is not a casual value, but it has been calculated with an average over the whole three-dimensional Germanium material(Fig. 1.3). As we are going to see in the results, this choice is quite good and comparable with the three-dimensional simulation. Another important approximation is the choice of the photons generated in the silicon part of the device, which reproduce the waveguide with the coupling to the germanium. The values in this part for what concern the optical generation must be for sure lower than the value used in germanium, in order to obtain a reasonable frequency response. Once we go below a value of 1e21 there is no more a great difference if we go further below, which corresponds exactly to the difference in the integral of the Optical Generation in the three-dimensional device. On the other hand, the input for the three-dimensional case is the OpticalGeneration coming from Synopsis RSoft FullaWAVE, in which we can see the resonance of the evanescent mode inside the germanium.

1.4.4 Properties of the optical input pulse

The model of the light signal in important in order to model the electro-optical response to a light impulse. It is necessary for Sentaurus to specify the time dependence, which scales the generation rate obtained by a stationary solution of the optical -problem- as a function of time. The dependency

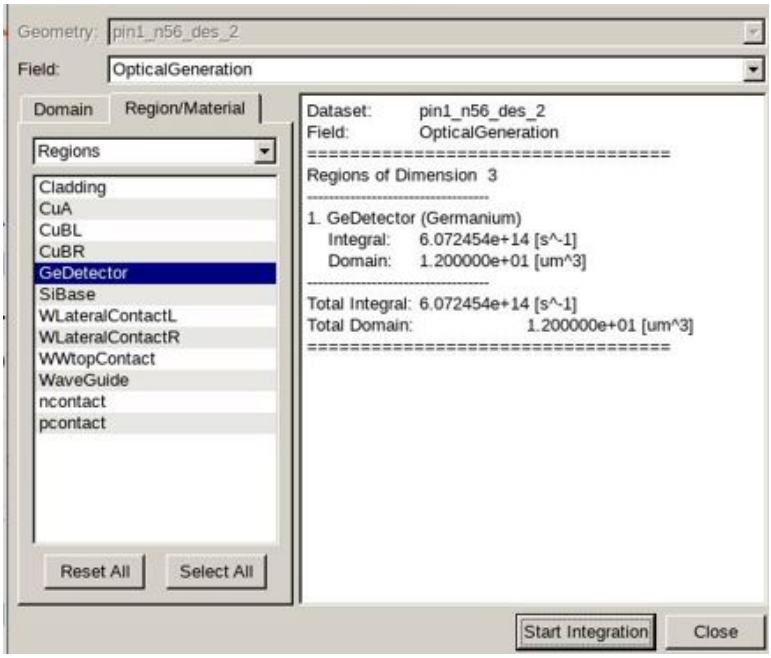


Figure 1.3: Integration of the optical generation over the whole Germanium.

we have used is the linear one, while we had the possibility to use also a Gaussian, exponential, cosine or arbitrary description of the pulse. We have to specify four different times in order to describe it: the first one t_0 tells us when the optical generation begin to rise and this happens until the second value t_1 is reached. At this point the value is kept constant at the relative scaling factor equal to 1 until the third time value t_2 is reached, which expresses the start of the falling edge. This part in which the generation rate is constant is described by $WaveTime = (t_1, t_2)$. This continue linearly until t_3 is reached and the optical impulse is another time equal to zero. The description can be expressed mathematically as the next formula:

$$F(t) = \begin{cases} \max\left(0, m\left(t - t_1\right) + 1\right) & t < t_1 \\ 1 & t_1 \leq t \leq t_2 \\ \max\left(0, m\left(t_2 - t\right) + 1\right) & t > t_2 \end{cases} \tag{1.45}$$

As we can see in this formula, m is a parameter we have not introduced by now: this is the slope of the rising edge and it is described in Sentuarus as WaveTSlope or as its inverse WaveTLin, that corresponds to t_1 - t_0 (see Fig. 1.4). In the next image it is possible to have a clearer vision of the description. It is important to specify that the rising edge and the falling edge must coincide, so the rate will increase from t_0 on as fast (with the same derivative) ad it decreases after t_2. It is easy to imagine that those two edges need to be as fast as possible. The lower limit to the rising and

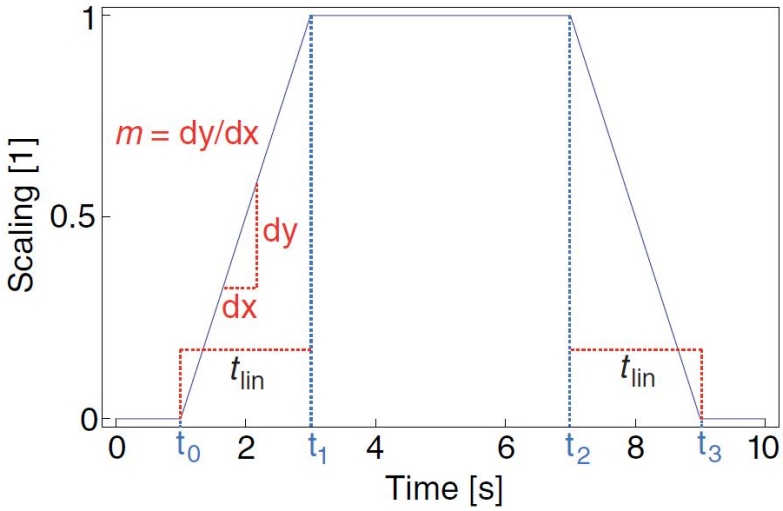


Figure 1.4: Linear optical impulse [6].

falling times is given by the convergence of the problem, which is not granted if we go beyond a certain value.

1.4.5 Generation term in the drift-diffusion model and explanation of the transient loop

The carrier transport models can all be expressed as continuity equations and then derived for each specific case or model Sentuarus is going to use. The important part is that the problem's time evolution of the solution of the transport problem can be generally written as the conservation of the charges in the form of these two equations, also known as the continuity equations [7, 8]:

$$\nabla \cdot \vec{J}_n = q\left(R_n - G_n\right) + q\frac{\partial n}{\partial t}, \tag{1.46}$$

$$-\nabla \cdot \vec{J}_p = q\left(R_p - G_p\right) + q\frac{\partial p}{\partial t}, \tag{1.47}$$

where $\vec{J}_n$ and $\vec{J}_p$ are the electron and hole current densities, and R_n, R_p, G_n, G_p are the electron and hole recombination and generation rates, respectively. There are two main points on which we have to focus to explain this formulation of the problem. The first one is the generation term, which is the one we are giving as a transient parameter and we explained just before. So, to resume, we have a solution of the optical problem through a FDTD method and then it is taken and scaled to obtain a linear impulse as we have described in the previous part. The combination of this generation term and the recombination one, which is related to various recombination

methods and events are happening in our device, will determine part of the solution. The other important term is the last one, which describes the electrons and holes population density variation in time. This is exactly what is changing in the transient analysis each time we are moving on by a Δt (also the generation which changes during the rise and fall of the edge, which is ideally vertical from a mathematical approach, but, as we have already seen, it is impossible from a computational and physical point of view). The solution of the drift-diffusion model is done each time we increment time. Therefore, the model we use to calculate the current densities is the drift-diffusion one. The model is described by two formula, one for the current density of electrons and one for the holes' one. The statistic used is the Fermi-Dirac one. This statistic is necessary because in our device there is a doping concentration quite high, which reaches even $10 \times 10^{20}\,\mathrm{cm}^{-3}$ in the silicon waveguide. So the two parameters can be described by the carriers distributions 1.4 with the use of Fermi-Dirac statistic 1.1.

1.4.6 Discretization method

The transient simulation is described by equations that can be written as a set of ordinary differential equations

$$\frac{d}{dt}q(z(t)) + f(t, z(t)) = 0 \tag{1.48}$$

which can be mapped to the DC and transient part of the partial differential equations [6]. What Sentaurus uses for the discretization of the transient equations are two different discretization schemes: the simple backward Euler (BE) method and the composite trapezoidal rule/backward differentiation formula (TRBDF) [9, 10, 11]. This scheme with the use of the two different schemes is the default one for all the transient simulations. The Backward Euler is a definitively stable method, but this is not so precise due to the fact it has only a first-order of approximation over the time-step h_n. This discretization can be written through the formula:

$$q\left(t_n + h_n\right) + h_n f\left(t_n + h_n\right) = q\left(t_n\right) \tag{1.49}$$

We can now think about the local truncation error (LTE) estimation, which is based on the comparison of the obtained solution $q\left(t_n + h_n\right)$ with the linear extrapolation from the previous time-step. The extrapolated solution is then written as:

$$q^{\mathrm{extr}} = q\left(t_n\right) - \frac{f\left(t_n\right) + f\left(t_n + h_n\right)}{2} h_n \tag{1.50}$$

In every point of we can estimate the relative error by the formula:

$$\left(q\left(t_n + h_n\right) - q^{\mathrm{extr}}\right) / q\left(t_n + h_n\right) \tag{1.51}$$

Using the two previous formulas in order to estimate the norm of the relative error, Sentaurus Device computes this value:

$$r = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{f\left(t_n + h_n\right) - f\left(t_n\right)}{\varepsilon_{\mathrm{R,tr}}\left|q_n\left(t_n + h_n\right)\right| + \varepsilon_{\mathrm{A,tr}}} h_n\right)^2} \tag{1.52}$$

with the sum over all the unknowns also known as all the free vertices of all the equations: $\varepsilon_{\mathrm{R,tr}}$ is the relative and $\varepsilon_{\mathrm{A,tr}}$ is the absolute transient errors, respectively. The next-time step is estimated as:

$$h_{\mathrm{est}} = h_n r^{-1/2} \tag{1.53}$$

This estimation is used for h_{n+1} computation, being it equal to the estimated value when the r is lower in value than 2 time f_{rej} . We can now describe the TRBDF method: for each time point t_n, The next time point which can be written as $t_n + h_n$ (with h_n current step size) is not directly reached. A step in between to $t_n + \gamma h_n$ is made improving the accuracy of the discretization method. The optimal value of gamma has been shown to be equal to $\gamma = 2 - \sqrt{2}$. Using this value the method reach two different nonlinear systems.For the trapezoidal rule (TR) step:

$$2q\left(t_n + \gamma h_n\right) + \gamma h_n f\left(t_n + \gamma h_n\right) = 2q\left(t_n\right) - \gamma h_n f\left(t_n\right) \tag{1.54}$$

while for the BDF2 step:

$$\begin{aligned}(2-\gamma)q\left(t_n + h_n\right) + (1-\gamma)h_n f\left(t_n + h_n\right) = \\ = (1/\gamma)\left(q\left(t_n + \gamma h_n\right) - (1-\gamma)^2 q\left(t_n\right)\right)\end{aligned} \tag{1.55}$$

The local truncation error (LTE) in this method is estimated after the double step as:

$$\tau = \left[\frac{f\left(t_n\right)}{\gamma} - \frac{f\left(t_n + \gamma h_n\right)}{\gamma(1-\gamma)} + \frac{f\left(t_n + h_n\right)}{1-\gamma}\right] \tag{1.56}$$

$$C = \frac{-3\gamma^2 + 4\gamma - 2}{12(2-\gamma)} \tag{1.57}$$

And then Sentaurus computes the norm of the relative error:

$$r = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\frac{\tau_i}{\varepsilon_{\mathrm{R,tr}}\left|q_n\left(t_n + h_n\right)\right| + \varepsilon_{\mathrm{A,tr}}}\right)^2} \tag{1.58}$$

another time the sum is over all the unknowns which are the free vertices of all equations, $\varepsilon_{\mathrm{R,tr}}$ and $\varepsilon_{\mathrm{A,tr}}$ the relative and absolute transient errors. Since TRBDF method has in this case second order approximation on h_n instead of first order (as in the BE one). The successive step can be computed as:

$$h_{\mathrm{est}} = h_n r^{-1/3} \tag{1.59}$$

The default method is this last one, the TRBDF. We can easily switch with a command to the BE. We can see the control mechanism to see if a time-step was successful and to provide an estimate of the next step size. There are a few simple rules: - if one nonlinear system cannot be solved, the step is rejected and it uses half the step $h_n = 0.5 \cdot h_n$ - otherwise, $r < 2f_{\text{rej}}$ is tested, and if fulfilled the transient assign to the next step the value of the esteem, or it is tried again with a value of 0.9 times the esteem - the LTE is checked only if an option (CheckTransientError) is selected, otherwise the next time-step is chosen only for the purpose of convergence of the nonlinear iterations.

Chapter 2

Photodetectors

In this chapter we are going to introduce the analyzed and studied device. It is a waveguide photodetector, and where the light is guided all the way through a waveguide (typically a silicon one) to the detector region, where a coupling happens. When the light reaches the absorber, thanks to its material properties, it is possible to see a great number of electron-hole pairs that have been generated in the detector. Hence, it is important to explain them from a general point of view to fully understand the specific one we are going to study.

The photodetectors are part of the receiver in a generic communication system. The main role they cover is to convert an optical signal in an electrical one, i.e. an optical power into a current. The use of light to transfer information in close-distance devices is crucial in order to overcome the actual limit in copper transmission lines and obtain a greater band. Photodetectors are meant to be at the end of the optical chain in order to convert the optical input in an electrical output: we obtain a voltage or a current with the appropriate use of an external circuit (Fig. 2.1). Since they depend on both material choice and on targeted application, device performances rely upon available fabrication and its integration technologies. An example of those devices can be found in [12, 13] where Ge-on-Si photodetectors are fabricated and characterized. The physics on which this conversion is based is the generation-excitement mechanism of electron-hole pairs due to a photon incident on the active region of the detector [14]. We want to convert

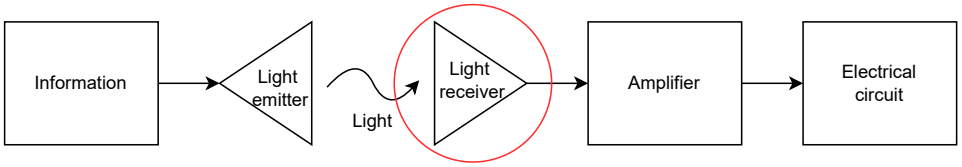


Figure 2.1: Receiver block in a optical link.

the optical excitement seen as a photon in an electrical signal, which is then read by an external circuit. The main parameter for an efficient conversion of photons with a certain energy is the correct choice of the semiconductor used as active material with a suitable band gap energy. [12] The photodetectors are meant to be designed in order to have both a great sensibility on the photons arriving and a even great electric field, in order to have a fast separation of the pairs to make the electron collected by the cathode and the holes by the anode. Literally, the measure of how fast the collection of the generated pairs happens tells us the frequency of the detector and its increase is exactly the main goal we want to achieve and the main reason we are studying the photodetectors. The easiest way to carry out a high field is the use of an external reverse bias, which makes the collection way quicker. As we are going to see, the movement of the carriers in the device can be explained with a drift-diffusion model: we are exploiting the drift component of the density current of the carriers, prerogative of the electric field. This is kinda necessary because of the greater value of this component respect to the diffusive one, which is way slower and it would only slow all our system (in absence of electric field). However, the presence of a regular and well-spread electric field over the whole volume of the photodetector is not trivial and needs some precautions, for example in the corners of a rectangular device. The current in case of a reverse bias is not zero even in case of a completely dark situation and this take the name of dark current i_d. As far as a depletion region is present in the device, there is this sort of leaked generation of carriers inside it, which are then collected and this generate a current with quite low values respect to the one obtained in a illuminated context, but it cannot be ignored. What it is experimentally observed [15][16] is the fact that the main contributions to the dark current are given by two different mechanisms: tunneling and trap-assisted recombination of the carriers [17][18][19]. It is interesting to better explain those two mechanisms and we will understand next why they are important in our structure. It can be considered as noise, because it sums to the photocurrent without being properly produced by the photons' creation of carrier's pairs and it has to be taken into account for every calculation we are going to do in the next chapters. One of the first parameters which can say the quality of the detector is the responsivity and this has the measure units of A/W. This is the proportionality factor between the photocurrent and the input optical power and this relation holds[14]:

$$i_L(t) = \Re p_{in}(t) \tag{2.1}$$

The first assumption is the linearity of this relation, but this soon is not anymore valid if we are considering larger input power, due to the screening of the consequent extremely high density of carriers. A saturation of the current is present and we have to think a little more over the absorption

of our material. It is necessary to introduce the dependence of the responsivity to the wavelength. The shape of its curve is like a band-pass and it is important to use wavelengths inside this range. To fully understand this statements it is important a rapid digression on the band structure of semiconductors. The plot of the bands of the Germanium (spoiler: the material of our detector) for example helps us to understand easily the meaning of the wavelength dependence(look at Fig. 2.2). In the case of a direct gap semiconductor (one symmetry point has the same k vector of the maximum of the valence band), the energy between the minimum of the conduction band and the maximum of the valence one is exactly the amount of energy required to make an electron go to the conduction band and a hole in the valence band. If there are no other effects such as defect or tail states inside the band gap, the lowest amount of energy is exactly that one. This happens also in the case of an indirect semiconductor such as Silicon, but the conservation of the momentum k has to be preserved and a single photon is not able to provide it: a photon has a great amount of energy, but an insignificant amount of momentum compared to the one required. This opens the world of phonons scattering. Going back to the main topic, the minimum amount of energy required can be related to the frequency with

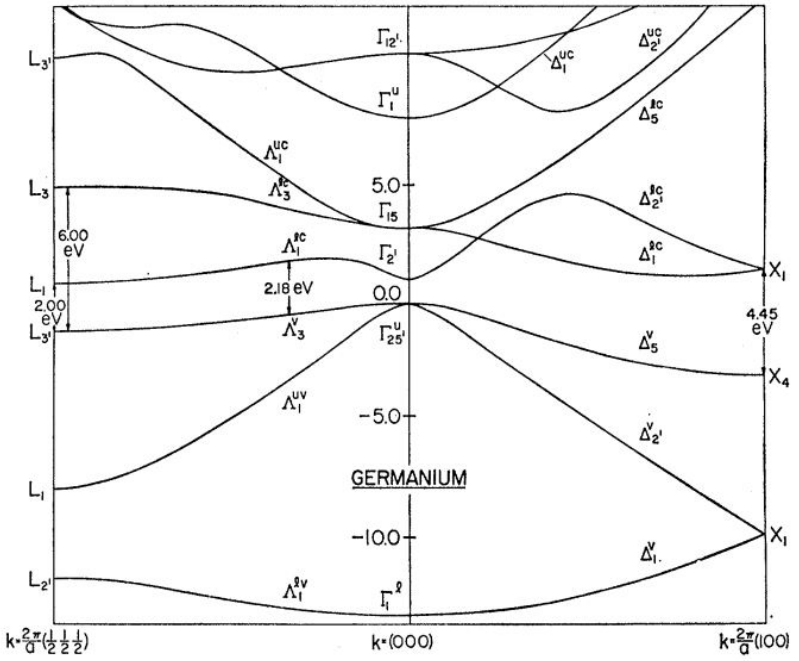


Figure 2.2: Energy bands of germanium calculated by the $\mathbf{k} \cdot \mathbf{p}$ method in the [111] and [100] directions of $\mathbf{k}$ space [20].

the simple property:

$$E_{ph} = h\omega \geq E_g \tag{2.2}$$

In other words, there is a maximum value of wavelength in the responsivity curve, thus we can see the dependence: there is ideally no current even in case of high optical power if we are over that range. The linear relation holds only in case the power change is slow in time (or if modulation frequency is lower than the cutoff one in case of a time-harmonic input), if not, the capacitance effects and the transit time effect come into play. In small-signal analysis, this is equivalent to a complex responsivity, which in this case relates both the amplitude and the phase of current and optical input power. If we look another time from above and we think about the whole receiver block, the current which exits from the detector needs to become a voltage, using usual converters such as a load resistance or as a transimpidance amplifier (TIA). From a structural point of view, the detector can be:

- bulk
- junction
- photodiode
- avalanche photodiode
- phototransistors

The differences between them is not only from a structural and construction point of view, but they differ also for what concern the speed of the device. There are huge differences in losses, noise (signal/noise ratio, comparing the photocurrent to the dark one for example), costs and frequency response. Both junction and photodiode types are based on a *pn* junction: the region devoted to collect the photons is the depletion one. This is created in the junction and it could be enlarged by the use of an external negative bias (which also increases the dark current). The main limit in those cases are both the lifetime of the carriers before they get recombined and the transit time, while the capacitances are not related to any frequency limitation. The first idea, in order to have better performances in these cases, is to use a pin junction, where the "i" stands for intrinsic. We want to increase the ratio between the intrinsic regions and the diffusion ones, which are the main reason of lag in the collection of carriers at contacts. This lightly doped region, as a neutral charge one, is able to absorb even more photons, maximizing the absorption pattern of the photodetector. As we are going to see, this is what happens in our device, being it a heterojunction between a highly doped Silicon waveguide, a contacting Germanium zone with quite low doping and a high doping one in the region close to the metallic contact. The fact that we can engineer easily this absorption region is an optimum point

in favor to this structure, because we can design it based on the absorption length of the material, increasing as much as we can the photogeneration in this region respect to the one in highly doped regions. Photogenerated carriers are removed quickly from an electric field mostly uniform applied on the intrinsic region with a movement which can be associated almost entirely to its drift component. Being the diffusion component way less important, the frequency response is not anymore limited by the lifetime of the carriers (being the collection at contacts way faster), but it is still limited by the transit time, even if with even lower limitations. The frequency values are quite higher, so some kind of RC effects can arise, in particular in devices particularly large (high) in the direction parallel to the illumination direction. If the speed can reach very high values (what we are going to see in our device), the sensitivity is not the best one out of every photodetector device we have mentioned, because there is a unit gain and, unlikely APD based photodiodes, there is not a multiplication of carriers, which makes the detection of even a single photon way easier. We have talked about the different structures and their advantages/disadvantages, we can talk now about how the materials of the detector influence the working point of the device. If we think to the generation of carriers, it happens when the energy of the photon is enough to overcome the band gap. This is related to the material absorption profile and to its threshold, which tells us the minimum energy necessary to create a carrier's pair. The relation is quite simple and because of the dependence of energy on the frequency of the incoming light, we can think it in an inverse way, looking from the point of view of the wavelength:

$$E_{ph} = h\omega \geq E_g \longrightarrow \lambda[\mu\text{m}] \leq \frac{1.24}{E_g[\text{eV}]} \tag{2.3}$$

The materials for these devices can have both direct and indirect gap, but the firsts have higher absorption values. In general, the exponential decrease of power in the semiconductor can be expressed both with the absorption coefficient or with the absorption length, which is the inverse:

$$P_{in}(x) = P_{in}(0)\exp(-\alpha x) = P_{in}(0)\exp\left(-x/L_\alpha\right), \tag{2.4}$$

If we think to the problem, to design a photodetector for a certain source we have to choose wisely the material, because we want to absorb most of what hits our apparatus. Its thickness has to be at least comparable with the absorption length, however we cannot use a value much higher in order to limit arising transit time problems. In junction-based detectors, the capacitance gain increasing importance because of the inverse proportionality between its value and the thickness of the device, which limit and cap the device speed (its frequency response) while the transit time problem fall off. This is then a trade-off in the design of the photodetector between these two limits: RC related and transit time.

2.1 Parameters and relations

If we think of the photodetector from an electrical point of view, this is a one-port with an optical input port. The input power is pin around the working point wavelength λ and the output current i_{PD} is composed by both dark i_d) and illuminated currents i_L. Once we have defined these quantities, we can model it with the following constitutive relation:

$$i_{PD}(t) = f\left(p_{in}(t), v_{PD}(t); \frac{\mathrm{d}}{\mathrm{d}t}, \lambda\right) \tag{2.5}$$

The only quantity we have not talked about yet is the voltage v_{PD}, but this is trivial. The time derivative inside the relation tells us that there is in some way memory because the memory-less condition holds only in case of slow variation in time of the power or a modulation frequency lower than the cutoff one. As we are going to see, especially in the transient analysis, we want to see the feedback of the detector in case of an optical impulse that goes from 0 to a certain value in very low time (in an ideal case we would want it to be an actual impulse, but this gives convergence problems when we use reduce too much the time necessary for the rise and falling edges). As already said, the output current is the sum of the two contributions:

$$i_{PD} = i_L + i_d, \tag{2.6}$$

To be clear another time, the dark current is the one intrinsically present in the detector due to tunneling effects and due to recombination events of carriers, while the illuminated component is present when the light is incident to the detector and the photons create electron-hole pairs, logically when the photons have enough energy to overcome the band gap of the material. Those two can be defined by the following relations:

$$\begin{aligned} i_d &= f\left(0, v_{PD}(t); \frac{\mathrm{d}}{\mathrm{d}t}, \lambda\right) \\ i_L &= f\left(p_{in}(t), v_{PD}(t); \frac{\mathrm{d}}{\mathrm{d}t}, \lambda\right) - i_d. \end{aligned} \tag{2.7}$$

In case of DC stationary state we have the exclusion of the derivative from the relation:

$$I_{PD} = f\left(P_{in}, V_{PD}; 0, \lambda\right) = I_L + I_d, \tag{2.8}$$

where $I_d = f\left(0, V_{PD}; 0, \lambda\right)$ is the DC dark current, $I_L = f\left(P_{in}, V_{PD}; 0, \lambda\right) - I_d$ is the DC photocurrent. The relation is in principle both nonlinear and dispersive, but, as we have seen, if the input power is varying slowly in time, we can approximate using the definition itself of responsivity, which still depends on the wavelength:

$$i_{PD}(t) = i_L + i_d \approx \Re\left(\lambda, v_{PD}\right) p_{in}(t) + i_d\left(v_{PD}\right) \tag{2.9}$$

In many cases the responsivity and the dark component of the current are not dependent on v_{PD}, because the voltage is quite small and overall negligible in the formula, leading to this expression:

$$i_{PD}(t) = \Re(\lambda) p_{in}(t) + I_d \approx \Re(\lambda) p_{in}(t) \tag{2.10}$$

After a certain optical power, there is both a decrease in responsivity and a saturation in the curve, due to the space-charge screening of the electric field cause by the collected photocarriers.

2.1.1 Responsivity

The photocurrent can be defined as the integration of the optical generation rate G_0 over all the active volume that absorb the radiation:

$$I_L = q \int_V G_o\left(\underline{r}, P_{in}\right) \mathrm{d}\underline{r} \tag{2.11}$$

And the responsivity is the ratio between the output current and the input optical power in both incremental and differential definitions:

$$\Re = \frac{I_L}{P_{in}} \text{ or } \Re_{\text{diff}} = \frac{\mathrm{d}I_L}{\mathrm{d}P_{in}}. \tag{2.12}$$

Even if the generation rate is evaluated with a numerical solver which performs FDTD, a simple derivation can be done to understand the analytic method with a consequent ideal solution. The goal is to relate the optical generation rate G_0 and the optical power, and this can be obtained by differentiating the optical power density $\tilde{P}_{in} = P_{in}/A$ $\left(\mathrm{W/m^2}\right)$ respect to the x axis. The area on which we have calculated the density is the detection area (intrinsic or neutral charge region) of our photodetector. What we obtain is:

$$\begin{aligned} \frac{\mathrm{d}\widetilde{P}_{in}(x)}{\mathrm{d}x} &= -\alpha \widetilde{P}_{in}(x) \rightarrow \\ \frac{\text{Energy lost due to absorption}}{t \cdot V} &= -\frac{\Delta \widetilde{P}_{in}}{\Delta x} = \alpha \widetilde{P}_{in}. \end{aligned} \tag{2.13}$$

We can now divide for the definition of the energy of the photon $E_{ph} = \hbar\omega$:

$$\begin{aligned} \frac{(\text{ Energy lost })/(t \cdot V)}{\text{Photon energy } \hbar\omega} &= \frac{\alpha \widetilde{P}_{in}}{\hbar\omega} = \frac{\text{Number of photons absorbed}}{t \cdot V} \\ &= \frac{\text{Number of e-h pairs generated}}{t \cdot V} = G_o \end{aligned} \tag{2.14}$$

With the final relation:

$$G_o = \frac{\alpha \tilde{P}_{in}}{\hbar\omega} \tag{2.15}$$

Due to the fact we can say that the optical power density decreases in an exponential way respect to the absorption direction, the behaviour of the optical generation rate will be the same:

$$G_o(x) = \frac{\alpha \widetilde{P}_{in}(x)}{\hbar\omega} = \frac{\alpha \widetilde{P}_{in}(0)}{\hbar\omega} \exp\left(-x/L_\alpha\right) = G_o(0) \exp\left(-x/L_\alpha\right) \quad (2.16)$$

Now, we can do some assumptions: all the incident power is absorbed and all the generated pairs are collected in form of current in the external circuit. We have in this way: $\frac{\text{Number of electrons in the external circuit}}{t} = \frac{I_L}{q} = V \cdot \frac{\text{Number of e-h pairs generated}}{t \cdot V} = \frac{\text{Number of photons absorbed}}{t \cdot V} = A \int_0^\infty G_o(x) \mathrm{d}x = A \int_0^\infty \frac{\alpha \widetilde{P}_{in}(x)}{\hbar\omega} \mathrm{d}x = -\frac{A}{\hbar\omega} \int_0^\infty \frac{\mathrm{d}\widetilde{P}_{in}(x)}{\mathrm{d}x}\, \mathrm{d}x \approx \frac{P_{in}(0)}{\hbar\omega}$,

$$\frac{I_L}{q} = \frac{P_{in}(0)}{\hbar\omega}$$

It follows that the photocurrent depends linearly on $P_{in}(0)$ through the responsivity $\Re$:

$$I_L = \frac{q}{\hbar\omega} P_{in}(0) = \Re P_{in}(0) \quad (2.17)$$

Using the density point of view of the same quantities, we can in parallel define $J_L = \Re \widetilde{P}_{in}(0)$. From the previous formula we can derive a different expression of the responsivity which is related to the energy of the photon:

$$\Re = \frac{q}{h\omega} = \frac{q}{E_{ph}} \quad (2.18)$$

This is valid only when all the incident photons are both absorbed and converted in electron-hole pairs which arrive at the contacts without any recombination phenomenon. In this best case possible, the responsivity has a maximum as a function of the photon energy: it is zero for photon energies below the absorption threshold and it assumes a sharp increase of absorption once we have reached the threshold value. We can assume the photon density equal to the band gap one $E_{ph} \approx E_g$:

$$\Re_{\max} \approx \frac{q}{E_g} = \frac{1}{E_g[\mathrm{eV}]} \approx \frac{\lambda[\mu\mathrm{m}]}{1.24} \quad (2.19)$$

and a first approximation of the behaviour for energies higher than the gap, we can introduce at first an inverse proportionality:

$$\Re\left(E_{ph}\right) \approx \Re_{\max} \frac{E_g}{E_{ph}}, \quad (2.20)$$

This is quite a good estimation when we look at the form of the responsivity respect to the absorption (Fig. edge2.3). The value of responsivity of a real device can be a little different, because not all the carriers are collected,

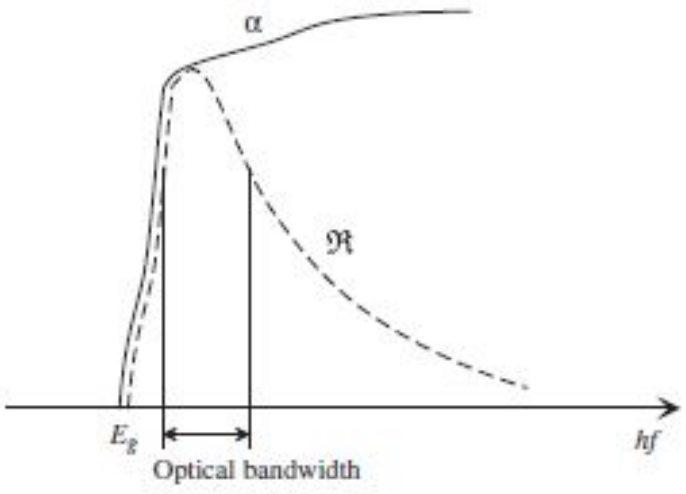


Figure 2.3: Behavior of absorption α and responsivity $\Re$ (in arbitrary units, normalized so that $\Re \approx \alpha$ near E_g) vs. the photon— energy [14].

due to many possible effects happening such as recombination events in correspondence of defects at the interface between different materials (silicon/germanium). The maximum responsivity can be very large in case of far infrared photodetectors, while in case of long-wavelength infrared detectors we have a maximum in the order of 1 A/W. Aside from the responsivity, we can talk also about the quantum efficiency: the internal one assume that all the photons are correctly absorbed, so we think about the number of generated carriers respect to the photons which actually reach the active region.

$$\eta_Q = \frac{\text{generated pairs}}{\text{photons reaching the active region}} \tag{2.21}$$

The external quantum efficiency is more general, because we now consider:

- the carriers which are collected and not anymore the generated ones
- the incident photons and not anymore the ones which effectively reach the absorption region.

We are in this way accounting two more real device problems: not every generated pair is correctly collected and not each photon incident on the device is properly absorbed by the active region of the detector.

$$\eta_x = \frac{\text{collected pairs}}{\text{incident photons}} = \frac{I_L/q}{P_{in}/\hbar\omega} = \frac{\hbar\omega}{q}\Re < \eta_Q. \tag{2.22}$$

If the internal one can be very close to 1, the second one is usually lower if we do not assume an ideal operation case, which would make the two

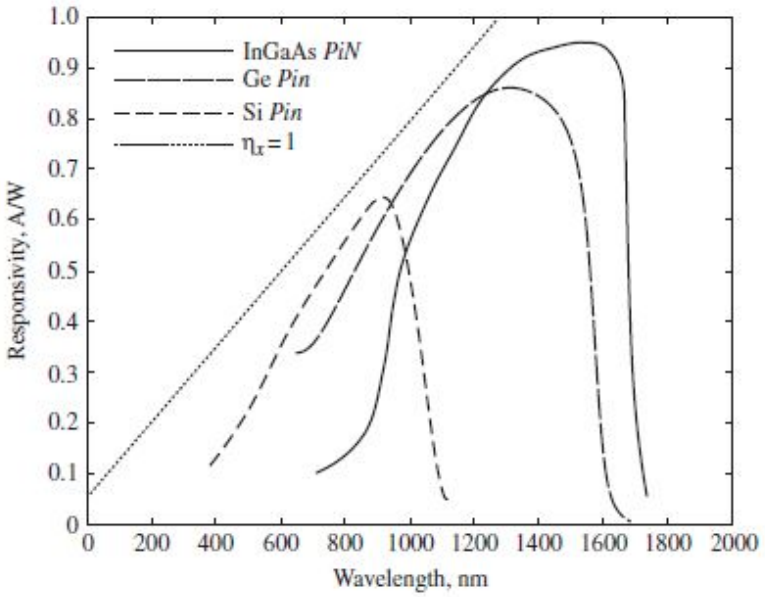


Figure 2.4: Responsivity versus wavelength for a Si homojunction, an InGaAs heterojunction high-speed pin photodiode, and a Ge-on-Si photodiode. The dotted curve is the ideal case with external quantum efficiency $\eta_x = 1$ and responsivity $\Re = q/(\hbar\omega)$ [14].

different definitions of quantum efficiency coincide. If we talk a little more about non-ideality mechanisms present in a real device we can sum them in this list:

- The optical power P_{in} (0) is incident on the photodetector.
- Part of the power is reflected at the PD interface due to dielectric mismatch.
- Part of the power is absorbed in regions where it does not contribute to the useful output current.
- Part of the power is transmitted through the PD without being absorbed.
- Finally, part of the power is absorbed and yields a useful current component.

We can also see where there is a good responsivity in some materials and for which wavelength we can use them (see Fig. 2.4). We can focus mainly on the Germanium, the leading actor of this thesis.

2.1.2 Electrical bandwidth

We start another time from the same constitutive relation we have used at the beginning of the discussion of the responsivity, the difference is the negligence of the wavelength dependence:

$$i_{PD}(t) = f\left(p_{in}(t), v_{PD}(t), \frac{\mathrm{d}}{\mathrm{d}t}\right) \tag{2.23}$$

We separate now the DC and AC components and for simplicity we denotes the DC components with a 0:

$$P_{in} = P_{in,0} + \hat{p}_{in}(t), \quad V_{PD} = V_{PD,0} + \hat{v}_{PD}(t), \quad I_{PD} = I_{PD,0} + \hat{\imath}_{PD}(t) \tag{2.24}$$

The AC component is a general signal, but for this demonstration we can assume a sinusoidal modulation of the light with the consequent association of the signal components:

$$P_{in} = P_{in,0} + \hat{p}_{in}(t), V_{PD} = V_{PD,0} + \hat{v}_{PD}(t), I_{PD} = I_{PD,0} + \hat{\imath}_{PD}(t), \tag{2.25}$$

ω is the light angular modulation frequency. We are now able to linearize around the DC working point:

$$\begin{aligned} &I_{PD,0} + \hat{\imath}_{PD}(t) = \\ &= \underbrace{f\left(P_{in,0}, V_{PD,0}, 0\right)}_{I_{PD,0}} + \left.\frac{\partial f(\ \mathrm{d}/\mathrm{d}t)}{\partial p_{in}}\right|_0 \hat{p}_{in}(t) + \left.\frac{\partial f(\ \mathrm{d}/\mathrm{d}t)}{\partial v_{PD}}\right|_0 \hat{v}_{PD}(t), \end{aligned} \tag{2.26}$$

where the second and third term are the two small-signal parts of the photodetector current: respectively photocurrent and dark current. For these two terms, we can another time express them with the phasor notation:

$$\hat{\imath}_{PD}(t) = \hat{\imath}_L(t) + \hat{\imath}_d(t) = \mathrm{Re}\left(\Re(\omega)\hat{P}_{in}\mathrm{e}^{\mathrm{j}\omega t}\right) + \mathrm{Re}\left(Y_{PD}(\omega)\hat{V}_{PD}\mathrm{e}^{\mathrm{j}\omega t}\right) \tag{2.27}$$

in which we introduce two relatively new terms in our discussion: the complex small-signal responsivity and the small-signal admittance. The associated phasor is:

$$\hat{I}_{PD}(\omega) = Y_{PD}(\omega)\hat{V}_{PD}(\omega) + \hat{I}_L(\omega) \tag{2.28}$$

where the second term can be related through the well-known and already discussed equation formulated in a phasorial way. In the case of $\hat{V}_{PD} = 0$ the current is called short-circuit photocurrent. While the complex responsivity R_ω describes the small-signal frequency response and it is a low-pass function of the modulation frequency. A normalized responsivity $\mathfrak{r}(\omega)$ can be defined from this division:

$$\frac{\hat{I}_L(\omega)}{\hat{I}_L(0)} = \frac{\Re(\omega)}{\Re(0)}\frac{\hat{P}_{in}(\omega)}{\hat{P}_{in}(0)} = \mathfrak{r}(\omega)\frac{\hat{P}_{in}(\omega)}{\hat{P}_{in}(0)}. \tag{2.29}$$

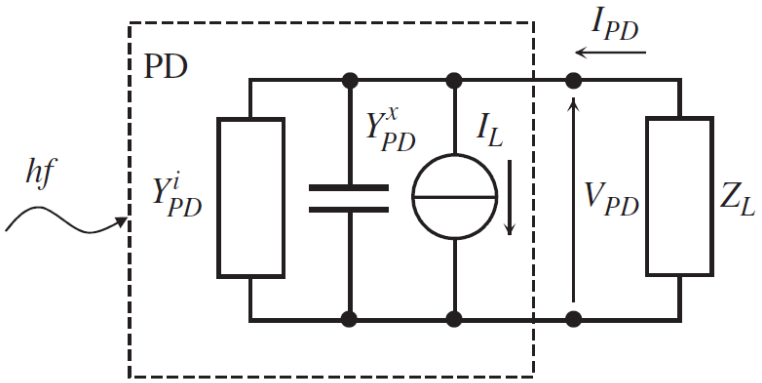


Figure 2.5: Small-signal equivalent circuit of loaded photodetector [14].

and assuming a constant complex input power we obtain:

$$\mathfrak{r}(\omega) = \frac{\hat{I}_L(\omega)}{\hat{I}_L(0)} = \frac{\Re(\omega)}{\Re(0)} \rightarrow |\mathfrak{r}(\omega)|_{\mathrm{dB}} = 20\log_{10}\left|\frac{\Re(\omega)}{\Re(0)}\right|. \tag{2.30}$$

For a typical low-pass detector, the cutoff frequency (or 3 dB bandwidth) is the frequency at which the normalized responsivity drops by 3 dB respect to its DC value. To this aim, it is usual to normalize the DC value to 0 and look and the value at -3 dB:

$$\begin{gathered} |\mathfrak{r}\left(\omega_{3\ \mathrm{dB}}\right)|_{\mathrm{dB}} = -3 \rightarrow \\ \rightarrow 20\log_{10}\left|\frac{\Re\left(\omega_{3\ \mathrm{dB}}\right)}{\Re(0)}\right| = -3 \rightarrow \Re\left(f_{3\ \mathrm{dB}}\right) = \frac{1}{\sqrt{2}}\Re(0) \end{gathered} \tag{2.31}$$

This cutoff frequency refers to short-circuit photocurrent and does not depend on the load. Many effects we have already talked about are crucial such as the transit time, the avalanche buildup delay in case of APD's photodetectors or current gain at high-frequency. The overall response considers also the load and the parasitic elements such as capacitances: this is the RC cutoff caused by the combination of both intrinsic and extrinsic capacitance effects with the external load resistance. A quantitative evaluation can be performed with the equivalent circuit represented in Fig. 2.5. and, in the frequency domain, the detector can be modelled by the current-voltage phasor relation:

$$I_{PD}(\omega) = \left[Y^i_{PD}(\omega) + Y^x_{PD}(\omega)\right] V_{PD}(\omega) + I_L(\omega) \tag{2.32}$$

If we have already talked about $I_L = \Re(\omega)P_{in}$ as the short-circuit component at *omega*, we can now distinguish the admittance in the "i" intrinsic one and the "x" parasitic (and in particular capacitive) one. The load impedance has to be considered for what concerns the bandwidth of the loaded detector as well as the capacitive and resistive loading. The frequency-dependent

responsivity takes into account only cutoff effects, while RC cutoff is handled at circuit level. Assuming both $Z_L = R_L$ and C_{PD} as total capacitance, the current flowing through the load $I_{R_L} = -I_{PD}$ is:

$$I_{R_L}(\omega) = -\frac{I_L(\omega)}{1 + \mathrm{j}\omega R_L C_{PD}} \rightarrow |I_{R_L}(\omega)| = \frac{|I_L(\omega)|}{\sqrt{1 + \omega^2 R_L^2 C_{PD}^2}} \tag{2.33}$$

From that formula we can derive also another formulation of the responsivity of the loaded detector:

$$|\Re_l(\omega)| = \frac{\Re}{\sqrt{1 + \omega^2 R_L^2 C_{PD}^2}} \tag{2.34}$$

The limitation in the cutoff frequency due to RC effects is:

$$f_{3\ \mathrm{dB}} = \frac{1}{2\pi R_L C_{PD}} \tag{2.35}$$

As a first approximation of the large-signal model, one can model an equivalent circuit with the parallel of a capacitive (sometimes nonlinear) admittance with two different current generators: one models the photocurrent I_L (linearly dependent on the optical power) and the dark one I_d (which is negligible in some context). Adding more elements, one can take into account other kinds of effects such as voltage-dependent photocurrent, nonlinear detector input admittance, more interconnected parasitic networks with series connector resistances, wire inductances, and distributed more than lumped elements.

2.1.3 Frequency response in pin

We can now focus a little more on pin detector, being the general structure of our analysed devices. There are four mechanisms that can limit the speed of the device when we are exciting it and one of these can be the dominant one in each specific case depending on the structure and the conditions. The main ones are:

- The effect of the capacitances, which are the sum of the one consequent of the depleted region, external parasitic ones, and the one consequent of the height of the device.
- transit time of carriers drifting in the depletion layer of width W (which is larger due to the intrinsic area).
- diffusion time of carriers that are generated outside of the depleted regions and this is even more important in homojunction.
- charge trapping in heterojunctions (as in our case).

Transit time is quite important in our devices, while it is almost negligible in *pn* devices, being their depleted region narrower. In our device as in every optimized pin photodiode, the two main limitations are the transit time and the *RC* cutoff.

2.2 Dynamic pin model

We can discuss now a general way to follow the path of photogenerated carriers in a pin photodetector, in order to give a generic description to fully understand the main effects that happen in a photodetector and what can actually limit the speed of the device. If we want to follow the dynamic of photogenerated carriers in the less doped region, we can model it as a 1D problem and we can write the continuity equations for electrons and holes in this way:

$$\begin{aligned} \frac{\partial p}{\partial t} &= -\frac{p-p_0}{\tau_b} + G_{op}(x,t) - \frac{1}{q}\frac{\partial J_h}{\partial x} \\ \frac{\partial n}{\partial t} &= -\frac{n-n_0}{\tau_b} + G_{op}(x,t) + \frac{1}{q}\frac{\partial J_n}{\partial x} \end{aligned} \tag{2.36}$$

the "0" means equilibrium value, while the τ refers to excess carrier lifetime. The electron and hole current densities are:

$$\begin{aligned} J_h &= q v_h(\mathcal{E}) p - q D_h \frac{\partial p}{\partial x} \\ J_n &= q v_n(\mathcal{E}) n + q D_n \frac{\partial n}{\partial x} \end{aligned} \tag{2.37}$$

where v_n and v_h are the field-dependent drift velocities. We can couple that with the Poisson equation:

$$\frac{\partial \mathcal{E}}{\partial x} = \frac{\rho}{\epsilon_s}, \tag{2.38}$$

Being the movement only due to drift component, when we apply a bias and we have a certain electric field with a direction, electrons and holes will have different velocities moduli and this is going to divide them. If we look also at its absolute value, this will be different and the charge has an overall asymmetry, which screens the field and change it. We assume this as a negligible effect: there is not a change in the total electric field we are using to solve the problem. When the field value is not too high we have:

$$\begin{aligned} J_h = q v_h(\mathcal{E}) p - q D_h \frac{\partial p}{\partial x} \approx q \mu_h \mathcal{E} p - q D_h \frac{\partial p}{\partial x} \\ \approx q \mu_h \frac{|V_A|}{W} p - q \frac{k_B T}{q} \mu_h \frac{\Delta p}{W} = \frac{q \mu_h}{W} \left(|V_A|\, p - \frac{k_B T}{q} \Delta p \right) \end{aligned} \tag{2.39}$$

This approximation holds as far as we assume the injected charge in the central region as linear with a slope equal to $\Delta p / W$ and we assume the

worst case scenario: $\Delta p \approx p$. The diffusion component of the current can be neglected when we have a $|V_A| \gg k_B T/q = 26\text{mVT}$ at ambient temperature and this is usually a correct inequality in operating devices. We will also assume that the transit time in the device of the photogenerated carriers is lower than the average time necessary to encounter a recombination event, so they will reach the contacts before having any kind of recombination process. The last assumption is that the electric field is high enough to make the velocity saturate inside the intrinsic part of the junction. The omission of space charge contribution can be assumed valid as far as the incident power is small. An higher value would lead to a higher space charge in the active region and a screening and non-uniformity of the field. To evaluate the impact of this effect we can assume $v_n \approx v_h \approx v$ or if we are in low field regime we can write the velocity as a linear dependence of the mobility $\mu_n \approx \mu_h \approx \mu$ and $n \approx p$. What we obtain is:

$$J_L = J_h + J_n \approx q\mu_h \mathcal{E}_0 p + q\mu_n \mathcal{E}_0 n \approx 2q\mu \mathcal{E}_0 n = 2qvn, \tag{2.40}$$

where $\mathcal{E}_0$ is the field of the mere external bias. The charge q multiplied by n, since $n = p$, is the charge density associated with both carriers:

$$|\rho| \approx \frac{J_L}{2\mu\mathcal{E}_0} = \frac{J_L}{2v} \tag{2.41}$$

If we put this charge inside the Poisson equation seeing it as an uncompensated density deriving from the displacement of generated carriers:

$$\frac{\partial \mathcal{E}}{\partial x} = \frac{\rho}{\epsilon_s} = \frac{J_L}{2\mu\epsilon_s\mathcal{E}_0} = \frac{J_L}{2\epsilon_s v} \tag{2.42}$$

where $\mathcal{E}$ is the extra field introduced due to this photocarriers. If we simply integrate on all the intrinsic layer (remembering we are doing this only in one dimension) we obtain:

$$\mathcal{E} = \frac{J_L W}{2\mu\epsilon_s\mathcal{E}_0} = \frac{J_L W}{2\epsilon_s v} \tag{2.43}$$

This electric field is negligible in two different cases: when we are in a low-field condition and when we are in a velocity-saturation regime.

$$\mathcal{E} = \frac{J_L W}{2\mu\epsilon_S\mathcal{E}_0} \ll \mathcal{E}_0 \rightarrow |\mathcal{E}_0| \gg \sqrt{\frac{J_L W}{2\mu\epsilon_S}} \tag{2.44}$$

$$\mathcal{E} = \frac{J_L W}{2\epsilon_s v} \ll \mathcal{E}_0 \rightarrow |\mathcal{E}_0| \gg \frac{J_L W}{2\epsilon_s v} \tag{2.45}$$

Thanks to these two inequalities, we can find an electric field which is sufficient to ignore the space charge contribution. We can obtain an estimation of the saturation power of the photodetector.

2.2.1 Transit time limitation

Considering only the drift currents with a constant electric field, the continuity equations are written in this way:

$$\begin{aligned} \frac{\partial p}{\partial t} &= G_o(x,t) - \frac{1}{q}\frac{\partial J_h}{\partial x} \\ \frac{\partial n}{\partial t} &= G_o(x,t) + \frac{1}{q}\frac{\partial J_n}{\partial x} \end{aligned} \tag{2.46}$$

where the density currents can be written as drift currents as:

$$J_h = q v_{h,\text{ sat}}\, p, \quad J_n = q v_{n,\text{ sat}}\, n. \tag{2.47}$$

The position $x = 0$ is positioned in correspondence of the pi junction, while we are assuming a length of $x = W$ of the intrinsic region. Being the system linear, we can use a harmonic optical incident power at an angular frequency of ω:

$$p_{in}(t) = P_{in}(\omega)\mathrm{e}^{\mathrm{j}\omega t} \tag{2.48}$$

If we substitute the density currents in the continuity equations we obtain, once we assume the electron density equal to zero at the interface between intrinsic and p doped region and the hole density equal to zero in the interface (minority carriers density at the interfaces is 0):

$$\begin{aligned} \mathrm{j}\omega p(x) &= G_o(x) - v_{h,\text{sat}}\frac{\mathrm{d}p(x)}{\mathrm{d}x} = G_o(0)\mathrm{e}^{-\alpha x} - v_{h,\text{sat}}\frac{\mathrm{d}p(x)}{\mathrm{d}x} \\ \mathrm{j}\omega n(x) &= G_o(x) + v_{n,\text{sat}}\frac{\mathrm{d}n(x)}{\mathrm{d}x} = G_o(0)\mathrm{e}^{-\alpha x} + v_{n,\text{sat}}\frac{\mathrm{d}n(x)}{\mathrm{d}x} \end{aligned} \tag{2.49}$$

What we are looking for is an exponential trial solution in the homogeneous associate:

$$\mathrm{j}\omega p'(x) = -v_{h,\text{sat}}\frac{\mathrm{d}p'(x)}{\mathrm{d}x} \tag{2.50}$$

written as:

$$p'(x) = A\mathrm{e}^{\mathrm{jkx}} \tag{2.51}$$

If we substitute this solution inside the formula we obtain:

$$\mathrm{j}\omega A\mathrm{e}^{\mathrm{j}kx} = -v_{h,\text{sat}}kA\mathrm{e}^{\mathrm{j}kx} \rightarrow k = -\frac{\mathrm{j}\omega}{v_{h,\text{sat}}} \tag{2.52}$$

We can now write the solution of the first complete equation with the sum of two terms$p_1 + p_2$, with p_2 particular solution of the same equation. We compose the solution as: $p_2 = B\exp(-\alpha x)$

$$p(x) = p_1 + p_2 = A\mathrm{e}^{-\frac{\mathrm{j}\omega x}{v_{h,\text{sat}}}} + B\mathrm{e}^{-\alpha x} \tag{2.53}$$

with A and B general constants. Substituting and using another time the condition $p(W) = 0$ we obtain the hole density:

$$p(x) = \frac{G_O(0,\omega)}{\mathrm{j}\omega - \alpha v_{h,\mathrm{sat}}} \mathrm{e}^{-\alpha W} \left[\mathrm{e}^{-\alpha(x-W)} - \mathrm{e}^{-\frac{\mathrm{j}\omega(x-W)}{v_{h,\mathrm{sat}}}} \right]. \tag{2.54}$$

While for the electron density we use the trial solution with the changed sign at the exponent of

$$n(x) = A\mathrm{e}^{\frac{\mathrm{j}\omega x}{v_{n,\mathrm{sat}}}} + B\mathrm{e}^{-\alpha x} \tag{2.55}$$

Substituting in the same way and imposing this time $n(0) = 0$, we have:

$$n(x) = \frac{G_o(0,\omega)}{\mathrm{j}\omega + \alpha v_{n,\mathrm{sat}}} \left[\mathrm{e}^{-\alpha x} - \mathrm{e}^{\frac{\mathrm{j}\omega x}{v_{n,\mathrm{sat}}}} \right] \tag{2.56}$$

From what we have obtained for $n(x)$ and $p(x)$, when we substitute this back in the drift currents we obtain:

$$\begin{aligned} J_h(x) &= q v_{h,\mathrm{sat}} p(x) = \frac{q v_{h,\mathrm{sat}} G_o(0,\omega)}{\mathrm{j}\omega - \alpha v_{h,\mathrm{sat}}} \mathrm{e}^{-\alpha W} \left[\mathrm{e}^{-\alpha(x-W)} - \mathrm{e}^{-\frac{\mathrm{j}\omega(x-W)}{v_{h,\mathrm{sat}}}} \right] \\ J_n(x) &= q v_{n,\mathrm{sat}} n(x) = \frac{q v_{n,\mathrm{sat}} G_o(0,\omega)}{\mathrm{j}\omega + \alpha v_{n,\mathrm{sat}}} \left[\mathrm{e}^{-\alpha x} - \mathrm{e}^{\frac{\mathrm{j}\omega x}{v_{n,\mathrm{sat}}}} \right] \end{aligned} \tag{2.57}$$

The total current will be the sum of these two components and of the displacement current:

$$J_t(\omega) = J_h + J_n + \mathrm{j}\omega\epsilon_s E(x,\omega) \tag{2.58}$$

where $E(\omega)$ derive from the harmonic input optical power we have chosen at the beginning of this section. We have now to integrate over the intrinsic region on both sides of the currents sum equation. If on the left side this is equivalent to a simple multiplication of the total current for the width, being it assumed not dependent, we can take care of the right one:

$$\begin{aligned} \int_0^W J_t(\omega)\mathrm{d}x =& W J_t(\omega) = \int_0^W \left[J_h(x) + J_n(x) + \mathrm{j}\omega\epsilon_s E(x,\omega) \right] \mathrm{d}x \\ =& \frac{q v_{h,\mathrm{sat}} G_o(0,\omega)}{\mathrm{j}\omega - \alpha v_{h,\mathrm{sat}}} \mathrm{e}^{-\alpha W} \int_0^W \left(\mathrm{e}^{-\alpha(x-W)} - \mathrm{e}^{-\frac{\mathrm{j}\omega(x-W)}{v_{h,\ \mathrm{sat}}}} \right) \mathrm{d}x \\ &+ \frac{q v_{n,\mathrm{sat}} G_o(0,\omega)}{\mathrm{j}\omega + \alpha v_{n,\mathrm{sat}}} \int_0^W \left(\mathrm{e}^{-\alpha x} - \mathrm{e}^{\frac{\mathrm{j}\omega x}{v_{n,\mathrm{sat}}}} \right) \mathrm{d}x + \mathrm{j}\omega\epsilon_s \int_0^W E(x,\omega)\mathrm{d}x \\ =& \frac{q v_{h,\mathrm{sat}} G_o(0,\omega)}{\mathrm{j}\omega - \alpha v_{h,\mathrm{sat}}} \mathrm{e}^{-\alpha W} \left[\frac{\mathrm{e}^{-\alpha(x-W)}}{-\alpha} - \frac{\mathrm{e}^{-\frac{\mathrm{j}\omega(x-W)}{v_{h,\mathrm{sat}}}}}{-\mathrm{j}\omega/v_{h,\mathrm{sat}}} \right]_0^W \\ &+ \frac{q v_{n,\mathrm{sat}} G_o(0,\omega)}{\mathrm{j}\omega + \alpha v_{n,\mathrm{sat}}} \left[\frac{\mathrm{e}^{-\alpha x}}{-\alpha} - \frac{\mathrm{e}^{\frac{\mathrm{j}\omega x}{v_{n,\mathrm{sat}}}}}{\mathrm{j}\omega/v_{n,\mathrm{sat}}} \right]_0^W + \mathrm{j}\omega\epsilon_s [-V]_0^W \end{aligned} \tag{2.59}$$

The last formulation of the current density is this one:

$$\begin{aligned} J_t(\omega) = &\frac{1}{W}\frac{qv_{h,\mathrm{sat}}G_o}{\mathrm{j}\omega - \alpha v_{h,\mathrm{sat}}}\mathrm{e}^{-\alpha W}\left(\frac{\mathrm{e}^{\alpha W}-1}{\alpha} + \frac{1-\mathrm{e}^{\frac{\mathrm{j}\omega W}{v_{h,\mathrm{sat}}}}}{\mathrm{j}\omega/v_{h,\mathrm{sat}}}\right) \\ +&\frac{qv_{n,\mathrm{sat}}G_O}{\mathrm{j}\omega + \alpha v_{n,\mathrm{sat}}}\left(\frac{1-\mathrm{e}^{-\alpha W}}{\alpha} + \frac{1-\mathrm{e}^{\frac{\mathrm{j}\omega W}{v_{n,\mathrm{sat}}}}}{\mathrm{j}\omega/v_{h,\mathrm{sat}}}\right) + \mathrm{j}\omega\frac{\epsilon_s}{W}[V(0)-V(W)]. \end{aligned} \tag{2.60}$$

One last step before the conversion of the density in the current quantity, we can introduce the transit times through the use of the space and of the velocity:

$$\tau_{dr,n} = \frac{W}{v_{n,\mathrm{sat}}}, \quad \tau_{dr,h} = \frac{W}{v_{h,\mathrm{sat}}} \tag{2.61}$$

and the complete expression for the optical generation rate which was never specified before in our discussion.

$$G_o(0,\omega) = \eta_Q\frac{(1-R)}{Ahf}\alpha P_{in}(\omega) \tag{2.62}$$

What we finally obtain is:

$$\begin{aligned} I_t(\omega) = &\alpha W\frac{q}{hf}\eta_Q(1-R)P_{in}(\omega) \\ \times&\left\{\frac{\mathrm{e}^{-\alpha W}-1}{\alpha W\left(\alpha W - \mathrm{j}\omega\tau_{dr,h}\right)} + \mathrm{e}^{-\alpha W}\frac{\mathrm{e}^{\mathrm{j}\omega\tau_{dr,h}}-1}{\mathrm{j}\omega\tau_{dr,h}\left(\alpha W - \mathrm{j}\omega\tau_{dr,h}\right)}\right. \\ +&\left.\frac{1-\mathrm{e}^{-\alpha W}}{\alpha W\left(\mathrm{j}\omega\tau_{dr,n}+\alpha W\right)} + \frac{1-\mathrm{e}^{\mathrm{j}\omega\tau_{dr,n}}}{\mathrm{j}\omega\tau_{dr,n}\left(\mathrm{j}\omega\tau_{dr,n}+\alpha W\right)}\right\} + \mathrm{j}\omega\frac{A\epsilon_s}{W}V_A(\omega) \\ = &-I_L(\omega) + \mathrm{j}\omega C V_A(\omega). \end{aligned} \tag{2.63}$$

In this formula we can see the small-signal short-circuit photocurrent $-I_L$ which is the bracket term and the current absorbed by the intrinsic layer geometric capacitance which is the last term. This in particular can be neglected when a DC bias is applied. In particular for $w \lim 0$ the equation simplify to:

$$I_t(0) = -I_L(0) = -\frac{q}{hf}\eta_Q(1-R)P_{in}(0)[1-\exp(-\alpha W)] \tag{2.64}$$

The small-signal photocurrent can be written as:

$$\begin{aligned} I_L(\omega) = &\alpha W\frac{q}{hf}\eta_Q(1-R)P_{in}(\omega) \\ &\times\left\{\frac{1}{\alpha W - \mathrm{j}\omega\tau_{dr,h}}\left[\frac{1-\mathrm{e}^{-\alpha W}}{\alpha W} + \mathrm{e}^{-\alpha W}\frac{1-\mathrm{e}}{\mathrm{j}\omega\tau_{dr,h}}\right]\tau_{dr,h}\right] \\ &-\frac{1}{\alpha W + \mathrm{j}\omega\tau_{dr,n}}\left[\frac{1-\mathrm{e}^{-\alpha W}}{\alpha W} + \frac{1-\mathrm{e}^{\mathrm{j}\omega\tau_{dr,n}}}{\mathrm{j}\omega\tau_{dr,n}}\right]\Bigg\} \end{aligned} \tag{2.65}$$

Even if the response is quite hard from an analytical point of view and it should be solved numerically in theory, there are a bunch of cases in which we can write an expression. For this discussion we shall start from the normalized responsivity, which is equal to:

$$\begin{aligned}\mathfrak{r}(\omega) =& \frac{I_L(\omega)}{I_L(0)} = \frac{1}{\alpha W - \mathrm{j}\omega\tau_{dr,h}}\left[\frac{1}{\alpha W} + \frac{1-\mathrm{e}^{\mathrm{j}\omega\tau_{dr,h}}}{\mathrm{j}\omega\tau_{dr,h}}\frac{1}{\mathrm{e}^{\alpha W}-1}\right] \\ &- \frac{1}{\alpha W + \mathrm{j}\omega\tau_{dr,n}}\left[\frac{1}{\alpha W} + \frac{1-\mathrm{e}^{\mathrm{j}\omega\tau_{dr,n}}}{\mathrm{j}\omega\tau_{dr,n}}\frac{\mathrm{e}^{\alpha W}}{\mathrm{e}^{\alpha W}-1}\right]\end{aligned} \tag{2.66}$$

We can now show two cases in which we have an overall simplification and we are able to achieve an analytical formulation: 1 diode is thick, so we impose $\alpha W \gg 1$, we have the absolute value of the responsivity:

$$|\mathfrak{r}(\omega)| \approx \left|\frac{\sin\left(\frac{\omega\tau_{dr,n}}{2}\right)}{\frac{\omega\tau_{dr,n}}{2}}\right| \tag{2.67}$$

and the 3 dB bandwidth condition with epsilon $\xi = \mathrm{j}\omega\tau_{dr,n}/2$

$$20\log_{10}\left|\mathfrak{r}\left(\omega_{3\ \mathrm{dB},tr}\right)\right| = 20\log_{10}\left|\frac{\sin(\xi)}{\xi}\right| = -3 \tag{2.68}$$

i.e.,

$$\frac{\omega_{3\ \mathrm{dB},tr}\tau_{dr,n}}{2} \approx 1.391 \rightarrow f_{3\ \mathrm{dB},tr} = \frac{2\times 1.391}{2\pi}\frac{1}{\tau_{dr,n}} = 0.443\frac{v_{n,\mathrm{sat}}}{W}. \tag{2.69}$$

What we have come to is the transit time-limited cutoff frequency, which is dependent on the transit time of minority carriers in the illuminated part of the device. This is because electrons are the minority ones close to the p+ surface being it illuminated on the front. If we have instead a back illumination, on the n+ side, we will have the same formula related to saturation velocity of holes, which are this time the minority carriers:

$$f_{3\ \mathrm{dB},tr} = 0.443\frac{v_{h,\mathrm{sat}}}{W}. \tag{2.70}$$

Being in general holes slower, a front illumination fits more this device to obtain a better overall response and speed. If we are assuming the same transit time for both electrons and holes, we will then obtain this formula:

$$f_{3\ \mathrm{dB},tr} \approx \frac{1}{2.2\tau_t} \tag{2.71}$$

2 the diode can be defined thin, so with $\alpha W \ll 1$: the generation is much more uniform in the active intrinsic region and the frequency response is

limited by both electrons and holes at the same time. An approximation in this case is the subsequent:

$$f_{3\ \mathrm{dB},tr} = \frac{3.5\bar{v}}{2\pi W}, \qquad \text{where } \frac{1}{\bar{v}^4} = \frac{1}{2}\left(\frac{1}{v_{n,\mathrm{sat}}^4} + \frac{1}{v_{h,\mathrm{sat}}^4}\right). \tag{2.72}$$

The difference between these two cases is not that high in case of equal saturation velocity between carriers.

2.2.2 Capacitance effect

We can design the photodetector with an equivalent circuit and from this we can have an idea of the RC-limited cutoff frequency. If C_p stands for the external diode parasitic capacitance and R_S the series parasitic resistance, R_D the parallel one and C_J the intrinsic capacitance, we have

$$R_D \gg R_s, R_L, \tag{2.73}$$

and the formula of the cutoff frequency is given by a resistance that is the sum of the series one with the one deriving form the illumination $R \approx R_S + R_L$ and the capacitance is the sum of the intrinsic capacitance and the external one $C \approx C_j + C_p$, $C_j = \frac{\epsilon_s A}{W}$. The total cutoff frequency can be evaluated at a circuital level and an approximation can be done in this way $f_{3\ \mathrm{dB},RC} \approx \frac{1}{2\pi RC}$. The total cutoff frequency resulting from the transit time and RC effect can be evaluated at a circuit level; an approximate expression is:

$$f_{3\ \mathrm{dB}} \approx \frac{1}{\sqrt{f_{3\ \mathrm{dB},RC}^{-2} + f_{3\ \mathrm{dB},tr}^{-2}}} \tag{2.74}$$

2.3 Bandwidth-efficiency trade-off

Because both the efficiency and the cutoff frequency depend on the measures of the detector, we can say that there is a trade-off between these two quantities. We have seen both RC-limited bandwidth and transit time limited bandwidth: when we increase the thickness W we increase the first one because of the decrease of the junction capacitance and, on the other hand, when we decrease it there is an increase of the second contribution. If we think about the total area, its increase does not affect the transit-time, but it surely increases the capacitance, decreasing the total bandwidth and in particular the RC component. If we keep the area value constant, we can see the direct proportionality of the RC cutoff frequency with the thickness W and the inverse proportion with the transit time. From the moment that the total bandwidth is influenced by the lower one between these two, it is influenced by the RC component in case of low W and from the transit time

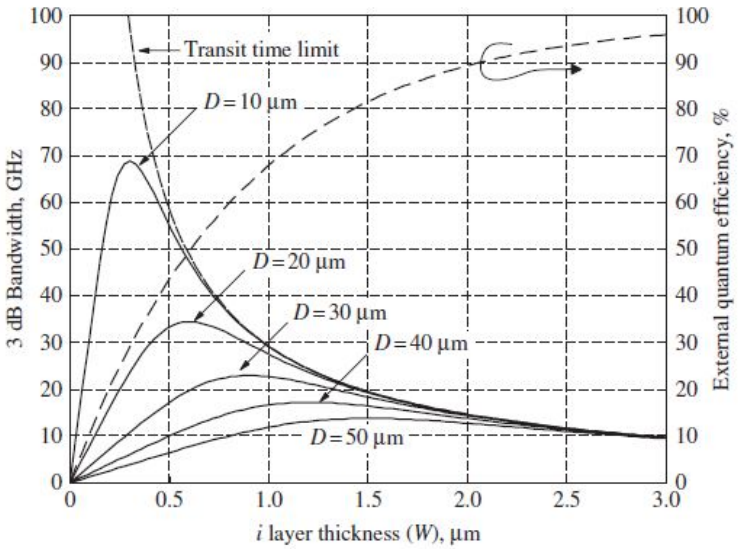


Figure 2.6: High-speed pin optimization: trade-off between speed and efficiency [14].

one in case of large W, so it increases with W and then decreases. We can suppose and imagine there is somewhere in the middle of these two different behaviors a maximum value whose position depends on the area. The maximum is going to move to lower W and larger cutoff frequencies for an increasing area and the efficiency will still increase with the width W. When we are working at high frequencies, the required aspect is a quite small area device with small W and poor efficiency. This trade-off can be shown with the Fig. figure2.6. is which we can better understand the trade-off between speed and efficiency and its meaning. This is a InGaAs pin photodiode with circular illuminated area and the D means diameter of this area. The total volume of the intrinsic region is exactly the area calculated as the area of a circle with that particular diameter multiplied for the thickness W of this region. We have considered for this plot only the intrinsic junction capacitance and not the parasitic one, which may have a fundamental role in many cases. For good bandwidth, we have to use small area devices, at least 20μm;. But, for decreasing illuminated area, we have to decrease also the W of the device in order to have both good efficiencies and speed, as far as for larger combination of measures the transit time limit become the limiting one making the bandwidth and the external quantum efficiency worse and worse.

2.4 Si-Ge photodiode, the choice of the materials and the heterostructure

The aim of all the research around photodetectors is to implement a reasonably low-cost and effective interconnections based on optical principles, due to the speed required to keep on the roadmap for the next evolution of computers. In fact, optical technologies are foreseen as the next step to overcome the bottleneck of metal interconnections, as we can assume the speed of light will provide to the device a much larger bandwidth than the slower physical interconnection. Those are not only meant to be cheaper and faster, but also consume less energy improving both thermal and power link budgets [12]. One of the most promising paths is the silicon photonics and germanium detectors are very attractive for many reasons: they are CMOS-compatible through mastered heteroepitaxy on silicon (refer to the figure with gap vs lattice constant) and a quite decent crystalline film quality, the absorption coefficient of Ge is very competitive for both 1.31 µm and 1.55 µm, being the absorption range high over a wide spectral range and the use of strain (compressive or tensile) can help the engineering of the gap. This is exactly what our device is based on: a waveguide photodetector. Those are able to optimize the frequency response and the responsivity. Another interesting aspect of germanium is the fast mobility value for both electrons and holes [12]. The central mechanism used for light detection is to transduce impinging photons with given energy into an electrical signals, typically current or voltage, which can then be collected by an electronic readout circuit. Indeed, an efficient photon-to-electron conversion depends on the electronic bandgap of the active material used (see Fig. 2.7), because if the energy is too low we cannot create an electron-hole pair and if it is too high there is a waste of energy with many other effects in order to relax the created pairs to an equilibrium value of energy in correspondence of the bands. We can do now a little digression of what this kind of heterostructure means from a different point of view. The heterostructure is defined when two different crystals with different lattice constants are grown one on top of the other. Due to this difference there are misfit dislocations, that are essentially traps for both electrons or holes and these make the performances of a generic device worse. Essentially, we can call two crystals a heterostructure when, in the process of growth, there is a very little difference in the lattice constant. We can take advantage of this differences from an electronic and optical point of view. The discontinuity at the interface can be used as a confinement for carriers from a band structure point of view because of the jump we have in one or in both bands. The heterstructures can be defined with three different types: straddling gap (type I), staggered gap (type II) or broken gap (type III). Depending from the type of heterojunction there could be, from the band diagram point of view,

the confinement of one or both carriers in one semiconductor respect to the other. Instead, if we look at it from an optical point of view, we can have the confinement of the radiation due to the difference in refractive index between the two different semiconductors. The lattice can be matched or it can have a little mismatch which causes some kind of strain (compressive or tensile). This little mismatch can be used for a confinement, that is especially useful in the photodetectors in order to have the generation of the pairs only in the desired intrinsic active region (in the case of pin). A double heterojunction such as AlGaAs/GaAs/AlGaAs creates a potential well in the band diagram and can be engineered to construct a multi quantum well or even a superlattice when the number of wells starts to increase dramatically, with even more quantum effects to be introduced and considered, which significantly change the electronic properties of the device. Now we focus now on the Si-Ge interface that is present in our device. The light arrives from the silicon waveguide, while the evanescent coupling happens in the germanium photodetector. We can show in Fig. 2.8 the difference of band gap in the two semiconductors. The main difference between the two is the fact that germanium is an almost direct material, while silicon is an

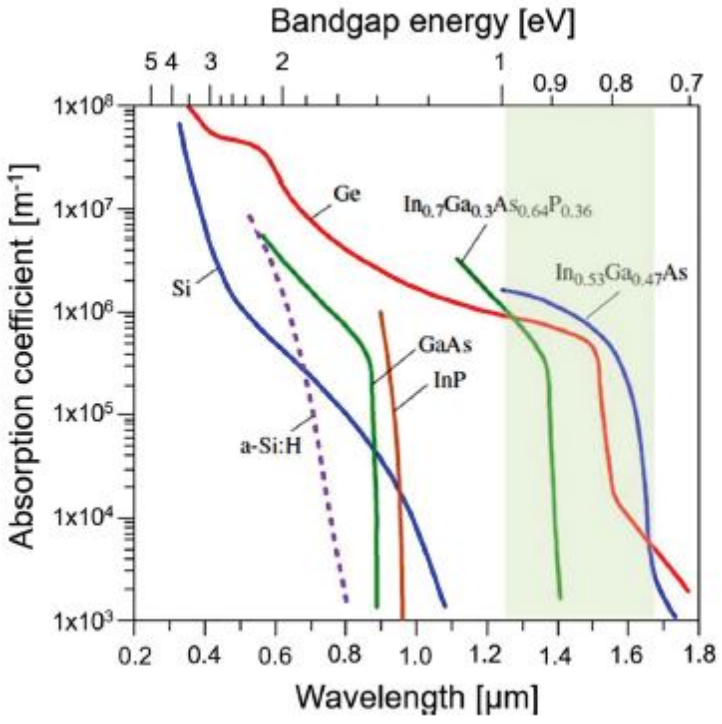


Figure 2.7: Absorption coefficient as a function of the wavelength and bandgap energy for several semiconductors typically harnessed in optical photodetectors.[12]

indirect one. We can easily understand why the semiconductors are used in this way. The germanium is commonly used as a detector for the above mentioned properties, for its easy and well-explored growth and integration with silicon and with the contemporary technology. The waveguide is done in silicon both for its transport properties and for the growth methods on silicon wafers. But how the growth can be made? What we want for our device is a direct growth of germanium over silicon for the fabrication of our detector. This is not trivial because of the difference in lattice constant between the two semiconductors, which is equal to 4.2 %. The lattice constant is equal to 5.431 Å in Si and 5.658 Å in Ge, determining a high density of misfit and threading dislocations in the germanium side of the interface. If a potential higher roughness can be a problem for the fabrication itself of the photodetector, all the consequent defects can impact the performances of the device by increasing the dark current and decreasing the mobility of both carriers. If one can think the solution is a graded layer of Si-Ge buffers with a percentage that makes the growth of germanium on silicon with no lattice mismatch and with the reduction of defects at the interface, this is not anymore the solution. In fact, the recent solutions were the use of a detector directly on a silicon waveguide in order to have, as in our case, an evanescent coupling. The efficiency of this connection is precluded by the use of buffer layers, so the growth need to be direct on the waveguide and a two-step deposition technique is used in order to mitigate as much as possible the mismatch of lattice constants. This approach exclude the creation of islands during the chemical vapor deposition (CVD), giving a smooth, high quality germanium. The first step is the formation of a seed layer at low temperature in the 320-450 °C range, making the defects accumulate in it and creating a regular array of misfit dislocations 10 nm apart, decreasing

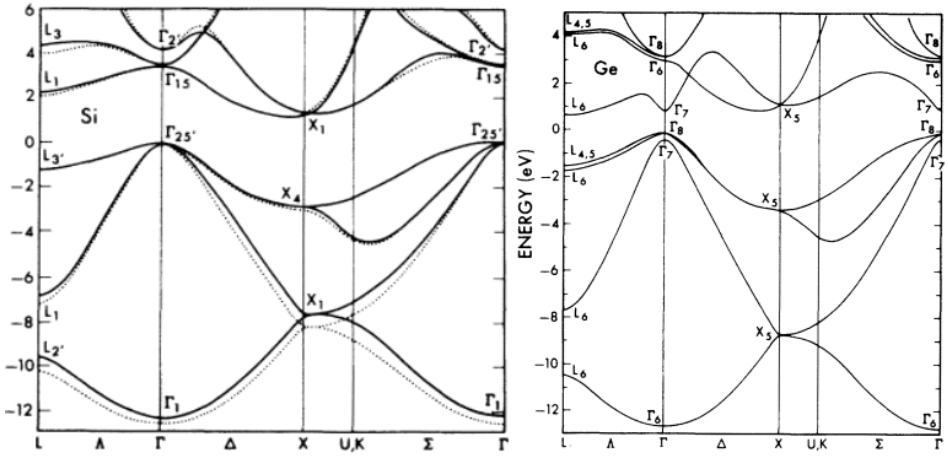


Figure 2.8: Band diagram comparison between germanium and silicon.

the roughness due to the elastic strain relaxation. In a second step, the used temperature is way higher and between 600 °C and 850 °C with the formation this time of germanium more rapidly and with an overall better quality. A thermal cycling can be even be used after these steps in order to further reduce the number of defects. Despite the initial design for this kind of detectors as discrete normal incidence devices, this is now an outrun idea for chip integration. The main problems were the thick layer needed to absorb the incident light making the dark current quite high, the footprint on the silicon substrate large, a high capacitance which limited the bandwidth and a limited responsivity-bandwidth product. As the capacitances limit the speed, the sensitivity is then limited by the dark current, making an overall bad performance for this kind of devices respect to other kind of photodetectors or materials. A turning point were the waveguide-integrated photodetectors, being their active area smaller, increasing speed, low-noise and sensitivity. In this configuration, the collection path are orthogonal one to each other, with a great advantage respect to normal incidence ones. We have in this way a new degree of freedom to optimize responsivity and bandwidth more independently. The coupling of the germanium detector with the waveguide can be done with two different couplings: evanescent or butt-coupling. The first one is the coupling used in our actual real device studied in this thesis and in particular the waveguide is positioned under the detector with a pin configuration of this heterojunction. In fact, the p doped part create an ohmic contact at the very end of the Germanium close to the metal part, while the intrinsic part is the active region where we have the absorption of the main part of the light coming from the coupling. The heterojunction is between the intrinsic part of the detector and the waveguide silicon. On the other hand, a butt-coupling is different and can also have better absorption per device length. In general, the design of Ge high-performance photodetectors are created with front-end-of-line ion implantation and epitaxy processes, while metal contacts on it, as in our case tungsten and copper, are fabricated with the back-end-of-line CMOS metalization (Fig. 2.9). There are two main configurations of Ge integrated on Si waveguides: the MSM (metal-semiconductor-metal) and the pin (see

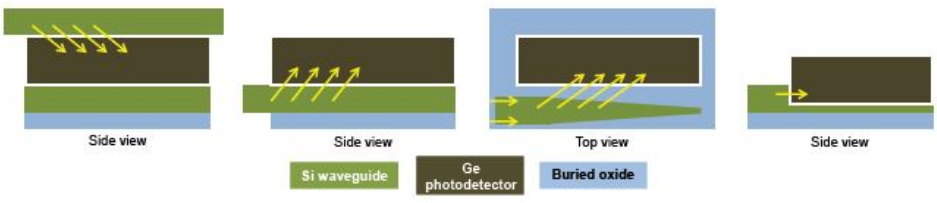


Figure 2.9: Evanescent-coupling schemes: top-to-down, down-to-top, and side-coupling. Butt-coupling scheme. [12].

Fig. 2.10). if the first is constructed as interdigitated metal layers on top of the Ge with Schottky barriers, its easiness in the fabrication and its low-bias operation are good points. But, despite improved responsivity and fast responses were achieved with time and research, they still show high dark currents due to low Schottky barriers: there is an long-term low signal-to-noise ratio, high power utilization and a relatively poor sensitivity when the optical input power is not so consistent. These are the main reasons why the pin device is more challenging and more studied now. There are two main possibilities: a lateral scheme and a vertical one, which is the one used in our device, and the one we will focus on. Another degree of freedom is the homo and hetero-junction in the detector. Our device belongs to the first category since we have only germanium for both the intrinsic active region and the doped contact one, but let us see the main concerns and advantages of this choice. As first, there is a possible lack of confinement due to the weak index contrast between the intrinsic and the doped regions, meaning there is also, due to the coupling, the generation of carriers' pairs in the doped region, leading to higher recombination of photogenerated carriers with consequent degradation of the responsivity. The heavy doping of the p region is a symptom of performance drop: lower responsivity, reduced speed, and higher noise. Even with higher dark current, reduced responsivity, and slower response, homojunctions are still a good choice for fast detection systems. The hetero-junction version with silicon on top as a doped region exploits the benefits of its processing, making it more simple to produce lowering wafer-level production costs. From a performance point of view, there is an higher optical confinement in the intrinsic region, leading to lower losses in the absorption because of the less significant recombination. Due to the second hetero-junction aside from the one with the waveguide, both the responsivity and the bandwidth are low when we are not applying any kind of bias due to the limited collection thanks to the energy barriers. A voltage has to be applied in order to have a good working point for these devices. In general, pin photodetectors have an insufficient electrical output levels and there is a need for other electronic stages attached to the chip. The receiver input-referred noise is prevailing on the intrinsic detector noise limiting the sensitivity [12].

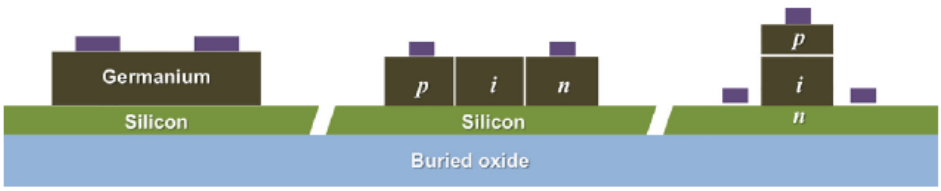


Figure 2.10: Common types of waveguide-integrated Si–Ge photodetectors. MSM structure, lateral, and vertical p-i-n diodes.[12]

Table 3.1: WPD geometry.

	W_{Ge}	H_{Ge}	W_{doping}
Device 1	$4\,\mu$m	$0.8\,\mu$m	$3\,\mu$m
Device 2	$4\,\mu$m	$0.8\,\mu$m	$3.5\,\mu$m

Chapter 3

Results

In this chapter we will focus on the results that we obtained during the thesis work, starting from the description of the geometry, the figures of merit and concluding with the transient analysis. The results are first obtained on the two-dimensional version of the devices, then moving on to the complete structure.

3.1 Structure

The structure of our photodiode is a Ge-on-Si device, with a vertical configuration. It is a waveguide photodetector, meaning that the silicon waveguide, below the Germanium block, illuminates the device with an evanescently coupled mode. The light-matter interaction that appears in the photodetector allows the photogeneration of carriers. In the Germanium absorber, it is important that the doping increases in a limited region and close to the contact in order to create an ohmic contact[21]. The width of this doped region is quite important for the frequency response of the device and the second device we have data on differs from the first one exclusively for this measure. It is necessary that the width of the metal contact is narrower than this doped region if we want something relatable and with a good response. The difference between the two devices (as we can see in Fig. 3.1) is exclusively the width of the doped region, which is equal to 3 µm in the first device and to 3.5 µm in the second one. Respect to what we have written in the photodetector general description, we can say that these devices have a waveguide coupled below it with an evanescent kind of coupling. From

a semiconductor point of view there is a heterojunction between the silicon waveguide and the germanium detector, but there is a homostructure between the top doped layer and the intrinsic active one. The absence of difference of reflective index between these two parts of the germanium detector make the spreading of the evanescent coupled light all over the device, causing diffusion currents in the highly doped region and even more recombination in this area with an overall increase of the dark current. For every preliminary simulation, as already said, we have used a two-dimensional structure derived from the complete one for both devices. This was done by deleting the z axis and maintaining each measure for what concern the x and y axis unaltered. In order to have some significant simulations, it is obvious that the two-dimensional cut (see Fig. 3.2)of the structure was performed in a point in which the metal of the contacts is present (the definition of the contacts is obligatory in order to perform any kind of electrical simulation). The main reason for the two-dimensional transformation of the project is done in order to study the transient, because of the excessive time needed for a three-dimensional simulation. As we are going to see, the study of this analysis and of the one of the break criteria was quite long and it required a huge number of simulations, for example, to push the study of the convergence of the problem due to the rise time to its limit. But, one important fact was to understand if the 2D simulation was on point or if this would have been not sufficiently precise and coherent with what we obtained in

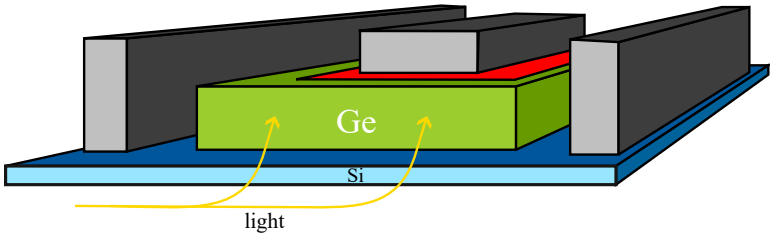


Figure 3.1: 3D photodetector geometry.

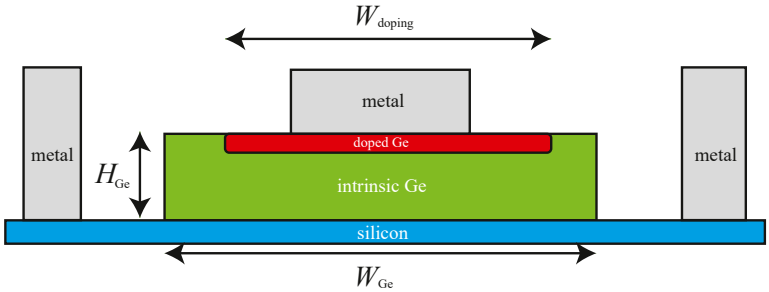


Figure 3.2: 2D transverse cross section of the 3D photodetector geometry. The length of the devices is 15 µm.

the simulations of the full device. If from the dimensions point of view, this is quite easy, and we just have to keep the same absolute values for what concern the doping concentrations, there is some more reasoning on the illumination of the device. In fact, Rsoft FullWave is not able to work on a two-dimensional device and this makes us think of another method to recreate the same photon density in the devices. For sure, we already know it is impossible to recreate the same mode as in the evanescent coupling through the waveguide, so we have explored two main paths:

- use the maximum value of photons present in the photodetector through the three-dimensional Rsoft FullWAVE coupling at 200 µW
- use a medium value of the same quantity in the same conditions

This last value has been obtained through the integration tool of *SVisual*, by simply integrating the OpticalGeneration quantity, which is derived from the RSoft FullWave simulation. This value, for the 1.31 µm is equal to a medium value of $4.675 \times 10^{21}\,\mathrm{cm^{-3}s^{-1}}$ photons, which is what we are going to use as a constant rate in all the Germanium of the 2D device as we are going to use the medium value instead of the maximum one.

3.2 Figures of merit and preliminary analysis

Once we have completed the two dimensional structure with all the aspects including dimensions, doping and illumination, we can now move on to some figures of merit and on some analysis on both the 2D case and the 3D on. This is done in order to understand both the closeness of the simplified device and the actual values of the detector to be then compared with our transient analysis. So, we can start from the description of the two-dimensional "try" devices in order to understand how to use the transient simulation properly and its relation compared to its properties. The first ever analysis we make is a *quasistationary* in dark conditions. This is useful to have a correct value of the dark current and of all the microscopic quantities when we apply different negative voltages to the device. The first thing this command will calculate is the equilibrium condition: in dark and with no applied bias, to study the built-in potential and other quantities. Once Sentuarus has this, it is able to perform a number of steps of the bias in a certain range of motion up to the desired value of voltage. We have to explain better what range of motion is, because these are not the appropriate words: each step of the bias to higher value, aside from the first one that is surely $1 \times 10^{-4}\,\mathrm{V}$, can have a certain value which goes from a minimum value called $MinStep$ to a maximum value called $MaxStep$ and in this analysis those are equal to $\mathrm{V}1 \times 10^{-6}\,\mathrm{V}$ and V0.1 V. The dark current value we obtain (Fig. 3.3) will be quite important also for the transient analysis, because we can use it for the break criteria. This can be an easy control parameter in order to understand

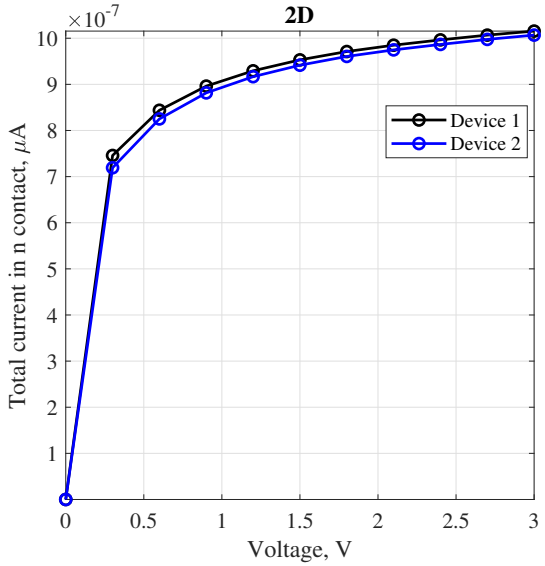


Figure 3.3: Dark current comparison in the two 2D devices.

when we can stop the simulation after the transient fully exhausted after the light input went off.

The dark current is negligible in the equilibrium case: a voltage is required in order to see some electric field in the Germanium region, some movement and some recombination event in dark conditions. The dark current is equal to $1.015\,483 \times 10^{-6}$ µA in device 1 and it is equal to $1.006\,763 \times 10^{-6}$ µA in device 2. The difference between the two devices in dark conditions is minimal.

The second aspect we can analyse is the illuminated current value of the device to actually know the maximum current value we can reach in case of an illumination with the power we have studied. This is rather important and it can also carry out the role of a control parameter to know exactly when the transient simulation current reaches its higher possible value, and we can stop the transient rise part and proceed with the falling one, reducing the computational cost immensely. We perform the same *Quasistationary* command with the optical generation rates set as described before, while we put actually 0 photons and 0 generated pairs in the Silicon. We had some studies to justify this choice, but since a generation value up to 1×10^{21} centim^{-3}s^{-1} did not make any difference in the electro-optical response of the device and since it has very small values also in the 3D simulation project with RSoft FullWAVE simulating the incoming light from the waveguide and the evanescent coupling.

As expected (Tab. 3.2) there is no major difference for different biases, but there is a great difference between the two devices. The W_{doping} doping

Table 3.2: Photocurrent in the 2D case with $G_0 = 4.675 \times 10^{21}\,\mathrm{cm}^{-3}\mathrm{s}^{-1}$.

Bias	Device 1 Photocurrent	Device 2 Photocurrent
0.0V	$2.807\,058 \times 10^{-3}\,\mathrm{cm}^{-}3\mathrm{s}^{-}1$	$2.807\,061 \times 10^{-1}\,\mu\mathrm{A}$
0.8V	$2.808\,033 \times 10^{-3}\,\mathrm{cm}^{-}3\mathrm{s}^{-}1$	$2.807\,163 \times 10^{-1}\,\mu\mathrm{A}$
1.5V	$2.808\,182 \times 10^{-3}\,\mathrm{cm}^{-}3\mathrm{s}^{-}1$	$2.807\,240 \times 10^{-1}\,\mu\mathrm{A}$
2.0V	$2.808\,273 \times 10^{-3}\,\mathrm{cm}^{-}3\mathrm{s}^{-}1$	$2.807\,299 \times 10^{-1}\,\mu\mathrm{A}$
3.0V	$2.808\,697 \times 10^{-3}\,\mathrm{cm}^{-}3\mathrm{s}^{-}1$	$2.807\,522 \times 10^{-1}\,\mu\mathrm{A}$

difference plays a huge role in the photocurrent for what concern the two-dimensional structures. This kind of simulation is not only interesting for the current value, but also because we can calculate the responsivity and we can have a measure of it for our device. We already described responsivity as the ratio between the output current and the input power and it is important to have a good value. The results we get are a $2.807 \times 10^{-5}\,\mathrm{AW}^{-}1$ for device 1 and $2.807 \times 10^{-3}\,\mathrm{AW}^{-}1$ for device 2, which is a reasonable number if we think the fact that the device is very small and the volume illuminated by the coupling is really little and uniform, so the real problem of this value is the fact that we are considering a two-dimensional structure. The difference is evident and it is exactly equal to the difference in photocurrent between the two devices. This problem is due to the two-dimensional device and its deal with the highly doped region close to the contact. We can now look at the bandwidth of the device done by the small-signal analysis in order to see its speed. The electro-optic response was performed for frequencies going form $1 \times 10^{8}\,\mathrm{GHz}$ to $1 \times 10^{11}\,\mathrm{GHz}$ with 40 points in the interval and this is sufficient in order to obtain a smooth result without the necessity of longer and more computational heavy simulations. What we could expect is to see differences when we apply different biases, because the higher the bias, the larger the depleted region and the electric field, leading to a faster drift motion of the pairs generated in the intrinsic region and a generally faster response. As first comparison (see Fig. 3.4), we can see how much a different bias change the response:

The bigger difference is between the 0 V case and the cases in which we have an applied voltage. This proves us another time that the correct working of the detector happen when there is an applied reverse bias of some measure. Another important fact is that the difference in GHz is always reducing between different curves as long as we consider higher biases, meaning that an even higher voltage value is not useful for a better speed of the photodetector and this also proves that the choice of $-2\,\mathrm{V}$ in all our 3D simulations is correct.

We can now think about the real three-dimensional device and look another time into all the figures of merit to better understand the device

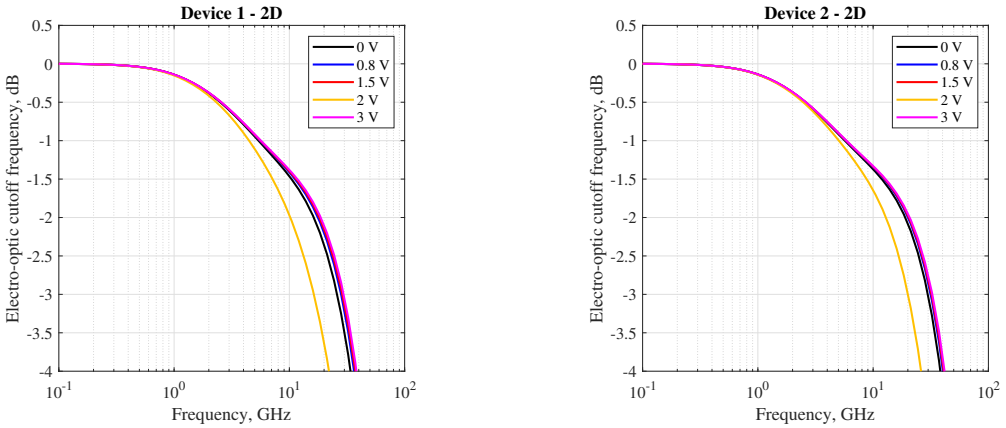


Figure 3.4: Frequency response for the two 2D devices for different biases for a constant generation rate of $4.675 \times 10^{21}\,\mathrm{cm}^{-}3\mathrm{s}^{-}1$.

Table 3.3: Frequency response in the 2D case with $G_0 = 4.675 \times 10^{21}\,\mathrm{cm}^{-3}\mathrm{s}^{-1}$.

Bias	Device 1 Cutoff frequency	Device 2 Cutoff frequency
0.0V	16.31 GHz	20.04 GHz
0.8V	25.94 GHz	29.81 GHz
1.5V	28.10 GHz	31.52 GHz
2.0V	28.81 GHz	32.09 GHz
3.0V	29.56 GHz	32.62 GHz

and then we will see if the results we are going to obtain in the transient analysis are comparable and coherent or not. We can follow the same path and order done before for the device with one dimension less. The The dark current is equal to $1.403\,477 \times 10^{-5}\,\mu\mathrm{A}$. We can now use the power-current file in order to have the value of the maximum photocurrent the device can reach. Those values are important to describe the device and to correctly run the transient simulation. This is a necessary value mainly for the break criteria, which allows us to perform much faster and more automatic simulations without the necessity of putting the correct times of the rising and falling edges in the physics section of the common file of the simulation. With these two values, we have the boundaries of the current curve in the transient simulation and this lets us understand exactly when the rise or fall are fully completed. The photocurrent has a value of $159.29\,\mu\mathrm{A}$ as we can see in Fig. 3.6. Another time, from the power-current curve we can easily calculate the responsivity as we have already explained multiple times and the value for the three-dimensional device is equal to 0.7963 A/W. As written in [12], a good value could be 0.7 A/W for devices

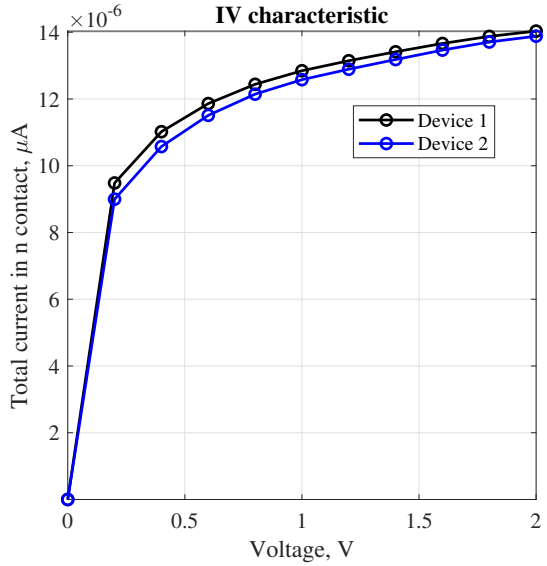


Figure 3.5: Dark current in the 3D device at a reverse bias voltage of 2 V.

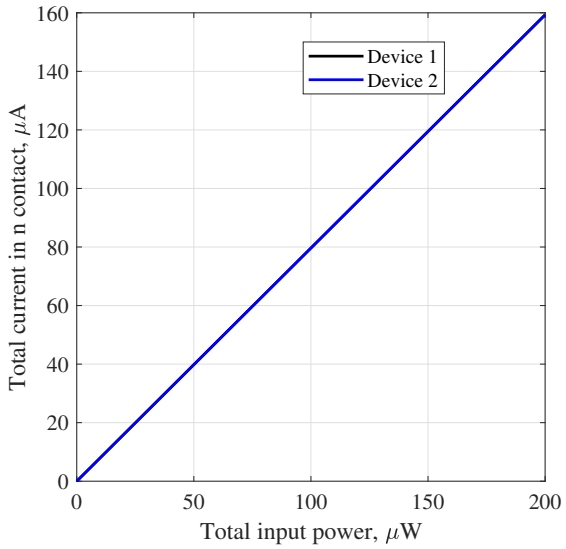


Figure 3.6: Photocurrent in the 3D device at reverse bias of 2 V for an increasing input optical power.

under 20 µm and this is coherent with what we get. We have to analyse another important figure of merit, the bandwidth and the cutoff frequency. This is important because it is comparable not only with the experimental results, but also with the values we are going to obtain in the transient simulation. The electro-optic response is shown in Fig. 3.7. It is interesting

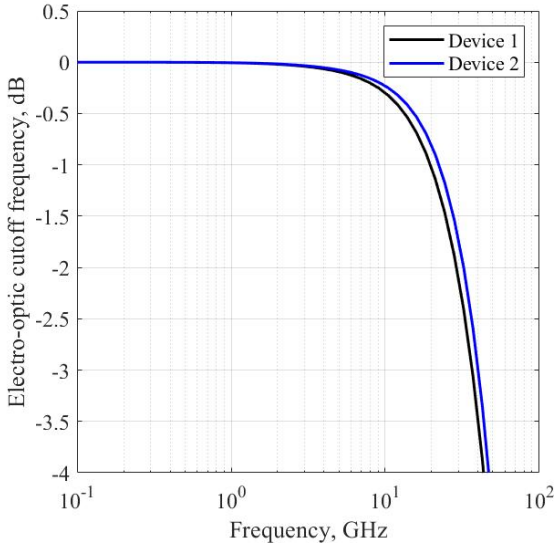


Figure 3.7: Frequency response in the 3D device at a reverse bias of 2V for an input optical power of 200 µW.

to see what is the limiting factor in this response. In order to do that, we used a in-house developed model using information from the structure and from the electrical response of the devices. The plots we have done are relative to the variation of one of the main measures that characterize the device and we can see if the choice of a 0.8 µm height is optimal. The values for the constants are took from bibliography of the Germanium, such as the mobility, the velocity of carriers. In order to evaluate the transit time, we need some values such as the height of the device, the absorption coefficient, the velocity of the carriers and the frequencies we are considering: we already have everything form the specifics of the device and from the characteristics of the germanium. For what concern the evaluation of the RC cutoff frequency, there is something more to evaluate. In fact, because of the fact we are using the analytic formula for the extraction of electrical parameters, we need the values of the capacitances and of the resistances to model our device. Even if the photodetector capacitance value can be calculated starting from its dimensions with the formula (we consider it as a capacitance with parallel plates):

$$C_{PD} = \epsilon_0 \epsilon_r \frac{W_{Ge} L_{Ge}}{H_{Ge}} \tag{3.1}$$

and the load resistance is equal to 50 Ω from the measurements, we have to find out the values of the parasitic resistance and capacitance of the device. Those are possible to obtain with an electrical AC analysis of the devices. This analysis is done with no optical excitation and what we ob-

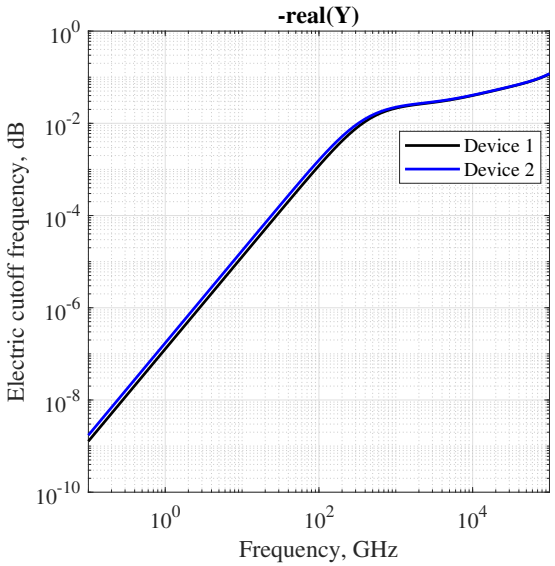


Figure 3.8: Real part of the admittance of the two devices.

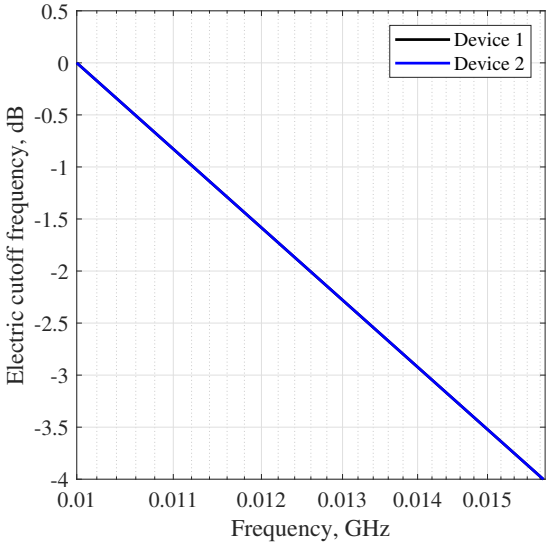


Figure 3.9: Electrical frequency response.

tain are the admittance Y and the frequency response with its own cutoff frequency(Fig. 3.8 and Fig. 3.9, respectively). As we can see, the response of the two devices is very close and seeking for an order of value for the parasitic capacitance and resistance, we can use 1 fF and 50 Ω, respectively. We are now able to compute the RC component of the cutoff frequency. Once we have both, we can finally see which is the predominant in our devices and

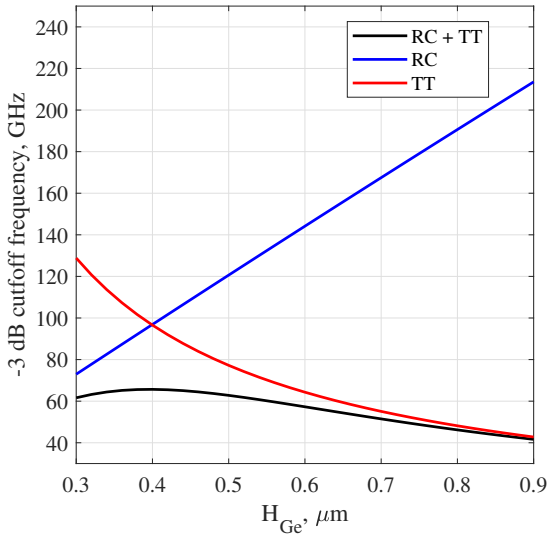


Figure 3.10: Cutoff frequency for different H_{Ge}.

if a change in the dimension can change them. We can, at first, modify the height of our device. We already know in advance that this will change the photodetector capacitance, because it changes the height of the capacitor with two parallel plates. On the other hand, this will change also the values of the constant of time tau used in the transit time cutoff frequency and the height value is also present in the formula itself. Therefore, there will be a change in both cutoff values and logically also in the combined one (see Fig. 3.10). The situation is a little different in case we enlarge the photodetector, increasing the dimension of W. This measure do not influence in any way the transit time of the carriers, because their lifetime is related only to the height of the device (as well as the characteristics of the material). The main difference is present in the RC of the device, because we are changing the area of the ideal capacitor: C_{PD} will be lower with a consequent difference in the cutoff frequency related to RC effects(Fig. 3.11). We can see that the transit time cutoff is effectively not influenced, while there is a change in the RC component, but with not significant changes in the total cutoff: the decrease is quite slow.

3.3 Transient analysis

After a first work on the study of certain parameters such as the velocity saturation (that we are reporting later), we started to effectively see the transient analysis in a two-dimensional framework. Once understood the values to use in order to have a converging simulation, we have then switched

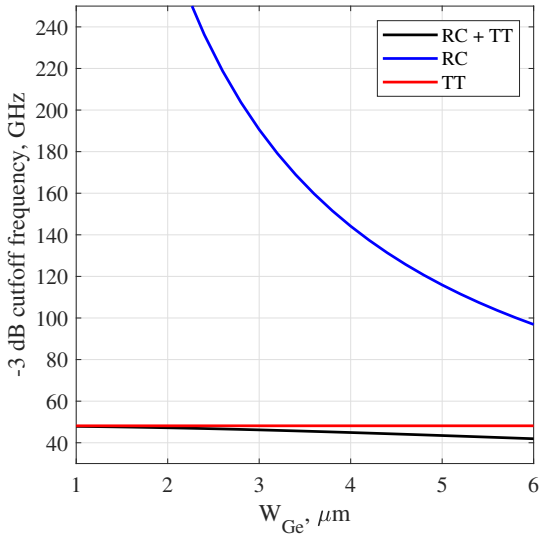


Figure 3.11: Cutoff frequency for different W_{Ge}.

to the three-dimensional one.

3.3.1 2D

The first results we are going to see regard the transient simulation for the two 2D devices. One of the bigger works of this thesis is the study itself of the analysis: due to the computational time in the three-dimensional device, the use of the two-dimensional structure was crucial. The initial tests were performed to understand the parameters we could control. We can represent in the next figure the current response and the optical impulse we are giving to the device in order to see their correlation. We are reporting the normalized plot of the current transient at the n contact deriving from the represented optical pulse with the shape of a step function. We could in principle increase the slope of the pulse, since, as we can clearly see, it is not vertical at all, but it would be meaningless because the current of the detector is now following at the maximum possible velocity. This time analysis gives us also many information from the frequency point of view as reported in the Fig. 3.12. Aside from all the control parameters we can use, we can focus on the main information we get to optimize the simulation. There are few important points on which we have to focus:

- shape of the response
- minimum value
- maximum value

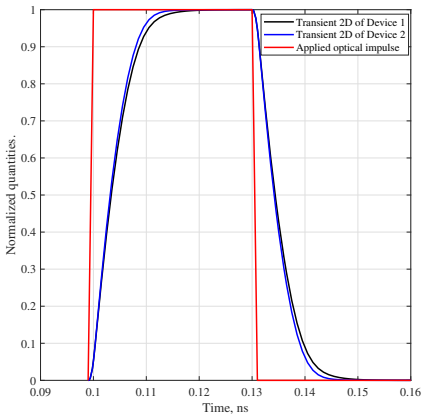


Figure 3.12: Shape of the transient respect to the optical pulse for the 2D devices.

This figure is made by the 2D version of the two devices. We can already see that there is a difference in the transit time between the two devices. The shape of the current is following the optical pulse with a certain delay due to the frequency limiting effects of the device, such as the transit time and the capacitance (this last one is not the main one even if the detector is quite high). Another important quantity is the minimum value of the current, which coincides with the dark current, which is the current of the device when there is no light impinging on it and that, as we are going to see in the next section, can be evaluated easily with a simple *Quasistationary* simulation. This is the value of both the beginning of the analysis and of the end, so when the transient has exhausted after the impulse is removed. This is important because when the current reach this value, we are sure the transient is completely finished and we can interrupt the simulation before it goes on until our time limit value. We can say something quite similar on the maximum value, in fact, when the current reach the maximum value it can reach under illumination, we know that the transient finished arriving to the maximum value corresponding to that kind of illumination and we can analyse the time necessary for the transient to follow up the light impulse. Another time this will be useful in the break criteria to stop the simulation and cut out the pulse and make the falling transient part, without losing computational time with the detector in a static situation after the illumination and with no source of information for our analysis. Some results can be shown and obtained already in the 2D simulation, such as the 10 % and 90 % current of the curve in the rising part, because it tells us the time constant of the device and consequently its cutoff frequency. The results of these 2D structures are not fully comparable with the one we are going to obtain in the 3D devices mainly because of the constant

optical generation rate used in the two-dimensional structure, even if the difference is not that important. But this is actually not relevant, being all the analysis performed with these structures only functional to the study of the simulation. The obtained figure reported before is the result of many simulations and the fact that it uses a quite wide optical input is merely done in order to clearly see what happens in general as a response of the device.

3.3.2 Break criteria

A big amount of work was done on the break criteria and its study was quite huge in terms of time. We have to remember that the use of the two-dimensional version of the two devices were functional also for this study. This is something we can introduce in the transient simulation for two main reasons: decrease the computational time and the computational cost. A first definition is the subsequent: it allows us to stop the simulation once some values you have pre-fixed are reached before the end of the simulation. In our specific simulation, we know that after a certain transient of time, the current will cap to a maximum, which is correspondent and equivalent to the photocurrent of the device. Before entering more into details, we have to explain the problem we encountered already in the first transient simulations. As we already specified in the introductory part of the transient simulation, the definition itself of the transient needs some values to be specified firsts. While in the *cmd* file of the sdevice we have to define the initial and the final time of the simulation and the time steps in order to have a coherent discretization of the problem, in the physics file we need to define many parameters for what concern the optical pulse used. In the definition of the optical pulse, as we have seen in the chapter on the transient analysis, we must specify the time length of the pulse and its slope in the rise and fall parts. The study of the $WaveTLin$ was done in order to have a light impinge as instantaneous as possible, but we could not push this value more than a certain limit for two main reasons: convergence problems and unnecessary computational cost when it was too fast. The simulation time step needs to be way lower than the rise time, in order to have many points in this range and the possibility to follow and converge. But, doing this we increase dramatically the computational cost having serious impact on the time necessary to perform the simulation in the 3D case. The study of this parameter in order it to be fast enough to see the faster possible transient of the current with an acceptable convergence time was long and necessary. On the other hand, the definition of the length of the pulse in order to exhaust the transient before the falling edge was quite a problem. In fact, we could not know before how much it would have taken and its definition was always larger to avoid problems with the simulation. The point was that, if in the 2D simulation we could change it manually many

times to obtain a correct pulse range because of the speed of the simulation, it is not possible to do the same thing int he 3D simulation. We wanted to automatize the simulation in order to obtain the correct one for each device we want to simulate without the necessity of many simulations to obtain a correct result. What we need are two values: the photocurrent one, which is the maximum the device can reach with the specified bias and illumination conditions and the dark current, which is the current of the device in the same bias conditions without any kind of illumination. The photocurrent can be calculated with the use of the $PIFR.plt$ file, which is the *Quasistationary* simulation with the costant illumination in the case of the 2D structure or with the one deriving from the RSoft simulation. The evaluation of these exact value from the *Quasistationary* simulation is implemented directly in Sentaurus with the support of an external code written in MATLAB (in this case specifically, but it could also be an other programming language) put in a Bash node. By simply running this bash command before the transient simulation we obtain, not only the photocurrent, yet it is easy to also achieve the dark one. Actually, using the $REVERSEBIAS.plt$ file, we have the electrical simulation of the device without any optical type of source, therefore by definition the dark current, which is the maximum current of the device when it is on in absence of any source of light in case of the reverse bias we are using. This second value is quite important because it represents the other control parameter for the transient break criteria. We can now introduce two break criteria in the transient simulation. The first one correspond to the maximum photocurrent, which is the maximum value the transient reaches after a certain time. This interrupts it the first time, stopping the optical pulse and setting up the simulation for the second part: the falling one. We can make it restart without any kind of illumination from the same value of current, so there is the fall. The second break criteria is put for the dark current value, which is then reached after the device is not anymore exposed to an illumination source. This part would in general continue until the final time of the simulation and another time, this will not give us any useful information. We are in this case able to stop the simulation before the time expires and we reduce dramatically the computational time of the simulation, especially when we are going to come back on the 3D devices in the next section.

3.3.3 3D results

We can start now the discussion about the results obtained for the two devices. But first, we have to better explain how we can calculate the frequency from a transient analysis. In order to have a time constant to see how fast the device is, we have to calculate the rise time of the rising edge. The rise time is defined as the time between the 10 % and 90 % of the curve, in this case, the photocurrent. In fact, for this experiment, we

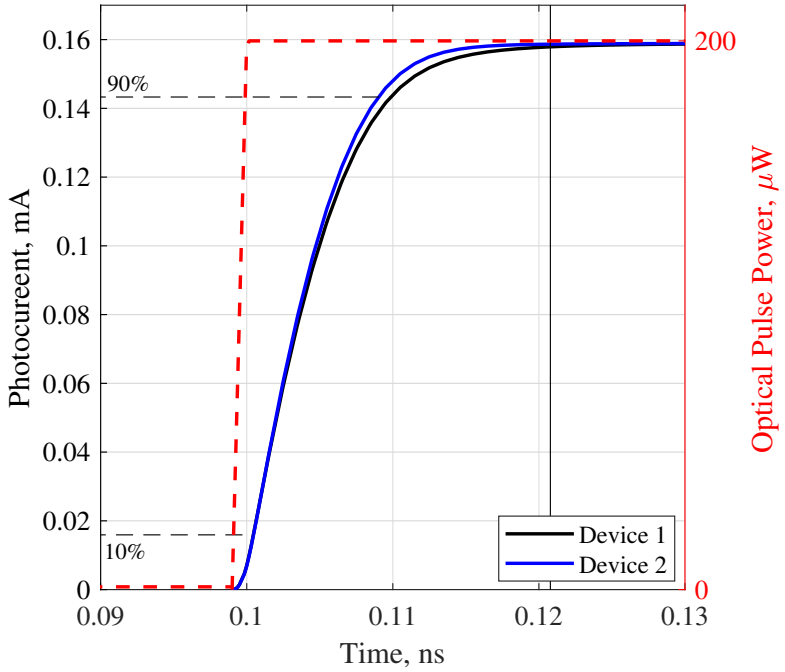


Figure 3.13: Transient simulations performed with a reverse bias voltage of 2 V [22].

simulated a simple rising signal applying a low input optical power signal and looking at the output current on a 50 Ω load resistor, which is the same used by Cisco Systems for their experiments. The rise time is particularly important not only for the time constant of the device, but also because from this value we directly derive the bandwidth of the device. There is a quite simple approximated formula that approximates quite well the single pole response of a device and tells us the cutoff frequency:

$$f_{3dB} \cong \frac{0.35}{\tau_r} \tag{3.2}$$

We can use this conversion through the rise time analysis to compare the transient analysis with both the small-signal analysis and the experimental data we have. The optical pulse has a power measured at the beginning of the detector of 200 µW, starting at 0.1 ns. This signal power is compatible with the 0 dBm optical power measured at the laser output, assuming −7 dBm for coupling and waveguide losses. So, we have described the conditions of the simulation and we have calculated the photocurrent rise time from Fig. 3.13, which is equal to 9.43 ps in the device 1 and to 8.58 ps in the device 2. As we have argued, we can tell that these two values correspond to electro-optic cutoff frequencies of a single-pole transfer function of 37.1 GHz

and 40.7 GHz, respectively. At first we have to reason a little more on the transit time we have obtained: is it a reasonable value? In order to understand this we can calculate a theoretical value using a quite simple formula, knowing that the time is the space divided the velocity, as we know from middle school.

$$t_{\text{transient}} = \frac{H_{\text{Ge}}}{v_{\text{Sat}}} \tag{3.3}$$

If we use the thickness of the intrinsic Germanium absorption layer as space to be traveled by the further away carriers and we assume the velocity to be capped at the saturation velocity value, we can substitute these values. For the height of the device $H_{\text{Ge}} \approx 0.8\,\mu\text{m}$, while the saturation velocity is equal to $v_{\text{sat}} \approx 7 \times 10^6\,\text{cm}\,\text{s}^{-1}$ [23]. What we obtain is:

$$t_{\text{transient}} = \frac{H_{\text{Ge}}}{v_{\text{sat}}} = \frac{0.8\,\mu\text{m}}{7 \times 10^6\,\text{cm}\,\text{s}^{-1}} \approx 11\text{ps} \tag{3.4}$$

This result is absolutely comparable and compatible and we have the certainty it is correct. For what concern the frequency point of view, we can now compare the obtained results with both experimental results and small-signal analysis ones, as reported in Fig. 3.14. These results are compared with the state-of-the-art small-signal electro-optic simulations and experimental data [24]. Measurements and simulations show a $-3\,\text{dB}$ cutoff frequency between 37 GHz and 37.5 GHz for Device 1, between 41 GHz and 41.5 GHz for Device 2 and this is very close to the estimated response from the transient simulation of 37.1 GHz GHz and 40.7 GHz, respectively. This difference is nominal, but we can account for it in the type of simulation carried out. In fact, the small-signal simulations are fundamentally limited to the cyclo-stationary steady state, they do not account for effects related for example to slower carriers, that can be studied starting from the current tails of the transient curve. And this is also one of the reasons why we are investing time in the transient: this leads us to deeper analysis in many aspects and the elaboration of the data coming from a time-domain simulation can actually give a new point of view and new information. This can improve the modeling of the devices and their optimization to a whole different and even more advanced level.

3.3.4 Fourier Transform and small-signal comparison

In this section, we are going to analyse the comparison between the transient analysis and the small-signal one for Device 1 only. We are moving forward concerning the mere analysis of the cutoff frequencies. What we want to compare is now the transient curve, to see if the small-signal is able to represent everything for small input power or if it is present some additional component as we could see from different time constants. The first step for this discussion is the creation of an ideal input signal, whose parameters were

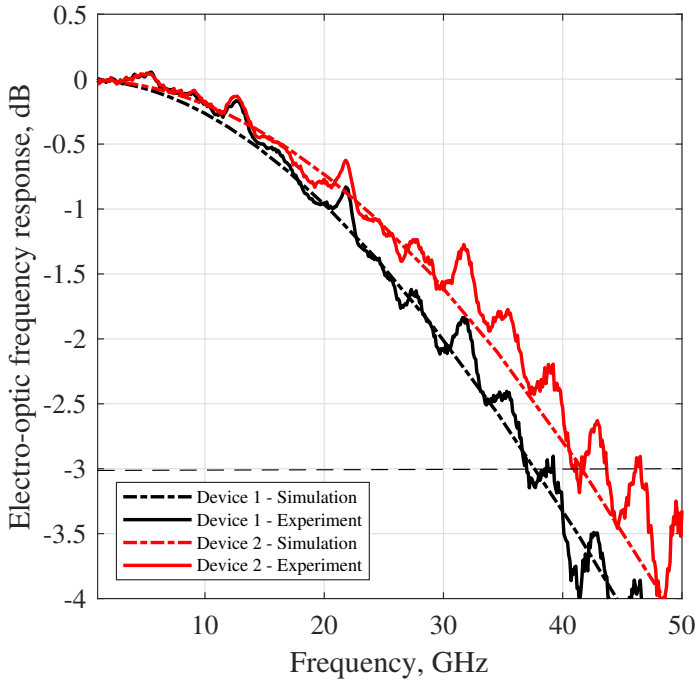


Figure 3.14: Measurements and simulations of the electro-optic frequency response of the two devices considered, performed at a reverse bias of 2 V and a laser power of 0 dBm.[22]

taken from the transient current file. From the time variable, we were able to obtain both the period of this signal and its duty-cycle. From the fact that we know also the frequency response, we can perform a correct choice of the number of periodicity of the signal, the sampling frequency which requires to be higher than the considered frequencies, and as a consequence, the time step is chosen as the inverse of the sampling frequency. The time step is little, so we have a time vector with a lot of points concerning the files we extract from Sentaurus. Once we have chosen the periodicity, we can apply it to the input signal and finally perform the fast Fourier transform to it. From the MATLAB and FFT manuals, we can perform the calculations for the single-sided representation of the Fourier transform we just got. We can now move on and focus on the frequency response. Since it is discrete, we are fitting this with a single-pole low-pass filter, making the operations we are going to do better and possible to perform. The curves are almost overlapping, so it is a good approximation. The frequency response has now to be multiplied with the Fourier transform of the input signal created similar to the one Sentaurus is giving as illumination (we are still using the linear function we have explained in the transient analysis). We have to better explain why are

we doing this multiplication. Starting from the transient analysis, when we apply the Fourier transform to the current respect to the time, we obtain a curve that is not comparable with the frequency response deriving from the small-signal analysis. This happens because we have not only the output component, but we have also the Fourier component of the input signal that is increasing the overall cutoff frequency we obtain. What we should do is the division of the Fourier transform (i.e.fast Fourier transform) of the transient current for the Fourier transform of the input optical signal, which is quite close to a step function. The methodology used in this analysis is different, in the sense that does not want to obtain the frequency response, but the transient current in the time domain. We have to multiply the FFT of the small-signal analysis with the FFT of the input signal we have created starting from the data we have from the *plt* file. Once we have multiplied these two, we have the Fourier transform of the transient starting from the frequency response of the device and by using the inverse fast Fourier transform (IFFT), we obtain the current in the time domain we can finally compare with the transient analysis we have obtained in Sentaurus. In the IFFT we are forcing the 'symmetric' command to avoid some numerical noise that could be present. We are allowed to do this because of the properties of the Fourier transforms, in fact, for real-valued $f(x)$, the equation:

$$\hat{f}(\xi) = \int_{-\infty}^{\infty} f(x) e^{-i2\pi\xi x} dx, \quad \forall \xi \in \mathbb{R} \tag{3.5}$$

own the symmetry property $\hat{f}(-\xi) = \hat{f}^*(\xi)$, with $\hat{f}^*(\xi)$ complex conjugate of $\hat{f}(\xi)$. The Fourier transform is now written as:

$$\hat{f}(\xi) = \begin{cases} \int_{-\infty}^{\infty} f(x) e^{-i2\pi\xi x} dx, & \xi \geq 0 \\ \hat{f}^*(|\xi|) & \xi < 0 \end{cases} \tag{3.6}$$

which means that negative values of the frequency are no more necessary for the description of the Fourier transform [25]. We then proceed with the normalization of the transient curve, its interpolation with more points for a better representation, and its periodization with the same number of periods as the created input signal (sufficiently high number). We are now able to compare the two normalized curves. What we have finally obtained is Fig. 3.15, in which we can see how much the comparison for a small optical input power between the transient curve and the curve derived from the Fourier transform of the small-signal frequency response. In the figure, there are shown only two periods, but the input signal is repeated more times, a number sufficient to make this comparison relatable and meaningful. The comparison is made possible by the normalization of the total current I_{tot}, that is the photocurrent exiting from the device after we use as input the declared input power, divided by its maximum value, so the maximum current the device achieves after a certain time the device is exposed to the

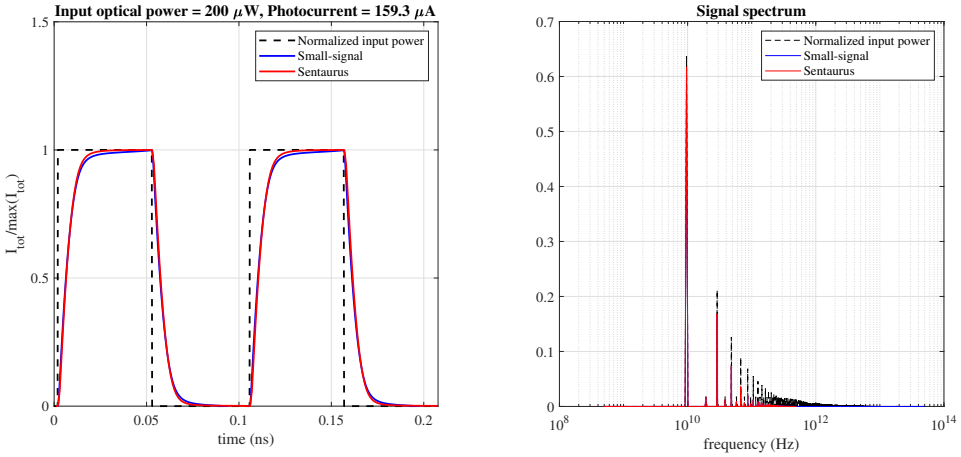


Figure 3.15: Transient comparison between the Sentaurus file and the Fourier transform of the small-signal file for an optical input power of 200 µW.

illumination. What we have also in the plot is the signal spectrum of the two different curves and we can compare them in order to see if there is a particular difference in some kind of harmonics or if the difference is distributed over the whole spectrum. Another important comparison we can perform to create a bridge with the next section is the same kind of comparison in a case in which the optical input power is higher than before. As expected, the difference between the two curves at low power is rather low, as the small-signal linearization is correct in the neighborhood of the working point. This is not anymore true as we are rising the input power value since there are many more parasitic and capacitive effects that are not taken into account with a small-signal analysis. The small-signal circuit and model are no more correct, and a large-signal analysis is more congruent. The procedure in order to obtain the same results is the same, but we have now upscaled the value of the optical input power to 2 µW in both the RSoft simulation, which is used by the transient, and by the frequency opto-response simulation, which value of input power ramps up to this higher value. What we have obtained is presented in Fig. 3.16. The main consideration we can do is the observation of the transient curve: as we are going to discuss in the next section, it is not so much different from the one obtained before, but the cutoff frequency we obtain between the 10 % and 90 % of the curve is present. On the other hand, the obtained cutoff value from the small-signal analysis is very different, since the decrease of its cutoff is rather higher than what we expected. The cutoff frequency is now 11.0 GHz, which makes the

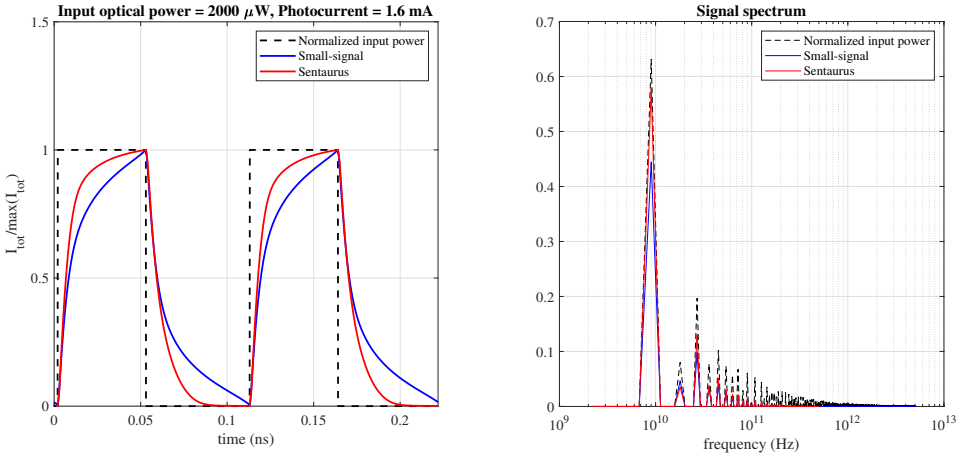


Figure 3.16: Transient comparison between the Sentaurus file and the Fourier transform of the small-signal file for an optical input power of 2 mW.

device impossible to use with a larger optical input power.

3.4 High optical input power

An interesting possible application for the transient analysis is to study devices in case of high powers. In fact, in this condition the small-signal analysis performed with the *QuasiStationary* command could become useless when we exaggerate the input power values and the transient can help us to solve this issue. The hypothesis behind the small-signal approximation in fact is a linearization in the neighbourhood of a working point, so we assume that the device response is linear. This is not anymore valid in some cases we are going to analyse, so a large-signal analysis in the time-domain, which has no constrain on the initial hypothesis for its validity, is a good choice to take into account lots of effect that make the detector work differently from before. This could look like a mere simulation to magnify and celebrate the transient analysis which is the topic of this thesis, as we have found a case in which the frequency response can not anymore be calculated and compared with a small-signal analysis, but that is not the case. In fact, we have analysed high-power applications after many studies present in literature on germanium over a silicon waveguide photodetectors, making this an interesting research field for the use of this kind of applications. Aside from the comparison with other materials, there is a significant upgrade in this kind of applications of a mode evolution for Ge-on-Si photodetectors over a butt-coupling scheme. This is base on a silicon bus waveguide directly

incident on the Ge-on-Si structure. With this coupling method, all of the light in the bus waveguide is transferred into the photodetector directly at the waveguide-detector interface, causing two main problems:

- butt-coupling excites modes in the detector with vastly different propagation constants: strong modal interference with high peak intensities occur.
- all of the light is transferred into the germanium at once, so most of the absorption occurs only in the first few micrometers of the detector.

Together, these two issues will produce discrete locations of high power densities in the germanium, leading to some saturation effects. Trap sites on the germanium-silicon interface are occupied for longer fractions of time due to a high generation rate of free carriers, leading to decreases of the recombination rate of minority carriers and of the responsivity at high input powers. Additionally, a large electron-hole pair density in a single location creates a large gradient of charge, inducing a strong electric field opposing the applied bias and lowering the actual electric field in the depleted region. This effect, called carrier screening, reduces the bandwidth at high powers because the carriers are not anymore efficiently swept out of the detector [26]. The total bandwidth of these type of detectors drops to values that make them unusable in cases of large powers, while the mode evolution ones have a far lower decrease of the cutoff frequency. To have an idea in term of numbers, as reported in the experiment performed in [26], the bandwidth of a certain Ge-on-Si detector is equal to 40 GHz for both the devices in the two configurations for an optical input power equal to 20 µW. For a increase up to 4 mW, the two detectors have very different behaviour and response to this increase. If in case of a mode evolution based detector the cutoff frequency drop to 31 GHz, which is still a solid value, in the case of a butt-coupling the bandwidth is equal to 0.7 GHz, not even comparable in terms of performance with the other device in this conditions. This improvement is due to the more uniform absorption of optical power in the germanium part, significantly reducing the effects of carrier screening observed in the butt-coupled one, maintaining its characteristics even at high input powers, with weaker saturation effects. Furthermore, the compact mode evolution-based coupler takes little additional space and also adds no complexity to the detector, so it can easily replace butt-coupled devices. All these characteristics indicate that this coupling scheme for Ge-on-Si detectors can be useful for many integrated optical systems in the fields of microwave photonics, optical communications, and optical sensing that demand both high-power and high-speed devices [26]. Going back to our analysis, the first analyzed thing is the relation between the optical input power and the electro-optic frequency response. Even if this device is designed and used for Silicon Photonics (SiPh) applications where the powers involved are limited due to

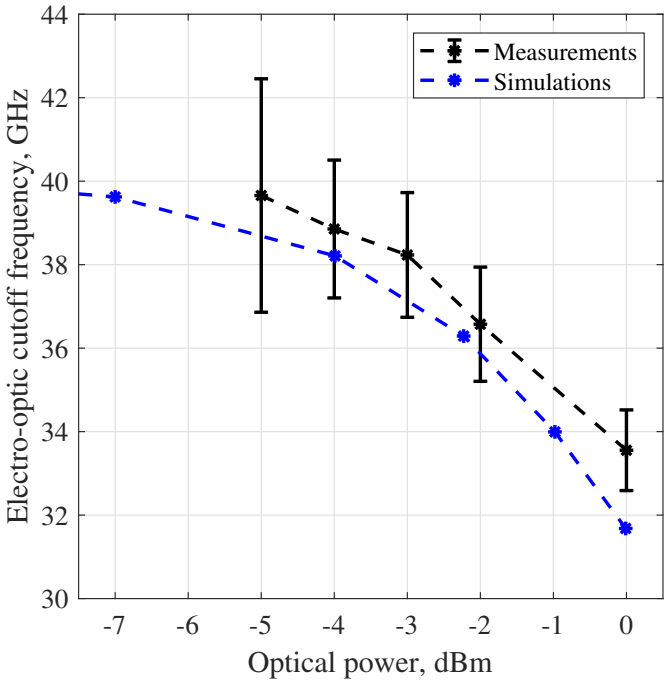


Figure 3.17: Comparison of simulations and measurement of the device with a reverse bias voltage of 3 V [29].

the low-power consumption requirement, these detectors can be also coupled directly with an optical fiber instead of a waveguide and the power we are achieving is much higher. In this analysis, we used a multiphysics modeling approach, where the solution of the optical problem is used to calculate an input optical generation rate term, which is then included as a source term in the drift-diffusion model. Simulations were performed with Synopsys TCAD Sentaurus [27] and Synopsys RSoft FullWave, and are validated against device fabricated and characterized by Cisco System with a Keysight LCA [28] with a bandwidth up to 50 GHz on 5 nominally identical devices from different parts of the wafer. Light is coupled to the device via a dielectric waveguide and a taper directly into the Si substrate. The measurements presented deembedded and they take into account both the loss in the waveguide and the loss of the coupling with the fiber used for the measurements. Fig. 3.17 shows a comparison between the measurements and the small-signal simulation for different optical powers. As we increase the input optical power considered from -7 dBm to 0 dBm, a reduction of the electro-optic cutoff frequency appears, reducing the cutoff frequency of almost 10 GHz, which is a huge value. This decrease can tell us that the velocity of the carriers decreased considerably in the Germanium region,

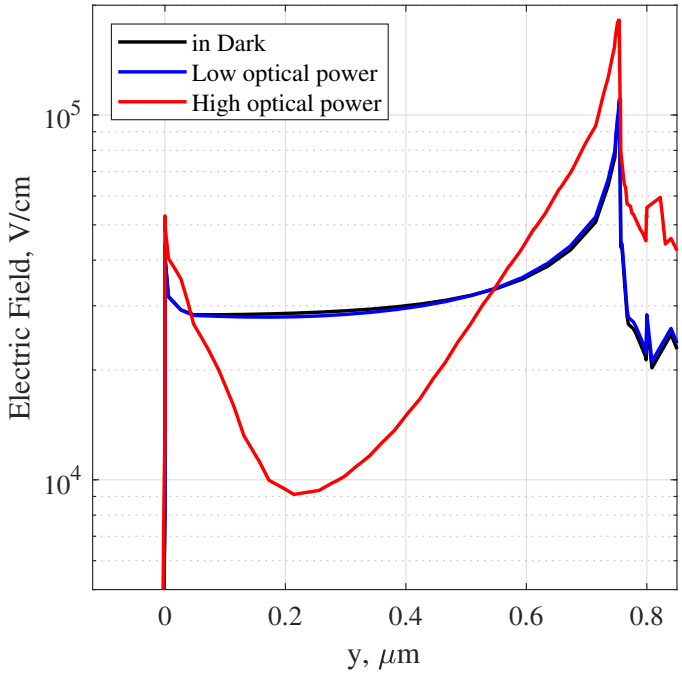


Figure 3.18: Vertical 1D cut at $z = 1\,\mu m$ of the absolute value of the electric field. The field for low input optical power (10 µW) is close to the in-dark condition, while the field decreases for higher input optical power (500 µW) [29].

therefore the first quantity we wanted to plot was the electric field in the device (see Fig. 3.18, which report a 1D cut). The figure reports the absolute value of the electric field in 1D vertical cut in the germanium and we can see that for 10 µW the electric field is close to the one in dark conditions, while for 500 µW the field is way lower and it is even reduced by three times, leading to slower drift component of the movement of the carriers and an overall slowdown of the carriers. The high power applications can be explored in another way. We can use the transient simulation in order to look at the rising edge of the current consequent to the optical pulse. In fact, when we increase the power way more than we have done before, the small-signal analysis makes no more sense and the transient simulation can be a good choice for the calculation of the cutoff frequency of a device. In this case the simulations are performed with a reverse bias of 2 V and very different input optical powers ranging from 200 µW, which is the one we have always used in all the other results, to 2000 µW. This choice has been done to better appreciate the differences between the different simulations in order to understand that the linear approximation we have assumed for

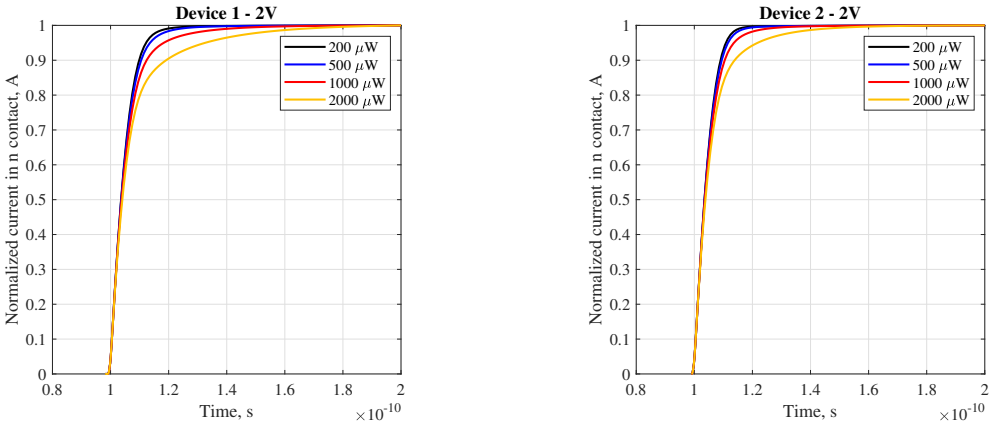


Figure 3.19: Transient simulation for higher input power values.

the quasi-stationary simulation is not possible anymore. What we obtain is the subsequent plot for the two three-dimensional devices (see Fig. 3.19): We see the transient become more and more flat and the main differences arise after half of the transient rising edge. This is not only important for the fact that the 10 % and 90 % difference is increasing for increasing input power, but also the transient rise is completed in much more time. If in our standard case it expires in 30 ps, in the case of 2000 µW it takes around 80 ps for device 1, while for device 2 the 200 µW takes 20 ps and the 2000 µW case takes up to 60 ps, which makes the device way slower and it requires a longer simulation to make the transient simulation start with the falling edge. We can now approach the calculation of the bandwidth. We compare the values obtained from the transit time with the ones deriving from a small-signal analysis of the same devices with the same optical conditions, so an input power of 200 µW, 500 µW, 1000 µW, 2000 µW, respectively. The meaning of the comparison in this part of the transient analysis is to prove and show that the results are reasonable, while the ones from the small-signal analysis is not. The transient analysis of high-power devices is possible. We can, at first, report the frequency responses of both the devices (Fig. 3.20). What we understand is the fact that a single-pole response in these illumination conditions is not possible and the results are completely wrong and far from reality. The measured cutoff frequencies, calculated in the same way we have done till now with the approximated formula from the current transient curve are reported in the next table (Tab. 3.4) alongside the small-signal ones: As we have already compared, the results for the 200 µW case are quite similar and both methods are even comparable with the experimental results given by Cisco Systems. For higher intensities, the small-signal analysis does not take into account many effects that shift the response far away from an ideal single-pole one, underestimating in each case the cutoff

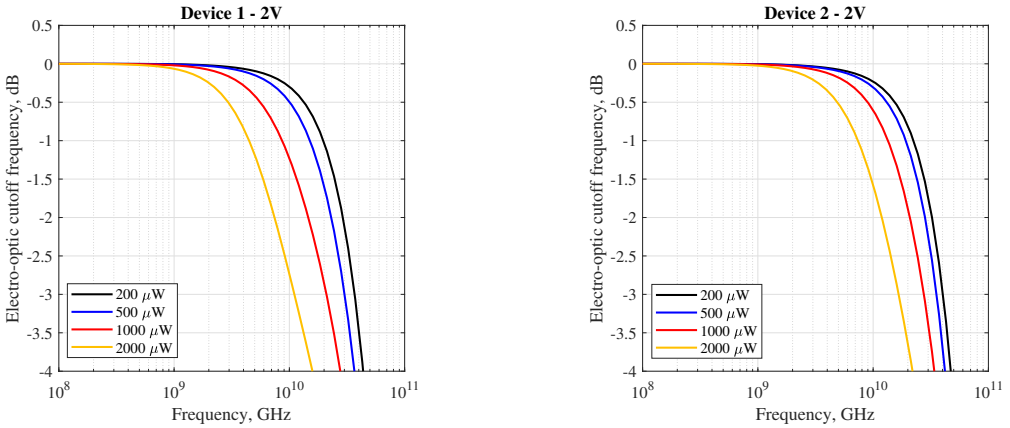


Figure 3.20: Frequency response for higher input power values.

Table 3.4: Frequency response for different power values for device 1.

	Device 1 Transient	Device 1 Small Signal	Device 2 Transient	Device 2 Small Signal
200 µW	37.1 GHz	36.8 GHz	40.3 GHz	40.3 GHz
500 µW	34.3 GHz	30.1 GHz	38.0 GHz	35.5 GHz
1000 µW	28.7 GHz	21.0 GHz	34.3 GHz	27.6 GHz
2000 µW	18.2 GHz	11.0 GHz	26.5 GHz	16.7 GHz

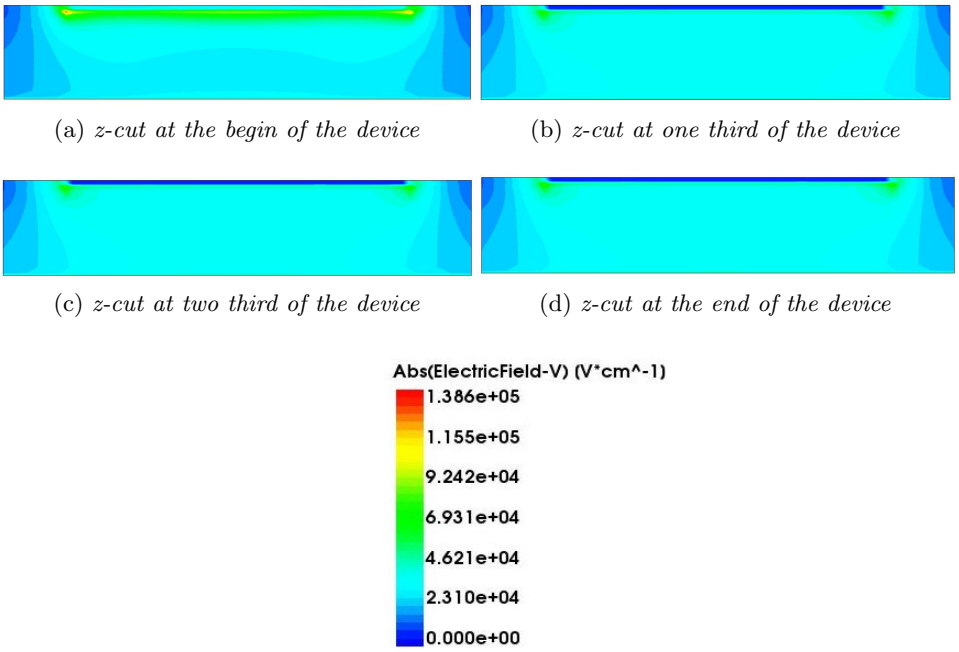
frequency The transient can be considered a large-signal analysis and it is able to correctly represent the problem due to the fact it is not linearizing in the neighbourhood of a working point. From these results it is obvious the drop in the bandwidth when the power increases leading even to a 15 GHz drop in the two furthest cases in both devices.

3.5 Optimization and exploration

In this part of the results chapter, we experimented with various changes in material values, models used, or device quantities. Since the transient analysis is not present in the literature, we tried to seek other applications and selling points of the transient analysis, trying to understand if it can be helpful in other contexts for a more complete analysis of some parameters.

3.5.1 Velocity saturation

The first analysis of the velocity saturation was done in order to gain confidence with the tools. We experimented all the simulations possible aside from the transient one, which was then performed once we had full confidence with the simulation. What we can report are for example the dark



(a) *z-cut at the begin of the device*

(b) *z-cut at one third of the device*

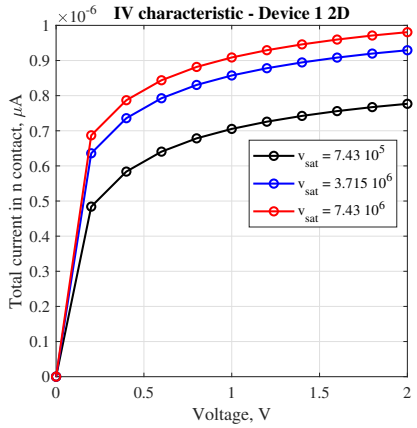
(c) *z-cut at two third of the device*

(d) *z-cut at the end of the device*

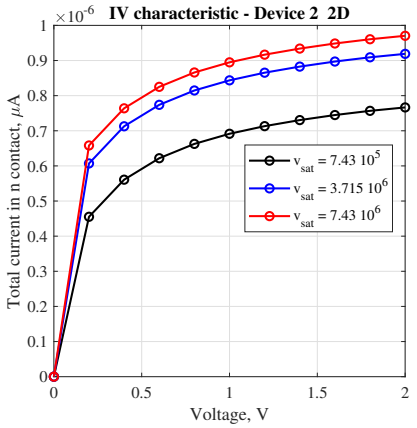
(e) *Electric field color map v*

Figure 3.21: Electric field for different cuts of the device.

current difference for different saturation velocities and how this last parameter can influence the working of the device in dark conditions, essentially when it does not receive any light. Since what we are considering is a dark condition with a reverse bias of 2 V, the electric field in the depleted region has a quite high value as we can see in the reported Fig. 3.21. Due to this, we can properly say that the current in the detector has a great drift component, which is the faster and main one. We can assume that the velocity of the carriers saturates quite easily to its maximum value in the middle of the germanium detector, so this is important not only because it can cap the speed of the carriers, but also because it limits the energy of the carriers and their velocity in this region, changing the type of generation or recombination events it can participate to. The first part of these plots and results concern the two-dimensional devices. There is substantially a change in the value of the current in dark conditions as we can see in the IV characteristic detailed in Fig. 3.22. Furthermore, what we obtained in the case of a small-signal analysis was exactly what we expected: for higher saturation velocity we have a higher maximum reachable velocity in the semiconductor. As carriers can move more rapidly, they are subject to less recombination events due to the lower time constant and they can be collected even more. The difference in the response is quite visible already for a drop of half of the value

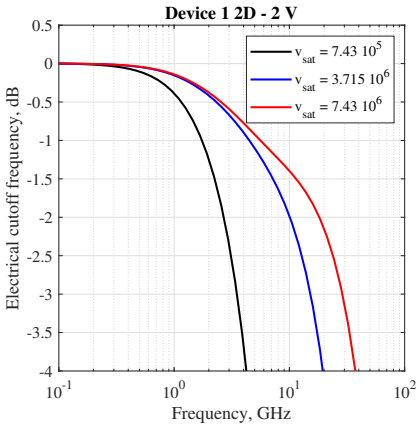


(a)Device 1.

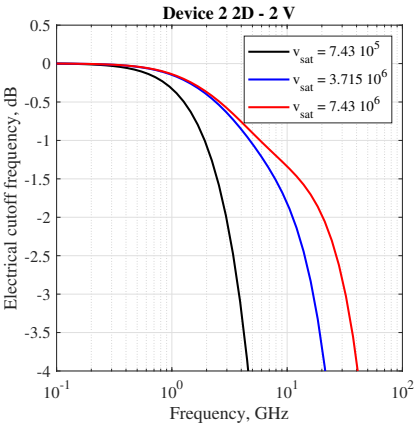


(b)Device 2.

Figure 3.22: Dark current for a reverse bias of 2 V for different velocity saturation values in 2D device 1 and device 2, respectively.



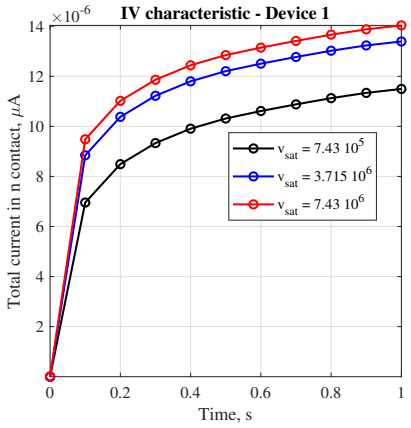
(a)Device 1.



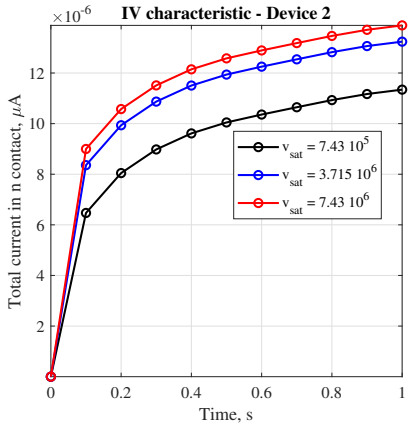
(b)Device 2.

Figure 3.23: Frequency response for a reverse bias of 2 V for different velocity saturation values in 2D devices.

of the velocity saturation parameter in the physics common file used in the simulation. The drop of 10 times shows even a higher decrease for the cutoff frequency quite close to an entire order of magnitude of difference, as shown in Fig. 3.23. Once we have seen the differences in the frequency response in the two-dimensional device and we know there is a significant difference in the working of the detector, we can study even deeper the influence of this parameter with the simulations of the real devices. Starting form device 1, we can start from the dark current because there should be difference due to the maximum velocity carriers can reach even in dark conditions. The



(a)Device 1.



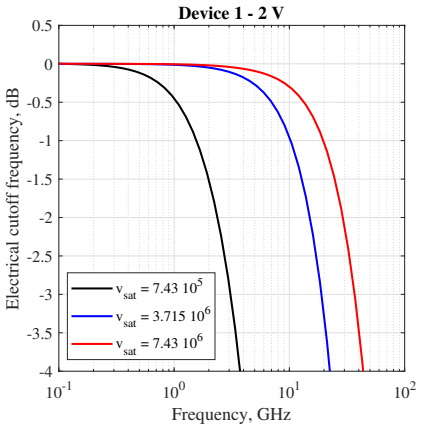
(b)Device 2.

Figure 3.24: Dark current for a reverse bias of 2 V for different velocity saturation values.

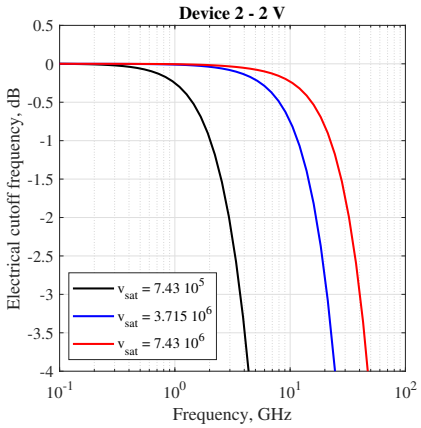
Table 3.5: Cutoff frequency for different values of saturation velocity.

v_{sat}	Device 1	Device 2
$7.430 \times 10^5\,\text{cms}^{-1}$	3.086 GHz	3.695 GHz
$3.715 \times 10^5\,\text{cms}^{-1}$	18.992 GHz	20.898 GHz
$7.430 \times 10^6\,\text{cms}^{-1}$	36.821 GHz	40.028 GHz

electric field is considerable in the 2 V case, because the bias enlarge the depletion region and both type of carriers can hit their maximum velocity in some regions of the Germanium photodetector. From the simulations, the difference in dark current is not overwhelming, but it has still an appreciable difference which tells us that the real value in the Germanium, even if it allows the device to go faster, increase the dark current (see Fig. 3.24). However, this is not the main effect of a different velocity saturation value in a detector. Another time, we can simulate the frequency response and we can see a drastic drop in the performances in relation to the parameter (Fig. 3.25). The cutoff value drops by almost half in case of half saturation velocity, but the most surprising result is given by when we decrease the value by a whole order of magnitude. In this case the value is more than a order of magnitude lower in GHz as reported in Tab. 3.8. We are now able to study the change in the velocity saturation value with the transient analysis. We have done it in both devices (see Fig. 3.26) and what we obtain is quite interesting. Once we half the value of the velocity saturation, the characteristic time defined as the 10 % to 90 % of the curve reduces significantly and become itself approximately half: for device 1 it becomes equal to 18.4 ps and equal to 16.4 ps in device 2. In the case of one order of difference, the

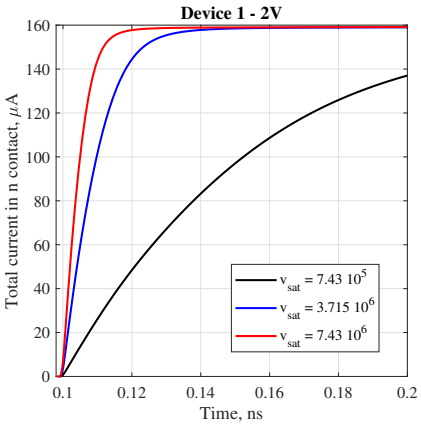


(a)Device 1.

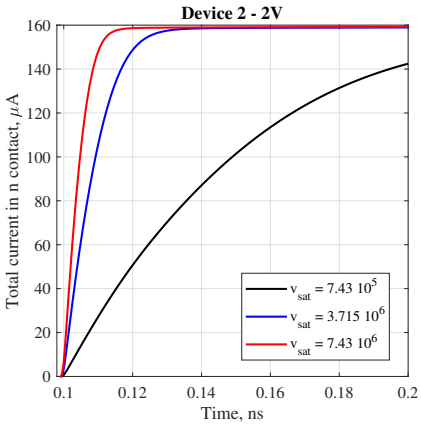


(b)Device 2.

Figure 3.25: Frequency response for a reverse bias of 2 V for different velocity saturation values.



(a)Device 1.



(b)Device 2.

Figure 3.26: Transient simulation for a reverse bias of 2 V for different velocity saturation values.

time follows the same trend going up to a little more than 94 ps for device 2 and breaking the 100 ps barrier for what concern device 1 (the simulation in this case stops before, but we can estimate it with an extension of the curve, assuming it will follow the same behavior). What can be understood through the transient simulation is the difference in microscopical quantities that a different velocity saturation can cause. The response of the device is different because in the regions where the saturation of their velocity value is reached, which is quite large due to the field exploited by some value of reverse bias. Carriers will then move to a velocity closer to the one of the

Table 3.6: WPD geometry.

	W_{Ge}	H_{Ge}	W_{doping}
Simulated Device 1	$4\,\mu$m	$0.8\,\mu$m	$2\,\mu$m
Simulated Device 2	$4\,\mu$m	$0.8\,\mu$m	$2.5\,\mu$m
Device 1	$4\,\mu$m	$0.8\,\mu$m	$3\,\mu$m
Device 2	$4\,\mu$m	$0.8\,\mu$m	$3.5\,\mu$m

Table 3.7: Dark Current for different WNG values.

	dark current
Simulated Device 1	$1.411\,32 \times 10^{-5}\,\mu$A
Simulated Device 2	$1.411\,25 \times 10^{-5}\,\mu$A
Device 1	$1.403\,48 \times 10^{-5}\,\mu$A
Device 2	$1.388\,19 \times 10^{-5}\,\mu$A

regions that are not well invested by the electric field, as we are going to see better in the next section, and the importance of a reverse bias is less predominant than before, even if it is still necessary for the correct working point of the device.

3.5.2 Ge doping profile

An experiment to prove even more the trend we have already seen is to further reduce the W_{doping} value. What we have seen from the results in the two devices in small-signal analysis, transient analysis and from experimental results is that the increase of the doped region inside the Germanium close to the contact, it makes the contact as ohmic [21] as possible. There is an increase of the bandwidth in the Device 2 that is associated solely to the increase of W_{doping} from 3 µm to 3.5 µm. So, we expect that a reduction of this width parameter should lead to a reduction of the cutoff frequency and we perform simulations to verify this. Because of the fact we have the two devices with those width of the Germanium doped regions, we decided to carry on two different simulations on devices unchanged ahead of the W_{doping} value, that is equal to 2 µm for the first simulation and equal to 2.5 µm for the second one. The comparison is made with the two devices and we want to extrapolate a general trend from the values reported in Tab. 3.6. What we compare are all the figures of merit we have already analysed for both Device 1 and Device 2 including the transient analysis. For what concern the dark current there is a slight increase of its value as we can see in Fig. 3.27, but this is not as much as significant as what we are going to obtain in the other simulations. We can report the values in the Tab. 3.7 anyway for completeness: Another aspect that is practically not influenced is the responsivity as we could expect from the beginning. The difference in the

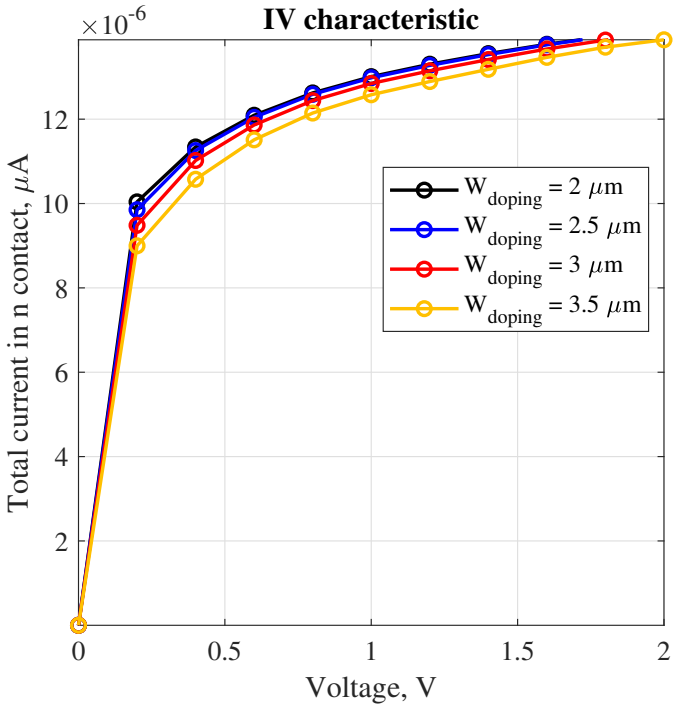


Figure 3.27: Dark current in the simulated devices for different W_{doping}.

power-current curve and its slope are minimal and it has no sense to talk about it. We move on directly to the frequency response in the small-signal analysis. In this case, the difference is huge and the bandwidth drop in case of a W_{doping} of 2 µm meets our expectations. We can also say that the more we decrease the W_{doping} value, the more the response will become worse with respect to the device with a higher W_{doping} (see Fig. 3.28). The variation between 2 µm and 2.5 µm is higher than the one between this last one and the 3 µm W_{doping} device case. The cutoff frequency in the last case is lower than 20 GHz and it makes the device practically useless if compared to the two real devices. There is a reduction in the cutoff frequency of almost half in the comparison with device 1 and more than one in the comparison with device 2. The explanation for this is the fact that the heavily doped region acts as the contact, so as long as we decrease this doping width in the device, we have two main phenomena:

- the electric field is lower in value
- the width of the germanium invested by a high electric field is narrower, leading to a more spread effect of edge effects

We can better see this by reporting the electric field in different W_Ge doping cases as reported in the Fig. 3.29. What are the consequences of these two

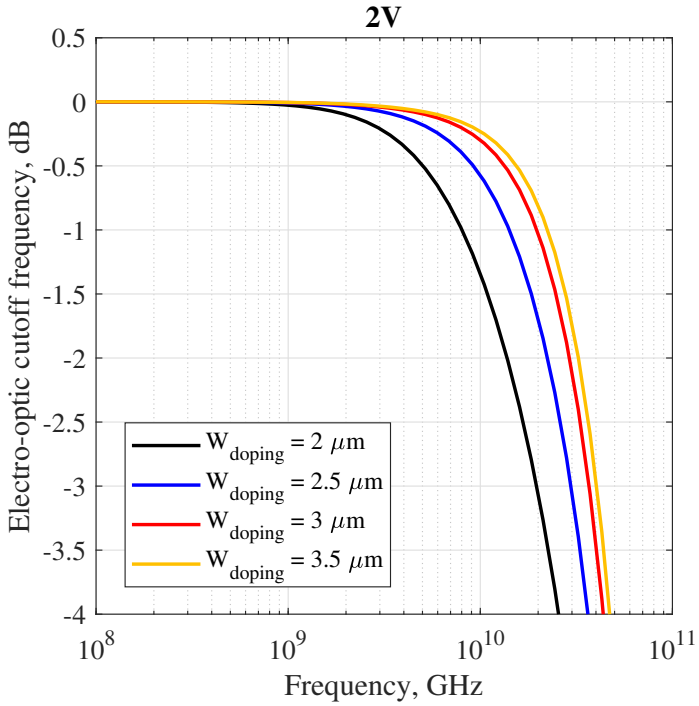
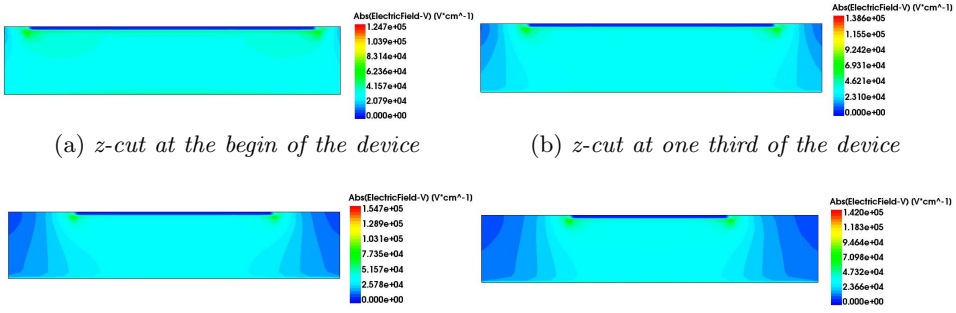


Figure 3.28: Frequency response in the simulated devices for different implantation region width.



(a) *z-cut at the begin of the device*

(b) *z-cut at one third of the device*

(c) *z-cut at two third of the device*

(d) *z-cut at the end of the device*

Figure 3.29: Electric field with the same x cut for different W_{doping} values.

effects? The first one is quite trivial, as long as the electric field is lower in value, the velocity of the carriers will be lower and carriers will spend more time before being collected. The velocity of the carriers can be expressed by the formula: $v_{\mathrm{d}} = \mu E$, lower the electric field, lower their velocity and the speed of the device. The second one is due to the rounding of the electric field, which lines becomes more rounded between the silicon and the high

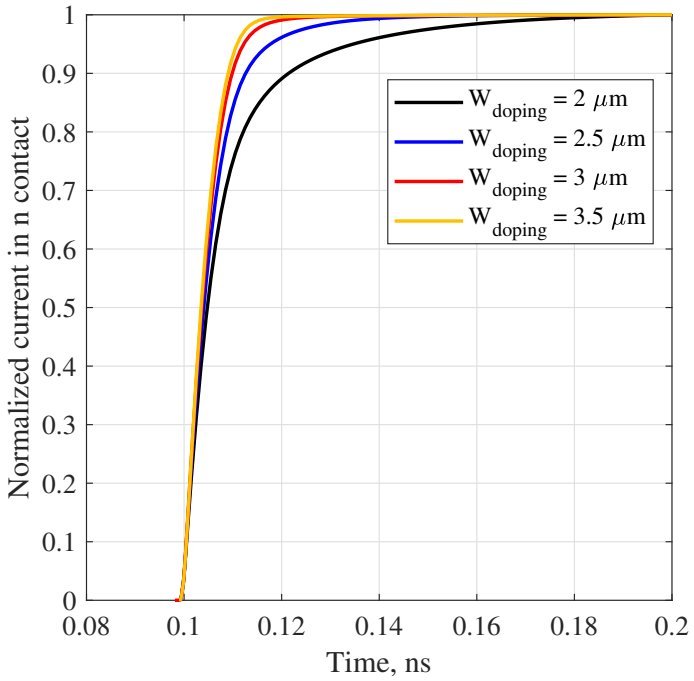


Figure 3.30: Transient in the simulated devices for different W_{doping}.

doped region and there are less vertical lines in the whole germanium directly between the top and the bottom of the germanium. This means that the path of the carriers is longer than before and their transit time will be higher due to the fact that they are not following the shortest possible path inside the device. We can see the higher transit time from the transient analysis and we can also appreciate another time the reduction of the bandwidth. We are plotting the normalized curves reported in Fig. 3.30in order to have a direct comparison on the time scale of the time necessary for the transient to be completed. As expected, the way the transient saturates is coherent with the small-signal analysis and the increase of the 10 % and 90 % difference of the curve makes sense with the results. What it is interesting to notice is that the time constants of the device are changing a lot while we are considering lower and lower width of the implanted region. The capacities and the parasitic effects are different and we can analyse this difference by the use of the transit time, correlating directly the equivalent circuit and the microscopic quantities to the time constants of the time-domain responses.

3.5.3 Mobility

We can refresh just a little on how the problem is posed. In order to design a waveguide photodetector, we need a multiphysical approach coupling elec-

tromagnetic (EM) and carrier transport numerical simulators. The Synopsis tool for the propagation of the optical wave is based on a finite-difference time-domain (FDTD) solver, where we put a certain optical power to a monochromatic wave of a certain wavelength, which is 1310 nm in our simulations and compared results. If the carrier transport is described by a drift-diffusion model [7] [8] with Fermi-Dirac statistics and incomplete ionization, heterointerfaces are included by adding the thermionic current contributions [30]. Many recombination processes are considered for the description of the problem (SRH, radiative, Auger), and an additional carrier generation term, driven by the photon density distribution determined by the EM simulator, couples the EM and transport models [31], [32]. To obtain the photodetector bandwidth it is necessary to have an accurate estimate of the carrier transit time and velocity saturation mechanism due to the high electric field in the intrinsic region of the detector (more specifically the depleted one) has to be taken into account. The standard model for mobility is the one proposed by Canali et al. [33] and this model the velocity saturation effect by putting an electric field dependence in the mobility formula:

$$\mu(\mathcal{E}) = \frac{\mu_0}{\left[1 + \left(\frac{\mu_0 \mathcal{E}}{v_{\text{sat},0}}\right)^{\beta}\right]^{\frac{1}{\beta}}} \tag{3.7}$$

Here μ_0 is the low-field mobility, $v_{\text{sat},0}$ is the saturation velocity, β is a phenomenological parameter, and the electric field magnitude $\mathcal{E}$ acts as carrier driving force. But the mobility depends also on the density of doping impurities and because of this fact Sentaurus adopted for the low-field mobility the Masetti model [34], [35]:

$$\mu_0 = \mu_{\text{min},1} \exp\left(-\frac{P_{\text{c}}}{N}\right) + \frac{\mu_{0,\text{i}} - \mu_{\text{min},2}}{1 + \left(\frac{N}{C_{\text{f}}}\right)^{\alpha}} - \frac{\mu_1}{1 + \left(\frac{C_{\text{s}}}{N}\right)^{\gamma}}, \tag{3.8}$$

where $\mu_{0,\text{i}}$ is the low-field mobility of the intrinsic material, $N = N_{\text{A}} + N_{\text{D}}$ is the sum of acceptor and donor concentrations, and $\mu_{\text{min},1}$, $\mu_{\text{min},2}$, μ_1, P_{c}, C_{f}, C_{s}, α, γ are fitting parameters. Why can we say that the Masetti model is good for our devices? Because, due to the high values of doping, the mobility degrades principally due to impurity scattering and a doping-dependent mobility model is a great choice. The carrier-carrier scattering can be said to have far less influence as well as the interface degradation, due to the lack of semiconductor-oxide interfaces in important regions for the transport as it happens in the channel region of the MOSFET. In general, the mobility model combines with Matthiessen's rule when more than one mobility model is used:

$$\frac{1}{\mu} = \frac{1}{\mu_1} + \frac{1}{\mu_2} + \dots \tag{3.9}$$

We decide to investigate other possible mobility models and see if different hypotheses and assumptions can change much the detector response. What we want to understand in this exploration is the possibility to understand if some effects that may be important are not considered in our model. We start from other doping-dependent mobility models, such as Arora and UbiBo. The Arora model [36] uses the subsequent formula:

$$\mu_{\text{dop}} = \mu_{\text{min}} + \frac{\mu_{\text{d}}}{1 + \left(\left(N_{\text{A},0} + N_{\text{D},0}\right)/N_0\right)^{A^*}} \tag{3.10}$$

with:

$$\mu_{\text{min}} = A_{\text{min}} \cdot \left(\frac{T}{300\ \text{K}}\right)^{\alpha_{\text{m}}}, \mu_{\text{d}} = A_{\text{d}} \cdot \left(\frac{T}{300\ \text{K}}\right)^{\alpha_{\text{d}}} \tag{3.11}$$

and:

$$N_0 = A_N \cdot \left(\frac{T}{300\ \text{K}}\right)^{\alpha_N}, A^* = A_{\text{a}} \cdot \left(\frac{T}{300\ \text{K}}\right)^{\alpha_{\text{a}}} \tag{3.12}$$

The other one, the University of Bologna bulk mobility model has been developed for a range between 25 °C and 973 °C. The model is completely based on the Masetti model, but with two extensions and substantial differences in the approach:

- attractive and repulsive scattering are separately considered, with a function of both donor and acceptor concentrations. This directly accounts for different mobilities for different carriers and ensures the continuity at the device junctions as long as impurity concentrations are continuous.
- a suitable temperature dependence is introduced, in order to predict the dependence of the mobility in a wider temperature range respect to the Masetti model.

The model for lattice mobility is the following one:

$$\mu_{\text{L}}(T) = \mu_{\text{max}} \left(\frac{T}{300\ \text{K}}\right)^{-\gamma + c\left(\frac{T}{300\ \text{K}}\right)} \tag{3.13}$$

with μ_{max} mobility at room temperature and c correction term to the mobility at higher temperatures. The model for bulk mobility is:

$$\mu_{\text{dop}}(T) = \mu_0(T) + \frac{\mu_{\text{L}}(T) - \mu_0(T)}{1 + \left(\frac{N_{\text{D},0}}{C_{\text{r}1}(T)}\right)^{\alpha} + \left(\frac{N_{\text{A},0}}{C_{\text{r}2}(T)}\right)^{\beta}} - \frac{\mu_1\left(N_{\text{D},0}, N_{\text{A},0}, T\right)}{1 + \left(\frac{N_{\text{D},0}}{C_{\text{s}1}(T)} + \frac{N_{\text{A},0}}{C_{\text{s}2}(T)}\right)^{-2}} \tag{3.14}$$

The two mobilities μ_0 and μ_1 are weighted averages of the limiting values for pure acceptor and pure donor-doping densities:

$$\mu_0(T) = \frac{\mu_{0\ \text{d}} N_{\text{D},0} + \mu_{0\text{a}} N_{\text{A},0}}{N_{\text{A},0} + N_{\text{D},0}} \tag{3.15}$$

$$\mu_1(T) = \frac{\mu_{1\ \mathrm{d}} N_{\mathrm{D},0} + \mu_{1\mathrm{a}} N_{\mathrm{A},0}}{N_{\mathrm{A},0} + N_{\mathrm{D},0}} \tag{3.16}$$

The last model we have used is the Philips unified mobility model, which unifies the description of majority and minority carrier bulk mobilities. Other than the temperature dependence of the mobility, it considers also the electron-hole scattering, the screening of ionized impurities by charge carriers, and the clustering of impurities. In this model, we actually use Matthiessan's rule because we consider both the contributions of phonon scattering $\mu_{i,\ \mathrm{L}}$ and bulk scattering mechanisms $\mu_{i,\ \mathrm{DAeh}}$. The total bulk mobility is calculated in this way:

$$\frac{1}{\mu_{i,\ \mathrm{b}}} = \frac{1}{\mu_{i,\ \mathrm{L}}} + \frac{1}{\mu_{i,\mathrm{DAeh}}} \tag{3.17}$$

where the i index has the e value for electrons and the h value for holes. The first mobility contributions is equal to:

$$\mu_{i,\ \mathrm{L}} = \mu_{i,\max} \left(\frac{T}{300\ \mathrm{K}} \right)^{-\theta_i} \tag{3.18}$$

While the second one is more complex:

$$\mu_{i,\ \mathrm{DAeh}} = \mu_{i,N} \left(\frac{N_{i,\mathrm{sc}}}{N_{i,\mathrm{sc,eff}}} \right) \left(\frac{N_{i,\ \mathrm{ref}}}{N_{i,\mathrm{sc}}} \right)^{\alpha_i} + \mu_{i,\mathrm{c}} \left(\frac{n+p}{N_{i,\mathrm{sc,eff}}} \right) \tag{3.19}$$

with the two mobility terms inside the formula which can be expressed as:

$$\mu_{i,N} = \frac{\mu_{i,\max}^2}{\mu_{i,\max} - \mu_{i,\min}} \left(\frac{T}{300\ \mathrm{K}} \right)^{3\alpha_i - 1.5} \tag{3.20}$$

$$\mu_{i,\mathrm{c}} = \frac{\mu_{i,\max}\mu_{i,\min}}{\mu_{i,\max} - \mu_{i,\min}} \left(\frac{300K}{T} \right)^{0.5} \tag{3.21}$$

We need also to express the charge densities in the formula, which is valid for both carriers, for electrons and holes:

$$N_{\mathrm{e,sc}} = N_{\mathrm{D}}^* + N_{\mathrm{A}}^* + p \tag{3.22}$$

$$N_{\mathrm{e,sc,eff}} = N_{\mathrm{D}}^* + G\left(P_{\mathrm{e}}\right) N_{\mathrm{A}}^* + f_{\mathrm{e}} \frac{p}{F\left(P_{\mathrm{e}}\right)} \tag{3.23}$$

$$N_{\mathrm{h,sc}} = N_{\mathrm{A}}^* + N_{\mathrm{D}}^* + n \tag{3.24}$$

$$N_{\mathrm{h,sc,eff}} = N_{\mathrm{A}}^* + G\left(P_{\mathrm{h}}\right) N_{\mathrm{D}}^* + f_{\mathrm{h}} \frac{n}{F\left(P_{\mathrm{h}}\right)} \tag{3.25}$$

The concentrations of donors and acceptors have a $*$ in order to remark the fact that they are both taking into account the effect of clustering. At

ultrahigh concentrations, those concentrations are described by 'clustering' functions Z_{D} and Z_{A}, which are defined as:

$$N_{\mathrm{D}}^{*} = N_{\mathrm{D},0} Z_{\mathrm{D}} = N_{\mathrm{D},0} \left[1 + \frac{N_{\mathrm{D},0}^{2}}{c_{\mathrm{D}} N_{\mathrm{D},0}^{2} + N_{\mathrm{D,ref}}^{2}} \right] \tag{3.26}$$

$$N_{\mathrm{A}}^{*} = N_{\mathrm{A},0} Z_{\mathrm{A}} = N_{\mathrm{A},0} \left[1 + \frac{N_{\mathrm{A},0}^{2}}{c_{\mathrm{A}} N_{\mathrm{A},0}^{2} + N_{\mathrm{A,ref}}^{2}} \right] \tag{3.27}$$

We have to express other two terms in the formula, which are kind of new, the analytic functions $G\left(P_i\right)$ and $F\left(P_i\right)$ that describe minority impurity and electron-hole scattering:

$$F\left(P_i\right) = \frac{0.7643 P_i^{0.6478} + 2.2999 + 6.5502\left(m^{*}{}_i / m^{*}{}_j\right)}{P_i^{0.6478} + 2.3670 - 0.8552\left(m_i^{*}{}_i / m^{*}{}_j\right)} \tag{3.28}$$

$$\begin{aligned} G\left(P_i\right) = 1 - \\ -a_{\mathrm{g}} \left[b_{\mathrm{g}} + P_i \left(\frac{m_0}{m^{*}} \frac{T}{300\ \mathrm{K}} \right)^{\alpha_{\mathrm{g}}} \right]^{-\beta_{\mathrm{g}}} + c_{\mathrm{g}} \left[P_i \left(\frac{m^{*}{}_i 300K}{m_0} \frac{\alpha}{T} \right)^{\alpha_{\mathrm{g}}} \right]^{-\gamma_{\mathrm{g}}} \end{aligned} \tag{3.29}$$

where $m^{*}{}_i$ and $m^{*}{}_j$ are fit parameters for both carriers. The screening parameter used for the analytic functions is calculated by a weighted harmonic mean of the Brooks-Herring approach and Conwell-Weisskopf approach:

$$\begin{aligned} P_i = \left[\frac{f_{\mathrm{CW}}}{3.97 \times 10^{13}\ \mathrm{cm}^{-2} N_{i,\mathrm{sc}}^{-2/3}} + \right. \\ \left. + f_{\mathrm{BII}} \frac{(n+p)}{1.36 \times 10^{20}\ \mathrm{cm}^{-3}} \frac{m_0^{*}}{i} \left(\frac{T}{300\ \mathrm{K}} \right)^{2} \right] \end{aligned} \tag{3.30}$$

For values of $P_i < P_{i,\min}, G\left(P_{i,\min}\right)$ substitutes $G\left(P_i\right)$, where $P_{i,\min}$ is the value at which $G\left(P_i\right)$ reaches its minimum. Once we have seen these possible alternatives to the Masetti standard method used from Sentaurus, we can compare the results with different models. Finally, we can compare some quantities between the different models. The first one can be for example the dark current in Fig. 3.31. As we already said, mobility is a fundamental value to see the velocity of the carriers, especially in the case it saturates to a maximum value where the electric field is particularly strong. A direct consequence is a difference in the electro-optical bandwidth there is between the analyzed models as we can see in Fig. 3.32. The cutoff frequencies have a major value downgrade and we can report the values to compare them: At the end we can compare also the transient simulation results for different models (Fig. 3.33).

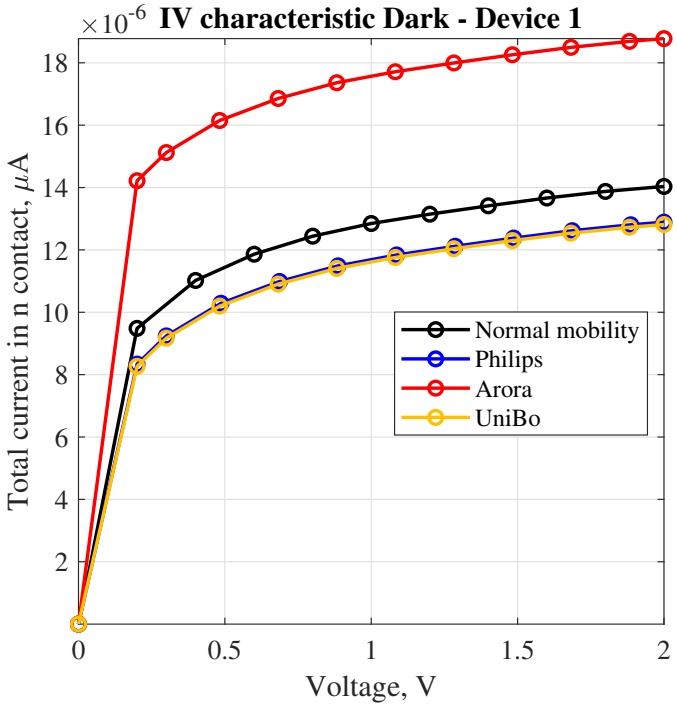


Figure 3.31: Dark current for different mobility models.

Table 3.8: Cutoff frequency for different values of saturation velocity.

v_{sat}	Device 1
Masetti	36.81 GHz
Arora	28.72 GHz
UniBo	21.89 GHz
Philips	21.71 GHz

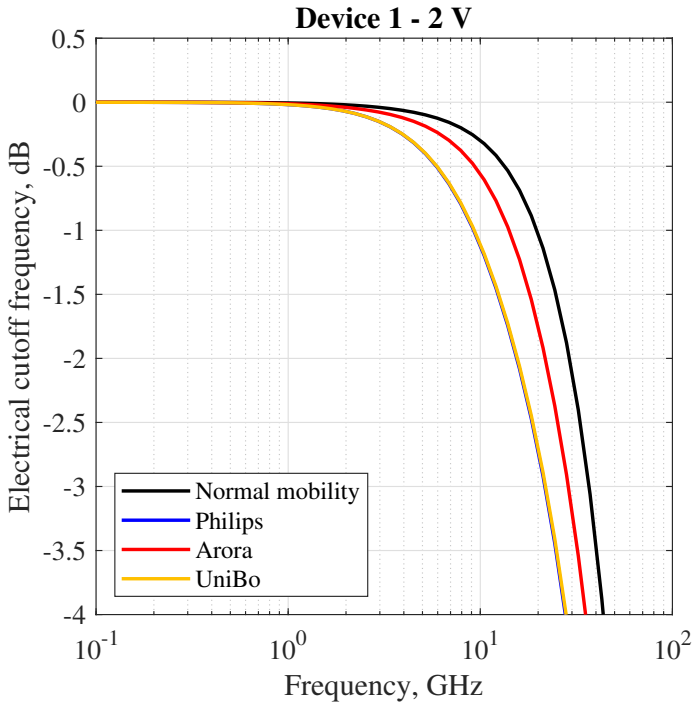


Figure 3.32: Frequency response for different mobility models.

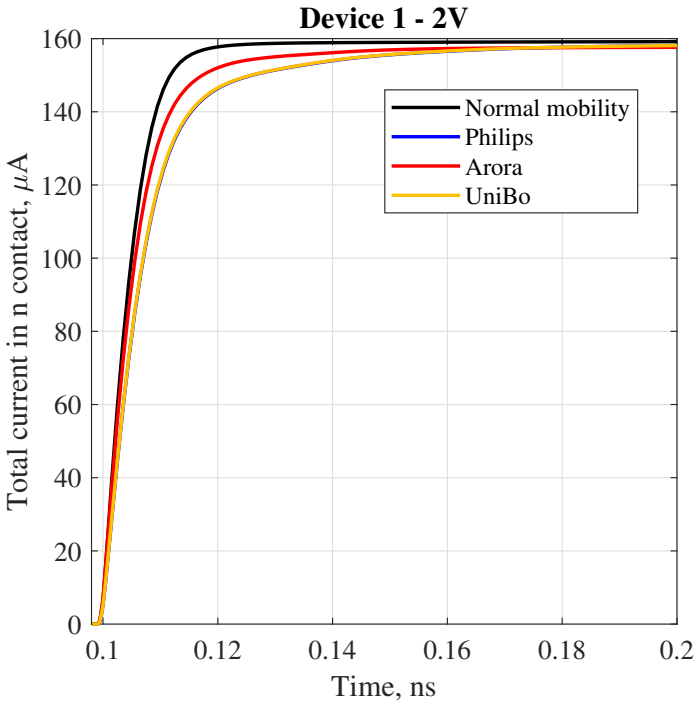


Figure 3.33: Transient analysis for different mobility models.

Chapter 4

Conclusions

Silicon photonics is an area for the development of very fast optical connections that is rapidly growing. The simulation of the devices constitutes a fundamental part of the maturation of this technology, making the creation of new devices quite cheaper and exploiting possible problems even before they are actually found in a finished device. Most of the simulations present in literature for this kind of purpose are the ones we have used for the preliminary analysis, while the transient simulations have not been studied so far and they can give a different point of view for the optimization of this type of device. One of the main cores of the thesis was the study of the two-dimensional device and the consequent optimization of the tool, in order to make its computational cost reasonable for this type of time-domain analysis. The study of the rise time of the optical input power curve and the correct implementation of the break criteria are a great part of the results and allowed us to move on and perform the simulations on the studied devices. The transient uses the FDTD solution as the source of the simulation. The pulsed optical generation rate term is put in the transport equations, Sentaurus then solves the device through the use of the Poisson equation coupled in a self-consistent way with the continuity equations of the carriers and with the constitutive relations of drift-diffusion for the current densities of electrons and holes, the transport problem is solved. The transient considers temporal variations of the applied signal and it increases the time step in a loop until the time of the simulation is up or the break criteria force the transient simulation to quit. The results obtained through the transient simulations are important and they are promising. Even if some of the obtained results are already a turning point for even more complete optimization of the devices, there is a lot of research to do to exploit the full potential of a time-domain analysis. The main result we have obtained is the correspondence for low optical input power between the transient and the small-signal simulation, while when we are increasing the input power, the achieved results are very different. In fact, the small-signal linearization

is not able to describe this condition, being it is far away from the working point and a simulation that includes large-signal effects. Small-signal simulations are by definition limited to the cyclostationary steady state and they are not able to describe effects linked to slower carriers, that can be studied starting from current tails. In conclusion, despite their much higher computational cost with respect to other simulations, transient ones have shown a significant, and still largely unexplored, potential contribution to the field of WPD modeling and optimization [22]. From the work of this thesis, it was possible to give a contribution to two articles for two different conferences, NUSOD 2022 and SIE 2022.

Bibliography

[1] L. Pavesi and D. J. Lockwood, *Silicon Photonics*. Berlin: Springer-Verlag, 2004.

[2] D. Thomson, A. Zilkie, J. E. Bowers, T. Komljenovic, G. T. Reed, L. Vivien, D. Marris-Morini, E. Cassan, L. Virot, J. Fédéli, J. Hartmann, J. H. Schmid, D. X. Xu, F. Boeuf, P. O'Brien, G. Z. Mashanovich, and M. Nedeljkovic, "Roadmap on silicon photonics," *J. Opt.*, vol. 18, no. 7, p. 073003, 2016.

[3] V. Sorianello, L. Colace, N. Armani, F. Rossi, C. Ferrari, L. Lazzarini, and G. Assanto, "Low-temperature germanium thin films on silicon," *Opt. Mater. Express*, vol. 1, pp. 856–865, Sept. 2011.

[4] M. Vallone, A. Palmieri, M. Calciati, F. Bertazzi, M. Goano, G. Ghione, and F. Forghieri, "3D physics-based modelling of Ge-on-Si waveguide *p-i-n* photodetectors," in *17th International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD 2017)*, (Copenhagen, Denmark), pp. 207–208, July 2017.

[5] K. Yee, "Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media," *IEEE Trans. Antennas Propagation*, vol. 14, pp. 302–307, May 1966.

[6] Synopsys, Inc., Mountain View, CA, *Sentaurus Device Electromagnetic Wave Solver User Guide. Version N-2017.09*, Sept. 2017.

[7] S. Selberherr, *Analysis and Simulation of Semiconductor Devices*. Wien: Springer-Verlag, 1984.

[8] F. Bertazzi, M. Goano, G. Ghione, A. Tibaldi, P. Debernardi, and E. Bellotti, "Electron transport," in *Handbook of Optoelectronic Device Modeling and Simulation* (J. Piprek, ed.), ch. 2, pp. 35–80, Boca Raton, FL: CRC Press, 2017.

[9] R. E. Bank, W. M. Coughran, W. Fichtner, E. H. Grosse, D. J. Rose, and R. Kent Smith, "Transient simulation of silicon devices and

circuits," *IEEE Trans. Computer-Aided Design*, vol. CAD-4, no. 4, pp. 436–451, 1985.

[10] M. E. Hosea and L. F. Shampine, "Analysis and implementation of TR-BDF2," *Applied Numerical Mathematics*, vol. 20, no. 1-2, pp. 21–37, 1996.

[11] G. Ghione and A. Benvenuti, "Discretization schemes for high-frequency semiconductor device models," *IEEE Trans. Antennas Propagation*, vol. 45, pp. 443–456, Mar. 1997.

[12] D. Benedikovic, L. Virot, G. Aubin, J.-M. Hartmann, F. Amar, X. Le Roux, C. Alonso-Ramos, É. Cassan, D. Marris-Morini, J.-M. Fédéli, *et al.*, "Silicon–germanium receivers for short-wave-infrared optoelectronics and communications," *Nanophoton.*, vol. 10, pp. 1059–1079, Dec. 2020.

[13] J. Liu, S. Cristoloveanu, and J. Wan, "A review on the recent progress of silicon-on-insulator-based photodetectors," *Phys. Status Solidi A*, vol. 218, p. 2000751, July 2021.

[14] G. Ghione, *Semiconductor Devices for High-Speed Optoelectronics*. Cambridge, U.K.: Cambridge University Press, 2009.

[15] L. Vivien, J. Osmond, J.-M. Fédéli, D. Marris-Morini, P. Crozat, J.-F. Damlencourt, E. Cassan, Y. Lecunff, and S. Laval, "42 GHz p.i.n germanium photodetector integrated in a silicon-on-insulator waveguide," *Opt. Express*, vol. 17, pp. 6252–6257, Apr. 2009.

[16] K. Ito, T. Hiraki, T. Tsuchizawa, and Y. Ishikawa, "Waveguide-integrated vertical pin photodiodes of Ge fabricated on p^+ and n^+ Si-on-insulator layers," *Japan. J. Appl. Phys.*, vol. 56, no. 4S, p. 04CH05, 2017.

[17] M. Auf der Maur, B. Galler, I. Pietzonka, M. Strassburg, H. Lugauer, and A. Di Carlo, "Trap-assisted tunneling in InGaN/GaN single-quantum-well light-emitting diodes," *Appl. Phys. Lett.*, vol. 105, no. 13, p. 133504, 2014.

[18] M. Vallone, M. Mandurrino, M. Goano, F. Bertazzi, G. Ghione, W. Schirmacher, S. Hanna, and H. Figgemeier, "Numerical modeling of SRH and tunneling mechanisms in high-operating-temperature MWIR HgCdTe photodetectors," *J. Electron. Mater.*, vol. 44, no. 9, pp. 3056–3063, 2015.

[19] M. Mandurrino, G. Verzellesi, M. Goano, M. Vallone, F. Bertazzi, G. Ghione, M. Meneghini, G. Meneghesso, and E. Zanoni, "Physics-

based modeling and experimental implications of trap-assisted tunneling in InGaN/GaN light-emitting diodes," *Phys. Status Solidi A*, vol. 212, no. 5, pp. 947–953, 2015.

[20] M. Cardona and F. H. Pollak, "Energy-band structure of germanium and silicon: The $k \cdot p$ method," *Phys. Rev.*, vol. 142, pp. 530–543, Feb. 1966.

[21] N. DasGupta and A. DasGupta, *Semiconductor devices: modelling and technology*. PHI Learning Pvt. Ltd., 2004.

[22] A. M. GC, P. Franco, A. Tibaldi, F. Bertazzi, S. Namnabat, D. Adams, P. Gothoskar, G. Masini, F. Forghieri, G. Ghione, *et al.*, "3d multiphysics transient modeling of vertical ge-on-si pin waveguide photodetectors," in *2022 International Conference on Numerical Simulation of Optoelectronic Devices (NUSOD)*, pp. 5–6, IEEE, 2022.

[23] R. Quay, C. Moglestue, V. Palankovski, and S. Selberherr, "A temperature dependent model for the saturation velocity in semiconductor materials," *Mater. Sci. Semicond. Processing*, vol. 3, no. 1–2, pp. 149–155, 2000.

[24] M. G. C. Alasio, M. Goano, A. Tibaldi, F. Bertazzi, S. Namnabat, D. Adams, P. Gothoskar, F. Forghieri, G. Ghione, and M. Vallone, "Bias effects on the electro-optic response of Ge-on-Si waveguide photodetectors," in *IEEE Photonics Conference*, (online), Oct. 2021.

[25] M. Rahman, *Applications of Fourier Transforms to Generalized Functions*. , Boston, MA: WIT Press, 2011.

[26] M. J. Byrd, E. Timurdogan, Z. Su, C. V. Poulton, N. M. Fahrenkopf, G. Leake, D. D. Coolbaugh, and M. R. Watts, "Mode-evolution-based coupler for high saturation power Ge-on-Si photodetectors," *Opt. Lett.*, vol. 42, pp. 851–854, Feb. 2017.

[27] Synopsys, Inc., Mountain View, CA, *Sentaurus Device User Guide. Version N-2017.09*, Sept. 2017.

[28] Keysight Technologies, Santa Rosa, CA, *Lightwave Component Analyzer application notes*, Dec. 2017.

[29] A. M. GC, P. Franco, A. Tibaldi, F. Bertazzi, S. Namnabat, D. Adams, P. Gothoskar, G. Masini, F. Forghieri, G. Ghione, *et al.*, "Bandwidth dependence of ge-on-si vertical pin photodetectors," in *Riunione Annuale dell'Associazione Società Italiana di Elettronica SIE 2022*, p. 1, SIE, 2022.

[30] D. Schroeder, *Modelling of Interface Carrier Transport for Device Simulation*. Computational Microelectronics, Wien: Springer-Verlag, 1994.

[31] M. Vallone, M. Goano, F. Bertazzi, G. Ghione, W. Schirmacher, S. Hanna, and H. Figgemeier, "Comparing FDTD and ray tracing models in the numerical simulation of HgCdTe LWIR photodetectors," *J. Electron. Mater.*, vol. 45, no. 9, pp. 4524–4531, 2016.

[32] A. Palmieri, M. Vallone, M. Calciati, A. Tibaldi, F. Bertazzi, G. Ghione, and M. Goano, "Heterostructure modeling considerations for Ge-on-Si waveguide photodetectors," *Opt. Quantum Electron.*, vol. 50, p. 71, Feb. 2018.

[33] C. Canali, G. Majni, R. Minder, and G. Ottaviani, "Electron and hole drift velocity measurements in silicon and their empirical relation to electric field and temperature," *IEEE Trans. Electron Devices*, vol. 22, pp. 1045–1047, Aug. 1975.

[34] G. Masetti, M. Severi, and S. Solmi, "Modeling of carrier mobility against carrier concentration in arsenic-, phosphorus-, and boron-doped silicon," *IEEE Trans. Electron Devices*, vol. 30, pp. 764–769, July 1983.

[35] G. Hellings, G. Eneman, R. Krom, B. De Jaeger, J. Mitard, A. De Keersgieter, T. Hoffmann, M. Meuris, and K. De Meyer, "Electrical TCAD simulations of a germanium pMOSFET technology," *IEEE Trans. Electron Devices*, vol. 57, pp. 2539–2546, Oct. 2010.

[36] N. D. Arora, J. R. Hauser, and D. J. Roulston, "Electron and hole mobilities in silicon as a function of concentration and temperature," *IEEE Transactions on electron devices*, vol. 29, no. 2, pp. 292–295, 1982.