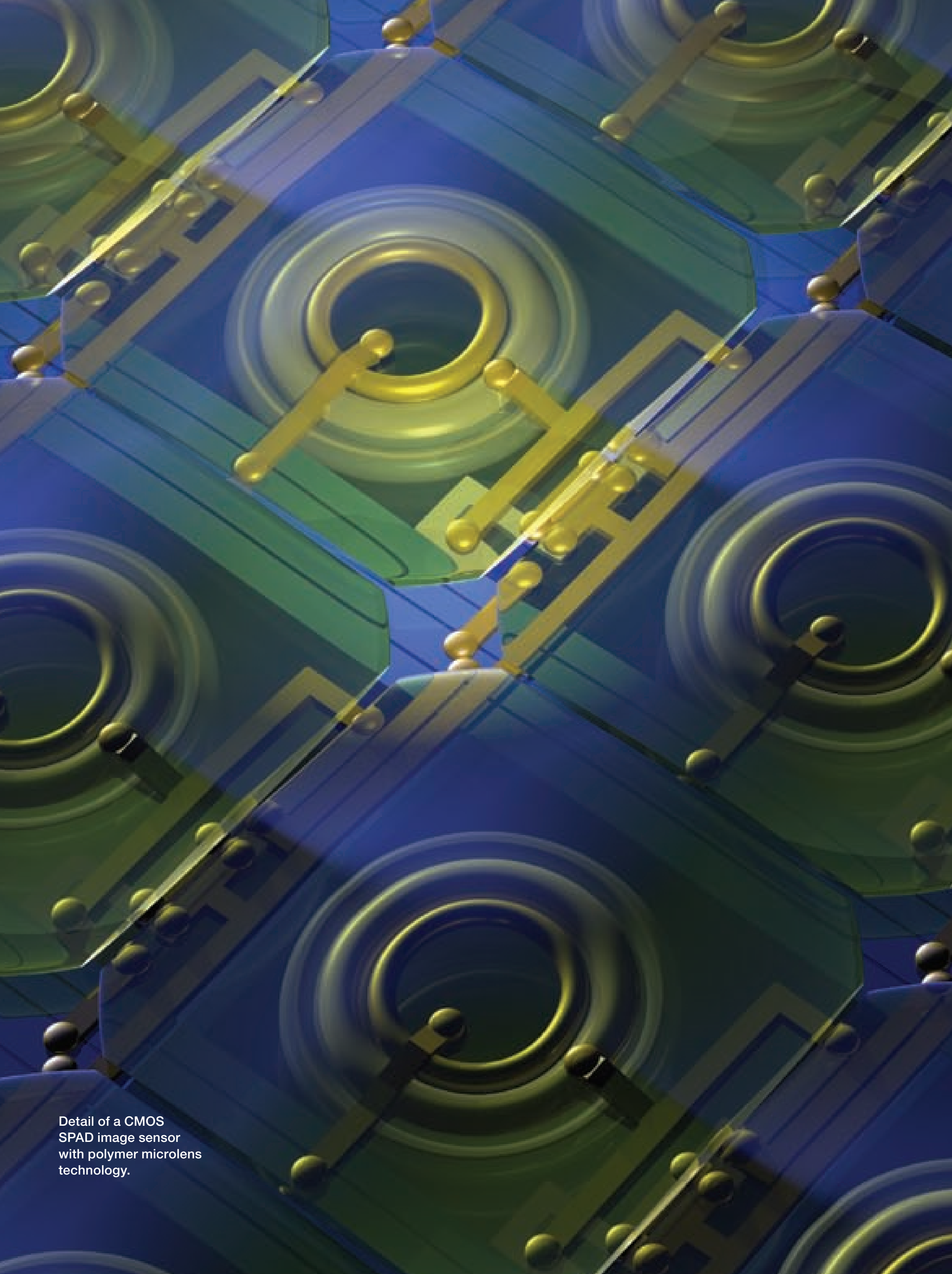
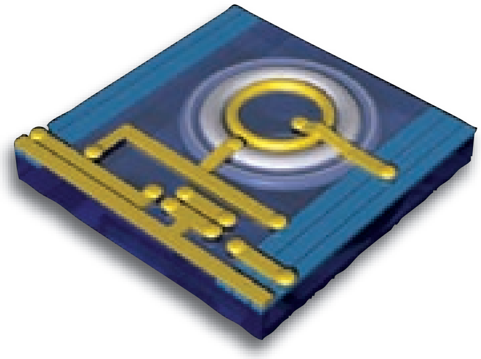


Detail of a CMOS SPAD image sensor with polymer microlens technology.



SPAD Sensors Come of Age

Edoardo Charbon and Silvano Donati

A unique light detector combines single-photon performance, multi-pixel image resolution and deep sub-nanosecond response.

In 1839, Alexandre-Edmond Becquerel set the stage for the development of photodetectors when he observed the photovoltaic effect. Then, Willoughby Smith furthered the field with his discovery of photoconductivity in selenium in 1873. Today, some 130 years later, light sensing continues to grow as a discipline, enriched of new ideas and achievements.

For example, over the past decade, several new devices have been introduced for far-infrared detection, including the micro-bolometer array for uncooled thermal-IR cameras and the quantum well infrared detector. In the ultrafast photodiode segment, improved structures such as the uni-travelling carrier and the wave-guide photodiode have dwarfed normal photodiode performance, pushing the upper frequency cutoff to a record 500-1,000 GHz and unveiling the field of T-waves.

In this quickly evolving field, only the 90-year-old photomultiplier tube (PMT) appeared to be invulnerable to change. For years, it quietly withstood the onslaught of semiconductor devices, none of which ever seemed to approach its single-photon detection capability, along with its wide dynamic range of linearity, its extended-blue response and, last but not least, its huge, no-extra-cost sensitive area.

However, the single-photon avalanche diode (SPAD) may be changing that. In the 1980s, S. Cova and his colleagues developed these new silicon devices, which were classifiable as single photon and which could enhance time-resolution. Although these devices were limited by dead time between detections and a small active area, research in the past 20 years has yielded important refinements to both the structure of the SPAD and the ancillary circuits—shortening the dead time from microseconds to nanoseconds.

Courtesy of Donati and Charbon. Illustration by Phil Saunders, SpaceChannel.org.

1047-6938/10/02/0034/8-$15.00 ©OSA

Investigators have also begun to realize that some of the disadvantages of SPAD compared to PMT technology can also be viewed as strengths. For example, the fast response combined with single-photon capability makes it possible to perform fast timing of light pulses (of low repetition rate) down to the sub-nanosecond time scale. More important, the small size of SPADs is ideal for integrating the device into an array of pixels. This opened the possibility that SPADs could be used for imaging applications such as 3-D cameras and biomedical applications that call for fast image gating. This is why, in 2003, we began to design SPADs and SPAD imagers in standard CMOS processes. This work led to the design of large SPAD imagers.

The science of SPADs

First let's review the science behind SPADs. To understand these devices, we must start by considering a normal avalanche photodiode (APD)—a junction semiconductor diode reverse-biased to a high voltage. In this device, charge carriers acquire enough energy from the electric field across the depleted junction to break covalent bonds, thus freeing new carriers in what is called impact ionization or avalanche multiplication. For a single detected photon, the avalanche process results in a collected charge much larger, with M carriers, or, the APD has an internal gain M.

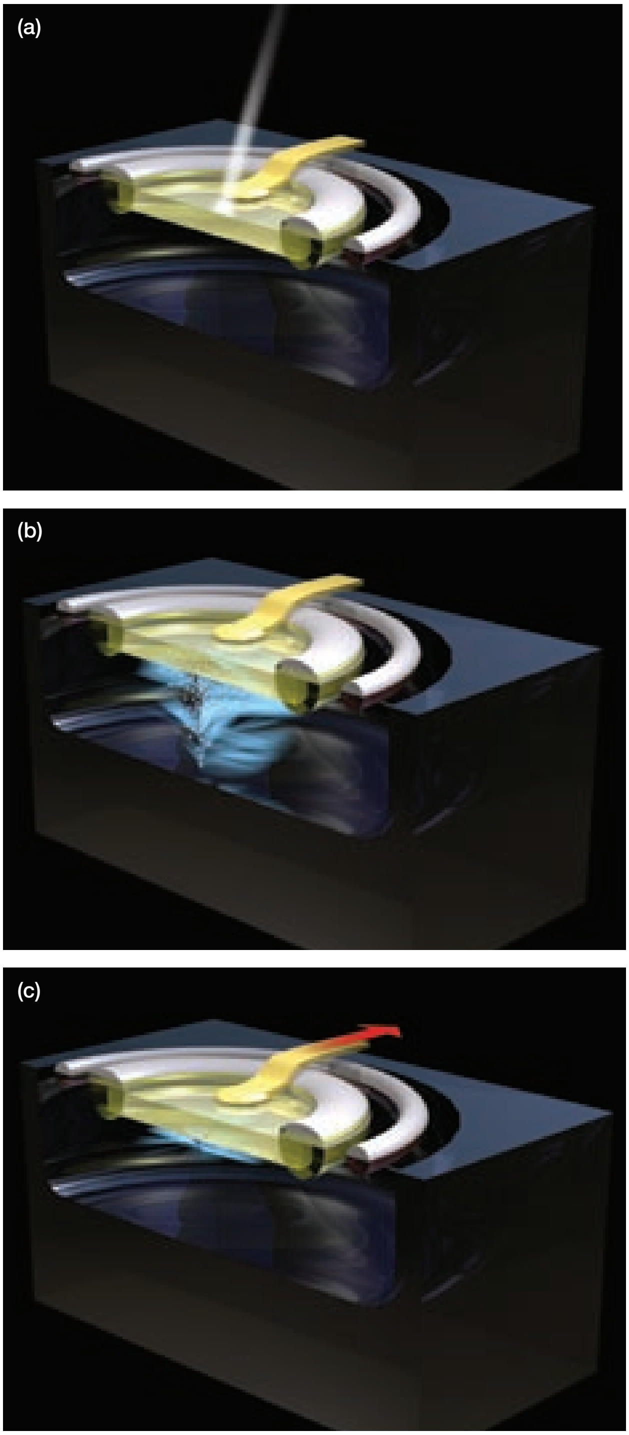
APD operation is linear with respect to input optical power multiplied by a factor M. Incidentally, the optimum value of M is the ratio α/β of the ionizing coefficients α and β of electrons and holes. Ionizing electrons produce both an electron and a hole along their path to the anode. The hole travels in the opposite direction to the cathode and produces new pairs back along the path.

Because of this positive feedback, the field or the width of the multiplication region is increased beyond a certain value; the gain also increases—but at the expense of a bandwidth decrease and an excess noise increase. This is typically not desirable in applications, so we stop at the value $M=\alpha/\beta$ of gain when using the APD in the usual, linear mode.

On the other hand, if we continue to increase the gain, we reach a point where it becomes infinite—a single initial electron starts an avalanche that doesn't stop, and the number of created carriers increases without a limit until an external circuit prevents more current from being drained by the device. This is the trigger or Geiger mode of operation of the APD, also known as single-photon avalanche diode.

The first carrier crossing the high-field region of the junction will generate an avalanche—that is, an exponentially increasing current—until a limit is set by the bias circuit. As a result, a spike of current is generated by the SPAD in response to a single photoelectron detected in the depletion layer as well as to a single "dark" photoelectron generated thermally. Following the pulse, there is a dead time for the device to recover from the high-current on state to the zero-current off state of the SPAD, which is ready to sense the next photoelectron.

[Photon absorption and avalanche propagation in a SPAD]



(a) A photon strikes the surface of the semiconductor and is absorbed at a certain depth. (b) Upon initiation, the avalanche develops towards the junction. (c) Upon reaching the junction, the avalanche carriers are collected by the doped region. Then they flow to the conductor for pulse shaping and processing.

Research in the past 20 years has yielded important refinements to both the structure of the SPAD and the ancillary circuits—shortening the dead time from microseconds to nanoseconds.

In this implementation, a SPAD is not fully equivalent to a PMT, because only one photon is detected at a time, and a packet of N-photons gives the same response as a single photon. In addition, a SPAD has a certain dead time between detections, whereas the PMT hasn't.

Another drawback is that the undesirable dark-current electrons (or those in the reverse current of the junction) are indistinguishable from photon-generated electrons. To minimize this effect, one must keep the active area as small as possible—10 µm in diameter or less. This value is dwarfed by the huge cm^2 area of PMTs. The small active area is often a serious limitation in applications based on SPADs.

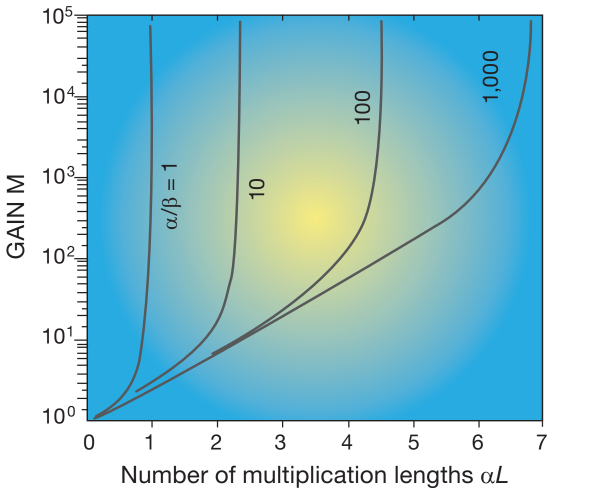
A big step forward happened in 2000, when V. Saveliev and V. Golovin realized that the statistics of the radiation could help circumvent the "Geiger" limitation and make the SPAD almost equivalent to a PMT in linearly detecting multi-photon packets. Indeed, think of an array of many (say *K*) individual small SPADs with the outputs wired in parallel or as a logical OR. In the array, each SPAD will trigger and give a pulse followed by a long recovery time when hit by a photon.

But, by virtue of the Poisson statistical distribution in a uniform beam, any new incoming photon will very likely fall on a different SPAD of the array, in the off-state, whereas the probability of falling on a previously triggered and yet on-state SPAD is negligibly small if *K* is high (say > 100).

In this way, the SPAD array yields a peak pulse amplitude of 1, 2, ..., *N* times the single photon amplitude in response to a packet of *N* photons, and, for example, with *K*=100, the performance is almost the same as a PMT. Also, the sensitive area is now *K* times larger than the individual SPAD; areas of over 1 cm^2 have been reported.

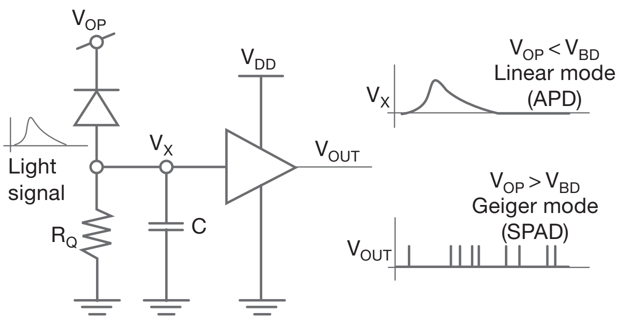
Within the past two decades, as researchers have looked to shorten dead time and integrate SPADs into an array of pixels, a new problem has emerged. By integrating circuits around the SPAD sensitive area with a common planar epitaxial technology, much space is used up by ancillary electronics, and it is subtracted to the function of detection, thus impairing the system efficiency. If $\rho=A_d/A_{pix}$ is the fill factor of the detector area A_d with respect to total pixel area A_{pix}, the effective

[Preferred range for APD operation]



The gain *M* of the avalanche photodiode (APD) plotted vs. α*L* (α=ionization coefficient, *L*= width of multiplication region) has first a Townsend-like dependence as in a gas discharge (the linear increase in log scale). This is the preferred range for linear operation of the APD. When the optimal value *M*=α/β (ratio of the ionization coefficient of electron and hole) is exceeded, *M* increases asymptotically to infinity because of positive feedback of electron and hole ionization. Biasing the APD at a value in excess of the asymptote results in an infinity gain—the Geiger mode regime of operation of the APD, in which each photoelectron may trigger an avalanche terminated only by the charge limitation of the external circuit, and a pulse is obtained for any photon detected.

[Biasing the APD]


V_{OP}
V_{DD}
V_X
V_{OUT}
Light signal
R_Q
C
$V_{OP} < V_{BD}$ Linear mode (APD)
V_X
$V_{OP} > V_{BD}$ Geiger mode (SPAD)
V_{OUT}

The APD is biased through a R_Q-C group to battery V_{OP}. If VOP is less than the infinity-gain voltage V_{BD}, the APD works in the linear regime, supplying an output voltage replica of the input light signal; if V_{OP} is larger than V_{BD}, the device works in the Geiger mode, and the output voltage is a sequence of pulses (after a comparator), one for each new photon detected. In this mode, the APD is known as a single-photon avalanche diode. The single photon output pulse has a time constant R_QC and there is a dead time (of the order of R_QC) to recover in the off state for the next photon detection.

quantum efficiency of the detector is decreased from η to ηρ, an unacceptable loss, usually, if A_{pix}>> A_d.

Again, optics come to rescue: We can fully recover the loss ρ by trading the increase of apparent area *A* with a decrease of the solid angle of acceptance of the detector, given that the acceptance AΩ is an invariant. Indeed, we can use a microlens of area A_{pix} focusing the incoming light (arriving from the objective lens of the system) in a smaller detector area A_d, provided the π steradian) solid angle of the detector field-of-view is reduced to π/C, where C= A_{pix}/A_d is the concentration factor. Of course, we need an array of microlenses, each of which is aligned with its SPAD photosensitive area. Fortunately, any one of several proposed microlens technology would apply to SPAD arrays.

SPADs in multi-pixel format CMOS technologies have advanced quickly in the past 40 years to reach the nanometer-scale feature sizes of today. Thus, the first step toward creating high-performance ancillary electronics on-pixel is to develop a SPAD in deep-submicron CMOS technology.

The challenge of migrating SPADs from submicron to deep-submicron technology is to ensure that the multiplication region can operate in Geiger mode and still exhibit a reasonably high photon detection probability. At the same time, dark counts must be kept to a minimum.

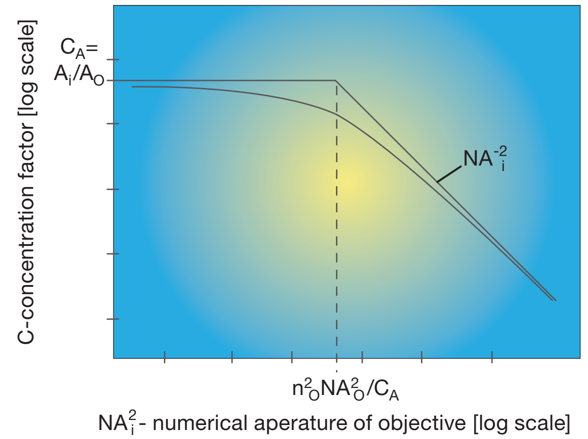
Preventing premature edge breakdown can be done in a number of ways. Among the most used in planar technologies is a combination of shallow- or medium-depth implants, which are generally available in standard CMOS technologies. The goal is to reduce the electric field at the perimeter of the p-n junction to ensure that only the center area is the preferential place for an avalanche.

Since our first CMOS SPAD imagers designed in 2003, we have followed up matching Moore's Law (but with a few generations of lag) moving to submicron and deep-submicron CMOS processes over recent years. To maximize miniaturization, we have introduced the use of p-surrounded shallow trench isolation (pSTI) as an effective and miniaturized guard ring for premature edge breakdown. Highly doped layers near the STI reduce the mean free path for minority carriers generated at the STI interface. More lightly doped layers are used near the multiplication region, thus providing the desired electric field reduction.

While optical gain is virtually infinite, the probability of a photocharge triggering an avalanche is less than 1; it is generally dependent on the voltage in excess of breakdown—which is known as excess bias voltage—and the wavelength of impinging photons. This probability, known as photon detection probability, in CMOS SPADs is usually maximal at 400 to 550 nm and covers the entire visible spectrum.

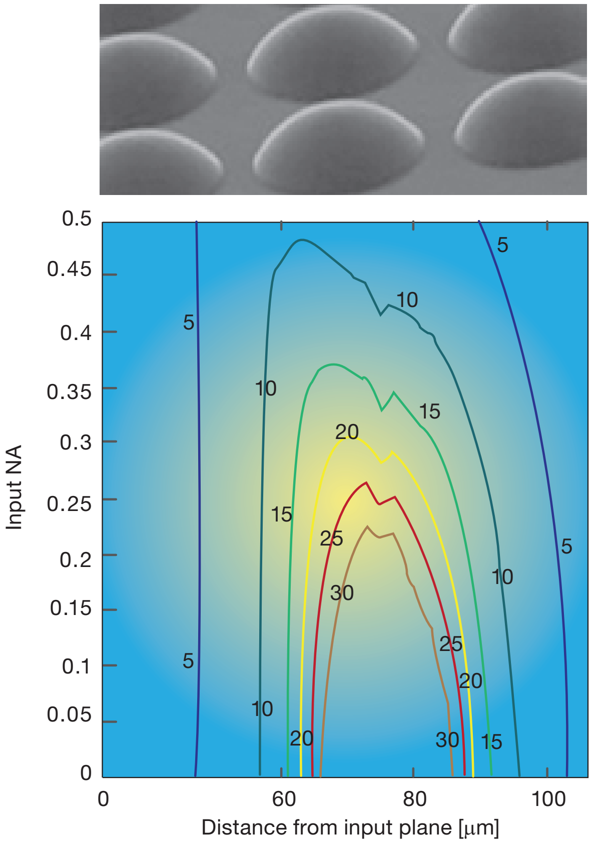
In SPADs, the avalanche must be quenched immediately once it is created to prevent destruction of the device. Quenching and recharge may be achieved with passive and active methods. One of the simplest passive methods is to use a ballast resistance. More complex approaches, involving

[Using a lens to recover loss in a SPAD]



The concentration factor C that can be obtained by a simple lens is limited by the ratio of receiving and collecting areas at small apertures. Then it rolls off as the inverse square of the numerical aperture (NA) of input rays.

[Concentration factor plotted in a NA-z diagram]



(Top) Microphotograph of lens array fabricated by polymer mold replica. (Bottom) The (experimental) concentration factor plotted in a NA-z diagram for a 6-μm diameter receiver and a collection diameter of 50 concentration factor plotted in a NA-z diagram.

pull-up or pull-down transistors controlled by a feedback loop, are indicated when the detector is kept large, thus causing a large parasitic capacitance C, or when the dead time must be kept within precise limits.

Noise can also be a problem. Noise is characterized in terms of the average frequency of the generation of Geiger pulses, which is known as the dark count rate. This rate is caused by thermal or tunneling-generated carriers and by carriers formed by using a combination of the two mechanisms. In deep-submicron CMOS processes, due to the high doping levels of shallow implants, tunneling is more prominent than in less advanced processes. Hence, specific measures besides the use of pSTI can be taken to reduce dark count levels. A SPAD we implemented in 130-nm CMOS technology exhibits a record dark count rate of a few tens of hertz, thanks to modified doping profiles in the multiplication region.

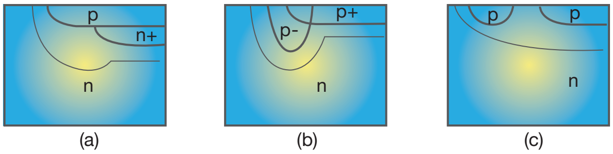
Once one develops a SPAD with satisfactory performance in the target CMOS process, one must couple it with the appropriate ancillary circuitry that implements the desired functionality. Such circuitry is defined by the application; generally it involves single-photon counting, whether time-correlated or not. This is due to the exceptional properties of SPADs in detecting photon time-of-arrival with picosecond accuracy. Thus, we decided to implement time discrimination on-chip even in the first SPAD imagers, leading to the first SPAD imager with an embedded array of time-to-digital converters in 2008. Time-to-digital converters are fast chronometers used to freeze and measure the time-of-arrival of photons with high resolution, typically in the picoseconds.

The MEGAFRAME consortium was created to bring this concept even further and to design SPADs with complex ancillary electronics at the pixel level. The team includes researchers from EPFL (École Polytechnique Fédérale de Lausanne) in Switzerland, the University of Pavia and FBK (Fondazione Bruno Kessler) in Italy, and the University of Edinburgh and ST Microelectronics Edinburgh in Scotland. The consortium built a pixel comprising a SPAD, gating functionality, a time-uncorrelated photon-counting facility, and a chronometer for time-correlated single-photon counting.

Incidentally, the gating and quenching strategy used in the chip corresponds to a technique proposed by Niclass and Charbon in 2009. In this approach, active quenching is used not to reduce dead time, thereby potentially introducing increased afterpulsing, but to control it to an exact time, so as to minimize afterpulsing to a known, controllable value.

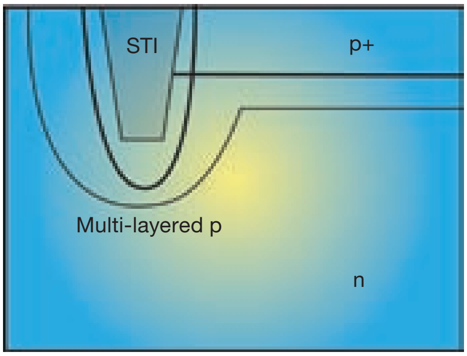
The main challenge of building such a pixel was to squeeze a chronometer with a resolution of better than 100 picoseconds (ps) in an area of less than 50 × 50 μm^2, requiring less than 100 μA bias current, and with a uniformity of less than 200 ps. In addition, the infrastructure of the sensor needed to ensure stand-alone operation with as many different applications as envisioned.

[Cross-section of a SPAD implemented in a planar process]



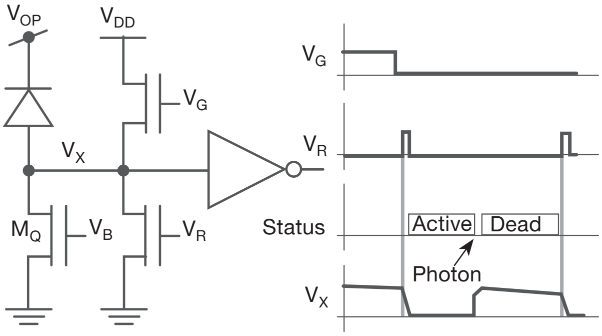
Three techniques are shown to prevent premature edge breakdown. (a) and (b): The edge electric field is reduced by using doping level grading. In structure (c), a lightly doped p-enclave is introduced between two more heavily doped regions.

[pSTI-based PEB prevention process]



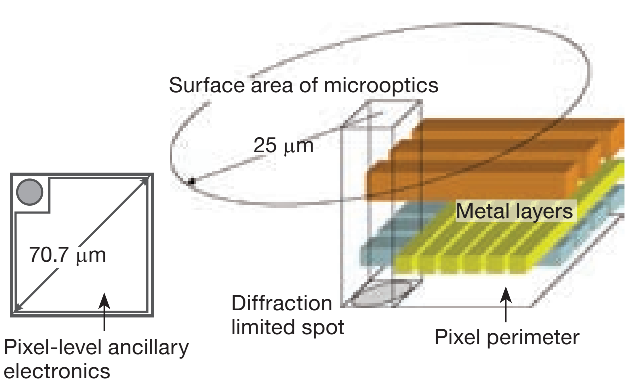
Free minority carriers generated in STI are absorbed by a multiple layer of variably doped regions.

[Gating and quenching mechanism]



Gating and quenching mechanism to control afterpulsing. V_G causes the pn junction of the SPAD to be brought below breakdown. When V_G returns to zero, V_R is used to force the p-n junction again above breakdown in a controlled time. When above the breakdown, the SPAD becomes active again. Upon photon detection, passive quenching is performed via MOS transistor M_Q that is controlled by bias voltage V_b. After quenching is complete, the same transistor begins a slow current-source-like recharge that is not sufficient to make the device active again. Dead time is thus forced upon the SPAD until sufficient relaxation time has elapsed, when the recharge is triggered again by V_R. Bias voltage V_b is used to ensure sufficiently high resistance in M_Q during quenching and sufficiently low current for a long passive recharge to avoid interfering with the active recharge circuit.

[A MEGAFRAME pixel]



(Left) The SPAD is placed at a corner, and its active area is approximately 4 percent of the pixel surface. (Right) Metal geometry and micro-optics arrangement in typical MEGAFRAME pixel.

Two different architectures were proposed for the pixel: one based on a time-to-digital converter and one on time-to-amplitude converter. Both occupy approximately 96 percent of the surface of the pixel to achieve the target performance. Horizontal and vertical metal lines are used in this design for data readout and control lines, thus posing an additional limitation to any micro-optical system placed on top of this pixel.

In our program, the microoptics array is based on an array of 32 × 32 microlenses overlaid on 32 × 32 pixels. The noise and sensitivity of the SPADs are consistent with the data published by Gersbach et al. The microoptics array ensures a concentration factor C of 20 at an input NA of 0.35, thus boosting the fill factor from 3 percent to an effective value of 60 percent. Thus, the product ρPDP becomes comparable to that of conventional detectors.

Applications of SPAD arrays and SPAD imagers

$g^{(2)}$ imaging

In Young's double slit interferometer two points "1" and "2" separated in space produce a fringe pattern with intensity

$$I_1 + I_2 + 2\sqrt{I_1 I_2}\,|g^{(1)}(x_{12},\tau)|\cos(\Delta\varphi_{12})\ ,$$

where $\Delta\varphi_{12}$ is the phase difference and τ the propagation time difference, while $g^{(1)}(x_{12},\tau)$ is the first-order correlation function between the two beams. When the light-state distribution is known, then the phase difference is the only required parameter to define the relationship between the field points. However, with unknown field distributions, measuring $g^{(1)}(x_{12},\tau)$ alone yields ambiguous results.

Thus, entagled photons, incoherent light or coherent and thermal light may be indistinguishable. This is a problem in many experiments—which is why researchers use another quantity, the second-order correlation function $g^{(2)}(x_{12},\tau)$, in these experiments. The second-order correlation function is defined as

$$g^{(2)}(x_{12},\tau) = \frac{\langle I_1(t) I_2(t+\tau)\rangle}{\langle I_1(t)\rangle\langle I_2(t+\tau)\rangle}\ .$$

Usually, $g^{(2)}(x_{12},\tau)$ is associated with a Hanbury-Brown-Twiss (HBT) type of interferometer, and the measurement is performed on the image plane. In order to compute $g^{(2)}(x_{12},\tau)$ over a large number of pairs of points in space, one must determine the time-of-arrival of each photon in the plane simultaneously. Using an array of SPADs placed in the image plane of an HBT setup, we showed that this technique was feasible over a small number of detectors. However, with larger arrays and more advanced on-pixel electronics, localization techniques for the computation of $g^{(2)}(x_{12},\tau)$ will be necessary to avoid the explosion in computation.

Fluorescence correlation spectroscopy

This technique is often used to measure transitional diffusion coefficients of macromolecules, to count fluorescent transient molecules, or to determine the molecular composition of a fluid being forced through a bottleneck or gap. In fluorescence-correlation spectroscopy, a femto-liter volume is exposed to a highly focused laser beam that causes the molecules in it to emit light in a well-defined spectrum and with a time-response that depends on the modality of the movement of the molecules to and from the detection volume.

For normal gap sizes, and most molecules, sub-nanosecond time resolutions are necessary. In addition, the availability of multi-pixel sensors with simultaneous, parallel operation enables better characterization of the diffusion processes underlying the experimental setup.

Lifetime imaging

Among time-correlated imaging methods, time-correlated single photon counting is perhaps one of the most used in bio-imaging. Multiple exposures are used to reconstruct the statistical response of matter to sharp and powerful light pulses. The study of calcium at the cellular level has made intensive use of fluorescent Ca^{2+} indicator dyes. Some of the heavily used dyes or fluorophores are Oregon Green Bapta-1 (OGB-1), green fluorescent protein and many others. Calcium concentration can be determined precisely by measuring the lifetime of the response of the corresponding fluorophore, when excited at a given wavelength, using fluorescence lifetime imaging microscopy. (Lifetime is generally characterized using fluorescence lifetime imaging microscopy, or FLIM.) There are several types of FLIM-based imaging techniques depending on how lifetime is characterized or based on the excitation mode.

In another application, we used a two-photon FLIM setup based on an SPAD array capable of a time resolution of 79 ps

1047-6938/10/0??/0040/6-$15.00 ©OSA

In 2000, V. Saveliev and V. Golovin realized that the statistics of the radiation could help circumvent the "Geiger" limitation and make the SPAD almost equivalent to a PMT in linearly detecting multi-photon packets.

at a system level. The sensor made it possible to fit the lifetime dependency of OGB-1 on Ca^{2+} using a triple exponential fit. Unlike previous approaches that exploit detectors with lower resolutions, our model required no calibration factors, nor corrections of any kind, thus proving the robustness of the measurement system.

Time-of-flight imaging

Time-of-flight is the time a light ray takes to propagate between two points in three-dimensional space. Several applications require a precise measurement of the time-of-flight to image particular properties of targets and environments. In 3-D imaging, for example, pulsed or continuously modulated light is used to determine the distance between the sensor and a reflecting target.

The distance is computed in every pixel using the relation $d = c/2$ TOF, where c is the speed of light and TOF the time of flight. Again, for a resolution of 1 mm, a time resolution of at least 6.6 ps is necessary, whereas statistical methods may be used to relax the resolution of a single measurement. A pulsed laser or LED source is used in combination with on-chip chronometers that convert two subsequent pulses onto a time measurement, and thus a distance. Depth maps can be used in human-computer interfaces as well as applications in security, transportation and a variety of other areas.

In positron emission tomography, the exact location of positron emission is found by monitoring all gamma radiation that reaches a pair of detectors on an axis at exactly the same time and then cross-correlate all estimated arrival times. The emission loci may be derived by measuring the time-of-flight of the particle with respect to a reference point of known coordinates.

Conclusions

With the introduction of CMOS single-photon avalanche diodes, it is possible to achieve great levels of miniaturization without compromising time resolution and overall speed. Not only are large arrays of photon counters now possible, but a very high dynamic range and timing accuracy have become feasible as well. Thanks to these advances, applications that require time-resolved single photon detection are now possible using low-cost CMOS detectors.

The main drawback of designing complex pixels is the increasing dominance of on-pixel ancillary area over a sensitive area; this is causing concern over the excessive loss of photons. Optics-based techniques are becoming the solution of choice in SPAD image sensors.

The authors wish to acknowledge the MEGAFRAME consortium, ST Microelectronics and Crolles.

OSA Member Edoardo Charbon (e.charbon@tudelft.nl) is with the Delft University of Technology in the Netherlands, formerly with EPFL in Switzerland, and **Silvano Donati** (silvano.donati@unipv.it) is with the University of Pavia in Italy.

[References and Resources]

>> The MEGAFRAME consortium: www.megaframe.eu.

>> S. Cova et al. Rev. Sci. Instr. **60**, 1104-10 (1989).

>> V. Saveliev and V. Golovin. Nuclear Instrum. Methods. A, **442**, 223-9 (2000).

>> S. Donati. *Photodetectors: Devices, Circuits and Applications*, Prentice Hall, Upper Saddle River, N.J., U.S.A. (2000).

>> J. Fisher et al. Opt. Lett. **29**(1), 71-3 (2004).

>> C. Niclass and E. Charbon. Proc. ISSCC'05, San Francisco (Feb. 6-10, 2005).

>> C. Niclass et al. J. Solid-State Circuits **40**(9), 1847-54 (2005).

>> W.T. Welford et al. "Optics of Nonimaging Concentrators," Academic Press, N.Y., U.S.A. (2005).

>> S. Donati et al. Proc. WFOPC'07, Taipei (Dec. 4-7, 2007).

>> S. Donati et al. Opt. Express **15**, 18066-74 (2007).

>> C. Niclass et al. J. Sel. Top. Quantum Electron. **13**(4), 863-9 (2007).

>> E. Charbon. J. Phys. D **41**, 9 (2008).

>> V. Fakhfouri et al. "Inkjet Printing of SU-8; A Case Study for Microlenses," MEMS, 407-10 (2008).

>> J-H Lee et al. Opt. Express **16**, 11044-51 (2008).

>> C. Niclass. "Single-Photon Image Sensors in CMOS: Picosecond Resolution for Three-Dimensional Imaging," Ph.D. Thesis 4161, EPFL (2008).

>> C. Niclass et al. J. Solid-State Circuits **43**(12), 2977-89 (2008).

>> D.L. Boiko et al. Opt. Express **17**, 15087-103 (2009).

>> M. Gersbach et al. Opt. Lett. **34**(3), 362-4 (2009).

>> M. Gersbach et al. Solid-State Electronics **53**(7), 803-8 (2009).

>> C. Niclass et al. J. Solid-State Circuits **44**(7), 1977-89 (2009).

>> J. Richardson et al. Proc. Custom Integrated Circuits Conference, (Sep. 13-16, 2009).

>> D. Stoppa et al. Proc. European Solid-State Device Conference, (Sep. 14-18, 2009).