Multipurpose, Fully Integrated 128 × 128 Event-Driven MD-SiPM With 512 16-Bit TDCs With 45-ps LSB and 20-ns Gating in 40-nm CMOS Technology

A Carimatto, A Ulku, S Lindner, Eric Gros-Daillon, B Rae, S Pellegrini, E Charbon

Multipurpose, Fully Integrated 128 × 128 Event-Driven MD-SiPM With 512 16-Bit TDCs With 45-ps LSB and 20-ns Gating in 40-nm CMOS Technology

A. Carimatto, A. Ulku, S. Lindner, E. Gros-Daillon, B. Rae, S. Pellegrini, and E. Charbon

***Abstract*—A multipurpose monolithic array of 2 × 2 multichannel digital silicon photomultipliers (MD-SiPMs) fabricated in 40-nm CMOS technology is presented. Each MD-SiPM comprises 64 × 64 smart pixels connected to 128 low-power 45-ps sliding-scale time-to-digital converters (TDCs). The system can operate in two different modes: 1) event-driven and 2) frame-based. The first is suited for positron emission tomography (PET) and the second for synchronous applications like LiDAR. The design includes electronics to capture gamma events by means of a scintillator. The digital readout is fully embedded in the sensor and it is reconfigurable by SPI. Data packets are sent following a simple protocol compatible with an external FIFO, therefore making use of an FPGA optional. Every MD-SiPM can deliver up to 64M time-stamps/s. The sensor can be arranged in any type of configuration through a dedicated synchronization input and can be used to operate jointly with an event generator, such as a pulsed laser, which is useful in many applications. Inherently compatible with 3-D-stacking technology, the sensor can serve as front-end electronics when it is used with a different SPAD silicon tear.**

***Index Terms*—3-D imaging, CMOS, positron emission tomography (PET), SPAD, time-to-digital converter (TDC).**

I. Sensor Architecture

This letter is an extended version of [1] presented in the Symposium VLSI 2018. A monolithic SPAD-based multipurpose system is presented. Fabricated in a new 40-nm CMOS process [2], it was designed to support applications where photon intensity and timing are required along with gating and synchronization. In particular, several features were included to support positron emission tomography (PET). A new time-to-digital converter (TDC) architecture is purposed to improve differential nonlinearity (DNL), integral nonlinearity (INL), and power consumption. Additionally, by making use of 3-D stacking technology, the chip is prepared to work as a common platform to test different SPAD back-ends, thus largely reducing development time in this bond-and-play system.

II. Description of the System

The systems is comprised of four multichannel digital silicon photomultipliers (MD-SiPMs) [3] organized in independent quadrants as depicted in Fig. 1(a). Every MD-SiPM has three main components: 1) a 64 × 64 dual SPAD pixel array; 2) a bank of 128 TDCs; and 3) a digital core that communicates commands and data from and to an external system. Three global skew-free signals (shutter, clock, and reset) are routed along the whole sensor. The SPAD array is divided into 16 panels of 4 × 64 dual SPADs plus the electronics to operate it. Columns in the panel are split in two semicolumns whose SPADs share one TDC to register the arrival time of the photons shown in Fig. 1(b). The electronics that serve the SPAD is composed of five subcircuits plotted in Fig. 2(a): 1) quenching and recharge transistors ($Q1$ and $Q2$); 2) a 1-bit memory that stores the SPAD state (fired or not fired); 3) a self-reset module; 4) a masking memory to shut off the SPAD in case its dark count rate (DCR) is too high; and 5) a transistor ($Q3$) to register the time of arrival through the timing line. The timing line remains in the high state thanks to a pull-up transistor to Vdd until any SPAD in the semicolumn pulls it down after an event occurs. The circuit is presented in Fig. 2(b).

Two modes of operation are possible: 1) frame based and 2) event driven. In frame-based mode, the global shutter signal is opened for a fixed time and the detector captures the events whose information is available after the shutter is closed. This mode is preferred when events are synchronous with the system clock, such as time-of-flight cameras. In event-driven mode, the shutter remains open until an event occurs. This event is defined by the ratio of photons per unit of time, which is an externally configurable parameter. If this condition is not achieved, the self-reset module resets the pixels and TDCs that have fired, thus rejecting DCR and background noise. The level of the latter is defined for the given application; when such levels are reached, an event is detected, the digital core closes the shutter after the predefined integration time [4] the information becomes immediately available. Event-driven operation reduces dead times, throughput, power, and is particularly effective when the events are uncorrelated with the system clock and the shutter (e.g., PET). The

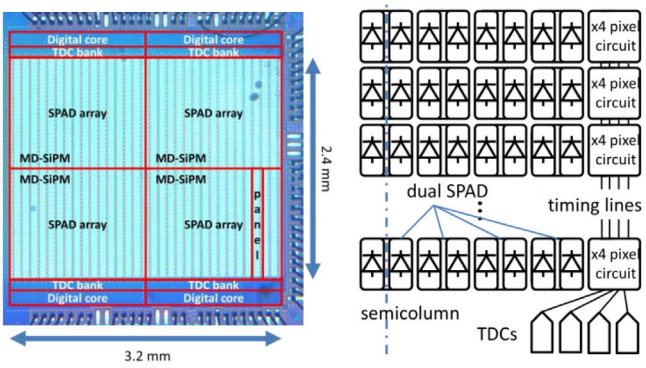

Fig. 1. (a) Micrograph of the sensor: main blocks are shown. (b) Panel.

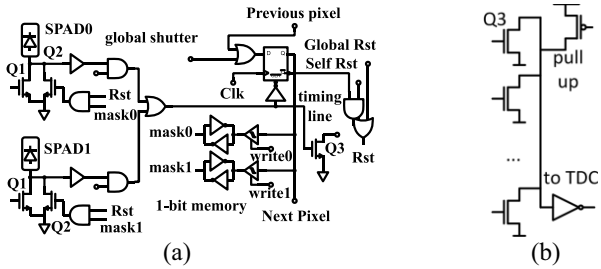

Fig. 2. (a) Pixel circuit: a flip flop is chained with the previous and next pixels in scan-mode fashion to configure the masking memory and read out the data. (b) Connection between SPADs and the TDC in a semicolumn.

Manuscript received December 1, 2018; revised January 24, 2019; accepted February 11, 2019. Date of publication April 26, 2019; date of current version May 17, 2019. This paper was approved by Associate Editor Alvin Leng Sun Loke. This work was supported by NWO under Grant L3SPAD. *(Corresponding author: A. Carimatto.)*

A. Carimatto is with AQUA, Delft University of Technology, 2628 CD Delft, The Netherlands (e-mail: a.j.carimatto@tudelft.nl).

A. Ulku, S. Lindner, and E. Charbon are with the Institute of Microengineering, Ecole Polytechnique Federale de Lausanne, 1015 Lausanne, Switzerland.

E. Gros-Daillon is with MINATEC, LETI, 38054 Grenoble, France.

B. Rae and S. Pellegrini are with Imaging, ST Microelectronics, Edinburgh EH3 5DA, U.K.

digital core performs the control and the readout of the MD-SiPM, which includes: masking, SPAD and TDC reset, configuration for window frame, frame mode, readout of pixels and TDCs, and synchronization operation. The core accepts the commands by serial communication and the readout is a 16-bit 2.5-V CMOS clocked bus that can be connected through a CMOS USB FIFO directly to the PC. A dedicated synchronization input is provided to work with several modules simultaneously. The maximum event rate depends on the mode of operation, the length of the frame, and the activity for the given application. However, it finds its upper limit in the transmission bandwidth of the system. For frame-based mode, the package to be transmitted comprised of TDCs (2560 bits) and pixels (4096 bits). At a frequency of 80 MHz, it takes about 5.2 μs for the bus to transmit these data packages, thus reaching 24.6M timestamps/s and 192 kframes/s. For the event-driven mode, TDCs (2560 bits), pixel addition (32 bits), and global time (32 bits) are sent. It takes about 2 μs to transmit the data packages, thus reaching 64M timestamps/s.

III. Time-to-Digital Converters

A. Design Principles

The TDCs were designed to achieve high time resolution while maintaining low power consumption. Several aspects of the MD-SiPM were considered to choose the architecture; the DCR and the time resolution of the SPADs being the most important. For low-level background light applications, like PET and indoors LiDAR, DCR represents the main source of hits for the TDCs. The probability of a TDC to be triggered by DCR is governed by the Poisson distribution as follows:

$$P(\text{hits} > 0) = 1 - e^{-\lambda(1-M)NT} = 0.89 \tag{1}$$

where λ is the mean DCR, M is the masking factor (estimated 5%), N is the number of SPADs (64), and T is the full range of the TDCs (1 μs). N and M are the two parameters that can be tuned at design time and operation time, respectively, to ensure the TDC availability at any given time during their maximum range. The bin size was chosen considering the time resolutions of state-of-the-art SPADs. The total jitter is given by

$$J_T \cong \sqrt{J_{\text{SPAD}}^2 + \frac{2.2Q^2}{12} + J_{\text{TDC}}^2} \tag{2}$$

where J_{SPAD} is the jitter of the state-of-the-art SPADs (estimated 100 ps), Q is the bin size, and J_{TDC} is the jitter of the TDC. The bin size was chosen to be 40 ps to make the second and third terms negligible with respect to the first term, which is the limiting factor.

B. Architecture

Each MD-SiPM is equipped with a bank of 32 VCOs based on ring oscillators (ROs) that are constituted by nine pseudo differential stages. The RO and stages are shown in Fig. 3(a) and (b), respectively. Every stage counts with two output buffers to prevent any influence of the latching process into the oscillation. The delay of the stages (Δt) that can be configured within 40 and 120 ps by means of an off-chip PLL (implemented in FPGA in this case). The range of the operation frequency [calculated as $1/(18\Delta t)$] extends from 0.46 to 1.38 GHz. The VCOs are phased coupled along the sensor with the main objective of palliating the phase noise as demonstrated in [5]. The phase that is coupled is properly sized to compensate for extra capacitive loads. The coupling, shown in Fig. 3(d), is made through two nMOS–pMOS transistor pairs that connect the internal nodes *inp* and *inn* of the first phase of each VCO(i) to VCO(i–2) and to VCO(i + 2). At the ends, the last VCO and the first one (the reference) make the bridge between the odd and even VCOs. The transistors can be externally controlled to provide the different degrees of coupling to achieve the different performance as explained in [5]. Four logic modules are attached to each VCO, thus totalizing 128 TDCs plus one extra reference TDC. The TDC architecture is shown in Fig. 3(c). The logic modules include five components: 1) a buffer; 2) a 10-bit LFSR counter; 3) a phase sampler; 4) a tri-state bus to read out the information; and 5) a buffer and 1-bit counter.

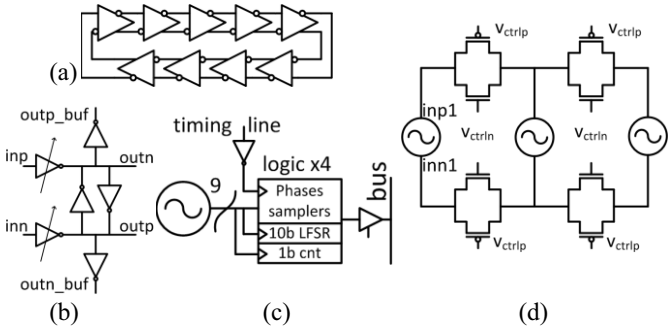

Fig. 3. (a) RO circuit. (b) RO pseudo-differential stage. (c) TDC block diagram. (d) Coupling method: only the first phase of the VCO is coupled.

C. Operation

The TDCs measure the time from the moment an event occurs to the end of the frame by means of a coarse LFSR counter and a fine scale formed by the phases of the VCO. Equation (3) shows the way every time stamp is calculated

$$T = (\text{cnt} + \text{CC}) * N + P_{\text{vcof}} - P_{\text{vcos}} \tag{3}$$

where cnt is the LFSR counter, CC is a correction applied to the counter, N is the number of phases (18), P_{vcof} is the final phase, and P_{vcos} is the start phase of the VCO. Although this is the most natural way to calculate the time stamp, it would double the area and the power of the samplers and the transmission time since every TDC requires two phases to obtain the time stamp. Taking advantage of the fact that the phases are coupled, only one phase (reference) is sampled at the end of the frame. Equation (3) can be rewritten as follows:

$$T = (\text{cnt} + \text{CC})N + (P_{\text{reff}} - \text{PS}(\text{vco})) - P_{\text{vcos}} \tag{4}$$

where P_{reff} is the final phase of the reference VCO and PS is the phase shift between a given VCO against the reference. Though the phases of the VCOs are tightly coupled, there is a small phase shift between every VCO(i) and VCO(i+2) that accumulates over the sensor. By firing all of the SPADs at the same time with a synchronous laser, those phase shifts can be found and stored in a look-up table (PS) for corrections. The detailed operation is as follows: after photon arrival, the timing line is pulled down, and the buffer activates the samplers that latch the phases of the VCO (P_{vcos}). The counter, fed by one of the phases of the VCO (*Pc*), starts running from that moment until the end of the frame when it is stopped by the shutter. The shutter signal is controlled by the digital core and is asynchronous with the VCO. As consequence, if those events happen simultaneously, it is not possible to distinguish whether the counter included the last VCO cycle or not (off-by-one error). In order to mitigate this problem, a 1-bit counter was added. Fed by a different phase (*Pb*), this 1-bit counter can be checked to know if the main counter should be odd or even; the variable CC in (4) will get the values 0 or 1 accordingly. The phases *Pc* and *Pb* are 180° apart to ensure that at least 1 counter is always correct. If *P*vcof is equal to *Pc*, the counter might have incurred in an off-by-one error and should be modified according to the value of the 1-bit counter. If *P*vcof is different from *PC*, the counter is correct and it does not require any correction.

D. Sliding Scale Study

The sliding scale technique is a proven method that has worked very well for ADCs; in this letter, it was used to reduce the DNL of the TDCs. The VCOs are asynchronous with the clock of the system and the window frame, therefore every time stamp taken is measured by a different phase of the VCO. This method can compensate any mismatch in the layout of the ROs, and furthermore any local and global transistor mismatch. In order to calculate the impact of the sliding scale, the VCOs were measured with a random pulsed laser in 300-ns frames. The minimum and maximum DNL were calculated for the start phases of the VCOs and for the timestamps. The results are shown in Fig. 4 exhibit an improvement of 6.25 times.

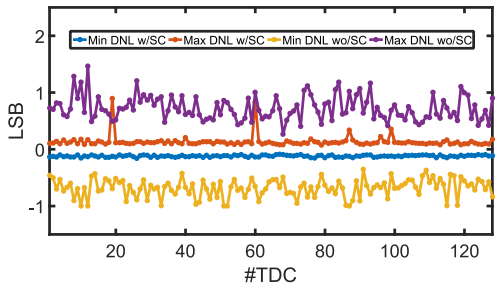

Fig. 4. DNL of the TDCs with and without sliding scale.

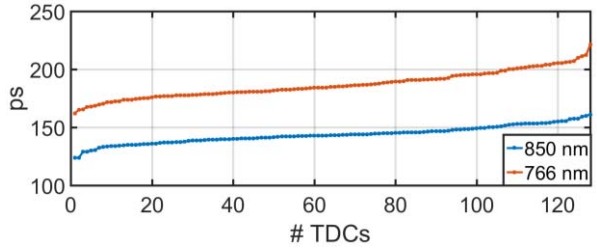

Fig. 5. Time resolution for every pair semicolumn-TDC.

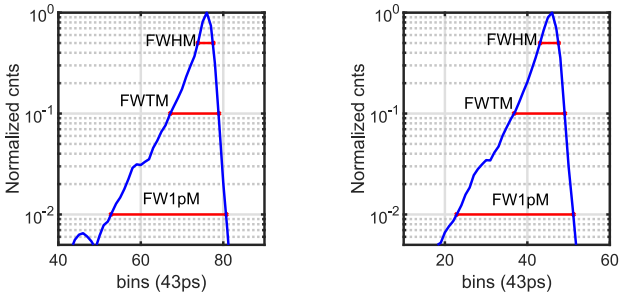

Fig. 6. (a) SPTR 161 ps@FWHM for 850-nm source. (b) SPTR 194 ps for 766 source.

The reader might notice that 2 TDCs have a missing code. This only happens when the maximum frequency is used due to a glitch in one specific bin in the TDC spectrum. However, it does not represent a problem since the probability to encounter this glitch is as low as 0.014% and it only happens for few TDCs. The maximum DNL of those TDCs is still at a level of 0.12 LSB when that problematic bin is discarded. The cause of this effect is an asymmetry in the layout that leads to uneven IR drops in the TDCs supply.

E. Timing Performance of the System

A laser, synchronous with the system, was employed to characterize the time response of every pair semicolumn-TDC. The hits originated by any SPAD in the semicolumn were used to build a histogram and the jitter was calculated at full-width at half maximum (FWHM). The results are shown in Fig. 5.

The total jitter calculated includes the jitter contributions of the laser, SPAD, timing line, TDC, and FPGA. The SPAD is the main contributor. In some applications, particularly in PET, the single photon time resolution (SPTR) is an important parameter to characterize the system as explained in [7]. It essentially describes the uncertainty in time of the whole system when a single photon impinges the sensor. Fig. 6(a) was obtained by measuring 1 million hits and shows the resolution for 850-nm source for FWHM, full-width at tenth maximum (FWTM), and full-width at 1% maximum (FW1pM). Resolution at FWHM is 161 ps.

The same procedure was used to obtain the results for 766-nm wavelength. The SPTR is 194 ps at FHWM, 529 ps@FWTM, and 1.12 ns@FW1pM. Results are shown in Fig. 6(b). Another important

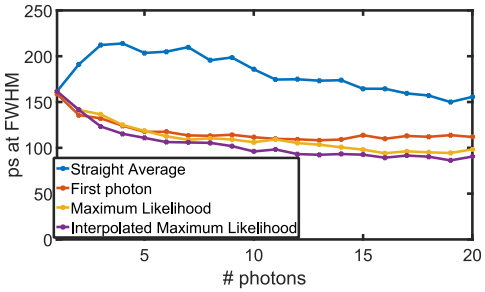

Fig. 7. MPTR: four different methods are shown.

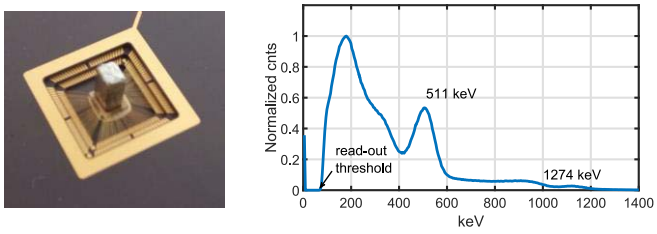

Fig. 8. (a) Sensor-scintillator. (b) ^{22}NA energy resolution.

parameter is the multiphoton time resolution (MPTR) that describes the resolution of the system when several photons impinge any combination of SPAD-TDC. Fig. 7 shows the resolution of the time of arrival @FWHM when multiple photons impact the sensor.

Four different methods were used to calculate the time of arrival. *Averaging* shows a poor result due to the non-Gaussian time response of the SPAD. *The first-photon method* considers only the first photon to calculate the time of arrival. Last, the maximum-likelihood (ML) method exhibits the best result as it accounts for the TDC and SPAD response. It finds the estimator with the highest probability given a set of values. Interpolated ML inserts the response of the TDCs to mitigate the quantization error.

IV. Time-of-Flight Applications

A. Positron Emission Tomography

PET is a noninvasive medical imaging technique to generate a 3-D image of the tissue of interest as explained in [7]. An image sensor can indirectly detect gamma photons (511 keV) by means of a scintillator that absorbs gamma radiation and generates a shower of visible photons that can be time-stamped by the sensor. Fig. 8(a) shows the sensor coupled with an LYSO scintillator.

1) Energy Resolution: The number of photons detected is proportional to the energy deposited by the gamma photon into the crystal. A lower energy than the initial energy signifies that the gamma photon lost energy by scattering meaning it cannot be used for calculation and should be dismissed. Hence, the importance of the estimation of the energy deposited into the crystal. The scattering process, fully described in [8], is ruled by the Compton's law and it can be deduced that the minimal energy that a gamma photon might lose equals 1/3, therefore the resolution must be within that limit. Fig. 8(b) shows the spectrum of a ^{22}Na source, exhibiting an energy resolution of 20%. The first and second peaks of ^{22}Na are shown.

2) Linearity: Some other applications, like spectroscopy, depend on the linearity of the system. This assessment was performed by measuring five different radiation sources with the four MD-SiPMs. A histogram like Fig. 8(b) was built for five radiation sources for each MD-SiPM and its peak coordinates were extracted to build the linearity plot shown in Fig. 9. The peak of ^{22}NA can be seen at ($x = 511$ and $y = 450$). The nonlinearity obtained is 2% and it can be further improved by applying the saturation-correction curve of the MD-SiPM.

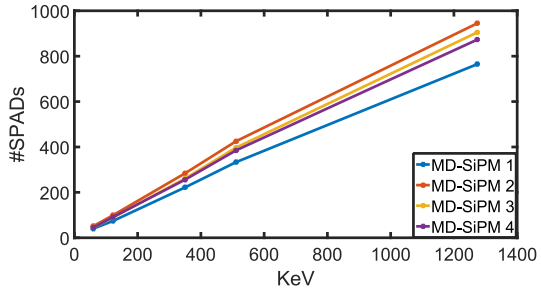

Fig. 9. Radiation linearity of the sensor-scintillator.

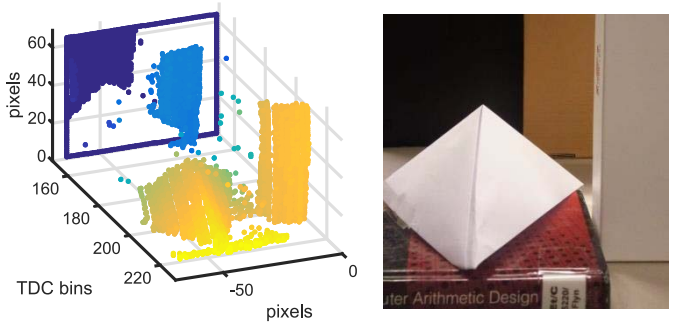

Fig. 10. (a) 3-D picture taken with flash technique. (b) 2-D view from the sensor position.

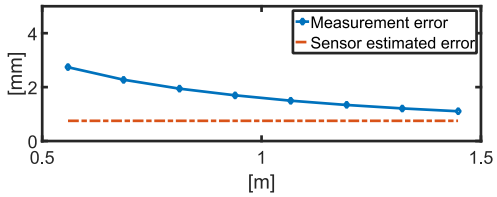

Fig. 11. Blue line shows the error in mm of the measurement and the dashed orange line show the estimated error of the sensor when $M = 1000$.

B. 3-D Imaging/LiDAR

3-D imaging is a topographic method to create a 3-D graphical representation of a physical target. The working principle is based on the illumination of a scene with a pulsed laser and the detection of the photons that reflect off the target. By calculating the time of arrival of these photons, it is possible to create a representation model with X, Y, and depth information. There are two main approaches: 1) flash and 2) scanning methods; both are explained in detail in [9] and [10]. Every photon absorbed by the sensor has a time stamp used to calculate the depth, and its position in the array is used to calculate X and Y in the scene. Fig. 10(a) shows a 3-D image generated by a flash technique; it displays a resolution of 6.5 mm.

For every single photon, the space resolution is about 2.4 cm (160 ps). For flash LiDAR, multiple measurements were performed to improve the resolution by averaging. Assuming that every measurement can be represented by a Guassian distribution with media μ and standard deviation σ, the average of M measurements has a Gaussian distribution with media μ and standard deviation $sqrt(M)/M^*\sigma$. As M increases, the resolution becomes finer and finer. In this example, 1 LSB (43ps -> 6.45 mm) was used to define the spatial resolution. If the previous equations are combined, $M = 13.8$; thus, at least 14 measurements per point are required. The system, working in the frame-based mode, can provide up to 128 depth data per 5.2 μs (1 frame).

C. Distance Measurements, Ranged Method

A pulsed nondiffused laser is pointed to the object whose distance to the sensor is wanted. In this experiment, the shutter is open for the whole range. Fig. 11 shows the error of the measurement for $M = 1000$. For short distances, the parallax problem between the laser and sensor dominates.

TABLE I
STATE-OF-THE-ART COMPARISON

	Braga et al. JSSC 2014	Carimatto et al. ISSC 2015	Frach et al. JSSC 2014	This work
SPAD/SiPM	180 ev/10 ns	416	6396	8192
TDC LSB (ps)	64.5	48.5	23	40
TDC DNL (LSB)	-0.24+0.28	-.75+.75		-0.1 +0.12
TDC INL (LSB)	-3.9+2.3	-2+4		<1
Pitch (μm x μm)	16.2x16.2	30x50	59x64	19x5
SPTR (ps)	266	327	153	162
Power/TDC (μW)	948	1500		171
Energy res (%)	10.2	15	11	20

V. SUMMARY

The performance of the sensor is summarized and compared to other state-of-the-art works in Table I.

VI. CONCLUSION

The first MD-SiPM in 40-nm technology has been designed and presented. Its performance was demonstrated for PET, 3-D vision, and light ranging applications. The new TDC architecture was proven to obtain the expected resolution (161 ps@FWHM) with a low power consumption (12-mW per bank) which makes it suitable for image sensors. The sliding scale technique reduced the DNL of the VCOs (from 0.75 to 0.12 LSB). The level of integration achieved by the digital core largely facilitates the usage of the system and the scaling for synchronous applications. The calibration by means of a synchronous laser that it is needed to find the phase relationship between the oscillators and the reference will be replaced by an electrical calibration in the next version of the chip for simplicity.

REFERENCES

[1] A. Carimatto *et al.*, "Multipurpose, fully-integrated 128×128 event-driven MD-SiPM with 512 16-bit TDCs with 45 ps LSB and 20 ns gating," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, 2018, pp. 73–74.

[2] S. Pellegrini *et al.*, "Industrialised SPAD in 40 nm technology," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, San Francisco, CA, USA, 2017, pp. 1–4.

[3] A. Carimatto *et al.*, "11.4 A 67,392-SPAD PVTB-compensated multi-channel digital SiPM with 432 column-parallel 48ps 17b TDCs for endoscopic time-of-flight PET," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, San Francisco, CA, USA, 2015, pp. 1–3.

[4] M. Chin, M. F. Bieniosek, B. J. Lee, and C. S. Levin, "Integration time window for pulse width modulation readout of silicon photomultipliers for 0.5 mm resolution 3-D position sensitive PET scintillation detectors," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, Seattle, WA, USA, 2014, pp. 1–2.

[5] A. R. Ximenes, P. Padmanabhan, and E. Charbon, "Mutually coupled time-to-digital converters (TDCs) for direct time-of-flight (dTOF) image sensors," *Sensors*, vol. 18, no. 10, p. 3413, 2018.

[6] B. Markovic, S. Tisa, F. A. Villa, A. Tosi, and F. Zappa, "A high-linearity, 17 ps precision time-to-digital converter based on a single-stage Vernier delay loop fine interpolation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 3, pp. 557–569, Mar. 2013.

[7] J. W. Cates and C. S. Levin, "Evaluation of a TOF-PET detector design that achieves ≤ 100 ps coincidence time resolution," in *Proc. IEEE Nucl. Sci. Symp. Med. Imag. Conf. (NSS/MIC)*, Atlanta, GA, USA, 2017, pp. 1–3.

[8] M. K. Nguyen, T. T. Truong, M. Morvidone, and H. Zaidi, "Scattered radiation emission imaging: Principles and applications," *Int. J. Biomed. Imag.*, vol. 2011, Jun. 2011, Art. no. 913893. doi: 10.1155/2011/913893.

[9] M. Beer *et al.*, "1×80 pixel SPAD-based flash LIDAR sensor with background rejection based on photon coincidence," in *Proc. IEEE SENSORS*, Glasgow, U.K., 2017, pp. 1–3.

[10] K. Ito *et al.*, "System design and performance characterization of a MEMS-based laser scanning time-of-flight sensor based on a 256×64-pixel single-photon imager," *IEEE Photon. J.*, vol. 5, no. 2, Apr. 2013, Art. no. 6800114.