Abstract—We report on the first implementation of a single photon avalanche diode (SPAD) in 130 nm complementary metal–oxide–semiconductor (CMOS) technology. The SPAD is fabricated as p+/n-well junction with octagonal shape. A guard ring of p-well around the p+ anode is used to prevent premature discharge. To investigate the dynamics of the new device, both active and passive quenching methods have been used. Single photon detection is achieved by sensing the avalanche using a fast comparator. The SPAD exhibits a maximum photon detection probability of 41% and a typical dark count rate of 100 kHz at room temperature. Thanks to its timing resolution of 144 ps full-width at half-maximum (FWHM), the SPAD has several uses in disparate disciplines, including medical imaging, 3-D vision, biophotonics, low-light illumination imaging, etc.

Index Terms—CMOS single photon detector, Geiger mode of operation, single photon avalanche diode, SPAD.

I. INTRODUCTION

The world of experimental sciences in biology, chemistry, and physics has in the last years tightened practically all performance requirements for most sensors. In addition, commercial applications are creating the demand for unconventional imaging techniques to achieve, for example, compact 3D cameras and high dynamic range vision. In this context, important advances have been made in optical imaging as well. Following feature size evolution, pixels have generally shrunk, and consequently, image size has expanded.

Imaging technology has advanced in speed as well. Charge-coupled device (CCD) and complementary metal–oxide–semiconductor (CMOS) active pixel sensor (APS) architectures have reached frame rates up to 1 Mbps [1] in burst mode and up to 250 kfps in continuous mode but with impractically small frame sizes [2]. Nonetheless, the number of scientific applications making use of these cameras has exploded, mostly in fluidics, physics, and biochemistry [3]–[6], while fields such as neuroscience and biomedicine are becoming increasingly dependent on high dynamic range and fast imaging [7]–[9].

Along with high frame rates, researchers have turned their attention to high timing resolution [10]. In this respect, relatively new imaging techniques involving low photodetection timing uncertainty have been proposed aimed, for example, at computing emission decay in fluorescent molecules. Other currently used time-correlated techniques are single and multispectral fluorescence lifetime imaging (FLIM) [11], [12], fluorescence correlation spectroscopy (FCS) [13], Förster resonance energy transfer (FRET) [14], etc.

Another important class of problems that use high timing resolution aims at computing time-of-flight (TOF) of a modulated or pulsed light source. Examples of applications based on TOF include rangefinding, 3-D vision, LIDAR, etc. While in TOF problems, both CCD and CMOS technologies have been used with some success [15]–[17], in bioimaging Photomultiplier Tubes (PMTs) remain the sensor of choice. This is due to the requirement of tens or hundreds of picoseconds timing accuracy and single photon sensitivity that PMT can ensure. In addition to their sensitivity to single photons, PMTs have several advantages in terms of noise and timing. A major disadvantage is cost, size, and the fact that large arrays of PMTs are impractical. More compact microchannel plate (MCP) devices have also been fabricated [18]. However, these devices generally require relatively bulky vacuum chamber apparatuses.

For reasons of cost and miniaturization, a solid state solution for single photon detection is highly desirable. In addition, the potential for massively parallel single photon detection could enable today novel applications, involving, for example, extensive bioanalysis microarrays with thousands of reactors.

Solid state single photon detectors have been known for decades [19]. Nonetheless, researchers have been successful designing fully integrated single photon detectors in CMOS only recently [20]–[22]. More recently, the emergence of multipixel arrays combined with time-correlated single photon counting (TCSPC) technique, has accelerated the international effort in designing single photon detectors in CMOS and other technologies [23]–[25].

CMOS single photon detectors are based on a device known as single photon avalanche diode (SPAD). The effectiveness of SPADs has been demonstrated in a number of applications, including rangefinding [24], [26], [27] fluorescence detection [21], FCS [28], high-speed imaging [29], one- [19] and two-photon [30] FLIM, and latchup/leakage test [31]. In some of these systems however, pitch and array size are still limited due to the technologies being used, hence the push to design SPADs in deep-submicrometer (DSM) technologies. DSM implementations are expected to enable larger arrays and more functionality on chip, thus relaxing the input-out I/O data throughput requirements, simplifying the overall design, and slashing power dissipation.
The main step towards this goal is the design and optimization of a DSM pixel, and, in particular, of its core, the SPAD. In this paper, we present the design details of a SPAD implemented in 130 nm CMOS technology and its characterization. As a result, it does not require any post-processing steps nor hybrid technologies such as 3-D integration.

The SPAD presented here is amenable to the design of large arrays and, in principle, it enables the choice of any readout architectures proposed in the literature for SPAD arrays [23], [27], [32], [33]. Due to the reduced breakdown voltages, the structure is interesting in the context of applications where only a few supply voltages are available and where power dissipation is a critical factor. Advanced CMOS technology provides a level of miniaturization that is important to design smaller front-end circuits. Thus fill factor can be improved and/or new functionality can be added in SPAD arrays.

The paper is organized as follows: SPAD principles are outlined in Section II. The design of the proposed SPAD is described in detail in Section III. Experimental results are presented and discussed in Section IV.

II. SINGLE PHOTON AVALANCHE DIODES

A SPAD is generally implemented as a p-n junction biased above breakdown. In this regime of operation, known as Geiger mode, photogenerated carriers may cause an avalanche by impact ionization. The number of carriers generated as a result of the absorption of a single photon determines the optical gain of the device, which in the case of SPADs may be virtually infinite.

An avalanche in the multiplication region causes a current pulse of appreciable amplitude but it needs to be stopped. This is generally accomplished via a quenching circuit. The avalanche current pulse may be converted into a digital voltage pulse by proper design techniques, thus enabling the direct conversion of photons onto digital signals compatible with CMOS low-voltage circuitries.

There exist several types of quenching circuits, divided in two main categories: active quenching and passive quenching. In active quenching, the avalanche is sensed and a feedback circuit provides a mechanism to force the reverse bias of the p-n junction below breakdown. The same circuit is generally used to actively recharge the device to its initial state, above breakdown, so as to enable the next detection cycle.

In passive quenching, the avalanche current is used to directly act on the reverse bias voltage by lowering it towards breakdown voltage, which eventually quenches the current. If this is achieved, for example, using a resistance in series to the photodiode, the effective capacitance of the junction must be passively recharged through the quenching resistance.

In SPADs, the detection cycle requires a total time known as dead time, which includes quenching and recharge. The dead time is also responsible for the upper limit of photon flux detectable by a SPAD.

The noise performance of SPADs is mainly characterized by spurious pulses in the dark, i.e., dark counts. Dark counts, quantified in terms of a frequency, or dark count rate (DCR), are caused by thermally or tunneling generated carriers [34]. The relative impact of the two effects can generally be evidenced with device analysis as a function of temperature. DCR is also strongly dependent upon the excess bias voltage, i.e., the voltage in excess of breakdown at which the SPAD is biased.

The sensitivity is characterized in SPADs as the probability of a photon impinging the device’s surface to cause a pulse. It is known as photon detection probability (PDP) and it is a strong function of photon wavelength and excess bias voltage.

The uncertainty of the time delay between photon impingement and the leading edge of the pulse generated by the sensor is known in the literature as timing resolution or timing jitter. In a small SPAD, the timing jitter mainly depends on the time a photogenerated carrier requires to be swept out of the absorption zone into the multiplication region. In large devices, timing jitter is also caused due to the fluctuations of the avalanche propagation across the active area.

Trapping centers in the multiplication region tend to capture carriers generated during an avalanche. As trapping centers are characterized by finite lifetimes, trapped carriers are released at a random later time, thus potentially retriggering a subsequent avalanche [34]. Such phenomenon causes so-called afterpulses, i.e., spurious pulses correlated to previous Geiger pulses. The parameter characterizing this effect is known as afterpulsing probability, or probability of afterpulsing, and it is also a function of the number of carriers involved in an avalanche, which in turn depends on the SPADs parasitic capacitance. In addition to the correlated noise introduced by afterpulsing, this phenomenon may limit the maximum rate of detectable photons as one photon may generate in average more than a single event.

III. SPAD STRUCTURE AND DESIGN CONSIDERATIONS

The design of avalanche photodiodes in DSM CMOS technology involves additional challenges than in larger feature size technologies. In order to operate in the so-called Geiger mode, a SPAD requires a design configuration that supports a planar and uniform multiplication region extending laterally and vertically underneath the area of the SPAD as much as possible [34]. Even though this requirement is mandatory to allow the creation of a reasonably large photosensitive or active area, it is not sufficient in general. For example, [35] reports the design of a SPAD fabricated in 0.18 μm CMOS that implements a planar multiplication region, according to simulations, but exhibits DCR levels of 1 MHz, unacceptable in most applications.

Noise performance becomes a major issue for SPADs in deep-submicrometer CMOS technologies. It is therefore, very important to keep a strict notion of noise performance when assessing potential design structures. The main sources of noise in SPADs are more significant in deep-submicrometer due to (i) higher doping levels, (ii) reduced annealing and drive-in diffusion steps, and (iii) the presence of shallow-trench isolation (STI).

Higher doping levels increase the effects of tunneling-induced dark counts and increase the parasitic capacitance. The increase of parasitic capacitance increases the number of carriers involved in an avalanche discharge and thus worsens afterpulsing probability.
Driven by miniaturization, state of the art fabrication processes reduce the strength and duration of annealing and drive-in diffusion steps to a minimum. The lack of effective annealing steps increases the concentration of impurities that introduce carrier recombination–generation and trapping centers, thus worsening both thermally generated dark counts and afterpulsing effects [34].

At and below the 0.25 \( \mu \text{m} \) mark, standard CMOS processes feature STI compulsorily. It is known that STI may dramatically increase the density of deep-level carrier generation centers at its interface [36], [37]. When an STI is close to or in contact with the multiplication region of a SPAD, such as in [35], one can expect high dark count rates.

Unfortunately, very often designers do not have enough flexibility to change or adapt a process parameter in order to better fit the SPAD requirements in CMOS technology. In order to address the issues described above, designers are left with a number of design layers, models and rules. It was the aim of this work to design, test, and characterize high-quality SPADs based on an existing and fixed 130 nm CMOS technology. This approach is obviously beneficial in terms of design time and fabrication costs.

Fig. 1 shows the cross section of the proposed SPAD. It consists of a p+ anode within an n-well cathode where p+ and n-well are the implantations of source/drain and bulk, respectively, of standard 1.2 V PMOS transistors. This configuration allows for a full isolation of the p+ anode from the p-substrate. In addition, the configuration enables coupling relatively high bias voltages necessary in SPADs to low-voltage CMOS logic, similarly to [20] and [23].

The planar multiplication region was enabled by means of a p-well guard ring [20], where p-well is the bulk of isolated 1.2 V NMOS transistors. A useful feature of this technology is the availability of a buried n-type isolation layer that allows for a full isolation of p-well within n-well from p-substrate. This layer was used to prevent a punchthrough of the p-well guard ring to p-substrate. The combination of n-well and buried n-isolation layer was the lowest doping concentration feasible in this technology for the cathode.

A major improvement in this design is the physical separation of the STI interface from the SPAD multiplication region, thus having a beneficial impact on DCR. In standard DSM CMOS, it is not possible to prevent STI by means of a drawn layer. As a general rule, STI is etched everywhere so that all the p+ and n+ implantations are surrounded by STI to improve isolation. It is possible, however, to draw a polysilicon gate of a standard transistor that represents a stop mask for n+ and p+ implantations. STI can therefore be effectively separated from the surrounding of the anode by drawing a superposition of polysilicon, thin-gate-oxide, p+, and diffusion layers around the p+ anode. In order to prevent the formation of a high-electric field within the thin-gate-oxide layer, the polysilicon gate is kept at the same potential as the p+ and p-well layers, ensured by ohmic contacts.

Since the polysilicon gate prevents the p+ to be implanted, the result of the fabrication process is a p-well extension of p+ completely free of STI, whose extension can be adjusted as desired. Around the p-well guard ring, there is still a STI ring. This STI interface, in particular at the depletion region between the p-well guard ring and n-well cathode, may induce a large density of generation centers. Nonetheless, the p-well guard ring lowers the electric field around the SPAD sufficiently to prevent impact ionization but it is enough to collect most of the carriers generated at the STI/p-well interface. As a result, this structure allows a small parasitic current to flow from cathode to anode without triggering avalanche events, thus reducing DCR.

IV. EXPERIMENTAL RESULTS

The photomicrograph of the proposed device is shown in Fig. 2. The structures visible in the figure include the octagonal anode, guard ring, and metal interconnect. The additional function of the metal is that of preventing the guard ring to be exposed to light for characterization purposes. The anode measures 10 \( \mu \text{m} \) in the picture. 30 \( \mu \text{m} \) structures were also integrated in the same technology for characterization purposes.

The diode was tested in a number of ways. First, the \( I–V \) characteristic was measured statically using a standard semiconductor analyzer. Fig. 3 shows the \( I–V \) characteristics of the diode in reverse bias. The picture shows that the reverse current close to breakdown voltage approaches 600 pA. This relatively large current would suggest that DCR tends to be high. For instance, if we suppose that all the carriers were collected by the multiplication region, the device would not properly operate in
Geiger mode as its DCR would be of the order of 3–4 GHz. In this section, it will be shown that the structure properly operates in Geiger mode and exhibits acceptable levels of DCR. As described in Section III, most of the reverse current is expected to be generated at the periphery of the SPAD, at the STI/p-well/n-well interface, where impact ionization is prevented by the p-well guard ring.

The diode was operated in Geiger mode using both passive and active quenching circuitries. The schematic setup of the passive quenching configuration is shown in Fig. 4.

The 20-kΩ quenching resistance $R_Q$, placed at the anode of the p-n junction, causes an increase of its potential in case of avalanche. If the reverse bias voltage across the junction decreases towards breakdown voltage, the avalanche current is reduced to a level of the order of tens of microamperes and eventually stops. Avalanche quenching is followed by an exponential recharge to allow the voltage across the junction to return to its initial value of $V_{OP}$. This voltage satisfies the following condition

$$V_{OP} = |V_{BD}| + V_e$$  \hspace{1cm} (1)

where $V_{BD}$, and $V_e$ are the breakdown and excess bias voltage, respectively.

The plot in Fig. 5 shows the recharge phase of the probed voltage as a function of time for different values of $V_{OP}$. The simple exponential behavior is due to the $RC$ recharge. $R$ accounts for the resistive path to ground and $C$ for the overall capacitance at the probing node.

Due to the fact that this device does not have integrated quenching circuitry, the term $C$ is dominated by the parasitic capacitance of external components. It has been estimated to be 10 pF, a factor 70–100 larger than the expected SPAD junction capacitance. The dead time under this condition is estimated to be 450 ns. As described in Section II, afterpulsing probability depends independently on dead time, due to trap lifetimes, and on the parasitic capacitance as it increases the number of carries traversing the multiplication region, thus filling up traps. Thus, a characterization of afterpulsing probability under this condition is irrelevant, since it gives no insight on the true potential of the device and of its internal capacitance when the SPAD is monolithically integrated with its quenching and recharge circuit. We assume that afterpulsing characterization under the present condition would be incorrect and thus irrelevant.

In order to precisely investigate DCR and PDP in this work independently of dead time and afterpulsing effects, an alternative setup involving the use of an external gated active recharge circuit combined with TCSPC was used. This technique is often used in the characterization of III–V SPADs, which exhibit significantly higher DCR and afterpulsing effects [39]. In most active quenching and recharge setups, an active circuit replaces the quenching resistance, thus allowing one to reduce the recharge time to a few tens of nanoseconds. Our experimental setup is based on a commercially available gated active recharge circuit [38] and is described as follows.

VOP is maintained below $V_{BD}$ at the beginning of each event measurement cycle. VOP is then quickly increased to its nominal level, according to Equation (1), so as to recharge the SPAD. The time interval between the SPAD recharge signal and the moment a first Geiger event occurs is measured using a high precision time-to-digital converter (TDC). VOP is subsequently kept below $V_{BD}$ during a hold-off time of the order of 500 µs. This hold-off time is chosen large to prevent any afterpulse. As this measurement cycle is repeated a large number of times, a histogram is built conforming to the TCSPC technique. The resulting histogram shows an exponential decay similarly to a typical florescence lifetime measurement. The inverse of the mean value of the histogram provides the desired counting rate. Any timing offset between full SPAD recharge and Geiger pulse
leading edge is removed prior to computing the counting rate. The active recharge circuit conveniently performs fast active recharge and also provides a trigger signal used to compute time interval measurements as described above. As detector dead time and afterpulsing do not impair the measurement even at high counting rates, this technique is used to measure DCR as well as PDP.

In order to correctly characterize the measurements presented hereafter, the breakdown voltage was firstly measured for the structure as a function of temperature. Hence, VOP was set for a given temperature to reflect the correct excess bias voltage according to (1). Fig. 6 shows a plot of the breakdown voltage as a function of temperature.

The PDP was measured for two excess bias voltages for the entire spectrum of interest (350–1000 nm). Fig. 7 shows a plot of the PDP at room temperature. PDP outperformed our expectations, as measurements showed values in the range of previously reported SPADs in near-micrometer CMOS technologies [20]–[23], whose multiplication regions are wider and deeper. The result of a shallower multiplication region was a shift of the maximum of detection probability from 550 nm in [20] to 450 nm in this work. We believe that this relatively good PDP performance partially resulted from the use of enhanced dielectrics for optical detection available in this imaging CMOS technology. In Fig. 2, it is possible to notice a darker region in the middle of the SPAD, where optimized dielectrics was used, if compared to the remaining area of the picture, where only partial optimization was used. This darker region suggests that the light reflection coefficient at the center of the SPAD was noticeably lower than it would have been if we utilized non-optimized passivation layers. Notice that, as DCR measurements were performed previously to the PDP characterization, the mean value of DCR contribution was suppressed from each counting rate used in the measurement of PDP. As a result, DCR is not responsible for an artificially increased PDP.

In near-micrometer CMOS SPAD implementations, DCR can be as low as a few tens of hertz [20]–[23] and it is a strong function of temperature and of excess bias voltage. Fig. 8 shows a plot of DCR as a function of temperature for two excess bias voltages, measured using the TCSPC method as described above. Besides its higher absolute values if compared to [20], DCR also exhibits weaker dependence on temperature. This suggests that DCR has a non-negligible tunneling contribution [40]. This behavior was expected due to relatively higher doping levels of both p+ and n-well layers available in this CMOS technology, as described in Section III. Fig. 9 shows a plot of DCR as a function of $V_e$ for four different temperatures.
Such cases, in order to improve PDP and increase overall signal-time-of-flight principle [16], [23], [24], is in general dominated. For instance, noise in a 3-D image sensor, based on the light in a given application, higher levels of DCR may be tolerated. Depending on the amount of parasitic errors in those measurements compared to the TCSPC method are significant for any DCR higher than a few tens of kHz.

As can be seen in Fig. 9, DCR reaches prohibitive levels as $V_e$ is chosen higher than 2 V. Depending on the amount of parasitic light in a given application, higher levels of DCR may be tolerated. For instance, noise in a 3-D image sensor, based on the time-of-flight principle [16], [23], [24], is in general dominated by the parasitic background light when it operates outdoor. In such cases, in order to improve PDP and increase overall signal-to-noise ratio, higher values of $V_e$ may be recommended.

Timing jitter was characterized in this work also based on the TCSPC technique. A fast laser source with pulse width of 40 ps and repetition rate of 40 MHz emitting a beam with a wavelength of 637 nm was used to illuminate the SPAD. The time interval between the laser output trigger and the leading edge of the SPAD signal, operated with the active recharge circuit, was measured via a high performance oscilloscope operating as a TDC. The oscilloscope, a LeCroy 8600 A, features 20 GS/s and 3 ps of uncertainty. A histogram was built as the time interval measurements yielded photon detection probabilities similar to other CMOS single photon detectors found in the literature. The dark count rate and timing jitter of this device have also been measured at various operating conditions. In the future, arrays of this device will be monolithically integrated with front-end and application circuits, to be used in a number of applications requiring high dynamic range and timing resolution.

V. CONCLUSION

The first single photon avalanche diode implemented in 130-nm CMOS technology is reported. Techniques to fabricate the device using available layers within standard design rules are described in detail. The characterization of the device yielded photon detection probabilities similar to other CMOS single photon detectors found in the literature. The dark count rate and timing jitter of this device have also been measured at various operating conditions. In the future, arrays of this device will be monolithically integrated with front-end and application circuits, to be used in a number of applications requiring high dynamic range and timing resolution.

ACKNOWLEDGMENT

The authors are grateful to A. Rochas of IdQuantique for providing us with the active quenching circuit used in our measurements and M. Lanz for technical support.

REFERENCES


Cristiano Niclass received the M.S. degree in Microtechnology from EPFL in 2003. In May 2003, he joined the Processor Architecture Laboratory (LAP) of EPFL and subsequently the Quantum Architecture Group (AQUA), where he is working toward the Ph.D. degree. His interests include high-speed and low-noise mixed-signal integrated circuits with emphasis on high-performance imaging. He is currently working on the design, implementation, and evaluation of fully integrated image sensors in CMOS and based on single photon avalanche diodes. He is also involved in the design of high-speed and high-resolution data converters implemented in conventional technologies.

Marek Gersbach was born in Durham, NC, in 1982. In 2004, he received the M.Sc. degree in Microtechnology from EPFL. He has since been working towards the Ph.D. degree in Electrical and Computer Engineering at EPFL.

As a research assistant at the Quantum Architecture Group (AQUA), he is currently working on single photon detectors for biological imaging applications such as fluorescence lifetime imaging.

Robert Henderson (S’87–M’89) received the B.Sc. and Ph.D. degrees in electronics and electrical engineering from the University of Glasgow, Glasgow, U.K., in 1986 and 1990, respectively. From 1989 to 1990, he was with the University of Glasgow as a Research Assistant, working in the area of switched capacitor filter design. He was a Research Engineer with the Swiss Centre for Micro-Technology (CSEM), Neuchâtel, Switzerland from 1990 to 1996, where he worked on low power A/D and D/A converters. From 1996 to 2005, he held the position of Principal VLSI Engineer at VLSI Imaging and since then has worked on CMOS Image Sensor (CIS) process technology. He is currently working on the design, implementation, and evaluation of fully integrated image sensors for biological imaging applications such as fluorescence lifetime imaging.

Lindsay Grant received the B.Sc. degree in physics at St. Andrews University in 1984. After an initial year with Electrotech, U.K., as a process engineer he spent 4 years working on BiCMOS technology as a device engineer, with STC Semiconductors in Harlow, Essex, U.K. He then joined Seagate Microelectronics in Livingston, Scotland and worked there for 12 years. In his time at Seagate he held positions in product, device, and process engineering finishing his time as the process section head for photolithography. In 1999 he joined MiroSensors Ltd, electronics, Edinburgh, Scotland, U.K. His research interests are in analog signal processing, CMOS IC Design, imaging, and biosensors. He has published over 35 articles and holds ten patents.

Edoardo Charbon (M’00) received the Diploma from ETH Zurich in 1988, the M.S. degree from UCSD in 1991, and the Ph.D. degree from UC-Berkeley in 1995, all in Electrical Engineering. From 1995 to 2000, he was with Cadence Design Systems, where he was responsible for analog and mixed-signal design automation tools and the architect of the company’s initiative for intellectual property protection. In 2000, he joined Canesta Inc. as its Chief Architect, leading the development of wireless 3-D CMOS image sensors. He has also held the position of Senior VLSI Engineer at VLSI Imaging and since then has worked on CMOS Image Sensor (CIS) process technology. He is currently working on the design, implementation, and evaluation of fully integrated image sensors for biological imaging applications such as fluorescence lifetime imaging.

Dr. Charbon has served as Guest Editor of the TRANSACTIONS ON COMPUTER-AIDED DESIGN OF INTEGRATED CIRCUITS AND SYSTEMS and the JOURNAL OF SOLID STATE CIRCUITS and is currently the Chair of technical committees in ESSCIRC, ICECS, and VLSI-SOC.