CMOS Photodetectors

Albert H. Titus¹, Maurice C-K. Cheung² and Vamsy P. Chodavarapu²

¹Department of Electrical Engineering, University at Buffalo, The State University of New York
²Electrical and Computer Engineering, McGill University, Montreal

¹USA
²Canada

1. Introduction

The inclusion of cameras in everything from cell phones to pens to children’s toys is possible because of the low cost and low power consumption of the imaging arrays that form the core of the cameras. However, these arrays are low cost and low power because they are CMOS-based; this allows for the devices to be made with the same processes and facilities that are used to make memory and computer chips. Yet, the continued surge in CMOS imager popularity goes beyond the lower cost to other factors such as ability to integrate the sensors with electronics, and the ability to achieve fast, customizable frame rates [1].

In this chapter, we will cover the design and layout of the various types of CMOS photodetectors and fundamentals of photo-conversion processes in these devices, including a brief review of CMOS photodetector history. We will then describe the emerging CMOS based technologies in photodetector design optimization to tune device responsivity, integration of micro-optics to achieve enhanced detection in low-light conditions, integration of photonic grating structures on photodetectors for spectral response selectivity, and bio-inspired CMOS imaging devices. We will conclude the chapter with some examples of applications of these technologies.

2. CMOS photodetector history

Since the mid-1960s, combinations of p-n (or n-p-n) junctions have been used to convert light into electronic signals [2, 3]. Work not only focused on the conversion of photons to electrons, but also on the ability to read the signals out from arrays of pixels. For example, Shuster and Strull reported in 1965 that they had developed a 2500 pixel array of phototransistors with 100 leads for readout [4, 5]. For these earliest devices, high gain had to be used at the pixels because no integration of light (or charge) was used. A year later, Weckler demonstrated how to operate a p-n junction photodiode in “photo flux integrating mode” which enabled pixels to be much simpler and ultimately smaller [6].
Photon integration required that the photodiode be turned on for a fixed amount of time to raise (or lower) the voltage level on a charge storage device; thus, the amount of charge remaining (or charge that was removed) is proportional to the light intensity over that integration time.

In the 1970s, CMOS detectors and imaging arrays began to lose popularity to Charge-Coupled Device (CCD)-based imagers because CCDs could achieve a higher fill-factor; the fill-factor was lower for CMOS imagers because of the need for transistors at the pixels for read out and gain. The ability to produce CCD imagers with the necessary number of pixels for applications (such as TV) gave CCDs a large advantage over CMOS imagers [7]. Noise in CCDs was also considerably less than in CMOS devices; it was generally regarded that fixed pattern noise in CMOS imaging devices was worse than in CCDs, which remained true into the late 1980s and early 1990s. However, improvements in CMOS fabrication technology and increasing pressure to reduce power consumption for battery operated devices began the re-emergence of CMOS as a viable imaging device.

It is generally regarded that the first all-CMOS sensor array to produce acceptable images is the active pixel sensor (APS) imager [8-10]. The APS design used the linear integration method [6] for measuring light because of the large output signal generated, as opposed to the logarithmic method [11-16]. The active column sensor (ACS) [17] is similar to the APS but has lower fixed pattern noise (FPN). As of 2011, we have CMOS image sensors that have 14.6 Megapixels, and higher.

While CMOS and CCDs continue to compete for a share of the image array sensor market, the ability to design custom integrated circuits (ICs) with photodetectors to perform specific functions is an enormous advantage over CCD arrays. These ICs are often used in applications that have specific requirements such as extremely low power consumption [18] or variable read-out (frame) rates [19] or very fast read-out rates [20]. Normal video rates of 30 or 60 frames per second are fine for standard definition videos, but for some applications, frame rates of more than a thousand frames per second are needed to capture extremely fast occurring events (for example see [20]).

CMOS photodetectors are the technology of choice in smart focal plane arrays. In the mid-1980s, Carver Mead and Misha Mahowald introduced the Silicon Retina that used a "vertical bipolar transistor" as the light detecting element [21]. This spawned a significant amount of research into bio-inspired vision chips that used CMOS photodetectors combined with CMOS signal processing circuitry [22-32]. Generally, these chips are arrays of smart pixels with significantly more transistors per pixel than the three or four found in typical APS-based arrays. However, the relative low cost of fabrication for prototyping CMOS ICs enables chips with one, two, dozens or thousands of detectors to be designed, fabricated and tested. For chips with few photodetectors, the remaining silicon die area can be used for signal processing, read-out, or digital interface logic, so there is no wasted space. Applications of these custom detector and imaging chips range from sub-retinal implant imagers [33] to glare detection [34] to fluorescence imaging [35, 36] and x-ray imaging [37], to name a few. A study of three common photodiode structures available in non-imager/standard CMOS processes provides valuable benchmark data for designers looking to use CMOS photodetectors [38]. While not an exhaustive study of all possible CMOS photodetectors, this chapter provides a useful starting point for selecting the best structure for the application.
3. Operation of CMOS photodetectors

The core of the sensing element of a CMOS detector is the photosensitive element of the circuit. Photogates, phototransistors, and photodiodes all can be used as the sensing element. In this section, the use of a photodiode is discussed. As its name implies, the photodiode is simply a junction between a p-type and an n-type semiconductor, commonly known as a p-n junction. Although a simple p-n junction can be used for light detection, the more sophisticated p-i-n junction with an intrinsic region between the p-type and n-type region is often used to improve the device efficiency. In this section, the basic working principle of a photodiode will be discussed, followed by a discussion on the p-i-n photodiode, and a method of signal amplification resulting in the avalanche photodiode.

3.1 Photogeneration and recombination in semiconductors

When a semiconductor is illuminated, a photon that has higher energy, \( h\nu \), than the bandgap energy, \( E_g \), may cause an excitation of an electron from the lower energy valence band to the higher energy conduction band. This results in a pair of the mobile charge carriers - electrons and holes. This process known as photogeneration can occur if the total energy and total momentum among the photogenerated electron-hole pair and the incoming photon is conserved.

The probability of photogeneration by a single photon is a property of the material. Macroscopically, this is described by the absorption coefficient, \( \alpha \). As a light beam propagates through a piece of homogeneous semiconductor, its power decreases exponentially as the semiconductor absorbs some of the power for photogeneration. The power that remains in the light beam after propagating through a depth of \( z \) is given by:

\[
P = P_0 \exp(-\alpha z)
\]

where, \( P_0 \) is the intensity at zero depth. Note that the rate of absorption, \( -dP/dz \), decreases exponentially with depth. Therefore, more photogeneration is expected to occur near the surface.

Because the photogenerated carriers exist in an excited state, the excess electrons and holes will recombine after a short period of time (~ picoseconds on average) to release the excess energy. This process is known as recombination, and it returns the carriers distributions to thermal equilibrium condition. These excess carriers are lost if they are not captured to create an electrical signal for light detection. Therefore, a semiconductor device structure is needed to facilitate the capturing of the photogenerated carriers. The simplest and most commonly used structure for this purpose is a diode structure known as photodiode (PD).

3.2 Quantum efficiency and responsivity

Before the discussion on photodiode, two important parameters that are used to characterize the effectiveness of detection by a photodetector should be discussed; these are quantum efficiency and responsivity. Quantum efficiency is defined as the probability that an incident photon will generate an electron-hole pair that will contribute to the detection signal, so it can be expressed as,

\[
\eta = (1 - \mathcal{R}) \zeta [1 - \exp(-\alpha d)]
\]
where, $\mathcal{R}$ is the surface reflectance, $\zeta$ is the probability that the generated electron-hole pair will become a contribution to the detection signal, and $d$ is the depth of the photo-absorption region. Therefore, quantum efficiency is affected by material properties and device geometry. The captured carriers are used to generate a signal either as a voltage or a current. The measure of signal strength to incident power of the device is called responsivity. If the output is a current, then it is related to the quantum efficiency by the following expression,

$$
R = \frac{q}{h \nu} \eta = \frac{q}{h c} \frac{\lambda}{1.24 \text{[nm⋅W/A]}} \eta
$$

where, $q$ is the electron charge and $\lambda$ is the wavelength in nanometers.

4. Photodiode

Many photodetectors utilize the formation of p-n junctions, and the simplest of these is a photodiode, because a photodiode is simply a p-n junction that is designed for capturing the photogenerated carriers. In CMOS sensing, a photodiode is usually made by forming an n-type region on a p-type semiconductor substrate, or vice-versa. This can be done by epitaxial growth, diffusion or ion implantation.

4.1 A quick review of P-N junction

Figure 1 shows the carrier distributions, charge distribution, built-in electric field and band-diagram of a typical p-n junction. The inhomogeneous charge and carrier distributions are the result of a state of equilibrium between diffusion, drift and recombination. The following is a review of the key features of a p-n junction,

1. An absent of the carriers in a region known as the depletion region or space charge region. It has layer width, $W$, and ionic space charge of the donors and acceptors are exposed in this region.

![Fig. 1. (a) Diffusion of carriers at the p-n junction, (b) the resulting electron and hole density distribution, where $N_a$ and $N_d$ are the acceptors and donors densities, (c) the resulting charge density distribution, (d) the resulting electric field, and (e) the band diagram of a p-n junction showing the alignment of the Fermi level, $E_f$, and shifting of the bands. $E_v$ is the top of the valence band and $E_c$ and the bottom of the conduction band.](www.intechopen.com)
2. Due the charge distribution, a built-in electric field, \( E \), has emerged.
3. Alignment of \( E_f \) between the p-type region and the n-type, as the conduction and valance bands shift in the depletion region.
4. There is a potential difference of \( V_0 \) between the p-side and the n-side.

### 4.2 Operating principle

The operation of a photodiode relies upon the separation of the photogenerated carriers by the built-in field inside the depletion region of the p-n junction to create the electrical signal of detection. Under the influence of the built-in electric field, the photogenerated electrons will drift towards the n-side, and photogenerated holes will drift towards the p-side. The photogenerated carriers that reach the quasi-neutral region outside of the depletion layer will generate an electric current flowing from the n-side to the p-side; this current is called a photocurrent. The generation of the photocurrent results in the shift of the I-V characteristic of the photodiode as shown in Figure 2. Therefore, the I-V characteristic of a photodiode is expressed as,

\[
I_L = I_s \left[ \exp \left( \frac{qV}{kT} \right) - 1 \right] - I_{ph} \tag{4}
\]

where, the first term is the Schottky Equation that described the ideal I-V characteristics with \( I_s \) being the saturation current, \( k \) the Boltzmann constant and \( T \) the operating temperature, and the second term, \( I_{ph} \), is the photocurrent.

---

![Fig. 2. Photogeneration in a p-n junction: (a) the built-in electric field driving the photogenerated carriers from the depletion region away from the junction, and (b) shifting of the I-V characteristic due to the photogenerated current, \( I_{ph} \).](#)
voltage is created across the photodiode, called the open circuit voltage, $V_{OC}$. In reference to Figure 2(a), the photodiode is operating at the point where the I-V characteristic curve intersects the x-axis.

![Diagram of PD in open circuit mode](image)

![Diagram of PD in short circuit mode](image)

![Diagram of PD in reverse-bias mode](image)

Fig. 3. Basic operating mode of a PD: (a) open circuit mode, (b) short circuit mode, and (c) reverse-bias mode.

In contrast, in the short circuit (SC) mode, the terminal of the photodiode is short-circuited. This allows the photogenerated current to flow in a loop as illustrated in Figure 3(b). In Figure 2(b), this is represented by the point at which the I-V characteristic curve intersects the y-axis. The current that flows in the loop in SC mode is also known as the short circuit, $I_{sc}$, and it has the same magnitude as $I_{ph}$.

In reverse bias mode, a reverse bias is applied across the photodiode as shown in Figure 3(c). Therefore it is operated in the lower left quadrant of Figure 2(b). Note that by applying a bias voltage, the potential difference across the p-n junction changes to $V_0 - V$, and the balance between drift and diffusion in the p-n junction also changes. This will affect the depletion width ($W$) and $E$ as well. The dependence of $W$ on the bias voltage can be described by,

$$W = K(V_0 - V)^{m_j}$$  \hspace{1cm} (5)

where, $K$ is a constant, and $m_j$ depends on the junction geometry ($m_j = 1/2$ for a step junction and $m_j = 1/3$ for a linear junction). Therefore, operating in reverse bias has the effect of increasing $W$. The increase in $W$ is not as great as the change in the potential difference, because $m_j < 1$, so $E$ should also increase. From the point of view of charge distribution and Gauss's Law, a wider depletion region exposes more of the ionic space charge, which in turn increases the electric field.

The widened depletion region under reverse bias creates a greater photogeneration region, while the stronger $E$ increases the drift velocity of the photogenerated carriers. In principle, the drift velocity increases in proportion with $E$, so even with an increase in $W$ given by Equation (5), the transit time (the average time that a drifting carrier to reach the end of the depletion region) is reduced. Therefore signal loss due to recombination in the depletion region is reduced. Because of these beneficial effects, reverse bias operation is often preferred.
4.4 Dark current

4.4.1 Saturation current – diffusion of minority carriers
As shown in Figure 2(b) there exist a small current under reverse bias in the I-V characteristic even in dark condition. This dark current is caused by saturation current, $I_s$. On the boundaries of the depletion region, minority carriers (electrons on the p-side and holes on the n-type side) can diffuse into the depletion region. Because of the built-in electric field, these diffused minority carrier may drift across to the depletion region; this is a source of the saturation current. Therefore, in photodiodes there exist a dark current; however, the diffusion process is not the only contribution the dark current.

4.4.2 Generation-recombination current
Apart from the diffusion contribution to the dark current, carriers that are generated by thermal excitation inter-band trap (defect) states in the depletion region can also have a contribution to the dark current. This trap-assisted process is essentially the reverse of Shottky-Read-Hall (SRH) recombination. Just like the photogenerated carriers, carriers created by trap-assisted generation in the depletion region can also be swept away from the depletion region by drift before they can recombine and form part of the dark current. This dark current contribution is known as the Generation-Recombination (G-R) current, and it can be more significant than the diffusion contribution.

4.4.3 Tunneling currents
Under sufficient reverse bias, the $E_C$ on the n-side can fall below $E_V$ on the p-side. In this condition, there is a finite possibility that an electron in the valence band of the p-side can tunnel through the bandgap into the n-side conduction band. This process is called direct tunneling or band-to-band tunneling. If sufficient numbers of direct tunneling events occur, its contribution to the dark current will be measurable.

Tunneling can also occur through an inter-band trap (defect) state. Due to thermal excitation, a carrier can be trapped in one of these states. If this state exists in the depletion region, and sufficient reverse biased is applied, a trapped electron from the valence band can have energy higher than $E_C$, and tunneling into the conduction band can occur. This is called trap-assisted tunnel.

4.4.4 Surface leakage current
Due to the interruption of the crystal lattice structure, there can be a high density of surface charge and interface states at the physical surface of a device. The surface charge and interface states can affect the position of the depletion region, as well as behaving as generation-recombination centers. Therefore, the surface of a device can introduce another contribution to the dark current called surface leakage current. Passivation of the surface can be used to control the surface leaking current. This is usually achieved by adding a layer of insulator such as oxide to the surface.

4.4.5 Frankel-poole current
When a sufficiently large electric field is applied to an insulator, it will start to conduct. This is called the Frankel-Poole Effect. This effect is caused by escape of electrons from their localized state into the conduction band. Therefore, the Frankel-Poole Effect can occur in a semiconductor as well. When a sufficiently large reverse bias is applied across a p-n
junction, electrons generated by the Frankel-Poole Effect can also have a contribution to the dark current.

4.4.5 Impact ionization current
Under reverse bias, the motion of carriers in the depletion can be described as a drift where the carriers are repeatedly accelerated by the electric field and collide with the atoms in the crystal lattice. Under strong reverse bias, the acceleration between collisions can be large enough for a carrier to obtain the energy required to dislodge a valance electron to create a new electron-hole pair. This process is known as impact ionization and it can generate new carriers that contribute to the reverse bias current. When the applied reverse bias is beyond the breakdown voltage, \( V_{bd} \), impact ionization becomes a dominant factor of the photodiode behavior, and the photodiode is said to be operating in avalanche mode.

4.4.6 Summary of dark current
Table 1 summarizes the different dark currents and their dependence. When these dark currents are taken into consideration, the photodiode no longer follows the ideal diode characteristic. A detailed discussion on dark current can be found in [39].

<table>
<thead>
<tr>
<th>Process</th>
<th>Dependence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diffusion</td>
<td>( \propto \exp \left( -\frac{E_g}{kT} \right) )</td>
</tr>
<tr>
<td>G-R</td>
<td>( \propto \sqrt{V} \exp \left( -\frac{E_g}{2kT} \right) )</td>
</tr>
<tr>
<td>Band-to-band tunneling</td>
<td>( \propto V^2 \exp \left( -\frac{a}{V} \right) )</td>
</tr>
<tr>
<td>Trap-assisted tunneling</td>
<td>( \propto \exp \left( -\frac{a^2}{V} \right) )</td>
</tr>
<tr>
<td>Surface Leakage</td>
<td>( \propto \exp \left( -\frac{E_g}{2kT} \right) )</td>
</tr>
<tr>
<td>Frankel-Poole</td>
<td>( \propto V \exp \left( -\frac{b}{T} \right) )</td>
</tr>
<tr>
<td>Impact Ionization</td>
<td>( \propto \exp \left( -\frac{c}{T} \right) )</td>
</tr>
</tbody>
</table>

Table 1. Summary of dark current dependence, where \( a, a', b \) and \( c \) are constants.

4.5 Noise
There are two basic noise generating mechanisms in a photodiode: the statistical fluctuation in the number of carriers and photons, and the random motion of the carriers. The statistical fluctuation in particle numbers is the cause of shot noise. The root mean square of the current fluctuation due to shot noise is

\[
i_{sh, \text{ rms}} = \sqrt{2 \ q \ I_{avg} \ \Delta f}
\]  

(6)
where, $I_{avg}$ is the average signal current and $\Delta f$ is the bandwidth of the signal. The signal-to-noise ratio (SNR) of shot noise is given by,

$$\text{SNR} = \frac{I_{avg}}{2q\Delta f}$$  \hspace{1cm} (7)

The random movement of carriers produces thermal noise also known as Johnson-Nyquist noise. The root mean square of the current fluctuation due to Johnson-Nyquist noise is

$$I_{sh, \, \text{rms}} = \sqrt{\frac{2kT\Delta f}{R_L}}$$  \hspace{1cm} (8)

where, $R_L$ is the load resistance.

4.6 Capacitance and dynamic response

In high-speed light detection application of the photodiode, the dynamic response of the photodiode is of the utmost importance. The dynamic response of the photodiode depends on the drift velocities of the photogenerated carriers, the junction capacitance that is associated with the space charge in the depletion region, and the diffusion of photogenerated carriers from the quasi-neutral regions into the diffusion region. The delay related to the drift can be characterized by a transit time, the amount of time that it takes a photogenerated carrier to reach the quasi-neutral region. It is simply given by,

$$t_{tr} = v_{\text{drift}} \cdot x = \mu E x$$  \hspace{1cm} (9)

where, $v_{\text{drift}}$ is the drift velocity, $x$ is the distance from the point of photogeneration to the quasi-neutral region, and $\mu$ is the mobility of the carrier. Then, the longest possible transit time is,

$$t_{tr}(\text{max}) = \mu E W.$$  \hspace{1cm} (10)

Another delaying factor is due to the bias voltage dependence of the depletion layer width. A change in bias voltage will change the depletion region width as described by $W = K(V_0 - V)^{m/2}$, Equation (5), which in turn changes the amount of exposed space charge. This change in the amount of space charge due to a change in the bias voltage can simply be modeled by the junction capacitance. The junction capacitance is given by,

$$C_{PD} = \frac{\varepsilon A}{W}$$  \hspace{1cm} (11)

where, $\varepsilon$ is the dielectric constant, $A$ is the cross-sectional area of the p-n junction.

Photogenerated carriers in the quasi-neutral region are normally lost by recombination. However, on occasion, the minority species of the photogenerated electron-hole pair can diffuse into the depletion region and contribute to the photogenerated current. Although the contribution to the overall signal by these carrier diffusion is small, a delay due to this diffusion process is observable [40, 41]. The time that takes a minority carrier in the quasi-neutral region to diffuse into the depletion region is approximately,

$$t_{\text{diff}} = \frac{x^2}{4D}$$  \hspace{1cm} (12)

where, $x$ is the carrier’s distance to the depletion region boundary, and $D$ is the diffusion constant.
5. Accumulation mode for signal integration in imaging

The basic operating modes are very useful for real time measurement of light intensity that falls on a photodiode. However, in imaging applications, an integrated signal from the photodiode is often preferred. The integrated signal from the accumulation of photogenerated charge provides a form of statistical binning. This statistical binning not only provides a stronger accumulative signal, but the accumulative signal is a more faithful representation of light intensity that falls on the pixel within the time period of signal integration than an instantaneous measurement, especially for weak or noisy signals. Moreover, in imaging applications, the integration period of the pixels can be synchronized using a shutter. In this case, the stored charge can then be read one by one after the exposure. The result is an image.

Charge accumulation can be achieved in the accumulation mode, and it is used in CCD and CMOS imaging.

![Diagram of accumulation mode operation of a photodiode](image)

**Fig. 4.** An abstract representation of a circuit for accumulation mode operation of a photodiode

Figure 4(a) shows an abstract representation of a circuit for operating a photodiode in accumulation mode. This circuit represents the circuit element of a pixel in an imaging array. The goal of the circuit is to accumulate the photogenerated carriers (as charge) within a set period of time and readout the amount of charge collection as the voltage $V_{out}$. Here is a description of an operation cycle, assuming that there is an optical shutter for exposure control:

1. To begin, the reset switch is closed. The photodiode is in reverse bias, and $V_D = V_{DD}$.
2. The reset switch opens, follow by the shutter. $V_D$ starts to drop.
3. After $t_{int}$, the shutter closes, followed by the read switch, and $V_{out} = V_D$.
4. Read switch opens, and then reset switch closes. Circuit returns to initial state.

To measure the light intensity of the pixel for single image, only one cycle of operation is needed. To measure a series of images (e.g. in video recording), the operation cycle can be repeated continuously. During the integration phase, the rate of voltage drop depends on $I_{ph}$, $I_d$ and $C_{PD}$, which can be described by the following differential equation,
Note that $C_{PD}$ varies with the bias voltage. As discussed in [42], this voltage drop is usually linear for a wide range of values below $V_{DD}$. Therefore, the voltage readout at the end of the integration period can be used as a measure of the amount of light that has fallen on the pixel during the integration phase.

6. CMOS active pixel sensing

There are many ways to implement CMOS pixel sensing using accumulation mode. The simplest active pixel sensing implementation is the T3-APS, as shown in Figure 5. In this implementation, a MOSFET is used to control the reset current. Another MOSFET is used as a source follower (SF) to keep the output voltage the same as $V_D$, and the control of the readout is controlled by yet another MOSFET (Select). The SF-FET prevents discharge during the readout; therefore it is possible to re-read the same pixel twice without losing $V_D$. Because of difficulty in suppressing thermal noise and other design limitations [43-45], the T3-APS has been superseded by other implementations such as T4-APS [43, 44, 46].

![Fig. 5. T3-APS implementation of CMOS pixel sensing.](image)

7. P-i-N photodiode

Increasing the active region where the signal generating photogenerated carriers originated from should in principle increase the collection efficiency. Previously, increasing the active region by increasing $W$ through reverse biased was discussed. To further increase the active region, the device geometry can be altered to include an intrinsic region between the p-type and n-type regions, as shown in Figure 6, this results is a p-i-n junction. Figure 6(e), shows a band diagram of a p-i-n junction under reverse bias. Under reverse bias, certain assumptions can be made because the external field has driven almost all the carriers from the intrinsic region [47]. These assumptions are,

- No net charge in the intrinsic region.
- A very narrow depletion region on the doped side of each of the doped-intrinsic junctions.

With these assumptions, the depletion layer width is simply the width of the intrinsic region; the electric field in the intrinsic region is simply,
and the junction capacitance is given by Equation 11. Consequently, the p-i-n photodiode under reverse bias can have a $C_{PD}$ smaller than a p-n junction photodiode, but the wider depletion region also implies a longer transit time. Therefore, transit time becomes the dominant limiting factor in the speed the device, and operating the device under sufficient reverse bias is essential, and quick estimation in [41] shows that the transit time of a carrier to cross the intrinsic layer is on the order 0.1 ns. Moreover, to optimize for quantum efficiency without sacrificing speed, it is common practice to design the intrinsic layer thickness to be larger than absorption length given by $\alpha^{-1}$, but not much more [41]. Further discussion on p-i-n photodiodes can be found in references [40, 41].

![Photodiode diagram](image)

**Fig. 6.** (a) Structure, (b) carrier distribution, (c) charge distribution, (d) electric field, and (e) band diagram of P-I-N junction under reverse bias.

### 8. Avalanche photodiode

The avalanche photodiode (APD) is the solid state equivalent of a photomultiplier tube. Its main application is for detection of weak optical signal such as single photon events. The APD exploits the impact ionization of carriers by photogenerated carriers under extremely high reverse bias. During its operation, carriers that trigger impact ionization and the carriers that are generated by impact ionization continue to drift to cause more impact ionization events to occur. The result is a cascade of impact ionization events that produces an avalanche effect to amplify the photogenerated current. The amplification of the current is given by a multiplication factor, $M = \frac{I_{mph}}{I_{ph}}$. It is fundamentally related to the ionization coefficients of the carrier by the following Equation,

$$1 - \frac{1}{M} = \int_{0}^{W} \alpha_n \exp \left[ - \int_{x}^{W} (\alpha_n - \alpha_p) dx \right] dx$$

(15)
where, $\alpha_n$ and $\alpha_p$ are the ionization coefficients of the electrons and holes respectively [48]. Empirically, it can be approximated by,

$$M = \frac{1}{1 - \left(\frac{V}{V_{BD}}\right)^{n}}$$

(16)

where, $n$ is material dependent parameter [49]. When taking the dark current into account, it becomes

$$M = \frac{1}{1 - \left(\frac{V-I_R}{V_{BD}}\right)^{n}}$$

(17)

where, $I$ is the total current following through the APD, and $R'$ is the differential resistance observed in junction breakdown [50]. Because the random nature of impact ionization, APD suffers from another form of statistical noise call excessive noise [41]. It is given by,

$$F = M \left[1 - (1 - k) \left(\frac{M-1}{M}\right)^2\right]$$

(18)

where, $k$ is the ratio $\alpha_n/\alpha_p$.

An APD can also be operated in Geiger mode under reverse bias beyond the breakdown voltage. In this case, an electron-hole pair generated by a single photon will trigger the avalanche effect that generates a large current signal. Therefore, photon counting (like particle counting with a Geiger counter) can be achieved in Geiger mode. An APD that can operate in Geiger mode is also known as a single photon avalanche diode (SPAD).

Fig. 7. The cross-section of a SPAD CMOS sensor [51] showing the guard ring surrounding the active region.

Because APDs operate in or near the breakdown region, a physical feature known as a guard ring around the active region is used prevent surface breakdown, as shown in Figure 7. Moreover, the high reverse bias voltage required to produce the avalanche effect had hindered the incorporation of CMOS technology with APD. In 2000, Biber et al. at Centre Suisse d’Electronique et de Microtechnique (CSEM) produced a 12x24 pixel APD fabricated...
in standard BiCMOS technology [52, 53]. Since that pioneering work, there has been a steady growth in the development of CMOS APD [54-56] and CMOS SPAD [51, 57-61] for application such as fluorescence sensing [51, 62, 63] and particle detection [64].

9. Mask-level layout and structure

Mask-level layout is the process in which integrated circuits (IC) are defined in terms of their fabrication layers. In this process, a circuit is defined in terms of the functional layers that are later fabricated at a foundry. The designer specifies sizes, locations, connections and spacings of all of the devices in a circuit. The term “mask-level” means that the design results in specifications for the lithographic masks used to make the IC.

Mask-level layout is done using specialized computer-aided design (CAD) tools, such as CADENCE, L-Edit, or the venerable Magic. The process is essentially one of drawing colored boxes or other shapes, and “connecting” them by drawing other boxes. The view of design is as one looks at an IC held in one’s hand; for example, see Figure 8 to compare the mask-level layout and the fabricated circuit itself.

The layout of a circuit involves the placement of shapes, where the shapes are predefined layers; the layers are represented as different colors or shadings or fill. For example, a P-type MOSFET is shown with the layers labeled in Figure 9. These layers, in addition to other metal layers and, in some processes, another poly layer are available for circuit design. The designer controls the sizes of all of the components (wires, capacitors, transistors, etc.), how they are connected and their location. How large or small these components can be, how close to other layers, and whether layers can overlap are all determined by the design rules for the particular technology. The CAD tools used for layout will have these rules available and can aid the designer by indicating when these rules are violated. The rules help to ensure that the circuit can be fabricated without the fabrication process itself causing the circuit to fail. These failures are a result of limitations in the lithography process that can create undesired short circuits or open circuits. However, the layout tools by themselves do not guarantee proper functionality; the designer must perform a circuit verification step in which the layout is compared to the original circuit to verify that the two match. See [65] for additional information on layout.

What is clearly missing in this process is the third dimension, which is the depth into the substrate or height above the substrate. Since designers have no control over depths or heights, it is “hidden” from the designer in the CAD tools and ignored. While, for integrated photodetectors, the depths of the p-n junctions are critical, the designer still does not have control over these in a standard CMOS fabrication process. Also, it should be noted that most processes are an n-well/p-substrate process, and we will assume that for the discussion of photodetector devices.

The simplest structure is the vertical p-n photodiode; it can be formed as a p+ region in an n-well (Figure 10) or as an n+ region in a p-substrate (Figure 11). The uncovered active area is the region that is intended to be the photon collection area. To prevent unwanted charge carrier generation, other regions of the IC should be covered in a metal layer (see Figure 12). It is also possible to create a p-n photodiode using n-well/p-substrate; the difference with this type of device is that the p-n junction is quite a bit deeper than the junction for the p+ active or n+ active devices. As discussed previously, this can affect the wavelength sensitivity of the device.
Fig. 8. (a) Mask-level layout of a circuit. (b) Microphotograph of the fabricated circuit shown in (a).
Fig. 9. Mask level layout with layers labeled.

Fig. 10. CMOS-based photodiode: p+/n-well.
In order to create a dense array of photodiodes, as needed for a high-resolution imaging device, the ratio of the area designated for the collection of light to the area used for control circuitry should be as high as possible. This is known as the fill-factor. Ideally, this would be unity, but this is not possible for an imaging device with individual pixel read-out. Thus, actual fill-factors are less than one. A layout of the APS pixel as shown in Figure 5 is shown in Figure 13. The fill factor of this 3 transistor pixel is 41% using scalable CMOS design rules. The metal shielding of the circuitry outside of the photodetector is not shown for clarity; in practice, this shielding would cover all non-photoactive areas and can also be used as the circuit ground plane.
In order to create an array of imaging pixels, the layout not only requires maximizing the active photodetector area, but also requires that the power (VDD), control and readout wires be routed so that when a pixel is put into an array, these wires are aligned. An example of this is shown in Figure 14.

Fig. 13. Layout of the T3-APS pixel as shown in Figure 5.

A slightly more complex structure is the buried double junction, or BDJ, photodiode [66]. The BDJ is formed from two vertically stacked standard p-n junctions, shown in Figure 15. The shallow junction is formed by the p-base and N-well, and the deep junction is formed by the N-well and P-substrate. As discussed previously, the depth of each junction is determined by the thickness of the p-base and n-well. Incident light will be absorbed at different depths, so the two junctions will produce currents based on the wavelength of the incident light. The current flow through two junctions is proportional to the light intensity at the junction depth. An example layout of the structure is shown in Figure 16.
Fig. 14. Simple 3x3 arrays of pixels shown in Figure 13. Notice that the wires align vertically and horizontally.

Fig. 15. Cross-sectional view of the BDJ (not to scale).
Fig. 16. Layout of a 24μm x 24μm BDJ photodiode in the AMI 1.5μm process.

The final structure we will discuss is the phototransistor. Typically, a phototransistor can produce a current output which is several times larger than a same area size photodiode due to the high gain of the transistor. However, a major drawback of these phototransistors is their low bandwidth, which is typically limited to hundreds of kHz. Additionally, the current-irradiance relationship of the phototransistor is nonlinear, which makes it less than ideal to use in many applications. Like the photodiode, there are a number of possible configurations for the phototransistor in a standard CMOS process, such as the vertical p-n-p phototransistor and lateral p-n-p phototransistor [67-71].

A cross-section of a vertical p-n-p phototransistor is shown in Figure 17 and an example layout is provided in Figure 18.

10.1 Tune device responsivity

In a standard CMOS technology, a photodiode can be formed using different available active layers, including n-active/p-substrate, p-active/n-well and n-well/p-substrate, to form a p-n junction. In a photodiode, the photo-conversion mostly takes place in the depletion region where an incident photon creates an electron and hole pair with the
electron passing to the n-region and hole to the p-region. Hence, varying the depth at which the depletion region forms in the silicon wafer would control the performance of the photodiodes in terms of responsivity and quantum efficiency. Also, varying the width of the depletion region by appropriately applying a reverse bias to the photodiode, one could control the response time of the detector. A wider depletion region reduces the junction capacitance of the p-n-junction and improves the response time of the detector.

Here, we will aim to understand the effect on responsivity and external quantum efficiency on the design of photodiode structures. Given that all materials in a standard CMOS process are set by the manufacturer, the external quantum efficiency, which takes into account only the photon-generated carriers collected as a result of the light absorption or, in other words, the useful portion of signal generated by interaction of light and photodetector, is more relevant. The external quantum efficiency depends on the absorption coefficient of the material, \( a \) (units: cm\(^{-1}\)) and thickness of the absorbing material. Assuming that the entire incident light is absorbed by the detector, if the photon flux density incident at the surface is \( \Phi_0 \), then the photon flux at depth, \( x \), is given by Beer’s law (Equation 1) [72].

The external quantum efficiency is also a function of wavelength of the incident light. Thus in a CMOS photodiode, one can strategically chose the depth of the location of the depletion region to which photons are likely to penetrate and thereby optimize the photodetector to provide high absorption for particular spectrum of wavelengths. In practical optoelectronic systems development, responsivity, that is defined as the output current divided by the incident light power, may be a more relevant performance metric. Responsivity is related to quantum efficiency by a factor of \( h\mu/q \), where, \( q \) is the electron charge, \( h \) is Planck’s constant, and \( \mu \) is the frequency of the incident photon. The spectral response curve is a plot of responsivity as a function of wavelength.

Thus, to optimize a silicon photodiode structure for detecting blue wavelengths, the depletion region should be near to the silicon surface. For red wavelengths, the depletion region should be placed deeper in the silicon substrate. Based on this idea, Yotter et al. [73] have compared photodiode structures (p-active/n-well and n-well/p-substrate) to develop photodiodes better suited for blue or green wavelengths for specific biosensing applications. The blue-enhanced structure used interdigitated p+-diffusion fingers to increase the depletion area near the surface of the detector, while the green-enhanced structure used n-well fingers to increase the depletion region slightly deeper within the substrate. Bolten et al. [74] provided a thorough treatment of the photodiode types and their properties. They reported that in a standard CMOS process n-well/p-substrate structure provides relatively better quantum efficiency for biosensors operating in visible electromagnetic spectrum.

Using the properties that external quantum efficiency varies as a function of wavelength of the incident light and Beer’s law, many research groups reported the use of buried double p-n junction (BDJ) and buried triple p-n junction structures, which can be implemented with a standard CMOS process, for monochromatic color detection [75, 76]. The BDJ structure has two standard p-n junctions (p-base/n-well/p-substrate) are stacked vertically in the CMOS chip. For the BDJ detector, we obtain \( I_{top} \) (only from top p-n junction) and \( I_{bottom} \) (sum of currents from top and bottom p-n junctions) from the detector. The current ratio, \( I_{top}/I_{bottom} \), \( I_{bottom} \) can be used for the color/wavelength measurements. The CMOS BDJ detector has been used for fluorescence detection in microarrays [77], and for the detection and measurement of ambient light sources [78]. The BDJ color detectors have been used in many chemical and biological sensors such as seawater pH measurement [79] and volatile organic compounds detection [80].
10.2 Monolithic integration of photonic devices on photodetectors
10.2.1 Microlens and microfilters
Most CMOS image sensors are monochrome devices that record the intensity of light. A layer of color filters or color filter array (CFA) is fabricated over the silicon integrated circuit using a photolithography process to add color detection to the digital camera. CFA is prepared by using color pigments mixed with photosensitive polymer or resist carriers. Many recent digital color imaging systems use three separate sensors to record red, green, and blue scene information, but single-sensor systems are also common [81]. Typically, single-sensor color imaging systems have a color filter array (CFA) in a Bayer pattern as shown in Figure 19. The Bayer pattern was invented at Eastman Kodak Company by Bryce Bayer in 1976 [82]. This CFA pattern has twice as many green filtered pixels as red or blue filtered pixels. The spatial configuration of the Bayer pattern is tailored to match the optimum sensitivity of human vision perception. Imager sensors also include microlenses placed over the CFA to improve the photosensitivity of the detection system and improve the efficiency of light collection by proper focusing of the incident optical signal over the photodetectors [83]. A microlens is usually a single element with one plane surface facing the photodiode and one spherical convex surface to collect and focus the light. Thus, as photons pass through the microlens and through the CFA filter, thus passing only wavelengths of red, green, or blue color and finally reach the photodetectors. The photodetectors are integrated as part of an active pixel sensor to convert the incident optical signal into electrical output [84]. The analog electrical data from the photopixels are then digitized by an analog-to-digital converter. To produce a full color image, a spatial color interpolation operation known as demosaicing is used. The image data is then further processed to perform color correction and calibration, white balancing, infrared rejection, and reducing the negative effects of faulty pixels [85, 86].

![Fig. 19. Bayer film pattern and microlenses integrated onto a device.](image-url)

One of the first example of a monolithic microlens array fabricated on the MOS color imager was done using photolithography of a polymethacrylate type transparent photoresist [87].
In commercial camera production, glass substrates are typically used as carrier and spacer wafers for the lenses and are filled with an optical polymer material which is photolithographically patterned to form the microlenses. A fairly straightforward method used in many microlens implementations is to photolithographically pattern small cylinders of a suitable resin on a substrate. The small cylinders are then melted in carefully controlled heating conditions. Hence, after melting they tend to form into small hemispheres due to surface tension forces. However, molten resin had a tendency to spread such that lens size and spatial location is difficult to control. A well-defined spherical surface for the microlens is required to achieve high numerical aperture which improves the image sensor efficiency. Different techniques are used to control the spherical shape and spatial location of the microlens including pre-treatment of the substrate to adjust the surface tension to control the reflow of the microlens [88] and use of microstructures such as pedestals to control the surface contact angle [83]. In more recent processes the glass substrates are eliminated and instead microlenses are made with polymer materials that are molded using master stamps. The molded polymer microlenses are cured with ultra violet exposure or heat treatment. By replacing the glass substrates, wafer-level system manufacturers face fewer constraints on the integration optics and imager integrated circuit enabling the production of compact and efficient imager sensors.

10.3 Waveguides, gratings, and couplers

In this section, we will concentrate on understanding device architectures that deal with monolithic integration of photonic waveguides, gratings and couplers with CMOS photodetectors for applications in optoelectronics to improve quantum efficiency, spectral response selectivity, and planar coupling and guiding of light signals to on-chip photodetectors. CMOS photodetectors operate only in visible and near infra-red region between 400nm and 1.1µm of the electromagnetic spectrum. There are applications in sensing and optical communications in this wavelength region where silicon or CMOS photodetectors can offer low-cost and miniaturized systems. Monolithic integration of photonic components with silicon/CMOS photodetectors started as a major research area since early 1980’s [89-92]. It is advantageous that a monolithic integrated optoelectronic system on silicon use materials typically employed in CMOS-technology. The dielectrics available in CMOS are favorable as the light guiding layer for wavelengths in the visible and near infrared region. The available materials in CMOS processing technology to develop the photonic devices include layers such as silicon nitride [3], Phospho-Silicate Glass (PSG) –SiO$_2$ doped with P$_2$O$_5$ [4] or silicon oxynitride layers deposited as insulating and passivation layers. Confinement of light is achieved by an increased refractive index in the light guiding film, compared to silicon oxide. The first proposed CMOS compatible devices were based on using silicon oxynitride waveguides sandwiched with silicon dioxide (SiO$_2$) layers [93-95].

System-level integration is commonly used in compact spectrometers with Lysaght et al. [96] [97] first proposing a spectrometer system in the year 1991 that would integrate silicon photodiode array with microfabricated grating structures for diffraction of the incident light signals and subsequent detection of the optical spectrum components by the silicon photodiode array. More recent and commercial available compact spectrometers use a CMOS line array. Csutak et al. [98] provided an excellent background for related work done
prior to their research article. After considering the absorption length of silicon and required bandwidth for high-speed optical communications, the improvement of quantum efficiency of the photodetectors remains an important challenge.

10.4 Biosensors on CMOS detectors

Many research groups are working on the idea of contact imaging systems for imaging or detection of a biological specimen coupled directly to the chip surface which was first proposed by Lamture et al. [99] using a CCD camera. As the photodetector components in biosensors, CMOS imagers are preferable to convert the optical signals into electrical signals because of monolithic integration of photodetection elements and signal processing circuitry leading to low cost miniaturized systems [100, 101]. In 1998, a system termed as bioluminescent-bioreporter integrated circuit (BBIC) was introduced that described placing genetically engineered whole cell bioreporters on integrated CMOS microluminometers [102]. In a more recent implementation of BBIC system includes sensing low concentrations of a wide range of toxic substances such as salicylate and naphthalene in both gas and liquid environments using genetically altered bacteria, *Pseudomonas fluorescens* 5RL, as the bioreporter [103]. BBIC system operates on the basis of using a large CMOS photodiode (1.47 mm$^2$ area using n-well/p-substrate structure) for detection of low levels of luminescence signals by integration of the photocurrent generated by the photodiode over time and a current-to-frequency converter as signal processing circuit to provide a digital output proportional to the photocurrent.

Recent implementations of contact imaging include using custom-designed CMOS imagers as platform for imaging of cell cultures [104] and DNA sequencing [105, 106]. Now researchers are working on the integration of molded and photolithographically patterned polymer filters and microlenses with CMOS photodetectors and imagers towards complete development of miniaturized luminescence sensors. Typically, luminescence sensors require an optical excitation source for exciting the sensor materials with electromagnetic radiation and a photodetector component for monitoring the excited state emission response from the sensor materials at a higher wavelength electromagnetic spectrum that is filtered from the excitation input. The next step towards convenient monolithic integration of filters, biological support substrates, and microfluidic interfaces create interesting challenges for engineers and scientists. A recent report discusses the approach of using poly(acrylic acid) filters integrated with custom-designed CMOS imager ICs to detect fluorescent microspheres [107]. Polydimethylsiloxane (PDMS) could offer a more versatile material to fabricate lenses, filters, diffusers and other components for optical sensors [108]. PDMS is a silicone-based organic polymer that is soft, flexible, biocompatible and optically transparent and well amenable to various microfabrication techniques. PDMS can be doped with apolar hydrophobic color dyes such as Sudan-I, -II or -III to form optical filters that work in different regions of visible electromagnetic spectrum [109]. The Authors group recently proposed a prototype compact optical gaseous O$_2$ sensor microsystem using xerogel based sensor elements that are contact printed on top of trapezoidal lens-like microstructures molded into PDMS that is doped with Sudan-II dye as shown in Figure 20 [110]. The molded PDMS structure serves triple purpose acting as immobilization platform, filtering of excitation radiation and focusing of emission radiation onto the detectors. The PDMS structure is then integrated on top of a custom design CMOS imager to create a contact
imaging sensor system. The low-cost polymer based filters is best suited for LED excitation and may not be able to provide optimum excitation rejection performance when laser radiation is used for excitation. As a more traditional alternative, Singh et al. [111] proposed micromachining a commercially available thin-film interference filter and gluing it to the CMOS imager die.

![Xerogel based oxygen sensors](image1)

**Fig. 20.** Fabricated microlenses and xerogel sensors.

### 11. Optical sensor chip with Color Change-Intensity Change Disambiguation (CCICD)

In this section we present a CMOS-sensor chip that can detect irradiance and color information simultaneously. Compared with other BDJ-based systems [112-114], this system includes an irradiance detection pathway that can be used in combination with the color information to provide color change -intensity change disambiguation (CCICD). The irradiance detection pathway is based on the work by Delbruck and Mead [23] with a single standard CMOS photodiode (see Figure 21). Thus, this pathway can function ambient light conditions without an additional light source, is more robust to background light changes, has a higher bandwidth for time-varying signals, and has the ability to emulate adaptation to background light levels, which is an important phenomena found in biological visual systems. The color detection pathway consists of a BDJ photodetector and subsequent processing circuitry to produce a single voltage as the chip output without additional external circuitry. The BDJ produces two currents which are used as inputs to individual logarithmic current to voltage converter circuits whose outputs are converted to a voltage.
difference using a differential amplifier. The output of this pathway is a single voltage that represents the color of the input signal, with better than 50nm resolution. Both pathways are integrated on the same IC.

Fig. 21. Block diagram of the sensor chip pathways (from [115]).

The irradiance detection pathway out is shown in Figure 21. The response of the irradiance detection pathway circuit is logarithmic over the measured irradiance range spanning nearly 3 orders of magnitude.
Fig. 22. Output of the irradiance detection pathway (from [115]).

Fig. 23. Output of the color detection pathway as a function of incident light wavelength (from [115]).
From the experimental results, we can see that the output voltage is larger for longer incident light wavelengths (see Figure 23). So, for a practical implementation based on this chip, a look-up table can be used to map the output voltage to the incident wavelength. Moreover, from Figure 24, the changes in the output voltage caused by irradiance change will not cause confusion between which color (primary wavelength) is detected; the R, G and B curves will not overlap for a normal operating range of irradiance. The reason for this performance is because the I2/I1 ratio from the BDJ is (ideally) independent of light intensity.

12. References


Photodiodes or photodetectors are in one boat with our human race. Efforts of people in related fields are contained in this book. This book would be valuable to those who want to obtain knowledge and inspiration in the related area.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:
