PHOTON-COUNTING APDs
A PRIMER
August 9, 2005

This document discusses avalanche photodiodes (APDs) in the context of photon-counting systems. It includes an overview of the basic photon counting problem, a review of the relevant noise theory, and an in-depth discussion of APD design.
PHOTON-COUNTING APDs

A Primer

THE TASK OF COUNTING PHOTONS

Photons are quanta of light – they are discrete unit excitations of the electromagnetic field. Although photons are often described as “particles” of light, one should not imagine them as little round balls or geometric points. As with all denizens of the quantum world, photons have no native shape – their distribution in space is determined by their environment and interactions. To the extent that photons have a shape, it is the shape of the optical “mode” they occupy. The concept of a photon occupying an optical mode that is a solution of Maxwell’s equations is very similar to the idea of an electron occupying an orbital in an atom or molecule that is a solution of Schrödinger’s equation (one distinction being that photons carry a spin of 1, so an indefinite number can occupy the same state; this is in contrast to electrons which obey Pauli’s exclusion principle).

From the standpoint of optical measurement and sensing, a single photon’s worth is the smallest amount of energy that can be extracted from the electromagnetic field; hence, a single photon is the smallest amount of light that can be detected by absorption. Of course, even relatively strong optical signals can be decomposed into a stream of “single” photons when viewed on a short time scale. Although the actual absorption of each individual photon is a random event that obeys Poisson’s distribution, a 1 W beam of 580 nm yellow light delivers one photon every $3.4 \times 10^{-19}$ s on average. Of course, technologically speaking, no detector responds on such a short time scale, so the granular nature of strong optical signals manifests itself as shot noise on a continuous signal rather than

---

1 Non-localized radiation and free space photon modes also exist, and are analogous to unbound electron states.

2 We draw a distinction between the smallest amount of light that can be extracted versus the smallest amount of light that exists, because the zero-point energy of the electromagnetic field effectively populates every photon mode with half a photon’s worth of energy. This zero-point energy cannot be extracted, but it is responsible for stimulating the “spontaneous” emission of light from noncoherent sources.
discrete pulsing of the signal itself. Only in the limit of weak signals – 300 pW at 580 nm corresponds to an average of one photon every 1.14 ns – can one speak of a detector “counting” the arrival of individual photons. Extremely weak signals on this order are characteristic of laser radar returns from distant targets, the fluorescence of rarified species in spectroscopic experiments, and data in certain quantum information applications.

Most photon counting schemes for the visible (400 – 720 nm), near infrared (NIR; 720 – 1800 nm), and short-wave infrared (SWIR; 1800 – 2500 nm) bands use the photon’s energy to liberate a mobile electron which is then accelerated in an electric field. The primary photoelectron eventually gains sufficient energy from the applied field to knock additional electrons free from a suitable medium, thereby creating a current pulse large enough to be detected by the circuit that monitors the detector.

In photomultiplier tubes (PMTs) and electron-bombarded charge-coupled devices (EBCCDs), the photon is absorbed in a photocathode material which ejects an electron into an evacuated chamber; a strong electric field accelerates this primary photoelectron through the vacuum and smashes it into a target. In the case of a PMT the target is the first of a series of dynodes which emits a shower of secondary electrons in response to the impact. The secondaries are driven into the next dynode by the electric field, producing an even greater number of tertiary electrons, and so on until a strong current pulse is generated. In an EBCCD the target is the focal plane array (FPA) of a CCD camera, and multiple electron-hole pairs are generated from the energy deposited in the silicon pixel struck by the primary photoelectron; these secondary carriers are then detected as if multiple photons had been received by the CCD pixel. In contrast, photon absorption in an APD produces two primary photocarriers – an electron and a hole – and both their acceleration and multiplication takes place inside the material itself over a distance of between 0.2 and 20 µm, depending upon the design. APDs are thus more compact than EBCCDs (both devices being smaller than PMTs), and – being monolithic – do not require vacuum integration of a separate photocathode and imaging focal plane. More shall be said about the operation of APDs shortly.
PHOTON-COUNTING APDs

A Primer

To summarize, the two major functions of a photon-counting detector are (1) absorption of the incident photon to generate a primary electron (or electron-hole pair), and (2) field-driven multiplication of the photocarrier(s). It is then the task of an external circuit to register the multiplied signal as a detection event. Successful photon counting thus depends upon the ability of the detector to convert the incident photon into a mobile electron, the size of the amplified signal resulting from high field multiplication of the primary photocarrier, the sensitivity of the readout circuit, and any noise characteristic of the system. Detector noise may give rise to spurious detection events as well as conceal legitimate signals; amplifier noise will limit the external circuit’s ability to register weak output from the detector. This is a suitable starting point to introduce figures of merit describing the sensitivity of various detectors and discuss the mathematics of APD photon detection.

Characterizing Photon Counting Detectors

The point of any figure of merit is to provide a uniform basis for comparing similar devices. We begin our discussion with responsivity ($\mathcal{R}$) and the related parameter quantum efficiency ($\eta$), both of which describe a detector’s ability to convert light into electricity. Then, following a review of the mathematics of noise, we will introduce noise equivalent power (NEP) and the related parameters specific detectivity ($D^*$) and noise equivalent input (NEI), which quantify how noise limits the sensitivity of a detector. Finally, having demonstrated that NEP and NEI are not the most natural way to treat noise in photon counting problems, we introduce methods to calculate the photon counting efficiency of an APD directly.

Responsivity and Quantum Efficiency

Responsivity measures the ratio of the detector’s output to its input, usually in units of Amperes per Watt, or

$$\mathcal{R} \equiv \frac{\sqrt{I^2_{\text{photo}}}}{\sqrt{P^2_{\text{signal}}}}.$$

3 The overbars indicate a time average, so this expression means $\mathcal{R}$ is defined in terms of root-mean-square (RMS) photocurrent and signal power. Also, it should be noted that $\mathcal{R}$ is normally a function of signal modulation frequency, as no detector can have an infinitely fast impulse response.
In general, responsivity will vary as a function of wavelength – both because the ability of a detector material to absorb photons varies with wavelength, and because $I_{\text{photo}}$ depends only upon the rate of photon absorption but $P_{\text{signal}}$ is scaled by the photon’s energy. An optical signal that delivers 500 nm photons at the same rate as an identical signal consisting of 1000 nm photons will have twice the power because each 500 nm photon carries twice the energy of a 1000 nm photon. Thus, a detector that converts both types of photons to electrons equally well will nonetheless have half the responsivity at 500 nm that it has at 1000 nm. For this reason, quantum efficiency ($\eta$) defined by

$$\eta \equiv \frac{\text{quanta out (electrons)}}{\text{quanta in (photons)}}$$

is often used in the place of $\Re$, because it allows direct comparison between a detector’s ability to receive photons at different wavelengths. Responsivity is related to quantum efficiency by

$$\Re = \frac{q}{\hbar \nu} \eta$$

where $q$ is the elementary charge in Coulombs, and $\hbar \nu$ – the product of Planck’s constant and the photon frequency – is the photon energy in Joules.

Although $\eta$ is technically defined for the detector as a whole, it is common to quote the pre-multiplication value of $\eta$. There are three reasons for this. First, the multiplication provided by a linear mode APD, PMT, or EBCCD will vary depending upon the bias across the multiplication stage; $\eta$ is a more useful basis for comparison between different detectors if it expresses the ability to deliver primary photocarriers to the multiplication stages of these devices rather than the total quantum yield at a given operating bias point. Second, the pre-multiplication value of $\eta$ is an absolute limit on a detector’s ability to detect single photons – it matters not what the eventual multiplication will be if no primary photocarrier is generated in the first place. Finally, Geiger mode APDs do not have a proportional response, so a Geiger mode APD’s post-multiplication $\eta$ is characteristic of the circuit that
operates the detector rather than the detector itself. Throughout this document, we will refer to the pre-multiplication value of $\eta$.

Quantum efficiency is the product of several probabilities:

$$\eta(\lambda) = T(\lambda) \times A(\lambda) \times \eta_i.$$  

The meaning of each probability is described next.

The power transmission coefficient of the detector’s input window $T(\lambda)$ is the probability that an incident photon will enter the detector. For the special case of normal incidence on a detector characterized by a homogenous refractive index $n_i$, $T(\lambda)$ can be modeled using the Fresnel formula:

$$T(\lambda) = \frac{4 \ n_i(\lambda)}{[n_i(\lambda)+1]^2}. \quad \text{4}$$

In practice, however, anti-reflection (AR) coatings should be employed, and $T(\lambda)$ must be calculated using software that takes internal reflections, losses, and the angle of incidence into account. Empirically, it is reasonable to assume a value of $T(\lambda) \sim 95\%$ at normal incidence for a device with an AR coating designed for optimum transmission at the desired wavelength.

The probability of absorption $A(\lambda)$ expresses the likelihood that a photon will generate a photoelectron (or electron-hole pair) once inside the detector. Two parameters – the absorber material’s absorption coefficient $\alpha(\lambda)$ and the optical path length through the absorber $L$ are a good guide to estimating $A(\lambda)$ through an application of Beer’s law:

$$A(\lambda) \approx 1 - \exp[-\alpha(\lambda) L]. \quad \text{5}$$

A more accurate estimate can be obtained using software that takes reflections into account; indeed, certain waveguide and resonant cavity structures rely upon reflections to dramatically increase $L$ by setting up multiple passes through the detector.

---

4 Here we have assumed the signal is incident from vacuum, for which the refractive index is 1. This is approximately true for air.

5 This formula expresses the power loss of a plane wave propagating a distance $L$ through a homogenous medium characterized by $\alpha(\lambda)$. To first order, one can use the thickness of the detector’s photocathode or absorber for $L$. 
absorber. However, all else being equal, \( \alpha(\lambda) \) is the main parameter that determines \( A(\lambda) \), and the dependence is exponential. Consider that at 1064 nm, \( \alpha_{\text{silicon}} = 13 \text{ cm}^{-1} \) and \( \alpha_{\text{InGaAs}} = 30,000 \text{ cm}^{-1} \) – clearly an InGaAs absorber will receive 1064 nm light much more efficiently than silicon.

The internal quantum efficiency \( \eta_i \) expresses the probability that a photocarrier generated in the absorber will make it to the multiplication stage of the detector. Photocarrier collection in an APD with a fully depleted absorber is quite efficient, and \( \eta_i \) approaches unity. In contrast, photoelectron collection from the photocathode in a PMT or EBCCD is a major limiting factor. Photoelectrons extracted from a photocathode are emitted over the Schottky barrier of a metallized contact; bias applied to this contact can lower the barrier to emission, but doing so also increases the emission of “dark current” carriers. Thus, although \( T(\lambda) \) and \( A(\lambda) \) are generally similar among PMTs, EBCCDs, and APDs, it is \( \eta_i \) that differentiates the photocathode devices from the APD. Practical operation of EBCCDs with NIR-sensitive photocathodes is normally limited to \( \eta \sim 30\% \), whereas NIR APDs routinely achieve \( \eta > 80\% \).

The Mathematics of Detector Noise

A plane light wave is fully specified if one knows its wave vector\(^6\), polarization, amplitude, and phase. Although the light pulses encountered in an application like laser radar come in the form of wave packets rather than pure plane waves, by their nature, wave packets can be decomposed into a summation of plane waves. Thus, these properties of plane waves are a convenient starting point for our discussion of noise.

Information can in principle be encoded in any of light’s properties, but in practice optical power – derived from the wave’s amplitude through its Poynting

\(^6\) The wavevector \( \vec{k} \) is oriented in the direction of propagation, and its magnitude is related to wavelength by \( |\vec{k}| = \frac{2\pi}{\lambda} \); the frequency of the light wave is related to its wave vector through its dispersion relation.
PHOTON-COUNTING APDs

A Primer

vector⁷ – is most commonly used. Optical power is more convenient to measure than field amplitude or any of the other plane wave characteristics as it is manifest directly in the photogeneration rate inside the detector. Optical power is also easier to modulate at a transmitter. However, as was previously mentioned, the power of an optical signal is really a surrogate for the time-averaged arrival rate of photons. Outside the regime of nonlinear optics, photon absorption events are independent and random – they obey Poisson statistics, which means that uniform illumination striking a detector doesn’t give rise to a sequence of absorption events that is uniformly spaced in time. Instead, if the number of absorption events averaged across an ensemble of detectors during a unit time interval is \( \langle n \rangle \), the probability \( P(n) \) of \( n \) absorption events occurring in any specific detector during such a span is:

\[
P(n) = \frac{\langle n \rangle^n \exp(-\langle n \rangle)}{n!}.
\]

Thus, the instantaneous rate of absorption in a single detector under uniform illumination will fluctuate randomly from moment to moment, even though the average rate is determined by the incident optical power level.⁸ This irregularity in the timing of photon absorption events constitutes a type of “shot” noise.

The photogeneration rate inside a photodiode is translated more-or-less directly into a photocurrent through its influence upon the minority carrier concentration in the diode’s depletion region. The photon shot noise – and its statistical distribution – is preserved. However, photogeneration is not the only thing affecting the current in the photodiode. Other important processes active in the diode which cause current fluctuations include “dark” current leakage and the

⁷ The Poynting vector \( \vec{S} = \frac{1}{\mu_0} \vec{E} \times \vec{B} \) is oriented in the direction of propagation, and its magnitude

\[
|\vec{S}| = \frac{1}{\mu_0 c} |\vec{E}|^2
\]

– which is proportional to the square of the wave’s amplitude – is the power flux delivered by the wave in units of Watts per square meter. We defer explicit discussion of the detector’s collection area for now, and consider power rather than power flux.

⁸ The assumption of ergodicity – that the distribution of a single fluctuating random variable over a large sample of time is the same as the distribution of an ensemble of identically-prepared random variables examined at a single instant – is implicit in this and subsequent applications of statistics in this discussion.
PHOTON-COUNTING APDs

A Primer

thermal motion of the carriers. In an APD, the random variation in the amount of avalanche multiplication experienced by individual carriers adds an additional source of noise. As will be established shortly, contributions to the total noise current add in quadrature, so long as their disparate sources are uncorrelated. The statistics of independent noise sources can therefore be analyzed separately and combined in a later calculation; we will proceed with our analysis of photon shot noise to illustrate some general features of noise statistics, and return to this point momentarily.

The information which originally arrived as an optical power level and which is subsequently conveyed by the photocurrent resides in its time average:

\[ \text{signal} = \frac{1}{\tau} \int_{t_0}^{t_0+\tau} dt \ I_{ph}(t) = \overline{I_{ph}} \equiv I_{signal}. \]

The “noise” in the photocurrent is related to the fluctuations of its instantaneous value \( I_{ph}(t) \) above or below the signal (the photocurrent’s average value \( \overline{I_{ph}} \)):

\[ \Delta I_{ph}(t) = I_{ph}(t) - \overline{I_{ph}}. \]

This instantaneous fluctuating quantity is not particularly useful because of its time-dependent and random nature. In order to characterize the size of these fluctuations with a single parameter, one might think to calculate the time average of \( \Delta I_{ph}(t) \); however, with a little thought, it should be apparent that

\[ \overline{\Delta I_{ph}} = \overline{I_{ph}} - \overline{I_{ph}} = 0. \]

In such cases that the first moment (or mean) of a distribution is zero, the distribution must be characterized by its second moment:\[10\]

\[ I_{noise} = \sqrt{\Delta I_{ph}^2} \text{defined as the square root of the noise is the standard deviation of } I_{ph}(t). \]

---

\[9\] This is the “first moment” of the photocurrent distribution – its expectation value. Technically, the integral should look like \( \overline{I_{ph}} = \int_{-\infty}^{\infty} dt \left[ I_{ph}(t) \times f(t) \right] \) where \( f(t) \) is a normalized distribution function. The form given above is appropriate for \( \overline{I_{ph}} \) and similar quantities when defined over a finite time span \( \tau \).

\[10\] The quantity given here is the second moment of \( \Delta I_{ph}(t) \); it is the “variance” or “second central moment” of \( I_{ph}(t) \). The noise current \( I_{noise} \) defined as the square root of the noise is the standard deviation of \( I_{ph}(t) \).
Having defined the noise of a fluctuating random variable as its variance, we can return to the question of the addition of noise sources. Consider a total current measured at the terminals of a detector which consists of several independent fluctuating components, each characterized by a mean and a variance:

\[ I_{\text{total}}(t) = \sum_i I_i(t). \]

The aim is to find the variance of \( I_{\text{total}}(t) \) in terms of the variances of its components. Conveniently, the variances of independent random variables are additive. To see this, one first applies the rule for integrating sums to show that the means are additive:

\[
\text{signal} = I_{\text{total}} = \frac{1}{t} \int_0^{t=\tau} dt \ I_{\text{total}}(t) = \frac{1}{t} \int_0^{t=\tau} dt \ \left[ \sum_i I_i(t) \right] = \sum_i \left[ \frac{1}{t} \int_0^{t=\tau} dt \ I_i(t) \right] = \sum_i \bar{I}_i.
\]

The additivity of the variances will follow so long as \( \bar{I}_i(\tau) \times \bar{I}_j(\tau) = \bar{I}_i \times \bar{I}_j \) for all \( i \neq j \); this is essentially the definition of statistical independence.\(^{11}\) The desired relationship between the variance of \( I_{\text{total}} \) and the individual variances of its components can be demonstrated explicitly:

\[
\text{noise} = \text{var}(I_{\text{total}}) = \left[ I_{\text{total}}(t) - \bar{I}_{\text{total}} \right]^2 = \left[ \sum_i I_i(t) - \sum_i \bar{I}_i \right]^2 = \left[ \sum_i \left[ I_i(t) - \bar{I}_i \right] \right]^2 = \sum_i \left[ I_i(t) - \bar{I}_i \right]^2 + \sum_{i,j} I_i(\tau) \times I_j(\tau).
\]

\(^{11}\) A formal statement of statistical independence says that the joint distribution function for two random variables can be factored into the individual distribution functions. In other words, the likelihood of measuring \( P(t) = I_i(t) \times I_j(t) \) and getting the result \( P(t_0) = I_i \times I_j \) is equal to the product of the probability of measuring \( I_i(t_0) \) individually and getting \( I_i \), and the probability of measuring \( I_j(t_0) \) individually and getting \( I_j \).
Thus, as promised, we have shown that independent noise sources can be treated separately, and their individual noise contributions can be added to find the total noise.

A little unit analysis and discussion is in order here. The expressions given above for the signal are in units of current; the expressions for noise have the units of current squared. In order to construct a dimensionless signal-to-noise ratio (SNR), we must either square the signal to obtain a power ratio\(^{12}\) or take the root of the noise to get a current ratio. In either case, we must keep in mind that we are comparing a first moment to a second moment – these statistical measures do not tell us much of anything unless we know what distribution we’re talking about. This would not be a significant problem if all noise sources obeyed the same well-behaved distribution, with individual trials clumped tightly and symmetrically about the mean. In that case, the mean would be broadly representative of most points in the data set, and the variance (or standard deviation) would be a clear measure of the likelihood of finding outlying data points. Unfortunately, in the case of APDs, two very different distributions come into play – and one of them isn’t easy to work with.

The common sources of electrical noise – shot noise\(^{13}\) and Johnson (or thermal) noise – all either obey or approximate a Gaussian distribution:

\[
P(n) = \frac{1}{\sqrt{2\pi \text{ var}(n)}} \exp\left[ -\frac{(n-\langle n \rangle)^2}{2 \text{ var}(n)} \right].
\]

Similar to the Poisson distribution, the Gaussian distribution has the useful property that \(\text{var}(n) = \langle n \rangle\), and fits the description of a “well-behaved” distribution posited above: the Gaussian distribution is largest at its mean, and it falls off rapidly and symmetrically to either side of \(\langle n \rangle\).

---

\(^{12}\) Assuming a resistance \(R\), \(R^2\) is power; however, we don’t need an explicit resistance, because \(R\) cancels out in a power ratio, and one is left with \(\text{SNR}_{\text{power}} = \frac{I_{\text{signal}}^2}{I_{\text{noise}}^2}\).

\(^{13}\) The Poisson distribution approaches a Gaussian distribution in the limit of a large number of trials.
Avalanche multiplication works somewhat differently. Instead of thinking of the avalanche process as an independent source of carriers (or current), we regard the process as amplifying an existing number of carriers and study the distribution of the total output over an ensemble of trials. McIntyre derived an expression that accurately models the distribution of the multiplied output \( n \) as a function of the input \( a \), the ensemble average of the gain \( M \equiv \langle n / a \rangle \), and the effective ratio of electron and hole ionization coefficients \( k \):

\[
P(n) = \frac{a \times \Gamma\left(\frac{n}{1-k} + 1\right)}{n \times (n-a)! \times \Gamma\left(\frac{k \times n}{1-k} + 1 + a\right)} \times \left(\frac{1+k(M-1)}{M}\right)^{\frac{k a}{1-k}} \times \left(\frac{(1-k)(M-1)}{M}\right)^{n-a}
\]

where the Euler gamma function is defined as

\[
\Gamma(z) = \int_0^\infty dt \, t^{z-1} \exp(-t).
\]

As may be apparent from its form, the McIntyre distribution does not resemble a Gaussian profile at all for small signals. For small \( a \) – the case most relevant to photon counting applications – the McIntyre distribution is very broad, and its maximum doesn’t even coincide with its mean. To help illustrate this point, both distributions are plotted in Figure 1 for an average output signal of \( 10^4 \) electrons. In the case of the McIntyre distribution, various combinations of input and average gain \((a \text{ and } M)\) resulting in \( 10^4 \) are plotted for \( k \) values typical of silicon and III-V APDs. To the extent that the application dictates \( a \), the supreme importance of keeping \( k \) small in order to get a tighter (less noisy) distribution should be apparent.
Although "mean" and "variance" have the same definition for any distribution, they don’t always imply the same thing. For instance, the probability of a trial giving a result within one standard deviation of the mean of a Gaussian distribution is 68.3%, with equal probability to either side of the mean. This property is true for all Gaussian distributions, independent of \( \langle n \rangle \). The same probability for the McIntyre distribution varies depending upon \( a, M, \) and \( k \) – and tends to be weighted toward values of \( n \) smaller than the mean. For instance, the probability of finding \( n \) within one standard deviation of \( \langle n \rangle = M \times a \) is 68.7% for \( a = 1000, M = 10, \) and \( k = 0.02 \) (the McIntyre distribution in Figure 1 that most resembles a Gaussian). As is evident from Figure 1, this “Gaussian-like” McIntyre distribution assigns roughly equal probability to values of \( n \) on either side of the mean. On the other hand, for \( a = 10, M = 1000, \) and \( k = 0.02, \) the standard deviation (~14,460) is larger than the...
mean itself (~10,093)\(^{14}\) – and of the 90.4% of the data points within one standard
deviation of the mean, fully 71.7% are smaller than the mean. Clearly, the physical
significance of any particular value of \(I_{\text{noise}}\) depends upon how the current is
distributed: a 1 nA contribution from Gaussian-distributed shot noise does not imply
the same thing as a 1 nA contribution from McIntyre-distributed multiplication noise
except in the limit of large \(a\) and small \(k\).

Apart from their statistical distribution, it is useful to study the spectral
distribution of noise sources. The measurement bandwidth of the electronics
monitoring a detector is always going to be limited, and sometimes the signal itself
is limited to a single harmonic component. For instance, binary telecommunica
tions data can be transmitted by modulating the output of a continuous-wave (CW)
semiconductor laser at a predetermined bit rate. Although the optical power levels
which encode the 1’s and 0’s of the bit stream will be noisy, only the noise within a
narrow frequency band centered on the bit rate will be measured if the receiver
circuit uses a filter to isolate the data. Thus, we are often interested in the noise
across a finite band of frequencies, as opposed to the total noise.

The spectral intensity \(S_{\text{f}}(f)\) gives the component of the noise power at a
frequency \(f\). If one integrate \(S_{\text{f}}(f)\) across all frequencies, the total noise power is
recovered:

\[
\text{noise power} = I_{\text{noise}}^2 = \left( \Delta I \right)^2 = \left[ I(t) - \bar{I} \right]^2 = \int_{0}^{\infty} df \, S_{\text{f}}(f).
\]

It is quite common to find the SNR of a detector expressed as the ratio between a
squared photocurrent and \(S_{\text{f}}(f)\) for a particular frequency. Several assumptions
are implied by this practice: first, that the signal itself is harmonic; second, that the
output of the detector is filtered to pass a 1 Hz band centered on the signal’s
frequency; and third, that the integrated noise across that 1 Hz band is adequately
approximated by a rectangle of width \(\Delta f = 1\) Hz and height \(S_{\text{f}}(f)\):

\[\]

\[^{14}\text{These numbers are approximate because they were arrived at numerically rather than analytically. They represent statistics on a 10,000-element data set generated from the McIntyre distribution, spanning possible outcomes from 10-200,000.}\]
We will take this up again after we have concluded our discussion of $S(f)$. 

The spectral intensity can be calculated in several ways. If a large set of $I(t)$ data is available – such as the output of a computer simulation – then $S(f)$ can be computed with the aid of the autocorrelation function $A(\tau)$ by an application of the Wiener-Khintchine theorem:

$$\int_{-\infty}^{\infty} df' \ S'(f') \approx 1 \text{Hz} \times S(f).$$

This is the most general approach, and the one we will take in analyzing our APD designs. However, widely-used spectral intensity theorems have been derived analytically for the major noise sources. In most cases, these theorems assume that the noise has a uniform spectral distribution (i.e., it is “white”) at frequencies much less than the inverse transit time for carriers across the junction. The spectral intensity theorems relevant to the noise processes mentioned previously – shot noise on the signal and dark currents, thermal noise from carrier velocity fluctuations, and multiplication noise – will be introduced next.

A particularly useful extension of Milatz’s theorem derived by van der Ziel says that for an arbitrary rate $r$ for which $\text{var}(r)$ exists, the low-frequency spectral intensity$^{16}$ of $r$ is:

$$S_r(f) = 4 \int_{0}^{\infty} d\tau \ [A(\tau) \times \cos(2\pi f \ \tau)],$$

where $A(\tau) = \int_{-\infty}^{\infty} dt \ [\Delta I(t) \times \Delta I(t + \tau)]$.\footnote{Although the limits of integration are technically required to go to infinity – and $\Delta I(t)$ is supposed to model a “stationary” (time-invariant) random process – the analysis still works for integration over a finite time span $T$ where the resolution of the resulting transform will be $1/T$. Also, in cases where it is impractical to isolate the noise component in the simulated data, a spectral intensity can be computed for the total current, and the signal component can be identified in $S(f)$ from its frequency (e.g. a peak at the modulation frequency for a harmonic signal).}

$^{15}$This is the most general approach, and the one we will take in analyzing our APD designs. However, widely-used spectral intensity theorems have been derived analytically for the major noise sources. In most cases, these theorems assume that the noise has a uniform spectral distribution (i.e., it is “white”) at frequencies much less than the inverse transit time for carriers across the junction. The spectral intensity theorems relevant to the noise processes mentioned previously – shot noise on the signal and dark currents, thermal noise from carrier velocity fluctuations, and multiplication noise – will be introduced next.

A particularly useful extension of Milatz’s theorem derived by van der Ziel says that for an arbitrary rate $r$ for which $\text{var}(r)$ exists, the low-frequency spectral intensity$^{16}$ of $r$ is:

$$S_r(f) = 4 \int_{0}^{\infty} d\tau \ [A(\tau) \times \cos(2\pi f \ \tau)],$$

where $A(\tau) = \int_{-\infty}^{\infty} dt \ [\Delta I(t) \times \Delta I(t + \tau)]$.\footnote{Although the limits of integration are technically required to go to infinity – and $\Delta I(t)$ is supposed to model a “stationary” (time-invariant) random process – the analysis still works for integration over a finite time span $T$ where the resolution of the resulting transform will be $1/T$. Also, in cases where it is impractical to isolate the noise component in the simulated data, a spectral intensity can be computed for the total current, and the signal component can be identified in $S(f)$ from its frequency (e.g. a peak at the modulation frequency for a harmonic signal).}

$^{16}$The spectral intensity is assumed to be flat, with $S_r(f) = S_r(0)$, up until $f$ approaches the inverse junction transit time. Note that although $S_r(\infty)$ is a notation used by convention to designate the low-frequency spectral intensity, it is not literally the spectral intensity of the DC component. The DC component of the detector’s response is actually part of the signal – it either represents the mean photocurrent resulting from CW illumination, or a predictable (and hence subtractable) mean value of dark leakage current.
PHOTON-COUNTING APDs
A Primer

\[ S_r(0) = 2 \var(r). \]

Previously we used \( n \) to represent the number of carriers generated in a unit period of time (a dimensionless quantity); to avoid confusion, we represent the generation rate (with dimensions of 1/time) by \( r \), even though its statistical behavior is the same as \( n \). In particular, if \( r \) obeys the Gaussian distribution, as in unmultiplied shot noise, then \( \var(r) = \langle r \rangle \), and:

\[ S_r(0) = 2 \langle r \rangle. \]

To obtain a spectral intensity for the noise power, one scales by a factor of \( q^2 \), since \( I = q \times r \), \( S_r(0) \) goes as \( r^2 \), and power goes as \( I^2 \):\(^{17}\)

\[ S_{I\text{-shot}} = q^2 \times S_r(0) = 2 q^2 \langle r \rangle = 2 q \langle I \rangle = 2 q I. \]

Nyquist derived the spectral intensity of thermal noise in a resistance \( R \) based upon thermodynamic arguments. He found that:

\[ S_{I\text{-thermal}} = \frac{4 k_B T}{R}. \]

The Burgess variance theorem is used in conjunction with the theorem for shot noise to find the spectral intensity of multiplied shot noise. Burgess showed that the variance of the total carrier output \( n \) that results from a primary carriers and a multiplication factor for individual carriers of \( m_1 \) is:

\[ \var(n) = \langle m_1 \rangle^2 \var(a) + \langle a \rangle \var(m_1). \]

When the process that generates the primary carriers obeys Poisson statistics – as it will for photogeneration and dark current – we can make the substitution \( \var(a) = \langle a \rangle \). Then the Burgess variance theorem looks like:

\[ \var(n) = \langle m_1 \rangle^2 \langle a \rangle \left[ 1 + \frac{\var(m_1)}{\langle m_1 \rangle^2} \right] = \langle m_1 \rangle^2 \langle a \rangle \left[ 1 + \frac{\langle m_1 \rangle^2 - \langle m_1 \rangle^2}{\langle m_1 \rangle^2} \right] = \langle m_1 \rangle^2 \langle a \rangle \left[ \frac{\langle m_1 \rangle^2}{\langle m_1 \rangle^2} \right]. \]

\(^{17}\) This particular result was first derived by Schottky in 1918.
The ratio $\frac{\langle m_i^2 \rangle}{\langle m_i \rangle^2}$ is normally referred to as the excess noise factor $F(M_1, k)$, where $M_1 \equiv \langle m_i \rangle$. McIntyre derived a local theory for $F(M_1, k)$ that applies to single carrier type injection. The theory is "local" in the sense that the ionizing behavior of a carrier is assumed to be a function solely of its immediate environment, independent of its past history; single carrier type injection will occur if the APD’s absorber is physically separated from its multiplication region, so that either electrons or holes – but not both – will be swept by the internal electric field into the avalanche layer. The formula for electrons is:

$$F(M_1, k) = M_1 [1 - (1 - k) \left( \frac{M_i - 1}{M_i} \right)^2].$$

This formula is accurate for APDs with thick, homogenous multiplication layers. However, for thin multiplying structures or those with major variations in internal composition and field strength, the excess noise factor must be calculated directly as the ratio $\frac{\langle m_i^2 \rangle}{\langle m_i \rangle^2}$. We will return to this issue when we discuss the simulation of our low-noise APD designs. However, to conclude this discussion of the spectral intensity theorem for avalanche multiplication noise, we note that by associating the average number of primary carriers injected in a unit time span $\langle a \rangle$ with the primary rate $\langle r_{\text{primary}} \rangle$, plugging in the Burgess variance theorem for $\text{var}(r)$, and using the usual factor of $q^2$ to convert from $S_\lambda(0)$ to $S_r(0)$, we find the following result:

$$S_{r-\text{multiplied shot}} = 2q^2 M_i^2 \langle r_{\text{primary}} \rangle F(M_1, k) = 2q T_{\text{primary}} M_i^2 F(M_1, k).$$

We will use these spectral intensity theorems momentarily when we discuss common measures of detector noise, including SNR, NEP, $D^*$, and NEI. First, however, we will touch on the issue of signal integration to complete our discussion of detector noise mathematics.

Up to this point, we have been examining the noise in the measured photon arrival rate – essentially the photocurrent. True photon counting corresponds to
measuring the charge delivered by the photocurrent over a given integration span. For a single trial, the integrated charge is:\(^{18}\)

\[
Q_i = \int_{t_0}^{t_0 + \tau} dt \, I_i(t) = \bar{I}_i \times \tau .
\]

For an ensemble of trials,

\[
Q_{signal} = \langle Q \rangle = \langle I \rangle \times \tau = I_{signal} \times \tau .
\]

Let \( \tau \) be the unit time for which the average number of carriers is \( \langle n \rangle \). Then,

\[
Q_i = q \times n_i ,
\]

so that

\[
\langle Q \rangle = q \times \langle n \rangle \quad \text{and} \quad \text{var}(Q) = q^2 \ \text{var}(n) . \quad ^{19}
\]

For a Gaussian-distributed process, \( \text{var}(n) = \langle n \rangle \). Substituting

\[
\langle n \rangle = \frac{\langle Q \rangle}{q} = \frac{I_{signal} \times \tau}{q} ,
\]

the charge noise can be written as:

\[
Q_{noise} = \sqrt{\text{var}(Q)} = q \times \sqrt{\langle n \rangle} = \sqrt{q \times I_{signal} \times \tau} .
\]

For a process with multiplication, one uses the Burgess variance theorem to find \( \text{var}(n) \), so that:

\[
\text{var}(Q) = q^2 \ \text{var}(n) = q^2 \ M_1^2 \ \langle a \rangle \ F(M_1, k) .
\]

Recalling that \( \langle n \rangle = M_1 \times \langle a \rangle \) and rewriting \( \langle a \rangle \) in terms of an average primary current:

\[\text{Here we are assuming that } \tau \text{ is too short to apply ergodicity: } Q_i / \tau \neq I_{signal} \text{ for any individual trial.}\]

\[\text{Here we have used the identity } \text{var}(c \times n) = c^2 \ \text{var}(n), \text{ where } c \text{ is a constant.}\]
PHOTON-COUNTING APDs

A Primer

\[ \langle a \rangle = \frac{I_{\text{primary}} \times \tau}{q}, \]

we have

\[ Q_{\text{signal-APD}} = q \times \langle n \rangle = M_1 I_{\text{primary}} \tau \quad \text{and} \]
\[ Q_{\text{noise-APD}} = \sqrt{\text{var}(Q)} = M_1 \sqrt{q I_{\text{primary}} F(M_1, k) \tau}. \]

Noise Equivalent Power, Specific Detectivity, and Noise Equivalent Input

The NEP of a detector is the optical power incident upon the detector that must be supplied to equal the noise power from all sources in the detector – in other words, NEP is the optical power that results in a SNR of unity. By convention, NEP is calculated for a 1 Hz window (i.e. the spectral intensity of the noise power is used in the denominator of the SNR, multiplied by a bandwidth \( B = 1 \) Hz). Generally speaking, one separates the noise contributions into three categories: multiplied shot noise sources such as the primary photocurrent and the component of the dark leakage current that passes through the bulk \((I_{db})\), un-multiplied shot noise sources such as the component of the dark leakage current that flows around the periphery of the device \((I_{dp})\), and noise sources like thermal noise that obey spectral intensity theorems of unique form. If these are the main noise mechanisms that operate in the APD, its SNR is:

\[ \frac{S}{N} = \frac{\left( I_{ph} M_1 \right)^2}{2 q \left( I_{ph} + I_{db} \right) M_1^2 F(M_1, k) + 2 q I_{dp} + \frac{4 k_B T}{R}} \times B. \]

Setting the SNR equal to unity and applying the quadratic formula, the noise equivalent photocurrent is:

\[ \frac{S}{N} = \frac{\left( I_{ph} M_1 \right)^2}{2 q \left( I_{ph} + I_{db} \right) M_1^2 F(M_1, k) + 2 q I_{dp} + \frac{4 k_B T}{R}} \times B. \]

A variety of other noise sources could be included in the denominator as required. These include shot noise resulting from stray background light, 1/f noise resulting from surface trapping phenomena, additional leakage from radiation damage, etc.
PHOTON-COUNTING APDs

A Primer

\[
\overline{I_{\text{NEP}}} = q F(M_i, k) B + \frac{1}{M_i} \sqrt{q^2 M_i^2 F^2(M_i, k) B^2 + 2B \left[ q I_{ab} M_i^2 F(M_i, k) + q I_{dp} + \frac{2k_BT}{R} \right]},
\]

for \( B = 1 \text{ Hz}. \)

This “primary” noise equivalent photocurrent is related to the noise equivalent optical power by the responsivity:

\[
\text{NEP} = \frac{1}{\eta} \times \overline{I_{\text{NEP}}} = \frac{h\nu}{q\eta} \times \overline{I_{\text{NEP}}}. 
\]

The specific detectivity \( D^* \) is a form of NEP that has been normalized for area \( A \):

\[
D^* = \sqrt{\frac{A}{\text{NEP}}}. 
\]

Specific detectivity is used in cases when the noise scales as the square root of the area, such as shot noise on currents that scale in proportion to detector area.

The letters NEI are used inconsistently in the detector community. In some instances NEI stands for “noise equivalent irradiance” and is given in units of photons per square centimeter per second; in other cases NEI stands for “noise equivalent input” and is quoted in photons. The definition in which “I” stands for “irradiance” is more closely related to NEP than the definition in which “I” stands for “input”. To go from a power to a photon irradiance, one simply divides out the detector’s collection area and scales by the photon energy. However, in order for NEI to make sense measured in quanta, the SNR must also be written for quanta rather than rates (such as current or power). It is incorrect to compute NEI from NEP by merely dividing NEP by the photon energy – that still results in a rate (and

\[21 \text{ Keep in mind that both the primary photocurrent appearing in the SNR, and the NEP derived from it, only apply to the bandwidth used in the calculation. Because of the way in which they were derived, } \overline{I_{\text{ph}}} \text{, } \overline{I_{\text{NEP}}} \text{, and NEP are all spectral components – they only correspond to the total signal when that signal is taken to be harmonic (as in a telecommunications application).} \]
PHOTON-COUNTING APDs

A Primer

a rate for a 1 Hz bandwidth at that). Instead, one applies the relations given earlier for $Q_{signal}$ and $Q_{noise}$\footnote{Terms for noise sources that do not integrate – such as thermal noise – are omitted. As with NEP, the shot noise is broken down into multiplied and un-multiplied components.}:

$$\frac{S}{N} = \frac{Q_{signal}}{Q_{noise}} = \frac{M_1 I_{ph} \tau}{\sqrt{q(I_{ph} + I_{db}) M_1^2 F(M_1, k) \tau + q I_{dp} \tau}} = \frac{\sqrt{q M_1 N_{input}}}{\sqrt{M_1^2 F(M_1, k) [q N_{input} + I_{db} \tau] + I_{dp} \tau}}$$

where the “primary input” carrier count $N_{input}$ is related to the primary photocurrent by:

$$N_{input} = \frac{I_{ph} \times \tau}{q}.$$ 

In analogy to the method for finding NEP, NEI is found by setting the SNR equal to unity and solving for $N_{input}$:

$$N_{NEI} = \frac{F(M_1, k)}{2} + \frac{1}{2} \sqrt{F^2(M_1, k) + \frac{4}{q} \left[ F(M_1, k) I_{db} \tau + \frac{I_{dp} \tau}{M_1^2} \right]}.$$ 

Finally, NEI is found from the noise equivalent primary input by scaling by the quantum efficiency:

$$NEI = \frac{1}{\eta} N_{NEI}.$$ 

Photon Detection Probability

It is not immediately obvious that these figures of merit apply to all photon counting applications. In the first place, the statistics upon which the noise theorems are erected apply to stationary random variables – i.e., they imply illumination that is uniform over the time span of the measurement. Secondly, NEP and related quantities are formulated to characterize the noise on the information carried by the photon arrival rate, in the limit of a large number of photons to which ensemble statistics can be applied. In contrast, laser radar returns and similar signals are...
pulsed, and the information they carry is often the fact of their arrival as opposed to
their exact magnitude.

The inapplicability of these figures of merit is particularly apparent in the case
of single photon counting. Insofar as dark current electrons are indistinguishable
from photogenerated electrons, dark current will manifest itself as a "dark count"
rate – it is this false detection rate which is of primary interest to a designer of
photon counting systems, not its variance (which is what the shot noise on the dark
current characterizes). Instead, the dark count rate can be used in conjunction with
the photon detection probability to compute the likelihood of detecting a signal
photon within a certain time span while simultaneously avoiding a spurious
detection event from dark current.

Both photon detection probability and dark count rate depend upon the
probability that the multiplied progeny of a single primary carrier exceed the
threshold of a readout circuit monitoring the detector, \( P_{th} \). If one assumes that the
readout circuit rejects all charge signals below a threshold \( n_{th} \), then the primary
carrier detection probability is found by integrating McIntyre’s distribution between
\( n_{th} \) and infinity:

\[
P_{th}(M, n_{th}, k) = \int_{n_{th}}^{\infty} dn \rho(M, n, k),
\]

where

\[
\rho(M, n, k) = \frac{\Gamma\left(\frac{n}{1-k} + 1\right)}{n(n-1)! \Gamma\left(\frac{k n}{1-k} + 2\right)} \times \left(\frac{1+k n}{M}\right)^{1+k n \frac{1-k}{k n}} \times \left(\frac{(1-k)(M-1)}{M}\right)^{n-1}
\]

is found by setting the primary carrier count \( a \) equal to unity. The probability of
detecting a single photon incident on the detector is just the product of \( P_{th} \) and \( \eta \); if
the bulk dark leakage current is \( I_{db} \), then the dark count rate is the product of \( P_{th} \)
and \( I_{db}/q \).

McIntyre’s single primary carrier pulse height distribution function is plotted in
Figure 3, and \( P_{th} \) is plotted versus discriminator threshold in Figure 2. A variety of
ionization coefficient ratios are investigated at an average gain of \( M = 100 \). The
importance of minimizing \( k \) should be obvious from its strong impact on \( P_{th} \).
PHOTON-COUNTING APDs
A Primer

Assuming a capacitance of roughly 30 fF, 100 μV of RMS voltage noise on the pre-amplifier input node will manifest itself as about 20 electrons of charge noise. For a signal-to-noise ratio (SNR) of at least 3, the discriminator threshold should be set to 60 electrons. As can be seen in Figure 2, under these conditions \( P_{in} \) varies from 55% for a ‘perfect’ \( k = 0 \) APD with no hole-initiated ionization to as little as 11% for a \( k = 0.4 \) material.

The ionization coefficient ratio \( k \) varies widely between semiconductors, and is a weak function of electric field strength. Under the field strengths relevant to APDs, \( k \sim 0.02 \) in bulk silicon, but is closer to 0.3 – 0.4 in bulk InAlAs and InP (common multiplication layer materials used for 1550 nm optoelectronics).

Clearly, in order to perform efficient photon counting with linear mode APDs at 1550 nm, it will be necessary to build APDs with \( k \) values much lower than what is characteristic of the bulk material. The design of multiplication layers to have lower \( k \) values than the bulk is called impact-ionization engineering (I²E). Voxtel has developed a linear mode APD design that is grounded in recent I²E research. Simulation results to be presented shortly indicate that our APD should be capable of efficient, high-speed photon counting at 1550 nm. It is the topic of the next section.

Low Noise Linear Mode APD Design

We begin this section with an introduction to basic III-V APD structure and modes of operation. A tutorial on designing low-noise APDs follows, with particular emphasis on features of the design being...
developed for AFRL. This document concludes with simulations of the design under development.

Introduction to Linear Mode SACM APDs

Voxtel's detector technology is based on InGaAs/InAlAs APDs grown lattice-matched to InP by metal-organic chemical vapor deposition (MOCVD). They are of the separate absorption; charge; multiplication (SACM) design which mitigates dark current leakage in the absorber by putting the low-bandgap material necessary to absorb NIR light in a single dedicated absorption layer. Doping is used in an adjacent charge layer to keep the potential across the absorption material low, so that only the multiplication layer experiences extreme fields (Figure 4).

The operation of a SACM APD can be understood by examining its $I$-$V$ characteristic (Figure 5). At low reverse bias, only a few photocarriers generated in the absorption layer manage to diffuse to the junction. Thus, the small reverse current is typical of the minority-carrier-controlled saturation current of any reverse-biased diode. As the voltage is raised, the depletion region that initially formed in the multiplication layer penetrates the charge layer and approaches the absorber. As the junction’s depletion region nears the absorber, more and more photocarriers are collected, and the reverse current climbs dramatically. “Punch-through” occurs when the depletion region first touches the absorber, and maximum collection of photocarriers follows soon after, once the absorber has been fully depleted. If the charge layer has been sized properly (a matter of choosing the correct amount of integrated space charge from doping), punch-through will occur before the field in the multiplication region climbs high enough for any substantial avalanche multiplication to take place. Thus, a plateau appears in the $I$-$V$ characteristic that marks unity gain. As the reverse bias is increased further, the field in the multiplication region rises to the point where primary photocarriers from the absorber gain sufficient kinetic energy to trigger impact ionization events. This -- and subsequent avalanche of the secondaries -- generates

![Figure 4: Layer schematic and band edge diagram showing how the SACM design (right) reduces tunneling leakage in NIR APDs.](image-url)
PHOTON-COUNTING APDs

A Primer

multiplication gain. Finally, beyond a critical threshold voltage, the counter-propagating electrons and holes multiply with sufficient frequency to establish a condition of positive feedback, and the avalanche current rises to a level limited only by the external circuit – this is avalanche breakdown.

Under conditions of extremely low illumination and very little dark current leakage, APDs can be momentarily biased above their breakdown voltage without actually undergoing avalanche breakdown. Instead, the APD exists in a metastable state in which the next carrier injected into its multiplication region – be it photogenerated or dark current – will trigger avalanche breakdown. Thus, an APD biased above breakdown is capable of generating a macroscopic current from a single photocarrier. At present, when APDs are used for photon counting, they are deployed in this “Geiger mode” of operation.

Afterpulsing is the fundamental problem associated with Geiger mode operation. It’s not that the sensitivity of a Geiger mode APD to upset is a show-stopper – any process capable of detecting the single photocarrier pair generated by a single photon will similarly be subject to spurious dark counts arising from individual dark current carriers. Neither, per se, is the cooling requirement – after all, any similar APD will have to contend with the same dark current sources, and will similarly benefit from some cooling to lower the dark count rate. Instead, it is the macroscopic size of the breakdown current that makes the presence of traps in combination with cooling fatal, for it is this macroscopic current that populates the traps in the first place. Consider instead an APD operated below its breakdown bias, but still with substantial multiplication gain. In response to a single photon, it will generate a few hundred secondary carriers at most – too few to fill the traps in its multiplication layer. Such a “linear mode” APD need neither be quenched nor reset – maximum count rates are limited by the RF characteristics of the packaging and circuitry rather than the detector itself. As an added benefit, dark count rates will be lower because less bias is required to operate an APD in linear mode, and

Figure 5: Illustration of a SACM APD’s I-V characteristic.
the leakage mechanisms that dominate an APD’s dark current have exponential dependence upon its internal field strength.

**Multiplication Noise Revisited**

The excess noise factor was derived by Robert McIntyre in 1966 through consideration of the statistical distribution of ionization probabilities, based upon the assumption that ionization probability was a function of local field strength only, and not carrier history. Multiplication noise increases when more than one carrier type causes ionizations (Figure 6). Whenever an ionizing collision takes place, both an electron and a hole are generated. In an ideal situation, only electrons will cause further ionizations – this corresponds to the impact ionization coefficient ratio \(k = 0\).\(^{23}\) Under these circumstances, the avalanche process moves through the multiplication region in an orderly fashion, and the resulting excess noise – uncertainty in the number of carriers generated by avalanche multiplication – only reflects the probabilistic nature of the ionization process for electrons. In reality, \(k \neq 0\), and holes can themselves generate new carrier pairs as they traverse the multiplication layer in the opposite direction. The result is a type of noisy feedback as forward-moving electrons generate an uncertain number of backward-moving holes which generate an uncertain number of forward-moving electrons, etc. This physics, in an idealized form, was captured in McIntyre’s original expression for the excess noise factor:

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

\(^{23}\) Alternately, APDs in which hole-initiated ionization dominates are also low-noise, with \(k \to \infty\). In this case, the excess noise factor has a different form. The point is that there is less excess multiplication noise when one carrier type dominates the ionizing collisions.
where \( k = \frac{\beta_{\text{hole}}}{\alpha_{\text{electron}}} \) and \( \beta_{\text{hole}} \) and \( \alpha_{\text{electron}} \) are ionization coefficients. This leads to a family of curves (Figure 7) for various values of \( k \) bounded by \( k = 0 \).

Obviously, negative values of \( k \) are non-physical, and so any noise data that is measured with \( F(M) \) below the \( k = 0 \) curve would be considered anomalous in the framework of McIntyre’s original theory. Some years ago, students in Professor J. C. Campbell’s group at UT Austin measured anomalous low-noise in GaAs-based APDs with very thin multiplication layers. Eventually it was determined that McIntyre’s noise theory does not apply to APDs when the multiplication layer is thin enough to allow statistical correlations between successive ionization events. These statistical correlations reduce the random fluctuation in ionization, and so result in lower multiplication noise. This phenomenon is the theoretical basis for the present work.

**Impact Ionization Engineering Concepts**

The goal of impact ionization engineering is to reduce excess multiplication noise by designing semiconductor structures in which the impact ionization events will naturally be correlated. In general, two tools are used: the so-called “dead-space” effect and localized enhancement of the ionization rate. Both reduce the number of possible ionization chains – and hence narrow the distribution of the multiplication gain – through spatial localization of the ionization events.

Detailed Monte Carlo simulations of avalanche multiplication confirm that SACM APDs with thin multiplication layers have reduced multiplication noise as a result of the dead-space effect. Ionization events tend to be localized inside thin multiplication regions because following each ionizing collision, carriers must pick up energy across a certain distance – the dead-space – before they are capable of causing subsequent ionizations. By effectively eliminating some of the places where impact ionization can occur in the structure, the dead-space effect reduces the number of possible ionization chains; this
is the source of the noise reduction observed for such APDs. However, the effect diminishes rapidly for thicker multiplication layers, as the dead space following any given collision becomes a progressively smaller fraction of the total multiplication layer thickness (Figure 8).