

Single Photon Avalanche Diode Arrays for Quantum Imaging and Microscopy

Francesca Madonini, Fabio Severini, Franco Zappa, and Federica Villa*

Quantum imaging and microscopy profit from quantum correlations and entanglement to image objects and samples with resolution and sensitivity that goes far beyond what can be achieved through classical optics. In order to carry out these techniques, suitable detectors with specific features must be employed. This paper aims to highlight the importance of sensors based on single photon avalanche diodes (SPAD) in quantum imaging and microscopy applications, paving the way for the next-generation ideal quantum imager. After reviewing the main techniques (based on quantum physics principles) for improving the resolution and sensitivity of a sample image, the pros and cons of different sensors, such as avalanche photodiodes (APDs), and the intensified and electron-multiplying charge coupled devices (ICCDs and EMCCDs), are identified. Then the analysis mainly focuses on SPAD-based detectors, identified as the best candidates for quantum imaging, critically discussing the requirements and performance, also in relation to already existing SPAD-based architectures with specific features fitting the application. Eventually, next-generation quantum imagers should integrate together all the best architectural choices herewith presented, so as to detect photon coincidences and to perform efficient event-driven readout, also by exploiting a suitable technology and SPAD design to optimize the discussed detection performance.

1. Introduction to Quantum Imaging and Microscopy

Measurement techniques enhanced by the laws of quantum mechanics have been proved in plenty of successful real-world applications, such as dynamic biological measurements,^[1] gravitational waves detection,^[2] quantum gravity tests,^[3] secure quantum-key distribution,^[4] phase-contrast microscopy,^[5]

and imaging.^[6,7] Taking advantage from quantum states and quantum correlations, quantum imaging successfully handles the issue of image formation, processing, and detection, thus overcoming the performance of classical light methods.


1.1. Noise Bounds of Classical Imaging

In its simplest way, classical imaging exploits several optical phenomena such as scattering, absorption, and reflection to detect variations in light amplitude and phase, caused by the interaction of light with the sample. Two possible approaches can be used to reconstruct the sample spatial features: point-by-point or wide-field scanning, to be preferred in static high z-resolution (i.e., depth-resolved) images or dynamic pictures, respectively. In any case, the amount of information carried by an image can be defined by two quantities: resolution and sensitivity. Resolution describes the minimum distance at which two points can be distinguished, and, in classical microscopy, it is limited by $R \cong 0.61\lambda/NA$

(Abbe or Rayleigh diffraction limit), where λ is the wavelength of light and NA is the numerical aperture of the imaging system. Sensitivity addresses the minimum measurable variation of the quantity under investigation in a certain point of the image and, exploiting classical light, it is limited by $S_{SNL} \cong 1/\sqrt{n}$ (shot-noise limit), where n is the mean number of photons in the light signal. Such a bound can be theoretically reduced arbitrarily by increasing the number of signal photons n , thus being relevant in circumstances where the optical power is limited, e.g., in the presence of either a limited light source or photosensitive samples prone to photobleaching.

The scaling of the shot-noise limit with the number of input photons n is equivalent to repeating the same measurement n independent times, meaning that classical photons behave independently of each other (i.e., Poissonian behavior). On the contrary, a quantum light source can feature sub-Poissonian fluctuations (e.g. with Fock states or Twin-Beams) or a high cooperation among photons such that the quantum state of each particle strongly depends on the state of the others. These latter states are called entangled states and imply quantum correlation.^[8] By using such non-classical light in microscopy, the image uncertainty scales with the number of photons involved in the correlation, reaching the so-called Heisenberg

F. Madonini, F. Severini, Prof. F. Zappa, Dr. F. Villa
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Via G. Ponzio 34/5, Milano 20133, Italy
E-mail: federica.villa@polimi.it

 The ORCID identification number(s) for the author(s) of this article can be found under <https://doi.org/10.1002/qute.202100005>

© 2021 The Authors. Advanced Quantum Technologies published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

DOI: 10.1002/qute.202100005

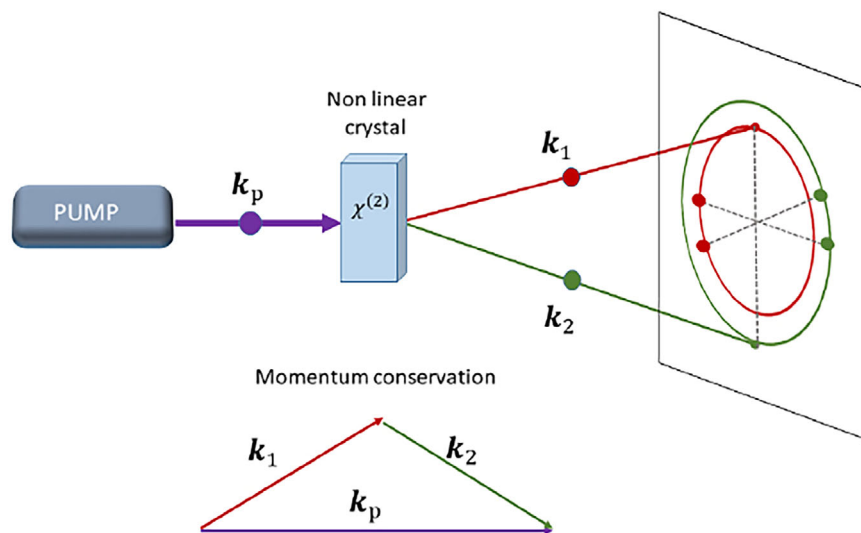


Figure 1. Generation of a TWB by SPDC.^[7] High-frequency photons from a strong pump field traverse a second-order nonlinear crystal giving rise to the emission of photon pairs at lower frequency. Under the laws of momentum and energy conservation, this process generates a pair of entangled photons diametrically opposed on a circle, with diameter function of their wavelength. Reproduced with permission.^[7] Copyright 2019, BIPM & IOP Publishing Ltd.

Limit $S_{HL} \cong 1/n$. Therefore, sensitivity improves by a factor \sqrt{n} with respect to the shot-noise limit or, in other terms, the same boundary can be reached with fewer photons, which is particularly interesting in case of delicate samples.

1.2. Quantum States Sources

The ability to produce entangled states, currently with two photons (such as NOON states with $N = 2$) but with an even higher number in the future, and the availability of single-photon sources and Fock states allowed to demonstrate quantum imaging potential.

The most efficient way to generate a two-photon entangled state is by spontaneous parametric down conversion (SPDC).^[9,10] As shown in **Figure 1**, in SPDC a nonlinear crystal generates a photon pair from a higher frequency photon pump. The produced photon pair at lower frequency obeys the laws of energy and momentum conservation. The stream of photon pairs can be exploited to generate the, so called, twin-beam (TWB) states largely exploited in quantum imaging applications, e.g., while one beam of the pair can interact with the sample, the other is used as a reference for quantum noise. By measuring the correlated (reference) beam, random noise in the probe beam can be known and subtracted, bringing advantages in many applications, ranging from imaging to metrology.^[11] Note that this differential approach is widely used in classical measurement schemes as well, but there only super-Poissonian fluctuations (e.g., thermal noise) becomes correlated, while shot noise remains uncorrelated and therefore cannot be eliminated. Hence, the key feature of TWBs stands in the sub-shot noise (SSN) fluctuation reduction thanks to the non-classical photon number correlation among the two beams.

On the other hand, quantum advantages can be obtained with single-photon sources, which are emitters able to deliver one pho-

ton at a time, upon user request. Whatever mechanism they arise from, their “antibunching” property provides a negligible probability of obtaining two or more photons at the same time,^[12] as represented in **Figure 2**. In principle, n ideal single-photon sources operating together could generate a Fock state source, which should provide on-demand a fixed number of photons indistinguishable from one another. Typically, an external control can put single-photon sources in a state of excitement that will emit a single photon when relaxing to a lower energy state (i.e., radiative decay). Due to imperfect optical out-coupling efficiency and, in some cases, non-negligible nonradiative decay mechanisms, single-photon sources are still far from being totally predictable and deterministic. However, they are already employed in quantum imaging applications, if coupled to photonic structures for enhancing photon collection efficiency.^[13]

1.3. Quantum Imaging Techniques

The antibunching property of single-photon emitters paved the way to super-resolution imaging.^[14] In microscopy applications, single-photon sources, such as nitrogen-vacancy (NV) color centers in diamond, dye molecules, and quantum dots, can be used as fluorophores and markers, preventing the presence of more than one photon at any emission process. Whenever observing a sample, a cluster of more emitters cannot be distinguished through standard intensity acquisitions if their centers are closer than the diffraction limit. Conversely, single-photon coincidences detection can reveal the presence of more than one single-photon source.

Indeed, a k -fold coincidences detection corresponds to the measurement of the Glauber correlation functions up to k -th order (i.e., $g^{(k)}$), which has been demonstrated to bring an improvement of a factor $1/\sqrt{k}$ to the resolution.^[15] Coupling the coincidence information, up to the highest measurable order,

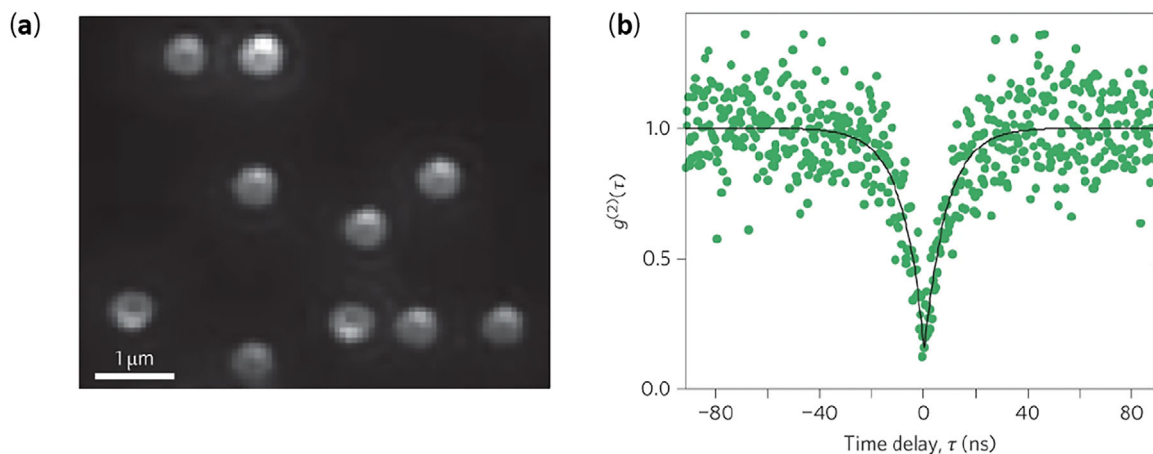


Figure 2. a) Single-molecule single-photon sources. b) Normalized second-order correlation function measurements of a single molecule under a continuous-wave strong excitation, confirming an almost null tendency of photons to arrive in pairs.^[13] Reproduced with permission.^[13] Copyright 2017, Springer Nature.

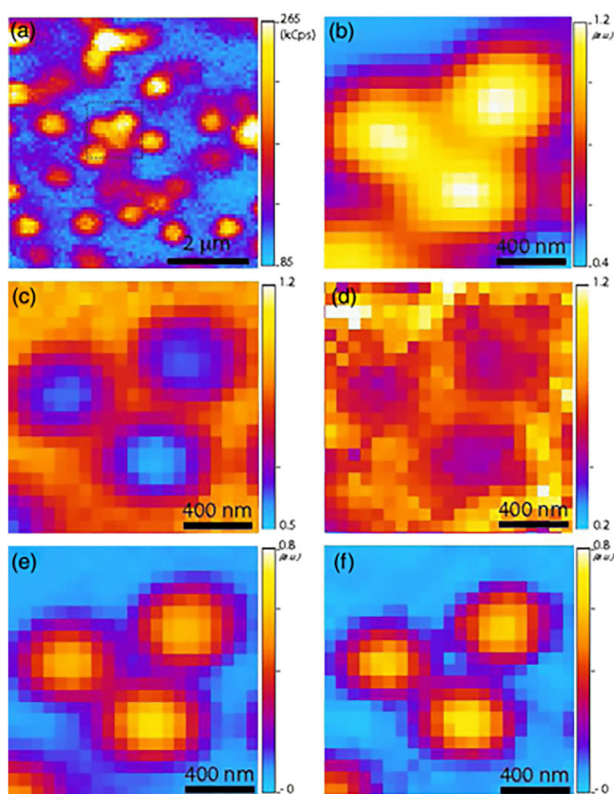


Figure 3. Example of super-resolution imaging applied to a cluster of 3 NV centers.^[15] a) Typical scan on a region of the sample obtained collecting the signals emitted by each center on a pixel-by-pixel basis via single-photon sensitive confocal microscope. b) Magnification of the area of interest. c) Map of $g^{(2)}$ function. d) Map of $g^{(3)}$ function. e) Super-resolved map for $k = 2$. f) Super-resolved map for $k = 3$. Reproduced with permission.^[15] Copyright 2014, American Physical Society.

with the standard intensity map allows to reconstruct super-resolved images, as in **Figure 3**. With the increasing number of emitters in the cluster, significant resolution enhancement can be obtained as long as detectors are able to reveal coinci-

dences up to the highest degree (at most equal to the number of single-photon emitters).

SSN imaging surpasses the shot-noise limit by using sub-Poissonian light, either single modes with sub-Poissonian photon statistics or correlations between modes. A possibility consists in using the two correlated modes of a TWB state in a differential scheme for noise compensation.^[16] Even further, the two-modes case has been extended to multiple spatial modes cases,^[17,18] since the generation of TWBs can be spatially broadband, consisting in a set of pairwise correlated modes in a unique plane. Wide-field SSN imaging can therefore be achieved by coupling multimode TWBs with high-sensitivity multipixels detectors. What happens is that each pair of correlated modes is entirely detected by a pair of pixels and the image noise can be subtracted pixel by pixel.^[19]

SSN photon correlations can also be used for quantum-enhanced displacement sensing of light beams, providing a higher accuracy compared to classical laser beams.^[20] In a 2D realization,^[21] this measurement technique consists of directing the light beam to a detector divided into quadrants, which measure the same average power when the beam experiences no displacement, while they sense an unbalanced signal when the beam moves in one direction, as shown in **Figure 4**. In the classical approach, resolution is limited by shot-noise, which is independent of quadrants. Conversely, a lower limit can be reached if the detector different quadrants are illuminated by spatially correlated beam portions.

Entangled photons generated by SPDC can also be exploited in ghost imaging,^[22,23] improving this technique quality compared to classical light operation. The concept is to separate the signal and the idler SPDC photons paths, then to detect the signal passing through the sample with a single-pixel detector and simultaneously to send the idler to a detector with spatial resolution. By measuring the correlations between the two detectors, it is possible to reconstruct the transmission/reflection profile of the sample. In other words, considering the spatially resolved detector, the pixel in temporal coincidence with the signal beam represents a transmission point of the object. Vice versa, if no coincidence is detected, the signal photon has been

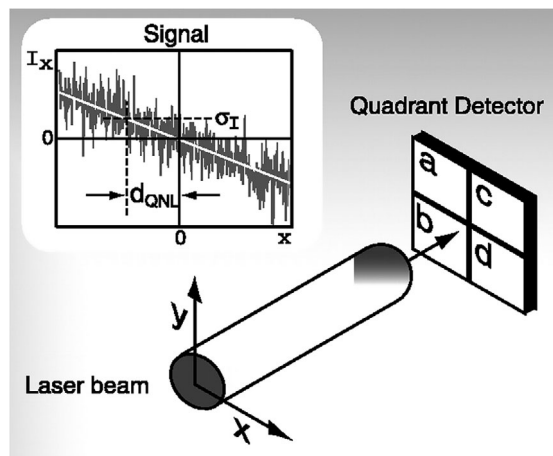


Figure 4. Measurement of laser beam displacement.^[21] For example, signal $I_x = I_a + I_b + I_c + I_d$ is reported, which is proportional to the horizontal direction displacement and can be obtained through simple calculations on the photocurrents I_a, I_b, I_c, I_d . The standard deviation σ_I defines the quantum noise-limited displacement d_{QNL} . Reproduced with permission.^[21] Copyright 2003, The American Association for the Advancement of Science.

absorbed or reflected by the object. As the correlation order increases, visibility and contrast of this technique improve, thanks to the increasingly efficient rejection of the uncorrelated background.^[24] Furthermore, by using quantum correlated pairs of photons, it is possible to illuminate the sample with a wavelength different from that of the idler beam.^[25] Therefore, investigations of objects in the infrared range can be achieved with cameras that operate in the visible region, as exemplified in **Figure 5**.

On the heels of ghost imaging, even more advanced techniques allow the so-called imaging without photon detection,^[26] such as through nonlinear interference in nonlinear crystals.^[27] In this case, due to a precise phase relationship between photons, detecting only the idler makes it possible to obtain the phase in-

formation on the signal passed through the sample, i.e. without signal detection.

The high correlation of a TWB can also be exploited in the quantum illumination method,^[28] experimentally realized for the first time in ref. [29]. In this case, the issue is to reveal a very weak signal reflected by a (potentially present) object surrounded by high thermal background. If one photon of the TWB is addressed to the object and the other is used to make correlation measurements between the two, it is possible to improve the signal-to-noise ratio (SNR). Refer to **Figure 6** for an example.

2. Detectors for Photon Coincidence Detection

In the last decades, several detection methods have been applied in quantum imaging experiments. One of the first imaging exploitations of a signal-idler entangled state generated by SPDC employed two ice dry cooled avalanche photo diodes (APDs), one for the signal beam and one for the idler beam, whose outputs were sent to a coincidence counting circuit with 1.8 ns coincidence windows.^[23] A 2D image (corresponding point-to-point to the sample) was reconstructed by scanning an aperture through a lens in front of the detector on the signal beam path, as shown in **Figure 7**.

Experiments based on the signal-idler quantum beams configuration paved the way for ghost imaging. In such a technique, the image is created by photons that never interact with the sample and that are detected by a pixelated device, while a single-pixel detector is enough to temporarily identify the idler photon (see Section 1.3). When dealing with low-intensity photon signals, sensitivity and time-resolution of pixelated cameras based on charge-coupled devices (CCDs) can be not enough; hence the need of optical or electrical amplification. Besides being significantly faster, intensified CCDs (ICCDs) (i.e., based on optical amplification) and electron-multiplying CCDs (EMCCDs) (i.e., based on electrical amplification) can be used in the photon-counting regime, when there is less than one photon per pixel per readout. A possible ghost imaging method involves the use of

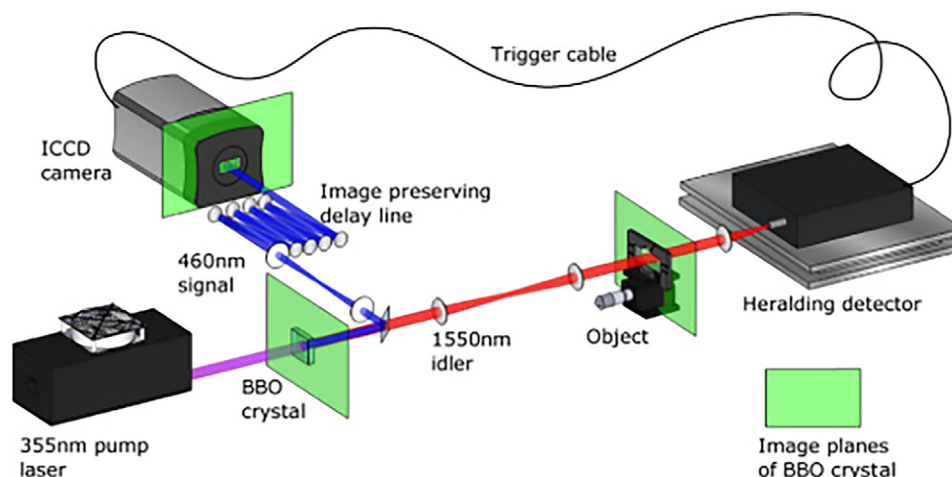


Figure 5. Experimental setup of two-color ghost imaging.^[25] A visible/infrared photon pair is generated at BBO crystal by SPDC and separated by a dichroic mirror. The infrared photon interacts with the object and is detected by a single-element heralding detector (i.e., InGaAs/InP SPAD), while the visible photon is detected by a pixelated camera (i.e., ICCD). In this case, the heralding detection triggers the ICCD, which detects the intentionally delayed correlated visible photon. Reproduced with permission.^[25] Copyright 2015, The Optical Society.

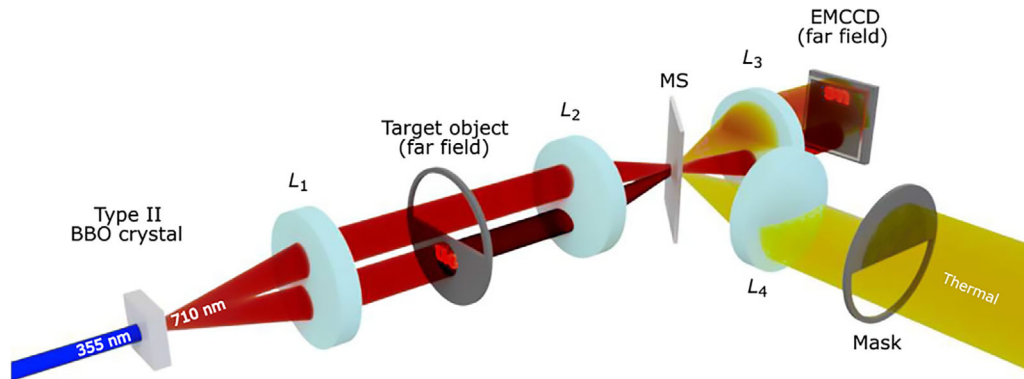


Figure 6. Quantum illumination experimental setup.^[30] A two-photon entangled state is produced at BBO crystal by SPDC and used as probe/reference beams. The probe beam illuminates the object and is consequently overlaid by a thermal light introduced to simulate an environmental background. Measuring pixel-by-pixel correlations between the two regions of the detector receiving either the probe or reference beam allows taking advantage of the quantum illumination. Reproduced with permission.^[30] Copyright 2020, The American Association for the Advancement of Science.

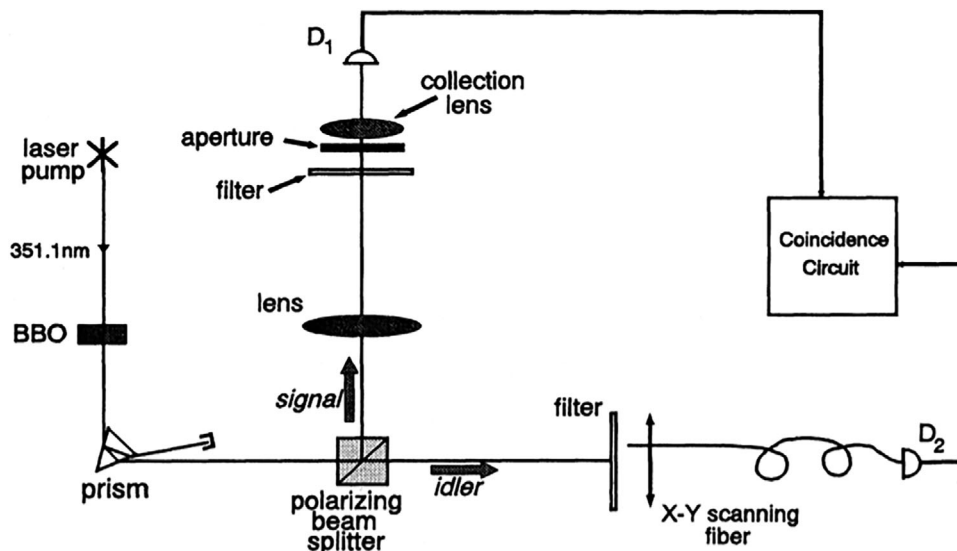


Figure 7. Schematic of the experimental setup exploiting a two-photon SPDC entangled state.^[23] Thanks to a lens, a plane image corresponding to the object (i.e., aperture) can be formed. Reproduced with permission.^[23] Copyright 1995, American Physical Society.

a pixelated ICCD camera, time-gated by the detection of the idler photon by a single-pixel embedding a single photon avalanche diode (SPAD).^[31] An image-preserving optical delay line before the ICCD detector ensures that the photon detected at the idler detector and the one at the camera are from the same correlated photon pair, as clarified in **Figure 8**. In these measurements, the coincidence time-window duration corresponds to the gate duration, in which the spatially resolved detector is enabled after the single-pixel detection. A similar measurement setup employing an ICCD camera can be found in a two-color ghost-imaging method,^[25] where visible/infrared photons pairs are employed, and the ICCD is coupled with an InGaAs/InP SPAD as infrared heralding detector. ICCDs were reported in literature also for experiments quantifying the amount of entanglement of two-photon states,^[32] real-time imaging of any spatial-mode entanglement,^[33] and determination of correlations up to fourth

order of degenerate and nondegenerate photons pairs produced by SPDC.^[34]

Having the drawback of long exposure times in the order of a few μs , EMCCDs have been used in applications where the generation rate of entangled states was reduced to minimize the probability of unpaired photons detection. Some notable examples concern the characterization of the spatial correlations of entangled photon pairs generated by parametric down conversion (PDC) with EMCCD description in terms of capabilities and limitations,^[35] and the entanglement observation of around 2500 spatial states with the demonstration of the Einstein-Podolsky-Rosen type correlation.^[36] Noteworthy is also the exploitation of an EMCCD camera for the split-detection of quantum correlations for the SSN position measurement of a shadow at the single-photon light level.^[37] This quantum displacement sensing experiment showed a sensitivity improvement that scales

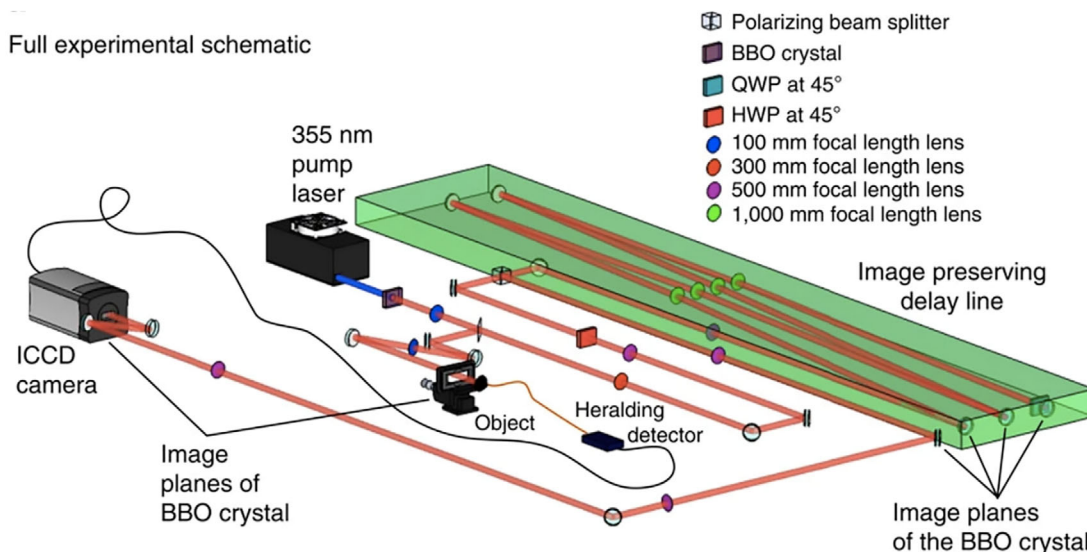


Figure 8. Schematic of a ghost imaging system.^[31] The heralding arm hosts the object, and each photon detection at the heralding detector triggers the ICCD camera. The image-preserving optical delay line is necessary to compensate for the electronic delays in the triggering mechanism. Reproduced with permission.^[31] Copyright 2015, Springer Nature.

with the detector efficiency. At last, EMCCDs have been used in quantum holography imaging to measure the spatial intensity correlations that reconstruct the phase images encoded in the polarization-entanglement of hyper-entangled photons.^[38]

All multipixel detectors operated in single-photon counting regime suffer from classical noise that can reduce or nullify the advantages of quantum states of light. Furthermore, EMCCDs suffer from clock induced charge (CIC), i.e., unwanted electrons generated during the shift process, especially if they are read out at high frame-rate and excess noise, due to the electron-multiplication process.^[39] These two additional noise sources have a deleterious effect on quantum signal acquisition. An EMCCD not operated in single-photon regime can be used for intensity correlation measurements aimed at the distillation of a quantum image starting from measured data composed by the superposition of both quantum and classical light.^[40]

A high quantum efficiency CCD camera is reported in a quantum illumination experiment, where the probe beam of a correlated pair may be partially reflected by the target object toward the camera, which also receives a high unknown thermal background noise.^[41] That implementation involved only photon-counting measurements of the second-order correlation between the probe and the reference beams, and the quantum superiority was demonstrated against the classical protocol. CCDs can also be exploited to develop an SSN wide-field microscope, exploiting thousands of spatial modes detected independently by the same number of pixels (i.e., each pixel detects thousands of photons per exposure window).^[42] Dynamic imaging was demonstrated with a noise reduction of 20% below the shot noise per pixel with 5 μm resolution in an array of about 8000 pixels.

The detection of nonclassical light generated by SPDC has also been carried out by a hybrid photo detector (HPD), which was proved to be capable of coincidence acquisition, thus opening the way to further applications in quantum optics.^[43] The HPD camera used in those experiments, whose cross-section is rep-

resented in **Figure 9**, consisted of a microchannel plate (MCP) tube having as anode readout a 2×2 array of photon-counting detectors “Timepix,” which are 256×256 pixelated application specific integrated circuit (ASIC), implemented in a complementary metal-oxide-semiconductor (CMOS) technology. Each pixel (55 μm size) contains a logic that, in correlation measurement mode, provides the arrival time of the electron pulses from the MCP, with a resolution of about 10 ns. Besides providing both space and time resolution, the HPD has the drawback of long readout times and consequently low frame rates.

3. SPADs for Photon Coincidence Detection

When very faint light signals must be detected, the forefront candidates are SPAD detectors. Compared to other single-photon sensitive devices, like photo multiplier tubes (PMTs) or MCPs, SPADs are smaller, insensitive to external magnetic fields, more reliable, operated at lower bias voltages, and can be integrated with CMOS electronics.^[44] Their relatively high photo-detection probability (PDP), along with the digital-like output and the absence of readout noise, makes them an arguably suitable choice for applications like single-photon coincidence detection.

Arrays made of a multiplicity of SPADs are the best candidates as multipixel sensors to detect photons correlations, since they allow single-photon sensitivity together with high frame rates.^[45] CMOS active pixel sensors (APS) are high-speed devices, but they do not perform internal amplification, thus suffering from low sensitivity. EMCCDs and ICCDs, although having high sensitivity and possibly millions of pixels, result to be bulkier and more expensive than SPAD arrays.

In literature, there are two main approaches to read out a multiplicity of SPAD pixels: either fully digital SPAD arrays or fully analog silicon photomultipliers (SiPMs). As depicted in **Figure 10**, a SPAD array is a multiplicity of completely independent SPAD pixels. In each pixel, the SPAD is integrated with

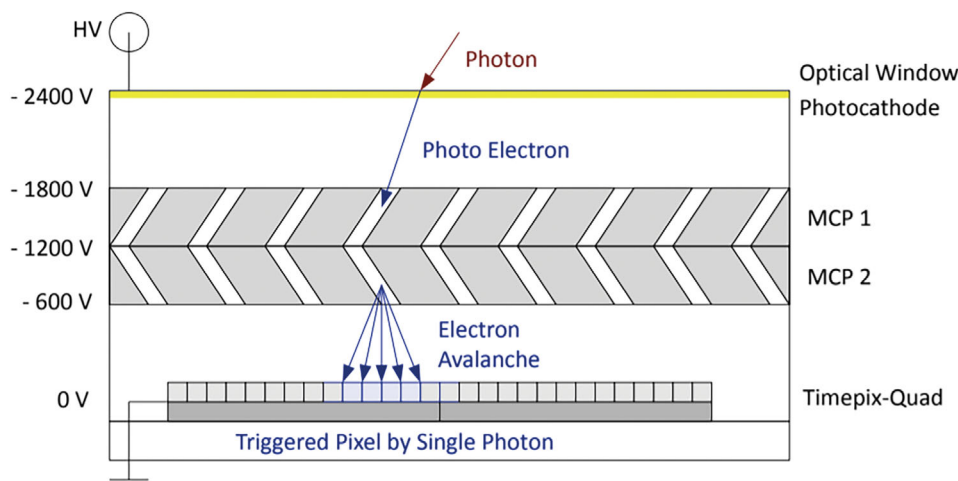


Figure 9. Cross-section of the hybrid photodetector tube.^[43] Below the optically transparent window and the photocathode, which emits a photo-electron when an optical photon is absorbed, there are two MCPs, one onto the other. Due to the applied voltages, multiple secondary electrons are emitted (about 10^6 from a single microchannel), which are then collected by the pixels of the Timepix-ASIC. Reproduced with permission.^[43] Copyright 2014, The Optical Society.

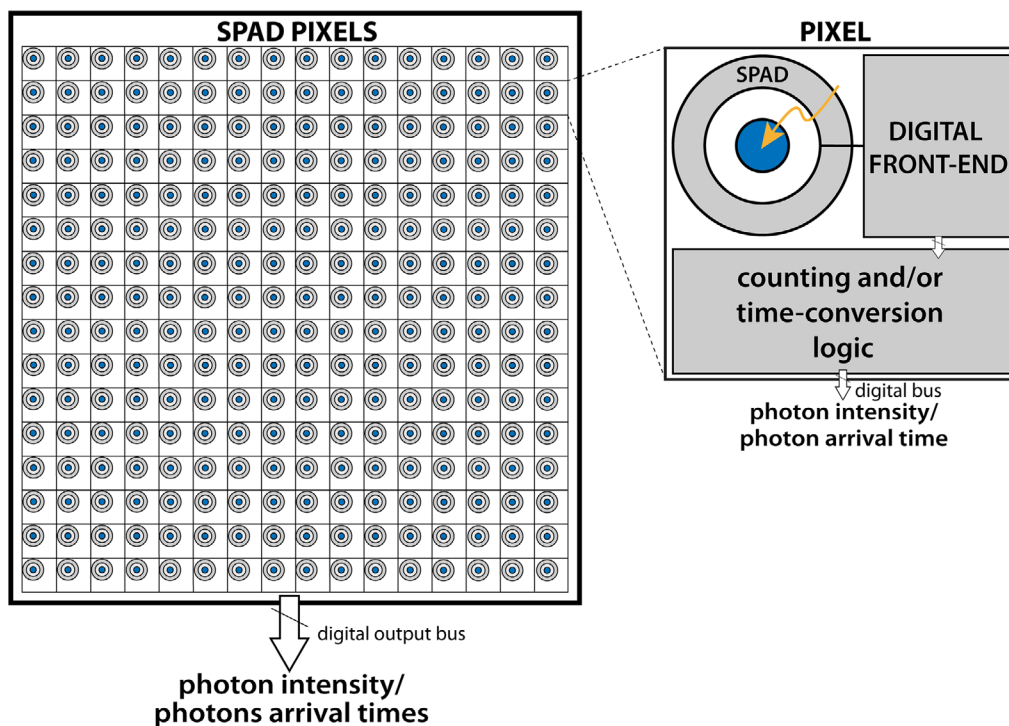


Figure 10. Typical architecture of a SPAD array, based on a multiplicity of pixels, each one with SPAD, digital front-end, and in-pixel electronics for either counting photons or measuring photon arrival time. The stored information is then transferred off-chip through a digital bus.

the digital front-end circuitry, and eventually with additional electronics to count the number of incoming photons or to time stamp their arrival time (i.e., so to provide the time-of-flight of the photon). The SPAD pixel output is fully digitally processed, preserving the advantage of SPADs to be immune to readout noise and effectively providing either the intensity (photon counting) or time-of-flight (photon timing) information.

The SPAD array is typically read out through a row-by-column scanning of all pixels, thus preserving the spatial $x-y$ information, i.e., the position of where photons got detected across the array. Fully digital SPAD arrays can hardly detect signal photon coincidences on-chip: typically, this is solved by integrating the timing electronics in each pixel (to time-stamp each photon arrival time) and then by off-chip postprocessing of the time

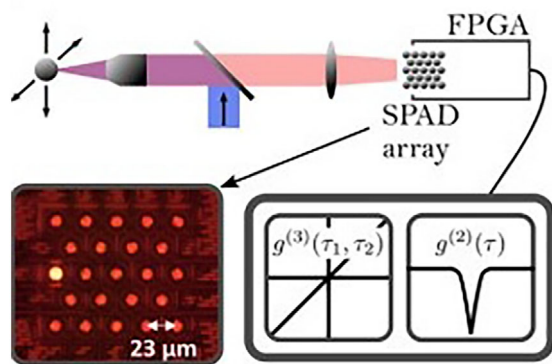


Figure 11. Photon correlation measurement setup.^[52] A 23 SPADs array is placed at the image plane of a scanning confocal microscope, measuring the second and third-order correlation functions (i.e., $g^{(2)}$ and $g^{(3)}$). Reproduced with permission.^[52] Copyright 2019, The Optical Society.

tags of all photons, so to identify coincident pixels with the same time-stamp. Therefore, quite complex on-chip timing electronics is required in each pixel, thus impairing the array fill-factor, when considering front-side illuminated sensors in standard planar technologies. This limitation is mitigated by stacked backside illuminated (BSI) sensors. In the simplest version, a tier implemented in a CMOS image sensor (CIS) process houses BSI SPADs, while a second tier integrates all the processing electronics, taking advantage of low-power and scaled down technologies. Recent progress in BSI SPADs have been reported.^[46–48] Furthermore, also off-chip data postprocessing is mandatory. When this approach was exploited for recording the spatial correlation function of a flux of SPDC entangled photons, high SNR measurements were achieved with the drawback of long data acquisition time,^[49] possibly solved by adding on-chip smart processing to increase the frame rate.^[50] The potential of SPAD cameras in terms of short data acquisition time has been also demonstrated in a work where a two-photons state entanglement dimensionality of 48 was certified in just 140 ns.^[51] Note that measuring these spatial joint probability distributions with a CCD camera could require many hours instead. As an interesting example, shown in **Figure 11**, SPAD arrays have also been reported to detect second and third-order photon correlations from a single quantum dot emitter, probing the quantum image scanning microscopy method.^[52]

A SiPM is an analog detector constituted by a multiplicity of microcells that behaves as a single detector, with just two pins (such as the anode and the cathode of a photodiode), so that spatial resolution of where photons hit the SiPM active area is lost. As shown in **Figure 12**, each microcell includes a SPAD and a resistor with one terminal connected to the SPAD and the other terminal connected to the common node, which acts as a single shared adder node for all SPAD avalanche currents. The SiPM output is the time-dependent analog sum of the currents of all microcells that got hit by at least one photon. Thus, it is not possible to know which SPADs got triggered and generated the current. In this way, the SiPM can be considered as a single-pixel detector, composed of many microcells and characterized by only one output, not spatially resolved. Unlike a single SPAD, a SiPM provides both a much larger active area (the sum of all SPADs

active areas), and a photon number-resolved capability, i.e., the ability to inform about how many photons (separately or concurrently) triggered different microcells. In fact, by measuring the amplitude of the output analog current, it is possible to know the number of photons impinging in coincidence (i.e., almost at the same time) on the device. Indeed, a SiPM is an event-driven detector, which provides an analog current every time one or more SPADs get triggered, resulting particularly suitable to detect photon coincidences. The main drawbacks of a SiPM are its analog nature, which makes SiPMs prone to readout noise, and the loss of spatial information across the multiplicity of pixels. Although the SiPM architecture is highly scalable, the increase of microcells severely limits device bandwidth and photon timing precision, because microcells' stray capacitances overload the common node and delay signal propagation across microcells.

The digital SiPM (dSiPM) is a subcategory of SiPMs that is a niche product in respect to the widespread analog SiPMs. Apart from the SPAD, the microcell of a dSiPM includes a digital front-end instead of the resistor, providing device immunity against readout noise. Each microcell digital output is fed to a digital adder node (typically an OR logic gate), which generates a digital pulse synchronous with the first detected photon. Just like its analog counterpart, the dSiPM behaves as an ensemble sensor that does not provide any spatial information about which SPADs got triggered. Conversely to analog SiPMs, dSiPM devices are not suited to signal photon coincidences because the output pulse is generated synchronously with the first photon only, thus no further information on the number of coincident photons is provided.

Whenever big sized detectors are required, for instance, in positron emission tomography (PET) applications, arrays of SiPMs can be exploited. They are spatially resolved at the pixel level, where each pixel consists of a single SiPM, with large active area and photon-counting capability, eventually with high counting range.

4. SPAD Requirements for Quantum Imaging and Microscopy

Section 3 displayed SPAD sensors superiority in the detection of quantum entanglement in imaging and microscopy applications and introduced the most generic SPAD array readout approaches. The present section drills down on the specific requirements that should be met for maximizing photons correlations detection.

Indeed, despite the peculiar measurement asset of each quantum imaging technique, they all rely on detecting temporal correlations of photons. To construct a wide-field image, each entangled state made by two or more photons can give the information on a point of the image, provided that the detection maintains the spatial information of where photons got detected. In other words, it is not enough to identify an occurred coincidence, but it is necessary to be able to reconstruct where that coincidence took place. Therefore, a pixelated detector is needed, and it goes without saying that the higher the number of pixels for a given area, the better the spatial resolution in the image formation.

Generating pairs of entangled photons is now a routine process using a wide variety of technological approaches, among

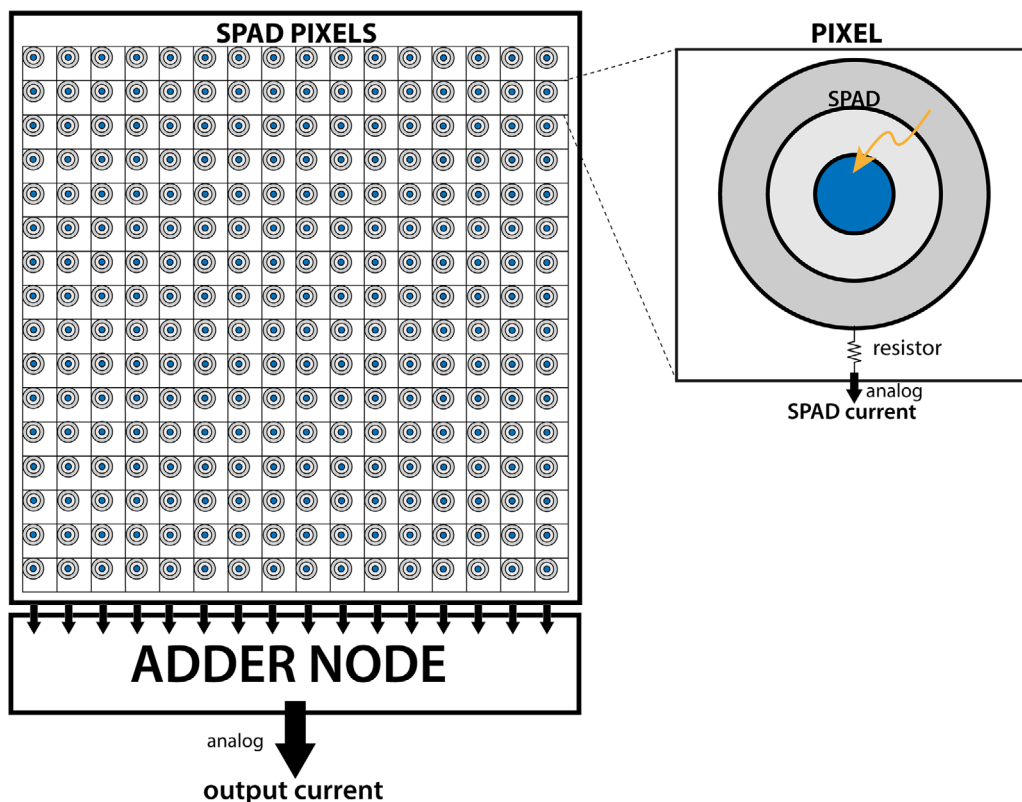


Figure 12. Typical architecture of a SiPM: an array of microcells with a SPAD and a quenching resistor each, characterized by a single analog output with the sum of all pixels current.

which SPDC represents the gold standard in terms of fiber coupling efficiency, entangled photon pair rates, and entanglement fidelity. Conversely, producing entangled states with many photons (i.e., above four) represents a current challenge. Thus, being the number of entangled photons in each measurement very small compared to the total pixels, the coincidence detection should be followed by an event-driven readout optimized to transfer ideally only the positions of pixels detecting coincident photons. In short, only an occurred coincidence should activate the readout procedure, transferring only useful information and avoiding the storage of meaningless data. Indeed, in quantum imaging applications, a fixed-frame rate readout of all pixels is a totally inefficient solution, requiring extremely long exposure times and impacting negatively on power consumption, processing capability, and storage memory availability.

Whatever solution is chosen to reveal coincidences and externally transfer their content, it must be scalable in terms of number of pixels. That is, the architecture should not be impaired by capacitive loads, poor time-resolution, high number of output pads, slow readout, and low frame rate when the number of pixels increases.

In large arrays with small pixel pitch, necessary for high-resolution wide-field imaging, the probability of crosstalk phenomenon becomes significant. Crosstalk shows up whenever a photon absorbed in a pixel triggers a detection also into a neighboring one. It can be virtually instantaneous with respect to the original signal, thus distinguishing between the signal photon and the crosstalk event is difficult. In case of low sig-

nal intensity, as in entangled photons experiments, crosstalk can completely hide useful information, mixing up with the real coincidence events and invalidating the measurement if coupled to low detection efficiency. Special isolation techniques among pixels, such as deep trench isolations (DTI), can be useful in decreasing such an issue, and postprocessing algorithms can help remove the dominant contributions, as reported in Section 6.3.

Even after crosstalk reduction expedients, signal correlations can still be hidden by accidental coincidences due to single photons being detected simultaneously. The quantum source could erroneously generate these singles, or they could come from detection losses of the entangled photons, or they could be dark count rate (DCR) events. The latter consists of intrinsic generations of carriers within the detector in absence of illumination, and it cannot be distinguished from useful photon detections, when light is impinging on the device. DCR has a Poissonian statistical behavior, and it is usually the most relevant noise contribution. In large-size arrays, a significant percentage of pixels could show a DCR much higher than the median DCR of the array, sometimes even exceeding the typical detected signal. This particularly high rate can be due to the presence of local defects in the device, compromising its activity. Such pixels are called “hot pixels” and have a negative impact on the surrounding ones as well, due to crosstalk mechanisms. Especially in presence of a low-intensity signal, it is mandatory to minimize the DCR effect, for instance by employing a high-quality fabrication technology that minimizes local defects

Table 1. Main characteristics of the considered SPAD detectors (N.A. indicates a not available information).

Detector name and reference	Category	Size [total sensitive area or #pixel]	# SPADs / pixel	Pixel fill-factor [%]	Peak PDP [% @ λ nm]	DCR [cps μm^{-2}]
CMOS analog SiPM ^[53]	Analog SiPM	3 × 3 mm ²	N.A.	75	48 @ 420	0.025
Philips' dSiPM ^[54]	Digital SiPM	3.8 × 3.3 mm ²	128 × 64	50	30 @ 450	0.19
MD-SiPM ^[55,56]	Array of digital SiPM	2 × 2 pixels	64 × 64	46	N.A.	N.A.
Linear dSiPM ^[57]	Array of digital SiPM	32 × 1 pixels	12	70	N.A.	6
SPADnet-I sensor ^[49,58]	Array of digital SiPM	8 × 16 pixels	720	42.6	19 @ 450	0.38
Direct TOF sensor ^[59]	Array of digital SiPM	64 × 64 pixels	8	26.5	N.A.	57
MONDO detector ^[60]	Array of digital SiPM	16 × 8 pixels	30	32.1	28 @ 500	1.5
Gated SPAD imager ^[63,64]	Array of SPADs	16 × 16 macropixels	4	9.6	60 @ 500	2.5
Row/frame skipping array ^[50]	Array of SPAD	32 × 32 pixels	1	19.48	N.A.	N.A.
Direct output SPAD array ^[65]	Array of SPADs	64 × 48 pixels	1	N.A.	26 @ 550	N.A.
AER detector ^[66]	Array of SPADs	20 × 20 pixels	1	20	N.A.	0.008
Decision tree imager ^[67]	Array of SPADs	8 × 32 pixels	1	31.3	31.8 @ 600	55.4
Coincidence tree imager ^[68]	Array of SPADs	256 × 128	1	N.A.	N.A.	N.A.

and thus thermally generated carriers. Eventually, in large-size detectors, it ought to be possible to selectively turn-off hot pixels.

The detection of temporally correlated photons can be fulfilled either by off-chip postprocessing of all pixels time tags (if each pixel is equipped with proper timing electronics) or through an on-chip analog mechanism based on the SiPM concept of sensing a current at an adder node common to all pixels. In the former case, the coincidence window duration is defined by the time resolution of the in-pixel timing electronics. Instead, in the latter case time resolution is mostly affected by the stray capacitance at the adder node that defines the sharpness of current edges, as well as by the arrival time delays between the closest and the furthest pixels to the adder node. Those parasitic effects can result in effective coincidence time windows in the order of a few nanoseconds. Since entangled SPDC photons are produced at the same instant, coincidences are expected with small time differences in the order of 500 fs, due to the coherence time of entangled photons. Therefore, a much longer coincidence time window directly leads to a higher probability of false coincidences (i.e., the probability of detecting more single photons within a certain time window, instead of true entangled photons). This is why particular attention should be paid in keeping the rate of unwelcome single-photon rate under control.

Finally, given the described noise sources, it is crucial to minimize the losses of the entangled photons signal. Therefore, further requirements for an ideal quantum imaging detector are high PDP, defined as the ratio between detected photons and incident photons across the active area, and high fill-factor, defined as the ratio of photoactive area compared to the total pixel area. The optimization of both these two parameters maximizes the photon detection over the entire detector area.

5. Review of SPAD Based Sensors

In this section, we summarize the main performance of some selected SPAD-based sensors (analog and digital SiPMs and SPAD arrays), which implement characteristics that can be potentially useful for quantum microscopy and imaging detec-

tors. While some of these detectors have been specifically designed for quantum applications, others aimed at different applications (such as high-energy-ray detection or light detection and ranging—LiDAR), though still presenting some distinctive features that can inspire a next-generation quantum imaging detector.

Table 1 shows at a glance the main figures of merit of the reviewed SPAD detector, namely: category (analog/digital SiPM, SiPM array, SPAD array), size (i.e., total sensitive area for SiPMs and number of pixels for arrays either of SPADs or SiPMs), number of SPADs per pixel (note that for SiPMs the number of microcells is reported, since a SiPM can be considered a single pixel), fill-factor, peak PDP and corresponding wavelength (λ), and DCR at room temperature per unity of area (in order to fairly compare detectors with different sizes). The following subsections briefly describe the architectures and main features of each detector.

5.1. CMOS Analog SiPM^[53]

The CMOS analog SiPM is a C-series analog SiPM developed by SensL, designed in an optimized CMOS process for SiPMs, with integrated comparator and time-to-digital converter (TDC). The CMOS analog SiPM features 48% PDP at 420 nm wavelength, 75% overall sensor fill factor, 3 × 3 mm² sensitive area, and 300 kcps overall DCR. The SiPM has been designed to optimize the timing performance, implementing the fast-output concept, which consists of a capacitive decoupling at SPAD level for promptly reading and processing each microcell output current. The SiPM fast output has a parasitic capacitance of about 12 pF and it is directly coupled to the comparator (with externally controlled threshold voltage), composed of two preamplifier stages followed by a complementary self-biased differential amplifier stage. The comparator output triggers the starting signal of the on-chip TDC. In this way, only if the incoming photon flux (i.e., coincident photons within a fixed temporal window) exceeds a user adjustable threshold, the timing information can be readout. Given the shape and duration of the microcell output current, the

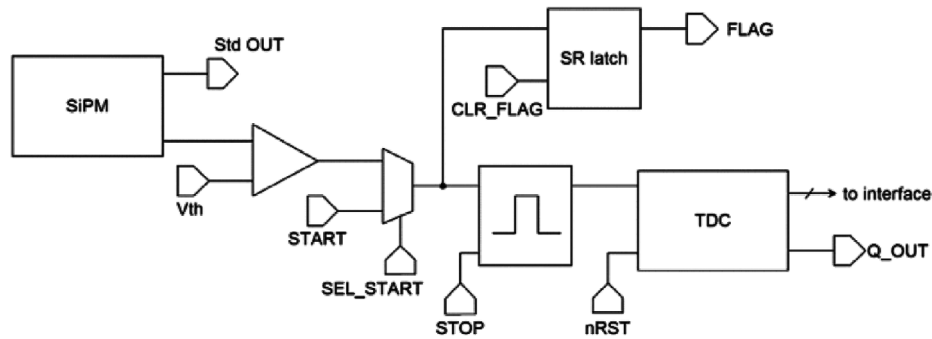


Figure 13. Block diagram of the CMOS Analog SiPM architecture by SensL.^[53] Reproduced with permission.^[53] Copyright 2018, IEEE.

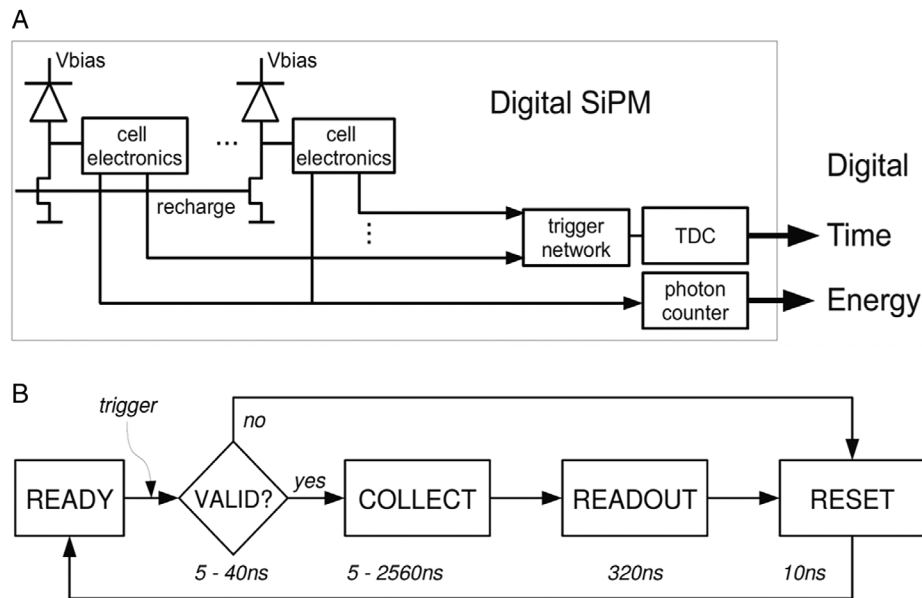


Figure 14. Philips' dSiPM implementation. a) Block diagram of the chip architecture (note the two outputs of the pixel, the fast trigger and the synchronous output data signal). b) Detector state machine implemented in FPGA.^[54] Reproduced with permission.^[54] Copyright 2009, IEEE.

coincidence window is shorter than 5 ns. An output flag is raised every time a TDC conversion is triggered, to communicate that data are ready to be read. Beyond the timing information, also the SiPM output current (standard output) can be readout to retrieve photon intensity (counting) information. The architecture of the CMOS analog SiPM is shown with a block diagram in **Figure 13**.

5.2. Philips' dSiPM^[54]

The Philips' dSiPM is one of the first implementations of a fully digital SiPM in a standard CMOS process. It was mainly designed for time-of-flight PET applications, in which it is required to measure the arrival time of the scintillation event and its energy level. The SiPM fill-factor is 50%, the peak PDP is about 30% at 450 nm and the average DCR of each microcell is 150 cps at room temperature. The block diagram of the Philips' dSiPM is shown in **Figure 14a**. Beyond the SPAD, the SiPM microcell includes sensing, active quenching, and recharge circuitry. Upon a photon detection, the microcell provides a fast-asynchronous

trigger signal and a slower synchronous data output signal. A balanced trigger network is used to propagate the trigger signal from all cells to the on-chip TDC, only one for the entire array. To reduce the number of triggers in the case of high DCR, the trigger network can be configured to start the TDC at the first photon or, alternatively, at higher photon levels (up to 4 photons). These thresholds are implemented in a fully digital way, by using logic operations on sub-arrays trigger lines. In particular, being the SiPM divided into four subarrays and assuming that each sub-array has a first-photon trigger signal, the trigger network can be configured to start the TDC if up to 4 among these trigger signals are present in a 10 to 15 ns observation window. A second higher threshold is used to validate the measurement and start the collection and readout phases. In particular, the measurement is validated only if all the 64 regions, in which the SiPM is divided, detected at least one photon, within an adjustable (5–40 ns) coincidence window. The double threshold is implemented for storing the timing information at low threshold (better timing precision) and validate it at a higher threshold, so to be more immune to dark counts. The collection time is user-defined

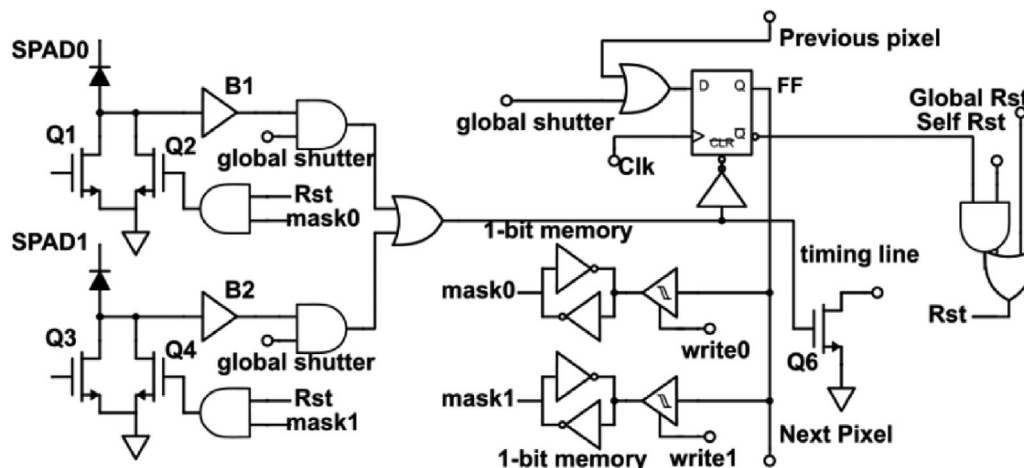


Figure 15. Pixel of the MD-SiPM, with the timing output and the 1-bit memory (FF) storing the firing information.^[56] Reproduced with permission.^[56] Copyright 2018, IEEE.

between 5 and 2560 ns, during which each microcell detects and stores the incoming photons. The microcell synchronous data output signal is sent to a line shared among the microcells of the same column. During readout, each line of the sensor is selected separately, and the number of photons detected in the line is added through a photon accumulator. The outputs of the SiPM are the timing information provided by the TDC, and the energy level, provided by the accumulator. The SiPM operation phases (validation, collection, readout and reset) are controlled by a state machine, implemented in a field programmable gate array (FPGA) and shown in Figure 14b.

5.3. MD-SiPM^[55,56]

The MD-SiPM is a multipurpose monolithic array of 2×2 multichannel digital SiPMs designed to indirectly capture gamma events in PET. Each MD-SiPM comprises 64×64 dual pixels (two SPADs in each pixel) connected to 128 low-power TDCs. In fact, each array column is connected to two TDCs to timestamp the photons arrival time, while the in-pixel circuit stores the fired SPAD information in a 1-bit memory. **Figure 15** shows the pixel circuit, with the timing output connected to the common column line and the 1-bit memory. The sensor can operate in both frame-based and event-driven mode. In the former one, the array is readout at a fixed 192 kfps frame rate, providing the TDCs timestamps and the firing map. In event-driven mode, the shutter is opened indefinitely until an event occurs. In order to detect the event, two (high and low) thresholds are used. The event is defined by the ratio of photons per unit of time, which is an externally configurable parameter through the high threshold, depending on the application. If this condition is not achieved (no event), but the lower threshold is exceeded because of dark counts, the system recovers the fired pixels and corresponding TDCs. Instead, if an event is detected, the digital core closes the shutter after the predefined integration time and gets ready to output the information, which consists of the TDCs conversions and the sum of fired pixels (with no spatial information), within a 2 μ s total readout time.

5.4. Linear dSiPM^[57]

The linear dSiPM is a linear array of dSiPM designed for time-of-flight distance ranging. The chip comprises 32 pixels, each consisting of 12 SPADs, a coincidence detection circuit, one shared TDC, and one associated circuitry to sample the number of involved photon events. The pixel reaches a very high fill-factor of 70%. In order to get rid of background incident light, the 12 front-end output signals within each pixel feed the coincidence detection circuit (**Figure 16**), which generates a trigger signal for the corresponding TDC whenever two or more photons are detected within 4 or 8 ns time windows. Such coincidence detection circuit is fully digital and includes four stages, made of full-adders, half-adders, and other standard logic gates. The array is readout at a maximum fixed frame rate of 12.5 Mfps. The final resolution of 340×96 pixels mentioned in the paper title is achieved by means of laser scanning.

5.5. SPADnet-I Sensor^[49,58]

The SPADnet-I sensor is an 8×16 pixels digital SiPM allowing photon counting and per-pixel time stamping of incoming photons with 265 ps resolution. Each pixel contains 720 SPADs, individually enabled/disabled, with 16.87 μ m diameter, the electronics required to count photons and two 12-bit TDCs each with 65 ps nominal time resolution. **Figure 17** shows the chip architecture. The peak PDP is 19% at 450 nm with 42.6% fill-factor and 7.9 Mcps overall pixel DCR.

The SiPM pixel implements the so-called “spatial compression” and “temporal compression”. “Spatial compression” consists in OR-ing together 3 SPADs, which is equivalent to having one single SPAD with three times the area of a single one, but better yield (in terms of defects that cause high DCR). “Temporal compression” consists in reducing the SPAD pulse duration down to 250 ps, in order to reduce the count losses in the OR tree that convey the outputs of all SPADs in a same pixel to the adder node. The sensor is synchronous with a global clock signal that can operate up to 100 MHz. For every clock bin, each

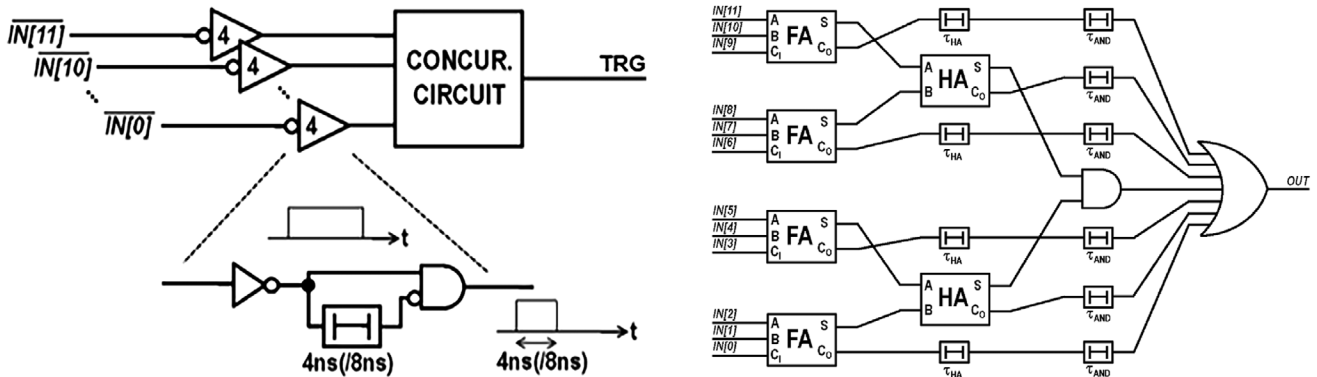


Figure 16. Coincidence detection circuit of the linear dSiPM, which combines the 12 outputs of the SPAD front-end circuits of the pixel to detect coincidences within 4 ns or 8 ns time windows.^[57] Reproduced with permission.^[57] Copyright 2013, IEEE.

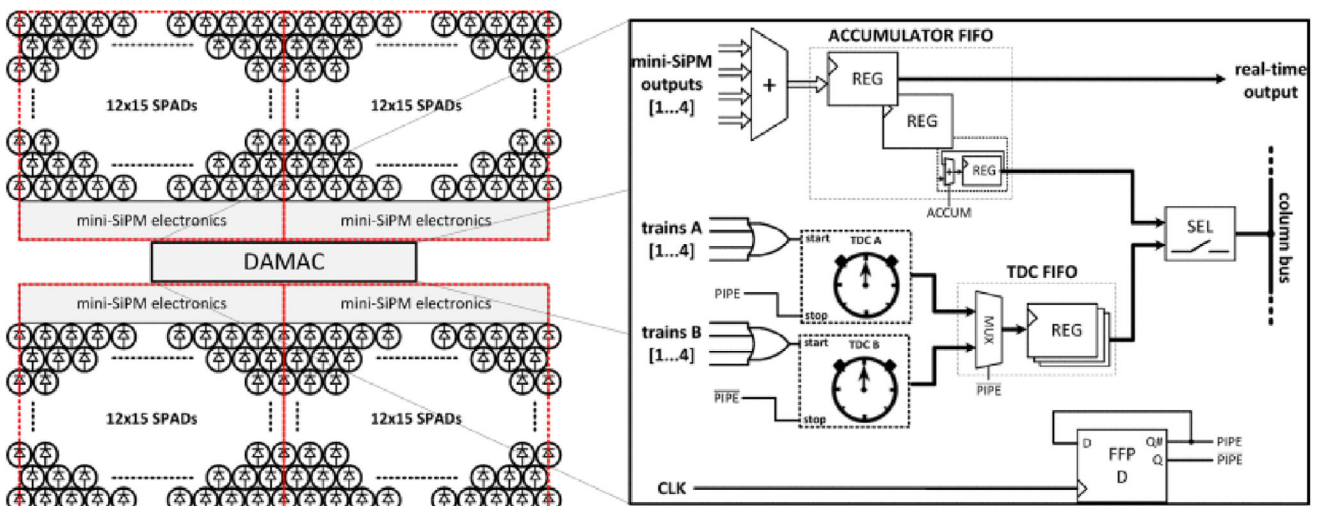


Figure 17. Block diagram of the SPADnet-I sensor architecture.^[58] Reproduced with permission.^[58] Copyright 2014, IEEE.

pixel generates a photon count (number of SPADs triggered in the current bin) and one timestamp of the first photon detected in the bin. At the same rate, a distributed network of adders computes the number of photons detected globally by the array. By means of a discrimination logic, if the derivative of this number exceeds a configurable double threshold, the integration phase (with user-adjustable duration) is started. A double threshold is used: the lower threshold (close to noise level) ensures to retain the timestamps of the very first photons of the event, which are the most important for a precise timing information; whereas the higher threshold allows to clearly distinguish between the desired event and dark counts. At the end of the integration phase the sensor is readout in 2.4 μ s.

The SPADnet-I sensor has been designed for PET applications,^[58] but has also been employed to measure a second-order correlation function for various non-collinear configurations of entangled photons.^[49] In the latter application, the discrimination logic is not used, and the array is readout at a fixed frame rate, generating 8×16 maps of photon timestamps at up to 500 kfps. These timestamps can be used in post-processing to detect the addresses of photon coincidence events.

5.6. Direct TOF Sensor^[59]

The direct TOF sensor is a 64×64 -pixel dSiPM array, for long-range time-of-flight applications. Each pixel, shown in **Figure 18**, includes 8 SPADs combined as a dSiPM, a triggering logic for photons temporal correlation, a 250 ps 16-bits TDC, and an intensity counter. The pixel fill factor is 26.5% and the DCR of a single SPAD is about 6.8 kfcs. Each of the 8 SPADs, connected to its respective front-end circuitry, feeds a monostable to shorten the pulse down to 260 ps and then an OR-tree. The resulting stream of pulses from the OR-tree is then processed by a fully digital triggering logic that identifies if the number of incoming photons exceeds a certain threshold within a rolling time window. This threshold depends on the settings of the number of correlated photons. If the condition is verified, a trigger validation signal is generated; otherwise, the whole pixel is reset and waits for another SPAD pulse. Eventually, the data processed and stored by the counter and the TDC are, respectively, the number of photons detected within the time window and the first photon arrival time. Counting and timing information is readout at a fixed frame rate of about 18 kfps.

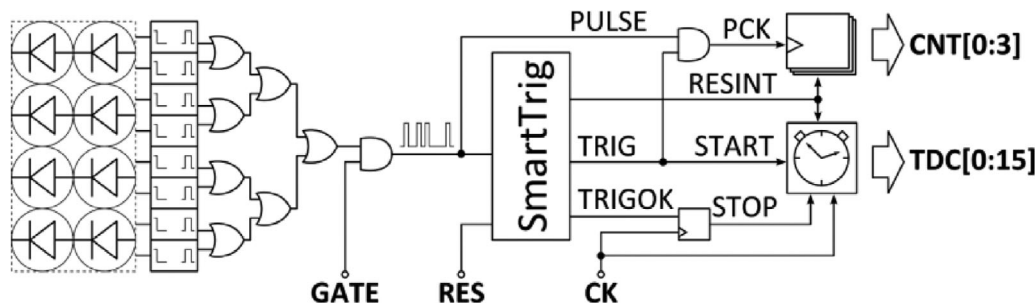


Figure 18. Block diagram of the direct TOF sensor pixel.^[59] Reproduced with permission.^[59] Copyright 2017, IEEE.

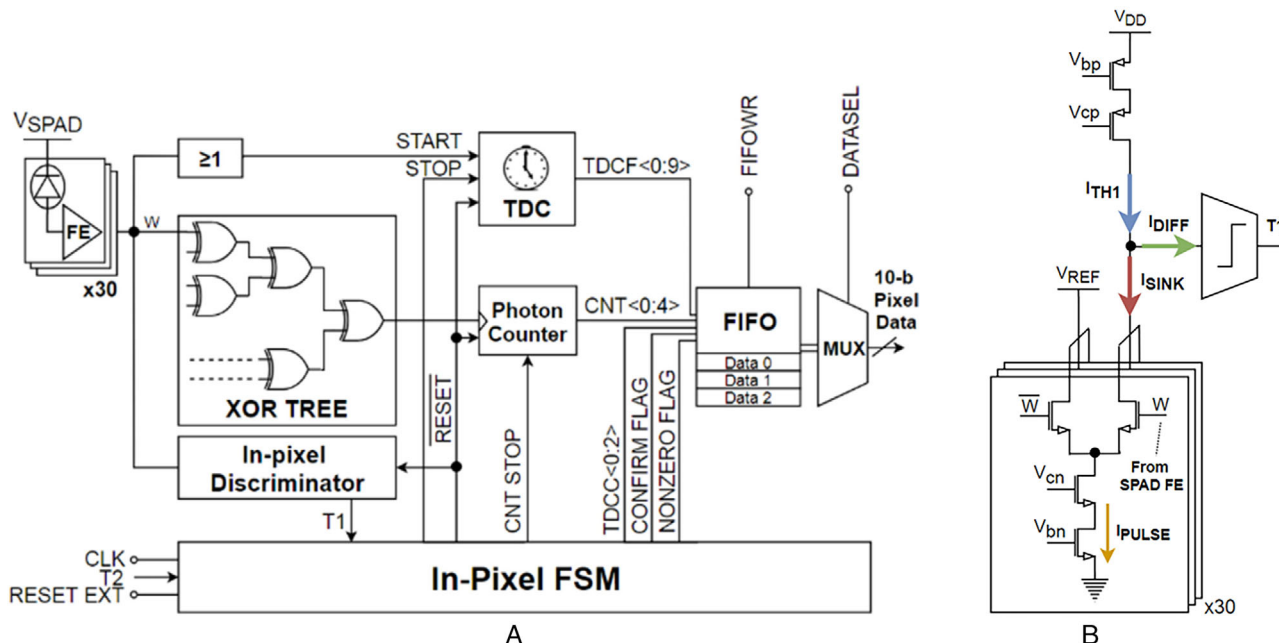


Figure 19. MONDO detector: a) Block diagram of the pixel and b) analog current-based discriminator.^[60] Reproduced with permission.^[60] Copyright 2019, IEEE.

5.7. MONDO Detector^[60]

The MONDO detector is an array of 16×8 dSiPMs, developed within the MONDO project to track secondary neutrons produced in particle therapy treatments.^[61] In respect to PET applications this sensor is designed to detect a low number of photons (<10) with a decay of few nanoseconds. Each pixel includes 30 SPADs with 500 cps DCR and 28% peak PDP at 500 nm.^[62] The pixel fill-factor is 32.1%. The block diagram of the pixel is shown in **Figure 19a**. The SPADs output edges are combined through an XOR-tree, along which different delay taps have been added to spread the signals generated by simultaneous photons. This approach allows to distinguish up to four simultaneous photons, while a standard XOR-tree implementation would reveal all of them as one. The XOR-tree output feeds a counter (for intensity measurement), and the first edge is timestamped with an in-pixel TDC. An analog current-based discriminator block (**Figure 19b**) generates a trigger T1, when the photons detected within the pixel during a $5 \div 10$ ns coincidence window exceed a certain threshold Th1. At the top-level all the triggers T1 are combined

through the XOR-tree, and a global discriminator generates a trigger T2, if the number of pixels that detected an event exceeds a second threshold Th2 within a few nanoseconds. When an event is detected, counters and TDCs contents are stored in the in-pixel memory and consequently readout.

5.8. Gated SPAD Imager^[63,64]

The Gated SPAD Imager is a time-gated array of 32×32 SPADs (16×16 macropixels), for photon timing, counting, and coincidence detection. The macropixel, whose architecture is shown in **Figure 20**, includes four separate SPADs with independent active time-gating and quenching circuit, a shared TDC, and four independent photon counters. The TDC is driven by a smart arbitration logic, which preserves spatial information among the four micropixel detectors and, in each frame, it converts the arrival time of the first photon detected in the macropixel. The fact that the spatial information within the macropixel is preserved distinguishes this SPAD imager from other SiPM arrays. An

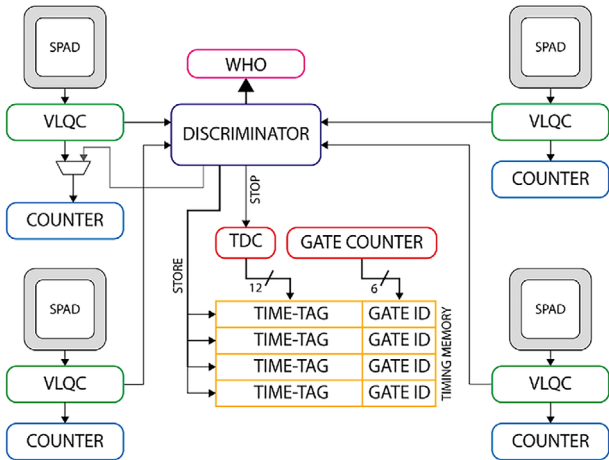


Figure 20. Pixel architecture of the gated SPAD imager.^[64] Reproduced with permission.^[64] Copyright 2018, SPIE.

alternative operation mode exploits photon-coincidence on multiple detectors of the same macropixel (with <1 ns coincidence window duration), to reduce the effect of high background levels in LiDAR applications. In such operation mode, the TDC conversion is triggered only when at least two photons are detected within the predefined coincidence time window, in order not to waste TDC conversions for single-photon events. The coincidence logic consists of a set of AND gates fed by every permutation of four signals synchronous with a photon detection event. The macropixel fill-factor is 9.6%, the peak PDP of each SPAD is 60% at 500 nm and their median DCR is 600 cps. The readout is performed in frame-based mode, reaching 90 kfps in full-readout mode (all timing, counting and spatial information are provided) and up to 360 kfps in fast-readout mode (only some selectable information is readout).

5.9. Row/Frame Skipping Array^[50]

The row/frame skipping array is specifically designed for recording the spatial correlation function of an entangled photons flux. It comprises 32×32 pixels with one SPAD and one TDC per pixel, and on-chip features aimed at increasing the conversion duty cycle. **Figure 21** shows the pixel architecture, with a 19.48% fill factor. In particular, two current-based mechanisms requiring only 3 transistors per pixel (M10, M11, and M12 in **Figure 21**) are used to speed up the array readout, considering the low-photon regime application this array has been conceived for. The first mechanism is used to sense the absence of SPAD activity in each row and reduce the readout time, by skipping empty rows. Instead, the second mechanism is implemented to skip entire frames when the total number of triggered pixels is below a user-defined threshold. Each pixel in the array contains a tunable current source (M11, M12) sinking a current from a global node only when the pixel is triggered. The sum of all pixel currents is compared with a reference current that corresponds to the requested minimum number of detected photons, and the result is given to the external controller which can eventually skip the readout phase and start a new acquisition. The sensor is able to open 50 ns long observation windows at up to 800 kHz; the readout of the entire array requires about 11 μ s and it is even shorter if not all rows are triggered.

5.10. Direct Output SPAD Array^[65]

The direct output SPAD array is an array of 64×48 SPAD pixels where sequential access has been replaced with column-wise parallel access (direct output) and row-wise non-sequential event-driven readout. As shown in **Figure 22**, the pixel consists of just the SPAD (with 26% peak PDP at 550 nm and 100 cps DCR), the front-end circuit, and some transistors to provide the pixel address. The output of the front-end circuit is connected, through

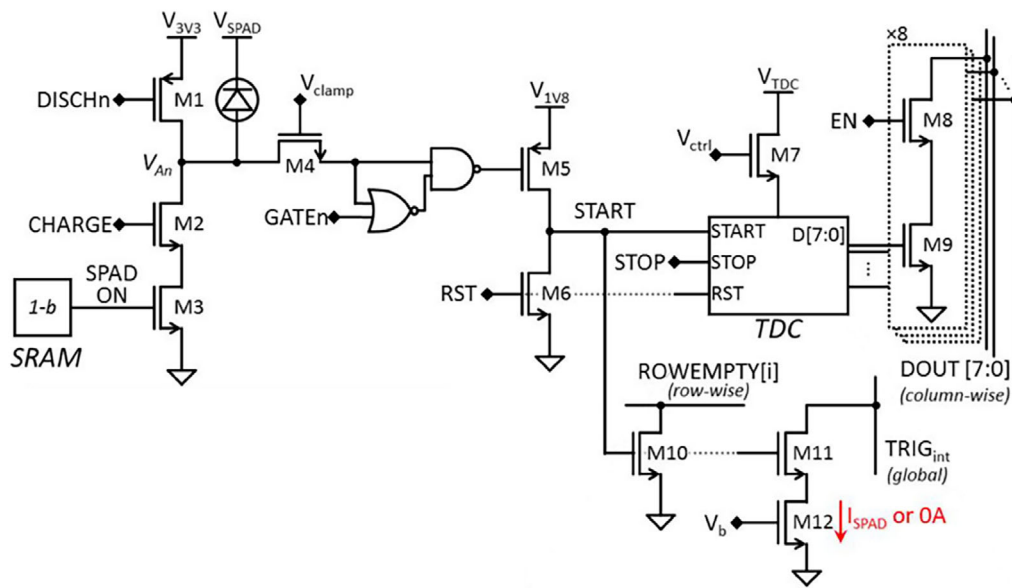


Figure 21. Pixel architecture of the row/frame skipping array.^[50] Reproduced with permission.^[50] Copyright 2018, IEEE.

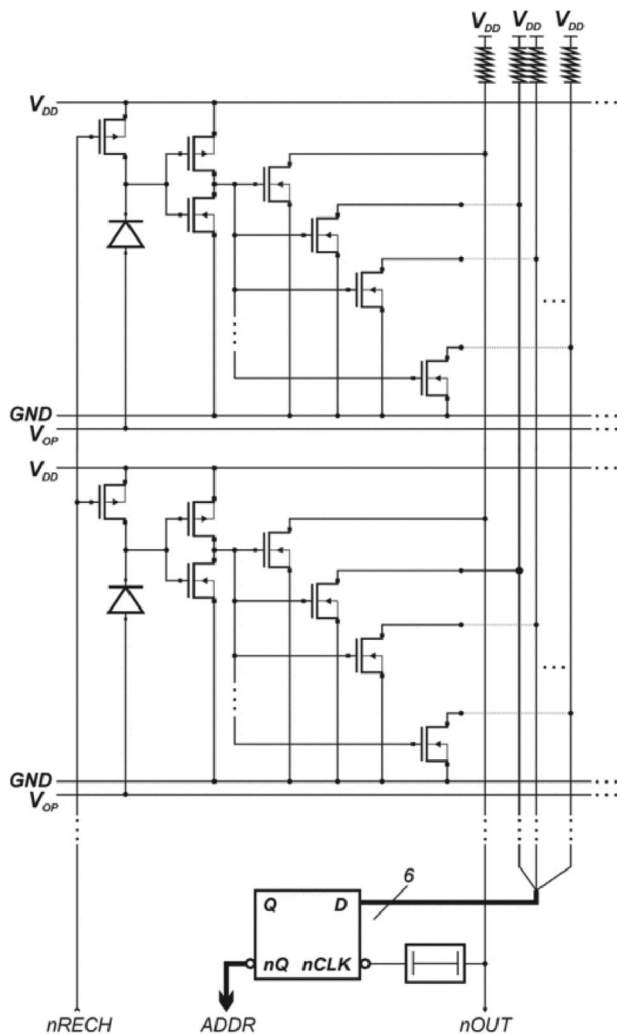


Figure 22. Column schematic of the direct output SPAD array, with the direct output signal (nOUT) and the row address (ADDR).^[65] Reproduced with permission.^[65] Copyright 2006, IEEE.

an open-drain, to a line shared among all the pixels of the same column, which provides an event-driven digital output pulse every time a photon is detected in the column. A column shared bus provides the row address of the fired pixel. This operation assumes that the event associated with photon absorption has a relatively low probability and thus that can be handled one at a time, independently for the entire column. The digital output pulse is routed to an output pad (direct output) for further processing (e.g., an external TDC). Simultaneously, the address associated with the pixel row is generated and sent onto the address line. Upon a photon detection, for the 600 ns column dead-time, the column is not able to detect other photons; for this reason, this array is not suitable for photon coincidences detection (at least if the coincidence may happen in the same column).

5.11. AER Detector^[66]

The AER (address-event representation) detector is a scalable 20×20 event-based imaging array. Each pixel consists of a

large area ($1200 \mu\text{m}^2$) SPAD with 10 cps DCR, active quenching circuitry, compact 9-bit counter (based on an analog integrator), a comparator for event detection and AER circuitry. Pixel-level analog counters are used to provide temporal integration of SPAD avalanche events. The output of the analog counter feeds a comparator that produces a digital pulse once a specified voltage threshold is exceeded. This latter signal then drives the AER block, which enables the external transfer of the pixel address through a set of row and column arbiters, that acknowledge the row and column to which the pixel belongs to. In fact, as shown in **Figure 23**, the comparator output is connected, through an open drain, to a shared row line and, when the threshold is exceeded, it sends out a row request signal for the particular row the pixel is in. An arbiter tree acknowledges the row and the pixels in the row are able to send out column requests to their respective column arbiters, avoiding any address ambiguity. Once the address is readout (at a maximum 100 MHz speed), the in-pixel analog counter is reset. The drawback of the AER implementation of this imager is the presence of errors in the timing information in case of coincident events.

5.12. Decision Tree Imager^[67]

The decision tree imager is a 3D-stacked image sensor devised as part of a LiDAR system. It comprises an array of 8×32 BSI SPADs fabricated in 45 nm standard CMOS process ($55.4 \text{ cps } \mu\text{m}^{-2}$ and 31.8% peak PDP) connected to a digital processing and communication unit (DPCU), fabricated in 65 nm standard low-power CMOS process.

The array architecture block scheme is shown in **Figure 24**. It is divided into two independent modules, each with a shared TDC at the end of a 6-level decision tree. Each node of the tree, or decision maker, acts both as an arbiter directing the SPAD pulse towards the next tree node by way of a “first-come-win-all” policy, and as a block generating the address of the first ignited SPAD. The decision makers are based on coupled D-type flip-flops (DFF) with reset. Upon photon detection, the state “1” is sampled at the output of the DFF corresponding to the event, while resetting its counterpart, thus blocking it from subsequent detection. The outputs of the DFFs are then combined using an OR gate to generate output “Q” and through a SR-latch to generate address “A.” While the “Qs” are combined in a 6-level tree (64-to-1) to generate the DTOF pulse, the “As” are used for generating the ID address, through a series of multiplexers. The DTOF pulse samples the TDC and the ID address is used to read the previous value stored in the corresponding pixel memory. The arithmetic logic unit (ALU) combines these timestamps and the result is stored in the next independent DTOF event. After each detection, the decision tree is reset and made available to the next event.

5.13. Coincidence Tree Imager^[68]

The coincidence tree imager is a sensor conceived for LiDAR with background light suppression capabilities. It is based on a 256×128 SPAD pixel array, implemented in 3D-stacked technology; the top tier houses the SPADs and it is fabricated in 45 nm

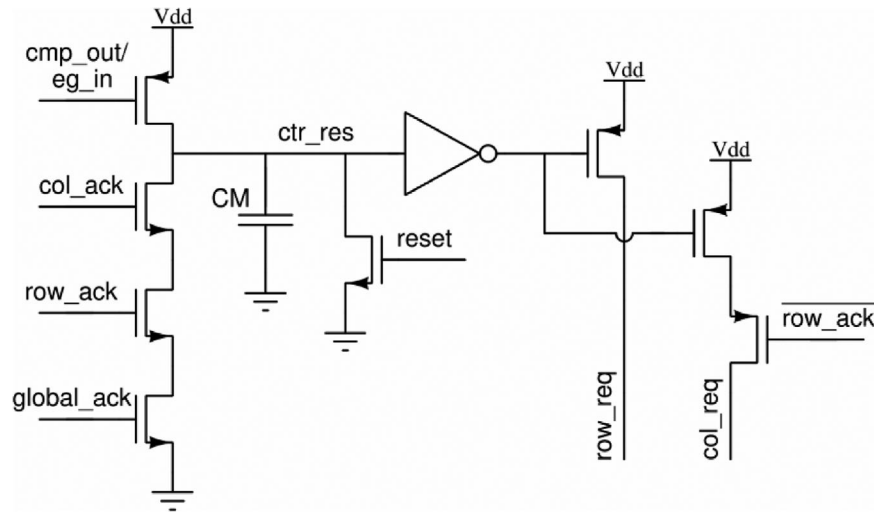


Figure 23. Schematic of the AER block of the AER detector.^[66] Reproduced with permission.^[66] Copyright 2015, IEEE.

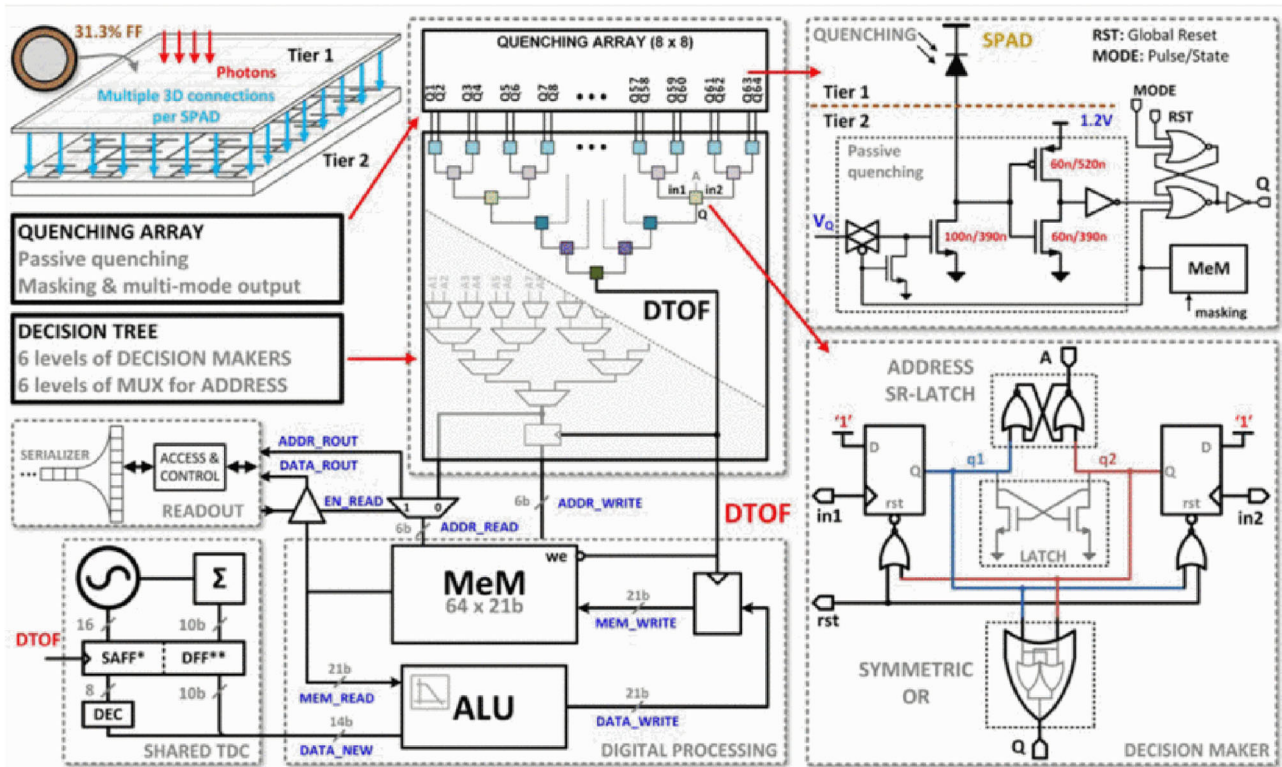


Figure 24. Block scheme of the decision tree imager^[67] architecture. Reproduced with permission.^[67] Copyright 2018, IEEE.

BSI CIS technology, while the bottom tier is fabricated in an advanced CMOS logic technology node. Figure 25 shows a block diagram of the sensor. The chip is clustered into 128 modules of 16×16 SPADs, organized in 2 submodules of 8×16 pixels, a TDC, and a seven-level coincidence detection with a tunable coincidence window t_{cw} (500 ps-to-2.2 ns). The coincidence detection mechanism is based on the concept that, when a SPAD detects a photon, the detection event is validated only if, inside the same

submodule, a predetermined number of secondary events within t_{cw} are detected, otherwise the event is ignored. More in detail, the pixel outputs are connected to the coincidence tree through the input samplers. When a photon bunch impinges on the sensor, the first photon detected triggers a pulse, which propagates to the TDC, along with the corresponding pixel address. Each pixel in the submodule hit within t_{cw} are accounted for and ranked; rank and corresponding addresses are then stored in the photon

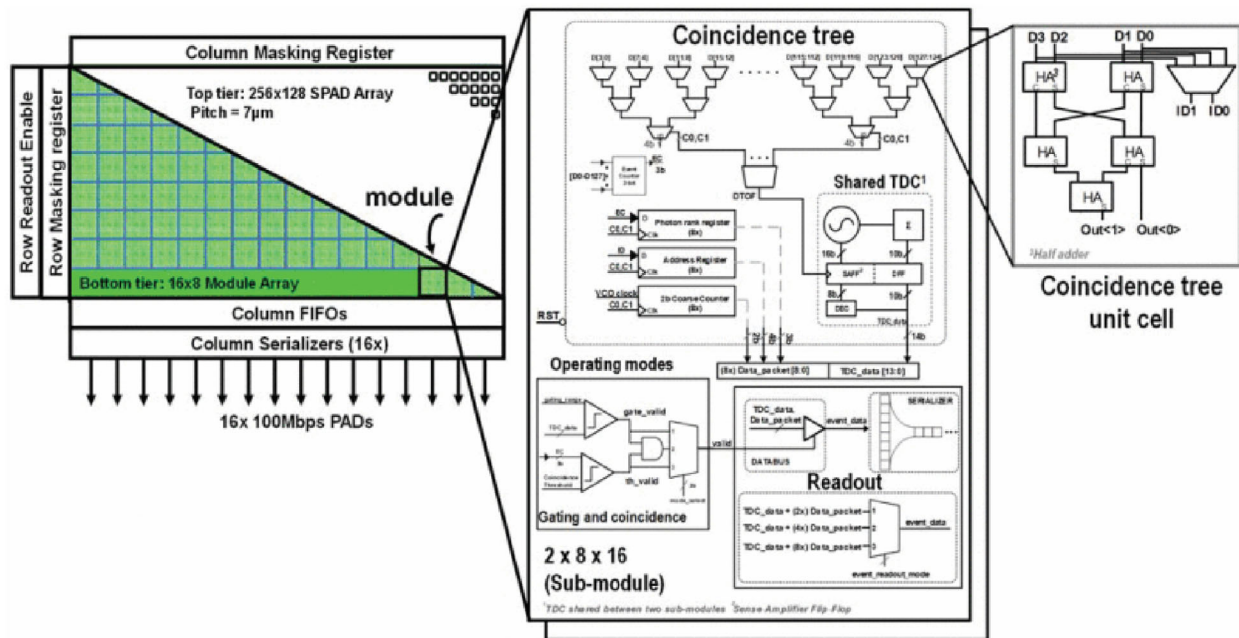


Figure 25. Block schematic of the coincidence tree imager.^[68] Reproduced with permission.^[68] Copyright 2021, IEEE.

rank register (PRR) and in the address register (AR) respectively. The coarse counter (CC) stores the timestamp MSBs of these 8 events, while the full timestamp for the first event is available from the TDC. The full data in PRR, AR, and CC is concatenated to form the complete output data packet. This data is eventually read out only if the number of detected photons within the coincidence window exceeds the coincidence threshold.

The sensor is also equipped with a mechanism known as progressive time gating. This technique is used when the distance of the objects in a scene is approximately known, thus maximizing the signal-to-noise-and-background ratio (SNBR). When progressive gating is enabled, a timestamp generated by the TDC is compared with a pre-defined gating range. Coincidence detection and progressive gating can be used concurrently.

6. Discussion on SPAD Performance for Quantum Imaging and Microscopy

Section 4 reviewed the requirements for an ideal detector for quantum imaging. This section analyzes each requirement and discusses how they have been attained by the detectors described in Section 5.

6.1. Photon Coincidence and Spatial Resolution

Photon coincidences must be detected in quantum imaging and in many other applications such as PET (to distinguish radiation events) and LiDAR (to filter background illumination). However, each application presents different peculiarities in terms of coincidence window duration and number of coincident photons to be detected. For instance, in PET the coincidence window can be tens of nanoseconds long and the expected number of coincident photons is in the order of tens or even hundreds of photons. Thus, given the high number of coincident photons and

their temporal distribution in tens of nanoseconds, it is not crucial if real coincidences in the sub-nanosecond time scale are missed. This makes PET detector requirements much different from those for a quantum imaging application, where coincidences typically involve only two photons with small time differences in the order of 500 fs.

The CMOS analog SiPM,^[53] as all analog SiPMs, is a photon number resolved detector and as such it is able to detect photon coincidences in an analog mode. Nevertheless, the coincidence window duration is not user-adjustable, since it is set by the value of the SPAD parasitic capacitance and the passive quenching resistance. Some issues typical of analog SiPMs are overcome by dSiPMs, in which the coincidence window duration can be digitally configured. For instance, in the Philips' dSiPM^[54] the coincidence window is adjustable between 10 ns and 15 ns. In the Philips' dSiPM the coincidence detection is implemented with some digital circuitry not able to detect coincidences among individual SPADs, but only among four regions in which the detector has been divided into. This strong limitation is not compatible with quantum imaging requirements.

Both analog and digital SiPMs are conceived as single point detectors, thus losing the spatial information. In order to preserve it, arrays of detectors (either SPADs or SiPMs) are then required. Typically, in order to combine the outputs of all the SPADs in a dSiPM (or in a pixel of an array of dSiPMs), balanced OR-tree networks are used. With this approach, real coincidences that happen in timescales shorter than the SPAD cell digital output duration are lost. In order to mitigate this issue, the SPADnet-I sensor,^[58] designed for PET applications, introduces the concept of "temporal compression", i.e. the SPAD cell output is shortened down to few hundreds picoseconds (250 ps in the presented implementation) before entering in the OR-tree network. The same concept is implemented in the direct ToF sensor,^[59] designed for LiDAR applications, in which the SPAD cell pulse is shortened

down to 260 ps. The MONDO detector,^[60] designed for low energy scintillation events detection, further improves the ability to detect real coincidences, by using an XOR-tree instead of the OR-tree. Indeed, the XOR tree is sensible to the signal edges instead of just to the signal status (high or low). Moreover, delay-cells have been added to the XOR-tree to further improve the probability to distinguish simultaneous photon detections. Note that, in the MONDO detector,^[60] differently from the SPADnet-I sensor^[58] and direct ToF sensor,^[59] the tree network is used to combine the output of different pixels of the SiPM array. In contrast, the SPAD cells' outputs within the same pixel are combined with an analog circuitry, which sums the digital current pulses generated by each cell and compares this value with a fixed threshold, in a coincidence window of 5–10 ns. The main limitation of this approach is that the information about the number of triggered SPADs within a pixel is lost, and consequently also the information about the total number of triggered SPADs across the entire array, since a digital trigger is generated if the pixel threshold is exceeded and all the pixels triggers are accumulated within just one global counter.

Similar tree-like structures based on either arbiters or multiplexers have been adopted in the decision tree imager^[67] and in the coincidence tree imager^[68] respectively, both designed for LiDAR applications. Among the two, just in the latter case the tree is able to discriminate the number of coincident events within a sub-module in a tunable window of 500 ps ÷ 2.2 ns, while in the former case the tree identifies the position of the first arrived photon within a sub-module of the array and triggers the TDC conversion.

A completely different approach to discriminate among coincidences has been implemented in the linear dSiPM,^[57] allowing to detect real sub-nanoseconds coincidences. A coincidence detection circuit made by full-adders, half-adders, and other logic gates signals if two or more photons have been detected in the same pixel (which includes 12 SPADs) within a 4 ns or an 8 ns coincidence window. A similar solution to detect coincidences of two or more photons has been proposed in the gated SPAD imager,^[62,64] yet resulting in a simpler (AND-gate network) logic, since the pixel includes only four SPADs.

In order to assure more flexibility, coincidences can be also detected off-chip through post-processing, exploiting the timing information provided by integrated TDCs, as it is done in the MD-SiPM^[55] and in the SPADnet-I sensor,^[58] when used for quantum imaging applications. With this approach coincidence windows as short as tens of picoseconds can be achieved (limited only by the TDC resolution) and no coincidence is lost, even if coincident photons arrive within few picoseconds. If the TDCs are integrated in each pixel, spatial information is preserved. The main drawbacks related to this approach are limited fill-factor, high-power consumption of many TDCs running in parallel, and slow readout. The latter causes a long measurement time, since SPADs are active only for a limited period, given by the TDC full-scale range (FSR), whereas for the rest of the frame they are not able to detect photons. This issue can be mitigated with the readout architecture of the row/frame skipping array,^[50] resulted as a refinement of previously presented optimal readout schemes,^[69] used in quantum imaging experiments.^[70] The TDC outputs are read out only if a photon coincidence is detected at array level, thus considerably reducing the amount of useless

data to be transferred from the SPAD array to the external processor. In order to discriminate the coincidence, each pixel generates a digital current synchronous with the photon detection, the currents of all pixels feed a common adder node, and the overall current is finally compared with a threshold current. Note that the precise coincidence information is provided by the TDC data analyzed in post-processing, whereas the on-chip coincidence discriminator is used only to start the event-driven readout. Nevertheless, we can assume that low (two photons) threshold cannot be implemented, since, in quantum imaging applications, this detector is used in frame-based mode.

A mixed digital/analog approach to distinguish photon coincidences, similar to the one presented in the MONDO detector and in the row/frame skipping array, is also covered by a United States patent.^[71]

6.2. Smart Readout

Event-driven operation, in contrast with frame-based operation, reduces dead times, throughput, power consumption, thus resulting particularly effective when the photon flux is very low (e.g., in quantum applications) and the events are uncorrelated with the system clock and the shutter (e.g., in PET imaging).

Analog SiPMs are intrinsically event-driven detectors, since an output current is provided synchronously with the photon absorption. In the CMOS analog SiPM,^[53] since a TDC is integrated on-chip, the event-driven readout of the TDC conversion is implemented through a discriminator and a flag for valid conversion to be read-out.

The Philips' dSiPM^[54] and SPADnet-I sensor^[58] implement an event-driven readout tailored for PET applications, but not applicable to two-photon coincidences. In fact, in the Philips' dSiPM collection and readout are performed only if at least one photon has been detected in each of the 64 zones, which the detector is divided into. Instead, the SPADnet-I sensor is based on the discrimination of fast-rising edges of the detected photon flux when used in PET, whereas it works at a fixed frame-rate in quantum applications.

The MONDO detector,^[60] designed for the detection of low energy scintillation events, better fits quantum imaging requirements, but the readout scheme is not described in detail in the paper. We can assume that the readout is performed when a coincidence at array level is detected; a “non-zero” flag in the pixel makes one assume that empty pixels can be skipped. A similar approach is clearly described in the row/frame skipping array,^[50] where the readout is performed only if an event at array level is detected and empty row are skipped. The entire array readout requires 11 μs, reduced to 1.2 μs if only 1 row must be read, highlighting a clear advantage in photon-starved applications.

In an ideal event-driven scheme, only useful information should be fed to the external processor. In case of relatively small arrays, such as with 4 × 4 pixels,^[72] or with 5 × 5 pixels,^[73] each pixel can be directly connected to an output pad, and a pulse synchronous with the photon detection is generated and analyzed in post-processing with great flexibility. An extended version of such arrays is the direct output SPAD array,^[65] which provides 64 × 48 pixels and a chip output pulse synchronous with the photon detection, associated with the address of the fired pixel. The

main limitation of this array is that coincidences cannot be managed, thus it cannot be used for quantum imaging applications. Another approach to optimize the amount of transferred data is presented in the AER detector,^[66] where the AER arbitration is exploited to output only the addresses of pixels exceeding a given counting threshold. Coincidences are managed by the AER arbitration, but precise timing information is lost.

6.3. Crosstalk

Crosstalk is a source of correlated noise in SPADs: it is due to the triggering of avalanche current pulses due to secondary dark counts generated by neighboring pixels. These secondary dark counts might be generated either by charge diffusion from the pixel electronics, causing electrical crosstalk, or by secondary photons spontaneously emitted (by hot carriers) during an avalanche event, which then trigger another detector, thus giving rise to optical crosstalk. The latter is typically the dominant contribution.

Generally speaking, crosstalk degrades the timing and detection performance of SPAD arrays and SiPMs and becomes even more important in quantum imaging applications, being crosstalk events temporally coincident just as entangled photons are, thus resulting in false coincident events. For this reason, crosstalk should be minimized during the detector design phase and then properly measured. Nevertheless, crosstalk is usually an unpredictable effect due to the presence of hidden paths through which both charges and photons might scatter. Moreover, it is also not straightforward to be measured and thus it is often unreported in literature. However, several ways to mitigate the effect of crosstalk events have been reported in literature and are summarized in the following.

Electrical crosstalk is prevented by the digital nature of in pixel signals and is further lowered by substrate isolated processes,^[73,74] thus constituting a negligible contribution. On the other hand, optical crosstalk constitutes the real limiting effect on optimal detection performance, and thus it must be minimized mainly through optical isolation. The usage of metal-coated trenches is an effective way to optically isolate SPADs,^[75] as well as the employment of heavily-doped regions as optical barriers and highlights the impact of indirect paths.^[76] The overall crosstalk coefficient decreases as the substrate thickness increases, because of a lower number of backside surface reflections.^[77] Deep trench isolation (DTI) proved to be effective in minimizing the impact of optical crosstalk, leaving only the residual contribution of secondary photons scattering mostly from metal layers around the SPADs.^[52] Crosstalk can be further lowered by reducing the avalanche charge, by integrating mixed passive-active quenching circuits in close proximity to the SPAD detector.^[78]

An effective way of measuring optical crosstalk is by operating photon-timing SPAD arrays in dark environment, and by arbitrarily choosing a pixel as “aggressor;” for each “victim” pixel a histogram of the difference between “aggressor” and “victim” triggering times is built.^[79,80] Moreover, the expected triangular cross-correlation in the absence of crosstalk was subtracted to eliminate spurious coincidences introduced by dark counts. The

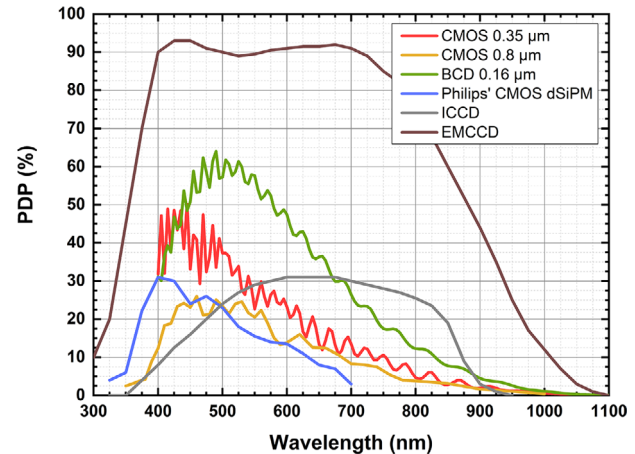


Figure 26. PDP comparison among SPADs fabricated in different technologies.^[54,65,81,82] For completeness, PDPs related to ICCD PI-MAX4-III Gen and EMCCD ANDOR iXon3 are added.

resulting histogram contains only the N_{xy} counts due to crosstalk, which were used to compute the crosstalk probability X as:

$$X = \frac{N_{xy}}{N_x + N_y} \quad (1)$$

where N_x and N_y are the total counts accumulated in “aggressor” and “victim” pixels, respectively.

When measuring quantum correlation, optical crosstalk can yield to false coincidences. Considering that crosstalk probability decreases with pixel distance, the problem can be mitigated splitting the photon pair, and imaging the two photons onto two distant parts of the sensor.^[49] However, it can also be eliminated through suitable post-processing techniques: exploiting its linear dependence on the number of detections, optical crosstalk probability can be precharacterized and then subtracted from any photon correlation measurement, by suitably redefining a corrected second-order correlation function.^[52]

6.4. Detection Efficiency

In quantum imaging, high detection efficiency values are desirable for improving sensitivity. The overall detecting capability of a SPAD based sensor can be evaluated as the product between the PDP at the wavelength of interest, which is a parameter intrinsically related to the SPAD, and the fill-factor, related to both SPAD and pixel geometries. The impact of these two factors will be herein discussed.

Concerning PDP, a first rough estimate can be made by looking at the peak values listed in Table 1. Nevertheless, a fair comparison among different sensors should be made examining the whole PDP spectral curve, in order to retrieve the actual SPAD detection efficiency at any given wavelength. In this regard, **Figure 26** represents some of the most representative PDP curves of SPAD sensors within arrays integrated into standard technologies.^[54,65,81,82] It can be seen that all of them have their peak performance in the visible spectrum, ranging from about 25% to 60%, and that even more than 20% PDP is possible in

the Near Infrared (NIR) range. It must also be pointed out that, at equal wavelength, the PDP spreads by no more than a factor 3.5, while fill-factor has a much broader variability, even up to a factor 20.

In point of fact, fill-factor of SPAD pixels implemented in CMOS technologies is partially limited by guard rings surrounding the sensitive area to prevent edge breakdown. Still, it is the in-pixel electronics to be the most area-consuming, causing fill-factor to drop even below 5% in cases of large in-pixel electronics,^[79] compared to cases above 70%,^[53,57] in which the in-pixel electronics was either limited to a simple low-area consuming quenching resistor,^[53] or moved aside the sensitive area (because of the low-number of pixels).^[57] Of course, low fill-factor values negatively impact on the overall sensor detection capability, but can be considered beneficial in terms of crosstalk, whose effect becomes less detrimental as the distance among detectors increases. A good solution to mitigate the problem, while maintaining the same pixel pitch, is employing microlens arrays mounted on top of the detector chip.^[83]

Microlenses are miniaturized lenses placed over each pixel to collect photons from the entire pixel area and convey them to the SPAD active area. The sensitivity enhancement due to microlenses is measured by means of a parameter known as concentration factor (CF), which is defined as the ratio of light intensity detected with and without microlenses. The possibility of achieving a uniform CF of 35 with a dispersion around 6% over large microlens arrays (32 × 32) with 50 μm pitch have been reported.^[84] The successful use of microlenses in even larger SPAD arrays have been successfully reported, where a peak CF of 14 is attained over 25 μm pitch SPAD arrays with 128 × 128 and 512 × 128 pixels.^[85] The development of microlens arrays with high CF presents many technological challenges, both in the fabrication of the microlens with optimal geometrical parameters and in the alignment with the detector procedure. The task becomes even more challenging for SPAD arrays with high number of pixels, low native fill-factor, and large pitch size.^[87] Since CF also increases with the f-number of the employed optical system,^[83,85,86,88] microlenses can significantly provide improvements when collimated light beams are employed, such as in most quantum microscopy applications. Of course, the employment of microlens arrays, increases fabrication complexity and cost, especially when considering large-format detectors.

6.5. Number of Pixels and Scalability

As Table 1 brings out, SPAD arrays typically comprise a number of pixels that ranges from a few tens up to thousands. In quantum imaging, SPAD detectors with high-pixel count over a large area are advisable, so to retrieve the position of incoming photons with the best resolution possible, and with the largest field-of-view (FOV) within reach. Nevertheless, large area detectors and high pixel density can somehow limit the sensor optimal performance, from both detection and electronics standpoints. On the one hand, enlarging the overall sensitive area of a SPAD array yields to higher dark-count triggered pixels probability, thus leading to false coincidence detections. To this end, SPAD dimension, number of pixels and pixel pitch should be optimized, so to maximize SNR. Besides, at equal architecture and number of pix-

els, worse performance is expected for a SPAD array spread over a larger area, because of greater components mismatches and stray capacitive effects, longer signal routes, and less uniformity on SPAD parameters.

On the other hand, pixel density is limited by the maximum achievable fill factor. In fact, to achieve higher pixel density SPAD dimension must be reduced, thus increasing the percentage of dead-area around the SPAD, being both guard-ring clearance and SPAD front-end electronics dimensions basically invariant. Signal routing over pixels with increasing density and complexity must be taken into consideration as well. Furthermore, also the overall amount of data produced by the pixel increments, leading to slower readout times, hence the need of conceiving smarter readout techniques for minimizing data throughput, as thoroughly highlighted in Section 6.2.

Scalability is a desirable feature in order to obtain large-format and high-performing SPAD arrays. Generally speaking, scalability is defined as the ability of a system to maintain its full functionality and optimal performance, even when area scales up indefinitely, without requiring more demanding processing hardware (i.e., both the amount of produced data and the readout time remain unchanged). The full-scalability concept of a SPAD array is then related to both the architecture itself and the technology node used for integration. On the one hand, ensuring constant data throughput and readout time is related to the conceived array architecture. On the other hand, ensuring identical performance for a larger array can be achieved through the hierarchical repetition of a modular core in which stray effects are negligible and in which electronics and routing complexity is moderate. Unfortunately, after a certain number of repetition steps, process variations and timings skews come into play, with less impact as the technology node scales down. As a matter of fact, the technology node used for detector integration is the ultimate bottleneck on large-format and high-pixel count SPAD arrays. At a given chip area, scaling down technology nodes allows the on-chip integration of denser and higher performance electronics and ever-more extensive variety of functionalities. At present days, SPAD arrays with high fill-factor have been integrated in 40 nm and 45 nm technologies.^[46–48] Moreover, employing stacked BSI SPADs and having a wafer dedicated to the sensors, makes it possible to have denser arrays at a given technology node.

6.6. Summary and Final Remarks

The SPAD array chips reviewed so far show many features that can be favorably exploited in a quantum imaging detector. In this section, we summarize which requirements have already been achieved in the presented detectors and which ones, instead, still require improved architectures.

Concerning the ability to detect coincidences among all array pixels, we can observe that all fully digital architectures do not fulfill the quantum imaging requirements. In fact, solutions based on OR/XOR trees (e.g., SPADnet-I sensor^[58] and direct ToF sensor^[59] for coincidences at pixel level, and MONDO detector^[60] for coincidences at array level) are not able to detect actual coincidences within sub-picosecond coincidence windows. Whereas solutions based on more complex logic architectures (e.g., linear dSiPM, gated SPAD imager^[64] and coincidence tree imager^[68])

are not scalable and are used only to detect coincidences within the same pixel or sub-parts of the array, as complexity drastically increases with the number of pixel to be combined.

On the other hand, fully analog solutions, such as in analog SiPMs, do not preserve spatial information, are prone to readout noise, and are hardly scalable (e.g., due to high capacitive load). The best approach applicable to quantum imaging is a mixed digital and analog architecture, such as the one presented in MONDO detector,^[60] for coincidences at the pixel level or in the row/frame skipping array^[50] for detecting an event at array level. These architectures preserve spatial information, immunity to noise, scalability, still allowing to detect sub-picosecond coincidences. The presented mixed digital and analog architectures are a good starting point toward the ideal quantum imager detector. However, they still need optimizations to make them more scalable in large arrays and able to precisely detect coincidences with low (two photons) threshold.

Considering the sparsity of the information in quantum imaging, another crucial aspect is readout optimization, which should be event-driven and only useful information (ideally the address of fired pixels) should be stored and output. A possible solution is presented in the row/frame skipping array,^[50] but still entire rows of the array are readout even if only one pixel carries useful information. Having direct outputs for each pixel of the array,^[72,73] is a purely event-driven approach, but it is not feasible for large SPAD array, because it would require one pad for each SPAD in order to identify the triggered one. In order to mitigate this issue, in the direct output SPAD array^[65] the pixels of each column share the same output, and the address of the row is sent out, together with the output pulse. Similarly, the AER detector^[66] outputs only the addresses of the fired pixels, but misses the temporal information and consequently the possibility to detect coincidences even in postprocessing. An ideal detector for quantum imaging should detect coincidences on-chip, for instance with a mixed analog and digital approach, and output only the addresses of the pixels involved in the coincidence event, for instance with an architecture inspired by the AER detector.

Detection efficiency and crosstalk are linked parameters, thus trade-offs between them must be considered. In fact, the detection efficiency is strongly impacted by fill-factor, which is intrinsically limited because of the need for in-pixel electronics. The introduction of microlens arrays on top of the SPAD detector is probably a must in quantum imaging, despite the increase of production costs and complexity.

Another important SPAD parameter to be optimized is the median DCR, since high DCR, together with a long coincidence time window, is another cause of false coincidences detection. As shown in Table 1, DCR is a parameter that varies a lot among fabrication technologies and architectures. On the other hand, the effect of hot pixels can be mitigated by implementing the possibility to disable every single SPAD independently, as shown in many of the presented architectures.^[50,55,58,62]

In quantum imaging applications, arrays of SiPMs do not present any clear advantage in respect to SPAD arrays, moreover, they have the drawback to lose the spatial information within the SiPM pixel. Thus, SPADs array should be the preferred architecture. Concerning the number of pixels, a higher pixel count allows to obtain better image resolution and wider FOV, but tradeoffs with overall performance must be considered. In

particular, increasing the total sensitive area also increases the overall DCR and consequently the false coincidence detection probability; higher number of pixels increases electronics and routing complexity, in order to preserve timing information and detect coincidence in distant pixels. Scalable architectures mitigate the latter issue, but still suffers from possible mismatches in signal paths or for process variations across the same chip.

In conclusion, SPAD arrays can achieve all requirements for quantum imaging and microscopy applications, representing the best candidates for most of those applications. Indeed, it is possible to develop relatively large SPAD arrays, with tens of thousands of pixels, able to detect on-chip two- or more-photons coincidences, exploiting mixed digital-analog architectures, which are highly scalable and with high noise immunity. Upon event detection, an event-driven readout should be performed, transferring to the external electronics only the addresses of those pixels involved in the photon coincidence event, thus optimizing data throughput. The probability to detect false coincidences can be limited through proper technology selection and SPAD pixel design, so to minimize DCR and crosstalk. Eventually, high photon detection efficiency is a key detector parameter, typically mostly affected by the chosen fabrication technology and the array fill-factor, which can be recovered by employing microlens arrays.

7. Toward Next Generation Quantum Imaging Detectors

New sensors designed for quantum imaging should consider the clear benefits, in terms of sensitivity and Signal to Noise Ratio (SNR) in the final application, that can be achieved reducing the coincidence window, enhancing the Photon Detection Efficiency (PDE; i.e. PDP and fill-factor) and limiting the DCR.

In fact, in quantum imaging, the measurement signal to noise ratio (SNR) can be defined as the ratio between real and false detected coincidences:

$$SNR = \frac{\mu_C}{\mu_{Cf}} \approx \frac{\mu_C}{\mu_S^2 \cdot \Delta t_{coinc}} \quad (2)$$

where μ_C is the flux of detected real coincidences, μ_{Cf} is the flux of detected false coincidences, μ_S is the flux of detected single photons and Δt_{coinc} is the coincidence window duration. The approximately equal is because we are not considering the Poisson statistics of the light. In this equation, we consider only the noise introduced by the detector, i.e. the false coincidences due to detectors nonidealities (limited PDE and finite coincidence window duration).

The flux of detected real coincidences and the flux of detected single photons can be computed as:

$$\mu_C = M \cdot \eta^2 \quad (3)$$

$$\mu_S = 2 \cdot M \cdot \eta \cdot (1 - \eta) + DCR \quad (4)$$

where M is the couple generation rate of the entangled photon source and η the overall detection efficiency (typically dominated by the detector PDE).

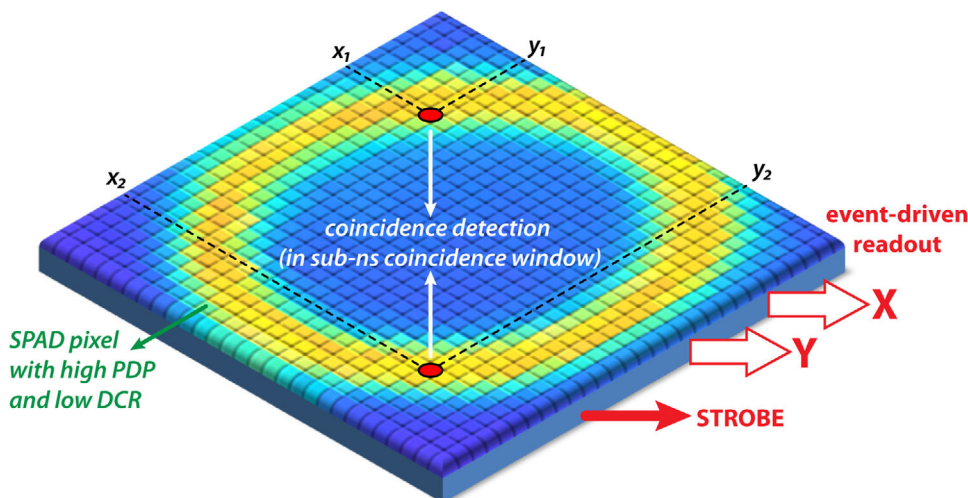


Figure 27. Illustrative example of next-generation SPAD detector with on-chip sub-nanosecond coincidence detection, event driven readout of the triggered pixel address and SPAD pixel with high detection efficiency and low DCR.

Substituting Equations (3) and (4) in 2 the SNR results:

$$\text{SNR} = \frac{M \cdot \eta^2}{[2 \cdot M \cdot \eta \cdot (1 - \eta) + \text{DCR}]^2 \cdot \Delta t_{\text{coinc}}} \quad (5)$$

Thus, we can observe that the SNR improves with $1/\Delta t_{\text{coinc}}$ and, if the DCR is negligible with $1/(1 - \eta)^2$, where $(1 - \eta)$ represents the photon losses, mainly due to limited detection efficiency (both fill-factor and PDP). For instance, considering 10^7 couple s^{-1} couple generation rate, 20% detection efficiency (η), 2 ns coincidence window duration and 10^6 cps DCR the SNR results 11.3. With such a high DCR, SNR is very sensible to DCR variations, for instance doubling the DCR ($\text{DCR} = 2 \times 10^6$ cps) the SNR almost halves ($\text{SNR} = 7.4$). When DCR contribution becomes negligible in respect to the generation rate, the sensibility of SNR with DCR reduces (e.g., considering 10^5 cps DCR the SNR results 18.4, with 2×10^5 cps DCR the SNR slightly reduces to 17.3), but the SNR is still very sensible to efficiency and coincidence window duration (e.g., halving the efficiency to 10% the SNR becomes 13.8 or doubling the coincidence window duration to 4 ns the SNR becomes 9.18).

Considering the sensitivity improvement in imaging schemes which exploit quantum entangled NOON states, losing single photons from a NOON states projects it into a classical state, canceling the quantum advantage.^[89] In ref. ^[90], it has been demonstrated that a real advantage of using NOON states is preserved, in a worst case scenario, only if $\eta^N \cdot \nu^2 \cdot N > 1$, with η the overall detection efficiency, ν the interference visibility (which depends on the light source) and N the number of entangled photons. So for instance, considering an ideal visibility of 100% and 2 entangled photons an efficiency as high as 70% is required to satisfy the inequality. Note that increasing the entangled photon number N , the inequality becomes harder and harder to be satisfied making it very difficult to practically implement high order NOON imaging techniques.^[3]

In twin beam detection schemes the theoretical noise reduction factor σ is equal to 0 (whereas with classical light it is equal to 1), but considering detection efficiency limitations the lower

bound of reduction factor results $\sigma_d = 1 - \eta$.^[7] Thus it is clear the strong impact of the detection efficiency in the noise reduction also in Twin Beam imaging systems.

Based on the remarks made in this paragraph and in Section 6.6, **Figure 27** shows an illustrative example of a next-generation SPAD array for quantum imaging, including on-chip coincidence detection in sub-nanosecond coincidence windows, event-driven readout that provides only the addressed of the triggered pixels and high-performance SPAD pixels.

8. Conclusion

This paper reviewed the main requirements of an optimal detector to be employed in quantum imaging and quantum microscopy. Many different SPAD-based architectures (i.e., analog SiPMs, digital SiPMs, and SPAD arrays) have been presented and discussed, with particular focus on their ability to detect photon coincidences on-chip, while preserving the spatial information, possibility to perform event-driven readout with minimum data overhead, maximized detection efficiency, low noise, high pixel number, and ease of scalability. Ultimate limits of SPAD-based sensors in CMOS technologies are represented by the limited PDP (particularly, in the near-infra-red range) and the minimum achievable coincidence window duration, typically in the order of a few nanosecond, albeit entangled photons are temporally distributed in a sub-picosecond timescale. Some of the analyzed architectures were specifically designed for quantum imaging, while others, although designed for different applications (PET, low energy scintillation detection, LiDAR), present some features that result advantageous for quantum measurements as well.

All in all, we may conclude that none of the detectors presented in literature embodies all the listed requirements. However, all the building blocks to develop next-generation detectors for quantum imaging and microscopy already exist, ready to be combined in a novel SPAD array architecture, optimized for the targeted application.

Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation programme under Grant 801060 ("Quantum-enhanced on-chip interference microscopy", Q-MIC).

The authors also acknowledge Ivo Pietro Degiovanni (Istituto Nazionale di Ricerca Metrologica) for the very useful discussion about quantum imaging applications and for having checked the introductory part of the paper.

Conflict of Interest

The authors declare no conflict of interest.

Keywords

event-driven readout, photon coincidence, quantum imaging, quantum microscopy, silicon photomultipliers, single-photon avalanche diode arrays

Received: January 15, 2021

Revised: April 21, 2021

Published online: May 14, 2021

- [1] M. Taylor, J. Janousek, V. Daria, J. Knittel, B. Hage, H. A. Bachor, W. P. Bowen, *Nat. Photonics* **2013**, 7, 229.
- [2] J. Aasi, J. Abadie, B. Abbott, R. Abbott, T. D. Abbott, M. R. Abernathy, C. Adams, T. Adams, P. Addesso, R. X. Adhikari, C. Affeldt, O. D. Aguiar, P. Ajith, B. Allen, E. Amador Ceron, D. Amariutei, S. B. Anderson, W. G. Anderson, K. Arai, M. C. Araya, C. Arceneaux, S. Ast, S. M. Aston, D. Atkinson, P. Aufmuth, C. Aulbert, L. Austin, B. E. Aylott, S. Babak, P. T. Baker, et al., *Nat. Photonics* **2013**, 7, 613.
- [3] I. Ruo Berchera, I. P. Degiovanni, S. Olivares, M. Genovese, *Phys. Rev. Lett.* **2013**, 110, 213601.
- [4] H. K. Lo, M. Curty, K. Tamaki, *Nat. Photonics* **2014**, 8, 595.
- [5] T. Ono, R. Okamoto, S. Takeuchi, *Nat. Commun.* **2013**, 4, 2426.
- [6] M. Genovese, *J. Opt.* **2016**, 18, 073002.
- [7] I. Ruo Berchera, I. P. Degiovanni, *Metrologia* **2019**, 56, 024001.
- [8] V. Giovannetti, S. Lloyd, L. Maccone, *Science* **2004**, 306, 1330.
- [9] Y. B. Zel'Dovich, D. N. Klyshko, *Sov. Phys. - JETP* **1969**, 9, 40.
- [10] D. C. Burnham, D. L. Weinberg, *Phys. Rev. Lett.* **1970**, 25, 84.
- [11] A. Meda, E. Losero, N. Samantaray, F. Scafirimuto, S. Pradyumna, A. Avella, I. Ruo-Berchera, M. Genovese, *J. Opt.* **2017**, 19, 094002.
- [12] B. Lounis, M. Orrit, *Rep. Prog. Phys.* **2005**, 68, 1129.
- [13] X. L. Chu, S. Götzinger, V. Sandoghdar, *Nat. Photonics* **2017**, 11, 58.
- [14] O. Schwartz, J. M. Levitt, R. Tenne, S. Itzhakov, Z. Deutsch, D. Oron, *Nano Lett.* **2013**, 13, 5832.
- [15] D. Gatto Monticone, K. Katamadze, P. Traina, E. Moreva, J. Forneris, I. Ruo-Berchera, P. Olivero, I. P. Degiovanni, G. Brida, M. Genovese, *Phys. Rev. Lett.* **2014**, 113, 143602.
- [16] A. Heidman, R. J. Horowicz, S. Reynaud, E. Giacobino, C. Fabre, *Phys. Rev. Lett.* **1987**, 59, 2555.
- [17] O. Jedrkiewicz, Y. K. Jiang, E. Brambilla, A. Gatti, M. Bache, L. A. Lugiato, P. Di Trapani, *Phys. Rev. Lett.* **2004**, 93, 243601.
- [18] J. L. Blanchet, F. Devaux, L. Furfaro, E. Lantz, *Phys. Rev. Lett.* **2008**, 101, 233604.
- [19] G. Brida, M. Genovese, I. R. Berchera, *Nat. Photonics* **2010**, 4, 227.
- [20] N. Treps, U. Andersen, B. Buchler, P. K. Lam, A. Maitre, H. A. Bachor, C. Fabre, *Phys. Rev. Lett.* **2002**, 88, 203601.
- [21] N. Treps, N. Grosse, W. P. Bowen, C. Fabre, H. A. Bachor, P. K. Lam, *Science* **2003**, 301, 940.
- [22] A. Belinskii, D. Klyshko, *J. Exp. Theor. Phys.* **1994**, 78, 259.
- [23] T. B. Pittman, Y. H. Shih, D. V. Strekalov, A. V. Sergienko, *Phys. Rev. A* **1995**, 52, R3429.
- [24] X. Chen, I. N. Agafonov, K. Luo, Q. Liu, R. Xian, M. V. Chekhova, L. Wu, *Opt. Lett.* **2010**, 35, 1166.
- [25] R. S. Aspden, N. R. Gemmill, P. A. Morris, D. S. Tasca, L. Mertens, M. G. Tanner, R. A. Kirkwood, A. Ruggeri, A. Tosi, R. W. Boyd, G. S. Buller, R. H. Hadfield, M. J. Padgett, *Optica* **2015**, 2, 1049.
- [26] G. Lemos, V. Borish, G. Cole, S. Ramelow, R. Lapkiewicz, A. Zeilinger, *Nature* **2014**, 512, 409.
- [27] M. V. Chekhova, Z. Y. Ou, *Adv. Opt. Photonics* **2016**, 8, 104.
- [28] S. H. Tan, B. I. Erkmann, V. Giovannetti, S. Guha, S. Lloyd, L. Maccone, S. Pirandola, J. H. Shapiro, *Phys. Rev. Lett.* **2008**, 101, 253601.
- [29] E. D. Lopaeva, I. Ruo Berchera, I. P. Degiovanni, S. Olivares, G. Brida, M. Genovese, *Phys. Rev. Lett.* **2013**, 110, 153603.
- [30] T. Gregory, P. A. Moreau, E. Toninelli, M. J. Padgett, *Sci. Adv.* **2020**, 6, eaay2652.
- [31] P. Morris, R. Aspden, J. E. C. Bell, R. W. Boyd, M. J. Padgett, *Nat. Commun.* **2015**, 6, 5913.
- [32] H. Di Lorenzo Pires, C. Monken, M. van Exter, *Phys. Rev. A* **2009**, 80, 022307.
- [33] R. Fickler, M. Krenn, R. Lapkiewicz, S. Ramelow, A. Zeilinger, *Sci. Rep.* **2013**, 3, 1914.
- [34] B. M Jost, A. V. Sergienko, A. F. Abouraddy, B. E. A. Saleh, M. C. Teich, *Opt. Express* **1998**, 3, 81.
- [35] L. Zhang, L. Neves, J. S. Lundeen, I. A. Walmsley, *J. Phys. B: At., Mol. Opt. Phys* **2009**, 42, 114011.
- [36] M. P. Edgar, D. S. Tasca, F. Izdebski, R. E. Warburton, J. Leach, M. Agnew, G. S. Buller, R. W. Boyd, M. J. Padgett, *Nat. Commun.* **2012**, 3, 984.
- [37] E. Toninelli, M. P. Edgar, P. A. Moreau, G. M. Gibson, G. D. Hammond, M. J. Padgett, *Opt. Express* **2017**, 25, 21826.
- [38] H. Defienne, B. Ndagano, A. Lyons, D. Faccio, presented at COSI, Washington, DC, USA, June **2020**, paper CTh3C.2.
- [39] M. Hirsch, R. J. Wareham, M. L. Martin-Fernandez, M. P. Hobson, D. J. Rolfe, *PLoS One* **2013**, 8, e53671.
- [40] H. Defienne, M. Reichert, J. W. Fleischer, D. Faccio, *Sci. Adv.* **2019**, 5, eaax0307.
- [41] E. D. Lopaeva, I. Ruo-Berchera, S. Olivares, G. Brida, I. P. Degiovanni, M. Genovese, *Phys. Scr.* **2014**, 2014, 014026.
- [42] N. Samantaray, I. Ruo-Berchera, A. Meda, M. Genovese, *Light: Sci. Appl.* **2017**, 6, e17005.
- [43] F. Just, M. Filipenko, A. Cavanna, T. Michel, T. Gleixner, M. Taheri, J. Vallerger, M. Campbell, T. Tick, G. Anton, M. V. Chekhova, G. Leuchs, *Opt. Express* **2014**, 22, 17561.
- [44] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, W. Brockherde, *J. Mod. Opt.* **2014**, 61, 102.
- [45] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. Dalla Mora, D. Contini, D. Durini, S. Weyers, W. Brockherde, *J. Sel. Top. Quantum Electron.* **2014**, 20, 364.
- [46] S. Yokagawa, presented at IEDM, San Francisco, CA, December **2019**.
- [47] G. Roherer, presented at ISSW, Edinburgh, Scotland, UK, June **2020**.
- [48] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, B. Mamy, presented at IEDM, San Francisco, CA December **2017**.
- [49] M. Unternährer, B. Bessire, L. Gasparini, D. Stoppa, A. Stefanov, *Opt. Express* **2016**, 24, 28829.
- [50] L. Gasparini, M. Zarghami, H. Xu, L. Parmesan, M. M. Garcia, M. Unternährer, B. Bessire, A. Stefanov, D. Stoppa, M. Perenzoni, presented at ISSCC, San Francisco, CA, February **2018**.
- [51] B. Ndagano, H. Defienne, A. Lyons, I. Starshynov, F. Villa, S. Tisa, D. Faccio, *npj Quantum Inf.* **2020**, 6, 94.
- [52] G. Lubin, R. Tenne, I. M. Antolovic, E. Charbon, C. Bruschini, D. Oron, *Opt. Express* **2019**, 27, 32863.

- [53] A. Muntean, A. Sachdeva, E. Venialgo, S. Gnechchi, D. Palubiak, C. Jackson, E. Charbon, presented at NSS/MIC, Sidney, Australia, November 2018.
- [54] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, R. Ballizany, presented at NSS/MIC, Orlando, FL, October 2009.
- [55] A. Carimatto, A. Ulku, S. Lindner, E. Gros-Daillon, B. Rae, S. Pellegrini, E. Charbon, *IEEE Solid-State Circuits Lett.* **2018**, 1, 241.
- [56] A. Carimatto, A. Ulku, S. Lindner, E. Gros-Daillon, B. Rae, S. Pellegrini, E. Charbon, presented at Symp. on VLSI Circuits, Honolulu, HI, June 2018.
- [57] C. Niclass, M. Soga, H. Matsubara, S. Kato, M. Kagami, *IEEE J. Solid-State Circuits* **2013**, 48, 559.
- [58] L. H. C. Braga, L. Gasparini, L. Grant, R. K. Henderson, N. Massari, M. Perenzoni, D. Stoppa, R. Walker, *IEEE J. Solid-State Circuits* **2014**, 49, 301.
- [59] M. Perenzoni, D. Perenzoni, D. Stoppa, *IEEE J. Solid-State Circuits* **2017**, 52, 151.
- [60] E. Manuzatto, L. Gasparini, M. Perenzoni, Y. Zou, L. Parmesan, G. Battistoni, M. De Simoni, Y. Dong, M. Fischetti, E. Gioscio, I. Mattei, R. Mirabelli, V. Patera, A. Sarti, A. Schiavi, A. Sciubba, S. M. Valle, G. Traini, M. Marafini, *IEEE Solid-State Circuits Lett.* **2019**, 2, 75.
- [61] S. M. Valle, G. Battistoni, V. Patera, D. Pinci, A. Sarti, A. Sciubba, E. Spiriti, M. Marafini, *Nucl. Instrum. Methods Phys. Res., Sect. A* **2017**, 845, 556.
- [62] H. Xu, L. Pancheri, G. Dalla Betta, D. Stoppa, *Opt. Express* **2017**, 25, 12765.
- [63] D. Portaluppi, E. Conca, F. Villa, *J. Sel. Top. Quantum Electron.* **2018**, 24, 1.
- [64] D. Portaluppi, E. Conca, F. Villa, F. Zappa, presented at SPIE Commercial + Scientific Sensing and Imaging, Orlando, FL, May 2018.
- [65] C. Niclass, M. Sergio, E. Charbon, presented at ESSCIRC, Montreux, Switzerland, September 2006.
- [66] A. Berkovich, T. Datta, P. Abshire, presented at ISCAS, Lisbon, Portugal, May 2015.
- [67] A. R. Ximenes, P. Padmanabhan, M. Lee, Y. Yamashita, Y. N. Young, E. Charbon, presented at ISSCC, San Francisco, CA, USA, February 2018.
- [68] P. Padmanabhan, C. Zhang, M. Cazzaniga, B. Efe, A. R. Ximenes, M. Lee, E. Charbon, presented at ISSCC, February 2021.
- [69] M. Zarghami, L. Gasparini, D. Stoppa, presented at PRIME, Giardini Naxos, Italy, June 2017.
- [70] V. Mitev, L. Balet, N. Torcheboeuf, L. Gasparini, M. Perenzoni, B. Bessire, A. Stefanov, A. Mikhalychev, D. Mogilevtsev, D. Boiko, presented at CLEO/Europe-EQEC, Munich, Germany, June 2019.
- [71] B. Nouri, B. Heights (Kiskeya Microsystems LLC), *US0259792A1* **2019**.
- [72] D. L. Boiko, N. J. Gunther, N. Brauer, M. Sergio, C. Niclass, G. B. Beretta, E. Charbon, *New J. Phys.* **2008**, 11, 013001.
- [73] M. Buttafava, F. Villa, M. Castello, G. Tortarolo, E. Conca, M. Sanzaro, S. Piazza, P. Bianchini, A. Diaspro, F. Zappa, G. Vicidomini, A. Tosi, *Optica* **2020**, 7, 755.
- [74] C. Veerappan, E. Charbon, *IEEE J. Sel. Top. Quantum Electron.* **2014**, 20, 299.
- [75] W. J. Kindt, H. W. van Zeijl, S. Middelhoek, presented at ESSDERC, Bordeaux, France, September 1998.
- [76] I. Rech, A. Ingarciola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, S. Cova, *IEEE Photonics Technol. Lett.* **2008**, 20, 330.
- [77] A. Ficorella, L. Pancheri, G.-F. Dalla Betta, P. Brogi, G. Collazuol, P. S. Marrocchesi, F. Morsani, L. Ratti, A. Savoy-Navarro, presented at ESSDERC, Lausanne, Switzerland, September 2016.
- [78] F. Zappa, A. Lotito, A. C. Giudice, S. Cova, M. Ghioni, *IEEE J. Solid-State Circuits* **2003**, 38, 1261.
- [79] E. Conca, I. Cusini, F. Severini, R. Lussana, F. Zappa, F. Villa, *IEEE Photonics J.* **2019**, 11, 1.
- [80] F. Villa, R. Lussana, D. Tamborini, A. Tosi, F. Zappa, *IEEE Photonics Technol. Lett.* **2015**, 27, 1261.
- [81] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, W. Brockherde, *J. Mod. Opt.* **2014**, 61, 102.
- [82] M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce, F. Zappa, *IEEE J. Sel. Top. Quantum Electron.* **2018**, 24, 1.
- [83] P. W. R. Connolly, X. Ren, A. McCarthy, H. Mai, F. Villa, A. J. Waddie, M. R. Taghizadeh, A. Tosi, F. Zappa, R. K. Henderson, G. S. Buller, *Appl. Opt.* **2020**, 59, 4488.
- [84] S. Donati, E. Randone, M. Fathi, J. Lee, E. Charbon, presented at CLEO/QELS, San Jose, CA May 2010.
- [85] J. M. Pavia, M. Wolf, E. Charbon, *Opt. Express* **2014**, 22, 4202.
- [86] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, E. Charbon, *IEEE Trans. Electron Devices* **2016**, 63, 57.
- [87] G. Intermite, A. McCarthy, R. Warburton, X. Ren, F. Villa, R. Lussana, A. J. Waddie, M. R. Taghizadeh, A. Tosi, F. Zappa, G. S. Buller, *Opt. Express* **2015**, 23, 33777.
- [88] S. Burri, Y. Maruyama, X. Michalet, F. Regazzoni, C. Bruschini, E. Charbon, *Opt. Express* **2014**, 22, 17573.
- [89] S. Slussarenko, M. M. Weston, H. M. Chrzanowski, L. K. Shalm, V. B. Verma, S. W. Nam, G. J. Pryde, *Nat. Photonics* **2017**, 11, 700.
- [90] K. J. Resch, K. L. Pregnell, R. Prevedel, A. Gilchrist, G. J. Pryde, J. L. O'Brien, A. G. White, *Phys. Rev. Lett.* **2007**, 98, 223601.



Francesca Madonini received her Master's degree (summa cum laude) in Electronics Engineering at Politecnico di Milano in 2019. She was double degree awarded at Politecnico di Torino, as a part of the Alta Scuola Politecnica program. Since 2019, she is enrolled in the research doctorate in Information Technology (area Electronics) at Politecnico di Milano. Her current research focuses on the design of microelectronic circuits for single photon avalanche diode (SPAD) arrays, specifically tailored for photon coincidences measurements in the field of quantum-enhanced microscopy.



Fabio Severini received the B.Sc. and M.Sc. degrees in Electronics Engineering from Politecnico di Milano in 2016 and 2018, respectively. Currently he is pursuing the Ph.D. degree in information technology at the Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano. His research activity focuses on the design and development of CMOS circuits containing SPAD detectors, targeting quantum imaging and quantum communication applications.



Franco Zappa was born in Milano (Italy) in 1965. Since 2011 he is a full professor of Electronics at Politecnico di Milano. His research deals with microelectronic circuitry for single-photon detectors (SPAD) and CMOS SPAD imagers, for high-sensitivity time-resolved measurements, 2D imaging, and 3D depth ranging.

In 2004 he co-founded “Micro Photon Devices”, producing SPAD modules and cameras for single photon-counting and photon-timing, and in 2020 “pioNIRS”, developing near-infrared spectroscopy instrumentation for non-invasive and continuative monitoring of biological tissues and food.



Federica Villa was born in Milan in 1986. She is assistant professor in Electronics Engineering since 2015 at Politecnico di Milano. Her academic research mainly focuses on the design of microelectronic integrated sensors and circuits in CMOS and bipolar-CMOS-DMOS (BCD) technologies for the development of single photon detectors such as single photon avalanche diodes (SPADs) and silicon photomultipliers (SiPMs), photon-counting array imagers with single-photon sensitivity, and high-resolution and linearity time-to-digital converters (TDCs).