

Avalanche Photodiode, SPAD, and SiPM

Multiplication by impact ionization is a very welcome mechanism in photodetectors, because it supplies an *internal gain* to circumvent to the bandwidth-noise tradeoff we have seen in Chapter 3. The *avalanche photodiode* or APD has been the first device to take advantage of the gain of impact ionization. Following the development and the technical maturity of APD, the SPAD (Single-Photon Avalanche Detector) has been invented, the first semiconductor device able to detect single photons but with a kind of Geiger-mode limitation. From it, recently the SiPM (Silicon Photomultiplier) emerged as a truly solid-state counterpart of the vacuum-tube technology PMT.

6.1 AVALANCHE PHOTODIODE

Avalanche multiplication by impact ionization is the gain mechanism used for amplification of the photogenerated charge internal to a photodiode junction. The avalanche is obtained by applying a high reverse bias to the photodiode, so as to bring it into the breakdown region usually avoided in normal operation. The preferable structure to start with is that of a pin photodiode (Fig.5-11) with a uniform electric field across the depletion layer. The high

electric field accelerates the photogenerated charges, raising them to a kinetic energy large enough to ionize the crystal lattice with the subsequent production of new electron-hole pairs. The new pairs are accelerated and multiplied as well, and so on, up to collection of charges at the anode.

If M is the total number of pairs produced by an initial photoelectron, the output signal current is $I = MI_{ph} = M\sigma P$, i.e., M is the internal gain of the avalanche photodiode.

Besides avalanche, a breakdown in reverse bias can also intervene by an alternative mechanism, namely the punch-through of the junction, when the depletion layer reaches one of the electrodes (as, for example, for $w=0$ in Fig.5-2). In this case, charge carriers are directly injected in the junction from the contact, and as there is no multiplication in this case, the punch-through must be avoided in avalanche photodiodes by keeping a safety margin in the thickness of p and n undepleted regions at the working voltage.

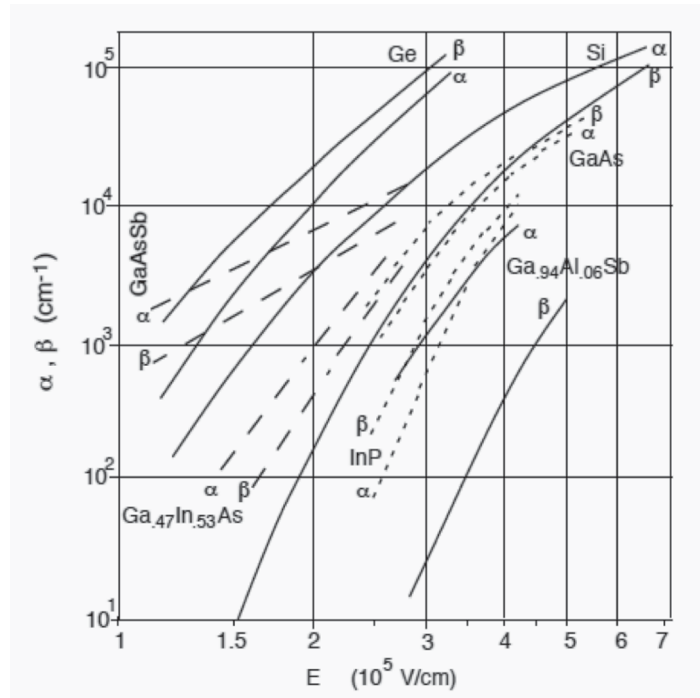


Figure 6-1 Ionization coefficients of electrons (α) and holes (β) as a function of the electric field E for some semiconductors used in avalanche photodiodes.

In a material, the charge-carrier multiplication process is described by the *ionization coefficients* α and β of electron and hole, defined as the number of new pairs generated for unit length by the carrier under a given electric field E . The inverse of α and β are the ionization lengths and represent the mean free path between successive multiplications.

Fig.6-1 shows the curves of α and β as a function of the electric field E for several semiconductors of importance in APDs. As we can see, starting from about a 1 or 2×10^5 V/cm, the ionization coefficients steeply increase with only a small increase of the electric field, while for $E < 10^5$ V/cm ionization is negligible in all of the known semiconductor materials. Also, in some materials electrons ionize more efficiently than holes (the case with Si, GaAsSb, GaInAs, for which $\alpha > \beta$), while in a few others the reverse is true (Ge, GaAs with $\beta > \alpha$). This difference is of little consequence, however, to the mechanism of operation of the avalanche photodiode.

With α and β not much different, both charge carriers are efficient to ionize and both will contribute to the gain. But, because they move in opposite directions, a positive feedback loop arises in the multiplication. For instance, consider a primary electron (Fig.6-2) in x that moves towards the anode and produces pairs from x to L . The new holes will be drawn toward the cathode and produce new electrons in positions behind the precursor in x , starting again the process of multiplication: this is a positive feedback.

From general feedback theory, it is well known that positive feedback degrades both the response bandwidth and the noise by a factor equal to the loop gain plus one.

Thus, to avoid this drawback, in an APD we will keep the loop gain appropriately small. This establishes a maximum for αL (or βL) and hence a limit to the maximum useful gain, as will be shown in detail later.

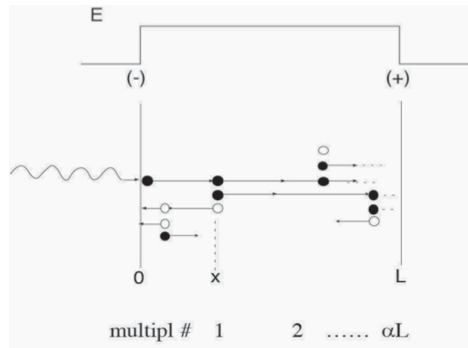


Figure 6-2 Avalanche multiplications starting from a primary electron injected at $x=0$. The first and second ionizations are indicated for illustration, to show how the back-going hole introduces a positive feedback mechanism.

6.1.1 Gain of the APD

To evaluate the dc gain $M = I/I_{ph}$ of the avalanche photodiode, we assume a multiplication structure of length L with a uniform field, as shown in Fig.6-2, so that $\alpha, \beta = \text{const}$. The primary electrons of the photogenerated current I_{ph} are injected at $x=0$, and the amplified cur-

rent I is collected at $x=L$. For simplicity of notation, let us take $\alpha > \beta$ in the following; in the opposite case, all equations still hold with α and β interchanged.

Let i_n and i_p be the electron and hole currents at point x (Fig.6-2). Their increments along an elemental path dx are proportional to multiplication rates α and β , and to the currents i_n and i_p . So, we can write the current balance as:

$$di_n/dx = \alpha i_n + \beta i_p, \quad di_p/dx = -(\alpha i_n + \beta i_p) \quad (6.1)$$

where the minus sign in the second equation occurs because the holes grow in the $-x$ direction. The boundary conditions for the set of Eq.(6.1), i.e., injection at $x=0$ and collection at $x=L$, are written as:

$$\begin{aligned} i_n(0) &= I_{ph}, \\ i_n(L) &= I, \quad \text{and} \quad i_p(L) = 0 \end{aligned} \quad (6.1A)$$

By differentiating the first Eq.(6.1) and inserting in the second, we obtain $d^2 i_n/dx^2 = (\alpha - \beta) di_n/dx$. By integration, this equation is solved as: $di_n/dx = C_1 \exp(\alpha - \beta)x$, whence $i_n = [C_1/(\alpha - \beta)] \exp(\alpha - \beta)x + C_2$; in addition, using the first of Eq.(6.1), we have: $\beta i_p = di_n/dx - \alpha i_n = -[\beta C_1/(\alpha - \beta)] \exp(\alpha - \beta)x - \alpha C_2$.

From the boundary conditions of Eq.(6.1A), we then get: $[C_1/(\alpha - \beta)] + C_2 = I_{ph}$, and $-[C_1/(\alpha - \beta)] \exp(\alpha - \beta)L - (\alpha/\beta)C_2 = 0$, which allows us to solve for C_1 and C_2 . Inserting the results in $M = i_n(L)/i_n(0)$ we obtain:

$$M = \frac{(\alpha - \beta) \exp(\alpha - \beta)L}{\alpha - \beta \exp(\alpha - \beta)L} \quad (6.2)$$

As special cases of Eq.(6.2), we note that for $\beta=0$, i.e., when only one carrier is capable of ionizing, the gain is:

$$M = \exp \alpha L \quad (6.2A)$$

or, it coincides with that of Townsend multiplication valid in a gas. For $\alpha=\beta$, i.e., when both carriers are equally capable of ionizing, we have from Eq.(6.2), after some algebra to solve for the indetermination form:

$$M = 1/(1 - \alpha L) \quad (6.2B)$$

and then it suffices that, in the multiplication length L , just one ionization length $1/\alpha$ is contained for the gain to become infinite.

For $\alpha > \beta$ we can put Eq.(6.2) in the expressive form:

$$M = A/(1 - hA) = \exp \alpha L / (1 - \beta/\alpha \exp \alpha L),$$

whence β/α assumes the meaning of feedback factor.

In general, for $\alpha \neq \beta$, the gain becomes infinite for $\exp(\alpha - \beta)L = \alpha/\beta$, that is, for sufficiently large feedback. This condition between the number of multiplication lengths αL and the ionization ratio α/β can be written as:

$$\alpha L = (\ln \alpha / \beta) / (1 - \beta / \alpha) \approx \ln \alpha / \beta \quad (\text{for } M = \infty) \quad (6.3)$$

At the opposite, we have only a (moderate) gain increase, of a factor of 2, due to positive feedback in respect to $\exp \alpha L$, when we set the gain equal to the ratio of the ionization coefficients, $M=\alpha/\beta$. This corresponds to the condition $\exp(\alpha-\beta)L=(\alpha/\beta)/(2-\beta/\alpha)\approx\alpha/2\beta$, or:

$$\alpha L \approx \ln \alpha/2\beta = \ln \alpha/\beta - 0.69 \quad (\text{for } M=M_0=\alpha/\beta) \quad (6.4)$$

This circumstance suggests that we can assume $M_0=\alpha/\beta$ as the *optimal gain* of the APD.

In Fig.6-3 we plot the curves of M , expressed by Eq.(6.2), as a function of the number of multiplication lengths αL , and with the ionization ratio α/β as a parameter.

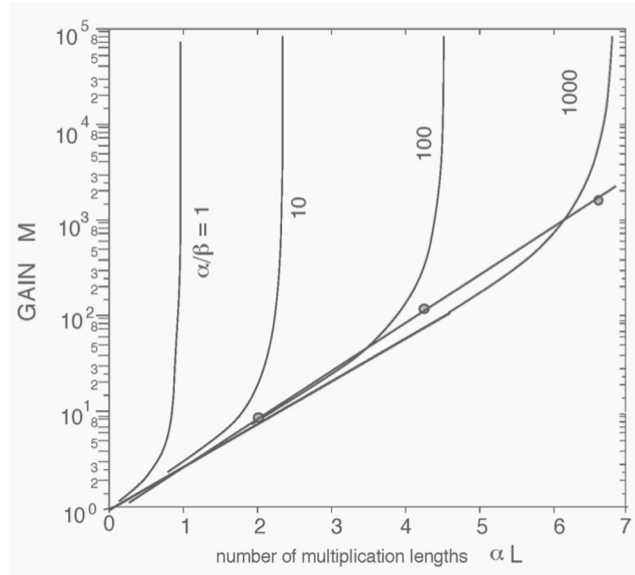


Figure 6-3 The dc gain M of the APD (avalanche photodiode) for electron injection at $x=0$, as a function of the number of multiplications αL and with the ionization ratio α/β as a parameter. Dots indicate the condition of optimal gain $M_0=\alpha/\beta$.

Further results on the gain analysis. Eq.(6.2) is for the injection of electrons at $x=0$, and it also holds for the injection of an electron-hole pair since the hole in $x=0$ is immediately collected and does not contribute to $i_n(L)$.

In general, it may be interesting to know the gain $M_n(x)$ for an electron injection at point x . Going back to Eq.(6.1) and with the new boundary conditions, we find:

$$M_n(x) = [\alpha \exp(\alpha-\beta)(L-x) - \beta \exp(\alpha-\beta)L] / [\alpha - \beta \exp(\alpha-\beta)L] \quad (6.5)$$

Analogously, the gain $M_p(x)$ for injection of holes in point x , always referred to the output $i_n(L)$ at $x=L$, is found by the same analysis as:

$$M_p(x) = [\beta \exp(\alpha-\beta)L - \alpha \exp(\alpha-\beta)(L-x)] / [\alpha - \beta \exp(\alpha-\beta)L] \quad (6.5A)$$

Note that the output at $x=0$ (hole current) would have a gain $M_n(x)-1$ for electron injection, and $M_p(x)+1$ for hole injection (duly taking account of the primary carrier).

Finally, for a photon that is absorbed at point x and generates an electron-hole pair, the gain $M_c(x) = M_n + M_p$ either at the electron output (at $x=L$) or at the hole output (at $x=0$) is found as:

$$M_c(x) = [(\alpha - \beta) \exp(\alpha - \beta)(L - x)] / [\alpha - \beta \exp(\alpha - \beta)L] \quad (6.5B)$$

$$= M_c(0) \exp(-(\alpha - \beta)x) \quad (6.5B')$$

Since Eqs.(6.5) have the same denominator, their trend with respect to the ionization ratio α/β is the same and Eqs.(6.3) and (6.4) still hold for these cases.

Using Eq.(6.5B), we can now easily calculate the APD gain in the case of an incoming radiation dissipated in the multiplication region with given distribution $I(x)$:

$$M_I = \int_{0-L} I(x) M_c(x) dx \quad (6.6)$$

For example, if the absorption is the negative exponential

$$I(x) = a \exp(-ax),$$

the case representative for a photodiode with a single region for radiation dissipation and carrier multiplication, we have from Eq.(6.6):

$$M_{I(a)} = M_c(0) [a/(a + \alpha - \beta)] [1 - \exp(-(a + \alpha - \beta)L)] \quad (6.7)$$

and if the absorption is uniform, $I(x) = 1/L$ for $0 < x < L$, the result becomes:

$$M_{I(u)} = M_c(0) [1/(\alpha - \beta)L] [1 - \exp(-(\alpha - \beta)L)] \quad (6.7')$$

Both expressions show that gain for distributed injection decreases in respect to point-wise absorption.

6.1.2 Frequency Response and Noise of the APD

Frequency response

The *extrinsic* cutoff of the APD is obviously the same as that of a normal pin photodiode of equal structure, and the results found in Section 5.2.4 are applicable as well as the circuit of Fig.5-9b.

The *intrinsic* cutoff is associated with the collection time of the charge generated in the injection by the photogenerated carriers. This time will again be the inter-electrode time of flight τ_d [see Eqs.(5.26) and (5.27)] if just one carrier is able to ionize (i.e., for $\alpha \neq 0$ and $\beta = 0$, or, more precisely, for $M < M_0 = \alpha/\beta$). Instead, if the gain derives from an appreciable positive feedback ($M > M_0 = \alpha/\beta$), there will be a back-and-forth generation of charges in the multiplication region, and the collection time will be much larger than τ_d .

To analyze the response of the APD, several approaches have been introduced: some of them are physical methods based on continuity and transport equations in the semiconductor [1], others are statistical methods based on the description of the multiplication process [2].

In both cases, the results are the same and the analytic developments required are rather involved. In the following, we limit ourselves to a discussion of the main results.

Defining the mean drift time τ_d as in Eq.(5.27):

$$\tau_d = (1/2)L(1/v_n + 1/v_p) = L/2v^*$$

where $1/v^* = 1/v_n + 1/v_p$ is the effective drift velocity of the carriers, the *intrinsic* cutoff frequency (-3dB point) f_{2d} of the APD is found, to a good approximation, as:

$$f_{2d} = 0.36 / \tau_d (1 + M/M_0) \quad (6.8)$$

As special cases of Eq.(6.8), we have that for moderate gains, not larger than the optimum value $M_0 = \alpha/\beta$, *bandwidth* is a constant and is determined by the transit time:

$$f_{2d} \cong 0.36 / \tau_d \quad (M < M_0) \quad (6.8')$$

while in the opposite case of large gain, bandwidth is inversely proportional to gain and therefore the *gain-bandwidth product* is a constant:

$$f_{2d} M \cong 0.36(\alpha/\beta) / \tau_d \quad (M > M_0) \quad (6.8'')$$

Eqs.(6.8) to (6.8'') confirms the significance of *optimum gain* attributed to M_0 , also as regard to the frequency response.

Derived for the injection electrons (or pairs) at $x=0$, Eqs.(6.8)-(6.8'') are also valid (with minor changes) for the injection of electrons or pairs at $0 < x < L$.

The frequency cutoff f_{2d} is the description, by a single parameter, of the APD frequency performance. Yet, more detailed information is provided by the full frequency response $M(\omega)$ of the gain, which can be calculated by statistical analysis [2].

The results are shown in Fig.6-4, where $M(\omega)$ is plotted as a function of the normalized frequency $\omega L/v^* = f/(v^*/2\pi L)$ and with α/β as a parameter. At high gains ($M > \alpha/\beta$), we can see the merging of all curves at high frequency and their fairly regular behavior, typical of the single-pole response; as β/α decreases the response moves to higher and higher frequencies, and the deviations from the single pole become significant.

The frequency cutoff f_{2d} is easily read from the diagrams of Fig.6-4 as the 3 dB-loss point with respect to the dc value $M_c(0)$. In Fig.6-5 we plot f_{2d} versus α/β and with values of the gain $M_c(0)$ as a parameter. The asymptotic cases Eq.(6.8), commonly used in practice as an approximation of f_{2d} , are depicted by straight lines in Fig.6-5. As can be seen, they provide a rough estimate which is generally adequate as a rule-of-thumb evaluation of APD performances.

Another interesting description of the APD dynamic behavior is obtained by the time-domain impulse (or Dirac-delta) response. The impulse response $i(t)$ is the waveform that could be observed following a very short light-pulse excitation., and it can be computed as the Fourier transform of the frequency response $M(\omega)$. Doing so, we obtain the results shown in Fig.6-6, with α/β as parameter and for several gains. From the diagrams, we can see the exponential leading edge (it starts at $t=0$ from the current $2v^*/qL$ induced by the electron injected at $x=0$) representing the avalanche multiplication increase in the number of carriers, followed at time $t=L/2v^*$ by a trailing edge representing the charge collection, which becomes slower and slower as β/α increases.

We can also note that, when only electrons ionize ($\beta=0$), the current ceases at $tv^*/L=1$: the collection time of a normal pin photodiode.

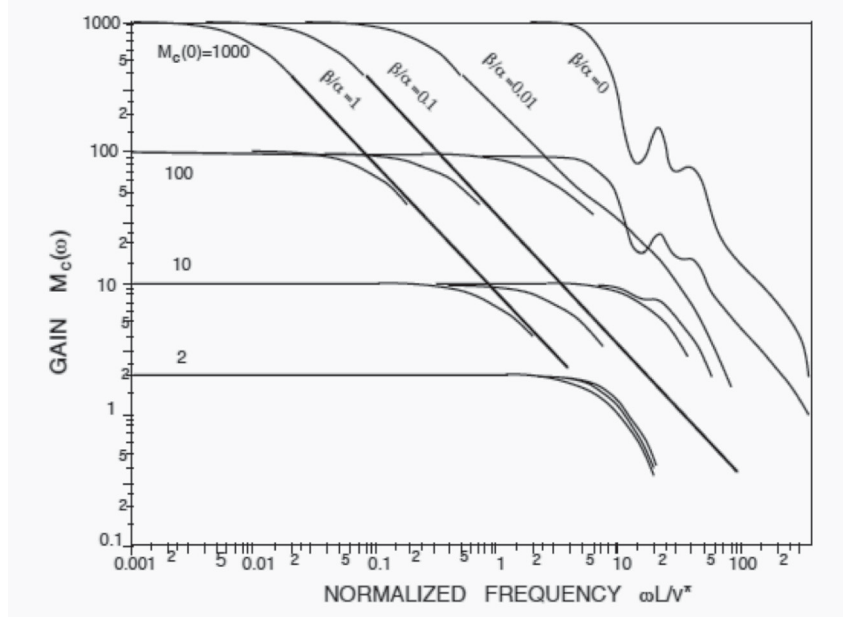


Figure 6-4 Frequency response of the APD gain $M_c(\omega)$ as a function of the normalized frequency $\omega L/v^*$ and with the ionization ratio β/α as a parameter (case for electron injection at $x=0$ and equal carrier velocities, $v_e=v_h$).

Noise

The multiplication process of the APD has a statistical nature, because the free path λ traveled by a charge carrier between successive multiplications is a random variable, to be described by a probability density, $p(\lambda)=\alpha \exp(-\alpha\lambda)$ for electrons and $\beta \exp(-\beta\lambda)$ for holes. As a consequence of the randomness of λ , the number of multiplications contained in L is random too, and the result is a fluctuation of the total charge generated by multiplication and collected at the output. Thus, the gain and the amplitude waveform of the impulse response also fluctuate from event to event, and have a mean value and a variance.

The quantities considered in previous sections, i.e., gain M , impulse response $i(t)$, etc., describe the APD on the basis of macroscopic relations. They now assume the meaning of average (or mean) values of their random counterparts. By a statistical approach [2], we can evaluate the random properties of the process without following the very intricate story of an initial carrier generating a series of branching multiplications up to the collection of all the charges at the output, and calculate the variance $\sigma_M^2(x)$ of the gain (or, the variance of the number of generated charges) for injection of a pair at x , as:

$$\sigma_M^2(x)=M_c^2(x)\{1+[\beta M_c^2(0)-\alpha M_c^2(L)]/(\alpha-\beta)M_c(x)\} \quad (6.9)$$

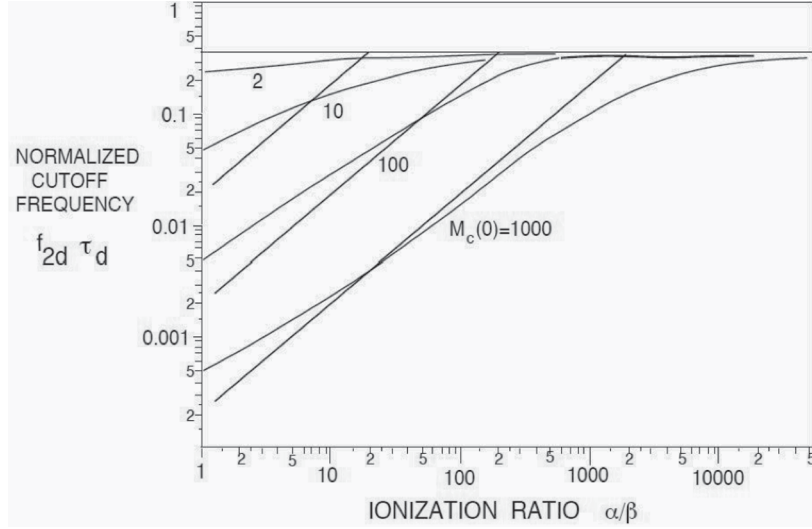


Figure 6-5 Frequency cutoff f_{2d} of the APD, normalized to the drift time $\tau_d = L/2v^*$, as a function of the ionization ratio α/β and with the gain M_c as a parameter (for $x=0$ and $v_e=v_h$).

where $M_c(x)$ is given by Eq.(6.5B). The relative variance $\sigma_M^2(x)/M_c^2(x)$ is a minimum when $M_c(x)$ is a maximum (that is, for injection at $x=0$) and the corresponding value is:

$$\sigma_M^2(0)/M_c^2(0) = 1 + \beta M_c(0)/(\alpha - \beta) - \alpha M_c^2(L)/M_c(0)(\alpha - \beta) \quad (6.10)$$

$$= 1 + M_c(0) \left\{ \beta/(\alpha - \beta) - [\alpha/(\alpha - \beta)] \exp(-2(\alpha - \beta)L) \right\} \quad (6.10')$$

For moderate gains M_c , the following asymptotic cases as a function of α/β are found:

$$\begin{aligned} \sigma_M^2(0)/M_c^2(0) &\approx 1 && \text{for } M_c(0) < \alpha/\beta \\ &\approx M_c(0)\beta/\alpha && \text{for } M_c(0) > \alpha/\beta \end{aligned} \quad (6.11)$$

This expression once again recognizes that $M_o = \alpha/\beta$ is the optimal gain.

With regard to the frequency spectrum of noise, the statistical analysis indicates that it is practically coincident with the transfer function (Fig.6-4), and that Eq. (6.10) for the cutoff and the diagram of Fig.6-5 still hold with a very good approximation.

From the coincidence of the spectra, it follows that in an avalanche photodiode the noise fluctuations superimposed on the mean waveform are of the *rigid* type, i.e., can be attributed to amplitude (or scale factor) variations rather than to shape variations of the waveform.

In the time domain, the output signal (with mean value MI_{ph}) has a current variance σ_I^2 given by the sum of two different contributions: (i) the shot noise already present in the signal I_{ph} and amplified by M to give $2eBI_{ph}M^2$ and, (ii) the gain fluctuation term given by $2eBI_{ph}\sigma_M^2$.

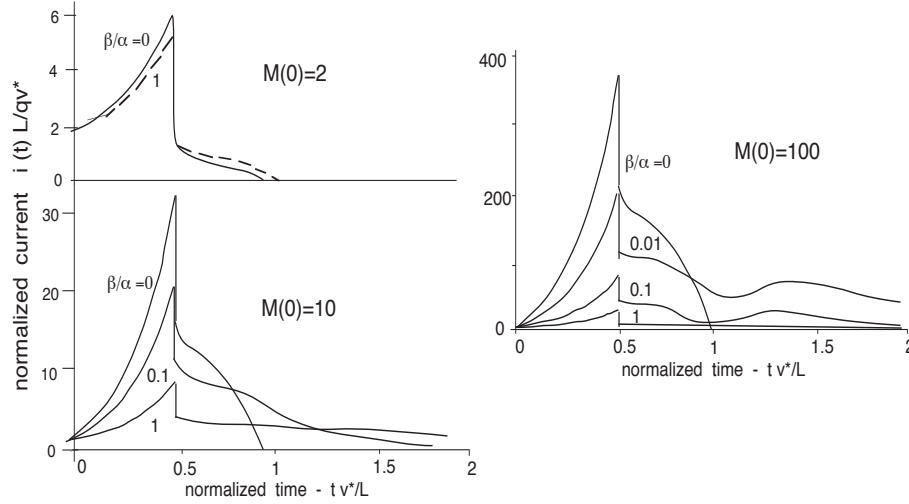


Figure 6-6 Impulse response of the APD (for injection at $x=0$). The ordinate scale gives the output current $i_U(t)$ in units of v^*/qL ; the abscissa scale is the time t in units L/v^* .

Summing up the two contributions (which are statistically independent) we have for the variance:

$$\sigma_I^2 = 2eB I_{ph} (M^2 + \sigma_M^2) = 2eB I_{ph} M^2 F \quad (6.12)$$

where F , the *excess noise factor* of the APD has been introduced. Comparing with Eq.(3.3'), we can see that F is coincident with the noise figure defined in Section 3:

$$F = 1 + \sigma_M^2/M^2 \quad (6.13)$$

Using Eqs.(6.10) and (6.11) in Eq.(6.13) we can write an explicit expression for F frequently quoted in literature and equivalent to Eq.(6.10') as:

$$F = (\beta/\alpha)M + (2-1/M)(1-\beta/\alpha) \quad (6.10'')$$

The trend of F versus β/α for several gains M is plotted in Fig.6-7 for pairs injected at $x=0$.

This result, as well as Eqs.(6.10) through (6.12), applies also to electrons injected at $x=0$ as the primary hole is collected immediately and does not contribute to variance.

We can now consider the total output noise of the basic circuit (Fig.5-24a) with an APD biased through a load resistance R . Of course, besides internal noise [Eq.(6.12)], we shall take into account the load resistance R and the dark current I_0 noise contributions.

Load resistance gives the usual $4kTB/R$ Johnson term. For the dark current $I_0 = I_{g-r} + I_d$ (see Section 5.2.2), one should distinguish the different multiplication gains of I_{g-r} and I_d .

The generation-recombination current I_{g-r} is thermally generated in the junction and thus is multiplied by $M_{I(u)}$ of Eq.(6.7'); the diffusion current I_d comes from the neutral regions and is multiplied by $M_n(0)$ or $M_p(L)$ according to which injection (electron or hole) is dominant. In addition, there might be a surface current I_s contribution to the total I_o , external to the junction and not multiplied.

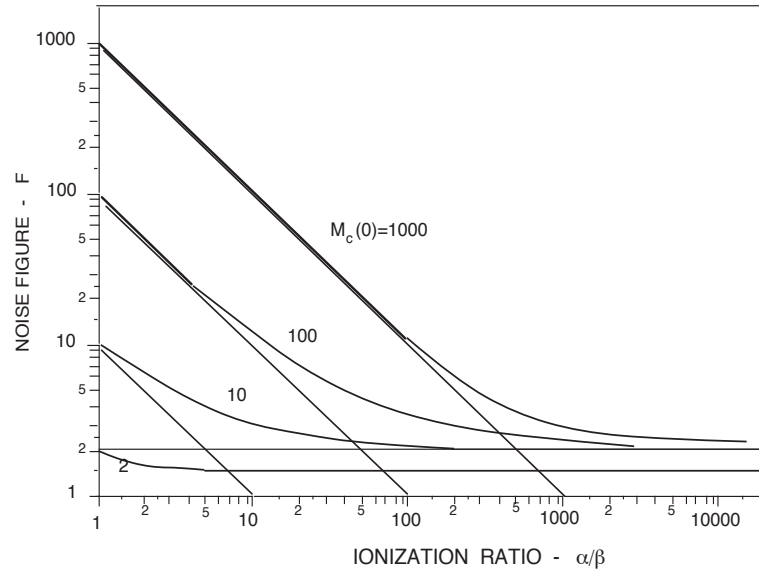


Figure 6-7 Excess noise factor F of the APD as a function of α/β for some values of gain (injection at $x=0$). For gain M_c less than the optimal M_0 , $F \approx 2$ whereas when gain is larger than $M_0 = \alpha/\beta$, F is the excess gain respect to the optimal gain, or M_c/M_0 .

However, with the conservative assumption that the dark current is multiplied with the same gain M of the signal, we can write the total output current noise of the APD as:

$$\sigma_I^2 = 2eB (I_{ph} + I_o) M^2 F + 4kTB/R \quad (6.14)$$

Dividing Eq.(5.88) by M^2 , we get the noise equivalent input at the level of current I_{ph} as:

$$\sigma_{in}^2 = 2eB (I_{ph} + I_o) F + 4kTB/RM^2 \quad (6.14')$$

This result points out that, in the avalanche photodiode, the weight of Johnson noise due to the load resistance is reduced by a factor M^2 . In other words, it appears as though we were using a resistance RM^2 , which makes the choice of low R values at the detector termination (a necessity for high-frequency response) much less critical, already at moderate optimum values of gain $M_0 = \alpha/\beta$ (e.g., 10–50).

It may sometimes be advantageous using a gain larger than the optimum one $M_0 = \alpha/\beta$, although the excess noise factor is not a minimum in this case [it is given by $F = M\beta/\alpha$, see Eq.(6.11)]. Indeed, as the output signal is $I_u = MI_{ph}$, we have for the signal/noise ratio:

$$(S/N)^2 = I_{ph}^2 / [2eB(I_{ph} + I_0) M\beta/\alpha + 4kTB/RM^2] \quad (6.15)$$

Because of the opposite trends of the terms in Eq.(6.15) containing M and $1/M^2$, the S/N is a maximum for a certain M , which is found as:

$$M = [(\alpha/\beta)(4kT/e) / R(I_{ph} + I_0)]^{1/3},$$

For instance, we have $M=125$ for $R=50 \Omega$, $I_{ph}+I_0=1nA$ and $\alpha/\beta=1$. The practical limit for M is then determined by the desired APD bandwidth, because at high M bandwidth decreases by a factor $M\beta/\alpha$ when $M > \alpha/\beta$ [Eq.(6.8")].

APD structures with non-uniform multiplication. Let us briefly report the extension of the above results to the case, very often found in practical structures, of a non-uniform multiplication. When the electric field is not a constant in the multiplying region, we must take the ionization coefficients $\alpha(x)$ and $\beta(x)$ as a function of position x . By repeating the calculations leading to Eq.(6.2) with the aid of the integration factor $\exp \int (\alpha-\beta)dx$, we obtain the gain for pairs injected at $x=0$ as:

$$M_c(x) = [\exp \int_{x-L}^x (\alpha-\beta) dx'] / [1 - \int_{0-L}^0 dx' \beta \exp \int_{x'-L}^{x'} (\alpha-\beta) dx'] \quad (6.16)$$

an expression that equivalently can be written:

$$M_c(x) = [\exp -\int_{0-x}^x (\alpha-\beta) dx'] / [1 - \int_{0-L}^0 dx' \alpha \exp -\int_{0-x'}^0 (\alpha-\beta) dx'] \quad (6.16')$$

The condition of infinity gain [Eq.(6.3)] becomes:

$$\int_{0-L}^x dx' \beta \exp \int_{x'-L}^{x'} (\alpha-\beta) dx'' = \int_{0-L}^0 dx' \alpha \exp -\int_{0-x'}^0 (\alpha-\beta) dx'' = 1 \quad (6.3')$$

In terms of gain for pairs injected at $x=0$, Eq.(6.16') can also be expressed as:

$$M_c(x) = M_c(0) \exp -\int_0^x (\alpha-\beta) dx' \quad (6.17)$$

to show the similarity of the trend with Eq.(6.5B). From a numerical comparison with Eqs.(6.2) through (6.5), valid for uniform field and $\alpha, \beta = \text{const}$, one can find that the diagrams of gain and frequency cutoff still hold with a good approximation using, in place of the ionization ratio α/β , the averaged value: $\langle \alpha/\beta \rangle = \int_{0-L}^0 \alpha dx / \int_{0-L}^0 \beta dx$.

Analogously, gain variance σ_M^2 and excess noise are still given by Eqs.(6.10) and (6.13) using the weighted value for α/β :

$$\langle \alpha/\beta \rangle = (K_2 - 1)/(K_1^2 - K_2) \quad (6.18)$$

where:

$$K_1 = \int_{0-L}^0 \alpha [\exp -\int_{0-x}^x (\alpha-\beta) dx'] dx / \int_{0-L}^0 \beta [\exp -\int_{0-x}^x (\alpha-\beta) dx'] dx \quad (6.18')$$

$$K_2 = \int_{0-L}^0 \alpha [\exp -2\int_{0-x}^x (\alpha-\beta) dx'] dx / \int_{0-L}^0 \beta [\exp -2\int_{0-x}^x (\alpha-\beta) dx'] dx \quad (6.18'')$$

Let us quote a final result, useful for the numerical simulation of cascaded multiplication regions, each having a constant internal field. In a structure with two multiplication regions 1 and 2, with gains $M_{1n}(0)$ and $M_{1p}(L)$ for the injection of electrons at $x=0$ and of holes at $x=L$ in region 1, and gains $M_{2n}(0)$ and $M_{2p}(L)$ in region 2, the compound $M_n(0)$ gain of the cascade is found as:

$$M_n(0) = M_{1n}(0) M_{2n}(0) / [1 - (M_{2n}(0) - 1) M_{1p}(0)] \quad (6.19)$$

and applying Eq.(6.19) recurrently to a structure with step-wise approximated field distribution, the gain can be calculated. Also, by computing $\sigma_M^2 = [\Delta M_n(0)]^2$, one can find that Eq.(6.18) still applies.

6.1.3 Experimental Evidence and Deviations

Experimental results match well the general trends of theory findings. The optimal gain $M_0 = \alpha/\beta$ is clearly confirmed as the key parameter of both frequency response and noise of the APD. However, surprisingly, in applications of APDs, gain-bandwidth product and excess noise factor are found serendipitously a little bit better than expected from theory, especially in thin junctions [3,4]. This mitigation of the positive feedback effects has been observed in GaAs [5], $\text{Al}_{0.8}\text{Ga}_{0.2}\text{As}$ [6], and in N_2 -cooled HgCdTe [7]. Such improved performance has been attributed to a number of effects that regularize the impact ionization, beyond the random exponential path distribution assumed to model the APD and presented in the previous Section. One effect is dead space between successive multiplications, i.e., the nonzero path the carrier shall cover before becoming able to ionize. Another is bandgap boundary effect encountered in heterojunction APDs, in which injected carriers arrive in the multiplication region with an initial energy [5].

The theory incorporating these nonlocal effects is rather involved and will not be reported here. However, in most cases the deviations amount to an improvement of a factor 1.5...2 of the gain-bandwidth product $f_{2D}M$, Eq.(6.8"), and of the excess noise factor F , Eq.(6.10"), when we are working at a gain exceeding the optimal gain ($M > \alpha/\beta$), whereas below the optimal gain ($M < \alpha/\beta$) there is no deviation from the results of the old theory.

Therefore, we can take the results of the 'memoryless' theory reported in previous Section as a conservative estimate of the APD performance.

6.1.4 APD Structures

With the normal structure of a pin (e.g., Fig.5-21 b,d) the APD does work, but does not yield good performance, because the depleted region serves both for the photon dissipation and the carrier multiplication. In this case, the length L of the multiplication region should be tied to the absorption length L_a to get a good quantum efficiency; moreover, the multiplication gain is that of M_I for distributed pairs, less than the maximum available $M_c(0)$ [Eq.(6.7)] and, even worse, with a large variance [Eqs.(6.9) and (6.10)].

A desirable structure for an APD would have two separate regions, as in Fig.6-1.8: one of thickness $L_1 \approx L_a$ for the photon absorption and the carrier drift, with a medium/high applied field E_d to reduce transit time, the other with a high field E_m for the multiplication of the most efficient ionizing carrier.

The ionizing carrier will be injected from the absorption region into the avalanche region, which is of a thickness L adjusted to get the desired gain $M = M(\alpha L)$. To obtain a high opti-

mum gain, materials with a high α/β ratio are preferred and, in the case of heterostructures, the optical entrance from the high field side is allowed, if a high bandgap material is used there (see Section 5.2.7), which is transparent to incident radiation at the λ of interest.

Not all these requirements can always be satisfied in a practical design, because of material-related limitations or fabrication difficulties.

The structures currently employed for APDs are illustrated in Fig.6-1.9, along with the schematic distributions of depleted charge and electric field, which may be compared (Fig.6-9a) with that of the pin shown as a reference.

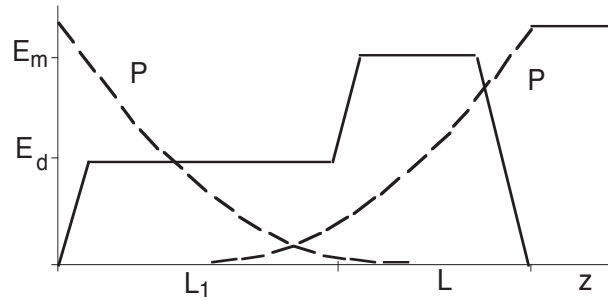


Figure 6-8 The desirable field distribution for an APD: left curve is for entrance from the drift region, right curve from the multiplication region.

The simple abrupt n^+np junction of Fig.6-9b employed in the first *germanium APDs* (in which $\beta > \alpha$) allows to obtain a reasonable quantum efficiency because of the absorption has a thick n -region available. Of the pairs produced in this region, electrons are at once collected by the cathode without multiplication, and holes are injected in the junction where they can contribute to the multiplication gain in correspondence with the peak of the electric field. This structure is not optimal, because the electric field in the multiplication region is not constant, nor can the drift region be made too thick; but it is simple and practicable and is used in many commercially available Ge APDs.

A substantial improvement is offered by the $n^+p^+\pi p^+$ structure, called *reach-through* (fig.6-9c) and commonly used in silicon APDs. Here, a first n^+p^+ junction of small thickness and high field provides the multiplication for the electrons coming from the thick undepleted π (i.e., lightly p -doped) region where photons are absorbed. Here, photons can cross the first junction with no loss because of its thinness, and the photogenerated holes in π are collected at the anode without multiplication.

The name *reach-through* comes from the way p regions are depleted at increasing bias voltage: first, the p^+ region is depleted and from here depletion becomes extended all through the region π reaching the substrate p^+ ; this reach-through condition is necessary to set-up the drift field in region π . A useful consequence of this mechanism is that the avalanche field at the n^+p^+ junction is less dependent on the external bias voltage compared, e.g., to an n^+np structure.

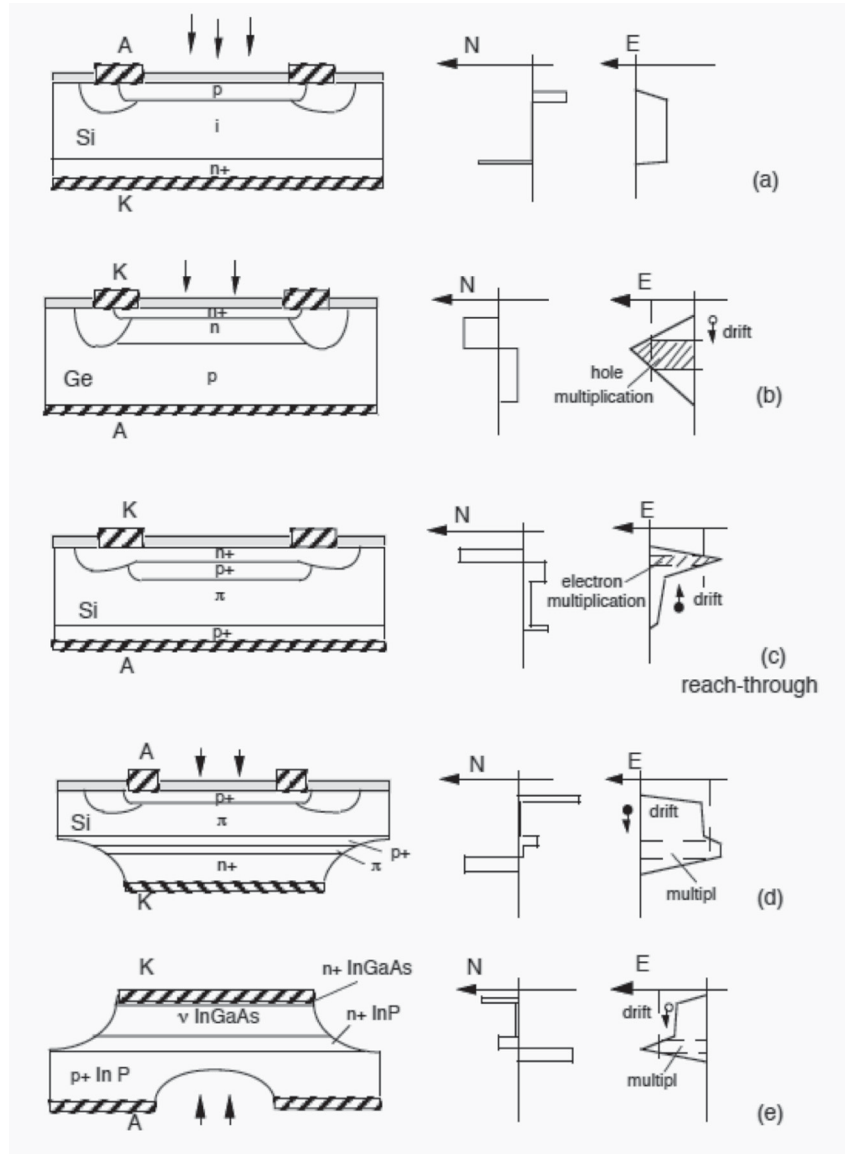


Figure 6-9 The pin structure (a) compared to several commonly used APD structures: (b) the Ge-APD n^+np ; (c) the popular *reach-through* structure; (d) the inverted reach-through, and (e) quaternary compound SAM. On the right, the schematic distributions of dopant charges and electric field are indicated.

An interesting variation for Si APDs, that looks like an inverted reach-through is the $p^+ \pi p^+ \pi n^+$ structure of Fig.6-9d. This offers a more uniform field in the drift and multiplication regions and has the optical access from the drift side.

The structure is more critical to fabricate because of the thick epitaxial π layer, but it offers a wider multiplication region and allows operation at moderate fields, where Si has the largest α/β ratio (see Fig.6-1).

A further improvement of the reach-through structure is obtained with the use of a *heterojunction* in compound semiconductors, because one can then choose the materials best suited for absorption and ionization in the two regions.

Fig.6-9e shows an example of a heterojunction APD in $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}/\text{InP}$, designed for the detection in the near IR up to $\lambda \approx 1.8 \mu\text{m}$ as required for fiber-optics communications.

Such a structure is called SAM (*separate absorption and multiplication*), to indicate that the absorption and ionization regions are separated and in different material or composition. Radiation is incident (Fig.6-9e) on the material with higher E_g , the indium phosphide of the substrate, and is dissipated in the thick region v , in indium-gallium arsenide; electrons are directly collected and holes drift to the p^+n^+ junction of multiplication.

We may remark that the $p^+n^+vn^+$ structure is complementary to the $n^+p^+\pi p^+$ reach-through of fig.6-9c, as it is $\beta > \alpha$ in InP (Fig.6-1) and as it uses the principle of a transparent incident layer (as described in Sect.5.2.7 and Fig.5-15).

A problem with the heterojunction is the presence of interface states at the InP/InGaAs boundary, which trap charge carriers. These states are nearly eliminated by sandwiching a layer with gradually varying composition, a structure called SAGM – *separate absorption, graded, and multiplication*.

An APD based on InGaAs/InP heterostructure has a useful spectral range $\lambda \approx 0.9\text{--}1.7 \mu\text{m}$, see Fig.5-2.16. It has a good bandwidth performance ($\approx \text{GHz}$), while the noise figure is relatively high because of the low ionization ratio β/α available in InP (see Fig.6-1.1).

Another interesting compound semiconductor for the APD heterostructure is gallium aluminum antimonide ($\text{Ga}_{0.94}\text{Al}_{0.06}\text{Sb}/\text{GaSb}$) which is grown on GaAs substrates and has a high ionization ratio $\alpha/\beta \approx 50$ (Fig.6-1.1) and a spectral range comparable to $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}$ (see Fig.5-17). The material is however difficult to grow and this hampers the commercial development of such a device.

6.1.5 Bandgap Engineered APDs

The excellent potentials of low-noise and large bandwidth of APDs as compared with a pin photodiode, have not yet been fully exploited. Even the best performances, offered by Si-APDs, are limited to a moderate internal gain (50–100 max) and cover a narrow spectral range (0.5–1.0 μm).

One direction is toward work on new materials, the best candidates being those with a very high ionization ratio α/β and a high carrier mobility μ , and an energy gap suitable for

the intended spectral range. Compound semiconductors and GaAsSb are an example of this trend, yet the difficulty is that a new material requires development of a new fabrication technology. The many steps like doping, oxidizing, etching, epitaxial growth, etc., are so costly that a single new component hardly justifies them.

Another approach is to develop the SAM concept further, in combination with heterostructures. If a single material cannot match both the spectral range and the α/β requisite, we can try to make a two-layer sandwich of different materials, one optimal for photon dissipation and the other for multiplication. Of course, the two layers shall be lattice-matched (Section 5.2.7), or very thin strained quantum wells.

This approach has been attempted with Si as the thick multiplication material, and a Ge/Si strained multilayer as the photon dissipation material. The multilayer is made of alternated Ge and Si layers, thin enough to match the Si bulk lattice constant (Fig.5-18). While Ge absorbs up to $\lambda \approx 1.7 \mu\text{m}$, silicon is transparent beyond its threshold $\lambda_s \approx 1.1 \mu\text{m}$ and only serves to keep memory of the lattice constant for the Ge layers. This Ge/Si APD has a p -doped Si region, and an n -doped Ge/Si layer, so that electrons are injected from Si into Ge/Si for multiplication; the optimum gain is that of Si (50–100) and the spectral response is that of Ge ($\lambda_s \approx 1.7 \mu\text{m}$), which is adequate for optical communications.

Two viable approaches for the synthesis of materials with unusual electronic properties have been demonstrated. Useful for APDs as well as for lasers and other photonic devices, these are the *bandgap engineering* and *superlattice* structures proposed by F. Capasso [8].

An example of a bandwidth engineering device is the *sawtooth* avalanche structure shown in Fig.6-10a. Many (30–50) layers of p -type $\text{Ga}_x\text{Al}_{1-x}\text{As}$ are grown one over the other, each with a thickness $< 1 \mu\text{m}$ and a composition x linearly increasing from a minimum of $x=0.55$ of $\text{Ga}_{0.55}\text{Al}_{0.45}\text{As}$ to the maximum $x=1$ of GaAs. The energy gap linearly increases in the layers from 1.4 to 1.8 eV, and, because of valence band alignment of the p -type doping, the potential energy curve resembles a sawtooth sequence. To complete the photodiode, an n^+ region is finally grown next to the p -sawtooth sequence.

By reverse-biasing the photodiode, at a certain voltage the depletion of the material tilts the potential distribution, and the working condition of Fig.6-10a is reached. Here, the field in the valence band is uniform and fairly high (e.g., $E \approx 10^5 \text{ V/cm}$) to give a high speed for hole drift, but not enough to yield substantial ionization by holes (as β is low, see Fig.6-1). In other words, holes do not contribute to multiplication and are swept off quickly.

In the conduction band, electrons drift along a sequence of alternately high and low fields. In the low-field regions electrons have a fairly high drift speed but α is small, while in the high-field regions α is very high and electrons strongly ionize. The regions at high fields occupy only a small fraction of the total width L , let us say ηL , but they sustain a high field $E=V/\eta L$ that makes α grow very large because α increases exponentially with the field (Fig.6-1). This more than compensates for the decrease in multiplication width, yielding a number of multiplications $\alpha\eta L$ much larger than the corresponding value of a uniform structure with the same total impressed voltage.

With the sawtooth layers we have obtained a synthetic material exhibiting an apparent α/β ratio much larger than the bulk material (by a factor 30–100), with a corresponding improvement in optimum gain and related performance.

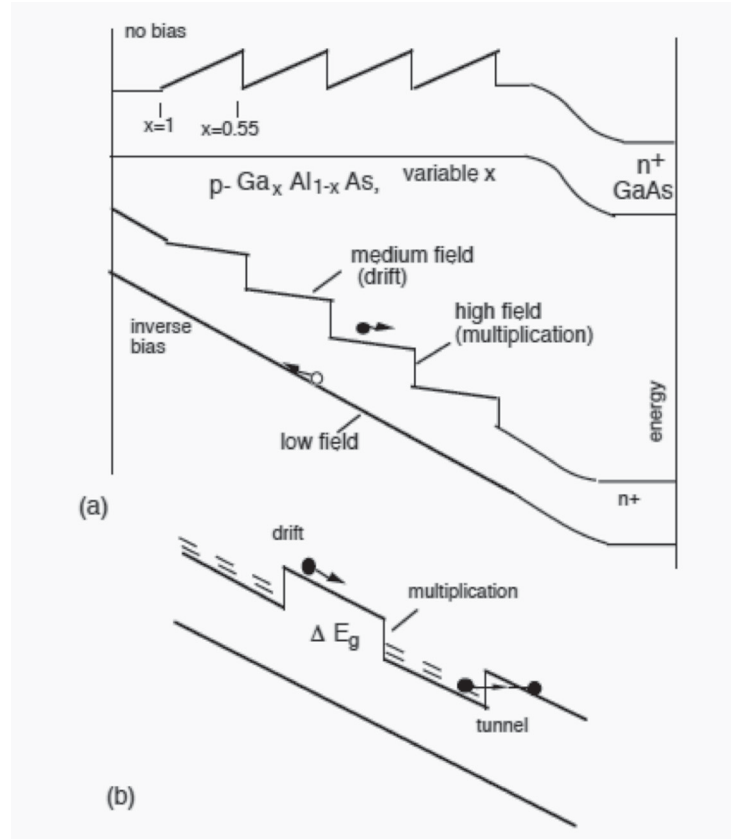


Figure 6-10 (a) Sawtooth APD in GaAlAs: distribution of the potential energy with no bias (above), and with reverse bias (below); (b) superlattice APD.

In *superlattices*, the structure is again a stack of several layers grown with composition (bandgap) alternatively high/low, but the layer thickness is now smaller than the mean free path of electrons (≈ 10 nm). This brings about the energy confinement of electrons in discrete levels (dashed lines in Fig.6-10b) inside the wells that are found in the conduction band as a result of valence band alignment. When properly biased, the superlattice stack presents a moderate field to holes, and a strong field to electrons corresponding to the potential step-down. Thus, an apparent α/β much larger (up to 300) than in the bulk material has been obtained for the system GaAs/Ga_{0.55}Al_{0.45}As.

To make a practical APD device with the high α/β synthetic p-regions, provision must be made for photogenerated electrons to enter from the left side in Fig.6-10. The inverted reach-through structure is adequate to this purpose, and therefore we will add two p^+i -regions to the $p\text{-}n^+$ junction of Fig.6-10 to complete the APD.

6.1.6 APD Biasing and Use Requisites

The bias voltage required in avalanche photodiodes is appreciably higher than in *pin* photodiodes because there are two regions (drift and multiplication) to be kept under a high electric field. Thus, bias voltages up to $V_{bb} \approx 100\text{--}300\text{ V}$ can be found for APDs.

A problem for the user is the strong dependence of the gain M on the bias voltage and temperature, a consequence of α and β strongly increasing with the electric field and temperature.

In Fig.6-11 we report the typical curves of $M=M(V_{ak}, T)$ for a reach-through Si-APD. As the junction thickness in reach-through is relatively large, the corresponding voltage is larger than in other APDs. As an example, by choosing in the diagram a quiescent point at $V_{ak}=205\text{ V}$ for a gain $M=100$ at $T=20^\circ\text{C}$, from the slope of the curve we find a voltage coefficient $\Delta M/(M \cdot \Delta V_{ak}) \approx +10\%/V$. Moving along the ordinate-axis we find a temperature coefficient of gain $\Delta M/(M \cdot \Delta T) \approx -8\%/^\circ\text{C}$ and, along abscissa-axis we find the voltage change needed for a constant gain with temperature as $\Delta V_{ak}/V_{ak} \cdot \Delta T \approx +0.3\text{ }^\circ\text{C}$, and $\Delta V_{ak}/\Delta T \approx +0.6\text{ V}/^\circ\text{C}$. Values not much different are found in APDs based on other semiconductors.

From these values, it follows that the supply voltage for APDs must be very well regulated (differently from *pin* photodiodes) and free from ripple, and that a temperature correction shall be provided to keep the gain reasonably constant in the range of operation.

Regarding supply voltage, requirements vary according to the intended application, e.g., from a typical $\Delta M/M \leq 0.5\%$ in analogue measurements to perhaps $<5\%$ in digital data detection. For a Si APD, these figures correspond to a regulation better than $\Delta V_{bb} \leq 0.05\text{ V}$ and 0.5 V , inclusive of all drift contributions and of the residual ripple from the rectification circuit of the power supply at $V_{bb}=205\text{ V}$.

Of course, the supply must have a low internal resistance R_s up to the maximum current $I=MI_{ph}$ set by the allowed power dissipation $P=MI_{ph}V_{bb}$; then, it will change to a current limitation mode (with $R_s=\infty$) to protect the APD.

Thermal control can be implemented by at least two methods. One very elegant method consists of tracking the APD temperature by means of a number of *pn* junctions connected in series, each of which yields a voltage drop $\approx 700\text{ mV}$ and a temperature drift of $-2.5\text{ mV}/^\circ\text{C}$ (in Si). The junctions are fabricated on the same chip, close to the APD, and their voltage is amplified and subtracted from the external voltage supply. For example, with $n=8$ junctions we have $-20.0\text{ mV}/^\circ\text{C}$ and by a $\times 30$ amplification we can cancel the $\Delta V_{ak}/\Delta T \approx +0.6\text{ V}/^\circ\text{C}$ coefficient quoted above.

A disadvantage of this method is that a trim is required in the subtraction circuit, because the temperature coefficient of individual devices has a (small) spread around the nominal value.

Another method, also very popular for temperature-sensitive devices like diode lasers, is to control the temperature of the APD chip by means of a *Peltier cell* or TEC (*thermoelectric cooler*).

The TEC is a series of junctions, made of dissimilar metals or semiconductors and sandwiched between two thin ceramic end plates with the *p* and *n* arranged to face the same plate. By passing a current through the cell, a thermal power P_{th} is brought from the *p* to the

n face because of the Peltier (or inverse Seebeck) effect. The thermal power is proportional to electrical power fed to the cell, $P_{th} = V \cdot I$, and reversing the current flow in the cell makes the cold and hot face interchange.

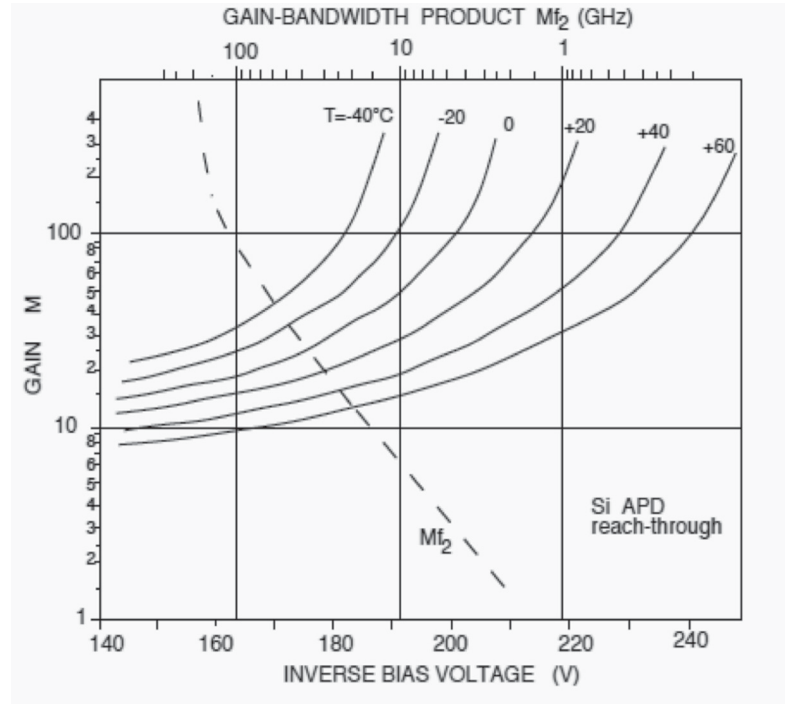


Figure 6-11 The gain M of a Si APD as a function of bias voltage and with ambient temperature as a parameter (full-line curves). Also shown is the gain-bandwidth product Mf_2 (dotted line).

Thus, if a thermistor is mounted close to the APD chip and a suitable feedback loop is arranged, we can actuate a thermostatic control of the device. Usually, the TEC is mounted internal to the APD package, with one face in thermal contact with the chip and the other coupled to the external heat sink.

If the APD is packaged in a small outline package SOP (typ. $10 \times 10 \times 4$ mm on a side), the temperature can be regulated within 0.1°C on a range of $\pm 20^\circ\text{C}$ from ambient temperature, at the expense of substantial power consumption (few Watt), however.

The general characteristics of APD are similar to those of normal photodiodes (Section 5.2.9), except for the smaller sensitive area, generally around 0.01-mm^2 , to limit the effects of edge defects and field non-uniformity. Recently, however, large-area APDs have been fabricated successfully (up to $A=10\text{--}50\text{ mm}^2$) with a proton bombardment technology.

6.2 SINGLE-PHOTON AVALANCHE DETECTORS (SPAD)

As we have seen in the previous Sections, the APD has superior performances of noise respect to the pin photodiode and also with respect to any other photodetector thanks to the internal gain. Internal gain mitigates the requirement of a high-value load resistance to keep the Johnson noise low, and so eases the tradeoff between noise and bandwidth (Sect.3.1).

Yet the APD is far from the reach of single-photon performance, because of the limited α/β ratio of carrier ionization available in known semiconductors. So, by operating the APD in the *linear mode*, that is with the output voltage (or current) proportional to the incident detected power, we obtain a minimum detectable signal $S_{\min} = \text{SNR } i_n$ where i_n , the total noise, is about one decade better than the *pin*-based detector but still one decade or more far from the PMT performance (Fig.6-12).

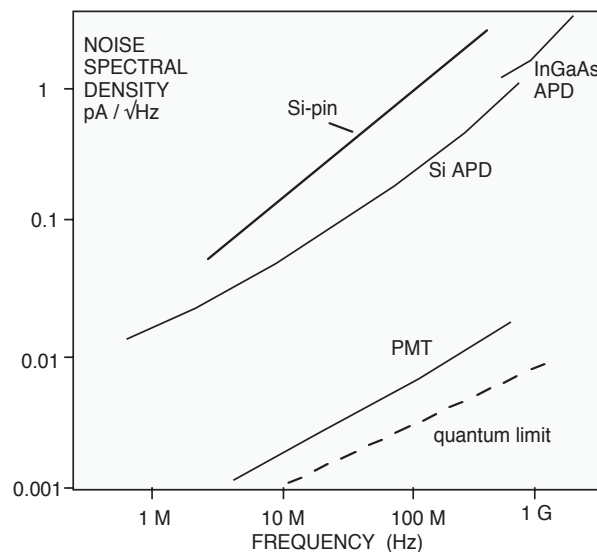


Figure 6-12 Typical noise performance of *pin*, APD and PMT receivers as a function of the maximum frequency of operation (or bandwidth) f_2 .

From another point of view, if we try to detect a packet of photons contained in a fast pulse, in the linear mode we reach a typical limit of $R \approx 30$ as the minimum number of photoelectrons that is measurable with $S/N=3$ (Probl.P6-2), better than the $R=3,000$ of a *pin*, but still far away from $R=1$ of the PMT.

6.2.1 The APD in Geiger mode

A way out to reach the single photon performance is to operate the APD in the *Geiger mode*, that is, biasing it at voltage *larger* than the infinity gain voltage V_∞ (see Fig.6-13).

Of course, in this situation, we must limit the current flowing through the device, and will do so by using a suitably large resistance R_b in series to the cathode (Fig.6-14) so that the maximum current through it and the dissipation are limited to a safe value.

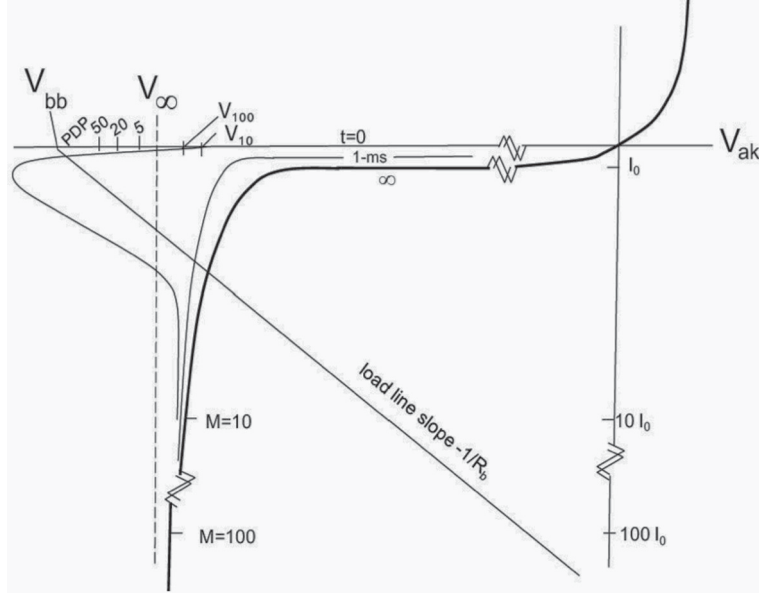


Figure 6-13 The bias voltage for an APD working in single-photon regime is chosen at $V_{bb} > V_{\infty}$. Also shown are the static load line corresponding to ballast resistance R_0 , the change in V-I characteristic from the supply voltage switch-on time ($t=0$) to dc ($t=\infty$) and the PDP (% probability that the avalanche is triggered).

Let's assume the voltage V_{bb} is switched on at time $t=0$, so that the initial current through the junction is also zero. Then, capacitor C (Fig.6-14) charges up gradually, with a time constant $R_b C$, to the battery value V_{bb} . As voltage V_{ak} increases, the diode passes through the increasing values of gain M (e.g., the V_{10} and V_{100} values in Fig.6-13) and then the infinity gain voltage V_{∞} .

By the way, the infinity gain can always be attained, as we can see from the asymptotes of gain as a function of αL in Fig.6-3. For example, for $\alpha/\beta=100$ as we increase αL from 4 at which $M=10^2$ to say, little bit less than 5, we get $M=\infty$.

At the time V_{ak} exceeds V_{∞} , the diode current is still $I=0$. But now, as soon as a single charge carrier pair is generated in the junction, either by photo-detection or thermal excitation at dark, the avalanche builds up, and current increases quickly with an exponentially increasing waveform, similar to those shown in Fig.6-6.

Actually, due to the probabilistic nature of the process, the avalanche is triggered with a certain probability called PDP (*photoelectron detection probability*).

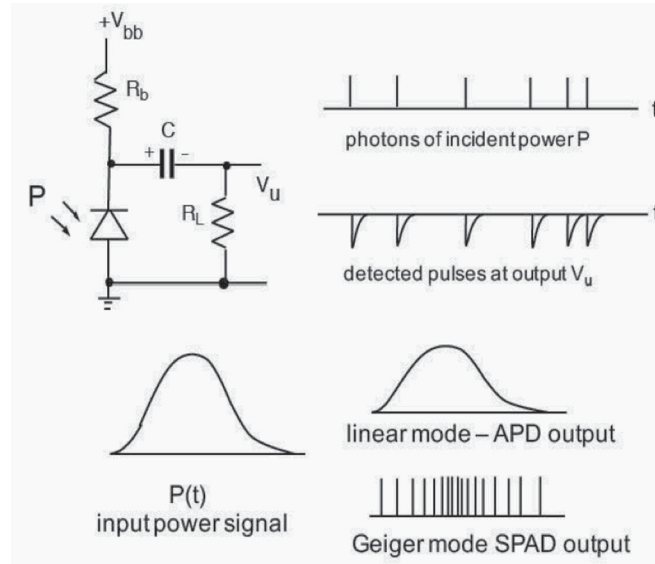


Figure 6-14 Basic circuit of the SPAD. The diode is biased at $V_{bb} > V_{\infty}$ through a large R_b , and a low resistance load R_L is coupled through capacitor C (top left). When a photon is detected, the SPAD switches to a high-current state and discharges C on R_L (top right). So, the sequence of detected photons is replicated with a sequence of pulses. Different from the linear mode of the APD supplying a linear replica of $P(t)$, the Geiger mode supplies a sequence of pulses whose rate is proportional to $P(t)$ (bottom).

The PDP increases at increasing overdrive voltage ΔV above V_{∞} , with the typical trend shown in Fig.6-15. Because of the PDP, the equivalent quantum efficiency of the device is decreased to $\eta_{eq} = \text{PDP} \eta$, with respect to the value η of the quantum detection process. Further, as the probability of triggering applies also to thermally generated carriers, at the increasing of the overdrive ΔV also dark current I_0 will increase. Therefore, we have to make a trade-off between a high PDP and a low I_0 by choosing the appropriate ΔV .

Once the avalanche has built up, the device is in a state of high current capability, and it draws all the current permitted by the external circuit, in this case the current of the discharge of capacitor C on the load resistance R_L . As this current is much larger than that drawn by R_b from the battery V_{bb} , now the diode voltage doesn't increase anymore, and V_{ak} is pinned to $V_{\infty} + \Delta V$.

Current I continues to flow until the voltage drop across resistance R_b (Fig.6-14) is such that the cathode voltage $V_{ak} = V_{bb} - R_b I$ falls below V_{∞} , and the gain returns to finite value. Then, current is no more sustained now and damps off, or, the avalanche is quenched.

Meanwhile, capacitor C has been partially discharged across the output resistance R_L , delivering at the output an exponential pulse, decaying with time constant $R_L C$ and with an amplitude given by the excess ΔV voltage, with respect to V_∞ , reached during the charging phase [in practice, up to $\Delta V \approx V_{bb} - V_\infty$].

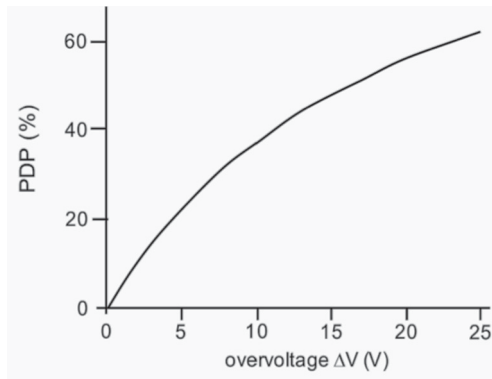


Figure 6-15 Typical PDP for a reach-through structure SPAD as a function of overdrive voltage ΔV in excess of V_∞ [for shallow junction, voltages are smaller].

In conclusion, for each detected photon (or thermally generated carrier) in the multiplication region when the SPAD is in the infinity gain state (at $V_{ak} \geq V_\infty$) a pulse is generated at the output with probability PDP. [This is as shown schematically in Fig.6-14, top, for the case $PDP=1$, while for $PDP<1$ some photons will have the corresponding pulse missing at the output]. Therefore, the SPAD has the single-photon detection capability.

There are important differences, however, with respect to a PMT.

First, following a detected pulse, there is a *dead time* for the SPAD to recover from the high current (on state) to the zero current (off state), and become ready to sense the next photoelectron. In the circuit of Fig.6-14, the dead time amounts to the time necessary for capacitor C to recharge from V_{on} , the voltage a little bit below V_∞ where the cathode ends during the discharge, to V_{off} the minimum voltage above V_∞ where the device is off and already infinity gain [$V_{off} < V_{bb}$ because we need to reach the infinity gain, not the supply voltage]. As the capacitor recharges through R_b (neglecting $R_L \ll R_b$) the dead time is easily calculated as $T_d = R_b C \ln [(V_{bb} - V_{on}) / (V_{bb} - V_{off})] \approx 0.8-1.5 R_b C$ in practice (see Fig.6-16).

The passive recharge provided by the simple RC circuit is rather slow, typically $T_d = 1 \mu s$ for $R_b = 100\text{-k}\Omega$ and $C = 10\text{-pF}$, but active circuits (see below) can shorten this time to a more reasonable $T_d = 5$ to 50-ns .

Second difference, during the dead time the SPAD *ignores* all detected photons, and also ignores, during the on-state, any further photon detected after the avalanche has been already triggered. So, if a short packet of R photoelectrons is detected by the SPAD, the out-

put will be just one pulse of the standard amplitude, like if it were $R=1$. Only if the packet is slower than the dead time we will have R pulses (or, more precisely, $R \times \text{PDP}$ pulses).

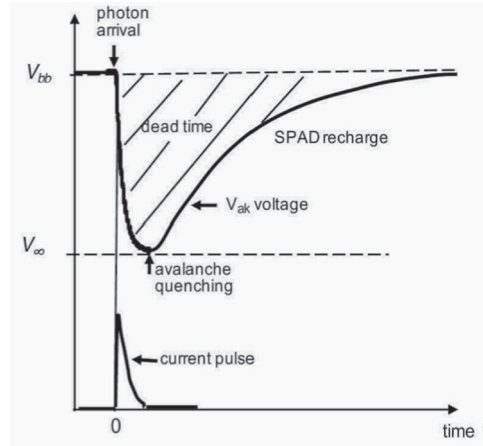


Figure 6-16 Illustration of the dead time after the detection of a photon and recovery.

In contrast, a PMT has a linear response down to times shorter than the SER (see Fig.4-29), and has no dead time. For this reason, the SPAD is not quite the equivalent of the PMT, it is rather the counterpart of the *Geiger Muller* tube, in which a particle or radiation quantum ionizes the gas and triggers an avalanche discharge delivering at the output a pulse of standard amplitude and duration, and where a dead time is needed after a detection to get ready for the next event. This is why the SPAD pulse regime is also called the *Geiger mode*, after the work pioneered by S. Cova and coworkers [9], in opposition to the linear mode of the APD.

The third feature of the SPAD is the relatively large *reverse current*, of the order of $J_0=1-5 \text{ pA/mm}^2$ in semiconductor junction with guard ring, which corresponds to a generation rate of $F = J_0/e = 6-30 \cdot 10^6 \text{ counts/mm}^2$. For a 1-mm^2 SPAD, this means that there is a dark count every $1/F=33-160 \text{ ns}$, a value problematic because it's comparable to the dead time T_d . There are two ways to fix the problem are: (i) use a *small active area* A , compatibly with the application – typically not cm^2 nor mm^2 of photodiodes, but rather $100 \times 100 \mu\text{m}^2$ in discrete device and $10 \times 10 \mu\text{m}^2$ in integrated SPAD; (ii) since for Si and other low intrinsic concentration n_i semiconductors the dark current is mainly due to the generation-recombination contribution $J_0 \propto W$ (Eq.5.17), also the junction *width* W will be made as *short* as possible, typically a few μm , thus renouncing to collect efficiently photons with long attenuation length $1/\alpha$, or equivalently, sacrificing the spectral response near the photoelectric threshold and in the red for low dark current. Equally important, all extra *sources of reverse current* (from defects, surface state, fringing field effects) shall be carefully controlled and

reduced as much as possible; also, increasing the carrier lifetime τ in lightly doped materials is another possible area of improvement of F.

Fourth, but this time, there is a positive consequence of the third point: as the width W of the device shall be made short, the transit time will be short too, and time response is *very fast* and considerably better than that of a PMT or a photodiode, with timing *jitters* consistently below 100-ps.

To finish, a disturbing feature of the SPAD is *afterpulsing*, that is, retriggering of the discharge some 50–200 ns after a main detection event, because of the charges blocked in trapping centers and released later, and then retriggering the discharge. Afterpulses generated with short release time can be eliminated by means of active quenching (see below), whereas states with long release time are worse and can be tackled only by improving the material.

In conclusion, the SPAD is not exactly the solid-state retrofit of the PMT, but it can be regarded as complementary to it [10]. In general, when the main requirements of the application are miniaturization and integration (as in multipixel arrays), or fast time response (for time-of-flight rangefinders, fluorescence spectroscopy, etc.), the SPAD is definitely preferable, whereas for large area (such as for the Kamiokande experiment, Fig.P1) or very low dark current applications (photon counting, nuclear radiometry, dating with radionuclides, Sections 4-8.4 to 4-8.6) does not yet fully achieve the PMT performance. Another difference is in the spectral range covered: the PMT can go deep into the UV but is limited to the NIR, while the SPAD, built as a semiconductor material, can potentially cover wavelengths up to the MIR and FIR.

6.2.2 SPAD Structures

Many APD devices intended for use below the breakdown voltage in the linear mode, can also be biased above the breakdown voltage and work as a SPAD in the Geiger mode.

For example, the popular C30921S silicon APD, the first introduced with the reach-through structure and data shown in Figs.6-9c and 6-11, works also fairly well as a SPAD. It has 0.1-mm dia. sensitive area, covers the full spectral range of Si (450 to 1050-nm) and 460-ps FWHM time resolution because of the thick junction, and with a simple passive self-quenching offers 300-ns dead time, and 20 kc/s dark count rate at PDP=5%.

The precursor of SPAD structures, considered by Haitz [11] in the study of contributions to dark current from microplasmas, was not much different from a n^+p junction with n guard ring (Fig.6-17 top left), and had problem of PDP uniformity and high dark current.

A more performant SPAD structure, pioneered by S. Cova and coworkers [9], is the double epitaxial structure of Fig.6-17 top right, grown inside a 12- μ m deep p-epitaxial layer, with a thin p^+n^+ junction acting as the avalanche region and a p^+ well to eliminate the slow tail of time response due to electrons diffusing from the substrate. The time response attains a record 43 ps FWHM, albeit at the expense of a reduced spectral width (400 to 900 nm). An improvement was later proposed, with the introduction of a double junction [12], by insert-

ing a patterned p^{++} buried layer beneath the junction (Fig.6-17 bottom left) so as to eliminate the contribution from the neutral layer and have a clean time resolution curve, without slow tail components, down to 10^{-3} of the peak.

Both the above-described structures are best performing with a medium-to-small active area (up to 100- μm diameter). For fabrication, they require good quality Si and the standard planar technology. So, they are the best for stand-alone SPADs intended for application in timing instrumentation.

But, to work properly, already the bare SPAD requires some ancillary circuit functions, like e.g., quenching, amplification, ADC or other preprocessing, and therefore it's advisable to adopt structures compatible with a standard IC process, so that the photosensor module (SPAD and its circuits) can be manufactured in batch as a single chip at low cost.

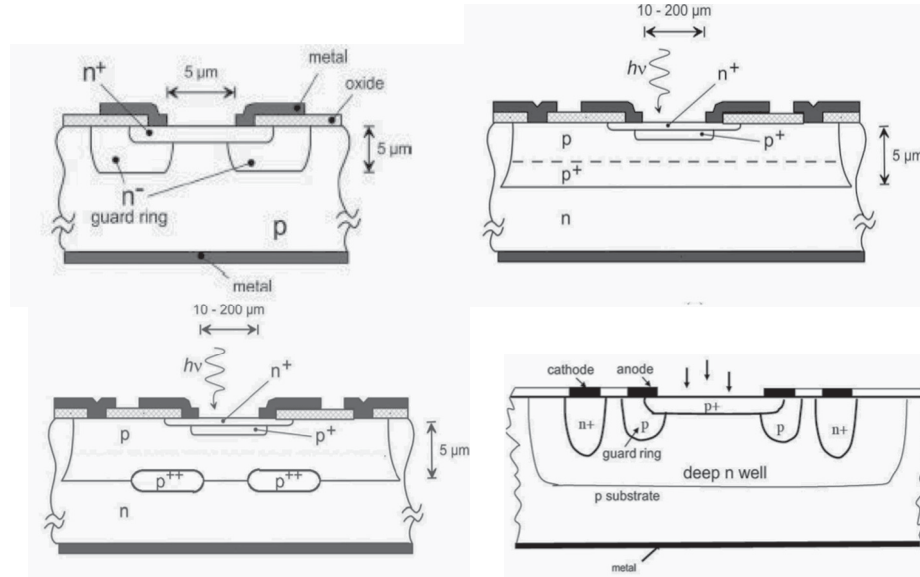


Figure 6-17 Typical structures of SPADs: top left, the early structure of Haitz, top right, the epitaxial reach structure of Cova et al. (from [9], by courtesy of IEEE). Bottom left, same, with buried layers, bottom right, a structure compatible with standard CMOS process (from [18]).

The first example of a SPAD structure integrated with a standard IC process was that shown in Fig.6-17 bottom right. Introduced by Rochas et al. [13], the structure used a 0.8- μm high-voltage standard CMOS technology, and used a p -guard ring to control the fringing field, thus eliminating the so called PEB (premature edge breakdown), and a deep n -well to reduce diffusion tail and shorten the time response.

With a 7- μm diameter active area, the SPAD had a 25-V V_∞ voltage, was passively quenched by the load resistance R_b , and had the following performance: a wide spectral response from 400 to 750-nm at 10% detection probability η_{eq} (and a significant 1% at 1000-nm), a time resolution of 60-ps FWHM; a dark count rate less than 1000 cps, and with a dead time of 75 ns, afterpulses were a total 7.5% of which 90% occurring in the first μs after the event.

After this device in the 0.8- μm technology, several others have been reported at deep submicron CMOS technology, from 0.3- μm to 180-nm, 130-nm, and 90-nm [14,15].

Of course, decreasing feature size allows to incorporate more circuit functions alongside the SPAD, and this is an important modern trend called *smart pixel* that will be discussed later. Also, not less important, it's the key to fabricate *SPAD arrays* for a number of applications.

Problems with submicron technology are: (i) a uniform multiplication region becomes more difficult to achieve and (ii) noise sources become more significant, because of the increased doping levels and the shallow trench isolation [13]. For these reasons, dark current and afterpulses rate do not decrease as expected with size.

Instead, the thinner junction has less absorption efficiency for red photons, and therefore the spectral response curve is considerably narrower. For example, the 0.8- μm SPAD has a FWHM spectral range $\Delta\lambda_{FWHM}$ of 300 nm, the 130-nm SPAD has 210 nm, and the 90-nm only 120 nm or, the photon detection efficiency extends from 450 to 570 nm at half the maximum.

6.2.3 SPAD Quenching

The SPAD can be operated in the Geiger mode by either *passive* or *active* quenching circuits. In the circuit of Fig.6-14 already discussed above, the SPAD is passively quenched with the aid of capacitor C discharging on the output load R_L .

Even simpler, the circuit of Fig.6-18 (left) offers a passive quenching, because when the SPAD turns on, current flows in the external resistances and produces a voltage drop that decreases the anode-to-cathode voltage below V_∞ , thus turning off the SPAD; the recovery then follows with the recharge of parallel capacitance C (sum of junction C_b plus stray C_p) to the battery voltage V_{bb} . The recharge will take a few time constants $R_b C = 200 \text{ k}\Omega \cdot 5 \text{ pF} = 1 \mu\text{s}$, a relatively long dead time - that however can be acceptable if we need to reject afterpulses with long (e.g., up to 1000-ns) trapping time.

With another variant (Fig.6-18 right) we may look at the V_{ak} voltage as the signal, because as the SPAD turns on, V_{ak} will switch down of $V_{bb} - V_\infty$, and after the SPAD goes off, it will recharge to V_{bb} . The comparator helps to shorten the output pulse duration, as voltage V_{ak} is longer than the C-R pulse. Active quenching adds more flexibility, and several circuit solutions have been proposed to implement it, see Ref. [16] for a good survey.

With active circuits, it is possible to initiate the quenching of the discharge as soon as it starts, so as to have a short recovery time for material with little afterpulses and, at the opposite, we can control the hold-off time so that the long-delayed afterpulses are eliminated.

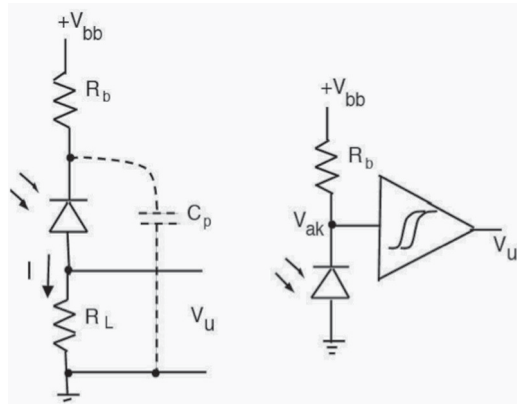


Figure 6-18 Passive SPAD quenching circuits. Left, sensing the current spike of stray capacitance discharging on R_L [typical values for the C30921S are $R_b=200\text{ k}\Omega$ and $R_L=220\text{ }\Omega$]; right, looking at the V_{ak} negative-going waveform, shortened in time duration by the comparator.

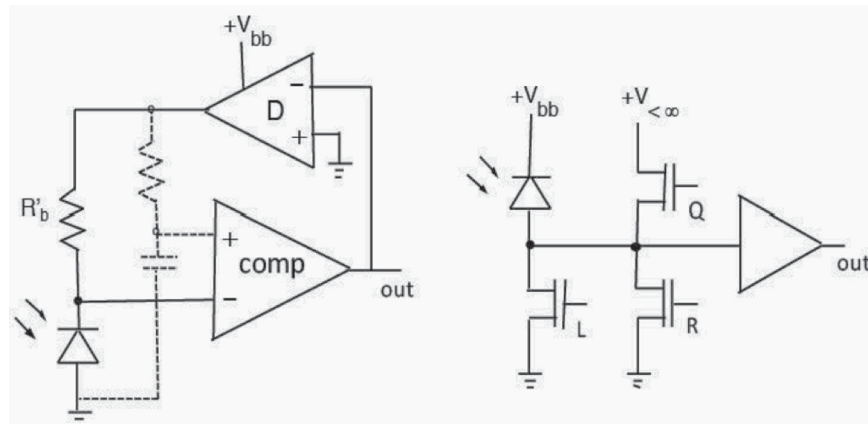


Figure 6-19 Circuits for active SPAD quenching: left, the coincidence quenching [16] for a stand-alone SPAD; right, the circuit for integrated SPAD with transistors ready to switch the SPAD below V_∞ to quench the discharge and thereafter recharge it to V_{bb} [10].

For a stand-alone SPAD, the coincidence quenching [16] circuit of Fig.6-19 (left) uses a differential comparator to sense that the avalanche is started, that is the condition $V_{ak} < V_{bb}$, and reacts switching off the SPAD with a negative-going step provided by the driver D , and thereafter the SPAD recharges it to the ready-to-detect state. Dotted line group is a compensator to cancel the transient of the negative-going step on the differential inputs of the com-

parator. A monostable multi inserted in the driver (not shown in figure) serves to reset the circuit when the SPAD has gone off.

For a SPAD integrated in CMOS technology, the circuit of Fig.6-19 right offers a similar sequence, with the MOS transistor Q quenching the avalanche (by imposing a V_{ak} , voltage $< V_{\infty}$), followed by transistor R that charges V_{ak} back to the battery voltage V_{bb} for the next photon detection, and with transistor L acting as the SPAD ballast resistance.

6.2.4 SPAD Performances and Parameters

In Fig.6-20 we show the typical spectral response of the *equivalent quantum efficiency* of a silicon SPAD device, $\eta_{eq} = \text{PDP } \eta$, also indicated in literature as PDE (photon detection efficiency) or even sometimes confused with PDP.

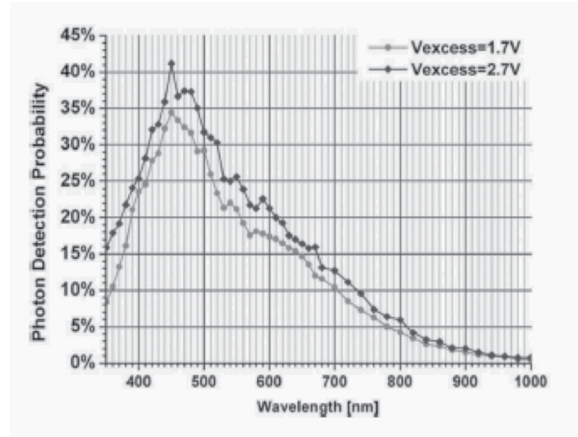


Figure 6-20 Typical spectral response of a silicon SPAD fabricated in 130-nm CMOS technology (from [14], by courtesy of the IEEE).

Two features are worth to point out: (i) curves are significantly narrower than the response of a plain Si *pn*-photodiode (Fig.5-6) and this is due to the thinner absorption layer of the SPAD, and (ii) efficiency η_{eq} increases with the applied overdrive voltage ΔV respect to V_{∞} , because of the PDP increase (see Fig.6-15), and it is therefore advisable to work at the maximum ΔV compatible with the dark current, that also increases with ΔV (see below).

The device described in Fig.6-20 is a SPAD with 10- μm diameter active aperture, has a $V_{\infty}=9.7$ V and can be biased at $V_{bb}=10\ldots12$ V supplying an output pulse, across a load resistance $R_L=50$ Ω , of 0.3–2.3-V peak amplitude.

About *DCR* (dark count rate), Fig.6-21 shows the dependence on temperature and on overdrive ΔV or excess bias voltage. As we can see, 1-V increase of ΔV corresponds to in-

crease the dark count rate by one decade, from 10^5 to 10^6 Hz (or counts per second); with temperature, the increase is a factor of 10 in 75°C .

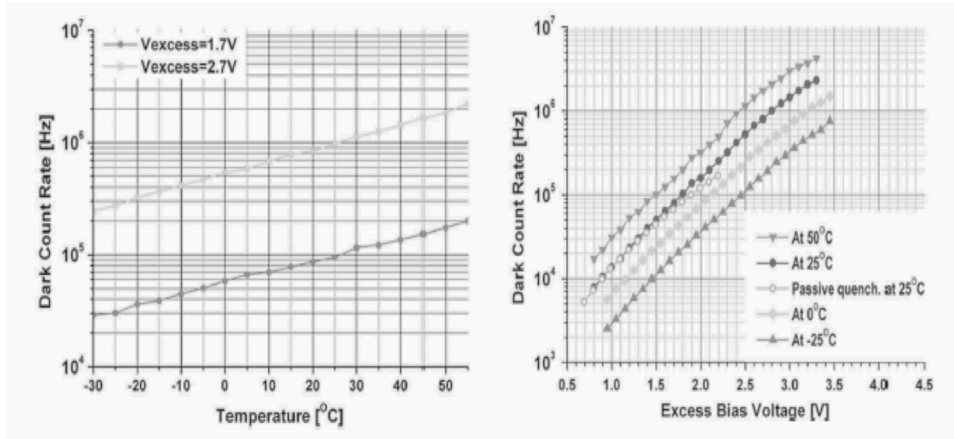


Figure 6-21 Dark count rate of a $10\text{-}\mu\text{m}$ dia. silicon SPAD fabricated in 130-nm CMOS technology as a function of temperature (left) and of over-drive voltage (or excess bias) ΔV (from [14], by courtesy of the IEEE).

The *time resolution* of the SPAD is shown in Fig.6-22. The distribution of time delay of the output pulse in response to a short (40-ps) optical pulse from a mode-locked laser was measured as 144 ps FWHM.

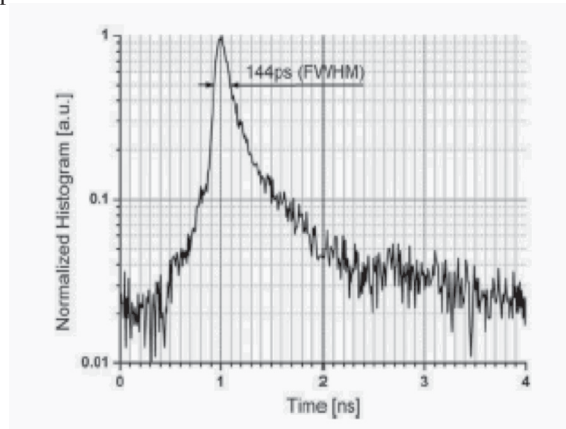


Figure 6-22 Time resolution of silicon SPAD fabricated in 130-nm CMOS technology. The full width at half maximum of the jitter distribution is 144 ps (from [14], by courtesy of the IEEE).

Assuming a Gaussian distribution, this amounts to a rms timing error of $\sigma_t = 144/2.36 = 61$ ps, very attractive for high resolution time-of-flight applications. About dependence of the time resolution from overdrive voltage ΔV , it is found that σ_t is inversely proportional to ΔV for small overdrive, and then saturates to a final value that can be 5–10 times less than for small ΔV . For example, a thin junction SPAD may have 120-ps FWHM at $\Delta V=1$ -V and 35 ps at 5-V. Similarly, for a thick junction we may have 350 ps at 10-V overdrive and 160 ps at 25-V [14]. Temperature dependence of σ_t is modest. Last, *afterpulses* due to deep levels in the depleted region of the SPAD that trap carriers and release them with a statistically distributed delay are disturbing in all applications because they cause the retriggering of the device, correlated to the last detected event. Apart from using a low-trap material, afterpulses can be somehow controlled avoiding large overvoltage and excessive on-state current. Also, active quenching can help to reduce afterpulses by hold-off, as we may empty trapping state by keeping $V_{ak} < V_\infty$ during quenching before recharging V_{ak} to V_{bb} .

6.3 SILICON PHOTOMULTIPLIERS (SiPM)

As we have pointed out in the last Section, the SPAD has two serious problems preventing to emulate the PMT performance; (i) the dead-time characteristic of the Geiger mode of operation, impeding to detect 1,2,...,N photons simultaneously, and (ii) the tiny active area, insufficient for most applications.

Both problems are solved by the clever idea introduced by Saveliev and Golovin [17] who proposed to segment the sensor receiving area (say 2-mm²) in a bidimensional X-Y array containing a large number M of individual SPADs (for example, M=1500 using 50×50 μm² individual SPADs, each representing a pixel).

The SPAD outputs currents are summed up into a single load R_L but each of the SPADs has its own separate quenching circuit (or resistance), integrated around the sensitive area as shown in Fig.6-23 [17].

Now, let us consider incoming photons that constitute the optical signal to be detected. Inside the optical beam hitting the detector area, because of the statistical nature of the optical field, each photon (Fig.6-23) arrives randomly at a different position of the X,Y array, independent from previous and next photon.

Therefore, the rate at which one particular SPAD of the array is hit by another photon is reduced by an average factor of M. If we take M larger than the product $T_R f_{ph}$ of recovery time and photon rate, each pixel has enough time to recover before the arrival of next photon.

Thus, *dead-time* effects are nicely *cleared*, and the response to a packet of m photons hitting the sensitive area at the same time is a peak of output current of amplitude 1, 2,..., m times the single photon amplitude (Fig.6-24). So, the SiPM has now a linear current-to-photon response like the PMT.

Only when the rate of arriving photons is very large, $f_{ph} > M/T_R$ there will be a pile-up of photons and some triggering events will be lost. Typically, this happens for $f_{ph} > 1000/1$ μs

or for ≈ 1 nW of incoming power, and this is the level of maximum input power or device saturation. As most developed and commercialized arrayed SPAD use silicon as a material, the device has been named *SiPM* (and sometimes SSPM, *solid state photomultiplier*).

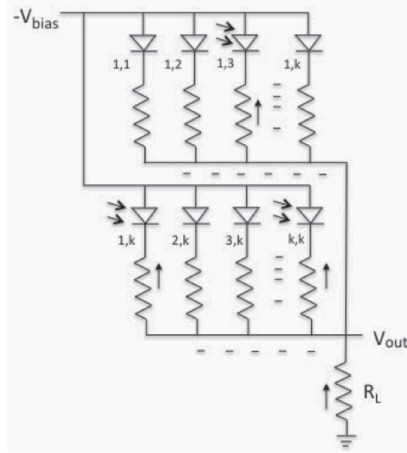


Figure 6-23 The SiPM arrays a large number of individual SPADs, each with its quenching resistance. Output currents are summed up in a common load R_L . If three photons are received simultaneously, because of statistics, they will hit separate pixels and trigger different SPADs, thus injecting into load R_L a current three times the single-pixel current (from [18]).

Different from the PMT, and better from the statistical point of view, the charge variance of the SPAD response is small, and individual packets of 1, 2, ..., R electrons are well resolved even when the incoming photon number is randomly distributed (Fig.6-24) [18].

Time Measurements. In the SiPM, transit time and avalanche build-up time are quite fast, especially for devices with a shallow junction, like in a normal SPAD, and typically the SER jitter is in the range of $\tau=100\text{--}200$ ps FWHM, corresponding to a $\tau=42\text{--}84$ ps rms error [17]. In applications like correlation and time-of-flight, we are interested in the variance σ_t^2 of the timing signal extracted from the photodetected waveform [19]. Usually, timing is performed with the aid of a threshold crossing circuit, giving an output pulse when a threshold I_0 is exceeded, $i(t) > I_0$, or by a centroid sensing circuit, one calculating the center of mass of the $i(t)$ waveform.

For the *centroid timing*, the time variance is found as [18,19]:

$$\sigma_t^2 = [\tau_0^2 + \varepsilon_A^2 (\tau_{\text{SER}}/2.36n)^2] / R \quad (6.20)$$

where:

- τ_0 is the jitter of the SER response, assumed of rigid shape (Sect.4.5.2), a value usually in the range 100–200 ps for PMT, while for the SiPM it is the jitter of the avalanche buildup, usually a fraction (0.2 to 0.4) of the 120-ps (typ.) FWHM, or 24 to 48 ps;
- ε_A^2 is the variance of the multiplier for PMT (Sect.4.5.1) ≈ 0.3 typ. and ≈ 0.05 –0.1 for SiPM, and n the number of dynodes (formally $n=1$ for the SiPM);
- τ_{SER} is the FWHM (full-width-half-maximum) of the SER, typically 2–5 ns in PMT and 0.5–2 ns in SiPM, depending on output circuit parasitic capacitance and quench circuit;

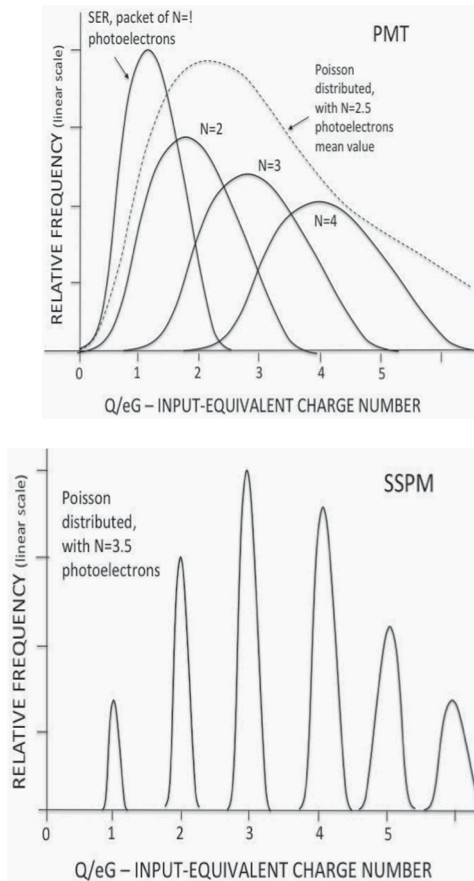


Figure 6-24 (top) in a PMT, the output charge for 1, 2,...,R photoelectrons has a rather large spread, due to the large multiplier variance ε_A^2 (Sect.4.5.1), and for a Poisson-distributed packet of photons we get a broad curve (dotted line); (bottom) in the SSPM, the charge variance is smaller (0.05 to 0.1) and we can resolve packets of N=1,2,...,R single photoelectrons even when N is small and Poisson distributed (from [18] by courtesy of IEEE).

- R is the number of detected photoelectrons, given by $R=\eta F$ in terms of the number of photons F and of the quantum efficiency η for the PMT, and $R=\eta F \text{ PDP}$ for the SiPM;
 - factor 2.36 is the FWHM to rms standard deviation ratio of the SER, assumed Gaussian.
- The dependence on $1/R$ of time variance allows us to make measurements well below the individual contributions τ_0 and τ_{SER} . Typically, with fast circuits, we can go down to a few ps for σ_t in both PMT [19] and SiPM [17], though requiring a larger photoelectron number for the former [i.e., we may have R of the order of $\approx 10^3$ (PMT) and ≈ 50 (SiPM)].

Amplitude measurement applications. In amplitude measurements, the SiPM can read a maximum signal frequency rate up to $\approx M/T_R$, where T_R is the recovery time, while the minimum signal is the dark count rate DCR; so, we have a dynamic range $\approx M/T_R \text{ DCR}$, of several 10^3 typically. In the PMT, the maximum signal is set by the saturation current of the last dynode (Sect.4.7.4) at typ. $\approx 100 \text{ mA}$, while the single-photon current in high gain operation has a peak $\approx 1 \text{ nA}$. Thus, at high gain ($G=10^8$), the ratio of the two currents sets the dynamic range at about a few 10^2 . But, if we lower the gain by a factor K in the PMT, the dynamic range will increase of K and then we may easily go up to 10^5 . Linearity is good in both devices up to about one decade below the dynamic range limit, i.e. the integral linearity error may be $<1\%$ up to 1 decade below and $<5\%$ half decade below the maximum amplitude set by dynamic range.

Dark current. This is the weak point of SiPM, presently orders of magnitude larger than the pA or fA per cm^2 of PMTs sensitive surfaces. Identification of the physical origin of dark current [17] is an important issue, and has already led to a sizeable improvement, and further progresses are expected in next years. In addition, we can think of shrinking the size of individual pixel to say, by ten times, to $5 \mu\text{m}$, and recover the useful area by means of a $C=100$ lens array concentrator (see Sect.6.4.1), thus obtaining in principle a reduction of the dark current by the concentration factor C (typically up to ≈ 50 – 100).

At any rate, as it is well known, we can extend the minimum detectable signal well below the dark current by using a *photon counting* technique, as seen in Sect.4.8.4, and this is true also for the SiPM.

With photon counting, we may go down to a minimum detectable rate $F_{\min}$ of photoelectrons given by Eq.4.72, that is, $F_{\min} = \sqrt{(2\eta_d I_d / eT) / \eta \eta_p}$.

Letting in this equation $\eta \eta_p = \eta_{\text{eff}}$ and expressing $\eta_d I_d$ in terms of the dark count rate per unit area F_{dark} , as $\eta_d I_d = eA F_{\text{dark}}$ where A is the active area, we may write

$$F_{\min} = \sqrt{(2F_{\text{dark}}A/T)/\eta_{\text{eff}}} \quad (6.21)$$

where η_{eff} is the effective quantum efficiency and T is the integration time.

Now, the typical dark count rate per unit area for an SiPM is typically $F_{\text{dark}} \approx 600 \text{ kHz/mm}^2$ [20], that is, about 10^7 times larger than the 10 Hz/cm^2 of a good PMT (see Figs.4-11 and 4-12). Therefore, in view of Eq.6.21, the minimum measurable photon rate of a SiPM at equal T is ≈ 3000 times larger than for a PMT.

Yet, the SiPM performance may be good enough for many applications, as Eq.5.95 provides, for example for $T=1\text{s}$, $A=1\text{mm}^2$ and $\eta_{\text{eff}}=0.4$, a minimum count rate $F_{\min}=2.7 \cdot 10^3 \text{ c/s}$, corresponding to a power of $\approx 10^{-15} \text{ W}$.

An example of a large area SiPM is shown in Fig.6-25 (left); also shown in Fig.6-25 (right) is a commercially available turnkey module of SPAD array smart pixels.



Figure 6-25 (left) A commercial 8x8 channels SiPM of $3 \times 3 \text{ mm}^2$ channel area with 3584 individual SPADs, $50\text{-}\mu\text{m}$ pitch, and 0.74 fill factor (courtesy of Hamamatsu); (right) a turnkey module with an SPAD array of 64×32 smart pixels comprising SPAD, front-end and processing circuits, with a $9.6 \times 4.8\text{-mm}^2$ sensitive area, and 96-kf/s frame rate (courtesy of Micro Photon Devices, Bolzano).

6.4 SPAD ARRAYS

The SiPM described in the last Section is already an example of SPAD array, as it incorporates a large number of individual sensors each with its own resistance for passive quenching. As active quenching would improve the performance further, it is immediate to think of adding this circuit function around the sensitive area of the SPAD, taking advantage of the integration capability offered by the CMOS technology already in place for the SPAD fabrication, and consuming little space additional to the SPAD.

Even more important and the real breakthrough, we may think of adding some pre-processing circuit to each SPAD element of the array, and organize the readout of each pixel information to be readily accessible to check information of any of the SPAD separately. In this new mode of operation, each pixel or SPAD is no more just a sensor of received power, like the pixels in a photography camera, but is able to process information associated to detected photons, like a measurement instrument connected to a sensor.

This is the innovative *smart pixel* concept, made possible by the union of IC technology (particularly, CMOS) and well-performing, small-size SPADs, is a new approach promising of important new applications and products in the fields of consumer, biomedicine and engineering applications.

A first example of smart pixel SPAD array has been the DSiPM (digital Silicon Photomultiplier) [22], consisting in a 64×32 array of SPADs backed by photon counter, trigger and TDC circuits integrated in a $50 \times 30 \mu\text{m}$ cell along with the 50% fill-factor SPAD.

Another important breakthrough development has been that reported by Stoppa et al. [23], a very general purpose 32×32 SPAD array sensor in 130-nm CMOS, for *time-resolved* measurement of photon arrival times at each pixel.

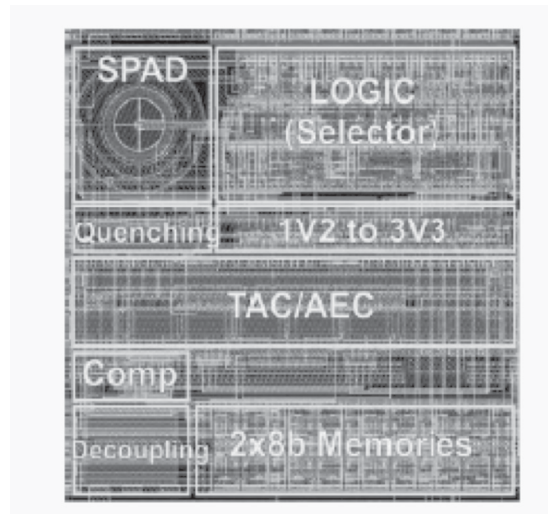


Figure 6-26 Structure of the $50\text{-}\mu\text{m}$ smart pixel of a time-resolved SPAD array fabricated in 130-nm CMOS technology. Along with the $8\text{-}\mu\text{m}$ diameter photodetector SPAD and its active quenching circuit, several circuit functions are integrated (from [22], by courtesy of the IEEE).

This is like saying that, in a $3 \times 3 \text{ mm}^2$ chip, we have integrated 1024 smart pixels, each comprising the full functions of a rangefinder instrument based on Time-of-Flight.

The pixel architecture is shown in Fig.6-26. It is a $50\text{-}\mu\text{m}$ chip on a side, with the sensitive area SPAD occupying an $8\text{-}\mu\text{m}$ diameter area (upper left in Fig.6-26).

The pixel comprises the active quenching circuit, a logic selector, a voltage supply converter, a Time-to-Amplitude-Converter (TAC), a comparator, and 2×8 -bit memory supplied by the TAC for the time delay. Some details of the functions are provided in Fig.6-27.

The complete SPAD array is shown in Fig.6-28. It is a $10 \times 13 \text{ mm}^2$ chip and embeds, in addition to the 32×32 -pixel array, all the necessary circuits to interface the user, namely programming registers, PLL for reference clock, bandgap voltage generators for supply, row and column decoders and column readout circuits. Throughput rate exceeds 4 Gb/s.

The above mentioned are just two examples of the many applications that will expectedly follow in next years, exploiting the potentialities of smart pixels. Presently, the largest array size fabricated so far is 1-Mega pixel [24].

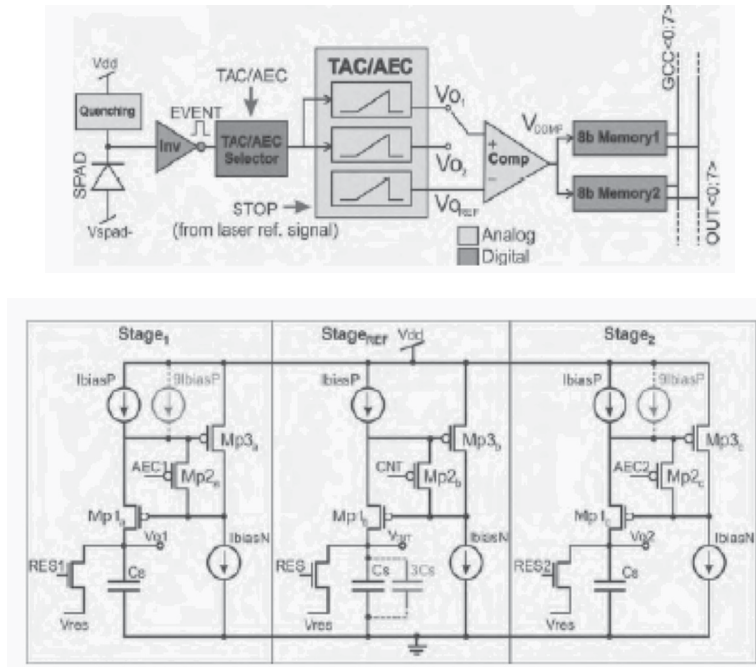


Figure 6-27 The functions realized by the smart pixel: top: block scheme of the pixel, bottom: circuit schematic of the TAC stage (from [22], by courtesy of the IEEE).

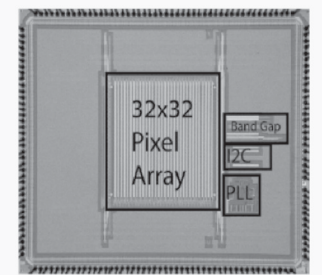


Figure 6-28 The 32x32 pixel SPAD array for time-sorting applications. Chip size is $13 \times 10 \text{ mm}^2$ (from [22], by courtesy of the IEEE).

6.4.1 Microlenses for SPAD Arrays

Till a few years ago, adding circuits for processing the detected signal at the pixel level, next to the SPAD, was considered not practicable, because of the chip area consumed by the

circuits which reduces the area fill-factor FF, and hence the effective sensitivity of the detector. Indeed, if the detector effective efficiency is η_{eff} and FF is the filling factor if the photo-sensitive area occupied by the SPAD, the actual detection efficiency is $\eta_{\text{eff}} \text{FF}$. For example, in the smart pixel of Fig.6-26, the FF is $(\pi/4)8^2/50^2 = 0.02$, so the actual detection efficiency drops to less than 1% from the SPAD top value of $\eta_{\text{eff}} = 40\%$

To limit the loss in detection efficiency, FF should be not so small (e.g. ≥ 0.8), but this is in contrast with integrating significant circuit functions. Nor miniaturization can help so much, because at the present 130 or 90-nm technology, the SPAD has already a shallow junction.

On the other hand, we can take advantage of an invariants of radiometry (Chapter 2), to recover the fill-factor. Recalling that acceptance $\mathcal{A} = A\Omega$ is invariant, we can trade a desired increase of area A for an equal solid-angle Ω decrease, or, we can recover the fill-factor FF with an *optical concentrator*, at the expense of a tolerable reduction of the numerical aperture NA_i (being $\Omega_i = \pi \text{NA}_i^2$) of the objective lens placed in front of the detector (see Fig.6-29) [25]. In details, let us consider the concentrator (Fig. 6-29) as the optical system that collects rays within an aperture of area A_i and solid angle Ω_i , and supplies them at the output in an aperture of area A_o and solid angle Ω_o . Let the photodetector entrance plane coincide with the output plane of the concentrator, and let n_i, n_o be the index of refraction at the input output planes.

Now, we can write the power collected by the detector as: $P_o = E_o A_o$, where E_o is the output irradiance. Introducing the concentration factor C as the ratio of output to input irradiances,

$$C = E_o / E_i \quad (6.22)$$

and substituting in P_o , we see that the collected power $P_o = C E_i A_o$ is increased of a factor C with respect to the placement of the detector at the input plane.

Of course, power P_o supplied at the output of the concentrator cannot be larger than the input power $P_i = E_i A_i$, and thus we have a first limit to the concentration factor C, set by the ratio of input to output areas:

$$C \leq A_i / A_o \quad (6.23)$$

and this is the horizontal asymptote in Fig.6-29 (bottom).

Second, recalling that acceptance $n^2 A \Omega$ is an invariant, we can write:

$$n_i A_i \Omega_i = n_o A_o \Omega_o$$

and, letting $n_i = 1$ for the medium outside the concentrator, and substituting in (6.23), we get the second limit to the concentration factor C, set by the ratio of input to output solid angles:

$$C \leq n_o \Omega_o / \Omega_i \quad (6.23')$$

and this is the asymptote of the roll-off of concentration at too high input numerical aperture, because $\Omega_{i,o} = \pi \text{NA}_{i,o}^2$ whence Eq.(6.23') can also be written as:

$$C \leq (n_o \text{NA}_o / \text{NA}_i)^2 \quad (6.23'')$$

The break point between the two asymptotes is at an objective lens numerical aperture given by $\text{NA}_i^2 = n_o^2 \text{NA}_o^2 / C$.

From Eq. (6.23'') we can see that the performance of the concentrator is worsened by a sub-unitary $NA_0 < 1$ output numerical aperture, whereas is improved by working in ‘immersion’, that is with an index of refraction $n_0 > 1$ for the medium of the output.

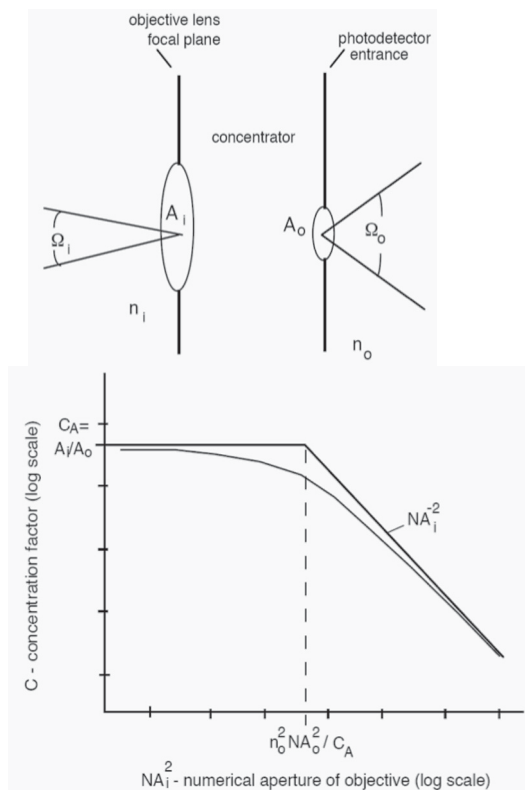


Figure 6-29 (top): light from the objective lens reaches the entrance aperture (of area A_i) under a solid angle Ω_i and concentrator collects it in a smaller aperture of area A_o within the solid angle Ω_o ; (bottom) horizontal asymptote is set by the area ratio, whereas the roll-off as NA_i^{-2} is due to the solid angle ratio, or, to the front objective numerical aperture NA_i becoming too large (from [23]).

This circumstance favors non-imaging concentrators in which NA_o approaches unity, respect to lens-based concentrators in which the lens numerical aperture NA_o is rarely in excess of, say, 0.5.

In conclusion, to recover the fill-factor, we will mount an *array of microconcentrators* in front of the SPAD array detector, as shown in Fig.6-30 [in the case of a truncated-cone concentrator], showing how the input numerical aperture is equal to the objective lens numerical aperture, $NA_i = NA_L$.

Microconcentrators with different geometry have been evaluated by a ray-tracing routine [23] to assess their concentration factor versus NA.

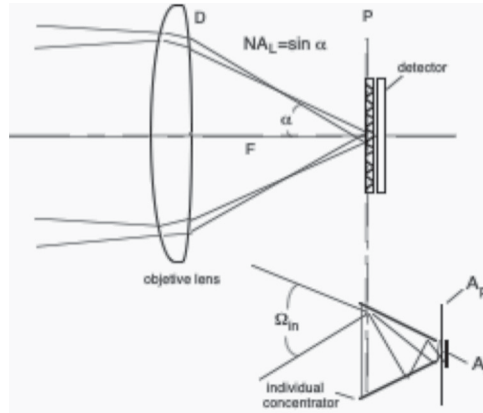


Figure 6-30 An array of microconcentrators can recover the FF of the SPAD array (from [25]).

Three kind of non-imaging concentrators have been considered: truncated-cone (TCC), paraboloid (PPC) and rotated paraboloid (CPC), and an imaging concentrator, a plano-convex lens (see Fig.6-31). As we can see from Fig.6-32 where concentration C is plotted versus NA of the objective lens in front of the device, the most performant geometry is that of a rotated paraboloid, also the most difficult to fabricate in the miniature format required for SPADs.

The plano-convex microlens array concentrator is more limited, yet useful to give a sizeable recovery of the FF. An array of PET-polymer molded microlenses, with 50- μm spacing and 32 $\times$ 32 elements has been fabricated [25] (see Fig. 6-33 left), and used in connection to a 32 $\times$ 32 SPAD array developed in [26].

A concentration factor of $C=30$ has demonstrated at $\text{NA}\approx 0$, and $C=20$ at $\text{NA}=0.1$ (see Fig.6-33, right). With $C=20$, about half of the loss of the $\text{FF}=0.02$ for an 8- μm SPAD in a 50- μm pixel is recovered and the detection efficiency returns to 16%, an acceptable value.

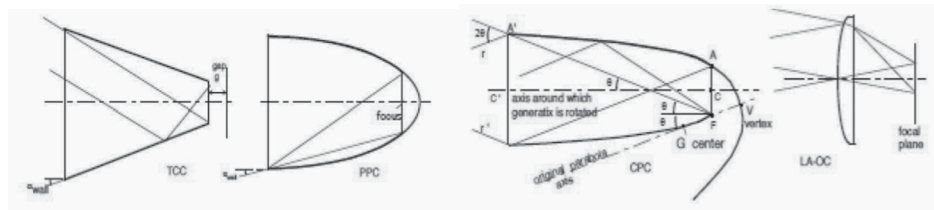


Figure 6-31 Types of microconcentrators: from left to right, truncated-cone, paraboloid, rotated paraboloid, and plano-convex lens (from [23]).

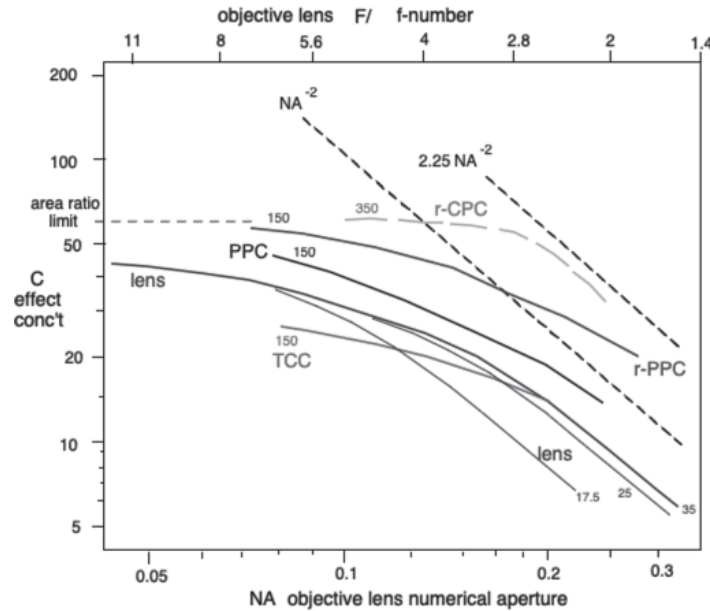


Figure 6-32 Performance of microconcentrators: the effective concentration C is plotted versus numerical aperture NA of the objective lens. Non-imaging type are reflective (prefix r -) or refractive (with $n=1.5$). Concentrator length L is 150 μm for TCC and PPC, and 350 μm for the CPC. Lens OC is plotted for dome heights of 17.5, 25 and 35 μm . All OC have a geometrical area ratio concentration of 60, and an input pixel size (or pitch) of 70 μm (from [25], by courtesy of IEEE).

Problems with the concentrator array technology are: (i) pitches of SPAD array and lens array shall be carefully matched, if they are fabricated in separate steps (or a technology shall be developed to make the lens array atop the SPAD array as the final operation), and (ii) uniformity of concentration shall be improved (was about $\pm 10\%$ in the example above).

6.4.2 Applications of SPAD Arrays

Multipixel capability and with time resolution below nanoseconds are the success factors of SPAD arrays in industrial, automotive and biological applications.

In robotics, even a modest array size (for example, 32×32 pixels) is adequate for 3D vision helping a robot machine to pick and assemble parts of complicated shape. To have the required spatial resolution, let's say 1 cm, we need a timing rms error of $\tau = 2\Delta l/c = 66$ ps, well in the reach of current technology (Fig.6-22).

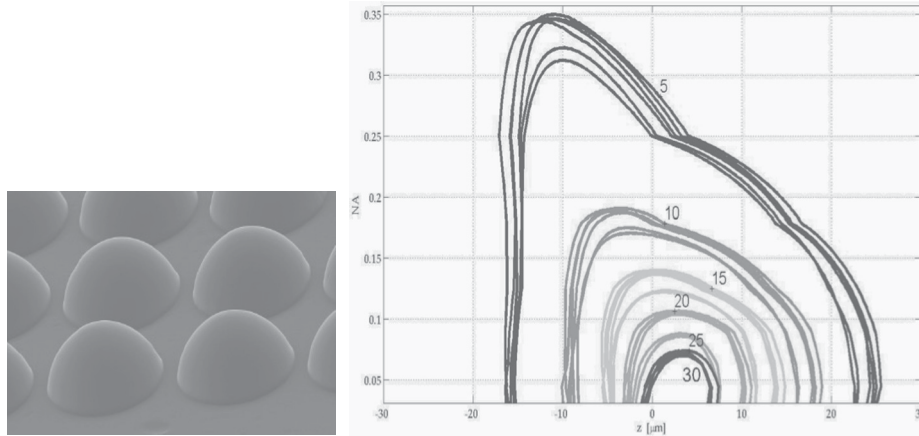


Figure 6-33 (left): microphotograph of plano-convex lens of the 32x32-element concentrator array, 50- μm pitch, 46- μm diameter, fabricated by polymer molding; (right): plot of iso-concentration lines versus distance along optical axis of the lens. The different curves indicate the spread of individual lenses in the array (from [25]).

Fig.6-34 left illustrates the capability of a 32x32 SPAD array, of the type shown in Figs.6-27 and 6.28, to pick up a 3D image on a scale of distance of some ten centimeter and size of several centimeter.

Another very important application of the time-sorting SPAD array is that of car automatic drive, or LIDAR system, where the 3D chip works as a collision avoidance aid. Here, the resolution specification is more relaxed, typ. 10 cm, while instead the range to be covered is up to 100 m, and in this case a high total detection efficiency η_{eff} is necessary, calling for a high PDP and a thick absorption layer.

In biology, the SPAD sub-ns time resolution is the key to a number of fluorescence spectroscopy applications [27].

The most developed is FLIM (fluorescence lifetime imaging microscopy) consisting in looking at the fluorescence emission of a sample illuminated by a short pulse source, for example, a Ti:sapphire mode-locked laser, (Fig.6-34, right) inducing a two-photon absorption.

The map of fluorescence decay time, at the image level, supplies important biological information and is the base for new diagnostic methods.

Also of importance are a few variants of the basic FLIM: FRET (Forster fluorescence resonance energy transfer), to probe the interaction of a donor and an acceptor molecule [28], FCS (fluorescence correlation spectroscopy) to measure concentration and diffusion constants of fluorophores, and STED (stimulated emission depletion) microscopy to create super-resolution images through the selective deactivation of fluorophores [27].

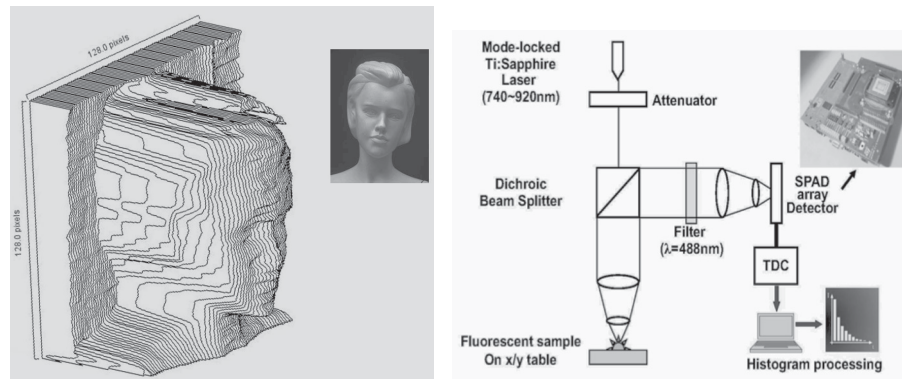


Figure 6-34 (left): example of 3D image pickup made by a SPAD array with a semiconductor mode-locked illuminator: resolution is 1-mm on 100 frames at 1 Hz (from [29], by courtesy of IEEE); (right): the layout of a FLIM measurement, using a Ti:sapphire illuminator to induce two photon absorption in the sample, whose time-resolved fluorescence, at image format, is looked at by the SPAD array detector (from [26]).

References

- [1] R.J. McIntyre: “*Multiplication Noise in Uniform Avalanche Junctions*”, IEEE Trans. Electr. Dev., vol.13 (1966), pp.164-168.
- [2] S. Donati, V. Svelto: “*Statistical Behavior of the Avalanche Photodiode*”, Alta Frequenza, vol.37 (1968), pp.476-486.
- [3] R. J. McIntyre: “*A new look at Impact Ionization-Part I: A Theory of Gain, Noise, Break-down Probability, and Frequency Response*”, IEEE Trans. Electr. Dev., vol. 46 (1999), pp. 1623-1631.
- [4] M.M. Hayat, Ü. Sakoglu, O.-H. Kwon, S. Wang, J. C. Campbell, B.E. A. Saleh, M. C. Teich: “*Breakdown Probabilities for Thin Heterostructure Avalanche Photodiodes*”, IEEE Journal Quant. Electr., vol.39 (2003), pp.179-185.
- [5] P. Yuan, K. A. Anselm, C. Hu, H. Nie, C. Lenox, A. L. Holmes, B. G. Streetman, J. C. Campbell, and R. J. McIntyre: “*A New Look at Impact Ionization—Part II: Gain and Noise in Short APDs*”, IEEE Trans. Electr. Dev., vol.46 (1999), pp.1632-1639.
- [6] B. K. Ng, J. P. R. David, R. C. Tozer, M. Hopkinson, G. Hill, G. J. Rees: “*Excess Noise Characteristics of $\text{Al}_{0.8}\text{Ga}_{0.2}\text{As}$ Avalanche Photodiodes*”, IEEE Phot. Techn. Lett., vol.14 (2002), pp.522-524.
- [7] J. Beck, M. Kinch, R. Scritchfield, M. Woodall, M. Ohlson, L. Wood, P. Mitra, J. Robinson: “*Properties of the HgCdTe Linear-Mode APD*”, Proc. 23rd Annual Meet. IEEE Phot. Soc., Denver, Nov.7-11 2010, paper TuC1.
- [8] F. Capasso: “*Multilayer Avalanche Photodiodes and Solid-State Photomultipliers*”, Laser Focus vol.20 (July 1984), pp.84-101; see also: “*Resonant Tunneling through Double-*

Barriers, Perpendicular Quantum Transport Phenomena in Superlattices and their Device Applications", IEEE J. Quant. Electr., vol.22 (1986), pp.1853-1869.

- [9] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, S. Cova: "Progress in Silicon Single-Photon Avalanche Diodes", IEEE Sel. Top. Quant. Electr. vol.13 (2007), pp.852-861.
- [10] E. Charbon, S. Donati: "SPAD Sensors Come of Age", OSA Opt. Phot. News, vol.21, No.2, (2010), pp.34-41; see also "Evolving CMOS Technology for High-Performance Single-Photon Detection", OSA Opt. Phot. News, vol.22, No.5, (2011), pp.14-15.
- [11] R.H. Haitz: "Mechanisms Contributing to the Noise Pulse Rate of Avalanche Diodes" J. Appl. Phys. vol.36 (1965) pp.3123-3131.
- [12] D. Spinelli A. Ghioni, S.D. Cova, L. M. Davis: "Avalanche Detector with Ultraclean Response for Time-Resolved Photon Counting", IEEE J. Quant. Electr., vol.34 (1998), pp.817-821.
- [13] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, N. Gisin: "Single Photon Detector Fabricated in a Complementary Metal-Oxide-Semiconductor High-Voltage Technology", Rev. Sci. Instr., vol.74 (2003), pp.3263-3270.
- [14] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology" IEEE Sel. Top. Quant. Electr. vol.13, (2007). pp.863-869
- [15] M.A. Karami, M. Gersbach, H.-J. Yoon, and E. Charbon: "A New Single-Photon Avalanche Diode in 90nm Standard CMOS Technology", Opt. Expr. vol.18 (2010), pp.22158-22166.
- [16] S. Cova, M. Ghioni, A. Lacaita, C. Samori, F. Zappa: "Avalanche Photodiodes and Quenching Circuits for Single-Photon Detection", Appl. Opt. vol.35 (1996), pp.1956-1976.
- [17] V. Saveliev, V. Golovin: "Silicon Avalanche Photodiodes on the Base of Metal-Resistor-Semiconductor (MRS) Structures", Nucl. Instr. and Meth., vol.A422, (2000), pp.223-229, see also "The Recent Development and Study of Silicon Photomultiplier", Nucl. Instr. and Meth. vol.535 (2004), pp.528-532.
- [18] S. Donati, T. Tambosso: "Single-photon Detectors: from Traditional PMT to Solid-state SPAD-based Technology", IEEE J. Select. Topics Quant. El., vol.20 (2014), DOI: JST-QE.2014.2350836
- [19] S. Donati: "Electrooptical Instrumentation", Prentice Hall, Upper Saddle River, N.J., 2004, Chapter 3.
- [20] A. G. Stewart, V. Saveliev, S. J. Bellis, D. J. Herbert, P. J. Hughes, J. C. Jackson: "Performance of 1sq-mm Silicon PMT", IEEE J. Quant. Electr. vol.44 (2008), pp.157-164.
- [21] T. Frach, G. Prescher, C. Degenhart, R. de Gruiter, A. Scmitz, R. Ballizany: "The Digital Silicon Photomultiplier - Principle of Operation and intrinsic Detector Performance", 2009 IEEE Nuclear Sc. Symp. Conf. Record, paper N28-5, see also: A. Rochas, et al.: "First Fully Integrated 2D-Array of Single Photon Detectors in Standard CMOS Technology" IEEE Photon. Tech. Lett., vol.15 (2003), pp.963-965.
- [22] D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R.K. Henderson, M. Gersbach, E. Charbon: "A 32x32-pixel Array with In-pixel Photon Counting and Arrival Time Measurement in the Analog Domain", Proc. ESSCIRC, 2009, pp. 204-207.
- [23] S. Donati, G. Martini, M. Norgia: "Microconcentrators to Recover Fill-factor in Image Photodetectors with Pixel On-board Processing Circuits", Optics Expr., vol.15 (2007), pp. 18066-18074.

- [24] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, E. Charbon: “Megapixel Time-gated SPAD Image Sensor for 2D and 3D Imaging Applications”, *Optica*, vol.7 (2020), pp.346-354.
- [25] S. Donati, G. Martini, E. Randone: “Improving Photodetector Performance by means of Micro-Optics Concentrators”, *IEEE J. Lightw. Techn.*, vol.29 (2011), pp.661-665.
- [26] Results of the EU Megaframe program, see <http://www.megaframe.eu/>
- [27] C. Bruschini, H. Homulle, I.M. Antolovic, et al.: “Single-photon Avalanche Diode Imagers in Biophotonics: Rreview and Outlook”, *Light Sci. Appl.*, vol. 8, (2019), doi.org/10.1038/s41377-019-0191-5.
- [28] S.P.Poland, R.K.Henderson et al.: “Time-resolved Multiphocal Multiphoton Microscope for High Speed FRET Imaging In vivo”, *Opt. Lett.* vol.39 (2014), pp.6013-6016.
- [29] C. Niclass, C. Favi, T. Kluter, M. Gersbach, E. Charbon: “A 128×128 Single-Photon Image Sensor with Column-Level 10-Bit Time-to-Digital Converter Array”, *IEEE J. Solid State Circuits*, vol.43 (2008), pp.2977-2989.

Problems and Questions

P 6-1 Design the width and voltage across the multiplication region of a Si APD with maximum attainable gain. For a reach-through structure, complete the APD design adding a consideration on the other doped regions.

P 6-2 Calculate the minimum number of photons that can be detected in a short illuminating pulse at a S/N ratio =3. Consider the case of (i) a pin photodiode; (ii) an APD photodiode.

P 6-3 A SPAD is working in the circuit of Fig.6-14 is biased at $V_{bb}=189$ V and has an infinity gain $V_{\infty}=185$ V. Following the avalanche triggered by a photon, what is the amplitude of the output pulse delivered across the load resistance $R_L=100\ \Omega$?

P 6-4 The signal indicating the switch-on of a SPAD can be taken as the cathode-to-anode voltage or the current delivered on a series resistance, as in Fig.6-18. Which of the two is preferable and why?

P 6-5 What is the dynamic range of the digital signal generated by the SPAD in the circuit of Fig.6-14 used with $C=10$ pF, $R_L=100\ \Omega$, and $R_b=100$ k Ω , when the dark count rate is 10 kHz?

P 6-6 How large are the amplitudes of the output pulse in the circuits of Fig.6-18?

P 6-8 For a time interval measurement, which SPAD is preferable, one with a time resolution $\tau=200$ ps FWHM and effective efficiency $\eta_{eff}=0.4$ or another with $\tau=100$ ps FWHM and $\eta_{eff}=0.1$?

P 6-9 For a SiPM made by M individual SPADs with the same data as in Problem 6-5, what is the dynamic range of amplitude measurement? What is then the real advantage of SiPM over the SPAD?

P 6-10 A SPAD array has a $64\text{-}\mu\text{m}$ pixel size and a sensitive (SPAD) size $w=8\ \mu\text{m}$. We want to recover the fill factor with an array of microlenses, plano-convex with $F=64\ \mu\text{m}$, $D=64\ \mu\text{m}$. Calculate the concentration factor we can obtain when the objective lens used in front of the array has a $NA_o=0.2$, and the NA_o necessary to obtain the concentration given by the area ratio.