A Single Photon Avalanche Diode Array Fabricated in Deep-Submicron CMOS Technology

Cristiano Niclass, Maximilian Sergio, and Edoardo Charbon

Ecole Polytechnique Fédérale de Lausanne (EPFL)
CH-1015 Lausanne, Switzerland

Abstract

We report the first fully integrated single photon avalanche diode array fabricated in 0.35μm CMOS technology. At 25μm, the pixel pitch achieved by this design is the smallest ever reported. Thanks to the level of miniaturization enabled by this design, we were able to build the largest single photon streak camera ever built in any technology, thus proving the scalability of the technology. Applications requiring low noise, high dynamic range, and/or picosecond timing accuracies are the prime candidates of this technology. Examples include bio-imaging at cellular and molecular level, fast optical imaging, single photon telecommunications, 3D cameras, optical rangefinders, LIDAR, and low light level imagers.

1. Introduction

Imaging fast, time-correlated, molecular processes in physics and the life sciences is placing increasing pressure on camera technology. Over the last two decades, conventional imagers based on Charge-Coupled Devices (CCDs) and CMOS Active Pixel Sensor (APS) architectures have consistently achieved improved speed and sensitivity. However, very low photon counts are generally undetectable in these technologies or require deep cooling and highly optimized ultra-low-noise read-out circuitries.

Emerging imaging techniques have recently developed the need for single or low photon count sensitivity. There are numerous examples, where bio-luminescence, assay, and bio-scattering methods, among others, are currently the techniques of choice. In bio-luminescence methods light is emitted as the result of a chemical reaction. The difficulty of detection is given by the extremely low emission intensity, typically in the microlux range [1]. Assay methods, based for example on fluorescence, may be used to determine the potential on a given cell *in vitro* or *in vivo* by means of Voltage-Sensitive Dyes (VSDs). A variety of techniques based on specific VSDs for example are utilized in neural activity analysis and research [2],[3]. The challenge in these methods is usually in the detection of small variability in the light intensity while maintaining high contrast in the presence of massive background illumination.

An increasing number of imagers also require timing accuracy. In time-correlated methods, for example, the arrival time of photons emitted by a synchronized source need be detected with picosecond accuracy. Examples of such need are in fluorescence decay measurements [4], Fluorescence Lifetime Imaging (FLIM) and Fluorescence Correlation Spectroscopy (FCS) [5],[6], Förster Resonance Energy Transfer (FRET) [7], flow cytometry, etc. Single photon sensitivity can also be beneficial in reducing the complexity of fluorometer systems by replacing laser stimuli by LEDs [8] and by applying frequency domain [9] or space domain signal processing [10].

Other, non-biological or medical imagers with high timing accuracy have also been sought over the years. Typical applications include time-of-flight based 3D cameras and fast cameras for natural or artificial gravity-driven flow analysis, fluid-dynamics modeling, and combustion optimization research, to name a few. However, the literature on the subject is very extensive [11],[12].

To meet sensitivity and timing requirements, researchers have increasingly turned to non-solid-state technologies, such as Photomultiplier Tubes (PMTs). In addition to their sensitivity to single photons, PMTs have several advantages in terms of noise, dynamic range, and timing accuracy. A major disadvantage is cost, size, and the fact that large arrays of PMTs are impractical. More compact multi-channel plate-based, multi-pixel devices have also been fabricated [13]. However, these devices generally require bulky vacuum chamber apparatuses.

Solid-state single photon detectors have been known for decades, however only recently researchers have succeeded in designing fully integrated single photon detectors in CMOS [14], thus triggering a renaissance in solid-state single photon detection. More recently, the emergence of multi-pixel arrays and new imaging techniques exploiting single photon detectors has accelerated the trend [15], [16]. These sensors are generally based on a device known as Single Photon Avalanche Diode (SPAD). A SPAD may be implemented as a reverse biased p-n junction where a lightly doped guard ring prevents premature discharge [14]. If the diode is biased above breakdown, the optical gain of the detector becomes virtually infinite, thereby ensuring Geiger mode of operation. In Geiger mode, when a photon triggers an avalanche in the multiplication region, a voltage pulse of appreciable amplitude is generated. By proper design, this pulse across the junction may be processed by conventional CMOS digital circuitry.

SPADs are useful in applications, where only a few photons must be detected and/or picosecond timing accuracy is required. Examples of SPAD based designs in this context include FCS [17], high-speed imagers [18], time-of-flight 3D cameras [15], [16] and latchup/leakage test [19]. In these systems however, pitch and array size are

3-9810801-0-6/DATE06 © 2006 EDAA

still limited due to the technologies being used and the dynamic nature of SPAD signals which make it difficult to access large arrays of pixels simultaneously.

To the best of our knowledge, all published monolithically integrated SPAD arrays rely on a sequential read-out paradigm, where pixels are accessed one at a time, each for an arbitrary integration time. However, many applications require simultaneous column and row pixel access. Moreover, in general, significantly larger array sizes are necessary to provide the full advantages of SPAD technology. For example, a larger array of 1000 or more Time-Correlated Single Photon Counting (TCSPC) pixels at high frame rate could significantly advance our understanding of neuro-transmitter biochemistry and pharmaceutical research [20]. In addition, to date CMOS SPAD imagers have been integrated only in near-micron processes. Deep sub-micron (DSM) technologies would enable larger arrays, more on-chip functionality, and overall better performance at significantly lower cost.

In this paper, we present the first SPAD ever integrated in a DSM technology. We demonstrate the lowest pitch ever achieved in a SPAD array *in any technology* and the largest SPAD streak camera ever designed. The array proposed in the paper has the lowest noise ever reported in a silicon single photon detector at room temperature.

Furthermore, the sequential read-out paradigm has been replaced with one that allows access to an arbitrary pixel in a column as soon as it detects a photon. In addition, all columns are independent from each other, thus enabling parallel column access. The novel readout scheme was implemented to carry the dynamic data generated by the pixels to the exterior of the array for processing. However, on-chip processing is possible with this scheme. The technique is scalable, thus the 4x112 pixel camera presented in this paper can be implemented in square or rectangular pixel arrays as well.

The paper is organized as follows: Section 2 describes the SPAD technology, while Section 3 describes the image sensor and camera prototype architecture. Finally, in Section 4, measurements are presented and summarized.

2. Single Photon Avalanche Diodes

SPADs are p-n junctions reverse-biased above breakdown voltage (V_{bd}) for single photon detection. When an avalanche photodiode is biased above V_{bd}, it remains in a zero current state for a relatively long period of time, usually in the millisecond range. During this time, a very high electric field exists within the p-n junction generating the avalanche multiplication region. Under these conditions, if a primary carrier enters the multiplication region and triggers an avalanche, several hundreds of thousands of secondary electron-hole pairs are generated by impact ionization, thus causing the diode's depletion capacitance to be rapidly discharged. As a result, a sharp current pulse is generated and can be easily measured. This mode of operation is commonly known as Geiger mode.

Conventional photodiodes, as those used in standard imagers, are not compatible with Geiger mode of operation since they suffer from premature breakdown when the bias voltage approaches V_{bd}. The main cause for premature breakdown is the fact that peak electric field is located only in the diode's periphery rather than in the planar region [14]. A SPAD is a photodiode specifically designed to avoid premature breakdown, thus allowing a planar multiplication region to be formed within the whole junction area.

Linear mode avalanche photodiodes are biased just below V_{bd}, thus they exhibit finite multiplication gain. Statistical variations of this finite gain produce an additional noise contribution known as excess noise. SPADs, on the contrary, are not concerned with these gain fluctuations since the optical gain is virtually infinite. Nevertheless, the statistical nature of the avalanche buildup is translated onto a detection probability. The probability of detecting a photon hitting the SPAD's surface, known as the *photon detection probability* (PDP), depends on the diode's quantum efficiency and the probability for an electron or for a hole to trigger an avalanche. Additionally, in Geiger mode, the signal amplitude does not provide intensity information since all the current pulses have the same amplitude. Intensity information is however obtained by counting the pulses during a certain period of time or by measuring the mean time interval between successive pulses.

Thermally or tunneling generated carriers within the p-n junction, which produce dark current in linear mode photodiodes, can trigger avalanche pulses. In Geiger mode, they are indistinguishable from regular photon-triggered pulses and they produce spurious pulses at a frequency known as *dark count rate* (DCR). DCR strongly depends on temperature and it is an important parameter for imagers since it defines the minimal detectable signal, thus limiting from below the dynamic range of the imager. Another source of spurious counts is represented by after-pulses. They are due to carriers temporarily trapped during a Geiger pulse in the multiplication region that are released after a short time interval, thus re-triggering a Geiger event. After-pulses depend on the trap concentration as well as on the number of carriers generated during a Geiger pulse. The number of carriers depends in turn on the diode's parasitic capacitance and on the external circuit generally used to quench the avalanche. Typically, the quenching process is achieved by temporarily lowering the bias voltage below V_{bd}. Once the avalanche has been quenched, the SPAD needs to be recharged above V_{bd} so that it can detect subsequent photons. The time required to quench the avalanche and recharge the diode up to 90% of its nominal excess bias is defined as the *dead time* (DT). This parameter limits the maximal rate of detected photons, thus producing a saturation effect. The dead time consequently limits from above the dynamic range of the image sensor.

The statistical fluctuation of the time interval between the arrival of a photon at the sensor and the output pulse leading edge is defined as the *timing jitter* or *timing resolution*. Timing jitter mainly depends on the time a

photo-generated carrier requires to be swept out of the absorption point into the multiplication region; in fully integrated SPADs it is generally a few tens of picoseconds [14].

During an avalanche, some photons can be emitted due to the electroluminescence effect [21]. These photons may be detected by neighboring pixels in an array of SPADs thus producing crosstalk. The probability of this effect is called optical crosstalk probability. This probability is much smaller in fully integrated arrays of SPADs in comparison to hybrid versions. This is due to the fact that the diode's parasitic capacitance in the integrated version is orders of magnitude smaller than the hybrid solutions, thus reducing the energy dissipated during a Geiger event. Electrical crosstalk on the other hand is produced by the fact that photons absorbed beyond the p-n junction, deep in the substrate, generate carriers that can diffuse to neighboring pixels. The probability of occurrence of this effect, whether it is optical or electrical, defines the *crosstalk probability*.

3. Sensor Architecture

3.1. Deep-submicron CMOS SPAD

The SPAD, whose cross-section is depicted in Figure 1, consists of a circular dual junction structure: p+ anode/deep n-well/p-substrate. The p+ anode/deep n-well junction forms the avalanche multiplication region where the so-called Geiger breakdown occurs. The deep n-well/p-substrate junction allows the p+ anode to be biased independently from the p-substrate. It additionally prevents electrical crosstalk due to minority carriers diffusing in the substrate and improves the timing jitter of the SPAD [16].

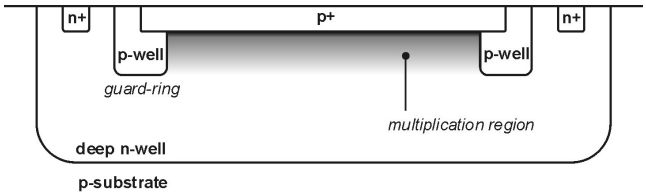

Figure 1. SPAD cross-section.

A p-well guard-ring surrounds the p+ anode to prevent premature breakdown [14]. Breakdown voltage measurements have been done on the SPAD structure. In the 0.35µm CMOS technology used, without the guard-ring, the breakdown voltage of a typical p+/deep n-well junction is about 11V. The use of the guard-ring allowed the breakdown voltage to be increased by approximately 13V. Conversely, the breakdown voltage of the deep n-well/p-substrate junction is typically 70V, thus allowing the active junction to be biased above V_{bd}.

3.2. Digital Pixel and Event-Driven Readout

The SPAD linear array consists of 4×112 pixels. Every pixel within a column can be accessed independently and simultaneously. The row readout is controlled directly by the pixels based upon an event, i.e. the detection of a photon. When a Geiger event occurs, a pixel generates a pulse synchronous with the photon arrival, thus enabling time-correlated measurements. A novel event-driven readout temporarily assigns the column readout bus to the pixel requesting it.

The pixel circuit consists of a SPAD, a quenching/recharge transistor (T_q), a CMOS inverter, and the pixel event-driven readout circuit as shown in Figure 2. The SPAD operates in passive quenching and passive recharge. The p+ anode is biased at a high negative voltage V_{op} equal to –23.4V. This voltage is common to all the pixels in the array. The deep n-well cathode is connected to the power supply V_{DD} = 3.3V through PMOS T_q, which eliminates the need for a quenching resistance [16]. The excess bias voltage (V_e), defined as the excess of bias voltage above V_{bd}, is thus equal to $|V_{op}| + V_{DD} - V_{bd} = 4.3V$.

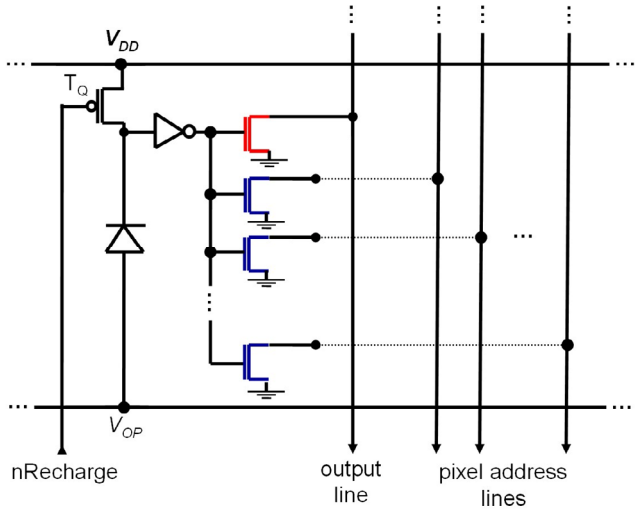

Figure 2. Pixel circuit

Upon photon arrival, the breakdown current discharges the SPAD's depletion region capacitance, reducing the voltage across the SPAD to close to V_{bd}. T_q is designed to provide a sufficiently resistive path to quench the avalanche process. After avalanche quenching, when the signal nRecharge is asserted, the SPAD recharges through T_q and progressively recovers its photon detection capability. The time required for quenching the avalanche and restoring the operating bias, i.e. DT, is typically less than 40ns for the digital pixel. Signal nRecharge allows additionally the pixel to go to an inactive state when nRecharge is not asserted.

At the cathode, an analog voltage pulse of amplitude approximately V_e [16] reflects the detection of a single photon. The inverter stage converts this analog voltage pulse into a robust digital pulse. The near-infinite internal gain inherent to Geiger mode operation leads to no further amplification.

The pixel event-driven readout circuit consists of an output NMOS transistor, a shared output line, N_{AT} address coding NMOS transistors, and N_{AT} shared address lines, where N_{AT} is given by $Ceiling(Log_2(N_{ROWS}))$, N_{ROWS} being

the number of pixels within a column (Figure 2). In this work, N_{ROWS} is 112, thus N_{AT} is 7. The source of the output transistor and the sources of all the address coding transistors are connected to *GND*. The drain of the output transistor is connected to the output line whereas the drains of the address coding transistors are either connected to the address lines or left unconnected so as to define a unique pixel address within the column.

Figure 3 shows the column readout circuitry based on the proposed event-driven topology, where only two pixels are sketched.

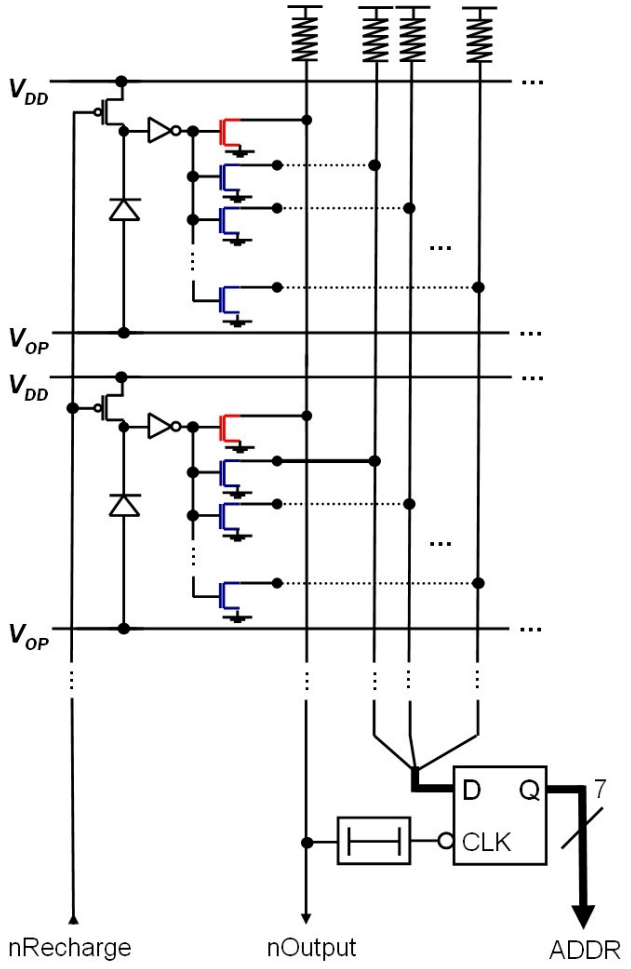

Figure 3. Column readout circuit

The output, address, and nRecharge lines forming the column readout bus are shared among all the pixels within a column and connected to V_{DD} via pull up resistors. Within the pixel, the output and address coding transistors are driven by the inverter stage. In idle state, nRecharge is asserted making all the SPADs recharged. The inverter output of all the pixels stay at *GND* until a photon is detected. During this time, the output and address coding transistors of all pixels are blocked, thus leaving the lines of the column readout bus at V_{DD} through the pull up resistors. Upon photon detection within a pixel, the output transistor of that pixel pulls the output line down asserting nOutput in the end of the column. At the same time, the address coding transistors pulls the address lines that are connected for that particular pixel down, making the address of the pixel available on the input of a set of D-latches at the end of the column. Signal nOutput is additionally used to clock the D-latches through a delay line to hold the address of the pixel that detected the photon. The sensor readout circuitry, triggered by nOutput, reads out the address of the pixel (ADDR) that is stable until the next photon detection. In order to respect the setup time of the D-latches and to reduce the probability of address collisions when multiple photons are detected, the delay line is actively programmed to match precisely the setup time of the D-latches. When a second photon is detected by another pixel within the column, for instance few tens picoseconds later, nOutput is already asserted and the address of the previous pixel is already latched, therefore the second photon is simply missed. Due to the statistical properties of photons, this technique demonstrated to be very robust when light levels are moderate. When high light levels are required, ADDR can be coded using a conflict-free coding, though increasing the number of address lines and transistors per pixel.

Operating in time-correlated mode, e. g. using a time-to-digital converter, nOutput is used as trigger signal and it reflects the photon arrival with picosecond precision. Signal nRecharge can be continuously asserted in run mode to use the maximum of the column bandwidth or, on the other hand, it can be deasserted to leave the whole column in waiting state. Figure 4 plots a comparison between the signal counts of a typical pixel within a column and the corresponding signal count that the pixel would have with ideally parallel implementation. As can be seen, since the readout cycle time is limited by the SPAD dead time (i.e. 40 ns), a saturation effect appears when the signal count is increased.

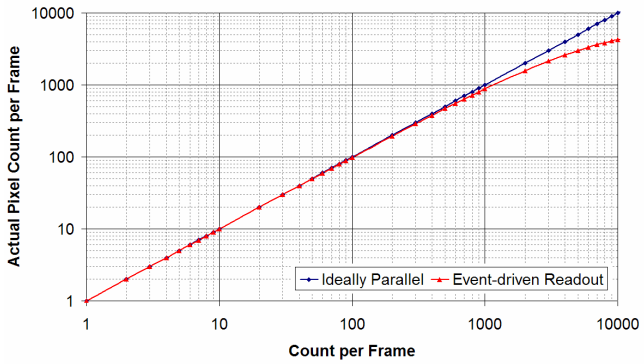

Figure 4. Signal count for a typical pixel vs. an ideal parallel readout for a frame rate of 30fps

A photomicrograph of the chip die is shown in Figure 5, which also shows the layout of a pixel in the inset. The 11T digital pixel occupies a square area of 25µm x 25µm. Each column is handled by a Column Control Unit (CCU). The CCU operates independently, thus an arbitrary number of columns *N* may be designed with a linear increase of complexity and power dissipation.

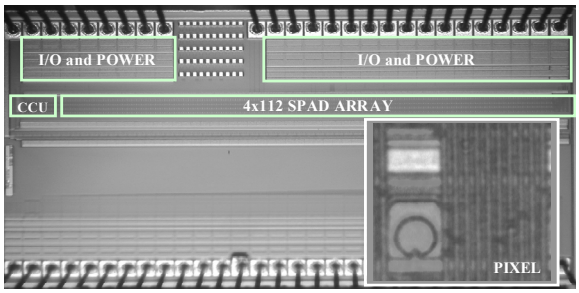

Figure 5. Image sensor photomicrograph. Inset: pixel detail.

3.3. Camera Prototype

A camera prototype based on the proposed image sensor was built. The camera prototype was designed using an Altera Excalibur FPGA embedding an ARM processor core. Within the FPGA, banks of 16-bit counters were implemented and associated to each image sensor pixel. A Finite State Machine (FSM) generates and controls interface signals between the FPGA and the SPAD-based image sensor. Signal ADDR combined with a column address is used to access the corresponding counter to be incremented for each detected photon. In addition, the FSM handles the interface between the counters and the readout software running on the ARM processor.

Volatile and non-volatile memories, communication links, power supply as well as optional measurement connectors were designed and implemented on the board level. Figure 6 shows the camera prototype. In order to use standard camera lenses, a c-mount thread was realized centered to the image sensor cavity.

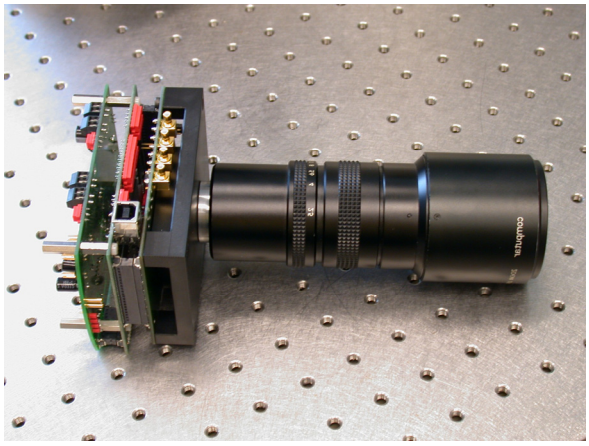

Figure 6. Camera prototype.

4. Sensor Characterization

The suitability of DSM CMOS SPAD arrays is demonstrated through a variety of performance measures. DCR is log-plotted in Figure 7, while a distribution of DCR across all the pixels is shown in the histogram of Figure 8.

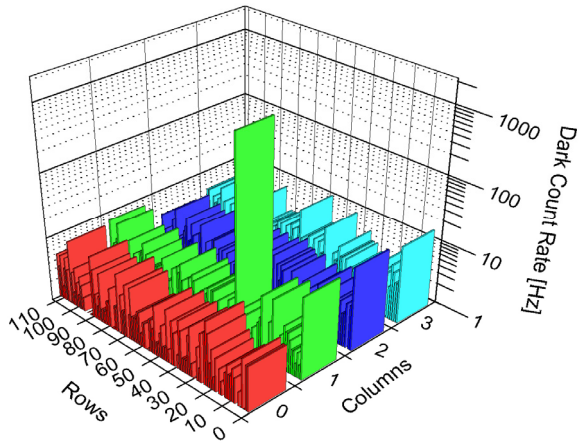

Figure 7. Pixel map of dark count rate across the array.

Note that at room temperature 98.9% of the pixels exhibit a DCR below 10Hz. To the best of our knowledge, this is the best noise performance ever reported for SPADs implemented in any solid-state technology.

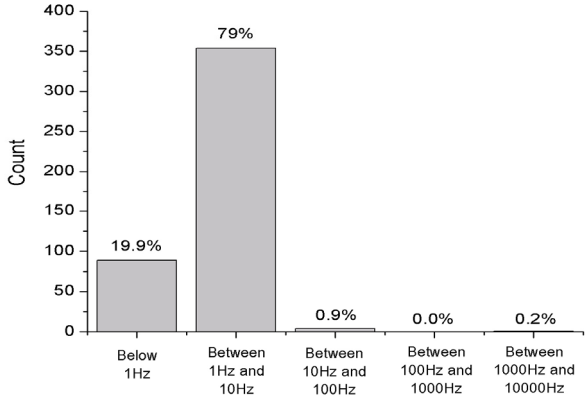

Figure 8. Histogram of dark count rate.

The sensitivity of the array is characterized in terms of its PDP as a function of photon wavelength and excess bias voltage V_e. (plot of Figure 9).

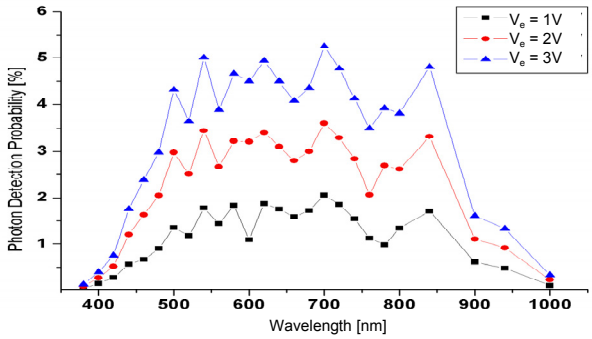

Figure 9. Photon detection probability vs. photon wavelength for different values of excess bias voltage V_e.

In this chip, the PDP varies from 0.1 to 5.3%. This performance measure was below expectations [16] due to reduced optical transparency in the passivation layers. The

problem had been caused by an error in the mask generation. We estimate that the layers attenuate light by a factor of 2~5, thus reducing the expected PDP from a peak of 10~25% to the measured values. Nonetheless, note the wide range of wavelengths for which the SPAD is sensitive, covering NUV, visible light, and NIR.

Other parameters completed the test, including pixel crosstalk, dead time, saturation, and image sensor power dissipation. Table 1 summarizes the overall performance of the imager.

Performance	Min.	Typ.	Max.	Unit
Pitch		25		μm
Array size		4×112		-
Avg. Dark Count Rate (DCR)		6		Hz
Dead Time (DT)			40	ns
Saturation			25	MHz
Wavelength range	350		1000	nm
PDP	0.1		5.3	%
Power dissipation			7	mW
Overall crosstalk @ 30μm		0.05		%

Table 1. Performance summary. All the measurements are reported at room temperature.

5. Conclusion

We have reported on the first implementation of an integrated SPAD array in DSM technology. Due to a significant reduction in feature size, if compared to state-of-the-art technologies, it has been possible to design an array of 4x112 single photon detectors. Thanks to a pitch of 25μm, a lateral dimension of less than 3mm was achieved. Another major innovation of this design is the readout scheme that allows simultaneous and independent pixel access upon photon arrival. The system is fully scalable, thus making a further expansion of SPAD based multi-pixel sensors realistic. Preliminary measurements suggest that DSM technology is suitable to yield SPADs with equivalent or better performance than current implementations. Thus, applications, where sensitivity and timing accuracy are critical, will significantly benefit from this technology.

Acknowledgments

This research was supported by a grant of the Swiss National Science Foundation. The authors are grateful to Zhen Xiao and Pierre-André Besse for helpful discussions and to Radivoje Popovic for continued encouragement.

References

[1] H. Eltoukhy, K. Salama, A. El Gamal, M. Ronaghi, R. Davis, "A 0.18μm CMOS 10^{-6} lux Bioluminescence Detection System-on-chip", *IEEE ISSCC*, pp. 222-223, Feb. 2004.

[2] A. Grinvald et al., "*In-Vivo* Optical Imaging of Cortical Architecture and Dynamics", Modern Techniques in Neuroscience Research, U. Windhorst and H. Johansson (Eds), Springer, 2001.

[3] S. Nagasawa, H. Arai, R. Kanzaki, and I. Shimoyama, "Integrated Multi-Functional Probe for Active Measurements in a Single Neural Cell", IEEE Intl. Conference on Solid-State Sensors, Actuators, and Microsystems, Vol. 2, pp. 1230-1233, Jun. 2005.

[4] J. C. Jackson et al., "Characterization of Geiger Mode Avalanche Photodiodes for Fluorescence Decay Measurements", *Proc. of SPIE*, Vol. 4650-07, Photonics West, San Jose, CA, Jan. 2002.

[5] A. V. Agronskaia, L. Tertoolen, H. C. Gerritsen, "Fast Fluorescence Lifetime Imaging of Calcium in Living Cells", Journal of Biomedical Optics, Vol. 9, N. 6, pp. 1230-1237, Nov./Dec. 2004.

[6] P. Schwille, U. Haupts, S. Maiti, W. W. Webb, "Molecular Dynamics in Living Cells Observed by Fluorescence Correlation Spectroscopy with One- and Two-Photon Excitation", *Biophysics Journal*, Vol. 77, pp. 2251-2265, 1999.

[7] W. Becker, K. Benndorf, A. Bergmann, C. Biskup, K. König, U. Tirplapur, and T. Zimmer, "FRET Measurements by TCSPC Laser Scanning Microscopy", Proc. of SPIE 4431**,** ECBO, 2001.

[8] N. Chen, Q. Zhu, "Time-resolved optical measurements with spread-spectrum excitation", *Optics Letters*, pp. 1806-1808, Vol. 27, N. 20, Oct. 2002.

[9] P. Herman et al., "Frequency-domain fluorescence microscopy with the LED as a light source", *Journal of Microscopy*, pp. 176-181, Vol. 203, Pt. 2, Aug. 2001.

[10] J. Qu et al., "Development of a Multispectral Multiphoton Fluorescence Lifetime Imaging Microscopy System Using a Streak Camera", *Proc. of SPIE*, Vol. 5630, pp. 510-516, Jan. 2005.

[11] E. Charbon, "Will CMOS Imagers Ever Need Ultra-High Speed?", IEEE International Conference on Solid-State and Integrated-Circuit Technology, pp. 1975-1980, Oct. 2004.

[12] T. G. Etoh et al., "An Image Sensor Which Captures 100 Consecutive Frames at 1,000,000 Frames/s", IEEE Trans. on Electron Devices, Vol. 50, N.1, pp. 144-151, Jan. 2003.

[13] J. McPhate et al., "Noiseless Kilohertz-frame-rate Imaging Detector based on Microchannel Plates Readout with Medipix2 CMOS Pixel Chip", *Proc. of SPIE*, vol.5881, pp.88-97, 2004.

[14] A. Rochas, "Single Photon Avalanche Diodes in CMOS Technology", Ph.D. Thesis, Lausanne, 2003.

[15] C. Niclass and E. Charbon, "A Single Photon Detector Array with 64x64 Resolution and Millimetric Depth Accuracy for 3D Imaging", *IEEE Intl. Solid-State Circuit Conference*, pp.364-365. Feb. 2005.

[16] C. Niclass A. Rochas, P.A. Besse, and E. Charbon, "Design and Characterization of a CMOS 3-D Image Sensor Based on Single Photon Avalanche Diodes", *IEEE Journal of Solid-State Circuits*, vol.40, n.9, Sep. 2005.

[17] M. Gösch et al., "Parallel Single Molecule Detection with Fully Integrated Single Photon 2x2 CMOS Detector Array", *Journal of Biomedical Optics*, Vol. 9, N. 5, 2004.

[18] C. Niclass et al., "CMOS Imager Based on Single Photon Avalanche Diodes", *IEEE Transducers*, June 2005.

[19] F. Stellari et al., "Testing and Diagnostics of CMOS Circuits Using Light Emission from Off-State Leakage Current", *IEEE Trans. on Electron Devices*, vol.51, n.9, pp.1455-1462, 2004.

[20] C. Petersen, *Personal communication*, Brain & Mind Institute (EPFL), 2005.

[21] R. H. Haitz, "Studies on optical coupling between silicon p-n junctions", Solid State Electronics, Vol. 8, pp. 417-425, 1965.