Detectors for Optical Communications: A Review

J. Katz
Communications Systems Research Section

The subject of detectors for optical communications in the visible and near-infrared regions of the spectrum is reviewed. The three generic types of detectors described are photomultipliers, photodiodes and avalanche photodiodes. Although the discussion is from the perspective of space-based applications, most of the information (excluding, perhaps, the material on photomultipliers) is applicable to other optical communications systems (e.g., fiber optics).

I. Introduction

The purpose of this paper is to describe three basic types of photodetectors which may be useful for the different channels of optical communications in the visible-near infrared region of the electromagnetic spectrum ($0.4\,\mu m \lesssim \lambda \lesssim 1.7\,\mu m$): Photomultiplier tube (PMT), photodiode (PD), and the avalanche photodiode (APD). These channels include both guided transmission (through optical fibers) and unguided transmission (e.g. among any combinations of spacecraft/satellites/airplanes/ground stations). The basic physical principle in all these detectors is the conversion of incoming photons directly into charge carriers. The detectors differ in the subsequent processing of the photo-generated carriers: In PMTs electrons are emitted into a vacuum tube which operates as an electron multiplier; in photodiodes the carriers (both electrons and holes) are swept out of the device into the subsequent amplifier circuit; and in APDs, generated carriers first undergo a random multiplication process within the device itself. Other detector types – both electronic and thermal – are not discussed here because of their speed and/or sensitivity limitations. The particular type of detector to be used depends on overall system considerations. PMTs, which are unique in their capability to detect individual photons, are the preferred choice in extremely low signal and low background noise channels. PDs can be used in heterodyne systems (where a local oscillator helps to overcome thermal noise) and in high background noise environments. Finally, APDs give better performance than PDs in systems which are thermal noise limited, and thus are used extensively in high-speed long-range optical fiber links.

In the following sections each of the three generic detector types will be described. Points to be discussed include the basic physical mechanism of operation, materials considerations and operational characteristics in terms of responsivity, noise, speed and limitations for space applications. Finally the last subsection will briefly review the subject of detector arrays which may be useful in acquisition/tracking/pointing subsystems of the communications link.

II. Photomultiplier Tubes (PMTs)

A. Principles of Operation

PMTs are the most sensitive detectors. In free space optical communications they should be used only under conditions

where they can produce a usable output to a single photon detection event. Only characteristics which are relevant to this operational mode will be discussed here. The generic structure of the PMT is shown in Fig. 1(a). It is basically an evacuated glass tube containing a photo-sensitive surface (photocathode), electron multiplier consisting of a chain of dynodes, and an output terminal (anode). Resistor bias network establishes positive potentials between each successive element of the PMT [Fig. 1(b)].

There are many types of PMTs. They differ by their size (PMTs with diameters ranging from roughly 1 to 50 cm are commercially available), structural arrangement and details of the photocathode and dynode chain, and materials of which they are fabricated. Special PMTs employ a magnetic field in addition to the electric field in order to reduce the temporal dispersion of the electrons, thus obtaining faster response. (These cross-field PMTs have lower gains than their all-electrostatic field counterparts.)

The photo-detection process in the PMT can be basically described as follows: a photon impinges upon the photocathode, and there is a certain probability (called "quantum efficiency") that a (primary) electron will be emitted. This electron (for larger light intensities several primary electrons can be emitted almost simultaneously) is focused and accelerated on the first dynode by the electric field established by the bias network. When it reaches the first dynode it has an energy of 200-500 eV, and it produces low-energy secondary electrons. (The characteristics of this process depend on the dynode material and on the primary energy.) These electrons are again focused and accelerated to the second dynode, where more electrons are emitted in a secondary emission process. This process is repeated at each dynode (there can be from several up to about 15 dynodes in a PMT), and as a result there is a large average number of electrons at the anode for each photoelectron emitted by the photocathode. It should be noted that there are some newer versions of PMTs where the electron multiplier is not a dynode chain but a microchannel plate (MCP). The inner walls of the channels of the MCP are coated with a material with a high secondary emission coefficient. The gain is thus distributed along the channels, and is usually smaller than that achievable with regular PMTs.

B. Basic Parameters

1. **Quantum efficiency**. Quantum efficiency (η) is a property of the photocathode of the PMT and is defined as the ratio of the number of photoelectrons emitted from the photocathode to the number of photons impinging upon it. The most popular class of photocathode materials is the multi-alkalis (Na-K-Sb-Cs) which has good response from the UV to near IR [Fig. 2(a)], with $\eta \simeq 20\%$ at $\lambda = 0.53\ \mu m$. The long wavelength response can be extended to 0.9 μm by special processing. GaAs activated in cesium is used in applications which require a higher sensitivity in the near IR. For this material $\eta \simeq 20\%$ at $\lambda = 0.8\ \mu m$ and $\eta \cong 25\%$ at $\lambda = 0.53\ \mu m$ in commercially available PMTs [Fig. 2(b)]. There are no photocathodes with good quantum efficiency for $\lambda > 1\ \mu m$. The only material is Ag-O-Cs, which has $\eta < 0.1\%$ at the Nd:YAG laser wavelength ($\lambda = 1.06\ \mu m$).

2. **Gain**. Gain (G) is the ratio of the number of electrons leaving the PMT anode to the number of photoelectrons emitted by the photocathode. Because of the statistical nature of the secondary emission process at each dynode, G is a random process. The probability density function of G has been calculated and found to be different from the Gaussian distribution, so it cannot be fully characterized only by its first two moments. Since in our application we want to be able to detect single photoelectron events, the magnitude of the average gain, $\bar{G}$, needed is determined by the condition that the peak current by a single electron response is larger than the RMS input current noise to the next stage of detection (e.g., pre-amplifier). This condition can be expressed as

$$\frac{q\bar{G}}{2t_r} \gg \sqrt{\frac{4KT_eB}{R}} \tag{1}$$

The term on the lefthand side of Eq. (1) is the approximate peak of the single photoelectron current pulse (t_r is the rise time of the PMT and q is the electron charge), and the term on the right hand side is the thermal (Johnson) noise current of the system to which the PMT is connected (K is Boltzman constant, B is the bandwidth, R is the input resistance and T_e is the equivalent noise temperature of the amplifier). For $B \simeq 1/t_r$, $R = 50\Omega$, and $T_e = 900$ K (corresponding to a typical commercial amplifier with a Noise Figure of 4.8 dB), Eq. (1) results in

$$\bar{G} \gg \frac{12500}{\sqrt{B\,[\mathrm{GHz}]}} \tag{2}$$

As an example, for $B = 140$ MHz (corresponding to $t_r \simeq 2.5$ ns), a PMT with $G \gg 3 \cdot 10^4$ is needed. This requirement can be easily met in commercially available PMTs which possess average gains of up to 10^6 to 10^7. For a given PMT, the average gain can be modified by varying the voltage V applied to it, approximately according to the following expression

$$\bar{G} = aV^b \tag{3}$$

where a and b are constants.

3. **Time response.** The time response of the PMT is characterized mainly by two parameters; both are affected by the dynode structure and the voltage applied to the PMT:

(1) Transit time is the delay between emission of a photoelectron from the photocathode and the instant when the anode output pulse reaches its peak amplitude. Typical transit times in regular PMTs are 10 to 100 ns (~1 ns for MCP-PMTs).

(2) Rise Time (t_r) is the time it takes the anode signal amplitude to rise from 10% to 90% of its peak value. Typical rise times are 2 to 20 ns for regular PMTs, and cross-field PMTs are up to an order of magnitude faster.

4. **Noise.** The relevant noise parameter in photon counting is the anode dark current, or, more specifically, the rate of dark counts. Since by using the high gain of the PMT we can virtually overcome effects of thermal noise in the system, the dark current is the only relevant system-generated noise (there are, of course, other noise sources, such as background radiation or the statistical nature of the signal radiation). The dark current increases with increasing PMT voltage.

Sources of dark current are:

(1) Thermionic emission of electrons from the photocathode: Since the rate of this process is proportional to $T^2\ e^{-E_W/KT}$ (where E_W is the work function of the photocathode material, and T is the absolute temperature), a modest amount of cooling of the PMT (to ~ -30°C) can virtually eliminate this mechanism.

(2) Unavoidable imperfections of the PMT: These include, for example, ionization of residual gases in the tube, field emission (mainly at high voltage levels), ohmic leakage and scintillation in the glass envelope or faceplate of the tube.

(3) Dark counts caused by Cherenkov radiation due to cosmic rays passing through the PMT faceplate (that part of the tube envelope facing the photocathode): This radiation is the ultimate lower limit of the dark current because cosmic rays can not be blocked. The flux of high energy cosmic rays is about $2 \cdot 10^{-2}$ cm^{-2}-sec^{-1} at sea level and about 1 cm^{-2}-sec^{-1} above earth atmosphere (exact values depend upon orientation of the PMT and other effects which are beyond the scope of this document). Each high energy particle can emit many photons (up to ~100) when passing through the PMT faceplate. Because of these large noise pulse levels, they can be effectively discriminated against the usually lower signal pulses.

C. Reliability

One of the major issues in using PMTs in space applications is their reliability. This issue is important when comparing PMTs to solid-state detectors, described later in this section, which have very long lifetimes. More and more data, albeit not under outer space conditions, are being accumulated mainly from particle physics experiments which use thousands of PMTs. Values of MTTFs (mean time to failure) are more than 2×10^7 hours for catastrophic failures (e.g., cracks in the glass envelope, short or open circuit of the electrodes) and $4 \cdot 10^6$ hours for partial failures (e.g., increased values of dark current, reduced photocathode quantum efficiency, gain drifts). These are long lifetimes even without taking into account the fact that, at least reliability-wise, a PMT can be considered as a system and not as a component. Of course, more testing needs to be done in order to evaluate PMT reliability in the space environment. It should be noted, however, that many space missions have included PMTs on board.

III. Photodiodes (PDs)

A. Principles of Operation

In contrast to PMTs, PDs are unity gain devices, i.e., there is no internal gain mechanisms in the device, so a subsequent electronic amplification of the resulting signal is always needed. The dominant detection process in PDs in our wavelengths region of interest is intrinsic absorption, schematically depicted in Fig. 3. Since this process involves a transition of an electron from the valence band to an unoccupied state in the conduction band of a semiconductor, the minimum photon energy required is Eg, the bandgap energy of the semiconductor, or

$$h\ \nu \gtrsim Eg \tag{4}$$

where h is Planck constant and ν is the frequency of the light radiation. Using $\nu = c/\lambda$ (c is the light velocity in vacuum), we obtain the cutoff wavelength λ_0, which is the longest wavelength that can be detected by intrinsic absorption (other absorption mechanisms are also possible but are of less relevance to our application):

$$\lambda_0\ [\mu m] = \frac{1.24}{Eg[eV]} \tag{5}$$

For $\lambda < \lambda_0$, the absorption coefficient α increases with decreasing wavelengths. The details of the $\alpha(\lambda)$ curve depend on the band structure of the specific semiconductor (including doping effects), on the temperature and on the electric field in the semiconductor (Franz-Keldysh effect). The lower limit on the wavelengths that can practically be detected is reached

when α is so large that most of the incoming photons are absorbed close to the surface, where several phenomena tend to reduce the efficiency of the detection process.

In order for the generated carriers (Fig. 3) to contribute to current in the external circuit, they have to be separated (otherwise they recombine). The usual method of obtaining this goal is to construct the detector as a pn diode. Thus carriers which are generated either in the depletion region, or roughly within one diffusion length from it will contribute to the photogenerated current. The pn diode is operated in a reverse bias, which improves the diode performance in terms of linearity, speed of response and quantum efficiency.

For high speed applications, two generic diode structures are used: PIN diode (Fig. 4) and Schottky diode (Fig. 5). In the PIN PD (Fig. 4) the region where most of the light is absorbed (I, or intrinsic, region) is designed to have low doping level, thus resulting in a fairly uniform electrical field in that region. The diode bias voltage is chosen such that the peak field (i.e., at the junction itself) is lower than the value needed for avalanche multiplication, and the field magnitude at the other edge of the intrinsic region is sufficiently large to maintain saturation drift velocity of the carriers. This assures the minimum possible transit time (see Section III.B) and thus helps to optimize the PD temporal response. The Schottky PD (Fig. 5) has a simpler structure: In its most basic structure it consists of a semiconductor on which a thin film ($\lesssim 500$ Å) of metal is deposited. Schottky PDs usually have lower quantum efficiency than PIN PDs, but they have faster response and extended response towards the short wavelengths region.

B. Basic Parameters

1. **Quantum efficiency**. Quantum efficiency (η) is defined as the ratio of the number of charge carriers contributing to the current (i_{ph}) and the number of photons impinging upon the diode surface per second.

$$\eta = \frac{i_{ph}/q}{P/h\nu} = \frac{hc}{q\lambda}\frac{i_{ph}}{P} \tag{6}$$

where P is the optical power. An ideal photodiode has $\eta = 1$ (or 100%), and practical PDs approach this ideal value. A related parameter is the *responsivity*, $\mathscr{R}$, of the photodiode, which has units of Ampere/Watt. The relation between η and $\mathscr{R}$ is

$$\eta = \frac{1.24\,\mathscr{R}\ [\text{A/W}]}{\lambda\ [\mu\text{m}]} \tag{7}$$

η (and $\mathscr{R}$) depends upon the diode material, structure, wavelength and temperature. Among the things done in order to optimize the structure of the PD are tailoring the depletion region to overlap as much as possible with the region where most of the light is absorbed (subject to speed constraints) and employing antireflection coating on the photosensitive surface of the PD (see Fig. 4).

2. **Speed of response**. Speed of response is determined by the combination of the following time constants:

(1) Carrier drift time in the depletion region, which is proportional to the width of the region and inversely proportional to the carrier velocity.

(2) Diffusion time constant of minority carriers which are generated within approximately one diffusion length from the depletion region.

(3) *RC* time constant associated with the diode shunt capacitance and series resistance (including contributions from the diode itself, and from the diode housing, wires and other parasitics). This time constant can be minimized by reducing the diode area and fabricating it on semi-insulating substrates (in GaAs).

A general rule of thumb is to minimize the diffusion time constant and to design the transit time to be approximately equal to the *RC* time constant. Usually there is a tradeoff between speed of response and the quantum efficiency. Both PIN and Schottky diodes with frequency response exceeding 20 GHz (and rise-times of 35 ps) have been demonstrated in the laboratory and some are commercially available. Recently a GaAs Schottky PD with a 3 dB bandwidth of 100 GHz has been demonstrated.

3. **Noise in PDs**. A typical detector circuit is shown in Fig. 6a, and the equivalent circuit, including all the noise sources, is shown in Fig. 6b. The signal source is given by

$$i_s = \frac{\eta q \lambda}{hc} P_s \tag{8}$$

where P_s is the optical power impinging upon the diode. The noise sources, are:

(1) The diode equivalent circuit contains three white spectrum shot-noise current sources which originate from carriers generated by signal photons $\left(\overline{i_s^2}\right)^{1/2}$, background photons $\left(\overline{i_b^2}\right)^{1/2}$, and dark current $\left(\overline{i_d^2}\right)^{1/2}$ (which is caused by generation recombination process in the depletion region, minority carrier diffusion and surface leakage). These three current sources are given by

$$\overline{i_s^2} = 2\frac{q^2\eta}{h\nu} P_s B \qquad [A^2] \tag{9a}$$

$$\overline{i_b^2} = 2\frac{q^2\eta}{h\nu} P_b B \qquad [A^2] \qquad (9b)$$

$$\overline{i_d^2} = 2 q i_d B \qquad [A^2] \qquad (9c)$$

where P_b is the background optical power.

(2) The bias and load resistances have a white spectrum thermal (Johnson) noise current source given by

$$\overline{i_{th}^2} = \frac{4KTB}{R_e} \qquad (10)$$

with R_e as the parallel combination of R_B and R_L.

(3) Since PDs are virtually always followed by an amplifier, its noise has to be taken into account. There are two main types of front-end amplifiers.

The first type is the transimpedance amplifier, shown on Fig. 7(a), where the thermal noise contribution is mainly determined by the feedback resistance R_f. The second type is the high impedance amplifier, whose first stage, which determines the overall behavior of the amplifier, is made of either a field-effect transistor (FET) (Fig. 7(b)) or a bipolar junction transistor (BJT). In the case of an FET, $\left(\overline{i_a^2}\right)^{1/2}$ [Fig. 6(b)] can be neglected. The voltage noise source is given approximately by

$$\overline{v_a^2} \simeq \frac{2.8KTB}{g_m} \qquad [V^2] \qquad (11)$$

where g_m is the FET transconductance. This noise source can be transformed into an equivalent noise source which appears in parallel to the PD noise sources and replaces all the other noise sources $\left(\text{i.e., } \overline{i_{th}^2}, \overline{v_a^2}\right)$:

$$\overline{i_{th}'^2} = \frac{4KTB}{R_e} + \frac{2.8KTB}{g_m} \times \left[\frac{1}{R_e^2} + \frac{[2\pi(C_d + C_L)]^2 B^2}{3}\right] \qquad [A^2] \qquad (12)$$

where C_d and L_L are the diode and the load capacitances, respectively. From Eq. (12) we see that low capacitances are important not only for high frequency but also for low noise performance.

For the BJT front-end amplifier we have

$$\overline{i_a^2} = \frac{2KTB}{r_e} \qquad (13)$$

$$\overline{v_a^2} = \frac{2KTB}{\beta} r_e \qquad (14)$$

where r_e is the AC input resistance of the BJT and ($r_e \simeq KT/qI_B$) and β is its current gain. As in the case with a FET, the amplifier noise dominates for large values of R_e. For the BJT high-impedance front-end amplifier there is an optimum bias point which yields the following value of the BJT thermal noise:

$$\overline{i_{th,\min}^2} = 4KTB \cdot \frac{2\pi(C_d + C_L)B}{\sqrt{3\beta}} \qquad (15)$$

By comparing Eqs. (12) and (15) we see that BJTs are better than FETs in high frequency operation (B^2 vs. B^3 dependence). The reason is that the transconductance of the FET is basically fixed, while that of the BJT can be increased by varying the bias point.

The signal to noise ratio at the PD-amplifier combination is found by combining Eqs. (8), (9) and (12) or (15)

$$\frac{S}{N} = \left(\frac{\eta q \lambda}{hc}\right)^2 \frac{P_s^2}{\left[2q(i_s + i_b + i_d) + \frac{4KT_e}{R_e}\right]B} \qquad (16)$$

where T_e is an equivalent noise temperature which depends on the circuit bandwidth, and whose value can be found from Eq. (12) or (15).

As an example, consider a PD/FET combination with the following parameters: T = 300 K, λ = 0.85 μm, P_s = 0.1 μW, i_d = nA, i_b = 0, B = 100 MHz, $(C_d + C_L)$ = 1 pF, R_e = 10 KΩ, g_m = 10 mS and η = 0.8. These parameters are achievable with a careful design (including hybrid or monolithic integration of the PD and the FET). Under these conditions the major noise mechanism is still the thermal noise of R_e. The resulting signal to noise ratio is $S/N \simeq 4.1$. The S/N can be improved by increasing R_e. Postamplification equalization is usually employed with these high-impedance (integrating) amplifiers.

An important parameter characterizing the overall sensitivity of the PD (and any detector) is the Noise Equivalent

Power (*NEP*) (or its inverse, *Detectivity*, $D \equiv 1/NEP$) which is the optical signal power for which $S/N = 1$. Dimensions of the *NEP* are Watt/$\sqrt{\text{Hz}}$. In many cases the Detectivity is normalized to the area of the detector and is denoted by D^*:

$$D^* \equiv D\sqrt{\text{Area}} \quad [\sqrt{\text{Hz}}\text{-cm/W}] \tag{17}$$

For optical communications, D^* is usually not a relevant parameter since the light distribution on the PD is not uniform. In addition, contribution from peripheral leakage currents in small-area PD does not scale as the area. Typical *NEP* of commercially available Si PDs (at their peak wavelength) is 10^{-14} to 10^{-13} W/$\sqrt{\text{Hz}}$ (corresponding D^* are approximately $0.5\text{-}5 \cdot 10^{12}$ cm-$\sqrt{\text{Hz}}$/W).

C. Materials Considerations

1. **Silicon**. Silicon (Si) is by far the preferred material because it is technologically the most mature. With an indirect bandgap of 1.1 eV it can be used for the wavelength region $0.4\ \mu\text{m} \lesssim \lambda \lesssim 1.1\ \mu\text{m}$, although special structures have to be used in order to extend the efficient range of operation beyond 1 μm. Quantum efficiencies at the peak wavelengths (0.8 to 0.9 μm) can exceed 80%. Quantum efficiencies of ~60% are available at the Nd:YAG laser wavelength (1.06 μm) and its second harmonic (0.53 μm). Dark currents are as low as 0.1 nA in commercially available Si PDs.

2. **Germanium**. Germanium (Ge) which can cover the $0.8\ \mu\text{m} \lesssim \lambda \lesssim 1.7\ \mu\text{m}$ region has the disadvantage of very high dark currents due to its smaller bandgap and higher surface recombination. Dark currents are typically 3 orders of magnitudes more than in Si. Quantum efficiencies are comparable to Si.

3. **GaAs**. GaAs basically covers approximately the same wavelengths region as Si ($0.4\ \mu\text{m} \lesssim \lambda \lesssim 0.9\ \mu\text{m}$), and thus it is used less extensively. The advantage of GaAs is that semi-insulating substrates are available, and thus devices with very low capacitance can be fabricated on the structures. Very fast Schottky PDs made in this fashion have been recently demonstrated. Other advantages of GaAs are its higher radiation resistance (20 to 30 times more than silicon) and its ability to operate at higher temperatures (because its bandgap is larger than in silicon) – both properties which are especially important in space application.

4. **Ternary and Quarternary Components/Hetrojunctions**. Different combinations of materials from the GaAlAs, InGaAsP and GaAlAsSb systems can be utilized to make PDs which can span at least the $0.6\ \mu\text{m} \lesssim \lambda \lesssim 1.7\ \mu\text{m}$ region. Once these materials and their fabrication processes become more technologically mature, PDs will have excellent performance characteristics. By including heterostructures in these materials, one could separately tailor the quantum efficiency of the PD, its spectral response and the speed of response. Thus $\eta \gtrsim 0.95$ is achievable even at high speed operation and at selected regions of the spectrum. Today performance of these PDs is better than Ge, but somewhat worse than Si, in particular in terms of higher dark currents.

IV. Avalanche Photo-Diodes (APDs)

A. Principles of Operation

The Avalanche Photo-Diode (APD) is a photodiode which contains an additional region where the photogenerated carriers are multiplied via the avalanche process. In this multiplication region the electric field is very high (several times 10^5 V/cm). Carriers are accelerated by the field, and upon acquiring enough kinetic energy (approximately 1.5 E_g) they can create new carrier-pairs by impact ionization. As this process repeats itself, more carriers are generated, which is an effective amplification process, as depicted schematically in Fig. 8.

A structure of a common version of an APD (Reach-Through APD) and the electric field distribution across it are shown in Fig. 9. The incoming light is absorbed mostly in the low doped (π) region. in a similar way to the drift region in a PIN diode. Holes are swept out the p-contact, and electrons enter the high field region near the np junction where the multiplication process takes place. As we will now see, it is important which type of carrier (electrons or holes) enters the multiplication region. The basic reason is electrons and holes have different ionization coefficients in a given material.

The ionization coefficient, denoted by α_n (for electrons) or α_p (for holes) is defined as the number of ionizing collisions per unit length for each carrier. Both α_p and α_n are strong functions of the electric field $\mathscr{E}$, and they are expressed by the empirical formula

$$\alpha_n = A_n e^{-(B_n/\mathscr{E})^u} \tag{18a}$$

$$\alpha_p = A_p e^{-(B_p/\mathscr{E})^u} \tag{18b}$$

where A_n, A_p, B_n, B_p, and u are material parameters. The ratio between the ionization coefficients

$$k \equiv \frac{\alpha_p}{\alpha_n} \tag{19}$$

(which by itself depends on the electric field) is a key parameter of the APD. $k = 0$ (or $1/k = 0$) means that only electrons (or only holes) can create secondary pairs via impact ionization. In this case, and assuming constant electric field in the multiplication region, the gain M is given by

$$M_{p,n} = e^{\alpha_{p,n} W} \tag{20a}$$

where W is the width of the multiplication region. The other extreme case is when $\alpha_p = \alpha_n$. Now the multiplication process is influenced by feedback mechanisms, and the resulting expression for the gain is

$$M_{p,n} = \frac{1}{1 - \alpha_{p,n} W} \tag{20b}$$

For general values of k we obtain

$$M_n = \frac{\left(1 - \frac{1}{k}\right) e^{\alpha_p W \left(1 - \frac{1}{k}\right)}}{1 - \frac{1}{k} e^{\alpha_p W \left(1 - \frac{1}{k}\right)}} \tag{21a}$$

and

$$M_n = \frac{(1-k)\, e^{\alpha_n W (1-k)}}{1 - k\, e^{\alpha_n W (1-k)}} \tag{21b}$$

When we can not assume a constant electric field in the multiplication region, the expressions are more complicated. In order to avoid numerical calculations, we define effective values so that

$$\int_{x_1}^{x_2} \alpha_{p,n}\, dx \equiv \overline{\alpha_{p,n}} \cdot \overline{W_a} \tag{22a}$$

$$\overline{k_1} \equiv \overline{\alpha_p} / \overline{\alpha_n} \tag{22b}$$

and

$$\overline{W_a} \equiv x_2 - x_1 \tag{22c}$$

where $\{x_1, x_2\}$ is the interval which contributes significantly to the multiplication process (i.e., the high-field region).

Inspection of Eqs. (21) shows that in order for the gain to be less sensitive to minor field variations, those carriers which have a higher ionization coefficient should be injected into the multiplication region. As we will see later, the k-parameter is very important in determining the noise characteristics of the APD. It should be noted that $\alpha_{p,n}$ represents average parameter values of the impact ionization process, and thus the gain mechanism of the APD is a random process. In Fig. 10 we see an example of the probability density function for the number of electrons leaving the multiplication region. It is interesting to note that even for high multiplication values, M, the most probable event is less than the average value.

B. Basic Parameters

1. Gain. Gain, or multiplication factor, M, of the APD is basically defined as the ratio of the actual current of the APD to the current that would have flown if no impact-ionization processes took place. The gain varies with the applied reverse voltage V_D on the APD approximately according to the empirical formula

$$M = \frac{1}{1 - \left(\frac{V_D}{V_B}\right)^u} \tag{23}$$

where V_B is the breakdown voltage of the diode and u is an empirical constant. As we will see later, the actual value of M is chosen to optimize the signal to noise ratio of the APD. These values are of the order of 100. In addition, there is a limit to the maximum gain that can be obtained from an APD in a regular bias circuit. Since

$$V_D = V_{KK} - I_D R_B \tag{24}$$

where V_{KK} is the voltage of the bias network, I_D is the APD current and R_B is the bias resistor, and

$$I_D = M(i_{ph} + i_d) \tag{25}$$

where i_{ph} and i_d are the photo-generated and the dark currents, respectively (before multiplication), then by setting V_{KK} to its maximum allowable value $V_{KK} = V_B$, using Eqs. (14) and (25) in Eq. (23) we obtain the maximum multiplication:

$$M_{\max} = \left[\frac{V_B}{(i_{ph} + i_d) R_B u}\right]^{1/2} \tag{26}$$

From Eq. (26) we can see the importance of reducing the dark current of the APD. In addition, we see the saturation effects caused by the signal that can introduce distortions which are

of great concern in analog applications. Typical operating gains of APDs are 100 to 500.

2. **Time response.** There are two basic limitations on the time response of APDs. At high multiplication factors, the time constant associated with the avalanche process is the dominant one. Because of the feedback effects (which are due to non-zero k values), this time constant is proportional to M. Thus at the high gain region the gain-bandwidth product is constant, given by

$$M_{p,n}(0)\,B \simeq \frac{1}{2\pi\chi\,\tau_a\,k,(1/k)} \tag{27}$$

where $M_{p,n}(0)$ is the DC value of the multiplication, τ_a is the effective transition time of carriers in the multiplication region, and χ is a slowly varying function of $k(1/k)$ (e.g., $\chi(1) = 1/3, \chi(10^{-3}) = 2$).

At low multiplication factors,

$$M(0) \ll \frac{W_d}{3\,k\,\chi\,W_a} \tag{28}$$

(where W_d is the width of the drift region), the dominant constraint is the drift time, similar to the PIN PD. Thus in this region the bandwidth is constant, and the gain-bandwidth product increases linearly with M:

$$M(0)\,B \simeq M(0) \cdot \frac{v_n}{2\,W_d} \tag{29}$$

where v_n is the drift saturation velocity of the electrons. The overall behavior of the gain-bandwidth product is shown in Fig. 11 for typical APD parameters: $(\tau_a\ \chi\ k) = 5 \cdot 10^{-13}$ s, $W_d = 50\ \mu$m and $v_n = 10^7$ cm/s. Since the gain of the PMT is usually determined from noise considerations, the bandwidth is determined by it too. Typical rise times are 0.1 to 1 ns, and typical gain-bandwidth products are 100 to 200 GHz.

3. **Noise.** The random nature of the gain process in the APD and in particular the feedback effects resulting from non-zero k values (i.e., both carrier types can initiate impact-ionization) cause an additional noise mechanism, called "Excess Noise," which increases at a faster rate than the rate at which the signal is amplified with increasing values of gain M. The excess noise prevents us from operating the APD at very high M values and thus virtually overcoming the effects of thermal noise, as is the case of PMTs.

Using basically similar analysis to that done for the PD, the expression for the signal to noise ratio of the APD is given by

$$\frac{S}{N} = \left(\frac{\eta q \lambda}{hc}\right)^2 \times \frac{(M P_s)^2}{\left\{2q\,(i_s + i_b + i_d - i_{ds})\,M^2\,F + \left[2q\,i_{ds} + \dfrac{4\,K\,T_e}{R_e}\right]\right\} B} \tag{30}$$

where i_{ds} is that component of the dark current that does not flow through the avalanche region and thus does not get multiplied (e.g., surface leakage current) and F is the excess noise factor of the APD which denotes the additional noise above that which is due to amplified shot-noise alone. (The other terms in Eq. (30) are explained in conjunction with Eq. (16).) F is a function of both k and M.

The excess noise factor for pure electron injection into the multiplication region is given by:

$$F_n = M_n - (1 - k)\frac{(M_n - 1)^2}{M_n} \xrightarrow[M_n \gg 1]{} kM_n \tag{31a}$$

and similarly for holes

$$F_p = M_p - \left(1 - \frac{1}{k}\right)\frac{(M_p - 1)^2}{M_p} \xrightarrow[M_p \gg 1]{} \frac{1}{k}\,M_p \tag{31b}$$

For $k = 1$, $F_n = F_p = M$, and for $k, (1/k) = 0$, $F_n = F_p = 2$. Generally, for a given operating region, F is sometimes approximated by

$$F = M^x \tag{32}$$

where x is a non-ideality factor.

From Eq. (30) we see that an ideal APD (i.e., $k(1/k) = 0$) can operate, in principle, like a PMT: By increasing M we can eliminate effects of thermal noise. However, in real APDs, the shot-noise is amplified more than the signal, so the optimum operating point M_{opt} is not at $M \to \infty$ but when the two noise terms in the denominator of Eq. (30) are equal. For a signal limited shot-noise, and assuming $F = kM$, we obtain

$$M_{opt} = \left[\frac{2\,k\,T_e}{q} \cdot \frac{h\,c}{\eta\,\lambda\,k} \cdot \frac{1}{R_e\,P_s}\right]^{1/3} \tag{33}$$

M_{opt} depends on the excess noise, load resistor and signal levels. In many applications we cannot neglect the dark and background currents. In this case the expressions become more complicated, though the calculations are straightforward. Figure 12 shows a graph of an *NEP* of an APD vs *M*. The optimum operating point is clearly seen. For lower values of *M* the thermal noise dominates, and for higher values of *M* the excess noise dominates. Typical values of *NEP* for Si APDs are of the order of 10^{-14} W/$\sqrt{\text{Hz}}$ in the near infrared region. Typical values of total dark current are 0.1 to 1 μA, which is mainly due to surface currents.

C. Materials Considerations

Wavelengths of operation, quantum efficiencies and dark currents for different materials are basically the same for PDs and APDs. For a brief review on these aspects the reader is referred to Section III.C. The material parameter that is unique to the APD operation is $k \equiv \alpha_p/\alpha_n$. Table 1 lists this factor for various materials. We see that for all the materials considered to be used in optical communications $0.01 \lesssim k \lesssim 100$. For $\lambda \lesssim 1.1$ μm silicon is the best material for APDs also from the excess noise factor aspect. The basic problem in APD technology today is that good materials which can be used for $\lambda \gtrsim 1.1$ μm (i.e., where Si can no longer be used) have $k(1/k)$ factors which are too close to 1, which implies large excess noise factors. Some novel structures have been devised to overcome this limitation by using heterostructure devices where the dynamics of electrons and holes can be separately controlled, thus yielding effective *k*-values which are much larger than the bulk values. In certain atomic systems (e.g., GaAlSb) one could also possibly utilize certain details of the band structure in order to enhance the *k*-factor.

D. The APD as a Solid State PMT

Under certain conditions, APDs can be operated in a "geiger-tube" mode so that they can respond to single electron events and thus can be used as photon counters. In order to operate in this mode, the APD must be cooled to liquid nitrogen temperatures in order to reduce the dark current (to approximately 100 electrons/s or less). The diode is biased at a voltage level above the avalanche breakdown voltage (which is lower than the dielectric breakdown voltage of the APD) by an amount of up to several tens of volts. A single carrier, which is either photo or thermally generated, has a high probability of causing sustained avalanche. At this point the diode voltage drops to the avalanche breakdown voltage, thus producing a voltage pulse whose magnitude is the difference between the bias and the avalanche voltages. This pulse can be easily detected without any further amplification. Then the avalanche has to be quenched and the device is ready for a new detection event. Since quantum efficiencies of solid-state detectors are higher than those of PMTs, and since for relatively high over-bias the probability that a single carrier will generate sustained avalanche is high (up to 50%), APDs operating in this mode may be an attractive alternative to PMTs, especially with the added bonus of their improved size, ruggedness and reliability.

V. Detector Arrays

In order to perform beam pointing, acquisition and tracking optically, it is necessary to use an optical detector that provides some position information from the received light signal. This information can be derived from single detectors ("position sensors"), small arrays, (e.g., quadrant detectors) or large arrays (e.g., CCD arrays) which are usually used for imaging but can also be used to perform functions such as one-step acquisition.

A. Position Sensors

Position sensors are usually PIN PDs with relatively large areas (~1 cm^2). Photogenerated current is collected from four electrodes, which are placed in pairs across opposite edges of the detector area. The position of the received light spot on the detector can be determined from the relative magnitudes of the currents at the individual electrodes. The absolute position accuracy in this device is very coarse (±10% approx.) but it can detect very minute changes in the position (on the order of

$$\Delta x[\text{Å}] \simeq \sqrt{\Delta f[\text{Hz}]/P_s[\mu\text{W}]}$$

where Δf is the system bandwidth).

B. Small Arrays

Solid state quadrant detectors are available in PDs and APDs. The common application of this segmented detector is to detect position deviations from the optical axis of a system, and thus they are best suited to perform tracking functions. Changes of the order of

$$\Delta x[\text{Å}] \simeq \sqrt{\Delta f[\text{Hz}] \times \text{spot area } [\text{cm}^2]/P_s[\text{W}]}$$

can be detected with quadrant detectors.

MCP-PMTs (see Section II.A) can utilize the fact that spatial information is preserved in the MCP amplification. PMTs with several anodes are available (from 4 X 4 to approximately 20 X 20 anode arrays). Several hundreds of photoelectrons are needed in order to achieve optimum performance in

terms of spatial resolution. For the 4 X 4 anode arrays $\Delta x \sim$ 0.2 mm which is about 1% of the photocathode area.

C. Large Arrays

Large arrays are usually used for imaging, similar to TV tubes. The most common type are CCD arrays, which are arrays of PDs which contain, on the same chip, a mechanism for charge read-out, similar to a "bucket-brigade" operation. A detailed description of CCD arrays is beyond the scope of this paper. These devices are made of silicon, so the quantum efficiency is comparable to Si PDs. There are additional noise mechanisms associated with the read-out process. Under optimized conditions (cooling to 77 K and readout at an optimum rate) RMS noise levels corresponding to 10 electrons per CCD cell can be achieved. Array sizes up to 800 X 800 cells are available.

Acknowledgment

The following figures were reprinted or adapted by permission:

Figs. 1(a), 2(a): Engstrom, R. W., "RCA Photomultiplier Handbook," RCA Corp., Lancaster, Penn. (1980);

Figs. 1(b), 2(b), 9: Various (uncopyrighted) RCA data sheets;

Figs. 4(a), 5(a), 6, 7(a), 8, 11, 12: Muller, J., "Photodiodes for Optical Communications," in *Advances in Electronics and Electron Physics, 55*, Marton and Marton, Editors, Academic Press, New York (1981);

Fig. 10: McIntyre, R. J., "The Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes: Theory," IEEE Trans. Electron. Dev. *ED-19*, pp. 702-713 (1972).

Bibliography

General

Yariv, A., "Introduction to Optical Electronics," 2nd Edition, Holt, Rinehart and Winston, New York (1976) (Chapters 10, 11).

Photomultipliers

Boutot, J. P., Nussli, J., and Vallat, D., "Recent Trends in Photomultipliers for Nuclear Physics" in *Advances in Electronics and Electron Physics*, Vol. 60, P. W. Hawkes, Editor, Academic Press, New York (1983).

Engstrom, R. W., "Photomultiplier Handbook," RCA Corp., Lancaster, Penn. (1980).

Reisse, D., Creecy, R., and Poultny, S. K., "Single Photon Detection and Sub-Nanosecond Timing Resolution with the RCA C31034 Photomultiplier," Rev. Sci. Instru. *44*, pp. 1666-1668 (1973).

Shockley, W., and Pierce, J. R., "A Theory of Noise for Electron Multipliers," Proc. IRE, *26*, pp. 321-333 (1938).

Tan, H. H., "A Statistical Model of the Photomultiplier Gain Process with Applications to Optical ... se Detection," *TDA Progress Report 42-68*, pp. 55-67, Jet Propulsion Laborat... asadena, Calif. (1982).

Young, A. T., "Photomultipliers: Their Cause and Cure," Chapter 1 in *Methods of Experimental Physics*, Vol. 12, Part A, N. Carleton, Editor, Academic Pre... New York (1974).

Photodiodes and Avalanche Photodiodes — General

Capasso, F., Tsang, W. T., and Williams, G. F., "Staircase Solid-State Photomultipliers and Avalanche Photodiodes with Enhanced Ionization Rates Ratio," IEEE Trans. Electron. Dev. *ED-30*, pp. 381-390 (1983), and references therein.

Melchior, H., "Detectors for Lightwave Communication," in *Physics Today*, Nov. 1977, pp. 32-39.

Müller, J., "Photodiode for Optical Communications," in *Advances in Electronics and Electron Physics, 55*, Marton and Marton, Editors, Academic Press, New York (1981).

Stillman, G. E., and Wolfe, C. M., "Semiconductors and Semimetals," *12*, Academic Press, New York (1977).

Sze, S. M., *Physics of Semiconductor Devices*, 2nd Ed., Wiley Interscience, New York (1981) (Chapter 13).

Wang, S. Y., and Bloom, D. M., "100 GHz Bandwidth Planar GaAs Schottky Photodiode," Electron Lett., *19*, pp. 554-555 (1983).

Webb, P. P., McIntyre, R. J., and Conradi, J., "Properties of Avalanche Photodiodes," RCA Rev., *35*, pp. 235-278 (1974)

Avalanche Photodiodes Statistics

Ingerson, T. E., Kearney, R. J., and Coulter, R. L., "Photon Counting with Photodiodes," Appl. Opt. *22*, pp. 2013-2018 (1983), and references therein.

McIntyre, R. J., "Multiplication Noise in Uniform Avalanche Diodes," IEEE Trans. Electron. Dev., *ED-13*, pp. 164-168 (1966).

McIntyre, R. J., "The Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes: Theory," IEEE Trans. Electron Dev., *ED-19*, pp. 703-713 (1972).

Personick, S. D., "New Results on Avalanche Multiplication Statistics with Applications to Optical Detection," Bell Sys. Tech. J., *50*, pp. 167-189 (1971).

Personick, S. D., "Statistics of a General Class of Avalanched Detectors with Applications to Optical Communication," Bell Sys. Tech. J., *50*, pp. 3075-3095 (1971).

Table 1. Ionization coefficient parameters of several semiconductors

Material	Parameters of Eq. (18)					k =	
	A_n	A_p	B_n	B_p			
	cm^{-1}	cm^{-1}	V/cm	V/cm		Low	n
	$\times 10^6$	$\times 10^6$	$\times 10^6$	$\times 10^6$	u	field	ield
Si	3.8	2.25	1.75	3.26	1	0.02	0.1
Ge	15.5	1	1.5	1.28	1	10	2
GaAs	12	360	2.3	2.9	1	3	5
GaSb	3.2	300	0.36	0.55	1	0.5	>1
InP	2000	30	1.5	2.5	1	0.2	0.4
InAs						10	20
GaAlSb	0.104	0.19	0.41	0.41	2	2	2
$Ga_xIn_{1-x}As$							
$x = 0.14$	1000	130	3.6	2.7	1	2	
$x = 0.2$						50	
$x = 0.47$						0.2	
GaAsSb	0.15	0.11	0.64	0.72	1	0.4	0.7
InGaAsP						0.3	>1

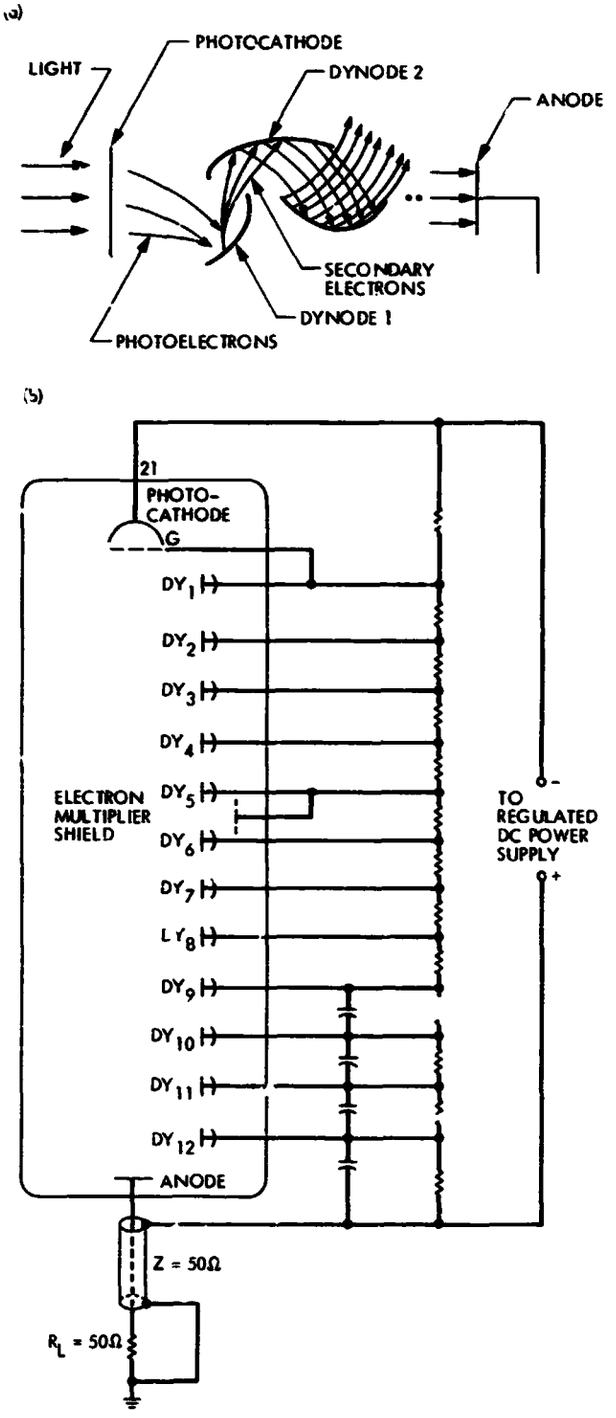

Fig. 1. Photomultiplier tube: (a) schematic structure; (b) including bias circuit

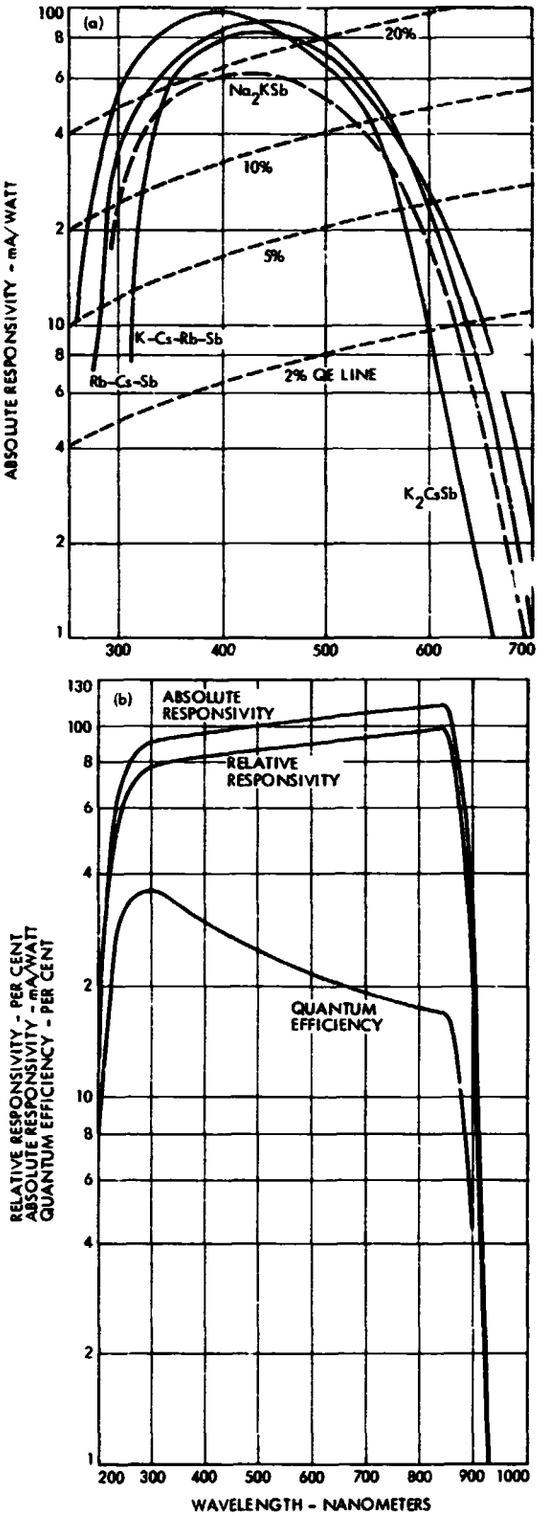

Fig. 2. Spectral response of PMT photocathodes: (a) multi-alkali photocathode; (b) GaAs (Cs) photocathode

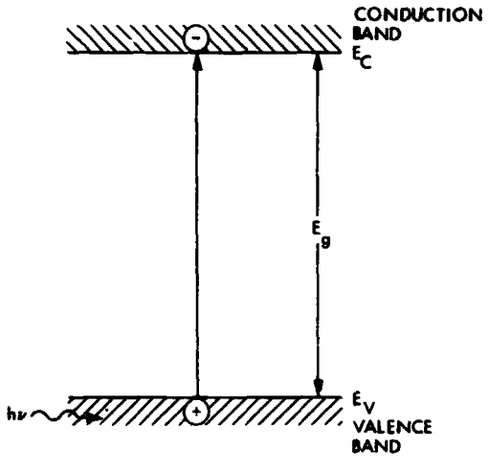

Fig. 3. Schematic band diagram of an intrinsic absorption process

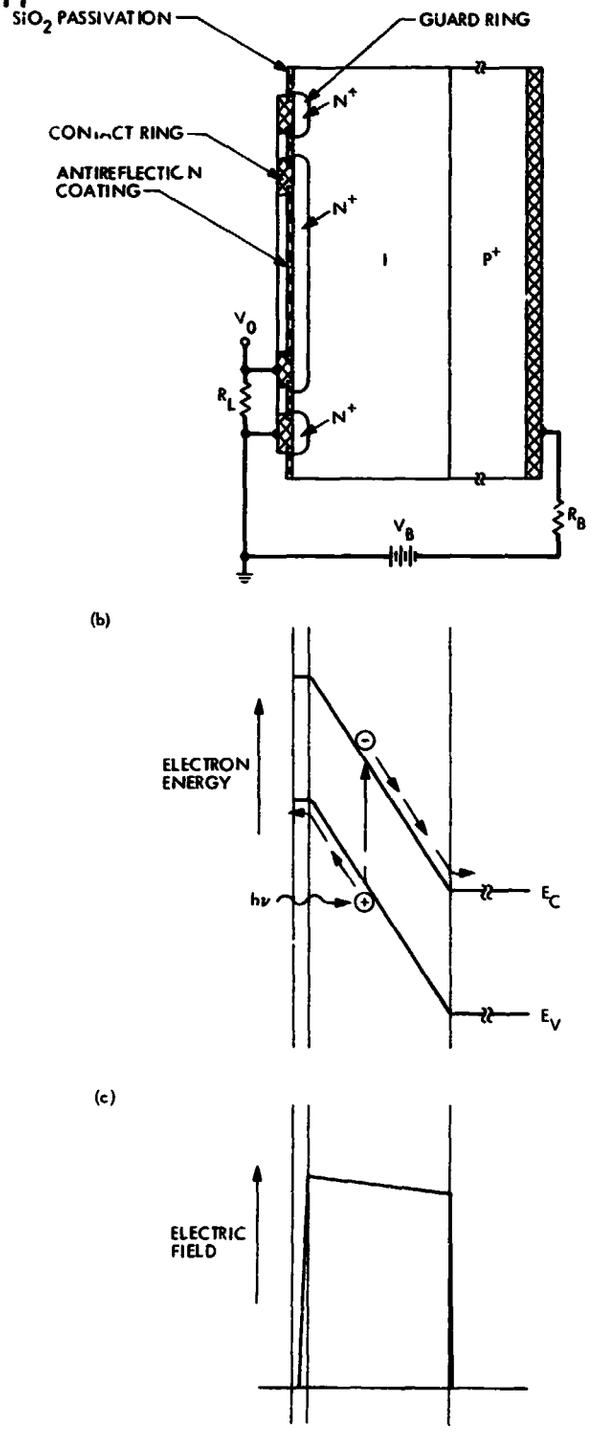

Fig. 4. PIN PD: (a) schematic structure of a Si device (the guard-ring shunts surface leakage currents from the load resistor); (b) band diagram; (c) electric field distribution (*I* region in this case is a lightly doped p-region, usually denoted π)

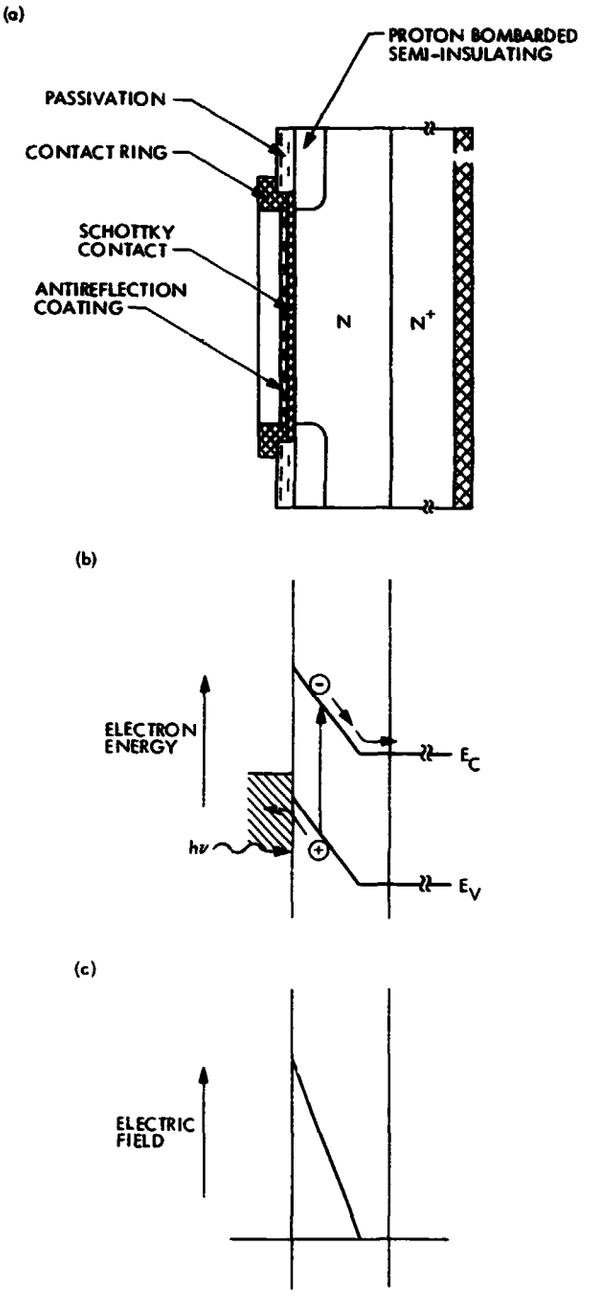

Fig. 5. Schottky PD: (a) schematic structure; (b) band diagram; (c) electric field distribution (details at nn^+ junction are neglected)

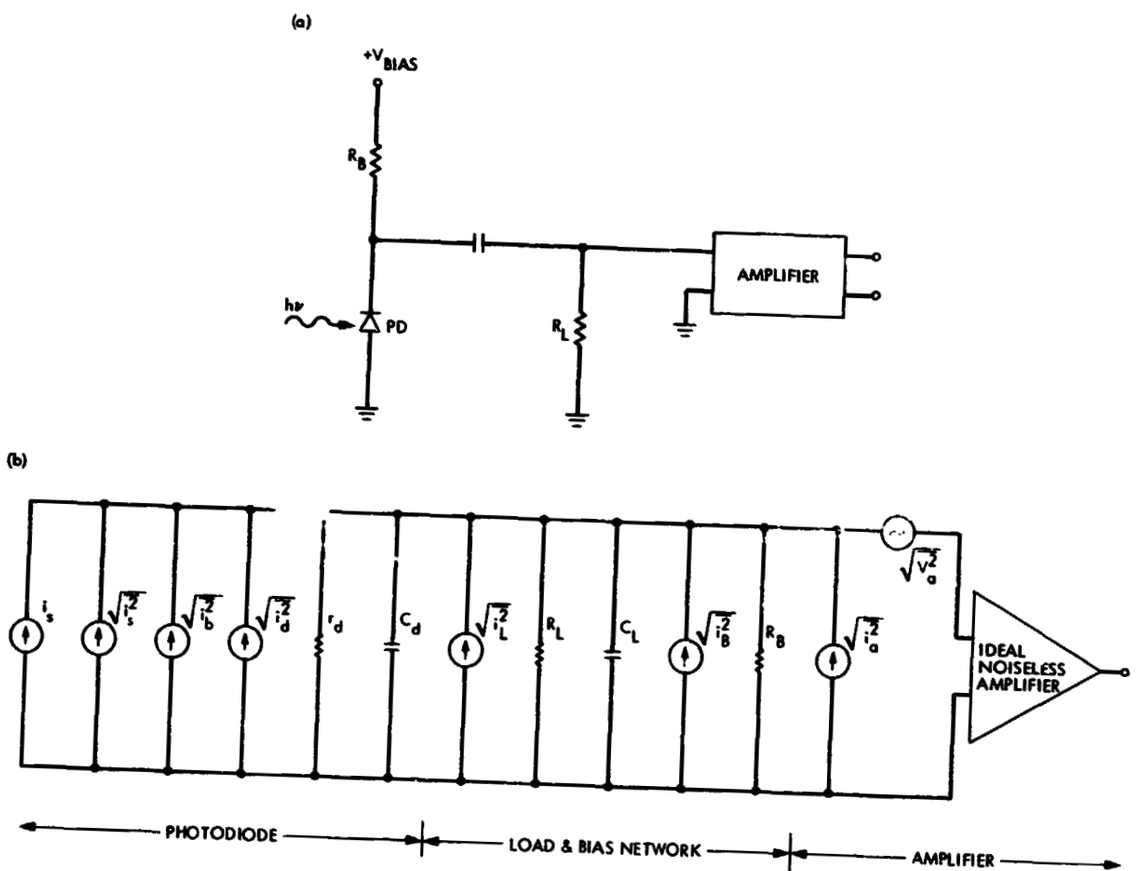

Fig. 6. Basic photodiode optical receiver: (a) detector circuit using a PD; (b) equivalent circuit (including noise sources) of the circuit in Fig. 6(a) (r_d is the PD dynamic resistance; the other parameters are explained in the text)

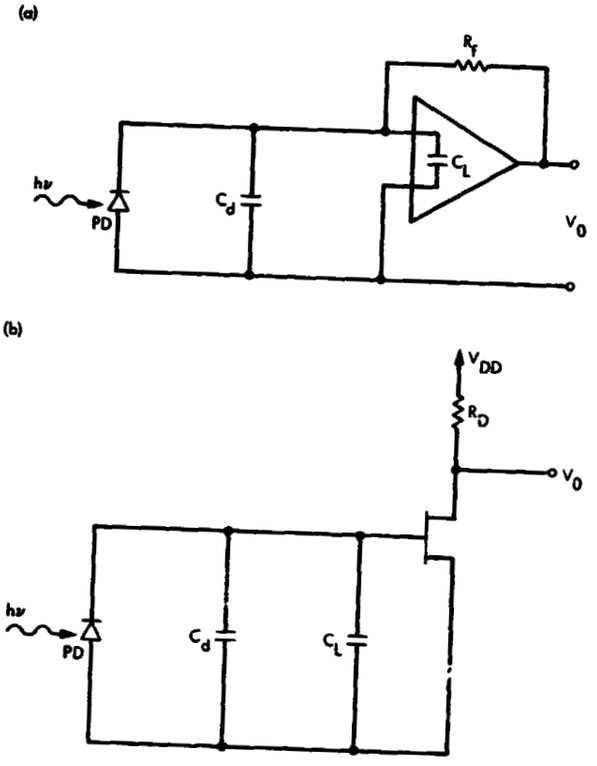

Fig. 7. Generic types of front-end detector circuits: (a) trans-impedance amplifier; (b) high impedance FET front end amplifier

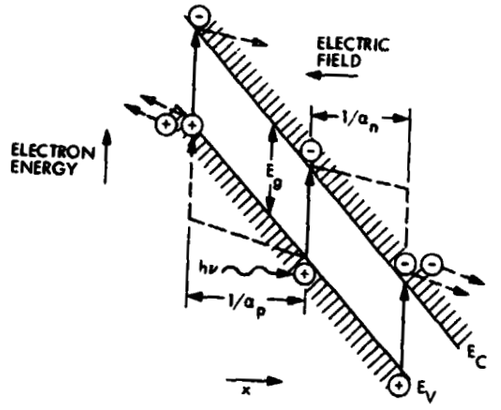

Fig. 8. Schematic depiction of the avalanche gain (impact ionization) process

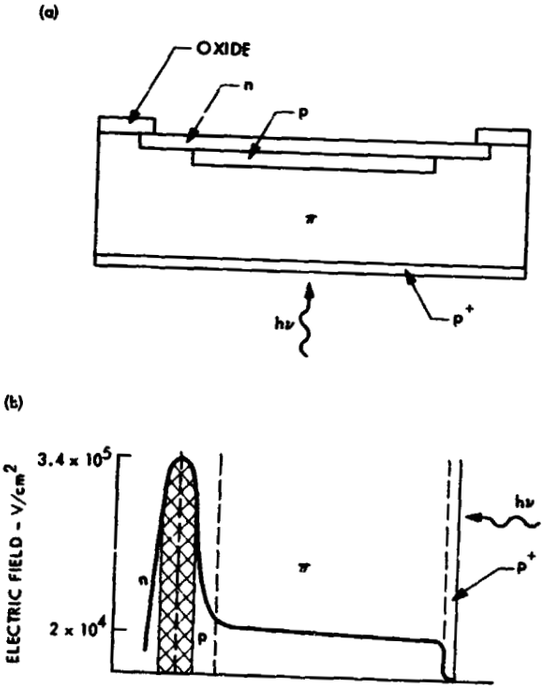

Fig. 9. Avalanche photodiode: (a) structure of a reach-through APD; (b) electric field distribution in the APD (the cross-hatched area is the multiplication region)

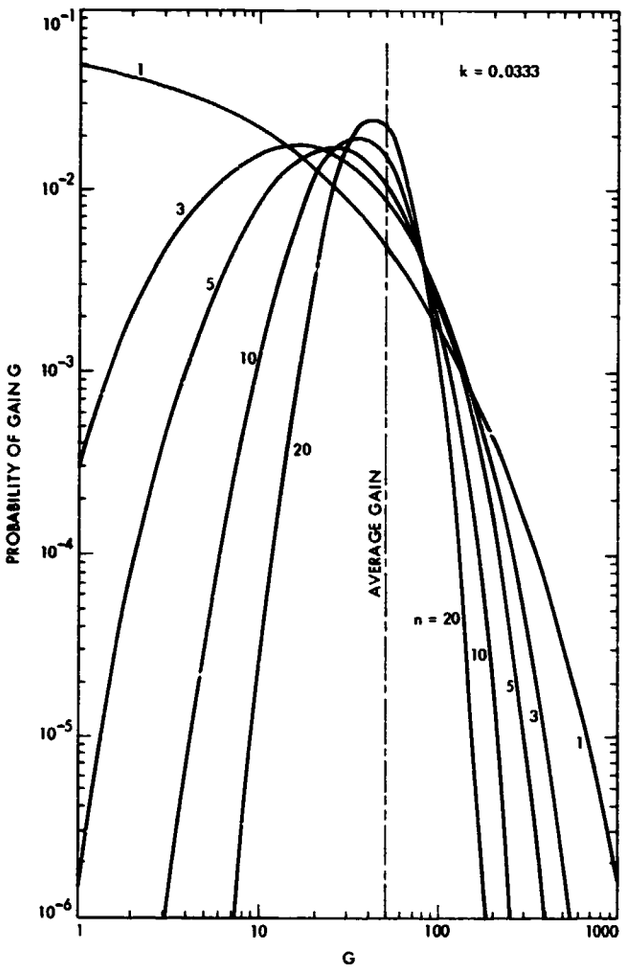

Fig. 10. Probability distribution of gain process of an APD with an average gain of 50. *n* is the number of carriers injected initially to the multiplication region

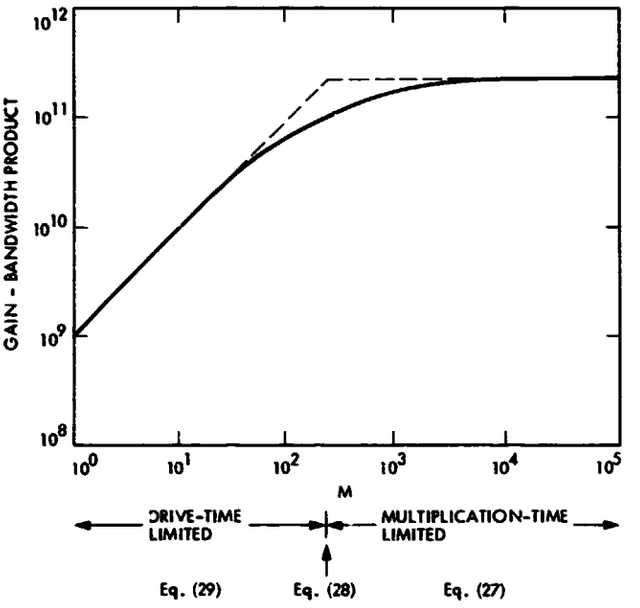

Fig. 11. Gain bandwidth product of an APD

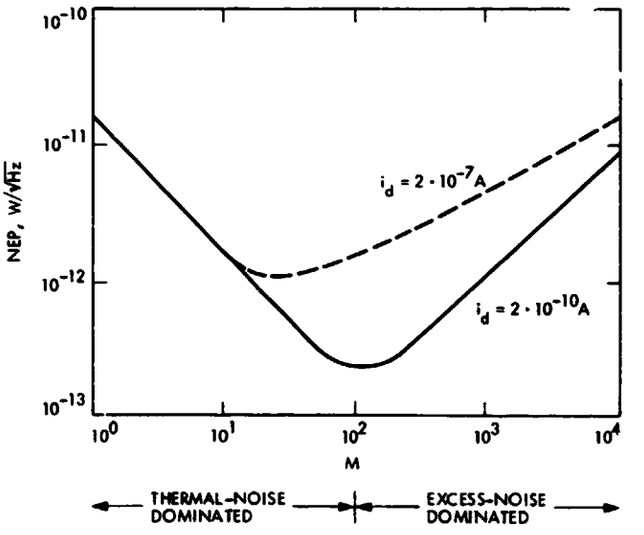

Fig. 12. *NEP* of APD versus gain *M* (k = 0.1, λ = 0.85 μm, *B* = MHz, R_e/T_e = 10.7 Ω/K, i_d/I_{ds} = 2;