Large-Area, Fast-Gated Digital SiPM With Integrated TDC for Portable and Wearable Time-Domain NIRS

Enrico Conca, *Member, IEEE*, Vincenzo Sesta, Mauro Buttafava, *Member, IEEE*, Federica Villa, *Member, IEEE*, Laura Di Sieno, Alberto Dalla Mora, Davide Contini, Paola Taroni, Alessandro Torricelli, Antonio Pifferi, *Member, IEEE*, Franco Zappa, *Senior Member, IEEE*, and Alberto Tosi, *Member, IEEE*

***Abstract*— We present the design and characterization of a large-area, fast-gated, all-digital single-photon detector with programmable active area, internal gate generator, and time-to-digital converter (TDC) with a built-in histogram builder circuit, suitable for performing high-sensitivity time-domain near-infrared spectroscopy (TD-NIRS) measurements when coupled with pulsed laser sources. We used a novel low-power differential sensing technique that optimizes area occupation. The photodetector is a time-gated digital silicon photomultiplier (dSiPM) with an 8.6-mm^2 photosensitive area, 37% fill-factor, and ∼300 ps (20%–80%) gate rising edge, based on low-noise single-photon avalanche diodes (SPADs) and fabricated in 0.35-μm CMOS technology. The built-in TDC with a histogram builder has a least-significant-bit (LSB) of 78 ps and 128 time-bins, and the integrated circuit can be interfaced directly with a low-cost microcontroller with a serial interface for programming and readout. Experimental characterization demonstrated a temporal response as good as 300-ps full-width at half-maximum (FWHM) and a dynamic range >100 dB (thanks to the programmable active area size). This microelectronic detector paves the way for a miniaturized, stand-alone, multi-wavelength TD-NIRS system with an unprecedented level of integration and responsivity, suitable for portable and wearable systems.**

***Index Terms*— Digital silicon photomultiplier (dSiPM), fast-gated single-photon avalanche diode (SPAD) array, photon counting, time-to-digital converter (TDC), time-domain near-infrared spectroscopy (TD-NIRS).**

I. Introduction

TIME-DOMAIN near-infrared spectroscopy (TD-NIRS) is a powerful technique for obtaining non-invasive, *in vivo* measurements of tissue constituents and structure [1]. This can be exploited in many scientific fields, from a clinical diagnosis of breast tumor to monitoring of hemodynamics and metabolism in muscle and brain [2]–[4], and even be brought into mass-market applications, such as improved monitoring of athletes' performances and non-destructive analysis of fruit maturity [5].

The main disadvantage of this technique is the complexity of the instrumentation since the state-of-the-art systems require the use of multi-wavelength, narrow (hundreds of picoseconds) laser pulses, relatively large-area single-photon detectors (typically few square millimeters photomultipliers), variable optical attenuators (VOAs) to equalize the signal, and a high-resolution time-tagger to reconstruct the optical waveform [6].

Classical TD instruments working in reflectance geometry (i.e., source and detector are placed on the same side of the scattering medium under investigation) require a large source–detector distance (e.g., 2–3 cm) to avoid the overwhelming burst of early photons at short distances (e.g., 0–5 mm) [7]. These photons represent an unwanted signal since they explored only the outer layers of the scattering medium, but they can saturate the detector preventing the detection of late photons, which instead probed deeper layers.

Such an issue can be overcome with time-gated single-photon detectors, thus improving spatial resolution and sensitivity [8], [9]. With a fast time-gated detector, the source and the detector can be almost in contact: keeping the detector blind during the first few hundred picoseconds after the laser firing, the part of the waveform corresponding to the direct path between the source and the detector (photons traveling in the shallower layers) is eliminated, and only late photons having traveled deeper into the tissue are detected. This also allows us to increase the laser power with respect to a non-gated system to increase the overall sensitivity and obtain deeper penetration inside the tissue. However, so far no detector combining both a large and programmable detection area and fast time-gating exists [1].

The detector presented in this work can be coupled with pulsed laser sources to characterize human tissues or other biological samples, by measuring their absorption and reduced scattering coefficients at different wavelengths, which are, respectively, linked to the sample chemical composition and microstructure [1]. In particular, compact laser sources based on integrated pulsed laser drivers have been recently

Manuscript received December 3, 2019; revised March 17, 2020 and June 4, 2020; accepted June 22, 2020. Date of publication July 13, 2020; date of current version October 23, 2020. This article was approved by Associate Editor David Stoppa. This work was supported by the European Union's Horizon 2020 Research and Innovation Programme under Agreement 731877. *(Corresponding author: Enrico Conca.)*

Enrico Conca, Vincenzo Sesta, Mauro Buttafava, Federica Villa, Franco Zappa, and Alberto Tosi are with the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20133 Milan, Italy (e-mail: enrico.conca@polimi.it; alberto.tosi@polimi.it).

Laura Di Sieno, Alberto Dalla Mora, Davide Contini, Paola Taroni, Alessandro Torricelli, and Antonio Pifferi are with the Dipartimento di Fisica, Politecnico di Milano, 20133 Milan, Italy.

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2020.3006442

reported [10], demonstrating performances in line with the requirements of TD-NIRS. The combination of these sources with microelectronic detectors can pave the way for the fabrication of wearable TD-NIRS systems. Such systems would allow us to have a performance level comparable to that of rack-mounted units [6], but with compact handheld (or even wearable) devices. The resulting solution will go well beyond what is available in the compact systems currently reported in the literature, like the one of [11].

Various microelectronic detectors compatible with the fabrication of wearable systems have been reported in the literature and validated in TD-NIRS applications, but none of those features, at the same time, a large collection area and fast time-gating. For instance, single-pixel single-photon avalanche diodes (SPADs) [12] have been integrated into wearable probes with fast time-gating capability (still with external photon-timing electronics), but with a maximum active area diameter of 200 μm [13], enabling the use of short source–detector separations (e.g., $<$15 mm) but at the expense of a limited collection of backscattered light. On the other hand, analog silicon photomultipliers (SiPMs) have been used in wearable probes, lacking time-gating capability but gaining more than 1 order of magnitude in terms of active area size, enabling their use only at large source–detector separations (i.e., $>$2 cm). The combination of these features would be highly beneficial in TD-NIRS [1], but it is challenging and the only attempt reported so far in literature has been done by placing side by side a time-gated SPAD and an analog SiPM [14]. However, in such configuration, the SPAD remains photon-starved and the SiPM can only operate at large source–detector separations.

The detector here reported has the advantage of featuring at the same time a wide collection area (maximizing the collection of backscattered photons) and time-gating capability (enabling the use of short separations between the source and the detector). Furthermore, the programmable active area and the integrated timing electronics eliminate the need for optomechanical signal attenuation stages (typically used for adjusting the photon-counting rate depending on the source–detector separation and on the signal strength) and for external timing electronics.

We designed our integrated detector for the EC-funded SOLUS project [15], whose ultimate goal is to develop a non-invasive, multimodal imaging system for the diagnosis of breast cancer, by combining multi-point TD-NIRS with traditional ultrasound imaging and shear-wave elastography. By exploiting the information coming from different source/detector locations, multi-point TD-NIRS allows reconstructing 3-D maps of optical properties (i.e., diffuse optical tomography [2]), while single-point TD-NIRS can monitor the average tissue composition and microstructure at a specific location inside the tissue. To this aim, different small optodes need to be placed around the ultrasound transducer in the probe, each one combining pulsed laser sources of different wavelengths and a large-area, time-gated detector. The small footprint ($\sim$1 cm^2) and volume (few cm^3) of the optode required an integrated circuit merging all the functions of the detection chain:

1) a digital SiPM (dSiPM) with large collection area;
2) the fast-gating capability of single SPADs;
3) the programmability of the active area for adjusting the collected signal to fit the single-photon counting statistics;
4) a time-to-digital converter (TDC);
5) a programmable signal generator.

The programmable active area feature can also be used to reduce the overall detector noise by disabling SPADs with a significantly higher than average dark count rate (DCR).

II. Gating Operation

It is well-known that a single-pixel SPAD can be operated in the so-called "fast-gated mode," that is, it can be swiftly turned from OFF to ON with a rising edge faster than 1 ns, while still being able to detect avalanches during the transition [16]. This is done by modulating the detector voltage from below breakdown (where no photon can trigger an avalanche) to few volts above it, where photons can trigger a self-sustained avalanche that can be easily read by the front-end circuit.

In addition, many wide SPAD arrays have been developed in the CMOS technology for counting and timing single photons [17]–[20]. However, CMOS SPADs are typically coupled with circuits not designed for fast-gating operation, being not capable of properly detecting photons arriving during (or just after) the gate rising edge. A common simple gating solution used in some SPAD arrays is to mask the photon detection output: the avalanche pulse can be read only within well-defined time intervals [21]. This solution keeps the SPADs in free-running operation but allows the digital avalanche pulse from the SPAD to reach the processing electronics only during the mask window. It is a simple and low-power solution, but it is not suitable when a very strong light pulse arrives just before the faint optical signal to be acquired (i.e., the typical scenario of TD-NIRS when the source and the detector are placed in close proximity): the strong pulse may blind the pixels, thus strongly limiting the number of SPADs available for detecting the few late photons of interest.

Other works presented in literature use actively gated detectors, but operated only in photon-counting mode, with no information on the photon arrival time within the gate [14], [20], [22], [23]. In such approaches, by shifting the gate position with respect to the excitation laser and acquiring multiple measurements, low temporal resolution waveforms can be acquired, but the measurement time is increased if the incoming photon flux is very low.

In conclusion, SPAD-based solutions currently available in the literature are not well-suited for TD-NIRS since either the area is too small or the gating is not effective, or they do not preserve the time of arrival of the photons, requiring longer acquisition times which become impractical in clinical applications.

A. Fast-Gating Techniques

Single-pixel fast-gated SPAD systems are mainly based on the SPAD–dummy approach, where a SPAD is coupled with a dummy structure (typically another SPAD biased below

its breakdown voltage) to mimic the parasitic capacitance of the detector and cancel the disturbances introduced by the gating operation by means of a differential readout [24]. This approach guarantees good temporal response, with no pile-up distortion in correspondence to the gate window rising edge.

Integrated circuits dedicated to the fast-gated operation and based on the SPAD–dummy readout have been already presented, with very good performance, but their high power consumption due to fast comparators and high-current pulsers impair their use in wide-area detectors [25].

In addition, the use of two SPADs as independent photodetectors while operating them in the gated mode has been presented in literature: two discrete InGaAs/InP SPADs were used, allowing to obtain two time-gated detection channels without introducing the dummy structures [26]. However, the hybrid junction there used cannot be used for building wide-area SPAD arrays with fast-gating capability.

B. Differential Sensing Technique

The wide-area fast-gated dSiPM presented here is based on pixels where the two SPADs needed for the fast-gated approach are both photosensitive and active. Its monolithic CMOS integration reduces the parasitic capacitance between the detector and the readout circuit, thus reducing the power consumption.

The use of two photosensitive elements to implement the cancellation of the gating disturbances also helps reducing the power consumption (by removing the power lost to gate a blind dummy structure) and saves the area occupation related to the dummy element, yielding a higher fill-factor.

Since both SPADs are able to detect photons during the gate window, we need a circuit able to distinguish a differential imbalance in either directions, that is, when either one or the other fires, and to reject the common-mode signals. A solution based on the two comparators with switched input terminals followed by an OR gate would be expensive both in terms of area occupation and power consumption, thus not being a realistic approach for a large array.

In our SPAD–SPAD implementation, we replaced the two comparators and the OR gate with a fully digital XOR gate, directly connected to the SPAD anodes. Since the XOR gate provides an output only when the two inputs are at different logic levels, it performs both the rejection of the common-mode gating disturbances (same signal at both anodes) and the extraction of the photon detection pulse from both SPADs, with a much smaller area and lower power consumption.

One drawback of the SPAD–SPAD gating technique is that if an avalanche starts in both SPADs within a temporal distance shorter than the time required for the logic to properly detect one avalanche, there would be no output because the voltage at both XOR inputs would be the same and the two SPADs' avalanches would not be quenched till the next gate period (thus causing an increase in the power consumption).

III. ASIC Design

The ASIC was developed in a 0.35-μm HV-CMOS technology and it is composed of five main blocks (as shown in Fig. 1): the photodetector, the TDC, the histogram builder, the signal generator, and the digital interface, laid out as shown in Fig. 2. Given the very small volume ($\sim$1 cm^3) of the optode where this chip will be mounted, we minimized the external connections: it requires two voltage supplies (SPAD bias, about 28 V, and a 3.3-V supply) and a total of eight I/O pads (four-wire SPI—serial peripheral interface—and two LVDS—low voltage differential signaling—pairs). Given the area constraints inside the optode, the pads have been limited in number and laid out only along three sides. The chip size is 6 $\times$ 7 mm^2 and most of it is dedicated to the photodetector (see the layout micrograph in Fig. 2), which is composed of 1728 pixels, organized in four quadrants of 432 pixels each, with a total active area of 8.6 mm^2 over an area of 4.9 $\times$ 4.7 mm^2 (i.e., fill-factor is 37%).

The detector is a dSiPM, meaning that the output is the logical OR of the outputs of all the pixels [27]. When a pixel

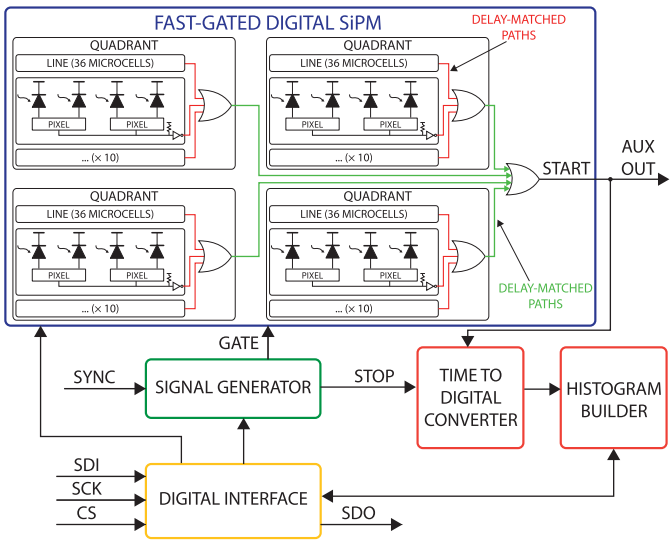

Fig. 1. Block diagram of the integrated circuit showing the main building blocks: fast-gated dSiPM, TDC with histogram builder, signal generator, and digital interface.

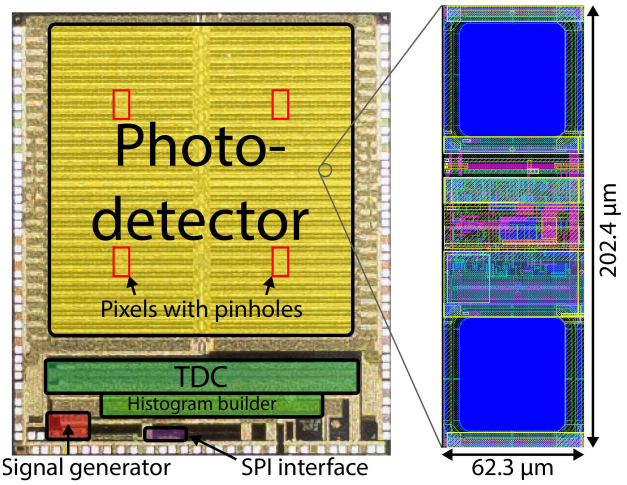

Fig. 2. Chip micrograph with highlighted main functional blocks: photodetector, TDC with histogram builder, signal generator, and digital interface (left). The chip size is 6 $\times$ 7 mm^2. Detail of a single-pixel layout showing the pixel electronics between the two SPADs (right).

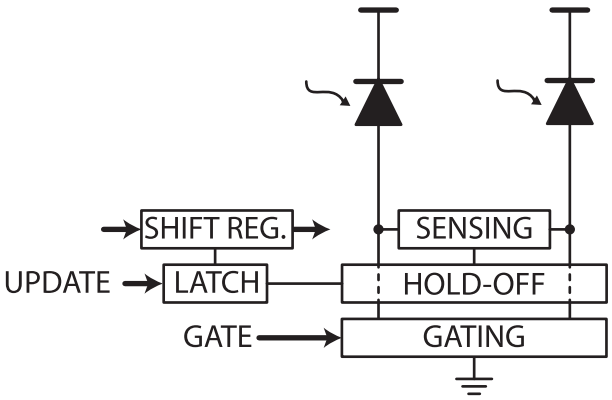

Fig. 3. Block diagram of one pixel.

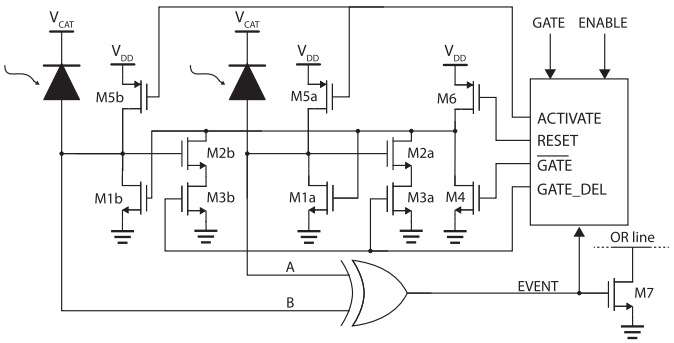

Fig. 4. Schematic of the pixel front-end and sensing circuits.

detects a photon, a digital pulse propagates over an OR-tree, designed to preserve the time of arrival of such photon, both in terms of signal integrity and propagation delay. Due to the finite duration of such a digital pulse, it is possible to detect one photon per gate window, whose duration is in the order of few (4–8) nanoseconds. If multiple pixels fire during the same gate window, only the first one's time of arrival is preserved, and all the others are lost. This is perfectly acceptable in TD-NIRS applications, where the detected photon rate is low enough that the probability of having two signal photons during one gate window is negligible. At the same time, this detector architecture masks the optical crosstalk events between different pixels, because the crosstalk avalanche is just slightly delayed (few hundreds of picoseconds) with respect to the photon avalanche so that the OR tree discards it. Therefore, crosstalk probability cannot be measured for this detector, but it is not important in TD-NIRS applications. The architecture also includes a 24-bit counter, in addition to the TDC, connected to the output of the OR-tree to measure the total number of photons detected by the SiPM, with a maximum count rate of 100 MHz (limited by one event per gate at the maximum detector gating frequency), which is readout and reset simultaneously with the TDC readout.

A. Photodetector

Each pixel of the dSiPM includes (see Fig. 3) the following:

1) a pair of SPADs connected to a gating circuit;
2) a differential sensing circuit based on a XOR gate;
3) a hold-off circuitry;
4) a latch for enabling/disabling the pixel.

To minimize power dissipation, we limited the SPAD excess bias to 3.3 V, instead of the 5-V excess bias typically used for operating SPADs in this 0.35-μm technology. The drawbacks are a detection efficiency reduced by about 30% over the entire range [e.g., peak detection efficiency (PDE) decreases from 50% to 35%] and a bit wider temporal performance (negligible when compared with other contribution of the overall TD-NIRS system). With the lower voltage, we could use thin oxide transistors, which are smaller and faster for better gating performance.

The overall layout of the pixel is reported in Fig. 2 right: SPADs are square with a side of $\sim$50 μm and rounded corners with a radius of 5 μm. The choice of the SPAD size resulted from a trade-off between higher fill-factor (requiring wider detectors) and fewer hot pixels (requiring smaller detectors, where the probability of having hot pixels is smaller), where a hot pixel is a pixel with a DCR more than 10 times higher than the average DCR. In addition, SPADs larger than 50 μm also show worse temporal response.

To dynamically adjust the detector sensitivity based on the incoming photon flux without the need of external VOAs, which would be impossible to fit inside the compact optode, we introduce an on-chip signal equalization capability: by selectively enabling/disabling each pixel through the SPI interface, the detector active area can be varied, thus changing the overall sensitivity of the detector over a range of 32 dB (1728/1). Furthermore, to expand even more the equalization dynamic range, a subset of pixels includes SPADs with the active area covered by metal pinholes of different diameters (5, 10, 20, and 40 μm) to reduce their effective collection area. Each quadrant contains eight pixels with metallic pinholes, two for each pinhole size, for a total of 32 pixels with pinholes in the SiPM. This way, the dynamic range over which the sensitivity can be adjusted is extended from 32 to 53 dB. SPADs with pinholes were preferred to small SPADs to preserve the same temporal response. It is possible to change the active area in real-time (in less than 500 μs), with the new configuration being updated simultaneously at the end of the programming sequence.

SPADs share the cathode well to reduce isolation distances [28] and are packed as tightly as allowed by design rules. Optical crosstalk between the two SPADs in the same pixel, which would cause a counting loss, is minimized by placing all the pixel electronics between the two SPADs [see Fig. 2 (right)], thus maximizing their distance for a given fill-factor.

The gating circuitry (whose schematic is reported in Fig. 4) allows activating the SPADs by simultaneously lowering both the anodes to ground and to quench them by bringing their anodes to V_{DD} (3.3 V). Furthermore, it allows to enforce the hold-off or disable the detector by keeping the anodes tied to V_{DD}.

The main components of the gating circuit are M1a–b ($W/L = 2.5/0.35$) and M5a–b ($W/L = 8/0.35$), which are sized to charge and discharge the anode capacitance (estimated at 150 fF per SPAD) in $\sim$300 ps to obtain sharp gating transitions. To minimize the power consumption, they are not

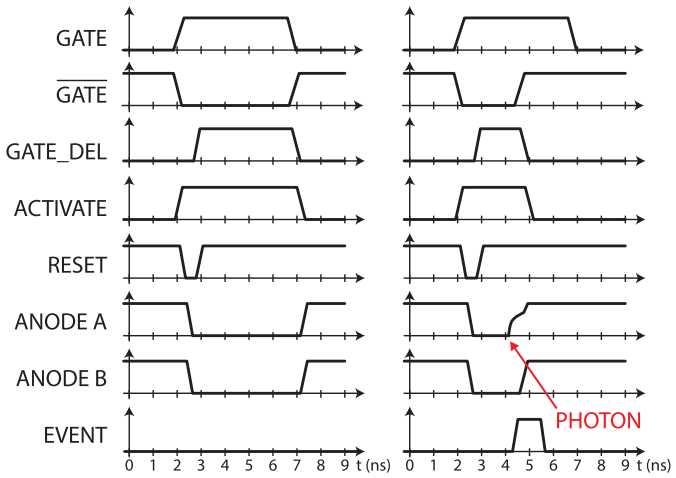

Fig. 5. Waveforms of the main signals inside the pixel when no photon detection occurs (left) or when a photon detection occurs (right). Time scale is an example of a typical operating condition.

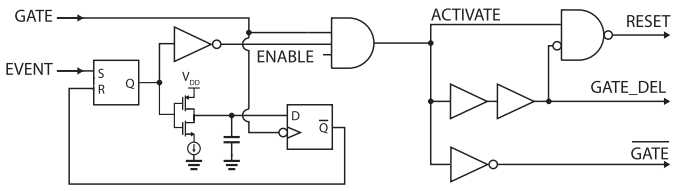

Fig. 6. Detail of the signal sequencing and hold-off logic.

controlled as a simple inverter, but they are driven with two asymmetric signals to have a dead time avoiding crowbar currents flowing directly from V_{DD} to GND. A simplified temporal diagram of the main control signals is reported in Fig. 5.

The gating M5a–b transistors are driven by the ACTIVATE signal, normally a copy of the GATE input signal, except during detector deactivation or hold-off, when it is kept at GND for having the SPAD reverse bias below breakdown. The gate terminals of M1a–b instead are brought to V_{DD} at the beginning of the gate window through M6 ($W/L = 4/0.35$) and remain there thanks to the stray capacitances. M3a–b ($W/L = 2/0.35$) are kept OFF during the first few hundreds of picoseconds of the gate window to prevent cross-conduction through M6 and M2a–b ($W/L = 2/0.35$), since M2a–b's gate terminals are connected to the anodes and start at V_{DD} at the beginning of the gate window. The purpose of M2a–b is to provide a prompt quenching of the avalanche current by quickly turning OFF M1a–b as soon as either of the anodes rises above M2's threshold voltage. M4 ($W/L = 2/0.35$) is activated during the gate OFF period to keep M1 OFF.

The hold-off time is set to ~50 ns by a constant current discharge of a capacitor. To avoid timing distortion due to detector reactivation in a random position of the gate window, a flip-flop is used to resynchronize the hold-off status with the gate signal, ensuring that all SPADs simultaneously activate at the gate opening (see Fig. 6). This feature is important mainly when the internal TDC is not used.

The sensing circuitry is made of an XOR gate whose inputs are the SPADs' anodes. To have identical propagation delays for both inputs, we implemented a 16-transistor XOR gate by splitting each pull-up and pull-down path in two paths, where the transistors' positions are swapped, to guarantee that the output transition always involves the same number of transistors and sees the same internal node capacitances [29]. This solution effectively removes any deterministic skew between the two inputs.

In other works (see [30]), a binary tree of XOR gates is used to combine the outputs of all pixels into a single digital line to maximize the achievable count rate. Our implementation is different, as the XOR gate is used to suppress common-mode disturbances and extract the signal from the SPADs and not to combine the output of more pixels, since our target application is photon-starved, with a count rate of few megahertz over the entire array. Indeed, with a XOR-tree, if two SPADs were to fire simultaneously over the entire array, the event would be lost, while in our approach the event is preserved unless the firing SPADs belong to the same pixel.

We adopted a hybrid approach to combine the outputs of all pixels, based on open-drain lines shared by each group of 36 pixels constituting a row within a quadrant, followed by an OR-tree with matched delay paths for the remaining stages (as shown in Fig. 1). Although a fully binary tree is the ideal approach to combine all pixels' outputs, its implementation in a time-gated chip is challenging, as high peak currents could affect differently the OR gates placed in different locations inside the chip, due to IR drops over the supply lines. Our approach allows us to place all the OR-related circuitry toward the center of the array, where it is powered with separate power supply lines to ensure a more stable voltage.

Similarly, the gate signal is distributed along the array with delay-matched paths to each detector row, where simple buffers distribute the signal from the center of the array toward the outer parts.

Decoupling capacitors between the 3.3-V supply and the ground have been placed wherever possible on the chip to help with high peak currents required for the gating operation (simulated to be more than 2 A when gating all the pixels).

B. Time-to-Digital Converter and Histogram Builder

TD-NIRS optical signals present a fast decay in intensity during the first few nanoseconds after the injected laser pulse and typically do not contain features faster than few hundreds of picoseconds. As such, a TDC specifically designed for this application typically requires a full-scale range (FSR) of few nanoseconds (<10 ns is enough) and a temporal resolution of about 100 ps. However, to develop a very compact optode, it should include a histogram builder circuit able to convert up to 10^7 events per second (with a laser repetition rate of up to 100 MHz).

We designed a TDC with FSR ~9 ns, a nominal resolution of 72 ps, and conversion time shorter than 100 ns to reconstruct the waveform with high precision and without distortion. The block diagram of the TDC is shown in Fig. 7. The fully differential architecture reduces disturbance and noise effects. The TDC core implements a Vernier delay line architecture, whose resolution is defined by the difference between

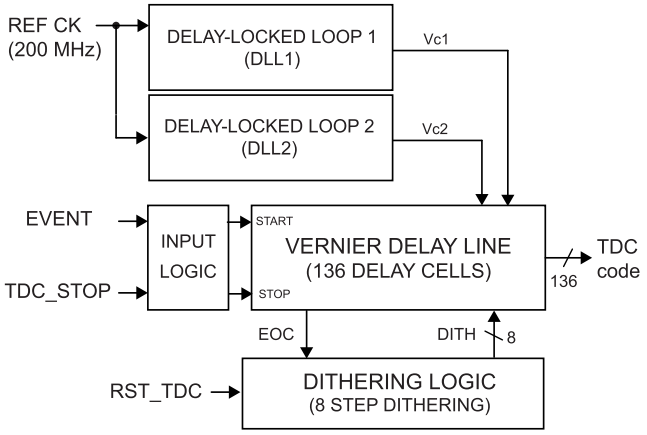

Fig. 7. Block diagram of the TDC with dithering logic.

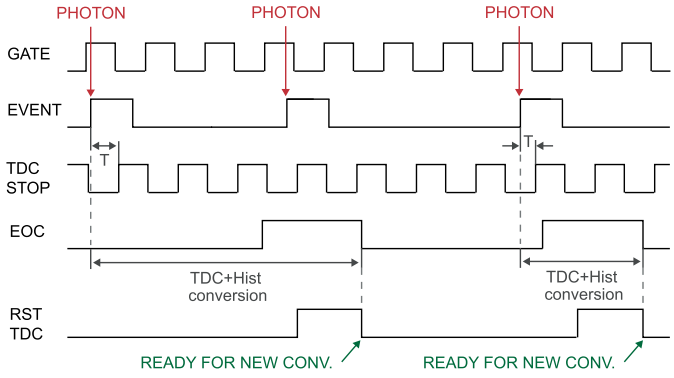

Fig. 8. Timing diagram of the TDC and the histogram builder.

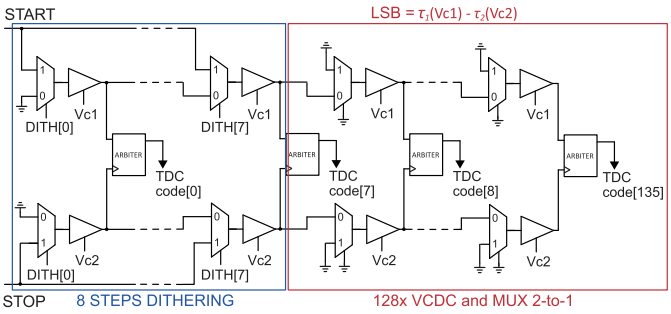

Fig. 9. Schematic of the Vernier delay line with the eight steps dithering implemented in the TDC.

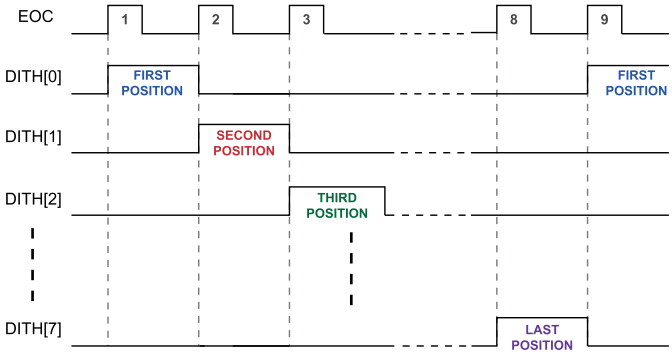

Fig. 10. Dithering signals' timing scheme. When the EOC is asserted, the successive START and STOP are injected into the next position of the previous eight delay cells. As soon as the last position is reached, the dithering cycle starts again with the first position.

propagation delays of two cells [31]. This approach is based on two delay lines in which the START and STOP pulses propagate through. The propagation delay of each cell in the START and STOP lines is fixed by the delay-locked loop (DLL) and the propagation delay of the delay cell in the STOP line is shorter than the propagation delay of the cell in the START line. The conversion is terminated when the START pulse arrives later than STOP. The conversion result is given by the number of delay cells needed to reach this condition.

The timing diagram of the TDC is reported in Fig. 8. The EVENT signal, coming from the SiPM, reaches the input logic, which starts the propagation of the EVENT pulse along the START delay line; the next STOP pulse, from the signal generator, is also allowed to propagate in the STOP delay line. The time interval between these two signals is converted by the TDC, with a conversion time that depends on the time interval. Once the conversion is complete, the TDC asserts its end-of-conversion (EOC) signal, whose rising edge updates the histogram; afterward, the TDC is reset and the circuit is ready for a new conversion. The total dead time varies between 30 and 115 ns, depending on the position along the delay line where the START pulse overcomes the STOP pulse. EVENTs arriving during a TDC conversion are ignored by the TDC, but they are still counted by the 24-bit photon counter.

The TDC Vernier delay line is composed of 136 voltage-controlled delay cells (VCDCs) and an arbiter circuit [32], [33], based on the symmetric fast latches, to properly sample the state of START delayed in each cell by means of the STOP rising edge delayed in each cell (see Fig. 9). The arbiter circuit is used instead of a classic D-type flip-flop to guarantee fast transition when the time delay between START and STOP signals becomes shorter than 1 least-significant-bit (LSB).

Given a large number of delay cells, the effect of cell-to-cell mismatch will be quite significant and would impair the converter linearity. As a solution, we implemented deterministic dithering, based on eight discrete dithering steps, to improve the linearity of the converter. Dithering is obtained by changing the injection point of the signal over eight steps to implement the sliding scale technique [33], [34].

Dithering is implemented by adding a 2-to-1 multiplexer at the input of each delay cell: the first 8 multiplexers are used to implement this feature, while the others are left to improve matching. This solution allows us to propagate the START and STOP pulses starting at eight different positions along the delay lines. At the end of each conversion, indicated by the EOC signal, the following START and STOP pulses are periodically injected into the next position of the first eight delay cells, as shown in Fig. 10. In this way, even if the same START–STOP time interval is converted, different portions of the START and STOP lines are exploited, and thus the linearity of the converter is significantly improved. Dithering is thus equivalent to adding a known delay to the input signal and subtracting the corresponding digital code from the conversion result. The additional TDC area due to

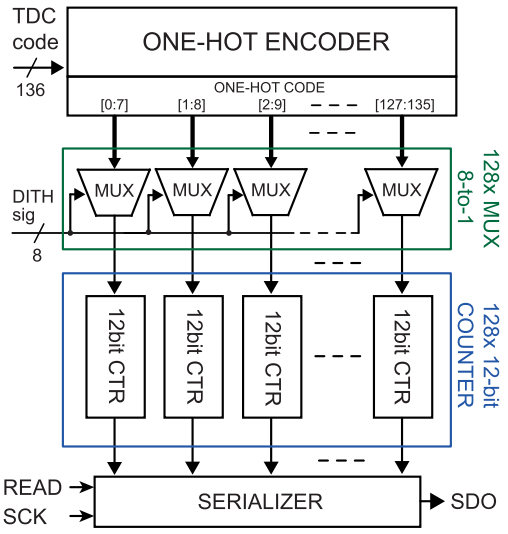

Fig. 11. Block diagram of the histogram builder circuit.

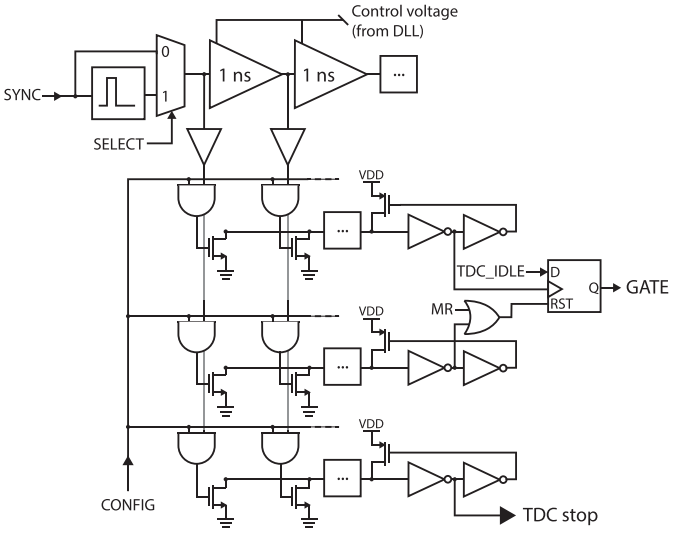

Fig. 12. Schematic of the internal signal generator.

the dithering circuit (mainly the multiplexers placed along the delay chain, as shown in Fig. 9, and at the output of the arbiters, as shown in Fig. 11) is about 0.68 mm^2, that is, about 1.6% of the die.

DLLs are used to guarantee stability against process–voltage–temperature variations (PVT) in the delay cells used in the Vernier delay line to fix the propagation delay of each VCDC [31].

The histogram builder circuit accumulates the TDC conversion to build a histogram on-chip. As shown in Fig. 11, the first step is an encoder to convert the TDC thermometric output code in one-hot code, followed by a bank of 8-to-1 multiplexers, used to select the range of the one-hot outputs corresponding to the dithering position to correctly implement the dithering. The conversion results are accumulated in 128 12-bit counters to build the histogram, similar to what is presented in [30] and [35]. An additional 24-bit counter is used to count the total number of events detected by the gated SiPM and the data are readout serially via the SPI interface.

C. Internal Signal Generator

The chip features a differential SYNC input (compatible with LVDS or LVPECL logic levels) connected to an internal circuitry where the gate window is generated with position and duration programmable in 24 steps of 1 ns (nominal) each, guaranteeing easy synchronization with the excitation signal. The TDC STOP signal can also be generated with a programmable delay with the same range and resolution. This allows the chip to generate its gate signal internally without the need for external delayers, thus minimizing the complexity of the complete system and allowing it to easily interface to a variety of pulsed laser sources.

The signal generation block (Fig. 12) is structured with a voltage-controlled delay line that receives its control voltage from a DLL included in the TDC circuitry, again to guarantee the stability of the delays. The SYNC signal enters the delay line, either directly or through a selectable monostable, and propagates along the chain of delay cells. The output of the delay line is tapped at each step and its buffered copy propagates to three circuits used to, respectively, set the position of the gate window rising and falling edges, as well as the TDC STOP delay.

The circuits are based on an open-drain line connected to a set of NMOS transistors that can be individually enabled to pull down the line in correspondence to the selected tap of the delay line; the falling edge is then used to generate the required signal. The open-drain line features an active reset to increase the maximum repetition rate, allowing up to 100-MHz SYNC frequency. The opening and closing edges of the gate window are set independently not to constrain gate window duration with respect to the selected delay. Exploiting the high level of integration of this chip, a low-power operation mode that inhibits gate window generation during TDC conversion is also available to optimize power consumption and timing accuracy.

D. Programming and Readout Circuitry

The chip is controlled with a simple four-wire communication interface, compatible with SPI for a direct interface with a simple microcontroller. The activation of each of the 1728 pixels can be programmed independently, with a synchronous update of the configuration to allow on-the-fly glitchless reprogramming during operation, as well as to configure the internal signal generator and TDC and to readout the timing histograms and the 24-bit cumulative event counter; the histogram output is double-buffered to allow concurrent acquisition and readout.

Auxiliary I/O pins are available in case the integrated circuit should be used in a different setup and have been exploited for debugging and characterization of the single components. The most relevant ones are the raw event output, gate input, gate output, START, and STOP of the TDC (all compliant with LVCMOS levels). An on-die unsilicided undoped polysilicon resistor (10-kΩ nominal at 25 °C) with a high negative temperature coefficient can also be used to monitor the chip temperature.

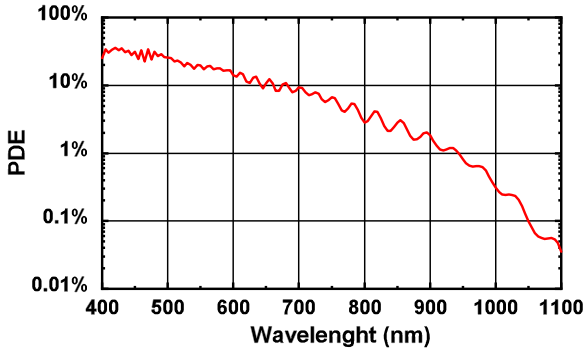

Fig. 13. PDE of the SPADs when operated at the reference 3.3-V excess bias.

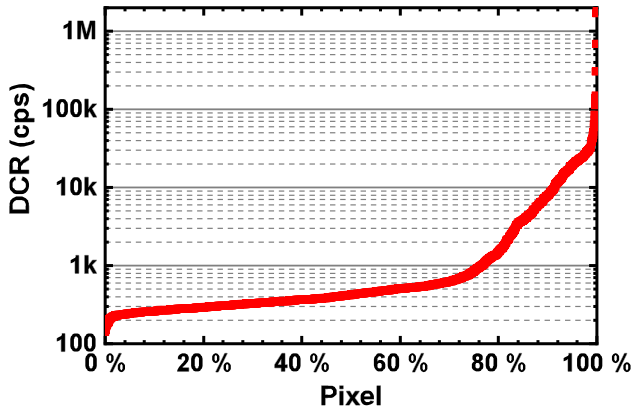

Fig. 14. DCR for each pixel, measured at room temperature and at the reference 3.3-V excess bias, and sorted in the ascending order. The median value is 424 Hz per pixel. A change in slope around the 70% mark shows the appearance of hot pixels, although the cumulative DCR can be kept below 250 kHz/mm^2 for active areas as large as 8 mm^2 at room temperature.

IV. Performance Assessment

We assessed the performance of the fabricated ASIC in terms of PDE, DCR, gating performance, temporal response, spectral responsivity, and time-gated instrument response function (IRF).

A. SPAD Performance

The SPADs fabricated in this 0.35-μm CMOS technology are well-assessed for their remarkably low noise, good temporal resolution [36], and for the fast exponential decay tail, which is of utmost importance for TD-NIRS.

The PDE for a single SPAD when the pixel is operated at a nominal voltage of 3.3 V is reported in Fig. 13. The DCR for each pixel, sorted in ascending order, is shown in Fig. 14. The DCR trend shows a distribution with a median value of 424 Hz per pixel when the chip is operated at room temperature and with a reference excess bias of 3.3 V, with a change in the slope of the curve around the 70% mark, followed by few extremely hot pixels (DCR greater than tens of kilohertz).

B. Integrated TDC Performance

We assessed the performance of the integrated TDC in terms of precision and linearity by exploiting dedicated test inputs. The measured FSR is just less than 10 ns and the average LSB is about 78 ps, with a maximum channel width (i.e., largest time-bin) of ~89 ps with dithering and ~106 ps without dithering.

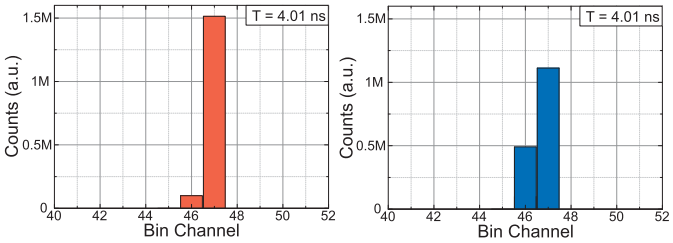

Fig. 15. Example of the TDC output when converting the same delay without dithering enabled (left, red) and with dithering enabled (right, blue).

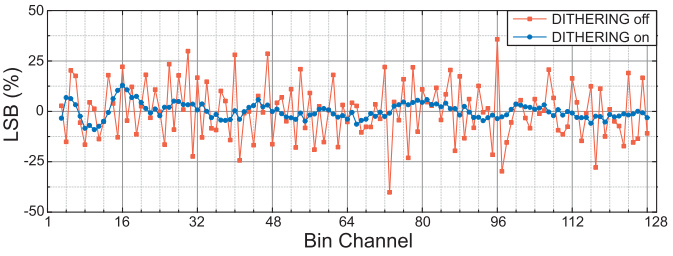

Fig. 16. DNL of the fabricated TDC. The measured rms value of DNL is about 3.9% of LSB with dithering enabled, while it increases to 14% with dithering disabled.

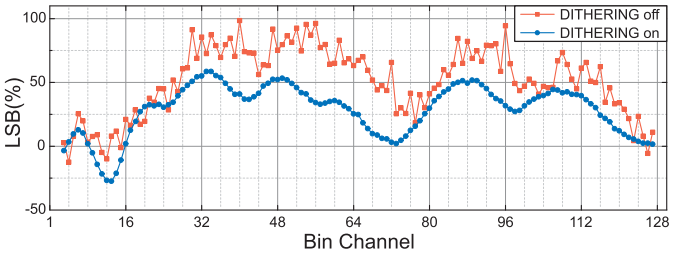

Fig. 17. INL of the fabricated TDC. The measured value is −0.27/+0.58 LSBs with dithering ON and −0.12/+0.98 LSBs with dithering OFF.

We characterized the timing precision of TDC, using external START and STOP signals. We measured the full-width at -maximum (FWHM) in the whole TDC range, sweeping externally the delay between START and STOP signals. The FWHM ranges between 78 ps (a single histogram bin) and a maximum value of 156 ps (2 time-bins) due to the quantization error, when the signal is across two channels. The measured mean value of FWHM is 97.6 ps with dithering disabled, increasing to 107 ps with dithering enabled (Fig. 15).

The linearity of the TDC has been measured, using uncorrelated START and STOP signals and building the histogram after many repetitions. The differential nonlinearity (DNL) of the TDC is shown in Fig. 16. With an average of 20-k events per channel (time-bin), the measured root-mean-square (rms) value of DNL is about 3.9% of LSB with dithering enabled, while increases to 14% with dithering disabled. The integral non-linearity (INL) of the TDC is reported in Fig. 17, and is within −0.27/+0.58 LSBs with dithering enabled and increased to −0.12/+0.98 LSBs without dithering. Dithering significantly improves TDC DNL, while it is less effective for INL due to limited dithering range.

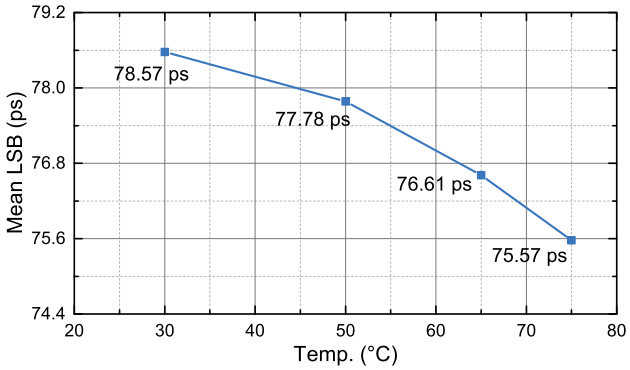

Fig. 18. Measured TDC average bin width (LSB) versus temperature.

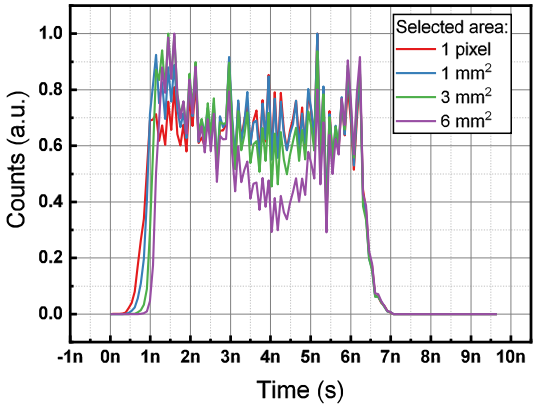

Fig. 19. Normalized count distribution with a 5-ns gate window as acquired by the internal TDC when different active areas are enabled.

The effectiveness of the DLL to stabilize the average LSB against temperature variations was measured by varying the chip temperature in a temperature-controlled environment, and the results are shown in Fig. 18. The variation in the average LSB remains within 3 ps within the expected chip operating in the temperature range of 30 °C–75 °C, showing the effectiveness of the approach.

Unfortunately, due to a race condition between the internal reset circuit and the dithering logic, sporadically the TDC input logic is not correctly re-armed after the end of conversion when dithering is enabled, causing the TDC to stop converting until a reset command is issued. The chip can still be used correctly when dithering is disabled and the deterministic non-linearity pattern can be corrected on the acquired data.

The SPI interface can operate up to 50 MHz if fast histogram readout is required, resulting in a maximum histogram transfer rate of 30 kHz (195 bytes per histogram at 50 MHz) with no dead-time between subsequent histograms, but typical operating conditions are an SPI frequency of 8 MHz and histogram transfer rate <1 kHz.

C. Time-Gated dSiPM Performance

The uniformity of the gate window has been characterized for different numbers of active pixels, with both the internal TDC (Fig. 19) and an external high temporal resolution TCSPC board (SPC-130, Becker & Hickl GmbH, Berlin, Germany) (Fig. 20), exploiting an additional output pad

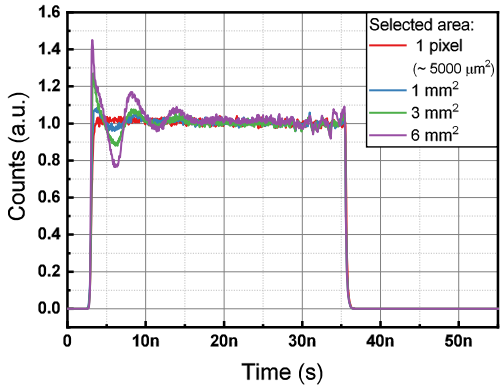

Fig. 20. Normalized count distribution with a long gate window as acquired by Becker & Hickl SPC-130. A significant ringing can be noticed during the first ~10 ns of the gate window, increasing with the active area as a result of the higher peak currents through the power supply connections.

exposing the output of the dSiPM. For all the measurements presented in this article, we kept the incoming light flux so that the overall SiPM count rate was limited to less than 5% of the laser repetition rate to have no distortion on the reconstructed waveform [37]. The bin-to-bin fluctuations visible in Fig. 19 are mostly due to TDC non-linearity since dithering is disabled (see TDC DNL in Fig. 16), which is the operating condition for the measurement. However, this non-linearity is a deterministic distortion, which can be compensated in post-processing. In Fig. 20, the single-pixel response has very fast transitions (300-ps rise time from 20% to 80%) and excellent uniformity. As the active area increases (1 and 3 mm^2), some oscillations start to become noticeable and become very evident for very large areas (see the 6-mm^2 curve). This behavior can be explained considering the parasitic inductance of supply voltage connections: due to constraints set by the compact optode, only few pads have been placed for the cathode voltage supply, which is affected by strong capacitive feedthrough due to the gating operation. This effect is also visible in the rising edge (20%–80%) of the gate window, which changes from 313 ps for the single pixel to 141 ps when 6 mm^2 are enabled, which can again be explained by the bond wire inductance causing a voltage sag at the gate window opening, followed by an overshoot which speeds up the transition.

We characterized the temporal response of the digital SiPM by shining a pulsed diode laser (at 780 nm) with ~50 ps (FWHM) pulse duration and a repetition rate of 40 MHz onto the detector at various positions within the gate window (Fig. 21). The single pixel shows a temporal response with an FWHM of 235 ps, uniform within the gate window, which is in line with what was expected given the low excess bias (3.3 V) and the relatively high threshold for sensing the avalanche due to the low-power fast-gating front-end. The temporal response degrades for larger active areas, reaching a value of ~300 ps due to the different propagation delays of pixels in different positions within the detector.

In addition, for large active areas, we can notice a non-uniform behavior within the gate window, which is related to the non-uniform sensitivity within the gate window (see Figs. 19 and 20).

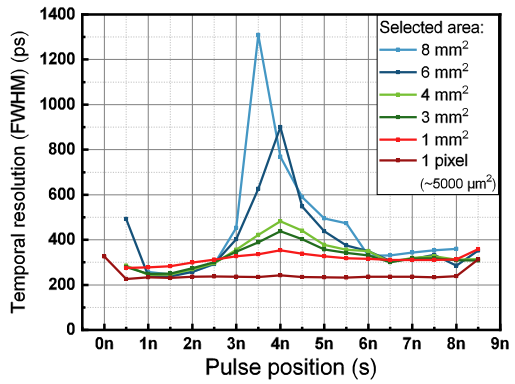

Fig. 21. FWHM of the temporal response versus pulse position inside the gate window, obtained with a pulsed diode laser (at 780 nm) with ∼50-ps (FWHM) pulse duration and a repetition rate of 40 MHz, measured with a Becker & Hickl SPC-130.

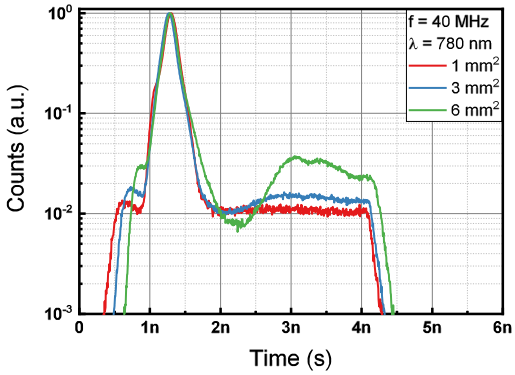

Fig. 22. Temporal response of the digital SiPM illuminated with a ∼50-ps FWHM 780-nm laser diode and measured with a Becker & Hickl SPC-130, for various active areas.

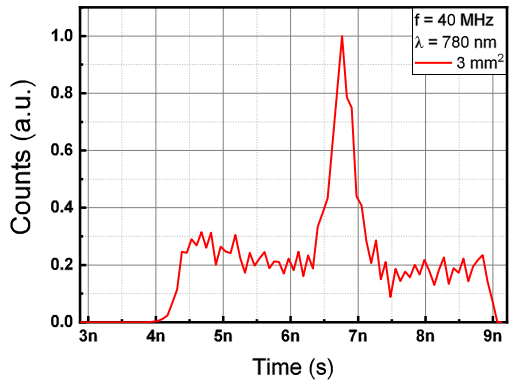

Fig. 23. Histogram as acquired by the on-chip TDC with 5-ns gate window.

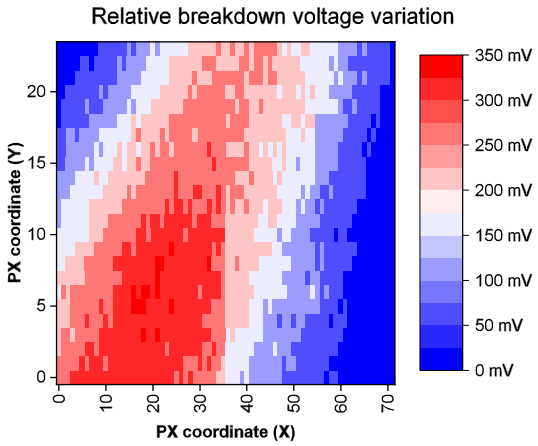

Fig. 24. Map showing the distribution of the breakdown voltage of each pixel in the array. The discontinuity visible after column 35 is due to a gap between quadrants along the vertical direction on the chip.

Indeed, due to the ringing of the power supply lines in response to the high peak current required for the gating operation, quick fluctuations of the instantaneous reverse voltage applied to the SPADs are noticeable and are the source of the degraded resolution after ∼3.5 ns, as shown in Fig. 21.

Indeed, the modulation of the excess bias causes different build-up times of the avalanche current, which affect both the jitter and the photon-to-output delay, both increasing during the undershoot (which is strongest ∼3.5 ns after the gate window opening), with negative effects on the overall temporal response of the detector. However, the first ∼3 ns of the gate window show a good temporal resolution even for large active areas and this makes it possible to effectively exploit this chip for time-gated diffuse optics, where most of the information lies in the first few nanoseconds after the laser pulse, which will be just before the gate rising edge.

An example of the SiPM's temporal response to a laser pulse in the first part of the gate window is reported in Fig. 22. As the active area increases, the disturbance in the gate uniformity becomes more visible as a non-uniform noise floor toward the middle of the gate window, which is due to the power supply oscillations (as shown in Fig. 20). However, the temporal response in the first part of the gate window is not significantly affected.

Finally, an example of the histogram as obtained with the internal TDC when the chip is operating in realistic conditions (3 mm^2 of the active area at 40-MHz repetition rate) and flood illuminated with a 780-nm pulsed laser (∼-50 ps FWHM) is reported in Fig. 23. The resulting temporal resolution is 366 ps (FWHM), which agrees with the values reported in Fig. 21 for the given active area and pulse position.

All the SPAD arrays from this fabrication run suffered from a uniformity issue, whose main impact is a non-uniform SPAD breakdown voltage across the detector chip. We characterized the breakdown voltage of each pixel by selectively enabling each pixel and acquiring its count rate when placed under faint constant illumination, at different cathode biases. By doing so, we were able to classify the breakdown voltage of each pixel with a 50-mV resolution. Fig. 24 shows the distribution of the breakdown voltage across the active area of a typical sample: a gradient is clearly visible, with a peak-to-peak variation exceeding 350 mV, that is, more than 10% of the maximum excess bias that this chip can operate at. Unfortunately, the different breakdown voltages result in a different temporal response of the pixels in the array when biased at the same voltage, which contributes to the performance worsening with a large active area. As a consequence, the reference excess bias of 3.3 V is indirectly set as the highest bias voltage that

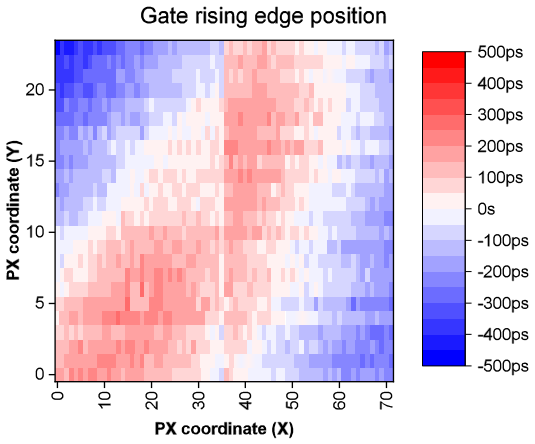

Fig. 25. Relative temporal position of the rising edge of the gate window for each pixel in the array, as obtained from the timing histograms of each pixel illuminated with weak uncorrelated illumination.

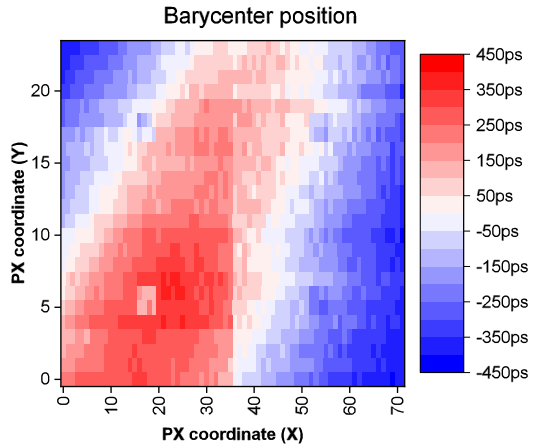

Fig. 26. Temporal shift of the barycenter of the pixels' response to an incoming light pulse.

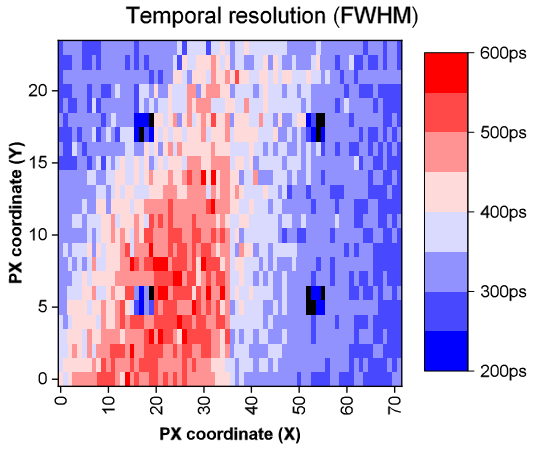

Fig. 27. Temporal resolution of each pixel, quoted as FWHM.

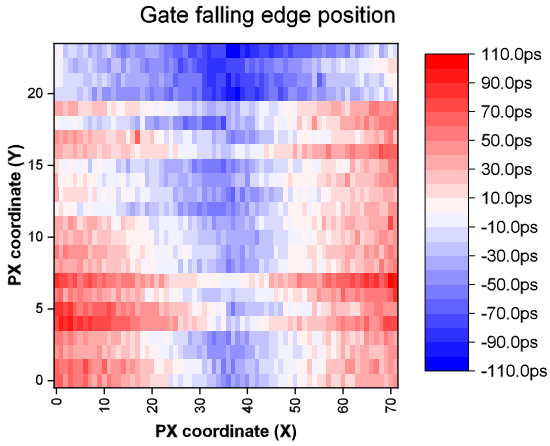

Fig. 28. Map of the temporal position of the gate window falling edge.

leads to correct device operation with proper quenching of each SPAD.

The gate window rising edge position (Fig. 25) is strongly correlated with the breakdown voltage distribution and shows a standard deviation of 137 ps. A similar behavior is also seen with the barycenter of the temporal response of the pixels when illuminated with a pulsed laser (shown in Fig. 26), having a 198-ps standard deviation, and in the temporal resolution, as reported in Fig. 27. All these results are explained with the slower avalanche buildup in pixels with lower excess bias (i.e., higher breakdown voltage) which causes an increase in the timing jitter and delayed detection of the avalanche.

The position of the gate window falling edge shows significantly better uniformity (Fig. 28), with a standard deviation of 37 ps, and includes the skews of the propagation of the gate signal toward each pixel and the skew between each pixel's digital output toward the detector output.

To better quantify the performance of the designed circuitry, decoupled from the SPAD uniformity issue, we performed a second set of measurement by scanning each pixel while adjusting the cathode bias to ensure almost the same excess bias for all the SPADs. In this case, we had to set the excess bias to 2.8 V, so that even when the cathode bias is increased by 350 mV (to bias the SPADs with highest breakdown voltage), the SPADs with lowest breakdown voltage would remain properly gated off. The measured temporal response resulted to be much more uniform, with a standard deviation of the barycenter of the temporal response reduced to 61 ps and a more uniform temporal resolution, as shown in Fig. 29.

We need to consider that since the power supply oscillations increase their effect when more pixels are enabled, the system does not behave linearly, and thus the performance of the final chip is not just the sum of the contributions of each pixel.

To mitigate the SPAD uniformity issue, an automatic procedure to map the breakdown voltage and the DCR of each pixel has been devised, so that both the parameters can be considered when choosing which SPADs to operate to achieve a given sensitivity.

D. Chip Performance in TD-NIRS Applications

A complete validation of any TD-NIRS system requires a thorough set of measurements, according to multiple conventionally agreed protocols (see [38]–[40]) which extend beyond the scope of this journal.

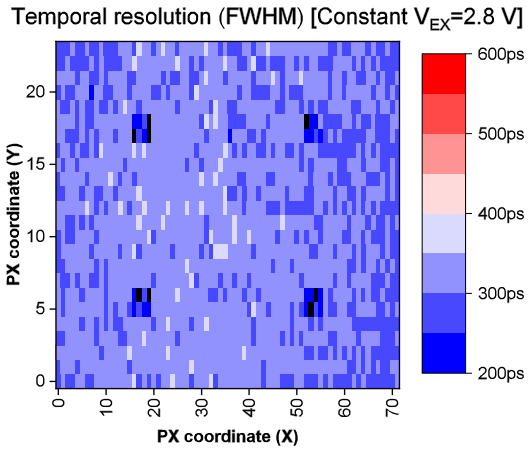

Fig. 29. Temporal resolution of each pixel obtained adjusting the cathode voltage bias for each pixel to compensate for the breakdown voltage variation. The much better uniformity helps quantify the performance of the designed electronics decoupled from SPAD fabrication issues.

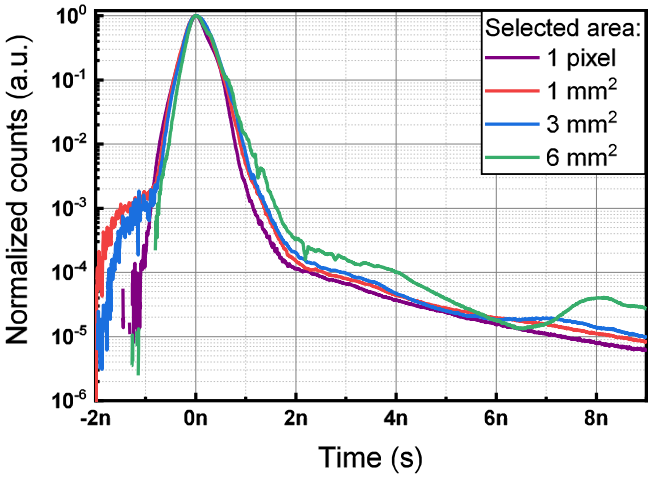

Fig. 30. Reconstructed IRFs acquired with an external TCSPC board for different sensitive areas of the detector (ranging from a single pixel to 6 mm^2).

As a preliminary check of the detector suitability for the targeted application, we focused on the IRF, which has a key role in demonstrating the suitability of the detector in TD-NIRS measurements [38], [41]. We acquired the IRF obtained by enabling a different number of pixels, corresponding to an active area of 5000 μm^2 (single pixel), 1, 3, and 6 mm^2.

The measurements were recorded using a pulsed laser at 690 nm. Light exiting from the laser fiber was attenuated using a VOA and then fiber-coupled again. The tip of the injection fiber was kept few centimeters far from the detector so as to illuminate the whole active area. A thin layer of Teflon was put on the fiber tip, thus allowing a homogeneous distribution of the light on the detector. Following the procedure described in [42], each portion of the TD curve (corresponding to a given gate delay) was acquired using a constant count rate (of about 1 MHz), and thus the laser power impinging onto the detector was adjusted by the VOA. The temporal distribution of the photons was recorded using an external TCSPC board (SPC-130, Becker & Hickl GmbH).

The synchronization between laser pulses and detector gate was ensured by a pulse generator (set at 40 MHz), which provided the trigger to both the laser and the detector, as well as to the sync input of the TCSPC board. The laser was thus emitting optical pulses at 40 MHz, while the opening of the detector gate could be delayed thanks to an external delayer with 25-ps resolution.

TABLE I

MEASURED POWER CONSUMPTION OF THE CHIP: THERE IS AN ALMOST LINEAR TREND OF POWER CONSUMPTION WITH RESPECT TO ACTIVE AREA AND GATE FREQUENCY

Detector active area	Gate frequency	Gating power	TDC power[a]	HV supply power[a]
1 mm^2	40 MHz	100 mW	35 mW	< 5 mW
3 mm^2	40 MHz	300 mW	35 mW	< 5 mW
6 mm^2	40 MHz	630 mW	35 mW	< 5 mW
1 mm^2	80 MHz	230 mW	40 mW	< 5 mW
3 mm^2	80 MHz	670 mW	40 mW	< 5 mW
6 mm^2	80 MHz	1300 mW	40 mW	< 5 mW

[a]Power consumption of the TDC and from the high voltage supply depend only on the count rate. Values are reported for a count rate of 5 % of the gate frequency.

To reconstruct the IRF, we acquired 81 delays at steps of 50 ps. For each delay, ten repetitions of 1 s were acquired and then summed in the post-processing to improve the signal-to-noise ratio of the measurements. Due to the gate non-uniformity, background measurements (i.e., without laser) were acquired and their shape was subtracted from the signal curves. The reconstruction of the high-dynamic-range temporal distributions was done following the procedure described in [42].

Fig. 30 reports the normalized reconstructed IRF, obtained using different values of the active area of the detector. The dynamic range is more than five decades and the limit is set by the so-called "memory effect," [43] an afterpulse-like noise arising when the detector in the OFF state is exposed to a strong illumination. The temporal response (quoted as FWHM) is smaller for the single pixel (380 ps), while, as expected, it increases for the larger area values where it is almost constant around 480 ps. Compared with other SiPMs, whose suitability for diffuse optics applications has been widely validated (see [44]–[46]), the dynamic range is nearly two decades larger. On the other hand, the FWHM is wider than the cited works, but, as demonstrated in [41], it is not supposed to compromise the performances of the system in TD-NIRS measurements.

The power consumption of the chip is reported in Table I, showing a linear increase in power consumption with respect to the active area and gating frequency, as expected. The power consumption of the TDC does not depend on the area or the gating frequency, but only on the count rate. The values are compatible with the integration in the compact optode. Keeping in mind the self-heating of the die, which affects the DCR of the device, a low-thermal resistance (<10 K/W) should be used to ensure good detector performance.

V. CONCLUSION

We presented the first implementation of a large-area, time-gated, all-digital single-photon detector with the variable active area, programmable gate signal generator, and on-chip TDC

TABLE II
COMPARISON OF THE ARRAY PRESENTED IN THIS WORK WITH OTHER SPAD ARRAYS WITH TIME-GATING, HISTOGRAM BUILDER, AND LARGE COLLECTION AREA

Ref.	Technology node	Fill Factor	Net SPAD active area Gated [Non gated]	Gating capability[a]	TDC resolution	Histogram builder architecture	Type of sensor
This work	**350 nm**	**37 %**	**8.6 mm²**	**Yes**	**78 ps**	**Counters**	**Digital SiPM**
[14]	350 nm	40 %	0.03 mm² [1 mm² SiPM]	Yes	No TDC	-	For TD-NIRS
[27]	N.A.	78 %	[19 mm² [b]]	No	23 ps	-	Digital SiPM
[30]	130 nm	43 %	[0.2 mm²]	No	105 ps	Counters	Digital SiPM
[35]	130 nm	49.3 %	0.6 mm² [0.4 mm²]	Yes	51.2 ps	Counters	Line sensor
[47]	350 nm	44.3 %	2.1 mm²	Yes	No TDC[c]	-	Line sensor
[48]	350 nm	21 %	0.9 mm²	Yes	No TDC[d]	-	Imager
[49]	180 nm	10.5 %	7.4 mm²	Yes	No TDC[c]	-	Imager

[a] Only detectors capable of actively bringing the SPAD below breakdown are reported as 'Yes'. [b] Per chip. A multi-chip module is presented. [c] This detector provides a 1b output per pixel. [d] This detector has an in-pixel analog photon counter.

with built-in histogram builder circuit, suitable for high-sensitivity TD-NIRS measurements with minimal external components. The chip has been designed for a miniaturized, multi-wavelength TD-NIRS system with an unprecedented level of integration and responsivity. The experimental characterization shows that its performance, mainly in terms of temporal response and gating, even though affected by SPAD non-uniformity, is adequate for TD-NIRS applications at short source–distance separation, showing a dynamic range better than other TD-NIRS systems based on analog SiPMs, and the small size and reduced power consumption make it possible to integrate the chip in an extremely compact photonics module.

Table II lists the comparison of this work with other SPAD arrays presented in the literature with time-gating capability, histogram generation, and/or large collection area. In Table II, we focus on the net SPAD active area offered by the chip, because although the addition of microlenses to SPAD arrays has demonstrated a considerable increase in the equivalent fill-factor [49], a TD-NIRS detector collects light from a near-Lambertian angular distribution, hindering the achievable concentration factor and overall making this solution ineffective. The work we present, to the best of our knowledge, is the first to combine a large (>1 mm^2) time-gated SPAD area together with TDC and histogram builder on the same chip, allowing it to be a complete single-chip detection channel for TD-NIRS.

REFERENCES

[1] A. Pifferi, D. Contini, A. D. Mora, A. Farina, L. Spinelli, and A. Torricelli, "New frontiers in time-domain diffuse optics, a review," *J. Biomed. Opt.*, vol. 21, no. 9, Jun. 2016, Art. no. 091310.

[2] T. Durduran, R. Choe, W. B. Baker, and A. G. Yodh, "Diffuse optics for tissue monitoring and tomography," *Rep. Prog. Phys.*, vol. 73, no. 7, Jul. 2010, Art. no. 076701.

[3] A. Torricelli *et al.*, "Time domain functional NIRS imaging for human brain mapping," *NeuroImage*, vol. 85, pp. 28–50, Jan. 2014.

[4] R. Cubeddu, A. Pifferi, P. Taroni, A. Torricelli, and G. Valentini, "Noninvasive absorption and scattering spectroscopy of bulk diffusive media: An application to the optical characterization of human breast," *Appl. Phys. Lett.*, vol. 74, no. 6, pp. 874–876, Feb. 1999.

[5] A. Torricelli, "Recent advances in time-resolved NIR spectroscopy for nondestructive assessment of fruit quality," *Chem. Eng. Trans.*, vol. 44, pp. 43–48, Sep. 2015.

[6] M. Renna *et al.*, "Eight-wavelength, dual detection channel instrument for near-infrared time-resolved diffuse optical spectroscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 1, pp. 1–11, Jan. 2019.

[7] A. Torricelli *et al.*, "Time-resolved reflectance at null source-detector separation: Improving contrast and resolution in diffuse optical imaging," *Phys. Rev. Lett.*, vol. 95, no. 7, Aug. 2005, Art. no. 078101.

[8] A. Pifferi *et al.*, "Time-resolved diffuse reflectance using small source-detector separation and fast single-photon gating," *Phys. Rev. Lett.*, vol. 100, no. 13, Mar. 2008, Art. no. 138101.

[9] A. Dalla Mora *et al.*, "Fast-gated single-photon avalanche diode for wide dynamic range near infrared spectroscopy," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 4, pp. 1023–1030, Jul. 2010.

[10] L. Di Sieno *et al.*, "Miniaturized pulsed laser source for time-domain diffuse optics routes to wearable devices," *J. Biomed. Opt.*, vol. 22, no. 08, p. 1, Aug. 2017.

[11] S. Saha, Y. Lu, F. Lesage, and M. Sawan, "Wearable SiPM-based NIRS interface integrated with pulsed laser source," *IEEE Trans. Biomed. Circuits Syst.*, vol. 13, no. 6, pp. 1313–1323, Dec. 2019, doi: 10.1109/TBCAS.2019.2951539.

[12] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, pp. 1956–1976, 1996.

[13] A. D. Mora *et al.*, "Towards next-generation time-domain diffuse optics for extreme depth penetration and sensitivity," *Biomed. Opt. Express*, vol. 6, no. 5, p. 1749, 2015.

[14] S. Saha, S. Burri, C. Bruschini, E. Charbon, F. Lesage, and M. Sawan, "Time domain NIRS optode based on Null/Small source-detector distance for wearable applications," in *Proc. IEEE Custom Integr. Circuits Conf. (CICC)*, Apr. 2019, pp. 1–8.

[15] *SOLUS*. Accessed: Oct. 5, 2019. [Online]. Available: http://www.solus-project.eu/

[16] G. Boso, A. Dalla Mora, A. Della Frera, and A. Tosi, "Fast-gating of single-photon avalanche diodes with 200ps transitions and 30ps timing jitter," *Sens. Actuators A, Phys.*, vol. 191, pp. 61–67, Mar. 2013.

[17] R. K. Henderson *et al.*, "A 192×128 time correlated SPAD image sensor in 40-nm CMOS technology," *IEEE J. Solid-State Circuits*, vol. 54, no. 7, pp. 1907–1916, Jul. 2019.

[18] C. Zhang, S. Lindner, I. M. Antolovic, J. Mata Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252 × 144 SPAD flash LiDAR with 1728 dual-clock 48.8-ps TDCs, and pixel-wise integrated histogramming," *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1137–1151, Apr. 2019.

[19] L. Gasparini *et al.*, "A 32×32-pixel time-resolved single-photon image sensor with 44.64μm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800 kHz observation rate for quantum physics applications," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2018, pp. 98–100.

[20] H. Ruokamo, H. Rapakko, and J. Kostamovaara, "An 80×25 pixel CMOS single-photon image sensor with sub-ns time gating for solid state 3D scanning," in *Proc. 13th Conf. Ph.D. Res. Microelectron. Electron. (PRIME)*, Jun. 2017, pp. 365–368.

[21] F. Villa *et al.*, "SPAD smart pixel for Time-of-Flight and time-correlated single-photon counting measurements," *IEEE Photon. J.*, vol. 4, no. 3, pp. 795–804, Jun. 2012.

[22] S. Burri, F. Powolny, C. Bruschini, X. Michalet, F. Regazzoni, and E. Charbon, "A 65k pixel, 150k frames-per-second camera with global gating and micro-lenses suitable for fluorescence lifetime imaging," *Proc. SPIE Opt. Sens. Detection*, vol. 9141, May 2014, Art. no. 914109.

[23] N. A. W. Dutton *et al.*, "A SPAD-based QVGA image sensor for single-photon counting and quanta imaging," *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 189–196, Jan. 2016.
[24] M. Buttafava, G. Boso, A. Ruggeri, A. Dalla Mora, and A. Tosi, "Time-gated single-photon detection module with 110 ps transition time and up to 80 MHz repetition rate," *Rev. Sci. Instrum.*, vol. 85, no. 8, Aug. 2014, Art. no. 083114.
[25] A. Ruggeri, P. Ciccarella, F. Villa, F. Zappa, and A. Tosi, "Integrated circuit for subnanosecond gating of InGaAs/InP SPAD," *IEEE J. Quantum Electron.*, vol. 51, no. 7, Jul. 2015, Art. no. 4500107.
[26] A. Tomita and K. Nakamura, "Balanced, gated-mode photon detector for quantum-bit discrimination at 1550 nm," *Opt. Lett.*, vol. 27, no. 20, p. 1827, 2002.
[27] Y. Haemisch, T. Frach, C. Degenhardt, and A. Thon, "Fully digital arrays of silicon photomultipliers (dSiPM)—A scalable alternative to vacuum photomultiplier tubes (PMT)," *Phys. Procedia*, vol. 37, pp. 1546–1560, Oct. 2012.
[28] L. Pancheri and D. Stoppa, "A SPAD-based pixel linear array for high-speed time-gated fluorescence lifetime imaging," in *Proc. ESSCIRC*, Athens, Greece, 2009, pp. 428–431.
[29] J. Zhu, R. K. Nandwana, G. Shu, A. Elkholy, S. J. Kim, and P. K. Hanumolu, "A 0.0021 mm21.82 mW 2.2 GHz PLL using time-based integral control in 65 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 8–20, Jan. 2017.
[30] T. Al Abbas, N. A. W. Dutton, O. Almer, N. Finlayson, F. M. D. Rocca, and R. Henderson, "A CMOS SPAD sensor with a multi-event folded flash Time-to-Digital converter for ultra-fast optical transient capture," *IEEE Sensors J.*, vol. 18, no. 8, pp. 3163–3173, Apr. 2018.
[31] T. E. Rahkonen and J. T. Kostamovaara, "The use of stabilized CMOS delay lines for the digitization of short time intervals," *IEEE J. Solid-State Circuits*, vol. 28, no. 8, pp. 887–894, Aug. 1993.
[32] P. Dudek, S. Szczepanski, and J. V. Hatfield, "A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line," *IEEE J. Solid-State Circuits*, vol. 35, no. 2, pp. 240–247, Feb. 2000.
[33] R. Sumner, "A sliding scale method to reduce the differential non linearity of a time digitizer," in *Proc. IEEE Nucl. Sci. Symp. Conf. Rec.*, Nov. 2014, pp. 803–806.
[34] B. Markovic, S. Tisa, F. A. Villa, A. Tosi, and F. Zappa, "A high-linearity, 17 ps precision Time-to-Digital converter based on a single-stage Vernier delay loop fine interpolation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 3, pp. 557–569, Mar. 2013.
[35] A. T. Erdogan *et al.*, "A CMOS SPAD line sensor with per-pixel histogramming TDC for time-resolved multispectral imaging," *IEEE J. Solid-State Circuits*, vol. 54, no. 6, pp. 1705–1719, Jun. 2019.
[36] D. Bronzi *et al.*, "Low-noise and large-area CMOS SPADs with timing response free from slow tails," in *Proc. Eur. Solid-State Device Res. Conf. (ESSDERC)*, Sep. 2012, pp. 230–233.
[37] D. V. O'Connor and D. Phillips, *Time-Correlated Single Photon Counting*. London, U.K.: Academic, 1984.
[38] H. Wabnitz *et al.*, "Performance assessment of time-domain optical brain imagers, part 1: Basic instrumental performance protocol," *J. Biomed. Opt.*, vol. 19, no. 8, Aug. 2014, Art. no. 086010.
[39] A. Pifferi *et al.*, "Performance assessment of photon migration instruments: The MEDPHOT protocol," *Appl. Opt.*, vol. 44, no. 11, p. 2104, Apr. 2005.
[40] H. Wabnitz *et al.*, "Performance assessment of time-domain optical brain imagers, part 2: NEUROPt protocol," *J. Biomed. Opt.*, vol. 19, no. 8, Aug. 2014, Art. no. 086012.
[41] A. Behera, L. Di Sieno, A. Pifferi, F. Martelli, and A. D. Mora, "Instrumental, optical and geometrical parameters affecting time-gated diffuse optical measurements: A systematic study," *Biomed. Opt. Express*, vol. 9, no. 11, p. 5524, 2018.
[42] A. Tosi *et al.*, "Fast-gated single-photon counting technique widens dynamic range and speeds up acquisition time in time-resolved measurements," *Opt. Express*, vol. 19, no. 11, p. 10735, 2011.
[43] A. Dalla Mora *et al.*, "Memory effect in silicon time-gated single-photon avalanche diodes," *J. Appl. Phys.*, vol. 117, no. 11, Mar. 2015, Art. no. 114501.
[44] R. Re, E. Martinenghi, A. D. Mora, D. Contini, A. Pifferi, and A. Torricelli, "Probe-hosted silicon photomultipliers for time-domain functional near-infrared spectroscopy: Phantom andin vivotests," *Neurophotonics*, vol. 3, no. 4, Oct. 2016, Art. no. 045004.
[45] L. Di Sieno *et al.*, "Time-domain diffuse optical tomography using silicon photomultipliers: Feasibility study," *J. Biomed. Opt.*, vol. 21, no. 11, Nov. 2016, Art. no. 116002.
[46] A. Farina *et al.*, "Time-domain functional diffuse optical tomography system based on fiber-free silicon photomultipliers," *Appl. Sci.*, vol. 7, no. 12, p. 1235, Nov. 2017.
[47] Y. Maruyama, J. Blacksberg, and E. Charbon, "A 1024×8, 700-ps time-gated SPAD line sensor for planetary surface exploration with laser Raman spectroscopy and LIBS," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 179–189, Oct. 2014.
[48] M. Perenzoni, N. Massari, D. Perenzoni, L. Gasparini, and D. Stoppa, "11.3 a 160×120-pixel analog-counting single-photon imager with sub-ns time-gating and self-referenced column-parallel A/D conversion for fluorescence lifetime imaging," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2015, pp. 1–3.
[49] A. C. Ulku *et al.*, "A 512×512 SPAD Image Sensor With Integrated Gating for Widefield FLIM," *IEEE J. Sel. Topics Quantum Electron.*, vol. 25, no. 1, Jan./Feb. 2019, Art. no. 6801212.

Enrico Conca (Member, IEEE) was born in Cremona, Italy, in 1992. He received the M.Sc. degree (*cum laude*) in electronics engineering and the Ph.D. degree in information technology from the Politecnico di Milano, Milan, Italy, in 2016 and 2019, respectively.

He is currently a Post-Doctoral Researcher with the Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano. His research activity focuses on the design and development of time-gated single-photon counting CMOS circuits.

Vincenzo Sesta was born in Ribera, Italy, in 1991. He received the M.Sc. degree in electronics engineering and the Ph.D. degree in information technology from the Politecnico di Milano, Milan, Italy, in 2016 and 2019, respectively.

He is currently a Post-Doctoral researcher with the Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano. His research activity focuses on the design and development of time-to-digital converters and timing electronics CMOS circuits for arrays of silicon single-photon avalanche diodes (SPADs).

Mauro Buttafava (Member, IEEE) received the M.Sc. degree in electronics engineering and the Ph.D. degree (*cum laude*) in information technology from the Politecnico di Milano, Milan, Italy, in 2013 and 2017, respectively.

He is currently a Post-Doctoral Researcher with the Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano. His research activity mainly focuses on the design and characterization of time-resolved, gated-mode single-photon counting systems. His expertise covers different applications in research, industrial, and biomedical fields (like optical spectroscopy, ultrafast time-of-flight imaging, and fluorescence microscopy systems design).

Federica Villa (Member, IEEE) was born in Milan, Italy, in 1986. She received the B.Sc. degree in biomedical engineering, the M.Sc. degree (*summa cum laude*) in electronic engineering, and the Ph.D. degree in information and communication technology from the Politecnico di Milano, Milan, in 2008, 2010, and 2014, respectively.

She has coauthored more than 100 articles. Her present research interests include the design and development of CMOS single-photon avalanche diode (SPAD) imagers for 2-D imaging via single-photon counting and 3-D ranging through direct time-of-flight photon-timing.

Laura Di Sieno was born in Varese, Italy, in 1987. She received the master's degree in electronics engineering and the Ph.D. degree in physics from the Politecnico di Milano, Milan, Italy, in 2011 and 2015, respectively.

She is currently a Research Fellow with the Department of Physics, Politecnico di Milano. Her activity is mainly focused on the study and application of a new approach and instrumentation for time-domain optical spectroscopy of highly scattering media using single-photon detectors.

Alberto Dalla Mora was born in Fiorenzuola d'Arda, Italy, in 1981. He received the M.Sc. degree (*summa cum laude)* in electronics engineering and the Ph.D. degree (*summa cum laude)* in information and communication technology from the Politecnico di Milano, Milan, Italy, in 2006 and 2010, respectively.

He is currently an Associate Professor with the Department of Physics, Politecnico di Milano. His research interests include time-resolved diffuse optics techniques and instrumentation for biomedical applications.

Davide Contini was born in Angera, Italy, in 1978. He received the master's degree in electronic engineering and the Ph.D. degree in physics from the Politecnico di Milano, Milan, Italy, in 2003 and 2007, respectively.

Since 2014, he has been an Associate Professor of physics with the Politecnico di Milano. His current research interests include the interaction of laser light with matter, and in particular the time-resolved spectroscopy of highly diffusive media for applications in biology and medicine.

Paola Taroni was born in Como, Italy, in 1963. She has been a Full Professor of physics with the Politecnico di Milano, Milan, Italy, since 2011. She works in the development and application of optical diagnostic methods, with special consideration for translational aspects. Her major research interest is in the use of diffuse optics for breast cancer diagnosis and management.

Alessandro Torricelli was born in Modena, Italy, in 1968. He received the M.Sc. degree (*summa cum laude)* in electronics engineering from the Politecnico di Milano, Milan, Italy, in 1994, and the Ph.D. degree in physics from the Politecnico di Torino, Turin, Italy, in 1999.

He is currently a Full Professor of physics with the Department of Physics, Politecnico di Milano. His research activity has been focused on radiation–matter interaction, on the development of innovative time-domain techniques based on the pulsed laser for monitoring and imaging biomedical applications.

Antonio Pifferi (Member, IEEE) received the M.S. degree in nuclear engineering and the Ph.D. degree in physics from the Politecnico di Torino, Turin, Italy, in 1991 and 1995, respectively.

He is currently a Full Professor with the Department of Physics, Politecnico di Milano, Milan, Italy. His research is directed toward the development of laser techniques and instrumentation for diagnosis and the study of light propagation in diffusive media, with applications to optical biopsy, optical mammography, and functional brain imaging.

Franco Zappa (Senior Member, IEEE) was born in Milan, Italy, in 1965. He received the master's degree in electronics engineering and the Ph.D. degree in communication technology from the Politecnico di Milano, Milan, in 1989 and 1993, respectively.

In 2004, he co-founded Micro Photon Devices, Bolzano, Italy, a company focused on the production of single-photon detector (SPAD) modules and cameras for single photon-counting and photon-timing. Since 2011, he has been a Full Professor of electronics with the Politecnico di Milano. He has coauthored more than 220 articles, published in peer-reviewed journals and conference proceedings, and eight text books on electronic design, electronic systems, and microcontrollers. His research interests include microelectronic circuitry for SPADs and CMOS and BCD SPAD imagers, for high-sensitivity time-resolved optical measurements, 2-D imaging, and 3-D depth ranging via single-photons' time of flight.

Alberto Tosi (Member, IEEE) was born in Borgomanero, Italy, in 1975. He received the master's degree in electronics engineering and the Ph.D. degree in information technology engineering from the Politecnico di Milano, Milan, Italy, in 2001 and 2005, respectively.

In 2004, he joined the IBM T. J. Watson Research Center, Yorktown Heights, NY, USA, as a Student, where he was working on optical testing of CMOS circuits. He was an Assistant Professor with the Politecnico di Milano from 2006 to 2014. Since 2014, he has been an Associate Professor of electronics with the Politecnico di Milano. He currently works on silicon and InGaAs/InP single-photon avalanche diodes (SPADs). His research activity includes arrays of silicon SPADs for 2-D and 3-D applications and time-correlated single-photon counting electronics.