CMOS SPAD Sensors for 3D Time-of-Flight Imaging, LiDAR and Ultra-High Speed Cameras

Zhang, Chao

DOI
10.4233/uuid:f2e8ac06-33c0-423e-9617-6eaa87f7abd8

Publication date
2019

Document Version
Final published version

Citation (APA)
Zhang, C. (2019). *CMOS SPAD Sensors for 3D Time-of-Flight Imaging, LiDAR and Ultra-High Speed Cameras*. [Dissertation (TU Delft), Delft University of Technology]. https://doi.org/10.4233/uuid:f2e8ac06-33c0-423e-9617-6eaa87f7abd8

CMOS SPAD Sensors for 3D Time-of-Flight Imaging, LiDAR, and Ultra-High Speed Cameras

CMOS SPAD Sensors for 3D Time-of-Flight Imaging, LiDAR, and Ultra-High Speed Cameras

Proefschrift

ter verkrijging van de graad van doctor
aan de Technische Universiteit Delft,
op gezag van de Rector Magnificus Prof.dr.ir. T.H.J.J. van der Hagen,
voorzitter van het College voor Promoties,
in het openbaar te verdedigen op
maandag 13 mei 2019 om 12:30 uur

door

Chao Zhang

Master of Engineering in Microelectronics
Jiangnan University
geboren te China

This dissertation has been approved by the promoter:

promotor: Prof.dr. E. Charbon

Composition of the doctoral committee:

Rector Magnificus	chairperson
Prof.dr. E. Charbon	Delft University of Technology, promoter
Independent members:	
Prof.dr. S. Hamdioui	Delft University of Technology
Prof.dr. L. K. Nanver	University of Twente
Prof.dr. G. Etoh	Osaka University, Japan
Prof.dr. D. Faccio	University of Glasgow, UK
Dr. D. Stoppa	AMS Inc., Switzerland
Dr. C. Jackson	On-Semiconductor Inc., Ireland
Prof.dr. K. Bertels	Delft University of Technology, reserve member

Keywords: Single-photon avalanche diode, time-of-flight, LiDAR, image sensor, high-speed sensor

There is only one heroism in the world:
to see the world as it is, and to love it.

-Romain Rolland

CONTENTS

NOMENCLATURE

ADAS	Advanced driver assistance system
ADC	Analog to digital converter
ALTDC	Address latch and TDC
AMCW	Amplitude modulated continuous wave
APD	Avalanche photodiode
BSI	Backside illumination
CDS	Correlated double sampling
CIS	CMOS image sensor
CMOS	Complementray metal oxide semiconductor
cps	count per second
DCR	Dark count rate
DNL	Differential nonlinearity
DTI	Deep trench isolation
FIFO	First-in-first-out
FMCW	Frequency modulated continuous wave
FOV	Field-of-view
FPGA	Field-programmable gate array
fps	frame per second
FWHM	Full width at half maximum

INL	Intergral nonlinearity
LiDAR	Light detection and ranging
LSB	Least significant bit
MCG	Multi-collection gate
MEMS	Micro-electro-mechanical system
PDE	Photon detection efficiency
PDP	Photon detection probability
PHR	Partial histogramming readout
PLL	Phase-locked loop
PVT	Process-voltage-temperature
QE	Quantum efficiency
QIS	Quanta image sensor
RO	Ring oscillator
SBNR	Single-to-background noise ratio
SOI	Silicon-on-insulator
SPAD	Single photon avalanche diode
STI	Shallow trench isolation
TAC	Time-to-amplitude converter
TCSPC	Time-correlated single photon counting
TDC	Time-to-digital converter
TIA	Transimpedance amplifier
TOF	Time-of-flight

1

SCOPE OF THE THESIS

This thesis mainly focuses on the design and implementation of high-speed image sensors and SPAD imagers. A high-speed image sensor based on multi-collection-gates is presented, targeting at frame rate of 1G fps. As a key enabling technique, 3D-stacking was used in the design, where the the sensor was implemented on the top chip with charge-coupled device (CCD) technology and the driver was on the bottom chip with complementary metal-oxide-semiconductor (CMOS) technology. In this thesis, the sensor architecture, operation principle and detailed driver chip design will be presented in detail.

Similarly, 3D-stacking could also be applied to single-photon-avalanche-diode (SPAD) sensors for time-correlated single photon counting. However, due to the limited accessibility of this technique, a planar technology with front-side illumination was used for SPAD sensors. In this case, challenges in pixel pitch, fill factor, TDC number and photon throughput are discussed. To overcome these challenges, new techniques including collision detection coding, dynamically reallocating time-to-digital converters (TDCs) and per-pixel partial histogramming were proposed and implemented.

Besides, introductions to 3D-stacking, high-speed imaging, 3D-imaging and SPADs will be given before the core chapters of the thesis. The author's perspective on the next generation sensors will conclude the thesis.

2

INTRODUCTION

Three-dimensional imaging is a key enabling technology for a wide range of applications, such as augmented and virtual reality (AR/VR), facial recognition, assembly line robotics, advanced driver assistance systems (ADAS), and light detection and ranging (LiDAR) systems in autonomous driving. Among various depth imaging technologies, time-of-flight (TOF) approach is emerging as a widely applicable method due to its versatility. This chapter reviews and discusses the principle, applications and sensor technologies of TOF imaging systems. For applications in 3D imaging, the motivation and challenges of developing time-resolved image sensor based on single-photon avalanche diodes (SPADs) are presented.

2.1. OVERVIEW OF HIGH-SPEED IMAGE SENSORS

Ultra-high speed solid state cameras, featuring low cost, high spatial resolution and high frame rate are a powerful tool for many applications, such as bio-imaging, physics and mechanics, which require ultra-fast phenomena analysis. The typical structure of high-speed image sensor is based on burst-capturing mode, where charges are captured and stored in the on-chip memory for a limited number of frames at high speed then read out at a slow speed. A large format (312-kpixel) high speed image sensor was reported in [1], equipped with high sensitivity and backside-illuminated (BSI) CCD pixels with *in-situ memory*, working at burst capturing mode and achieving a frame rate of 16.7 Mfps. The sensor architecture is shown in Fig. 2.1. Similarly, active pixel based CMOS image sensor (CIS) with on-chip capacitive memory can also be used in burst mode for high-speed imaging [2], achieving a maximum frame rate of 20 Mfps with a resolution of 400 × 256.

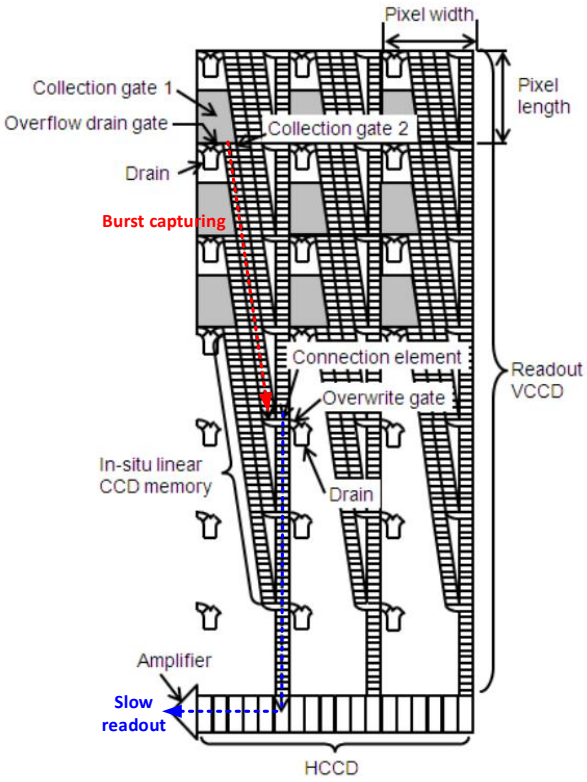


Fig. 2.1 Structure of high-speed image sensor in [1]. The sensor works in burst capturing and slow readout modes, achieving a maximum frame rate of 16.7 Mfps for 139 consecutive frames.

Apart from the in-situ memory image sensors, other types of image sensors based on computational imaging approaches with more than 20 Mfps have been reported. For instance, a multi-aperture CMOS imager was reported in [3], which achieved 200 Mfps. In this sensor, an array of 5 × 3 apertures was implemented, where each aperture comprises an array of 64 × 108 pixels and optical signals are evenly distributed to each aperture. The apertures are synchronized with a common clock, but work independently as a set of temporally-coded binary shutters. Since the shutter pattern of the apertures are slightly different from each other, consecutive images with temporal difference can be captured with the imaging operation. The final time-resolved images can be reconstructed by solving the image-capturing process with the known shutter pattern. The number of frames it can record is dependent on the number of apertures, N. However, for a given total number of pixels, K, the spatial resolution of each aperture will be reduced to K/N. Besides, due to the light spreading, the light intensity to each aperture is only $1/N$ of the total incoming light, making its application in low light level environment limited.

2.2. 3D-STACKING TECHNOLOGIES

For planar technology based image sensors, the readout and processing circuits are normally placed outside the core detection region. However, with the increase in array size, it becomes challenging to read out the sensor array and process the large amount of data at high frame rate, due to the increased propagation delay in pixel-to-circuit connection. To solve this problem, 3D-stacking technologies have been widely proposed, which stack multiple chips vertically with dense connections. Therefore, pixel array and processing circuits can implement in different chips, which significantly shorten and simplify the connection. Moreover, different technologies can be applied to each of the chips, e.g. device optimized technology for the pixel array design while small technology nodes for circuit design. With such a combination, each independent chip can be designed in its optimal technology, so as to achieve the highest performance with the stacked sensor.

To stack multiple chips, two main technologies are available, comprising through-silicon-via (TSV) [4, 5] and micro-bump junction [6]. The cross-sections of these two connections are shown in Fig. 2.2. From this figure, we can see part of the substrate of top tier needs to be etched away to form TSV connections, implying it is not suitable for pixel connection due to reduced active silicon area. On the contrary, micro-bump utilizes a face-to-face connection without impacting the substrate. Besides, compared to the multi-tier connectivity with TSV, micro-bump can only stack 2 tiers. Therefore, to build 3D-stacked image sensors, pixel array is normally connected via micro-bumps, while the

bonding pads are with TSVs. Given all the benefits of 3D-stacking, it is essential to understand the challenges that are holding the technology back from completely disrupting the semiconductor industry. The most critical challenges of 3D integration include heat removal, reliability, yield, power delivery and cost,etc, which are further discussed in [7].

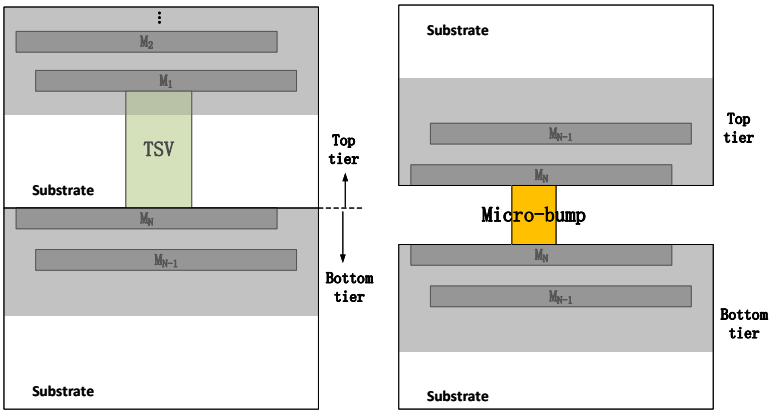


Fig. 2.2 Cross-sections 3D-stacking with TSV and micro-bump.

2.3. GENERAL CLASSIFICATION OF 3D IMAGING TECHNOLOGIES

Range detection techniques are well known and applied in many applications. According to the sensing mechanism, there are three major approaches, comprising microwave, ultrasonic and optical techniques. By comparing the performance and constraints, different ranging techniques are evaluated in this chapter. Common performance criteria include detection range, resolution, accuracy, field-of-view (FOV), and frame rate, while constraints include cost, size, power consumption, operation condition, robustness, and hazard level. Microwave based radar technology has been highly developed and applied in military, industry and consumer fields, which features long detection range, high immunity to environmental conditions, matured technology, low cost, etc. However, the spatial and depth resolution are poor. Ultrasonic sensing technology achieves low power, compact size and high depth resolution, but suffers great losses in air and can only achieve short detection range. Optical sensing technology involves long range, large FOV, the highest spatial and depth resolution, which has been used as one of the main sensing technologies in emerging applications, such as automotive driving, AR/VR, robotics. Figure. 2.3 comprises a taxonomy diagram of optical ranging technologies [8],

where the highlighted route shows the central area of the research in this thesis.

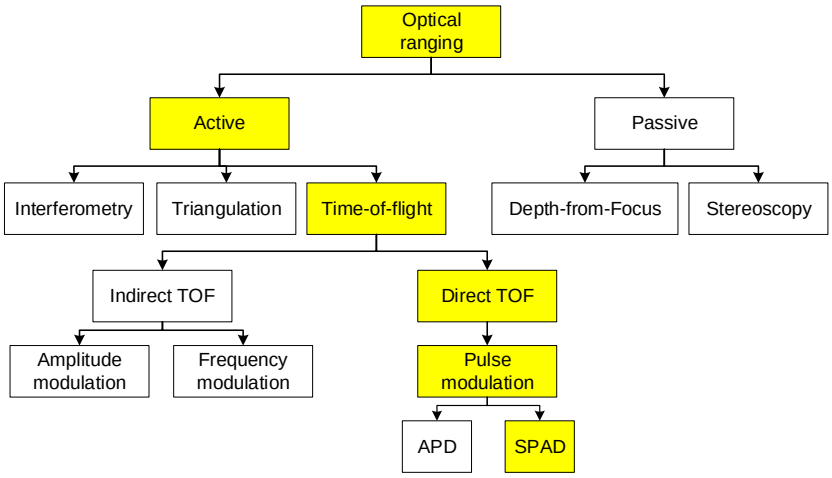


Fig. 2.3 Classification of the optical ranging technologies.

The computational stereo imaging method, which is similar to the human visual system, extracts the 3D structure of a scene from two or more images taken from distinct cameras by means of triangulation [9]. Since no active illumination is required, the eye-safety criteria doesn't have to be considered, which simplifies the system design. The distance extraction is based on the displacement of an object in two images, so feature correspondence is of importance, which determines the locations of the same object in two camera images. However, ambiguous matches could happen in some situations, such as occlusion (features are visible in one camera but not the other), lack of texture and specularities. This limits the computational stereo method to some specific applications, in which a variety of constraints are made, e.g. image brightness and surface smoothness [9].

The concept of depth-from-focus involves distance calculation by modeling the quality of images with the variation of the camera's focal parameters [10, 11]. By scanning the scene with different focal settings, the sharpest image for each point can be decided, and the distance determined. Compared with stereo imaging, only one camera is required, which avoids the correspondence and occlusion problems. However, it is still challenging to detect scenes with textureless regions, such as a flat white wall. On the other hand, with the scanning operation in the depth-of-field, a tradeoff has to be made between depth resolution and frame rate. For both stereo and depth-from-focus imaging methods, large format and commercialized cameras can be employed directly, offering

extremely high spatial resolution at the expense of more computational power.

In contrast to passive imaging, active imaging employs a light source, e.g. laser or LED, to illuminate the scene. Among these ranging methods, the interferometry method provides the highest depth resolution at the level of nanometers, which measures the interference fringe generated by the backscattered laser beam interfering with the reference beam. However, the detection is heavily limited in range, typically within several hundred of millimeters even with the technique of multiple-wavelength interferometry [12, 13]. Besides, since the laser wavelength can be affected by the operating environment, calibration needs to be applied by monitoring the environmental parameters, including temperature, atmospheric humidity and pressure, which limits the application in consumer field.

Similar to computational stereo imaging, the triangulation method is applied to structured light vision systems in an active way, where known light patterns are projected to the object and the 3D profile can be obtained by solving the deformation of the object image with triangulation computation. The simplest pattern can be a 1-D stripe light. While in order to image the entire scene, a scan operation is required, which is typically based on a scanner, e.g. mechanical scanner, micro-electro-mechanical system (MEMS) mirror. This reduces the imaging frame rate and the robustness, due to the mechanical mechanism. For one-shot 3D image acquisition, a variety of coded structured light approaches were proposed, where a known 2-D pattern is projected to the scene and, each coded pixel has its own codeword in terms of color [14], spatial coding position [15], or hybrid patterns [16]. A well known application of the structured light imaging is the iphone-X facial recognition, in which more than 30000 infrared dots are projected onto the face to build an unique facial map. However, a major drawback of the system is the ranging distance that is limited by the length of the camera-to-projector baseline. In order to achieve longer distance imaging, a larger baseline is required, thus resulting in the enlargement of the system size.

2.4. TIME-OF-FLIGHT TECHNIQUES

Taking the limitations of the previously mentioned methods into account, the time-of-flight approach provides more configurable features in range, resolution, system size and cost, which have received significant attention in the last decade. The TOF technique is based on active illumination, where the light travel time from the source to the object, then back reflected to the photo-detector, is measured indirectly (iTOF) or directly

(dTOF). Since the speed of light is 3×10^8 m/s, for a normal p-i-n photodiode, it is difficult to measure the light propagation time directly and accurately, because of the limitation in gain and response time. Instead, the TOF can be resolved indirectly by measuring the phase or frequency shift of the reflected signal with respect to the illumination signal. Meanwhile, with the improvement of the gain and timing performance of photodiodes, dTOF systems have been demonstrated based on linear-mode avalanche photodiodes (APDs). However, if the photodiode bias voltage is further increased and exceeds the breakdown voltage, the photodiode will work in a so called Geier-mode with a virtually infinite optical gain and fast response time, enabling single-photon detection. Such a photodiode is normally referred to as a single-photon avalanche diode (SPAD). With SPADs, dTOF systems based on time-correlated single photon counting (TCSPC) can be built. A detailed classification and analysis of the TOF techniques is reported in the following sections.

2.4.1. PHASE SHIFT BASED INDIRECT TIME-OF-FLIGHT

The principle of the phase-shift based iTOF is illustrated in Fig. 2.4, where a sinusoidally-modulated light is used to illuminate the scene. This kind of system is normally referred to as amplitude modulated continuous wave (AMCW) LiDAR. The phase shift of the received signal is measured at $\Delta\varphi$ and the distance of the object d can be calculated with equation (2.1) [17, 18].

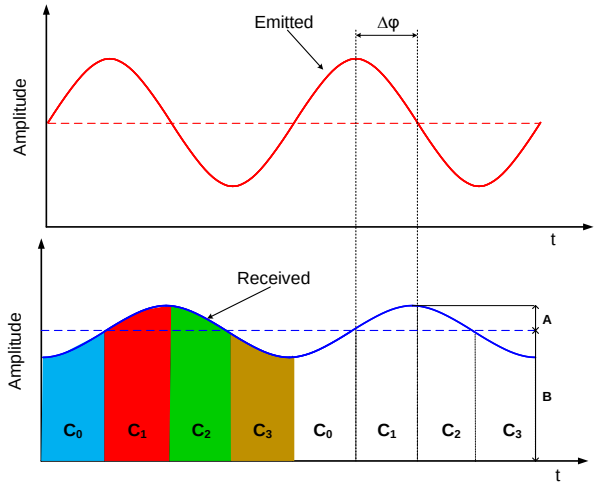


Fig. 2.4 Phase-shift based iTOF operation diagram. A sinusoidal modulated light is used for the active illumination. The reflected signal is sampled with 4 accumulation windows, then the phase delay $\Delta\varphi$ can be calculated accordingly.

$$d = \frac{c}{2f} \cdot \frac{\Delta\varphi}{2\pi} = R_D \cdot \frac{\Delta\varphi}{2\pi}, \tag{2.1}$$

where c is the speed of light and f is the modulation frequency, which defines the maximum unambiguous range (R_D) that the system can achieve with a single modulation frequency. To retrieve the phase delay, the received signal is sampled with 4 accumulation windows, providing signals of C_0, C_1, C_2 and C_3. Then, the phase delay $\Delta\varphi$ can be calculated with (2.2).

$$\Delta\varphi = arctan\frac{C_3 - C_1}{C_0 - C_2}. \tag{2.2}$$

From (2.1) we can see that for a longer ranging distance, a slower modulation frequency f is required. However, this will reduce the distance precision δ_d which is distance dependency as well, as is shown in (2.3) [17].

$$\delta_d = \frac{R_D}{\sqrt{8}\pi}\frac{B}{A}, \tag{2.3}$$

where B is the background light intensity and A is the signal intensity. For a given object and background light, since the reflected signal intensity reduces exponentially with distance, an exponentially degrading precision can be expected and was verified in [17–20]. Multiple modulation frequency method was used in [21], which partly solved this trade-off, but at the expense of the system complexity. Besides, since the light illuminates the scene continuously, to satisfy eye-safety criteria, a relatively low peak optical power has to be used, which results in a low signal amplitude as well as the signal-to-background noise ratio (SBNR). With all these concerns, the AMCW technique is more suitable for short range detection, e.g. less than 20 m.

Furthermore, a fundamental limitation of the AMCW technique is multi-path interference [22]. Since the system utilizes a single frequency illumination, a single phase delay is measured. If two spatially separated objects are detected by one pixel, a reflection signal with mixed phase information is received, which will lead to significant errors. Due to this limitation, it is challenging to sense a complicated scene with AMCW technique or to image the same scene with multiple AMCW systems.

Nevertheless, since iTOF is based on in-pixel photodemodulators, high resolution becomes a major advantage, e.g. a 1 Mpixel 3D imaging sensor was reported in [23]. In this case, the light source illuminates the scene in a flash manner, which removes the mechanical scanning mechanism, thus resulting in a compact system with high robustness. The system cost is also lower compared to dTOF, due to the lower speed requirement to

both the illuminator and the receiver.

2.4.2. FREQUENCY SHIFT BASED INDIRECT TIME-OF-FLIGHT

Similar to radar, frequency modulated continuous wave (FMCW) techniques have been widely used for distance measurements in different applications. The principle of FMCW LiDAR system is illustrated in Fig. 2.5, where the optical frequency is linearly modulated in time. The time delay between the emitted light and the reflection (τ_R) causes a frequency difference f_R. Therefore, a beat tone at this frequency can be retrieved by post-processing the photodiode signal in the frequency domain, and the object distance R can be calculated as follows:

$$R = \frac{\tau_R}{2}.c = \frac{f_R}{2\gamma}.c, \tag{2.4}$$

where γ is the slope of the frequency modulation and c is the speed of light. High resolu-

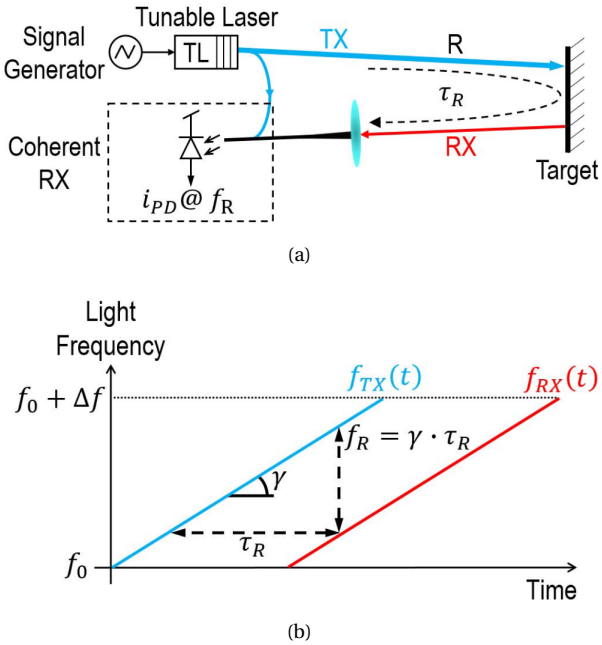


Fig. 2.5 FMCW operation diagram [24]. (a) the basic system architecture and (b) waveform frequency difference between the transmitted light and received light.

tion at short distance imaging has been demonstrated in [24], where a depth resolution of 8μm was reached at a distance of 5 cm. Apart from the distance, the velocity of the object can be measured in one shot with the Doppler effect, which can be a big advantage in automotive applications. Compared to AMCW, FMCW can offer better tolerance

against environmental disturbances. For instance when multi-path reflection from different distant targets is detected by FMCW, multiple beat tones can be resolved, thus determining the distance of each object. However, the maximum measurable distance of FMCW LiDAR system is typically limited to tens of meters, due to the laser phase noise that determines the spectral linewidth [8]. On the other hand, similarly to AMCW, continuous waveform light yields high illumination power. Due to the eye-safe limitation, this reduces the detection range.

2.4.3. PULSED DIRECT TIME-OF-FLIGHT

The operating diagram of dTOF is shown in Fig. 2.6, where a laser pulse with a picosecond to nanosecond duration, is transmitted, reflected and detected by a photodetector. A 'stopwatch' circuit is used to measure the elapsed time, which is started co-incidently with the laser pulse and stopped with the detection of signal, or vice versa. The dTOF technique is very straightforward, and the distance d can be calculated as (2.5), where c is the speed of light. Despite the simplicity, the dTOF approach became feasible only at the end of the 60's [25], due to the stringent speed requirements to the photodetector, light source and time-measurement associated circuitry.

$$d = c \cdot t \cdot \frac{1}{2}, \tag{2.5}$$

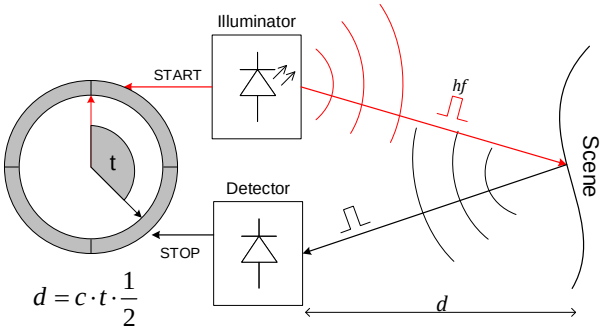


Fig. 2.6 Direct time-of-flight operating diagram.

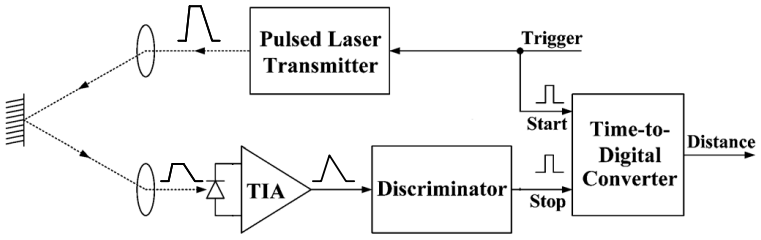
In order to detect the short laser pulse, photodetectors with fast timing response and high gain are required. Linear-mode avalanche photodiodes (APDs) with high gain (50-100) and high quantum efficiency (>80% at 800 nm), have been widely used in dTOF measurement [26–29]. A basic block diagram of APD based dTOF system is shown in Fig. 2.7(a), where the time-to-digital converter (TDC) is started with the synchronization of laser pulse. At the receiver, the photocurrent generated by the APD is converted into

voltage with a transimpedance amplifier (TIA). A discriminator, typically a voltage comparator, outputs a STOP signal that freezes the TDC operation when the output signal of the TIA exceeds a certain threshold, V_{TH}. However, due to the difference of object reflectivity, the amplitude of the TIA output will be changing. For the voltage comparison with a single threshold, a large timing error can be generated, which is known as a timing walk error and is shown in Fig. 2.7(b). Since the walk error is typically at the level of ns, calibrations at both circuitry and system level have to be applied to ensure accurate detection. On the other hand, in order to detect the fast and weakly reflected laser signal, low-noise and high bandwidth with the analog front-end circuit is required. To satisfy these requirements, more power must to be dissipated, e.g. 79 mW per channel in [28] and 180 mW in [29], which limits the implementation of these systems to single point or linear format. Furthermore, as the pitch of these APD pixels is typically hundreds of micro-meters, it is challenging to implement large arrays. Therefore, to perform complete imaging, a scan mechanism has to be used, which results in bulky and less robust systems.

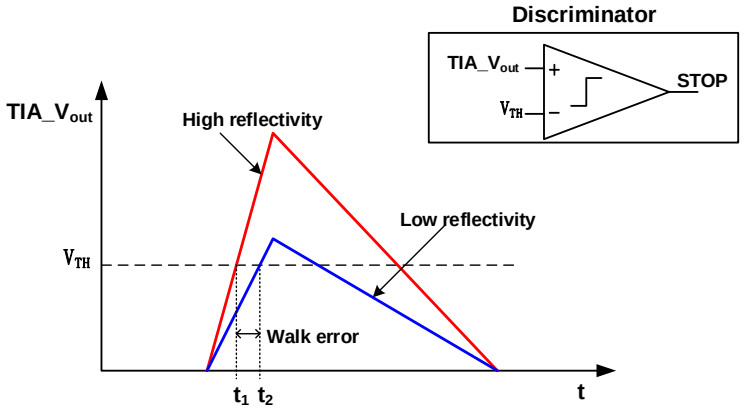
With pulse modulated light, a low duty cycle illumination can be employed, enabling illuminating light with short pulse width and high peak optical power to be used while maintaining the average eye-safe exposure. Due to the high peak power, the SBNR is significantly improved, which extends the detection range to hundreds, even thousands of meters [30]. Besides, multi-path reflection can be detected and recognized easily by multi-event measurement.

2.4.4. TIME-CORRELATED SINGLE-PHOTON COUNTING

TCSPC relies on the similar concept as dTOF, while the photodetector is replaced with a single-photon avalanche diode (SPAD) based sensor, which is the main topic of this thesis and will be further discussed in the following chapters. The difference is that in APD dTOF the TDC triggering signal is generated by the conversion of a TIA, while in TCSPC a digital triggering signal can be generated directly by a SPAD. Thus, in terms of the functionality, a SPAD can be treated as a high-speed binary switch that can be triggered with single-photon detection. Low timing jitter at the level of tens of picoseconds can be achieved, resulting in high TOF precision. On the other hand, due to the device dead time, and in order to avoid pile-up , SPADs are typically operated in photon starving, where on average less than one photon is detected in each detection cycle. In order to improve the detection accuracy and reliability, the TOF histogram typically involving a large number of detections in TCSPC is built. Figure 2.8 illustrates the basic TCSPC principle and the histogram based on the detection in multiple cycles.



(a)



(b)

Fig. 2.7 (a) Conventional dTOF LiDAR architecture based on APD. (b) A diagram of walk error at the same distance due to the amplitude difference.

In comparison with APD based dTOF, apart from single-photon sensitivity and high accuracy, SPADs feature additional properties, such as small pixel pitch(tens of μm and even smaller) and CMOS compatibility, that enable chip-level and highly integrated LiDAR systems to be constructed with an array of SPAD pixels. Therefore, instead of scanning the scene, a diffused beam is used to illuminate the scene in flash mode, enabling dTOF measurement at each pixel in parallel. As in a FLASH, this method is known as FLASH LiDAR. Moreover, in a APD dTOF system, due to the bandwidth limitation of the TIAs, the laser pulse width is typically limited to several nenoseconds, which leads to a lower SBNR. On the contrary, for a SPAD sensor, due to the single-photon detection property, the laser pulse can theoretically be infinitely narrow. Under a constant eye-safe criteria, with the decrease in the pulse width the optical peak power can be increased accordingly. This gives an improved SBNR, resulting in a longer detection range.

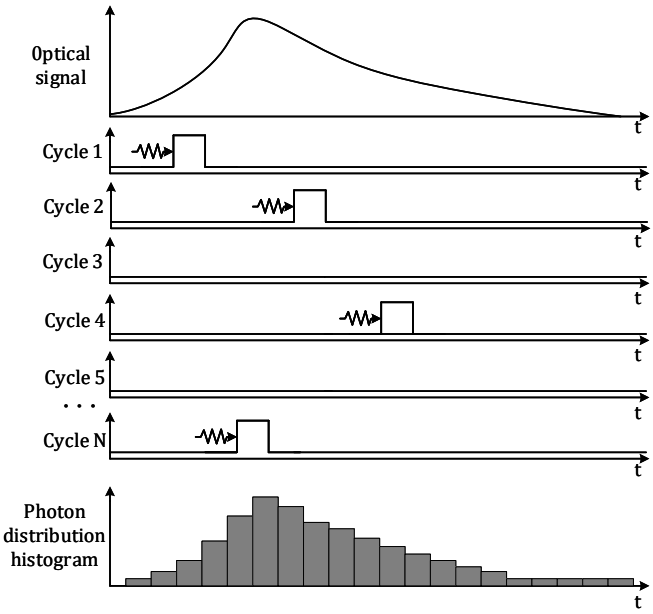


Fig. 2.8 TCSPC principle.

2.4.5. ToF SUMMARY

As is discussed in previous sections, ToF distance detection can be implemented in different approaches. To summarize, parameters of these ToF techniques are compared in Table. 2.1, in terms of detection distance, resolution, precision, power consumption and data throughput, etc. In general, AMCW presents high resolution 3D imaging, such as QVGA, in a medium range. However, the precision will degrade dramatically with distance; high calibration effort is required to align all the pixels; poor anti-interference performance in situations, such as multi-path reflection, interference between multiple AMCW, etc. FWCM exhibits the highest precision, with high interference immunity. However, a tradeoff between the precision and maximum range has to be made for a given light source, and it is challenging to generate linear light-wave in frequency. Long distance detection with centimeter-level precision can be achieved with pulsed ToF, but with limited resolution, typically at the format of single point or linear array, such as 1×16.

Compared to AMCW, TCSPC exhibits millimeter-to-centimeter precision, long detection range, improved anti-interference performance. But from this table, we also can see the main challenges of TCSPC are the resolution, detection power consumption, data

throughput and background light suppression. The first three challenges drive the author to work on this thesis and will be discussed extensively in chapter 5 and **??**. To improve the background light suppression, one sensor architecture based on collision detection bus is proposed in section 5.4.

Table 2.1: Comparison table of ToF techniques

Parameters	AMCW	FMCW	Pulsed ToF	TCSPC
ToF type	indirect	indirect	direct	direct
Detector type	PD	PD	APD/SiPM	SPAD
Light modulation	pulsed/ sinusoidal CW	sinusoidal CW	pulsed	pulsed
Range	tens of meters	tens of meters	hundreds of meters	tens-to-hundreds of meters
Resolution	high	high	low	medium
Precision	mm	μm	cm	mm to cm
Illumination peak power	tens of mW	tens of mW	tens of Watt	tens of Watt
Detection power	low	low	medium/high	high
Calibration effort	high	high	medium	low
Data throughput	low	low	low	high
Anti-interference	low	high	medium	medium
Background light suppression	high	high	Medium	low

2.5. CHALLENGES

For a flash TCSPC system, it always involves time-resolved measurements with a large array of pixels in parallel. In the last decade, in-pixel-TDC architectures have been widely used in SPAD sensor designs [31–33]. In these sensors, the intrinsic high gain of the SPAD is fully utilized, where each pixel is time-stamped with its own TDC upon photon detection. However, due to the circuit complexity, a large silicon area is occupied by the TDC, which results in a low fill factor even with a large pixel pitch, e.g. 1% fill factor for 50 μm in [31], 3.14% for 30 μm in [32], or 19.84% for 44.64 μm in [33].

More TDCs have to be implemented with the scaling of the pixel array, which brings challenges in power consumption and uniformity. To reduce the power consumption, localized ring oscillator (RO) based TDCs have been used in [31], where the RO is started

by the photon detection and stopped by a reference signal. Since the TDC will keep in an idle state when there is no SPAD firing, low power consumption can be achieved. Nevertheless, as the ROs are controlled with a common bias voltage, frequency non-uniformity will be generated due to the device mismatch between ROs, leading to increased TDC non-linearity with the accumulation of the RO oscillation. To improve the uniformity, TDCs based on multi-phase interpolation have been implemented in some sensors [34, 35], in which multiple clock phases, typically 8 or 16, are generated with a delay locked loop (DLL) circuit, and then distributed to every TDC via multiple balanced clock trees. Since all the TDCs share the same clock phases, high uniformity can be achieved, but at the cost of high power dissipated in the the always-on clock trees. Similarly, TDCs based on mutually coupled-ROs achieved high uniformity by jointing one of the phases of all the ROs, which synchronizes the frequency of each RO and improves the phase noise [36]. However, to maintain the coupling mechanism, all the ROs have to keep oscillating, resulting in high power consumption.

For a SPAD sensor, except for the increased spatial resolution, a large pixel array also implies massively parallel TOF measurements, resulting in a large amount of data for readout and processing. For example, if a 252 × 144 array operates at 1% pixel activity with a 40 MHz laser frequency, the photo detection rate can be 14.5 Gcps. Assuming each event comprises 20 bits, including both the TDC and address data, a required output data bandwidth will be 290 Gbps. This is impractical for a number of reasons, including high power consumption and large number of data pins. To solve this problem, on-chip histogramming was implemented in [37, 38] to accumulate photons for each TDC bin in memory. However, due to the large overhead area of the memory, it is impractical to implement full range histogramming for a large pixel array, thus limiting the these sensors to be single point or line formats.

In this thesis, we investigate the challenges of performing high resolution TCSPC imaging with SPAD imagers, targeting at scalable sensor architectures with small pixel pitch, high fill factor, low power consumption and high photon throughput. To achieve these goals, instead of per-pixel TDC architecture, a TDC-sharing approach was proposed and two sensors were designed and implemented, which can be used in applications including near-infrared optical tomography (NIROT), gesture recognition, and industrial robotics in light-starved, short-to-long-range scenarios. However, with further increase on the sensor functionalities, it is challenging to achieve all the goals in planar technologies. Therefore, backside illumination (BSI) and 3D-stacking technologies have drew significant attention due to the flexibility it brings to the sensor architecture design. To investigate these two technologies, an ultra-high speed image sensor based on multi-

collection-gate (MCG) pixels and bump-to-bump stacking was developed, targeting at a frame rate of 1 Gfps.

2.6. CONTRIBUTIONS

Three sensors have been designed and the major contributions of this thesis include:
(1) Nanosis: The limitation to the frame rate of image sensors was discussed, revealing the driving capability of the sensor is a major bottleneck. To overcome this limitation, a localized XNOR driver based on 3D-stacking technology was proposed for ultra-high speed image sensors. A minimum output pulse width of 1ns was achieved, leading to a frame rate of 1 Gfps.

(2) Piccolo: A collision detection architecture was demonstrated to increase the pixel fill factor, i.e. a fill factor of 28% was reached with a pixel pitch of 28.5 μm. Besides, analysis of photon throughput was conducted, which shows the chip readout bandwidth is the main limitation to the photon throughput, rather than the TDC number. Driven by this conclusion, a dynamically reallocating TDC architecture was proposed, which achieves the same photon throughput as that of per-pixel TDCs architectures but with much less number of TDCs. This gives a new way for designing large array sensors.

(3) Ocelot: To further improve the photon throughput, an on-chip TOF data compression technique was proposed. This is achieved by exploiting the intrinsic timing-bin distribution of the TOF histograms, where a two step approach was utilized comprising peak searching and partial histogramming. With this scheme it enables, for the first time, per-pixel integrated histogramming for a large 2D array, and achieves a 14.9-to-1 data compression factor.

2.7. THESIS ORGANIZATION

The thesis is organized as follows. Chapter 4 presents the architecture of the stacked sensor, the driver chip measurement results and the failure analysis. In Chapter 3, the background of SPADs is introduced. In particular, some useful performance parameters are defined. Moreover, a number of SPAD sensors and front-end circuits are reviewed and compared, exhibiting the variety of options in SPAD sensor design for different applications. Chapter 5 begins with the challenges in the design of time-resolved SPAD imagers. A new architecture is presented, based on which a 32 × 32 pixel sensor was implemented and the measurement results are reported. In Chapter 6, this architecture is extended to a larger sensor with 252 × 144 pixels, with the implementation of partial histogramming readout to achieve high photon throughput. In Chapter 7, conclusions

are drawn and an outlook for the future of SPADs and MCG sensors is presented.

REFERENCES

[1] T. Arai, J. Yonai, T. Hayashida, H. Ohtake, H. Van Kuijk, and T. G. Etoh, *A 252- V/, 16.7-Million-frames-per-second 312-kpixel back-side-illuminated ultrahigh-speed charge-coupled device,* IEEE Transactions on Electron Devices **60**, 3450 (2013).

[2] R. Kuroda, Y. Tochigi, K. Miyauchi, T. Takeda, H. Sugo, F. Shao, and S. Sugawa, *[Paper] A 20Mfps Global Shutter CMOS Image Sensor with Improved Light Sensitivity and Power Consumption Performances,* ITE Transactions on Media Technology and Applications **4**, 149 (2016).

[3] F. Mochizuki, K. Kagawa, S. I. Okihara, M. W. Seo, B. Zhang, T. Takasawa, K. Yasutomi, and S. Kawahito, *Single-shot 200Mfps 5×3-aperture compressive CMOS imager,* Digest of Technical Papers - IEEE International Solid-State Circuits Conference **58**, 116 (2015).

[4] D. Henry, J. Charbonnier, P. Chausse, F. Jacquet, B. Aventurier, C. Brunet-Manquat, V. Lapras, R. Anciant, N. Sillon, B. Dunne, N. Hotellier, and J. Michailos, *Through Silicon Vias technology for CMOS image sensors packaging: Presentation of technology and electrical results,* 10th Electronics Packaging Technology Conference, EPTC 2008 , 35 (2008).

[5] B. M. Motoyoshi, *Through silicon via (TSV),* **97**, 43 (2009).

[6] M. Motoyoshi, T. Miyoshi, M. Ikebec, and Y. Arai, *3D stacked SOI-CMOS pixel detector using Au micro-bump junctions,* 2016 SOI-3D-Subthreshold Microelectronics Technology Unified Conference, S3S 2016 , 6 (2017).

[7] V. Kumar and A. Naeemi, *An overview of 3D integrated circuits,* 2017 IEEE MTT-S International Conference on Numerical Electromagnetic and Multiphysics Modeling and Optimization for RF, Microwave, and Terahertz Applications, NEMO 2017 , 311 (2017).

[8] T. Bosch, *Laser ranging: a critical review of usual techniques for distance measurement,* Optical Engineering **40**, 10 (2001).

[9] M. Z. Brown, D. Burschka, and G. D. Hager, *Advances in computational stereo,* IEEE Transactions on Pattern Analysis and Machine Intelligence **25**, 993 (2003).

[10] S. K. Nayar and Y. Nakagawa, *Shape from Focus,* IEEE Transactions on Pattern Analysis and Machine Intelligence **16**, 824 (1994), arXiv:92 [0-8186-2855-3] .

[11] T. E. Bishop and P. Favaro, *The light field camera: Extended depth of field, aliasing, and superresolution,* IEEE Transactions on Pattern Analysis and Machine Intelligence **34**, 972 (2012).

[12] R. Dandliker, Y. Salvad, and E. Zimmermann, *Distance measurement by multiple-wavelength interferometry,* J. Opt **29**, 105 (1998).

[13] F. Li, J. Yablon, A. Velten, M. Gupta, and O. Cossairt, *High-depth-resolution range imaging with multiple-wavelength superheterodyne interferometry using 1550-nm lasers,* Applied Optics **56**, H51 (2017).

[14] K. L. Boyer and A. C. Kak, *Color-Encoded Structured Light for Rapid Active Ranging,* IEEE Transactions on Pattern Analysis and Machine Intelligence **PAMI-9**, 14 (1987).

[15] P. M. Griffin, L. S. Narasimhan, and S. R. Yee, *Generation of uniquely encoded light patterns for range data acquisition,* Pattern Recognition **25**, 609 (1992).

[16] Y. Zhang, Z. Xiong, Z. Yang, and F. Wu, *Real-time scalable depth sensing with hybrid structured light illumination,* IEEE Transactions on Image Processing **23**, 97 (2014).

[17] R. Lange and P. Seitz, *Solid-state time-of-flight range camera,* IEEE Journal of Quantum Electronics **37**, 390 (2001).

[18] C. Niclass, C. Favi, T. Kluter, and F. Monnier, *Single-Photon Synchronous Detection,* IEEE Journal of Solid State Circuits **44**, 1977 (2009).

[19] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, and W. Brockherde, *100 000 Frames/s 64 x 32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging,* IEEE Journal of Selected Topics in Quantum Electronics **20**, 354 (2014).

[20] N. A. Dutton, L. Parmesan, S. Gnecchi, I. Gyongy, N. J. Calder, B. R. Rae, L. A. Grant, and R. K. Henderson, *Oversampled ITOF Imaging Techniques using SPAD-based Quanta Image Sensors,* International Image Sensor Workshop , 1 (2015).

[21] C. S. Bamji, P. O'Connor, T. Elkhatib, S. Mehta, B. Thompson, L. A. Prather, D. Snow, O. C. Akkaya, A. Daniel, A. D. Payne, T. Perry, M. Fenton, and V. H. Chan, *A 0.13 µm CMOS System-on-Chip for a 512 × 424 Time-of-Flight Image Sensor with Multi-Frequency Photo-Demodulation up to 130 MHz and 2 GS/s ADC,* IEEE Journal of Solid-State Circuits **50**, 303 (2015).

[22] F. Remondino and D. Stoppa, *TOF range-imaging cameras,* TOF Range-Imaging Cameras **9783642275**, 1 (2013).

[23] D. Snow, R. Mccauley, M. Mukadam, I. Agi, S. Mccarthy, Z. Xu, T. Perry, W. Qian, V.-h. Chan, P. Adepu, G. Ali, M. Ahmed, and A. Mukherjee, *1Mpixel 65nm BSI 320MHz Demodulated TOF Image Sensor with 3.5µm Global Shutter Pixels and Analog Binning,* IEEE International Solid-State Circuits Conference - Digest of Technical Papers , 94 (2018).

[24] B. Behroozpour, P. A. Sandborn, N. Quack, T. J. Seok, Y. Matsui, M. C. Wu, and B. E. Boser, *Electronic-Photonic Integrated Circuit for 3D Microimaging,* IEEE Journal of Solid-State Circuits **52**, 161 (2017).

[25] W. Koechner, *Optical ranging system employing a high power injection laser diode,* IEEE transaction on aerospace and electronic systems **4**, 81 (1968).

[26] T. Ruotsalainen, P. Palojärvi, and J. Kostamovaara, *A wide dynamic range receiver channel for a pulsed time-of-flight laser radar,* IEEE Journal of Solid-State Circuits **36**, 1228 (2001).

[27] J. Nissinen, I. Nissinen, and J. Kostamovaara, *Integrated receiver including both receiver channel and TDC for a pulsed time-of-flight laser rangefinder with cm-level accuracy,* IEEE Journal of Solid-State Circuits **44**, 1486 (2009).

[28] H. S. Cho, C. H. Kim, and S. G. Lee, *A high-sensitivity and low-walk error LADAR receiver for military application,* IEEE Transactions on Circuits and Systems I: Regular Papers **61**, 3007 (2014).

[29] M. Hintikka and J. Kostamovaara, *A 700 MHz laser radar receiver realized in 0.18 µm HV-CMOS,* Analog Integrated Circuits and Signal Processing **93**, 245 (2017).

[30] M. Perenzoni, D. Perenzoni, and D. Stoppa, *A 64×64-pixel digital silicon photomultiplier direct ToF sensor with 100Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6km for spacecraft navigation and landing,* 2016 IEEE International Solid-State Circuits Conference (ISSCC) **52**, 118 (2016).

[31] C. Veerappan, J. Richardson, R. Walker, D.-u. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, *A 160×128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter,* ISSCC , 312 (2011).

[32] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. Dalla Mora, D. Contini, D. Durini, S. Weyers, and W. Brockherde, *CMOS imager with 1024 SPADs and TDCS for single-photon timing and 3-D time-of-flight,* IEEE Journal on Selected Topics in Quantum Electronics **20** (2014), 10.1109/JSTQE.2014.2342197.

[33] L. Gasparini, M. Zarghami, H. Xu, L. Parmesan, M. M. Garcia, M. Unternahrer, B. Bessire, A. Stefanov, D. Stoppa, and M. Perenzoni, *A 32×32-pixel time-resolved single-photon image sensor with 44.64µm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800kHz observation rate for quantum physics applications,* ISSCC , 98 (2018).

[34] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, *A 100-m range 10-Frame/s 340×, 96-pixel time-of-flight depth sensor in 0.18-µm CMOS,* IEEE Journal of Solid-State Circuits **48**, 559 (2013).

[35] D. Portaluppi, E. Conca, and F. Villa, *32 × 32 CMOS SPAD Imager for Gated Imaging, Photon Timing, and Photon Coincidence,* IEEE Journal of Selected Topics in Quantum Electronics **24** (2018), 10.1109/JSTQE.2017.2754587.

[36] A. Ximenes, P. Padmanabhan, and E. Charbon, *Mutually Coupled Time-to-Digital Converters (TDCs) for Direct Time-of-Flight (dTOF) Image Sensors,* Sensors **18**, 3413 (2018).

[37] 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 Journal of Solid-State Circuits **49**, 315 (2014).

[38] A. T. Erdogan, R. Walker, N. Finlayson, N. Krstajic, G. O. Williams, and R. K. Henderson, *A 16.5 giga events/s 1024 × 8 SPAD line sensor with per-pixel zoomable 50ps-6.4ns/bin histogramming TDC,* IEEE Symposium on VLSI Circuits, Digest of Technical Papers , C292 (2017).

3

OVERVIEW OF SINGLE-PHOTON AVALANCHE DIODES

As the core imaging device in this thesis, an overview of SPADs is presented in this chapter from basic operation principle up to SPAD sensor architectures. Section 3.1 discusses the general SPAD structure and operation principle. While in section 3.2, key parameters for SPADs, including PDP, DCR, dead time, jitter, afterpulsing and crosstalk, are presented with emphasis on the trade-offs between different parameters. At the system level, pixel circuits and sensor architectures are described in section 3.3, including the quenching and recharge approaches, digital and analog photon counters, time gating and time-to-digital converters. Compared to conventional CMOS image sensors, apart from the single-photon detection and high timing resolution, SPADs present a non-linear counting response which is discussed in section 3.4. Finally, section 3.5 concludes the chapter.

3.1. SINGLE-PHOTON AVALANCHE DIODE OPERATION

A photodiode is a reverse biased p-n junction. Depending on the reverse bias voltage, photodiodes can operate in three different modes, namely linear, proportional and Geiger mode. Figure 3.1 shows the I-V characteristics of a diode working in different bias conditions.

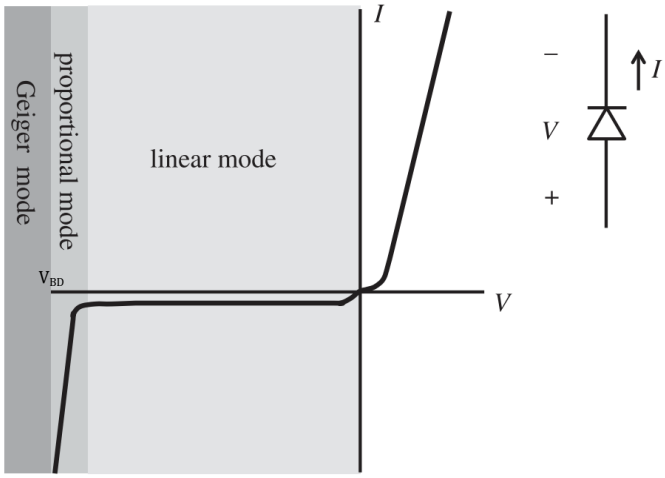


Fig. 3.1 I-V characteristics of a photodiode. A conventional photodiode operates in linear mode with a unity photon-electron gain. APDs and SPADs operate, respectively, slightly below and above breakdown voltage, V_{BD}, where the photon-electron gain ranges from tens of units to infinity.

A SPAD is a p-n junction reverse biased above its breakdown voltage, V_{BD}, in so-called Geiger mode. When a photon is absorbed in the depletion region, it can generate an electron-hole pair which is split and accelerated by the electric field. If the energy of the electron or hole is sufficiently high, more electron-hole pairs can be generated by impact ionization, triggering a self-sustaining avalanche. This avalanche phenomenon takes place when the electric field strength is higher than that of the critical field, E_{cr}, at which the impact ionization of carries happens [1]. In silicon, $E_{cr} \approx 3 \times 10^5$ V/cm. Once the avalanche is initiated, a large current, at the level of milliamperes, can flow through the device until its destruction. For this reason, a resistor R_q is generally connected in series with the SPAD, as shown in Fig. 7.1(a). This resistor is typically in the order of kilo-ohms; it quenches the avalanche by reducing the current to less than 100 μA when the anode voltage of V_A increases toward to the excess bias voltage, V_{EB} [2]. After quenching is complete, the recharge process starts, bringing the device back to idle mode.

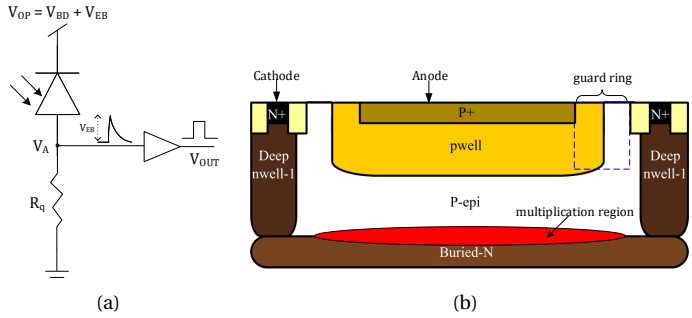


Fig. 3.2 (a) Simple SPAD front-end; (b) Cross section of a CMOS SPAD from [3].

In Geiger mode of operation, the optical gain of the photodiode is virtually infinite, limited only by the number of carriers involved in the avalanche. Thus, a large signal with an amplitude of a few volts or milli-amps can be generated in a short time with a single photon detection. If the output of the SPAD is connected to a voltage discriminator, e.g. a buffer, a digital signal V_{OUT} rising from logic '0' to '1' is generated, indicating the arrival of a single photon. By connecting this digital output signal to a time-to-digital converter (TDC), we can measure the photon arrival time directly.

A major advantage of the SPAD sensors is the CMOS compatibility that both the SPADs and circuits can be implemented on the same wafer. This enables the SPAD sensors to benefit from the scaling of CMOS technologies, including array size, power consumption, TDC resolution, low cost and massive production. An example of a SPAD designed in a 180 nm CMOS process [3] is illustrated in Fig. 7.1(b), which has been implemented in Piccolo and Ocelot. In this design, the pwell (PW) functions as anode and the buried-nwell (BN) as cathode. The BN layer assures substrate isolation, while the nwell-1 provides connection between N+ and BN. In order to avoid premature edge breakdown, a guard ring was implemented, consisting of a pwell lateral diffusion and a lightly doped p-epi. The characterization of this SPAD is illustrated in Chapter 5.

3.2. KEY PROPERTIES OF SPADS

This section describes the key properties of a SPAD to facilitate the understanding of the challenges and trade-offs for developing dTOF imaging systems. More detailed physics analysis can refer to [1, 4].

3.2.1. PHOTON DETECTION PROBABILITY AND FILL FACTOR

In conventional CMOS and CCD image sensors, the optical sensitivity of a detector is usually expressed by means of quantum efficiency (QE). This simply indicates the av-

erage percentage of photons incident the active area of a detector that produces an electron-hole pair. Since the penetration depth of light is wavelength-dependent, as shown in Fig. 3.3, the QE varies with the depletion width. From this figure, we can see the penetration depth can be over 30 μm for wavelength of 900nm. With a small absorption coefficient, the total photon absorption is approximately proportional to the depletion width, i.e.

$$P_{abs} = 1 - e^{-\alpha L} \approx -\alpha L. \tag{3.1}$$

where P_{abs} is the photon absorption probability; α is the absorption coefficient and L is the depletion width. Taking 900 nm wavelength as an example, since the absorption coefficient $\alpha \approx 306cm^{-1}$, for a depletion width of 5 μm, only about 15% of the photons will be absorbed, indicating that a wide depletion region is required to improve QE in near-infrared region.

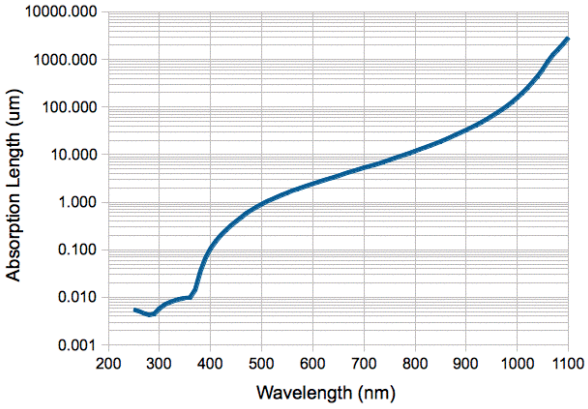


Fig. 3.3 Mean penetration depth in silicon as a function of wavelength.

For a SPAD, only photons that give rise to an avalanche are counted as valid events. Therefore, the turn-on probability is governed by both, the availability of carriers for triggering governed by the QE and the probability of a self-sustaining avalanche initiated by an electron-hole pair, which is known as breakdown probability [5]. In SPAD devices, photon detection probability (PDP) is typically used to indicate the percentage of photons triggering avalanche events over the number of photons illuminated on the multiplication region of the SPAD. In a pixel, we define fill factor as the ratio between the active area of detection and overall area of the pixel. Therefore, one term, photon detection efficiency (PDE), is typically defined as the percentage of photons triggering avalanche events over the number of photons illuminated on the entire SPAD area, which can be

mathematically calculated as the product of the PDP and fill factor.

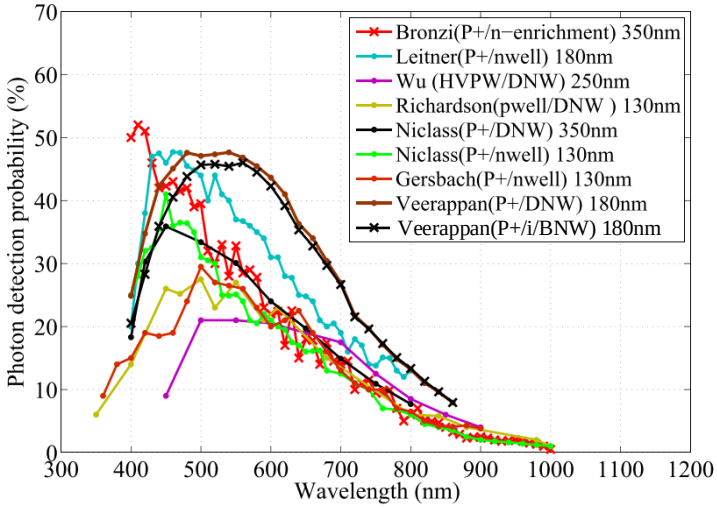


Fig. 3.4 PDP variation as a function of wavelength with different SPAD structures.

Theoretically, PDP can be defined as the product of the QE and breakdown probability. The PDP as a function of wavelength with different SPAD structures is shown in Fig. 3.4. In order to improve QE, n+/p type SPADs with deep junction and thick depletion region were presented in [6, 7], where in [7] a PDP of 40% at 800 nm was achieved at an excess bias of 20 V. Nevertheless, since the anode of these SPADs is shared with the p-type substrate which needs to be biased at ground voltage, poly resistors have to be used for quenching operation and the avalanche needs to be detected via capacitive coupling approach [6]. This limits the performance of pixel front-end circuits. Besides, the electric field across the depletion region determines the ionization rate as well as the breakdown probability. The PDP of a SPAD can be improved by increasing the SPAD excess bias, but at the cost of increased DCR. Recently, a novel SPAD based on nano-textured structure and silicon-on-insulator (SOI) implementation was presented in [8]. Instead of increasing the depletion width, photon absorbtion length is increased by diffracting the vertical incident light into a horizontal waveguide mode, which improves the PDP with a thin SPAD. With this technique, up to 3x PDP improvement was achieved, i.e., 32% at 850 nm, but the DCR also increased almost in the same order of magnitude and severe crosstalk would be expected. A similar concept was presented in [9], where it uses shallow trench isolation (STI) for light diffraction. Even though this process is fully CMOS compatible and without DCR degradation, compared to [8], a relatively low PDP improvement of

25% at 850 nm was reached.

3.2.2. DARK COUNT RATE

Dark counts are the avalanche events triggered by the carriers without incident photons hitting the SPADs. The generation of these carriers can be categorized into two factors, comprising the thermal generation and the tunneling effect [2]. In thermal equilibrium the carrier generation and recombination processes are in dynamic equilibrium. However, due to the high electric field, these carriers could trigger avalanches before recombination, resulting in dark counts. Moreover, when the system is supplied with additional energy, such as the influence of temperature, more carriers are generated, indicating a strongly thermal dependency. A special thermal mechanism is the trap-assisted carrier generation and recombination, where the traps are formed due to the crystalline defects and the impurities. Large number of carriers can be generated within the traps, resulting in the SPAD DCR orders of magnitude higher than the normal value [10]. In an array, such SPADs can reach up to 0.5% of the entire population ('hot pixels'). The pixels with at least 10x more DCR than the median are generally amounting to about 10-25% of the entire population of SPADs. They are usually referred to as 'laughers' [11]. One example is shown in Fig. 3.5, presenting the DCR distribution measured with a 128 × 512 SPAD array.

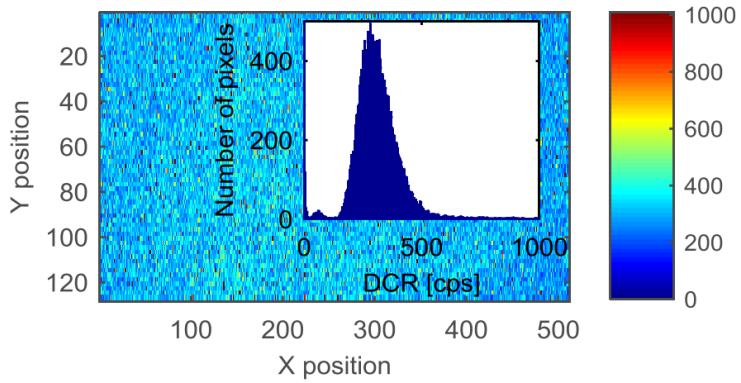


Fig. 3.5 DCR distribution over a 128 × 512 SPAD array, revealing an average of 1169 cps and a median of 302 cps [11].

Carrier generation due to band-to-band tunneling effect occurs when the doping concentration of the p and n sides of the junction are very high. In this case, the depletion region is very narrow and the electrons in the valence band have a probability of tunneling across the band gap to the conduction band. Tunneling effect depends strongly

on the electric field, whilst it is insensitive to temperature variation. Therefore, we can characterize the tunneling related DCR by cooling down the SPAD until its DCR becomes weakly dependent on the temperature, as shown in Fig. 3.6 [11].

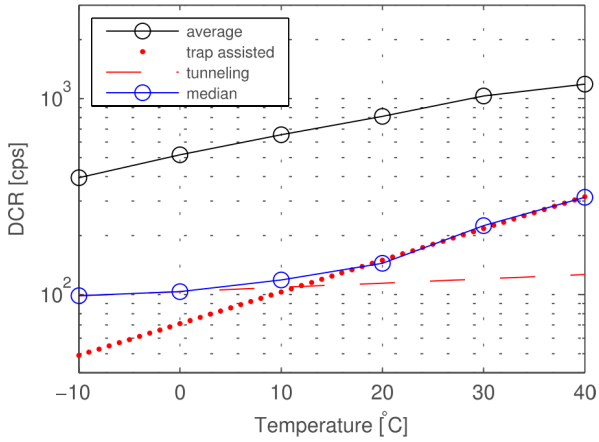


Fig. 3.6 DCR over temperature. The crossing of the dashed and dotted line indicates the cutoff temperature, at which band-to-band-tunneling DCR becomes dominant over trap-assisted DCR [11].

3.2.3. DEAD TIME

As previously mentioned in the Section 3.1, an avalanche event needs to be quenched to prevent the SPAD destruction. As shown in Fig. 7.1(a), the quenching operation is typically performed with a resistor to quench the avalanche. After the quenching, a recovery phase follows to discharge the anode, which brings the SPAD back to the original biasing condition for the next photon detection. Since the quenching resistor is typically hundreds of kilo ohms and the capacitance seen on the quenching node is in the order of a few tens of femto-farads, the discharge time can be from tens of ns to 1 μs. As a result, the photon detection ability of the SPAD is dramatically reduced or even disabled during this time, which is defined as dead time.

Active quenching and recharge is an alternative mechanism that may further reduce or at least control dead time more precisely. It is usually performed using a feedback loop, which detects the avalanche and actively recharges the SPAD by bringing the anode or cathode to the idle position. This feedback loop is generally controllable, so as to obtain a fixed dead time, during which the SPAD is completely inactive, whereas in passive quenching the SPAD is partially active during the recharge.

Dead time comprises the quenching and recharge times, which is a unique characteristic compared to conventional photodetectors, e.g. CIS, CCD. For these devices, photons are detected continuously without any dead time. For a SPAD, since the dead time restricts the maximum photon detection rate as well as the dynamic range, it should be as short as possible. However, as will be discussed in Section 3.2.5, dead time also has an impact on afterpulsing probability, which typically determines the limit of the SPAD dead time.

3.2.4. TIMING JITTER

One of the major features of SPADs is the jitter of the timing response. The leading edge of a SPAD output indicates the arrival time of photons. For a given flight time, the statistical fluctuation of the arrival time from a SPAD is defined as the timing jitter or timing resolution, which is typically characterized as the full-width-at-half-maximum (FWHM) of the underlying Gaussian distribution, as shown in Fig. 3.7 [3]. Due to the avalanche multiplication process, a sharp leading edge can output from the SPAD, leading to an extremely small timing jitter. Sub-hundred picoseconds timing jitter at FWHM has been reported in literature, where a jitter of 29.8 ps and 7.8 ps was achieved in [12] and [13], respectively.

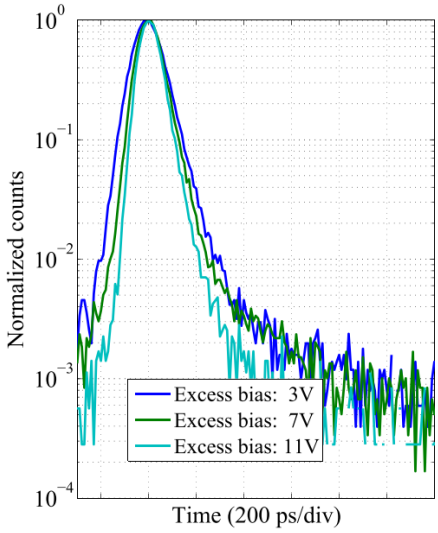


Fig. 3.7 SPAD jitter measurement at different excess bias voltage at 405 nm wavelength, where the FWHM jitter improves from 133 ps at 3 V excess bias to 97.2 ps at 11 V excess bias [3].

The contributing factors of the timing jitter include the avalanche process and the tim-

ing pick-up circuit. The lower limitation of timing jitter is given by the carrier transit delay uncertainty that starts from the generation points in the depletion region to the multiplication region. Moreover, the ionization coefficient difference with the electrons and holes as well as the fluctuation of the avalanche buildup process broaden the timing jitter. Similar to PDP, the timing jitter typically can be improved by raising the SPAD excess bias. In the meantime, the timing pick-up circuit plays an important role in the jitter measurement. The avalanche propagation process fluctuates with respect to its starting location and the avalanche propagation. While at the beginning of the avalanche, less fluctuation is accumulated. Therefore, a better jitter performance can be obtained when using a low threshold voltage with a discriminator [13, 14].

3.2.5. AFTERPULSING

As is discussed in Section 3.2.2, traps, such as the defects in the lattice and the impurities, produce uncorrelated noise as well as correlated noise in the form of afterpulsing. In a SPAD, when an avalanche occurs, a large volume of carriers will flow through the depletion region and some of them may be captured by the trapping centers. The release of these carriers follows a statistically fluctuating delay, depending on the traps involved. If the subsequential release of the carriers occurs after the dead time, a secondary avalanche could be triggered, generating afterpulses correlated with the previous avalanche pulse. The number of carriers captured during the avalanche increases with the total number of carriers crossing the depletion region. Therefore afterpulsing probability increases with the SPAD junction capacitance, i.e., SPAD pitch, as well as the excess bias. Techniques, such as active quenching, have been reported and is discussed in Section 3.3.2. Meanwhile if the carriers are released during the SPAD dead time, no afterpulses would be produced. This indicates an increased dead time reduces the afterpulsing, but at the cost of photon detection rate.

Due to the correlation between primary and afterpulses, afterpulsing probability (AP) can be characterized by measuring the inter-arrival time between adjacent SPAD events. In a ideal situation where there is no afterpulsing effect, the histogram of the inter-arrival time with uncorrelated events follows a single exponential decay shape due to the Poisson nature of light and of dark pulses as detected by a SPAD. Afterpulsing manifests itself as a super-exponential behavior. An example of SPAD response with afterpulsing is shown in in Fig. 3.8 from [3]. By fitting the histogram to a single exponential decay, the events in excess represent afterpulses. In this example, an AP of 7.2% was measured at a dead time of 300 ns and 11 V excess bias.

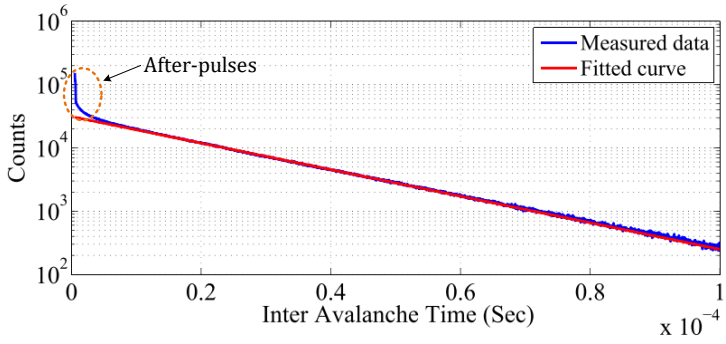


Fig. 3.8 An Afterpulsing measurement example from [3].

3.2.6. CROSSTALK

Another correlated noise source is crosstalk from the avalanche which is triggered by neighboring SPADs. The mechanism of crosstalk can be categorized into electrical and optical crosstalk, as shown in Fig. 3.9.

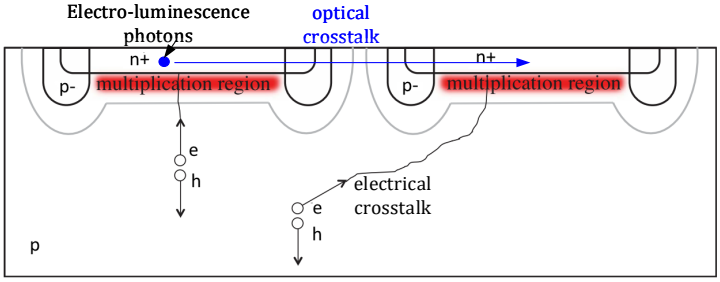


Fig. 3.9 The two mechanisms of crosstalk, including electrical and optical crosstalk.

Electrical crosstalk is due to carrier exchange. When carriers are generated in the deep quasi-neutral region, such as the substrate, they might diffuse laterally and trigger avalanches in a neighboring SPAD. Since the light penetration depth strongly depends on the wavelength, a higher electrical crosstalk is expected in the red and near-infrared ranges. Electrical crosstalk can be reduced by isolating the SPADs from the substrate. A widely used SPAD structure is p+/p-well or n-well/deep n-well, where the deep n-well is used as the cathode and isolates the carriers from the substrate [3, 12, 15, 16]. However, since the carriers generated in the deep substrate can't trigger avalanche events, the PDP performance is decreased in the near infrared, e.g. <5% PDP at 900 nm.

Optical crosstalk in SPADs is owed to the avalanche triggered by electro-luminescence photons. During an avalanche, some photons may be excited by the impact of high-energy electrons, leading to the electron-luminescence phenomenon. These photons may propagate laterally and be detected by the neighboring pixels. Since the electro-luminescence intensity is proportional to the the number of carriers generated in an avalanche, optical crosstalk can be reduced by reducing the diode capacitance and excess bias voltage. In the meantime, avalanche related carriers also can be reduced by employing an active quenching scheme. Besides, if each SPAD is isolated laterally with light absorbing materials, the propagation of emitted photons will be stopped before reaching the neighboring SPADs, thus reducing the optical crosstalk. This method is applied in all our designs by using a deep trench isolation (DTI) layer, where 0.09% optical crosstalk is achieved.

The crosstalk performance is characterized by crosstalk probability, which is defined as the probability of an avalanche event triggered by neighboring SPADs. It usually can be evaluated by measuring the correlation between SPADs. For example, one can measure the DCR fluctuation of a SPAD when turning off/on an adjacent SPAD [17]. Similar to afterpulsing characterization, time correlation method also can be used to characterize crosstalk probability, where the inter-arrival time is measured between two SPADs. Again in this case, the inter-arrival distribution should be exponential.

3.3. SPAD SENSOR CIRCUITS AND ARCHITECTURES

As is discussed in Section 3.2, a major advantage of SPADs is CMOS compatibility. Thus we can fully take the advantage of Moore law to scale up the design to achieve higher spatial and timing resolution, lower power consumption, more functionality at a lower price. In this section, an overview on the architectures and circuits based on SPADs are given. A number of factors are illustrated, including sensor architectures, quenching and recharge, time gating, photon counting techniques.

3.3.1. SPAD SENSOR ARCHITECTURES

To design a SPAD sensor, an optimal sensor architecture should be chosen depending on the applications and technologies. A basic architecture comprises several factors that need to be considered, including the SPAD choice, pixel circuit, pixel grouping, timing method, TDC architecture, readout mechanism and on-chip processing. These factors and its variations are summarised in Table. 3.1, where part of them are discussed in detail in the following sections.

Table 3.1: Factors in the design of SPAD sensors

Factors	Variations
SPAD choice	P-on-N type SPAD N-on-P type SPAD
Pixel circuits	Active or passive quenching and recharge Directly connected or capacitive coupling SPADs Pixel enabling and disabling Digital or analog photon counter Hold-off circuits
Time acquisition	iTOF, Time gating approach dTOF, Time-to-digital converters dTOF, Time-to-amplitude converters
Pixel grouping	Per-SPAD TDC architecture TDC sharing with a bus/multiplexer/dynamic reallocation approach TDC sharing with an OR/XOR tree
Readout	Event-driven readout Frame based readout
Interference	Phase coded pulses
Performance enhancement	On-chip histogramming Delta-sigma averaging Filtering Spatial and temporal correlation Weighted photons Processing based on artificial intelligence algorithms

3.3.2. QUENCHING AND RECHARGE CIRCUITS

The quenching and recharge circuit can be as simple as a resistor. However, to improve the performance of SPADs, such as the dead time, afterpulsing and crosstalk, quenching and recharge circuits based on active elements, in which transistors have been widely used in the literature. According to the operating mode, there are basically four combinations of quenching and recharge circuits, including:

- Passive quenching, passive recharge (PQPR)
- Passive quenching, active recharge (PQAR)
- Active quenching, passive recharge (AQPR)
- Active quenching, active recharge (AQAR)

PASSIVE QUENCHING PASSIVE RECHARGE

Passive quenching and passive recharge is the most commonly used technique in pixel design, due to the simplicity and low area occupancy. It is usually implemented with a poly resistor, a NMOS or a PMOS transistor, as shown in Fig. 3.10. The main advantage of using a poly resistor is that it can tolerant a high bias voltage. A typical application is in the quenching of a n+/n-well/p-sub type SPADs [6, 18], shown in Fig. 3.10 (a). To maintain the CMOS compatibility, the substrate as well as the anode of the SPADs have to be biased at ground voltage, requiring the cathode connecting to a high bias voltage, VOP. Since the VOP is typically at the level of 20 V, which exceeds the working voltage range of a normal transistor, a poly resistor has to be employed in such a situation. Meanwhile, to pickup the avalanche signal, a capacitive coupling method is used as an interface between high voltage and CMOS logic voltage [6, 18]. As a rule of thumb, the quenching resistance can be 50 kohm per 1 V of excess bias voltage[2]. However, once the resistor is implemented on silicon, the resistance is fixed, leading to a non-controllable dead time.

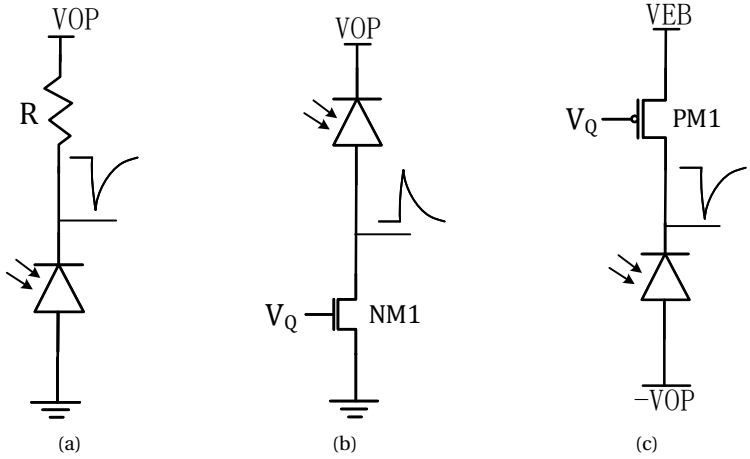


Fig. 3.10 Passive quenching and recharge based on (a) a resistor, (b) a NMOS transistor and (c) a PMOS transistor.

For some types of SPADs, such as a p+/p-well/deep n-well SPAD, a transistor can be used for PQPR [3], as shown in Fig. 3.10(b) and (c), where a NMOS and a PMOS is used, respectively. To extend the range of excess bias voltage, thick oxide transistors are preferable than thin oxide transistors. Compared to the poly resistor based PQPR, the resistance of the transistors can be controlled by the gate voltage V_Q. In this case, a variable dead time can be obtained, which is very useful to achieve a better trade-off between dead time and afterpulsing probability. Furthermore, if both the NMOS and PMOS transistor are appli-

cable, a NMOS is preferred, owing to area occupancy. Since the implementation of a n-well has to comply with the design rule check (DRC), including the spacing between n-wells at different potential and the spacing between n-well and SPADs, a larger spacing margin has to be made, which increases the pixel pitch and reduces the fill factor. However, due to the transistor voltage tolerance constraint, the maximum excess bias is limited. This limit can be improved by cascoding transistors; a 4.4-V excess bias is achieved with cascoded 2.5-V thick oxide transistors in [19].

ACTIVE QUENCHING ACTIVE RECHARGE

As is discussed in section 3.2.5 and 3.2.6, both the afterpulsing and crosstalk probability increases with the number of carriers crossing through the depletion region during an avalanche. To reduce the number of carriers, one can quench the avalanche as quickly as possible with the assistance of additional electronics. An example is shown in Fig. 3.11(a). Upon avalanche detection, the anode V_{AQPR} is gradually charged, through a low impedance PMOS transistor for a certain time τ_Q [20]. Similarly, in Fig. 3.11(b), the active recharge circuit discharges the anode V_{PQAR} back to the idle state to re-activate the SPAD for the next detection. The time τ_R can be determined by comparing V_{PQAR} against a threshold voltage [21] or by a delay element [22]. A combination of active quenching and active recharge circuit is shown in Fig. 3.11(c). For an extensive study of quenching recharge mechanism and circuits, one can refer to [2].

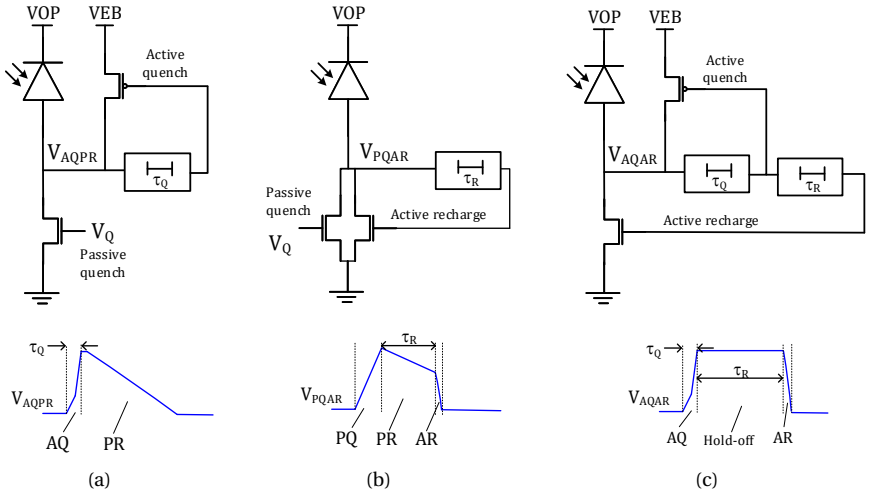


Fig. 3.11 Circuit block diagrams and anode voltage waveforms of (a) active quenching passive recharge, (b) passive quenching active recharge, and (c) active quenching active recharge.

In spite of the benefits of AQ, it has not been widely used in recent works, especially in large array designs. The main reason is the area limitation in planar technologies and the reduction of fill factor. Nevertheless, this could be improved by leveraging new technologies, such as BSI and 3D stacking technology. On the other hand, AQ is more suitable for large SPADs [2], where a large diode capacitance is involved during the quenching and recharge. On the contrary, AR has been implemented in multiple works [19, 21–23]. With the scaling of the SPAD pitch, diode capacitance can be reduced significantly. High SPAD performance, in terms of dead time and afterpulsing probability, has been achieved with PQAR. For example, in [19] a dead time of 8 ns and 0.08% afterpulsing probability was achieved at a SPAD pitch of 18.36μm at 4.4 V excess bias. Moreover, apart from the reduced dead time, non-paralyzable detection can be realized with AR, which further improves the photon detection rate and is discussed in section 3.4.

3.3.3. PHOTON COUNTERS

When a SPAD is connected via a logic cell, such as an inverter, it becomes a truly digital imaging device, where every detected photon is represented as a digital pulse. To realize 2D imaging, different kinds of circuits have been developed to count the pulses in either digital or analog approaches. The most compact pixel structure to perform photon counting is using a single bit memory, where a capacitor or a latch is used to retain the status of the SPAD output. With such a compact pixel structure, a 512 × 512 SPAD array was reported in [24], which achieves, so far, the highest spatial resolution of SPAD sensors. The pixel circuit is shown in Fig. 3.12, where a NMOS capacitor T9 is used as a 1-bit memory. To perform the detection, T9 is firstly reset via T8 with 'Reset'=1. Then, within the gating time that 'Gate'=1, T9 is charged and maintained at a high voltage by the first detected photon, until it is read out and reset once again. Compared to conventional pixels in active pixel sensor (APS) technology where the major noise source is the readout, a SPAD sensor has virtually no readout noise, due to pixel level digitalization. However, this structure limits only one event being detected in each reset-exposure-readout cycle. This binary feature has been studied and a Quanta Image Sensor (QIS) concept has been proposed by Fossum [25], where a SPAD sensor [26] offers the first QIS demonstration.

To achieve multi-bit counting, the simplest way is to build a digital counter with flip-flops, which counts the photons at the leading edge of the pulses. However, due to the flip-flop area overhead, this techniques is usually implemented in a per-pixel TDC architecture, where the counter records either the number of clock periods in photon-timing mode or the number of photons in photon-counting mode[27]. As an alternative, analog counters have been investigated in literature, where the charge in a capacitor is accumulated a number of timings corresponding to the counts, resulting in a voltage propor-

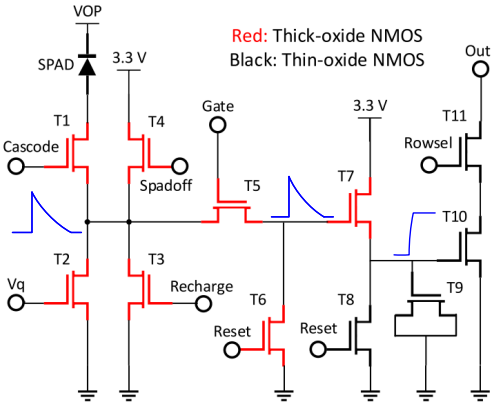


Fig. 3.12 Pixel circuit of [24], where a NMOS capacitor T9 is used as a 1-bit memory.

tional to that number [26, 28–30]. Take [30] as an example, where the pixel circuit and timing diagram are shown in 3.13(a) and (b), respectively. Initially, during the RESET phase, the NMOS capacitor C_{M6} and node B are precharged to VDD. If there are no photon events, M3 and M4 are at the off-state which holds the voltage of node D, while M5 is at the on-state that capacitor C_p and node C are charged to V_{REF2}. At the leading edge of photon detection that the SPAD signal IN rises from low to high, M3 and M4 are switched on, while M5 is changed to the off-state, which opens a charge transfer path from C_{M6} to C_p through M3 and M4. Therefore, the voltage step variation of node D is proportional to amount of charges transferred to C_p, which is determined by C_{M6}, C_p, V_{REF1} and V_{REF2}. The analog counter approach is very similar to the conventional APS working operation where the charges are transferred to the floating drain (FD) and then converted to a voltage output, and the voltage step variation can be treated as the conversion gain. In APS approach, since only one charge is generated at a photon detection, the conversion gain is simply limited by the FD capacitance. Due to this limited conversion gain, to achieve single photon detection and readout, an extremely low read noise circuit is required [31], which complicates the readout design. In contrast, a variable and much larger conversion gain can be achieved with SPAD-based analog counter. For example, a conversion gain in a range of 1 mV to 16 mV was reached in [30], which significantly reduces the complexity of the readout circuit.

3.3.4. TIME GATING

Benefiting from the fast time response, accurate photon timing information can be obtained by SPAD sensors without using a timer. This approach is known as time gating, and it has been widely used in applications such as fluorescence lifetime imag-

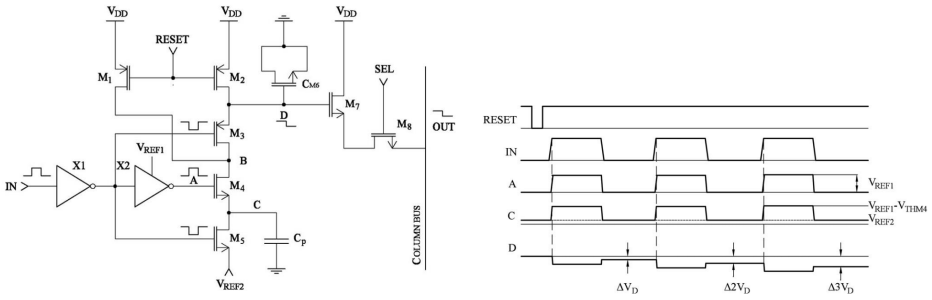


Fig. 3.13 (a) Analog readout of [30] based on charge transfer principle and (b) the operating timing diagram, where a conversion gain of ΔV_D is generated at every photon detection.

ing (FLIM)[17, 24, 32]. To illustrate this approach, one example is shown in Fig. 3.12, where T5 is used to switch on/off the photon detection path according to the input signal "Gate". It is normally a pulsed window signal with a length of tens of ns, where photon detection is only enabled during this time window. By shifting the window with time, a full range time response can be acquired with a scan operation, as shown in Fig. 3.14. Since the shifting step Δt can be as small as tens of ps, high timing resolution can be achieved.

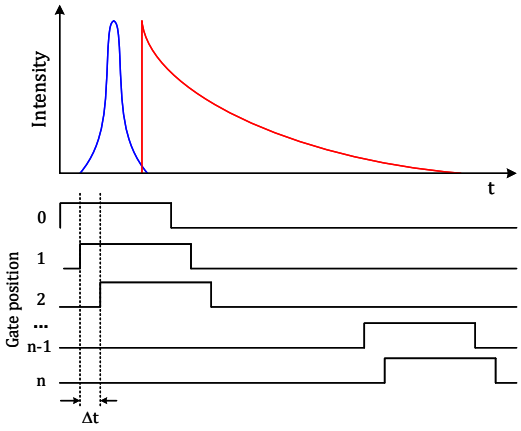


Fig. 3.14 Timing diagram of gating operation in FLIM. The blue signal represents the illumination light, while the red line is the fluorophore excitation signal. During each measurement, the time window is shifted by N times at a time step of Δt. To achieve sufficient SNR, this process will be repeated multiple times, in which the resulting histogram will closely match the intensity decay of the fluorophore.

Since the gating operation is performed within the pixel circuit and only one additional

transistor is required for the implementation, time-resolved SPAD sensors with large pixel array and pixel-wise gating operation have been reported in [17, 24, 32]. However, the drawbacks are also very obvious. Since only a small proportion of the photons can be accessed at one time, it reduces the photon efficiency. Therefore a long acquisition time is required for the window shifting operation, potentially resulting in low imaging speed and loss of sensitivity comparable with a reduced fill factor.

3.3.5. TIME-TO-DIGITAL CONVERTERS

To improve photon efficiency and directly measure the single photon timing feature, time-to-digital converters (TDCs) have been intensely used in SPAD sensors. Similar to an analog-to-digital converter (ADC), a number of properties needs to be considered when designing a TDC, including resolution or bin size (LSB), nonlinearity, dynamic range, conversion rate, power and area. Moreover, for a SPAD image sensor, it normally also includes a SPAD pixel array, e.g. 32 × 32 [33], 160 × 120 [34], 252 × 144 [35]. In order to acquire the time information of all the SPADs, a large number of TDCs are required, indicating factors, such as scalability, uniformity, easy-calibration and power line I-R drop immunity, need to be taken into account as well.

According to the operating principle, TDCs can be categorized into two types comprising analog and digital TDCs [36]. In an analog-type TDC, generally known as time-to-amplitude converter (TAC), a capacitor is charged with a constant current or a voltage ramp at the detection of a photon and stopped by the leading edge of a reference clock. This will produce a voltage variation over the capacitor, which is linearly proportional to the time interval between the photon and the reference clock. In order to improve the TAC resolution, time stretching method is normally utilized. The basic principle of this method is that a capacitor is charged with a current I_1 and then discharged back to the initial voltage with another current I_2 which is N times smaller than I_1. During this process, the discharging time is stretched N times compared to the charging time. Therefore, a digital counter, which is synchronized with a reference clock, can be used to digitize the discharging time. A larger stretching factor can be used to achieve a finer resolution, e.g. a stretching factor of 5000 is used in [37] which achieves a resolution of 7.8 ps. However, the larger the stretching factor the longer of the conversion time, which results in a low conversion rate. Furthermore, since it is based on an analog approach, many challenges need to be handled, such as the nonlinearity of the MOS capacitor over voltage, the analog noise sources (e.g. kT/c, thermal, etc), the charging/discharging current mismatch over temperature, power consumption and circuit area. One TAC based SPAD sensor was reported in [33], where a 32 × 32 SPAD array with per-pixel TAC architecture was presented. In this design, a stretching factor of 40 was applied, achieving a

resolution of 160 ps on a 20 ns time range with a uniformity across the array within ±2 LSB. The pixel pitch is 50 μm, while a fill factor of less than 5% was achieved. Besides, the pixel power consumption is also high, reaching 200 μW and 100 μW with the analog and digital components respectively.

Digital TDCs are generally more preeminent in SPAD sensors arrays due to a more compact implementation and potentially better performance per μm^2. This is due to the use of digital logic exclusively. A widely used digital TDC is based on a tapped delay line (TDL) consisting of a number of delay stages of buffers or inverters. The start signal propagates through the delay line, and the status of each delay stage is recorded by a flip-flop upon the arrival of the stop signal. However, the resolution is limited by the delay of an inverter, which has the shortest delay in CMOS technology. To improve resolution, a vernier-delay line (VDL) can be used, which consists of two delay chains with different delay units [38]. Subgate delay resolution can be achieved, e.g. a 17 ps resolution was achieved with a 350 nm technology in [39], at the expense of doubled number of delay elements and long conversion time. On the other hand, the dynamic range is limited by the number of delay stages in a line structure. In TOF 3D imaging applications, to achieve a maximum 150 m detection capability, a time range of 1 μs is required. To extend the dynamic range, ring-oscillator (RO) based TDCs have been widely used in SPAD sensors, where a basic diagram is shown in Fig. 3.15 [34]. The photon detection signal enables the operation of the RO, which drives a digital counter at a certain frequency determined by the biasing condition. The oscillation is stopped by a reference clock, which freezes the counter and the phases of each delay stage of the RO. After decoding the phases, the time of arrival can be obtained by combining the counter and the decoder outputs, where the resolution is defined by the delay of one RO stage. Similar to the VDL TDCs, a vernier ring-oscillator also can be adopted to achieve subgate resolution[40].

From the diagram we can see that the dynamic range can be easily extended just by increasing the width of the counter, without impacting other factors of the TDC. Moreover, a phase-locked loop (PLL) or delay-locked loop (DLL) can be used to latch the RO at a pre-defined frequency, leading to an enhanced immunity to process, voltage and temperature (PVT) variations. Based on this, an array of TDCs can be built, where every TDC can be driven by a set of phase-shifted clocks from a PLL or DLL, which results in a high uniformity among TDCs [27, 41]. Similar concepts have been implemented in different approaches, e.g. a TDC array based on coupled ROs in [42], a dual-clock TDC array in Ocelot in Chapter 6, where the core idea is to use one or a set of global clocks as a reference for all the TDCs, so as to synchronize all the TDCs to achieve a high uni-

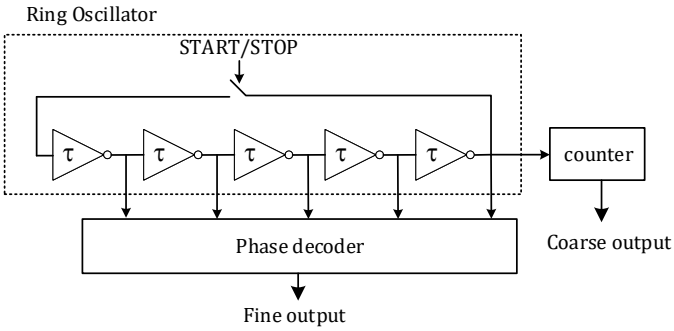


Fig. 3.15 Diagram of a basic RO based TDC.

formity. While a large power consumption can happen on the clock distribution trees, which might limits the scalability of the TDC array. This issue is partially solved with the dual-clock approach in Chapter 6, which achieves a trade-off among uniformity, linearity and power consumption. Nevertheless, compared to analog circuits, apart from the high immunity to noise, another property of digital circuits is the perfect compatibility with the scaling of CMOS technology, which brings significant benefits in terms of cell delay, power consumption, area, etc. All of these improvements can be applied in digital TDC array design, which makes it becoming a major option of TDCs in SPAD sensors.

Instead of designing a TDC with ASIC, field-programmable gate array (FPGA) is an alternative approach to implement TDCs using the carry chain of an adder to form a delay line. Different architectures of TDCs have been implemented and reported in literatures, e.g. TDL [43, 44], VDL [45]. Since FPGAs are usually designed with advanced technologies, a small cell delay as well as a fine TDC time resolution can be realized, e.g. an LSB of 17 ps was achieved in [43] with a 65 nm Xilinx Virtex-5 FPGA. However, the linearity is relatively poor due to the delay difference of the carry units. At the same time, they are susceptible to the supply voltage and temperature variations, implying that a careful calibration has to be applied. To solve these problems, instead of using the carry chain, delay unit based on the metal routing delay has been proposed in [46], which gives the designer more flexibility to adjust the delay of each stage by changing the routing paths. Meanwhile, since the properties of metal wires are insensitive to temperature and voltage, an improved stability can be achieved. In [46], a 7.4 ps resolution FPGA-based TDC with 1024 delay units was presented, which achieves a DNL (INL)of -0.74(-1.52) to +0.74(+1.57) LSB with a 65 nm Xilinx Virtex-5 FPGA. For a SPAD sensor, since a sharp digital output can be instantly generated at the photon detection due to the intrinsic digitalization, it provides the possibility to combine a SPAD to an external FPGA-based

TDC for TOF measurement. A SPAD camera system with FPGA-based TDCs has been reported in [44], where the 64 TDCs with a 50 ps resolution are shared by 256 SPADs via multiplexers. Even though FPGA offers a cost effective solution, it also constraints the SPAD array size and the photon detection rate due to the limited I/O pads and FPGA resources.

3.4. SPAD PHOTON COUNTING RESPONSE

In conventional CMOS image sensors, since photon-generated carriers are collected by the detector continuously, a linear voltage response with respect to the illumination power is obtained. In comparison, as a binary photon counting device, SPAD presents a non-linear response. The reason of the nonlinearity is due to a fundamental feature of SPADs, the dead time. More specifically, if a SPAD creates a photon-carrier during an existing avalanche process or in the dead time, this photon-carrier will not be detected. This kind of detection missing mechanism leads to a non-linear response between the number of detected avalanche and the total amount of photons impinging on the SPAD [47].

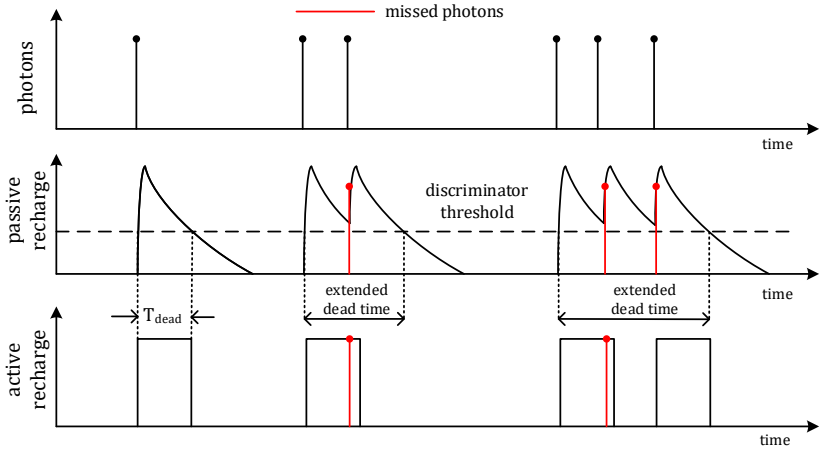


Fig. 3.16 A SPAD recharge diagram example with different mechanisms. In passive recharge, dead time extension can be observed, resulting in 3 missed photons. While with active recharge, 2 photons are missed.

Furthermore, researchers have shown that SPADs present different counting responses according to the recharge mechanism, as shown in Fig. 3.16. As has been discussed in Section 3.3.2, SPADs can be recharged passively or actively. In passive recharge, since the SPAD excess bias increases gradually from 0 V to the default value, there is a probability to trigger a second avalanche during the recharge process. If this happens, the SPAD will

be quenched again and the dead time is extended subsequently, resulting in a so-called paralyzable response. While in active recharge, since the excess bias is maintained at 0 V, no avalanche will be triggered before the SPAD is actively recharged. In this case, a constant dead time is achieved, leading to a non-paralyzable response. As a result, the models of paralyzable and non-paralyzable response can be described by the equations in (3.2) and (3.3) respectively [48].

$$m = n \times exp(-nT_{dead}), \tag{3.2}$$

$$m = \frac{n}{1 + nT_{dead}}, \tag{3.3}$$

where m is the measured photon rate, n is the true photon rate and T_{dead} is the SPAD dead time without extension. Based on the models, a simulated photon counting response is shown in Fig. 3.17. With a dead time T_{dead} of 50 ns, the maximum photon counting rate a SPAD can achieve is $1/T_{dead}$ = 20 Mcps. From Fig. 3.17 we can see that the curves start to diverge from the linear response at about 10% of $1/T_{dead}$. For passive recharge, due to the dead time extension, it can't achieve the expected maximum count rate and a decreasing count rate is observed when n is higher than $1/T_{dead}$. In comparison, the count rate in active recharge keeps rising and gradually saturates at $1/T_{dead}$. Compared to linear mode in which m saturates when n reaches $1/T_{dead}$, the SPAD nonlinear response offers an extended dynamic range, and a detailed analysis can be found in [47, 49, 50]. This response is similar to the one observed by Fossum in the QIS approach, where tiny CIS pixels (jots) are used for single photon detection with higher dynamic range [25, 31]. Besides, considering the decreased response in passive recharge, active recharge is preferable when the count rate is at the same order of $1/T_{dead}$, but at the expense of circuit area and fill factor.

3.5. CONCLUSION

As discussed in this chapter, some circuit blocks are required for a SPAD sensor to achieve time correlated single photon counting, including quenching and recharge circuits, masking and TDCs. Compared to conventional CMOS image sensors where only 3 or 4 transistors are required in one pixel, more transistors are required in SPAD pixels. However, as an imaging device, small pixel pitch is always required to improve image resolution, which brings a big challenge for SPAD sensors to reduce the pixel pitch without sacrificing the fill factor. To solve this problem, a promising option is to design SPADs in a backside-illumination technology, in which a 100% fill factor can in principle achieved

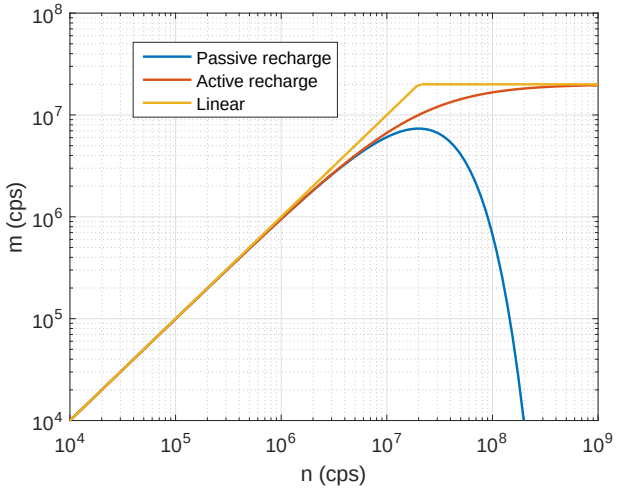


Fig. 3.17 Simulated SPAD photon counting response with passive and active recharge at 50 ns dead time, where active recharge shows a preferable response than the passive recharge at high count rate.

regardless of pixel pitch. Moreover, BSI is usually combined with 3D-stacking which can be another critical boosting technology for SPAD sensors. In this context, advanced technologies, e.g. 28 nm and 14 nm, have a significant advantage in speed, area and power compared to conventional image sensor processes, such as 180 nm, 130 nm and 45 nm. However, these technologies may not be indicated for SPAD pixel design. Therefore, an optimal combination, in the author's opinion, could be a hybrid, by means of 3D-stacking. This idea has been realized and reported in [51], where a 45 nm CIS technology is used for SPAD design and a 65 nm for circuits design.

REFERENCES

[1] E. Charbon, *Single-Photon imaging in complementary metal oxide semiconductor processes,* Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **372** (2014), 10.1098/rsta.2013.0100.

[2] S. Cova, M. Ghioni, a. Lacaita, C. Samori, and F. Zappa, *Avalanche photodiodes and quenching circuits for single-photon detection.* Applied optics **35**, 1956 (1996).

[3] C. Veerappan and E. Charbon, *A low dark count p-i-n diode based SPAD in CMOS technology,* IEEE Transactions on Electron Devices **63**, 65 (2016).

[4] M. W. Fishburn, *Fundamentals of CMOS Single-Photon Avalanche Diodes* (2012).

[5] S. R. R. Oldham W O and P. Antognetti, *Triggering phenomena in avalanche diodes,* IEEE Trans.Electron Devices **19**, 1056 (1972).

[6] 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 Letters **33**, 1589 (2012).

[7] A. Gulinatti, I. Rech, F. Panzeri, C. Cammi, P. MacCagnani, M. Ghioni, and S. Cova, *New silicon SPAD technology for enhanced red-sensitivity, high-resolution timing and system integration,* Journal of Modern Optics **59**, 1489 (2012).

[8] K. Zang, X. Jiang, Y. Huo, X. Ding, M. Morea, X. Chen, C. Y. Lu, J. Ma, M. Zhou, Z. Xia, Z. Yu, T. I. Kamins, Q. Zhang, and J. S. Harris, *Silicon single-photon avalanche diodes with nano-structured light trapping,* Nature Communications **8**, 1 (2017).

[9] L. Frey, M. Marty, S. Andre, and N. Moussy, *Enhancing near-infrared photodetection efficiency in SPAD with silicon surface nanostructuration,* IEEE Journal of the Electron Devices Society **6** (2018), 10.1109/JEDS.2018.2810509.

[10] E. A. Webster and R. K. Henderson, *A TCAD and spectroscopy study of dark count mechanisms in single-photon avalanche diodes,* IEEE Transactions on Electron Devices **60**, 4014 (2013).

[11] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe, and E. Charbon, *Nonuniformity analysis of a 65-kpixel CMOS SPAD imager,* IEEE Transactions on Electron Devices **63**, 57 (2016).

[12] 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 Journal of Selected Topics in Quantum Electronics **24** (2018), 10.1109/JSTQE.2017.2762464.

[13] F. Nolet, S. Parent, N. Roy, M.-O. Mercier, S. Charlebois, R. Fontaine, and J.-F. Pratte, *Quenching Circuit and SPAD Integrated in CMOS 65 nm with 7.8 ps FWHM Single Photon Timing Resolution,* Instruments **2**, 19 (2018).

[14] G. Acconcia, M. Ghioni, and I. Rech, *37ps-precision Time-Resolving Active Quenching Circuit for High-Performance Single Photon Avalanche Diodes,* IEEE Photonics Journal **0655**, 1 (2018).

[15] J. A. Richardson, L. A. Grant, and R. K. Henderson, *Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology,* IEEE Photonics Technology Letters **21**, 1020 (2009).

[16] H. Xu, L. Pancheri, G.-f. D. Betta, and D. Stoppa, *Design and characterization of a p + / n-well SPAD array in 150nm CMOS process,* Optics Express **25**, 77 (2017).

[17] I. Gyongy, N. Calder, A. Davies, N. A. Dutton, R. R. Duncan, C. Rickman, P. Dalgarno, and R. K. Henderson, *A 256×256 , 100-kfps, 61% Fill-Factor SPAD Image Sensor for Time-Resolved Microscopy Applications,* IEEE Transactions on Electron Devices **65**, 547 (2018).

[18] C. Niclass, H. Matsubara, M. Soga, M. Ohta, M. Ogawa, and T. Yamashita, *A NIR-Sensitivity-Enhanced Single-Photon Avalanche Diode in 0.18µm CMOS,* International Image Sensor Workshop , 1 (2015).

[19] 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 Letters **38**, 1547 (2017).

[20] A. Eisele and R. Henderson, *185 MHz Count Rate, 139 dB Dynamic Range Single-Photon Avalanche Diode with Active Quenching Circuit in 130nm CMOS Technology,* IISW , 6 (2011).

[21] C. Niclass and M. Soga, *A miniature actively recharged single-photon detector free of afterpulsing effects with 6ns dead time in a 0.18??m CMOS technology,* Technical Digest - International Electron Devices Meeting, IEDM , 340 (2010).

[22] S. Tisa, F. Guerrieri, and F. Zappa, *Variable-load quenching circuit for single-photon avalanche diodes.* Optics express **16**, 2232 (2008).

[23] D. Portaluppi, E. Conca, and F. Villa, *32 × 32 CMOS SPAD Imager for Gated Imaging, Photon Timing, and Photon Coincidence,* IEEE Journal of Selected Topics in Quantum Electronics **24** (2018), 10.1109/JSTQE.2017.2754587.

[24] A. C. Ulku, S. Member, C. Bruschini, S. Member, and I. Michel, *A 512 × 512 SPAD Image Sensor with Integrated Gating for Widefield FLIM,* IEEE Journal of Selected Topics in Quantum Electronics **PP**, 1 (2018).

[25] E. R. Fossum, *Modeling the performance of single-bit and multi-bit quanta image sensors,* IEEE Journal of the Electron Devices Society **1**, 166 (2013).

[26] N. A. Dutton, L. Parmesan, S. Gnecchi, I. Gyongy, N. J. Calder, B. R. Rae, L. A. Grant, and R. K. Henderson, *Oversampled ITOF Imaging Techniques using SPAD-based Quanta Image Sensors,* International Image Sensor Workshop , 1 (2015).

[27] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. Dalla Mora, D. Contini, D. Durini, S. Weyers, and W. Brockherde, *CMOS imager with 1024 SPADs and TDCS for single-photon timing and 3-D time-of-flight,* IEEE Journal on Selected Topics in Quantum Electronics **20** (2014), 10.1109/JSTQE.2014.2342197.

[28] L. Pancheri, E. Panina, G.-f. D. Betta, L. Gasparini, D. Stoppa, and F. B. Kessler, *Compact analog counting SPAD pixel with 1.9 % PRNU and 530ps time gating,* 2013 Proceedings of the ESSCIRC (ESSCIRC) , 295 (2013).

[29] E. Panina, L. Pancheri, G.-f. Dalla, L. Gasparini, D. Stoppa, E. Panina, L. Pancheri, and G.-f. D. Betta, *Compact CMOS analog readout circuit for photon counting applications,* SPIE optics and optoelectronics **877305** (2013), 10.1117/12.2020975.

[30] E. Panina, L. Pancheri, G. F. Dalla Betta, N. Massari, and D. Stoppa, *Compact CMOS analog counter for SPAD pixel arrays,* IEEE Transactions on Circuits and Systems II: Express Briefs **61**, 214 (2014).

[31] E. R. Fossum, *CMOS Image Sensors and the Quanta Image Sensor,* OMN (2018), 10.1117/12.2309630.

[32] I. M. Antolovic, S. Burri, R. A. Hoebe, Y. Maruyama, C. Bruschini, and E. Charbon, *Photon-counting arrays for time-resolved imaging,* Sensors **16** (2016), 10.3390/s16071005.

[33] D. Stoppa, F. Borghetti, J. Richardson, R. Walker, L. Grant, R. K. Henderson, M. Gersbach, and E. Charbon, *A 32×32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain,* ESSCIRC 2009 - Proceedings of the 35th European Solid-State Circuits Conference , 204 (2009).

[34] C. Veerappan, J. Richardson, R. Walker, D.-u. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, *A 160×128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter,* ISSCC , 312 (2011).

[35] S. Lindner, C. Zhang, I. M. Antolovic, M. Wolf, and E. Charbon, *A 252 × 144 SPAD pixel FLASH LiDAR with 1728 Dual-clock 48 . 8 ps TDCs , Integrated Histogramming and 14 . 9-to-1 Compression in 180nm CMOS Technology,* Symposium on VLSI Circuits Digest of Technical Papers , 69 (2018).

[36] Z. Cheng, X. Zheng, M. J. Deen, and H. Peng, *Recent developments and design challenges of high-performance ring oscillator CMOS time-to-digital converters,* IEEE Transactions on Electron Devices **63**, 235 (2016).

[37] K. Park and J. Park, *Time-to-digital converter of very high pulse stretching ratio for digital storage oscilloscopes,* Review of Scientific Instruments **70**, 1568 (1999).

[38] P. Dudek, S. Szczepański, and J. V. Hatfield, *A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line,* IEEE Journal of Solid-State Circuits **35**, 240 (2000).

[39] 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 Transactions on Circuits and Systems I: Regular Papers **60**, 557 (2013).

[40] J. Yu, F. foster Dai, and r. c. Jaeger, *A 12-Bit Vernier Ring Time-to-Digital Converter in 0.13um CMOS technology,* IEEE Journal of Solid State Circuits **45**, 830 (2010).

[41] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, *A 100-m range 10-Frame/s 340×, 96-pixel time-of-flight depth sensor in 0.18-µm CMOS,* IEEE Journal of Solid-State Circuits **48**, 559 (2013).

[42] A. Ximenes, P. Padmanabhan, and E. Charbon, *Mutually Coupled Time-to-Digital Converters (TDCs) for Direct Time-of-Flight (dTOF) Image Sensors,* Sensors **18**, 3413 (2018).

[43] C. Favi and E. Charbon, *A 17ps time-to-digital converter implemented in 65nm FPGA technology,* Proceeding of the ACM/SIGDA international symposium on Field programmable gate arrays - FPGA '09 , 113 (2009).

[44] S. Burri, C. Bruschini, and E. Charbon, *LinoSPAD: A Compact Linear SPAD Camera System with 64 FPGA-Based TDC Modules for Versatile 50 ps Resolution Time-Resolved Imaging,* Instruments **1**, 6 (2017).

[45] R. Narasimman, A. Prabhakar, and N. Chandrachoodan, *Implementation of a 30 ps resolution time to digital converter in FPGA,* 2015 International Conference on Electronic Design, Computer Networks and Automated Verification, EDCAV 2015 , 12 (2015).

[46] M. Zhang, H. Wang, and Y. Liu, *A 7.4 ps FPGA-based TDC with a 1024-unit measurement matrix,* Sensors (Switzerland) **17** (2017), 10.3390/s17040865.

[47] I. M. Antolovic, C. Bruschini, and E. Charbon, *Dynamic range extension for photon counting arrays,* Optics Express **26**, 22234 (2018).

[48] sang hoon Lee and R. P. Gardner, *a new G-M ounter dead time model,* applied radiation and isotops **53**, 731 (2000).

[49] G. Bondarenko, B. Dolgoshein, V. Golovin, A. Ilyin, R. Klanner, and E. Popova, *Limited Geiger-mode silicon photodiode with very high gain,* Nuclear Physics B - Proceedings Supplements **61**, 347 (1998).

[50] F. Yang, L. Sbaiz, E. Charbon, S. Süsstrunk, and M. Vetterli, *Image reconstruction in the gigavision camera,* 2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009 , 2212 (2009).

[51] A. R. Ximenes, P. Padmanabhan, M.-j. Lee, Y. Yamashita, D. N. Yaung, and E. Charbon, *A 256×256 45/65nm 3D-Stacked SPAD-Based Direct TOF Image Sensor for LiDAR Applications with Optical Polar Modulation for up to 18.6dB Interference Suppression,* ISSCC , 27 (2018).

4

3D-STACKING AND ITS APPLICATION TO THE MCG SENSOR

In chapter 2, we have introduced the concept of 3D-stacking as a technique to maximize fill factor while combining two potentially different (CMOS) technology nodes for system performance optimization. In this chapter, we describe the use of a 3D-stacking technology provided by IMEC to achieve just that. The goal of this design was the creation of 1Gfps sensor based on a multi-collection gate (MCG) pixel design. According to an analysis by Prof. Etoh, MCG image sensors could achieve speeds up to 90.1 Gfps when exposed to light at 550nm of wavelength in excess of the mere 200Mfps currently available [1]. This achievement is enabled by the use of a charge-coupled device (CCD) on the top chip and a fast CMOS driver on the bottom chip of the stack. This chapter is based on the results presented at the International Image Sensor Workshop 2015, C. Zhang etc. 'Pixel parallel, localized driver design for a 128 × 256 pixel array 3D 1Gfps image sensor' [2].

4.1. INTRODUCTION

The speed of the *in-situ* memory sensors is mainly limited by two factors: the transit time of electrons and the rising time of the driving signals. In [3], the electrons are collected in a sequential way so that in each frame the electrons are generated at different positions on the back side, transmitted to the collection gate and then transferred to the in-situ memory. This process takes approximately 10 ns, which limits the maximum frame rate to 100 Mfps. On the other hand, the electron collection and transfer are driven by externally generated signals. With the scaling of the pixel array, the load resistance and capacitance will be increased, which limits the maximum operation frequency of the sensor due to the RC delay.

In order to improve the frame rate, while maintaining the high sensitivity, single-shot property and the scalability, a CCD image sensor is presented in this chapter, which was designed in a 130 nm technology targeting at a frame rate of 1 Gfps. The sensor integrated several concepts and techniques, including the multi-collection gate BSI pixels, 3D stacking and pixel parallel drivers. In order to reduce the transmit and transfer time of electrons, a pixel structure with multi-collection gates was proposed. Since the pixel was designed by Son V.T. Dao and is out of the scope of this thesis, a brief introduction will be presented in section 4.3. To reduce the RC constant, a 3D stacked architecture was proposed. The sensor was implemented on the top chip, while an array of drivers were implemented on the bottom chip. Each driver controls a subset of pixels with reduced parasitics. Pixel-wise connection between the top and bottom chips was achieved with 3D bump stacking at a fine pitch of 18 μm [2].

4.2. 3D-STACKING TECHNOLOGY

The sensor was implemented in a 3D two-tier stacking technology. In this technology, the top and bottom chips are bonded face-to-face utilizing micro-bumps (diameter of 11 μm), as shown in Fig. 4.1. The pitch of the micro-bumps is 18 μm in horizontal direction and 24 μm in vertical. Each 12 micro-bumps are shared by 32 pixels with 6 signals, comprising 5 collection gate signals and 1 drain gate signal (red arrows). To improve the connectivity of the stacking, each signal is connected with 2 micro-bumps. While for other signals that the top chip requires (black arrows), such as STGs, TRGs, etc, signals are firstly transmitted to the bottom chip via bonding wires, then propagate to the top chip through metal wires and micro-bumps. After the stacking, the top chip is thinned down to 30 μm, which enables the sensor to operate in BSI mode.

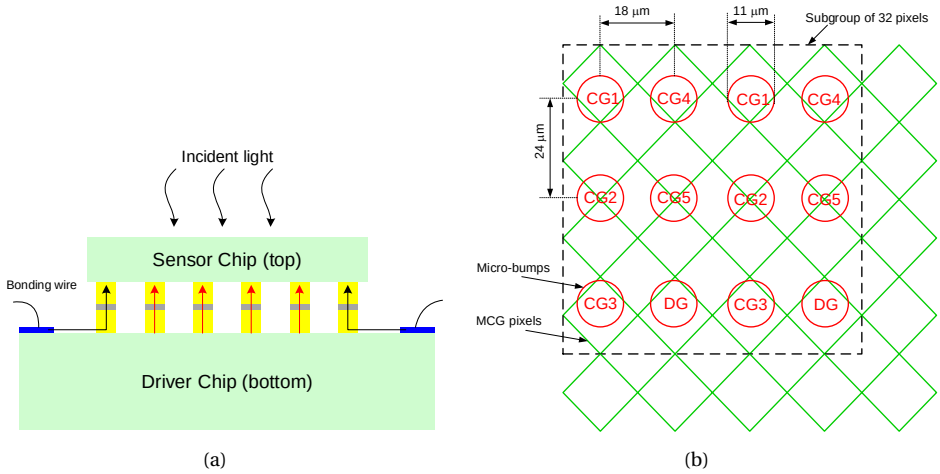


Fig. 4.1 (a) Cross-section of the 3D stacking technology based on micro-bumps; (b) stacking diagram of a subgroup of 32 pixels which are driven by one driver at the bottom chip [2].

4.3. SENSOR DESIGN

4.3.1. MULTI-COLLECTION-GATE PIXEL STRUCTURE

For BSI in-situ memory imagers, the travel time of the charge from the generation site to a collection gate, t_1, is around 1 ns, which is much shorter than that of an accumulated charge packet fully transferred from the collection gate to the attached storage area, t_2 [4]. To solve this problem, a pixel structure with MCGs was proposed with a pixel pitch of 18 μm. The layout of the pixel is shown in Fig. 4.2 (a), where a group of collection gates (A2-A6) and storage gates (B2-B6) are surrounding the center. A drain gate (A1) is used to drain out the electrons when the sensor is not in operation. The pixel is based on p-/n- double epi-layers with a total thickness of 30 μm, where the cross-section is shown in 4.2 (b). With a thick silicon layer, the absorption efficiency of the incident light at 700 nm wavelength is more than 99.9%, which reduces the direct inteference of light to the memory area, so as to generate noise.

While in operation, electrons can be collected at the collection gates by applying a high voltage to the collection gates (CGs). The timing diagram of the operation is shown in Fig. 4.2 (c), where a pulsed signal with a width of t_F is applied sequentially to each CG. After the operation of one collection gate, the charge packet will be transferred to the storage gate during the collection operation of other collection gates. If the number of CG is N and the transfer time t_2 is equal to or less than $(N\text{-}1)t_1$, the collection gate can be empty before the start of the next collection operation. In this case, the minimum frame

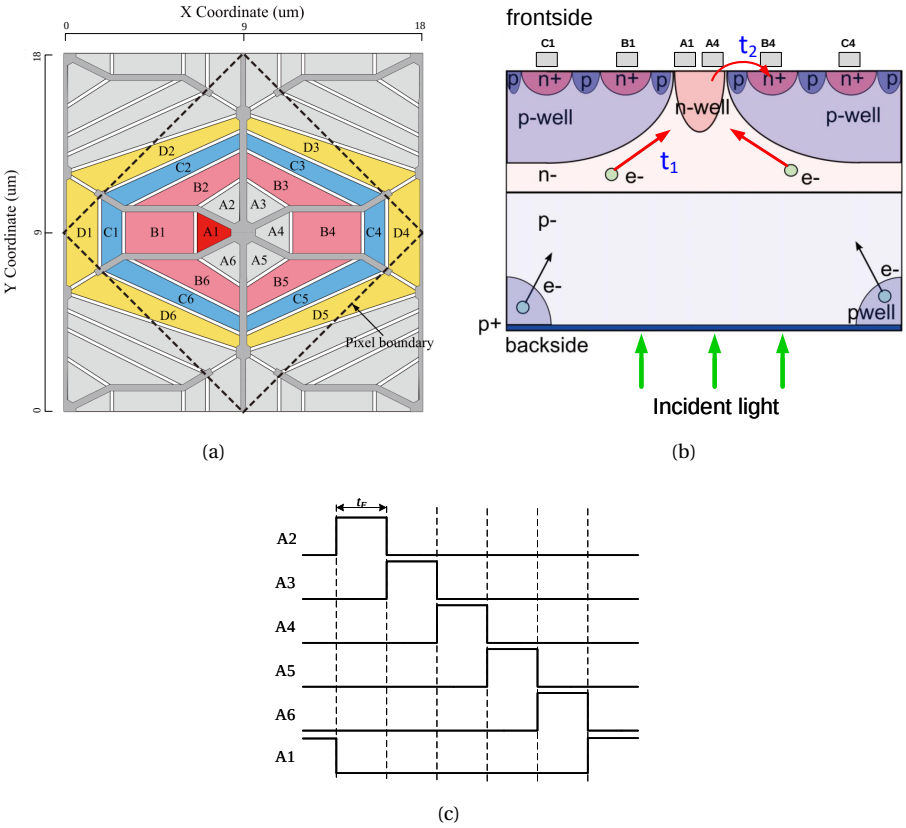
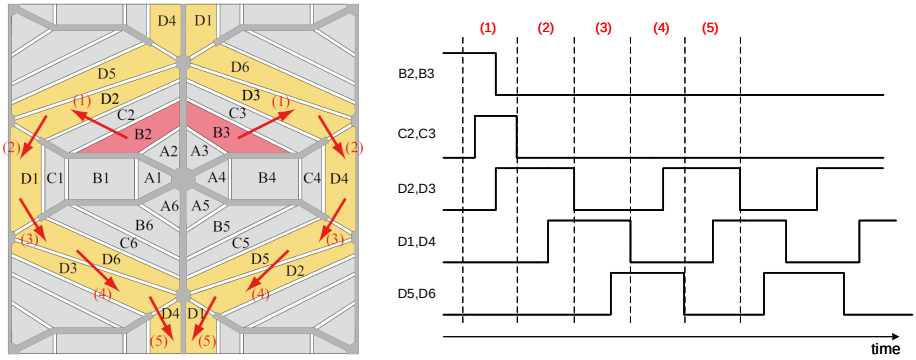


Fig. 4.2 (a) Hexagonal BSI MCG image sensor: A1: Drain gate; A2-A6: Collection gates; B1: Drain; B2-B6: Storage gates; C: Barrier gates; D: Transfer gates. (b) pixel cross section at C1-C4. (c) timing diagram of the pixel operation at electron collection.

interval, t_F, is equal to $t_1 = t_2/(N-1)$. For a conventional in-situ memory image sensor[3], there is only one collection gate, thus $t_F = t_2$. So with the MCG structure, the frame rate can be improved by N-1 times, which is a 4x improvement in this design. If t_F = 1ns, a frame rate of 1 Gfps can be achieved. On the other hand, more frames can be captured by connecting multiple storage gates to each collection gate. Since this is a conceptual design, only one storage gate is implementation for each CG, enabling 5 frames to be recorded at a high frame rate.

In comparison to the image capturing operation, the readout operates at a relatively slow speed. The charge transfer path and timing diagram of the storage gate 2&3 in the read-out is shown in Fig. 4.3. The charges are firstly transferred from B2&3 to the transfer gate

D2&3. Then a three-phase clocking approach is carried out on the transfer gates (D1-D6) to transfer the charges to the bottom of the chip. Since the charge transfer path of B2 and B3 is symmetric, B2 and B3 share one group of readout signals. By repeating this process three times, all the charges stored in the storage gates can be read out. For more information about the MCG pixel, please refer to [4–7].



(a) (b)

Fig. 4.3 (a) the charge transfer path of storage gate 2&3 in the readout. (b) timing diagram of the signals in the readout operation.

4.3.2. MCG PIXEL ARRAY AND READOUT

A sensor chip with an array of 128 × 128 × 2 pixels was designed. Since the shape of the MCG pixel is hexagonal, pixels are placed in an interleaved format, which explains why the pixel format is multiplied by 2. To read out the signals off-chip, each column of pixels is connected to a set of CCD gates to transfer the charges vertically and horizontally, referred to as VCCD and HCCD. The 1st and 2nd frame readout flow diagram is shown in Fig. 4.4, which comprises 5 steps:

(1) Charge is transferred from the storage gates to the transfer gate (TRG1).

(2) Charge transfer by one cycle from the prior TRG1 to the next TRG1 with a 3-phase clocking method.

(3) In parallel, VCCDs, working as an interface, transfer charges to the HCCDs, driven with a 4-phase clocking scheme.

(4) When HCCDs get the charge package from each VCCD column, the charge transfer flow in the TRGs and VCCDs will be stopped until all the charges in the HCCDs are read off-chip. The charges will be transferred from left to right to the floating diffusion (FD) node. Meanwhile, an output gate (OG) is used to enable/disable the readout pro-

cess.

(5) Correlated double sampling (CDS) is applied at the output, and the FD is reset prior to the signal measurement. A source follower based buffer is implemented to convert the charge signals to voltages, and an external analog-to-digital converter (ADC) is employed for the voltage digitalization. After the readout of all the charges in the HCCDs, the process restarts from (2) until all the charges are read out.

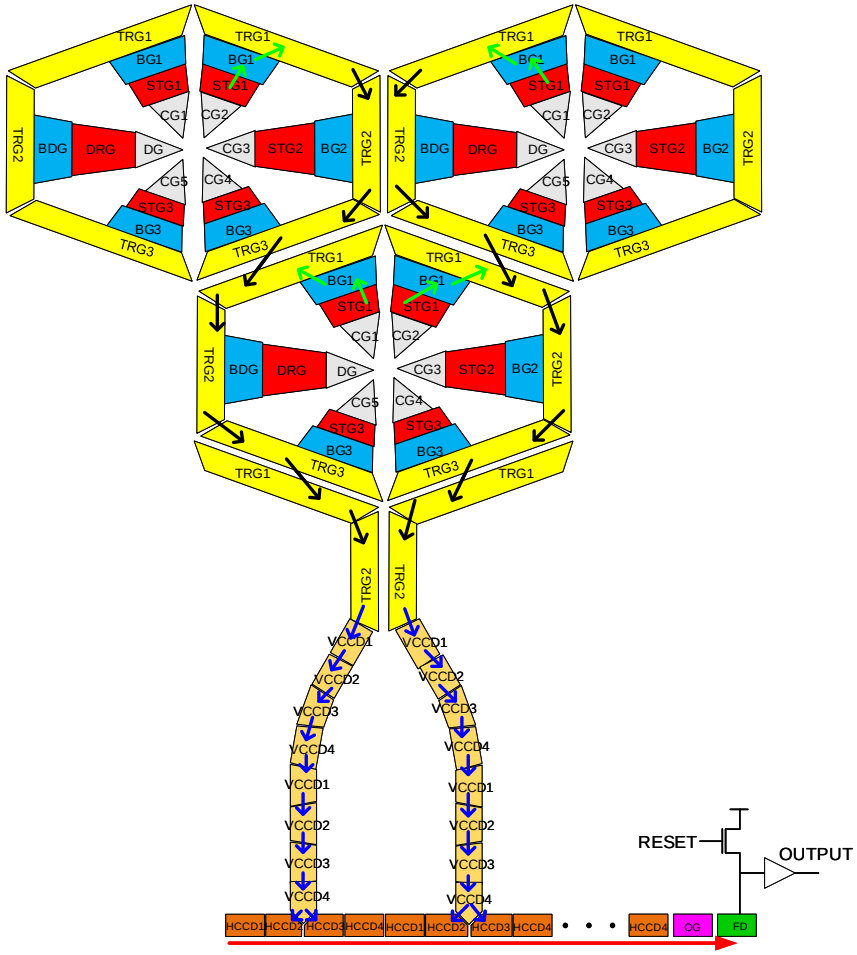


Fig. 4.4 The 1st and 2nd frame readout flow diagram.

For the readout of other frames, a similar process is applied. For the whole array, two parallel output channels are implemented, each handling one half of the array, or 64 × 128 × 2 pixels. Since the readout control signals, e.g. STGs, BGs, TRGs, VCCDs and

HCCDs, are shared by all the pixels and generated externally, a large load capacitance is expected, which limits the readout speed to several MHz.

4.4. DRIVER CHIP DESIGN

4.4.1. DRIVER ARCHITECTURE

As is discussed in section 4.3, to achieve 1 Gfps, a pulse train with 1 ns width is required for all the MCG pixels at the same time. It is impractical to input a such narrow pulse externally, due to the large load capacitance and propagation length. To reduce the load capacitance, an array of 32 × 32 drivers was designed and is located evenly at the bottom chip. Therefore, each driver will only drive a subgroup of 32 pixels, and the load capacitance for every pulse output is about 200 fF. The driver array architecture is shown in Fig. 4.5, where the array is divided into 4 quadrants for independent operation, and a phase-locked loop (PLL) is designed to bias the drivers at the desired pulse width.

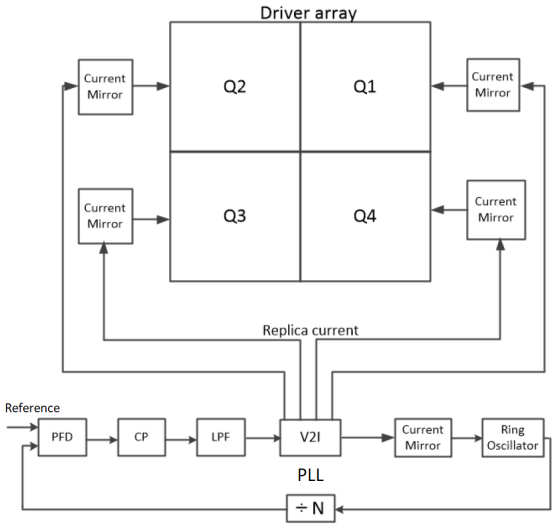


Fig. 4.5 The driver chip architecture, where the driver array is divided into 4 quadrants and biased with a PLL.

The PLL is based on a conventional integer-N (from 1 to 63) architecture, where a phase-frequency detector (PFD) detects the phase/frequency offset between the divided oscillator output and the reference clock. This error information then drives the charge pump, presented in section 4.4.4, to adjust the biasing voltage through a low pass filter (LPF). The voltage is then converted to current with a voltage-to-current (V2I) converter to drive the current-starved voltage controlled oscillator (VCO), presented in sec-

tion 4.4.3, which achieves the full tuning range for a low VCO gain, and a low jitter. This current is finally replicated and distributed to 5 current mirrors that bias the 4 quadrants of the driver array and the VCO of the PLL, which enables the frequency synchronization of the entire array and the PLL. In order to cover a wide range of pulse width, a reference clock of 10 MHz is utilized to ensure that the PLL can lock at low frequencies, such as 2.8 ns pulse width at 30 MHz. A low loop bandwidth of approximately 200 kHz was chosen to achieve PLL stability.

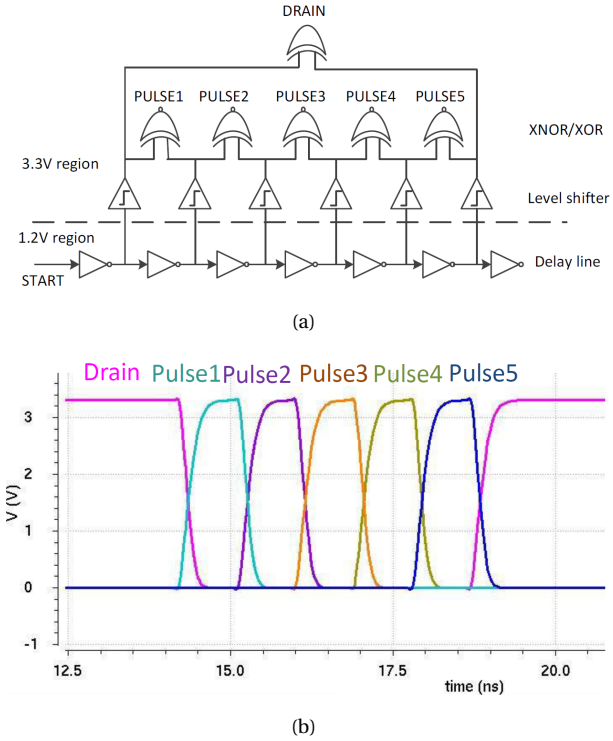


Fig. 4.6 (a) Simplified schematic and (b) simulation of the driver.

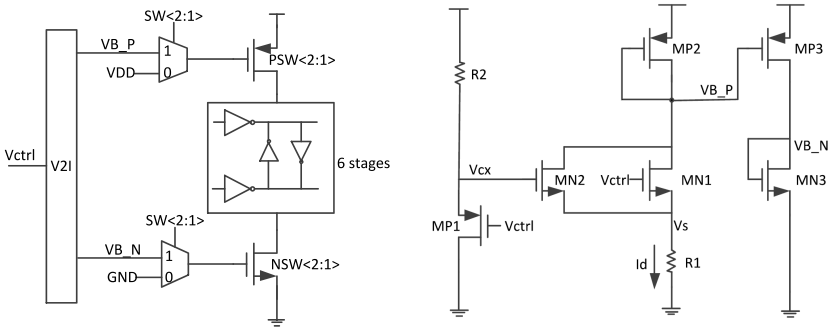
4.4.2. XNOR DRIVER

Since only one storage gate is connected to each collection gate, a delay line based XNOR driver is designed to generate equal delay stages, which are then level shifted from 1.2 V to 3.3 V with a set of level shifters. The XNOR gates generate continuous pulses and drain signal by combining adjacent delay taps. While each tap is a replica of the delay cell of the VCO in the PLL. In this way, the cell delay, thus the pulse width, can be controlled by adjusting the PLL frequency. The simplified schematic and the simulation are shown

in Fig. 4.6. The driver is enabled by applying a step signal to the *START*. In order to trigger all the pixels for the image capturing at the same time, a delay balanced clock tree is utilized for the *START* signal distribution, which achieves a maximum skew of 15 ps from the post-layout simulation.

4.4.3. RING OSCILLATOR

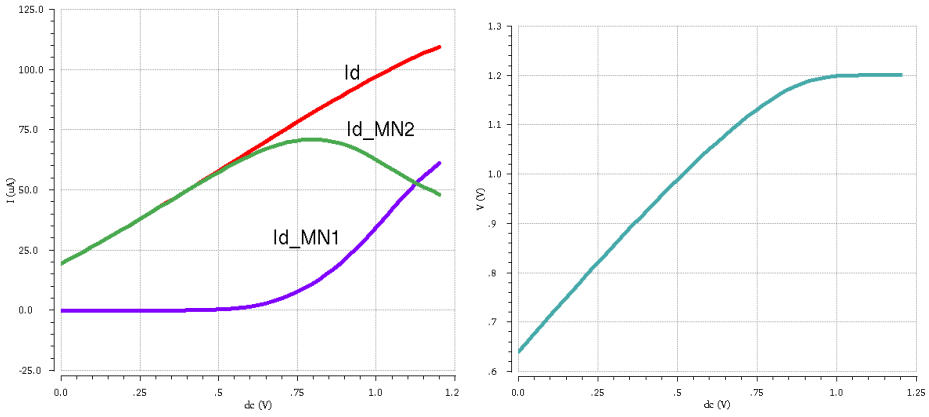
Since the pulse width is based on the delay of each delay cell, where the delay can be either rise or fall delay, to improve the linearity of pulse width at different oscillation frequencies, the rise and fall time of each delay cell has to be equalized, which requires an accurate 50% duty cycle. Besides, low power consumption is required to ensure the scalability up to several Mpixels. Considering these conditions, a 6-stage current-starved pseudo-differential RO was designed, as is shown in Fig. 4.7(a), where both the source and sink current are biased by *V_BP* and *V_BN* to guarantee the same rise and fall time, as well as the 50% duty cycle. If the expected pulse width is t_p, the VCO has to operate at a frequency of $12t_p$. To explore the limit of the pixel speed, the pulse width needs to be programmable in a wide range from 300 ps to 3 ns, corresponding to a frame rate from 3.3G to 330M fps and the VCO frequency from 270MHz to 27MHz. To cover a wide tuning range, while with a low VCO gain, the VCO frequency range is divided into 3 sub-ranges by switching on/off different pairs of pmos/nmos current sources, where the W/L ratio between PSW<2>(NSW<2>) and PSW<1>(NSW<1>) is 2:1. For a current starving VCO, the frequency is linearly proportional to the current flowing through the current sources. So the frequency range by switching on SW<2> will be twice wider than that of SW<1>. In this way, three frequency ranges are available with a ratio of 1:2:3 by setting SW<2:1> as '01', '10' and '11'.



(a) (b)

Fig. 4.7 The schematic of the (a) VCO and (b) voltage-to-current converter.

To reduce the VCO gain and the PLL jitter, a full-range tuanable and high linearity voltage-to-current (V2I) converter is proposed in Fig. 4.7(b), where MN1 and R1 are connected in a common-drain manner [8]. When *Vctrl* is greater than *Vth*, *Vs* increases linearly, which generates a linear current *Id* through R1. To achieve full tuning range, a parallel MN2 is added to the current path, which is biased with another common-drain structure comprising MP1 and R2. In such a way, with the increase of *Vctrl* from 0 to the supply voltage of 1.2 V, when *Vctrl* is less than the threshold of MN1, *Id* is only generated by MN2. While with the further increase of *Vctrl*, both MN1 and MN2 are closed and *Vcx* gradually gets saturated at 1.2 V. Therefore, the current flowing through MN2, *Id_MN2*, is decreasing due to the increased *Vs*, whilst the current of MN1, *Id_MN1*, keeps rising due to the increased overdrive voltage of MN1. By combining the two currents, the V2I converter with a full tunable voltage range can be realized. The I-V simulation of the current *Id*, *Id_MN1* and *Id_MN2* is shown in Fig. 4.8.



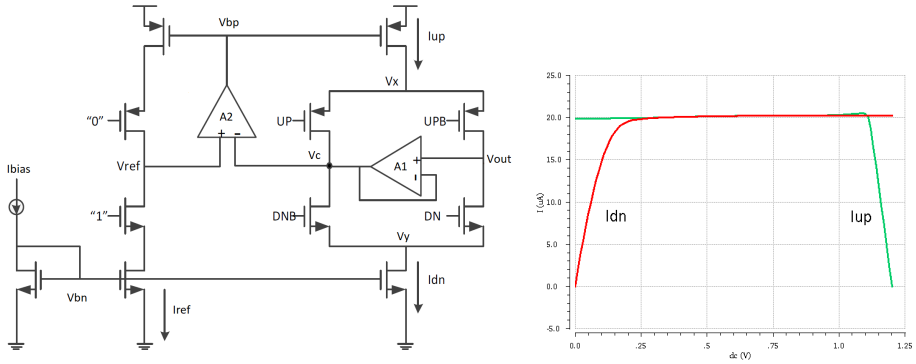
(a) (b)

Fig. 4.8 Simulation results of the V2I converter with (a) the current of MN1, MN2 and R1, where a full range and high linearity *Id* is achieved; (b) voltage of Vcx saturates at the supply voltage of 1.2 V.

4.4.4. CHARGE PUMP

To drive a full range voltage-to-current converter, a charge pump with wide output range was designed. The schematic is shown in Fig. 4.9 (a), where a bias current, *Ibias*, is used to generate a bias voltage, *Vbias*. With this voltage, a 20 μA charge pump reference current, *Iref*, is generated. In order to achieve high current matching between *Iup* and *Idn*, two error amplifiers, A1 and A2, are employed. At the right side of the design, a conventional charge pump is implemented based on a unity gain amplifier A1, which solves the charge sharing problem at *Vx* and *Vy* [9]. Meanwhile, with the implementation of A2, the voltages are regulated as *Vout* = *Vc* = *Vref*. This ensures that the source/sink current

(*Iup* and *Idn*) is equal to the reference current *Iref*, and one can achieve nearly perfect source/sink current matching regardless of the charge pump output voltage [10]. The simulation result of the charge pump current is shown in Fig. 4.9 (b). Assume the tolerable current mismatch is +/-10%, the dynamic range of this charge pump is from 0.145 V to 1.12 V, which covers 81% of the 1.2 V supply voltage. While large current mismatch occurs at the lower and higher voltages, which is mainly due to the gain degradation of the amplifiers and the source/sink transistor overdrive voltage limitations.



(a) (b)

Fig. 4.9 (a) The schematic of the charge pump; (b) simulation result of the charge pump current Idn and Iup, where +/-10% current mismatch is achieved in 81% of the voltage range.

4.4.5. TIME-TO-DIGITAL CONVERTER

Sub-nanosecond pulses cannot be observed directly with an oscilloscope. In order to characterize the pulse width uniformity, a vernier-delay-line TDC is designed, which has 120 stages of voltage controlled delay cells based on current-starving buffers [11, 12]. The schematic of the TDC is shown in Fig. 4.10, where the START is triggered by the rising edge of the pulse, the STOP by the falling edge and the cell delay is controlled by adjusting the bias voltage *Vbias*. Since the time resolution is defined by the delay difference of $\tau_s - \tau_f$, sub-gate-delay resolution can be achieved.

4.4.6. CHIP REALIZATION

Both the top and the bottom chip were fabricated in 130 nm CMOS technology, while extra layers with customized profiles were implemented in the top chip for the MCG pixels. An array of 128 × 128 × 2 pixels was implemented in the top chip, which is driven by a bank of 32 × 32 drivers in the bottom chip via a totally number of 12288 micro-bumps. The microphotograph of the stacked sensor is shown in Fig. 4.11, where the

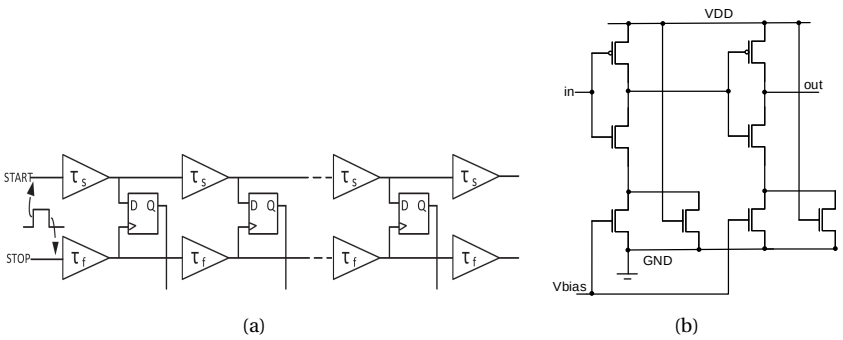


Fig. 4.10 The schematic of (a) the vernier-delay-line TDC and (b) the voltage controlled delay cell.

bottom and top chip dimensions are 5 mm × 5 mm and 3 mm × 3 mm, respectively. Four bonding wires are connected to the backside of the top chip, and they are used for the pixel substrate biasing. Since the driver array is connected to the pixels directly, a replica driver is designed and placed outside the array for driver characterization.

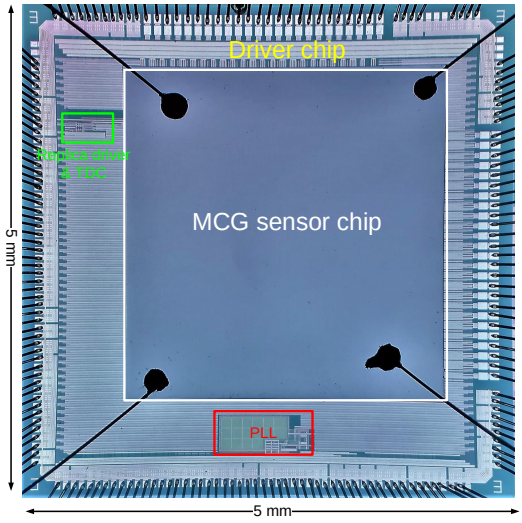


Fig. 4.11 Microphotograph of the stacked sensor.

4.5. RESULTS

4.5.1. PLL CHARACTERIZATION

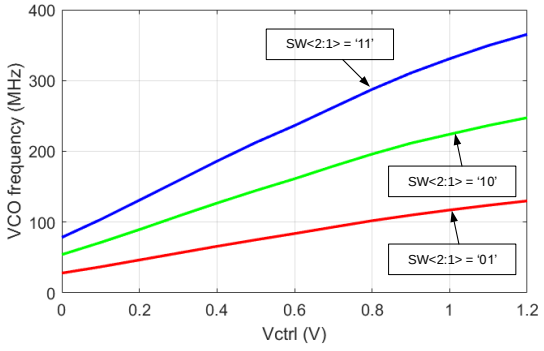
The VCO frequency characterization with 3 tuning ranges is shown in Fig. 4.12. Due to the linear property of the V2I converter, the VCO shows high linearity in the full tuning range and a relatively small gain. The VCO tuning frequency range is from 28MHz to 365MHz, where :

sub-range 1 is 28MHz - 123MHz, with SW<2:1> = '01';

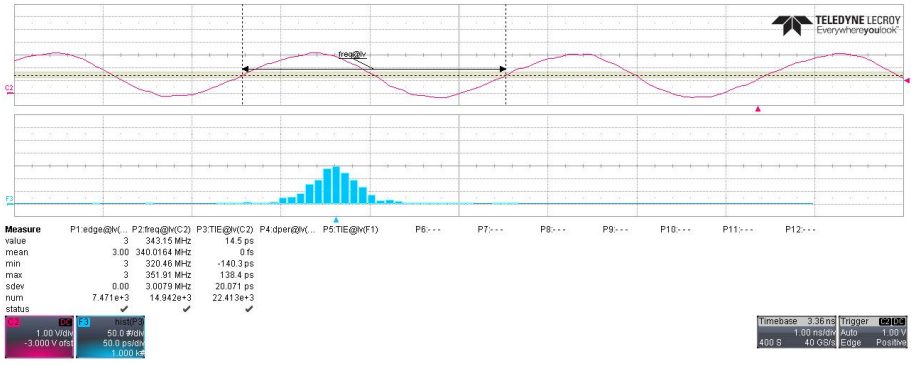
sub-range 2 is 54MHz - 247MHz, with SW<2:1> = '10';

sub-range 3 is 78MHz - 365MHz, with SW<2:1> = '11';

The PLL time interval error (TIE) jitter was measured at 20 ps rms at 320MHz with a power consumption of 4.3 mW.



(a)



(b)

Fig. 4.12 (a) VCO frequency measurement in 3 sub-ranges with the bias voltage *Vctrl* increasing from 0V to 1.2 V, where full tuning range and high frequency linearity is achieved. (b) PLL jitter performance, where a 20 ps rms jitter was achieved at 320 MHz.

4.5.2. TDC CHARACTERIZATION

To characterize the linearity of the TDCs, a code density test method was utilized in the measurement. Unlike the planar chips Piccolo and Ocelot, which will be described in chapter 5 and 6, where uncorrelated signals were generated with SPADs, a FPGA based approach is proposed for Nanosis. In this approach, the two inputs of the TDC, START and STOP, are triggered by a 50 MHz clock, START_CLOCK, and one oscillator output, STOP_CLK, which is constructed by connecting an odd number of inverters in a loop configuration in the FPGA (Opal Kelly, BRK7360). The oscillator frequency is unlocked and working at approximately 47.437531MHz with 353 loop-connected inverters. Since the phase offset of the two clocks is an infinite decimal of about 1.080355 ns, with the accumulation of clock cycles, the delay between the clock rising edges can be any value within one period of START_CLOCK (20ns). Therefore, an evenly distributed time interval can be created and applied for the TDC nonlinearity characterization. The simulation result of the time interval distribution with accumulation of 10M clock cycles is shown in Fig. 4.13. With this method, a -0.18/+0.26 LSB DNL and -1.2/+0.28 LSB INL were achieved, as is shown in Fig 4.14. In order to measure the TDC time resolution (LSB), the IODelay property of the FPGA IObuffers were utilized, where one tape delay is set to 78.125 ps. By adding known tap delay between the START and STOP signals, the TD time resolution can be characterized, where a minimum LSB of 28.58 ps was obtained.

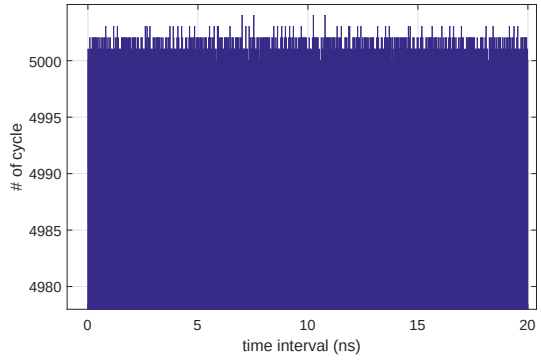
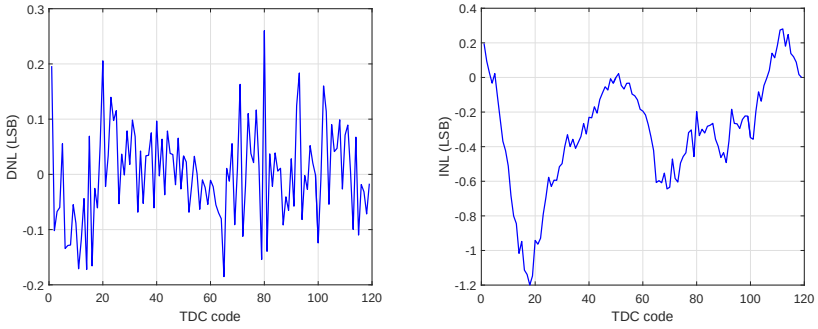
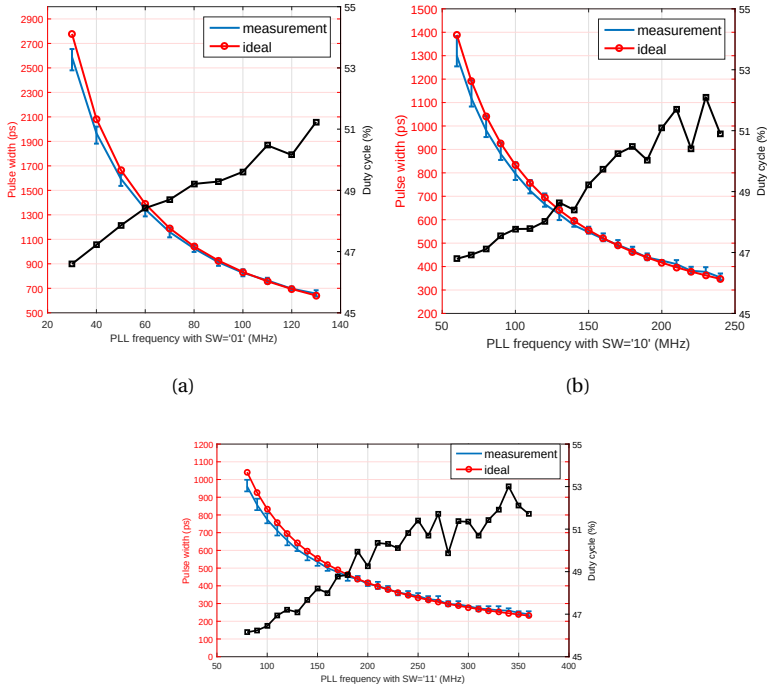


Fig. 4.13 Simulation result of the time interval distribution with two clocks working at 50MHz and 47.437531MHz.



(a) (b)

Fig. 4.14 TDC nonlinearity characterization based on code density test method with two clocks, which gives (a) -0.18/+0.26 LSB DNL and (b) -1.2/+0.28 LSB INL.



(a) (b)

(c)

Fig. 4.15 Pulse width measurement, where the PLL was working at (a) sub-range 1 with SW='01', (b) sub-range 2 with Sw='10' and (c) sub-range 3 with SW='11'. Pulse width offset can be observed due to the mismatch of the rise and fall cell delay.

4.5.3. PULSE WIDTH MEASUREMENT

Pulses were measured with the TDC at different PLL frequency, where the ideal pulse width would be (clock period)/12. Since the driver needs to be triggered by a step signal, which can be either rising or a falling, thus the pulse width is determined by the rise or fall delay of each delay cell. A 50% duty cycle is expected to create equal rise and fall delay that the pulse width can be linearly controlled by the PLL. The pulse width measurement result with a rising start signal is shown in Fig. 4.15, where the PLL was operated in all the 3 sub-ranges at a frequency step of 10 MHz. From the measurement, we can see a pulse width offset from the expected value, in which the pulses are shorter than the ideal value at lower frequencies and are longer at higher frequencies. This offset is mainly due to the mismatch of the rise and fall delay, where the duty cycle increases from 46% to 53%. Even though the delay cells are biased with a current mirror, a source-to-sink current mismatch could still be present due to the limited impedance of the biasing transistors (PSW and NSW in Fig. 4.7), which results in the delay mismatch and incorrect duty cycle.

The nonlinearity between the 5 pulses at different pulse width is shown in Fig. 4.16, where an approximately 4% nonlinearity is achieved at pulse width longer than 500 ps. Nevertheless, for shorter pulses, the nonlinearity presents an increasing trend, due to the limited TDC resolution. Since the minimum TDC LSB is 28.58 ps, the TDC quantization error would have a significant impact on the measurement of short pulse nonlinearity. Therefore, a TDC with finer resolution should be used for this characterization. Figure 4.17 shows the measurement result of pulse trains with an expected width of 1.042 ns. Due to the limited number of available probes, the drain and MCG driving signals were measured in two groups, with each group of 4 signals. The 5 pulses were measured with an oscilloscope to be 1.149, 1.021, 1.024, 1.069, and 1.024 ns, respectively. In comparison with the pulse width measured with the TDC of 1.026, 0.997, 1.027, 1.054 and 1.021 ns, a slight difference is observed, due to the pulse shrinking/enlarging through the IO buffers.

4.5.4. STACKING TECHNOLOGY EVALUATION

During the measurement of the stacked sensor, unexpected behavior was observed and no signals were generated from the CCD output. The reason for this behavior was identified in the imperfect micro-bumps and poor connectivity between the two chips can be observed in Fig. 4.18. Cracks were found between the bump-to-bump and bump-to-chip connections. Due to these cracks, signals generated by the bottom chip or external inputs cannot be transmitted to the top chip. The sensor cannot be characterized before resolving the connectivity problem, and the reason of the crack generation is still

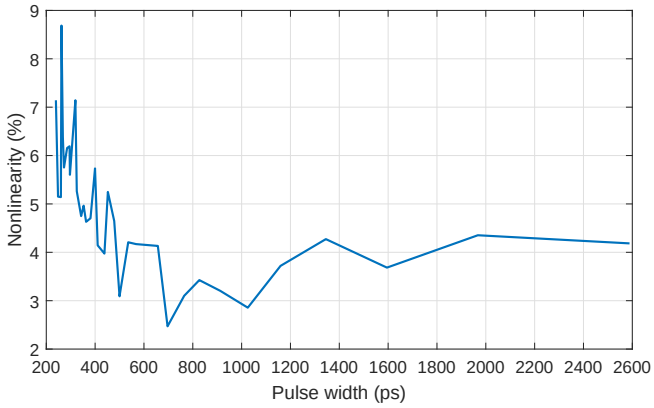


Fig. 4.16 Nonlinearity result of the 5 pulses with the pulse width in a range from 200 ps to 2.6 ns. The large nonlinearity at short pulses is due to the limited TDC time resolution.

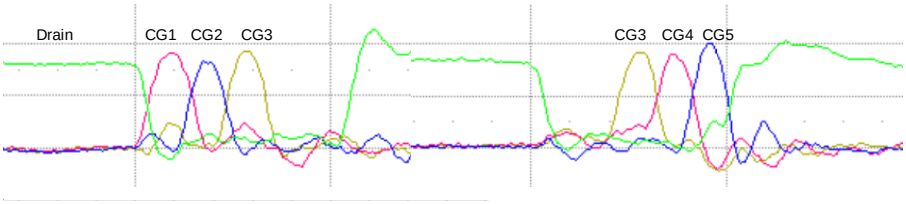
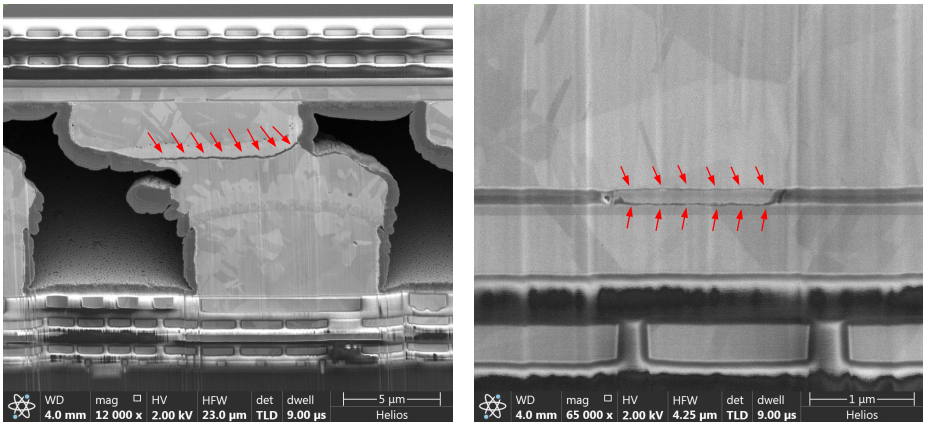


Fig. 4.17 Pulse output measured via an oscilloscope with the width of 1.149 ns, 1.021 ns, 1.024 ns, 1.069 ns, and 1.024 ns respectively.

under investigation. However, the MCG pixels were also implemented in another sensor, in which each collection gate is equipped with 305 folded in-pixel memory units. The sensor is backside illuminated and all the signals are driven externally. A maximum frame rate of 25 Mfps was achieved with 32 × 32 pixels and 1220 in-pixel memory. For the details of pixel characterization and high speed imaging results, please refer to [13].

4.6. DLL BASED DRIVER ARCHITECTURE

As discussed in section 4.5.3, the pulse width offset is generated due to the inaccurate duty cycle. It is difficult to generate a perfect 50% duty cycle clock with a PLL in full-range frequencies. To solve this problem, instead of biasing the delay cells with a PLL, a delay-locked-loop (DLL) could be used. If a number of N buffers are implemented in the delay line and a reference clock at frequency of f_{REF} is applied, the delay of each buffer will be locked at $1/Nf_{REF}$. Since only the rise delay of the buffers is considered in the DLL operation, no specific duty cycle is required. In this case, it can be much easier



(a) (b)

Fig. 4.18 SEM cross-section image of the stacked bumps, where cracks were generated at (a) bump-to-bump and (b) bump-to-chip connections.

to precisely generate matched delay taps, so as the pulses. However, if more storage gates are attached to the collection gate for per-pixel multiple frame capturing, more continuous pulses have to be generated. Compared to an inverter based ring oscillator, a ring architecture cannot be built with a buffer chain, which means a more complicated triggering mechanism should be designed for the proposed method.

4.7. CONCLUSION

An ultra-high speed image sensor based on MCG CCD technology was proposed in this chapter targeting a frame rate of 1 Gfps. The sensor comprises a top sensor and a bottom driver chip, where the two chips were stacked together via micro-bumps with a horizontal pitch of 18 μm and vertical pitch of 24μm . The top chip comprises an array of 128 × 128 × 2 interleaved BSI MCG pixels, where the collection gates and the drain gates are driven by the driver chip and other signals are controlled externally. The driver chip comprises a 32 × 32 driver matrix, which is biased with an on-chip PLL. Fully tuning range and high frequency linearity were achieved with the PLL, with a configurable pulse and a range from 234 ps to 2.78 ns. Measurement results showed that the pulse variation is about 4% with long pulses, while larger variation is observed with short pulses due to the limited TDC time resolution. The current vernier-delay-line TDC is based on current-starving buffers, achieving a LSB of 28.58 ps, with a DNL and INL of -0.18/+0.26 LSB and -1.2/+0.28 LSB, respectively. Higher TDC resolution could be achieved with the optimization in the delay cell design. Even though the stacked sensor cannot be char-

acterized with imaging experiments due to the failure in the chip-to-chip stacking, we believe the sensor architecture proposed in this chapter is highly suitable for ultra-high speed image sensors. Extensive exploration is carried out on the stacking technologies, and new results can be expected in the near future.

REFERENCES

[1] F. Mochizuki, K. Kagawa, S. I. Okihara, M. W. Seo, B. Zhang, T. Takasawa, K. Yasutomi, and S. Kawahito, *Single-shot 200Mfps 5×3-aperture compressive CMOS imager,* Digest of Technical Papers - IEEE International Solid-State Circuits Conference **58**, 116 (2015).

[2] C. Zhang, V. T. S. Dao, T. G. Etoh, K. Shimonomura, and E. Charbon, *Designing pixel parallel, localized drivers of a 3D 1Gfps image sensor family,* ICHSIP-31 , 6 (2015).

[3] T. Arai, J. Yonai, T. Hayashida, H. Ohtake, H. Van Kuijk, and T. G. Etoh, *A 252- V/, 16.7-Million-frames-per-second 312-kpixel back-side-illuminated ultrahigh-speed charge-coupled device,* IEEE Transactions on Electron Devices **60**, 3450 (2013).

[4] V. T. S. Dao, K. Shimonomura, Y. Kamakura, and T. G. Etoh, *Simulation Analysis of a Backside-illuminated Multi-collection-gate Image Sensor,* ITE Transactions on Media Technology and Applications **2**, 114 (2014).

[5] T. G. Etoh, V. T. S. Dao, K. Shimonomura, E. Charbon, C. Zhang, Y. Kamakura, and T. Matsuoka, *Toward 1Gfps: Evolution of ultra-high-speed image sensors -ISIS, BSI, multi-collection gates, and 3D-stacking-,* Technical Digest - International Electron Devices Meeting, IEDM **2015-Febru**, 10.3.1 (2015).

[6] H. D. Nguyen, V. T. S. Dao, and T. G. Etoh, *Design of folded CCDs for ultra high speed imaging,* 2014 IEEE 5th International Conference on Communications and Electronics, IEEE ICCE 2014 , 327 (2014).

[7] A. Q. Nguyen and K. Shimonomura, *Crosstalk Analysis of an Image Sensor Operating At 1 Gfps,* IEEE international conference on communications and electronics , 176 (2016).

[8] R. J. Baker, *CMOS circuit design, layout, and simulation 3rd edition,* IEEE international conference on communications and electronics , 564 (2010).

[9] W. Rhee, *Design of high-performance CMOS charge pumps in phase-locked loops,* ISCAS'99. Proceedings of the 1999 IEEE International Symposium on Circuits and Systems VLSI (Cat. No.99CH36349) **2**, 545.

[10] J. Y. Lee and S. N. Hwang, *Charge pump with perfect current matching characteristics in phase-locked loops,* Transactions of the Korean Institute of Electrical Engineers **57**, 982 (2000), arXiv:0504102 [arXiv:Rheephysics] .

[11] P. Dudek, S. Szczepański, and J. V. Hatfield, *A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line,* IEEE Journal of Solid-State Circuits **35**, 240 (2000).

[12] 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 Transactions on Circuits and Systems I: Regular Papers **60**, 557 (2013).

[13] V. Dao, N. Ngo, A. Nguyen, K. Morimoto, K. Shimonomura, P. Goetschalckx, L. Haspeslagh, P. De Moor, K. Takehara, and T. Etoh, *An Image Signal Accumulation Multi-Collection-Gate Image Sensor Operating at 25 Mfps with 32 × 32 Pixels and 1220 In-Pixel Frame Memory,* Sensors **18**, 3112 (2018).

5

A 32 × 32 TIME-RESOLVED SPAD SENSOR

A 32 × 32 SPAD sensor, referred to as Piccolo, was implemented in a 180 nm CMOS technology. Piccolo was a collaborative design with a division of labor among the different circuit blocks. The author was responsible for the design of the address latching chain, Section 5.2.3, data readout and the firmware for the measurement. Scott Lindner designed the TDC, Section 5.2.4, whilst Ivan Michel Antolovic designed the pixel array, Section 5.2.1.

This chapter is based on results presented in C.Zhang et al. "A CMOS SPAD Imager with Collision Detection and 128 Dynamic Reallocating TDCs for Single-Photon Counting and 3D Time-of-Flight Imaging", in MDPI Sensors, 18(11), 2018 and S.Lindner et al. "Column-Parallel Dynamic TDC Reallocation in SPAD Sensor Module Fabricated in 180nm CMOS for Near Infrared Optical Tomography", in International Image Sensor Workshop, 2017.

5.1. INTRODUCTION

SPAD sensors with time-to-digital converters (TDCs) have been reported in [1–6]. Some of them present a per-pixel TDC architecture, in which each SPAD is connected to a TDC in the pixel layout. For instance, [1] reports a large SPAD array of 160 × 128, while a fill factor of 1% was achieved at a large pixel pitch of 50 μm due to the significant area occupation of the in-pixel TDCs. On the other hand, TDC sharing architecture have been presented in [4–6], where one TDC is shared among a subset of pixels. With this architecture, TDCs can be placed outside the pixel array region, in which high fill factor of 57% was achieved in [6] with a pixel size of 30 μm × 50 μm. However, when one SPAD detects a photon, it will occupy the TDC until the conversion is finished, which prevents other SPADs in the same subset from being detected. In order to improve the data throughput, a TDC dynamic reallocation scheme 5.2.3 was implemented, where 4 TDCs in one column were shared by 32 pixels and each pixel can be detected by any one of the TDCs at a time.

With the proposed sharing approach, event overflow could happen with high illumination due to the limited number of TDCs. However, in many sensors, instead of the TDC number, the bottleneck is the readout IO bandwidth. Take the per-pixel TDC sensor [1] as an example, since the TDC dynamic range is 55 ns, implying a pixel activity of 18 Mcps could be reached, while the maximum achieving count rate per pixel is limited to 256 kcps due to the limited IO bandwidth of 51.2 Gbps which is already one of the fastest interfaces in literature [7]. In Piccolo, the TDC number of each column was analyzed and determined according to the IO bandwidth, which is able to detect the same amount of events as per-pixel TDC architecture. Pixels in one column are shared via address and timing buses, when more than one pixel fires at the same time, a collision event will be generated with an invalid address. To distinguish the collision from the address, a so called winner-take-all (WTA) circuit was implemented, which achieves collision event detection details in 5.2.2.

5.2. SENSOR ARCHITECTURE

The block diagram of the sensor is presented in Fig. 5.1. At the top, a shared bus architecture is employed, where 32 pixels in each column share a single address bus and timing line, enabling a fill factor of 28% with a 28.5 μm pixel pitch. Pixels are coded in an anti-collision approach with a WTA circuit, where collision events lead to an invalid address output, thus allowing collisions to be identified. In the sharing architecture, each event occupies the bus for a set period, the bus dead time. To reduce this duration, a

monostable circuit is employed in the pixel. Furthermore, the shared architecture also implies that noisy or 'hot' pixels could occupy the bus for long periods due to the dark count, thus reducing the sensitivity of the column to real photon arrivals. Therefore, a set of row and column masking shift registers were implemented to shut down these noisy pixels according to the DCR level.

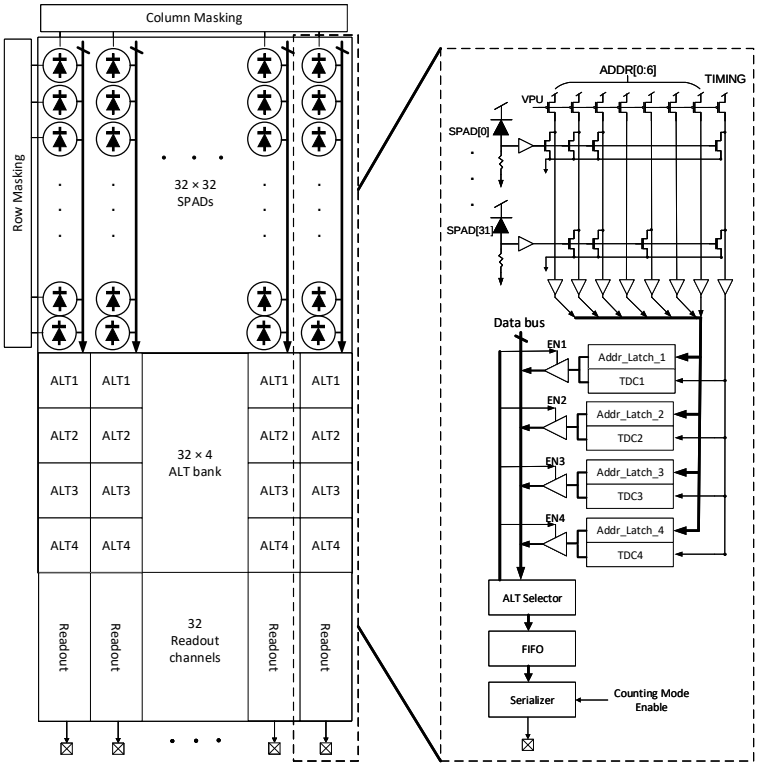


Fig. 5.1 Sensor architecture

At the bottom of the pixel array, a bank of 128 address latch and TDC (ALTDC) blocks are implemented to capture the pixel address and to measure photon arrival time. The 4 ALTDCs in one column are themselves connected in a daisy chain fashion, where a single ALTDC in the chain is available to capture events at any one time and events are distributed sequentially to improve the detection throughput. The TDCs employ an architecture based on a ring oscillator (RO), the frequency of which is set via an external voltage. The TDC has a 12-bit range with a temporal resolution of 50 ps, where the RO oscillates at 2.5 GHz.

At the output stage, each column has a dedicated readout block which serializes the data and streams it off-chip via a 160 MHz GPIO pad. The readout block works in an event-driven readout approach that only the ALTDCs which have detected photons will be read out through a tri-state bus. The data is firstly pushed into a first-in-first-out (FIFO) block and then a serializer reads the events out in UART format. With 32 GPIOs, a total output data bandwidth of 5.12 Gbps is achieved. In comparison with in-pixel TDC architectures, no null data is processed, which improves the efficiency of data transmission. The readout can operate in either photon timestamping (PT) or photon counting (PC) modes. In PT mode, both the TOF information and pixel address is read out from the sensor. A transmitted event comprises 23 bits including 1-bit start flag, 2-bit TDC identification number, 12-bit TDC code, 7-bit address code, and 1-bit stop flag. While in PC mode, the sensor only transmits 1-bit start flag , 7-bit address code, and 1-bit stop flag, so that the data length is reduced to 11 bits. As such, a maximum photon throughput of 222 Mcps and 465 Mcps can be achieved in PT and PC mode, respectively.

5.2.1. PIXEL SCHEMATIC

The sensor employs a SPAD with a p-i-n structure reported in [8]. A schematic of the pixel is shown in Fig. 5.2. 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 VOP = VBD + VEX, where VBD is the breakdown voltage and VEX the excess bias of the SPAD. As is reported in [8], the PDP and timing performance can be improved greatly by increasing the excess bias voltage, with only a small degradation in DCR. In order to achieve high excess bias voltage, a capacitance coupling method with area intensive resistors [9] and quenching circuit fabricated with high voltage process [10] have been proposed. In this design, high VEX was achieved with a cascoded passive quenching circuit [11], 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. The VEX limitation is mainly due to the large current of the avalanche that rises the voltage VA in a short time, whereas the source of M1 increases slowly due to the resistance of M1, thus generating a large voltage difference at VDS of M1. To prevent any device from breaking down, VEX of 5.2 V was obtained from the schematic simulation in Fig. 5.3, without exceeding the 3.6 V maximum voltage tolerance of any device. Due to the compact circuit implementation and the absence of in-pixel TDCs, a fill factor of 28% with a pixel pitch of 28.5 μm was achieved.

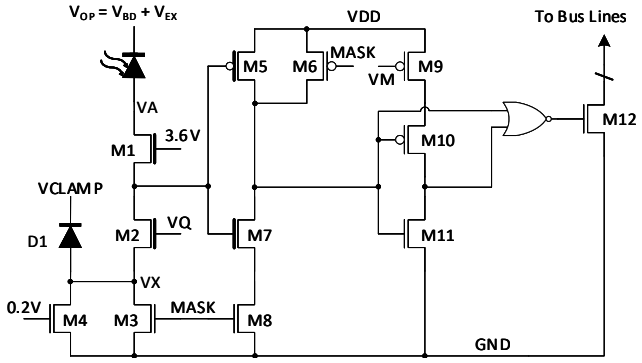


Fig. 5.2 Pixel schematic based on cascoded quenching scheme

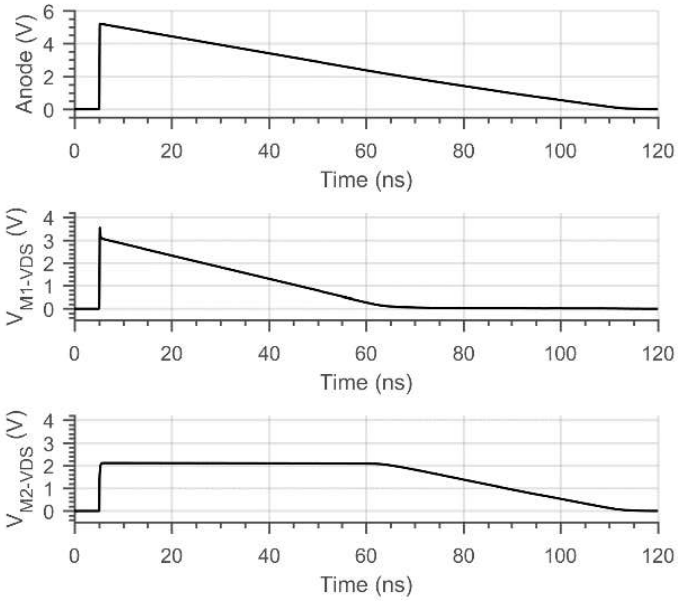


Fig. 5.3 Simulation of cascode quenching circuit. The VEX is limited to 5.2V with the drain-source voltage VDS of M1 reaching the tolerance of 3.6V.

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, whilst the configuration is managed by row and column shift registers on the array periphery. If voltage MASK is set as high, M3 operates in cut off region and the impedance is typically at the level of giga ohm, thus preventing the SPAD from recharging. However, if the leakage current from SPAD accumulates over time at the anode, the voltage of VA may increase over the tolerant limit, which could cause breakdown in M1. To ensure the safety of the pixel,

a parallel transistor M4, with its gate biased at 0.2 V, is used to provide a lower impedance path to drain out the leakage current and to prevent VA from increasing. Furthermore, a diode D1 clamps VX at a safe voltage VCLAMP, normally at 1.8 V, to protect M3 and M4 from high voltage in any case.

Since all pixels in a half-column are connected to a shared bus, each firing pixel will occupy the bus for a set period, referring to the 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 configurable monostable circuit comprising M9, M10, M11 and a NOR gate was implemented to reduce the output pulse duration time, thus the bus dead time. Post-layout simulations indicate that pulse widths in the region 0.4-5.5 ns can be achieved through proper adjustment of VM. This allows photons from multiple pixels to be detected during the same cycle.

5.2.2. COLLISION DETECTION BUS

The shared bus is similar to that implemented in [5], 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.

In the event that two pixels detect an event at the same time, the code present on the address lines will be a collision of the two pixel addresses, with every line pulled down by an active NMOS. This is an issue for binary coding, as these collisions will result in an incorrect address and thus, an invalid detection. For example, if pixel 1 and 2, coded as '110' and '101', respectively, fire simultaneously, the merged output code would be '100', thus providing invalid detection. For this reason, a collision coding scheme is implemented, where each address has 7 bits, consists of 3 '1's and 4 '0's. When collisions occur in this coding scheme, the detected address will have more than 3 ones. Therefore, collisions can be detected and distinguished at the data processing step. The diagram of the collision detection bus is shown in Fig. 5.1, which employs a WTA circuit. The address of each pixel is presented in Table 5.1. The collision detection bus shows good scalability that the maximum number of possible codes m for a the bus with n lines is given by (5.1), where k is the integer closest to $n/2$. An example of the code number with bus line increased from 1-bit to 12-bit is shown in Fig. 5.4.

$$m = \frac{n!}{k! * (n-k)!} \tag{5.1}$$

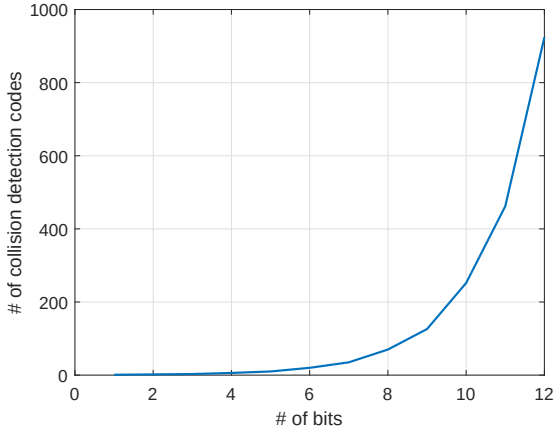


Fig. 5.4 Collision detection code number v.s. code bits

Table 5.1: Collision detection code table for 32 pixels

SPAD	ADDR	SPAD	ADDR	SPAD	ADDR	SPAD	ADDR
0	1110000	8	1000011	16	0011001	24	0001110
1	1100100	9	1000101	17	0011010	25	0101100
2	1100001	10	1000110	18	0010011	26	0101001
3	1100010	11	1010100	19	0010101	27	0101010
4	1101000	12	1010001	20	0010110	28	0100011
5	1001100	12	1010010	21	0000111	29	0100101
6	1001001	14	1011000	22	0001011	30	0100110
7	1001010	15	0011100	23	0001101	31	0110100

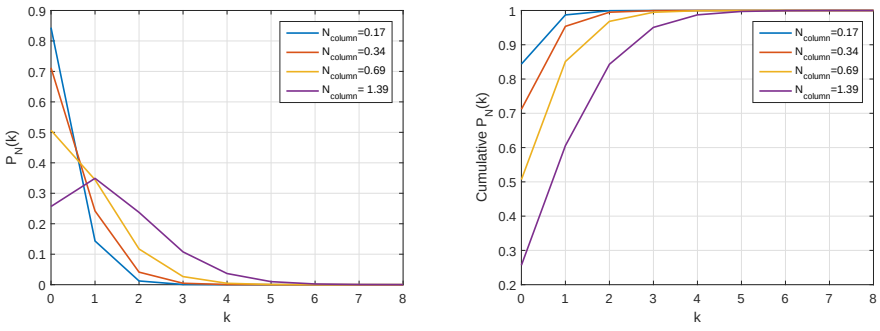
5.2.3. DYNAMIC REALLOCATION AND ADDRESS LATCH

To perform time-resolved measurements from a large array of pixels in parallel, many sensors have employed a TDC-in-pixel approach. In front-side illumination (FSI) technologies, this results in large pixel pitch and low fill factor, e.g. 19.48% fill factor for 44.64 μ m pitch [2] or 1% for 50 μm [1]. To improve the fill factor, instead of in-pixel TDC architecture, a TDC sharing scheme is employed in Piccolo. Due to the SPAD dead time and readout throughput limitations, in many applications pixel activity rates must be restricted to 1-3% to limit distortion due to photon pileup. With a limited pixel activity, in Piccolo, TDC bank is sized to achieve the same detection efficiency as that of in-pixel

TDC architecture. Since the activity of each pixel in one column is statically independent, the light incident can be modeled with a Poisson distribution, given by (5.2)

$$P_N(k) = \frac{N^k * e^{-N}}{k!}, \tag{5.2}$$

where $P_N(k)$ is the probability of k photons detected in one cycle, N represents the column activity that the average number of pixels firing in a column during one detection cycle. In Piccolo, N is determined by the IO bandwidth. Since each column has a dedicated IO working at 160 MHz and the TOF event data length is 23 bits, a maximum N of 0.17, 0.34, 0.69 and 1.39 can be obtained at 40, 20, 10 and 5 MHz illumination frequency, respectively, which covers the complete TDC dynamic range. If N is higher than these numbers, the readout block cannot stream out the data in time due to the limited IO bandwidth, which will cause data overflow and distortion. The probability distribution and cumulative distribution of $P_N(k)$ is shown in Fig. 5.5. We can see that more than 95% of the events can be detected with only three TDCs per column in all the cases, indicating a same photon throughput can be achieved with the TDC sharing approach compared to per-pixel TDC architectures.



(a) (b)

Fig. 5.5 (a) Poisson distribution and (b) cumulative distribution of $P_N(k)$ at column activity of 0.17, 0.34, 0.69 and 1.39, where more than 95 % of the events can be detected with 3 TDCs per column.

TDC sharing architectures have been implemented in some works, where one TDC is shared or multiplexed with a number of pixels [4–6]. 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, we implemented a dynamic

reallocation architecture, to perform pixel address latching and TOF measurement.

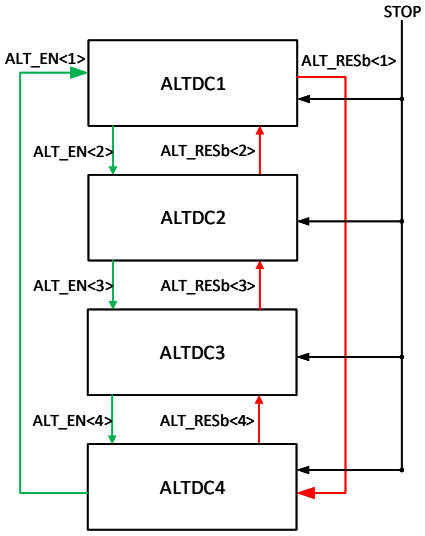


Fig. 5.6 ALTDC daisy chain block diagram

The diagram of the architecture is shown in Fig. 5.6, comprising 4 address latches and TDCs (ALTDCs) connected in a daisy chain configuration and enabled sequentially at the photon detection. At any one time, only one ALTDC is enabled for address latching and timing measurement. A simplified schematic of the ALTDC is shown in Fig. 5.7, where each ALTDC is enabled by ALT_EN<i-1> from the previous block and reset by ALT_RSTb<i+1> from the subsequent block. To prevent the entire ALTDC chain being reset by detection of 4 events, there is always one ALTDC keeping inactive in every cycle, which limits the maximum number of photons that can be detected in one cycle to 3. Fig. 5.8 shows the timing diagram associated with photon detection by ALTDC<1> which is enabled after a global reset with extra peripheral logic, T0. When a pixel in the column detects an event, T1, 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 to achieving fast address latching. At the falling edge of TIMING, T2, ALT_EN<1> rises to logic high, which 1) enables ALTDC<2> for photon detection; 2) latches the address to ADDR_L<1>; 3) triggers VALID signal to begin event-driven readout process. At the end of the cycle, T3, 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 the EOC signal can be sampled correctly. With EOC<1> high signaling the capture of TOF data, Tri_EN<1> is asserted by column readout block and ALTDC<1> is readout through two tri-state buses, O_ADDR and O_TDC, T4. The data on the tri-state buses is finally written into a FIFO memory for read out off-chip via a per-column IO pad. With the event-driven readout method applied, no power is dissipated communicating null events, which is the typical case for TDC in-pixel architectures [1–3, 7].

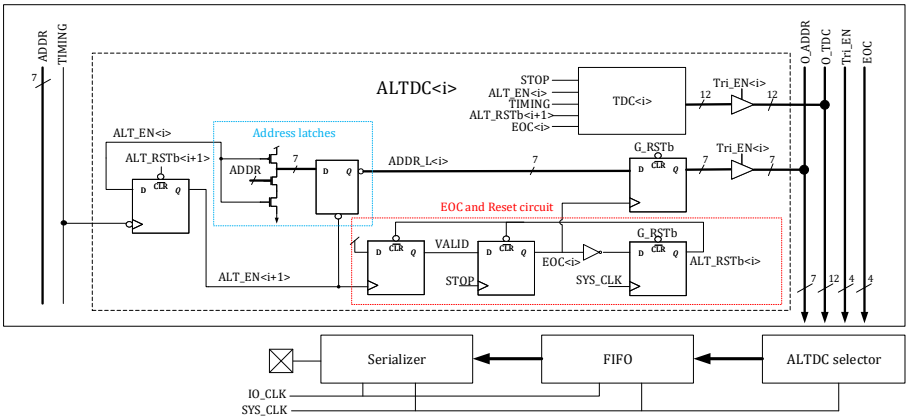


Fig. 5.7 Simplified ALTDC schematic

The minimum time between photons that can be detected is limited by two factors: ADDR/TIMING pulse width, known as the bus dead time, and the propagation delay of the ALTDC. Due to the load capacitance mismatch between TIMING and ADDR buses, pulses will be skewed in time, which limits the minimum pulse width that can be used to latch the addresses correctly. A minimum photon interval of 1.5 ns is achieved from post-layout simulation.

5.2.4. TIME-TO-DIGITAL CONVERTER

A major benefit of SPAD sensor is the compatibility with CMOS technology, which allows the integration of TDCs and other circuitry on the same silicon. This opens the possibility of time-resolved system with a large number of channels, where TDC is the key component for photon arrival time measurement. For the design of a TDC for large sensor arrays, some key requirements need to be considered. Firstly, in terms of temporal resolution, sub-gate resolution can be achieved with techniques, e.g. Vernier-delay-line [12–15], pulse shrinking [16, 17]. However, large area and low conversion rate limit the application in real-time TOF measurements with a large number of TDCs. In contrast,

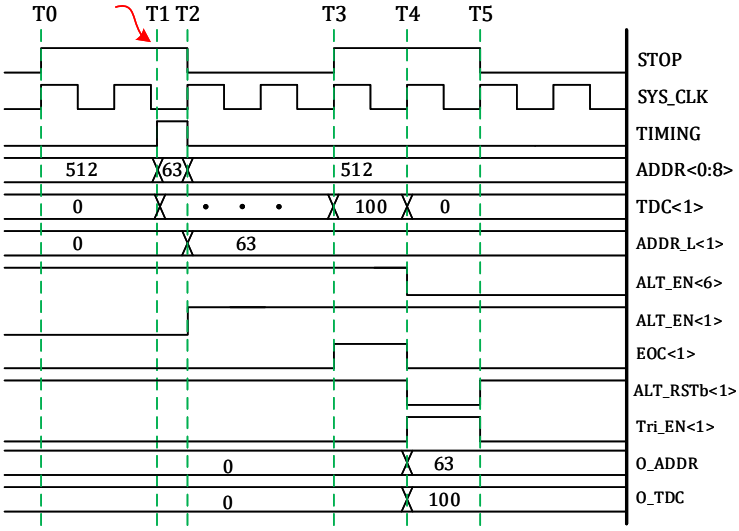


Fig. 5.8 ALTDC timing diagram with a photon detection

ring oscillator (RO) based TDCs have been widely used, featuring moderate resolution (<100 ps), compact size and flash conversion. A second aspect is the power consumption, since a small power increase at a single TDC can result in large power consumption when implementing a large number of channels. For the pixels, high power consumption can heat up the sensor, thus increasing the dark count rate. While for the circuits, IR drop can be a serious problem in a large TDC bank. Besides, for multi-channel systems, the matching between different channels is important in the measurement. Techniques, such as phase-coupled TDC [18] and multi-phase sharing TDC [3, 7, 19], have been explored to improve the matching among a large number of TDCs. The downside of these approaches is that the power consumption can be very high. In [7], 16 phases were distributed over the sensor and a total TDC power of 2.5 W was reached with 1024 TDCs. For the phase-coupled TDC, in order to achieve the phase alignment, all the ROs have to keep oscillating regardless of the photon detection, which reduces the power efficiency. Even though the increase of temperature, due to the power consumption, can be solved with active cooling, the complications in the system design implies it is not a desirable solution.

In Piccolo, a 4-stage pseudo-differential RO based TDC architecture was employed, as shown in Fig. 5.9. A thick oxide NMOS transistor M1 is used to regulate the voltage supply for the RO to mitigate against frequency variations due to IR drops in the ALTDC array. A 9-bit counter is connected to the RO clock which operates at 2.5 GHz, generat-

ing a coarse resolution of 400 ps. A phase discriminator resolves the 8-bit thermometer-coded phases and converts them to a 3-bit binary code, leading to a fine resolution of 50 ps. Synchronizers, similar to the design in [5], were designed to reduce the metastability when the asynchronous signals TIMING and BUSY switch from low to high. The 128 column TDCs, sharing one common control voltage VBIAS, are externally biased, where process-voltage-temperature (PVT) compensation can be implemented off chip via an on-chip replica RO. On the other hand, due to the random device mismatching between different ROs, the open-loop oscillation frequency could be different. With long time accumulation, this frequency offset could generate a big difference in the TDC output. For instance, if two ROs operate at frequency of 0.99 × 2.5 GHz and 1.0 × 2.5 GHz with 1% frequency difference, a maximum code difference of 40.96 LSB could be reached with a 12-bit dynamic range. To reduce this offset error, per-TDC frequency calibration should be applied with the measurement.

The timing diagram of the TDC operation is shown in Fig. 5.10. The TDC is enabled by the previous ALTDC slice once it has detected a photon and asserted ALT_EN at T1 in Fig. 5.10. With a photon detected by a SPAD, T2, the rising edge of TIMING starts the TDC conversion by setting EN as high, where the RO oscillates at a nominal frequency of 2.5 GHz. The conversion is stopped by the rising edge of the reference signal STOP, T3. The least significant bits TDC<0:2> is obtained from the phase decoder and the most significant bits TDC<3:11> from the coarse counter. The TDC output will be read out via a shared tri-state bus, as is discussed in 5.2.3.

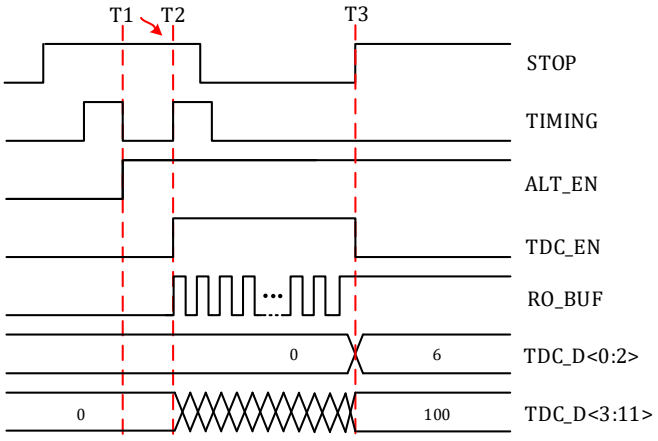


Fig. 5.10 TDC operation timing diagram with a photon detection

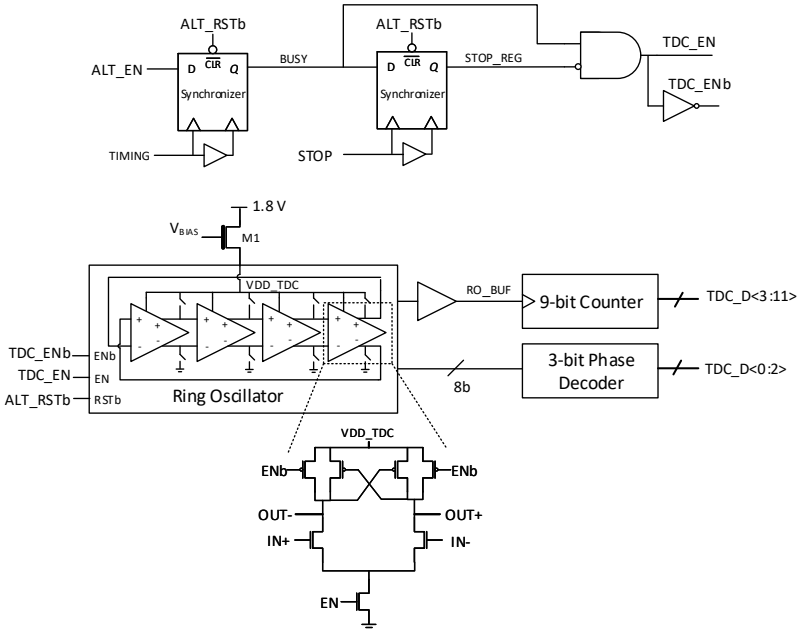


Fig. 5.9 TDC architecture based on a 4-stage RO

5.2.5. CHIP REALIZATION

The sensor was fabricated in a CMOS 180 nm CMOS image sensor (CIS) technology, and a microphotograph of the chip is shown in Fig. 5.11 with dimension of 5 mm x 2 mm . An array of 32 × 32 pixels was implemented, where 3 pixels are not connected to the main array and only used for SPAD test and pixel debugging purposes. From the microphotograph, we can see that the ALTDCs occupy the largest area. However, with a sharing architecture, the ALTDCs are moved outside of the pixel region, which doesn't impact the fill factor.

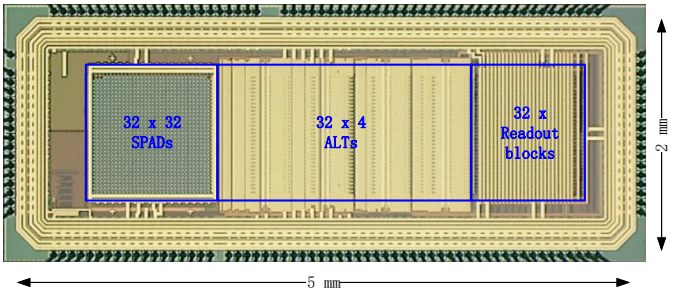
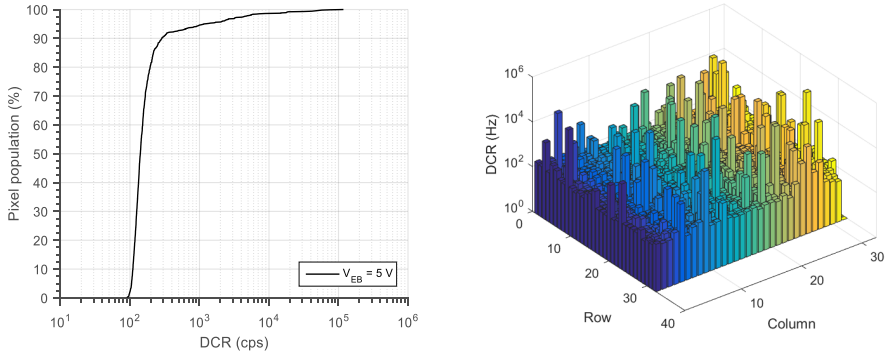


Fig. 5.11 Microphotograph of the sensor

5.3. RESULTS

5.3.1. DARK COUNT RATE

The pixel, of 28.5 μm pitch and 28% fill factor, is achieved with low DCR, as was reported in [8]. The breakdown voltage was measured at 22 V. DCR measurement at 5V excess bias voltage of the whole array is shown in Fig. 5.12, where the median value is 114 cps with an active area of 231 μm^2, which corresponds to a DCR of 0.49 cps/μm^2 at 20 °C temperature. From the DCR proportion distribution in Fig. 5.12(a), high DCR uniformity is achieved, where 94% of the SPADs have a DCR below 1 kHz.



(a) (b)

Fig. 5.12 DCR characterization of Piccolo pixel array at room temperature. (a)cumulative DCR population density of the whole array at 5V excess bias voltage, where 93% of pixels have a DCR lower than 1 kcps;(b) the DCR map of the pixel array

5.3.2. PHOTON DETECTION PROBABILITY

The PDP is measured with an integrating sphere, which is illuminated with monochromatic light. The number of photon counts detected by the SPAD are then compared to the photocurrent from a reference diode which also measures light intensity at an output port. The PDP characterization is shown in Fig. 5.13, where a peak value of 47.8% was achieved at a wavelength of 520nm with 5 V excess bias. PDP of 8.49%, 4.7% and 2.8% was achieved at 840nm, 900nm and 940nm respectively, which provides more flexibility for 3D imaging at near infrared wavelengths. With the cascoded quenching circuit, the maximum excess bias voltage for reliable operation is 5.2 V. More than 50% peak PDP was achieved at 7 V excess bias voltage. However, the long term operation without device breakdown or performance degradation cannot be guaranteed.

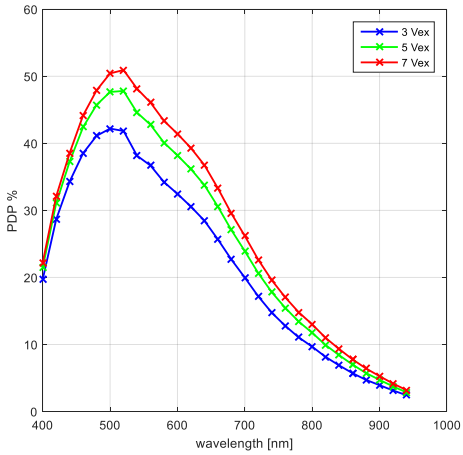


Fig. 5.13 PDP characterization for excess bias voltage in a range of 3-7V.

5.3.3. SPAD JITTER

The SPAD jitter was measured with a 700 nm laser, with a pulse width of 20 ps. The result was presented in Fig. 5.14 achieving 140 ps, 106 ps, 100 ps at full-width-half-maximum (FWHM) for the excess bias voltage of 3V, 5V and 7V respectively. The measurement is matching with the result reported in [8]. Even though the cascaded transistor, M1 in Fig. 5.2, in series is expected to increase the resistance of the quenching path, while the integrated quenching circuit reduces the load capacitance, the jitter performance is maintaining the same level as reported in [8]. Since the laser jitter is 40 ps, the pixel jitter can be extracted as 134ps, 110ps and 98ps at excess bias voltage of 3V, 5V and 7V respectively.

5.3.4. AFTERPULSING PROBABILITY

The afterpulsing probability (AP) was measured with 25 ns dead time at 5 V excess bias voltage, as shown in Fig. 5.15, where no afterpulsing is observed. Whilst the AP was measured at 7.2% from the same SPAD in [8] at excess voltage of 11 V and 300 ns dead time, without an integrated quenching circuit. Large capacitance was introduce during the measurement when probing the anode directly with an oscilloscope. Therefore, the reason of this improvement, we believe, 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 [20].

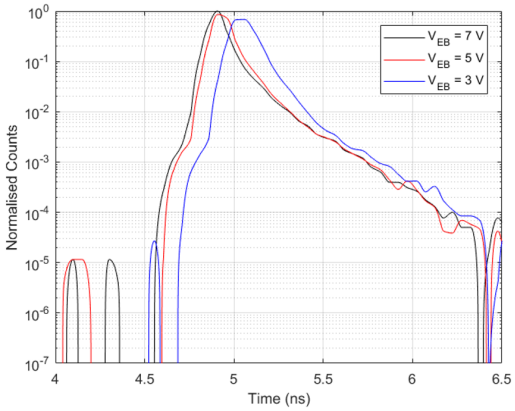


Fig. 5.14 SPAD timing jitter histogram for excess bias voltage of 3V, 5V and 7V at 637 nm.

5

5.3.5. TDC NONLINEARITY

To measure the nonlinearity 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 code density test method, where the STOP signal is generated with the FPGA. For the entire sensor, the temporal resolution (1 LSB), differential non-linearity (DNL) and integral non-linearity (INL) of each TDC can be calculated with code histogram statistics. The nominal LSB of the TDC is 50 ps, where the RO operates at 2.5 GHz. A good uniformity among TDCs was achieved with a LSB standard deviation of 0.48 ps over all the 128 TDCs, Fig. 5.16, which implies 99.7% TDCs are running at a frequency difference within +/-1.4%.

The DNL and INL measurement results obtained from code density test are shown in Fig. 5.17, where -0.07/+0.08 LSB DNL and -0.38/+0.75 LSB INL were achieved with a 20 MHz reference signal. From the measurement, a periodic DNL/INL nonlinearity component can be observed; this behavior is due to a weak coupling of the readout clock to the RO bias voltage. Fig. 5.18 shows the peak-to-peak (p2p) DNL and INL cumulative distribution of all the TDCs. As expected, the p2p INL is proportional to the TDC conversion time, since more noise is coupled and accumulated. Even so, a median p2p DNL and INL of 0.21 LSB and 0.92 LSB were achieved at 20 MHz reference signal, which shows high homogeneity across the image sensor despite the fact that no PVT compensation was applied to the TDCs.

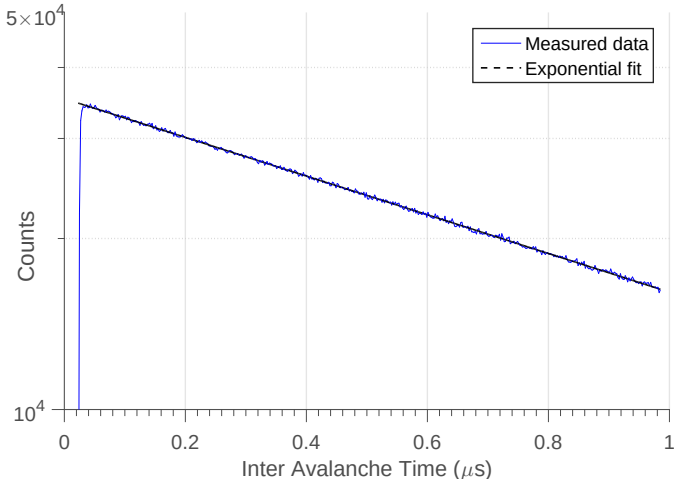


Fig. 5.15 Inter-arrival time measurement with uncorrelated illumination at 5 V excess voltage with 25 ns dead time. No obvious afterpulsing is observed.

5.3.6. TIMING RESPONSE

The system timing response was measured by illuminating the sensor with a pulsed laser with a wavelength of 637 nm. In the signal propagation path, photons were firstly detected by the SPADs; the triggering signal then travel through the collision detection bus and ALTDC chain, starting the conversion of TDCs; finally the TDC is stopped by a STOP signal, which is synchronized with the laser. At each node of the delay path, jitter could be introduced. With a 5 V excess bias voltage applied to the SPADs, a minimum FWHM jitter of 2.28 LSB (114 ps) was achieved, as shown in Fig. 5.19(a). Since the FWHM jitter of the laser, pixel and the TDC quantization noise is 40 ps, 98 ps and 32 ps respectively,

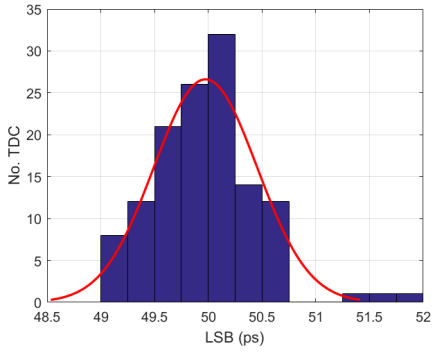
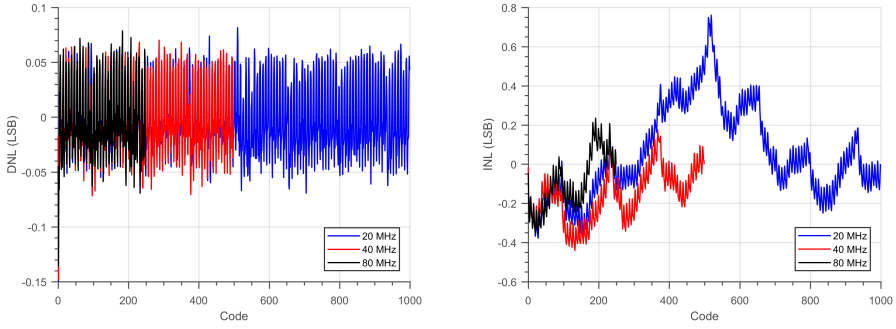
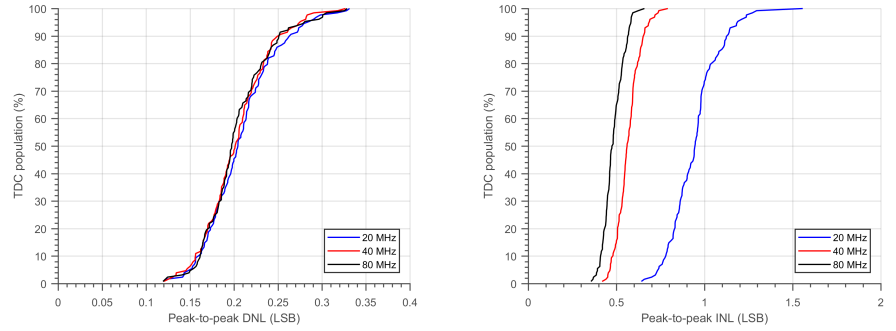


Fig. 5.16 LSB distribution of the 128 TDCs shows a standard deviation of 0.48 ps



(a) (b)

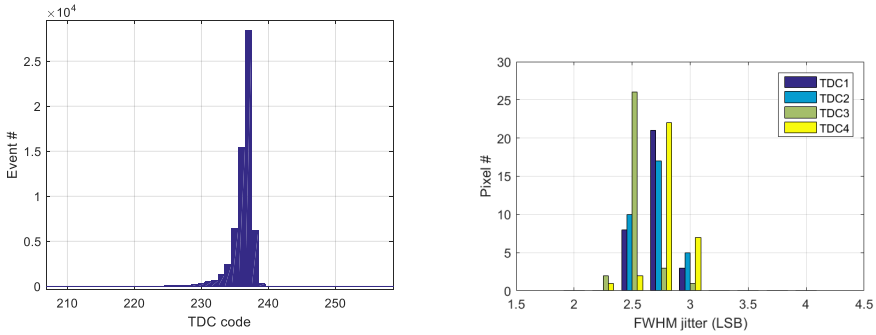
Fig. 5.17 DNL (a) and INL (b) of the TDC at STOP frequency of 20, 40 and 80 MHz.



(a) (b)

Fig. 5.18 Peak-to-peak DNL (a) and INL (b) of the TDC at STOP frequency of 20, 40 and 80 MHz.

we can calculate the average jitter from collision detection bus and ALTDC daisy chain is only 28 ps. The low jitter performance of the sharing architecture is also proved by the good timing uniformity of the 32 pixels from one column, Fig. 5.19(b), where the average and standard deviation of the jitter was achieved at 2.68 LSB (134 ps) and 0.15 LSB (7.5 ps) respectively. This also implies that larger arrays could be implemented in sharing architecture without degrading the array timing performance.



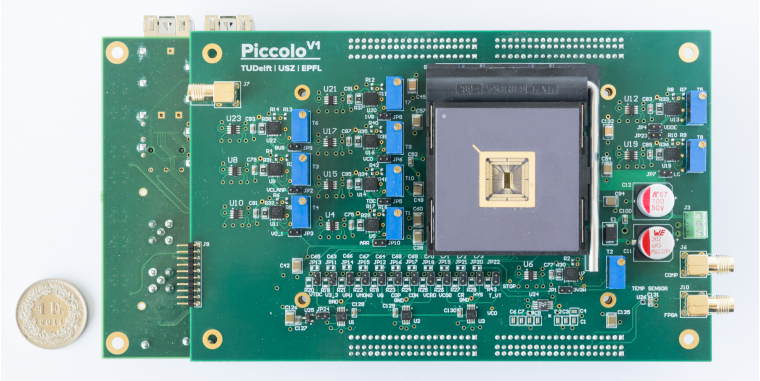
(a) (b)

Fig. 5.19 (a) Single shot SPAD-TDC timing jitter measurement with a minimum FWHM of 2.28 LSB (b) jitter distribution of all the pixels at each TDC measurement, leading to the average and standard deviation of 2.68 LSB and 0.15 LSB respectively.

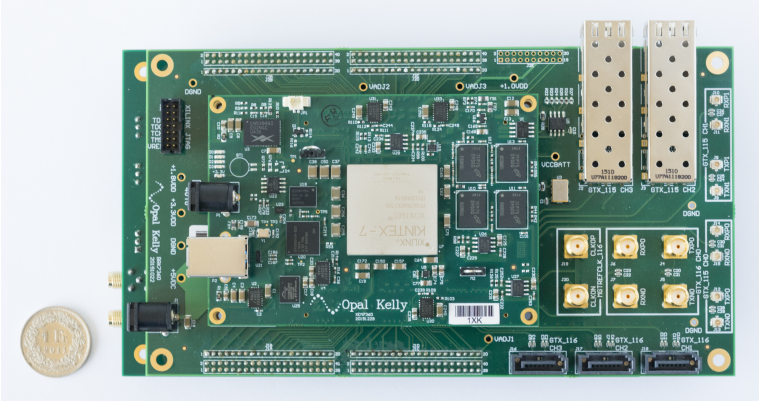
5.3.7. PICCOLO CAMERA SYSTEM

The Piccolo measurement system is shown in Fig. 5.20. The system comprises 3 printed circuit boards(PCBs), one daughterboard with the sensor, LDOs and other components mounted on top, one break-out board (BRK7360, Opal Kelly) to provide debugging interface of the sensor through a set of pin headers, and one FPGA board (XEM7360 based on Kintex-7, Opal Kelly) which is used for data readout and processing. The data transmission between the FPGA and the computer is carried out with an universal serial bus (USB) 3.0 interface.

Piccolo was firstly demonstrated in a flash imaging system, where a lens was placed in front of the sensor, achieving a field-of-view (FOV) of 40 degree × 40 degree. The spatial resolution is limited in flash imaging mode, due to the limited number of pixels. To extend the resolution, a scanning LiDAR system was built, as is shown in Fig. 5.21, comprising a dual-axis galvanometer scanner (Thorlabs GVS012) which is driven by an arbitrary waveform generator (AWG, Keysight 33600A), and a 637 nm pulsed laser. In this system, all the SPADs working as one component performs single point scanning and imaging. System control and synchronization is performed in the FPGA. In order to achieve real-time measurement, the timing histogram of each scanning point is constructed in the FPGA. Because the size of histogram data is much smaller than that of raw data, the bandwidth requirement of the USB interface is alleviated. The AWG generates two channels of step signals, driving the dual axis scanner to perform a raster scanning of the scene, thus allowing high resolution imaging to be carried out.



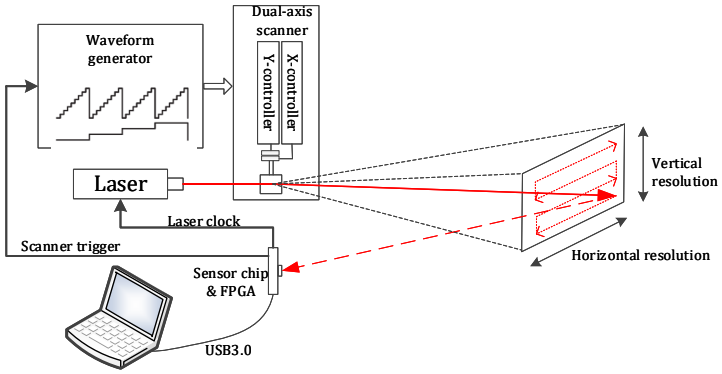
(a)



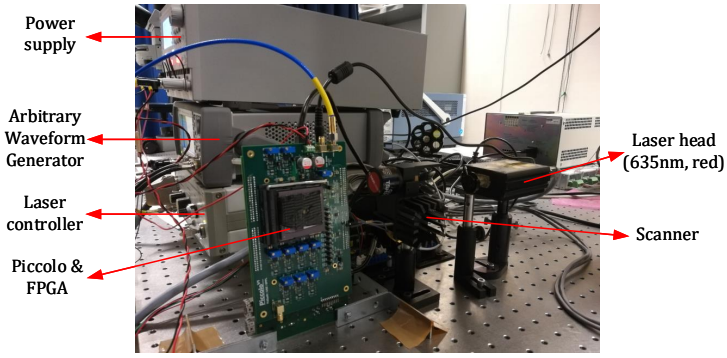
(b)

Fig. 5.20 (a) Frontside and (b) backside of Piccolo camera system

However, the photon collection efficiency is relatively low in the scan setup, because there is no objective placed in front of the sensor. Due to the SPAD quenching mechanism, each pixel can only detect one event at a time. For SPAD based scanning system, in order to increase the detection effectiveness, all the pixels has to be illuminated by the reflection light from different scanning points on the scene, which requires the FOV of each pixel to cover the entire scene. However, with a conventional optical setup, each pixel in the array only stares at a small fraction of the scene. So in this experiment, objectives are not mounted, which limits the photon collection efficiency. To improve the photon collection efficiency, more complicated optical setup is required, but this is out of the scope of this thesis. One possible solution was reported in [21], in which the laser diode is aimed coaxially at a 6-facet polygonal mirror. Since each facet of the polygonal mirror has a slightly different tile angle, 2D scanning can be achieved by rotating the



(a)



(b)

Fig. 5.21 Piccolo scanning measurement setup with (a) system diagram and (b) camera system photograph

polygonal mirror. At the receiver side, back-reflected photons are collected by the same facet and imaged onto the sensor at the focal plane of a concave mirror, so the FOV of each pixel can be much smaller.

5.3.8. FLASH IMAGING MEASUREMENT

To validate the Piccolo sensor, a flash 3D imaging measurement was performed where a target was illuminated with a diffused laser and the reflected light collected on a per-pixel basis. TDC calibration was applied for LSB variations among different TDCs, as well as time offset due to the skew of STOP clock. A 3D image can be constructed by histogramming the TOF data of each pixel then calculating the distance. As is shown in Fig. 5.22, a 3D image was obtained, where a person with the right hand raised standing at a distance of 0.7 m away from the sensor. The target was illuminated with a super-continuumlaser (SuperK Extreme, NkT Photonics) and an Acousto-optic tunable filter

(AOTF), operating at a wavelength of 650 nm. Due to the limited laser power, the measurement was performed at dark conditions and with an exposure time of a few seconds. Millimetric detail can be observed thanks to the low timing jitter of the system and high single-to-background noise ratio (SBNR).

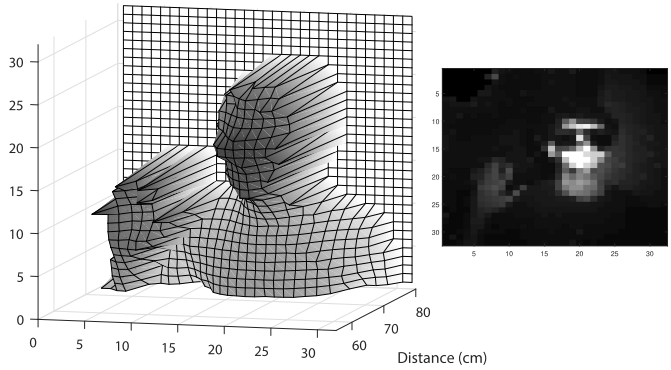


Fig. 5.22 Flash imaging measurement of a human subject at distance of 0.7 m with 2D intensity image inset.

5.3.9. DISTANCE CHARACTERIZATION

To perform scan imaging, the entire pixel array was used as a single detection component, where the mismatching between TDCs is accumulated with time. To improve the accuracy of the measurement, calibration has to be applied to each TDC and SPAD. The single shot timing response of the whole array, Fig. 5.23, was acquired by electrically triggering all the TDCs with a 40 MHz STOP signal at a phase shift step of 25 ps. As expected, the jitter is proportional to the TDC value, as more nonlinearity error is accumulated. Whilst with the calibration the FWHM jitter is stabilized and reduced from 10.63 LSB to 5.87 LSB in average. However, as is discussed in section 5.3.6, for a single pixel, the average jitter from a single TDC is 2.68 LSB which is smaller than that the system jitter of 5.87 LSB. Two main factors contribute to the calibration degradation. The first one is the calibration quantization error. In order to achieve real time imaging, the histogram statistic was built on the FPGA, which requires the calibration has to be performed inside the FPGA. However, to reduce the complexity of the firmware, the calibration process was based on integer instead of fixed point decimal, which reduces the accuracy due to the accumulation of quantization error from each TDC. Another reason is the RO frequency stability. Since all the ROs operate in an open-loop, the frequency is varying over temperature, leading to an unstable TDC linearity which is difficult to calibrate with a constant

coefficient.

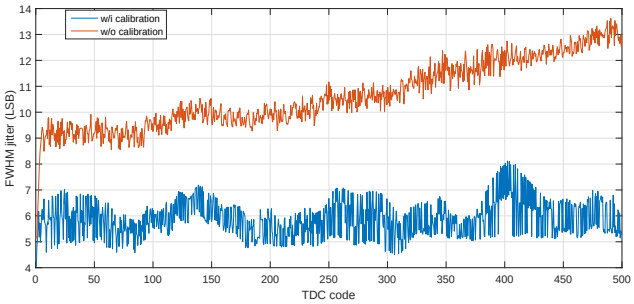


Fig. 5.23 Jitter measurement of the system w/i and w/o calibration in full range, where the average jitter reduces from 10.63 LSB to 5.87 LSB.

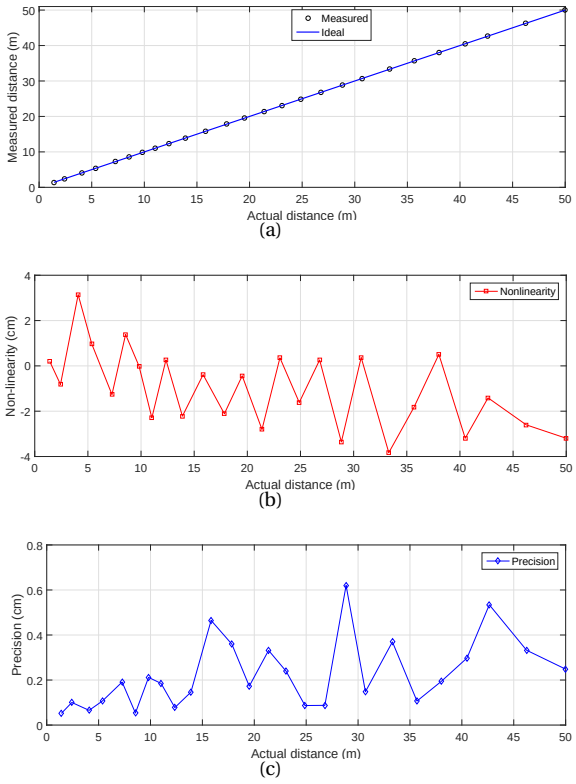


Fig. 5.24 (a) Measured distance up to 50 m as a function of the actual distance; (b)The accuracy and (c) precision at each measurement distance, where a peak-to-peak non-linearity and worst-case precision (σ) of 6.9 cm and 0.62 cm were achieved respectively.

A single point telemetry was performed with a 637 nm pulsed laser at 40 MHz repetition rate, 2 mW average power, 0.5 W peak power and 40 ps pulse width 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 of the system was characterized and shown in Fig. 5.24. A 60% reflectivity target was measured up to 50 m, where each distance was measured with a 50k photons histogram for 10 repeated measures, achieving a peak-to-peak non-linearity and worst-case precision (σ) of 6.9 cm and 0.62 cm respectively, over the entire range. In order to perform long range characterization, the measurement was performed in dark conditions to reduce the interference from the background light, due to the limitation of the laser power and photon collection efficiency of the setup.

5.3.10. SCAN IMAGING MEASUREMENT

The operation diagram of the scanning system is shown in Fig. 5.25, where the control and most of the processing operations are performed in the FPGA, achieving real-time imaging. To improve the accuracy, an averaged value of the histogram in a range of +/-8 bins to peak was used instead of the absolute peak value, which gives a millimetric detail depth resolution, as is shown in Fig. 5.26. In this experiment, the scanner was configured to operate at a resolution of 128 × 128, where both depth and grayscale images can be obtained at the same time. Facial image of a mannequin, placing at 1.3 m away from the sensor with curved background, was obtained with a scanning frequency of 1 kHz. Dark operating conditions and 1 ms acquisition time ensured that more than 10 K photons can be acquired at each point, enabling high SBNR and measurement accuracy.

By increasing the scan frequency, real-time imaging was achieved with 50 lux indoor light conditions at a resolution of 64 × 64 with the same laser, as is shown in Fig. 5.27. A person (reflectivity of about 40 %) standing 10 m away to the sensor, waving a hand and turning around, was recorded at 6 frames/s. In order to obtain high spatial resolution images, the FOV was adjusted to be 5 × 5 degrees, which gives a fine angular resolution of 0.078 degree in both X and Y direction, corresponding to a scanning step of 1.36 cm per point. Each point required 40 μs for data acquisition, where a narrow optical bandpass filter with FWHM of +/-10 nm was used to suppress the background light. Thanks to the high PDP and photon throughput, sharp images were recorded with an average and peak laser power of only 2 mW and 500 mW, respectively.

The system performance can be further improved mostly in three factors, including pho-

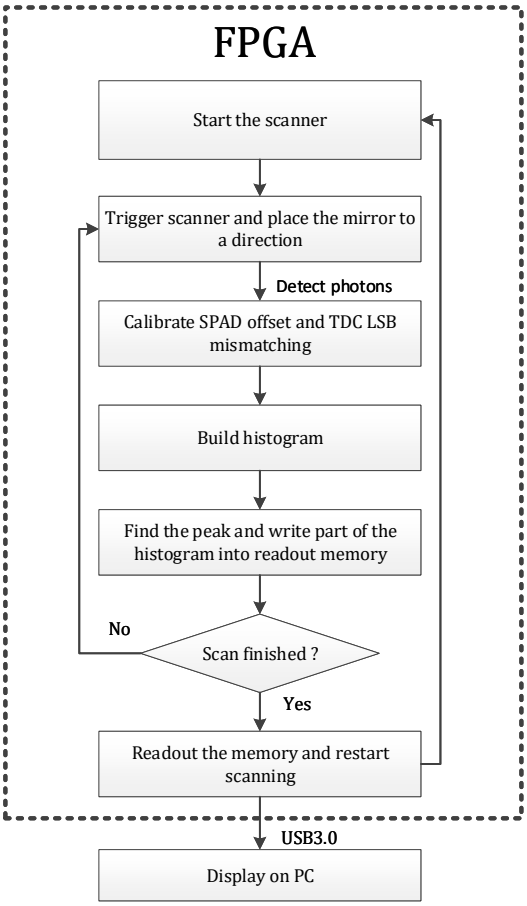
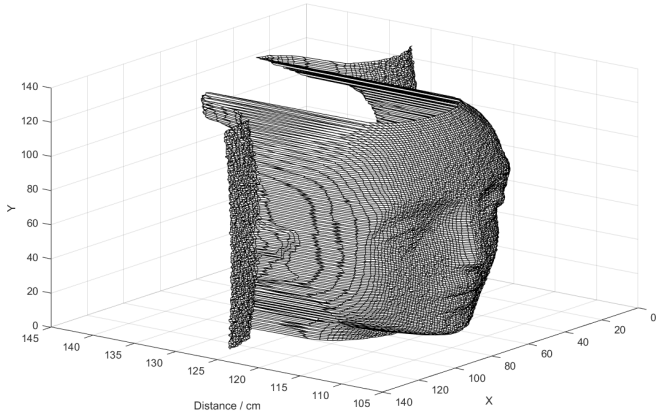
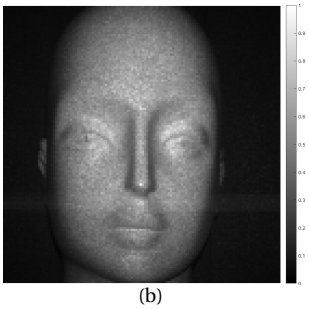


Fig. 5.25 The operation diagram of the scanning LiDAR system. Real-time imaging is achieved by performing most of the operations in the FPGA.

ton collection efficiency, laser power and background light suppression. More specifically, a more efficient photon collection optical setup can be designed, which will significantly increase the photon detection rate. The current laser power is limited at a low level, which reduces SBNR, so as the measurement distance. With a higher laser power, more photons can be reflected back to the sensor in each laser cycle, and less cycles are required to build reliable statistics. As a result, the scanner can operate at a higher speed, which improves the frame rate and spatial resolution. On the other hand, SBNR can be improved by suppressing the background light. Even though bandpass filter is used in the experimental setup, more techniques can be implemented in the future. For example, in [21–23] spatiotemporal correlated events are used to detect photons within a short coincidence window, which improves the rejection of detection events which are



(a)



(b)

Fig. 5.26 (a) 3D and (b) 2D image of a mannequin with a 128 x 128 resolution at distance of 1.3m, acquired with scan ranging measurement.

due to DCR or background light. The photon gating method in [24] allows only photons from a specified distance range to be detected. This reduces the sensitivity to photons outside the range of interest. However, these techniques also bring negative impacts at the same time, e.g. fill factor, array size, and power consumption. Further investigation is required to achieve the best balance between these characteristics.

5.3.11. POWER CONSUMPTION AND PERFORMANCE SUMMARY

The power consumption of the sensor, which is strongly dependent on the operating environment, was measured during the LiDAR experiment with a total photon throughput of 35.5 Mcps. At this activity, the total power consumption is 0.31 W, corresponding to a detection power of 8.7 nW/photon. The contribution of each components was shown

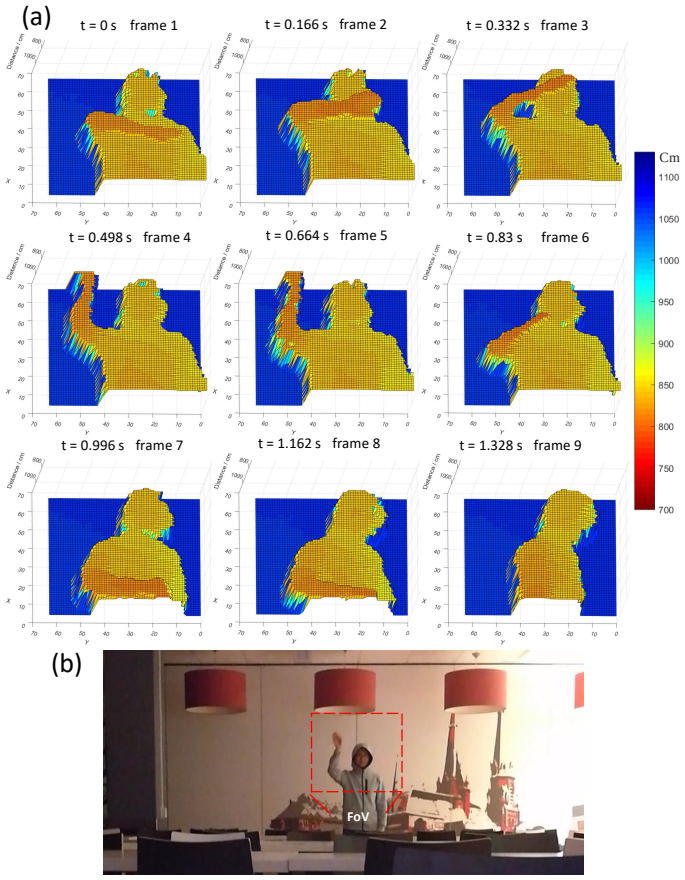


Fig. 5.27 (a) 9 consecutive frames recorded at 6 frames/s with resolution of 64 x 64 at 10 m distance; (b) image captured with a conventional camera

in Table 5.2. With the increase of event rate, the power consumption of IO would be expected to scale linearly. Increases would also be expected for the ALTDCs and the readout blocks, however, since a large proportion of this power is dissipated by the clock network, including STOP and SYS_CLK, the increases would not be linearly proportional to the photon detection. The summary of the sensor, including the architecture characteristics and measurement results, is presented in Table .5.3.

Table 5.2: Piccolo power consumption

Components	Power (mW)	Contribution (%)
ALTDCs	93	30
Readout	86.1	27.8
I/O	82.7	26.7
Pixel array	34.4	11.1
Test circuit	13.8	4.4
Total	310	

5.4. COLLISION DETECTION BUS BASED BACKGROUND LIGHT SUPPRESSION ARCHITECTURE

One big concern about LiDAR application is the background light suppression. With strong background light, the shared TDCs will be saturated by the ambient photons, which prevents the TDCs from being triggered by the laser signals. In order to improve the tolerance to the background light, a commonly used method is the detection of coincidence photons, which have been applied in [21–23], where coincidence events with be detected when more than one photon are detected in a coincidence time window. Since the uncorrelated background photons exhibit random trigger times, they tend to be uniformly distributed in the histogram. Whilst the energy of the laser is concentrated in a short pulse, the probability of coincidence detection from the laser pulse is higher than that of the background light. With this method, uncorrelated photons can be filtered out, thus the TDCs can be prevented from saturation and improved SBNR can be achieved. In order to distinguish coincidence photons, multi-levels of adders were implemented in [21, 22], referred to as method 1, which count the number of photons detected by a set of SPADs in a predefined time window. Due to the large area occupance of the adders and the requirement for the path delay compensation between the signal bit and carrier bit, it is difficult to be implemented in a large array. Another method was applied in [23], referred to as method 2, where a group of SPADs were combined to a single output via a balanced OR tree. This output is then connected to a set of shift registers. With the detection of photons, a pulse train will be generated and triggers the shift registers outputting different patterns according to the photon number. In this approach, a smaller coincidence detection circuit is constructed, which enables a larger array being implemented. Furthermore, the threshold of coincidence photon number can be easily configured by selecting the output of the shift registers, which enables the optimal coincidence detection according to the background light intensity. However, a drawback of this approach is the event missing detection problem. Since the pixels are connected to an OR tree, when two or more photons are detected at the same time, only one pulse will be gen-

erated. In this case, the shift registers will be triggered only once, so the coincidence event will not be able to be detected, which reduces the SBNR. Another problem is that the TDCs will be triggered by any event, while in case of uncorrelated events the TDCs will be reset after the coincidence window. So a high TDC power consumption can be expected with high background light.

With the intrinsic capability of coincidence photon detection, collision detection bus can be used for background light suppression. Since coincidence events will generate invalid address with different number of '1's, by counting the number of '1's on the bus, these events can be recognized easily. A proposed sensor architecture is shown in Fig. 5.28, where a group of 32 pixels are employed for coincidence detection. As is discussed in section 5.2.2, collision events will generate an address output with more than 3 bits of '1's. Therefore, when coincidence events are detected, the most-significant bit (MSB) of the adder, Z<2>, will rise to high, which can be directly used for the triggering of the TDC.

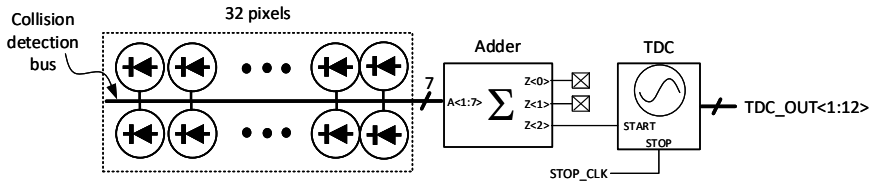


Fig. 5.28 Proposed sensor architecture with coincidence event detection among 32 pixels, based on collision detection bus.

In comparison with method 1 and 2, instead of 32 bits, only 7 bits have to be processed, so a much smaller coincidence detection circuitry can be constructed. More specifically, to perform coincidence detection with 32 pixels, 13 full-adders, 4 half-adders, one AND gate and one 18 input OR gate are required in method 1, while 31 NAND/NOR gates are required in method 2. Instead, the proposed approach only needs 3 full-adders and 4 half-adders. On the other hand, since it is based on an adder, the event-miss problem in method 2 would not happen in this approach. Furthermore, since the TDC can only be triggered by the output Z<2>, low power consumption can be achieved. In comparison to the current sensor architecture, only a minor modification with the implementation of the adders is required, which prevents the features of the pixel array from being affected. Therefore, with the proposed approach, a SPAD sensor with a large pixel array, higher fill factor and high background light suppression is expected.

5.5. CONCLUSION AND DISCUSSION

In this chapter, a 32 × 32 SPAD imager is presented, which was fabricated in a 180 nm CMOS image sensor technology. Each 32 pixels in one column were connected to a shared collision detection bus. With this intense resource sharing architecture, a fill factor of 28% is achieved with a pixel pitch of 28.5 μm. To improve the detection throughput, a scalable latch chain mechanism was used to dynamically reallocate TDCs for TOF detection. For the time arrival measurement, an array of 128 12-bit TDCs was implemented, operating with a resolution of 50 ps. The readout is through a tri-state bus, working in an event-driven readout method that only output valid timing data. This avoids the low readout efficiency due to the null communication of in-pixel TDC architecture or high power consumption of datapath mechanism. The GPIO, working at a speed of 160 MHz, provides a maximum throughput of 222 Mcps and 465 Mcps in timestamping and single-photon counting mode, respectively. Ranging measurements at distance of 50 m achieved 6.9 cm nonlinearity and 0.62 cm precision. Based on the sensor, a scanning LiDAR system, achieving depth imaging up to 10 m at 6 frames/s with a resolution of 64 × 64, has been demonstrated in 50 lux background light condition, with limited optical power of 2 mW and 500 mW in average and peak, respectively. Featuring high fill factor and PDP, multi-channel TOF measurement, high speed readout as well as easy-to-use property, this sensor is suitable for scanning LiDAR with low background light.

To improve the background light suppression, a new sensor architecture based on the concept of collision detection bus is proposed. Compared to other methods in literature, the proposed method has the benefit of reduced coincidence detection circuitry area and low TDC power consumption, which provides an approach of designing SPAD sensors with a large pixel array and high fill factor for TOF imaging applications in high background light environment, such as automotive LiDAR.

Table 5.3: Performance summary of the sensor and LiDAR system

Parameter	Value	Unit
Chip characteristic		
Array resolution	32 x 32	
Technology	180 nm CMOS	
Chip size	5 x 2	mm^2
Pixel pitch	28.5	μm
Pixel fill-factor	28	%
SPAD break down voltage	22	V
SPAD median DCR	140 (Vex = 5V)	cps
SPAD jitter	106 (Vex = 5V)	ps
SPAD PDP	50.93 (Vex = 7V @520nm)	%
SPAD crosstalk	0.09	%
SPAD afterpulsing	0.1	%
TDC resolution	50	ps
TDC depth	12	bit
No. TDC	128	
TDC area	4200	μm^2
Readout speed	160	MHz
Readout bandwidth	5.12	Gbps
Maximum photon throughput	222 (PT mode)	Mcps
	465 (PC mode)	Mcps
Distance measurement		
Measurement distance range	50	m
Accuracy (Non-linearity)	6.9 (0.14%)	cm
Precision (σ)(Repeatability)	0.62 (0.01%)	cm
LiDAR experiment		
Illumination wavelength	637	nm
Illumination power	2 (average), 500 (peak)	mW
Frame rate	6	fps
Image resolution	64 x 64	
Field of view (H x V)	5 x 5	degree
Target reflectivity	40	%
Distance range (LiDAR)	10	m
Background light	50	Lux
Chip power consumption	0.31 (@ 35.5 Mcps photon throughput)	W

REFERENCES

[1] C. Veerappan, J. Richardson, R. Walker, D.-u. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, *A 160×128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter,* ISSCC , 312 (2011).

[2] L. Gasparini, M. Zarghami, H. Xu, L. Parmesan, M. M. Garcia, M. Unternahrer, B. Bessire, A. Stefanov, D. Stoppa, and M. Perenzoni, *A 32×32-pixel time-resolved single-photon image sensor with 44.64µm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800kHz observation rate for quantum physics applications,* ISSCC , 98 (2018).

[3] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. Dalla Mora, D. Contini, D. Durini, S. Weyers, and W. Brockherde, *CMOS imager with 1024 SPADs and TDCS for single-photon timing and 3-D time-of-flight,* IEEE Journal on Selected Topics in Quantum Electronics **20** (2014), 10.1109/JSTQE.2014.2342197.

[4] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, *A 128x128 single-photon imager will on-chip column-level 10b time-to-digital converter array capable of 97ps resolution,* Digest of Technical Papers - IEEE International Solid-State Circuits Conference **51**, 44 (2008).

[5] 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 Journal of Solid-State Circuits **50**, 2406 (2015).

[6] A. Carimatto, S. Mandai, E. Venialgo, T. Gong, G. Borghi, D. R. Schaart, and E. Charbon, *A 67,392-SPAD PVTB-compensated multi-channel digital SiPM with 432 column-parallel 48ps 17b TDCs for endoscopic time-of-flight PET,* IEEE International Solid-State Circuits Conference - Digest of Technical Papers , 202 (2015).

[7] 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 Journal of Solid-State Circuits **49**, 867 (2014).

[8] C. Veerappan and E. Charbon, *A low dark count p-i-n diode based SPAD in CMOS technology,* IEEE Transactions on Electron Devices **63**, 65 (2016).

[9] 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 Letters **33**, 1589 (2012).

[10] 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,* Optics Express **24**, 17819 (2016).

[11] 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 Letters **38**, 1547 (2017).

[12] P. Dudek, S. Szczepański, and J. V. Hatfield, *A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line,* IEEE Journal of Solid-State Circuits **35**, 240 (2000).

[13] L. Vercesi, A. Liscidini, and R. Castello, *Two-dimensions vernier time-to-digital converter,* IEEE Journal of Solid-State Circuits **45**, 1504 (2010).

[14] J. Yu, F. foster Dai, and r. c. Jaeger, *A 12-Bit Vernier Ring Time-to-Digital Converter in 0.13um CMOS technology,* IEEE Journal of Solid State Circuits **45**, 830 (2010).

[15] 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 Transactions on Circuits and Systems I: Regular Papers **60**, 557 (2013).

[16] R. Szplet and K. Klepacki, *An FPGA-integrated time-to-digital converter based on two-stage pulse shrinking,* IEEE Transactions on Instrumentation and Measurement **59**, 1663 (2010).

[17] P. Chen, S. I. Liu, and J. Wu, *A CMOS pulse-shrinking delay element for time interval measurement,* IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing **47**, 954 (2000).

[18] A. R. Ximenes, P. Padmanabhan, and E. Charbon, *Mutually Coupled Ring Oscillators for Large Array Time-of-Flight Imagers,* Intl. Image Sensor Workshop , 258 (2017).

[19] B. Markovic, D. Tamborini, F. Villa, S. Tisa, A. Tosi, and F. Zappa, *10 Ps Resolution, 160 Ns Full Scale Range and Less Than 1.5 Differential Non-Linearity Time-To-Digital Converter Module for High Performance Timing Measurements,* Review of Scientific Instruments **83** (2012), 10.1063/1.4733705.

[20] S. Cova, M. Ghioni, a. Lacaita, C. Samori, and F. Zappa, *Avalanche photodiodes and quenching circuits for single-photon detection.* Applied optics **35**, 1956 (1996).

[21] 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 Journal of Solid-State Circuits **49**, 315 (2014).

[22] C. Niclass, M. Soga, H. Matsubara, S. Kato, and M. Kagami, *A 100-m range 10-Frame/s 340×, 96-pixel time-of-flight depth sensor in 0.18-µm CMOS,* IEEE Journal of Solid-State Circuits **48**, 559 (2013).

[23] M. Perenzoni, D. Perenzoni, and D. Stoppa, *A 64×64-pixel digital silicon photomultiplier direct ToF sensor with 100Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6km for spacecraft navigation and landing,* 2016 IEEE International Solid-State Circuits Conference (ISSCC) **52**, 118 (2016).

[24] I. Gyongy, N. Calder, A. Davies, N. A. Dutton, P. Dalgarno, R. Duncan, C. Rickman, and R. K. Henderson, *256×256, 100kfps, 61% Fill-factor time-resolved SPAD image sensor for microscopy applications,* Technical Digest - International Electron Devices Meeting, IEDM , 8.2.1 (2017).

6

A 252 × 144 TIME-RESOLVED SPAD SENSOR WITH PIXEL-WISE INTEGRATED HISTOGRAMMING

For SPAD sensors, a major benefit is the potential for designing large pixel arrays, which provide the possibility of designing SPAD based flash LiDAR. However, a large pixel array implies massively parallel time-resolved measurements resulting in a large volume of data. Since the data is typically transmitted off-chip for further processing, the output data bandwidth of the sensor can heavily limit the speed of measurements. This chapter presents a design of 252 × 144 SPAD sensor, called Ocelot, which employs per-pixel partial histogramming to improve the photon throughput aiming for flash LiDAR applications in low ambient light environment.

Ocelot was a collaborative design carried out with Scott Lindner and Ivan Michel Antolovic with a division of labour among the different circuit blocks. The author was responsible for the full design of the partial histogramming readout(PHR), comprising peak detection and partial histograming, where the concept of PHR was conceived jointly with Scott Lindner. The author also designed the complete firmware for the sensor measurement. Scott Lindner was also responsible for the dual-clock TDC and PLLs. IvanMichel Antolovic was responsible for the design of the pixel array and masking scheme. He also co-designed the collision detection bus with Scott Lindner contributing the concept of the

bus repeater scheme. This chapter is based on results presented at IEEE VLSI Symposium 2018, S. Lindner et al. "A 252 ×144 SPAD pixel FLASH LiDAR with 1728 Dual-clock 48.8 ps TDCs, Integrated Histogramming and 14.9-to-1 compression in 180nm CMOS Technology". C. Zhang et al. "A 30 fps, 252 × 144 SPAD Flash LiDAR with 1728 Dual-Clock 48.8 ps TDCs, and Pixel-wise Integrated Histogramming". [Submitted to Journal of Solid State Circuits].

6.1. INTRODUCTION

Image sensors with a larger array are normally preferred, due to the superior performance in spatial resolution. Whilst the challenges of scaling the pixel array involve in several aspects, including fill factor, data throughput, power consumption, and sensor area. To the best of the author's knowledge, the largest SPAD camera with on-chip TDCs was reported in [1], with an array size of 160 × 128. Even though a large pixel pitch of 50 μm was implemented, a fill factor of only 1% was achieved, due to the large area occupied by the in-pixel circuitry such as TDC. With the similar architecture, a 32 × 32 SPAD sensor with a fill factor of 19.48% was reported in [2] with a pixel pitch of 44.64 μm. Due to the large pixel pitch, it is expensive in terms of silicon area for scaling up.

To reduce the amount of the circuitry in pixels, time-gated SPAD sensors [3–7] based on SPAD were designed for time-resolved measurement. In this approach, the SPADs are only enabled for a narrow time window, typically from hundreds of picosecond to nanoseconds. The photons are stored by an in-pixel memory or analog counter, typically based on a capacitor, then readout after the illumination cycle. Full range of time-resolved photon counting can be achieved by temporally sweeping the window at a small step, e.g. 10 ps. In this approach, the pixel circuitry can be very compact, which is possible to achieved large sensor format, small pixel pitch and high fill factor at the same time. A 512 × 512 SPAD sensor with 10.5% fill factor at a pixel pitch of 16.38 μm was achieved in [3]. Higher fill factor of 61% with a pixel pitch of 16 μm sensor was reported in [4], achieving an array size of 256 × 256. However, the downside of this approach is obvious that the acquisition time is multiplied by the number of gates required for the measurement. This limits the applications of time-gated photon counting approach to the fields which can tolerant long measurement time, such as fluorescence life-time imaging (FLIM), super-resolution microscopy.

Along with the increase of pixel number, higher speed data throughput is required, in order to achieve short acquisition time and high frame rate. However, in raw data readout mode, where every event is read out directly via the IO pads, the power consumption is proportional to the data throughput. A data path approach was implemented in [8], achieving output bandwidth of 42 Gbps with total power consumption of 8.79 W. To solve this problem, instead of streaming out the raw data, on-chip histogramming was implemented in [9–11], to accumulate photons for each bin of the TDC. Since the size of the histogrammed data is much smaller than that of the raw format, high compression efficiency can be achieved. This alleviates the requirements for high data throughput, so increasing the power consumption. However, with the large area occupied by memory

in the full range histogramming, these sensors have been limited to single point or line formats.

The architecture of Piccolo, illustrated in Chapter 5, provides an alternative way for implementing a large format time-resolved SPAD sensor. By placing TDCs outside the pixel array and connecting SPADs to a sharing bus, a relatively high fill factor (28%) was achieved with a medium pixel pitch (28.5 μm), referring to Section 5.2.1 and 5.2.2. High photon detection throughput was guaranteed through dynamic reallocation ALTDC chains, Section 5.2.3, where the number of TDCs is determined by the bandwidth of the readout and the pixel activity. However, a major challenge is still unsolved in scaling Piccolo architecture to a large format, the data throughput. In Piccolo, events are read out directly through the GPIO pads in raw format. For the new sensor, if the number of pixel is N times that of Piccolo and the column activity is the same as that of Piccolo, the number of GPIO would be increased by N, so as to achieve the same photon throughput of each column. This is not a desired for sensors with a large number of columns, e.g. 256, in terms of IO pad number, power consumption and system complexity. It is clear that there is a stringent need for on-chip data processing which can reduce the volume of data before transmission off-chip. This is achieved by so called partial histogramming readout (PHR) scheme, which will be presented in Section 6.2.3. To the best of the authors knowledge, this is the first implementation of integrated histogramming for a full array via 3.32 Mb SRAM, providing a 14.9-to-1 compression.

6.2. OCELOT ARCHITECTURE

The block diagram of the sensor is shown in Fig. 6.1. The pixel array is divided into 4 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 6 address latch and time-to-digital converters (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 pulse width.

At the bottom of the half-column, the shared bus lines are connected to all 6 ALTDCs

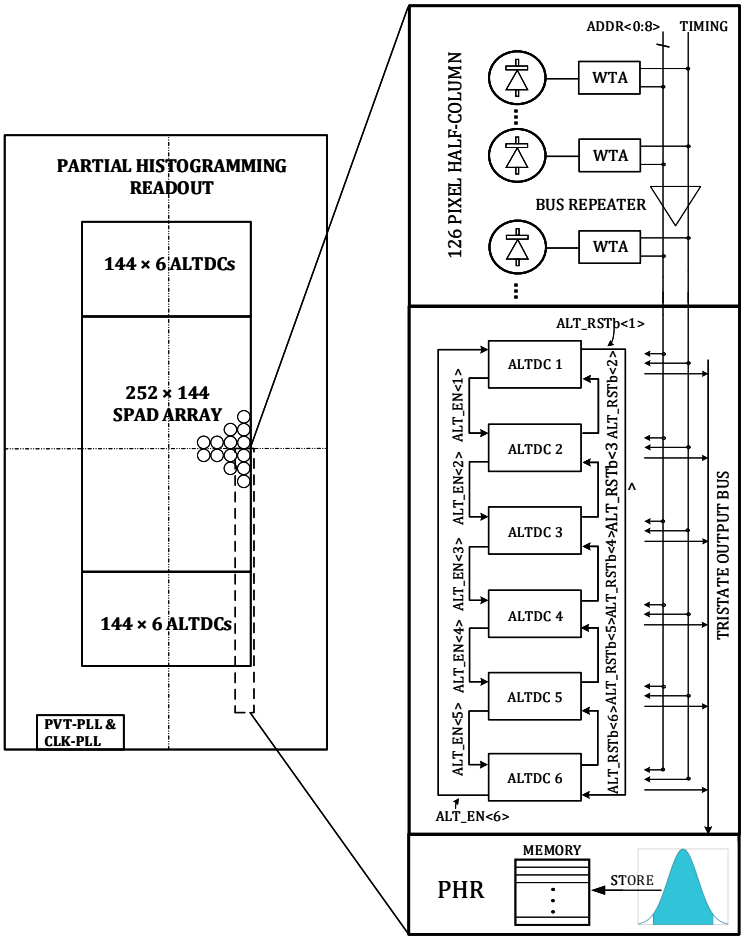


Fig. 6.1 Ocelot architecture

in the bank, which are connected in a daisy chain fashion. Events are distributed to the ALTDCs using a dynamic reallocation approach [12], 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 performed by an open-loop ring-oscillator (RO) TDC based on a dual-clock architecture. PVT compensation of the TDC is performed by a 2.56 GHz PLL whilst the on-chip clocks of 320 MHz and 240 MHz are generated by a separate 960 MHz PLL.

In time-resolved SPAD sensors, digitizing the pixel address and timestamp of every pho-

ton 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 [9], we have implemented a partial histogramming readout (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 is read out from the sensor via 72 160-MHz GPIO pads for a total bandwidth of 11.52 Gbits/s.

6.2.1. ARRAY SCALING FROM PICCOLO TO OCELOT

As discussed in Section 6.1, some concepts from Piccolo are implemented in Ocelot, comprising the pixel, the collision detection bus and dynamic reallocation ALTDC. With the scaling of the array, design of these components have to be optimized, due to the difference in the bus load capacitance, column activity rate and readout scheme.

SCALABLE PIXEL ARRAY

The pixel design in Ocelot is the same as that of Piccolo, which is already discussed in Section 5.2.1. An array of 252 × 144 pixels were implemented in the design, where 126 pixels in each half-column were connected to a collision detection bus. The pixel number of 126 is determined by the collision detection coding method, with the maximum number of possible codes m for a the bus with n lines given by (5.1), where k is the integer closest to $n/2$. For a 9-bit bus, the maximum number of pixels it can accommodate is 126, where each address consists of 5 '1's and 4 '0's.

REPEATERS

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 Ocelot bus repeaters were used to divide the bus into 8 sections where the section closest to the ALTDC array has 14 pixels and the remaining 7 sections have 16 pixels. A single section including bus repeaters is pictured in Fig. 6.12. 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

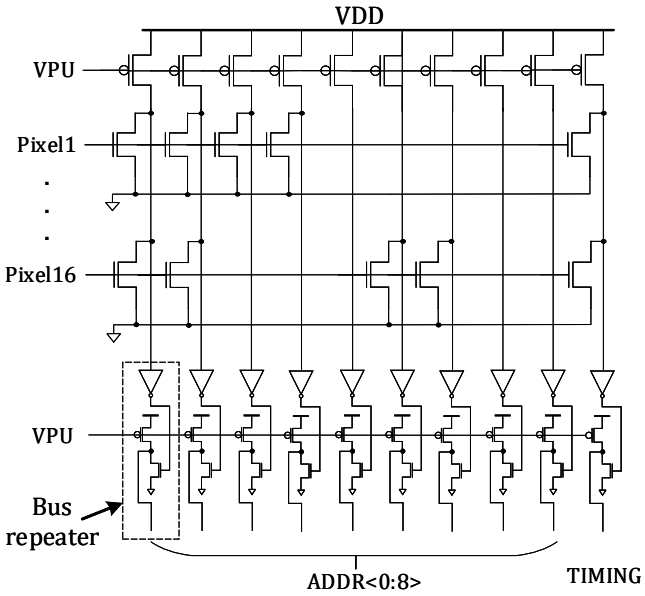


Fig. 6.2 Schematic of the first section of the collision detection bus and repeaters.

8, whilst requiring only a single bus repeater per line. Thus, this method maximizes the fill-factor.

For efficient area use, a small space, which is reserved in each pixel for decoupling capacitors in Piccolo, is used for the placement of 1 bus repeater. By embedding the repeaters into the pixels, no extra silicon area is required in the array implementation, which achieves the same fill factor (28%) as Piccolo at the same pixel pitch (28μm) with a much larger pixel number of 252 × 144. 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 6 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.

ALTDC INTERFACE

Similar with Piccolo, referring to the Section 5.2.3, dynamic reallocating address latches and TDCs (ALTDCs) were employed in Ocelot, where 6 TDCs were shared by 126 pixels in each half-column. In this case, totally 1726 ALTDCs were implemented, which are

shared by 144 columns. The block diagram of ALTDC is presented in Fig. 6.3, where each 4 half-columns were grouped and interfaced to one column selector which reads out events in an event-driven fashion and then writes the data into a FIFO for PHR processing. The ALTDC operation timing diagram can refer to Fig. 5.8, while The reason of column grouping will be explained in Section 6.2.3.

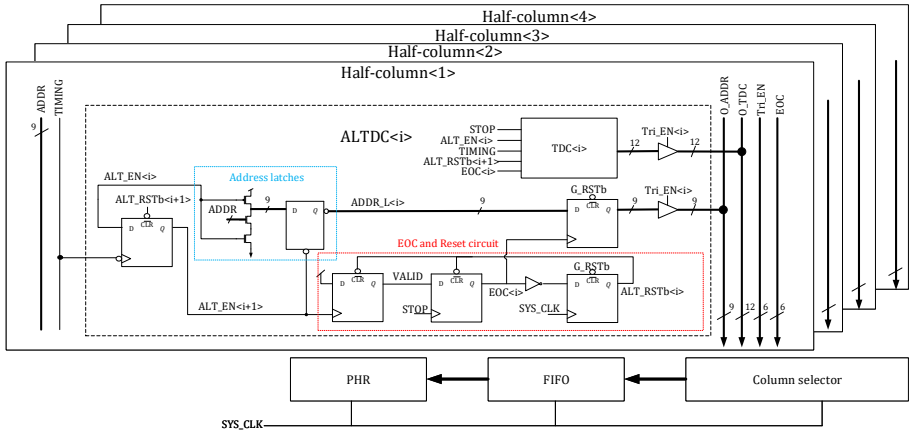


Fig. 6.3 ALTDC block diagram, where each each 4 half-columns were grouped and interfaced with one PHR block.

6.2.2. DUAL-CLOCK TDC

Due to the large number of on-chip TDCs and extensive logic employed in the PHR, power consumption of the TDC is critical. As such, an open-loop ring-oscillator (RO) based architecture is employed to mitigate against the need for distributing multiple phases of a clock [8, 13], across the sensor. However, the PHR constrains the design of the TDC in comparison to conventional RO based approaches [1, 14], because each partial histogram is constructed with data from 6 TDCs, i.e. 6 individual timing histograms. For a given time-of-arrival then, frequency mismatch in the open-loop oscillators will result in a deviation in the codes from the 6 TDCs. For longer measurement periods, these deviations will accumulate and the peaks of the 6 timing histograms disperse in time. This would pose a challenge for the peak location and partial histogramming functions.

To reduce the code dispersion of the TDCs, a dual-clock TDC architecture is implemented in Ocelot, which is based on a 3-stage interpolation approach. The first stage is driven by a global reference clock STOP_HF, working at 320 MHz, leading to a coarse resolution of 3.125 ns. The second stage is based on a local open-loop RO, which oscillates at a frequency of 2.56 GHz, providing a medium resolution of 390.6 ps. The RO

starts the oscillation at the photon detection, and it is stopped by the closest rising edge of STOP_HF, implying a maximum on-time of a single period of STOP_HF. This reduces the TDC mismatch due to the short accumulation time of RO operation. The third stage is carried out by decoding the phases of the RO, achieving a fine resolution of 48.8 ps.

A circuit schematic of the dual-clock TDC and timing diagram are shown in Fig. 6.4 and 6.5, respectively. The TDC is enabled by ALT_EN from the previous slice in the ALTDC chain, T1 in Fig. 6.5. This signal also gates the STOP_HF clock into the block, enabling TDC nodes to be toggled by STOP_HF_INT. At the next rising edge of TIMING, T2, 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 ring-oscillating frequency. A 4-bit counter counts the rising edges of the RO until the first rising edge of STOP_HF_INT, T3. 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, 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 may exceeds 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.

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 6 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.

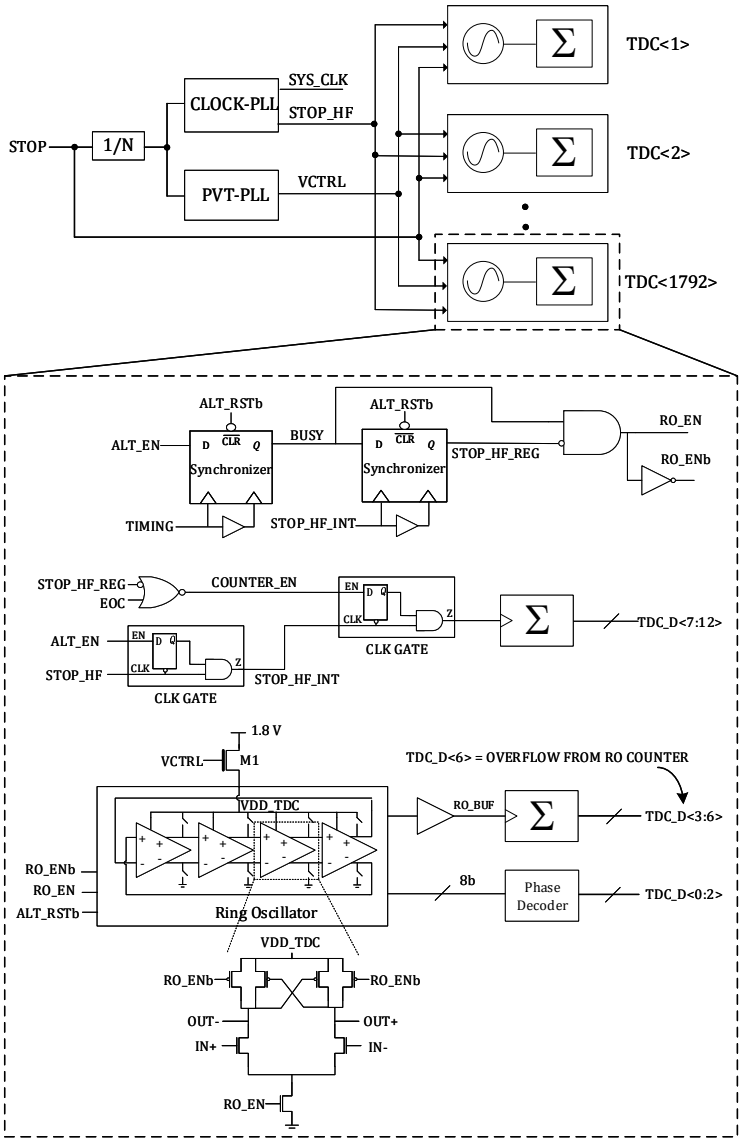


Fig. 6.4 Dual clock TDC architecture and schematic.

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 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

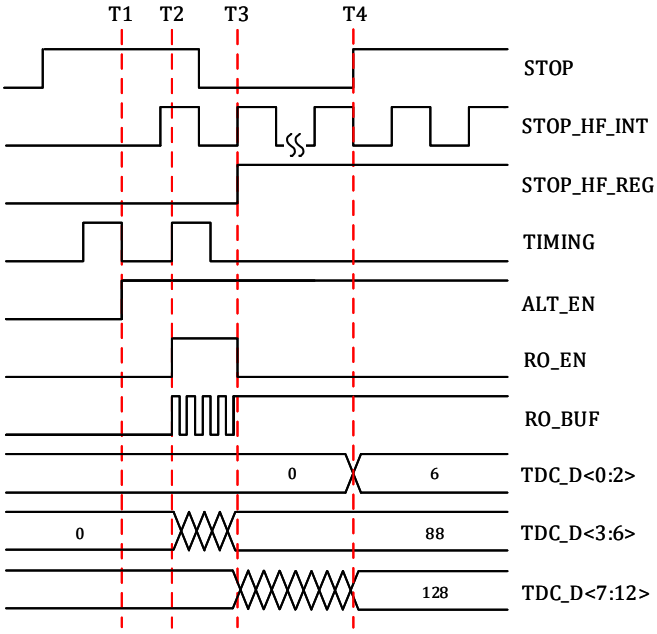


Fig. 6.5 TDC operation timing diagram.

frequencies can be generated with STOP frequencies of 80, 40, 20, 10, 5 and 2.5 MHz, by selecting the appropriate divide ratio. This scheme limits the loop bandwidth of the PLL, in this case a bandwidth of 125 kHz is used, and results in increased accumulation of jitter due to phase noise from the voltage-controlled oscillator. For this reason, a second PLL was designed, CLOCK-PLL in Fig. 6.4, to generate the STOP_HF and SYS_CLK clocks. This relaxes the jitter requirement of the RO in the TDC and PVT-PLL, as this RO will only accumulate jitter when it is enabled in the TDC with a maximum on-time of 3.125 ns. A separate RO was designed for the CLOCK PLL, where the jitter of STOP_HF sums in quadrature with the other components of the system jitter.

6.2.3. 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 is 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 Gbps. 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 range histogramming has been implemented in [9–11], 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 [11]. 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 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 comparison to SRAM based hitogramming [9], histogramming TDCs [10, 11] are even more pronounced due to the use of ripple counters to implement the histogram memory.

In Ocelot, we implemented an on-chip SRAM based histogramming method, which we refer to as the partial histogramming readout (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 partial histogramming (PH). The PD mode detects the histogram peak location for each pixel, whilst PH mode is used to build the partial histogram. 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. 6.6.

In the PHR scheme, the 12-bit TDC code is shorten into 10 bits with 3 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. The peak detection is a 3-step approximation process, where the searching resolution of each step is improved until the peak is located. In each step, the range is subdivided into 8 sections and a histogram is built with the HIST_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 peak detection, photons are accumulated in this histogram considering only the most-significant bits Q<9:7>, thus locating the section peak T1. With a TDC LSB of 48.8 ps, this results in a resolution of 6.2

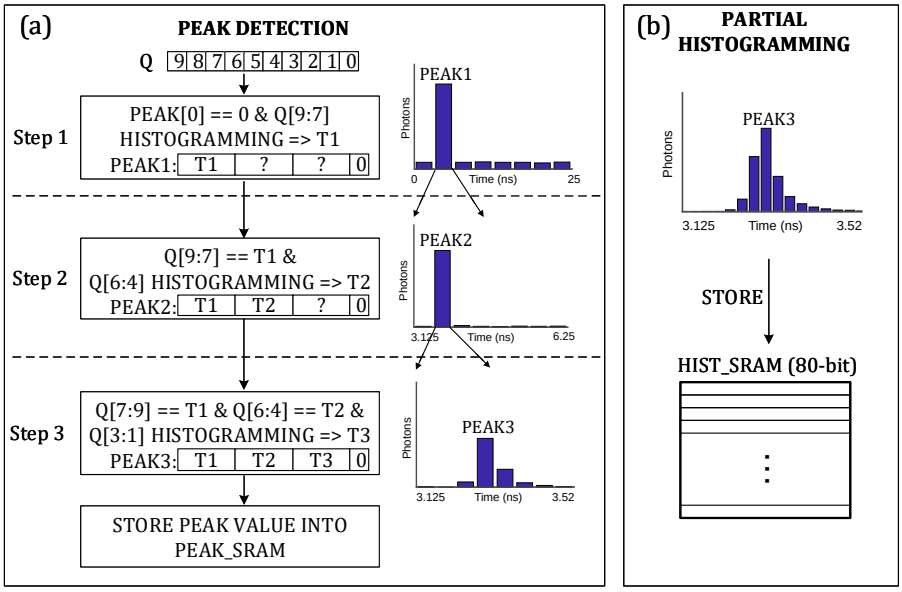


Fig. 6.6 Partial histogramming readout block diagram. (a) Peak detection requires a 3-step successive estimation of histogram peak. (b) Partial histogramming stores a configurable window of 16 bins around the peak.

ns per bin. In the second step, the region T1 is inspected with a resolution of 780 ps considering Q<6:4>, 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<3:1>. The peak is determined with Q<0> as '0' and is stored in the PEAK_SRAM for readout and partial histogramming. In comparison to peak detection based on the phase domain Δ-Σ approach [15], 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 Δ-Σ approach the peaks will be averaged leading to a significant error.

Once the peak is located, the PHR can be operated in partial histogramming 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 whilst the PHR is accumulating events. By combining the in-range partial histogram 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, NIROT,

etc [10, 11].

In the PHR, the PD is the most critical process that decides efficiency of histogramming. However, in order to detect the peak correctly, the background light should be limited due to the shot noise nature of light. Among the 3 steps of peak detection, the first step examines the full TDC range with a limited histogram depth of 10-bit, which is more susceptible to background noise. Therefore the success of peak detection is determined by the first step, in which the full TDC range is divided into 8 sections, with each section combined with 128 TDC bins. If there is only one peak in the histogram which is located in any one of the subsections, photons in other sections are all from the background light and are evenly distributed. The average number of photons in these sections can be represented by *N*. So, the noise floor, which is defined by the average photon count of each TDC bin is *N/128*. However, since the photons follows a poisson distribution, shot noise of $\sqrt{N}$ can be expected. In order to distinguish the peak signal from shot noise, let us assume a convenient peak signal with an amplitude of $3\sqrt{N}$ is required. Assume the timing response follows a gaussian distribution with a fair FWHM jitter of 3 LSB, corresponding to a standard deviation (σ) of 1.27 LSB, the proportion of peak signal bin over the entire response can be obtained at 0.31 from equation (6.1), where $\mu = x = 0$.

$$P(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2 / 2\sigma^2} \tag{6.1}$$

By knowing the peak signal and the noise floor, the signal-to-background noise ratio *SBNR* then can be defined in (6.2).

$$SBNR = 20log_{10}\left(\frac{0.31 * 3\sqrt{N}}{N/128}\right) = 20log_{10}(\frac{119}{\sqrt{N}}) \tag{6.2}$$

$$N + 0.31 * 3\sqrt{N} = 2^b - 1 \tag{6.3}$$

The *N* of 993.6 can be derived from equation (6.3), where *b* is the depth of the histogram bins which is 10-bit in PHR. Therefore, $2^b - 1$ represents the maximum number of photons each bin can accumulate. This leads to a *SBNR* of 11.5 dB, where the peak signal and noise floor are 29.4 and 7.7 respectively. This is the worst case *SBNR* that the peak can be detected with PHR by exploiting the entire depth of the histogram. Whereas, for the full range statistics with per-bin 10-bit histogramming, the requirement to the *SBNR* can be reduced by 128 times, reaching -30.6 dB, since the noise bins will not merge with the signal bins. Besides, the time for peak detection in full range histogramming can be faster that of PHR, as it is a single step process. However, as discussed at the beginning

of this section, the biggest downside in the area occupation prevents its implementation in large format arrays.

For the PHR implementation, 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 4 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 block were implemented, comprising 2.95 Mb HIST_SRAM and 0.37 Mb PEAK_SRAM. The layout of one PHR block is shown in Fig. 6.7, with a dimension of 1.3 mm x 0.94 mm, where the SRAM memory occupies 68% of the area.

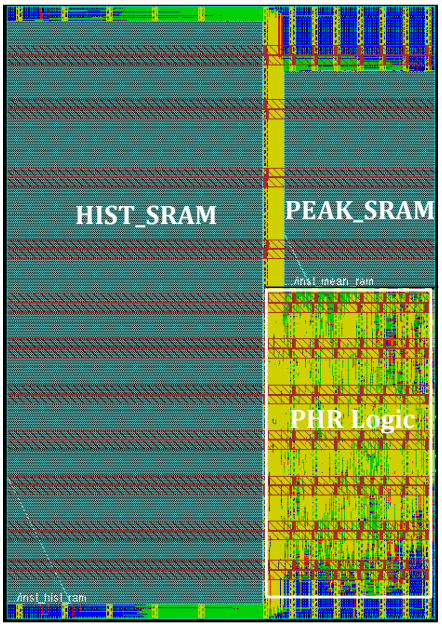


Fig. 6.7 Layout of one PHR block, where 68% of the area is occupied by the SRAM memory.

6.2.4. 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. 6.8. 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 down significantly in more advanced nodes.

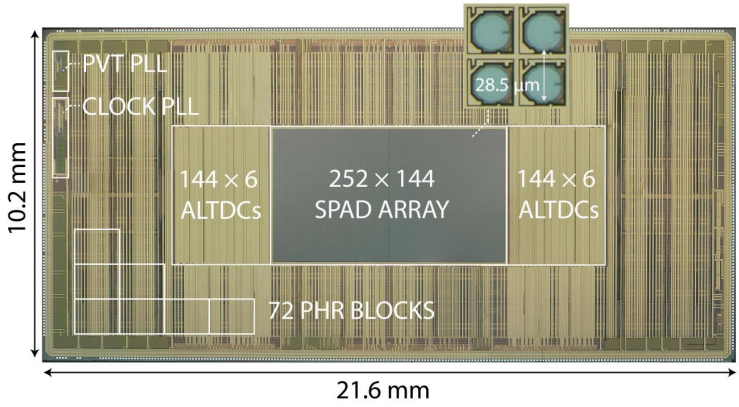
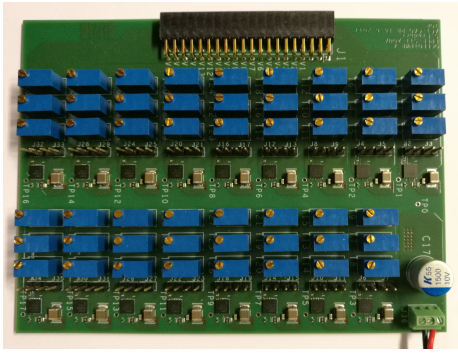
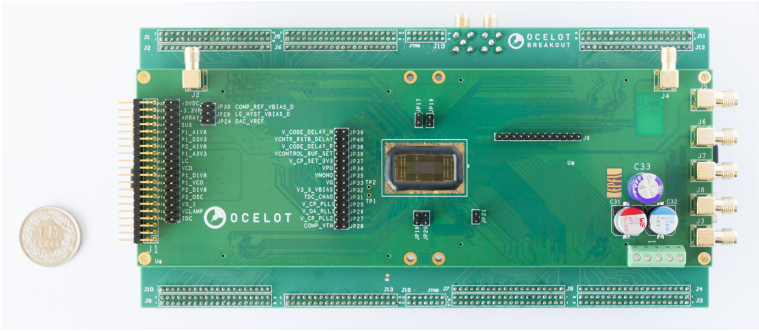


Fig. 6.8 Chip microphotograph with inset of 2 × 2 cluster of pixels.

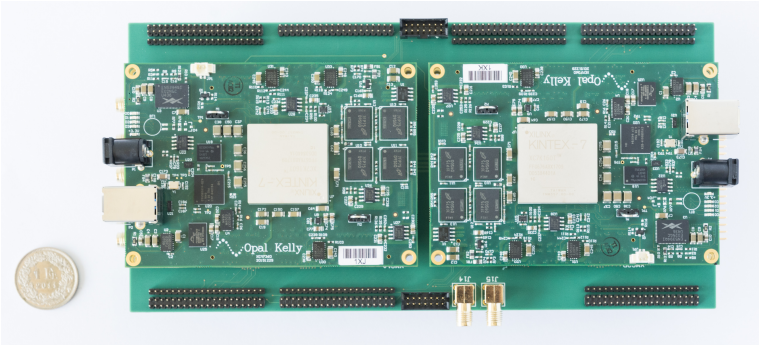
In order to characterize the sensor, a measurement system was designed, Fig. 6.9, comprising 5 printed circuit boards (PCBs), including the mainboard with the sensor mounted as chip-on-board, a power board, a breakout board for signal probing and debugging, and two field-programmable gate array (FPGA) boards (Opal Kelly XEM7360 based on Kintex-7), where each one handles half of the chip. The power board contains a number of voltage regulators (LT3081) to generate all the power signals, except the supplies with high voltage and current, including SPAD cathod, PHR logic and IOs, which are connected to external power supplies. 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.



(a)



(b)



(c)

Fig. 6.9 (a) The power board. (b) Front side of the system, where the sensor is mounted on the mainboard. (c) Back side of the system, where two FPGAs are used for readout and processing. Each one handles half of the sensor. The breakout board is mounted in between the mainboard and FPGA boards for debugging.

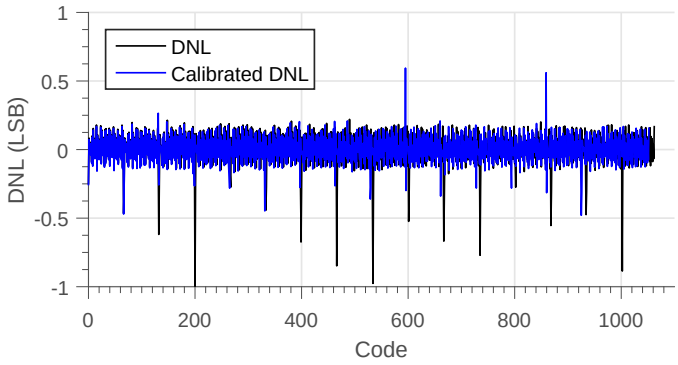
6.3. RESULTS

Since the SPAD architecture in Ocelot is the same with that of Piccolo, pixel characterization can refer to section 5.3. In this section, characterization of the dual-clock TDC, PHR and flash imaging at 2D and 3D will be presented.

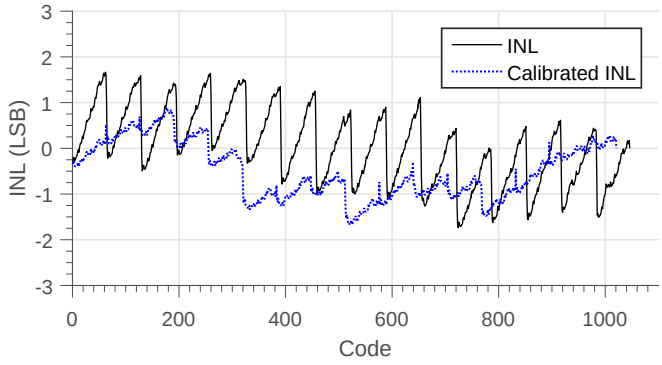
6.3.1. TDC CHARACTERIZATION

As is discussed in the Section 5.3.5, code density approach was used for the TDC non-linearity measurement. The TDCs operate with a nominal LSB of 48.8 ps, where STOP_HF = 320 MHz and VCOs oscillate at 2.56 GHz. The TDC non-linearity was measured with a 20 MHz reference signal, as is shown in Fig. 6.10. From the measurement, a periodic DNL/INL non-linearity error is observed every 64 bins, at the transition between the 4-bit and 6-bit counters. This is due to two factors. Firstly, although the ROs are biased by VCTRL from the PVT-PLL, small frequency offsets are present in the ROs due to random device mismatch. Secondly, 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 +0.6/-0.48 (+0.89/-1.67) LSB after calibration.

With the dynamic reallocation scheme, each photon impinging a SPAD can be detected by any TDC in the half-column. Since the mixed TDC response is used for peak detection and partial histogramming 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. 6.11, where the sensor was illuminated with a short pulsed laser with 40 ps full-width-at-half-maximum (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. 6.11(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. 6.11(b), where excellent uniformity is achieved with the average and standard deviation of 3.03 LSB (148 ps) 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.

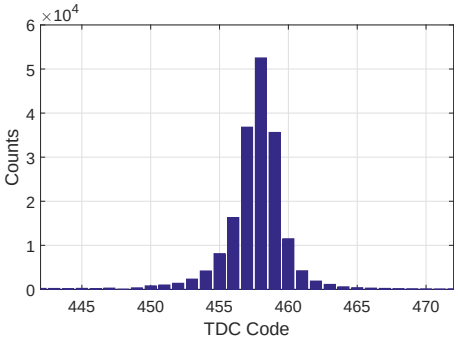


(a)

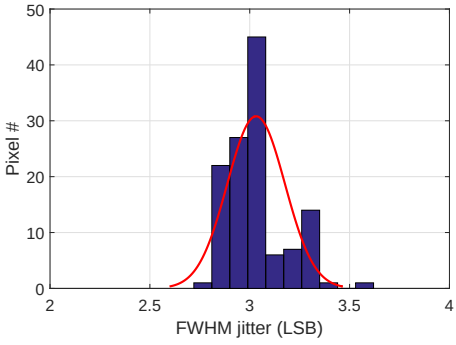


(b)

Fig. 6.10 DNL (a) and INL (b) of the TDC at STOP frequency of 20 MHz. The measurement was carried out by Scott Lindner.



(a)



(b)

Fig. 6.11 (a) Single-shot FWHM jitter of 3.01 LSB with mixed response of 6 TDCs in one half-column. (b) Jitter distribution among 126 pixels in a half-column, a standard deviation of 0.14 LSB was achieved.

6.3.2. PIXEL DELAY OFFSET

With the insertion of bus repeaters in the collision detection bus, event signals will propagate at different delays to the ALTDCs, regarding to the location in the bus. This delay difference has to be measured and calibrated in order to achieve precise TOF measurement at each pixel. The offset was characterized by measuring the photon arrival time of each pixel when illuminating the sensor with a uniformly distributed laser beam. Fig. 6.12 shows the pixel offset from 128 columns of the top-half sensor referring to the pixel in the center of the sensor (row 63, column 77), where an average delay of the bus repeater of 3.65 LSB (178 ps) was obtained. The offset deviation of pixels from the same section at different columns was observed, which is mainly due to the device mismatching in the bus and the TDC nonlinearity.

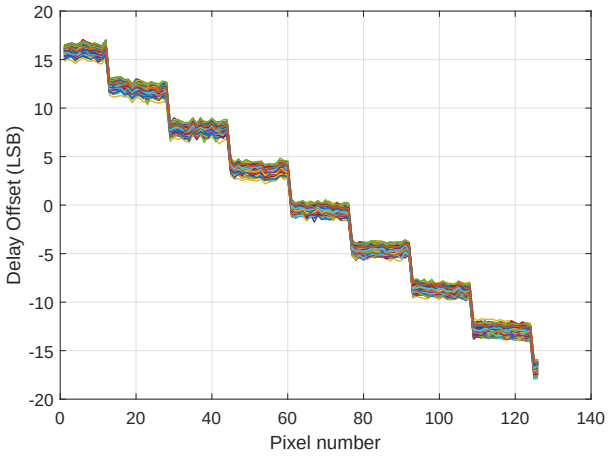


Fig. 6.12 Delay offset of pixels in top-half of the sensor referring the pixel at row 63 and column 77.

6.3.3. 2D INTENSITY IMAGING

In order to enable both the 2D and 3D imaging capability of the sensor, a camera system was built with a 25 mm objective (f/1.5) placed in front of the sensor, achieving a FOV of 20 degree × 40 degree. In 2D imaging mode, the PEAK_SRAMs are configured as 10-bit resolution counters per pixel. The sensor works in global shutter mode with an I/O speed of 160 MHz, leading to the readout time of 32 μs, thus a maximum frame rate of 32 kfps. To achieve both high speed and low light level imaging, an integration time of 1.5 ms was used, achieving a frame rate of 666 fps. An experiment was carried out with a 3-blade fan rotating at 1300 r/min, where the rotation was recorded with half of the chip at illumination levels of 0.1 lux and 10 lux. As is shown in Fig 6.13, although the image

quality at 0.1 lux is poisson-limited, the edge of the blades can still be ascertained thanks to the high PDE and single-photon detection capability.

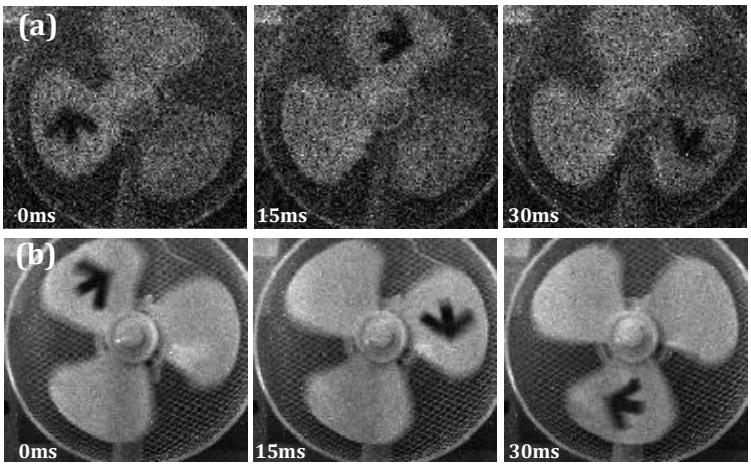
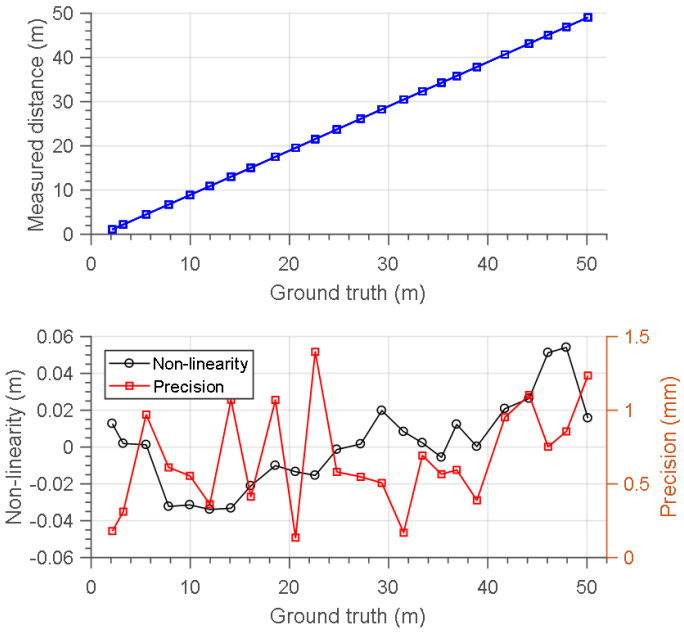


Fig. 6.13 2D movies of a rotating fan at 1300 r/min with background light of (a) 0.1 lux, (b) 10 lux. The gray scale for (a) is 0-30 counts, and 0-300 count for (b).

6.3.4. 3D FLASH IMAGING

Similarly as in Piccolo, time-resolved ranging measurement was performed with the same setup, as is discussed in section 5.3.9. The linearity of the system was characterized and shown in Fig. 6.14, where a subset of 16 × 128 pixels were used for the detection. A 60% reflectivity target was measured up to 50 m, where each distance was measured with a 30k photons histogram for 10 repeated measures, revealing a peak-to-peak nonlinearity and worst-case precision (σ) of 8.8 cm and 1.4 mm respectively, over the entire range. Compared with Piccolo, better precision was achieved due to the improved matching between TDCs.

In order to demonstrate depth imaging with the PHR scheme, including PD and PH processes, a 252 × 128 flash 3D image was acquired at a distance of 1 m with intensity data superimposed, as seen in Fig. 6.15. A diffuser was placed in front of the laser to illuminate the scene uniformly with a 20-degree diverged circular beam. Since the sensor FOV is 20 degree × 40 degree, the measurement was performed in a sequence of 8 exposures, illuminating different sections of the mannequin. Due to the limited laser power, the image was obtained in dark conditions to maximize the SBNR. The profile of cross section A-A' is shown in Fig. 6.15 (b), where 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



(a)

Fig. 6.14 The upper figure shows the measured distance as a function of the actual distance, while the lower one shows the non-linearity and precision 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. A peak-to-peak non-linearity and worst-case precision (σ) of 8.8 cm and 1.4 mm were achieved respectively

averaging the partial histogram of each pixel, which resolves the image with millimetric detail. One example of the partial histogram is shown in Fig. 6.15 (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 265 kcps, where the sensor was illuminated directly by the laser. To avoid high photon collision due to the simultaneous laser shot, only one pixel is enabled. However the data volume is not changed because the entire HIST_RAM has to be read out, which takes 0.256 ms to read off-chip via a 160 MHz GPIO. In order to avoid histogram bin overflow, the histogram was only integrated for 0.5 ms, in which about 130 events were detected by PHR in average. So a photon throughput of 177 kcps was achieved for one pixel. If we assume all the pixels have the same activity, the totaly photon throughput of Ocelot would be 6.4 Gcps. In raw readout mode, every event comprises 27 bits, consisting of 1-bit start flag(constant '0'), 2-bit column number, 10-bit TDC value, 3-bit

TDC number, 9-bit address, 1-bit overflow flag and 1-bit stop flag (constant '1'). If this 6.4G events are read out in raw mode, it would take 14.9 seconds. Therefore, in comparison with raw mode readout, a compression factor of 14.9-to-1 is enabled by PHR scheme. 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.

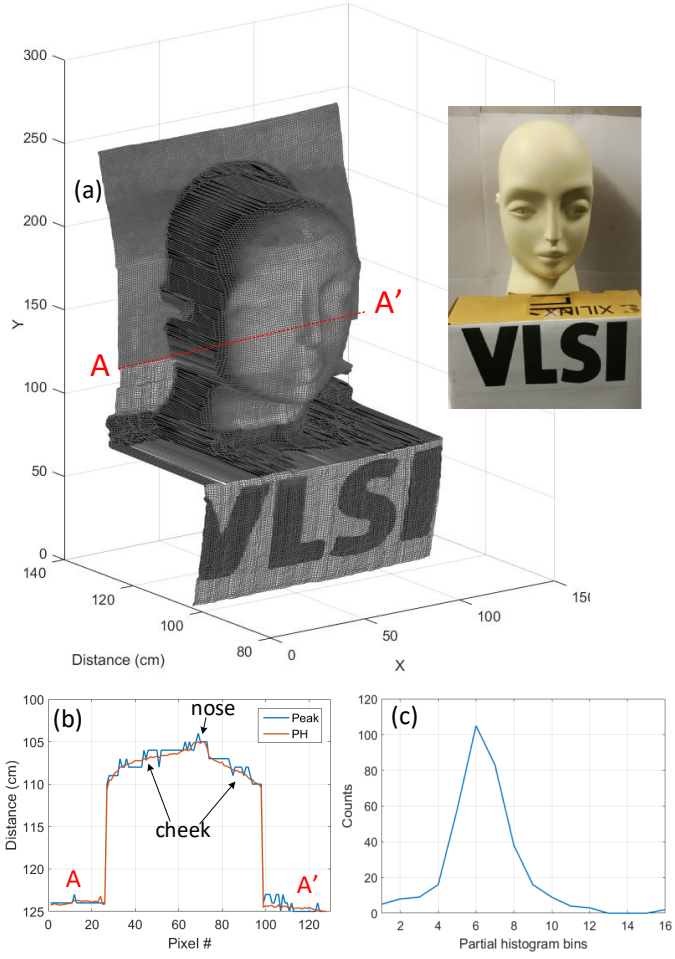


Fig. 6.15 (a) a 252 × 128 3D flash image using PHR scheme with intensity superimposed. Note the insensitivity of depth map from the reflectivity of the writing 'VLSI' in the pedestal. Median filtering with a neighborhood size of 2 × 2 was applied. (b) profile of cross section of A-A', drawn with the peak (coarse resolution) and averaged partial histogram (fine resolution). (c) partial histogram of a pixel on the nose.

Since the time-of-arrival statistics of each pixel are built on-chip, the workload of the

FPGA is reduced dramatically, allowing 3D imaging to be performed in real-time. Fig. 6.16 represents 6 successive frames of a 3D movie acquired at 30 fps at 0.7 m distance with half of the array, in which a hand is clenching and unclenching. The sensor was operated in PHR mode, with peak detection 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.

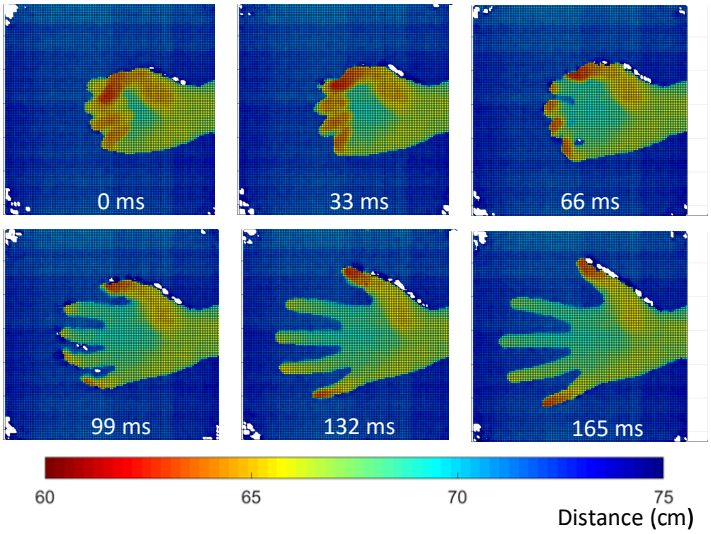


Fig. 6.16 Six successive frames from a 3D movie at 30 fps, in which a hand is clenching and unclenching. The movie has been denoised with median filtering.

6.3.5. STATE-OF-THE-ART COMPARISON

The power consumption of each component in the sensor is detailed in Table 6.1, showing a total measured consumption of 2.54 W with a photon throughput of 156 Mcps in PHR mode, which is equivalent to a detection power of 16.3 nW/photon. As expected, the digital core dissipates the largest proportion of the power, due to the significant logic of the PHR whilst operating at a 240 MHz clock. The second most power hungry block is the ALTDC array, where the TDCs, address latches, 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, 13], the RO based dual-clock architecture in this work 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 6.1: 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	

Table 6.2 summarizes the performance of the sensor in comparison with state-of-the-art time resolved SPAD LiDAR systems. Ocelot 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 [16, 17]. [9, 13] report a better linearity, but with a much lower resolution and a much higher TDC power consumption [13]. In [9], full range histogramming 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 partial histogram 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 [17] 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 SBNR and thus the range. The ranging performance can be significantly improved by employing a high power NIR laser, without affecting other aspects of the sensor.

6.4. CONCLUSION AND DISCUSSION

In this chapter, Ocelot is presented, comprising 252 × 144 SPAD pixels for time-resolved imaging applications, including flash LiDAR. To achieve a fill factor of 28% with a pitch of 28.5 μm, heavy use of resource sharing through a collision detection bus was made. The architecture is highly scalable, and no obvious timing degradation was observed in the scaling up. 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 peak detection from multi-reflections and a compression factor of 14.9-to-1 thanks to partial histogramming. To the best of the authors knowledge, this is the first implementation of fully integrated histogramming on a per-pixel basis for a full 2D 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. 2D images were captured operating the sensor in SPC mode at an optical level of 0.1 lux and a frame rate of 666 fps. 3D images were captured operating the sensor in TCSPC mode from a minimum of 0.7 m at 30 fps using an illumination power as low as 2 mW (average). The maximum ranging operation was 50 m, where a non-linearity 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. Whilst due to the absence of background light suppression techniques, Ocelot is more suitable in low background light level detection.

As is discussed in 5.5, background light suppression can be achieved with the collision detection bus. However, to achieve coincidence detection with a flash imaging sensor, a in-pixel TDC architecture and digital SiPM type of pixel would be required, such as [16] where each pixel consists of 8 SPADs sharing one TDC. This would reduce the fill factor and array size, due to the large pixel pitch. A better solution would be employing the 3D stacking technology, where the SPAD array is implemented on the top chip and the circuits on the bottom chip. The two chips are connected via micro-bumps or through-silicon-vias (TSVs). In this case, different technologies can be selected for the implementation of the two chips to achieve the optimal performance of each design.

For the partial histogramming, except for the requirement of the SNBR, one assumption that the background light is distributed uniformly in the histogram, may not be satisfied in some extreme situations, e.g. fog, dust. In these environments, instead of a flatten background histogram, an exponential distribution is expected [18, 19], which may results in wrong peak detection and histogramming. So a full range histogram and data post processing would be preferable.

Table 6.2: Ocelot state-of-the-art comparison

Parameters	Ocelot	[17]	[16]	[9]	[13]
		Sensor characteristics			
Technology	180 nm	45/65 nm	150 nm	180 nm	350 nm
Integrated histogramming	Per-pixel, partial hist	N/A	N/A	Per-pixel, full hist	N/A
Pixel array	252×144	16×8	64×64(5)	32×1	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
DCR(cps) @ Vex	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 LSB(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
DNL(LSB)	+0.6/-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
Illum. wavelength(nm)	637	532	470	870	750
Illum. frequency(MHz)	40	1	N/A	0.133	N/A
Illum. mean power(mW)	2	6	N/A	21	90
Illum. peak power(W)	0.5	N/A	N/A	39.5(4)	N/A
Max. distance(m)	50(2)	150 - 430	367 - 5862(6)	128	8
Imaging 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(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(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(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;

REFERENCES

[1] C. Veerappan, J. Richardson, R. Walker, D.-u. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, *A 160×128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter,* ISSCC , 312 (2011).

[2] L. Gasparini, M. Zarghami, H. Xu, L. Parmesan, M. M. Garcia, M. Unternahrer, B. Bessire, A. Stefanov, D. Stoppa, and M. Perenzoni, *A 32×32-pixel time-resolved single-photon image sensor with 44.64µm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800kHz observation rate for quantum physics applications,* ISSCC , 98 (2018).

[3] A. C. Ulku, C. Bruschini, X. Michalet, S. Weiss, and E. Charbon, *A 512×512 SPAD Image Sensor with Built-In Gating for Phasor Based Real-Time siFLIM,* International Image Sensor Workshop , 1 (2017).

[4] I. Gyongy, N. Calder, A. Davies, N. A. Dutton, P. Dalgarno, R. Duncan, C. Rickman, and R. K. Henderson, *256×256, 100kfps, 61% Fill-factor time-resolved SPAD image sensor for microscopy applications,* Technical Digest - International Electron Devices Meeting, IEDM , 8.2.1 (2017).

[5] I. M. Antolovic, S. Burri, C. Bruschini, R. A. Hoebe, and E. Charbon, *SPAD imagers for super resolution localization microscopy enable analysis of fast fluorophore blinking,* Scientific Reports **7**, 1 (2017).

[6] E. Panina, L. Pancheri, G. F. Dalla Betta, N. Massari, and D. Stoppa, *Compact CMOS analog counter for SPAD pixel arrays,* IEEE Transactions on Circuits and Systems II: Express Briefs **61**, 214 (2014).

[7] M. Perenzoni, N. Massari, D. Perenzoni, L. Gasparini, and D. Stoppa, *A 160 × 120 pixel analog-counting single-photon imager with time-gating and self-referenced column-parallel A/D conversion for fluorescence lifetime imaging,* IEEE Journal of Solid-State Circuits **51**, 155 (2016).

[8] 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 Journal of Solid-State Circuits **49**, 867 (2014).

[9] 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 Journal of Solid-State Circuits **49**, 315 (2014).

[10] N. A. W. Dutton, S. Gnecchi, L. Parmesan, A. J. Holmes, B. Rae, L. A. Grant, and R. K. Henderson, *A time-correlated single-photon-counting sensor with 14GS/S histogramming time-to-digital converter,* IEEE International Solid-State Circuits Conference - Digest of Technical Papers **58**, 204 (2015).

[11] A. T. Erdogan, R. Walker, N. Finlayson, N. Krstajic, G. O. Williams, and R. K. Henderson, *A 16.5 giga events/s 1024 × 8 SPAD line sensor with per-pixel zoomable 50ps-6.4ns/bin histogramming TDC,* IEEE Symposium on VLSI Circuits, Digest of Technical Papers , C292 (2017).

[12] 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 180nm CMOS for Near Infrared Optical Tomography,* International Image Sensor Workshop , 86 (2017).

[13] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. Dalla Mora, D. Contini, D. Durini, S. Weyers, and W. Brockherde, *CMOS imager with 1024 SPADs and TDCS for single-photon timing and 3-D time-of-flight,* IEEE Journal on Selected Topics in Quantum Electronics **20** (2014), 10.1109/JSTQE.2014.2342197.

[14] 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 Journal of Solid-State Circuits **50**, 2406 (2015).

[15] 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 0.13µm CMOS Imaging Technology,* IEEE International Solid-State Circuits Conference - Digest of Technical Papers **21**, 1020 (2011).

[16] M. Perenzoni, D. Perenzoni, and D. Stoppa, *A 64×64-pixel digital silicon photomultiplier direct ToF sensor with 100Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6km for spacecraft navigation and landing,* 2016 IEEE International Solid-State Circuits Conference (ISSCC) **52**, 118 (2016).

[17] A. R. Ximenes, P. Padmanabhan, M.-j. Lee, Y. Yamashita, D. N. Yaung, and E. Charbon, *A 256×256 45/65nm 3D-Stacked SPAD-Based Direct TOF Image Sensor for LiDAR Applications with Optical Polar Modulation for up to 18.6dB Interference Suppression,* ISSCC , 27 (2018).

[18] T. E. Laux and C.-I. Chen, *3D flash LIDAR vision systems for imaging in degraded visual environments,* Proc. SPIE, Degraded Visual Environments: Enhanced, Synthetic, and External Vision Solutions **9087**, 908704 (2014).

[19] I. Ashraf and Y. Park, *Effects of fog attenuation on LIDAR data in urban environment,* Proc. SPIE, Smart Photonic and Optoelectronic Integrated Circuits **1053623**, 77 (2018).

7

Conclusion

SPADs, with single photon counting capability and extremely high time resolution, have drawn great attention in the past decade. The primary goal of this thesis was to develop time-resolved SPAD sensors for high resolution dTOF imaging. In order to meet the requirements, including high fill factor and PDP, easy scalability, high photon throughput, high TDC resolution and uniformity, and low power consumption, efforts were focused on the investigation of the sensor architecture and system development. This goal is achieved in two steps with the development of two sensors: Piccolo (Chapter 5) and Ocelot (Chapter 6).

In Chapter 5, high excess bias as well as PDP was achieved with a cascoded passive quenching circuit. To improve the fill factor, instead of per-pixel TDC architecture, a new TDC sharing architecture was introduced. By placing the TDCs outside of the pixel array, the pixel fill factor is greatly increased. Besides, each pixel in a column is connected to a collision detection bus via a WTA circuit, which enables the recognition of collision events among different pixels while also simplifying the readout of data from the sensor with an event-driven approach. A 32 × 32 SPAD sensor, referred to as Piccolo, was implemented in a 180 nm CMOS technology. The sensor was fully characterized and demonstrated with a scaner-based imaging system, where depth imaging upto 10 m at 6 frames/s with a resolution of 64 × 64 has been achieved with a limited optical power of 2 mW and 500 mW in average and peak, respectively.

Based on the design of Piccolo, a larger sensor, referred as Ocelot, with 252 × 144 SPAD pixels was developed and described in Chapter 6. One of the inevitable challenges is to scale up the array size without impacting other pixel parameters, such as the fill factor, bus dead time, etc. This is achieved by breaking a long collision detect bus into several sections with bus repeaters, which allows the reduction in area occupation of the collision detection bus while maintaining the same fill factor and a narrow bus dead time. With this scheme, it is theoretically possible to scale the array architecture up to infinite resolution, e.g. megepixel. On the other hand, more TDCs are required with increase in the array size, giving rise to challenges in the TDC uniformity and power consumption. Therefore, a dual-clock TDC architecture is proposed, which uses a globally distributed low-speed clock as a reference for the coarse counter whilst reaching a high resolution via a local high-speed RO with phase interpolation. Due to the constrained operating range of the RO, a low power consumption was achieved. Meanwhile, all the TDCs are synchronized to the reference clock, leading to a good uniformity. On the other hand, for high resolution SPAD sensors, the readout bandwidth is always a big challenge due to the large volume of data generated by timestamping photons. To reduce the bandwidth requirement, a PHR scheme was implemented to compress the raw TOF data into

a partial histogram, which achieves a 14.9-to-1 data compression ratio. To the best of the author's knowledge, Ocelot is the first time-resolved SPAD sensor which includes a per-pixel basis on-chip histogramming for a large 2D array. To verify the sensor, an imaging system was built and demonstrated in a flash mode, where depth images were acquired at 30 fps from 0.7 m distance with an illumination power as low as 2 mW in average.

The other goal of this thesis was to develop a CCD-based ultra-high speed camera targeting at a frame rate of 1 Gfps, which is referred to as Nanosis in Chapter 4. In order to reduce the electron transmit and transfer time, a pixel structure with multi-collection gates was implemented in a 130 nm BSI technology, which provides the probability of reaching 1 Gfps at device level. Meanwhile, to drive all the pixels simultaneously, a 3D stacking architecture was proposed. The sensor was implemented on the top chip, while an array of drivers were implemented on the bottom chip. Therefore, each driver only drives a subset of 32 pixels with reduced parasitics to achieve a high imaging rate. Pixel-wise connection between the top and bottom chips was achieved with 3D bump stacking at a fine pitch of 18 μm. In the measurement, each individual chip was characterized successfully. However, the stacked sensor failed in the functionality due to the failures in the stacking technology. Extensive exploration is currently being carried out on the stacking technology, and new results can be expected in the near future.

As a conclusion, the two SPAD sensors presented in this thesis provide a solution for time-resolved and high resolution imaging. Compared to per-pixel TDC architectures with raw data readout, the proposed TDC sharing architecture and PHR scheme hold the advantages in array scalability, pixel fill factor, photon throughput and power consumption, which also can be applied in the sensor development for applications such as bio-imaging, robotics, etc. For Nanosis, although failures in the connection, it is the first trial with a single solid state imager aiming 1 Gfps. If the stacking problem is solved, we believe many exciting results can be obtained. Moreover, the concepts and ideas involved in Nanosis also can be useful for the future ultra-high speed imager development.

SUMMARY AND PERSPECTIVE

In conventional applications, such as bio-imaging and microscopy, SPAD is typically used as a single-photon counter. However, this advantage has been challenged by other photon-counting technologies, especially from CMOS-based QIS. Comparatively, apart from single-photon counting capability, QIS is superior to SPAD in terms of intrinsic multi-photon counting capability, quantum efficiency, dark noise, pixel size and fill factor. All these features indicate a low cost and high resolution photon counting imager can be built with QIS, which can be a great competition to SPADs. Moreover, QIS has been demonstrated with 1 Mjot array, 0.175e- rms read noise and 1000 fps at less than 20 mW power consumption.

Nevertheless, single-photon counting is not the only feature of SPAD, a unique property of SPAD is the picosecond-level time resolution, which is the highest time resolution that can be achieved with a CMOS image sensor, thus enabling accurate depth detection in short-to-long ranges. On the other hand, CMOS compatibility enables both the SPAD sensor and circuits to be implemented on the same chip, which dramatically reduces the cost of the imaging system. Therefore, SPADs have drawn great attention in the past few years in many applications ranging from single point proximity sensor to automotive flash/scan LiDAR. However, along with these advantages, it is necessary to clarify the disadvantages and trade-offs between different parameters, including:

- PDP in SPAD is lower than the equivalent QE in CIS or APD, especially at the near-infrared wavelength which is the most widely used band in LiDAR applications. This is particularly important for long range detection, since the reflected signal typically degrades exponentially with the distance.

- DCR in SPAD is also higher than the equivalent dark current in CIS, and it increases exponentially with the excess bias and the temperature. Moreover, the SPAD DCR distribution has a non-uniformity challenge. In Piccolo, even though the median DCR is 114 cps and only 6% of SPADs have a DCR below 1 kcps, the maximum DCR can be two orders of magnitude higher than the median DCR. Therefore, it is required to disable these hot pixels to prevent the disturbance to the entire system. However, for n-on-p type SPAD which benefits from a higher PDP, challenges exist

in disabling the SPAD since a poly resistor is normally used for quenching and recharge.

- Dead time, which is a certain time period that the SPAD is insensitive to light, leads to a maximum dynamic range a SPAD can achieve. Combining with the single photon triggering ability, it is challenging to operate a SPAD sensor in high ambient light environment. The dynamic range can be improved by reducing the dead time, but at the risk of increased afterpulsing. Besides, this problem also can be addressed by grouping a cluster of SPADs as one macro pixel, in which the optical power can be spread over the whole pixel to reduce the photon rate of each SPAD, but at the cost of reduced resolution.

- To prevent edge breakdown, a guard ring is normally required, which basically determines the maximum fill factor of a SPAD. Whilst compared to the APS in CIS where typically 3 or 4 transistors are included, more transistors are required in a SPAD pixel, which further decreases the fill factor. Even though fill factor can be improved with microlens, a price to pay is the cost.

- Since TDCs can be triggered by a single photon, a high TDC activity and data rate can be generated when operating the whole SPAD array in parallel. Addressing the high number of generated events will require a large sensor area and consume high power, e.g. an area of 21.6 mm × 10.2 mm was occupied by Ocelot with a power consumption of 2.54 W. If we consider the SPAD as a photon-to-digital converter, almost all the sensor circuits can be implemented in digital domain. Due to this digital compatibility, circuit chips designed with advanced processes can directly interface with SPADs via stacking technologies, which can dramatically reduce the sensor area overhead and power consumption. However, the stacking technologies are normally very expensive.

Therefore, before designing a SPAD sensor, all these factors should be taken into account. While from the author's point of view, the dominant factor is the ambient light level, which determines the sensor architecture. For example, in low ambient light conditions, a TDC sharing architecture could be applicable due to the low photon rate, e.g. architectures of Piccolo and Ocelot. However, when the sensor working environment is in high ambient light environment, such as in automotive applications, macro pixels consisting of multiple SPADs are required to extend dynamic range. In the meantime, each macro pixel may be shared by multiple TDCs to ensure that true signal photons being timestamped. Spatial and temporal correlation techniques could also be applied to improve the SNR of detection. Besides, on-chip data processing, such as SRAM-based histogramming, could be required to reduce the bandwidth requirement due to the high

photon throughput. Nevertheless, when implementing these functionalities, a large silicon area would be required, implying only a limited format size, e.g. single pixel, line format or a small pixel array, could be feasible when using matured FSI technologies. In large pixel formats, BSI and 3D stacking would be necessary for the implementation of high resolution SPAD sensors.

From the author's point of view, for a SPAD sensor every photon is useful, depending on how we use it. Therefore, to improve the imaging performance, efforts should not only be spent on the hardware, including illumination source, optics and sensor design, but also in the algorithms. These algorithms may include depth calculation, resolution extension, frame interpolation, image denoising and reconstruction. Moreover, since SPADs give every photon a meaning, massive information will be generated during imaging. To handle this high volume data, AI-based algorithms could be good option, given that they are vastly exploited in solving big-data challenges. As an assumption, AI has massive potential in boosting the SPAD sensor development process.

SAMENVATTING

In conventionele toepassingen, zoals bio-beeldvorming en microscopie, wordt een SPAD meestal gebruikt als een teller van enkele fotonen. Dit voordeel is echter in twijfel getrokken door andere foton-teltechnologieën, vooral door op CMOS gebaseerde QIS. Behalve voor het enkele foton-telvermogen is QIS, vergeleken met SPAD, beter in termen van intrinsieke multi-foton-telmogelijkheden, kwantumefficiëntie, donkere ruis, pixelgrootte en vulfactor. Al deze kenmerken wijzen erop dat een foton-telcamera met een lage kostprijs en een hoge resolutie kan worden gebouwd met QIS, wat een grote concurrent voor SPAD's kan zijn. Bovendien is de QIS ontwikkeld door Prof. Fossum en Gigajot Inc. gedemonstreerd met 1 Mjot array, 0.117 elektron effectieve uitleesruis en 1000 beelden per seconde bij een energieverbruik van minder dan 20 mW, wat een grote zorg kan zijn voor de SPAD-gemeenschap.

Niettemin is het tellen van enkele fotonen niet het enige kenmerk van SPAD; een uniek kenmerk van SPAD is de picoseconde-niveau tijdswaarneming , wat de hoogste tijdresolutie is die bereikt kan worden met een CMOS-beeldsensor, waardoor een nauwkeurige afstandsdetectie mogelijk is in het korte-tot-lange afstandsbereik. Aan de andere kant zorgt compatibiliteit met CMOS ervoor dat zowel de SPAD-sensor als de elektrische schakelingen op dezelfde chip kunnen worden geïmplementeerd, waardoor de kosten van het beeldvormingssysteem drastisch worden verlaagd. Om deze reden hebben SPAD's vanwege de groeiende toename van 3D-weergave toepassingen de afgelopen jaren veel aandacht getrokken in veel toepassingen, variërend van een enkel-punts-bewegingssensor tot flits/scan LiDAR voor de automobielindustrie. Naast deze voordelen is het echter noodzakelijk om de nadelen en de afwegingen tussen de verschillende parameters te verduidelijken, waaronder:

- Fotondetectiekans (PDP) in SPAD is lager dan de equivalente kwantumefficiëntie (QE) in CIS of APD, vooral bij de nabij-infrarode golflengte die de meest gebruikte band is in LiDAR-toepassingen. Dit is vooral belangrijk voor langeafstandsdetectie, aangezien het gereflecteerde signaal typisch exponentieel afneemt met de afstand.
- Donkere stroom (DCR) in SPAD is ook hoger dan de equivalente donkere stroom in CIS en neemt exponentieel toe met de aangeboden excessieve spanning en de

temperatuur. Bovendien vormt de niet-uniforme verdeling van de SPAD DCR een uitdaging. In Piccolo, hoewel de mediaan van de DCR 114 cps is en slechts 6% van de SPAD's een DCR van minder dan 1 kcps heeft, kan de maximale DCR twee orden van grootte hoger zijn dan de mediaan van de DCR. Daarom is het vereist om deze zogenoemde 'hete'-pixels uit te schakelen om verstoring van het gehele systeem te voorkomen. Voor SPAD's van het n-op-p-type, die profiteren van een hogere PDP, bestaan er echter uitdagingen bij het uitschakelen van de SPAD, aangezien nomaliter een polyweerstand gebruikt wordt voor het uitdoven en herladen.

- Dode tijd, wat een bepaalde tijdsperiode is waarin de SPAD ongevoelig is voor licht, leidt tot een maximaal dynamisch bereik dat een SPAD kan bereiken. In combinatie met het enkele foton-telvermogen is het een uitdaging om een SPAD-sensor te gebruiken in omgevingen met veel omgevingslicht. Het dynamisch bereik kan worden verbeterd door de dode tijd te verkorten, maar met het risico van verhoogd na-pulsen. Bovendien kan dit probleem ook worden aangepakt door een cluster van SPAD's als ëë n macropixel te groeperen, waarbij het optische vermogen over de gehele pixel kan worden gespreid om het fotontempo van elke SPAD te verminderen, maar ten koste van een verminderde resolutie.

- Om randdoorslag te voorkomen, is gewoonlijk een afschermring vereist, die in feite de maximale vulfactor van een SPAD bepaalt. In vergelijking met de APS in CIS, waar doorgaans 3 of 4 transistoren zijn opgenomen, zijn er meer transistoren nodig in een SPAD-pixel, waardoor de vulfactor verder wordt verlaagd. Hoewel de vulfactor met een micro-lens kan worden verbeterd, zijn de verhoogde kosten de prijs die betaalt moet worden.

- Aangezien TDC's kunnen worden geactiveerd door een enkel foton, kunnen een hoge TDC-activiteit en datasnelheid worden gegenereerd wanneer de hele SPAD-array in parallel wordt gebruikt. Om het hoge aantal gegenereerde gebeurtenissen te adresseren, is een groot sensorgebied nodig en een hoog vermogen, bijv. een gebied van 21.6 mm × 10.2 mm werd ingenomen door Ocelot met een vermogensverbruik van 2.54 W. Als we de SPAD beschouwen als een foton-naar- digitaalomzetter, kunnen bijna alle sensorschakelingen in het digitale domein worden geïmplementeerd. Vanwege deze digitale compatibiliteit kunnen chips die zijn ontworpen in geavanceerde processen, direct worden gekoppeld aan SPAD's via stapeltechnologieën, waardoor de overhead van de sensor en het stroomverbruik drastisch kunnen worden verminderd. De stapeltechnologieën zijn echter doorgaans erg duur en beperkt in aantallen.

Daarom moet er vöö r het ontwerpen van een SPAD-sensor rekening worden gehouden

met al deze factoren. Vanuit het oogpunt van de auteur is de dominante factor het omgevingslichtniveau, dat de sensorarchitectuur bepaalt. Bijvoorbeeld, bij omstandigheden met weinig omgevingslicht kan een TDC-delende architectuur toepasbaar zijn vanwege de lage fotonsnelheid, bijv. architecturen van Piccolo en Ocelot. Wanneer de sensor zich echter in omgevingen met veel omgevingslicht bevindt, zoals automobieltoepassingen, zijn macropixels bestaande uit meerdere SPAD's vereist om het dynamische bereik te vergroten. In de tussentijd kan elke macro-pixel gedeeld worden door meerdere TDC's om te verzekeren dat alle signaalfotonen tijdstempels hebben. Ruimtelijke en temporele correlatietechnieken zouden ook kunnen worden toegepast om het ruisniveau (SNR) van de detectie te verbeteren. Bovendien kan de verwerking van de gegevens op de chip, zoals met op SRAM-gebaseerde histogrammen, vereist zijn om de bandbreedtevereisten als gevolg van de hoge fotondoorvoersnelheid te verminderen. Desalniettemin, wanneer deze functionaliteiten worden geÏmplementeerd, zou een groot siliciumoppervlak vereist zijn, wat een beperkte omvang impliceert voor bijv. ëë n pixel, lijnformaat of een kleine pixelarray, wat haalbaar zou zijn bij gebruik van volwassen FSI-technologieën. In grote pixelindelingen zouden BSI en 3D-stapeling nodig zijn voor de implementatie van SPAD-sensoren met een hoge resolutie.

Vanuit het oogpunt van de auteur is voor een SPAD-sensor elk foton nuttig, afhankelijk van hoe we het gebruiken. Om de beeldprestaties te verbeteren, moeten inspanningen dus niet alleen worden besteed aan de hardware, inclusief verlichtingsbron, optica en sensorontwerp, maar ook aan de algoritmen. Deze algoritmen kunnen zijn: dieptebepaling, resolutie-uitbreiding, frame-interpolatie, verlaging van de beeldruis en reconstructie. Bovendien, aangezien SPAD's betekenis geven aan elk foton, zal tijdens de beeldvorming enorme hoeveelheden informatie worden gegenereerd. Om deze gegevens met hoge aantallen te verwerken, kunnen op kunstmatige intelligentie-gebaseerde algoritmen een goede optie zijn, aangezien deze vooral worden gebruikt bij het oplossen van vraagstukken met grote hoeveelheden gegevens. Het wordt aangenomen dat kunstmatige intelligentie een enorm potentieel heeft in het stimuleren van het ontwikkelingsproces van de SPAD-sensor.

ACKNOWLEDGEMENTS

Time flies. As I look back on my PhD time, it'd not be possible to make this journey without the help, support and encouragement of many people. Taking this opportunity, I'd like to express my appreciation to them.

First of all, I'd like to thank my supervisor Prof. Edoardo Charbon for giving me the chance to work on such an exciting project that I am really interested in. Regular meetings with him guided me to correct directions in the research. One impressive thing I remember was the kick-off meeting of Piccolo and Ocelot. He reserved a quiet and beautiful place outside the Delft and invited everyone who was involved in this tape-out to participate. It was a long and fruitful day that we went over every aspect of the design, which built a solid foundation for the development of these two sensors. His profound knowledge and rigorous attitude played a crucial role during this process. In the daily life, he always joined the lunch together with us while bringing a croissant for share, which created a very pleasant atmosphere and tightened the group as a family. All in all, I'd like to express my sincere appreciation to Prof. Charbon for his supervision, encouragement and great help in my research and many other aspects.

I'd like to thank Prof. Goji Etoh for the guidance in the collaborative project, Nanosis. He was very patient and explained me many details about high-speed sensors. His hard-working and dedication in high-speed imaging set a model to me. This project also provided an opportunity to collaborate with excellent people. I'd appreciate Paul Goetschalckx, Luc Haspeslagh, Philippe Soussan, Bivragh Majeed and Dr. Piet De Moor for the technology development; Dr. Quang nguyen anh, Iguchi-san, Hayashi-san and Mitsui-san for the help in the measurement. Especially, I thank Dr. Son Dao for the collaboration in the sensor development.

There were many happy moments with Dr. Myung-Jae Lee, Dr. Masahiro Akiyama and Dr. Ivan Michel Antolovic, who were my officemates. I'd like to thank MJ for sharing his expertise in SPAD devices, measurement setup and the beer selection. We had a nice collaboration in the SiPM project. Thanks to Masahiro for the warm dinner invitation at his place in the new year of 2014 which was my first new year spent in Delft. Michel started his PhD just about one month before me. I'd like to thank him for his great com-

pany and deep discussion on SPAD pixels, FLIM, microscopy, QIS as well as some social topics such as drugs, Taiwan issue, communism, democracy, etc. He was the witness of all my three times drunk, and thanks to his encouragement in drinking.

I met Esteban Venialgo firstly during my PhD interview. It was the first time to discuss Dutch food during the lunch with him in the canteen. Obviously, this discussion lasts and becomes a precious memory. Also we had nice experience in the soccer table. Dr. Chockalingam Veerappan shared many fundamental principles about SPAD design. His kindness and honesty impressed me more than once. Augusto Carimato was the source of laugh and humor, where his laugh even can be heard at the end of a 50m-long corridor. Thanks to the jokes and the optimistic attitude, I always got positive impact from him. I'd like to thank Dr. Pengfei Sun for sharing knowledge about the SPAD process and SOI technology. He is the guy who can offer humor and surprise. Augusto Ximenes was my last officemate. We had a lot of interesting discussion about LiDAR and other SPAD-based applications.

I'd like to express my special appreciation to two online colleagues: Scott Lindner and Preethi Padmanabhan. Scott was based in Zurich, and we had a wide-ranging collaboration in designs of my first tape-out, Piccolo and Ocelot via skype. His British accent was so different, just as unique as his creative ideas in the sensor development. I'd like to give him a huge thanks for his dedication, patience and great help in the collaboration. Preethi was a master student in TUD and continued her research with Prof. Charbon as a PhD student in EPFL. I have been involved in her first and the latest projects. I'd like to thank Preethi for the frequent discussion, the willingness to share new ideas, calm attitude and persistence during the tape-out. She has a strong wish to develop the best SPAD sensors, which always encourages me moving forward.

I thank Dr. Fei Wang, Zhao Chen and Ting Gong for the knowledge sharing on CIS, SPADs and analog circuits. Everyday we ride back home with a lot of fun. The 10-minute riding time was the most relaxed moment during a day.

I'd like to thank to my colleagues: Harald Homulle, Dali Zhang, Claudia Damiani, Dr. Shingo Maindai, Arin Ulku, Andrea Ruffino, Andrada Muntean, Ashish Sachdeva, Francesco Gramuglia, Andrei Ardelean, Ekin Kizilkan, Jiuxuan Zhao, Yatao Peng, Kzuhiro Morimoto, Simone Frasca, Dr. Samuel Burri, Dr. Claudio Bruschini, Bishnu Patra, Rosario Marco, Jeroen van Dijk, Jiang Gong, Lizzy Hatfield, Pascal't Hart, Gerd Kiene, Milad Mehrpoo for the great help. I'd also like to thank my friends from TUD: Yan Xie, Sining Pan, Chao Chen, Hui Jiang, Long Xu, Weihan Hu, Mingliang Tan, Yang Liu, Xiaoliang Ge, Yu Xin,

Zuyao Chang, Lukasz Pakula, Yiyu Shen, Zhe Hou, Kefei Hei, Wenjie Pei, Daiting He, Shanshan Ren, Xiaoping Xu. Thanks to Antoon Frehe for the management of the computer system. I'd like to thank Minaksie Ramsoekh and Joyce van Velzen for the logistics management. Also I had good company with my housemates: Tiantian Yao, Dr. Jianbin Fang, Dr. Weichen Mao, Dr, Xi Zhang, Qiang Liu, Meng Li.

Finally, I wish to thank my parents for supporting me all the time and the unconditional love. I am grateful to my girlfriend for the understanding and love, which inspired me during the whole PhD period. It has been a long time I am not with them. Probably it is time to go back and fix this up.

The people, experience, moments as well as the country have became a special symbol and great treasure of my life. May the friendship and love last forever.

Chao Zhang
2019-1-10

LIST OF PUBLICATIONS

Journals

1. **C. Zhang**, S. Lindner, I.M. Antolovic, M. Wolf, and E. Charbon, "A CMOS SPAD Imager with Collision Detection and 128 Dynamically Reallocating TDCs for Single-Photon Counting and 3D Time-of-Flight Imaging", Sensors, Nov, 2018. [Shared first authorship]

2. **C. Zhang**, S. Lindner, I.M. Antolovic, J.M. Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252 × 144 SPAD Flash LiDAR With 1728 Dual-Clock 48.8-ps TDCs, and Pixel-Wise Integrated Histogramming", Journal of Solid State Circuits, DOI: 10.1109/JSSC.2018.2883720, 2018. [Shared first authorship]

Conferences

1. S. Lindner, **C. Zhang**, I.M. Antolovic, M. Wolf, J.M. Pavia, and E. Charbon, "A 252 × 144 SPAD pixel FLASH LiDAR with 1728 Dual-clock 48.8 ps TDCs, Integrated Histogramming and 14.9-to-1 Compression in 180nm CMOS Technology", Symposium on VLSI Circuits, 2018. [Shared first authorship]

2. S. Lindner, **C. Zhang**, I.M. Antolovic, A. Kalyanov, J. Jiang, L. Ahnen, A. di Costanzo, J. Pavia, S. Majos, E. Charbon, and M. Wolf, "A Novel 32 ×32, 224 Mevents/s Time Resolved SPAD Image Sensor for Near-Infrared Optical Tomography," in Biophotonics Congress: Biomedical Optics Congress 2018 (Microscopy/Translational/Brain/OTS), OSA Technical Digest (Optical Society of America, 2018), paper JTh5A.6. [Shared first authorship]

3. 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 180nm CMOS for Near Infrared Optical Tomography", Proc. Int Image Sensor Workshop, June 2017. [Shared first authorship]

4. P. Padmanabhan, **C. Zhang**, E. Charbon, "Analysis of a modular SPAD-based direct time-of-flight depth sensor architecture for wide dynamic range scenes in a LiDAR system", accepted by IISW 2019.

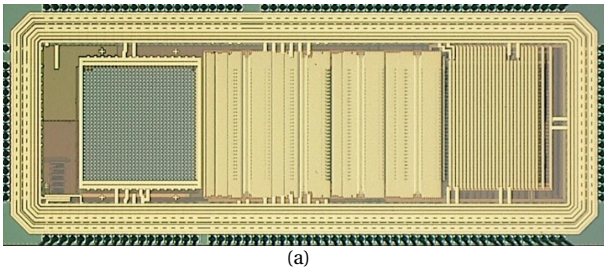
5. S. Lindner, **C. Zhang**, A. Kalyanov, M. Wolf, C. Bruschini, E. Charbon, "A Close-in LiDAR for Diffusive Media based on a 32 × 32 CMOS SPAD Image Sensor", accepted by IISW 2019.

6. **C. Zhang**, V.T.S. Dao, T.G. Etoh, and E. Charbon, "Pixel parallel localized driver design for a 128 × 256 pixel array 3D 1Gfps image sensor", 31st International Congress on High-Speed Imaging and Photonics, 2017.

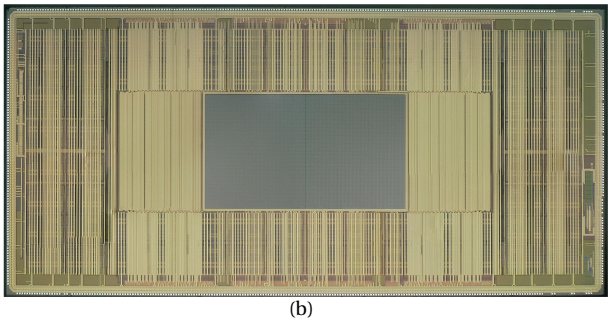
7. **C. Zhang**, V.T.S. Dao, T.G. Etoh, K. Shimonomura, and E. Charbon, "Designing pixel parallel, localized drivers of a 3D 1Gfps image sensor family", Proc. Int Image Sensor Workshop, 2015.

8. T.G. Etoh, V.T.S. Dao, K. Shimonomura, E. Charbon, **C. Zhang**, Y. Kamakura and T. Matsuoka, "Toward 1Gfps: Evolution of ultra-high-speed image sensors -ISIS, BSI, multi-collection gates, and 3D stacking", International Electron Devices Meeting (IEDM), 2014.

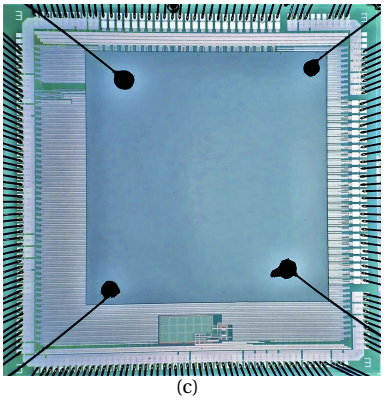
CHIP GALLERY



(a)



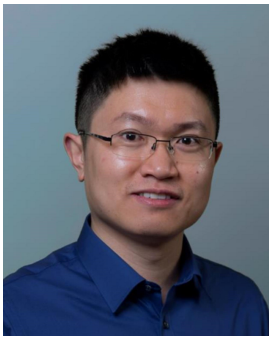
(b)



(c)

Fig. 7.1 (a) Piccolo; (b) Ocelot; (c) Nanosis.

About the author



Chao Zhang was born in Qingdao, China, on 27th, October, 1986. He received his M.S. degree from Jiangnan University, Wuxi, China, in 2011. From 2011 to 2012, he was a digital designer in nVIDIA, Shanghai, China. From 2013, he joined Prof. Edoardo Charbon's group in Delft University of Technology, the Netherlands, where he was working on the development of SPAD sensors and high-speed imagers. His research interest is focused on TOF imaging and processing, digital and analog circuits, high speed imaging, and relevant applications such as LiDAR, AR/VR, AI, etc.