AIP | Review of Scientific Instruments

High-efficiency integrated readout circuit for single photon avalanche diode arrays in fluorescence lifetime imaging

G. Acconcia, A. Cominelli, I. Rech, and M. Ghioni

Citation: Review of Scientific Instruments **87**, 113110 (2016); doi: 10.1063/1.4968199
View online: http://dx.doi.org/10.1063/1.4968199
View Table of Contents: http://scitation.aip.org/content/aip/journal/rsi/87/11?ver=pdfcov
Published by the AIP Publishing

REVIEW OF SCIENTIFIC INSTRUMENTS **87**, 113110 (2016)

High-efficiency integrated readout circuit for single photon avalanche diode arrays in fluorescence lifetime imaging

G. Acconcia, A. Cominelli, I. Rech, and M. Ghioni
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

(Received 22 July 2016; accepted 8 November 2016; published online 23 November 2016)

In recent years, lifetime measurements by means of the Time Correlated Single Photon Counting (TCSPC) technique have led to a significant breakthrough in medical and biological fields. Unfortunately, the many advantages of TCSPC-based approaches come along with the major drawback of a relatively long acquisition time. The exploitation of multiple channels in parallel could in principle mitigate this issue, and at the same time it opens the way to a multi-parameter analysis of the optical signals, e.g., as a function of wavelength or spatial coordinates. The TCSPC multichannel solutions proposed so far, though, suffer from a tradeoff between number of channels and performance, and the overall measurement speed has not been increased according to the number of channels, thus reducing the advantages of having a multichannel system. In this paper, we present a novel readout architecture for bi-dimensional, high-density Single Photon Avalanche Diode (SPAD) arrays, specifically designed to maximize the throughput of the whole system and able to guarantee an efficient use of resources. The core of the system is a routing logic that can provide a dynamic connection between a large number of SPAD detectors and a much lower number of high-performance acquisition channels. A key feature of our smart router is its ability to guarantee high efficiency under any operating condition. *Published by AIP Publishing.* [http://dx.doi.org/10.1063/1.4968199]

I. INTRODUCTION

Fluorescence Lifetime IMaging (FLIM) is a powerful tool to investigate the complex interactions between biomolecules in many fields:[1,2] a deep understanding of these processes, for example, is fundamental for gaining a better insight into key biological issues like the origin and growth mechanisms of cancer cells. Time Correlated Single Photon Counting (TCSPC) is the technique of choice in FLIM measurement.[3,4] TCSPC basically consists in the repeated excitation of a sample by means of a pulsed laser source and in the measurement of the arrival time of a single photon belonging to the re-emitted fluorescence signal with very high sensitivity, accuracy, and timing resolution.[3] In order to properly reconstruct the light waveform in the time domain from the histogram of the photon arrival times, TCSPC requires that photons remain distinguishable from each other, avoiding the so called pile-up distortion.[3] To this aim, depending on the application the excitation power is typically regulated to obtain an average photon count rate much lower than 10% of the excitation rate: with modern pulsed laser, repetition rates up to 80 MHz are possible, corresponding to a pile-up limited count rate lower than 10 Mcps. The collection of a statistically significant number of events, thus, requires long measurement times, especially when only one detection channel is available, and the acquisition of an image is obtained by means of a scanning point system. The exploitation of many channels in parallel could in principle mitigate this issue, and at the same time it opens the way to a multi-parameter analysis of optical signals adding, for example, spectral or spatial dimensions. In the past decade, several multichannel TCSPC acquisition systems have been reported in literature, and some of them are currently commercially available. Commercial multi-module TCSPC systems, though, rely on the simple parallelization of single conversion chains made of discrete components:[5,6] these systems definitely provide the best-in-class performance especially in terms of timing jitter and Differential Non-Linearity (DNL), but the high power consumption combined with large area occupation has limited the number of available acquisition chains to 4 or 8 channels so far. On the other hand, the exploitation of standard CMOS technologies has been leading to the development of high-density TCSPC acquisition systems able to integrate on the same chip both large arrays of Single Photon Avalanche Diodes (SPADs) and all the electronic circuits necessary to extract the timing information from these sensors.[7–11] Unfortunately, the constraints set by the use of a single technology have prevented the designer from having the necessary degrees of freedom to pursue the best performance. Concerning the detector, the inherent features of CMOS technologies (high doping levels, low thermal budget, thin p- and n-well layers, etc.) conflict with SPAD detector performance especially in terms of Photon Detection Efficiency (PDE), Dark Count Rate (DCR), and afterpulsing.[12,13] To the same extent, the need to design compact and relatively low power timing circuits to be placed in each pixel led to circuits featuring relatively poor performance, especially in terms of DNL.[7,9] Furthermore, the push to increase the number of channels did not lead to a comparable increase in the overall operating frequency thus strongly reducing the advantages of having a multichannel system with respect to simpler solutions like scanning point systems. In most cases, in fact, each pixel can store one digitized photon time of arrival, and after an arbitrary dwell time the whole array is sequentially readout,[7,8,10,14] resulting

into a strong reduction of the maximum achievable operating rate of each pixel. Moreover, in a TCSPC multichannel acquisition system the number of required conversion channels is intrinsically much lower than the number of detectors. As previously stated, indeed, the excitation power has to be tuned in order to guarantee that each detector has an average count rate much lower than 10% of the laser pulse rate. This means that in a hundred-pixel array, we would have less than ten triggered detectors on average per excitation cycle. For this reason, building a system that features an acquisition channel for each SPAD is not only too expensive in terms of area and power dissipation, but it is also unnecessary. In principle, a limited number of acquisition channels can be shared amongst a much larger number of detectors. In this scenario, the multiplexing strategy plays a key role in determining the performance of the acquisition system. In this paper, we present a novel routing algorithm, able to effectively provide a smart connection of a dense array of SPAD detectors to few external, high-performance acquisition channels. The number of output channels can be dimensioned after the sustainable transfer rate towards the PC, and the designed routing mechanism is able to provide a highly efficient exploitation of the resources under any operating condition.

This paper is organized as follows: in Section II the architecture of the overall system is presented; in Sections III and IV the new routing algorithm is described, while the structure of a low-jitter delay line, which is an essential part of the routing electronics, is presented in Section V. In Section VI preliminary experimental results are reported. Conclusions are drawn in Section VII.

II. SYSTEM ARCHITECTURE

The full exploitation of a TCSPC system featuring a dense array of SPAD detectors and a much lower number of timing acquisition chains requires a routing algorithm designed on purpose. The simplest solution consists in partitioning the array into N independent clusters of pixels, each one statically multiplexed to one of the N conversion channels. This approach has been widely exploited with different read-out mechanisms,[11,15–17] although it does not ensure the maximum efficiency of the system in all possible operating conditions. If the illumination is mainly concentrated on some part of the array, for example, the converters associated to the remaining part of the circuit are not used, resulting in a loss of recorded events that could have been avoided.

Moreover, the presence of a hot pixel with higher probability to generate a signal can mask the other pixels in the same group, resulting into a distortion of the overall acquisition.

In order to overcome these issues, we developed a novel routing algorithm: it relies on the selection of a limited number of signals from the whole array of detectors to be routed towards the timing electronics. An arbitration logic is responsible for providing a dynamic connection between the two groups, performing a selection among all the signals at each excitation cycle. In principle, the routing logic should also check if the timing channels are effectively ready to perform the conversion. In fact, most of the currently available timing circuits – either based on a Time to Amplitude Converter (TAC) or on a Time to Digital Converter (TDC) architecture – do typically exhibit a dead time in the order of 100 ns[3,11,18,19] that is well above the excitation period of the laser (typically 12.5 ns). As a result, a single time measurement circuit is not able to start a conversion at each excitation cycle. Nevertheless, we recently demonstrated that the combination of an array of TAC converters and a sequential router can reduce to a negligible value the dead time associated to the electronics.[20]

This architecture, known as Fast TAC (F-TAC), can lead to a significant simplification of the routing logic tasks since the converters can operate at the same rate of the laser.

In order to build a compact, scalable system where each detector can access all the conversion channels, the designed algorithm is based on a digital bus that crosses the whole system, featuring a number of lines equal to the amount of time-measurement circuits. Considering F-TAC-based conversion channels followed by an Analog to Digital Converter (ADC) and at least two bytes to encode the timing information, a remarkable throughput of 6.4 Gb/s can be reached employing only 5 time-measurement circuits. Conversely, it would take 64 converters operating at 4 Mcps to obtain the same result.

The exploitation of F-TACs, then, allows us to substantially reduce the number of interconnections needed to route the signals towards the conversion units, which is one of the major issues in high-density systems. Finally, since interconnection issues lie also in the communication between the arbitration unit and each pixel of the array, the presented routing algorithm relies on a pixelated architecture for the selection logic and on 3D-stacking techniques to connect each SPAD detector to its dedicated circuitry.

III. ROUTING ALGORITHM

A simplified schematic of the routing circuit associated to each pixel is reported in Fig. 1. The circuit provides the pixel with the ability to determine when it may have access to one of the shared output conversion channels (F-TAC 5-1). The selection process starts at the end of each excitation period, as shown in Fig. 2: in this way, all the pixels that have detected a photon in that period concur for the control of the conversion channels; this avoids any discrimination based on the arrival time of the photon within the period, which could result into a distortion of the reconstructed histograms. While the routing algorithm is performed the timing information is preserved by means of a low-jitter delay line.

A. The priority generator

In order to ensure an equal readout probability for each detector, the selection mechanism is based on different routing priorities, consisting in digital words dynamically associated to each pixel at the beginning of every excitation cycle. For instance, considering a 32×32 detector array, a 10-bits

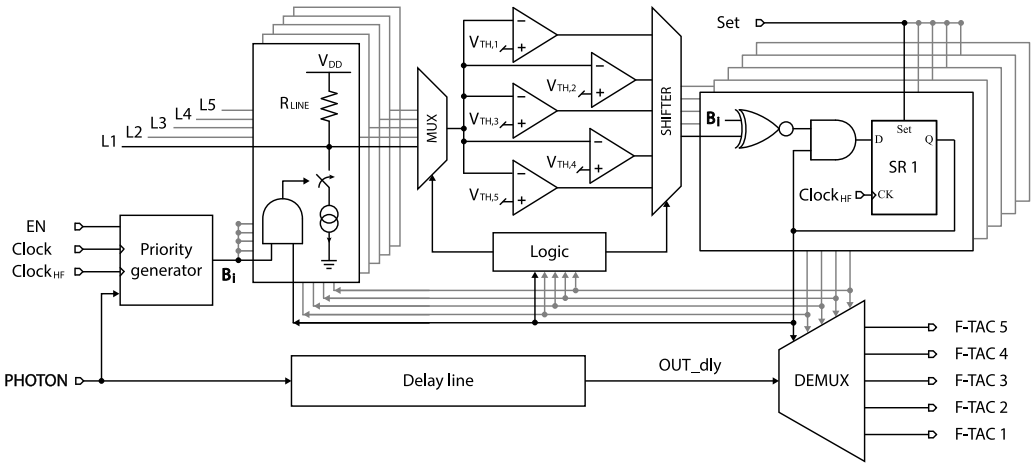

FIG. 1. Routing logic associated to each single pixel, except for RLINE that is shared among all pixels. The circuit determines if the pixel can have access to one of the shared output channels (F-TAC1-5), while a dedicated delay line handles the signal coming from the detector (PHOTON) until the routing process is ended (see text for a more detailed description).

priority code is required in order to have a different priority for each pixel. To this aim, each pixel features a priority generator: it basically consists of a counter that is incremented by a clock synchronous to the excitation source while a startup phase, preceding the measure, provides a different initial condition for each pixel. In particular, the in-pixel counters are enabled in sequence starting from an external enable signal ("EN_ext") as shown in Fig. 3. When a photon in an excitation cycle is detected, a simple logic circuit stores the output of the counter in a shift register, but with a reversed order of bits. In this way, the most significant bit varies at the highest frequency, and no pixel can exhibit a high priority for more than a consecutive period. An example of the priorities evolution for the first three pixels of the array is reported in Table I.

During the selection process, the priority bits (Bi in Fig. 1) are extracted sequentially from the shift register with a dedicated clock ($Clock_{HF}$) and fed to the logic circuit that performs the selection, as will be described later in this section.

B. The comparison lines

The routing is performed on the basis of a bitwise comparison between priorities, which requires a communication among different pixels. To this purpose, we introduced a set of lines shared amongst the whole array. In order to minimize the number of physical interconnections, the priorities are compared considering only one bit at a time, in sync with a dedicated clock signal ($Clock_{HF}$ in Fig. 1) featuring a frequency much higher than the excitation rate, as will be explained later. Each comparison line is able to exploit more than two voltage levels, leading to a significant increase in the amount of carried information with respect to a typical binary encoding. This result is achieved by means of an integrated pull-up resistor for each line (R_{LINE}) and relying on in-pixel current generators to change the voltage drop across the resistor (see Fig. 1). By accurately choosing the number of levels to encode the information, it is possible to perform the comparison with a single line for each conversion channel. When considering five external channels, then, five comparison lines (L5-L1 in Fig. 1) are required, each one exhibiting five significant voltage levels. In particular, each comparison line is used to request the access

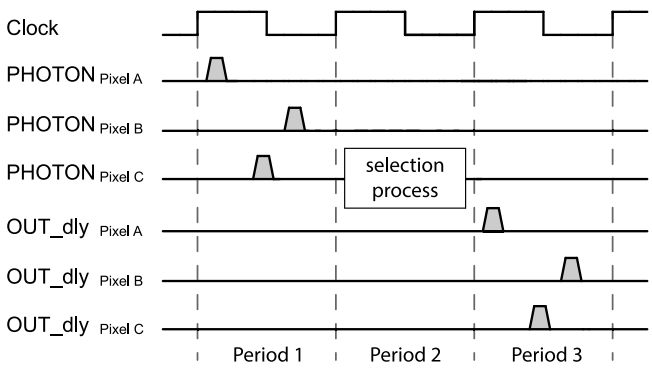

FIG. 2. Timing diagram of the routing process.

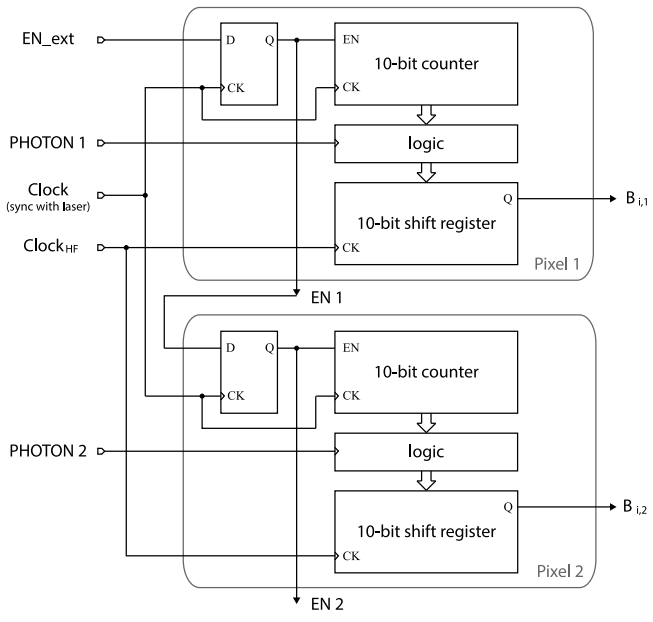

FIG. 3. Schematic view of the priority generator for two consecutive pixels.

TABLE I. Example of priorities evolution for the first three pixels of the array.

	Pixel 1 priority	Pixel 2 priority	Pixel 3 priority
At start-up	0000000000	0000000000	0000000000
Period 1	1000000000	0000000000	0000000000
Period 2	0100000000	1000000000	0000000000
Period 3	1100000000	0100000000	1000000000
Period 4	0010000000	1100000000	0100000000
Period 5	1010000000	0010000000	1100000000
Period 6	0110000000	1010000000	0010000000

to the corresponding conversion channel, e.g., L5 for F-TAC 5. Overall, the routing mechanism requires ten shared lines that cross the whole system: five lines to route the signals towards the time-measurement circuits and five paths (L5-L1) to perform the comparison between priorities.

C. The finite-state machine

The routing logic behaves like a finite-state machine (FSM), synchronized with the high-frequency clock ($Clock_{HF}$). To this aim, each pixel integrates a state register (SR5-1), featuring one bit for each external F-TAC: in particular, the Most Significant Bit (MSB) corresponds to F-TAC 5 while the Least Significant Bit (LSB) to F-TAC 1. As long as a bit in the register is high, the routing towards the associated converter is possible. Therefore, each state represents the possibility to route the signal towards a subset of the available channels. The routing algorithm determines a gradual update of the state registers content until each triggered pixel can be routed towards a single F-TAC at most. The behavior of the FSM is explained in details in the following.

When the first photon in an excitation cycle is detected by a pixel, a high value is stored in each flip flop of its state register (SR5-1 = "11111"). Every pixel that detects a photon during this laser period takes part in the comparison starting from the same state. The comparison is organized in many consecutive steps, considering one priority bit at a time (signal Bi in Fig. 1), starting from the most significant one. During the first step of the comparison each pixel request accesses to the same external converter, i.e., F-TAC 5. This request is done on the corresponding comparison line L5, by connecting to it a current generator if the priority bit of the pixel is high. The same holds for each line that corresponds to a high bit in the state register. An analog multiplexer, then, selects L5, and its voltage level is translated into a thermometric digital code by means of five comparators with different thresholds: when N current generators are connected to the line, the thermometric code will exhibit N "1" (for N higher than five, the thermometric code will be equal to "11111"), meaning that N high-priority pixels are requesting access to the same channel. The output of the comparators is, then, used to update the state registers content: only the N most significant bits of the state registers of pixels featuring a high priority will remain set to "1" while the others are forced to "0," meaning that these pixels will continue to have access to N consecutive comparison lines starting from L5. On the contrary, pixels that exhibit a low priority bit during this step of the comparison, immediately lose the possibility to access to the first N channels, and their state register most significant bits are forced to "0." An example of the evolution of the comparison among six triggered pixels is sketched in Fig. 4, and the relative evolution of the state registers content for each pixel is reported in Table II.

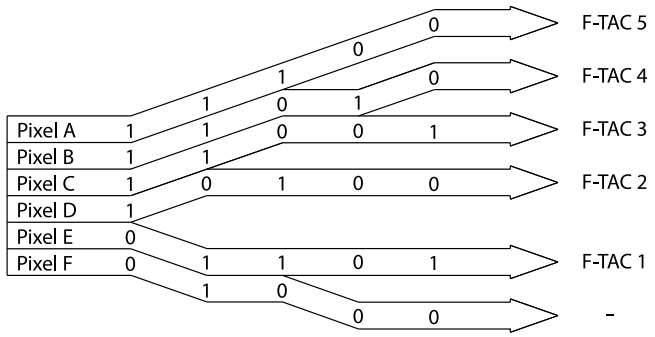

FIG. 4. Bitwise comparison between priorities of six pixels. Five pixels are selected on the basis of their priority, while the remaining signal is lost. In the diagram, the priority bits are reported into arrows: each time two pixels in the same state exhibit a different priority bit, the content of both their state registers is updated, and the pixels follow different paths.

The update of the state register is performed by a dedicated combinatorial logic consisting of a XNOR gate followed by an AND gate for each flip flop of the state register (see Fig. 1). Considering the example in Fig. 4, at Step 1 four pixels exhibit a high priority MSB: each of them will, then, connect its current generator to the comparison line, giving rise to a thermometric code equal to "11110." These high priority pixels will continue the comparison on four lines (L5-L2): the future state of their FSMs is obtained by a bitwise multiplication of their current state with the thermometric code. On the other hand, the low-priority pixels lose the possibility to access the same four lines (L5-L2) and can continue the competition only on L1, so the future state of their FSMs must be equal to "00001." This result is obtained by means of the XNOR and the AND gates that perform a bitwise multiplication of the current state with the complement of the thermometric code:

Current state	11111
Thermometric code	(11110)
Complement of thermometric code	00001
Future state	00001

In general, before the described operation can occur, the thermometric code should be right shifted until its MSB is

TABLE II. Evolution of the state register for each pixel, referring to the example of Fig. 4.

	Step 1	Step 2	Step 3	Step 4	Step 5	Step 6
Pixel A	**11111**	**1111**0	**111**00	**1**0000	**1**0000	**1**0000
Pixel B	**11111**	**1111**0	**111**00	0**1**100	0**1**000	0**1**000
Pixel C	**11111**	**1111**0	**111**00	0**1**100	00**1**00	00**1**00
Pixel D	**11111**	**1111**0	000**1**0	000**1**0	000**1**0	000**1**0
Pixel E	**11111**	0000**1**	0000**1**	0000**1**	0000**1**	0000**1**
Pixel F	**11111**	0000**1**	0000**1**	00000	00000	00000

aligned to the first "1" in the state register. Referring again to the example in Table II, during Step 4 Pixel B and C are competing on L4-L3, which correspond to a FSM state equal to "01100." At Step 5 one of them (Pixel B) exhibits a high-priority bit resulting into a thermometric code equal to "10000." The future state of Pixel B, then, is obtained by the bitwise multiplication of the current state, and the thermometric code right shifted by one position:

Current state	01100
Thermometric code	(10000)
Shifted thermometric code	01000
Future state	01000

To the same extent, the future state of Pixel C is obtained as follows:

Current state	01100
Thermometric code	(10000)
Shifted thermometric code	(01000)
Complement of shifted thermometric code	10111
Future state	00100

D. The pipeline architecture

The time required to carry out the overall comparison depends directly on the number of bits in the priority code. For instance, 10 steps are required when 1024 detectors are present. In order to exploit the maximum working frequency of the F-TACs each comparison must be, in principle, carried out in one laser period, i.e., 12.5 ns. However, this would imply the use of a step duration of only 1.25 ns, corresponding to a high frequency clock ($Clock_{HF}$) equal to 800 MHz. To mitigate this issue, a pipeline-like architecture has been introduced in order to exploit lower frequencies to operate the FSM, at the expense of a higher number of shared multilevel lines. In particular, while a selection phase progresses, a new one can start at each laser period, providing that a different set of five comparison lines is available. In this case the in-pixel logic is slightly more complex: each time a photon is detected, indeed, a free set of lines is selected to perform a comparison that lasts for more than one excitation cycle, depending on the number of pipeline stages.

The logic provides the proper connection of both the in-pixel current generators and the five comparators to the chosen set of lines, and it guarantees that the state register is not reset during the whole duration of the comparison, e.g., due to another photon arrival.

IV. ADDRESS COMMUNICATION

The exploitation of a shared set of time measurement circuits in an imaging system requires that every pixel is identified by an address. In principle, at the end of the selection process the arbitration unit should communicate the address of the routed pixels to the following processing electronics. In order to avoid the introduction of an additional set of interconnections, instead, we developed a decoding algorithm able to extract the position of the routed pixels from the same set of lines used for the comparison. It has been previously highlighted that starting from a given initial condition, the priority codes vary in sync with the excitation frequency, following a known sequence. Therefore, the address of the routed pixels can be deduced from their priorities, once the time elapsed from the beginning of the startup phase is known.

The priority bits can be extracted by following step by step the voltage variations on the comparison lines. At the beginning of the selection process, all the pixels have access to the whole set of comparison lines. In this scenario, three cases are possible, depending on the number of current generators that are connected to the lines. If the voltage drop across the line resistor is zero, then the MSB of all the routed pixels is equal to "0." In fact, only if all the pixels have a low priority bit, no current generator is connected to the line, resulting into a zero voltage variation. On the contrary, if five current generators or more are connected to the lines, then there are five high-priority pixels that will be routed to the output channels. In this case, then, the decoding algorithm extracts "1" as MSB of all the routed pixels. Finally, if the number N of current generators connected to the lines is between zero and five, then the MSB of N routed pixels is equal to "1," while the MSB of the others is equal to "0." In this case, the lines are divided into two subgroups for the next step: one cluster for the high-priority pixels that will concur for the assignment of the first N F-TAC and a group of lines for the remaining low-priority pixels.

Every time a split of a group of lines into two subsets occurs, the decoding algorithm is repeated independently for each group of lines.

During a given step, the decoding algorithm performs two operations on a generic group of M comparison lines: it extracts the priority bit of the pixels that will be routed towards the output channels corresponding to the M lines, and it splits the group into two subsets if necessary. Consider the example in Fig. 5 with M equal to five (Step a), corresponding to the initial configuration of the system until the first separation into subgroups occurs. The comparison lines are identified by a progressive number I (with I ranging from 1 to M) that defines an order inside the group: the pixel with the highest priority in this group will be routed to the F-TAC corresponding to the comparison line with I = 1, the second one in order of priority will be routed to the F-TAC of the line with I = 2, and so on. This can be done since the routing mechanism assigns the lines to the pixels on the basis of their priorities in a descending order. In the example, there are 3 high-priority pixels competing in this group, corresponding to three current generators connected to the lines. Therefore, this group has to be divided into two subsets: a subgroup H_P with 3 lines and a subgroup L_P with the remaining two lines. Starting from the separation of the lines into these two subsets, the extraction of the priority bits at this step is straightforward: the pixels corresponding to subgroup H_P, indeed, feature a "1," the others have a "0." The decoding algorithm determines how many lines are included in subgroup H_P as follows: the first line belongs to subgroup H_P if there is at least one current generator connected to the lines, which corresponds to having one high-priority pixel in the group. To the same extent, the second line belongs to subgroup H_P if there are at least two current generators connected to the lines, and so on. In general,

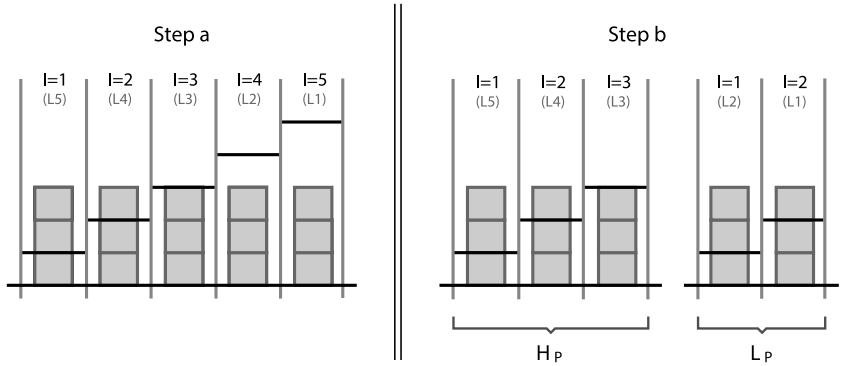

FIG. 5. Example of division of one group of 5 lines into two subsets, HP and LP. The histograms represent the number of current generators connected to the lines. The thresholds exploited by the decoding algorithm are reported for each line.

a line I of the initial group belongs to the subgroup H_P if the number of current generators connected to that line is equal or higher to I, otherwise the line belongs to subgroup L_P.

By using a mechanism based on progressive thresholds from 1 to M on the lines of the group (see Fig. 5), the structure of the two subsets can be derived: in the example, lines 1, 2, and 3 belong to group H_P since their values exceed the thresholds, while lines 4 and 5 belong to group L_P. Once the separation is performed, the decoding algorithm can be reiterated on the two subgroups separately. At the next step, the identification numbers of each line in group L_P are reassigned: lines four and five of the initial group become lines one and two of the subgroup L_P, and the thresholds are scaled accordingly as shown in Fig. 5 (Step b).

In the actual implementation, each line is provided with a set of comparators with different fixed thresholds, and the value to be used in each step of the decoding is selected by means of a multiplexer. The total number of required comparators is equal to 15, ranging from a single comparator for L5 (that always uses the same threshold) to five comparators for L1.

In Fig. 6 the circuitry associated to L3 is reported. The result of the comparison between the value on the line L3 and the selected threshold at a given step is stored in a flip flop (FF_1). This result has to be compared with the result obtained on the adjacent line L2 to find out if a separation into two subgroups is needed: the XNOR gate performs this comparison and stores the result into a flip flop (FF_2): as previously stated, a separation occurs if the comparison result on L3 is high while the result on L2 is low. Finally, FF_2 and the AND gate mask the XNOR output when the two adjacent lines belong to different groups. Based on the results of the separation into groups, a simple combinatorial logic provides the control signal to the multiplexer in order to properly select the threshold to be used at each step. The decoding circuit operates at the same clock frequency of the comparison ($Clock_{HF}$ in Fig. 6), and it sequentially extracts the priority bits of the routed pixels to be fed to the following electronics.

V. THE DELAY LINE

Each time a pixel detects a photon, the access to a conversion channel can be provided only after the end of the selection phase. Therefore, the signal has to be delayed until the arbitration process is ended. To this aim, a delay line is required to be integrated along the signal path, as shown in Fig. 1.

The main challenge in the design of a delay line for this purpose is the minimization of the introduced jitter, since it directly impacts on the overall timing performance of the system. On the other hand, the necessity to integrate a delay line for each pixel poses tight constraints on power dissipation and area occupation, two aspects that conflict with jitter minimization.[21]

We decided to follow an approach based on a cascade of an odd number of inverting stages closed in a ring-oscillator topology, as shown in Fig. 7. The oscillation starts when a photon is detected: this result is achieved by means of a flip flop that stores a high-bit upon this event and a AND gate (AND_1), included in the ring-oscillator loop, that enables the oscillation only when the output of the flip flop is high.

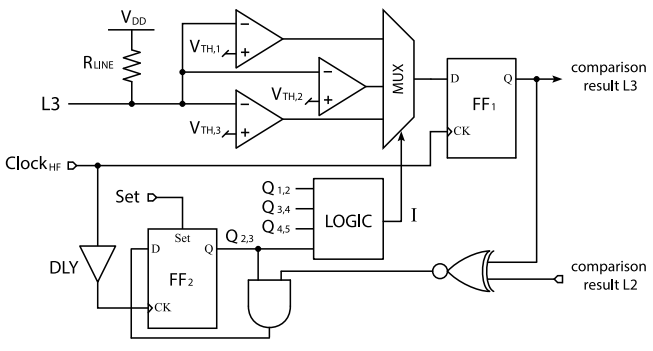

FIG. 6. Decoding circuit associated to L3.

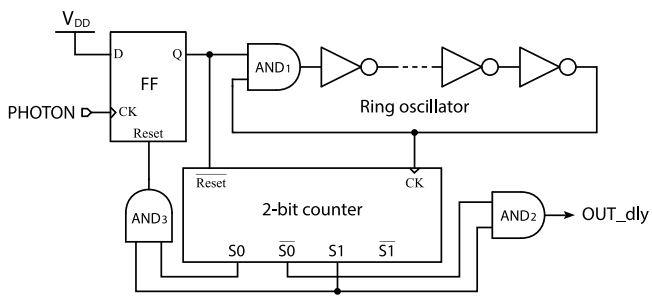

FIG. 7. Structure of the delay line integrated in each pixel.

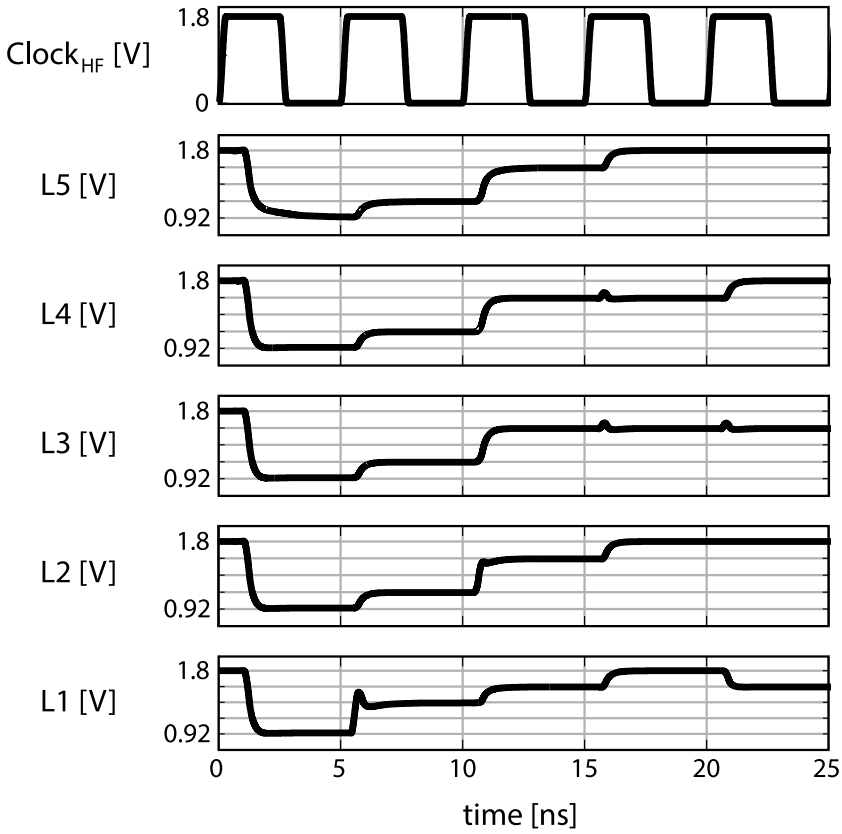

FIG. 8. Simulated evolution of the voltage on the comparison lines during the selection process described in Fig. 4.

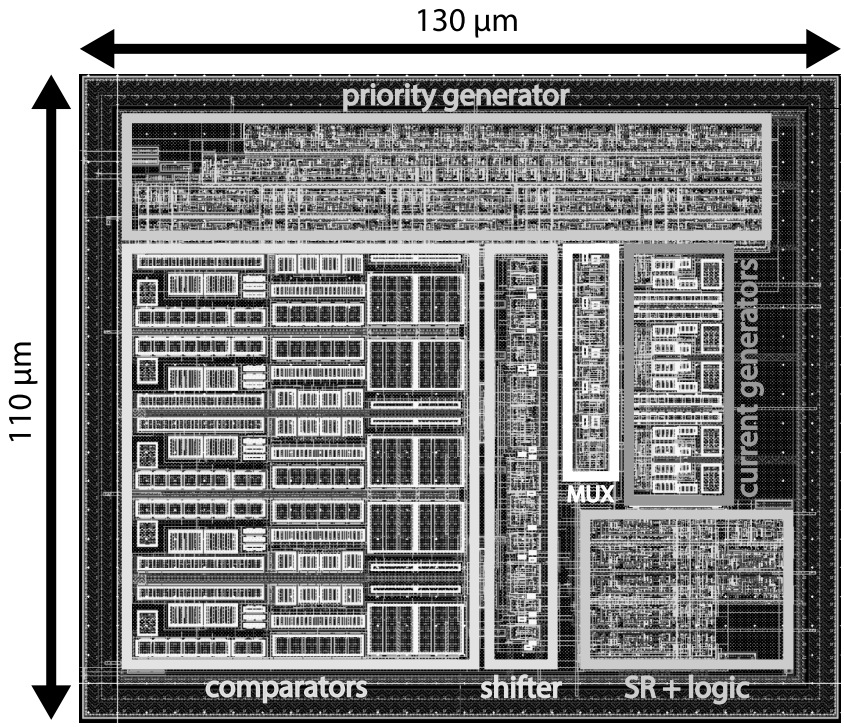

FIG. 9. Layout of a pixel of the routing circuit prototype; the main blocks are highlighted.

The control logic relies on a counter, which increments its output in sync with the ring-oscillator output, while a simple combinatorial circuit defines the length of the delay. For instance, when a delay of only two cycles is needed, a simple AND gate (AND_2 in Fig. 7) is sufficient to control the commutation of the out signal, while another AND gate (AND_3) resets the input flip flop, after a further cycle has passed.

A prototype of delay line has been designed and fabricated exploiting a 0.18 μm high-voltage CMOS technology (Austria Micro Systems AMS H18) in order to evaluate the tradeoff that the described approach imposes. The circuit exploits fully differential stages biased at a constant current in order to achieve both high immunity to disturbances and low noise. The described architecture allowed us to obtain a measured timing jitter as low as 27 ps FWHM, superimposed on a mean delay of 28 ns by exploiting two cycles of a nine-stage ring oscillator. The area occupation is 0.021 mm^2, and the power consumption equal to 2.34 mW.

Simulations on the variability of the introduced delay have been performed, and the results show a variation lower than 1 ns, which has been confirmed by measurements on different samples of the prototype. Furthermore, it is important to note that, in this application, a deterministic variation among the introduced delays of different pixels can be corrected by properly calibrating the system. To this aim, the calibration algorithm reported in Ref. 20 can be exploited. Starting from the results obtained with the prototype, the duration of the introduced delay in future implementation of the circuit could be adjusted acting on the number of stages or on the number of cycles in order to match the duration of the selection process.

VI. EXPERIMENTAL RESULTS

In order to validate the behavior of the routing algorithm, a first integrated prototype has been fabricated in 0.18 μm High Voltage technology (AMS H18). Given five conversion channels, a sufficiently higher number of pixels, namely seven, have been integrated in the chip, each one with the same structure described in Fig. 1. The comparison lines feature a pull-up resistor of 300 Ω, and the in-pixel current generators provide a constant current of 730 μA. Considering the example of Fig. 4, the evolution of the voltage on the comparison lines has been simulated, and it is reported in Fig. 8: as can be seen, the connection of each current generator to one of the lines corresponds to a voltage drop of about 220 mV.

Aiming at the future realization of a 32×32 array, the circuit exploits 10-bits priorities. In this case, a single excitation period is not sufficient to perform a complete comparison, and, in principle, the introduction of a pipeline-like architecture is necessary. Since the introduction of a pipeline mechanism has no impact on the behavior of the algorithm, though, this has not been included in the prototype.

For test purposes, the decoding algorithm has been implemented by means of a Field Programmable Gate Array (FPGA) and 15 external high-speed commercial comparators (TI LMH7324):[22] at this stage, indeed, this approach guarantees a higher flexibility with respect to a complete integration of the system.

The functionality of the circuit has been tested by means of a dedicated Printed Circuit Board (PCB), featuring the 15 comparators and the Field-Programmable Gate Array (FPGA) that controls the operations: the detection of photons is emulated by means of digital pulses generated by the FPGA and provided at the input of the routing logic, while the evolution of the voltage levels on the comparison lines is converted into digital words, by means of the 15 comparators, and fed back to the FPGA. An automatic routine provides for the activation of different pixels at different times: the routing algorithm selects five signals and routes them to the output channels. The decoding algorithm, then, extracts the positions of the routed pixels by looking at the lines and verifies if there is a proper correspondence with the position of the pixels triggered by the FPGA.

The test proved the validity of the routing algorithm, varying the excitation frequency up to 20 MHz. In the described setup, the frequency limitation is mainly due to a relatively high settling time of the comparison lines, since the direct connection with the external discrete comparators introduces a high capacitive load. The future integration of the whole decoding circuit on the same chip of the routing electronics will allow us to mitigate the impact of these capacitive loads on the operating frequency of the system.

The area of a single pixel is 130×110 μm^2, and the layout is reported in Fig. 9. The power consumption of this first prototype is about 8.5 mW/pixel; this value does not depend on the operating frequency since the dynamic contribution is negligible. Future developments will aim at the reduction of the dissipation of each pixel in order to limit the power consumption of a routing circuit with a higher number of pixels.

VII. CONCLUSIONS

In this paper we presented a new readout architecture for dense arrays of SPAD detectors, aimed at breaking the tradeoff between occupied area, dissipated power and performance that currently limits TCSPC acquisition systems. A smart routing entity allows us to dynamically connect a large number of SPAD detectors to a limited number of high performance external conversion channels. Such a circuit is able to guarantee high efficiency in every operating condition, and it opens the way to the exploitation of different technologies in a 3D stacked architecture.

The proposed algorithm exploits five shared conversion channels operating at 80 MHz, able to provide an overall throughput up to 10.4 Gb/s including 2 bytes for the timing information and 10 bits to address the pixels of a 32×32 detector array.

[1]H. Wallrabe and A. Periasamy, Curr. Opin. Biotechnol. **16**, 1 (2005).
[2]K. Suhling, P. M. W. French, and D. Phillips, Photochem. Photobiol. Sci. **4**, 1 (2005).
[3]W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques* (Springer, Berlin, 2005).
[4]W. Becker, J. Microsc. **247**, 2 (2012).
[5]See https://www.picoquant.com/products/category/tcspc-and-time-tagging-modules/hydraharp-400-multichannel-picosecond-event-timer-tcspc-module for PicoQuant, "HidraHarp 400."
[6]See http://www.becker-hickl.de/products.htm for Becker & Hickl GmbH, "TCSPC-Modules."
[7]C. Veerappan, J. Richardson, R. Walker, D. U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, in *Proceedings of IEEE International Solid-State Circuits Conference* (IEEE, 2011), pp. 312–313.
[8]D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, and W. Brockherde, IEEE J. Sel. Top. Quantum Electron. **20**, 6 (2014).

[9]R. M. Field, S. Realov, and K. L. Shepard, IEEE J. Solid-State Circuits **49**, 4 (2014).
[10]L. Parmesan, N. A. Dutton, N. J. Calder, N. Krstajic, A. J. Holmes, L. A. Grant, and R. K. Henderson, "A 256 x 256 SPAD array with in-pixel Time to Amplitude Conversion for Fluorescence Lifetime Imaging Microscopy," paper presented at International Image Sensor Workshop, Vaals, Netherlands, Memory, 900(M4), M5 (2015).
[11]C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, IEEE J. Solid-State Circuits **43**, 12 (2008).
[12]M. Ghioni, A. Gulinatti, I. Rech, F. Zappa, and S. Cova, IEEE J. Sel. Top. Quantum Electron. **13**, 4 (2007).
[13]D. Bronzi, F. Villa, S. Tisa, A. Tosi, and F. Zappa, IEEE Sens. J. **16**, 1 (2016).
[14]D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R. K. Henderson, M. Gersbach, and E. Charbon, in *Proceedings of IEEE European Solid-State Device Conference* (IEEE, 2009), pp. 204–207.
[15]C. Niclass, M. Sergio, and E. Charbon, Proc. SPIE **6372**, 63720S (2006).
[16]J. Frechette, P. J. Grossmann, D. E. Busacker, G. J. Jordy, E. K. Duerr, K. A. McIntosh, D. C. Oakley, R. J. Bailey, A. C. Ruff, and M. A. Brattain, Proc. SPIE **8375**, 63720S (2012).
[17]D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt, and R. Henderson, in *Proceedings of IEEE International Solid-State Circuits Conference* (IEEE, 2012), pp. 122–123.
[18]M. Crotti, I. Rech, and M. Ghioni, IEEE J. Solid-State Circuits **47**, 3 (2012).
[19]B. Markovic, S. Tisa, F. A. Villa, and F. Zappa, IEEE Trans. Circuits Syst. **60**, 3 (2013).
[20]P. Peronio, G. Acconcia, I. Rech, and M. Ghioni, Rev. Sci. Instrum. **86**, 11 (2015).
[21]T. C. Weigandt, B. K. B. Kim, and P. R. Gray, in *Proceedings of IEEE International Symposium on Circuits and Systems* (IEEE, 1994), pp. 27–30.
[22]See http://www.ti.com/product/LMH7324/description Texas Instrument, "LMH7324 datasheet."