Figures of Merit for CMOS SPADs and arrays

D. Bronzia, F. Villa, S. Belli, S. Tisa, G. Ripamonti, A. Tosi

ABSTRACT

SPADs (Single Photon Avalanche Diodes) are emerging as most suitable photodetectors for both single-photon counting (Fluorescence Correlation Spectroscopy, Lock-in 3D Ranging) and single-photon timing (Lidar, Fluorescence Lifetime Imaging, Diffuse Optical Imaging) applications. Different complementary metal-oxide semiconductor (CMOS) implementations have been reported in literature. We present some figure of merit able to summarize the typical SPAD performances (i.e. Dark Counting Rate, Photo Detection Efficiency, afterpulsing probability, hold-off time, timing jitter) and to identify a proper metric for SPAD comparison, both as single detectors and also as imaging arrays. The goal is to define a practical framework within which it is possible to rank detectors based on their performances in specific experimental conditions, for either photon-counting or photon-timing applications. Furthermore we review the performances of some CMOS and custom-made SPADs. Results show that CMOS SPADs performances improve as the technology scales down; moreover, miniaturization of SPADs and new solutions adopted to counteract issues related with the SPAD design (electric field uniformity, premature edge breakdown, tunneling effects, defect-rich STI interface) along with advances in standard CMOS processes led to a general improvement in all fabricated photodetectors; therefore, CMOS SPADs can be suitable for very dense and cost-effective many-pixels imagers with high performances.

Keywords: Single-photon avalanche photodiode, photodetectors, photon counting, photon timing, CMOS technology.

1. INTRODUCTION

Since 1960s single-photon avalanche diodes (SPADs) have been deeply studied and used in several fields where single-photon sensitivity is required such as fluorescence correlation spectroscopy (FCS), fluorescence lifetime imaging (FLIM), adaptive optics, quantum cryptography, positron emission tomography, as well as laser (LIDAR/LADAR) and 3-D optical ranging. Although many single-photon sensitive devices already exist, SPADs have gained attention because of some advantages over electron-multiplying charge-coupled devices (EMCCDs), photomultiplier tubes (PMTs) and multi-channel plates (MCPs). They all can reach sensitivity at close to single-photon level but EMCCDs needs cooling and long integration time, whilst PMTs and MCPs require high bias voltage, are bulky and sensitive to magnetic fields, and cannot be integrated with complex complementary metal-oxide semiconductor (CMOS) electronics. Conversely, SPAD devices are small, rugged and insensitive to magnetic fields. Till few years ago, SPAD could be fabricated solely with custom processes whose flexibility allow to address the most demanding requirements in terms of efficiency, noise and timing performance; unfortunately, the cost of dedicated processes and the impossibility to manufacture large array due to a lack of integration with proper electronics prevent SPADs to emerge as main actors in the aforementioned applications.

From the early 2000s onwards many research groups have leveraged the advances in CMOS technologies to fabricate SPADs in standard CMOS processes. The main advantage is the feasibility to monolithically integrate on the same substrate the photodetector and the electronics needed to implement smart photon-counting or photon-timing techniques. As a consequence, researchers started to develop compact and cost-effective multi-pixel SPAD-based image sensors that are comparable in terms of single-photon sensitivity, dynamic range and acquisition speed to most of high-end CCD cameras.

After a decade of research, many groups worldwide have developed their own SPAD structures in different technology nodes, coping in a unique way with all the problems related to the CMOS SPAD fabrication (premature edge breakdown,
tunneling effects, electric field uniformity, junction depth). Moreover each group performed measurements in different experimental conditions. In such a maze of variables and parameters, it is generally difficult to make a fair comparison between different research activities and results, unless a subset of shared representative parameters is found.

To this purpose, we now present different figures of merit elaborated taking into account the typical SPAD performances (i.e. noise, efficiency, hold-off time, timing jitter) in order to identify a proper metric for SPAD comparison and to define a practical framework within which it could be possible to rank detectors based on their performances in specific experimental conditions, for either photon-counting or photon-timing applications.

The paper is organized as follows: in the second section, we will describe the working principle and the main parameters of a SPAD; in the third section, we will define the figures of merit and in the fourth section we will show and comment the results of experimental characterization; finally, we will draw some conclusions.

2. MAIN SPAD PARAMETERS

The main process that determines the characteristic of a SPAD is the impact ionization mechanism; in fact, a SPAD is a p-n junction biased at a voltage $V_{\text{POL}}$, well above its breakdown voltage; at this bias, the electric field in the depletion layer of the p-n junction is high enough that, upon the absorption of a photon, a single photo-generated carrier is able to trigger a macroscopic self-sustaining avalanche current. This operation mode is generally referred to as Geiger-mode. After the ignition of the avalanche, the current keeps flowing until a proper front-end circuit lowers the bias voltage below the breakdown voltage; in these conditions, none of the carriers crossing the high field region impact ionizes so the current cannot self-sustain any longer, and the avalanche is then quenched. The front-end circuit after a pre-set time (called hold-off time) swiftly restores the bias condition to allow the detection of another photon (re-bias phase).

Before discussing about figures of merit, we now briefly introduce the most important parameters that can be used to evaluate the SPAD performances: photon detection efficiency, dark counting rate, hold-off time and timing jitter.

2.1 Photon Detection Efficiency

The photon detection efficiency (PDE) is the ratio between the number of detected photons and the number of photons impinging the SPAD:

$$\text{PDE} = \frac{N_{\text{DETECTED}}}{N_{\text{IMPISING}}}$$  

(1)

In order to be detected a photon must be first absorbed and then it must trigger the avalanche. It follows that the PDE is the product of the absorption efficiency and the triggering probability. The absorption efficiency can be expressed by the following equation:

$$\eta_{\text{ASS}} = (1 - R) \exp(-aD) \left[ 1 - \exp(-aW) \right]$$  

(2)

Where $R$ is the reflection coefficient at the interface between the air and the detector’s surface, $D$ and $W$ are the depth and the width of the depleted layer, $\alpha$ is the silicon absorption coefficient which is a wavelength-dependent parameter. The triggering probability instead is given by the following equation:

$$P_t = 1 - \exp\left( -\frac{V_{\text{EX}}}{V_C} \right)$$  

(3)

Where $V_{\text{EX}}$ is the excess bias and $V_C$ is a weighted mean of the hole and electron ionization coefficients.

2.2 Dark Counting Rate

In analog photodetectors (PIN, APD) the fluctuations of the photo-generated current or the dark current are generally negligible with respect to the noise contribution of the front-end electronics; conversely, SPADs are capable of tagging photon arrivals with the generation of a macroscopic current that can be processed to obtain a digital pulse; consequently, the SPAD output signal is not affected by the noise of the read-out electronics and the main noise contributions are the fluctuations of the optical signal and the dark counts. Dark counts can be either uncorrelated or
correlated to signal photons. The former contribution is due to ignitions caused by carriers generated through either thermal or Shockley-Read-Hall processes. Instead the correlated noise is caused by carriers that get trapped during an avalanche ignition and are then released igniting an “afterpulse”.

Generally the term “dark counting rate” (DCR) is used to refer to the uncorrelated noise source, while the term “afterpulsing” is used to describe the effects of the afterpulses. To understand the effect of the afterpulsing, consider the following example: if $N_{DET}$ photons are detected and $N_{AP}$ correlated counts are generated then we can define the afterpulsing probability $P_{AP}$ as:

$$P_{AP} = \frac{N_{AP}}{N_{DET}} \tag{4}$$

Of course, the afterpulses generation is a cascade process and $N_{AP}$ counts will generate $N_{AP} \cdot P_{AP}$ counts and so on. Therefore, for $N_{DET}$ detected photons the number of measured counts is:

$$N_{MEAS} = N_{DET} + N_{DET} P_{AP} + N_{DET} P_{AP}^2 + \ldots = N_{DET} \sum_{n=0}^{\infty} P_{AP}^n = \frac{N_{DET}}{1 - P_{AP}} \tag{5}$$

So the number of detected photon is always less than the measured number of photons because of the afterpulsing probability.

### 2.3 Hold-off time

As described before in the paper, the hold-off time ($T_{HO}$) is the time interleaved between the avalanche quenching and the restoration of the bias condition. The hold-off time is an important parameter because it fixes the maximum count rate of the SPAD, which is roughly equal to the inverse of the hold-off time. Moreover afterpulsing can be mitigated if the hold-off time is long enough so most of trapped carriers are released before re-biasing the photodetector.

### 2.4 Timing jitter

The timing jitter of a single-photon avalanche diode is the statistical distribution of the delays between the true photon arrival and the output signal generation. The distribution has a gaussian component arising from the statistics of the avalanche build-up and an exponential component caused by diffusive processes within the neutral layer of the SPAD. The most important parameter of this distribution is its full-width at half maximum (FWHM) generally measured in ps.

### 3. FIGURES OF MERIT

To properly define the figures of merit, first of all, we must consider that SPADs can be used in both photon-counting and photon-timing applications; therefore, it is important to identify the parameters that influence most the performances of a SPAD when it is operated in counting or timing mode.

#### 3.1 Figure of merit for photon-counting applications

The maximization of the signal-to-noise ratio (SNR) is a fundamental requirement in every application. For a SPAD, the signal-to-noise ratio is expressed by the following equation:

$$SNR = \frac{PDE \cdot \Phi_s}{\sqrt{PDE \cdot \Phi_s + DCR}} \sqrt{T_{INT}} = \frac{PDE \cdot \Phi_s}{\sqrt{DCR}} \sqrt{T_{INT}} \tag{6}$$

Where $\Phi_s$ is the signal photon flux, and $T_{INT}$ is the integration time. So Eq. (6) shows that to maximize the SNR, at a given integration time, a SPAD should have high photon detection efficiency and low dark counting rate. In a photon-counting application, besides, the photodetector must have a high dynamic range; therefore, the saturation should occur at very high illumination level. Ideally, a photodetector should never saturate ($\Phi_{S,MAX} = \infty$); a real SPAD, of course, will always saturate at a certain photon flux whose value is equal to the inverse of the hold-off time ($\Phi_{S,MAX} = 1/T_{HO}$). Indeed, the saturation flux is depending on the afterpulsing probability:

$$\Phi_{S,MAX} = \frac{l - P_{AP}}{T_{HO}} \tag{7}$$
A further discussion on the SPAD area must be done: large area devices with respect to small area ones enable the fabrication of arrays with higher fill-factor; in addition, the dark counting rate of a SPAD depends quadratically on the area of the device. For these reasons the SPAD area must be considered in the definition of the figure of merit.

Finally, a last parameter must be taken into account: the technology node, as more scaled technologies enable the fabrication of denser arrays.

Starting from these considerations, we elaborated the following figure of merit for photon-counting applications:

\[
FOM_{\text{COUNTING}} = \frac{PDE_{\text{PEAK}} \cdot \phi \cdot (1 - P_{\text{AP}})}{\sqrt{\text{DCR} \cdot T_{\text{HO}} \cdot \text{TechNode}}}
\]  

(8)

The DCR is put under square root as it appears in the SNR formula, and for this reason we used the SPAD diameter (\(\phi\)) instead of the SPAD area, otherwise we would overestimate the area contribution. Moreover, we chose to evaluate the PDE at its peak, as it represents the most significant value of the efficiency waveform. The FOM thus obtained is dimensionless.

3.2 Figure of merit for photon-timing application

The considerations made before about SNR, area and technology node are valid for photon-timing applications, as well. A different discussion must be done about the hold-off time: in photon-timing applications (e.g. time-correlated single-photon counting, TCSPC) the light flux is limited well below the saturation of the photodetector to avoid distortions in the measurement; so, hold-off time is usually not a concern and the timing jitter of the detector plays a more important role. Consequently, we modified the photon-counting figure of merit to account for the previous considerations and we elaborate the following figure of merit for photon-timing applications:

\[
FOM_{\text{TIMING}} = \frac{PDE \cdot \phi \cdot (1 - P_{\text{AP}})}{\sqrt{\text{DCR} \cdot \text{FWHM} \cdot \text{TechNode}}}
\]  

(9)

RESULTS

Table I summarizes the performances of most of works available in literature about CMOS SPADs. Other works could not be considered because they miss some of the analyzed parameters. PDE, DCR and afterpulsing probability (AP) are rated at the same excess bias voltage; if performances are rated at more than an excess bias voltage we consider the overvoltage that gives the best FOM and when not specified by authors we consider AP to be negligible (-0%). Some authors fabricated in the same technology more devices and we report all of them. In addition, the wavelength of the peak PDE and the wavelength at which the timing response was characterized are also specified. For most of the reviewed works, the PDE has a peak in the blue wavelength range (450 – 500 nm) as expected; [12] and [25] have a near UV-enhanced PDE with a peak at 410 nm, whereas for [6] and [26] the peak (600 – 690 nm) is shifted towards the red wavelength. The difference in peak PDE wavelength does not represent a disadvantage, but set the validity of the FOM in different ranges. The timing performances were characterized mainly at 470 nm and 637 nm, with some characterizations performed at 710 nm and 820 nm. This is disadvantageous especially for those SPADs whose timing jitter was characterized at a suboptimal wavelength and therefore the FWHM measured does not represent the best timing performance achievable. Moreover, not all the timing measurements are performed at the same excess bias as PDE, DCR and afterpulsing measurements.

Nevertheless, the trend of both counting and timing FOMs shows that SPAD performances improve as the technology node scales. This could be explained considering that the variation of the peak PDE in different technologies is negligible, whereas the ratio between DCR and SPAD area (which appears in our FOM as \(\sqrt{\text{DCR}/\phi}\)) decreases in more scaled technologies. This is a consequence of the fact that with the technology scaling it is possible to benefit from lower excess bias voltages (without impairing PDE) and from the miniaturization of SPAD and electronics; both conditions help in reducing the DCR and afterpulsing probability of the fabricated devices and in improving timing performances because on smaller devices the avalanche build-up has a narrower statistical spread.
Table I. Summary of the main parameters for several CMOS SPADs and their respective counting and timing figures of merit. For [6] we estimated the afterpulsing probability from the data reported in the paper. The value of diameter is the one of a circular structure with the same area of the considered SPAD.

<table>
<thead>
<tr>
<th>Reference</th>
<th>PDE_{PEAK} (%)</th>
<th>Ø (µm)</th>
<th>DCR, @ RT (cps)</th>
<th>T_{THO} (ns)</th>
<th>AP (%)</th>
<th>FWHM (ps)</th>
<th>Tech. Node (nm)</th>
<th>FOM_{SPAD} (dB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1] – Rochas et al., 2003</td>
<td>28 (470)</td>
<td>7</td>
<td>900</td>
<td>75</td>
<td>7.5</td>
<td>60 (710)</td>
<td>800</td>
<td>120.1</td>
</tr>
<tr>
<td>[2] – Zappa et al., 2004</td>
<td>40 (500)</td>
<td>12</td>
<td>3500</td>
<td>55</td>
<td>2.6</td>
<td>36 (820)</td>
<td>800</td>
<td>115.1</td>
</tr>
<tr>
<td>[3] – Niclass et al., 2005</td>
<td>26 (460)</td>
<td>7</td>
<td>350</td>
<td>40</td>
<td>~0</td>
<td>115 (?)</td>
<td>800</td>
<td>129.7</td>
</tr>
<tr>
<td>[4] – Pancheri et al., 2007</td>
<td>33 (470)</td>
<td>11</td>
<td>1400</td>
<td>32</td>
<td>0.3</td>
<td>144 (480)</td>
<td>700</td>
<td>132.9</td>
</tr>
<tr>
<td>[5] – Niclass et al., 2006</td>
<td>40 (460)</td>
<td>10</td>
<td>750</td>
<td>40</td>
<td>23</td>
<td>80 (?)</td>
<td>350</td>
<td>138.1</td>
</tr>
<tr>
<td>[6] – Mosconi et al., 2006</td>
<td>11 (600)</td>
<td>22.6</td>
<td>5000</td>
<td>150</td>
<td>4*</td>
<td>80 (670)</td>
<td>350</td>
<td>112.1</td>
</tr>
<tr>
<td>[8] – Niclass et al., 2008</td>
<td>35 (460)</td>
<td>7</td>
<td>646</td>
<td>100</td>
<td>~0</td>
<td>-</td>
<td>350</td>
<td>128.8</td>
</tr>
<tr>
<td>[8] – Stoppa et al., 2009</td>
<td>33 (450)</td>
<td>22.6</td>
<td>300</td>
<td>500</td>
<td>4.5</td>
<td>160 (470)</td>
<td>350</td>
<td>127.4</td>
</tr>
<tr>
<td>[9] – Pancheri et al., 2009</td>
<td>32 (450)</td>
<td>17.8</td>
<td>1000</td>
<td>200</td>
<td>6</td>
<td>70 (470)</td>
<td>350</td>
<td>127.7</td>
</tr>
<tr>
<td>[10] – Guerrieri et al., 2010</td>
<td>42 (450)</td>
<td>20</td>
<td>4900</td>
<td>100</td>
<td>27</td>
<td>39 (820)</td>
<td>350</td>
<td>128.0</td>
</tr>
<tr>
<td>[12] – Bronzi et al., 2012</td>
<td>52 (410)</td>
<td>30</td>
<td>50</td>
<td>20</td>
<td>3.9</td>
<td>113 (390)</td>
<td>350</td>
<td>169.6</td>
</tr>
<tr>
<td>[13] – Faramarzpour et al., 2008</td>
<td>2.5 (470)</td>
<td>10</td>
<td>60k</td>
<td>30</td>
<td>~0</td>
<td>-</td>
<td>180</td>
<td>105.5</td>
</tr>
<tr>
<td>[14] – Niclass et al., 2010</td>
<td>20 (470)</td>
<td>8</td>
<td>180</td>
<td>6</td>
<td>~0</td>
<td>80 (?)</td>
<td>180</td>
<td>160.9</td>
</tr>
<tr>
<td>[15] – Isaak et al., 2010</td>
<td>17.4 (470)</td>
<td>10</td>
<td>13k</td>
<td>40</td>
<td>~0</td>
<td>-</td>
<td>180</td>
<td>126.5</td>
</tr>
<tr>
<td>[16] – Pancheri et al., 2011</td>
<td>26 (470)</td>
<td>10</td>
<td>160</td>
<td>230</td>
<td>30</td>
<td>1.3</td>
<td>2.1</td>
<td>60 (470)</td>
</tr>
<tr>
<td>[17] – Niclass et al., 2007</td>
<td>35 (450)</td>
<td>10</td>
<td>100k</td>
<td>100</td>
<td>~0</td>
<td>144 (637)</td>
<td>130</td>
<td>118.6</td>
</tr>
<tr>
<td>[18] – Gersbach et al., 2009</td>
<td>30 (480)</td>
<td>8.6</td>
<td>670</td>
<td>220</td>
<td>180</td>
<td>&lt;1</td>
<td>125 (637)</td>
<td>128 (637)</td>
</tr>
<tr>
<td>[19] – Richardson et al., 2009</td>
<td>28 (500)</td>
<td>8</td>
<td>60</td>
<td>100</td>
<td>~0</td>
<td>200 (470)</td>
<td>130</td>
<td>146.9</td>
</tr>
<tr>
<td>[20] – Field et al., 2010</td>
<td>30 (425)</td>
<td>4.75</td>
<td>230</td>
<td>-</td>
<td>~0</td>
<td>198 (408)</td>
<td>130</td>
<td>-</td>
</tr>
<tr>
<td>[21] – Veerappan et al., 2011</td>
<td>27.5 (?)</td>
<td>5.6</td>
<td>160</td>
<td>100</td>
<td>~0</td>
<td>140 (637)</td>
<td>130</td>
<td>139.4</td>
</tr>
<tr>
<td>[23] – Gersbach et al., 2012</td>
<td>25 (?)</td>
<td>8.6</td>
<td>100</td>
<td>100</td>
<td>0.1</td>
<td>61 (?)</td>
<td>130</td>
<td>144.4</td>
</tr>
<tr>
<td>[22] – Webster et al., 2012a</td>
<td>25 (560)</td>
<td>8</td>
<td>18</td>
<td>-</td>
<td>1</td>
<td>88 (654)</td>
<td>130</td>
<td>-</td>
</tr>
<tr>
<td>[24] – Karami et al., 2010</td>
<td>12 (470)</td>
<td>8</td>
<td>8</td>
<td>10k</td>
<td>32</td>
<td>398 (637)</td>
<td>90</td>
<td>108.2</td>
</tr>
<tr>
<td>[25] – Henderson et al., 2010</td>
<td>36 (410)</td>
<td>2</td>
<td>250</td>
<td>-</td>
<td>~0</td>
<td>107 (470)</td>
<td>90</td>
<td>-</td>
</tr>
<tr>
<td>[26] – Webster et al., 2012b</td>
<td>44 (690)</td>
<td>8</td>
<td>100</td>
<td>15</td>
<td>0.375</td>
<td>51 (470)</td>
<td>90</td>
<td>168.3</td>
</tr>
</tbody>
</table>

SPAD performances not only benefit from the advantages of more scaled technologies, but also from global advances in standard CMOS processes. This is reflected into a general improvement that crosses all technology nodes; for instance, [12] (0.35 µm) and [26] (90 nm) have the best counting FOMs and both were published in 2012. Also, the best four timing FOMs belongs to works published in 2012 which present SPADs fabricated in different technologies ([12] (0.35 µm), [22]-[23] (130 nm), and [26] (90 nm)). Also in this case, the DCR density improves in more recent technologies. This could be connected to the adoption of new solutions to counteract issues related to tunneling effects, premature edge breakdown or (especially in scaled technologies) to the presence of shallow trench isolations (STI) that introduce traps and defects thus increasing DCR; moreover, starting materials with higher quality and cleaner processes could have played an important role in improving SPAD performances.
4. CONCLUSION

In this paper we presented two figures of merit to rank SPAD performances. The figures of merit were elaborated by analyzing the main SPAD parameters that influence most the performances when the photodetector is used for timing or counting applications. We found that FOMs increase as the technology scales down, thanks to miniaturization of SPADS and electronics that allow to reduce DCR and improve timing performances; we also found that the efforts made by researches to find new solutions and the progress of the CMOS processes led to improvements crossing all fabricated devices, even if manufactured in different technologies.

ACKNOWLEDGMENT

This work was supported by the “MiSPiA” project, under the ICT theme of the EC 7th Framework Program (FP7, 2007-2013), grant agreement n. 257646.

REFERENCES


