Single-photon Imaging in CMOS

E. Charbon[*a]
[a]Delft University of Technology, Mekelweg 4, 2628 CD Delft, Netherlands

ABSTRACT

We report on the architectural design and fabrication of medium and large arrays of single-photon avalanche diodes (SPADs) for a variety of applications in physics, medicine, and the life sciences. Due to dynamic nature of SPADs, designs featuring a large number of SPADs require careful analysis of the target application for an optimal use of silicon real estate and of limited readout bandwidth. This paper describes the main trade-offs involved in architecting such chips and the solutions adopted with focus on scalability and miniaturization.

Keywords: single-photon avalanche diode, avalanche photodiode, complementary metal-oxide semiconductor, SPAD, APD, CMOS

1. INTRODUCTION

1.1 Single-photon Avalanche Diodes

Solid-state single-photon detectors have existed for decades and, while several flavors of solid-state detectors exist in various technologies and ranges of operation, from cryogenic to room temperature detectors, silicon avalanche photodiodes (APDs) have emerged as the most versatile and easy to use among them [1]. A class of APDs operating above breakdown, in so-called Geiger mode and known as single-photon avalanche diodes (SPADs), is of particular interest due to their amenability to integration in planar silicon processes in combination with conventional digital and analog circuitries. The first SPADs implemented in a planar technology have emerged relatively recently [2],[3]. But, while the physics of solid-state SPADs is well understood [4], it is only with the advent of devices integrated in conventional CMOS processes [5], that the evolution onto smaller and smaller feature sizes has rapidly advanced to the point that it has now become possible to envision large imaging systems based on SPADs.

A SPAD is essentially a pn junction biased above breakdown and equipped with avalanche quenching and recharge mechanisms. Upon photon detection, an avalanche may be triggered. The quenching circuitries stop the avalanche thus preventing the destruction of the device, while the recharge circuitries prepare the SPAD for the next detection cycle by raising the bias voltage again to the initial stage. Two modes for quenching and recharge exist: active and passive. In active mode, active circuitries are used to control the process. In passive mode, the avalanche current is passively controlling the process by way of a ballast resistive device. The literature on the subject is extensive and it is beyond the scope of this paper to describe it further. In the remainder of the paper we focus on passive quenching and recharge that represent an adequate solution in terms of miniaturization and performance.

1.2 SPAD Implementation in Planar Processes

Implementing a SPAD in a planar process first involves finding way to prevent premature edge breakdown (PEB). Several techniques exist to implement PEB prevention. In essence, the techniques have in common the reduction of the electric field at the edges and everywhere else in the device, so as to maximize the probability that the avalanche is initiated in the center of the multiplication region. This is the region where the critical electric field for impact ionization is reached and, possibly, exceeded. In Figure 1 four of the most used structures are shown. In a) the n+ layer maximizes the electric field in the middle of the diode. In b) the lightly doped p- implant reduces the electric field at the edge of the p+ implant. In c) a floating p implant locally increases the breakdown voltage. With a polysilicon gate one can further

[*] e.charbon@tudelft.nl; phone +31 15 278-3667; fax +31 15 278-6190; cas.et.tudelft.nl

extend the depletion region (gray line in the figure). Finally, in a process with trenches it is possible to decrease the electric field using the geometry of solution d). When trenches are used one needs to adopt techniques to prevent that traps accumulated in the trench during fabrication can induce PEB. An effective technique proposed in [6] consists of using several layers of doped semiconductor material with decreasing doping levels from the trench to the multiplication region. The purpose is to achieve short mean-free paths close to the trench, thereby forcing carriers generated there to recombine before reaching the multiplication region.

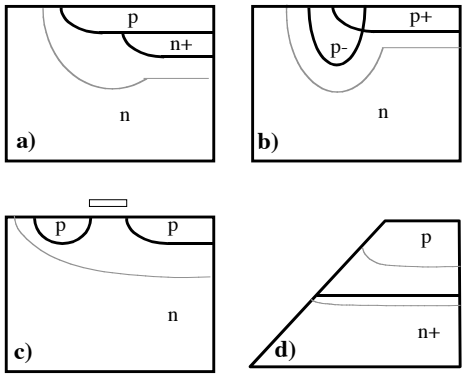

Figure 1. Premature edge breakdown prevention mechanisms in planar and semi-planar processes.

The structures of Figure 1 (a)-(c) are indicated in a number CMOS technologies, while a trench based structure (d) is mostly appropriate in deep-submicron CMOS technologies, where deep and medium tubs are not available without a major change in the fabrication process. In the remainder of the paper we focus our attention to schemes (b) and (d), because they require, in general no modifications to the process and thus enable the design of large SPAD array chips in standard CMOS technologies.

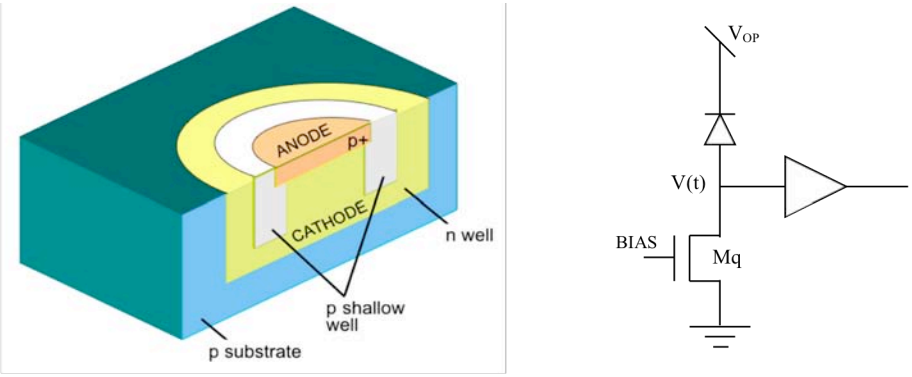

Figure 2. SPAD cross-section in a conventional CMOS process (left); passive quench and recharge circuitries, as well as pulse shaping (right).

1.3 Biasing a SPAD

Upon design of a pn junction capable of withstanding relatively high voltages and whereby PEB is practically prevented, the device may be biased above breakdown in Geiger mode. In this mode of operation the avalanche must be quenched to prevent destruction of the device. There exist a variety of avalanche quenching techniques, partitioned in active and passive methods. In active methods, the avalanche is detected and stopped by acting on the bias. In passive methods the pn junction bias is self-adjusted e.g. by a ballast resistor.

Recharge methods can also be active and passive. In active methods, the bias across the diode is re-established by a switch activated by an avalanche detector. In passive methods the recharge occurs through the ballast. Figure 2 shows the

cross-section of a SPAD and simple circuitry to perform passive quench and recharge. Upon photon detection, the device generates a current pulse that is converted to a digital voltage level by means of a pulse shaping circuitry, also shown in the figure. The pulse shaper is also acting as an impedance adapter to drive the load of the column readout often employed in a SPAD matrix. The ballast may be implemented as a resistor or as an active element acting as a non-linear resistor (M_q in Figure 2). Assuming that the latter can be approximated as a linear element, recharge voltage V(t) yields

$$V(t) \approx V_E e^{-\frac{t-t_0}{RC}}, \quad t \geq t_0 \tag{1}$$

where R and C are estimated overall resistance and capacitance to the ground at the cathode and V_E is the excess bias voltage, i.e. the voltage above breakdown at which the SPAD is biased. Time t_0 is the instant at which the quenching is complete. Upon proper selection of the bias (the signal BIAS in Figure 2), M_q may also be operated in active mode, thus causing a controlled recharge. When controlled recharge is applied, V(t) becomes

$$V(t) \approx V_E - \frac{I_R}{C}(t - t_0), \quad t_0 \leq t \leq t_0 + \frac{CV_E}{I_R} \tag{2}$$

where I_R is the current discharged by M_q of Figure 2. In reality, the transistor may not be operated in strong inversion throughout the recharge and thus Equation (2) is only an approximation. The accuracy of this approximation is discussed in the literature.

The advantage of using an active recharge is a better control of the detection cycle and in particular of the overall time spent in the quenching and recharge, known collectively as dead time. Furthermore, active recharge can also be performed in multi-slope mode to allow for a precise control of dead time over larger SPAD arrays, thus improving overall detection uniformity, especially in high illumination regimes. Figure 3 shows three typical recharge profiles. Passive recharge is the most commonly used technique, while active recharge is used in many devices whereby the recharge process has to respond to specific requirements. Figure 3 also shows two types of active recharge known as single- and double-slope active recharge. Single-slope recharge is simple to implement requiring only one bias per SPAD. In double-slope recharge [7], the SPAD's dead time is effectively controlled by time t_R at which the second slope is activated. If the voltage V_R achieved at this point still disables the avalanche, then it is guaranteed that the device is still in dead time regime. Thus the dead time can be triggered by one properly timed control signal and thus it is independent of R.

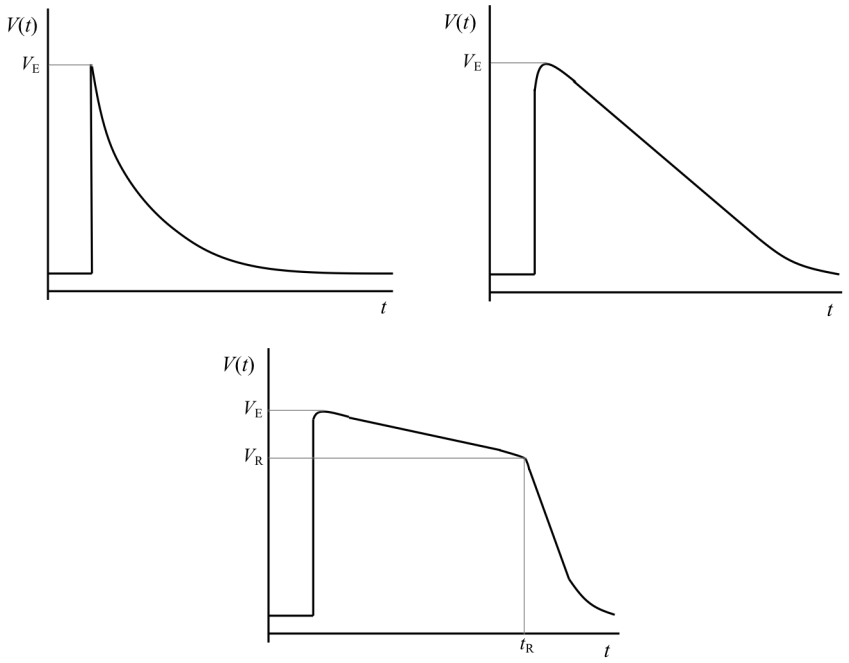

Figure 3. SPAD recharge mechanisms: passive (left), single-slope active (right), and double-slope active (center).

Dead time is an important parameter, as it determines the maximum count rate of a detector and thus the saturation intensity. A variety of active quenching and recharge circuits can be found in the literature whereby the differentiating factors are complexity as well as dead time programmability and stability. One of the most important considerations in the selection of the best possible recharge mechanism is simplicity, when it comes to miniaturization.

2. SPAD CHARACTERIZATION IN IMAGE SENSORS

2.1 Basic SPAD Parameters

Individual SPADs are characterized by their sensitivity, measured as **photon detection probability** (PDP), the noise performance, measured a rate of spurious pulses due to thermal events or **dark count rate** (DCR). Other parameters include **timing jitter**, also known somewhat inappropriately as **timing resolution**, **afterpulsing probability**, and, as mentioned earlier, **dead time**. These parameters have appeared in the literature for individual SPADs implemented in a variety of CMOS processes [8],[9],[10],[11],[12],[13],[14],[15],[16],[17].

Some performance parameters found in individual SPADs are described in Table 1 for four different implementations in CMOS submicron and deep-submicron processes on which the imagers described in this work were based.

Measurement	0.8μm [8]	0.35μm [12]	130nm [6]	130nm [18]	Unit
Timing jitter (FWHM @ 637nm)	82*	80	125	200	ps
DCR (mean at 300K)	350 (V_E: 5.0V)	750 (V_E: 3.3V)	220 (V_E: 2.0V)	1221 (V_E: 0.6V)	Hz
Active area	38	78	58	50	μm^2
Mean DCR per active area	9.2	9.6	3.8	24.4	$Hz/\mu m^2$
Breakdown (V_{BD})	25.5	17.4	12.8	14.4	V
Dead time	<40	40	100	100	ns
PDP @ 460nm	26 (V_E: 5.0V)	40 (V_E: 3.3V)	26 (V_E: 2.0V)	17.5 (V_E: 0.6V)	%
EM spectrum (PDP > 1%)	380~900	350~1000	380~900	350~1000	nm
Technology	CMOS	CMOS	CMOS	CMOS	-

Table 1. Comparison of CMOS SPAD performance for a variety of published devices.

Afterpulsing is a process by which a primary avalanche is followed by other ones unrelated to photons. The physical process underlying afterpulsing has been thoroughly researched in the literature. Afterpulsing characterization can be found in [6],[12], and [19] for the devices described in this work.

2.2 Characterization of SPAD Arrays in Image Sensors

When implemented in an array, other performance measures become relevant to the quality of the imager. Besides the aforementioned dead time uniformity, timing jitter uniformity and PDP uniformity, as well as DCR uniformity and crosstalk have to be accounted for and properly characterized. Figure 4 shows the dead time and PDP uniformity achieved in a 32x32 pixel array implemented in a CMOS process [8]. In the remainder of the paper we will refer to these parameters as a function of excess bias voltage and temperature.

* Recently improved measurement.

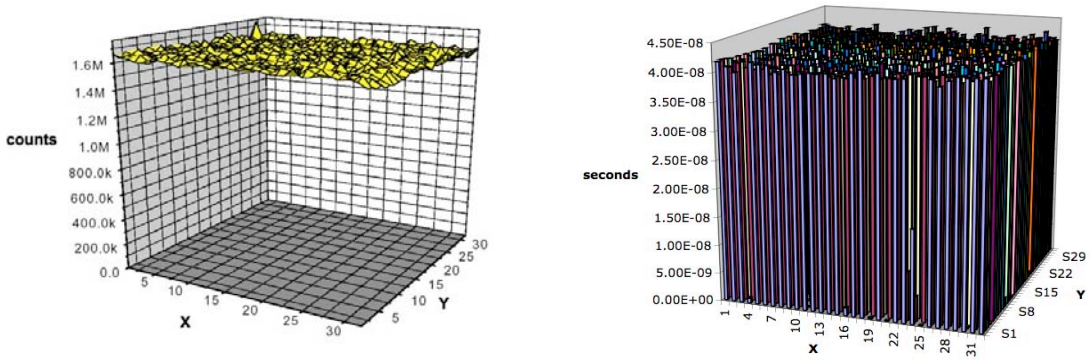

Figure 4. PDP and dead time uniformity in a 32x32 array of low-pitch passively recharged pixels.

PDP of course will also be a function of the input wavelength. In CMOS SPAD implementations, the sensitivity range is mostly in the visible spectrum, with somewhat reduced near infrared and near ultraviolet PDP.

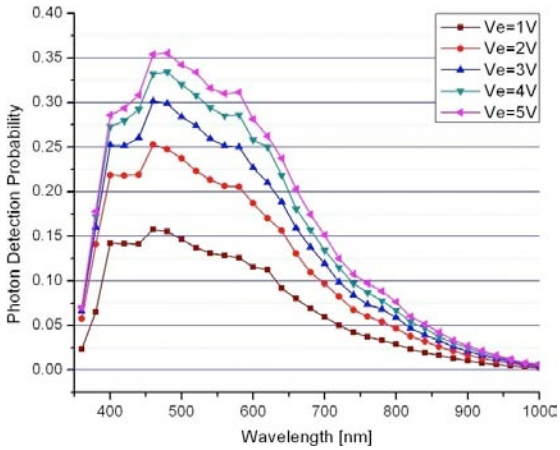

Figure 5. PDP in a SPAD implemented in 130nm CMOS process as a function of wavelength and excess bias [6].

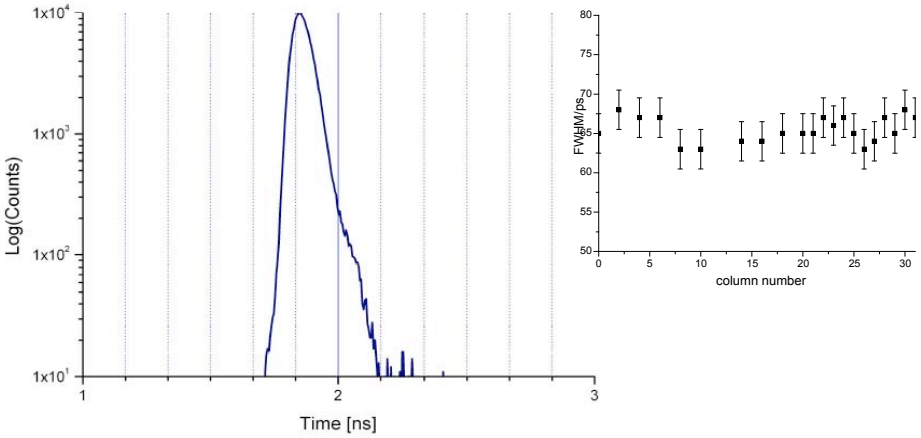

Figure 6. Timing jitter performance reported in [12]. The inset shows the FWHM timing jitter over an array of SPADs.

Figure 6 shows the timing jitter on a SPAD implemented in 0.35μm CMOS technology. In the inset of the figure the uniformity of such jitter can be seen in an array of pixels integrated on the same chip measured by exposing the chip to a cone of light from a pulsed laser source. In this case a femtosecond Ti:Sapphire laser source doubled to achieve a wavelength 488nm was employed. The uniformity in time is also very important, especially in photon-starved applications whereby long measurements may be needed to reach good accuracy. As an illustration, consider a time-resolved measurement of photons emitted by a fluorophore upon excitation. Due to the relatively low efficiency of the fluorescence process, one generally uses techniques such as time-correlated single-photon counting (TCSPC) in combination with single-photon detection and photon time-of-arrival (TOA) evaluation. Figure 7 shows the histogram of a TCSPC experiment performed on a sample with high-affinity nonratiometric Ca^{2+} indicator Oregon Green BAPTA-1 (OGB-1) in a solution of calcium ions in various concentrations [20]. Using a SPAD based setup with overall impulse response function (IRF) of 79ps, the fluorescence dynamics of OGB-1 was found to follow a triple exponential decay, thus providing an accurate model of the relation between concentration and decay parameters.

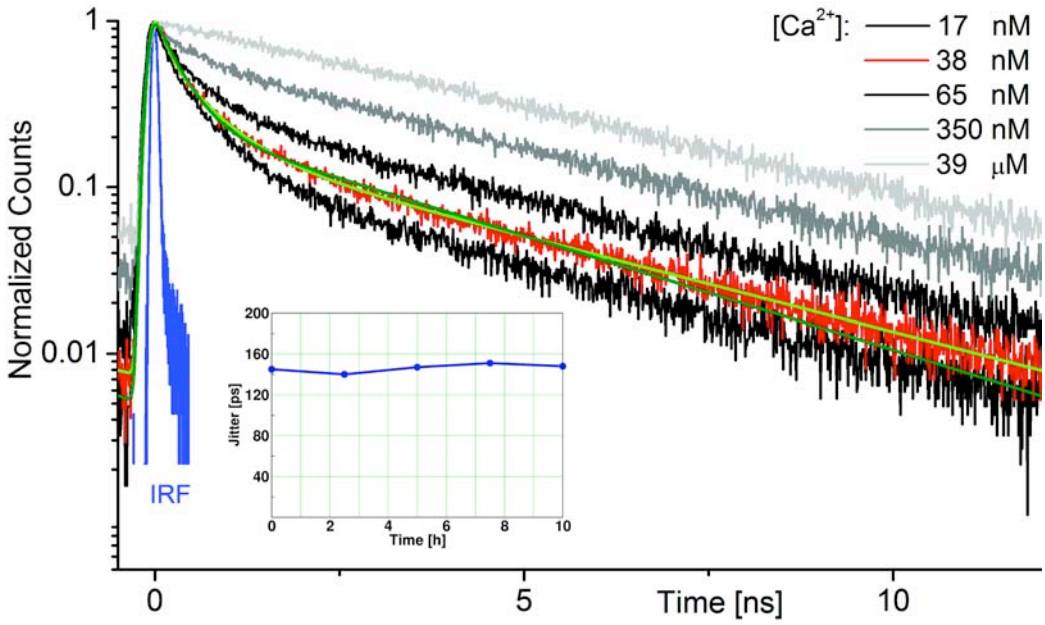

Figure 7. Histograms of the response of OGB-1 molecules to repeated excitation in presence of Ca^{2+} ions at various concentrations [20]. The inset shows the long-term stability of the timing jitter.

Crosstalk may be electrical and/or optical. Electrical crosstalk is due to the interference between pixels. It may be caused by a temporary drop of sensitivity and DCR in a victim pixel due to the drop of excess bias voltage. The latter, in turn, may be caused by a neighboring aggressor pixel as an avalanche is triggered. Similarly, substrate noise originated in one or more pixels may be picked up by the victim pixel and a spurious avalanche may thus be triggered. Optical crosstalk may occur when an avalanche is triggered in the aggressor pixel. By impact ionization, several photons may be emitted, thus causing a victim pixel to detect it. While electrical crosstalk is strongly dependent on the design of supply lines and of substrate noise rejection measures, optical crosstalk may only be influenced by the number of carriers involved in an avalanche and by pixel pitch. The reduction of the number of avalanching carriers may be best achieved by reducing the active area of a SPAD, and thus its capacitance at a cost of lower fill factor if the pixel pitch is kept constant. Figure 8 shows the impact of a pixel near saturation onto the neighboring pixels at their nominal DCR values, indicating unmeasurable crosstalk in the array structure of the design in [8].

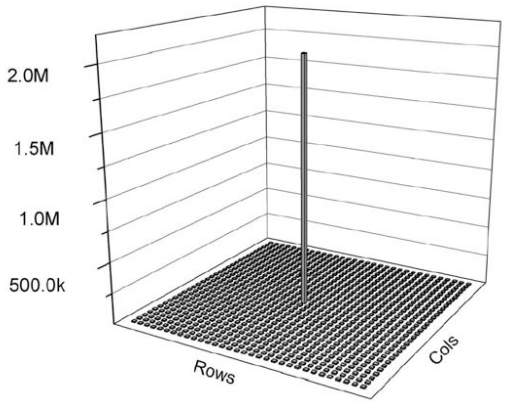

Figure 8. Crosstalk measured in a 32x32 array of low-pitch passively recharged pixels, when the center pixel is selectively illuminated near saturation with an optical fiber.

3. ARCHITECTURE SELECTION

3.1 Architecture vs. Application

SPADs are dynamic devices generating a digital pulse upon detection of a photon. Unlike conventional diodes, they cannot hold a charge proportional to the overall photon count. Photopulses must be counted *in situ* or read outside the image sensor and counted externally. Due to their reaction speed and low timing uncertainty, SPADs are most appropriate for photon TOA evaluation. However, even this operation must be performed upon photon detection. To address this problem, researchers have adopted a number of architectures that take advantage of the low propagation delay or high level of miniaturization achievable in standard submicron and deep-submicron CMOS technologies [1].

The available architectures are (1) in-pixel, (2) in-column, and (3) on-chip counting or TOA evaluation. When in-pixel architectures are used all the operations are performed and saved locally; the stored value is read out later in random access or sequential mode. In-column or cluster counting implies the sharing of operations of all the pixels on the column or the cluster whereas the result is stored in a column-based memory and read out on the column-by-column basis. When sharing is used, trade-offs between pixel utilization, column/cluster size, and detection bandwidth are generally to be foreseen. In these cases, understanding application specifications is key to an appropriate use o the available techniques. On-chip counting or TOA is essentially an extension of the in-column architecture, whereas the working cluster is the entire chip. Similar trade-offs are also used in this case.

Hereafter we describe a few designs based on the above architectures, from on-chip to in-pixel styles. The first design demonstrating the feasibility of large SPAD arrays comprised a matrix of 32x32 pixels, each with an independent SPAD, a quenching mechanism, a pulse shaping and column access circuitry [8]. The readout scheme was based on random access whereby all time-sensitive operations had to be performed off-chip and an overall jitter as low as 70ps was measured on a pixel while the entire array was operating [10]. The main drawback of this design is the fact that only one pixel can be read out at any time while photons falling outside that pixel are lost. The block diagram of the imager and the pixel schematic is shown in Figure 9.

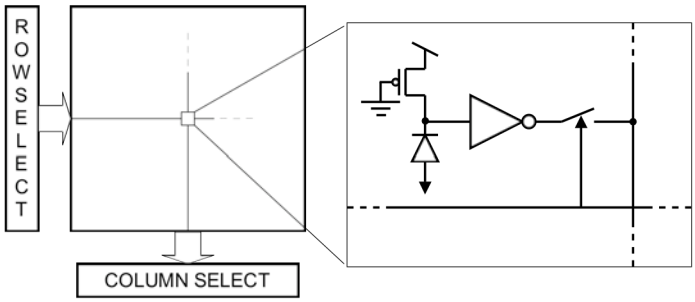

Figure 9. Block diagram and pixel schematic of the 32x32 SPAD array with random access readout.

Note that the SPAD is connected to a negative voltage at the anode and the quenching is placed at its cathode. The negative voltage is chosen so as to ensure that the device operates above breakdown by an excess voltage V_E. Thus the avalanche pulse must be inverted by an inverting component, in this design a simple inverter. The micrograph of the chip is shown in Figure 10.

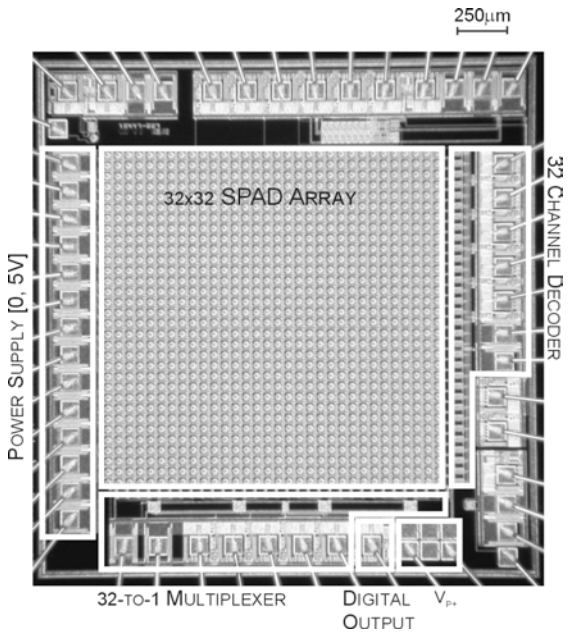

Figure 10. 32x32 SPAD array with random access readout. The chip was implemented in 0.8μm CMOS technology.

To address the readout bottleneck, two approaches were devised. The first, known as event-driven readout, consists of using the column as a bus that is addressed every time a photon is detected. The address of the row where the photon was detected is sent to the bottom of the column where the TOA is evaluated, either off chip [10],[12] or on chip [22]. The second approach, known as latchless pipelined readout, consists of using the column as a timing-preserving delay line. Every photon triggers a pulse that is injected onto the pipeline at a precise location that corresponds to the physical place where the pixel is situated. The row information is thus encoded in the timing of the pulse arrival at the end of the pipeline, thus it can be sequentially reconstructed by a single time-to-digital converter (TDC), a sort of miniaturized high-speed chronometer, at the bottom of the column. The TDC will also detect the exact TOA of the photon within a predefined window of time.

Figure 11 shows a schematic of the injection mechanism at each pixel. The avalanche current is sensed and converted onto a digital voltage pulse, as before, by a properly designed inverter. The L to H transition at the inverter's output pulls down node "X" through transistor T_{PD} and resistor R_{PU}, provided that gating transistor T_G is enabled by

signal “GATE”. The anode of the diode is intentionally set to a negative voltage, as before. T_q was sized for a dead time τ_{DT} of 40ns and by choosing a gating window τ_G that satisfies inequality $\tau_G < \tau_D < \tau_{DT}$. When there is no activity on the preceding delay line, signal “VIN_j” is at logic level L, hence the gate of source-degenerated transistor T_{PP} is L, thus the impedance at node “X” is dominated by the impedance at the drain of T_{PP}. When a photon is detected, a pulse is originated at this point and it is propagated towards the remainder of the delay line. When there is activity on the delay line, a logic transition L to H on “VIN_j” occurs, thus causing “X” to become a low impedance node. During this time any photon detection in this stage will have no effect on traveling pulses but it will inject spurious pulses onto the line when it is at logic level L, hence the need for gated SPAD operation. To avoid ghost pulses, an appropriately sized NMOS was added to the cathode of the diode. A simplified timing diagram to operate the 8-stage delay line is shown in the inset of the figure. Controls “BIAS” (transistor T_B) and “TUNE” are used for coarse- and fine-tuning of the delay line, respectively. The goal is to compensate for technological variations and temperature.

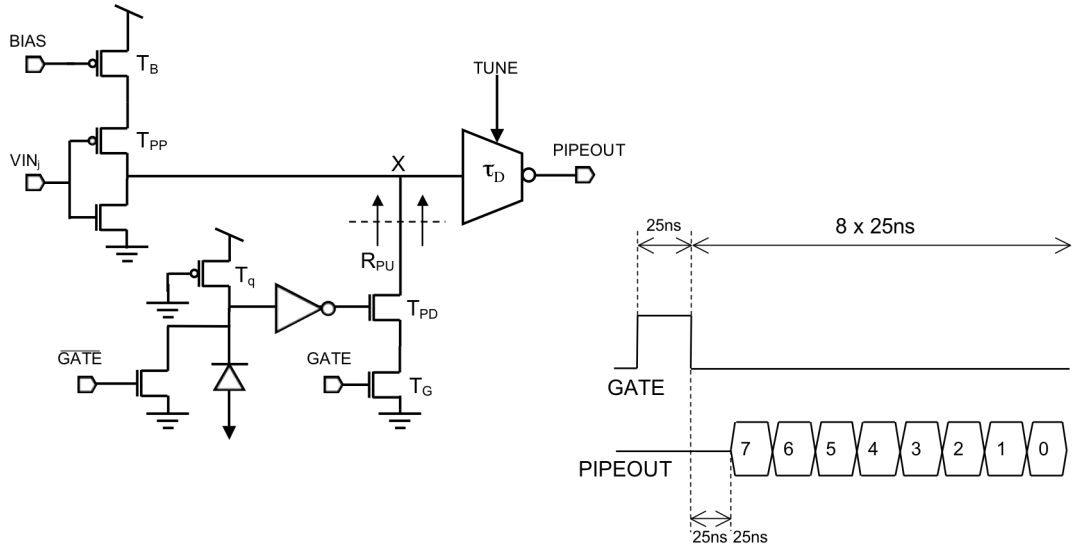

Figure 11. Schematic diagram of the latchless pipelined readout.

A chip implementing the concept for a 128x2 SPAD array is shown in Figure 12; the architecture was implemented in 0.35μm CMOS [21]. The chip also includes a single SPAD line for 8-bit time-uncorrelated photon counting.

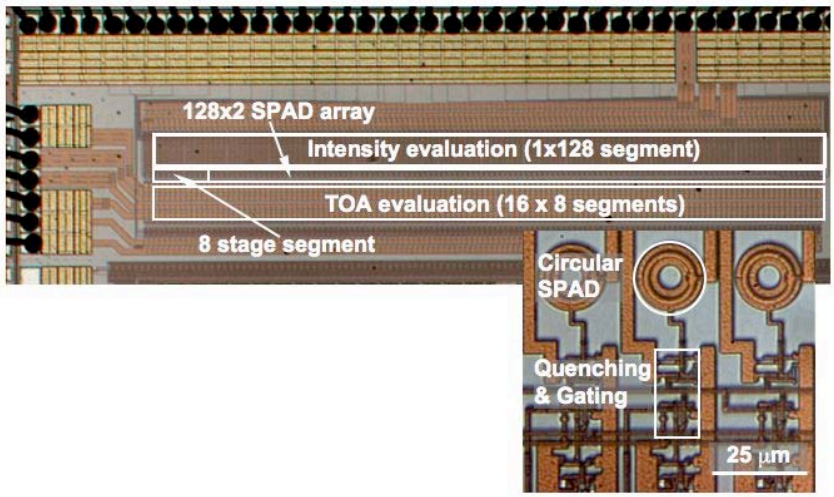

Figure 12. Integrated version of the latchless pipelined readout implemented in 0.35μm CMOS technology [21].

The first design implementing parallel on-chip time discrimination was LASP [22], a 128x128 SPAD array, where a row of 128 SPADs can be randomly selected and processed at high speed. Figure 13 shows the block diagram of LASP. A row of 128 SPADs can be randomly selected for TOA processing. A bank of 32 TDCs shared on a 4-to-1 basis is used for the time conversion to digital code. Each TDC can generate 10MS/s with a time resolution of 97ps.

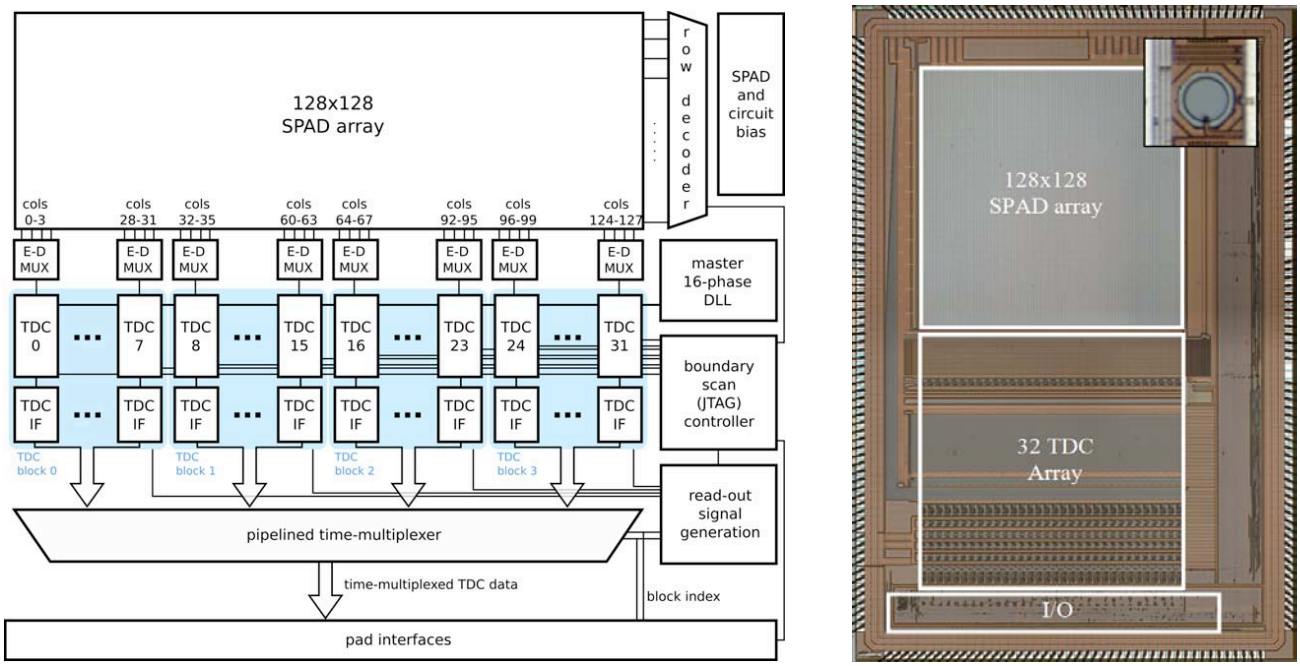

Figure 13. Block diagram of LASP, a fully integrated SPAD array with a bank of on-chip TDCs (left); Photomicrograph of the chip implemented in 0.35μm CMOS technology (right). The inset shows a detail of the pixel [22].

Each TDC in LASP operates in cascade mode, generating the 2 MSBs via a clocked counter, the 4 intermediate bits with a phase interpolator controlled by a temperature-compensated DLL, and the 4 LSBs by means of a Vernier line. The total time resolution of 10bits is subsequently routed to the exterior of the chip through a high-speed digital network operating at 3.2Gb/s. The differential and integral non-linearity of the TDCs was evaluated in detail in [22] to be in a range of +/-0.2LSB and +/-1.2LSB, respectively. The dead time was fixed to 100ns to allow a complete time-to-digital conversion at reasonable afterpulsing levels. The uniformity of PDP and its spectral behavior as well as the chip DCR were consistent with the data reported in [12]. The main drawback of the LASP architecture is the partial parallelism that is essentially column-based, thus limiting the efficiency of photon counting over the entire chip.

In the third approach, time discrimination, photon counting, and any additional functionality, including local storage, are performed on pixel. The advantage of this approach is the massive parallelism that can be achieved, thus potentially improving the amount of photons that can be detected and processed at the same time at reasonable power consumption. There exist many embodiments of this approach depending on the level of complexity implemented on pixel. The simplest one, shown in the schematic diagram of Figure 14, demonstrates the detection and storage of photons at pixel level.

In this design, the avalanche voltage is sensed by M2 that forces the latch to logic level “1”. Transistor M7 acts as pull-down of the column line that is kept high by resistor R_{PU}, while M6 is the row selection switch, controlled by “RowSEL”. When the column is pulled down, a buffer (not shown in the schematic) controls a pad and the output of the chip for that column is interpreted as a photon detected in the previous interval of time. Transistors M4 and M5 are controlled respectively by column line (“ColSET”) and row line (“RowSET”) to force the static memory of a specific pixel to logic level “1”, irrespective of the SPAD state, for testing purposes. M8 is used to operate a global or row based reset via signal “gRESET”, whereas M3 prevents memory conflicts in case of a SPAD firing during reset. SPAD

quenching and recharge are performed by transistor M1 that can be adjusted globally via signal “BIAS”, so as to select a proper trade-off between dead time and afterpulsing probability. The pixel comprises ten NMOS and only two PMOS transistors, thus enabling minimization of NWELL surface, hence ensuring a pitch of 30μm.

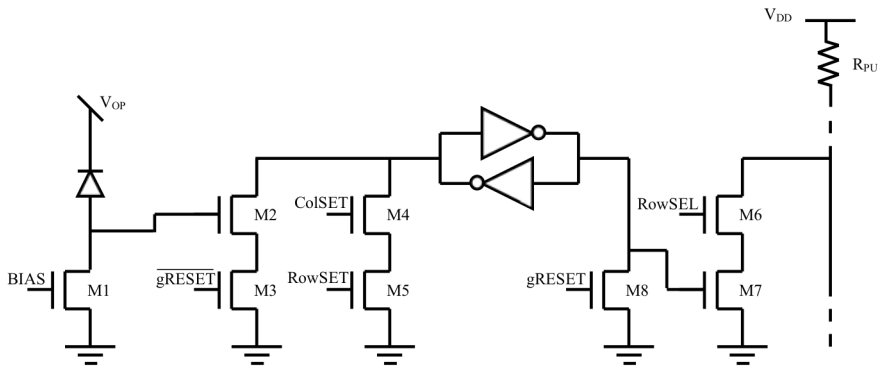

Figure 14. Schematic diagram of the pixel with embedded 1-bit counter. The counter is implemented as a static memory. The content of the counter is read out using a simple pull-down transistor and it may be set and reset using appropriate controls.

The chip micrograph is shown in Figure 15. The inset also shows a detail of the pixels and their column data readout interconnect and row-wise control lines. To construct images with multi-bit gray levels, a high-frequency readout was put in place capable of reading an entire 1-bit frame in 2.88μs [23]. Thanks to the speed of this architecture, moderate time-resolution techniques, such as fluorescence correlated spectroscopy (FCS) are possible on a much larger pixel scale than earlier attempts [24].

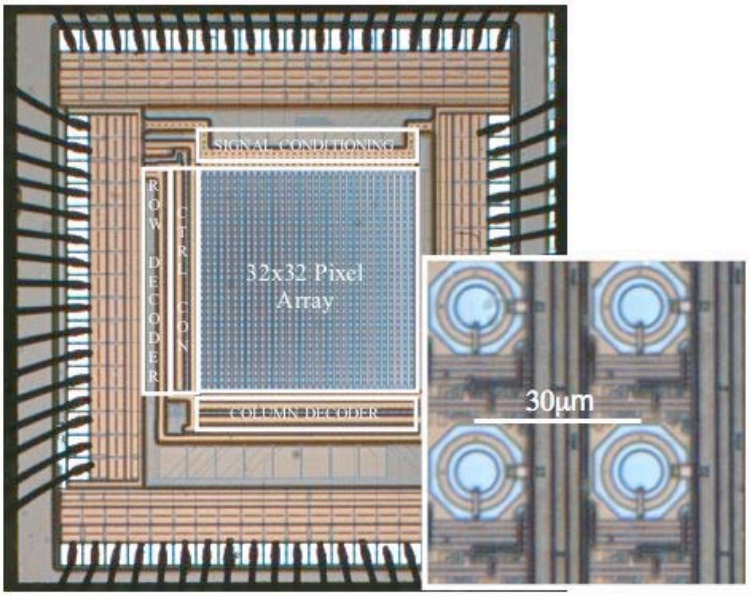

Figure 15. Photomicrograph of RADHARD2, a 32x32 parallel-counting pixel array implemented in 0.35μm CMOS technology. The inset shows a zoom of 4x4 pixels [23].

More recently, with the implementation of the first SPADs in 130nm CMOS technologies [13],[6] and 90nm [25], it has been possible to integrate more functionality on pixel. The pixels of the array in the MEGAFRAME project for

example comprise a multi-bit counter and a picosecond resolution TDC [19],[26],[27]. One of the available implementations of the MEGAFRAME concept comprises an array of 32x32 pixels *each* of which is capable of performing TOA measurements with picosecond resolution and digital photon counting; it was conceived to operate both in TCSPC and time-uncorrelated photon counting (TUPC) modes. In TCSPC mode, the TDC in each pixel is enabled; it can determine and store the first of 10 TOA measurements in every frame of a length of a microsecond. In TUPC mode the counter in each pixel is enabled; it can count up to 64 photon arrivals per microsecond. Figure 16 shows a photomicrograph of the implementation of MEGAFRAME reported in [19] and [28]. The design includes a phase-lock loop (PLL) frequency synthesizer that generates the clock signals necessary to operate the TDCs. A I^2C block also integrated in the chip manages the various modes of operation seamlessly. The overall chip and pixel architectures are described in more detail and fully characterized in [19],[28].

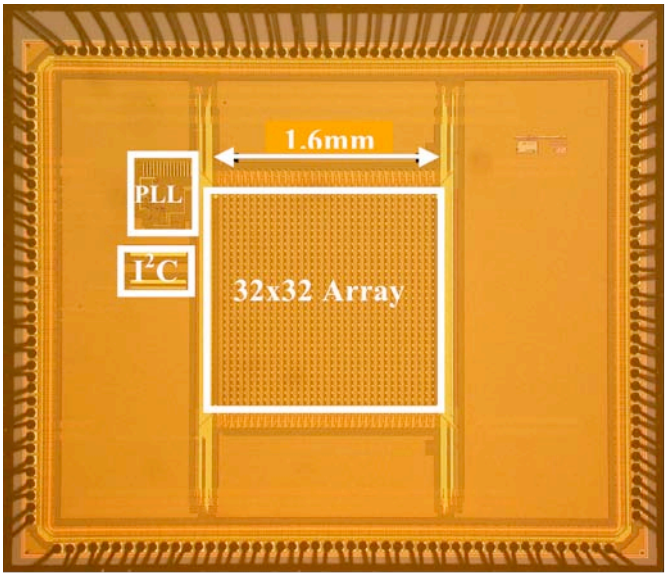

Figure 16. Photomicrograph of MEGAFRAME, a 32x32 pixel array, capable of performing 1 million TOA evaluations per pixel per second at 119ps time resolution.

Table 2 is a summary of the performance of four representative image sensors characterized by random-access, event-driven readout and on-pixel TOA evaluation. The imagers were implemented in a variety of CMOS processes and thus a fair comparison is not possible. Since the first two architectures in Table 2 do not include integrated TACs/TDCs, we report the timing uncertainty as it is evaluated externally, using a commercial TDC, while timing resolution and differential/integral non-linearity are reported elsewhere. The overall pixel bandwidth refers to the maximum symbol rate that the image sensor can generate per pixel (irrespective of whether TOA is computed on- or off-chip). When TOA is computed off-chip, we assumed that processing speed is not limited by the TAC/TDC used but by intrinsic I/O speed. This is why the design of [22] is penalized in the table with respect to the two previous designs, as the integrated TDCs are the bottleneck. In this design in fact, only one row can be operational at each time, while each four columns share a TDC. Thus, the overall TDC bandwidth of 10MS/s must be divided by 4 times 128, to reach the reported value.

In the design of [19] a bandwidth of 1MS/s in TCSPC mode can be achieved, while a much higher count rate is possible thanks to an on-pixel 6-bit counter. Thus, the maximum count rate is limited by the dead time of 100ns. The timing uniformity, wherever measured, is expressed in % or LSB depending on the presence of TOA evaluation on-chip.

	Measurement	Min	Typ	Max	Unit
Random-access readout [8] Architecture style (3)	Format		32x32		-
	Fill factor		1.1		%
	Timing uncertainty or jitter (FWHM)		115		ps
	Pixel pitch		58		μm
	Overall pixel bandwidth		10		kS/s
	Count rate			10	kc/s
	Timing uniformity			1	%
	Power dissipation			6	mW
Event-driven readout [10],[12] Architecture style (2)	Format		4x112		-
	Fill factor		12.5		%
	Timing uncertainty or jitter (FWHM)		80		ps
	Pixel pitch		25		μm
	Overall pixel bandwidth		223		kS/s
	Count rate			223	kc/s
	Timing uniformity		-		%
	Power dissipation			7	mW
In-column TOA evaluation [22] Architecture style (2)	Format		128x128		-
	Fill factor		6		%
	Timing resolution (LSB)		97.68	70	ps
	Differential/integral non-linearity (DNL/INL)			0.08/1.89	LSB
	Pixel pitch		25		μm
	Overall pixel bandwidth		19.5	2,500	kS/s
	Count rate		19.5	2,500	kc/s
	TOA uniformity		-		LSB
	Power dissipation		33	150	mW
On-pixel TOA evaluation [19] Architecture style (1)	Format		32x32		-
	Fill factor		1		%
	Timing resolution (LSB)		119	111	ps
	Differential/integral non-linearity (DNL/INL)		0.4/1.2		LSB
	Pixel pitch		50		μm
	Overall pixel bandwidth		1,000		kS/s
	Count rate			10,000	kc/s
	TOA uniformity		2		LSB
	Power dissipation		93.6		mW

Table 2. Performance of CMOS SPAD imagers for three representative architectures.

4. CONCLUSIONS

In this paper we have reviewed the most important architectures available today in the context of SPAD image sensors implemented in CMOS technologies. For the architecture selection it was shown how critical the target application is, while proper circuit design techniques can be used to reduce the impact of supply and substrate noise.

In summary, with this research we have demonstrated that it is possible to implement significant functionality together with single-photon detection capability in deep-submicron CMOS chips with a performance comparable to that of state-of-the-art single-pixel detectors implemented in dedicated technologies, but with a massive number of pixels operating simultaneously. The applications are endless, from biomedicine to chemistry, from engineering to entertainment.

ACKNOWLDGEMENTS

The author is grateful to his current and former graduate students and post-doctoral fellows that made this research possible. Special thanks go to Lucio Carrara, Marek Gersbach, Cristiano Niclass, and Maximilian Sergio who were responsible for the designs outlined here, as well as Fausto Borghetti, Claudio Favi, Matthew Fishburn, Robert Henderson, Mohammad Karami, Theo Kluter, Estelle Labonne, Yuki Maruyama, Justin Richardson, David Stoppa, and Richard Walker who co-designed the chips. The author acknowledges Giordano Beretta, Claudio Bruschini, Dmitri Boiko, Neil Gunther, Lindsay Grant, David Li, and Luciano Sbaiz for useful discussions.

REFERENCES

[1] E. Charbon, “Will CMOS Imagers Ever Need Ultra-High Speed?”, *IEEE International Conference on Solid-State and Integrated-Circuit Technology*, 1975-1980 (2004).

[2] S. Cova, A. Longoni, and A. Andreoni, “Towards Picosecond Resolution with Single-Photon Avalanche Diodes”, *Rev. Sci. Instr.*, 52(3), 408-412 (1981).

[3] R.J. McIntyre, “Recent Developments in Silicon Avalanche Photodiodes”, *Measurement*, 3(4), 146-152 (1985).

[4] A. Spinelli and A. L. Lacaita, “Physics and Numerical Simulation of Single Photon Avalanche Diodes”, IEEE Trans. on Electron Devices, 44(11), 1931-1943 (1997).

[5] A. Rochas *et al*., “Single Photon Detector Fabricated in a Complementary Metal–oxide–semiconductor High-voltage Technology”, *Rev. Sci. Instr.*, 74(7), 3263-3270 (2003).

[6] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, E. Charbon, “A Low-Noise Single-Photon Detector Implemented in a 130 nm CMOS Imaging Process”, *Solid-State Electronics*, 53(7), 803-808 (2009).

[7] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, “Single-Photon Synchronous Detection”, *IEEE Journal of Solid-State Circuits*, 44(7), 1977-1989 (2009).

[8] 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*, 40(9), 1847-1854 (2005).

[9] S. Tisa, F. Zappa, I. Labanca, “On-chip Detection and Counting of Single-photons”, *IEEE International Electron Device Meeting*, 815-818 (2005).

[10] C. Niclass, M. Sergio, and E. Charbon, “A Single Photon Avalanche Diode Array Fabricated in Deep-Submicron CMOS Technology”, *IEEE Design, Automation & Test in Europe*, 1-6 (2006).

[11] H. Finkelstein, M. J. Hsu and S. C. Esener “STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology,” *IEEE Electron Device Lett*., 27, 887 (2006).

[12] C. Niclass, M. Sergio, E. Charbon, “A Single Photon Avalanche Diode Array Fabricated in 0.35μm CMOS and based on an Event-Driven Readout for TCSPC Experiments”, *SPIE Optics East*, Boston, (2006).

[13] C. Niclass, M. Gersbach, R.K. Henderson, L. Grant, E. Charbon, "A Single Photon Avalanche Diode Implemented in 130nm CMOS Technology", *IEEE Journal of Selected Topics in Quantum Electronics*, 13(4), 863-869 (2007).
[14] D. Stoppa, L. Pacheri, M. Scandiuzzo, L. Gonzo, G.-F. Della Betta, A. Simoni, "A CMOS 3-D Imager Based on Single Photon Avalanche Diode", *IEEE Trans. on Circuits and Systems*, 54(1), 4-12 (2007).
[15] L. Pancheri and D. Stoppa, "Low-noise CMOS Single-photon Avalanche Diodes with 32ns Dead Time", *IEEE European Solid-State Device Conference*, (2007).
[16] N. Faramarzpour, M.J. Deen, S. Shirani, and Q. Fang, "Fully Integrated Single Photon Avalanche Diode Detector in Standarrd CMOS 0.18-um Technology", *IEEE Trans. on Electron Devices*, 55(3), 760-767 (2008).
[17] C. Niclass, M. Sergio, and E. Charbon, "A CMOS 64x48 Single Photon Avalanche Diode Array with Event-Driven Readout", *IEEE European Solid-State Circuit Conference*, (2006).
[18] J. Richardson, L. Grant, R. Henderson, "A Low-dark Count Single-photon Avalanche Diode Structure Compatible with Standard Nanometer Scale CMOS Technology", *International Image Sensor Workshop*, (2009).
[19] M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, R. K. Henderson, F. Borghetti, D. Stoppa, E. Charbon, "A Parallel 32x32 Time-to-Digital Converter Array Fabricated in a 130nm Imaging CMOS Technology", *IEEE European Solid-State Device Conference*, (2009).
[20] M. Gersbach, D. L. Boiko, C. Niclass, C. Petersen, E. Charbon, "Fast Fluorescence Dynamics in Nonratiometric Calcium Indicators", *Optics Letters*, 34(3), 362-364, (2009).
[21] M. Sergio, C. Niclass, E. Charbon, "A 128x2 CMOS Single Photon Streak Camera with Timing-Preserving Latchless Pipeline Readout", *IEEE Intl. Solid-State Circuits Conference*, 120-121 (2007).
[22] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128x128 Single-Photon Image Sensor with Column-Level 10-bit Time-to-Digital Converter Array", *IEEE Journal of Solid-State Circuits*, 43(12), 2977-2989 (2008).
[23] L. Carrara, C. Niclass, N. Scheidegger, H. Shea, E. Charbon, "A Gamma, X-ray and High Energy Proton Radiation-Tolerant CMOS Image Sensor for Space Applications", *IEEE Intl. Solid-State Circuits Conference*, 40-41 (2009).
[24] A. Rochas, M. Gösch, A. Serov, R.S. Popovic, T. Lasser, and R. Rigler, "First Fully Integrated 2-D Array of Single-Photon Detectors in Standard CMOS Technology", *IEEE Photonics Technology Letters*, 15(7), (2003).
[25] M. Karami, M. Gersbach, E. Charbon, "A New Single-photon Avalanche Diode in 90nm Standard CMOS Technology", *SPIE Optics+Photonics, NanoScience Engineering, Single-Photon Imaging*, (2010).
[26] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, R. K. Henderson, "A 32x32 50ps Resolution 10 bit Time to Digital Converter Array in 130nm CMOS for time Correlated Imaging", *IEEE Custom Integrated Circuits Conference*, (2009).
[27] D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R.K. Henderson, M. Gersbach, E. Charbon, "A 32x32-Pixel Array with In-Pixel Photon Counting and Arrival Time Measurement in the Analog Domain", *IEEE European Solid-State Device Conference*, (2009).
[28] M. Gersbach, R. Trimananda, Y. Maruyama, M. Fishburn, D. Stoppa, J. Richardson, R. K. Henderson, and E. Charbon, "High Frame-rate TCSPC-FLIM Readout System Using a SPAD-based Image Sensor", *SPIE Optics+Photonics, NanoScience Engineering, Single-Photon Imaging*, (2010).