A 64 × 64 SPAD Array For Quantum Ghost Imaging with Integrated TDCs and event-driven readout in a 40 nm CMOS Technology

Davide Moschella
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Milan, Italy
davide.moschella@polimi.it

Davide Berretta
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Milan, Italy
davide.berretta@polimi.it

Alberto Tosi
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Milan, Italy
alberto.tosi@polimi.it

Federica Villa
Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano
Milan, Italy
federica.villa@polimi.it

2024 19th Conference on PhD Research in Microelectronics and Electronics (PRIME) | 979-8-3503-8630-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/PRIME61930.2024.10559678

***Abstract*— The exploitation of quantum phenomena has become a pioneering frontier allowing imaging in scenarios where direct illumination is challenging or impractical. Quantum Ghost Imaging (QGI), among various techniques, stands out as a robust and promising approach. While in traditional imaging light interacts directly with the scene, QGI makes use of two separate quantum-correlated beams, one illuminating the object and later being detected by a non-spatially resolved sensor, the other one being straight detected by a pixelated sensor. The scene is then reconstructed by comparing the time tags associated to the detected photons. This work focuses on the design of the spatially solved imager, a 64 × 64 pixel Single-Photon Avalanche Diode (SPAD) sensor. It is implemented with a 3D-stack approach employing the STMicroelectronics C40 technology node for the bottom tier containing the processing electronics, a dedicated imaging technology for the top die embedding the SPAD array, and a microlens array for enhancing the collection efficiency. The integrated 12-bit Time-to-Digital Converters (TDCs) resolution is 200 ps and their stability to process-voltage-temperature (PVT) variations is guaranteed by a shared on-chip wide-tuning range Phase-Locked-Loop (PLL). The event-driven jump readout approach maximizes the acquisition duty-cycle allowing the chip to operate at up to 1.2 Mframe/s.**

***Keywords*— Quantum Ghost Imaging (QGI), SPAD array, 40 nm technology, Time-To-Digital Converter (TDC), Event-driven Readout**

I. Introduction

Measurement techniques that leverage the principles of quantum mechanics have demonstrated their effectiveness in a wide range of practical applications. By harnessing quantum states and exploiting photon correlations, quantum imaging has emerged as a powerful approach addressing challenges in image formation, processing, and detection, often surpassing the capabilities of classical light-based methods. As a matter of fact, quantum sources can feature sub-Poissonian fluctuations or quantum correlation between particles enhancing sensitivity beyond the classical limit imposed by shot noise [1]. One of the most effective methods to produce quantum correlated photons is to generate a two-photon entangled state through Spontaneous Parametric Down Conversion (SPDC). In this process, a non-linear crystal produces a photon pair from a higher frequency photon pump respecting the energy and momentum conservation principles [1].

Entangled photons generated through SPDC have found applications in QGI, a technique that introduces a distinctive separation between object illumination and image acquisition, a concept pioneered by Aspden et al. [2] and gaining significant traction. Notably, QGI operates with an exceptionally low number of photons and allows spectral differentiation between imaging and illumination wavelengths. This technique involves separating the two beams generated by SPDC, called signal and idler beam respectively, and then detecting the idler photons after they pass through the sample using a single-pixel detector. Simultaneously, the signal photons are redirected away from the scene and are sent to a spatially resolved detector. By analyzing the correlations between the two detectors, it

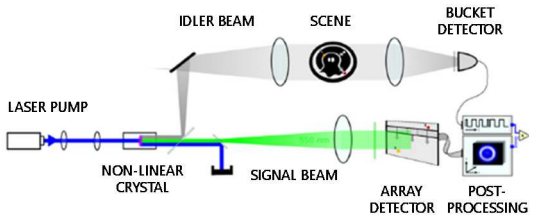

Figure 1: Asynchronous QGI setup [3].

This project has received funding from the European Defense Fund (EDF) under grant agreement EDF-2021-DIS-RDISADEQUADE (n°101103417).

becomes possible to reconstruct either the transmission or reflection profile of the sample. Typical QGI approaches rely on signal and idler photons synchronization through triggering the signal camera upon detecting the idler photon. This kind of setup demands a temporal delay within an image preservation system, impacting accuracy. Moreover, challenges arise in implementing a delay line, especially for extensive imaging processes or remote sensing, as the minimum length scales with complexity and scene distance. Finally, adjustments for varying imaging distances become intricate due to constraints imposed by image preservation.

Reference [3] demonstrated QGI benefits achieved with the so called "asynchronous" approach, shown in Figure 1. This setup requires a coincidence window to be established, both with post-processing comparison of the arrival times of the photons ([3], [4]), or with a proper on-chip logic ([5]). In both cases, the idler photons interacting with the scene are detected by a single-pixel Near-InfraRed (NIR) bucket detector and their arrival time is recorded. The I-CCD camera typically employed in QGI systems is replaced with a silicon SPAD array, able to detect and precisely time-stamp individual photons in the visible spectrum with ps-scale timing precision, eliminating the need for image intensifiers. Time-tagging for each pixel enables post-processing comparison of arrival times between idler and signal photons, distinguishing entangled photon pairs from background and noise.

Albeit satisfactory, the SPAD sensors used in such a setup until now ([3], [4]) had some efficiency limits. First, they are planar arrays, so each pixel contains the SPAD and its electronics and thus both suffer from low fill factor. The system in Ref. [3] even needs a scan process for 2D scenes since it has a linear geometry. Second, they achieve a low maximum duty cycle, defined as the ratio between the observation window and the frame period. This is due to their long readout time, which compels to perform the data acquisition phase after the actual acquisition, thus extending the minimum frame duration and reducing the overall detection efficiency.

Within the European project ADEQUADE (Advanced, Disruptive and Emerging QUAntum technologies for DEfence) we aim to develop a chip suitable to perform quantum imaging measurements, where the SPAD array will include a timing circuit to tag the photons detected in each pixel. The limitations discussed above are overcome by implementing a high resolution 2D array able to reconstruct larger images without the need for scanning the scene and a novel smart event-driven readout architecture, with advantages in terms of measurement time. In addition, fill factor and resolution are enhanced thanks to a 3D-stacking technology. The final chip will be part of a reflection-based QGI setup similar to the one in Figure 1.

II. Proposed SPAD array architecture

The complete chip consists of a 64 × 64 10 µm pitch SPAD array, designed in a dedicated SPAD imaging technology and vertically integrated on a 40 nm bottom die. The bottom tier includes the array processing, organized in 32 × 32 macro-pixels, each connected to 2 × 2 SPADs sharing the timing and readout electronics. The structure of a macro-pixel is shown in Figure 2. The overall chip size is 2.5 mm × 1.3 mm, including the input/output ring and will be fabricated in a 40 nm 3D-stacked STMicroelectronics technology.

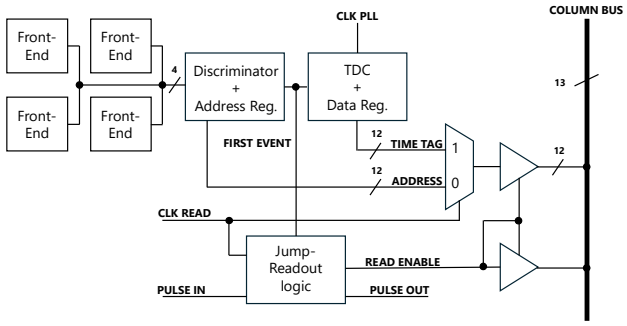

Figure 2: Macro-pixel architecture.

In each macro-pixel, the four SPAD anodes are bonded to their corresponding quenching circuits, which propagate the photon arrival signals to a discrimination logic acknowledging the first event and discarding the next ones. This logic block validates the event as well (if it arrives during the observation gate window). The leading event information is used to generate the address of the hit SPAD and to start the TDC conversion. In correspondence of the gate window closure, the TDC is disarmed, and its 12-bit output is then sampled. The readout is performed in an event-driven fashion (called "jump readout") by means of a specific logic block whose architecture will be explored in the *IV paragraph*. This structure controls the buffers driving the common output high-impedance lines and it is responsible for switching the reading between TDC data and SPAD address. Photon acquisition in one frame and readout of the previous frame happen simultaneously, to maximize the array count rate. The 12-bit column line is complemented by an enable flag indicating to the external tri-state buffer array whether to consider valid or not the value on the bus, as shown in the overall array scheme in Figure 3.

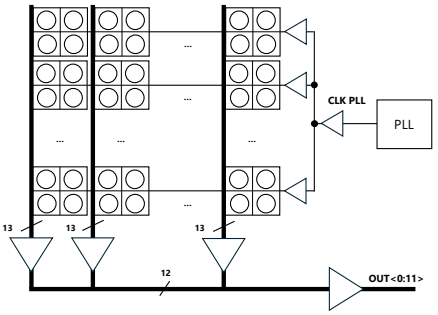

Figure 3: Overall SPAD array organization.

Since in QGI the capability to detect each photon is essential in order not to impair the coincidence detection process, it is important not to mask any event (which would lead to the so-called pile-up distortion). This results in the need for a reduction of the photon rate. A statistical simulation has been conducted and an average array operation of 10% (i.e., only 10% of the macro-pixels should be triggered in each frame and thus only ~100 TDCs) resulted enough to reduce the chance of

pile-up to 0.5%. The proposed sharing of electronics between different SPADs and the choice of an event-driven approach were corroborated by the photon-starved nature of the expected signal. This choice not only reduced power consumption, but also made it possible to implement the discussed architecture within the 10 µm SPAD pitch, maximizing the fill factor.

III. Timing Tag and Address Generation

A highly optimized version of the pixel architecture has been implemented to achieve a 10 µm pitch compatible with the SPAD tier. This optimization involves eliminating any functions unnecessary for the pixel and minimize the layout area when possible.

The front-end is a simple passive quenching circuit. A single thick oxide 3.3 V nMOS is connected to a global analog voltage and is responsible for the quenching of the corresponding SPAD. The sensing node is connected to a level-shifter logic, which converts the anode voltage swing to a 1.1 V digital pulse. The front-end is configurable via shift register to switch off the hot pixels.

The discrimination logic cell interfacing the SPAD front-ends to the Time-to-Digital Converter (TDC) and the address management structure are depicted in Figure 4. The 4 front-end outputs are connected to a gating logic which masks the events arrived outside the observation window and saves the valid events until the gate closure. This is followed by a NOR gate linked to a reset logic whose output switches to *1* (*VALID EVENT* signal) in correspondence of the first event detected and goes to *0* once a new gate window is opened. Four arbiters compare the *VALID EVENT* signal with the 4 gated *EVENT* signals from the gating logic, and their output one-hot code is converted to the fine 2-bit address via a priority encoder. Since the *VALID EVENT* signal is generated by a logic block with an intrinsic delay, it would be possible for multiple photons to arrive in a short time window and trigger more than one arbiter at a time. To avoid contentions, the encoder assigns different priorities to its inputs, thus always granting a stable output. The resulting 2-bit fine address is saved and read along with the 10-bit macro-pixel-hardcoded address.

The NOR gate output *TDC EN* is used as the start signal of the time conversion. When the gate window is opened and a valid event occurs *TDC EN* is *1* activating the conversion process. When the gate goes to *0 TDC EN* goes to *0* as well and the TDC halts. After a small time interval needed to stabilize the TDC output nodes, the frame ends, the TDC conversion is sampled and the data is stored for the duration of the following frame. Once the data is sampled, the gate window can be re-opened and the TDC is reset by *VALID EVENT* going to *0*. The 200 ps resolution TDC core is a 12-bit counter realized with a first dynamic and the following 11 static D-Flip Flops driven by a global 5 GHz clock generated by an on-chip analog PLL. The PLL is a Voltage-Controlled Ring Oscillator (VCRO) based architecture with a large passive RC loop filter. To reduce area occupation, it has been placed externally to the array and shared among all the macro-pixels. The high frequency clock is distributed to the cells via a 2-level buffer tree driving the 32 columns of the array as shown in Figure 3. Albeit being relatively power consuming, this choice has been considered advantageous since the buffer bank consumption is negligible compared to the entire array one during standard operation. To compensate for the clock skew (signal reaching the last column before the first ones), the gate signal (which is the global stop for all TDCs) is distributed similarly, so that the two mismatches are alike and overall there is no skew along the array.

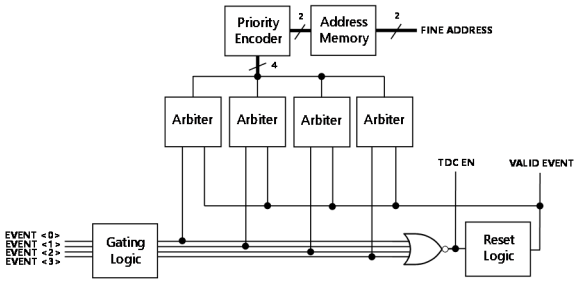

Figure 4: Discriminator block diagram.

Post-layout simulations were run in the most critical process corners. Figure 5 shows an example of regular operation of the data generation blocks of the last macro-pixel, with an address range “4092” - “4095”. The cell is hit by multiple photons, which, since all 4 SPADs are configured to be on, trigger avalanches thus varying the anode voltages. All the photons arrive inside a gate window, therefore they are all considered valid and the first among the four is the one triggering a conversion, while the others are discarded. Since the hit SPAD is the second one, the “01” code is added to the coarse address thus generating the displayed “4093”. When the gate window closes, the TDC is halted and its output stabilizes to “197” in this example, which corresponds to 39.4 ns. In correspondence of the frame pulse, both the conversion and the fine address are sampled and stored until the end of the following frame.

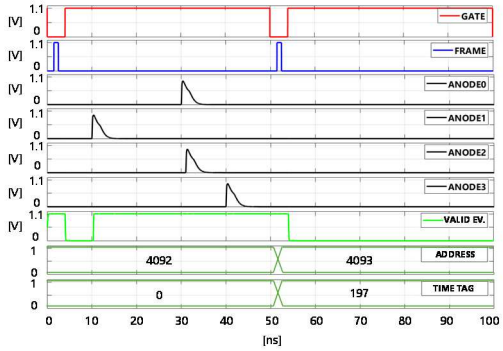

Figure 5: Macro-pixel operation.

Simulations showed that the predicted coincidence window in which multiple arbiters could be triggered exists and its post-layout duration is around 500 ps. Thus, photons arriving during this interval are masked by the one with higher priority, which may not be the actual first. However, the low photon rate expected and the low duration of this window with respect to the full-scale range (FSR ~ 820 ns) make this possible distortion negligible.

IV. Event-Driven Jump Readout

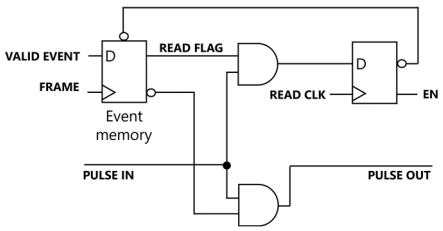

Figure 6: Jump readout approach.

In order to maximize the duty cycle, the readout and the acquisition are performed at the same time. Thus, it is fundamental to adjust the readout time to match the FSR. In addition, the expected photon rate is quite low, thus reading the entire array each frame would bring to acquire mainly worthless data, about 90% of the total. To minimize the readout time and avoid useless data processing, an event-driven readout has been implemented. The proposed architecture, called "jump readout", exploits an event-driven approach involving a scan of the entire array. The empty cells are asynchronously skipped, while the meaningful ones are read synchronously to an externally generated 120 MHz read clock. The architecture of the readout logic inserted in every macro-pixel is shown in Figure 6. The *VALID EVENT* bit is stored during the current acquisition and then stored in another register as a *READ FLAG* for the duration of the following frame. When the frame starts, a voltage step (*INIT*) is generated and sent to the *PULSE IN* input of the first macro-pixel of the array. If the macro-pixel contains data to be read the *READ FLAG* is high thus the bottom AND gate is disabled preventing the *INIT* signal from passing to the next cell. Effective reading happens only when *INIT* is blocked in a cell (thus *PULSE IN* is high while *PULSE OUT* is low) and in correspondence of the rising edge of *READ CLK*. Such a coincidence makes *EN* flag to go high activating the 12 tri-state buffers writing on the column-common bus, while *READ CLK* controls the MUXes choosing whether to output the TDC data or the address. The *EN* flag is buffered and brought out from the array as well to properly control the final readout buffers. The *EN* signal is used to unlock the corresponding AND gate as well, thus allowing *INIT* to go on from cell to cell until it finds another macro-pixel to be read. At the end of the chain composed of all the SPAD array readout cells, the last *PULSE OUT* resets the *INIT* generation logic block thus ending the readout.

This architecture not only prevents useless information from being read and processed, but also greatly reduces the readout time, thus making it possible to implement a pipeline approach without needing for a longer FSR or resorting to stronger power-consuming output buffers.

An example of readout of a row of macro-pixels is shown in Figure 7. The frame window opening triggers the generation of the voltage step *INIT* propagating through the AND cascade. The first 3 cells are read and bring to '1' their corresponding *EN* signal, while at 30 ns multiple macro-pixels are skipped, as it happens again at 60 ns. Eventually *INIT* reaches the end and quenches itself. The architecture has proved to be robust and able to comply with the 120 MHz read clock frequency.

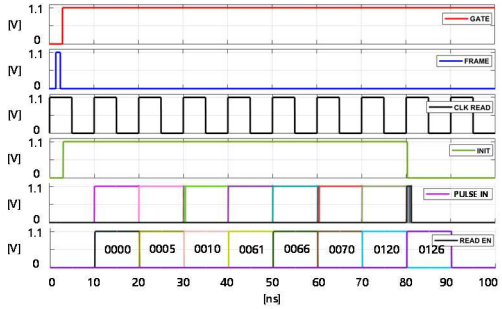

Figure 7: Readout waveforms.

V. Conclusions

We presented a 64 × 64 3D-stacked SPAD array with high fill factor, spatial resolution of 10 µm and time resolution of 200 ps, designed for quantum ghost imaging. In Table 1, the main features of the presented work are compared against other solutions used in QGI setups. The choice of the 3D-stacked 40 nm technology node allowed to reach a high level of integration and good performances. The adoption of various sharing techniques, justified by the low expected photon count rate, granted the achieved low power consumption. Finally, the adoption of a longer FSR and the resort to a smart event-driven approach contributed to the achievement of a high frame rate.

Table 1: Comparison among state-of-the-art chips for QGI

Reference	This work	[3]	[4]
Process (nm)	40	350	150
Pixel pitch (µm)	10.00 × 10.00	40.5 × 200	44.64 × 44.64
Array Size	64 × 64	2 × 192	32 × 32
SPADs/TDC	4	1	1
LSB (ps)	200	312.5	210
FSR (ns)	820	80	50
Readout time (µs)	0.8	1.2	6.9
Acquisition rate (fps)	1.2M	-	250k

References

[1] F. Madonini, F. Severini, F. Zappa, e F. Villa, «Single Photon Avalanche Diode Arrays for Quantum Imaging and Microscopy», Adv Quantum Tech, vol. 4, fasc. 7, p. 2100005, lug. 2021, doi: 10.1002/qute.202100005.

[2] R. S. Aspden *et al.*, «Photon-sparse microscopy: visible light imaging using infrared illumination», *Optica*, vol. 2, fasc. 12, p. 1049, dic. 2015, doi: 10.1364/OPTICA.2.001049

[3] C. Pitsch, D. Walter, S. Grosse, W. Brockherde, H. Bürsing, e M. Eichhorn, «Quantum ghost imaging using asynchronous detection», *Appl. Opt.*, vol. 60, fasc. 22, p. F66, ago. 2021, doi: 10.1364/AO.423634.

[4] C. Pitsch, D. Walter, L. Gasparini, H. Bürsing, e M. Eichhorn, «3D quantum ghost imaging», *Appl. Opt., AO*, vol. 62, fasc. 23, pp. 6275–6281, ago. 2023, doi: 10.1364/AO.492208.

[5] M. Gandola et al., "A 100x100 CMOS SPAD Array with In-Pixel Correlation Techniques for Fast Quantum Ghost Imaging Applications," ESSCIRC 2023- IEEE 49th European Solid State Circuits Conference (ESSCIRC), Lisbon, Portugal, 2023, pp. 105-108, doi: 10.1109/ESSCIRC59616.2023.10268722