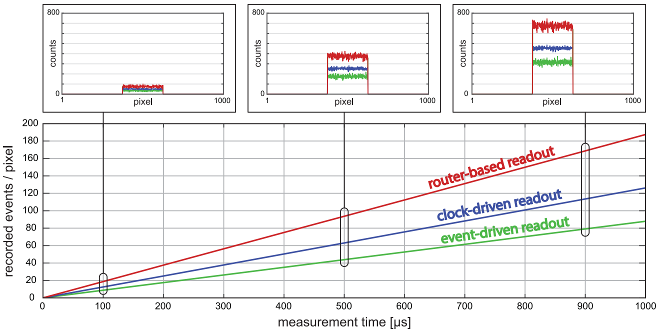
IEEE Photonics Journal
An IEEE Photonics Society Publication

Readout Architectures for High Efficiency in Time-Correlated Single Photon Counting Experiments—Analysis and Review

Volume 9, Number 3, June 2017

Alessandro Cominelli
Giulia Acconcia
Pietro Peronio
Ivan Rech, *Member, IEEE*
Massimo Ghioni, *Senior Member, IEEE*



DOI: 10.1109/JPHOT.2017.2695519
1943-0655

Readout Architectures for High Efficiency in Time-Correlated Single Photon Counting Experiments—Analysis and Review

Alessandro Cominelli, Giulia Acconcia, Pietro Peronio,
Ivan Rech, ***Member, IEEE,***
and Massimo Ghioni, ***Senior Member, IEEE***

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano
20133, Italy

DOI:10.1109/JPHOT.2017.2695519

Manuscript received March 31, 2017; accepted April 12, 2017. Date of current version May 19, 2017.
Corresponding author: Alessandro Cominelli (e-mail: alessandro.cominelli@polimi.it).

Abstract: In recent years, time-correlated single photon counting (TCSPC) has become the technique of choice in many life science analyses, where fast and faint luminous signals are recorded with picosecond accuracy. Nevertheless, the maximum operating frequency of a single TCSPC acquisition channel limits the measurement speed, especially when scanning point systems are exploited. In order to increase the speed of TCSPC experiments, many multichannel systems based on single photon avalanche diode arrays have been proposed in the literature, which integrate thousands of pixels on the same chip. Unfortunately, the huge number of data generated by this kind of system can easily bring to the saturation of the transfer bandwidth to the external processing unit. For this reason, several different readout architectures have been proposed in the literature, attempting to exploit at best the limited bandwidth under TCSPC operating conditions. In this paper, some typical readout approaches, namely clock-driven and event-driven readouts, are discussed and compared, along with a recently-introduced router-based algorithm that is specifically designed to obtain maximum bandwidth exploitation under any condition. Quantitative comparisons are performed starting from imager response of the systems, which is the rate of recorded events in the case of uniform illumination of the detector array.

Index Terms: Single photon avalanche diodes, single photon avalanche diode (SPAD), time correlated single photon counting (TCSPC), readout comparison, imager response, smart router.

1. Introduction

Time-correlated single photon counting (TCSPC) is recognized as a valuable technique for efficient recording of extremely fast and low-intensity luminous signals.

In life sciences, time-resolved imaging by means of TCSPC is the enabling technology for several powerful analytical techniques, such as fluorescence lifetime imaging microscopy (FLIM), Förster resonance energy transfer (FRET), or fluorescence lifetime correlation spectroscopy (FLCS) [1]–[3].

In a typical TCSPC experiment, a biological sample is excited by means of a periodic laser source and the arrival times of fluorescence photons are recorded and processed in order to measure the temporal behavior of the re-emitted waveform and extract the relevant information, e.g., the decay time-constants.

The main limitation of this kind of measurement is related to the maximum operating frequency of a TCSPC acquisition channel. In fact, at high detection rate the system undergoes a significant loss of events, resulting into a distortion of the recorded waveform [4]. In particular, the use of an independent measurement channel at relatively high detection frequency can lead to two different kind of distortion: classic pile-up and counting loss. Pile-up distortion occurs when more than one photon is detected during the same laser period. In that case only the first photon is successfully recorded, resulting in an artificial speedup of the measured fluorescence decay [4].

Counting loss is related to the combined dead time of detector and conversion electronics, which makes the system blind for a fixed time interval after a photon has been recorded.

In order to limit the distortion under a tolerable level, the power of the laser source is adjusted to keep the detection rate well below 10% of the excitation frequency, thus leading to a negligible probability of observing more than one photon in a period [4].

Unfortunately, the collection of a significant amount of data at low count rate involves a relatively long measurement time, especially when a large, time-resolved image is obtained by means of a scanning point system with a single acquisition channel.

Over the past few years, much effort has been made in order to overcome this speed limitation leading, in particular, to the design of many systems based on the parallelization of a large number of TCSPC acquisition channels.

High-performance multi-module commercial systems are nowadays available but, given the high power consumption and area occupation of each module, the degree of parallelization is still limited to only 4 or 8 channels [5], [6].

On the other hand, the exploitation of complementary metal–oxide semiconductor (CMOS) technology has paved the way to the integration of thousands of acquisition chains based on single photon avalanche diodes (SPADs) on the same chip, [7]–[11] thus leading, in principle, to a proportional increment of the measurement speed.

Nevertheless, the presence of a huge number of detectors gives rise to a considerably high rate of events, which can easily lead to a saturation of the transfer bandwidth to the external processing unit.

When considering a typical laser frequency of 80 MHz and a detection probability of 5%, a mean detection rate of 4 Mcps is obtained for each channel. When 1024 pixels are exploited and 2 bytes are used to encode each timing information, a mean throughput of 64 Gb/s is reached. As a result, a much higher bandwidth is needed to transfer the timing data without a significant loss of events.

Unfortunately, the real-time elaboration of such a high-throughput data stream involves a considerable complexity of the system design. As a result, bandwidth saturation is, to date, one of the main limitation to the speed of TCSPC measurements.

Recently, many different readout architectures have been proposed in literature to cope with a limited bandwidth of the bus to the PC, trying to maximize its exploitation under typical operating conditions.

The goal of this paper is a quantitative comparison between the main readout approaches proposed in literature for SPAD imagers used in time-correlated experiments.

In Section 2, the working principles of the main readout architectures are described, along with a recently-introduced router-based architecture [12], specifically designed to ensure maximum bandwidth occupation. The systems are compared on the basis of their imager response and both qualitative and quantitative considerations are reported. In Section 3 some simulations are performed to validate the quantitative results, while conclusions are drawn in Section 4.

2. Imager Response

The key parameter to compare different readout architectures is the time needed to acquire a significant number of events from each pixel. In general, this time is a function of the detection frequency of each single channel, and therefore, it strongly depends on the experiment.

Nonetheless a quantitative comparison between different readout systems can be performed on the basis of the rate of recorded events as a function of the rate of impinging photons, in the simple case of uniform illumination on the whole detector array.

In general, only a portion of the photons impinging on the system is effectively detected due to the limited collection efficiency of the imager, that results from a combination of the detectors photon detection efficiency (PDE) and system fill factor.

In this scenario, the presence of circuitry specifically dedicated to data readout could, in principle, be a significant limitation to the overall collection efficiency. Nevertheless, this limitation is not directly related to the readout approach, but rather to the practical implementation of the algorithm.

Moreover, the introduction of 3-D-stacking techniques can effectively solve this issue and have been already proposed for different readout architectures [12], [13].

The impact of the sole readout algorithm on the measurement speed can be evaluated starting from the imager response function $r_{rec} = f(r_{det})$, that is the average rate of recorded events, r_{rec}, as a function of the detection rate, r_{det}.

In the case of ideal detector and readout system every detected event is recorded and the imager response shows a linear behavior

$$r_{rec} = r_{det}. \tag{1}$$

Due to the non-ideal readout process the response function approaches the ideal line only at low detection rate. Conversely, at high frequencies the response curve saturates to a level that depends on the ability of the readout architecture to efficiently exploit the limited bandwidth of the system. It follows that the same bandwidth must be considered to perform a fair comparison of different systems.

In a real TCSPC system the presence of a finite dead time of the detector and the conversion electronics introduce an additional source of event loss. Nevertheless, the readout architecture typically represents the main limitation, especially in large arrays. For this reason, the recorded event rate can be estimated considering only the loss of events due to the readout approach, as will be clarified later in this section.

We can define the counting efficiency as the ratio between the recorded event rate and the detected event rate, i.e., the fraction of detected events which are also recorded by the system. The counting efficiency can be used as a figure of merit to rank the different readout architectures.

2.1 Clock-Driven Readout

In high-density systems, a widely used architecture is based on the integration of each detector with all the necessary electronics, including a time-measurement circuit and a memory to store the timing data [11], [14], [15]. Each memory location is read out periodically in synch with a clock signal (hence the name clock-driven readout), while the pixel is able to measure new events [11], [14].

In this case each clock period is used simultaneously as readout and dwell time and only the first event in each period is registered in the memory and recorded at the output. The event loss mechanism is described in Fig. 1.

In this kind of architecture the clock period must be sufficiently long to ensure a complete readout of the array and in systems presented so far it ranges from about 1 to 10 μs for one thousand pixels [7], [8], [11], [14].

The transfer bus is designed to sustain the maximum possible throughput, which is reached only in the case of maximum occupation of the channels, i.e., when one event per dwell time is recorded for each pixel. When considering a number of pixels equal to N_{pix} and N_{bit} bits to encode the timing data, the bandwidth of the system can be expressed as follows:

$$BW = N_{pix} \cdot N_{bit} \cdot \frac{1}{T_{dwell}}. \tag{2}$$

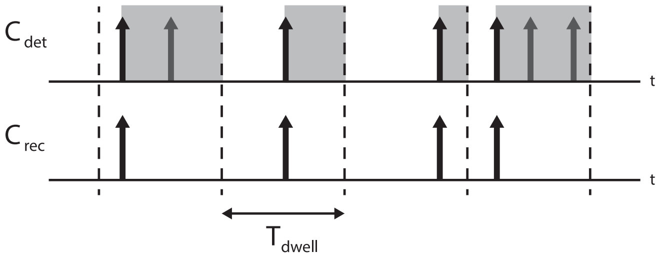


Fig. 1. Event loss mechanism with a clock-driven readout. Only one of the detected events C_{det} (the first one) is recorded during a dwell time.

As a result, given 10 bits to encode the timing data and one thousand pixels, the bandwidth ranges between 1 and 10 Gbit/s.

Usually, both the excitation period of the laser and the dead time of the single channel are well below 1 μs and the analytical expression of the imager response can be computed considering only the limitation imposed by the finite bandwidth of the system. As can be inferred from Fig. 1, a single event is registered at the output only if at least one photon impinges on the detector during a period of the clock (T_{dwell}).

In a typical TCSPC experiment the detection rate varies periodically with time, since a fluorescence pulse is emitted during each excitation cycle. Nevertheless, in clock-driven readouts T_{dwell} is far longer than a single excitation period and the probability to observe one or more events during a dwell time can be evaluated as the product between T_{dwell} and the average detection rate r_{det}. It is worth noting that this result also holds for a dwell time that lasts for an integer number of excitation cycles.

Given a Poisson distribution of the light, the probability to detect at least one photon during a dwell time is given by $1 - e^{-T_{dwell}}$, resulting in a mean recorded rate equal to

$$r_{rec} = \frac{1 - e^{-T_{dwell} \cdot r_{det}}}{T_{dwell}}. \quad (3)$$

As expected, the curve saturates at high detection rates and the maximum frequency is equal to $1/T_{dwell}$, corresponding to one recorded event per dwell time.

In Fig. 2 a comparison between imager responses with different dwell times is reported, along with the fraction of detected events which is successfully recorded by the system (r_{rec}/r_{det}). It is worth noting that a faster readout frequency leads to a lower loss of events, thus bringing to higher recorded rate and so to higher counting efficiency.

On the other hand, the probability to detect a photon in a channel (equal to $1 - e^{-T_{dwell} \cdot r_{det}}$) reduces with the clock period resulting into an incomplete exploitation of the converters which becomes evident especially at high bandwidth.

As a result, the use of a clock-driven architecture keeps the speed of a TCSPC measure below the limit imposed by the bandwidth of the system. This can be mathematically shown by looking at the imager response as a function of the system bandwidth, which results from a manipulation of (2) and (3):

$$r_{rec} = \left(1 - e^{-\frac{N_{pix} \cdot N_{bit} \cdot r_{det}}{BW}}\right) \cdot \frac{BW}{N_{pix} \cdot N_{bit}}. \quad (4)$$

It is clear that an increase in the bandwidth does not lead to a proportional increment of the recorded event rate, thus resulting in an inefficient exploitation of the bus, especially at high bandwidth.

For example, considering 1000 pixels and N_{bit} equal to 10, a bandwidth of 5 Gbit/s is obtained with a dwell time equal to 2 μs.

If the detection rate is 800 kcps (i.e. 1% of an 80 MHz laser) approximately half the events are not measured, resulting in a mean recorded rate equal to 400 kcps for each pixel, that is 80%

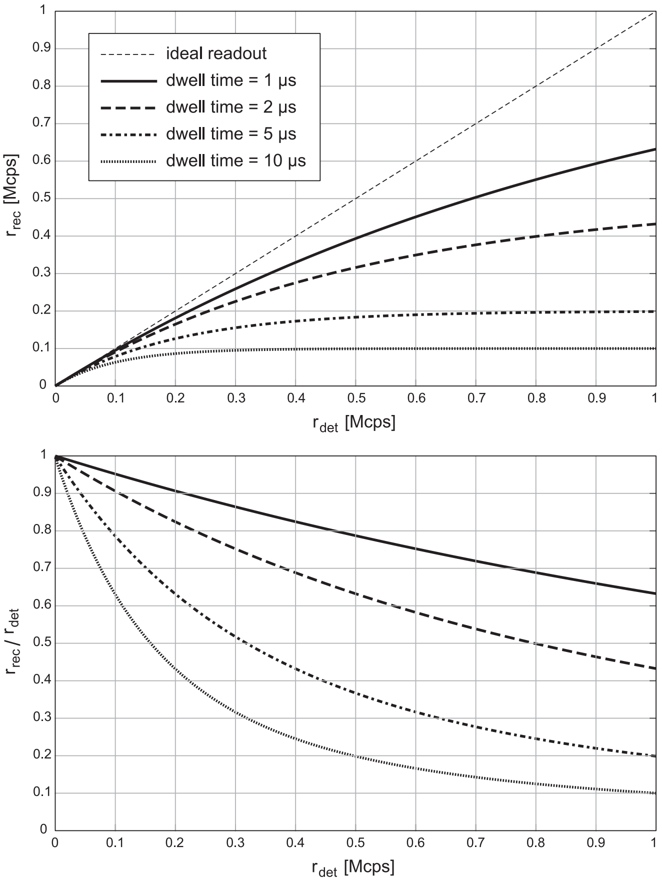


Fig. 2. Imager response of a clock-driven architecture featuring different dwell times. The counting efficiency r_{rec}/r_{det} is also reported.

bandwidth occupation. Instead, a bandwidth of 10 Gbit/s (T_{dwell} of 1 μs) leads to a recorded rate up to 550 kcps per pixel over a maximum of 1 Gcps, and therefore, only 55% of the bandwidth is exploited.

2.2 Event-Driven Readout

Typical TCSPC operating rates does not ensure a complete exploitation of the detectors during every excitation period. Therefore, integrating a complete conversion chain for each pixel can result in an overall inefficiency of the system in terms of both area occupation and power dissipation.

An alternative is represented by systems where a limited set of conversion channels is shared within a much larger number of detectors.

In event-driven architectures [13], [16], [17], the system is usually subdivided in independent clusters of pixels, each one sharing a single line directed towards an external time-conversion circuit. The line behaves like a digital bus and every pixel of the same cluster can access it; in particular, when a photon is detected the line changes its status and it cannot be accessed by other pixels for a fixed time T_{line}.

An event loss mechanism is responsible for the saturation of the recorded rate, since no events in the cluster can be measured during the line dead time T_{line}. The mechanism is shown in Fig. 3.

Like in clock-driven systems, the bandwidth limitation is related to the saturation of the recorded event rate.

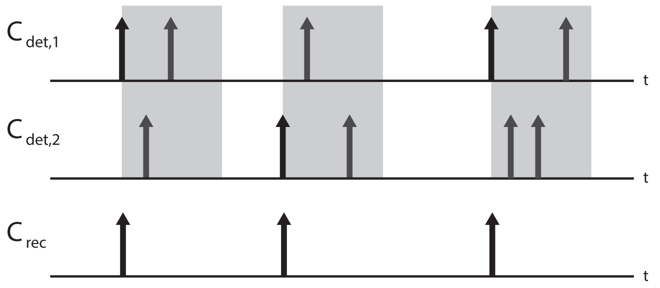


Fig. 3. Event loss mechanism with an event-driven readout. Here, two pixels share the same timing line and each time a photon is recorded the system remains blind for a fixed time $T_{line.}$

When working at saturated data rate, timing measurements are generated in each cluster at a frequency equal to $1/T_{line}$. Furthermore, in order to properly build a time-resolved image starting from timing data, an address comes along each timing information, identifying the pixel within the cluster [16].

As a result, the transfer bandwidth is equal to

$$BW = N_{clusters} \cdot [N_{bit} + log_2(N_{pix,c})] \cdot \frac{1}{T_{line}} \tag{5}$$

where $N_{pix,c}$ is the dimension of a cluster, while $N_{clusters}$ is the number of clusters, that is the ratio between the total number of pixels in the array and the cluster dimension. N_{bit} bits are exploited to encode the timing information, while $log_2(N_{pix,c})$ is the minimum number of bits needed to identify the position of different pixels in a cluster.

If T_{line} is sufficiently higher than the dead time of a single SPAD a closed-form expression can be found for the imager response of the system.

As can be inferred from Fig. 3, during the whole measurement time the system is blind for an interval T_{line}, starting from each recorded event. As a result, the whole dead time of the system during a long measure can be calculated as

$$T_{dead,TOT} = N_{rec} \cdot T_{line} \tag{6}$$

where N_{rec} is the total number of recorded events during the measure.

When considering the same illumination for all the detectors N_{rec} can be expressed as a product between the cluster dimension $N_{pix,c}$ and the number of recorded events in one pixel n_{rec}:

$$T_{dead,TOT} = N_{pix,c} \cdot n_{rec} \cdot T_{line}. \tag{7}$$

Given a random distribution of photons over time, the fraction of detected photons which are also recorded (i.e. r_{rec}/r_{det}) is equal to the fraction of time during which the system is able to record data, that is

$$\frac{r_{rec}}{r_{det}} = \frac{T_{measure} - T_{dead,TOT}}{T_{measure}} = 1 - N_{pix,c} \cdot T_{line} \cdot \frac{n_{rec}}{T_{measure}} \tag{8}$$

where, by definition, the ratio between n_{rec} and $T_{measure}$ is equal to r_{rec}. Rearranging (8), the following result is obtained:

$$r_{rec} = \frac{r_{det}}{1 + N_{pix,c} \cdot T_{line} \cdot r_{det}}. \tag{9}$$

It is worth noting that the imager response does not only depend on the dead time T_{line}, but on the cluster dimension as well. In fact, in contrast to the case of clock-driven readout architectures, pixels are not independently read out, but the registered event rate originated from a detector depends on the presence of other pixels sharing the same converter. In particular, only one event in the cluster can be recorded at a time, leading to a high loss of events due to pile-up and counting loss effects

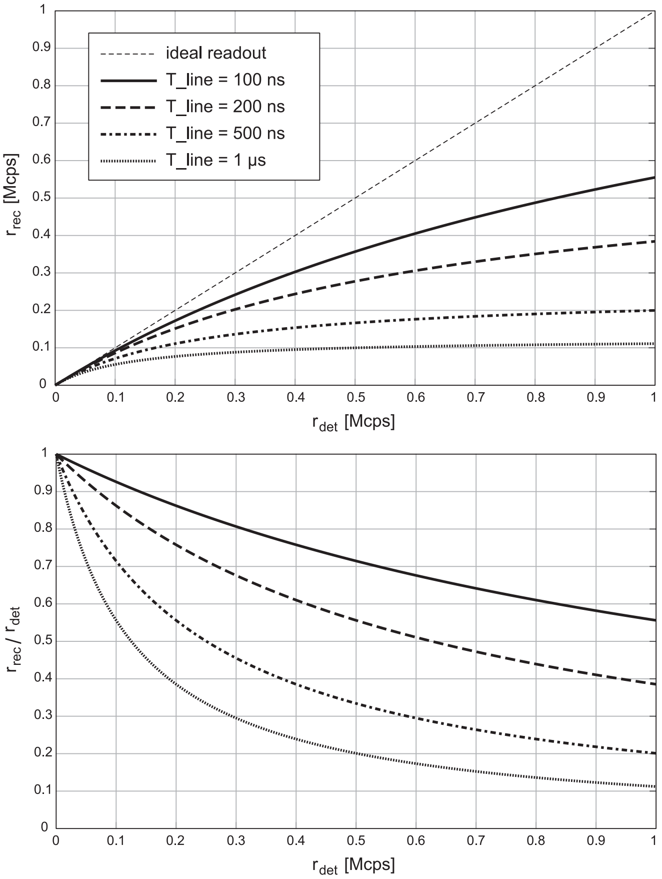


Fig. 4. Imager response of an event-driven architecture featuring 8-pixel clusters and dead times T_{line}. The counting efficiency r_{rec}/r_{det} is also reported.

in large clusters [13]. For this reason, $N_{pix,c}$ is usually maintained low (4 or 8 are typical values) [13], [17], [18].

The imager response with an event-driven readout architecture is shown in Fig. 4, considering groups of 8 pixels and different dead times.

Also in this case a lower dead time, i.e., a larger bandwidth, leads to higher counting efficiency and once again typical TCSPC detection rates do not allow the whole bandwidth occupation.

Rearranging (5) and (9), the recorded event rate can be expressed as a function of the bandwidth BW:

$$r_{rec} = \frac{r_{det} \cdot BW}{BW + N_{pix,array} \cdot [N_{bit} + log_2(N_{pix,c})] \cdot r_{det}} \tag{10}$$

where the number of pixels in the array, $N_{pix,array}$, results from the multiplication between $N_{clusters}$ and $N_{pix,c}$.

It is worth highlighting that this result is strictly valid only in case of constant illumination. Conversely, a typical TCSPC experiments involves a pulsed distribution of detected photons over time. As a result, the recorded rate follows the same trend expressed in (9) and (10), but r_{rec} decreases in discrete steps when increasing T_{line}.

In any case, (9) and (10) represent a good approximation of the average recording rate r_{det} if T_{line} is much higher than the laser period.

A comparison between the recorded data rate using an event-driven or a clock-driven architecture is shown in Fig. 5. It is evident that the counting efficiency is higher with a clock-driven readout,

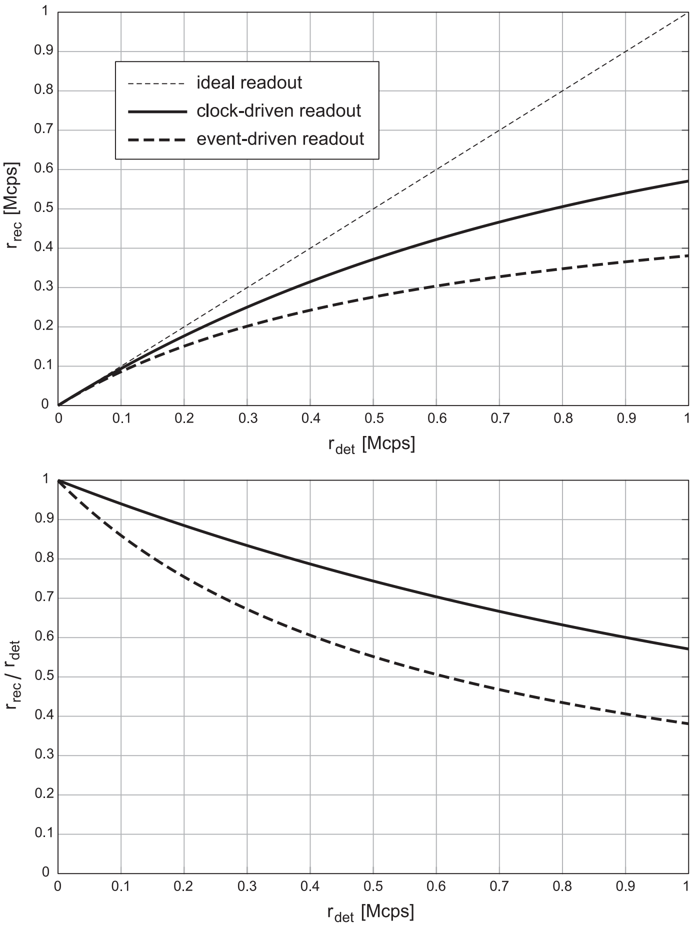


Fig. 5. Comparison between a clock-driven and an event-driven architecture in terms of imager response (top) and counting efficiency (bottom). The imager responses follow (4) and (10) respectively. The two systems have the same bandwidth, equal to 8 Gbit/s, considering arrays of 1000 pixels and 10 bits to encode timing information. 8-bits clusters are considered for the event-driven architecture.

resulting in faster measurement speed. This applies for every detection rate, even when the whole bandwidth is exploited, that is at saturation. In fact, in event-driven systems part of the bandwidth is always used to transfer the pixel position, resulting in a lower rate of events that can be recorded.

On the other hand, event-driven architectures feature a better exploitation of resources, thus allowing less area occupation and power dissipation of the converters.

2.3 Router-Based Readout

The readout architectures proposed so far do not guarantee the full exploitation of transfer bandwidth to the processing unit.

Instead, we recently proposed a novel readout algorithm, specifically designed to take maximum advantage of the data communication bandwidth under any operating condition [12].

In the proposed architecture a limited set of L timing lines is shared with the whole array, each one directed to an external converter able to operate at laser repetition frequency [19]. The number of lines is evaluated starting from the maximum throughput the elaboration unit can handle, thus guaranteeing the maximum exploitation of the bandwidth even at low detection rate. In particular, each time the number of pixels detecting a photon is higher than or equal to the number of lines the full bandwidth is exploited.

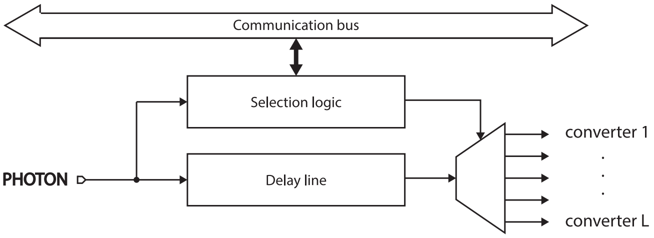


Fig. 6. Simplified schematic of the routing logic dedicated to a pixel of the array. A dedicated bus is needed to allow a communication between different pixels, while an in-pixel selection logic performs the routing of the signal towards L external converters. A delay line is included to keep the timing information during the selection process.

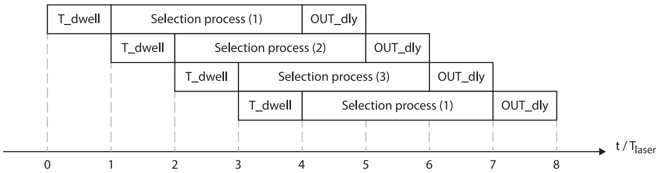


Fig. 7. Timing diagram of the routing algorithm: Each time a photon is detected during a dwell time a selection process starts, lasting for more than one period (for instance, here three periods are needed). At the end of this phase the delayed signals are measured by means of the external converters. Each excitation period can be exploited as dwell time, since a pipelined mechanism has been designed on purpose. In particular, more than one communication bus is present, and therefore, different selections can be performed at the same time.

Like in event-driven architectures an address comes along with each timing information, in order to track the position of the pixel within the whole array, thus demanding for a bandwidth equal to

$$BW = L \cdot [N_{bit} + log_2(N)] \cdot \frac{1}{T_{laser}} \tag{11}$$

where N_{bit} is the number of bits needed to encode the timing measure, while $log_2(N)$ bits are required to identify the pixel within N pixels composing the array.

A simplified schematic of the routing logic associated to one pixel is represented in Fig. 6.

Like in clock-driven systems, a dwell time is used to accumulate events into the array but, instead of a converter followed by a memory, a low-jitter delay line is integrated in each pixel to keep the timing information.

Each excitation cycle is used as dwell time; at the end of every period a smart selection logic makes a comparison between pixels which detected a photon and some of them are chosen to be routed towards the converters. Only after the selection process is over, the pixels are connected to the lines by means of a demultiplexer and the delayed signals can be processed by the external electronics.

In general, the selection mechanism lasts for more than one laser period, therefore a pipeline-like architecture has been included; in this way a new selection can start, while another is being carried out. A timing diagram of the algorithm is reported in Fig. 7, including the pipeline mechanism.

Three different event loss mechanism arise in the algorithm described so far. First of all, like in classical clock-driven architectures, only one event (the first one) is recorded during the dwell time, while the others are lost. At this point, the pixel participates to a selection process, starting from the end of the dwell time and for an integer number, α, of excitation cycles, which includes the selection process and the period used by the delay line to output the signals, no photon impinging on the pixel can trigger a new comparison, resulting in a second source of counting loss.

Finally, each time a high number of pixels detects a photon during the same excitation period, only a limited set is selected by the router to be connected to the external converters, while all the other events are discarded.

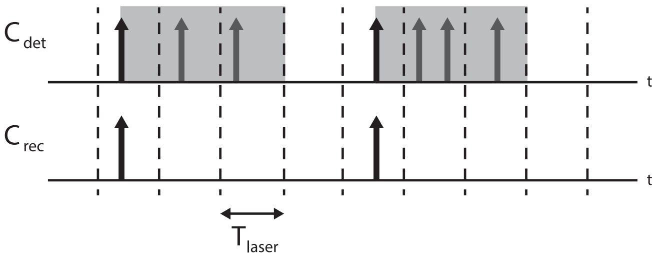


Fig. 8. Event loss mechanism with a router-based readout and a timing line dedicated to each pixel. For instance, here, the selection process lasts for a single laser period, so $\alpha = 2$.

The combination of the different event-loss mechanisms leads to a saturation of the imager response.

For the sake of simplicity, it is possible to separate the analysis of the response in two parts. Initially, only the first two event loss mechanisms are taken into account. As a result, the rate of participations to a comparison is found, that is the mean number of times a pixel enters a selection process during the whole measure. Finally, the third event loss mechanism, due to the presence of other pixels taking part to the selection process, is included in the analysis.

The rate of participations to a selection process, r_{sel}, is derived by neglecting the third event loss mechanism, that is equivalent to calculate r_{rec} for a router-based system with a number of timing lines equal to the pixel number. In this case the selection logic does not play any role and each pixel exhibits only a dead time whose end is synchronous with the laser pulse. The situation is shown in Fig. 8.

The whole measurement time, $T_{measure}$, can be expressed as the sum of two contributions: an integer number of laser periods used by the logic to perform the routing (including the selection processes and the time taken by the delay line to send the timing signal to the external conversion electronics) and the remaining time, consisting in dwell times where either zero or one event is recorded.

Considering the number of times a pixel participates to a selection during the measurement (N_{sel}) the first contribution has a duration $T_{sel,TOT}$, which can be expressed as

$$T_{sel,TOT} = \alpha \cdot T_{laser} \cdot N_{sel} \tag{12}$$

where $\alpha \cdot T_{laser}$ is the time needed to perform the routing.

On the other hand, the remaining time is equal to the difference between the measurement time, $T_{measure}$, and $T_{sel,TOT}$.

The remaining time can be considered as the effective measurement time to calculate the participation rate to the selection process. The situation is totally equivalent to the one reported in Fig. 1, where T_{dwell} is equal to T_{laser}. As a result

$$r_{sel} = \frac{1 - e^{-T_{laser} \cdot r_{det}}}{T_{laser}} \cdot (1 - \alpha \cdot T_{laser} \cdot r_{sel}) \tag{13}$$

where the same expression of (3) is reduced by the factor $(1 - \alpha \cdot T_{laser} \cdot r_{sel})$, that is the probability that an event occurs in an excitation period of the remaining time.

It is worth noting that the equation was obtained for a constant r_{det}; however it also applies to the case of pulsed illumination. In fact, the dwell time coincides with a period, so the average number of photons detected in a dwell time does not depend on the distribution of detected photons within a period, but only on its integral, so only on the average value.

Rearranging (13), the following result is obtained:

$$r_{sel} = \frac{1 - e^{-T_{laser} \cdot r_{det}}}{1 + \alpha \cdot (1 - e^{-T_{laser} \cdot r_{det}})} \cdot \frac{1}{T_{laser}}. \tag{14}$$

As expected, the curve approaches the ideal line for low detection rates and reaches a saturation at high frequencies due to event loss. Numerical simulations have confirmed the validity of (14), when considering the system of Fig. 8, i.e., a timing line dedicated to each detector.

Nevertheless, a router-based approach features a limited number of lines L, along with a much higher number of pixels N. To this aim, a pixel selection is performed during each comparison, thus introducing an additional source of counting loss, which must be considered in the imager response of the system.

Considering an excitation cycle, the probability P that a pixel participates to a selection starting from the following period can be simply evaluated from (14):

$$P = r_{sel} \cdot T_{laser}. \tag{15}$$

From a probabilistic point of view, during each period a binomial experiment is performed: a pixel can either enter a selection phase (with probability P) or not (with probability $1 - P$). As a result, the probability that a pixel enters in a selection phase where exactly n pixels ask for an access to the external channels is equal to

$$P_n = P^n \cdot \binom{N-1}{n-1} \cdot (1-P)^{N-n} = \frac{n}{N} \cdot P^n \cdot \binom{N}{n} \cdot (1-P)^{N-n} \tag{16}$$

where the binomial coefficient can be expressed in the form

$$\binom{N}{n} = \frac{N!}{n! \cdot (N-n)!}. \tag{17}$$

Given a random selection mechanism, the probability of a pixel to be routed towards one of the converters depends on the number of pixels taking part to the comparison. In the simple case of a single timing line shared with N pixels, if only one pixel detects a photon its probability to be recorded is 1. Conversely, if n competitors are present the probability drops to $1/n$.

In the general case of L timing lines shared with the array, each time n is less or equal than L all the competitors are routed towards an external converter and no event is lost. On the contrary, if a higher number of pixels participate to a comparison ($n > L$) only L pixels are randomly selected by the routing logic, resulting in a recording probability P_{rec} equal to $P_n \cdot L/n$:

$$P_{rec}(n) = \begin{cases} 1 \cdot P_n, & n \le L \\ \frac{L}{n} \cdot P_n, & n > L. \end{cases} \tag{18}$$

The overall probability to record an event is equal to the sum of the probabilities $P_{rec}(n)$ with n ranging from 1 to N, that is the number of pixels in the array. The imager response of the system is obtained dividing this probability by the laser period

$$r_{rec} = \frac{1}{T_{laser}} \cdot \sum_{n=1}^{N} P_{rec}(n) = \frac{1}{T_{laser}} \cdot \left[\sum_{n=1}^{L} P_n + \sum_{n=L+1}^{N} \frac{L}{n} \cdot P_n \right]. \tag{19}$$

It is now possible to simplify the expression:

$$\begin{aligned} r_{rec} &= \frac{1}{T_{laser}} \cdot \left[\sum_{n=0}^{L} P_n + \sum_{n=0}^{N} \frac{L}{n} \cdot P_n - \sum_{n=0}^{L} \frac{L}{n} \cdot P_n \right] \\ &= \frac{1}{T_{laser}} \cdot \left[\sum_{n=0}^{N} \frac{L}{n} \cdot P_n + \sum_{n=0}^{L} \left(1 - \frac{L}{n}\right) \cdot P_n \right] \\ &= \frac{L}{N \cdot T_{laser}} \cdot \left[\sum_{n=0}^{N} P^n \cdot \binom{N}{n} \cdot (1-P)^{N-n} - \sum_{n=0}^{L-1} \frac{L-n}{L} \cdot P^n \cdot \binom{N}{n} \cdot (1-P)^{N-n} \right]. \end{aligned} \tag{20}$$

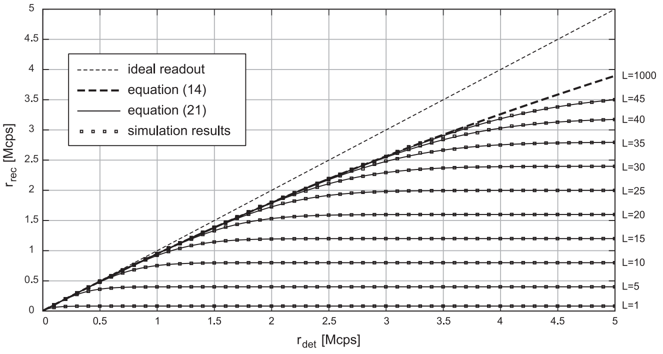


Fig. 9. Response of a 1000-pixels imager featuring a router-based readout architecture with the timing diagram of Fig. 7 ($\alpha = 4$). The curves have been obtained from (21) for different numbers of timing lines, ranging from 1 to 45. The response with a number of lines equal to the dimension of the array ($L = 1000$), expressed in (14), is also shown.

By applying the binomial theorem on the first summation, the imager response is obtained as

$$r_{rec} = \frac{L}{N \cdot T_{laser}} \cdot \left[1 - \sum_{n=0}^{L-1} \frac{L-n}{L} \cdot P^n \cdot \binom{N}{n} \cdot (1-P)^{N-n} \right]. \tag{21}$$

The formula has been validated by means of numerical simulations. In particular, the whole routing algorithm presented in [12] has been simulated at different r_{det} (considering a Poisson distribution of events) and the corresponding r_{rec} has been extracted and compared with the theoretical result expressed in (21).

In Fig. 9 the response of an imager with 1000 pixels is shown for a number of shared lines ranging from 1 to 45, along with simulated results. It is worth noting that the imager response depends mainly on three elements: the duration of the comparison phase, expressed by the α parameter, the number of pixels N and the number of shared lines L.

It can be shown that the presence of a number of pixels much higher than L ($N >> L$) gives rise to an imager response which is almost independent from the value of α. In particular, the curve saturates to a value close to $L/(N \cdot Tlaser)$, which corresponds to the limitation set by the bandwidth, despite the duration of the comparison phase.

A comparison based on the imager response between the router-based system and other readout architectures is shown in Fig. 10.

It is worth noting that the imager response of a router-based architecture approaches saturation, that is maximum bandwidth occupation, at lower detection rates if compared with clock-driven and event-driven readouts, thus leading to higher counting efficiency at low frequencies. For instance, in the example of Fig. 10 a detection rate of 800 kcps (1% of the laser rate) is sufficient to provide a maximum exploitation of the bus bandwidth.

The comparison shows that a router-based approach results faster than clock-driven solutions only at frequencies well below 1% of the excitation rate (see Fig. 10). In fact, the maximum bandwidth occupation comes along with a large number of address bits, which are not required with clock-driven architectures. Concerning event-driven readouts, the maximum measurement speed is relatively low at pile-up-limited detection rates, even if the number of address bits can be limited to 2 or 3.

3. Simulations Under Typical TCSPC Conditions

The imager response of a readout architecture provides a useful tool to compare different systems in the case of uniform illumination of the array.

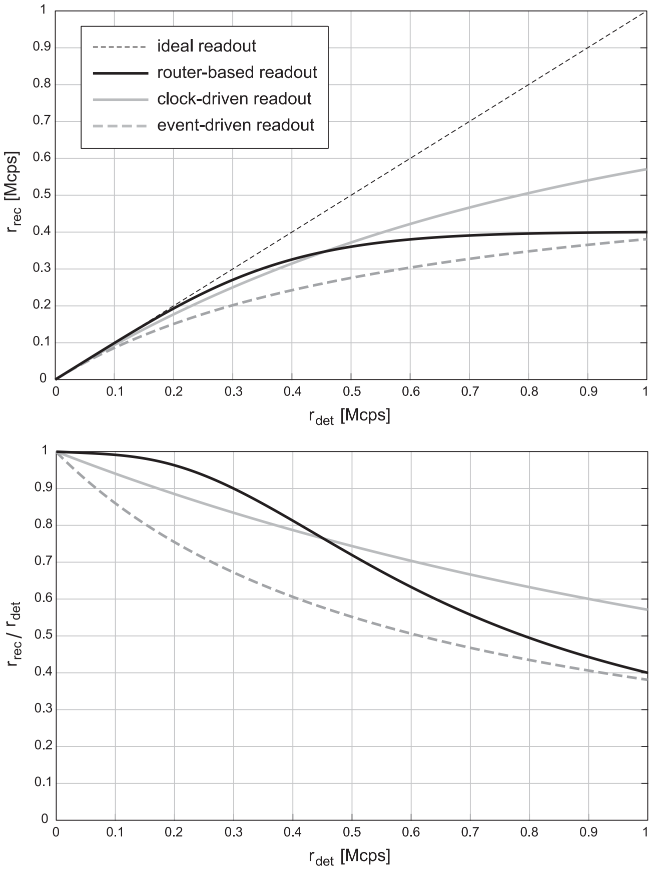


Fig. 10. Comparison between the different readout architectures described in the article, both in terms of imager response (top) and counting efficiency (bottom). The three systems have the same bandwidth, equal to 8 Gbit/s, considering arrays of 1000 pixels and 10 bits to encode timing information. 8-bits clusters are considered for the event-driven architecture, while 5 timing lines and $\alpha = 4$ have been considered for the router-based system.

Nonetheless, this situation does not reflect the spatial distribution of fluorescence phenomena in real TCSPC experiments. In fact, illumination is usually confined to some parts of the array, thus impairing the performance of both clock-driven and event-driven architectures.

In particular, clock-driven systems involve the readout of every pixel location, irrespective of the presence of timing data, resulting in a reduction of the throughput every time illumination is not uniformly distributed on the array. Similarly, in event-driven systems the counting efficiency drops each time a cluster is not illuminated.

The effect of a non-uniform illumination is shown in Fig. 11 for different readout systems: Here, only a portion of the array, namely a quarter, is illuminated at 800 kcps detection rate, while no photon impinges on other pixels. In particular, the illumination on each detector has been emulated by means of a Poisson process, while events are recorded using the different approaches described in Section 2.

It is evident that, in this case, the router-based architecture records a higher number of events, since the converters are shared amongst a lower number of detectors if compared to the case of uniform illumination, thus resulting in higher counting efficiency.

It is worth highlighting that in clock-driven and router-based architectures the spatial distribution of fluorescence signals does not have any impact on the counting efficiency of the system, since each pixel receives the same treatment. It follows that the results shown in Fig. 11 are valid also for different illumination patterns. Conversely, in event-driven systems, a sparser illumination leads

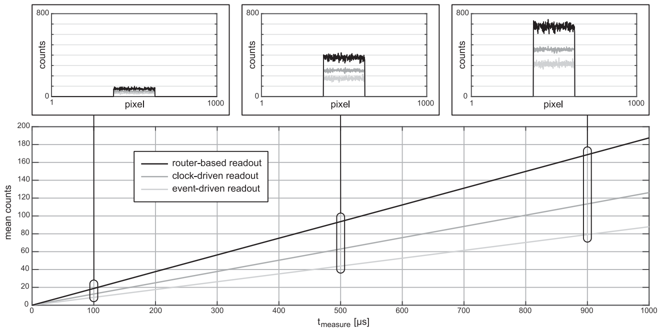


Fig. 11. Three 1000-pixels imagers, featuring the same bandwidth but different readout architectures (same systems considered in Fig. 10), are illuminated with a rectangular pattern: A quarter of the array (pixels from 376 to 625) detect photons at an average rate of 800 kcps, while no event is registered on other detectors. The events accumulated after a different time $t_{measure}$ elapsed from the beginning of the experiment are reported for each pixel (top), along with the mean number of registered counts over the array (bottom).

to a better exploitation of resources and, therefore, to higher r_{det}. In any case, the event-driven architectures features the lowest counting efficiency, as can be inferred from Fig. 10.

4. Conclusion

The imager response of different readout architectures, namely clock-driven, event-driven and the recently introduced router-based approach, has been mathematically derived, thus allowing a comparison in terms of measurement speed in the simple case of uniform illumination of the whole array.

In this case, clock-driven architectures exhibit the highest counting efficiency, thus leading to faster TCSPC measurements. Nevertheless, in typical TCSPC experiments fluorescence phenomena do not affect the whole set of detectors and the presence of a converter dedicated to each pixel is, in principle, unnecessary.

Numerical simulations show that the router-based architecture features the highest counting efficiency in the case of illumination limited to a subset of the detectors, thus representing a valid alternative to both clock-driven and event driven-readouts and paving the way to a significant speedup of TCSPC experiments.

References

[1] W. Becker, “Fluorescence lifetime imaging–techniques and applications,” *J. Microsc.*, vol. 247, pp. 119–136, 2012.
[2] H. Wallrabe and A. Periasamy, “Imaging protein molecules using fret and flim microscopy,” *Current Opinion Biotechnol.*, vol. 16, pp. 19–27, 2005.
[3] P. Kapusta, M. Wahl, A. Benda, M. Hof, and J. Enderlein, “Fluorescence lifetime correlation spectroscopy,” *J. Fluorescence*, vol. 17, pp. 43–48, 2007.
[4] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*. Berlin, Germany: Springer, 2005.
[5] Becker & Hickl GmbH. Tcspc modules. [Online]. Available: http://planning.cs.cmu.edu/humanoids07/p/85.pdf
[6] PicoQuant. Hidraharp 400. [Online]. Available: https://www.picoquant.com/products/category/tcspc-and-time-tagging-modules/hydraharp-400-multichannel-picosecond-event-timer-tcspc-module
[7] C. Veerappan *et al.*, “A 160 × 128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter,” in *Proc. IEEE Int. Solid-State Circuits Conf. Dig. Tech. Papers*, 2011, pp. 312–314.
[8] F. Villa *et al.*, “CMOS imager with 1024 SPADS and TDCS for single-photon timing and 3-D time-of-flight,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 6, pp. 364–373, Nov./Dec. 2014.
[9] L. Parmesan, N. Dutton, N. Calder, L. Grant, and R. Henderson, “A 256x256 SPAD array with in-pixel time to amplitude conversion for fluorescence lifetime imaging microscopy,” in *Proc. Int. Image Sensor Workshop, Vaals, Netherlands, Memory*, 2015, vol. 900, Paper M5.
[10] R. Field, S. Realov, and K. Shepard, “A 100 fps, time-correlated single-photon-counting-based fluorescence-lifetime imager in 130 nm CMOS,” *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 867–880, Apr. 2014.

[11] M. Gersbach *et al.*, “A time-resolved, low-noise single-photon image sensor fabricated in deep-submicron CMOS technology,” *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1394–1407, Jun. 2012.

[12] G. Acconcia, A. Cominelli, I. Rech, and M. Ghioni, “High-efficiency integrated readout circuit for single photon avalanche diode arrays in fluorescence lifetime imaging,” *Rev. Sci. Instrum.*, vol. 87, 2016, Art. no. 1131107.

[13] J. Pavia, M. Scandini, S. Lindner, M. Wolf, and E. Charbon, “A 1 × 400 backside-illuminated SPAD sensor with 49.7 ps resolution, 30 pj/sample TDCS fabricated in 3-D CMOS technology for near-infrared optical tomography,” *Opt. Exp.*, vol. 50, pp. 2406–2418, 2015.

[14] D. Stoppa *et al.*, “A 32 × 32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain,” in *Proc. ESSCIRC*, 2009, pp. 204–207.

[15] F. Villa *et al.*, “SPAD smart pixel for time-of-flight and time-correlated single-photon counting measurements,” *IEEE Photonics J.*, vol. 4, no. 3, pp. 795–804, Jun. 2012.

[16] C. Niclass, M. Sergio, and E. Charbon, “A single photon avalanche diode array fabricated in 0.35-μm CMOS and based on an event-driven readout for TCSPC experiments,” in *Proc. Opt. East*, 2006, pp. 63 720S–63 720S.

[17] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, “A 128 × 128 single-photon image sensor with column-level 10-bit time-to-digital converter array,” *IEEE J. Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, Jan. 2008.

[18] C. Niclass *et al.*, “Design and characterization of a 256 × 64-pixel single-photon imager in CMOS for a MEMS-based laser scanning time-of-flight sensor,” *Opt. Exp.*, vol. 20, pp. 11 863–11 881, 2012.

[19] P. Peronio, G. Acconcia, I. Rech, and M. Ghioni, “Improving the counting efficiency in time-correlated single photon counting experiments by dead-time optimization,” *Rev. Sci. Instrum.*, vol. 86, 2015, Art. no. 113101.