A 30-frames/s, 252 x 144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming

Zhang, Chao; Lindner, Scott; Antolovic, Ivan Michel; Mata Pavia, Juan; Wolf, Martin; Charbon, Edoardo

DOI
10.1109/JSSC.2018.2883720

Publication date
2019

Document Version
Final published version

Published in
IEEE Journal of Solid-State Circuits

Citation (APA)
Zhang, C., Lindner, S., Antolovic, I. M., Mata Pavia, J., Wolf, M., & Charbon, E. (2019). A 30-frames/s, 252 x 144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming. *IEEE Journal of Solid-State Circuits*, *54*(4), 1137-1151. Article 8595429. https://doi.org/10.1109/JSSC.2018.2883720

A 30-frames/s, 252 × 144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming

Chao Zhang, Scott Lindner, Ivan Michel Antolović, Juan Mata Pavia, Martin Wolf,
and Edoardo Charbon, *Fellow, IEEE*

***Abstract*— A 252 × 144 single-photon avalanche diode (SPAD) pixel sensor, called Ocelot, is reported for light detection and ranging (LiDAR). The sensor, fabricated in the 180-nm CMOS technology, features 1728 12-bit time-to-digital converters (TDCs) with 48.8-ps resolution (LSB). Each 126 pixels in a half-column are connected to six TDCs through a collision detection bus, which enables effective sharing of resources, and consequently a fill factor of 28% with a pixel pitch of 28.5 μm. The column-parallel TDCs, based on dual-clock architecture, exhibit a DNL of +0.48/−0.48 LSB and an INL of +0.89/−1.67 LSB; they are dynamically reallocated in a scalable daisy chain approach that enables a maximum of five photon detections per illumination cycle per half-column. The sensor can operate in time-correlated single-photon counting (TCSPC) and single-photon counting (SPC) modes, while peak detection (PD) and partial histogramming (PH) are included in the operation of the sensor. The PD and PH modes are enabled by the first implementation of integrated histogramming for a full array via 3.32-Mb SRAM-based PH readout (PHR) scheme providing a 14.9-to-1 compression. Telemetry measurements up to 50 m achieve an accuracy of 8.8 cm and worst-case precision of 1.4 mm (σ). A flash LiDAR using direct time of flight (dTOF) and based on Ocelot is demonstrated, achieving depth imaging at short distances with a frame rate of 30 frames/s, employing an ultra-low power laser with an average power of 2 mW and peak power of 0.5 W.**

***Index Terms*— 3-D imaging, advanced driver-assistance systems (ADAS), autonomous driving, CMOS, direct time of flight (dTOF), histogramming, image sensor, indirect TOF (iTOF), light detection and ranging (LiDAR), self-driven cars, single-photon avalanche diodes (SPADs), time-to-digital converter (TDC), TDC sharing architecture, TOF.**

I. Introduction

Light detection and ranging (LiDAR) systems are a key enabling technology for a wide range of applications, such as augmented reality and virtual reality (AR/VR), facial recognition, assembly line robotics, advanced driver-assistance systems (ADAS), and autonomous driving [1].

LiDAR based on time of flight (TOF) is emerging as a widely applicable method due to its versatility. In this method, the scene under study is illuminated by the pulses of light, some of which will be detected after reflecting from objects in the scene. The distance between the detector and the object can be determined directly [direct TOF (dTOF)] or indirectly [indirect TOF (iTOF)], by measuring the time or the phase between the emission and detection, respectively. In single-carrier iTOF [2]–[5], the illuminator modulation frequency is proportional to the detection range, while inversely proportional to the detection precision. This limits the applicable fields to short range imaging, usually within 20–50 m. In contrast, dTOF [6]–[14] detects the arrival of photons with high sensitivity detectors and measures the photon traveling time with accurate circuitries, such as time-to-digital converters (TDCs), which extend the detection range up to kilometers. Theoretically, the maximum distance is only limited by the optical power.

LiDAR systems can be classified into two categories, scanning, and flash [1]–[14]. The former typically comprises one or more laser and detector pairs, which are mounted on a rotating or vibrating scanner [1]. Due to the increased optical power that can be applied when scanning a laser point across the field of view (FOV), scanning-based systems benefit from improved signal-to-noise ratio (SNR) in comparison to flash. However, the use of a mechanical scanner increases the system size and complexity, while posing concerns for long-term reliability of the moving parts. These concerns are particularly acute for operation in extreme environments, such as in automotive applications. In contrast, flash LiDAR, Fig. 1, illuminates the entire scene in a single shot and detects the light reflected from the scene with an array of detectors. Since there are no moving parts, solid-state LiDAR systems can be built with enhanced reliability. With collimated arrays of lasers, the expected optical echo power reduces as the square of the object distance, which implies single-photon level reflection could be reached at long distances, given sufficient illumination power [9]. In order to detect

Manuscript received August 16, 2018; revised November 9, 2018; accepted November 16, 2018. Date of publication December 28, 2018; date of current version March 25, 2019. This work was supported in part by Netherlands Organization for Scientific Research under Project 12807 and in part by NCCBI (CH). This paper was approved by Guest Editor Ken Chang. *(Chao Zhang and Scott Lindner contributed equally to this work.) (Corresponding author: Chao Zhang.)*

C. Zhang is with the Department of Quantum and Computer Engineering, Delft University of Technology, 2628CD Delft, The Netherlands (e-mail: c.zhang-10@tudelft.nl).

S. Lindner is with the Biomedical Optics Research Laboratory, University of Zurich, 8006 Zürich, Switzerland, and also with the Advanced Quantum Architecture Laboratory, École Polytechnique Fédérale de Lausanne, 2002 Neuchâtel, Switzerland.

I. M. Antolović and E. Charbon are with the École Polytechnique Fédérale de Lausanne, 2002 Neuchâtel, Switzerland.

J. M. Pavia and M. Wolf are with the University of Zurich, 8006 Zürich, Switzerland.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org provided by the author. This consists of three videos in Windows Media Player format. The total size of the files is 17.1 MB.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2018.2883720

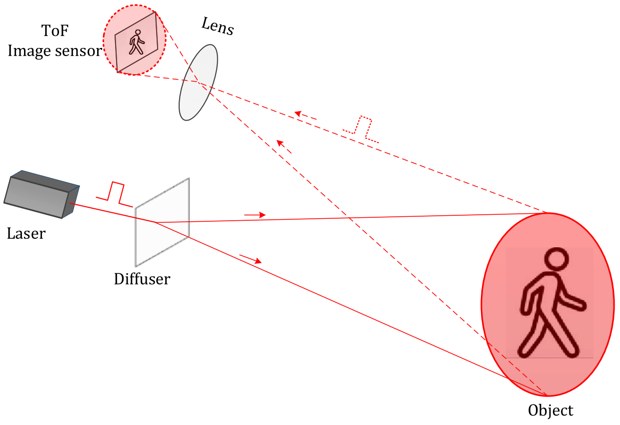

Fig. 1. Flash LiDAR operation diagram in the dTOF mode.

this ultra-low intensity reflection, single-photon avalanche diodes (SPADs) [15]–[19] have been proposed for TOF, as well as other applications [4]–[11], due to their ability to produce a digital pulse from a single detected photon, excellent timing response, and CMOS process compatibility.

In comparison to iTOF, where the distance can be calculated based on the demodulation of the signal intensity, dTOF requires specific circuit blocks, such as a TDC, to measure the time of arrival of incident photons. A range of SPAD sensors with per-pixel TDC architectures have been reported [7], [11], which achieved fully parallel photon detection but at the expense of a large pixel pitch (>40 μm) and low fill factor (<20%). Another major challenge for dTOF with a large pixel array is the considerable volume of data created. Without on-chip processing, this requires a large off-chip data bandwidth or the reconstruction frame rate will be heavily limited by the readout bandwidth. In this paper, we present an image sensor comprising 252 × 144 SPAD pixels and 1728 column-parallel TDCs, fabricated in a 180-nm CMOS technology. Each 126 SPADs in a half-column connect to an address bus with a collision detection mechanism. A TDC sharing architecture was implemented in a dynamic reallocation scheme, where six dual-clock TDCs are shared per half-column, achieving a fill factor of 28% and a pixel pitch of 28.5 μm. To overcome the challenge of memory area overhead we employ a new histogramming scheme, called partial-histogramming readout (PHR), which exploits the intrinsic structure of TOF histograms. In comparison to the existing integrated histogramming solutions [12], [27], which implemented histogramming for lines of macro-pixels, the memory requirements for each pixel are greatly reduced. The PHR scheme enables, for the first time, per-pixel integrated partial histogramming (PH) for a full 2-D array, achieving 14.9-to-1 data compression factor. The sensor was validated in a flash LiDAR system, yielding millimetric detail depth measurement at 30 frames/s at 0.7-m distance with an FOV of 20° × 40°, employing an average and peak illumination power of 2 mW and 0.5 W, respectively.

This paper is organized as follows. In Section II, the sensor architecture and function blocks are explained in detail. Section III shows the sensor characterization and imaging experiments in both 2-D and 3-D. Finally, the conclusion is drawn in Section IV.

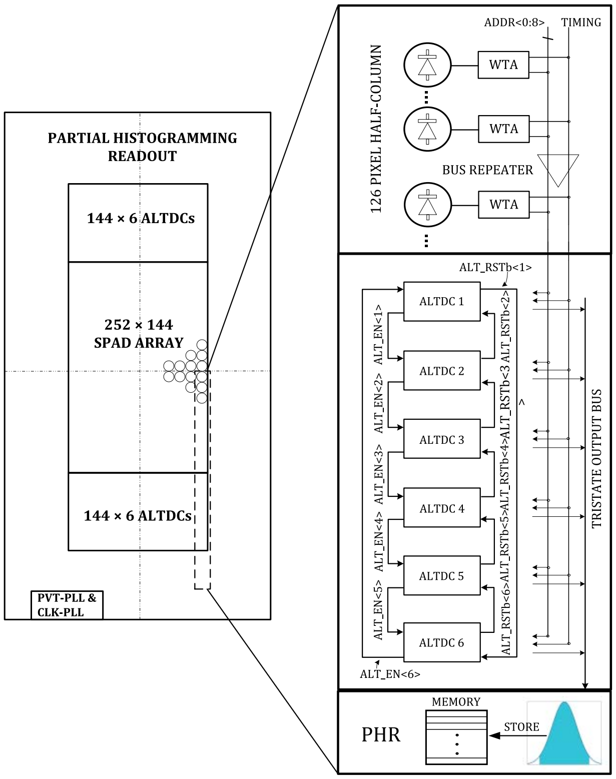

Fig. 2. Sensor architecture.

II. Sensor Architecture

The block diagram of the sensor is shown in Fig. 2. The pixel array is divided into four quadrants, with each sub-array allocating its own timing circuitry, PHR blocks, and data pads. The 126 pixels, which make up a half-column, are connected to a bank of six address latch and TDCs (ALTDCs) to capture the pixel address and to perform the timing conversion for each event. Events are transmitted from the pixels to the ALTDC bank via a shared bus, which employs a winner-take-all (WTA) circuit with a collision-detection coding scheme. This means that when two or more coincident events occur in different pixels, the address present on the bus is invalid, thus allowing collisions to be identified and then rejected by the readout. To reduce the rate of collisions, the output pulses from the pixels are temporally compressed to minimize the amount of time that each event occupies the bus, which is known as the bus dead time. Bus repeaters are distributed throughout the bus to maintain a narrow pulsewidth.

At the bottom of the half-column, the shared bus lines are connected to all six ALTDCs in the bank, which are connected in a daisy chain fashion. Events are distributed to the ALTDCs using a dynamic reallocation approach [20], where a single ALTDC slice in the chain is available to capture events at any one time. The capture of an event results in the next slice in the chain being activated, thus enabling multiple events to be captured on each cycle. The timing conversion is

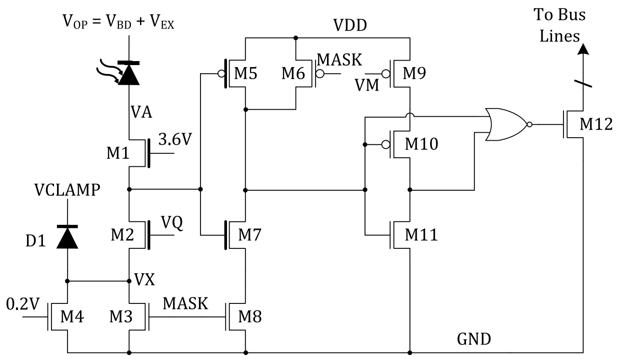

Fig. 3. Pixel schematic with cascoded quenching circuit, which enables up to 5.2-V excess bias with only transistors. Compact layout was implemented, achieving 28% fill factor at a pixel pitch of 28.5 μm.

performed by an open-loop ring-oscillator (RO) TDC based on a dual-clock architecture. Process, supply voltage, and temperature (PVT) compensation of the TDC is performed by a 2.56-GHz phase-locked loop (PLL), while the on-chip clocks of 320 and 240 MHz are generated by a separate 960-MHz PLL.

In time-resolved SPAD sensors, digitizing the pixel address and timestamp of every photon generates a large volume of data to process and read out from the sensor. Therefore, achieving fast measurements with a large number of pixels is a major challenge. Instead of histogramming the full range of the TDC on-chip [12], we have implemented a PHR scheme, which exploits the intrinsic structure of time-correlated single-photon counting (TCSPC) data to perform integrated histogramming for every pixel in the sensor array. Data are read out from the sensor via 72 160-MHz general purpose input/output pads for a total bandwidth of 11.52 Gbits/s.

A. Pixel Design and Collision Detection Bus

The sensor employs an SPAD with a p-i-n structure reported in [19]. A schematic of the pixel is shown in Fig. 3. It consists of a circular SPAD with 17.15-μm-diameter active area and circuitry for SPAD quenching, pixel masking, and pulse shrinking. The cathode of the SPAD is connected to a high voltage bias $V_{\mathrm{OP}} = V_{\mathrm{BD}} + V_{\mathrm{EX}}$, where V_{BD} is the breakdown voltage, and V_{EX} is the excess bias of the SPAD. For the SPAD, in this paper, the photon detection probability (PDP) and timing performance can be improved greatly by increasing the excess bias voltage, with only a small degradation in dark count rate (DCR) [19]. In this design, high V_{EX} was achieved with a cascoded passive quenching circuit [21], implemented with M1 and M2. M1 is a thick oxide NMOS, biased at 3.6 V, which allows the SPAD to operate at excess bias voltages of up to 5.2 V without exceeding the 3.6-V maximum voltage tolerance of any device. This scheme extends the maximum excess bias voltage of the SPAD without the need for area intensive resistors [22] or high voltage process options [23]. Due to the compact circuit implementation and the absence of in-pixel TDCs, the above reported fill factor and pitch were achieved.

High noise or "hot," pixels are disabled by configuring the MASK signal to low. The value of MASK is set independently for each pixel with an in-pixel 6T-SRAM, while the configuration is managed by row and column shift registers on the array periphery. As well as masking the electrical output of the SPAD with M6 and M8, M3 increases the SPAD quenching resistance, thus greatly increasing the recharge time. However, due to the very large resistance of the quenching branch formed by M1–M3, leakage current from the SPAD could cause the voltage at VA to rise above the maximum voltage tolerance of M1. To limit the voltage at VA and prevent the breakdown of M1, a parallel discharge path is provided by M4, which, with its gate at 0.2 V, is biased in the sub-threshold region. A diode D1 acts as a clamp should the voltage at VX to rise above 1.8 V, preventing the breakdown of M3 and M4. The advantage of this scheme is that optical crosstalk from hot pixels is reduced due to their greatly diminished count rate.

Since all pixels in a half-column are connected to a shared bus, each firing pixel will occupy the bus for a set period, the aforementioned bus dead time. For valid event detection, only one pixel can occupy the bus at a time. This implies that the bus dead time must be minimized. As such, a monostable circuit, consisting of M9–M11 and a NOR gate, is included in each pixel to reduce the bus dead time below that of the SPAD. By adjusting the voltage VM, the pulsewidth at the output of the monostable can be set in the range of 0.4–5.5 ns from a post-layout simulation in the typical corner. In a system with an optical illumination frequency of 40 MHz, the cycle period is 25 ns. Since the SPAD dead time is typically on the order of tens of nanoseconds, e.g., 50 ns, one SPAD could occupy the bus for a time longer than the cycle period. Therefore, by reducing the SPAD output onto the bus with the monostable circuit, the pulsewidth of each event is reduced and can occupy the bus for less than 1 ns. This enables multiple photons to be detected in each half-column, per cycle.

The shared bus is similar to that implemented in [24], where a number of shared address lines and a single shared timing line are used to transmit events to the ALTDCs. When no pixels have fired and the bus is idle, all bus lines are pulled up to "1" via a set of PMOS pull-up transistors, which have a fixed gate bias, VPU. Each pixel includes a set of NMOS pull-down transistors to transmit the pixel address and timing signal by pulling the bus lines low. This pull-down of the bus results in a current flow for the duration of the pulsewidth and hence a bus power consumption which is dependent on pixel activity. Upon the reception of a photon, a current of 1.1 mA and 700 μA flows in each bus section for the TIMING and each ADDR line, respectively. Despite having different drive strengths, the worst-case skew between the TIMING and ADDR lines is limited to ±250 ps.

In the event that a second pixel detects an event during the bus dead time, the code present on the address lines will be a collision of the two pixel addresses, with every line pulled low via an active NMOS. This is an issue for binary coding, as these collisions will result in an incorrect address and thus, an invalid detection. Using the example of a 3-bit bus, if two pixels with addresses "110" and "101" fire simultaneously, the merged output code would be "100," a pixel that did

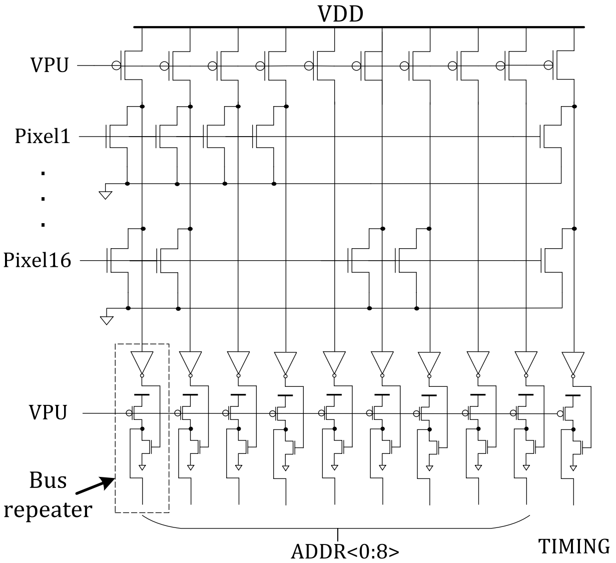

Fig. 4. Schematic of the first section of the collision detection bus and repeaters.

TABLE I

CODE TABLE FOR 16 PIXELS IN THE FIRST SECTION

Pixel	ADDR<0:8>	Pixel	ADDR<0:8>
1	000011111	9	110000111
2	001001111	10	101000111
3	100001111	11	011000111
4	010001111	12	001010111
5	000101111	13	100010111
6	001100111	14	010010111
7	100100111	15	000110111
8	010100111	16	001110011

not fire. For this reason, a collision detection coding scheme is implemented as in [24]. The total number of collision detection codes m for n-bits lines is given by the following equation, where k is the number of "1"s:

$$m = \frac{n!}{k!(n-k)!}. \tag{1}$$

In order to obtain the maximum number of codes, k is chosen to be the integer closest to $n/2$. In this design, address buses with nine lines were implemented, where each address consists of five "1"s and four "0"s, enabling 126 pixels being coded. When collisions occur in this coding scheme, the detected address will have more than 4 zeros. Therefore, collisions can be detected and are then rejected in the PHR readout. In contrast with binary coding, where only 7 bits are required to address 126 pixels, the coding efficiency is reduced with the collision detection coding scheme.

With 126 pixels per half-column, the total length of each bus is 3.6 mm which presents a large capacitance to the pull-up and pull-down transistors which drive the bus. However, to achieve low jitter and a short bus dead time, sharp and narrow pulses are required to propagate through the bus. Of course, the pull-up and pull-down transistors could be scaled to minimize the rise and fall times of the bus; however, the large transistor sizes would severely impact pixel fill factor. Therefore, in this design, bus repeaters were used to divide the bus into eight sections where the section closest to the ALTDC array has 14 pixels, and the remaining seven sections have 16 pixels. A single section including bus repeaters is shown in Fig. 4, with the pixel addresses coded in Table I. By replicating the pull-down behavior of the pixels and including another pull-up transistor, the bus repeaters divide the larger bus into a set of mini-buses. Since the capacitance of each section is reduced by a factor of 8, for a given signal transition time, the size of the pull-up and pull-down transistors also decreases by 8, while requiring only a single bus repeater per line. Thus, this method maximizes the fill factor.

For efficient area use, a small space is reserved in each pixel for the placement of 1 bus repeater. The bus lines are then repeated in sequence as the signals propagate through each section. With only 10 bus lines required to encode the address and timing information, the remaining six reserved spaces are occupied by decoupling capacitance. This also illustrates the scalability of the method in that the bus can be increased to greater pixel numbers with very little overhead. Each additional address line requires one more bus repeater per section as well as an extra pull-down transistor per pixel.

B. Dynamic Reallocation ALTDCs

To perform time-resolved measurements from a large array of pixels in parallel, many sensors have employed a TDC-in-pixel approach. In monolithic technologies, this results in large pixel pitch and low fill factor, e.g., 19.48% fill factor for 44.64-μm pitch [11] or 1% for 50 μm [7]. In addition, in many applications pixel activity rates must be restricted to 1%–3% to limit distortion due to photon pileup [25]–[29]. In this case, valid data are only sparsely distributed throughout the array. This is a major challenge for the readout.

To overcome these problems, some works share a single TDC with a number of pixels [6], [28], [29]. This method has the benefit of reduced area occupation by timing circuitry and simplified readout of data. However, in these architectures, the detection throughput is limited due to the fact that on each cycle only the first event per sub-group is detected as the TDC is occupied by this event until the conversion is complete. To improve the throughput, in this paper, we implement a dynamic reallocation architecture [20], to perform pixel AL and TOF measurement.

The block diagram of the ALTDC bank is shown in Fig. 2, where six ALTDCs are connected in a daisy chain configuration and enabled sequentially according to the bus activity. As such, all six ALTDCs are connected to the shared address and timing signals from the bus. At any one time, only one ALTDC is enabled for AL and timing measurement. Thus, in this sharing architecture, the six ALTDCs can detect multiple photons from the half-column on every cycle.

A simplified schematic of the ALTDC is shown in Fig. 5, where each ALTDC is enabled by ALT_EN$\langle i-1\rangle$ from the previous block and reset by ALT_RSTb$\langle i+1\rangle$ from the

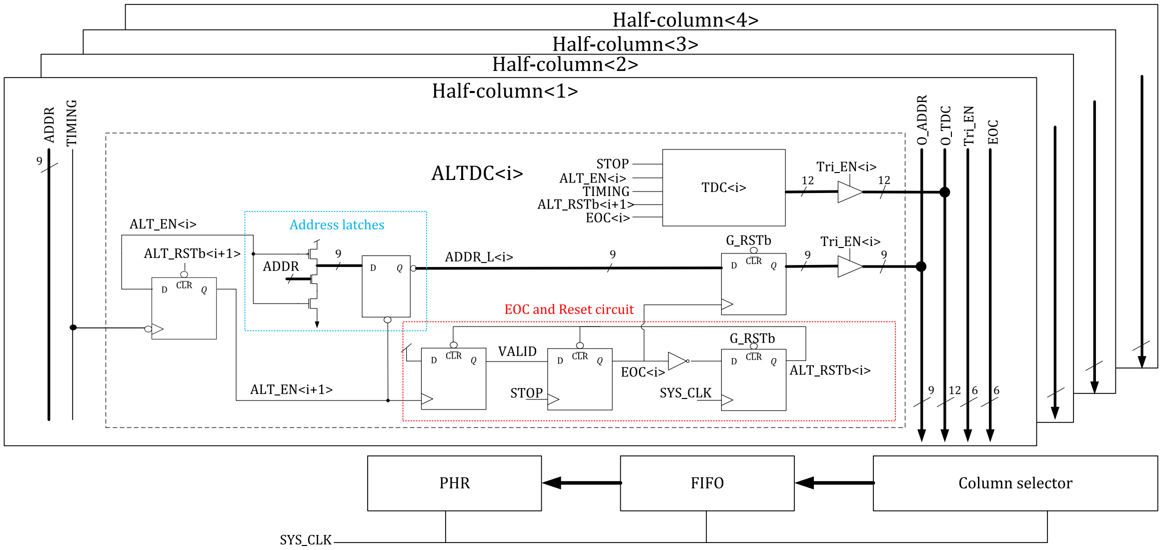

Fig. 5. ALTDC block diagram, where each PHR is shared by four half-columns.

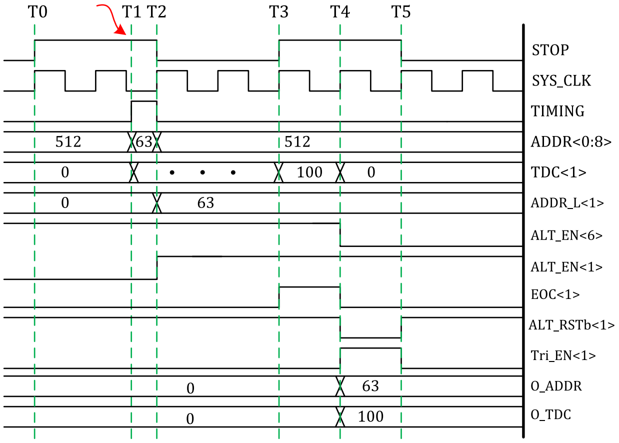

Fig. 6. ALTDC timing diagram. After global reset, ALT_EN⟨6⟩ is set to high, which enables ALTDC⟨1⟩ for photon detection (T0). Upon a photon impinging, TDC⟨1⟩ starts conversion and ADDR is captured by the dynamic logic (T1). At the falling edge of TIMING, ALTDC⟨2⟩ is activated by ALT_EN⟨1⟩ and is ready for detection; ADDR is latched in ADDR_L⟨1⟩ (T2). At the rising edge of STOP, EOC⟨1⟩ is generated to stop TDC⟨1⟩ conversion and signal data ready to the readout (T3). At the rising edge of SYS_CLK, Tri_EN⟨1⟩ is asserted to enable ADDR_L⟨1⟩ and TDC⟨1⟩ readout through the tri-state buses O_ADDR and O_TDC, respectively (T4). The data on the tri-state buses is written into FIFO for PHR processing (T5).

subsequent block. Half-columns are grouped into sets of 4, where each group is accessed by a column selector which reads out events in an event-driven fashion and then writes the data into an FIFO for PHR processing. Fig. 6 shows the timing diagram associated with photon detection by ALTDC⟨1⟩ which is enabled after a global reset with some extra peripheral logic. When a pixel detects an event, a short pulse is generated on the TIMING line; the rising edge of the pulse then begins the conversion of TDC⟨1⟩. The pixel address, ADDR, propagating through the bus together with the TIMING signal, is captured by a set of dynamic logic. At the falling edge of TIMING, ALT_EN⟨1⟩ rises to logic high, which: 1) enables ALTDC⟨2⟩ for photon detection; 2) latches the address to ADDR_L⟨1⟩; and 3) triggers VALID signal to begin event-driven readout process. At the end of the cycle, the TDC conversion is completed by the rising edge of STOP; signal EOC⟨1⟩ is generated to indicate the availability of valid data and latches the address and TDC data into registers for readout. The readout block is synchronized with the system clock SYS_CLK, which is phase aligned with STOP to make sure that the EOC signal can be sampled correctly. With EOC⟨1⟩ high signaling the capture of TOF data, Tri_EN⟨1⟩ is asserted by column selector, and ALTDC⟨1⟩ is readout through two tri-state buses, O_ADDR and O_TDC. Since the event-driven readout method is applied, no power is dissipated communicating null events, which is the typical case for TDC in-pixel architectures [7].

To prevent the entire chain being reset, there is always one inactive ALTDC, which limits the maximum number of photons that can be detected in 1 cycle to 5. To the best of our knowledge, the first instance of TDC dynamic reallocation was demonstrated in [30], which used a “token-passing” scheme to handle TDC allocation. In this method, 16 TDCs, which are arranged in pairs enabled on interleaved cycles of the STOP clock, measure up to eight photons per cycle. This interleaved operation is required to ensure continuous operation while accommodating for the TDC dead time. The TDCs and token passing registers are reset with common reset signals. In our implementation, we extend the dynamic reallocation to also capture the pixel addresses while the ALTDC chain enables TDCs to be reset individually, only once they have performed a conversion. As such, it is not required to use pairs of TDCs to operate continuously.

The minimum time between photons that can be detected is limited by two factors: ADDR/TIMING pulsewidth, equal to the bus dead time, and the propagation delay of ALTDC. Due to the load capacitance mismatch between TIMING and ADDR buses, pulses will be skewed in time, which limits the minimum pulsewidth that can be used to latch the addresses correctly. The performance of the monostable circuit was confirmed with indirect electrical measurements at a minimum width of 0.8 ns, where both the TIMING and ADDR signals are sampled correctly. Both the rising and falling edges of a pulse were captured with on-chip flip-flops and the time-delay between the edges measured with an oscilloscope. The propagation delay of the ALTDCs is simulated in post-layout in the typical corner, in which the ALT_EN⟨1⟩ propagates from the bottom ALTDC to the top one, giving a maximum delay of about 700 ps. Therefore, a minimum photon interval of 1.5 ns is achieved.

C. Dual Clock Time-to-Digital Converter

Due to the large number of on-chip TDCs and extensive logic employed in the PHR, the power consumption of the TDC is critical. As such, an open-loop RO-based architecture is employed to mitigate against the need for distributing multiple phases of a clock [8], [25], across the sensor. However, the PHR constrains the design of the TDC in comparison to conventional RO-based approaches [7], [24], because each PH is constructed with data from six TDCs, i.e., six individual timing histograms. For a given time of arrival, then the frequency mismatch in the open-loop oscillators will result in a deviation in the codes from the six TDCs. For longer measurement periods, these deviations will accumulate, and the peaks of the six timing histograms disperse in time. This would pose a challenge for the peak location and PH functions.

To reduce the code dispersion of the six TDCs, a new dual-clock TDC architecture is implemented here, which decreases the ON-time of the open-loop RO to a single period of a second reference clock, STOP_HF, which is an integer multiple of the laser frequency, STOP. This is in contrast to conventional RO-based approaches [7], [24], where the open-loop RO oscillates for the period of STOP in the worst case. A circuit schematic of the dual-clock TDC and timing diagram are shown in Figs. 7 and 8, respectively. The TDC is enabled by ALT_EN from the previous slice in the ALTDC chain, T1 shown in Fig. 8. This signal also gates the STOP_HF clock into the block, enabling TDC nodes to be toggled by STOP_HF_INT. This reduces the power consumption of those TDCs in the column without a valid ALT_EN bit. At the next rising edge of TIMING, T2 shown in Fig. 8, the RO_EN signal enables the RO, which runs at a nominal frequency of 2.56 GHz. An NMOS source follower is connected in series between the TDC power supply and the RO to reduce the impact of IR drops on the RO frequency. A 4-bit counter counts the rising edges of the RO until the first rising edge of STOP_HF_INT, T3 shown in Fig. 8. This asserts STOP_HF_REG after which RO_EN is driven low and a 6-bit counter is then used to count the rising edges of STOP_HF_INT. On the next rising edge of STOP, T4 shown in Fig. 8, the EOC signal is asserted, signaling the end of conversion, and the 6-bit counter stops counting. The final TDC code is obtained from the frozen phase of the RO (3 bits), the edges of the RO counted by the RO counter (4 bits), and the edges counted by the STOP_HF counter (6 bits). Although this code is 13 bits, the most significant bit of the RO counter, TDC_D⟨6⟩, is an "overflow" bit. Due to mismatches in the RO frequency, some TDCs will run faster than 2.56 GHz. In such cases, the RO counter will exceed 7 in a single period of STOP_HF_INT. As such, an extra bit is required to capture the minority of codes that exceed this count.

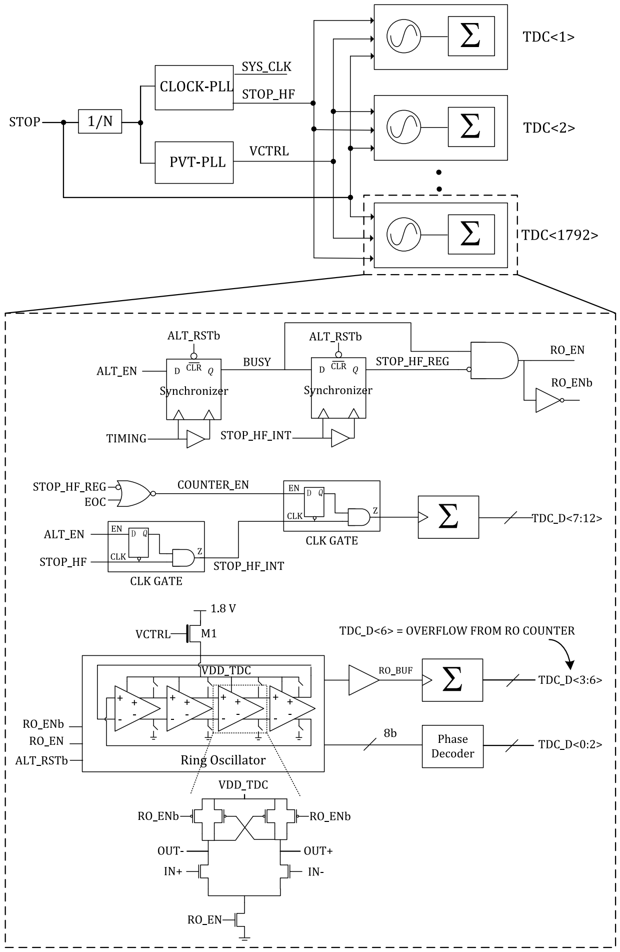

Fig. 7. Dual clock TDC architecture and schematic.

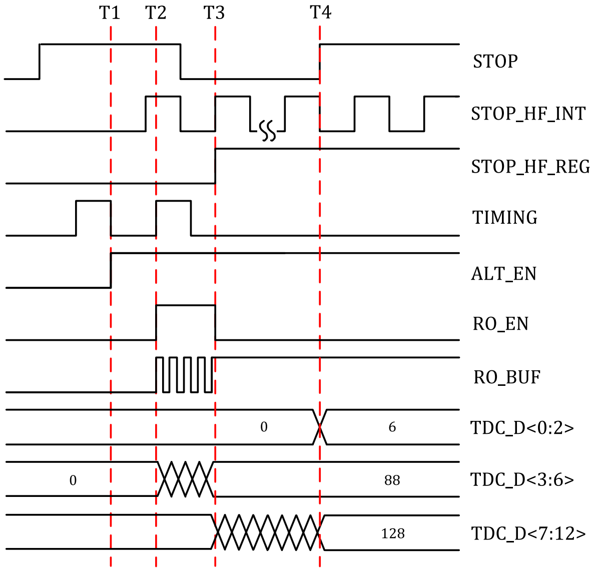

Fig. 8. TDC operation timing diagram.

With STOP_HF = 320 MHz, two open-loop ROs with frequencies of 0.99 × 2.56 GHz and 1.01 × 2.56 GHz, will accumulate a maximum code difference of 1.28 LSBs. As well as limiting the code dispersion of the six TDCs, this TDC architecture can achieve lower power consumption of the TDC array in comparison to the single-clock RO architecture in high activity cases. When TDC activity levels are high enough to compensate for the power dissipated in the STOP_HF clock tree, this dual-clock architecture benefits from lower power consumption per conversion due to the reduced ON time of the RO.

To compensate the frequency of the open-loop ROs for PVT variations, the control voltage of the RO in all TDCs, VCTRL, is generated by a PLL with a replica RO locked at 2.56 GHz. In principle, this same PVT-PLL could have been used to generate the 320-MHz STOP_HF clock. However, to maximize the range of STOP frequencies the sensor can operate at, a configurable frequency divider is employed to divide the STOP signal down to 2.5 MHz to be used as the reference for the PLL. This means that the desired clock frequencies can be generated with STOP frequencies of 80, 40, 20, 10, 5, and 2.5 MHz, by selecting the appropriate divide ratio. With a type-II charge pump PLL, this setup restricts the PLL loop bandwidth which should be at most 1/10th of the PLL reference frequency. In this case, a conservative bandwidth of 125 kHz was chosen. With this loop bandwidth, the phase noise of the RO (−81.36 dBc/Hz at 1-MHz offset from the carrier at 2.56 GHz, simulated) results in an rms jitter of 38 ps, which is of the same order as the SPAD, and will sum in quadrature with the other components of the system timing response, e.g., SPAD jitter and laser full-width at half-maximum (FWHM). Although not strictly critical for LiDAR, a narrow system timing response FWHM is desirable for many applications. Improving the jitter performance of the RO would require increased power consumption and since the RO is also embedded in the TDCs this is undesirable. Thus, to minimize the system timing response of FWHM, a second PLL was designed, CLOCK-PLL in Fig. 7, to generate the STOP_HF and SYS_CLK clocks. The PVT-PLL then performs PVT compensation of the TDC RO while the CLOCK-PLL generates the STOP_HF and SYS_CLK clocks with a dedicated RO with reduced phase noise (−98.61 dBc/Hz at 1-MHz offset from the carrier at 960 MHz, simulated).

D. Partial-Histogramming Readout

For SPAD sensors, a major benefit is the potential for designing large pixel arrays. This is a fundamental requirement for flash LiDAR. However, a large pixel array implies massively parallel time-resolved measurements resulting in a large volume of data. Since the data are typically transmitted off-chip for further processing, the output data bandwidth of the sensor can heavily limit the speed of measurements. For example, a 252×144 pixel operating at 1% pixel activity with a 40-MHz laser frequency would result in a required output data bandwidth of approximately 300 Gbits/s. This is impractical for a number of reasons, including high power consumption and high number of data pins.

To overcome this bottleneck, rather than streaming out the full raw data, full histogramming readout (FHR) has been implemented in [12], [26], and [27], to accumulate photons for each bin of the TDC on-chip. Since the size of the histogrammed data is much smaller than that of the raw format, high compression efficiency can be achieved and photon rates up to 16.5 GS/s have been reported [27]. However, these sensors have thus far been limited to single point or line formats. This is due to the large memory overhead required to capture all bins in the TDC for a large number of pixels. For example, with a 6T-SRAM cell size of 4.65 μm^2 in a 180-nm technology, 1024 5-bit bins for every pixel in a 252×144 array would require an impractically large silicon area of 864 mm^2. In the case of histogramming TDCs [26], [27], this problem is even more pronounced due to the use of ripple counters to implement the histogram memory.

In this design, we implement an on-chip SRAM-based histogramming method, which we refer to as the PHR. This readout exploits the fact that the events, which are time-correlated with the laser, are confined within a narrow range of histogram bins. Rather than building a histogram of the full TDC range, high compression efficiency can be achieved by only histogramming photons within this narrow range. Due to the greatly reduced memory requirements of this method, per-pixel histogramming can be implemented for a large format sensor.

The PHR operation can be divided into two processes, peak detection (PD) and PH. The PD mode detects the histogram peak location for each pixel, while the PH mode is used to build the PH. The PD and PH processes employ two SRAMs, referred to as PEAK_SRAM (10 bits per pixel) and HIST_SRAM (80 bits per pixel), respectively. A block diagram of the two processes is shown in Fig. 9.

The PD is a three-step approximation process, where the searching resolution of each step is improved until the peak is located. In each step, the range is sub-divided into eight sections, and a histogram is built with the PHR SRAM, which is configured as 8 bins of 10 bits per pixel. Assuming ambient light is uncorrelated, it will distribute uniformly across all bins in the histogram. The region containing correlated photons reflected from the scene will have a greater count, allowing it to be detected. In the first step of PD, photons are accumulated in this histogram considering only the most-significant bits $Q\langle 9:7\rangle$, thus locating the section peak T1. With a TDC LSB of 48.8 ps, this results in a resolution of 6.2 ns per bin. In the second step, the region T1 is inspected with a resolution of 780 ps considering $Q\langle 6:4\rangle$, thus locating the section peak T2. Finally, T3 is located by constructing a histogram of region T2 with a resolution of 97.6 ps by considering $Q\langle 3:1\rangle$. The peak is determined with $Q\langle 0\rangle$ as "0" and is stored in the PEAK_SRAM for readout and PH. Since the photons acquired in each step of the peak location cannot be reused in the following steps, three exposure periods are required to locate T1, T2, and T3. This stands in contrast to full-range histogramming (FRH), where all photons acquired can be accumulated into the histogram. Another tradeoff in comparison to FRH is that the background noise tolerance is degraded in the PD stage as the large number of bins which

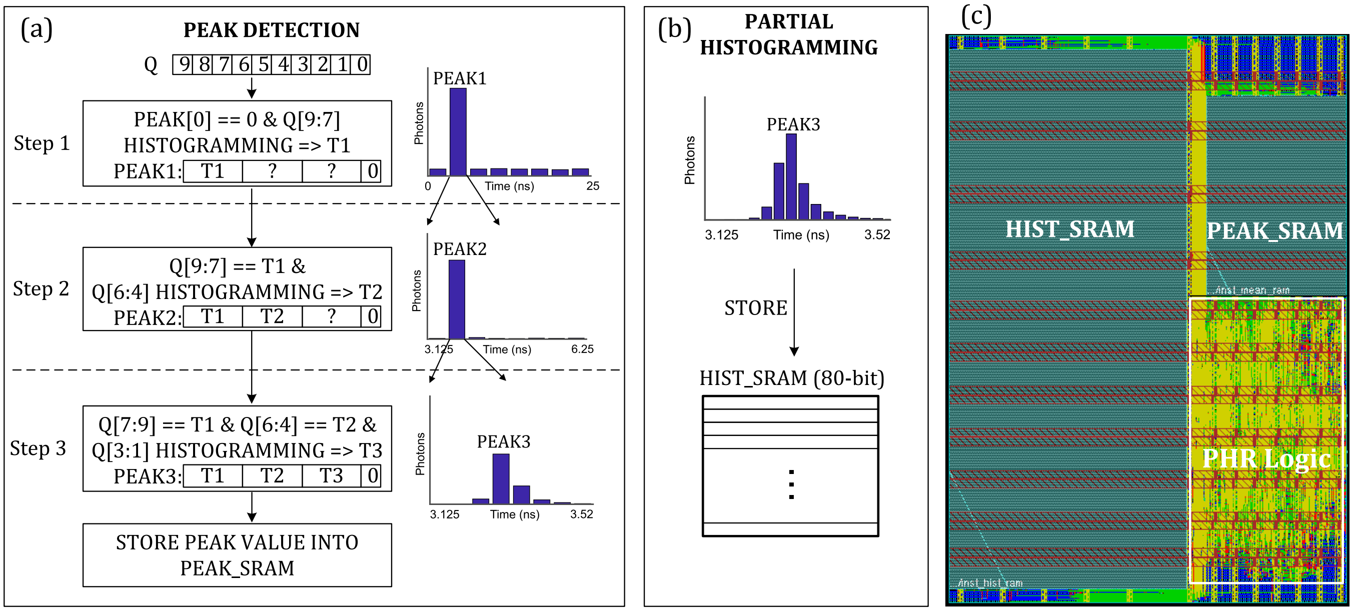

Fig. 9. PHR block diagram. (a) PD requires a three-step successive estimation of the histogram peak. (b) PH stores a configurable window of 16 bins around the peak. (c) PHR block layout, where 68% of the area is occupied by the memory.

capture background light and noise are merged with those containing the signal. While in comparison to PD based on the phase-domain $\Delta-\Sigma$ approach [31], this method has the benefit of detection reliability. If multiple peaks exist in the histogram, e.g., due to multiple reflections, the largest peak will be detected. In contrast, in the $\Delta-\Sigma$ approach, the peaks will be averaged leading to a significant error.

Once the peak is located, the PHR can be operated in the PH mode with the HIST_SRAM configured as 16 bins of 5 bits for each pixel, where a configurable 16-bin window is formed around the peak. Events with a timestamp within the window are stored in the corresponding bins of the histogram in SRAM and read out periodically before the bins overflow. At the same time, photons lying outside with range will stream out as raw data via the I/O pads while the PHR is accumulating events. By combining the in-range PH and out-of-range events, the full-range histogram can be reconstructed. Therefore, the sensor is also suitable for applications requiring the complete timing response, which may span over a range of nanoseconds, e.g., FLIM and NIROT [25]–[27], [32].

Since the number of correlated photons returning to the camera decreases exponentially with distance, the SNR also decreases with distance. A higher SNR can be achieved with a coarser TDC LSB, due to the merging of bins. For this reason, the 12-bit TDC code is shorten into 10 bits with three different ranges and LSBs, including short range of 50 ns with 48.8-ps LSB, medium range of 100 ns with 97.6-ps LSB, and long range of 200 ns with 195.2-ps LSB.

Since the SRAM peripheral circuits, such as sense amplifiers and row/column decoders, are shared by all the memory cells, the memory density is increased with the capacity. To reduce the chip level overhead for the SRAM peripherals, instead of PHR block per half-column, one PHR block is shared by four half-columns with 504 pixels, which employs one 40-kb HIST_SRAM and one 5-kb PEAK_SRAM. So, for the entire sensor, in total, 72 PHR blocks were implemented, comprising 2.95-Mb HIST_SRAM and 0.37-Mb PEAK_SRAM. The layout of one PHR block is shown in Fig. 9(c), with a dimension of 1.3 mm × 0.94 mm, where the SRAM memory occupies 68% of the area.

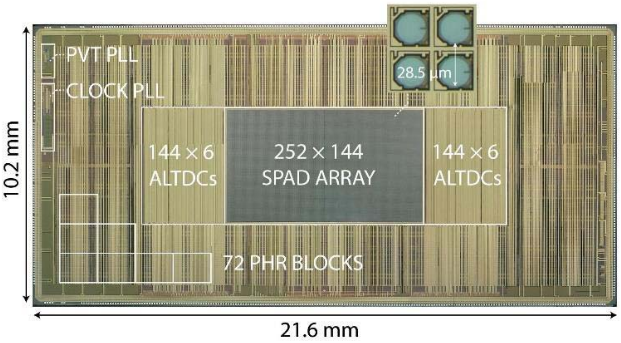

Fig. 10. Chip microphotograph. Inset: 2 × 2 cluster of pixels.

E. Chip Realization and Measurement System

The sensor was fabricated in a 180-nm CMOS technology and occupies an area of 21.6 mm × 10.2 mm. The microphotograph of the chip is shown in Fig. 10. Approximately, 70% of the area is occupied by the PHR blocks, which is due to the SRAM and large amount of logic implemented with a mature technology. Although area intensive in this design, since the PHR is entirely digital, the architecture can scale very well in more advanced nodes.

In order to characterize the sensor, a measurement system was designed, comprising five printed circuit boards (PCBs), including the mainboard with the sensor mounted as chip-on-board, a power board, a breakout board for signal probing and

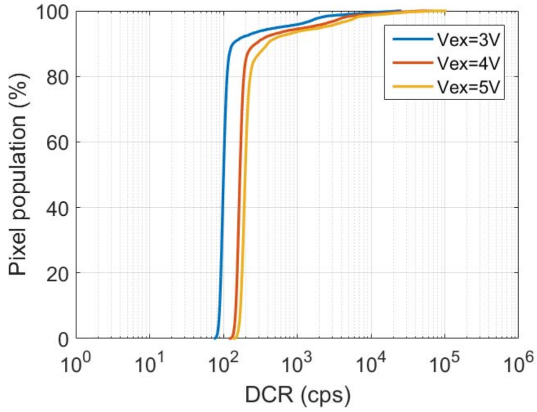

Fig. 11. DCR population density at different excess voltages.

debugging, and two field-programmable gate array (FPGA) boards (Opal Kelly XEM7360 based on Kintex-7), where each one handles half of the chip. FPGAs were used for sensor configuration, readout, and data transmission to the computer via high-speed universal serial bus (USB) 3.0. To reduce data acquisition time and processing complexity, 128 out of 144 columns were read out and processed.

III. Results

A. Pixel Characterization

The breakdown voltage of the SPAD was measured at 22 V, confirming the result published in [19]. The DCR of the entire array has been characterized at room temperature without an external cooling system. The DCR was measured with a dead time of 50 ns, and the population density with excess voltage in the range of 3–5 V is shown in Fig. 11. A median DCR of 195 Hz was achieved with 5-V excess bias voltage, corresponding to a DCR density of 0.84 Hz/μm^2. Based on the results published in [19], a median DCR of about 80 Hz would be expected in the same condition. We believe the DCR increase is mainly due to the self-heating of the sensor under normal operation. Low DCR variation is achieved, with 93.8% of pixels having a DCR of less than 1 kHz at 5-V excess voltage.

The PDP has been reported in [19], which achieved wide spectral sensitivity with PDP greater than 40% from 460 to 600 nm at 11-V excess voltage. For the near-infrared region, high PDP of 9.32%, 5.27%, and 3.19% was achieved with 5-V excess voltage at 840, 900, and 940 nm, respectively, thus providing more flexibility for LiDAR applications. Although photo-response non-uniformity (PRNU) has not been characterized for this array, measurements on a 32×32 array with the same SPAD and pixel [20] have shown a PRNU of 1.18% with 5.3% of noisy pixels masked, which is likely dominated by breakdown voltage non-uniformity [33]. The afterpulsing probability (AP) was measured with 25-ns dead time at 5-V excess bias voltage, as shown in Fig. 12, where no afterpulsing is observed. The negligible AP, in this design with respect to [19] is due to the integrated quenching circuit that significantly reduces the capacitance of the anode and thus the number of carriers crossing the device during the avalanche [34].

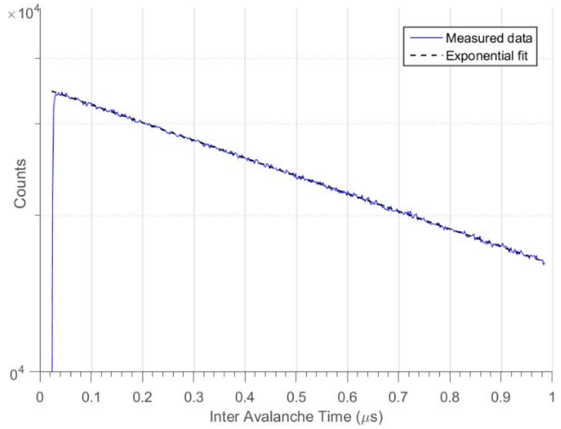

Fig. 12. AP measurement at 5-V excess voltage with 25-ns dead time, where no obvious afterpulsing is observed.

B. Time-to-Digital Converter Characterization

To measure the non-linearity of the TDCs, the sensor was illuminated with uncorrelated light, ensuring the probability of receiving a photon is less than 1 per cycle. Under these conditions, events are uniformly distributed over the full range of the TDC. Triggered by SPADs, the TDCs were characterized with the code density test method, where the STOP signal is generated with the FPGA. The DNL and INL can be calculated with code histogram statistics. The TDCs operate with a nominal LSB of 48.8 ps, where STOP_HF = 320 MHz and voltage-controlled oscillators (VCOs) oscillate at 2.56 GHz. The TDC non-linearity was measured with a 20-MHz reference signal, as is shown in Fig. 13. From the measurement, a periodic DNL/INL non-linearity error is observed every 64 bins, at the transition between the 4- and 6-bit counters. This is due to two factors. First, although the ROs are biased by VCTRL from the PVT-PLL, small frequency offsets are present in the ROs due to random device mismatch. Second, there is a non-negligible jitter associated with STOP_HF. These issues result in some bins with very few events. TDC calibration was performed by redistributing photons from the regions, where the photon counts are less than half of the median count in the TDC histogram, to the closest earlier bin. The worst-case DNL (INL) was reduced from +0.22/−1 (+2.39/−2.6) LSB to +0.48/−0.48 (+0.89/−1.67) LSB after calibration.

With the dynamic reallocation scheme, each photon impinging, an SPAD can be detected by any TDC in the half-column. Since the mixed TDC response is used for PD and PH in the PHR processing, the uniformity of TDCs is important to have accurate TOF measurement. The characterization of SPAD-TDC timing response of one half-column is shown in Fig. 14, where the sensor was illuminated with a short pulsed laser with 40-ps FWHM at 637-nm wavelength. A 3.01 LSB (146 ps) FWHM jitter of mixed TDC response was achieved with the pixel at the center of the array, Fig. 14(a), where the jitter of each individual TDC was in the range of 2.50 LSB (122 ps) to 2.87 LSB (140 ps). The jitter distribution of 126 pixels in a half-column is shown in Fig. 14(b), where excellent uniformity is achieved with the average and standard deviation of 3.03 LSB (148 ps)

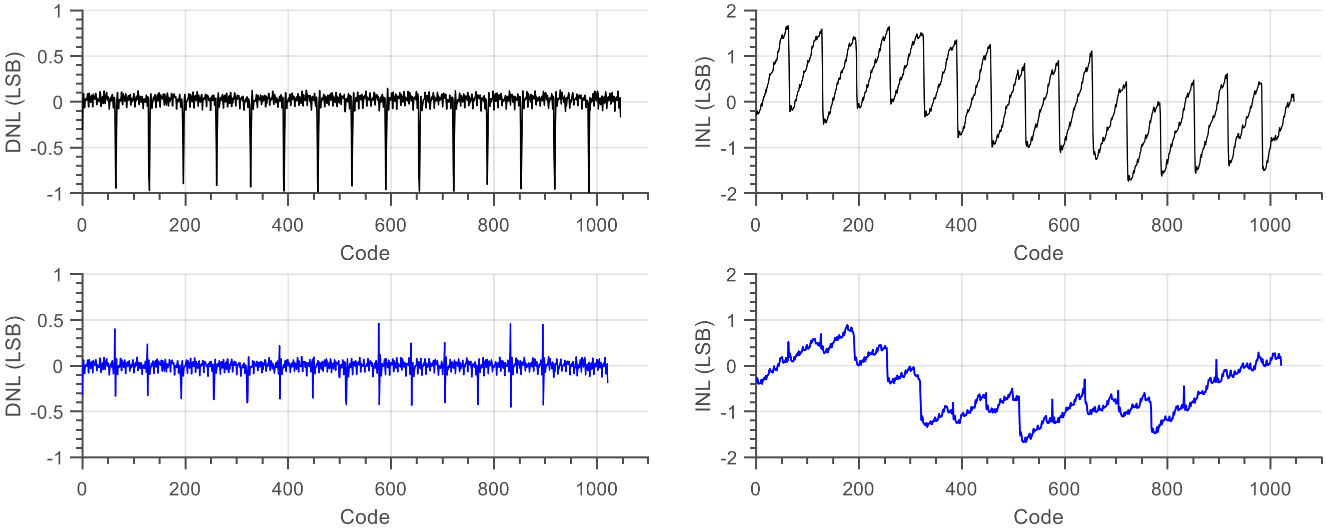

Fig. 13. DNL and INL of one TDC in a half-column which resulted in the worst-case peak-to-peak INL after calibration. Raw DNL and INL are shown in black (top), while calibrated DNL and INL are shown in blue (bottom).

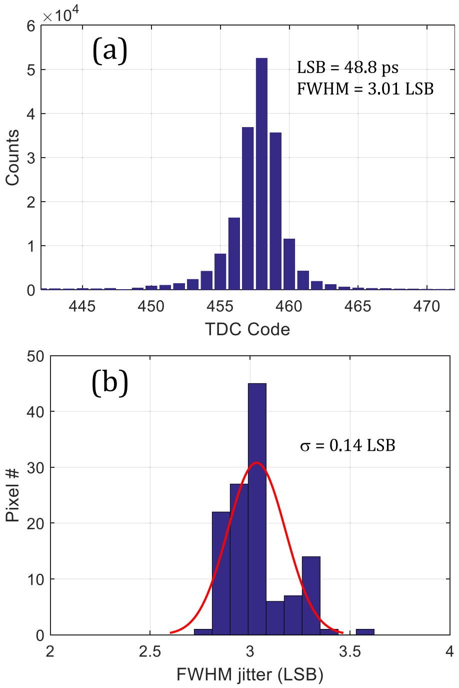

Fig. 14. (a) Single-shot FWHM jitter of 3.01 LSB with mixed TDC response. (b) Jitter distribution among 126 pixels in a half-column, a standard deviation of 0.14 LSB was achieved.

and 0.14 LSB (6.8 ps), respectively. No obvious degradation of jitter is observed due to the signal propagation through the complete length of the collision detection bus and ALTDC chains. This result indicates that the shared bus architecture could be extended to larger formats without significantly degrading the timing performance of the sensor.

C. 2-D and 3-D Imaging

In order to enable both the 2-D and 3-D imaging capability of the sensor, a camera system was built with a 25-mm objective ($f/1.5$) placed in front of the sensor. In the 2-D imaging mode, the PEAK_SRAMs are configured as 10-bit resolution counters per pixel. The sensor works in the global shutter mode with an I/O speed of 160 MHz, leading to a readout time of 32 μs, thus a maximum frame rate of 31.25 kframes/s. To achieve both high speed and low light level imaging, a frame time of 1.5 ms was used, achieving a frame rate of 666 frames/s. As is shown in Fig. 15, a fan with three blades rotating at 1300 r/min was recorded with half of the chip at illumination levels of 0.1 and 10 lx. Although the image quality at 0.1 lx is Poisson-limited, the edge of the blades can still be ascertained due to the high photon detection efficiency.

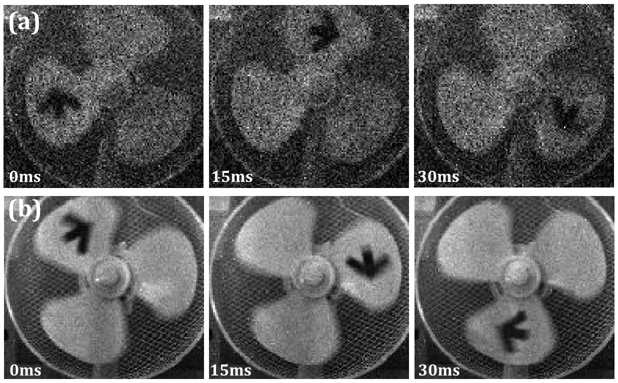

Fig. 15. 2-D movies of a rotating fan at 1300 r/min with light of (a) 0.1 and (b) 10 lx. The gray scales for (a) 0–30 counts and (b) 0–300 counts.

Time-resolved ranging measurements were performed with a 637-nm pulsed laser at 40-MHz repetition rate, 2-mW average power, 0.5-W peak power, and 40-ps pulsewidth at FWHM. At this laser frequency, the unambiguous range that can be measured is 3.75 m. However, the sensor still can be characterized with a larger range, by exploiting prior knowledge of the distance offset. In such a way, the TDCs traversed multiple times and the linearity and precision of the system

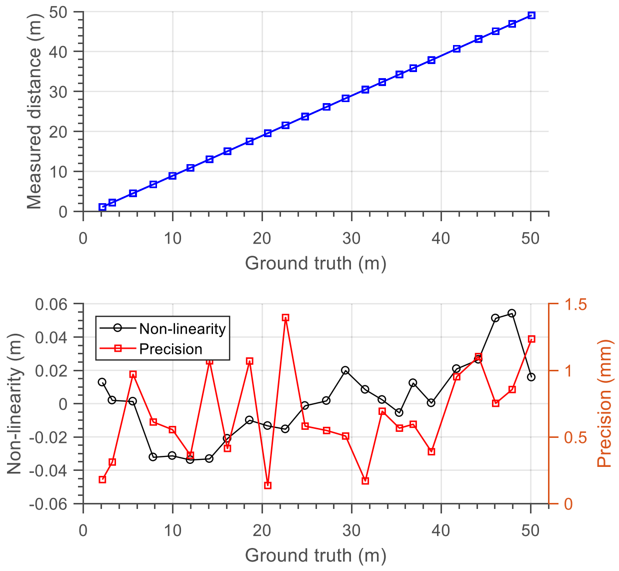

Fig. 16. Non-linearity of depth measurement with a 60% reflectivity target up to 50 m, using 16 × 128 subset pixels from the same section of the collision detection buses of the main array.

was characterized and shown in Fig. 16. A 60% reflectivity target was measured up to 50 m, where each distance was measured with a 30 000 photons histogram for 10 repeated measures, revealing a maximum (peak-to-peak) non-linearity of 8.8 cm, and worst-case precision (σ) of 1.4 mm, over the entire range.

In order to demonstrate the depth imaging with the PHR scheme, including PD and PH processes, a 252 × 128 flash 3-D image was acquired at a distance of 1 m with intensity data superimposed, as shown in Fig. 17. A diffuser was placed in front of the laser to illuminate the scene uniformly with a 20° diverged circular beam. Since the sensor FOV is 20° × 40°, the measurement was performed in a sequence of eight exposures, illuminating different sections of the mannequin. Due to the limited laser power, the image was obtained in dark conditions to maximize the SNR. The profile of cross-sectional A–A′ is shown in Fig. 17(b). The coarse curve is drawn with the peaks acquired in PD step with a minimum spatial resolution of 1 LSB; fine resolution is achieved by averaging the PH of each pixel, which resolves the image with millimetric detail. The time offset associated with the bus repeaters is calibrated by performing an initial ranging measurement with a flat white board in front of the sensor. The TOF difference between different bus sections is measured and then used to calculate the delay associated with each bus repeater. The information is stored in a look-up table to perform a depth calibration in 3-D imaging. At present, this calibration does not account for any temperature drift that could occur, e.g., for the most distant pixel, the simulations show a temperature rise from 25 °C to 85 °C results in an increase in bus propagation delay of 78 ps. In the future, this drift could be included in the calibration using measurements from a temperature sensor.

One example of the PH is shown in Fig. 17(c), which is taken in a pixel from the nose. The compression factor of the PHR scheme was measured at high pixel activity of approximately 200 kHz, where the sensor was illuminated directly by the laser. To avoid bin overflow, the histogram was read out every 0.5 ms, which enables a compression factor of 14.9-to-1, based on the comparison with the equivalent data size transmitted in the TCSPC mode. This high compression factor can be employed to reduce the power consumed by the I/O pads for data transmission and to increase the image acquisition speed of the sensor II.

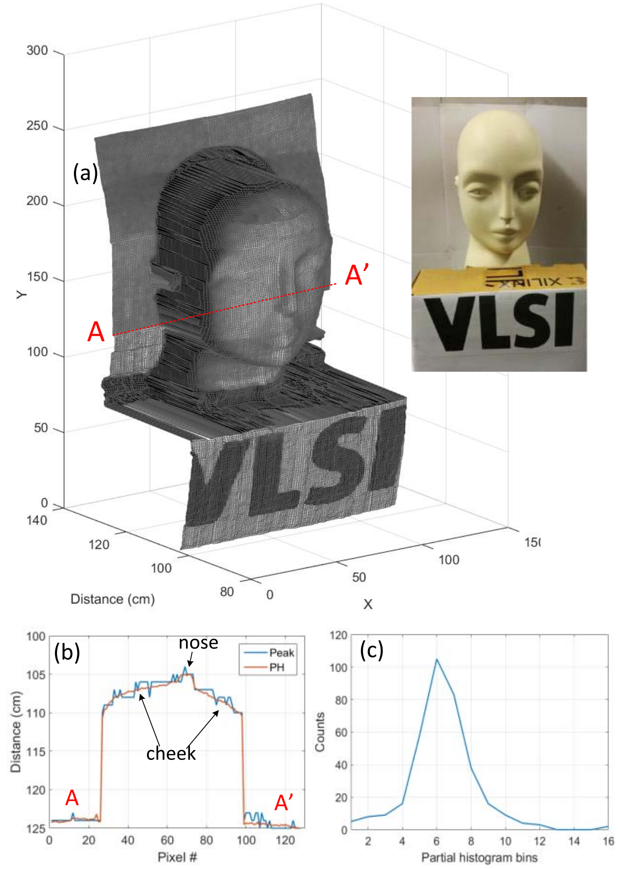

Fig. 17. (a) 252 × 128 3-D flash image using PHR scheme with intensity superimposed. Note the insensitivity of depth map from the reflectivity of the writing "VLSI" in the pedestal. Median filtering with a neighborhood size of 2 × 2 was applied. (b) Profile of cross section of A–A′, drawn with the peak (coarse resolution) and averaged PH (fine resolution). (c) PH of a pixel on the nose.

TABLE II
POWER CONSUMPTION OF THE SENSOR IN PHR MODE

Components	Power (mW)	Contribution (%)
PHR digital core	1252	49.3
ALTDCs	838	33
I/O	198	7.8
PLLs	176	2.9
Pixel array	74	6.9
Total	2538	

Since the time-of-arrival statistics of each pixel are built on-chip, the workload of the FPGA is reduced dramatically, allowing 3-D imaging to be performed in real-time. Fig. 18 represents six successive frames of a 3-D movie acquired at 30 frames/s at 0.7-m distance with half

TABLE III

COMPARISON OF PERFORMANCE WITH STATE-OF-THE-ART SPAD SENSORS

Parameters	Unit	**This work**	Ximenes et al, ISSCC, 2018 [13]	Perenzoni et al, JSSC, 2017 [9]	Niclass et al, JSSC, 2014 [12]	Villa et al, JSTQE, 2014 [8]
			Sensor characteristics			
CMOS Technology	-	**180 nm**	45/65 nm	150 nm	180 nm	350 nm
Integrated histogramming	-	**Per-pixel, partial histogram**	N/A	N/A	Per-pixel, full histogram	N/A
Pixel array	-	**252×144**	16×8	64×64$^{(5)}$	16 × 1 TOF pixels$^{(5)}$ 32 × 1 intensity pixels	32×32
Pixel pitch	µm	**28.5**	19.8	60	21	30
Fill factor	%	**28**	31.3/50.6$^{(3)}$	26.5	70	3.14
Median DCR @ V_{EX}	cps	**195 @5V**	5.3k@2.5V$^{(4)}$	6.8k@3V	2.65k@N/A	120@5V
TDC depth	bit	**12**	14	16/15	12	10
TDC resolution	ps	**48.8**	60 - 320	250 - 20000	208	312
TDC power	mW	**0.3**	0.5 - 0.1	N/A	N/A	N/A
TDC area	µm^2	**4200**	550	N/A	N/A	N/A
TDC number	-	**1728**	1	4096	64	1024
TDC linearity	DNL(LSB)	**+0.48/-0.48$^{(1)}$**	+0.8/-0.7	+1.2/-1	+0.15$^{(4)}$/-0.17	+/-0.06
	INL(LSB)	**+0.89/-1.67 $^{(1)}$**	+3.4/-0.8	+4.8/-3.2	+0.32/-0.56	+/-0.22
			LiDAR measurement			
Image resolution	-	**252×144**	256×256	64×64	202×96	32×32
Imaging type	-	**Flash**	Scanning	Flash	Scanning	Flash
Illumination wavelength	nm	**637**	532	470	870	750
Illumination repetition rate	MHz	**40**	1	N/A	0.133	N/A
Illumination power	Mean(mW)	**2**	6	N/A	21	90
	Peak(W)	**0.5**	N/A	N/A	39.5$^{(4)}$	N/A
PDP at Illum. Wavelength @ V_{EX}	%	**33.7@5V**	21@2.5V$^{(4)}$	20@3V$^{(4)}$	N/A	10/5V$^{(4)}$
Max. measured distance	m	**50$^{(2)}$**	150 - 430	367 – 5862$^{(6)}$	128	8
Imaging distance range	m	**0.7**	4.5	N/A	100	8
FOV	degree	**40 × 20**	N/A	N/A	55 × 9	N/A
Frame rate	fps	**30**	N/A	7.68 - 7.16$^{(6)}$	10	13
Accuracy (non-linearity)	m(%)	**0.088(0.17)**	0.07(0.3)-0.8(0.4)	1.5(0.37) - 35(1.9)$^{(6)}$	11(0.11)	N/A
Precision (σ)	m(%)	**0.0014(2.8e-3)**	0.15(0.1)-0.47(0.11)	0.2(0.13) - 0.5(0.14)$^{(6)}$	15(0.14)	N/A
Background light	-	**dark**	N/A	100 Mph/s/pix$^{(4)}$	70 klux	dark
Target reflectivity	-	**white**	white	N/A	9%	N/A
Power consumption	W	**2.54**	N/A	0.0935	0.53	2.8

(1) After TDC calibration; (2) Measured with prior knowledge of the scene; (3) Without and with micro-lens; (4) Estimated results; (5) Macro-pixels; (6) Emulated results using a fiber instead of free space;

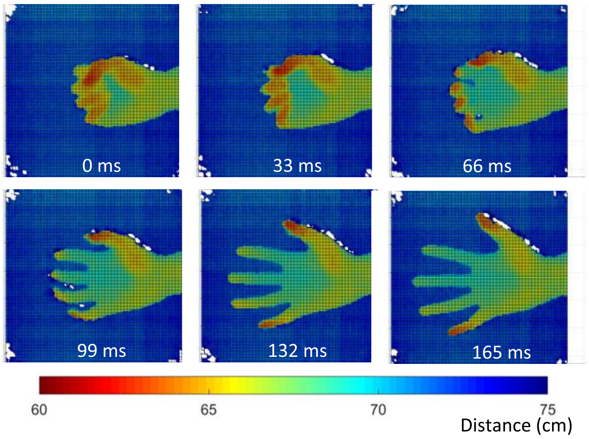

Fig. 18. Six successive frames from a 3-D movie at 30 frames/s, in which a hand is clenching and unclenching. The movie has been denoised with median filtering.

of the array, in which a hand is clenching and unclenching. The sensor was operated in the PHR mode, with PD time of 16 ms and histogramming time of 17 ms for each frame. Complete images were obtained by using median filtering with a neighborhood size of 2 × 2.

In order to measure the power consumption of the sensor in the PHR mode, a white board was placed in front of the sensor and illuminated with the same red laser at 40 MHz. After the peak is located, the PH was constructed and read out at a frame rate of 30 frames/s. The power consumption of each component in the sensor is detailed in Table II, showing a total measured consumption of 2.54 W with a histogrammed photon throughput of 156 MHz, and achieving an overall detection power of 16 nW per photon. The core voltage of the design, including the pixel circuitry, PHR, ALTDCs, and PLLs, is 1.8 V, while the IO voltage is 3.3 V. As expected, the digital core dissipates the largest proportion of the power, due to the significant logic of the PHR while operating at a 240-MHz clock. For the ALTDC array, the TDCs, ALs, and VCOs contribute 63%, 35%, and 2%, respectively. A major proportion of the TDC power consumption is due to the globally distributed clock network, STOP_HF, running at 320 MHz, which is a static value and would not increase significantly with pixel activity. In comparison with the multi-phase sharing TDCs in [8] and [25], the RO-based dual-clock architecture in this paper has only one clock distributed across the sensor, which dramatically reduces the TDC power consumption, thus improving the scalability for building larger TDC arrays. The I/O power consumption was limited due to the high compression factor achieved.

Table III summarizes the performance of the sensor in comparison with the state-of-the-art time resolved SPAD LiDAR systems. This design achieves the highest PDP with low DCR due to the superior SPAD performance and cascoded quenching circuit. The TDC achieves superior performance in resolution, power, and linearity when compared to [9] and [13]. Villa *et al*. [8] and Niclass *et al*. [12] report a better linearity, but with a much lower resolution and a much higher TDC power consumption [8]. In [12], FHR was integrated on chip for all the 16 pixels, which limits the scalability of the array due to SRAM area overhead. In our design, per-pixel PH was implemented for a 252 × 144 pixel array, which enables compression for the entire array with improved memory area efficiency. For ranging performance, our sensor achieves the highest spatial resolution and frame rate among all the listed sensors, except for [13] which extended the resolution by scanning the scene at the expense of frame rate. Although the measured distance in our design is relatively short compared with that of other systems, it should be noted that a low laser power and visible wavelength were employed, which limits the SNR and thus the range. We believe that the ranging performance can be significantly improved by employing a high power near-infrared response (NIR) laser, without affecting other aspects of the sensor.

IV. Conclusion

In this paper, we presented Ocelot, an image sensor comprising 252 × 144 SPAD pixels for time-resolved imaging applications, including flash LiDARs. To achieve a fill factor of 28% with a pitch of 28.5 μm, we made heavy use of resource sharing through a collision detection bus. The architecture is highly scalable, while at the same time reducing the I/O requirements even in highly complex 3-D scenes. This is achieved by timestamping events with TDCs that are shared in a dynamic reallocation method. RO-based TDCs were implemented in a dual-clock architecture with only one clock distributed across the sensor, which significantly reduces the power consumption while maintaining high uniformity timestamp processing. In order to increase photon throughput, an integrated histogramming scheme was implemented via 3.3-Mb SRAM memory. The scheme enables true PD from multi-reflections and a compression factor of 14.9-to-1 thanks to PH. To the best of our knowledge, this is the first implementation of fully integrated histogramming on a per-pixel basis for a full 2-D array and a design with one of the largest fill factors for a smaller than 30-μm pitch.

To demonstrate the suitability of Ocelot, a complete imaging system was designed with large FOV. The 2-D images were captured operating the sensor in the SPC mode at an optical level of 0.1 lx and a frame rate of 666 frames/s. The 3-D images were captured by operating the sensor in the TCSPC mode from a minimum of 0.7 m at 30 frames/s using an illumination power as low as 2 mW (average). The maximum ranging operation was 50 m, where a nonlinearity of 8.8 cm was measured with the same laser. Further improvements on this system will significantly extend the ranging distance by employing a high power NIR laser. Target applications include gesture recognition, AR/VR, automotive safety, and industrial robotics in light-starved, short- to-long-range scenarios.

Acknowledgment

The auhtors would like to thank E. Venialgo for the printed circuit board (PCB) design support. They would also like to thank R. van der Werf for assistance with technology and design environments.

References

[1] R. Halterman and M. Bruch, “Velodyne HDL-64E lidar for unmanned surface vehicle obstacle detection,” *Proc. SPIE, Unmanned Syst. Technol. XII*, vol. 7692, p. 76920D, May 2010, doi: 10.1117/12.850611.

[2] M.-W. Seo *et al.*, “A 10 ps time-resolution CMOS image sensor with two-tap true-CDS lock-in pixels for fluorescence lifetime imaging,” *IEEE J. Solid-State Circuits*, vol. 51, no. 1, pp. 141–154, Jan. 2016.

[3] C. S. Bamji *et al.*, “IMpixel 65 nm BSI 320 MHz demodulated TOF image sensor with 3 μm global shutter pixels and analog binning,” in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 94–96.

[4] D. Bronzi *et al.*, “100 000 frames/s 64 × 32 single-photon detector array for 2-D imaging and 3-D ranging,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 6, pp. 354–363, Nov./Dec. 2014.

[5] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, “Single-photon synchronous detection,” *IEEE J. Solid-State Circuits*, vol. 44, no. 7, pp. 1977–1989, Jul. 2009.

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

[7] C. Veerappan *et al.*, “A 160 × 128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter,” in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2011, pp. 312–314.

[8] F. Villa *et al.*, “CMOS imager with 1024 SPADs and TDCS for single-photon timing and 3-D time-of-flight,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 6, Nov./Dec. 2014, Art. no. 3804810.

[9] M. Perenzoni, D. Perenzoni, and D. Stoppa, “A 64 × 64-pixel digital silicon photomultiplier direct ToF sensor with 100 Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing,” in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Jan./Feb. 2016, pp. 118–119.

[10] B. F. Aull *et al.*, “Geiger-mode avalanche photodiodes for three-dimensional imaging,” *Lincoln Lab. J.*, vol. 13, no. 2, pp. 335–350, 2002.

[11] L. Gasparini *et al.*, “A 32 × 32-pixel time-resolved single-photon image sensor with 44.64 μm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800 kHz observation rate for quantum physics applications,” in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 98–100.

[12] C. Niclass, M. Soga, H. Matsubara, M. Ogawa, and M. Kagami, “A 0.18-μm CMOS SoC for a 100-m-range 10-frame/s 200 ×96-pixel time-of-flight depth sensor,” *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 315–330, Jan. 2014.

[13] A. R. Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D. N. Yaung, and E. Charbon, “A 256 × 256 45/65 nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6 dB interference suppression,” in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2018, pp. 96–98.

[14] C. Niclass *et al.*, “Design and characterization of a 256 × 64-pixel single-photon imager in CMOS for a MEMS-based laser scanning time-of-flight sensor,” *Opt. Express*, vol. 20, no. 11, pp. 11863–11881, 2012.

[15] M.-J. Lee *et al.*, “High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm CMOS technology,” *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, Dec. 2018, Art. no. 3801809.

[16] M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce, and F. Zappa, "Single-photon avalanche diodes in a 0.16 μm BCD technology with sharp timing response and red-enhanced sensitivity," *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 2, Mar./Apr. 2018, Art. no. 3801209.

[17] C. Niclass, H. Matsubara, M. Soga, M. Ohta, M. Ogawa, and T. Yamashita, "A NIR-sensitivity-enhanced single-photon avalanche diode in 0.18 μm CMOS," in *Proc. Int. Image Sensor Workshop*, 2015, pp. 1–4.

[18] H. Xu, L. Pancheri, G.-F. D. Betta, and D. Stoppa, "Design and characterization of a p+/n-well SPAD array in 150 nm CMOS process," *Opt. Express*, vol. 25, no. 11, pp. 12765–12778, 2017.

[19] C. Veerappan and E. Charbon, "A low dark count p-i-n diode based SPAD in CMOS technology," *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 65–71, Jan. 2016.

[20] S. Lindner, C. Zhang, I. M. Antolovic, J. M. Pavia, M. Wolf, and E. Charbon, "Column-parallel dynamic TDC reallocation in SPAD sensor module fabricated in 180 nm CMOS for near infrared optical tomography," in *Proc. Int. Image Sensor Workshop*, 2017, pp. 86–89.

[21] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-PDE, backside-illuminated SPAD in 65/40-nm 3D IC CMOS pixel with cascoded passive quenching and active recharge," *IEEE Electron Device Lett.*, vol. 38, no. 11, pp. 1547–1550, Nov. 2017.

[22] E. A. G. Webster, L. A. Grant, and R. K. Henderson, "A high-performance single-photon avalanche diode in 130-nm CMOS imaging technology," *IEEE Electron Device Lett.*, vol. 33, no. 11, pp. 1589–1591, Nov. 2012.

[23] G. Acconcia, I. Rech, A. Gulinatti, and M. Ghioni, "High-voltage integrated active quenching circuit for single photon count rate up to 80 Mcounts/s," *Opt. Express*, vol. 24, no. 16, pp. 17819–17831, 2016.

[24] J. M. Pavia, M. Scandini, S. Lindner, M. Wolf, and E. Charbon, "A 1 × 400 backside-illuminated SPAD sensor with 49.7 ps resolution, 30 pJ/sample TDCs fabricated in 3D CMOS technology for near-infrared optical tomography," *IEEE J. Solid-State Circuits*, vol. 50, no. 10, pp. 2406–2418, Oct. 2015.

[25] R. M. Field, S. Realov, and K. L. Shepard, "A 100 fps, time-correlated single-photon-counting-based fluorescence-lifetime imager in 130 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 867–880, Apr. 2014.

[26] N. A. W. Dutton *et al.*, "A time-correlated single-photon-counting sensor with 14 GS/S histogramming time-to-digital converter," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 58, Feb. 2015, pp. 1–3.

[27] A. T. Erdogan, R. Walker, N. Finlayson, N. Krstajic, G. O. S. Williams, and R. K. Henderson, "A 16.5 giga events/s 1024 × 8 SPAD line sensor with per-pixel zoomable 50 ps-6.4 ns/bin histogramming TDC," in *Proc. Symp. VLSI Circuits*, Jun. 2017, pp. C292–C293.

[28] A. Carimatto *et al.*, "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*, Feb. 2015, pp. 1–3.

[29] S. Mandai, V. Jain, and E. Charbon, "A 780 × 800 μm^2 multichannel digital silicon photomultiplier with column-parallel time-to-digital converter and basic characterization," *IEEE Trans. Nucl. Sci.*, vol. 61, no. 1, pp. 44–52, Feb. 2014.

[30] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt, and R. Henderson, "A 100Mphoton/s time-resolved mini-silicon photomultiplier with on-chip fluorescence lifetime estimation in 0.13 μm CMOS imaging technology," in *IEEE ISSCC Dig. Tech. Papers*, Feb. 2012, pp. 122–124.

[31] R. J. Walker, J. A. Richardson, and R. K. Henderson, "A 128 × 96 pixel event-driven phase-domain $\Delta\Sigma$-based fully digital 3D camera in 0.13 μm CMOS imaging technology," in *IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, vol. 21, no. 14, Feb. 2011, pp. 1020–1022.

[32] J. M. Pavia, M. Wolf, and E. Charbon, "Single-photon avalanche diode imagers applied to near-infrared imaging," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 6, Nov./Dec. 2014, Art. no. 3800908.

[33] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, and E. Charbon, "Nonuniformity analysis of a 65-kpixel CMOS SPAD imager," *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 57–64, Jan. 2016.

[34] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl. Opt.*, vol. 35, no. 12, pp. 1956–1976, 1996.

Chao Zhang received the M.S. degree from Jiangnan University, Wuxi, China, in 2011. He is currently pursuing the Ph.D. degree with the Delft University of Technology, Delft, The Netherlands, with a focus on single photon imaging with single-photon avalanche diode (SPAD) sensors.

From 2011 to 2012, he was a Digital Design Engineer with Nvidia, Shanghai, China. His research is focused on the design of SPAD sensors for time-resolved imaging in light detection and ranging (LiDAR) applications.

Scott Lindner received the M.Eng. degree in electronics and electrical engineering from the University of Edinburgh, Edinburgh, U.K., in 2011, and the Ph.D. degree from the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland, in 2018.

From 2011 to 2013, he was a trainee with the Microelectronics Section, European Space Research and Technology Centre, Noordwijk, The Netherlands, where he worked on analog IC design for space applications. From 2013 to 2018, he was a Ph.D. student with the EPFL, working in collaboration with the Biomedical Optics Research Laboratory, University Hospital Zürich, Zürich, Switzerland, where he focused on time-of-flight (TOF) image sensor design for near-infrared optical tomography. Since 2018, he has held a post-doctoral position at EPFL. His research is focused on the design and application of single-photon avalanche diode (SPAD)-based TOF image sensors for use in biomedical applications.

Ivan Michel Antolović received the M.S. degree (*cum laude*) from the University of Zagreb, Zagreb, Croatia, in 2012, and the Ph.D. degree from TU Delft, Delft, The Netherlands, in 2018.

He has held a post-doctoral position at EPFL, Neuchâtel, Switzerland, since 2018. His research focuses on large-format photon counting single-photon avalanche diode (SPAD) imagers and small-format time-correlated SPAD imagers for microscopy applications.

Juan Mata Pavia received the M.S. degree in electrical engineering from the Royal Institute of Technology (KTH), Stockholm, Sweden, the Diploma degree in telecommunications engineering from the Polytechnic University of Valencia (UPV), Valencia, Spain, in 2004, and the Ph.D. degree from the Swiss Federal Institute of Technology in Lausanne (EPFL), Lausanne, Switzerland, in 2015.

In 2005, he joined Philips Semiconductors, where he was involved in the design of application processors for smartphones. In 2009, he joined EPFL, where he focused on the design of single-photon detectors for near-infrared applications.

Martin Wolf received the Ph.D. degree from ETH Zürich, Zürich, Switzerland.

He is currently the Head of the Biomedical Optics Research Laboratory, ETH Zürich, which specializes in developing techniques to measure and quantitatively image oxygenation of brain, muscle, tumor, and other tissues. He is a Professor of biomedical optics with the University of Zürich. His aim is to translate these techniques into clinical application for the benefit of adult patients and preterm infants.

Edoardo Charbon (M'92–SM'11–F'17) received the Diploma in electrical engineering and computer science (EECS) from ETH Zurich, Zurich, Switzerland, the M.S. degree in EECS from the University of California at San Diego, San Diego, CA, USA, and the Ph.D. degree in EECS from the University of California at Berkeley, Berkeley, CA, USA, in 1988, 1991, and 1995, respectively.

He has consulted with numerous organizations, including Bosch, X-Fab, Texas Instruments, Maxim, Sony, Agilent, and the Carlyle Group. From 1995 to 2000, he was with Cadence Design Systems, where he was the Architect of the company's initiative on information hiding for intellectual property protection. In 2000, he joined Canesta Inc., as the Chief Architect, where he led the development of wireless 3-D CMOS image sensors. From 2008 to 2016, he was the Chair of very large scale integration design with the Delft University of Technology, Delft, The Netherlands. Since 2002, he has been a member of the Faculty of EPFL, Lausanne, Switzerland, where has been a Full Professor since 2015. He has been the driving force behind the creation of deep-micron CMOS single-photon avalanche diode (SPAD) technology, which is mass-produced since 2015 and is present in telemeters, proximity sensors, and medical diagnostics tools. He has authored or co-authored over 300 papers and two books, and he holds 21 patents. His interests range from 3-D vision, light detection and ranging (LiDAR), FLIM, FCS, NIROT to super-resolution microscopy, time-resolved Raman spectroscopy, and cryo-CMOS circuits and systems for quantum computing.

Dr. Charbon is a Distinguished Visiting Scholar of the W. M. Keck Institute for Space at Caltech, a fellow of the Kavli Institute of Nanoscience Delft, and a Distinguished Lecturer of the IEEE Photonics Society.