Time-resolved Single-photon Detector Arrays for High Resolution Near-infrared Optical Tomography

THÈSE N^{O} 8815 (2018)

PRÉSENTÉE LE 31 AOÛT 2018

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR

LABORATOIRE D'ARCHITECTURE QUANTIQUE

PROGRAMME DOCTORAL EN GÉNIE ÉLECTRIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Scott Anthony LINDNER

acceptée sur proposition du jury:

Prof. D. Atienza Alonso, président du jury
Prof. E. Charbon, Prof. M. Wolf, directeurs de thèse
Prof. A. Torricelli, rapporteur
Prof. W. Uhring, rapporteur
Prof. P. Seitz, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

Acknowledgements

As I reflect on my time as a PhD student, there are many people I'm grateful to for making this all possible. I'd like to thank my two advisors Martin Wolf and Edoardo Charbon for giving me the chance to work on such a rewarding project and for their personal support when things went awry. I'd like to thank Martin for his scientific advice and for cultivating the friendly and collaborative environment at BORL which makes it a great place to come and work. I'm also very grateful to Edoardo, for fruitful scientific discussions, providing great opportunities to design chips and then encouraging us to pursue bold ideas.

I'd like to thank everyone in the team at BORL for their help along the way. Juan Mata Pavia, for meeting me at the airport over 5 years ago, providing guidance as I got to know the project, and karaoke nights in Selnau. I wish to thank Linda Ahnen for being a great office mate, runs along the limmat, and many chats over a cup of tea. Damien de Courten also shared an office with us and contributed to a friendly working environment, giving us many laughs over the years. Rashmi Hegde joined the project as a master student at a critical time and was a great help in turning Piccolo into a robust system. The Pioneer team at BORL, Sasha, Jinjing and Aldo, for their enthusiasm for the project and patience as I iron out the hardware. I'd like to thank Sasha in particular for assisting with the flash measurement with Piccolo. Iris Suter was always extremely helpful for administrative issues and has my gratitude. I'd also like to express thanks to all past and present members of neonatology at USZ who've contributed to my work, among them, Helene Isler, Daniel Ostojic, Dominik Wyser, Stefan Kleiser, Mark Adams, Nassim Nasseri, Raphael Zimmerman, Andreas Metz, Salvador Sanchez Majos, Felix Scholkmann, Simon Christen, Matthias Heinzmann, Flavia Wehrle, Manya Hendriks, Thi Dao Nguyen, Jean-Claude Fauchère, Dirk Bassler, Claudia Knöpfli and Tanja Karen.

Despite being based in Zurich, I worked for a lot of the time with colleagues from Edoardo's AQUA labs in Delft and then later in Neuchatel. Although I existed mainly as a voice through skype, whenever I worked with colleagues in these groups either online or when I visited, I found people eager to collaborate and who shared their knowledge freely. The Ocelot and Piccolo sensors were co-designed with Chao Zhang and Ivan Michel Antolovic. I'd like to

give a huge thanks to Chao Zhang, for his dedication, calm attitude under pressure, and for sharing his expertise of digital circuits. Michel's calm demeanor and attention to detail were massively appreciated and crucial to the success of the sensors. The 3D IC pixels were designed as part of the POLIS project, where I worked with Augusto Carimatto and Augusto Ximenes. Augusto Carimatto was a great collaborator in stressful situations and provided great company whilst watching football whenever I was in Delft. I'd like to thank August Ximenes for sharing his knowledge of IC design which helped improve the later designs of Piccolo and Ocelot. Esteban Venialgo, MJ Lee and Chockalingam Veerappan, with whom I never worked directly, deserve a special mention for sharing their extensive expertise. I always enjoyed these technical discussions, which improved my own work greatly. I'd like to thank Preethi Padmanaban for frequent discussions both inside and out of work, and for nights singing karaoke in Japan. Francesco Gramuglia wrote firmware for the Piccolo and Ocelot boards, which was a great help in bringing the system into the lab that much sooner. Claudio Bruschini and Brigitte Khan have also been a great help with organizational issues as I've come towards the end of my PhD. I'd also like to express thanks to all past and present members of AQUA who've contributed to my work, among them, Arin Ulku, Samuel Burri, Harald Homulle, Ting Gong, Kazuhiro Morimoto, Andrea Ruffino, Andrada Muntean, and Andrei Ardelean.

Finally, I would like to thank my family for the inspiration and support they've given me throughout the years. To my brother, Mark, whose determination and dedication is a daily motivator. To my mum, Doreen, for filling my head from a young age with stories of her travels and then cheering me on as I went on my own journey. And to my dad, Tony, for showing me the patience that only a lifetime of fishing can teach, and for instilling in me the strength to see things through to the end.

Abstract

Oxygenation is an important marker in many clinical settings, e.g. diagnosing ischaemic brain injuries in preterm infants or determining treatment effectiveness in cancer patients. Despite significant efforts to determine the oxygenation state of the human tissue with conventional imaging modalities, e.g. positron emission tomography (PET) and magnetic resonance imaging (MRI), a method which is fit for continuous, routine use in clinics does not exist.

Near-infrared optical tomography (NIROT) is a compelling alternative which can be employed to measure tissue oxygenation. This method is based on the illumination and subsequent detection of the tissue response to light in the near-infrared (NIR) spectrum. NIROT is non-invasive and safe for continuous monitoring of patients. However, conventional NIROT systems have demonstrated a limited spatial resolution, in the 1-2 cm range, and prevented widespread clinical uptake. A major reason for this limited resolution is the low numbers of sources and detectors in conventional systems.

In an effort to improve the spatial resolution of NIROT, researchers have recently applied time-resolved cameras based on single-photon avalanche diodes (SPADs) to NIROT phantom measurements, achieving a resolution of 5 mm. Despite these promising results, conventional SPAD cameras are unsuitable for clinical measurements due to a slow image acquisition time.

In this thesis, time-resolved cameras were developed which have the potential to perform image acquisitions in NIROT measurements for a multi-source multi-wavelength system in a number of minutes. The main objective was to develop a large format, 252×144 pixel, time-resolved SPAD camera capable of wide field measurements in a clinical setting.

Two new pixel circuits were developed in a backside-illuminated (BSI) 3D IC technology to increase the signal-to-noise ratio (SNR) and dynamic range (DR) in time-resolved measurements. The first circuit demonstrates, for the first time, a technique to increase the excess bias range, and thus the photon detection efficiency (PDE) and timing performance of conventional SPAD pixels. Coupled to an active recharge circuit, the pixel achieves a minimum dead time of 8 ns, and is thus suitable for high DR measurements. The second presents the first pixel circuit able

to interface with a SPAD via the anode or the cathode terminal, thus enabling the possibility of a general purpose time-correlated single-photon counting (TCSPC) die connecting to multiple application specific photodetector dies.

A new time-resolved SPAD sensor architecture is presented which employs a time-to-digital converter (TDC) sharing architecture to achieve both high PDE and high throughput parallel measurements. A 32×32 sensor based on this architecture is produced in a 180nm CMOS technology. At the maximum throughput, 10^6 photons can be obtained for every pixel in the array in parallel in 4.6 seconds.

Finally, a wide-field, 252×144 pixel time-resolved sensor is presented. To maintain a fast acquisition speed the architecture includes a per-pixel integrated histogramming readout. This enables the compression of the readout data by up to a factor of 14.9. To the best of the author's knowledge, this is the first implementation of integrated histogramming on a per-pixel basis for a full sensor array. This new sensor has game changing potential for NIROT, opening the door to high resolution wide-field clinical measurements.

Key words: Near-infrared spectroscopy (NIRS), near-infrared optical tomography (NIROT), optical tomography (OT), diffuse optical tomography (DOT), diffuse optical imaging (DOI), single-photon avalanche diode (SPAD), single-photon imaging, time-to-digital converter (TDC), time-correlated single-photon counting (TCSPC), integrated histogramming

Zusammenfassung

Die Oxygenierung ist ein wichtiger Biomarker in vielen klinischen Situationen, insbesondere bei der Diagnose von ischämischen Hirnverletzungen bei Frühgeborenen oder der Bestimmung der Behandlungwirksamkeit bei Krebspatienten. Trotz erheblicher Anstrengungen, den Oxygenierungszustand des menschlichen Gewebes mit konventionellen bildgebenden Verfahren, wie z.B. der Positronen-Emissions-Tomographie (PET) und der Magnetresonanztomographie (MRT), zu messen, gibt es bisher keine Methode, die für den kontinuierlichen, routinemäßigen Einsatz in Kliniken geeignet ist.

Die optische Nah-Infrarot-Tomographie (NIROT) ist eine überzeugende Methode zur Messung der Gewebeoxygenierung. Sie basiert darauf, dass Gewebe mit nahinfrarotem (NIR) Licht beleuchtet und die Intensität des aus dem Gewebe wieder austretenden Lichtes gemessen wird. NIROT ist nichtinvasiv und harmlos sogar bei einer kontinuierlichen Überwachung von Patienten. Herkömmliche NIROT-Systeme verfügen jedoch nur über eine begrenzte räumliche Auflösung im Bereich von 1-2 cm, was eine weite Verbreitung in der klinischen Anwendung verhindert. Ein wesentlicher Grund für diese begrenzte Auflösung ist die geringe Anzahl von Quellen und Detektoren in herkömmlichen Systemen.

Um die räumliche Auflösung von NIROT zu verbessern, haben Forscher kürzlich zeitaufgelöste Kameras mit Single Photon Avalanche Dioden (SPADs) auf NIROT mit Phantommessungen untersucht und eine Ortsauflösung von 5 mm erreicht. Trotz dieser vielversprechenden Ergebnisse sind herkömmliche SPAD-Kameras aufgrund der langen Aaufnahmezeit nicht für klinische Messungen geeignet.

In dieser Arbeit wurden zeitaufgelöste Kameras entwickelt, die das Potenzial haben, diese Aufnahmzeit bei NIROT-Messungen inklusive Messungen bei mehreren Wellenlängen und Lichtquellenpositionen auf wenige Minuten zu reduzieren. Das Hauptziel war die Entwicklung einer großformatigen, zeitaufgelösten SPAD-Kamera mit 252 × 144 Pixeln, die in der Lage ist, hochauflösende Messungen im klinischen Umfeld durchzuführen.

Zwei neue Pixelschaltungen wurden in einer rückseitig beleuchteten (BSI) 3D-IC-Technologie entwickelt, um das Signal-Rausch-Verhältnis (SNR) und den Dynamikbereich (DR) bei zeitauf-

gelösten Messungen zu erhöhen. Die erste Schaltung demonstriert zum ersten Mal eine Technik zur Erhöhung der Sperrspannung und damit der Photonendetektionseffizienz (PDE) und der Zeitauflösung herkömmlicher SPAD-Pixel. Kombiniert mit einer aktiven Wiederaufladeschaltung erreicht das Pixel eine minimale Totzeit von 8 ns und ist somit für hohe DR-Messungen geeignet. Die zweite Schaltung stellt die erste Pixelschaltung dar, die über den Anoden oder Kathodenanschluss mit SPADs verbunden werden kann und somit einen universell einsetzbaren zeitkorrelierten Einzelphotonenzähler (TCSPC) Chip ermöglicht, der an mehrere anwendungsspezifische Photodetektorchips angeschlossen werden kann.

Es wird eine neue zeitaufgelöste SPAD-Kamera vorgestellt, die eine Architectur verwendet bei der Time-to-Digital Converter (TDC) gemeinsam verwendet werden, so dass sowohl eine hohe PDE- als auch parallele Messungen mit hohem Durchsatz erreicht werden. Ein darauf basierender 32 × 32 Pixel Sensor wurde in einer 180nm CMOS-Technologie hergestellt. Bei maximalem Durchsatz können für jedes Pixel des Chips 10^6 Photonen parallel in 4,6 Sekunden gewonnen werden.

Schließlich wird ein großer, zeitaufgelöster 252 × 144 Pixel Sensor vorgestellt. Um eine hohe Aufnahmegeschwindigkeit zu gewährleisten, verfügt die Architektur über eine integrierte Histogrammberechnung für jedes Pixel. Dies ermöglicht die Komprimierung der ausgelesenen Daten um bis zu einen Faktor von 14,9. Nach bestem Wissen des Autors ist dies die erste Implementierung einer integrierten Histogrammberechnung auf Pixelbasis für einen gesamten Sensorarray. Dieser neue Sensor eröffnet ein grosses Potential für NIROT, und wird zu hochauflösenden klinischen Messungen führen.

Stichwörter: Nah-Infrarot-Spektroskopie (NIRS), Nah-Infrarot-Tomographie (NIROT), optische Tomographie (OT), diffuse optische Tomographie (DOT), diffuse optische Bildgebung (DOI), Einzelphotonen-Avalanche-Photodioden-Detektoren (SPAD), Einzelphotonen-Bildgebung, Zeit-Digital-Wandler (ZDW), zeitkorrelierte Einzelphotonen-Zählung, integrierte Histogrammberechnung

Contents

List of Figures

List of Tables

List of Acronyms

ALTDC	address latch and time-to-digital converter
AOTF	acousto-optical tunable filter
BSI	backside-illuminated
CCD	charge-coupled device
CMOS	complementary metal-oxide-semiconductor
CP	charge pump
cps	counts per second
CW	continuous wave
DAC	digital-to-analog converter
DCR	dark count rate
DFF	D-type Flip-Flop
DLL	delay-locked loop
DNL	differential nonlinearity
DOI	diffuse optical imaging
DOT	diffuse optical tomography
DR	dynamic range
FD	frequency domain
FLIM	fluorescence lifetime imaging
FOV	field of view

FPGA field-programmable gate array

FSI frontside-illuminated

FWHM full width at half maximum

GPIO general purpose I/O

Hb hemoglobin

HBP hybrid bonding pad

HHb deoxygenated hemoglobin

INL integral nonlinearity

IRF instrument response function

LET light emission test

LF loop filter

LiDAR light detection and ranging

LSB least significant bit

LVDS low-voltage differential signalling

MC Monte Carlo

MRI magnetic resonance imaging

NIR near-infrared

NIROT near-infrared optical tomography

NIRS near-infrared spectroscopy

O_2Hb oxygenated hemoglobin

OT optical tomography

PCB printed circuit board

PDE photon detection efficiency

PDP photon detection probability

PET positron emission tomography

PFD phase-frequency detector

PHR	partial-histogramming readout
PLL	phase-locked loop
PMT	photomultiplier tube
PVT	process, voltage and temperature
rms	root mean square
RO	ring oscillator
RTE	radiative transfer equation
SiPM	silicon photomultiplier
SNR	signal-to-noise ratio
SPAD	single-photon avalanche diode
SRAM	static random access memory
TCSPC	time-correlated single-photon counting
TD	time domain
TDC	time-to-digital converter
ToF	time-of-flight
TPSF	time point-spread function
VCO	voltage-controlled oscillator

1 Introduction

Oxygen is crucial for our continued survival, yet methods for measuring its concentration in our bodies are severely lacking. This chapter begins by outlining the importance of oxygen and discusses some clinical applications where there is a pressing need for better imaging techniques. Near-infrared optical tomography (NIROT), a technique based on the propagation of light, has the potential to address this need. The state-of-the-art in NIROT methods and instrumentation is reviewed. The spatial resolution, and thus the clinical relevance, of existing systems is constrained by hardware. Thus, a motivation for developing new time-resolved hardware based on single-photon avalanche diodes (SPADs) is presented.

1.1 The Clinical Need for Oxygen Measurement

Oxygen is vital for human life. Every time we take a breath, oxygen is extracted from the air filling our lungs and then transported around our bodies via the hemoglobin in the blood. At the cellular level, oxygen is used to convert nutrients into adenosine triphosphate (ATP), which provides the energy we need to function in our daily lives, e.g. movement, cognition, etc. As well as its presence, the concentration and distribution of oxygen in the body are also of critical importance. An inadequate supply of oxygen to the body, hypoxia, can affect the whole body but also localised regions. This thesis focuses on two applications in particular.

Every year approximately 15 million babies worldwide are born prematurely, with a gestational age of less than 37 weeks [2]. For these infants, a lack of oxygen in the brain after birth can lead to severe complications such as hypoxic-ischaemic and haemorrhagic brain injuries. This can result in life altering disabilities, e.g. cerebral palsy and epilepsy, and in the worst cases, death [3]. A study of surviving extreme preterm infants (< 28 weeks gestation), who

account for approximately 5.2% of the premature births [2]), showed almost 50% of the cohort demonstrated significant cognitive delay [4] at school age. Although there are treatments which can be applied, the time window within which these treatments are effective is typically small. For example, hypothermia has been shown to attenuate brain damage, however, it must be administered within 6 hours [5]. Thus, there is a pressing need to identify those infants which require treatment as early as possible.

Another medical setting where hypoxia is important is in the prognosis and treatment of cancer. For cancer tumors, hypoxia is the result of an imbalance between the consumption and supply of oxygen. This can be caused by inadequate blood supply to the tumor and also an increased oxygen demand due to tumor growth. Tumor hypoxia has been shown to indicate a greater resistance to radiation [6] and chemo- [7] therapies. A 1996 study [8], Figure 1.1, investigated the effect of tumor oxygenation on treatment effectiveness and patient outcomes by sampling the partial oxygen pressure (pO_2) of advanced uterine cervical tumors at multiple points spaced 0.7 mm apart with a polarographic probe. The study showed that patients whose tumors had a median pO_2 of less than 10 mmHg had a less than 40% 5-year survival rate. In contrast, patients whose tumors exhibited a median pO_2 of greater than 10 mmHg had a 5-year survival rate of more than 70%. These results tell two stories. Firstly, oxygen is an important indicator for treatment effectiveness and prognosis for cancer patients. Secondly, the necessity of sampling many points over a cancer tumor indicates that heterogeneity is crucial. Thus an effective method for determining the oxygenation state in this setting must provide spatially resolved information of the entire tumor, rather than just a point value.

With the pressing need for quantification of oxygen within the human body, it should come as no surprise that a number of methods have already been investigated, e.g. polarographic probes [9] as used in [8], paramagnetic resonance imaging [10], magnetic resonance imaging [11], and PET [12, 13]. Of these, PET is generally regarded as the gold standard imaging method for *in vivo* oxygen quantification due to its high specificity and sensitivity enabled by hypoxia specific radiolabelled agents. However, the use of these radiotracers raises concerns about radiation exposure to patients. As such, measurements are generally restricted to two time points [14]. Furthermore, long measurement times, high costs and the requirement for an on-site cyclotron have limited the uptake of PET for oxygen measurements.

For the two applications already mentioned, monitoring cerebral oxygenation in neonates and tumor hypoxia, there remains a clear need for improved quantification of oxygenation in human tissue. Such a method must firstly be safe. Although PET has been applied for the measurement of glucose metabolism in term infants [15], the safety of the radiation dose is

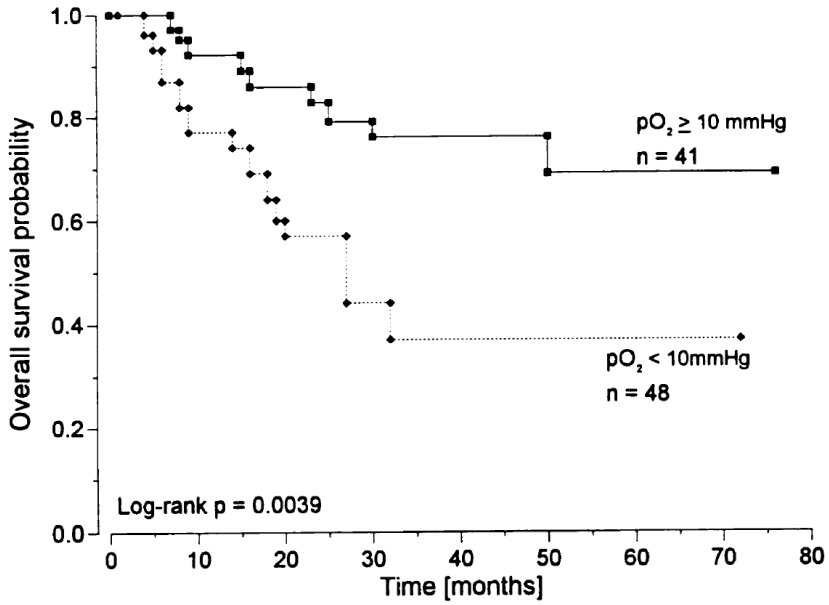


Figure 1.1 – A 1996 study [8] demonstrated that the 5-year survival rate for uterine cervical cancer reduced from greater than 70% with a median $pO_2 > 10$ mmHg to less than 40% with a $pO_2 < 10$ mmHg, with pO_2 sampled at multiple points with a polarographic probe. *Picture credit: M. Höckel, Copyright ©1996, American Association for Cancer Research.*

dependent on the administered radiotracer and measurement frequency. Thus, continuous monitoring is a concern as it is with tumor hypoxia monitoring [14]. Additionally, an effective oxygen monitoring method should have a high spatial resolution. This is important to localise regions of low oxygenation in the brain of preterm infants and also to detect oxygenation heterogeneities in cancer tumors. Finally, low cost is a crucial aspect for widespread uptake. Both of these applications would likely require screening of a large number of patients as well as continuous monitoring to monitor the treatment effectiveness.

Relatively new methods to measure the oxygenation state of biological tissue are near-infrared spectroscopy (NIRS) and near-infrared optical tomography (NIROT). In these methods, the biological tissue under study is illuminated with light in the NIR region of the spectrum, from 600-1000 nm. By detection of the light leaving the tissue via a suitable photodetector, the oxygenation state can be obtained by exploiting differences in the optical properties of the various substances in the tissue. Since this method employs only light, it is non-ionizing, completely safe and can be applied continuously. Furthermore, the hardware required to illuminate and detect light are inexpensive when compared to systems such as PET and MRI. Therefore, imaging methods based on this principle are highly promising for clinical settings.

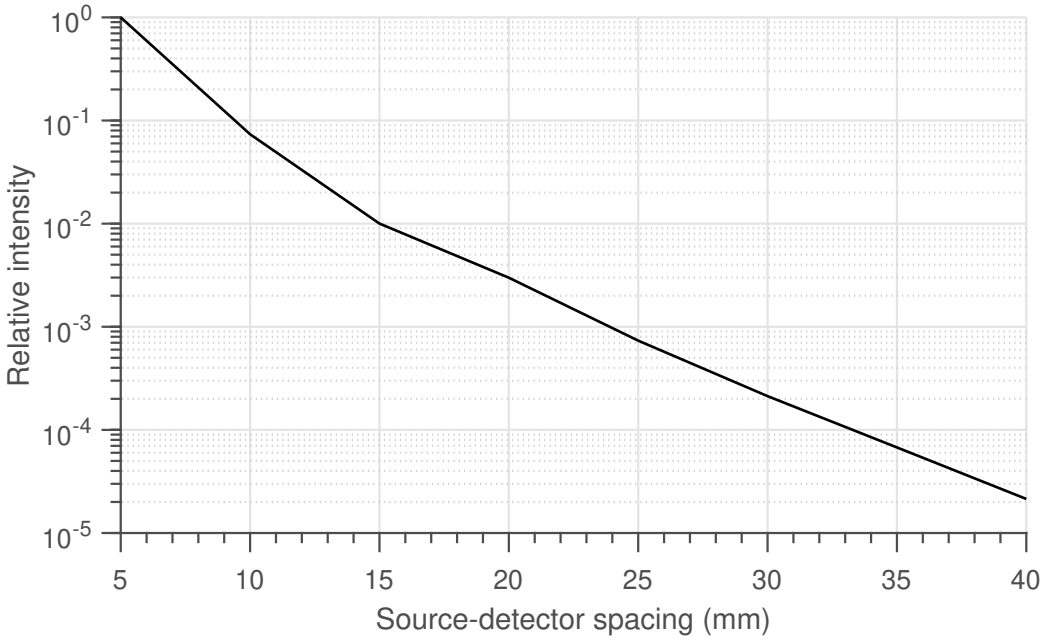


Figure 1.2 – NIRFAST [18] simulation of relative intensity in reflectance mode NIROT measurement as a function of source-detector separation. Simulation in 40 mm depth homogeneous medium with $\mu_a = 0.01$ mm^{-1} and $\mu_s' = 1$ mm^{-1}, i.e, typical properties of adult human tissue.

1.2 Light Propagation in Tissue

To better understand how NIROT works, we must first look at light propagation in tissue, which is largely dictated by two processes, scattering and absorption. Scattering is due to the interaction of photons with structures within the medium [16], e.g. cells, nuclei, organelles, [17] etc. This interaction alters the direction of the incident photons and can be expressed by the scattering coefficient, μ_s, which gives the probability that a photon will be scattered per unit length. An alternative measure, the reduced scattering coeffient μ_s', gives the probability per unit length that a photon loses its initial direction. As photons scatter inside the medium, collisions with molecules can result in the incident photon being absorbed. The probability of absorption per unit length is given by the absorption coefficient, μ_a. As a result of absorption, the light intensity detected at the surface of the medium decreases with an exponential dependence as the distance between the light source and detector, the source-detector (s-d) separation, is increased, Figure 1.2.

Figure 1.3 [19] shows the absorption spectra of many of the molecules found in human tissue over the wavelength range 100-10000 nm. Critical to the function of NIRS and NIROT, there exists an 'optical window' in human tissue between approximately 650 and 950 nm [19], at

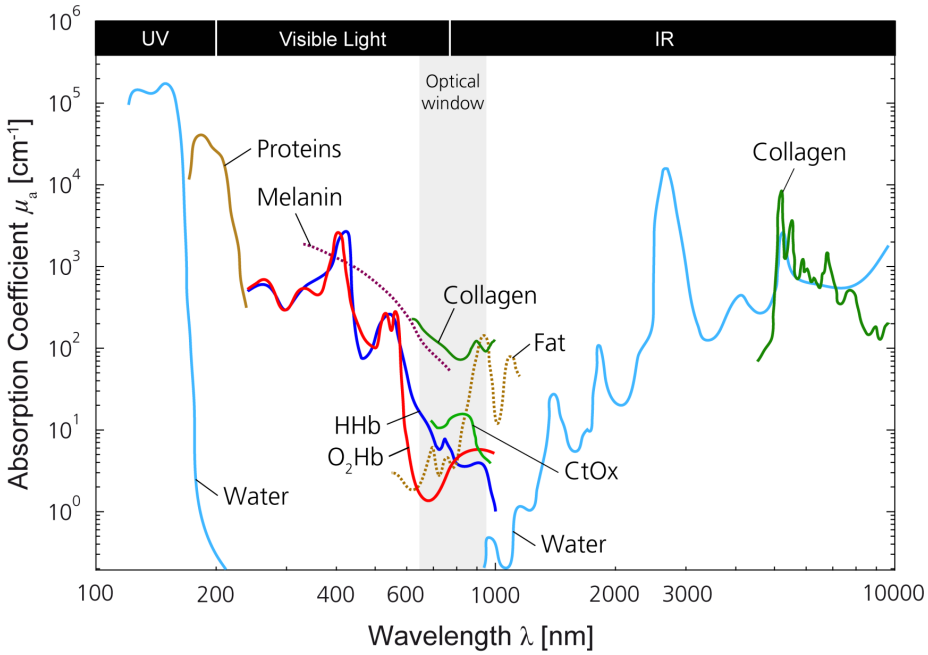


Figure 1.3 – Absorption spectra of molecules in human tissue in the wavelength range 100-10000nm [19]. *Picture credit: F. Scholkmann, Copyright ©2013 Elsevier Inc. All rights reserved.*

which absorption and scattering is relatively low. In this window, light is able to penetrate several centimetres without being absorbed, allowing the study of relatively deep regions of tissue. The main absorbers in the optical window are oxygenated hemoglobin (O_2Hb) and deoxygenated hemoglobin (HHb). Hemoglobin (Hb) is contained in the red blood cells which are mainly responsible for oxygen delivery from the lungs to the different parts of the body. Although there are molecules with a greater absorption coefficient between 650 and 950 nm, their low concentrations in human tissue mean that their contribution to total absorption is small, thus they are generally neglected in the calculation of oxygenation state.

The absorption spectra of the various molecules within the optical window are each unique and show a wavelength dependency. Therefore, measurements at multiple wavelengths can be performed to obtain the absorption contributions from individual molecules.

In human tissue, scattering is dominant in comparison to absorption, μ_s' values 100 times greater than μ_a are typical [20]. Light propagation in tissue then is a diffusive process, and any calculation of the oxygenation state must take into account the contributions to the detected signal from both absorption and scattering.

1.3 Near-infrared Spectroscopy (NIRS)

NIRS was first demonstrated in 1977 by Frans Jöbsis [21], where the oxygenation status of both hemoglobin and cytochrome oxidase were measured. This initial discovery led to the development of a range of different instruments applied in settings such as breast tumor diagnosis [22], functional activation of the cerebral cortex in neonates [23] and muscle oxygenation monitoring [24].

The earliest measurements were carried out with continuous wave (CW) NIRS devices [25]. In this method, the tissue under study is illuminated with light at a constant intensity and the detector measures the attenuation in the light that exits. The simplicity of this technique means that it can be implemented with relatively unsophisticated hardware that is both inexpensive and can be made portable. Furthermore, CW instruments have a high time resolution, e.g. a sampling rate up to 100 Hz [25], which means that they can monitor rapidly changing physiological signals. This property has enabled, for example, measurements of muscle oxygenation in sport activities [26] and has led to the CW method being by far the most popular among NIRS techniques. Unfortunately, the CW method cannot determine the scattering properties of the tissue under study. As such, only relative changes in O_2Hb and HHb can be measured without applying more advanced measurement procedures or instrumentation [19].

To obtain information on the scattering properties of the tissue, the time duration required for the light to travel from the source to detector must be known [25]. There are two methods which can be applied for this purpose. Frequency domain (FD) NIRS devices obtain the time information by illuminating the tissue with frequency modulated light and the detector measures the amplitude and phase of the light that exits [27]. Time domain (TD) instruments on the other hand illuminate the tissue with very short pulses of light, typically on the order of tens of picoseconds, and the detector measures the dispersion in time of the exiting light [28]. Due to the speed of the response, the detected light is typically measured by single-photon detectors coupled to time-correlated single-photon counting (TCSPC) hardware [29]. In TCSPC, the arrival time of individual photons are measured in relation to a reference signal, typically the electrical trigger from a pulsed light source. Photons are accumulated over many cycles to achieve a certain statistical precision. The data is then output in the form of a histogram, where the bins on the x-axis represent arrival time windows and the y-axis is the number of photons to arrive in that window, this histogram is termed the time point-spread function (TPSF). Of the two, FD instruments can be implemented with simpler, cheaper hardware, and provide better SNR and time resolution [30]. However, to provide

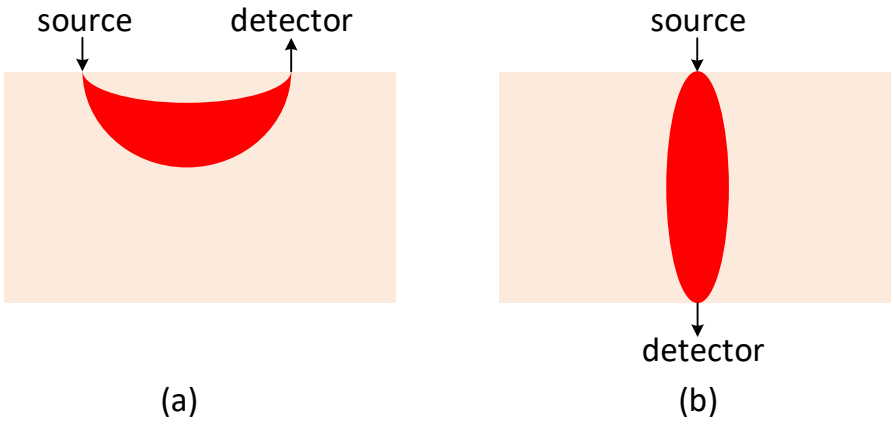


Figure 1.4 – Measurement geometry of NIRS in (a) reflectance and (b) transmission modes. The red region illustrates where the trajectories of most detected photons will be confined.

equivalent information to the TPSF in the TD method, an FD instrument must scan through a range of frequencies from 50 MHz to 1 GHz [31], thus increasing the system complexity and degrading the time resolution.

A crucial limitation of NIRS in comparison to established imaging methods such as MRI or PET is the depth penetration. NIRS instruments are typically employed with the source and detector in contact with the patient in either reflectance, Figure 1.4a), or transmission, Figure 1.4b), modes. In both of these figures, the region in which most of the photon trajectories are confined is illustrated in red, showing the diffuse nature of photon propagation. In reflection mode, due to the exponential decay of light intensity, Figure 1.2, a source-detector separation in the region of 1-4 cm is typically employed [32]. Since the source-detector spacing is proportional to the depth penetration in CW and single frequency FD methods, a source-detector separation of 4 cm translates to a depth penetration of approximately 2-3 cm. This can be increased by measuring the medium in transmission mode, where the source and detector are on opposite sides of the tissue under study. In this configuration a depth penetration of up to 6 cm can be achieved. In practice, transmission mode is applied in relatively few cases such as optical mammography [33] whilst the reflectance mode is highly versatile and applicable to a much larger number of situations.

In the TD method, the source-detector separation is not proportional to the depth penetration since the TPSF allows the discrimination between early and late photons. Thus, in principle the TD method can be used at very small source-detector separations to achieve up to 6 cm penetration depth [34]. In practice, these small source-detector separations can only be implemented by gating the detector on for small windows of time, [35], due to the large

dynamic range requirement. The power of the laser source can then be adjusted as the gate window is moved to encompass the entire TPSF, e.g. low power for the earliest window and increasing for late windows. Thus, the dynamic range required of the detector is reduced.

1.4 Near-infrared Optical Tomography (NIROT)

Whilst NIRS is a widely employed tool for measuring the oxygenation state of human tissue, resolving localized changes in oxygenation with millimetric accuracy requires a tomographic approach. This tomographic approach is termed NIROT here, but is also commonly referred to as diffuse optical tomography, or diffuse optical imaging. NIROT is enabled by multi-channel hardware, illuminating the tissue and sampling the diffuse light at different points on the tissue boundary. Due to the multi-source, multi-wavelength nature of the measurement system, NIROT has a lower time resolution in comparison to NIRS. To obtain 3D images of the oxygenation state, e.g. concentrations of O_2Hb and HHb within the tissue volume, a two-step reconstruction [36] is performed. In the first step, a forward model for light propagation in the medium under study is defined. This forward model can be based on the radiative transfer equation (RTE) [37], the diffusion approximation, which is a simplified version of the RTE, or alternatively on numerical models such as the finite element method [18] and Monte Carlo methods [38]. The problem is then inverted, in what is known as the inverse problem, and measurements from the tissue boundary are used to reconstruct the absorption and scattering properties of the medium. An extensive coverage of image reconstruction in NIROT can be found in [36, 39, 40].

A major difference between NIROT and conventional imaging modalities such as MRI or PET, is that the image reconstruction problem is ill-posed due to the absorbing and scattering nature of tissue [41]. Furthermore, it is often underdetermined since there are more voxels to reconstruct than there is measurement information [42]. The ill-posed nature of the problem can be reduced by providing the image reconstruction with more information. This can mean including prior information of the tissue under study [39], employing TD instruments which provide the richest datasets, or increasing the number of sources and detectors [42].

The spatial resolution of NIROT systems can be divided into lateral and depth resolution. Due to photon migration in tissue, the lateral resolution in NIROT measurements degrades with depth penetration and decreased depth selectivity [32]. As such, a fair comparison of lateral resolution should be made at a fixed depth. Lateral resolution in NIROT reconstructions can be improved by reducing the ill-posed nature of the reconstruction, as stated previously.

Depth resolution can be improved by employing a time-domain instrument with a narrower instrument response function (IRF) although the influence of the IRF can be partially overcome with deconvolution algorithms [43, 44]. Thus, from a hardware perspective, it appears that a time-resolved system with a large number of narrow IRF detectors is the most promising instrumentation for achieving NIROT reconstructions with a high lateral and depth resolution.

1.5 Time-resolved Instrumentation

Early time-resolved systems employed photomultiplier tubes (PMTs) as photodetectors [45, 46, 47], whilst the time-resolved information was acquired with dedicated TCSPC hardware. Due to the size of the PMTs, light is coupled from the target to the photodetectors via a fibre bundle. This results in some loss of light as well as a small temporal dispersion of the TPSF [45]. Furthermore, this coupling of light into and out of the measurement subject makes scaling the system to large numbers of sources and detectors very difficult. For example, a PMT based system with 32 sources and 32 detectors coupled via fibres [45, 48] is already cumbersome. This low number of sources and detectors results in a spatial resolution in the 1-2 cm range in a NIROT application [49], limiting uptake in clinical settings. Furthermore, it is difficult to imagine scaling such a system to several hundreds of sources and detectors. Additional complexities of PMTs are the requirement of a high voltage, 3.2 kV in [45], cooling elements for temperature stability and high cost (> 5 k€ per unit) [50].

Recently, systems based on silicon photomultipliers (SiPMs) have been developed which overcome many of the shortcomings of PMTs [50, 51]. SiPMs are solid-state detectors where the outputs of many SPADs are summed together to form a macrocell. Thus, despite consisting of a large number of detectors, SiPMs typically have few macrocells, e.g. 4×4. However, their large photosensitive area, which can be several mm^2, means that they can be placed on a small printed circuit board (PCB) and placed in direct contact with measurement subject. This avoids the complexity and losses associated with fibre coupling. Such detectors are available for a relatively low cost, e.g. < $100 per unit, and do not require an extremely high voltage, as is the case with PMTs. Furthermore, TDCs are employed to acquire the time-resolved data, which can be purchased for less than $1000 dollars per channel [51] in comparison to a TCSPC board at > 8 k€ per unit. The combination of SiPMs with TDCs represents a leap forward for time-resolved hardware. Despite these improvements, however, the application of discrete sensor probes attached to the measurement subject means that scaling up the system to a large number of sources and detectors would remain cumbersome.

An approach to NIROT which does not involve fibre coupling or probes based on discrete hardware acquires time-resolved data with a gated intensified charge-coupled device (CCD) camera [52, 53]. This method has the benefit of being able to acquire images over a wide field of view (FOV) with a large number of detectors, thus it can in principle be used for a high-resolution NIROT system. Gated operation implies that the measurement time is increased, since the complete TPSF must be acquired in several windows. On the other hand, the power can be increased or decreased depending on whether the window is early or late. There is thus, a delicate balance between the laser power and gate window length. A disadvantage of the gated CCD approach is the minimum length of the gate window with 300 [52] and 500 [53] ps, which is inferior to the state-of-the-art IRF of the SiPM at 130 ps [50], and 80-150 ps of the PMT [45].

A compelling alternative to gated CCD cameras, which still employ an array of detectors but exploit the TCSPC approach are image sensors based on CMOS SPAD. SPADs first came to prominence in the 1970s and 1980s, as researchers identified their potential for observing scientific phenomena on the picosecond scale [54]. As early as 1989, SPADs with full width at half maximum (FWHM) down to 20 ps [55] had been reported. Despite the speed of such devices, the hardware required to implement multi-channel systems was prohibitively bulky and expensive. A major breakthrough was made, with the demonstration of the first SPAD in CMOS [56]. Soon multi pixel SPAD arrays could be produced [57] and finally, SPAD arrays with integrated TDCs [58, 59], achieving a timing response FWHM in the range of 100-250 ps.

In [42] an early 128×128 time-resolved SPAD camera [58] was applied in NIROT phantom experiments, demonstrating a spatial resolution of 5 mm. A number of other systems have been developed employing individual SPADs with TCSPC systems in both free-running and gated modes, a review of which can be found in [60]. Time-resolved cameras based on SPADs are a compelling prospect for two reasons. The possibility for a large number of detectors with a narrow timing resolution could lead to unprecedented spatial resolution in NIROT, whilst the future possibility to gate such a camera would lead to improve depth sensitivity.

1.6 Motivation and Aims of the Thesis

Oxygen is an extremely important clinical marker. For example, preterm infants commonly suffer from brain injuries caused by a lack of oxygen in the brain. The treatment window for such injuries is very small, typically just 6 hours. Similarly, hypoxia in cancer tumors is an indicator of increased resistance to radiation and chemo- therapies. The oxygenation state of

tumors could lead the way to more effective treatments and accurate prognosis. Therefore, there is a stringent need for a safe, non-invasive, reliable and inexpensive method to measure the oxygenation state of the human body. Despite this need, no existing method is widely applied in clinics. In particular, the current gold standard, PET suffers from high costs, poor time resolution, as well as the use of ionizing radiation.

NIROT presents a promising alternative for measuring the oxygenation state of human tissue. It is safe, as it uses only non-ionizing radiation, can be employed for continuous monitoring and is relatively cheap in comparison to established imaging modalities, e.g. PET or MRI. However, the major obstacle preventing widespread uptake of NIROT in clinical applications is a limited spatial resolution. This is important, for example, to localise regions of the preterm brain with an oxygen deficit and to measure hypoxia heterogeneity across cancer tumors. To increase the spatial resolution of NIROT, the information available for the reconstruction should be maximised. Practically, this means employing a large number of time-resolved detectors in a multi-wavelength measurement.

Up to now, most TD NIROT systems have employed discrete photodetectors such as photomultiplier tubes with bulky TCSPC boards for the timing conversion [45, 46, 47]. Although more recent systems have been based on more compact and less costly SiPMs and TDCs, it remains difficult to scale systems based on discrete hardware to large numbers, e.g. > 500 detectors.

In this thesis we advocate the use of time-resolved image sensors based on SPAD arrays for NIROT. Thanks to SPAD integration in CMOS, a very large number of single-photon detectors is possible. The power of applying a time-resolved SPAD sensor to the NIROT application was demonstrated in [42], where a spatial resolution of 5 mm was achieved in a laboratory experiment. However, whilst these cameras are an exciting prospect for implementing a NIROT system with millimetric spatial resolution, the image acquisition time with existing sensors is prohibitively slow [42]. In clinical measurements a long acquisition time will result in motion artifacts and decreased patient comfort. The slow acquisition time is due to two issues, the low sensitivity of the image sensor and rate at which photons can be acquired, e.g. the sensor throughput. This thesis aims to address both of these issues and produce instruments which are suitable for high resolution clinical NIROT measurements.

For a fast data acquisition, the sensitivity of the sensor, the SNR, should be maximised. Similarly, to handle the large range of intensities across the sensor as seen in Figure 1.2, a high dynamic range is required. The first aim of this thesis was to investigate new SPAD pixels for

improved SNR and dynamic range.

A major factor in the slow acquisition speed of current time-resolved SPAD sensors is the low throughput of conventional circuit architectures. The second aim was to design and implement a 32 × 32 time-resolved SPAD camera for reflection mode NIROT measurements which overcomes the low throughput of existing architectures.

A compelling prospect for NIROT is the possibility of capturing wide-field images without requiring any mechanical scanning of the detectors across the FOV. To enable this, the third aim was to develop a larger format, 252 × 144 time-resolved SPAD camera which extends the architecture already developed whilst maintaining a fast data acquisition time.

1.7 Thesis Organization

In Chapter 2, a background to SPADs and the impact of design decisions on system performance is discussed. Two new SPAD pixel structures for improved SNR and dynamic range in time-resolved measurements are then presented. Chapter 3 begins with a review of time-resolved SPAD imagers and a discussion of the challenges involved in their design. A new architecture which is targeted at the NIROT application is presented. In Chapter 4, a 32 × 32-pixel sensor based on this architecture is produced and measurement results reported. This architecture is extended in Chapter 5 to a 252 × 144-pixel format sensor, with circuit additions made to maintain a high event throughput. In Chapter 6, conclusions are drawn and an outlook presented for the future time-resolved SPAD sensors applied to NIROT.

1.8 Thesis Contributions

Due to the challenging optical environment in NIROT, pixels with high **SNR and DR** are required. In Chapter 2, two new pixels designs, produced in a BSI 3D IC technology, are presented. The first pixel employs a new quenching and recharge structure based on the cascode technique which can operate the SPAD at increased excess bias voltages, thus increasing the sensitivity and timing performance. Implemented using only transistors, the circuit is compact and remains compatible with active quenching and recharge schemes. To the best of the author's knowledge, this is the first quenching and recharge circuit which is capable of operating above the maximum voltage tolerance of a single transistor without the use of large polysilicon resistors. Implemented with an active recharge scheme, the pixel achieves the lowest afterpulsing at 8-ns for a BSI SPAD to date. The second pixel design exploits the

flexibility of 3D IC technologies, to demonstrate the first SPAD pixel which is capable of interfacing to a SPAD via the anode or cathode terminal. This opens the door to new 3D IC sensors which exploit a general purpose data processing tier which can be interfaced with different application specific photodetector tiers.

Chapter 3 presents a new sensor architecture for **high-PDE time-resolved SPAD imagers**. Motivated by the short acquisition time required by the NIROT application, a new architecture is proposed which overcomes the low fill factor and throughput of conventional TDC-in-pixel architectures. By employing an event-driven bus and dynamic TDC reallocation, a small number of TDCs can be shared by a large number of pixels. A collision detection bus is included to detect events which are the result of coincident photons on the shared bus. Analysis of missed photons due to the shared architecture with dedicated GPIO and LVDS pads per column shows that it is easily scalable to higher bus activities. Analysis of collision rates with data from NIROT forward simulations showed less than 3.5% of colliding photons at maximum bus activity. Since the maximum bus activity, and therefore the number of shared TDCs required, is limited by the output data rate, the number of pixels in the array can be extended to large numbers without an impact on pixel fill factor or overall throughput.

In Chapter 4, a 32 × 32-pixel time-resolved sensor based on the architecture in Chapter 3 is presented. Produced in a 180 nm CMOS technology, the sensor includes a wide-spectral range SPAD for low DCR and high PDP. The cascoded quenching and recharge circuit developed in Chapter 2 is employed to increase PDE. Power consumption is minimised with the design of a ring oscillator (RO) based TDC. The manufactured sensor achieves an unprecedented 28% fill-factor in a pixel pitch of 28.5μm with a maximum throughput of 220 Mevents/second. At this event rate, 10^6 photons can be acquired for every pixel in the array in 4.6 seconds, thus it is highly relevant for a clinical NIROT setting.

With the main goal of the thesis being to develop a **high throughput wide-field time-resolved SPAD sensor**, the architecture in Chapter 3 is extended to a 252 × 144 pixel sensor in Chapter 5. A bus repeater scheme is implemented to make the collision detection bus scalable. This addition allows the column to be extended without adverse impact on the bus dead time, timing jitter or pixel PDE. A per-pixel integrated histogramming scheme was developed to increase the photon throughput and thus minimise the image acquisition time in NIROT clinical measurements. To the best of the author's knowledge this is the first integrated histogramming scheme implemented for a full array. Modifications were made to the RO based TDC from Chapter 4, to continue employing the low power RO based technique whilst remaining compatible with the integrated histogramming readout. A clock generation scheme was

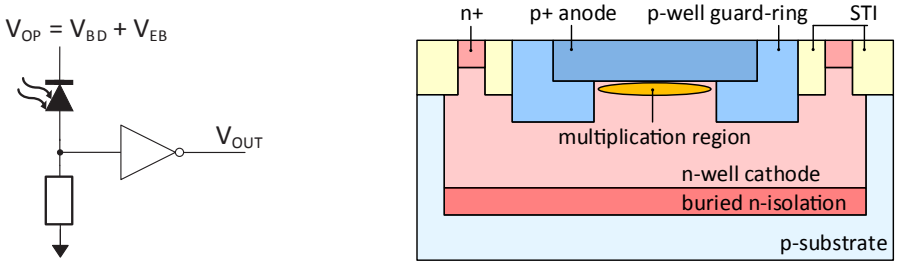
developed to achieve operation with a large range of laser frequencies without impacting the system timing response. The manufactured sensor retains the unprecedented 28% fill factor and narrow 110 ps FWHM timing response of the sensor in Chapter 3, whilst the integrated histogramming scheme enables a compression factor of up to 14.9. This compression factor can be employed to reduce the power consumed by the I/O pads or to increase the image acquisition speed of the sensor. Therefore, the manufactured sensor is highly suitable for NIROT measurements, opening the door to high resolution clinical systems.

2 Pixels for Improved SNR and Dynamic Range

NIROT is an extremely challenging optical environment for time-resolved image sensors. Due to the requirement for multiple source positions, which then need to be measured at a number of different wavelengths, many individual acquisitions must be made for a single reconstructed image. Thus, it is critical that each acquisition be completed in a short time, a number of seconds if possible. A high SNR and dynamic range are critical to achieving this goal. This chapter begins with a section dedicated to SPAD operation before discussing the SPAD characteristics and system level metrics which impact the duration and quality of NIROT measurements. Two new pixel designs are then demonstrated which can improve the SNR and dynamic range in a NIROT measurement. Results from the cascoded active recharge pixel are published in, S. Lindner et al., "A High-PDE, Backside-Illuminated SPAD in 65/40-nm 3D IC CMOS Pixel With Cascoded Passive Quenching and Active Recharge," in *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547-1550, Nov. 2017.

2.1 Single-photon avalanche diode (SPAD) Operation

A SPAD is a p-n junction which is reverse biased at a voltage V_{OP}. This is composed of the device breakdown voltage V_{BD}, which is determined by the device doping profiles and the excess bias voltage, V_{EB}, a parameter chosen by the user dependent on desired performance. When the reverse bias of a SPAD is set above its breakdown voltage the device is said to operate in Geiger-mode. When a photon arrives at or in close proximity to the depletion region, it can create an electron-hole pair, which with some probability can ignite a self-sustaining avalanche of mobile carriers via ionization of both holes and electrons. The region where 95% of the ionization occurs is termed the multiplication region. Once an avalanche has been initiated, the device would maintain this state until its destruction. For this reason,



(a) (b)

Figure 2.1 – (a) Simple SPAD front-end, (b) Cross section of CMOS SPAD. Structure reported in [62]

a SPAD may be connected in series with a ballast resistance, shown in Figure 2.1a. This resistance quenches the avalanche by reducing the current flowing to less than $\approx 100\mu$A [61] as the voltage across the SPAD decreases towards V_{BD} due to the space charge dynamics in the depletion region. Below this current, the avalanche is no longer self-sustaining and the depletion region becomes free of carriers once again. The voltage across the SPAD can then be recharged to V_{OP} through the ballast resistance.

The almost infinite optical gain implied by the creation of this avalanche of carriers means that SPADs are sensitive to the arrival of individual photons. Indeed, the output of the SPAD is connected to some discrimination electronics, which senses when the SPAD output voltage has passed a set threshold, this can be a comparator or as simple as a logic inverter, as depicted in Figure 2.1a. In this manner, a transition of the inverter output, V_{OUT}, from logic '1' to '0' signifies an avalanche event, or the arrival of at least a single photon. Therefore, SPADs are suitable for single-photon counting and TCSPC applications.

Whilst SPADs have been under investigation for a number of decades, it was only in 2003 that the first SPAD was fabricated in a CMOS technology [56]. An example of a SPAD is illustrated in Figure 2.1b. The depletion region and guard ring region, which ensures the electric field is highest in the multiplication region, are formed from layers already available in the CMOS process. Therefore, the performance of CMOS SPADs is typically inferior to those produced in custom processes, since the doping profiles are not been optimized. However, the possibility of integrating SPADs and CMOS electronics on the same silicon die is a great advantage, and enables the production of SPAD arrays with circuits for photon timing or counting. Due to this complex functionality, and the overhead for the SPAD, the pixel fill-factor, the area of the pixel which is sensitive to light, tends to be small.

2.2 SPAD Characteristics

To understand the challenges for developing pixels for a NIROT system it is instructive to first discuss the various characteristics of SPADs which will determine the performance. Particular attention is given to pixel design considerations. A more device physics oriented review can be found in [63].

Dark Count Rate

Dark counts are the occurrence of uncorrelated avalanche events in the absence of any photons impinging on the active region. The frequency is given by the dark count rate (DCR), measured in counts per second (cps) or Hertz (Hz). Although there are many physical mechanisms which contribute to DCR [64], they can be said to be due to either tunneling or thermal generation. A major source of noise is band-to-band tunneling [65], which is moderately dependent on temperature and highly dependent on doping concentrations. Furthermore, a high dependence on the electric field strength in the depletion region means this noise is very sensitive to the SPAD excess bias, V_{EB}. As process nodes used to develop SPAD imagers have shrunk below 100nm, band-to-band tunneling noise has become dominant due to the heavy doping concentrations of the native implants. To improve the noise performance at these nodes, custom layers have been defined for SPAD production [66]. Thermally generated noise is highly dependent on the density of lattice defects or "traps", which capture and then release carriers after some time duration dependent on the implantation and annealing processes in fabrication [63]. The contribution due to this trap assisted noise is exponentially dependent on temperature.

Photon Detection Probability

The photon detection probability (PDP) defines the likelihood of producing an avalanche for each photon incident on the active region of the SPAD. When a photon is incident on the detector, the depth at which it is absorbed and creates an electron-hole pair is dependent on the optical properties of the detection material. This dependency as a function of wavelength is given by the mean penetration depth, shown for silicon in Figure 2.2. Electron-hole pairs which are not created in the immediate vicinity of the multiplication region, e.g. too shallow or too deep in the silicon, will likely recombine before triggering an avalanche. As such, the peak PDP appears at the wavelength which has a mean penetration depth corresponding to the multiplication region depth. This is in the range of 400-560 nm in frontside-illuminated (FSI)

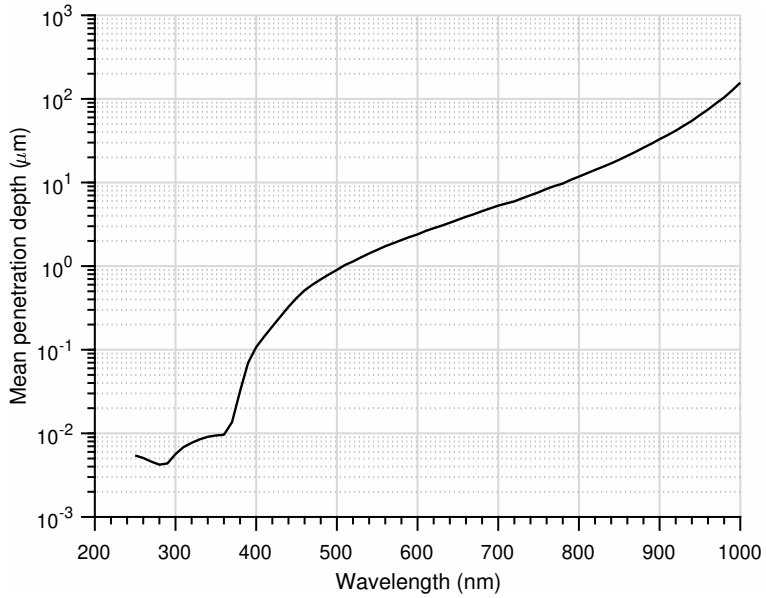


Figure 2.2 – Mean penetration depth in silicon as a function of wavelength, data taken from [1].

planar SPADs [67], corresponding to a multiplication region depth in the range of 100-1700 nm. Importantly, the electric field across the depletion region determines the ionization rate within the multiplication region. This means that the PDP of a SPAD can be improved by increasing the SPAD excess bias, V_{EB}. Practically, the excess bias cannot be increased indefinitely as PDP improvement diminishes at higher bias voltages and the DCR is also increased. This diminishing improvement does however mean that higher excess bias voltages can be employed to improve PDP uniformity in SPAD arrays [67]. This is illustrated in Figure 4.12, which shows that the gain in PDP as a result of increasing excess bias reduces as the voltage is increased and the PDP spectrum reaches the compression point.

Timing Jitter

One of the major benefits of SPADs is their timing response, characterized by a low timing uncertainty, or jitter, between a photon being absorbed and the output voltage of the SPAD reaching the comparison threshold of the discrimination circuitry. The SPAD has jitter because the build up and spread of each avalanche is dependent on the position [68] of the avalanche ignition and the statistics of the ionization process. The statistics of the avalanche process are responsible for the characteristic gaussian component of the timing response, usually expressed as the full width at half maximum (FWHM). Similar to the PDP, the timing jitter of a SPAD can be improved by increasing the ionization rate in the multiplication region by

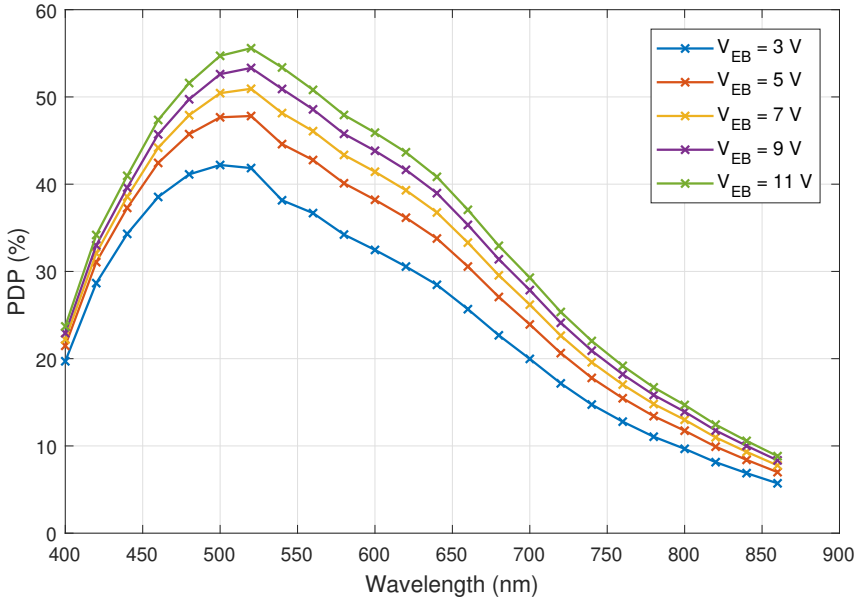


Figure 2.3 – Compression of PDP spectrum with increasing V_{EB}.

raising V_{EB}. Photons which are absorbed in the neutral region beneath the SPAD can also trigger avalanches but only after diffusing. Thus, the time to trigger the avalanche can take orders of magnitude longer than those photons which are absorbed in the depletion region. This mechanism is responsible for the slow exponential tail seen in the timing response, an example of which is given in Figure 2.4, where the tail has a time constant of 250 ps.

Afterpulsing

As well as being a major source of uncorrelated noise, as discussed in section 2.2, traps in the silicon lattice can also produce correlated noise in the form of afterpulsing. Carriers involved in the avalanche process can become captured by traps with an energy close to the energy bands of semiconductor. The release of these carriers can happen after a number of nanoseconds. If the SPAD excess bias has been sufficiently recharged by this point then the carrier release can trigger a second avalanche. Thus, the afterpulsing probability, P_{AP}, indicates the probability of an avalanche event resulting in a secondary avalanche via afterpulsing. Afterpulsing can be reduced by decreasing the number of carriers which flow during an avalanche or by minimising V_{EB}, and thus the probability of an avalanche, in the period that these secondary carriers will be released. These techniques are known as active quenching and active recharge, respectively.

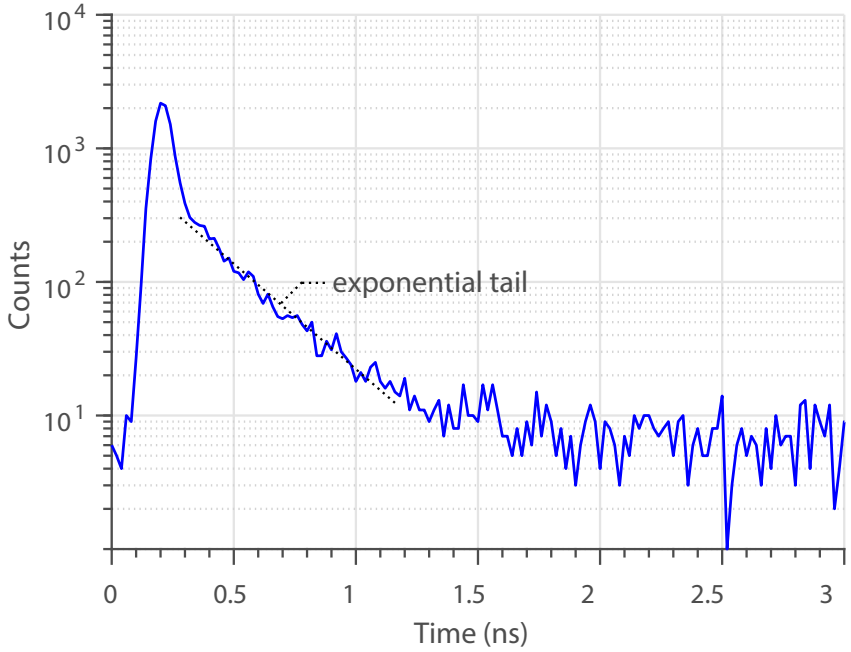


Figure 2.4 – Timing response of a SPAD with slow exponential tail, highlighted. The time constant of the tail is 250 ps.

Crosstalk

Crosstalk is correlated noise which is the result of an initial avalanche resulting in one or more secondary avalanches in neighboring SPADs. It can be categorised as being electrical or optical.

Electrical crosstalk can occur as the result of capacitive coupling from the anode or cathode terminals of neighboring devices. It can be reduced to a negligible level with careful layout of the sensor and PCB. As such, this effect is not widely reported in the literature.

Optical crosstalk is due to the release of secondary photons resulting from the recombination of carriers which flow during the initial avalanche. These secondary photons can then trigger avalanches in the surrounding SPADs. Optical crosstalk becomes more prevalent with decreasing spacing between neighbouring pixels. As such it is a major challenge when producing dense arrays. Crosstalk can be reduced by minimising the current flow during the avalanche, increasing the spacing between devices [69], or using isolation trenches between neighbouring SPADs [70].

2.3 System Level Metrics

As discussed in Section 1.6, the goal of this thesis is to produce time-resolved SPAD image sensors which massively increase the photon throughput in a NIROT system. From section 2.2, it is clear that there are several tradeoffs which must be considered at the pixel level to achieve this goal. To understand these tradeoffs more clearly, we must consider the impact of design decisions on some system level metrics.

2.3.1 Photon detection efficiency (PDE)

Due to the absorption and scattering of the biological media under study in NIROT, the detectable photon flux at the boundaries can be heavily attenuated in comparison to the source signal. As seen in Figure 1.2, moving away from the source by a few centimetres results in a drop in intensity of a few orders of magnitude. The likelihood that a photon incident on a SPAD array will be detected is given by the photon detection efficiency (PDE), where PDE = PDP $\cdot$ FF, and FF is the pixel fill factor. Thus, for monolithic image sensors there is a tradeoff between pixel functionality and PDE. More advanced functionality such as active quenching and recharge [71], which would reduce the impact of afterpulsing and crosstalk is only achieved at the expense of reduced PDE.

One approach to reduce the footprint of the pixel circuitry is to use only NMOS transistors, which mitigates against the large spacing requirement for the PMOS transistor nwell diffusions. However, whilst small pixel pitch, high fill-factor arrays [72, 73] have been produced with this technique, the pixel functionality is limited. Another approach to reducing the amount of area occupied by in-pixel circuitry is to design sensors in more scaled technologies. However, due to the high doping concentrations of the standard layers in scaled technologies, e.g. 65 nm, it is difficult to produce high performance SPADs without custom process layers [74].

2.3.2 Signal-to-noise ratio (SNR)

As discussed in section 2.2, the choice of excess bias is a balance between PDP, or PDE if we consider also fill factor, timing jitter and DCR. Both PDP and timing jitter are improved with increasing V_{EB}, however DCR will also increase. The SPAD metric which links these three parameters to the measurement quality is the signal-to-noise ratio (SNR), which for timing

measurements is given by [75]:

$$SNR = \frac{S_{PEAK}}{\sqrt{S_{BGND}}} = \frac{PDE \cdot \Phi_S}{k \cdot FWHM} \cdot \frac{1}{\sqrt{DCR}}, \tag{2.1}$$

where Φ_S is the signal photon rate, $k = \pi/(4\ln 2)$. This derivation defines the SNR as the ratio of the peak of the timing response, S_{PEAK} divided by the square root of the background noise, S_{BGND}. Whilst this may be valid for an application like light detection and ranging (LiDAR), where we wish to measure only the distance to an object, in NIROT information from the medium is encoded in the entire TPSF. Furthermore, information from depth is obtained from late arriving photons, thus to accurately reconstruct objects at depth, we also care about the SNR in the tail of the response.

Often Equation 2.1, does not tell the complete story at the system level. For example, in distance ranging measurements the ambient light can be many times greater than the DCR. Similarly, a clinical NIROT system will likely have some ambient light, e.g. light from the displays of other medical devices. Assuming this ambient light is poissonian, which is very often the case [76], Equation 2.1 becomes

$$SNR = \frac{PDE \cdot \Phi_S}{k \cdot FWHM} \cdot \frac{1}{\sqrt{DCR + PDE \cdot \Phi_A}}, \tag{2.2}$$

where Φ_A is the rate of ambient photons. Therefore, as the ambient light level increases, the influence of DCR reduces and the ratio $(PDE)/(FWHM \cdot \sqrt{PDE})$ can be maximised to reach the optimum SNR.

2.3.3 Dynamic range (DR)

The dynamic range (DR) is the ratio of the maximum detectable signal, S_{MAX}, to the minimum detectable signal, S_{MIN}. Since the NIROT measurements can require a dynamic range of more than 5 orders of magnitude [35], its maximisation is critical. The DR is given by [75]:

$$DR = \frac{S_{MAX}}{S_{MIN}} = \frac{\Phi_{MAX} \cdot T_{INT}}{\sqrt{DCR \cdot T_{INT}}}, \tag{2.3}$$

where

$$\Phi_{MAX} \approx \frac{1 - P_{AP}}{T_{DEAD}}. \tag{2.4}$$

T_{INT} is the integration time for the measurement, whilst Φ_{MAX} is the maximum photon rate where P_{AP} is the afterpulsing probability and T_{DEAD} is the SPAD dead time. The appearance of T_{INT} in both the numerator and denominator indicates that the dynamic range can be improved by increasing the measurement period. Whilst this is true, and also relies on adequate background subtraction, long measurement times in the clinical NIROT application will introduce additional sources of errors, e.g. motion artifacts. Furthermore, Equation 2.4 demonstrates how the dynamic range is degraded by afterpulsing and the benefit of a short dead time. Therefore, there appears to be a clear benefit of active quenching and recharge circuits for high dynamic range pixels. Since the inclusion of this functionality implies the reduction of fill factor in monolithic image sensors, there is a pressing need for technological solutions which enable such functions without a negative impact.

2.4 Stacked 3D BSI Image Sensors

A recent trend towards improving SPAD imager performance and functionality is to produce image sensors in 3D IC technologies. This means that the SPADs and data processing circuitry can be manufactured on separate dies and then bonded together with dense metal interconnects between the two dies. The first implementation of a 3D IC SPAD imager was achieved by Lincoln Labs [77], which included a FSI SPAD, or geiger-mode avalanche photodiode (GmAPD), with two tiers of circuitry for SPAD interface and timing circuitry. This technique was not, however, scalable to pixel pitches less than 50 μm. The first implementation of a BSI 3D IC SPAD imager was achieved in [78] with a 130 nm 3D IC technology achieving a pixel pitch of 11.75 μm. This technology is illustrated in Figure 2.6. In this case, the die with the photodetectors is thinned to 4.2 μm to maintain sensitivity at visible wavelengths. It is the BSI 3D IC technology which is now being most heavily developed.

Opportunities

There are several advantages for designing SPAD image sensors in BSI 3D technologies. Firstly, the fill factor can be much higher in comparison to monolithic FSI sensors as most, if not all, of the pixel circuitry can be placed on the data processing tier. Additionally, many applications,

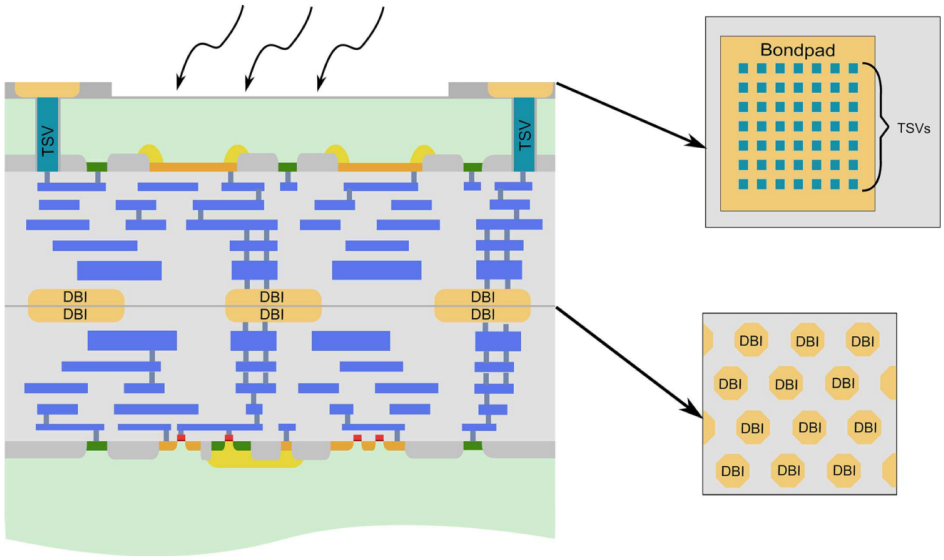


Figure 2.5 – Cross section of BSI 3D IC 130nm CMOS technology from [78]. Connections from wirebonds to I/O pads are made with through-silicon vias. The photodetector and data processing tiers are bonded together face-to-face via a 4 μm pitch grid of direct bond interfaces (DBI) on both dies. The photodetector wafter is thinned to 4.2 μm to increase sensitivity in visible wavelengths. *Picture credit: J. Mata Pavia, ©2015 IEEE.*

such as NIROT and LiDAR utilize wavelengths in the red and near infrared range, 650-900nm. In BSI sensors, the peak PDP is shifted to longer wavelengths, since photons must travel through a thicker silicon region before reaching the SPAD multiplication region. Indeed, with a BSI substrate thinned to 4.2 μm and a multiplication region of 4 μm, a peak PDE at 700 nm is achieved [78].

The photodetector tier could also be designed in an optimized SPAD technology [75], which typically outperform CMOS processes. The option of having a dedicated photodetector tier also raises the possibility of having a single data processing tier which can be bonded to different photodetector tiers targeting different applications. For example, InGaAsP SPADs are being investigated as a possible technology for airborne LiDAR [79] whilst silicon should suffice for consumer LiDAR, e.g. camera autofocus, gesture recognitition, etc.

There are also clear benefits for the data processing tier. Although this was not the case in [78], different technologies could be chosen for the photodetector and data processing tiers. This allows the pixel interface and timing circuitry to be implemented in a low power, high density, digital technology, e.g. 65nm CMOS, where SPAD performance is typically worse, complicating the design of FSI imagers. Leveraging an advanced CMOS technology on the data processing tier will allow advanced functionalities, such as on-chip histogramming [80]

or active recharge [81, 71], to be implemented on a much larger scale.

Challenges

Despite the myriad of benefits for designing a SPAD imager in 3D IC BSI technologies, there are also a number of challenges posed. Firstly, although decoupling the selection of the SPAD and data processing technologies is certainly an advantage, the choice of technology for the data processing tier will in most cases place a constraint on the excess bias voltage of the SPAD. Due to the advantages of using a high density digital process for the data processing tier, e.g. lower power consumption, reduced cost per function etc., the choice of technology in most, if not all, cases will be a highly scaled technology. Thus, V_{EB} will be limited to the maximum voltage tolerance of the thickest oxide transistors in that technology. At nodes from 65nm and below, this is typically just 2.75 V. Furthermore, although it is not within the scope of this thesis, sensitivity at shorter wavelengths, e.g. 400 nm, will be reduced due to the shift in PDP. Finally, due to the expected dense arrangement of SPADs as array formats target megapixel resolutions, crosstalk is likely to be more prevalent. This will require novel circuit and technology solutions.

2.5 A High-PDE Pixel with Cascoded Quenching and Active Recharge

The shift in peak PDP towards the NIR spectrum and additional functionality possible in a high density digital process makes a BSI 3D IC technology an outstanding candidate for a NIROT system. However, as discussed briefly in section 2.4, the gains in PDE when designing in such a technology may not be as large as expected due to the limitation on V_{EB}. That is, the gate-source (V_{GS}), gate-drain (V_{GD}), and drain-source (V_{DS}) voltages of a transistor, which is used to quench the SPAD, should not exceed their long term reliability voltage. For thick oxide transistors in highly scaled technologies this is usually just 2.75 V. There are several promising SPAD designs which can still achieve a significant improvement in PDP and timing jitter above V_{EB} = 2.75V with only a moderate rise in DCR [82, 83]. Thus, there is a pressing need for quenching and recharge circuits which exceed this limit. This section reports the design of a SPAD pixel, which can operate at excess biases up to 4.4 V, thus increasing the SPAD PDE, and timing performance, without high voltage process options [84] or area intensive polysilicon resistors [82].

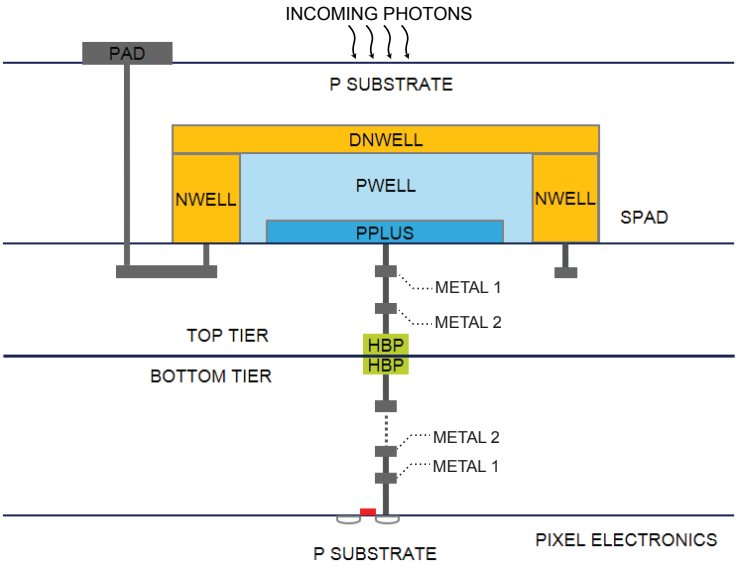


Figure 2.6 – p-well/deep n-well SPAD [86] in 65/40 nm 3D IC CMOS technology. ©2017 IEEE.

2.5.1 SPAD Structure and 3D IC Technology

The pixel is implemented in a BSI 3D IC CMOS technology, where the data processing tier and image sensing tiers are implemented in 40 nm CMOS and a standard 65 nm BSI image sensor process technology, respectively. Inter-tier connections are made with hybrid bonding pads (HBPs) using a wafer level hybrid bonding process, as described in [85]. The SPAD multiplication region is formed between the p-well and deep n-well implants, as implemented in [86], see Figure 2.6. The choice of a 65 nm technology for the photodetector tier will likely result in worse DCR, as discussed in section 2.2. However, the 65 nm BSI imaging technology is expected to suit a wider range of high volume applications, many of which will have requirements for smaller pixel pitches or alternative photodetectors. Better SPAD performance could be expected from a less scaled or custom SPAD technologies. Indeed, there is no inherent incompatibility in selecting these technologies, thus future designs may see custom SPADs coupled to high density digital data processing tiers.

2.5.2 Pixel Design

The pixel circuit is shown in Figure 2.7. M_1-M_5 are thick oxide transistors which operate with a nominal supply voltage of 2.5 V. The voltage range of the interface is extended by connecting a cascode transistor, M_2, in series with M_1, as implemented for other applications in [87, 88]. To ensure the range of this interface is maximised, the gate of M_2 is biased at $V_{OX,MAX}$. This is

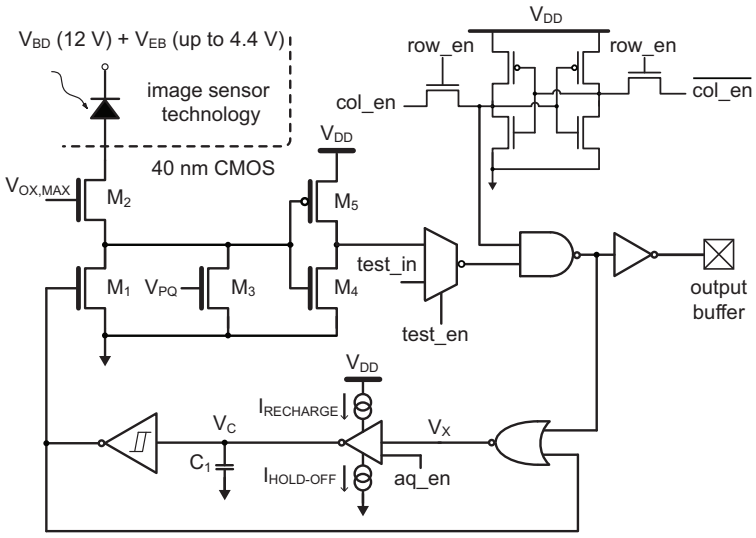


Figure 2.7 – Cascoded active recharge SPAD pixel schematic. ©2017 IEEE.

the maximum voltage the gate oxide can handle reliably, in this case $V_{OX,MAX} = 2.75V$. Passive quenching functionality for testing purposes is provided by M_3.

As well as utilizing thick oxide transistors to implement the cascode, this quenching scheme could also be designed with thin oxide devices. In this case, the maximum excess bias would be close to that of a single thick oxide device but with a reduced circuit area due to the smaller transistor sizes and avoiding the large spacing requirement between oxides of different thicknesses.

The feedback loop consisting of a NOR gate, an inverter charging a MOSCAP $C_1 = 5.7$ fF and a schmitt trigger implements active recharge functionality. Whilst this loop is similar to [71], here, the charging ($I_{RECHARGE}$) and discharging ($I_{HOLD-OFF}$) currents of the inverter are supplied via two independently controllable wide-swing cascode current mirrors, see Figure 2.8. An off-chip reference defines the input currents to the current mirrors, $I_{BIAS,N}$ and $I_{BIAS,P}$. This allows $I_{HOLD-OFF}$ to be configurable over a range 42 nA - 1 μA, resulting in a dead time range of 8-100 ns. This range enables us to explore small dead times and low afterpulsing.

The main advantage of charging and discharging C_1 with switched current mirrors is that no area intensive resistors are used to set the hold-off time [71]. Furthermore, static power dissipation [81] is limited to the bias distribution for the current mirrors, which can be shared among a large number of pixels. Therefore, it is highly suitable for large arrays which have a small pixel pitch.

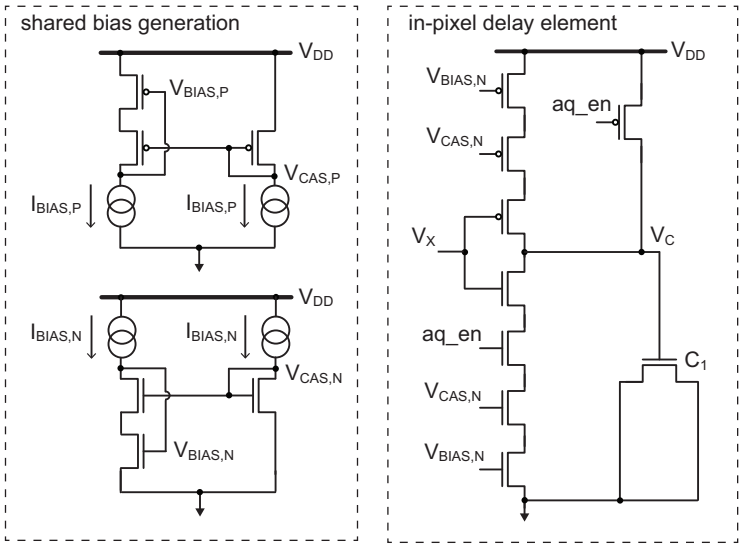


Figure 2.8 – Hold-off and recharge time, delay generation.

Simulation results illustrating the operation of the circuit in Figure 2.7 are shown in Figure 2.9. Upon the detection of a photon, the fast rising voltage at the drain of M_1, which is off, forces M_2 into the cut-off region. The SPAD anode is charged very quickly to the excess bias voltage, V_{EB}. The anode voltage can be up to 4.4 V, since it is distributed between M_1 and M_2, ensuring that the gate-source (V_{GS}), gate-drain (V_{GD}) and drain-source (V_{DS}) voltages of both are kept under 2.75 V. Thus, the inverter formed from M_4 and M_5 and used to detect the rising edge at the drain of M_1 is also protected. The anode voltage is held at V_{EB} due to the high impedance of M_1, M_2 and M_3, which quenches the avalanche.

Upon photon detection, $I_{HOLD\text{-}OFF}$ is used to discharge C_1. When the voltage crosses the threshold of the schmitt trigger, $V_{GS,M1}$ rises to V_{DD} = 1.1 V. $V_{DS,M1}$ then begins to drop, increasing $V_{DS,M2}$ until $V_{GS,M2}$ is greater than its threshold voltage and turns the device on. The result is a peak in $V_{DS,M2}$ which then determines the maximum achievable excess bias, assuming a hard device reliability limit where V_{DS} must be less than 2.75 V. The SPAD anode is then discharged to ground through M_1 and M_2 until C_1 is charged to the upper threshold of the schmitt trigger with $I_{RECHARGE}$. Now $V_{GS,M1}$ falls to zero and the SPAD is recharged and ready for another detection. If a photon detection occurs during the M_1 "on" period, the avalanche will cause the anode voltage to rise again towards 4.4 V. The avalanche will be quenched once the schmitt trigger crosses the upper threshold and M_1 is turned off. After M_1 is switched off, the loop immediately enters another hold-off period before recharging the SPAD again.

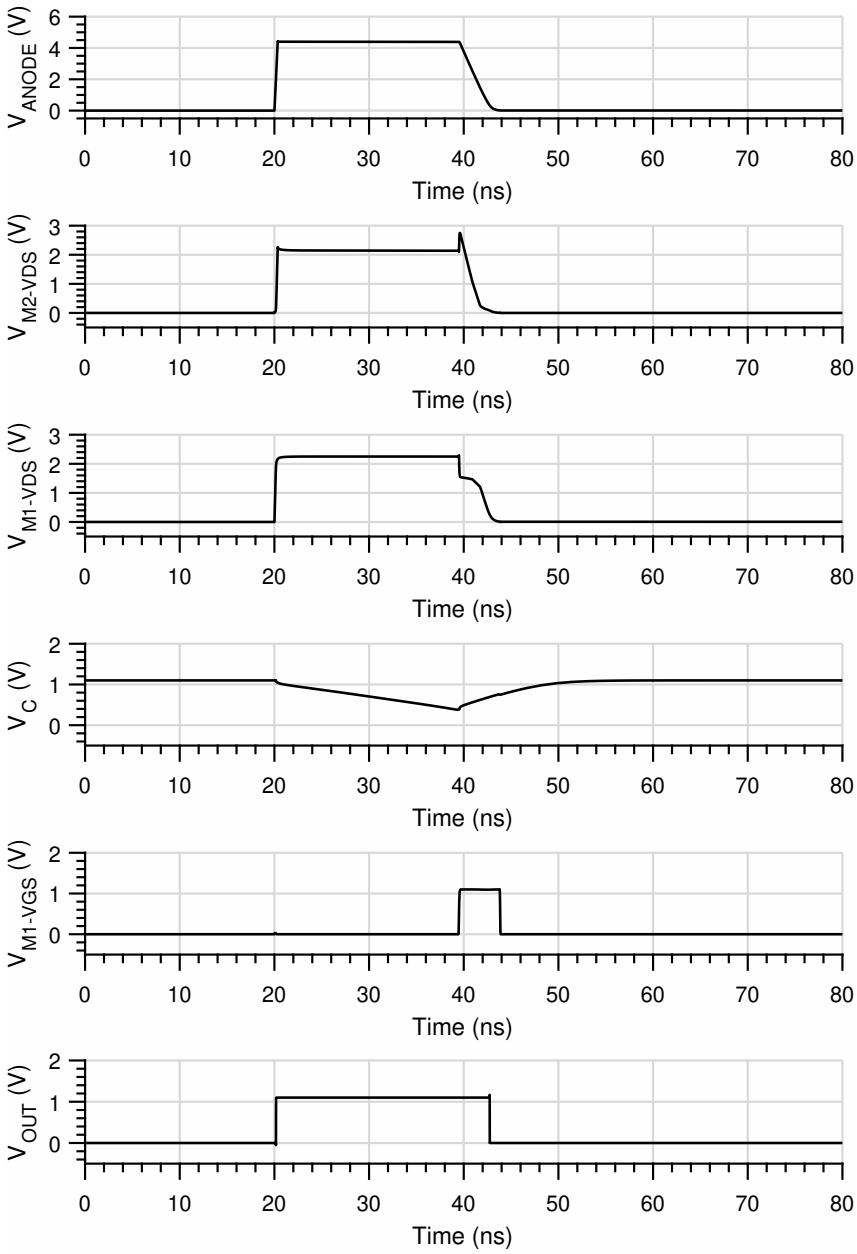


Figure 2.9 – Avalanche quenching and recharge simulation showing critical circuit voltages. ©2017 IEEE.

2.5.3 Results and Discussion

The pixel was manufactured in the BSI 3D IC technology discussed in Section 2.5.1. The data processing and photodetection tiers are fabricated in 40 nm CMOS and 65 nm BSI image sensor processes, respectively. Figure 2.10a shows a micrograph of the BSI SPAD pixel and an adjacent test pixel with a separating gap of 11 μm. The metal connections to the anode and

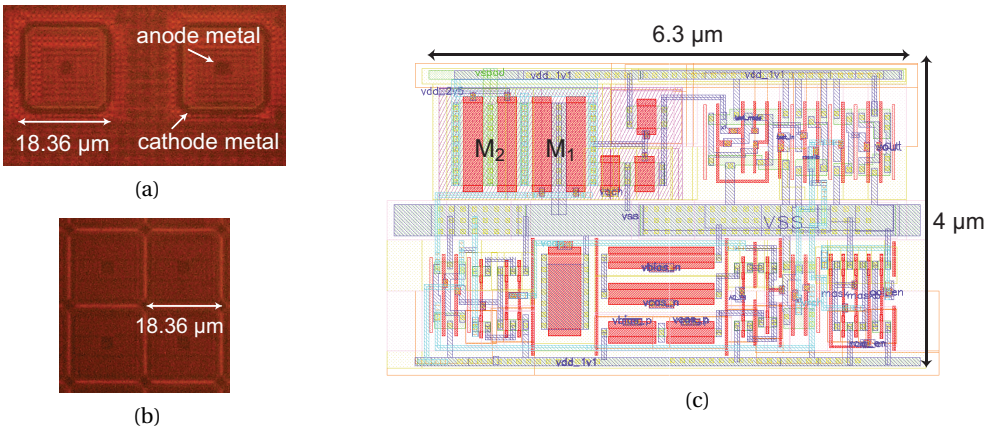


Figure 2.10 – (a) Pixel test structures micrograph showing two 18.36 μm pitch pixels separated by 11 μm gap, (b) Array format, fill-factor is 74.37% (c) Pixel circuit layout. ©2017 IEEE.

cathode correspond to the darkened rings around the SPAD and the square in the middle of the device, respectively. The pixel circuit consumes an area of 4 μm × 6.3 μm and is connected to a 250.7 μm^2 active area "Fermat" [89] shaped SPAD. When laid out in array format with an 18.36 μm pixel pitch, see Figure 2.10b, a fill factor of 74.37% is achieved. The layout of the pixel circuit can be seen in Figure 2.10c.

The DCR was measured at 25 °C on 4 separate devices with V_{EB} increasing to 4.4 V, see Figure 2.11. Although the DCR is high in comparison to state-of-the-art devices, process refinement is expected to improve this performance significantly. Furthermore, as discussed in 2.3.2, the background noise contribution in certain applications, e.g. LiDAR, means that DCR is less of an issue, even when using narrow near-infrared optical bandpass filters.

Afterpulsing probability is investigated by measuring the inter-avalanche times, shown in Figure 2.12. The dead time of the SPAD is set to 8 ns, whilst V_{EB} is at the maximum of 4.4 V. The probability of afterpusling is then calculated by determining the deviation from the exponential fit, also shown in Figure 2.12. This is calculated to be 0.08%, the lowest reported afterpulsing with 8 ns dead time for a BSI SPAD to date. This demonstrates that photon counting rates up to 125 Mcps could be achieved with this pixel and technology.

PDP measurements for the 400-950 nm wavelength range are shown in Figure 2.13. This demonstrates an improvement of approximately 27% by increasing V_{EB} from 2.75 V to 4.4 V. The estimated improvement at 450 nm is even more pronounced at 316%. The peak PDP of 29.5% at 660 nm results in a peak PDE of 21.9%. To the best of our knowledge, this is the highest reported for a BSI SPAD in a 3D IC sensor to date.

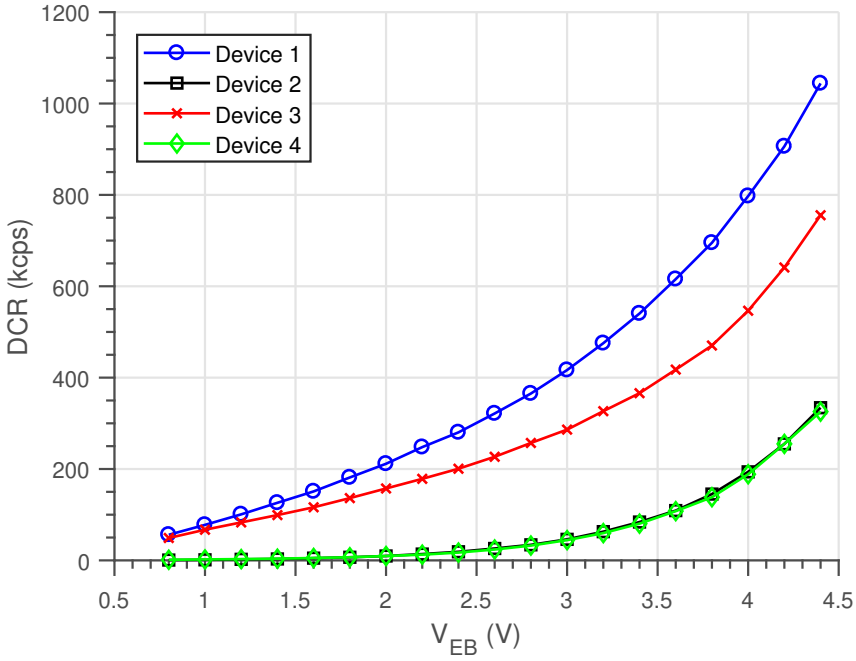


Figure 2.11 – Dark count rate vs excess bias. ©2017 IEEE.

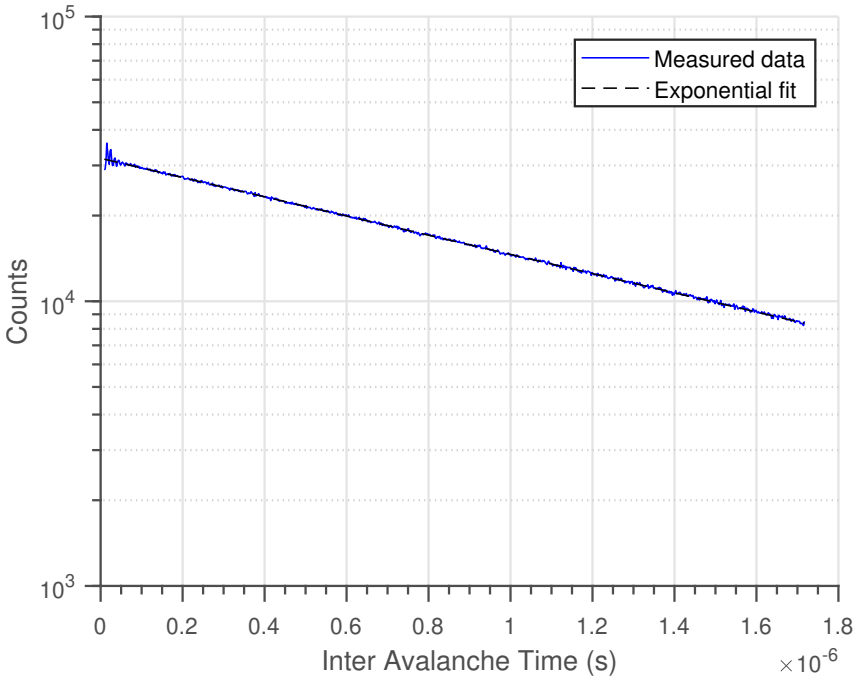


Figure 2.12 – Inter avalanche arrival time with 8 ns dead time. ©2017 IEEE.

Timing jitter measurements were performed by illuminating the pixel with a combination of a supercontinuum laser and acousto-optical tunable filter (AOTF), configured to output

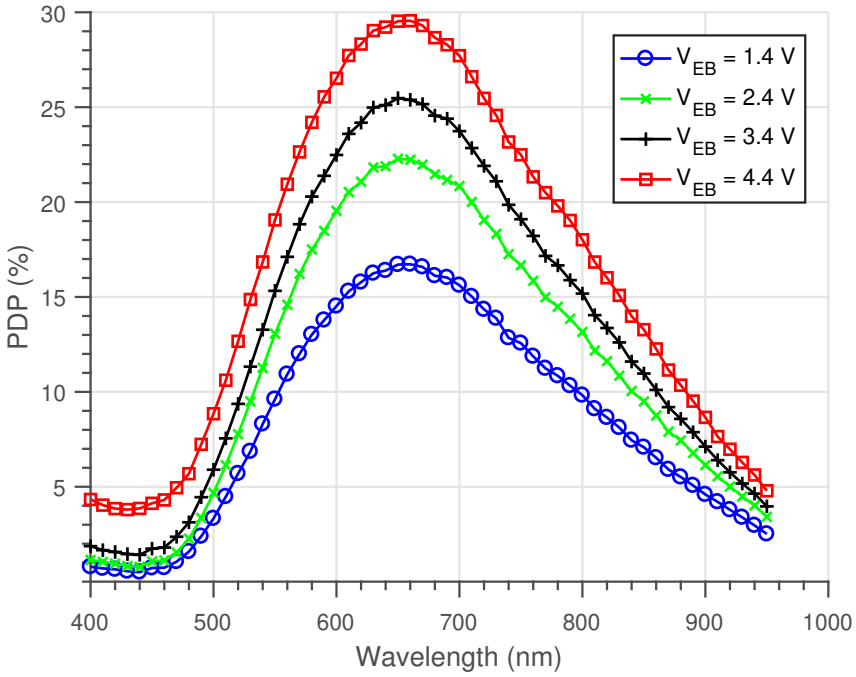


Figure 2.13 – PDP vs wavelength for varying excess bias. ©2017 IEEE.

pulsed laser light at 700 nm. An oscilloscope (Lecroy Waverunner 204MXi-A), was employed to measure the time interval between the SPAD pulses and the edges of the electrical trigger signal from the laser. SPAD activity was limited to less than 1 photon detected per 100 laser pulses to avoid pile-up. Figure 2.14 shows the timing response of the pixel for a range of excess biases. The increase in DCR as V_{EB} is increased can be clearly observed in the noise floor for the different measurements. The FWHM of the timing responses in ps are 122 (1.4 V), 114 (2.4 V), 104 (3.4 V) and 95 (4.4 V). This includes the FWHM of the laser (= 20 ps) and the jitter of the output buffer (FWHM = 55 ps). Subtraction of these additional sources of jitter results in FWHM timing responses in ps of 107 (1.4 V), 98 (2.4 V), 86 (3.4 V) and 75 (4.4 V).

In comparison to the existing 3D IC BSI SPAD pixels [78, 90], which include only passive quenching and recharge and a thick oxide inverter to shift the voltage level, this pixel contains significantly more devices. The increase in number of devices, however, is largely due to the active recharge functionality rather than the extension of the V_{EB} range. The cascode scheme could also be implemented with passive quenching and recharge, in which case it only requires one extra transistor in comparison to conventional schemes.

To the best of our knowledge this is the first all-transistor SPAD pixel capable of operating at excess biases above the voltage rating of a single transistor without exceeding the voltage reliability limits of any device. Although this improves the PDP and timing jitter, for the

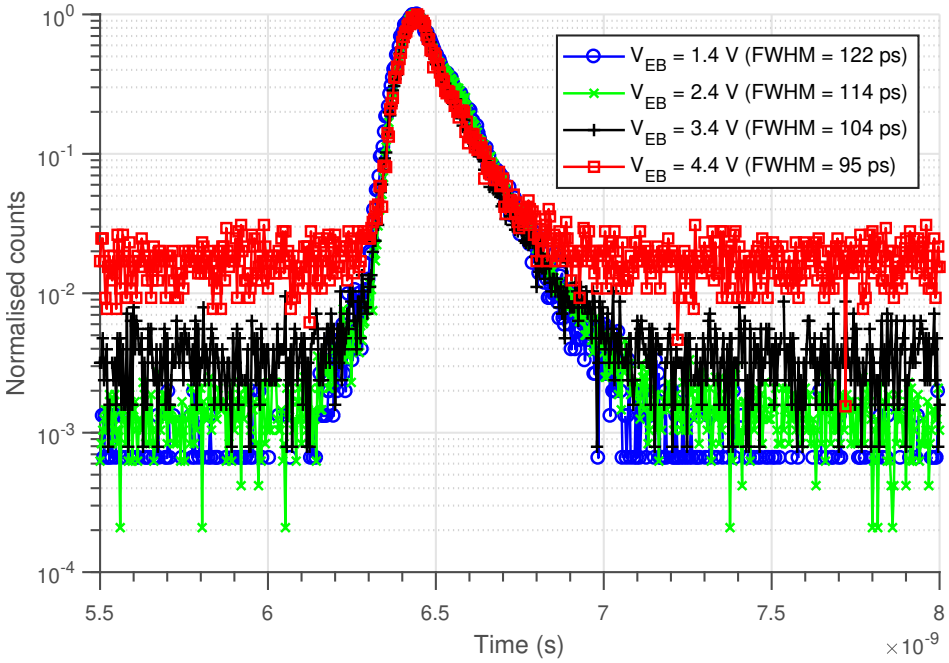


Figure 2.14 – Timing response for varying excess bias at 700 nm. ©2017 IEEE.

employed SPAD the DCR also increases significantly. There are however higher performance SPADs where this technique could be highly beneficial for improving SNR. Most importantly, this circuit is by no means limited to implementation in 3D IC technologies, it is applicable to all SPAD imagers.

The active recharge capability enables higher count rates than can be achieved with passive recharge. The dead time of 8 ns enables a maximum count rate of up to 125 Mcps. Thus, this circuit could enable a dynamic range of over 5 orders of magnitude as mentioned without an excessively long measurement time. Furthermore, the extremely low afterpulsing indicates that the dead time could be reduced further. Despite its higher complexity and extra features, e.g. electrical masking with a NAND and 6T-SRAM, this pixel design would still leave 59% of the pixel area free for circuitry such as counters when using the 7.83 μm pixel pitch from [90].

2.6 A Bidirectional Pixel for 3D IC Technologies

One of the great opportunities for SPAD imagers designed in 3D IC technologies is the possibility of interfacing a single data processing tier with different sensor tiers, each one with an SNR optimized for a specific wavelength range. The challenge with this approach is that SPADs with high sensitivity at longer wavelengths [82, 91] require a different quenching interface to those

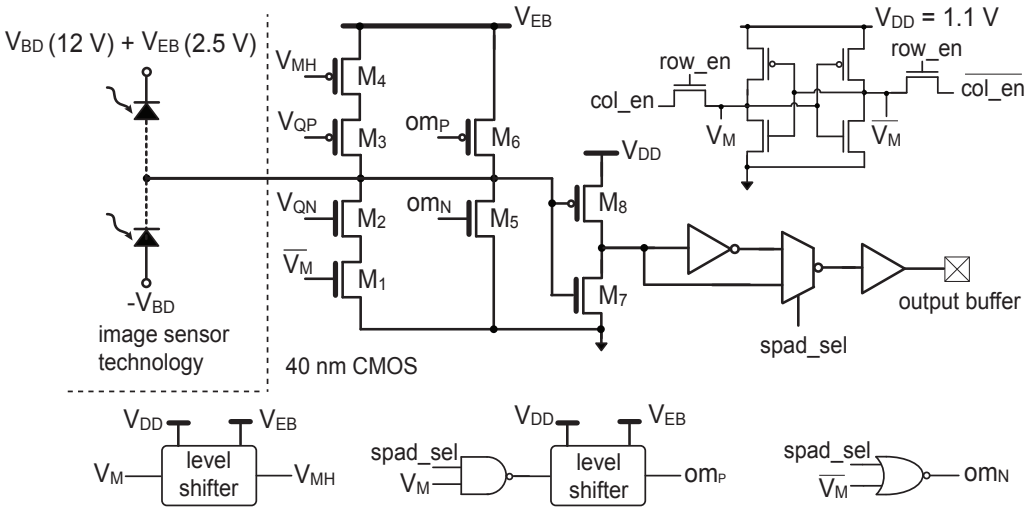


Figure 2.15 – Bidirectional passive quenching circuit schematic.

with high sensitivity at shorter wavelengths [86]. For the former, the substrate of the sensor tier should be biased at a negative voltage and the SPAD is quenched with a PMOS transistor. Although the latter can often be quenched with a PMOS transistor due to the isolated n-well, if a shared well is used to maximize fill-factor [90] these devices can only be quenched with an NMOS transistor. This section reports the design of the first SPAD quenching and recharge circuit for 3D IC technologies which can interface with SPADs in any biasing configuration.

2.6.1 Pixel Design

The pixel circuit is shown in Figure 2.15. For a SPAD which is biased at the anode (cathode), an avalanche is quenched by the branch consisting of M_1 and M_2 (M_3 and M_4). The dead time of the SPAD can be controlled by changing the resistance of M_2 (M_3) through V_{QN} and V_{QP}. If the SPAD is very noisy in comparison to the rest of the SPAD population in an array, it is often desirable to reduce the voltage across the SPAD to be equal to its breakdown voltage such that it does not produce an avalanche; it is optically masked and does not contribute to optical crosstalk.

The NAND and NOR gate in Figure 2.15 take as input a signal spad_sel, which represents the SPAD biasing configuration, spad_sel = '1' for anode biasing and '0' for cathode biasing. The remaining input is based on the masking configuration of the SPAD, which is stored locally in a 6T-SRAM. When the SPAD is not masked, both M_5 and M_6 are off and appear as a high impedance to the SPAD. When the SPAD is masked, both outputs om_P and om_N are set to either V_{EB}, or ground depending on the SPAD biasing configuration, which turns one of the

transistors on. The voltage at the gates of M_4 and M_6 are produced by a level shifter as their wells are connected to V_{EB} (up to 2.75 V), to avoid negative voltages when biasing a SPAD by its cathode. In the active masking condition, the transistors M_1 and M_4 are switched off to avoid a static current leakage path through the primary quenching transistors, M_2 and M_3. This functionality is very important for multi-megapixel arrays, where the static current consumption due to masked pixels through the quenching transistors without a series device could reach some 100s of milliamps. A thick oxide inverter formed by M_7 and M_8 scales the output signal to V_{DD}. The spad_sel signal is used to select whether to output the inverted or non-inverted output signal, such that the edge produced for a photon arrival is not dependent upon the SPAD biasing configuration.

2.6.2 Results and Discussion

The circuit occupies an area of 6.5 μm ×4.3μm, and was implemented in the same technology as in section 2.5. Two instances of the circuit were fabricated, each coupled to its own 17.78 μm active area diameter backside-illuminated p-well/deep n-well SPAD. One version was connected to the SPAD via the anode whilst the other was connected to its SPAD at the cathode. Due to the already extensive characterisation of the SPAD in section 2.5, the results shown here are merely a check of functionality for the dual biasing configuration. A sweep of V_{QN} and V_{QP}, whilst the opposing branch is off ($V_{QP} = V_{EB}$ or $V_{QN} = 0$ V), is shown in Figure 2.16. This demonstrates the range of achievable SPAD dead-times in each configuration.

As seen in Figure 2.16, down to 50 ns dead time can be achieved in both biasing configurations. To the best of our knowledge, this is the first quenching and recharge circuit which is capable of interfacing to a SPAD in both anode and cathode configurations. Thus, rather than requiring a data processing tier redesign when targeting a different wavelength, e.g. when an alternative quenching scheme is required, it is possible to use a circuit such as Figure 2.15 for both. Each sensor could then use a photodetector tier whose SNR is optimised for its own target application. Since the mask set for a high density digital process represents a large non-renewable expenditure, this may prove an attractive avenue for producing 3D sensors in the future.

2.7 Conclusions

The presented cascoded quenching and recharge pixel is, to the best of our knowledge, the first all transistor circuit capable of operating at excess bias voltages above the voltage tolerance

of a single transistor. The all transistor implementation makes the circuit compact whilst at the same time remaining compatible with both passive and active quenching and recharge schemes.

Indeed, implementation with an active recharge scheme yielded a pixel capable of dead times down to 8 ns, with just 0.08% afterpulsing probability. With static power dissipation of the active recharge circuit limited to bias circuitry, this combination of cascoded passive quenching with active recharge paves the way for BSI 3D IC imagers with unprecedented SNR and dynamic range.

To the best of our knowledge, the bidirectional passive quenching circuit is the first implementation of a SPAD quenching and recharge circuit which can interface to a SPAD in both configurations. It raises the possibility of interfacing a single data processing tier with a number of different application specific photodetector tiers.

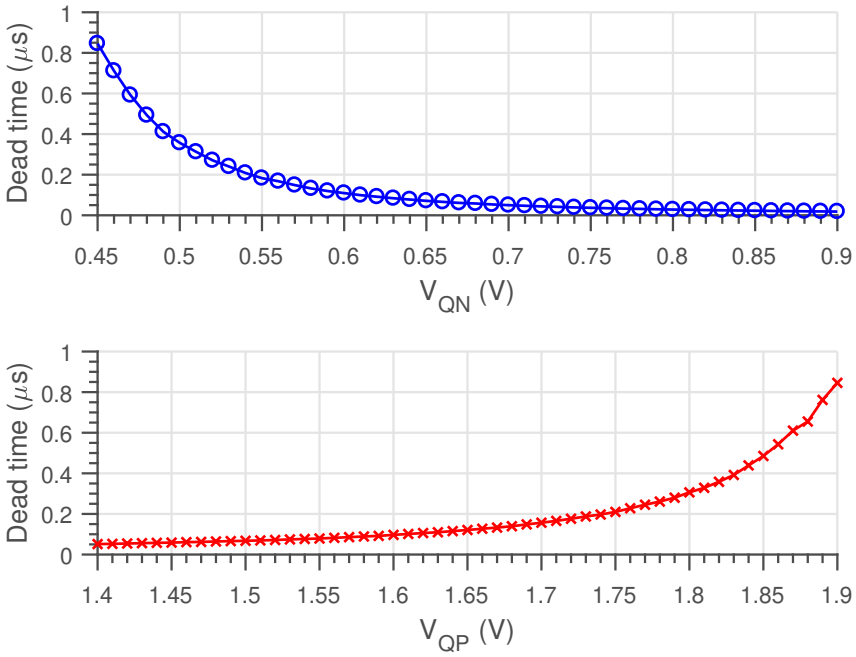


Figure 2.16 – SPAD dead time versus quenching voltage for anode (top) and cathode (bottom) biasing, V_{EB} = 2.5 V.

3 Event-driven Time-resolved SPAD Sensors

A major attraction of using CMOS SPADs in the NIROT application is the possibility of acquiring time-resolved data with a high number of parallel channels. This large volume of time-resolved data is key to improving the spatial resolution of the image reconstruction. Thus, time-resolved SPAD cameras could lead to powerful new clinical NIROT systems which are widely adopted. A major stumbling block in the road to achieving this goal is the design of the SPAD sensor. Acquiring huge volumes of time-resolved data on a timescale that is compatible with clinical measurements is not trivial. This chapter begins with an introduction to time-resolved SPAD cameras with the goal of identifying the major issues facing the development of high throughput sensors. Requirements for the design of a sensor dedicated to NIROT are then outlined. A new highly parallelised sensor architecture is proposed and then analysed in comparison to the dominant TDC-per-pixel architecture.

3.1 Time-resolved SPAD Image Sensors

Despite the relative simplicity of TDC circuits, e.g. in comparison to an analog-to-digital converter, the silicon area occupied by these circuits is generally thousands of μm^2 at nodes above 90nm. It is not surprising then, that the first SPAD sensor to include TDCs on the same die as the SPADs allocated just 32 TDCs to an array of 128×128 pixels [58]. In this case, an image from the entire sensor is acquired by measuring from every row in a serial fashion. Thus, the obvious downside to this architecture is a long acquisition time.

3.1.1 TDC-per-pixel Sensor Architectures

The natural progression from this first implementation was to design SPAD sensors with a TDC in each pixel [59, 92], with later designs such as [93, 94] and finally [95], which didn't include a TDC inside each pixel but each pixel had a dedicated TDC outside of the pixel. These sensors have the advantage that they are fully parallel. This means that when a SPAD detects a photon there is a TDC in close proximity to measure the time of arrival. The downside of this approach is that the large area consumed by the TDC results in a low fill factor. In the reported designs, fill-factors in the range 1-3% are achieved in pixel pitches in the range of 50-150 μm. The low fill factor can be improved by using microlenses, to concentrate the incoming light onto the active area of the pixel, however even with a concentration factor of ×5 [92], the fill factor remains low.

A further disadvantage of the TDC-per-pixel approach is that for any given clock cycle, the detected photons are sparsely distributed across the array. This is a side effect of the fact that for TCSPC experiments, the average number of photons detected per cycle should be less than one. The actual limit on activity is dependent on the application. A study of pile-up in [96], quotes an activity percentage of 10-20% but assumes that a few percent error in the calculation of lifetime in a fluorescence lifetime imaging (FLIM) experiment can be tolerated. The general consensus is that pileup becomes noticeable at activity rates of a few percent [29]. Assuming an activity rate of 3%, this means that for every cycle of the laser, 97% of TDCs in an array would detect nothing.

As well as being an inefficient use of circuit area, data sparsity presents a major challenge for data readout since most of the data stored in the array at the end of a cycle is null, i.e. *no photon detection was recorded*. If these null events are not eliminated before communication off chip, then the achievable photon throughput is much reduced, since most of the I/O data bandwidth is consumed by null events. In [95] an event driven datapath is designed to eliminate these null events, achieving a theoretical maximum throughput of 2 Gevents/sec with an output data bandwidth of 42 Gbps. However, power consumption in this case is a massive issue. The reported I/O and core power consumption figures are 2 W and 6 W, respectively, the latter largely due to the 1 GHz clock required for the event-driven datapath. Since this implementation is for a 64 × 64 array, scaling to larger formats would increase the core power consumption further. The resulting system requires a water cooling system to reduce DCR to an acceptable level during measurements.

Finally, a problem which is not isolated to TDC-per-pixel sensors but is rather a problem for

time-resolved sensors in general, is that they have the potential to create a huge volume of data. For example, consider the 160 × 128 sensor presented in [59]. When operating with an 80 MHz laser, as would ideally be used in the NIROT application, at 3% pile-up activity limit, with 10 bits of data representing each photon, the data throughput, D_{TP}, is

$$\begin{aligned} D_{TP} &= (160 \times 128) \cdot (80 \times 10^6) \cdot 0.03 \cdot 10 \\ &= 491.5 \text{ Gbps.} \end{aligned}$$

With 320 I/O pads operating at 160 MHz, the total data bandwidth in [59] is 51.2 Gbps, almost an order of magnitude less than would be required to reach the activity limit for pileup. Assuming a higher bandwidth protocol such as LVDS was available, using [95] as an example, 491.5 Gbps would require 983 signal pairs with a total I/O power consumption of 22.3 W, both highly impractical numbers. It is clear then, that if the data must be transmitted off-chip in its uncompressed raw format, the I/O bandwidth presents a bottleneck which limits the pixel activity, and also the dynamic range of the sensor.

Thus, when considering the design of a high throughput SPAD imager for NIROT, there are three key lessons to learn from the TDC-per-pixel approach which is common in monolithic sensors:

1. TDCs whilst being relatively simple, occupy a significant silicon area. When included in the pixel, the resulting sensor will have either a very low pixel fill factor, or very large SPADs and pixel pitch.
2. The pile-up limit in TCSPC systems means that at the end of a laser illumination period, very few pixels will have detected a photon. This is not only an inefficient use of silicon area but also poses a major challenge for the readout, since real timing data is only sparsely distributed amongst a mass of null events.
3. Operating at pixel activity levels which approach the pile-up limit for large sensors and high frequency lasers is impractical due to the required bandwidth and power to transmit the raw data. For sensors which transmit raw data in this way, the pixel activity in many cases will be determined by the I/O bandwidth.

3.1.2 TDC Sharing Architectures

In recent years, a number of groups have begun to explore alternative architectures where a TDC can be addressed by multiple pixels. In [97], the outputs of each pixel drive one or more

NMOS transistors. These pull to GND a number of common bus lines which are set at VDD when idle through pull up resistors, see Figure 3.1. Thus upon the detection of a photon, the pixel transmits its timing signal nOUT and address, which in this case is binary encoded. The address is captured at the end of the bus via a latch, which is clocked by a delayed version of nOUT. The simplicity of this scheme enables a number of pixels to share a TDC whilst achieving a high pixel fill-factor. The main disadvantage of this scheme is that since there is no pulse-shrinking of the SPAD output signal before driving the bus, each event will occupy the bus until the firing SPAD voltage recharges to the threshold of the inverter. Furthermore, there is some ambiguity in applying binary coding to such a bus. For example, assuming a 5-bit address bus, should the pixels with address codes '10111' and '01111' fire in close proximity, the latched address will be '00111', a pixel that didn't fire. This aspect of the bus is improved in [78], which uses a sharing bus with a collision detection coding scheme. This scheme requires more bus lines to encode the same number of pixels, however, code collisions between two or more pixels always result in invalid codes. In this case, the maximum number of pixels that can be encoded by a bus with n lines is given by

$$\text{number of codes} = \frac{n!}{k!(n-k)!} \tag{3.1}$$

where k is the integer closest to $n/2$.

The other main technique for sharing a TDC with a number of pixels is the use of CMOS static logic trees, generally constructed from OR [98] or XOR [80] gates, the latter achieving a higher dynamic range [99]. Whilst these tree based sharing schemes have demonstrated efficient sharing of TDCs with pixel fill-factors of up to 43% [80], they have typically been limited to digital-SiPMs sensors, where the timing data is accumulated to construct a single histogram for an array of SPADs. Indeed, for architectures which utilize this tree structure, each signal which must propagate to the TDC requires (N – 1) 2-input logic gates, where N is the number of pixels sharing the TDC. Thus, to implement a 32-pixel column with the collision free coding scheme reported in [78], requiring 7 address lines as per Equation 3.1, needs 8 balanced trees with a total of 248 2-input gates. Furthermore, since the tree must be embedded into the array, assuming a fixed pixel pitch, the fill factor will reduce as the array resolution is increased due to the larger number of gates required for the signal trees.

To the best of the authors knowledge, only two architectures have been presented which both share TDCs among a number of pixels and preserve pixel address information. In [100], an architecture is proposed which mitigates the need for transmitting an address along with the timing information from each pixel as in [97, 78]. Here, an OR tree is required in both column

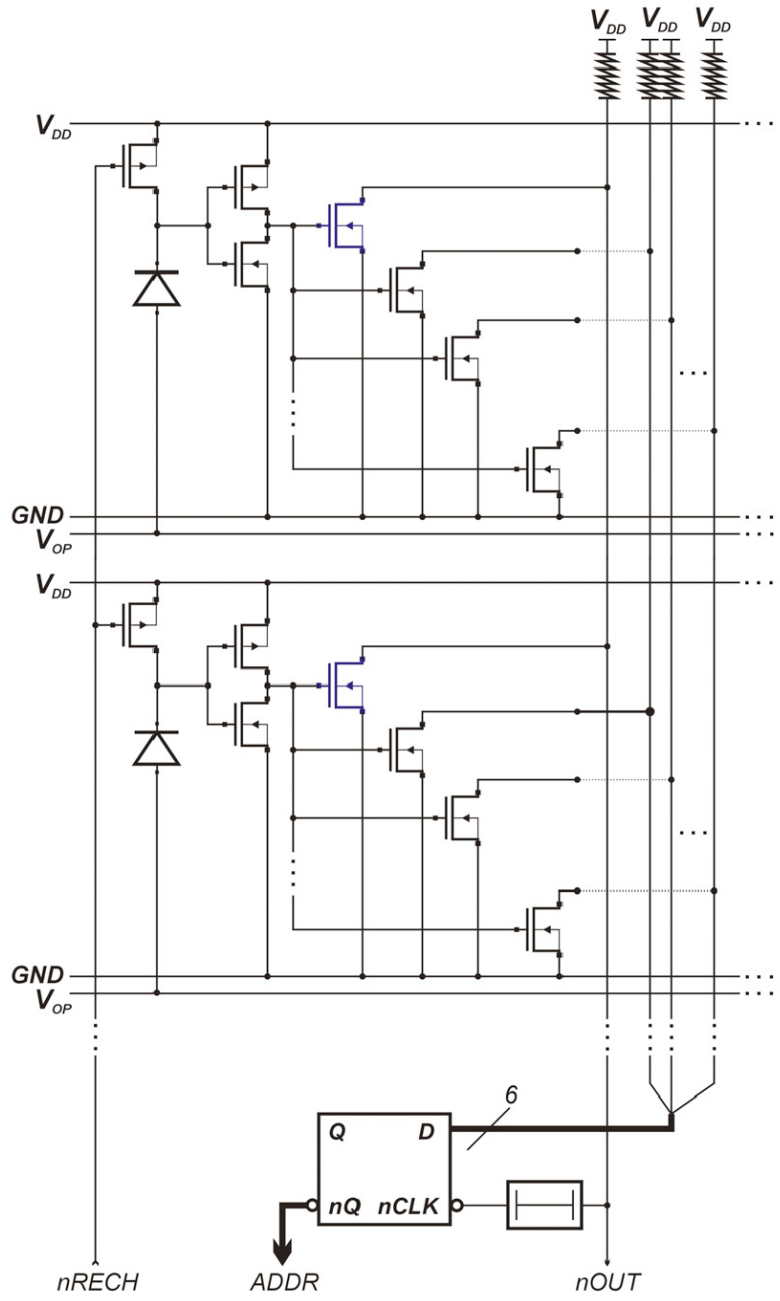


Figure 3.1 – Shared bus architecture with binary coding from [97]. *Picture credit: C. Niclass ©2006 IEEE.*

and row directions. The readout must then combine information from both the column and row decoders with the TDC code. Simulated in 130nm CMOS, the design claims a fill-factor of 5.7%, however no pixel pitch is mentioned. In [101] a shared TDC architecture is implemented in a 3D technology with a 65nm CMOS data processing tier. In this case, the timing signal is

propagated via dual D-type Flip-Flop (DFF) based decision makers, whilst the addresses are decoded using several levels of multiplexors.

3.1.3 Overcoming the I/O Bandwidth Bottleneck

Due to the increasing requirement for improved dynamic range, there has been significant activity towards developing architectures which can overcome the I/O bandwidth bottleneck discussed in Section 3.1.1. One approach involves compressing the data into a histogram [80] with a direct-to-histogram TDC before transmission from the chip. In this case, the size of the data to be transmitted can be reduced by a large factor. However, implemented in a 130nm CMOS technology, the direct-to-histogram TDC occupies an area of 30,000 μm^2. Thus, whilst this architecture has been extended to a line sensor [102], extending further to a large array would be difficult without migrating to a much more advanced technology node.

In contrast, data compression can also be achieved by integrating less, not more, on the sensor. In [103], a high fill-factor line sensor is produced which outputs buffered SPAD signals from the sensor IC directly to an field-programmable gate array (FPGA), where 64 TDCs are implemented. This scheme has the advantage, like 3D IC sensors, that the photodetector and data processing technologies are decoupled. FPGAs are produced in leading edge technologies and thus contain many resources, e.g. SRAM blocks, DDR interfaces, which would require a significant effort to include on a custom sensor. Furthermore, the FPGA gives a high degree of flexibility to adapt to different applications. The major downside to this scheme is that the number of channels is limited by the number of I/O pads available on the sensor and FPGA. Therefore, producing a sensor with many thousands of channels with this method is currently outside the limits of available technology.

3.2 NIROT Sensor Requirements

With discussions of SPAD characteristics and time-resolved system metrics in Chapter 2, and now the issues surrounding SPAD sensor design, some requirements for a high-speed NIROT sensor can be defined:

1. The sensor should employ a high performance SPAD design. For example, a PDP of greater than 10% at 800nm, a DCR density of less than 1 cps/μm^2, and negligible afterpulsing and crosstalk.
2. The SNR should be maximised.

3. To enable an image reconstruction with a high spatial resolution, the sensor should have an array resolution of at least 32 × 32 pixels.
4. The sensor must be able to measure from all pixels in parallel.
5. High data throughput will require a large output data bandwidth, e.g. a data pad per column or similar.
6. Power consumption should be limited to less than 200 mW such that an extensive cooling system is not required to keep DCR at the level desired. Although there is nothing that prohibits use of a cooling system, keeping system complexity at a minimum will ease the path to clinical measurements.

3.3 Case Study: Piccolo, A High-PDE Event-driven Time-resolved SPAD Sensor

To meet the requirements outlined in Section 3.2, the column architecture in Figure 3.2 is proposed for a new NIROT sensor, hereafter referred to as Piccolo. A full image sensor can be constructed by abutting columns in an array configuration. To improve the pixel fill-factor, the TDC is *not* included on a per-pixel basis. All Q pixels in a column are connected to an event-driven shared bus as in [97], however the coding scheme is able to detect collisions as in [78], where the number of bus lines, R, required is calculated according to Equation 3.1. In contrast to [78] however, each one of the Q pixels in the column are able to trigger the first available TDC in a bank containing S TDCs, where $S < Q$. Therefore, more than one photon may be detected per column, per cycle as long as the bus dead time, t_{bus}, is much less than the laser period. Data is read out from the sensor in serial via an output data pad which is dedicated to data from one column.

3.3.1 Sizing the TDC Bank

The purpose of having multiple TDCs per column is that once the first photon is detected, subsequent photons from the same column will not be missed because there is not an available TDC. Thus, in determining how many TDCs to have, the metric that concerns us is, *what percentage of photons are not detected due to insufficient timing resources?*

Since the light impinging on the detector can be modelled by a poisson process, this quantity can be calculated if we know the mean rate of photons detected in the column. From Section 3.1.1, we can expect that the data rate of the output pad, B_{io} measured in Hz, will create a

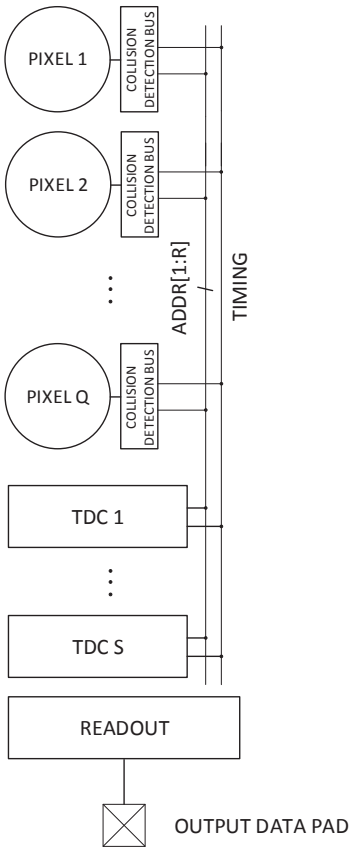


Figure 3.2 – Piccolo sensor architecture concept.

throughput bottleneck for the column. Thus, the worst case mean rate for the bus is given by $N = B_{io}/\mathrm{L}_C$, where L_C is the code length for one photon arrival and N is the mean event rate. The probability of missed photons, p_m is then given by

$$p_m = 1 - \sum_{k=0}^{S} \frac{N^k e^{-N}}{k!}, \tag{3.2}$$

where the quantity subtracted from 1 is the sum of the probabilities given by the poisson distribution for cycles where there are no missed photons. For example, when $S = 4$ the probability of having a cycle with no missed photons is the sum of the probabilities for 0, 1, 2, 3 and 4 events per cycle. With Equation 3.2, it is possible to calculate the percentage of missed photons versus the number of TDCs in a column for a given laser frequency. To implement the 32×32 array, as discussed in Section 3.2, 7 address lines are required. Assuming the TDC has 12-bit resolution and there is some overhead to indicate data transmission and TDC location, we can estimate the code length at 23-bits. Figure 3.3 shows the percentage of missed photons

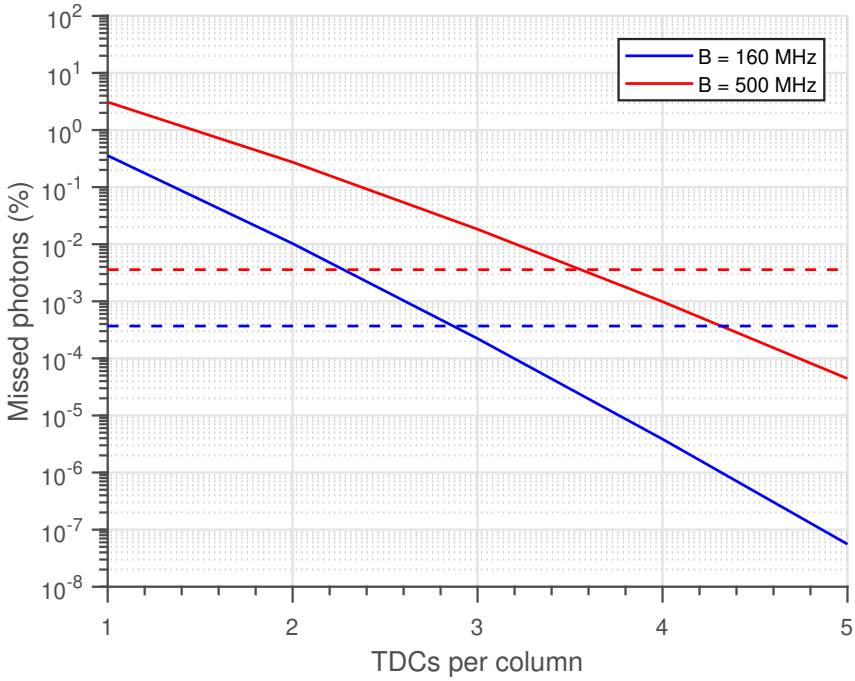


Figure 3.3 – Missed photons (%) vs number of TDCs per column with an 80 MHz laser frequency, missed photons for TDC-per-pixel case at each output data speed is indicated by horizontal dashed line.

in the case that the column data is transmitted via a GPIO pad operating at 160 MHz, and via an LVDS pair at 500 MHz. For comparison, the percentage of missed photons in the case that each pixel was allocated its own TDC at a pixel event rate N/Q, where $Q = 32$, is plotted as a dashed line. It should be noted however, that this calculation assumes an event driven readout of the TDC-in-pixel array as in [95], rather than a shift register and serialiser as in [59].

As seen in Figure 3.3, the shared architecture outperforms the TDC-in-pixel approach with just 3 TDCs in the column for the case of a GPIO output, and just 4 in the case of an LVDS pair. Thus, in terms of the percentage of missed photons, the architecture is *highly scalable*; a factor 3 increase in the detection rate requires only 1 extra TDC to reach the same performance as the TDC per pixel case. Furthermore, the percentage of missed photons is very low. With 2 TDCs in the bus with a GPIO pad, the percentage of missed photons is 0.01% of the total detected on the bus. This illustrates that in some cases it may be advantageous to employ less TDCs than required to meet the TDC-in-pixel standard, as the saving on die area may be more significant than an improvement in percentage of photons missed, a number which may already be very small.

Of course, a laser operating at 80 MHz may be too fast for many applications, e.g. fluorescence

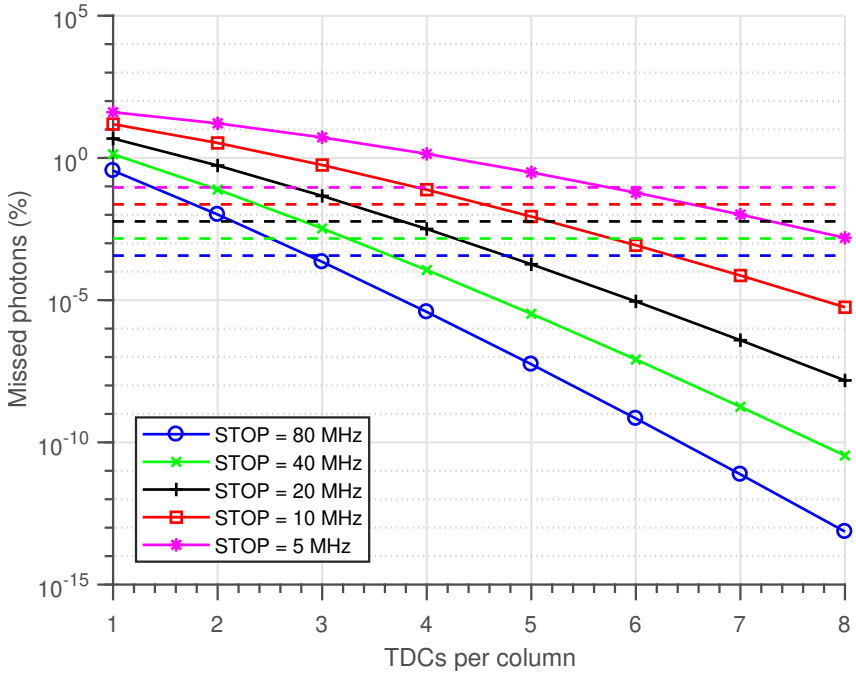


Figure 3.4 – Missed photons vs number of TDCs per column for GPIO case with varying laser frequency. The dashed horizontal lines represent the percentage of missed photons in the TDC-in-pixel approach.

lifetime imaging, LiDAR, etc. As the laser frequency is lowered, the mean event rate per cycle will increase causing a corresponding increase in the percentage of photons missed in comparison to Figure 3.3. Figure 3.4 displays this change for the GPIO case for a number of frequencies in the range 80-5 MHz. The horizontal dashed lines represent the percentage of missed photons in the case of the TDC-in-pixel approach, thus illustrating the increasing number of TDCs per column required to reach the same performance.

Finally, to evaluate the scalability of the system in bandwidth and frequency it is instructive to plot the same results for the LVDS case, shown in Figure 3.5. This illustrates that despite the increasing number of TDCs required to reach equivalence with the TDC per-pixel results, this can still be achieved with just 10 TDCs in the worst case scenario. In this case, around 0.5% of photons are missed.

There are a few conclusions that can be drawn from the preceding analysis. Firstly, in general the percentage of photons missed is low, particularly for high clock frequencies and slower output data speeds. Under these conditions, the proposed architecture seems very powerful. For flexibility for more applications however, the number of TDCs in the bank must be significantly increased, however even with 10 TDCs per column, this is still a major advantage in

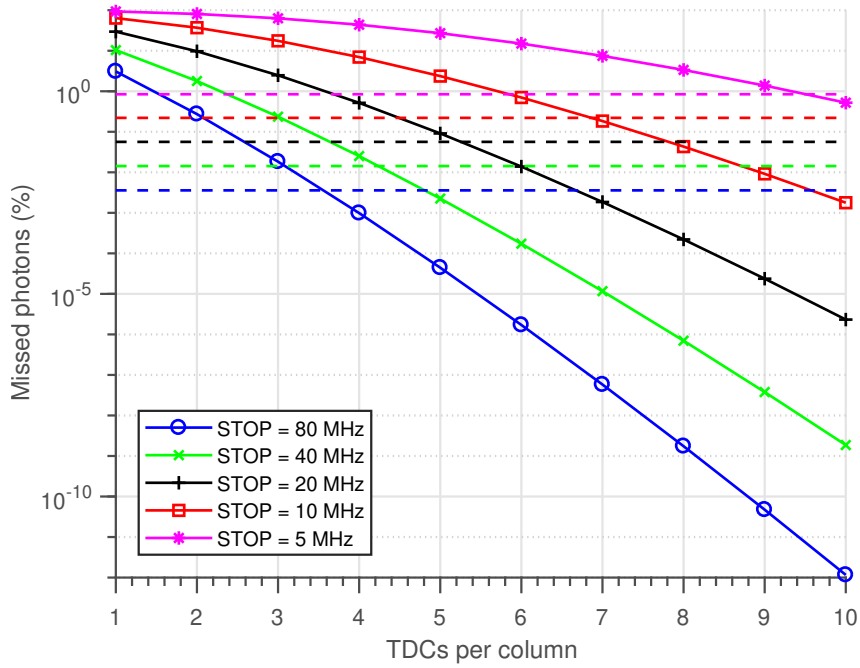


Figure 3.5 – Missed photons vs number of TDCs per column for LVDS case with varying laser frequency. The dashed horizontal lines represent the percentage of missed photons in the TDC-in-pixel approach.

comparison to the TDC-in-pixel approach. Secondly, the calculations for the shared architecture are completed on the basis of bus activity. Therefore, for a given bus activity, if the percentage of photons missed is at an acceptable level, the pixel column could be increased to very large numbers, e.g. megapixel, whilst achieving a high pixel fill-factor. Finally, the analysis of both GPIO and LVDS data bandwidths, shows that the proposed architecture has good scalability to high event rates.

3.3.2 Photon Collision Analysis

As discussed in Section 3.1.2, use of a shared bus for sharing one or more TDCs among a number of pixels will result in code collisions. This occurs due to overlapping bus dead time periods produced on the bus as a result of two or more photons being detected by different SPADs. There are two techniques employed in the architecture in Figure 3.2 to reduce the rate of incidence and effect of bus collisions. Firstly, as implemented for sharing via NOR trees [98], the pulse width which is generated by the SPAD pixel for a photon detection is reduced with a monostable. With this method the output pulse of the pixel due to the SPAD dead time can be reduced to a few 100s of picoseconds. Of course, the SPAD still remains inactive for its dead

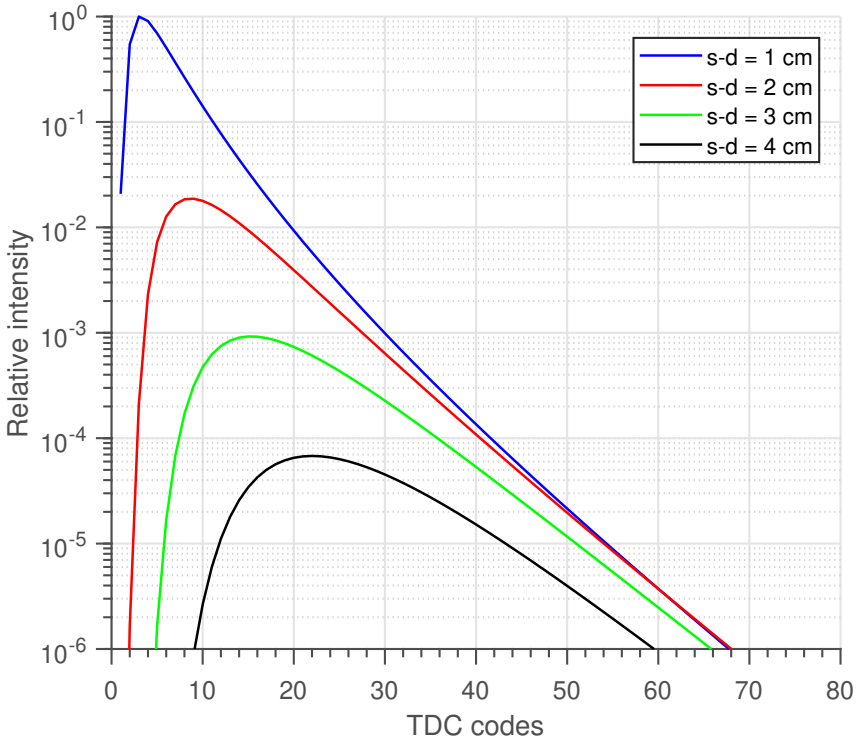


Figure 3.6 – NIROT TPSF of homogeneous media with varying source-detector (s-d) spacing simulated with NIRFAST, $\mu_a = 0.01$ mm^{-1}, $\mu_s' = 1$ mm^{-1}, and 1 time bin = 50 ps.

time. Therefore, the bus dead time can be reduced to some 100s of picoseconds per photon detected. Secondly, when collisions do occur, the coding scheme implemented, which is the same as in [78], produces invalid codes. This results in the two photons which are involved in the collision being discarded.

The incidence of photon collisions on the bus can be analysed for the NIROT application with TPSFs produced in NIRFAST [18] forward simulations. TPSFs are simulated for the cases of a homogenous medium with absorption coefficient of $\mu_a = 0.01$ mm^{-1} and reduced scattering coefficient $\mu_s' = 1$ mm^{-1}. These are typical optical properties for human tissue. The source-detector (s-d) separation in the simulation is varied from 1-4 cm in steps of 1 cm to demonstrate a range of TPSFs that may be seen in a typical NIROT measurement. As the separation becomes larger, the FWHM of the TPSF broadens due to the increased path lengths for photon propagation. These TPSFs are shown in Figure 3.6. In these simulations one time bin in the TPSF is equal to 50 ps. Thus, the TPSFs have approximate FWHMs in picoseconds of 200 (1 cm), 550 (2 cm), 750 (3 cm) and 900 (4 cm).

.

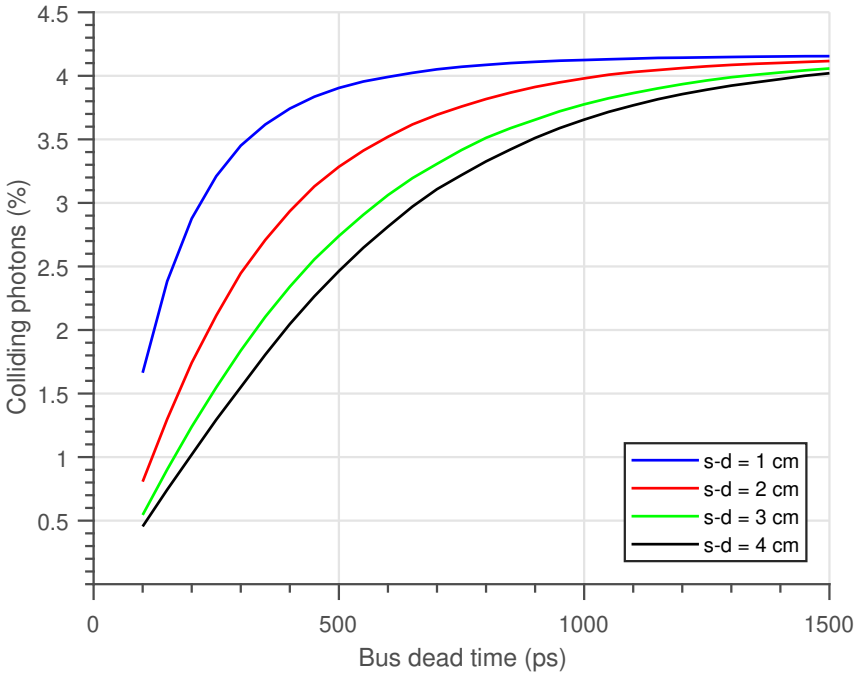


Figure 3.7 – Colliding photons vs bus dead time for varying source-detector (s-d) spacing. Simulated using randomized arrival according to poissonian statistics of NIROT TPSF of homogeneous medium simulated with NIRFAST, $\mu_a = 0.01$ mm^{-1} and $\mu_s' = 1$ mm^{-1}.

To calculate the incidences of photon collisions on the shared bus, the given TPSF from the simulation is used to generate a histogram with 10^6 photons. The histogram can then be converted into a vector with 10^6 integers between 1 and 250, representing all photons in the TPSF. The order that those photons arrive in is then randomized, at which point the vector can be segmented into groups based on how many photons arrive per cycle according to poisson statistics. For example, if we assume a bus with 32 pixels, 3 TDCs and a 160 MHz data pad for data read out, the photon vector can be divided into 3 groups as the probability of 4 or more photons arriving per cycle is negligible. In the groups where 2 or 3 photons were received per cycle, we find the time distance between any two photons detected in a cycle. If the time distance is less than the bus dead time, a collision is registered for both photons at which point they would be discarded. Simulation of both the mediums in Figure 3.6 for bus dead times ranging from 100-1500 ps are shown in Figure 3.7.

There are a number of key points to take away from these results. Firstly, a reduced bus dead time results in an improvement in the incidence of collisions. This motivates the design of the monostable and the bus in such a way that the pulse width at the output is minimised as far as is practically possible. Furthermore, photon collisions may result in the loss of many more photons than a lack of TDCs, as discussed in Section 3.3.1. Of course, in this case the TPSF

transmitted along the bus was only signal, *no noise was added*. Therefore, all of the counts in the TPSF are contained within a narrow range of the total time bins, and photon collisions are maximised. Whilst this may approximate the case in NIROT, with a small source detector separation, in other applications, e.g. LiDAR, background noise will be much stronger.

3.4 Conclusions

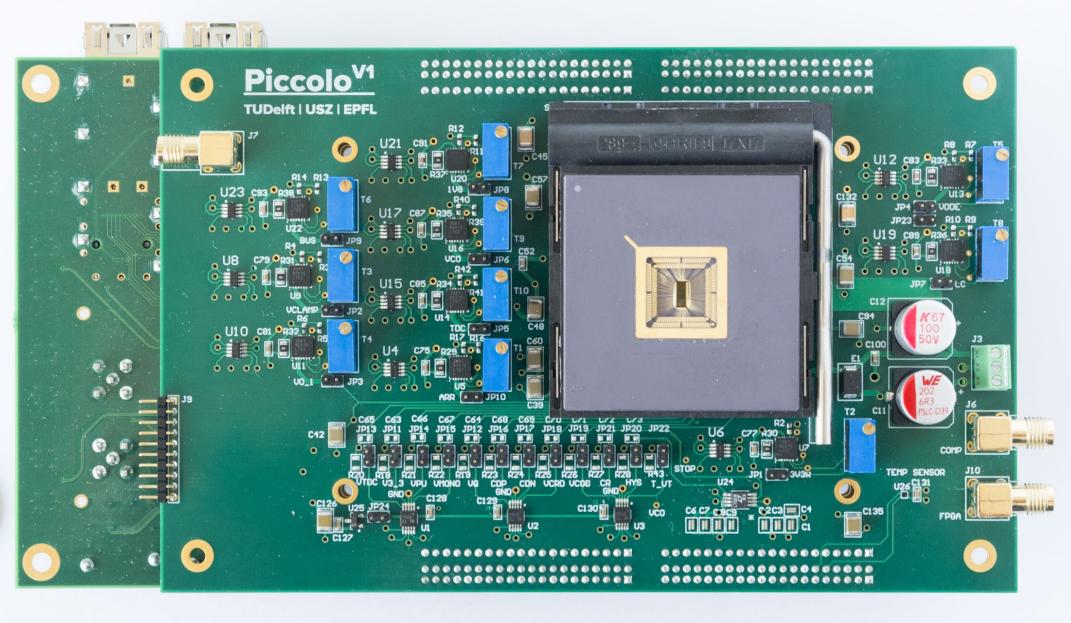
The sensor architecture presented here overcomes many of the shortcomings of conventional monolithic time-resolved sensor architectures. By implementing a dynamic reallocation of TDCs among a large number of pixels, the circuit area occupied by timing circuitry can be drastically reduced whilst missing fewer photons than an equivalent TDC-per-pixel architecture. Calculating the percentage of missed photons for the case of dedicated GPIO and LVDS pads per column indicates demonstrates the scalability of the system to higher bus activities. Additionally, an analysis of photon collisions in the column shows that less than 4.16% of photons will be involved in a collision at maximum throughput with a bus dead time of 1.5 ns. To the best of our knowledge, this is the first time-resolved SPAD sensor which implements a TDC sharing scheme with more than 1 pixel to more than 1 TDC without employing a static logic tree structure. Since the percentage of missed photons is dependent on the output data rate, and thus the bus activity, the pixel column can be increased to larger sizes without requiring additional timing resources. Thus, the architecture is highly suited to implementing a wide-field time-resolved sensor for clinical NIROT applications.

4 A 32 × 32 Event-driven Time-resolved SPAD Sensor

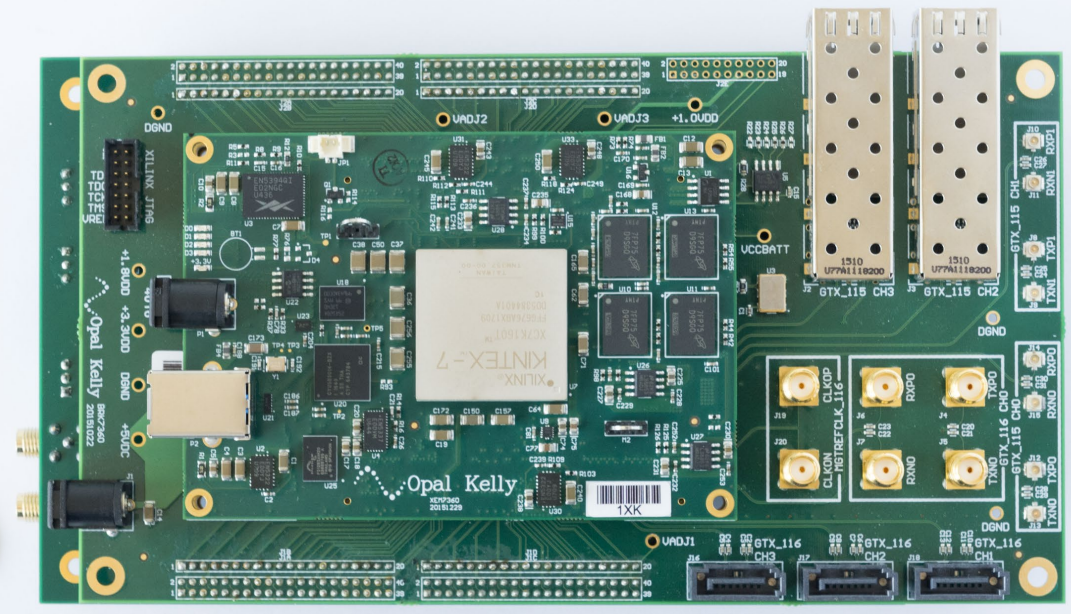
A 32 × 32-pixel sensor based on the architecture presented in Chapter 3 is implemented in a 180 nm CMOS technology. The Piccolo camera system is introduced in Section 4.1, prior to a characterisation of the SPAD and then system level measurements. The chapter concludes with silicon phantom and timing stability measurements to validate the system for NIROT measurements.

This chapter is based on results presented in, S.Lindner et al. "Column-Parallel Dynamic TDC Reallocation in SPAD Sensor Module Fabricated in 180nm CMOS for Near Infrared Optical Tomography", in *Proceedings of 2017 International Image Sensor Workshop,* C.Zhang et al. "A CMOS SPAD Imager with Collision Detection Bus and 128 Dynamic Reallocating TDCs for Single Photon Counting and 3D Time-of-Flight Imaging", [Submitted to Optics Express] and Alexander Kalyanov et al. "Time-Domain Near-Infrared Optical Tomography with Time-of-Flight SPAD Camera: The New Generation", Biophotonics Congress: Biomedical Optics Congress 2018 (Microscopy/Translational/Brain/OTS), OSA Technical Digest (Optical Society of America, 2018), paper OF4D.5.

Piccolo was a collaborative design effort with a division of labor among the different circuit blocks. The author was responsible for the design of the TDC, Section 4.1.4, whilst Chao Zhang developed the address latching structure, Section 4.1.3, and data readout, and Ivan Michel Antolovic designed the pixel array, Section 4.1.2, collision detection bus and pixel masking scheme. The PDP and light emission test (LET) measurements in this chapter were carried out by Ivan Michel Antolovic.



(a) front



(b) back

Figure 4.1 – (a) Front of the Piccolo camera system displaying Piccolo sensor bonded to a 208-pin CPGA package. The package is mounted in a zero-insertion-force socket on a daughterboard, containing auxiliary components necessary for sensor operation. (b) Back of the Piccolo camera system displaying XEM7360 integration module (Opal Kelly, USA), mounted on a BRK7360 PCB. The system dimensions are 188 × 102 mm.

4.1 Piccolo Camera System

The Piccolo camera system is shown in Figures 4.1a (front) and 4.1b (back). The system comprises of 3 PCBs, a daughterboard custom designed for the sensor operation, a break-out board (BRK7360, Opal Kelly, USA) to provide a debug facility through the large header connections to the daughter board, and an integration board (XEM7360, Opal Kelly, USA) which includes a Kintex-7 FPGA (XC7K160T-1FFG676C, Xilinx, USA) and USB 3.0 interface. The sensor die, shown in Figure 4.8 is bonded in a 208-pin CPGA package and then mounted in a zero-insertion-force socket on the front side of the daughterboard, Figure 4.1a. The daughterboard contains a number of quad digital-to-analog converters (DACs) (DAC7554, Texas Instruments) for bias voltage generation, and low-dropout regulators (LT3081, Linear Technology) and digital potentiometers (TPL0501, Texas Instruments) for on-board supply voltage generation and distribution. In time-resolved mode, the system can operate with either the FPGA or the laser as master. In the latter case, a fast comparator (LTC6752, Linear Technology) which can accept input signals over the range 0-5 V converts the trigger from the laser into a CMOS output compatible with the 3.3 V I/O pads on Piccolo. The daughterboard takes two supply voltages, a high voltage, $V_{OP} = 25 - 27$ V, for the SPAD bias, and 5 V for all remaining components.

On the back side of the PCB stack, Figure 4.1b, the XEM7360 interfaces to Piccolo and the auxiliary components on the daughterboard through GPIO pads on the Kintex-7 FPGA. Sensor and board control, and data acquisition are performed through the USB 3.0 interface to the Kintex-7 on the XEM7360 board. It is interesting to note that even with a USB 3.0 interface, the maximum data rate is 340 MB/sec or 2.72 Gb/sec. With 32 columns operating at maximum speed, Piccolo would output photons with a total data rate of 5.12 Gb/sec. Therefore, to achieve this maximum speed, the raw data from Piccolo must first be accumulated into histograms in the FPGA block RAM to compress the data into a size which can be managed by the USB 3.0 link.

4.1.1 Sensor Architecture

The circuit architecture of the Piccolo sensor is pictured in Figure 4.2. The sensor is implemented in a 180 nm CMOS technology and employs a wide spectral range p-i-n SPAD [67]. The sensor includes a 32×32 SPAD array where each column of SPADs are allocated to a bank of 4 address latch and time-to-digital converters (ALTDCs), see Figure 4.2b. The ALTDC block is responsible for the capture of addresses from the bus when photons are detected, as well

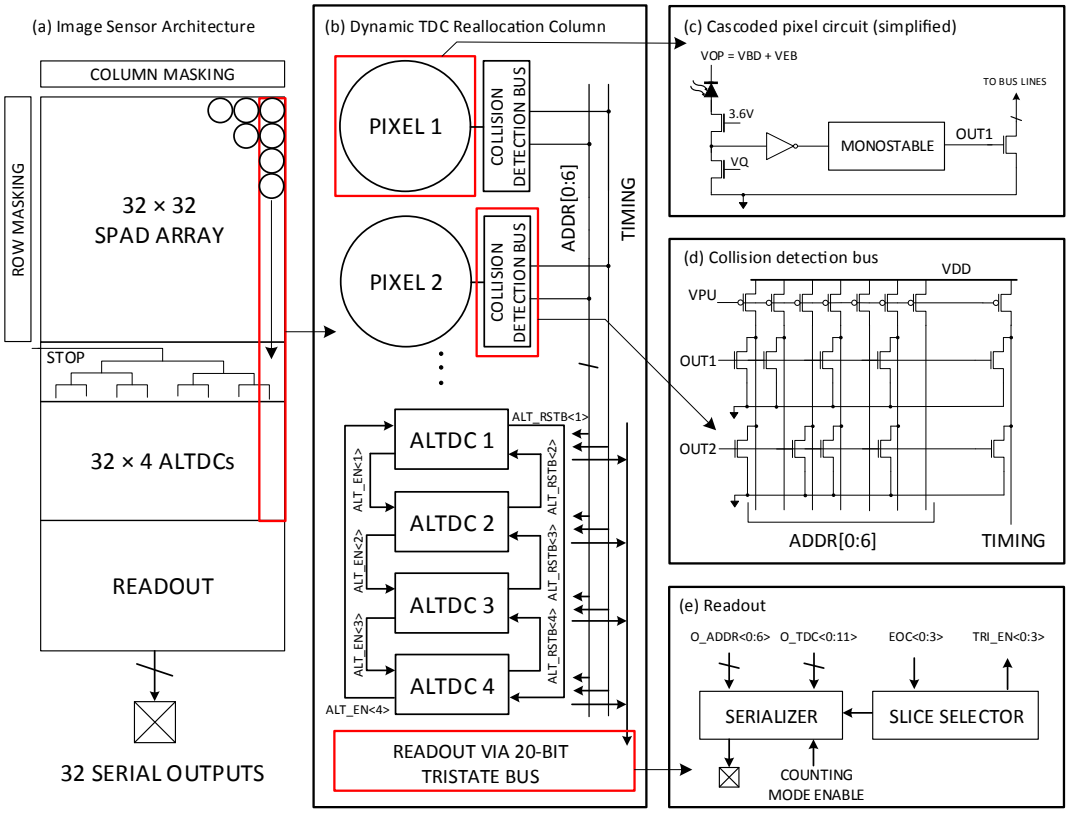


Figure 4.2 – Piccolo sensor architecture.

as the time-to-digital conversion and dynamic reallocation of the active ALTDC block. This sharing scheme enables a 28% fill factor in a 28.5 μm pixel pitch, in this case the diameter of the SPAD active area is 17.02μm. This is unprecedented in comparison to FSI designs which include a TDC in each pixel.

The pixels, simplified in Figure 4.2c, employ the cascoded quenching and recharge configuration developed in Chapter 2, to improve the sensor PDE. A monostable is also included to reduce the bus dead time for each photon detection. A shared collision detection bus, Figure 4.2d, is implemented where the bus is ordinarily pulled up to the supply voltage, V_{DD} = 1.8 V, when idle and then pulled down by a set of NMOS transistors in pixel when a photon is detected. The bus uses the collision detection coding scheme from [78], which requires 7 address lines to encode a 32 pixel column such that collisions can be detected.

A critical feature of the architecture is the masking scheme implemented via column and row masking registers which are used to write to an internal memory cell in each pixel, Figure 4.2a. Since all pixels in a column share the same bus lines and timing resources, very noisy or 'hot'

pixels may heavily degrade sensor performance if appropriate masking is not employed. The main concern is consuming a large portion of the TDC resources, whilst the dense layout of the SPAD array means that crosstalk from hot pixels could also be substantial.

Readout is performed via a 20-bit tristate bus in the column. Each ALTDC slice includes a signal, EOC, which signifies that the slice has detected a photon and is ready for readout. This is detected by the readout circuitry, which then begins the readout of the block on the next rising edge of the readout clock, which runs at 160 MHz. The tristate bus has the advantage that only valid photons are read out from the TDC array, there is no *null* data which would reduce the efficiency of the data transmission. Once in the read out, data from a column is serialized and then transmitted from the sensor via a dedicated 160 MHz GPIO pad. With 32 GPIOs, Piccolo has a total output data bandwidth of 5.12 Gbps. As each detection event is encoded with 23-bits, this corresponds to a maximum photon throughput of 220 Mevents/sec.

The trigger signal from the laser, the STOP signal in Figure 4.2a, is distributed to the ALTDC blocks through a balanced binary clock tree inserted between the SPAD array and the ALTDC blocks. This ensures that the TIMING signal, which travel along the bus lines in the array down to the ALTDC blocks, maintains a constant phase relationship with the STOP signal. This ensures that for any given pixel, there is not a large offset in TDC codes for an identical time of arrival depending on which ALTDC slice performs the measurement. For example, if the STOP signal was inserted from the bottom of the column then the code produced by the top ALTDC slice would be shifted in comparison to the same time of arrival at the bottom slice. Of course, there is nothing which prevents these shifts being calibrated for in post-processing. Saying that, it does complicate the design of systems which perform all processing on FPGA, where calibration is more time consuming to implement, and is thus avoided here.

4.1.2 The Pixel

The pixel design for Piccolo is pictured in Figure 4.3. A cascoded passive quenching and recharge is performed by the series combination of M_1 and M_2, which are both thick oxide transistors tolerant up to 3.6 V. Assuming this is a hard limit on the gate-source (V_{GS}), gate-drain (V_{GD}) and drain-source (V_{DS}) voltages, this quenching and recharge scheme can tolerate excess biases up to 5.2 V. This is slightly less than could be achieved with the active recharge scheme in Section 2.5, where a gain of 60% was obtained in comparison to 44% here. The reason for this is that in a passive quenching scheme, V_{GS} of M_2 must be set, via V_Q such that the SPAD recharges in an acceptable time. Thus the resistance of M_2 is lower in the passive recharge case than in the active recharge case, where the SPAD is recharged by switching the

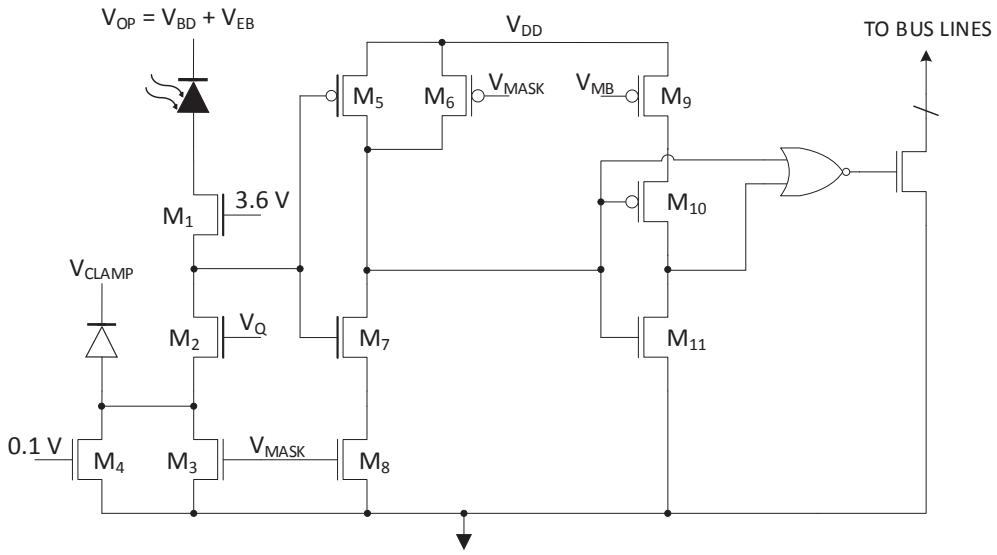


Figure 4.3 – Cascoded passive quenching pixel schematic.

gate of M_2 to a higher voltage for the recharge period. The lower resistance of M_2 is reflected in the decreased source voltage of M_1 required to bring it to an equivalent resistance. Increase of the threshold voltage of M_1 due to the body effect further decreases this voltage. An active recharge scheme was not implemented here since the I/O data rate limits the maximum pixel activity, rather than the SPAD dead time. This is the bottleneck for the dynamic range of the sensor. Furthermore, the expected low afterpulsing of the SPAD implies that the utility of active recharge would be minimal without the requirement of a short dead-time, e.g. 10 ns.

The pixel masking functionality is achieved with a 6T-SRAM internal to every pixel which contains the masking state. Masking is performed via two methods. Firstly, the electrical output of the pixel is deactivated when the masking signal, V_{MASK}, is set to ground. This pulls the output of the inverter formed by M_5 and M_7 to V_{DD}, thus the pixel is unable to toggle any further nodes. Although with this electrical masking the pixel may not transmit any avalanche events onto the bus, avalanche events are not prevented by it. Therefore, hot pixels which are masked may still cause secondary avalanches through crosstalk. As discussed in Section 2.2, this effect is particularly prevalent for densely packed arrays.

The second method of masking is employed here to prevent, or at the very least reduce, the incidence of avalanches in masked pixels. Typically, these type of schemes utilize a PMOS transistor to reduce the voltage across the SPAD to zero [104], as demonstrated in Section 2.6. Such a technique is not trivial to include with the cascoded quenching and recharge scheme.

The masking PMOS device would not only need a cascode transistor to ensure operation at higher voltages, but it would also require a gate voltage which could switch between V_{EB} when the pixel is unmasked and $V_{EB} - V_{MASK}$, where $V_{MASK} > 0$ V if $V_{EB} > 3.6$ V. The masking scheme implemented here employs two transistors M_3 and M_4, connected in parallel, where the parallel combination of these devices is in series with the quenching transistors. When $V_{MASK} = 0$ V, M_3 is in the cut-off region and thus is seen as a very large resistance, which could be many GΩ. Thus, when an avalanche is triggered, the SPAD anode rises to V_{EB} but recharges at a much slower rate, thus drastically reducing the number of avalanches triggered. In fact, the danger is that the resistance of M_3 is so high that the leakage through the reverse biased SPAD junction is larger than that through the series combination of M_3, M_1 and M_2. In this case, the SPAD anode voltage may drift to greater than V_{EB} which could cause a long term breakdown of the quenching transistors. To prevent this, M_4 is connected in parallel to M_3 with it's gate biased at a low non-zero voltage, nominally 0.2 V which places the transistor in the subthreshold region. This provides a leakage path for the quenching transistors, preventing the SPAD anode from drifting to high voltages. Finally, since M_3 and M_4 are thin oxide devices, a small clamping diode is connected to their drains with the remaining terminal at a voltage $V_{CLAMP} = 1.8 - V_D$, where V_D is the diode forward voltage. Thus, if their drain voltage was to exceed 1.8 V, the protection diode turns on and limits the voltage to approximately 1.8 V.

Finally, a monostable is formed from the combination of an inverter formed by M_{9-11} and a NOR gate. The monostable pulse width is set by the signal propagation through the delayed inverter path. A global bias voltage V_{MB} which is distributed to each pixel controls the rate at which the rising edge output of the inverter charges the input of the NOR gate. In this manner, monostable pulse widths in the range 0.4-5.5 ns can be produced. The NOR gate drives four NMOS transistors which pull-down the bus to transmit the address and timing signals onto the bus.

4.1.3 Address Latch and Dynamic Reallocation

Address capture and timing measurements are performed by four ALTDC slices connected in a daisy chain configuration, see Figure 4.2b. Each ALTDC slice is activated by the previous slice in the chain through the signals ALT_EN, a process called dynamic reallocation. This activation is triggered by the TIMING signal, ensuring that the time between successive photon detections is reduced to a minimum. Once a slice has completed its data readout, it resets the previous slice in the chain with ALT_RSTB. This implies that, for any given cycle, there is always one ALTDC which has not been reset. Therefore, with four ALTDC slices up to three

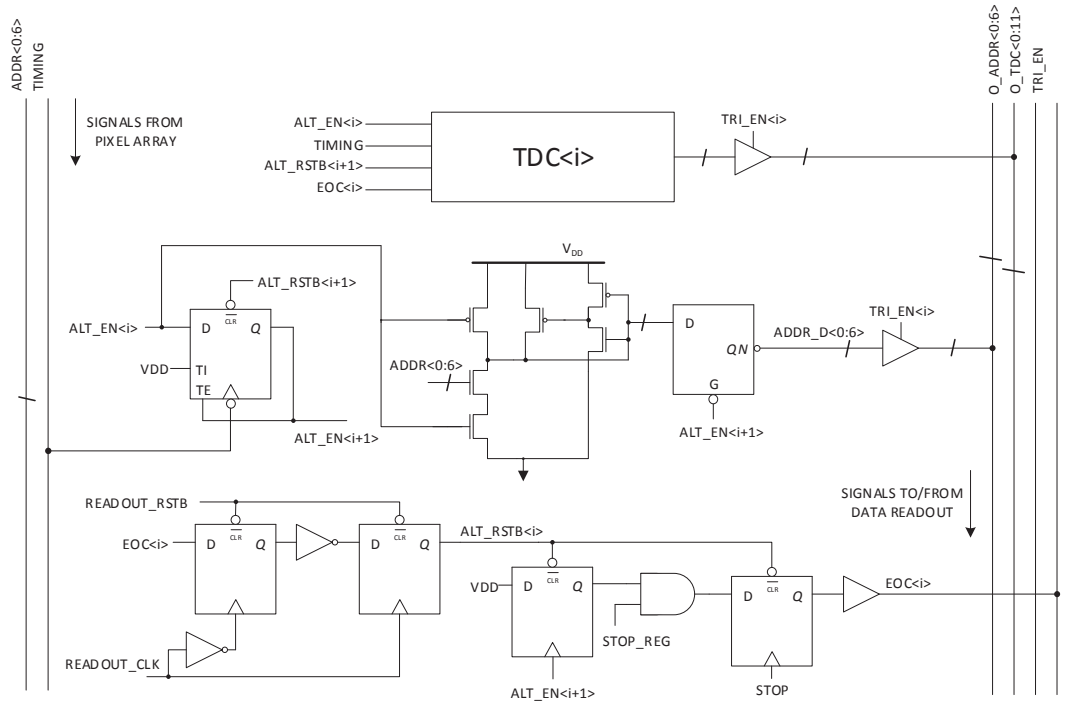


Figure 4.4 – Simplified ALTDC block showing a single slice and connections from pixel array and between the ALTDC and data readout.

photons may be detected per cycle.

Each column includes four ALTDC slices, Figure 4.4, as well as an initialization circuit to ensure that one slice is active after a global reset. A timing diagram showing a single photon detection is shown in Figure 4.5. Upon photon detection a TIMING pulse is produced in the pixel, which then travels down the bus towards the ALTDCs. The rising edge of TIMING triggers the start of the TDC conversion, at T1. Concurrently, ADDR<0:6> is captured by the dynamic logic since its reset is deasserted by ALT_EN<1>. On the falling edge of TIMING, T2 in Figure 4.5, the ALTDC asserts ALT_EN<2>, thus enabling the next slice in the chain, ALTDC<2>, for detection. At the same time, the address captured by the dynamic logic is latched, thus any further events on the bus will not change the captured code. The time-to-digital conversion is performed in reference to the STOP signal, which defines the end of the measurement. Upon the rising edge of STOP, time T3, the end of conversion EOC<1>signal is asserted. This assertion indicates to the data readout that both the address and timing data are valid for readout. On the next rising edge of READOUT_CLK at T4, the readout enables the tri-state buffers in the ALTDC block through TRI_EN<1>such that the address and timing data are transmitted on the shared data output lines, O_TDC<0:11> and O_ADDR<0:6>. At the same time, ALTDC_RSTB<1> resets

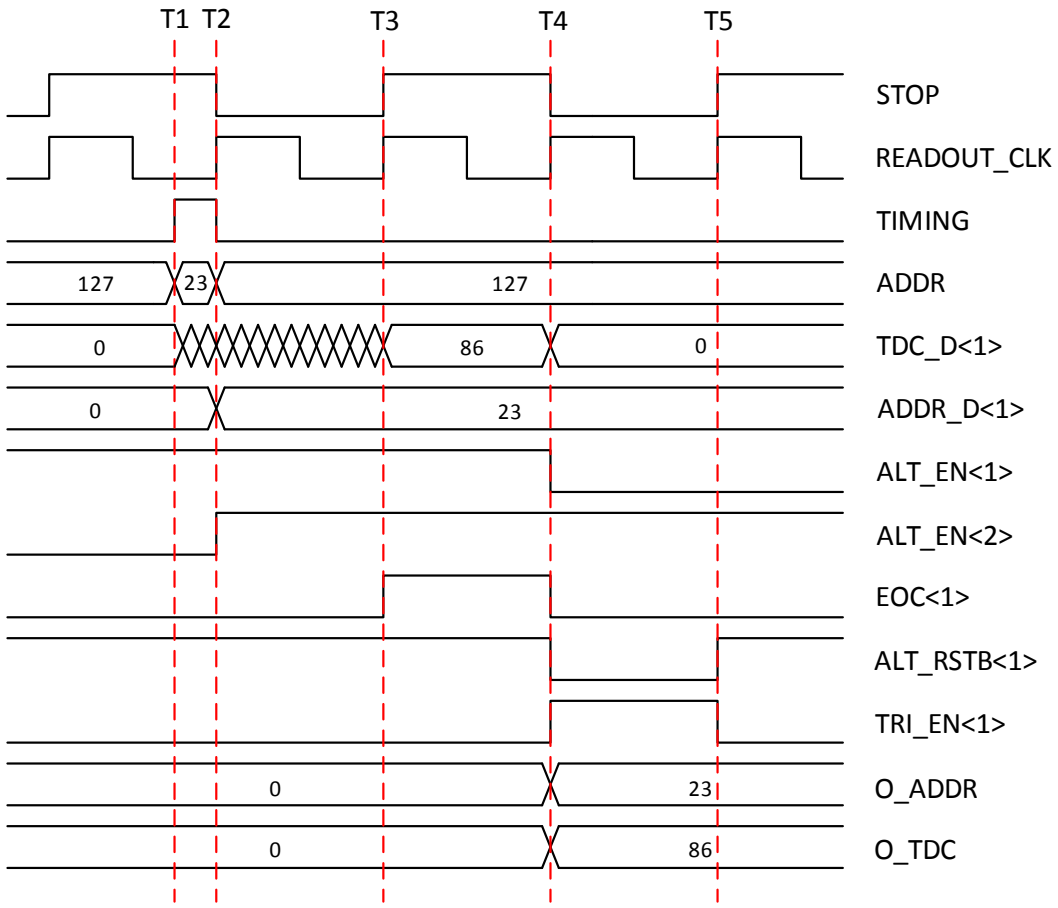


Figure 4.5 – ALTDC timing diagram. Photon detection triggers TDC conversion and capture of ADDR in dynamic logic (T1). On falling edge of TIMING, captured address is latched and the next ALTDC slice activated via ALT_EN<1> (T2). Next rising edge of STOP completes TDC conversion and asserts EOC for end of conversion (T3). On the next rising edge of READOUT_CLK, TRI_EN<1> enables tri-state buffers for data output to readout. At the same time ALT_RSTB<1> resets TDC<1>, EOC<1> and ALT_EN<1> (T4). Finally, on the next rising edge of READOUT_CLK the timing data, O_TDC, and address data, O_ADDR, are read by the readout block (T5).

TDC<1>, EOC<1> and ALT_EN<1>, the activation signal from the previous slice. Only after ALTDC<4>has completed a detection will ALTDC<1>be able to detect another photon.

With this method, the minimum time distance between successive photon detections (1.5 ns) is dependent both on the time taken to capture the address in the dynamic logic, and the time taken for the ALT_EN signal to propagate to the next slice. Both of these elements would be expected to reduce by moving to a more advanced node, e.g. 65nm, as the trend for CMOS logic is to increase in speed as the feature sizes are reduced.

4.1.4 The TDC

The integration of TDCs on the same silicon substrate as SPADs has opened up the possibility of time-resolved imaging systems with a large number of channels, e.g. > 1000. The design of a TDC for such a system has some key requirements. Firstly, the requirements for good SPAD performance implies that a more mature CMOS technology, e.g. 350-130nm, should be selected. Additionally, the need for a large number of TDCs to be integrated on-chip means that the circuit area should be minimized such that the photosensitive area can be maximised. Taken together, these two requirements suggest that TDCs for monolithic SPAD imagers should employ simple architectures, such that other aspects of the sensor performance are not sacrificed. A second aspect which drives architecture selection is power consumption as small increases in power dissipation at the block level can result in large power dissipation when implementing a large number of channels. This impact of this power dissipation on sensor temperature can be solved with active cooling, however the resulting complications in system design and use mean this is not a desirable solution.

Besides, area and power, there are some specifications of the TDC which will have a large effect on the sensor performance via the timing response. In TD NIROT, a narrowing of the timing response of the system, which is typically referred to as the IRF, results in improved depth sensitivity, penetration depth, depth selectivity and contrast [32]. A narrower system FWHM is also beneficial in other applications such as LiDAR. For example, it was discussed in Section 2.3.2 how the SNR in timing measurements improves as the system FWHM is decreased. Additionally, given N photons in a measurement, the FWHM of the system, FWHM_{SYS}, determines the root mean square (rms) precision of a timing measurement, σ_m, given by

$$\sigma_m = \frac{\sigma_{SYS}}{\sqrt{N}}, \tag{4.1}$$

where σ_{SYS} is the rms jitter of the system. The rms jitter of the system is related to the FWHM by, $\sigma_{SYS} \approx (\text{FWHM}_{SYS}/2.355)$. This mean that a narrower FWHM improves the precision of the system. Clearly, the timing response of the system should be minimised.

We can observe the influence of the TDC on system jitter through the following equation,

$$\sigma_{SYS} = \sqrt{\sigma_q^2 + \sigma_{START}^2 + \sigma_{STOP}^2 + \sigma_{TDC}^2} \tag{4.2}$$

where σ_q is the quantization error of the TDC given by

$$\sigma_q = \frac{LSB}{\sqrt{12}} \tag{4.3}$$

and σ_{START} and σ_{STOP} are the respective rms jitter contributions of the SPAD and STOP signals of the TDC whilst σ_{TDC} is the contribution from the TDC itself. Therefore, the jitter of the system can be improved by achieving a smaller LSB or decreasing the jitter of the various elements in the TDC itself. Since the SPAD for Piccolo has a FWHM of approximately 100 ps [67] and the laser contributes 20 ps FWHM, an LSB of 50 ps would achieve a $\text{FWHM}_{SYS} \approx 107$ ps, if σ_{TDC} is neglected momentarily. This LSB is chosen on the basis that the system performance should not be overly degraded in comparison to the intrinsic performance of the SPAD, and there are diminishing returns for targeting an LSB much less than this. Even with a small LSB, the system performance is still contingent on the contribution from σ_{TDC}. Although the logic transitions in CMOS, even in mature processes, are very fast and contribute very little jitter, the time-to-digital conversion always relies on the generation of some high frequency clock which typically dominates σ_{TDC}. Thus the architecture and design must be carefully chosen to minimise its contribution.

There are a few different TDC architectures which have been explored for time resolved SPAD imagers. To achieve a small circuit area with a fixed LSB, several works [95, 93, 105, 106] distribute multiple phases of a high frequency clock to all TDCs. The phase of the photon detection is then measured by an interpolator in reference to the multiple clock phases. The jitter accumulated by this TDC can be minimised with proper design as the clock phases are generated with either a delay-locked loop (DLL) or phase-locked loop (PLL). The downside to this approach is that the power consumption can be very high, particularly for a sub-100 ps LSB. In [95], a 62.5 ps LSB is achieved by distributing sixteen phases of a 1 GHz clock signal. Lower power designs can be based on delay line [92] or ring oscillator (RO) based [107, 59, 94] architectures.

Delay line based architectures are a simple and robust alternative. Here, the photon detection signal propagates along a delay line and its position interpolated upon the arrival of the STOP signal from the laser. However, achieving a 50 ps resolution in 180nm CMOS would require a delay line constructed from inverters, resulting in degraded differential nonlinearity due to unequal rise and fall times. A possible solution to this would be to employ a sub gate delay architecture, however the increased area required makes this option unfeasible.

Finally, RO based architectures occupy a small area whilst simultaneously achieving a sub-100

ps LSB. A photon detection begins the oscillation of a high frequency oscillator, the edges of which can be counted by a counter. An LSB smaller than the period of the oscillator is obtained by freezing the state of the oscillator on the arrival of STOP, and decoding the phase [107, 59]. For short laser periods, e.g. 80 MHz, this architecture can achieve a small σ_{TDC} as the jitter accumulates in the RO for a very short period. Furthermore, since the oscillator only runs once a photon is detected, this method is intrinsically low power. We can see the benefit of this by considering the Piccolo architecture. Assuming an 80 MHz laser frequency and a maximum throughput of 220 Mevents/sec, at maximum throughput there will, on average, be less than 3 TDCs active in the entire array per cycle.

The downside of this open-loop RO operation is that the LSB will be subject to some variation, it is not locked to a multiple of a reference frequency. A recent technique to mitigate this involves injection locking an array of mutually coupled oscillators [108]. This method has a

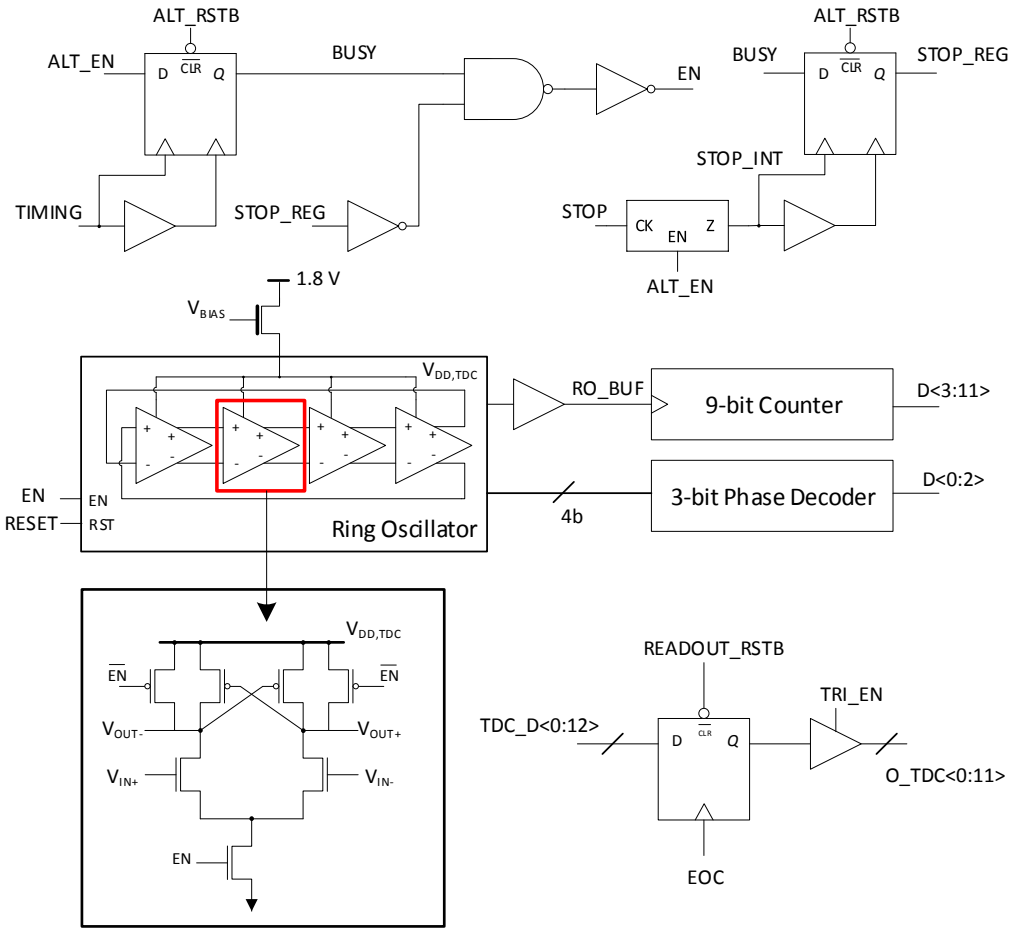


Figure 4.6 – Piccolo ring oscillator TDC architecture.

fixed LSB and low jitter, however since the oscillators must be continuously running, the low overall TDC activity of the Piccolo architecture means that it would consume considerably more power in comparison to the non-continuous open-loop method. For these reasons, a TDC topology based on an open-loop ring oscillator is selected here.

The TDC, see Figure 4.6, is based on a pseudo-differential four stage RO, as in [78]. However, in this implementation, a thick oxide NMOS source follower, M_1, regulates the voltage for the RO. This configuration reduces the effects of IR drops in the ALTDC array on the TDC non-linearity. The frequency of the RO can be controlled with an external voltage bias, V_{BIAS}.

A timing diagram of the TDC operation is shown in Figure 4.7. The TDC is active for detection once the previous slice in the ALTDC chain has detected a photon and asserted ALT_EN, time T1 in Figure 4.7. With ALT_EN high, the input synchronizer, which employs a design [78] with reduced metastability, will assert the BUSY signal, and thus EN, on the next rising edge of TIMING. Similarly, the clock gate is gated on and the internal STOP signal, STOP_INT is able to

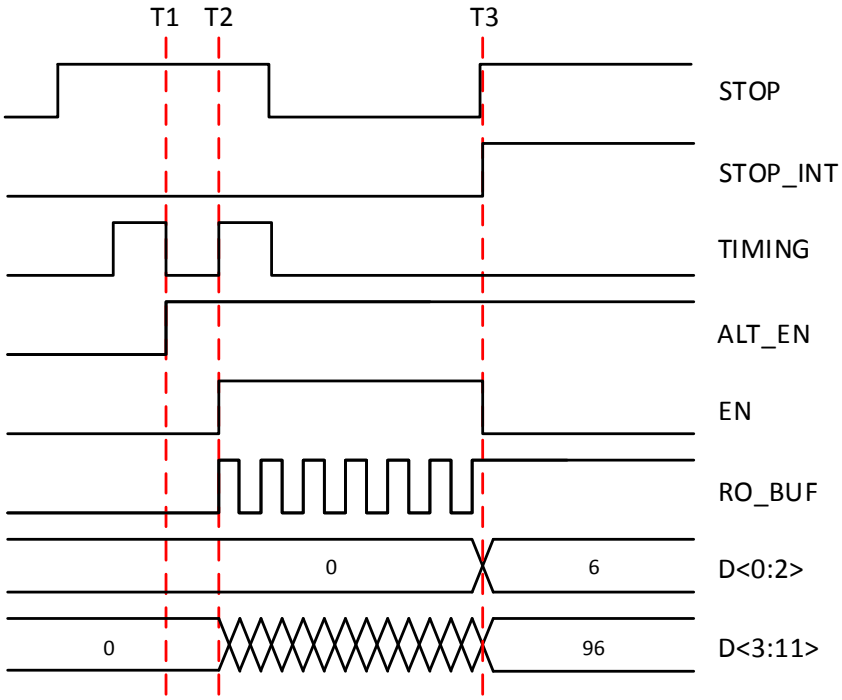


Figure 4.7 – Piccolo TDC timing diagram. TDC is enabled by photon detection in previous slice, which asserts ALT_EN on falling edge of TIMING (T1). Rising edge of timing begins conversion, EN goes high, RO oscillates at 2.56 GHz where the oscillations are counted by the 9-bit counter (T2). On the rising edge of STOP, EN set to low, phase encoder obtains three least significant bits from frozen state of RO, D<0:2>, most significant nine bits obtained from counter, D<3:11> (T3).

toggle the internal nodes of the TDC. This clock gating scheme reduces the power consumption of the TDC array since on average approximately 1/4 of the internal nodes toggled by STOP will be gated on. At the next rising edge of TIMING, time T2, the BUSY signal is asserted, which also asserts EN. This causes the RO to begin oscillating at a nominal frequency of 2.56 GHz. A high speed 9-bit counter counts the rising edges of the RO, RO_BUF, until the first rising edge of STOP, time T3. This triggers the deassertion of EN, freezing the state of the RO. A 3-bit phase decoder obtains the three least significant bits of the TDC code from the frozen phase of the oscillator whilst the 9-bit counter provides the nine most significant bits. A nominal frequency of 2.56 GHz enables the TDC to achieve a LSB of 48.8 ps. Furthermore, since the RO frequency can be configured through an external voltage, it can be set at any value in the range 40-100 ps, depending on the application requirements. A replica RO is also placed on-chip outside of the ALTDC array, with its output divided by a factor of 128 and connected to an output pad. By implementing a feedback loop with external hardware, the frequency of this replica oscillator could be controlled during the sensor operation through V_{BIAS}. Thus, the LSB of the on-chip TDCs can be compensated for process, voltage and temperature (PVT) variations.

4.1.5 Complete Sensor

The sensor was manufactured in a 180nm CMOS technology and occupies an area of 2×5 mm^2. A micrograph is pictured in Figure 4.8, showing how the silicon area is divided between the different blocks. Interestingly, the ALTDC structures occupy over twice as much area as the 32×32 SPAD array. Thus, an argument could be made that the architecture is not very area efficient; the TDCs have been moved outside of the array but still dominate the circuit area.

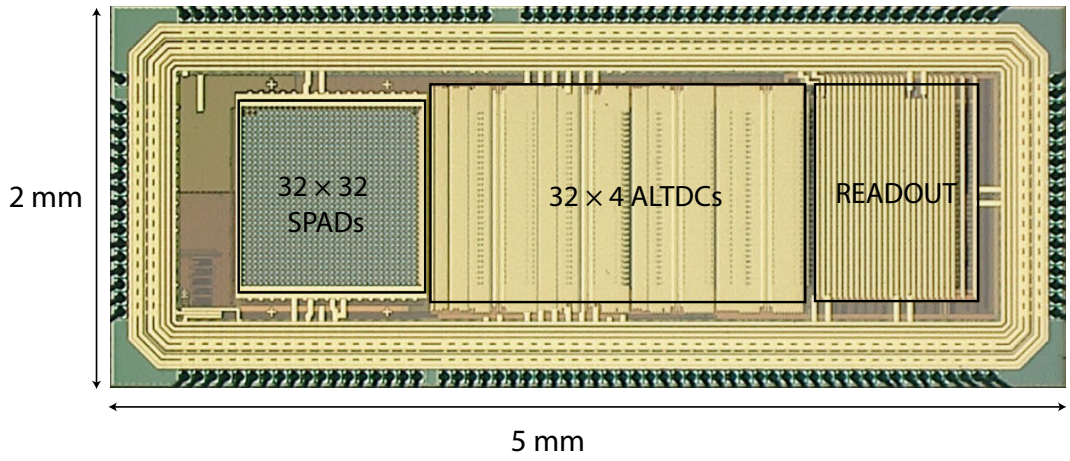


Figure 4.8 – Photomicrograph of Piccolo sensor.

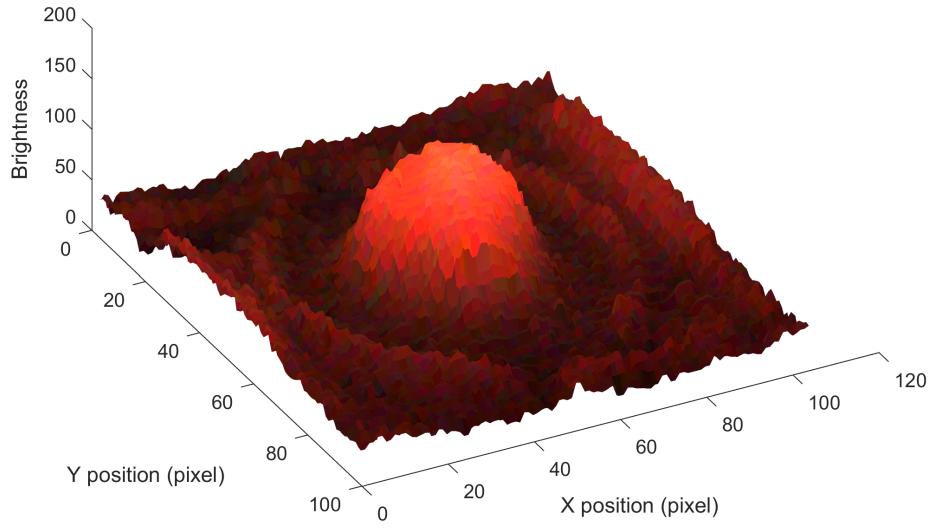


Figure 4.9 – Light emission test of p-i-n SPAD. RGB light emission image is overlayed onto a 3D surface where the height of each pixel represents its brightness.

However, even combining the area of the ALTDCs into the pixel array would result in a fill factor of approximately 9%. Furthermore, from the analysis of the TDC bank sizing in Section 3.3.1, the pixel column could be extended without any adverse effects on fill factor whilst maintaining the same photon throughput for the column. The pixel column could thus be easily extended to 64, 128, or more pixels to provide a much more favorable area comparison.

4.2 Results

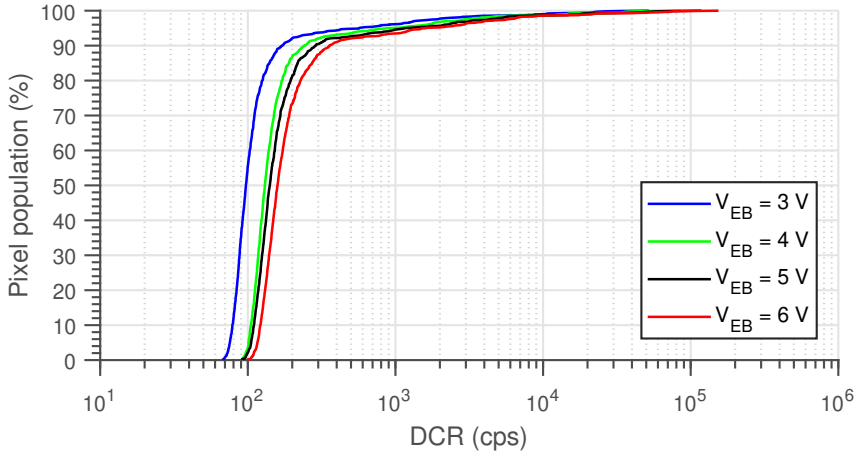
The following section reports both the optical and electrical characterisation of the Piccolo sensor. Although a full characterisation of the p-i-n SPAD is available in [67], elements of the characterisation are repeated here for completeness.

4.2.1 Light Emission Test

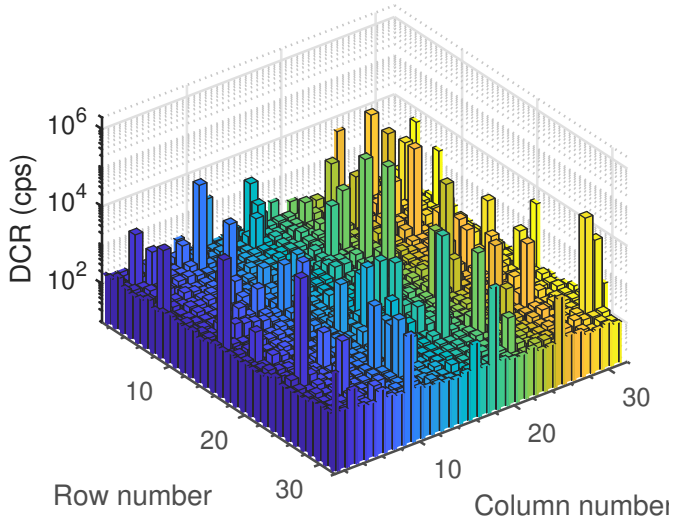
A light emission test of the p-i-n SPAD is shown in Figure 4.9, which has been taken from a micrograph of a separate test structure. The image is obtained by overlaying the RGB light emission image onto a 3D surface, where the height of the pixel is given by the brightness in the grayscale image from the same test. Thus, metal interconnects surrounding the SPAD which reflect light cause a slight raising of the surface. Figure 4.9 confirms a multiplication region focused in the active region of the SPAD.

4.2.2 Dark count rate (DCR)

The DCR has been characterised for the entire array at room temperature with no external cooling applied to the sensor. The population density for excess bias voltages in the range 3-6 V can be seen in Figure 4.10a, whilst the spatial distribution within the array at V_{EB} = 5 V can be observed in Figure 4.10b. Based on the results published in [67], scaling the active area from 113.1μm^2 to the active area used here (227.4μm^2) should result in a median DCR of approximately 80 cps at V_{EB} = 5 V. The median measured on Piccolo at this voltage is 140.9 cps. This increase is almost certainly due to heating of the sensor under normal operation.



(a) Population density



(b) DCR map, V_{EB} = 5 V

Figure 4.10 – DCR characterisation of Piccolo array.(a) Population density shows 90% of pixels exhibiting a DCR less than 290 cps at V_{EB} = 5 V. (b) DCR map shows hot pixels are randomly distributed throughout the array.

The exponential dependence on temperature means even a few degrees rise in temperature will impact DCR significantly. The population density shows a low hot pixel rate, with 90% of pixels exhibiting a DCR less than 290 cps at $V_{EB} = 5$ V.

4.2.3 Afterpulsing

Afterpulsing is measured with the same measurement process as in Section 2.5.3. Whilst afterpulsing has previously been measured for this SPAD in [67], the result of $P_{ap} = 7.2\%$ at $V_{EB} = 11$ V with 300 ns dead time was measured without an integrated quenching circuit and buffer. This means that the capacitance on the SPAD output node is much larger in comparison to the pixel in Piccolo. The result of a larger capacitance is that the charge that flows during an avalanche is much greater, increasing the probability of afterpulsing. The inter-arrival times for the Piccolo pixel are measured from an individual pixel outside of the main array. Due to the event-driven bus and readout scheme, it is impossible to determine exactly which clock cycle a photon was detected in. Therefore, a characterisation of afterpulsing over the entire array is unfortunately not feasible. The inter-arrival times at $V_{EB} = 5$ V with a dead time of 50 ns is shown in Figure 4.11, with the exponential fit of the measured data also plotted. This figure shows that afterpulsing is negligible, there is no deviation from the exponential fit

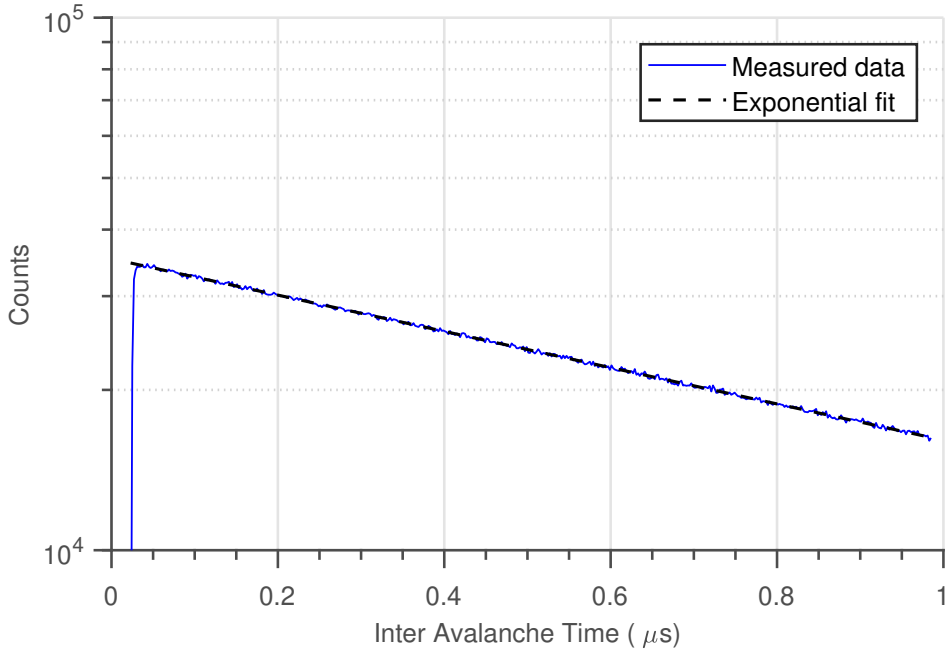


Figure 4.11 – Inter arrival times of photon detections with uncorrelated illuminated and 50 ns dead time. Afterpulsing probability is difference in the area under the plot between the measured data and an exponential fit of the data.

as was seen in Figure 2.12.

4.2.4 Photon detection probability (PDP)

The PDP is measured by placing the SPAD at an output port of 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 at an output port. The plot of PDP versus wavelength of the p-i-n SPAD, measured from a different array employing the same quenching scheme, can be seen in Figure 4.12, for excess bias voltages in the range 3-11 V. It is important to mention, however, that achieving an excess bias voltage of 11 V requires the quenching circuit to be operated at a voltage outside its reliable operating limit, V_{EB} = 5.2 V. In other words, the PDP up to V_{EB} = 11 V can be measured, however, long term operation without any performance degradation cannot be guaranteed. Despite this limitation on the excess bias, a PDP of approximately 12% at 800 nm with V_{EB} = 5 V is in line with the current state-of-the-art [109]. Therefore, the use of the cascode quenching scheme is responsible for an approximate 16% improvement in PDP in comparison to the conventional single transistor quenching and recharge circuit.

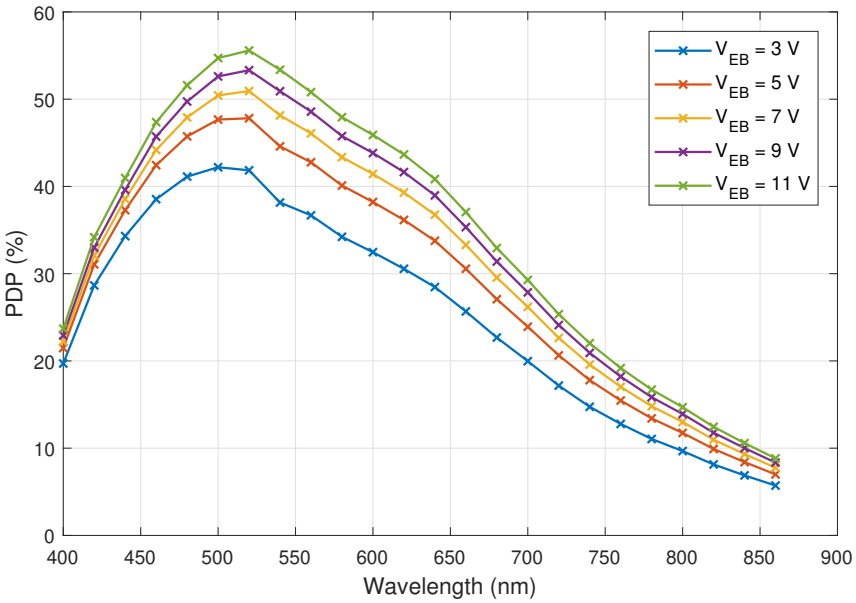


Figure 4.12 – PDP versus wavelength for excess bias voltages in the range 3-11 V. Measurements were carried out by Ivan Michel Antolovic on a different array employing the same SPAD and quenching and recharge structure.

4.2.5 TDC Characterisation

The nonlinearity of the TDC is typically expressed in terms of differential nonlinearity (DNL) and integral nonlinearity (INL). The DNL is the deviation of any given bin in the histogram response from its ideal size, and for a TDC is typically communicated as maximum deviations from the ideal size, e.g. $+0.2/-0.3$ LSB. The INL is the cumulative sum of the DNL and expresses the deviation of the TDC response in LSBs from the ideal response. The standard method for measuring DNL is with the code density test, where the sensor is illuminated with uncorrelated light, thus the photons are randomly distributed in time. Assuming each TDC receives, on average, less than one photon per cycle then in the ideal case each code would receive the same number of photons. The DNL is then the deviation in the number of photons collected per bin in comparison to the ideal case. Figure 4.13 shows the DNL (top) and INL (bottom) for STOP frequencies of 80, 40 and 20 MHz with a LSB of 50 ps.

The result for DNL shows a very good uniformity among bins, a conclusion which is reinforced by a characterisation of the peak-to-peak DNL for all TDCs on the chip, see Figure 4.14 (top). This indicates a very good matching of the capacitances between the stages of the RO and the comparators. The INL achieves good linearity in comparison to the literature [107, 59, 78], however there is a clear periodic behaviour. The INL with STOP = 80 MHz appears to have 2 peaks, whilst at 40 MHz there are 4 peaks. Therefore, it appears that the periodic behaviour is due to coupling of the READOUT_CLK, which switches at 160 MHz, onto the bias voltage for the NMOS source follower, V_{BIAS}. Indeed, there are two factors which make degradation of the TDC performance particularly susceptible to coupling to V_{BIAS}. Firstly, changing V_{BIAS} controls the oscillation frequency of the RO with a "gain", K_{RO}. Due to the RO structure, K_{RO} is very large, approximately 1.87 GHz/V. Secondly, since V_{BIAS} is a voltage which is input to an NMOS gate, this node is high impedance, thus it is susceptible to coupling. Thus, whilst the INL results are good, the DNL suggests it could be improved markedly with better shielding of the V_{BIAS} line. Another interesting feature to observe in the INL is the degradation when a 20 MHz clock is used. The likely cause of this is coupling of the STOP clock onto V_{BIAS} due to a single large peak in the period. This would be consistent with coupling of a signal which is high for half of the period and low for the remaining half. The effect of STOP coupling to V_{BIAS} could explain the increase in peak-to-peak INL as the frequency of STOP decreases.

Since the RO in the TDC is open-loop, that is, its operating frequency is not locked in any way, the frequencies of the ROs in the ALTDC array will run at slightly different frequencies. This is largely because of variations in the fabricated transistor sizes due to manufacturing tolerances. The degree of mismatch between different TDCs, which appears as a variation in the TDC

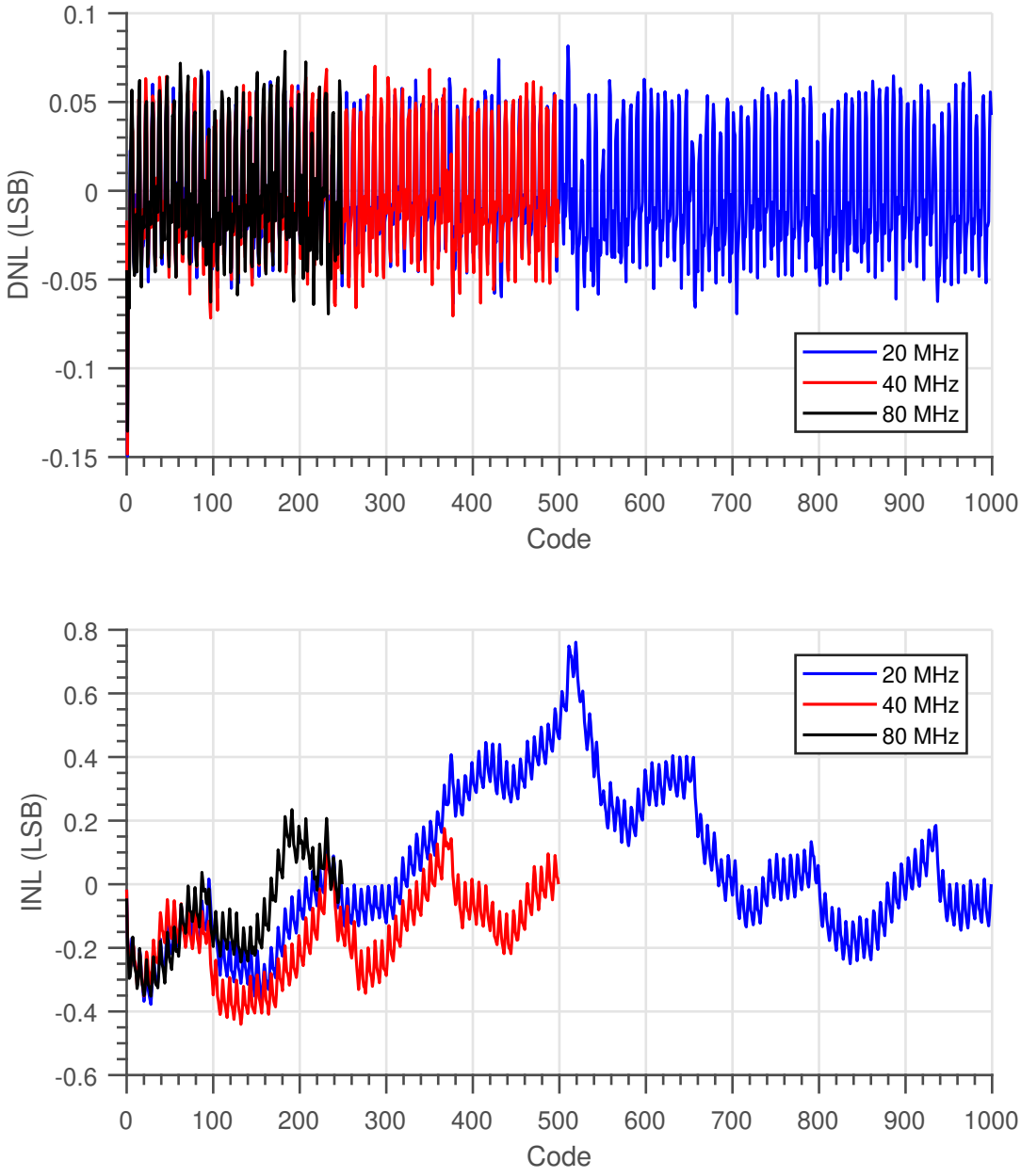


Figure 4.13 – Typical DNL (top) and INL (bottom) at STOP frequencies of 80, 40 and 20 MHz.

LSB, can be characterised by illuminating the sensor with pulsed light at two different known distances, d_a and d_b. The LSB is then calculated as follows,

$$LSB = \frac{d_b - d_a}{c(m_{1,b} - m_{1,b})} \tag{4.4}$$

where c is the speed of light ($\approx 3.0 \times 10^8$), and $m_{1,a}$ and $m_{1,b}$ are the first moments of the TPSFs

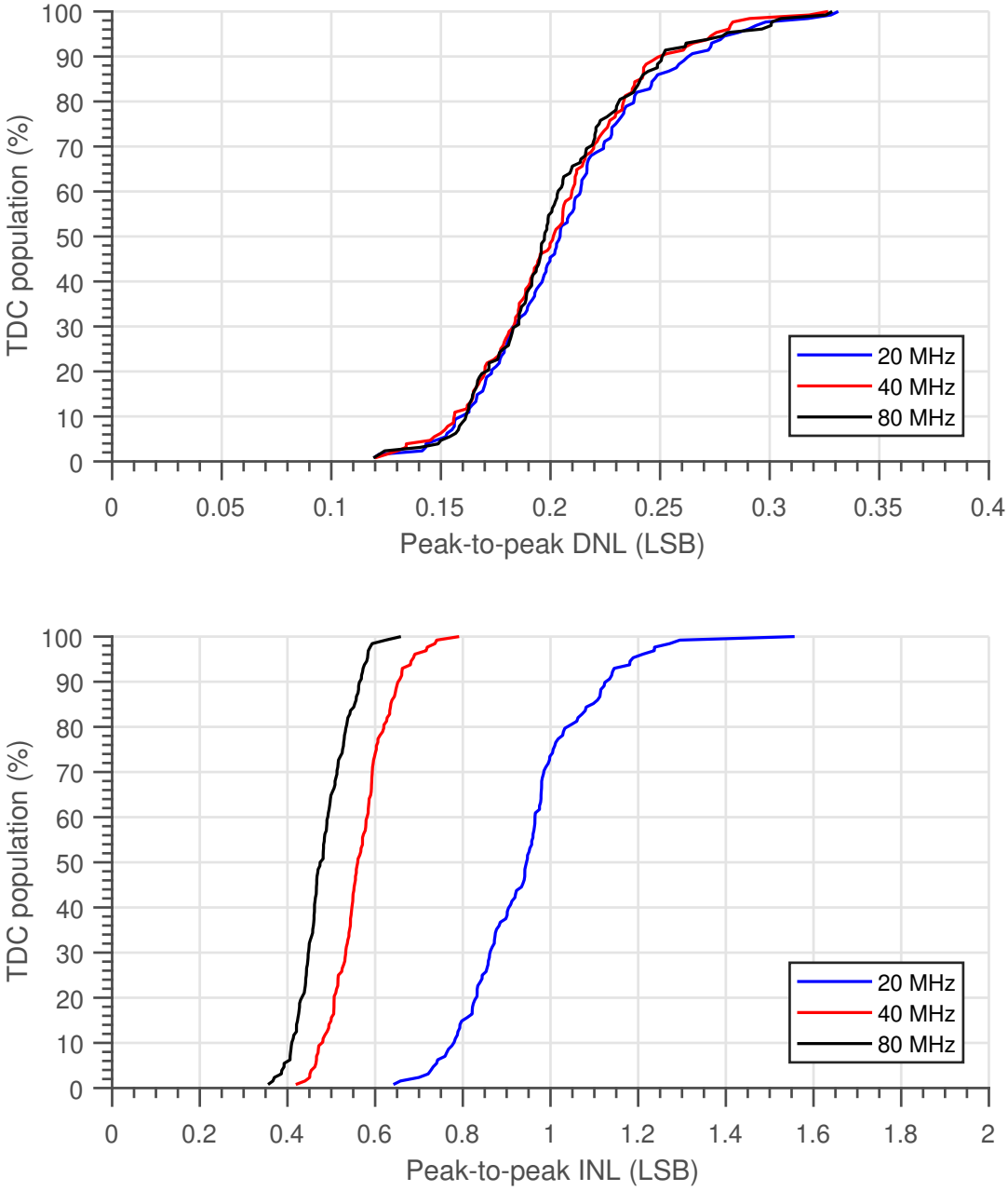


Figure 4.14 – Peak-to-peak DNL and INL at STOP frequencies of 80, 40 and 20 MHz.

of d_a and d_b, respectively. The first moment is given by $m_1 = \Sigma_{i=k}^{l}(i \cdot N_i/N_t) \cdot LSB$, where N_i is the number of photons in the ith bin, N_t is the total number of photons in the TPSF between the kth and lth bins, and the limits k and l define a region where the TPSF signal is above 1% of the maximum signal level. The population density of the LSB in the array is shown in Figure 4.15a, whilst Figure 4.15b displays the spatial distribution of the LSB within the array. With

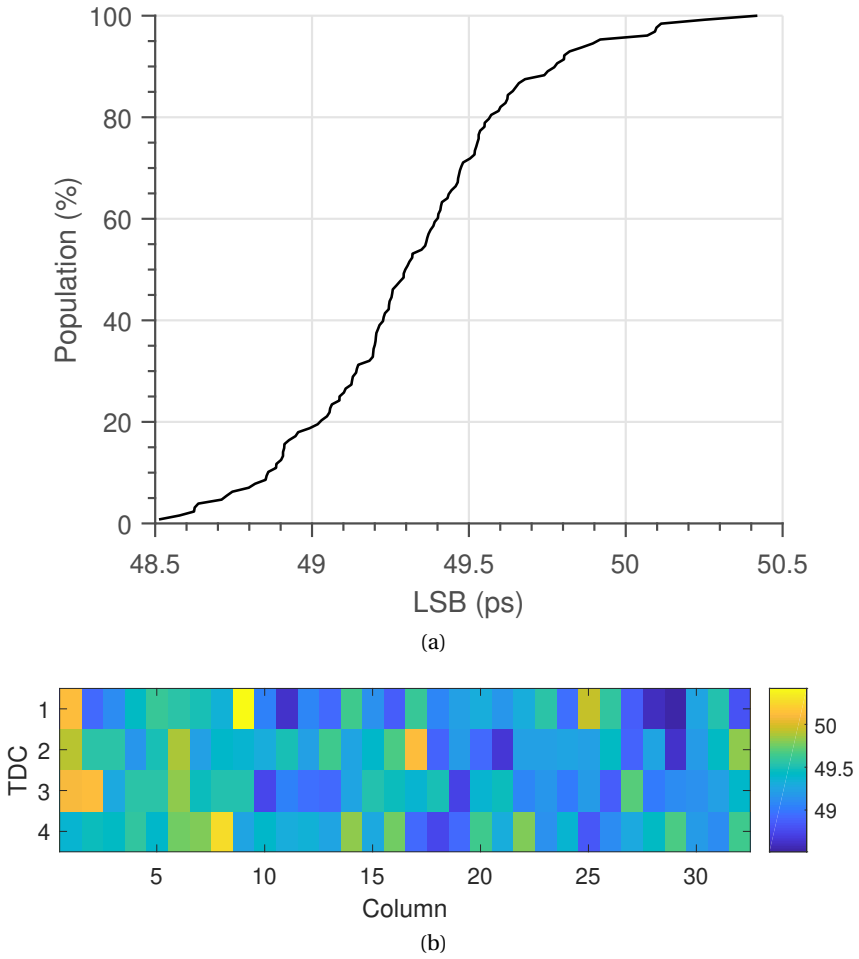


Figure 4.15 – (a) Population density of LSB within the ALTDC array. (b) Spatial distribution of LSB within the ALTDC array.

a bias voltage of $V_{BIAS} = 2.68V$, the median LSB is 49.3 ps whilst 84% of TDCs have an LSB which is within ±1% of this. Figure 4.15b shows a random distribution of the LSB within the ALTDC array.

4.2.6 Timing Response

The system timing response is measured with the electrical trigger from the laser providing the STOP signal to the TDCs. To minimise the jitter contributed by this STOP signal, the trigger from the laser is converted to a GPIO compatible voltage level via a fast comparator. A SuperK Extreme (NKT Photonics) and AOTF illuminate the sensor with pulses of light at a wavelength of 700 nm. The laser power is limited such that, for a given pixel, photon detections occur on less than 1% of laser pulses to limit pileup distortion. Timing jitter is measured from the pixel

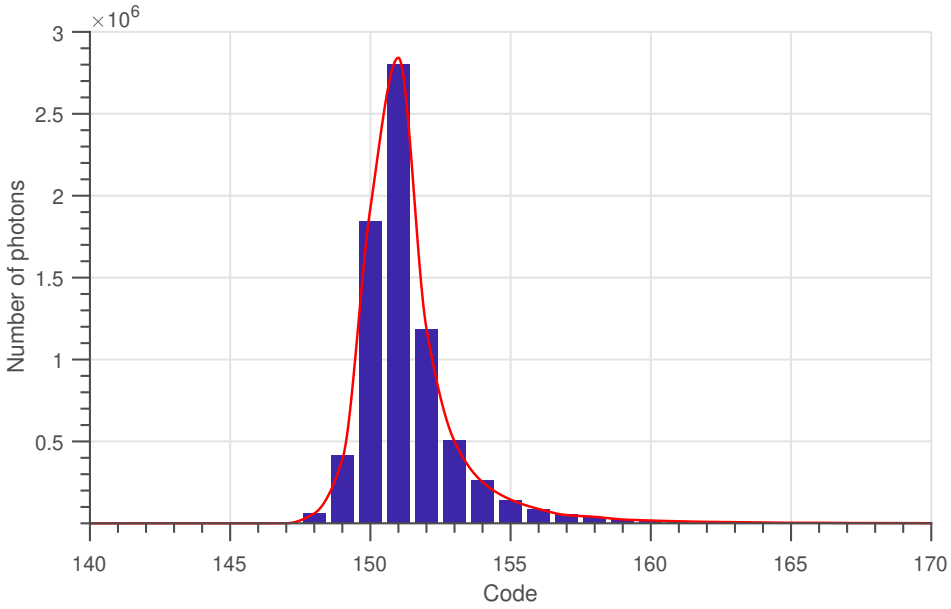


Figure 4.16 – Timing response of Piccolo system at 700 nm, FWHM measured from linear fit (red) is 110 ps.

at the top of the bus to capture degradation of the system timing response due to the shared bus architecture. The timing jitter for V_{EB} = 5 V is shown in Figure 4.16, where the FWHM measured from the linear fit of the histogram is 110 ps. The previous characterisation of this SPAD [67], reports a FWHM of 139.5 and 100.8 ps at an excess bias of 3 and 11 V, respectively. Therefore, it appears that the addition of TDC quantization noise and any jitter contribution from the TDC and bus has not noticeable degraded the timing response. Indeed, at V_{EB} = 3 V a FWHM of 127 ps is achieved, thus the results from Piccolo are marginally better than the previous results. Possible explanations for this are that the laser employed in this experiment has a very low jitter, approximately 20 ps FWHM, and the results in [67] does not subtract the laser jitter from the FWHM measurements so a comparison cannot be made. A further explanation could be the utilization of a discrete quenching and recharge configuration, which implies a much larger capacitance at the SPAD output. This can have a low-pass filtering effect on the fast rising edge of the SPAD, requiring a greater spreading of the avalanche before the discriminator, in this case an oscilloscope, threshold is reached.

4.2.7 Signal-to-noise ratio (SNR)

With measurements for PDP, DCR and timing jitter, it is possible to plot the SNR, computed according to Equation 2.1. The relative SNR is plotted in Figure 4.17 for V_{EB} equal to 3, 4 and 5 V, where the calculated values have been normalised to the value at V_{EB} = 3 V. With the

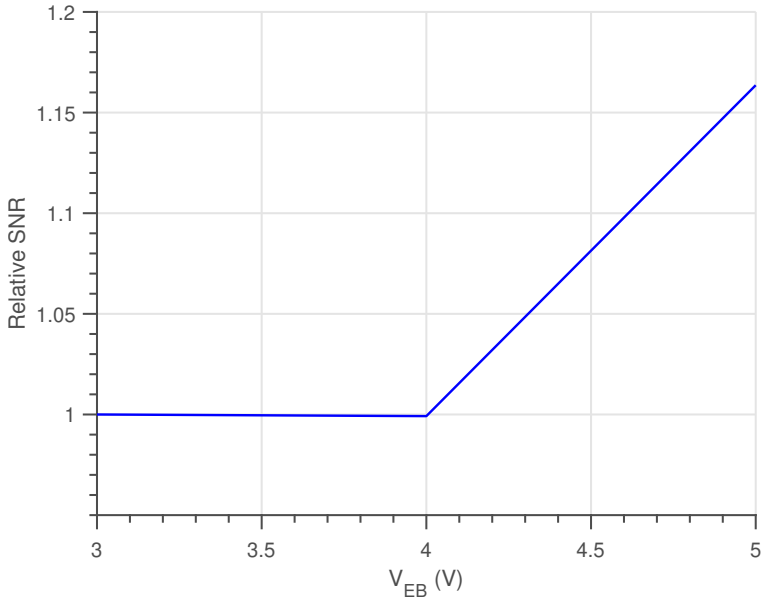


Figure 4.17 – Relative SNR as a function of V_{EB}.

absence of a measured value of PDP at V_{EB} = 4 V, an average of the results at V_{EB} = 3 V and V_{EB} = 5 V is employed. The FWHM timing measurements at each excess bias voltage given in picoseconds are 127 (3 V), 121.6 (4 V) and 110 (5 V). The figure displays a flat SNR up to 4 V, at which point a gain of approximately 16% is observed at 5 V. Although this would appear to validate the additional excess bias achievable with the cascode quenching, almost all of this gain is due to improvement in FWHM. Thus, the improvement is certainly valid for LiDAR, since there will likely be a large background noise contribution in this application. For NIROT the benefit is less clear since the FWHM of the TPSF is dominated by the response of the medium, and NIROT systems can be shielded from ambient light, e.g. with a tube as in [110]. To properly evaluate the optimum biasing point, the contrast-to-noise ratio [32] of NIROT system should be measured over a range of excess biases.

4.2.8 Flash Ranging Measurement

In NIROT measurements, tomographic reconstructions are achieved by acquiring a TPSF for each pixel in the array. To validate the Piccolo sensor for operation in this mode, a flash LiDAR measurement is perfomed where a target object is illuminated and a TPSF of the reflected light collected on a per-pixel basis. The data is corrected for LSB variations among the different TDCs, as well as an additive time offset to compensate for clock skew in the STOP clock distribution network. A 3D image can be constructed by calculating m_1 for each pixel. 32 × 32

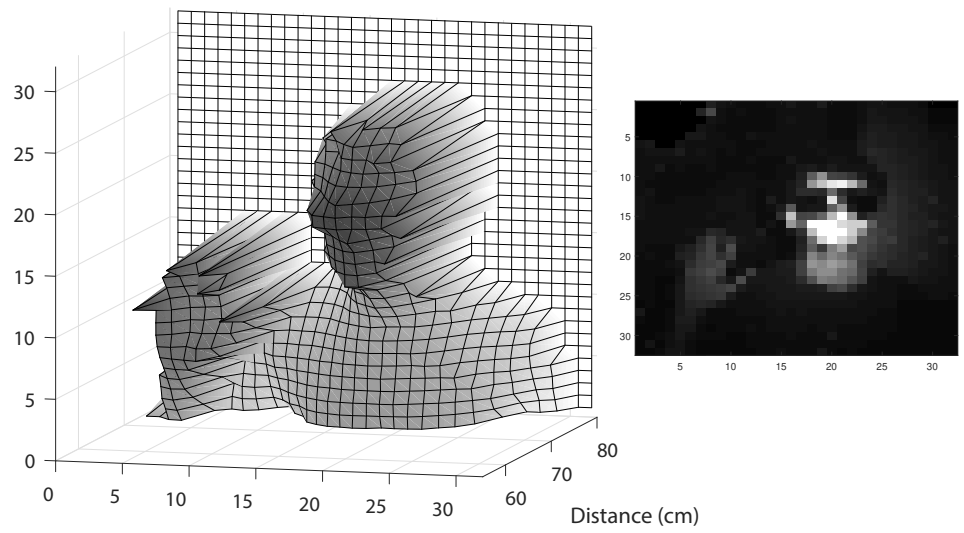


Figure 4.18 – 32 × 32 flash ranging measurement of a human subject at a distance of 0.7 m with 2D intensity image inset.

flash images are shown of a person with the right hand raised at a distance of 0.7 m in Figure 4.18. The target is illuminated with a supercontinuum laser (SuperK Extreme, NkT Photonics) and an AOTF, operating at a wavelength of 650 nm. The 3D image accurately reconstructs the target object whilst millimetric detail can be observed due to the calculation of m_1. The ability to perform 3D ranging also raises the possibility of using the camera for co-registration of the measurement surface in a clinical NIROT application.

4.2.9 Phantom Validation

To validate the Piccolo sensor for the NIROT application in reflection mode, time resolved measurements were carried out on a silicon phantom, Figure 4.19, with $\mu_a = 0.131$ cm^{-1} and $\mu_s' = 7.0$ cm^{-1} at 806 nm. Pulsed light from a super continuum laser and AOTF is guided into a fibre switch with 1 input channel and 24 output channels. In the current system, 11 of these output channels are connected to fibres (GIF625, Thorlabs, USA) which are embedded into a ring which makes contact with the object under study, in this case a 4 cm thick silicon phantom. The ring is black and has a diameter of 45 mm with a surface covered in biocompatible soft silicon [111]. Transparent windows in the ring allow light to be coupled from the source fibres to the object under study. A cylindrical tube with diameter of 25 mm runs through the middle of the ring to the sensor, encompassing the ultra-wide CS-mount lens which projects the exiting light onto Piccolo. The tube shields the sensor from ambient light from outside, and is

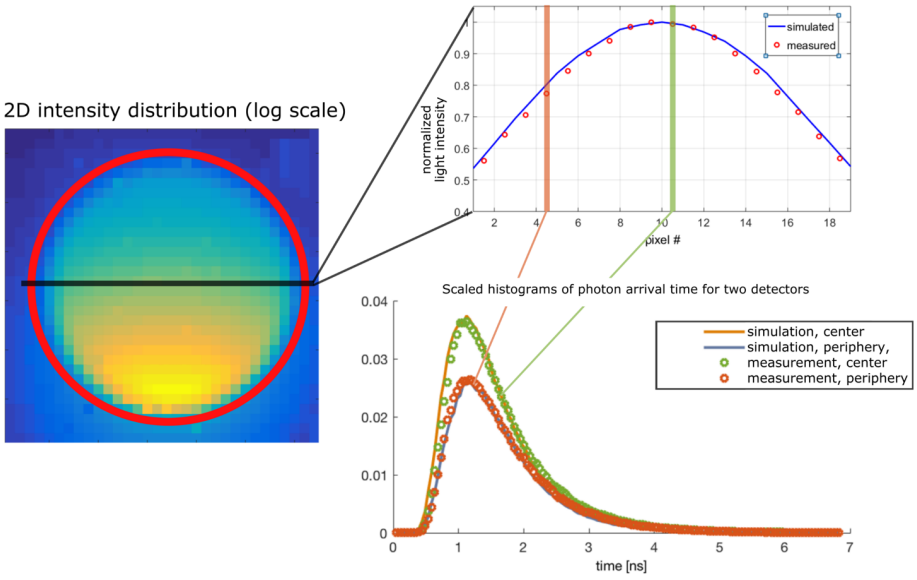


Figure 4.19 – TPSFs of silicon phantom measurement with optical properties $\mu_a = 0.131$ cm^{-1} and $\mu_s' = 7.0$ cm^{-1} at 806 nm, compared to MC simulation with identical optical properties.

coated with an anti-reflective coating (Vantablack, UK) to minimise reflections from the inside. This configuration results in a FOV of 25 mm. More complete details of the system can be found in [110]. Figure 4.19 shows the TPSF from two different detectors, at the center and the edge of the FOV. The TPSFs acquired by the Piccolo camera show good agreement with simulations obtained by a MC forward simulation [112, 113, 114]. This minimal validation shows that the system is suitable for application to NIROT measurements. Before the system can be employed in the clinic, however, extensive testing and image reconstructions on homogeneous [115] and non-homogenous [116] phantoms will be required.

4.2.10 Timing Stability

An important characteristic of any clinical system is the stability of measurements over time. To verify the stability of time-resolved measurements, the camera is illuminated by a pulsed laser with an optical diffuser and sheet of white paper in between the laser and the camera to even the illumination over all pixels. The stability of the system is characterised by measuring the deviation in the first moment, Δm_1, of the TPSF in 91.5% of pixels, e.g. 937 SPADs, over a period of 8 hours. To isolate the response of the camera system, the laser was run at the measurement power for a period of 1 hour prior to beginning the stability measurement. The camera was then only powered on at the beginning of the experiment. The results of the

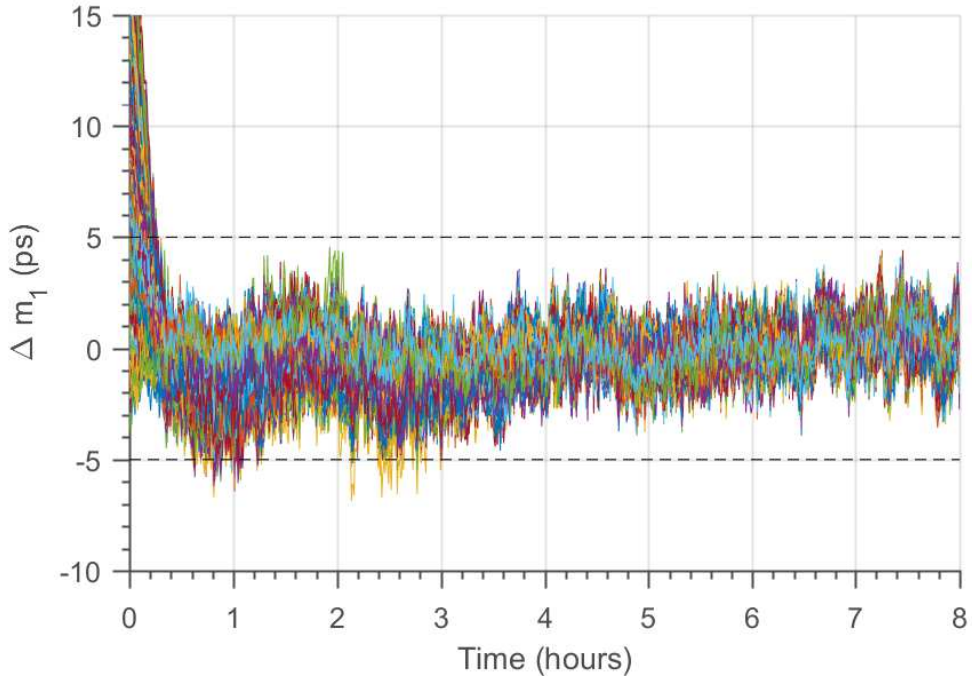


Figure 4.20 – Stability of m_1 for 91.5% of pixels in the Piccolo system over 8 hours at 800 nm. m_1 is well bounded after an initial warmup period.

stability measurements are shown in Figure 4.20. After an initial period of 20 minutes, m_1 settles into a stable regime, at points up to 3 hours there are minor deviations outside the ±5 ps bounds employed for other sensors [50], and after this period it is well bounded. It is important to note that whilst the LSB of the TDC can be compensated for temperature variations, this has not been implemented here. Therefore, to operate the camera in clinical measurements with less than 20 minutes warm up time, it is likely that compensation of the TDC and/or temperature stabilisation of the camera system would be required.

4.2.11 Power Consumption

The power consumption of the Piccolo sensor was measured during a LiDAR ranging measurement with a global throughput of 35.5 Mevents/s. At this activity the total power consumption, P_{total} = 310 mW, with the breakdown between the individual circuit blocks shown in Table 4.1. The core component is almost entirely consumed by the readout from the ALTDC array to the I/O pads. As the event rate is increased to greater activity levels, the IO power consumption would be expected to scale linearly. Increases would also be expected for the ALTDC array and the core circuitry, however, since a large proportion of this power consumption is due to the

Table 4.1 – Piccolo power consumption

	Power (mW)	Contribution (%)
ALTDC array	93	30
Core	86.1	27.8
I/O	82.7	26.7
Pixel array	34.4	11.1
Test circuits	13.8	4.4
P_{total} (mW)	**310**	

clock distribution, the increases would not be proportional to the throughput.

4.2.12 State-of-the-art Comparison

A comparison of Piccolo against the state-of-the-art in time-resolved SPAD sensor arrays is shown in Table 4.2. The most notable feature of Piccolo is the unprecedented PDE at 800 nm, 3.4% in comparison to 0.19% of the nearest competitor. This is enabled by the high PDP of the SPAD, but is mostly due to the massive improvement in fill factor as a result of the event-driven architecture. The DNL and INL are bettered only by [93] which employs an interpolating TDC with the sliding-scale technique. The maximum throughput of the sensor, TP_{max}, is exceeded only by [95], which achieves 2 Gevents/s at the expense of a large I/O bandwidth and 8.79 W power consumption. Also, the SPAD array in Piccolo is a factor of 4 smaller than that of [95], thus the maximum pixel activity rates of the two sensors differ by a factor of just 2.3.

4.3 Conclusions

The event-driven TDC sharing architecture from Chapter 3 is used in the design of a 32 × 32 time-resolved SPAD sensor, Piccolo. With just 4 ALTDC slices per column, a fill-factor of 28% is achieved with a pixel pitch of 28.5μm. The combination of high PDP from the p-i-n SPAD and 28% fill-factor results in a PDE of 3.4% at 800 nm, which is approximately 17 times larger than conventional time-resolved SPAD sensors. Furthermore, since the data readout is through a tri-state bus which receives only valid timing data, it avoids the problems of low efficiency due to the readout of null data, or high power consumption of event driven readout schemes. A GPIO pad dedicated to each column enables a maximum throughput of 220 Mevents/second. At this event rate, 10^6 can be accumulated for every pixel in the sensor in approximately 4.6 seconds. Validated with phantom and timing stability measurements, the sensor is highly relevant for clinical NIROT measurements.

Table 4.2 – Piccolo state-of-the-art comparison

	[95]	[93]	[59]	**Piccolo**
Array size	64 × 64	32 × 32	160 × 128	**32 × 32**
Fill factor (%)	0.77	3.14	1	**28**
PDP at 800 nm (%)	4	6	6	**12**
PDE at 800 nm (%)	0.03	0.19	0.06	**3.4**
DCR (cps/μm^2) at (V_{EB})	30.7 (2.5 V)	0.14 (6 V)	2 (0.73 V)	**0.62 (5 V)**
TDC LSB (ps)	62.5	312	55	**48.8**
DNL (LSB)	-2/+2	-0.06/+0.06	-0.3/+0.3	**-0.2/+0.2**
INL (LSB)	-4/+4	-0.22/+0.22	-2/+2	**-0.47/+0.77**
Timing FWHM (ps)	200	609	140	**110**
Technology (CMOS)	130 nm	0.35μm	130 nm	**180 nm**
Die area (mm × mm)	9 × 4.1	9 × 9	12.3 × 11	**5 × 2**
Data bandwidth (Gbps)	42	1.024	51.2	**5.12**
TP_{max} (Gevents/s)	2	0.064	0.102	**0.220**
P_{total} (W)	8.79	0.43(2.8)[1]	0.55	**0.31**[2]

[1] Gating functionality, where the P_{total} is given for 1 gate (50 gates) per cycle.
[2] Measurement performed with a global throughput of 35.5 Mevents/s.

5 A 252 × 144 Event-driven High-throughput Time-resolved SPAD Sensor

Whilst time-resolved measurements present exciting possibilities for a wide range of applications, the path to producing sensors with similar resolutions to their 2D counterparts is extremely challenging. As well as the difficulty of designing efficient sensor architectures which can scale to large resolutions, the data bandwidth bottleneck means that often increased resolution is only obtained at the expense of decreased pixel activity. This is a major problem for the NIROT application, since both large array size and fast acquisition speed are required. This chapter begins with a review of large format time-resolved SPAD based sensors and the major issues faced when trying to scale existing architectures. A 252 × 144 high-throughput event driven sensor, which builds on the architecture in Chapter 3, is presented.

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". Ocelot was a collaborative design carried out with Chao Zhang and Ivan Michel Antolovic with a division of labour among the different circuit blocks. Ivan Michel Antolovic was responsible for the design of the pixel array and the digital design of the pixel masking scheme. He also co-designed the collision detection bus with the author contributing the concept of the bus repeater scheme. The author was responsible for the dual-clock TDC and PLLs. The concept of the partial-histogramming was conceived jointly with Chao Zhang, with Chao carrying out the full digital design of the partial-histogramming readout (PHR), conception of the peak detection method, and also design of the address latch structure.

5.1 Large Format Time-resolved SPAD Sensors

The prospect of large format, e.g. greater than 128 × 128, time-resolved SPAD sensors will likely prove very powerful for a number of applications. In NIROT, it raises the possibility of obtaining image reconstructions over a larger FOV without the need for mechanical scanning or sacrificing spatial resolution. Designing large format time-resolved sensors is, however, not trivial. To the best of the authors knowledge, the largest camera to date with on-chip TDCs was reported in [59], with an array size of 160 × 128. With a pixel pitch of 50μ m, scaling this architecture to larger formats is very expensive in terms of the silicon area required. Furthermore, the fill factor of 1% results in a low PDE and implies that the size of the circuitry cannot be reduced to reach smaller pixel pitches. Finally, the large area occupied by the array in TDC-per-pixel architectures means that there is not as much silicon area to innovate in other areas of the circuit architecture, e.g. implementing an event driven datapath such as [95] on a large scale.

Time-gated sensors [117, 118, 119, 120] based on SPADs are an alternative method for time-resolved measurements which can achieve a much smaller pixel pitch. In this approach, the SPAD pixel only registers photon events during a narrow time window, typically some 100s of picoseconds [120] to nanoseconds [117, 120]. The photon event is stored by an in-pixel capacitor registering either the binary presence of a photon [117, 118, 119], or the number of photons detected over a number of frames [120]. By temporally shifting the window in small increments, a TPSF can be obtained. With this approach, sensor resolutions up to 512 × 512 pixels [118] have been designed, in this case occupying a total area of 9.5 × 9.6 mm^2. For comparison the 160 × 128 TDC-in-pixel sensor in [59] occupies an area of 12.3 × 11 mm^2. Therefore, the time-gated approach appears to offer a compelling alternative for high resolution time-resolved measurements. However, the major downside to the time-gated approach is that the acquisition time is multiplied by the number of gates required for the measurement. Since in NIROT the TPSF can measure some nanoseconds in duration, this approach will result in a prohibitively long measurement time.

The architecture for the Piccolo sensor in Section 3.3 then represents a major leap forward for implementing a large array format high throughput time-resolved SPAD sensor. By placing the TDCs outside of the pixel array and enabling a dynamic reallocation of events to the ALTDC bank, the number of TDCs and pixels can be independently chosen. Furthermore, as seen in Section 3.3.1, the number of TDCs required is determined by the data bandwidth of the readout and the STOP frequency of the laser. Therefore, for a given bus activity, the sensor array can be scaled to larger numbers of rows without requiring additional area occupation by

TDCs. Thus, the resulting sensor can have a fill-factor which is competitive with time-gated approaches, if in a slightly larger pixel pitch, whilst avoiding the long acquisition time resulting from multiple gate windows.

Despite the major benefits of the Piccolo architecture, there remains a major challenge in scaling the architecture for NIROT. Consider the Piccolo sensor, a 32×32 sensor with a 160 MHz GPIO pad dedicated to each column. Since the bus activity rate is given by the GPIO pad speed, increasing the number of pixels per column by a factor N, decreases the pixel activity rate by the same factor N. Thus, increasing the number of pixels results in a decrease in the acquisition speed. As discussed already in Section 3.3.1, the throughput can be improved by utilizing a higher speed I/O standard, e.g. LVDS. However, the pads for such standards are not commonly available in standard foundry libraries. Furthermore, dedicating a data output pad on a per column basis for a large number of columns, e.g. 256, requires a complex system design since many off-the-shelf FPGA evaluation boards do not have enough FPGA I/O pins available to interface to such a sensor. It is clear then that there is a stringent need for data processing methods which can reduce the volume of data to be transmitted between the sensor and FPGA, thus increasing the pixel activity rate without requiring additional output data bandwidth.

5.2 Sensor Architecture

The remainder of this chapter reports the design and testing of a 252×144 time-resolved SPAD sensor, hereafter referred to as Ocelot, which expands on the Piccolo architecture presented in Chapter 3. The same 180nm CMOS technology and wide spectral range p-i-n SPAD [67] as Piccolo are employed. As such, the focus is given to those aspects of the sensor design which are required to successfully scale the Piccolo architecture without sacrificing other aspects of system level performance, e.g. acquisition speed, in the NIROT application.

The architecture for the Ocelot sensor is shown in Figure 5.1. The 252×144 SPAD pixel array is divided into two 126×144 sub-arrays, with each pixel sub-array allocated its own ALTDC array, readout and data pads, Figure 5.1a), as in the Piccolo architecture, Section 3.3. Due to the array size, 4 half-columns share a single 160 MHz GPIO data pad. This sharing of data pads is necessary to reduce to the total number of pads on the sensor such that a commercial FPGA module can be utilized for sensor control and data processing.

Each half-column of 126 pixels, share 6 ALTDCs, Figure 5.1b), utilizing the dynamic reallocation scheme from Section 4.1.3. The TDC architecture employs a dual clock RO-based

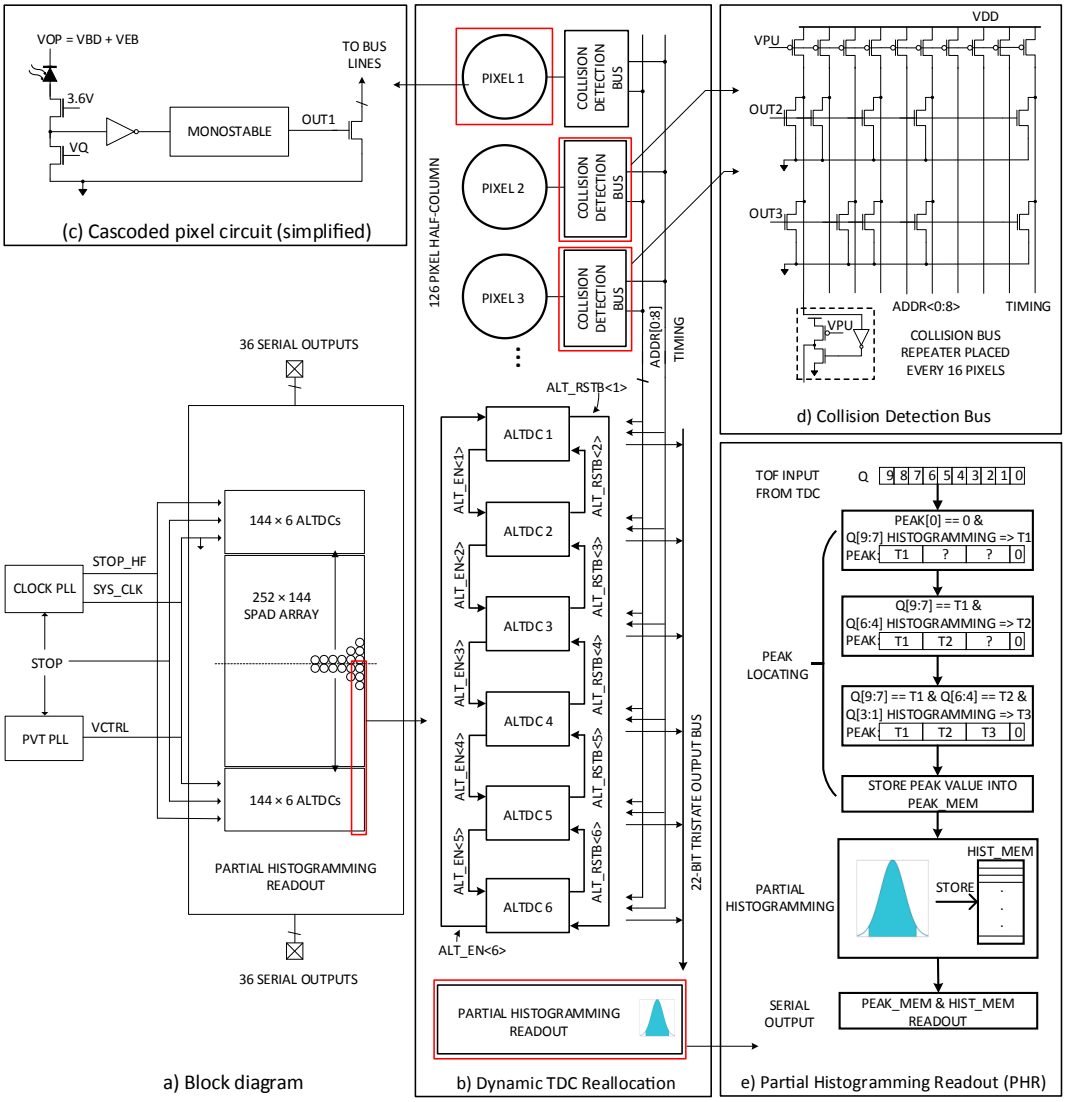


Figure 5.1 – Ocelot architecture.

architecture which builds on [78] and the TDC circuit already presented in Figure 4.1.4. A TDC slice is made active via ALT_EN once the previous slice in the chain has detected a photon, and reset once the subsequent slice has readout a detection. This scheme enables up to 5 photons to be detected per cycle, per column.

The pixel, Figure 5.1c), employs the same design as in Section 4.1.2. The output signal from the SPAD is shrunk via a monostable, where the pulse width can be controlled by a bias voltage within the range 0.4-5.5 ns. This narrow pulse is transmitted to the ALTDC bank via a collision detection bus, Figure 5.1d), which requires 9 address lines, ADDR<0:8> and a TIMING line to encode the half-column. With a pixel pitch of 28.5 μm, a 126 pixel half-column requires a bus

of length 3.6 mm. At this length, the parasitic capacitance of the bus lines heavily degrades the transition times of the address and timing lines. As such, a repeater scheme was developed which could maintain the collision detection functionality of the bus. The line repeaters are placed in sequential fashion throughout the column, occupying a reserved space alongside the pixel circuitry, where only one line is repeated per pixel.

Ocelot is capable of operating in TCSPC, single-photon counting, peak-detection and partial histogramming modes. The latter two modes are enabled by a static random access memory (SRAM) based partial-histogramming readout (PHR) block, Figure 5.1e), which provides histogram peak detection and 16 5-bit bins for histogram storage on a per pixel basis. This readout is based on the principle that in SPAD based time-resolved sensors, the overwhelming majority of the data is contained within a narrow range of time bins located in the immediate vicinity of the histogram peak. Therefore, the requirements for on-chip memory can be relaxed if only the photons belonging to the most-populated bins are stored in the SRAM before readout. As such, the readout process works in two phases. In the first, the detected photons received by each pixel are sampled and the peak of the histogram located. The readout can then switch to partial-histogramming mode whereby a window of 16 bins around the peak is formed, and photons arriving in these time bins are stored in memory in a compressed histogram rather than transmitted directly as raw data.

To compensate for process and temperature variations, the bias voltage which sets the frequency of the individual ROs, V_{CTRL}, is generated by a PLL before distribution across the array [107]. Since the oscillator included in the PLL architecture is a replica of the RO in the ALTDC array, the frequency of the ROs in the ALTDC array will remain almost identical to that of the PLL over process and temperature changes. A second PLL is included to generate the STOP_HF and READOUT_CLK clock signals necessary for the dual-clock TDC and PHR.

5.3 Scalable Collision Detection Bus

As discussed at length in Chapters 3 and 4, there are some major advantages in opting for an TDC sharing architecture, namely higher fill factor, more efficient use of silicon, and easier data readout. Unfortunately, placing a shared bus in between the pixels and the TDCs adds a potential source of signal degradation. In the TDC-per-pixel approach, conversion from a photon arrival to a digital code is performed directly in the pixel, thus the data leaving the pixel is already a binary code which has a large noise immunity. Whilst this is also true for the address lines in the collision detection bus, the TIMING line, in contrast, is highly sensitive

to noise as the photon time-of-arrival is encoded in its transition. Noise coupling onto this signal will result in poorer timing performance. Thus, for the TDC sharing approach to be a valid alternative to the TDC-per-pixel approach for large arrays, the timing response of the system should not be heavily degraded by the shared bus.

Bus jitter

The transition time of any digital logic gate displays some stochastic jitter. This is due to the noise sampled onto the output capacitance whilst the gate is idle, and the thermal noise of the transistor driving the output to the transition threshold of the next gate. In general, the jitter of digital gates is not significant providing the edges transition at a reasonably fast rate, e.g. in 180 nm 100-300 ps is satisfactory. Indeed, this fact enables the design of TDCs with many internal logic transitions without adding a large amount of jitter to the timing response of the system.

Coupling

Due to the dense layout of the pixel array to achieve a high fill-factor, the TIMING signal must share the same metal layer with some address lines. Since the length of the bus with 126 pixels is 3.6 mm, the fringing capacitance between the TIMING line and an adjacent address line will be large. For example, in 180nm CMOS two metal-4 lines 3.6 mm long at minimum spacing have a fringing capacitance of approximately 140 fF. This will lead to significant coupling of the signals from the address lines to TIMING. This will likely result in a longer bus dead-time due to the time for the line to settle after a detection. Additionally, coupling would change the fall time of the TIMING signal on the bus, with the degree of coupling dependent on which bus lines were activated at the same time as the TIMING line. To avoid these negative effects, the TIMING line is shielded by grounded metal interconnects which run parallel to the TIMING line. The disadvantage of this approach in the Ocelot architecture is that the dense pixel array requires a minimum spacing between the shield lines and TIMING, thus the TIMING line has a larger capacitance.

Bus Line Capacitance

Due to the shielding lines adjacent to the TIMING line, its capacitance is large, approximately 720fF for a 3600 μm line. As such, achieving a fast falling edge on the detection of a photon requires a large current drive for the NMOS transistors which drive the lines. For example, for

a falling edge of 200 ps, an average current of 6.48 mA is needed. This current presents a major problem for the array. Since the current in a transistor scales linearly with the device size, large transistors would be required in pixel, with one large NMOS for each line. This would reduce the pixel fill factor, or require a larger pitch to maintain it. Furthermore, this problem will only increase as the bus is extended further.

Bus Repeater Scheme

To reduce the size of the in-pixel NMOS transistors which drive address and timing lines, a bus repeater scheme is implemented, see Figure 5.2. The bus line repeater consists of an inverter, a PMOS transistor to pull the next section of the bus to V_{DD}, and an NMOS transistor to pull the bus down. Thus, these bus line repeaters can be distributed throughout the column to sharpen the signal edges and divide the bus lines into sections. This has the effect of reducing the capacitance seen by the transistors in the pixel by a factor of approximately $M+1$, where M is the number of times each line is repeated along the length of the column. Since the capacitance is reduced by a factor of $M+1$, for a given signal transition time, the driving current and thus the size of the driving transistors also decreases by a factor of $M+1$. Therefore, the bus repeater scheme maximises the pixel fill factor by reducing the total driving transistor area required for a given transition time.

For the Ocelot architecture, $M = 7$, is the optimum number of stages. For numbers greater than M, the area required for an extra bus line repeater is not offset by the area saved in the pixel due to a reduced load capacitance. The 126 pixel half-column is thus divided into 8 different sections, where the section closest to the ALTDC array has 14 pixels. For efficient area use, a section of each pixel is reserved for use by the bus repeaters. In a section of 16 pixels, 10 of these reserved spaces are used for bus line repeaters whilst the remaining spaces contain decoupling capacitors. Thus, the architecture is highly scalable. For example, with only three more bus lines, up to 924 pixels could be addressed in a single column with the same pulse width and fill factor as the 126 pixel version.

5.4 Partial Histogramming Readout

A major benefit of time-resolved SPAD based sensors is the potential for massively parallel time-resolved measurements. However, as discussed in detail already, the rate at which these measurements can be acquired is heavily limited by the data bandwidth of the sensor. Achieving fast measurements with a large number of pixels is a major challenge. To the best

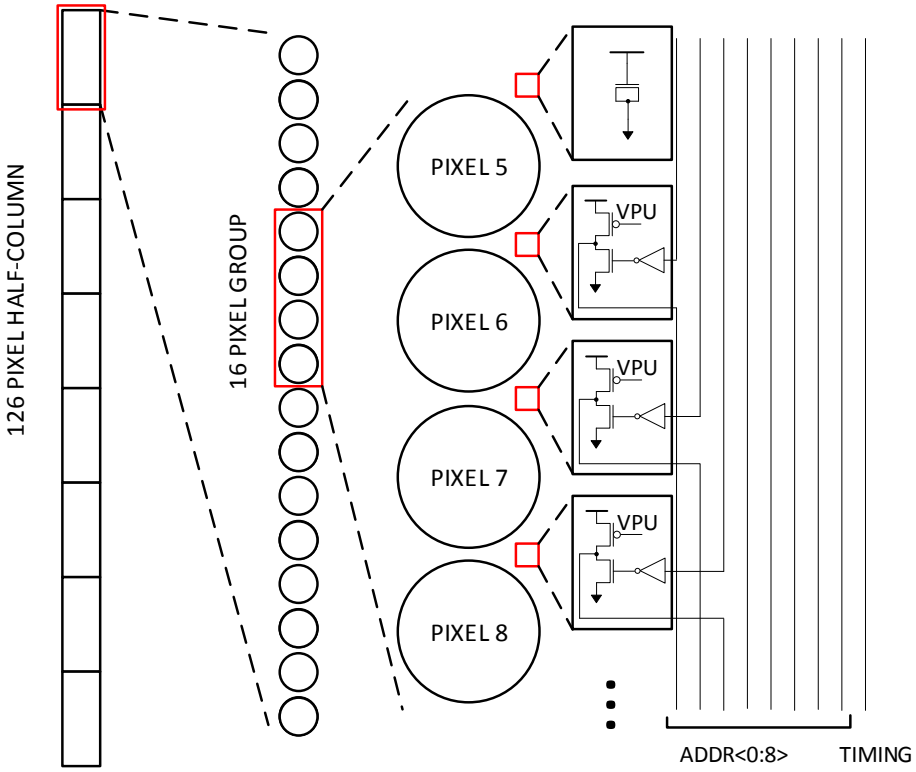


Figure 5.2 – Bus repeater distribution within the 126 pixel half-column. Half-column is divided into 8 groups, where each bus line is repeated once within the group. Bus repeaters are placed in reserved space within the pixel, where only one bus repeater is placed per pixel. Reserved spaces without bus repeaters are filled with decoupling capacitors.

of the author's knowledge, the fastest SPAD based TCSPC systems to date have implemented direct-to-histogram TDCs. In this case, the TDC employs a ripple counter for each TDC bin which is incremented on the arrival of a photon. This method can achieve extremely high photon rates, since the compressed histogram data is far smaller than comparative size of the raw data, it can be said to have a high compression efficiency. However, thus far these sensors have been restricted to point [80] and line [102] formats due to the large area overhead for the direct-to-histogram TDC. Scaling to a large array format would appear to only be feasible in a 3D implementation with the data processing tier in a highly scaled digital technology, thus reducing the TDC size.

The advantage of the TDC-sharing architecture for on-chip histogramming is that the area occupation for the timing measurements is reduced, thus in principle there is a larger area available for additional circuitry. In Ocelot, a per-pixel SRAM based on-chip histogramming scheme is implemented, which, to the best of the author's knowledge, is the first such scheme

for a full array. Despite the high memory density of SRAM, the area requirements to store a full 256 or even 128 bin histogram for each pixel is impractical. For example, with a 6T-SRAM cell size of 4.65 μm^2, 256 8-bit bins for every pixel in a 144×252 pixel array would require a die area of 345 mm^2, without any overhead for memory interfaces, sense amplifiers and digital logic.

Fortunately, although the TDC has a large number of bins, the majority of data in the TPSF for many applications is confined within in a narrow range. Thus, with a readout scheme which stores the photons within this narrow range in a compressed histogram format prior to transmission off-chip, a large compression factor, F_c can be achieved. The compression factor is defined as the equivalent raw data size divided by the compressed data size. Thus, a high compression factor can result in a large increase in photon throughput if the light

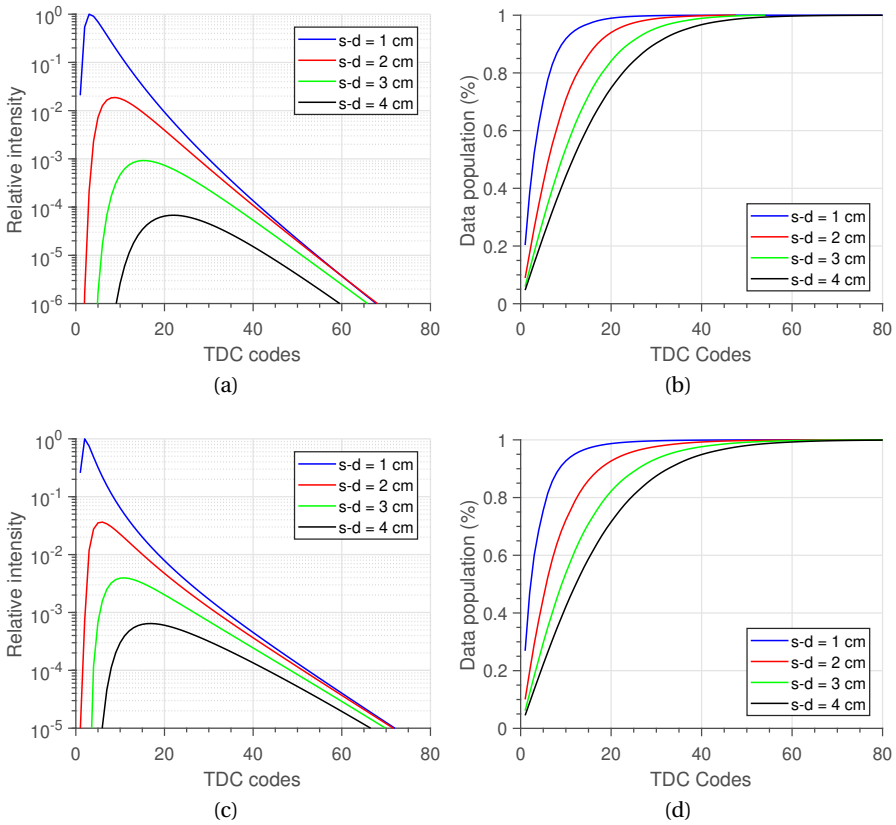


Figure 5.3 – TPSF of homogenous medium from NIRFAST forward simulation at varying source-detector (s-d) separations, a) $\mu_a = 0.01$ mm^{-1}, $\mu_s' = 1$ mm^{-1} and c) $\mu_a = 0.005$ mm^{-1}, $\mu_s' = 0.5$ mm^{-1}. Cumulative distribution function of the TPSFs at varying s-d separations, showing photons concentrated in a narrower range of bins as s-d separation decreases, b) $\mu_a = 0.01$ mm^{-1}, $\mu_s' = 1$ mm^{-1} and d) $\mu_a = 0.005$ mm^{-1}, $\mu_s' = 0.5$ mm^{-1}.

intensity is sufficient. Figure 5.3a shows simulated TPSFs at source-detector separations of 1-4 cm in simulations from NIRFAST [18] with $\mu_a = 0.01$ mm^{-1} and $\mu_s' = 1$ mm^{-1}, typical for an adult human brain. The TPSFs have been normalized such that each TPSF contains 10^6 photons such that the response shapes can be compared. Figure 5.3b displays the cumulative distribution function of these TPSFs. From both figures, it is clear that the majority of photons are contained within a narrow range, and also that this range is reduced as the source-detector spacing decreases. This is due to the reduced number of scattering events the photons have undergone to reach the detector. At 1 cm, 97.6% of the data in the TPSF is contained within 16 50 ps TDC bins. Although, this percentage decreases to 65% at 4 cm spacing, this increase in FWHM is not critical for the NIROT application due the reduction in light intensity. As seen in Chapter 1, the light intensity in a reflectance mode measurement decreases with an exponential dependence with increasing source-detector spacing. At 4 cm s-d spacing, the incident light flux is sufficiently reduced that a lower compression efficiency can be tolerated. Therefore, with this method, a high compression factor is achieved precisely where it is needed. This conclusion is confirmed also for, $\mu_a = 0.005$ mm^{-1} and $\mu_s' = 0.5$ mm^{-1}, which are typical bulk optical properties of the neonate brain, see Figures 5.3d and 5.3d. A system based on this compression system should thus be suitable for a range of NIROT measurements.

The readout implementing this scheme is pictured in Figure 5.4, which is referred to as the PHR. Four half-columns, or 504 pixels, share a PHR circuit which employs separate 5 kb and 40 kb SRAM blocks. The histogramming readout is split into two processes, peak-detection and partial-histogramming. For peak-detection, the 40 kb SRAM is configured as 8 bins of 10 bits per pixel. The peak detection, Figure 5.4, is a 3-step process, where the peak among 8 bins is successively located in the ranges Q<9:7>, Q<6:4> and Q<3:1>. This method has the benefit that if there exists a second smaller peak in the timing response, e.g. due to a reflection, the larger peak is selected. Once the peak of the histogram is located, it can be read out directly in peak-detection mode or stored on a per-pixel basis in the 5 kb SRAM. With the peak location known, the PHR can be operated in partial histogramming mode, whereby a configurable 16 bin window is formed around the peak, where each bin is 5 bits. Photons which arrive at the input to the PHR and lie within this range are then stored in the 40 kb SRAM and periodically read out before the SRAM overflows.

Since depth information in the TPSF in the NIROT application is encoded in late arriving photons, valuable information from the TPSF can exist outside of the 800 ps window defined by the PHR. If these photons are not read out from the sensor, the information from the greatest depths would be lost. For this reason, the readout also has the option to stream out

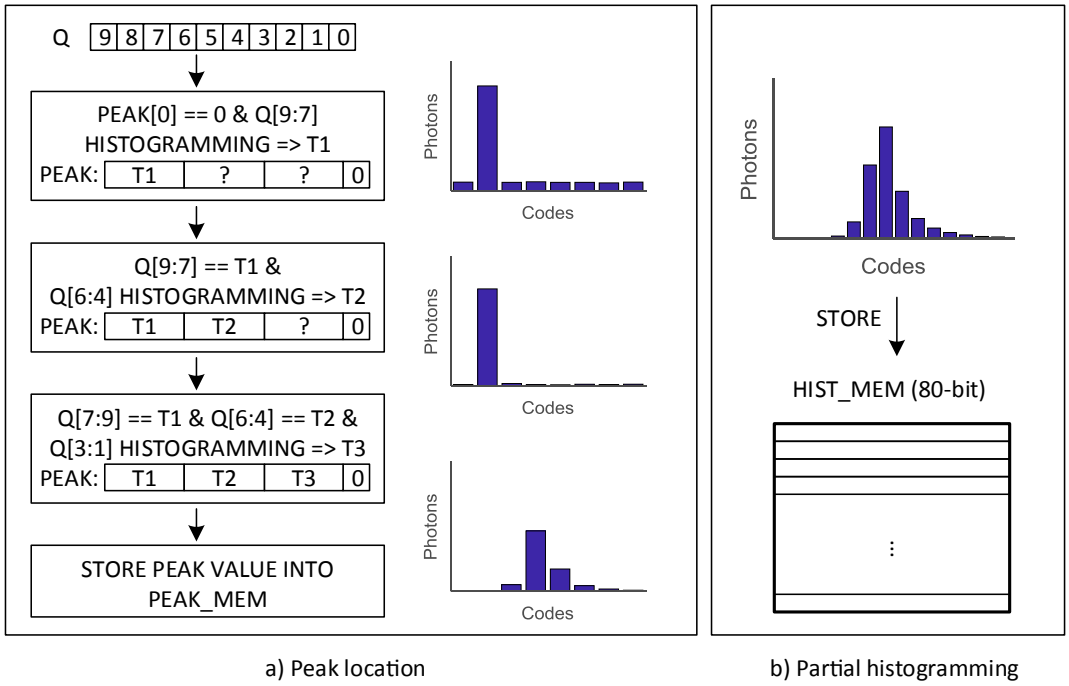


Figure 5.4 – Ocelot partial histogramming readout block diagram. (a) Peak detection requires three stage successive estimation of histogram peak. (b) Partial histogramming stores a 16 bin configurable window of bins around the histogram peak.

the low bandwidth data outside of this 800 ps window via the I/O pads whilst the PHR is accumulating events. Thus, the entire TPSF is captured.

The readout is capable of operating in a low power 60 MHz and fast 240 MHz modes. As well as the peak-detection and partial histogramming modes, the sensor can also operate in TCSPC and single-photon counting modes as was implemented in the Piccolo sensor. In TCSPC, the pixel address, TDC code and TDC identifier bits are serialized and read out via a GPIO pad, where one GPIO is shared with 4 half-columns. In single-photon counting mode, the TDC code and identifier are not transmitted and the photon throughput is increased.

5.5 Dual-clock RO TDC

The inclusion of the PHR block in the sensor places an additional constraint on the TDC design in comparison to the Piccolo sensor. This is due to the fact that, per pixel, the single compressed histogram from the PHR is formed from 6 histograms, one for each TDC. Thus, for a given time-of-arrival, the deviation between the codes from the 6 TDCs must be minimised to ensure the peaks of all 6 TDCs are not dispersed widely in terms of TDC codes. This

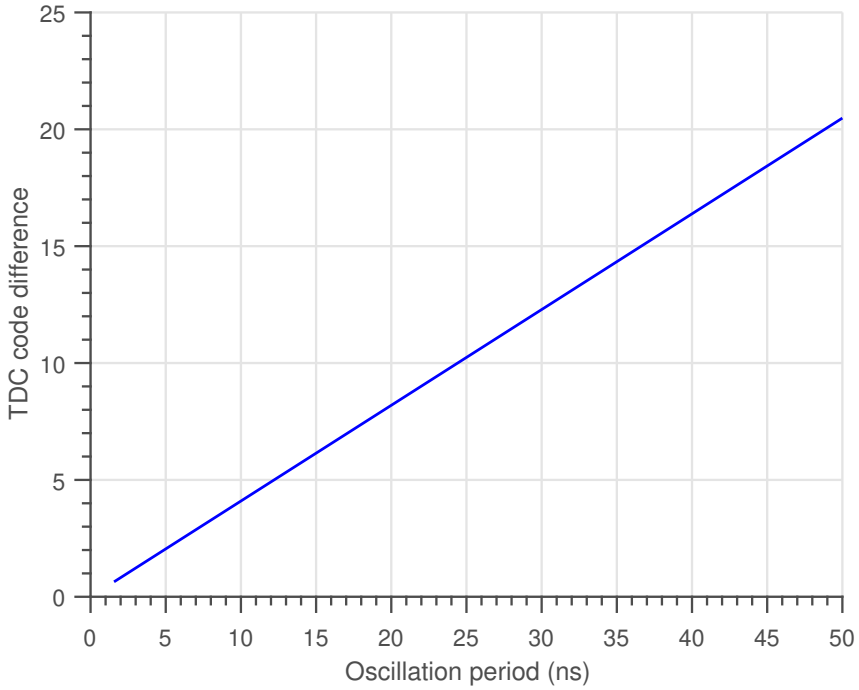


Figure 5.5 – Code difference accumulated between two TDCs, with a fixed measurement period. The TDCs have open-loop oscillating frequencies of 0.99 × 2.56 GHz and 1.01 × 2.56 GHz.

requirement appears to be incompatible with the TDC architecture employed in Piccolo, where the open-loop ROs accumulate code differences in relation to each other as a result of LSB variation. For example, consider two TDCs each based on an open loop oscillator with nominal frequencies of 0.99 × 2.56 GHz and 1.01 × 2.56 GHz, that is, 2.56 GHz ± 1%. Over an oscillation period of 50 nanoseconds, the measured code varies between the two TDCs by over 20-bits, see Figure 5.5. This plot also demonstrates that even for an 80 MHz STOP frequency, as could be applied in the NIROT application, the code difference is over 5 LSBs. Therefore, if a RO is to be used in combination with the PHR block it should operate for less than the laser period.

Despite this accumulated difference in codes between neighboring TDCs, the RO architecture is still preferred due to the large power consumption required to have 1728 'always-on' TDCs [95, 93, 105]. As such, a dual-clock RO-based TDC is developed which improves on the work in [78]. This architecture employs a second clock, STOP_HF, which allows the maximum on-time of the RO to be reduced, thus decreasing the accumulated code error between TDCs in the same column.

A circuit schematic of the dual-clock TDC and timing diagram are shown in Figures 5.6 and

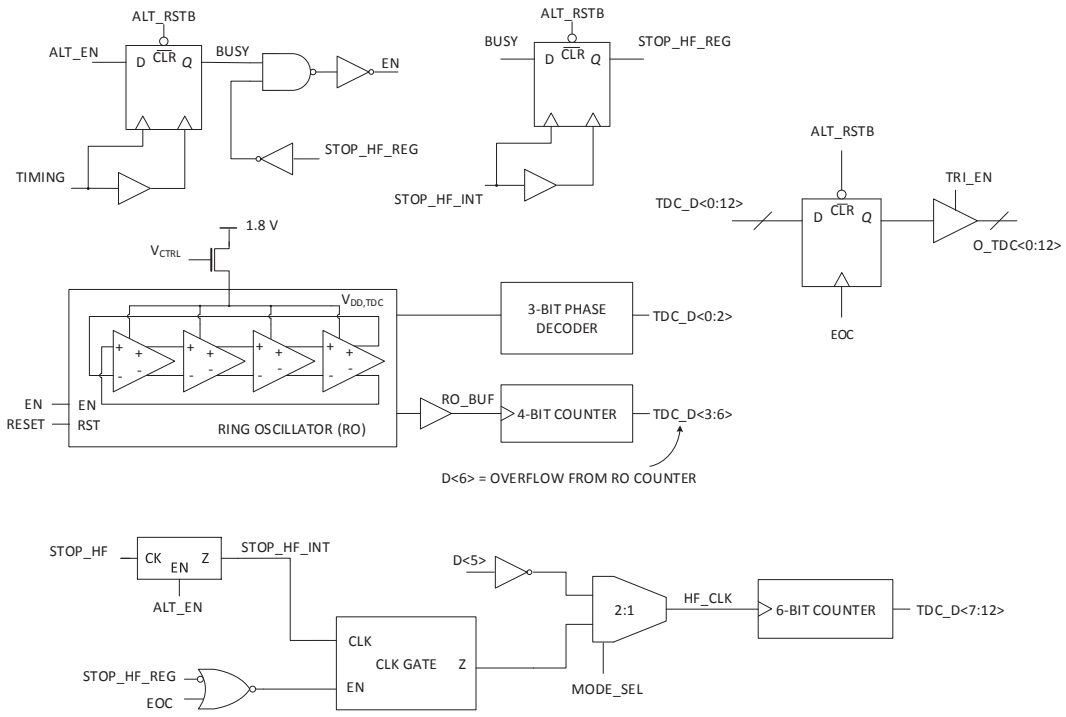


Figure 5.6 – Dual clock TDC schematic.

5.7, respectively. The TDC is enabled by the assertion of the ALT_EN signal from the previous slice in the ALTDC chain, T1 in Figure 5.7. This also gates the STOP_HF clock 'on' allowing the nodes in the TDC to be toggled by STOP_HF_INT. On the next rising edge of timing, T2 in Figure 5.7, the EN signal goes high enabling the RO to run at a nominal frequency of 2.56 GHz. As in the Piccolo architecture, an NMOS source follower is connected between the power supply and the RO to reduce the effect of IR drops on the RO frequency. The rising edges of the RO are counted by a 4-bit counter until the first rising edge of STOP_HF_INT, T3 in Figure 5.7. This asserts the STOP_HF_REG signal, drives the enable signal to low and the TDC begins counting edges of the STOP_HF_INT signal with a 6-bit counter. On the next rising edge of STOP, T4 in Figure 5.7, the EOC signal is asserted and the 6-bit counter stops counting. The complete TDC code is then obtained from the frozen phase of the RO (3-bits), the rising edges of the RO counter (4-bit), and the rising edges of STOP_HF_INT (6-bits). Although this would appear to indicate 13-bits of TDC resolution, the most significant bit of the RO counter is only intended as an 'overflow' bit. Due to mismatches in the frequency of the RO, there will be TDCs where the RO runs faster than 2.56 GHz. In these instances, the count will exceed 2^3 in a single period of STOP_HF_INT. Therefore, an extra bit is required to capture the minority of codes that exceed this.

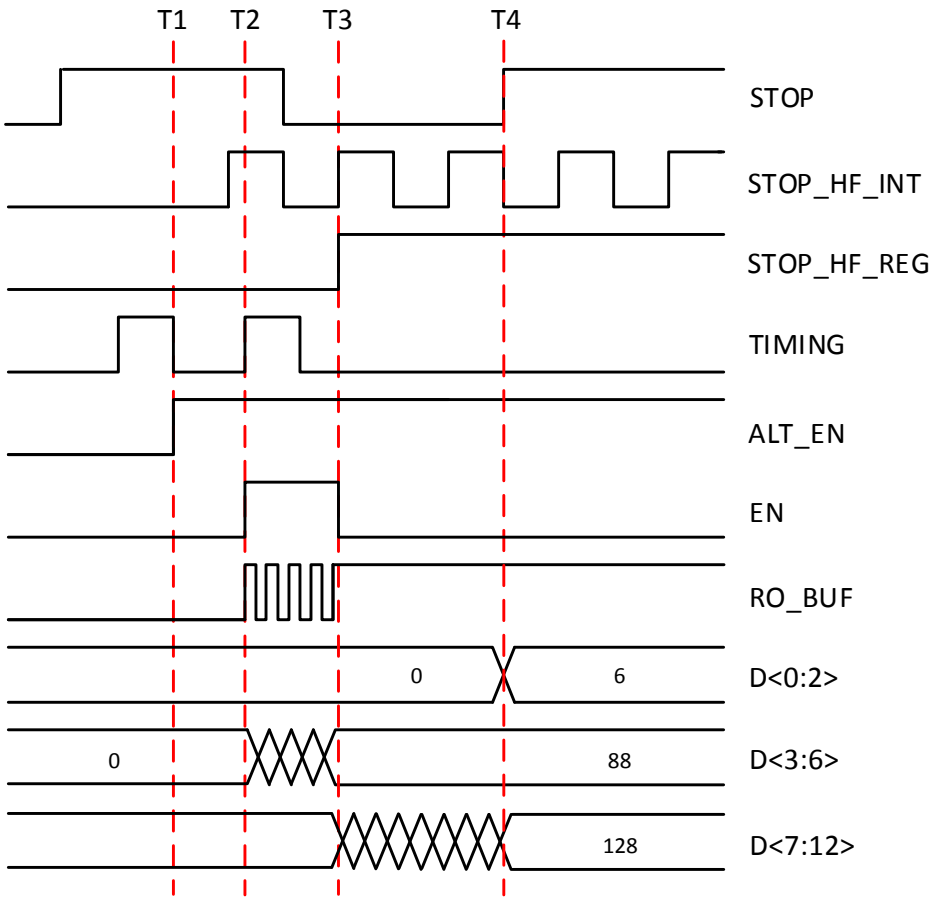


Figure 5.7 – Dual clock TDC timing diagram. Falling edge of timing from previous photon asserts ALT_EN and enables clock gate for STOP_HF_INT, TDC is active for detection (T1). Rising edge of timing enables high frequency RO at a nominal frequency of 2.56 GHz, rising edges of a buffered output from a single phase, RO_BUF, are counted by a 4-bit counter (T2). First rising edge of STOP_HF_INT asserts STOP_HF_REG and freezes state of RO, 6-bit counter begins counting rising edges of STOP_HF_INT (T3). Rising edge of STOP ends conversion by asserting EOC (not shown). Final TDC code obtained from the frozen phase of RO, D<0:2>, 4-bit RO counter, D<3:6>, and 6-bit STOP_HF_INT counter, D<7:12> (T4).

To determine the frequency of STOP_HF, it is important to consider the other half of the trade-off. As seen in Figure 5.5, increasing the frequency of STOP_HF reduces the accumulated error between ROs in the same column, however, this is achieved at the expense of increased power consumption. For a comparison, consider first the single-clock ring-oscillator from Piccolo, Section 4.1.4, adapted for the Ocelot architecture, e.g. larger clock tree and column, The power consumption, P_{sc}, is the sum of the static, $P_{sc,static}$ and dynamic, $P_{sc,dynamic}$ components. Static power consumption is independent of TDC activity, and is due to the power dissipated in distributing the STOP signal to the TDCs. From post-layout simulation, it is possible to

extract design parameters, $P_{n,tree}$ and $P_{n,col}$, which are the respective power consumption of the clock tree and column distribution networks for the STOP clock, normalised to a frequency of 1 Hz. The static power consumption for the single clock case is given by,

$$P_{sc,static} = P_{n,tree} \cdot \mathrm{STOP} + P_{n,col} \cdot \mathrm{STOP}. \tag{5.1}$$

The dynamic component of power consumption is dependent on the number of TDCs which are active per cycle, due to the average power consumption of the RO, $P_{av,ro}$ and the frequency normalised power consumption of the counter, $P_{n,cnt}$. It is given by,

$$P_{sc,dynamic} = \mathrm{N}_{cyc} \cdot \left(P_{av,ro} \cdot 0.5 + \left(\frac{2^{n_1} - 1}{2^{n_1}} \right) \cdot P_{n,cnt} \cdot \mathrm{f}_{osc} \right), \tag{5.2}$$

where

$$n_1 = \log_2 \left(\frac{\mathrm{f}_{osc}}{\mathrm{STOP}} \right).$$

For the dual-clock power consumption, P_{dc}, the static component, $P_{dc,static}$, now includes a clock distribution network for STOP_HF, and is given by

$$P_{dc,static} = P_{n,tree} \cdot (\mathrm{STOP} + \mathrm{STOP\_HF}) + P_{n,col} \cdot (\mathrm{STOP} + \mathrm{STOP\_HF}). \tag{5.3}$$

The dynamic component, $P_{dc,dynamic}$ is then given by

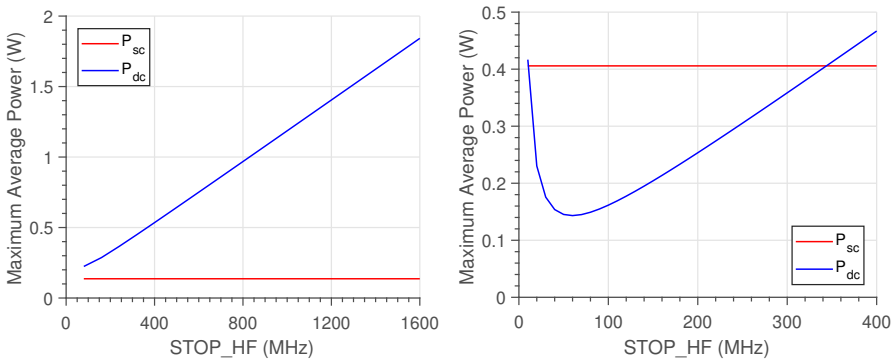
$$\begin{aligned} P_{dc,dynamic} = \mathrm{N}_{cyc} \cdot \Bigg(P_{av,ro} \cdot 0.5 \cdot \frac{\mathrm{STOP}}{\mathrm{STOP\_HF}} + \left(\frac{2^{n_2} - 1}{2^{n_2}} \right) \cdot P_{n,cnt} \cdot \mathrm{f}_{osc} \cdot \frac{\mathrm{STOP}}{\mathrm{STOP\_HF}} \\ + \left(\frac{2^{n_3} - 1}{2^{n_3}} \right) \cdot P_{n,cnt} \cdot \mathrm{STOP\_HF} \Bigg) \end{aligned} \tag{5.4}$$

where

$$n_2 = \log_2 \left(\frac{\mathrm{f}_{osc}}{\mathrm{STOP\_HF}} \right), \text{ and } n_3 = \log_2 \left(\frac{\mathrm{STOP\_HF}}{\mathrm{STOP}} \right).$$

Thus, the dynamic component has been reduced in comparison to the single-clock case as the on-time of the RO is reduced, and part of the conversion now involves counting rising edges of the STOP_HF clock, rather than f_{osc} for the complete cycle.

To calculate the maximum average power consumption, the only remaining unknown is the



(a) (b)

Figure 5.8 – Maximum average power consumption in the dual clock TDC for varying STOP_HF frequencies. a) $N_{cyc} = 38$ at STOP = 80 MHz, b) $N_{cyc} = 304$ at STOP = 10 MHz.

maximum activity per cycle, which is given by

$$N_{cyc} = \frac{B_{io} \cdot N_{pads} \cdot F_c}{L_C}, \tag{5.5}$$

where B_{io} is the bandwidth of a data pad, N_{pads} is the total number of data pads and L_C and F_c are the code length of one photon detection and the data compression factor, respectively. Maximum average power consumption for $STOP = 80$ MHz is shown in Figure 5.8a where the following values apply, $B_{io} = 160$ MHz, $N_{pads} = 72$, $L_C = 26$, $F_c = 6$, $P_{n,tree} = 41.6$ pW/Hz, $P_{n,col} = 1.05$ nW/Hz, $P_{av,ro} = 2.2$ mW, and $P_{n,cnt} = 79$ fW/Hz.

At a frequency of 80 MHz, events occur in relatively small numbers per cycle. With $F_c = 6$, the maximum number of events per cycle is 38. Therefore, the contribution from static power consumption dominates at all STOP_HF frequencies, and the increase in power consumption as the STOP_HF frequency scales is almost completely linear. Furthermore, the increased power consumption of the clock distribution networks due to STOP_HF means that the dual-clock TDC consumes more power than the single-clock structure at all STOP_HF frequencies. For comparison, the maximum average power consumption in the case of STOP = 10 MHz is also plotted, see Figure 5.8b. At this lower STOP frequency, the power consumed by the ROs due to the increased activity rate is much more significant and the maximum average power consumption can be reduced by a factor of 2.8 with a STOP_HF frequency of 60 MHz. Unfortunately, referring back to Figure 5.5, this STOP_HF frequency will result in a significant error accumulation, likely causing difficulty for the PHR. Thus, at lower STOP frequencies and high activity it may be possible if the activity level is high enough to save a significant amount of power with this architecture. Since the focus of this thesis is performing fast time-resolved

measurements with a short laser period, e.g. 80 MHz, STOP_HF is chosen to equal 320 MHz. This frequency is seen as an appropriate balance between restricting the accumulated error due to LSB mismatch and an increase in static power consumption due to the distribution of the STOP_HF clock.

5.6 Clock Generation

Although the dual-clock TDC has many advantages, the added complexity in comparison to the single-clock approach used in Piccolo is that there is a requirement for an additional clock, STOP_HF. Since this clock must maintain a constant phase relationship with the STOP signal from the laser, it should be generated by a frequency synthesis circuit such as a PLL. The main requirement for such a PLL is that the jitter contribution to STOP_HF should be minimal, as it adds to the jitter from the STOP clock in Equation 4.2. The jitter of the system when the dual-clock architecture is employed is given by,

$$\sigma_{SYS} = \sqrt{\sigma_q^2 + \sigma_{START}^2 + \sigma_{STOP}^2 + \sigma_{STOP\_HF}^2 + \sigma_{TDC}^2}. \tag{5.6}$$

A further requirement for an on-chip PLL is that the extra power consumption (in comparison to the Piccolo sensor) from the PHR implies that there could be significant heating of the die during sensor operation. As such, it is desirable to compensate the TDC RO frequency, f_{osc}, for PVT variations. The established method for PVT compensation is to embed a replica of the TDC RO into a PLL and distribute the control voltage of the PLL to the TDCs.

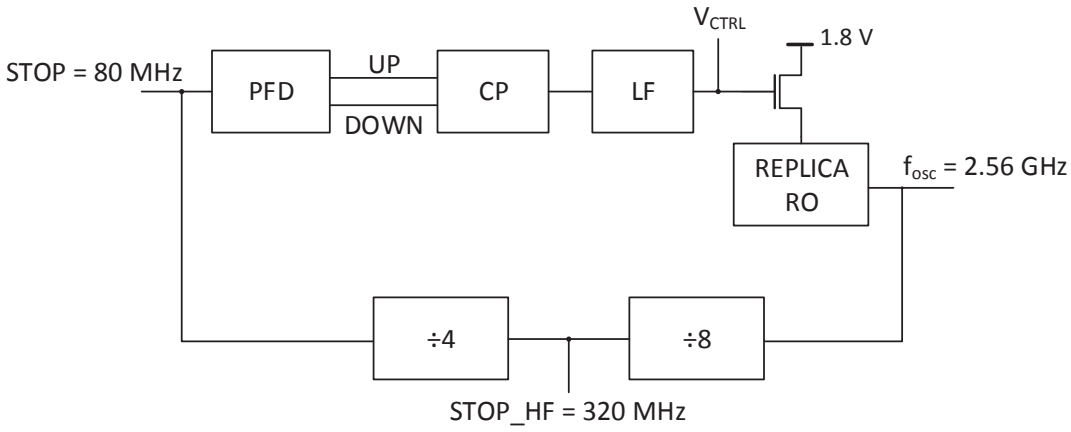


Figure 5.9 – Ideal clock generation for Ocelot sensor. PLL generates the highest frequency, f_{osc} and derives STOP_HF via a frequency divider. The control voltage to the replica RO is distributed to the array of TDCs on the sensor.

In the ideal case, both of these frequencies, f_{osc} and STOP_HF, would be generated by a single PLL, Figure 5.9. To differentiate the RO embedded in the TDC from that in the PLL, the latter will be referred to as a voltage-controlled oscillator (VCO). The VCO control voltage, V_{CTRL}, is set by adding or subtracting charge from a charge pump (CP) onto a loop filter (LF). The amount of charge is determined by the difference in time the STOP signal and the feedback signal spend in the logic high state, as determined by a phase-frequency detector (PFD). When the feedback signal spends longer in the high state, indicating a lower frequency, the UP pulse output from the PFD is longer than the DOWN output, which causes the CP to add charge to the LF, increasing the frequency of the VCO. When the frequency and phase of the feedback signal are locked to that of the reference, the UP and DOWN pulses are almost identical. STOP_HF and f_{osc} are thus synthesized by placing frequency dividers in the feedback path of the PLL, between the VCO output and the input to the PFD. Extensive coverage of PLL theory can be found in [121]. Frequencies of f_{osc} = STOP × 4 × 8 and STOP_HF = STOP × 4 are generated according to the PLL block diagram in Figure 5.9.

The main disadvantage of the PLL in Figure 5.9 is that these divider ratios are compatible only with a STOP frequency of 80 MHz. For example, if STOP = 40 MHz, f_{osc} would attempt to lock at 1.28 GHz. Even if the VCO could operate at this lower frequency, it implies a doubling of the TDC LSB. Whilst a STOP frequency of 80 MHz is generally reasonable for NIROT, there are many applications where an 80 MHz laser is not suitable or does not exist for the other performance requirements, e.g. size and power. At first glance, it appears that there is an easy solution to this problem, simply place a configurable frequency divider in between the STOP signal and the PFD input. The frequency division could then be set according to the operating frequency, where the input to the PFD always receives the lowest frequency in the range, $STOP_{min}$. For example, to be configurable with both 80 MHz and 10 MHz STOP frequencies, the configurable divider should be able to divide by 8, in the case of an 80 MHz clock, and 1, in the case of a 10 MHz clock.

The disadvantage of this technique is due to the way phase noise from the various PLL components contributes to the jitter of the generated clock signals. The main contributors to the PLL phase noise are from the PFD and the VCO, which in the case of Ocelot would be the RO from the TDC. Phase noise from the PFD is low-pass filtered by the PLL whilst phase noise from the VCO is high-pass filtered, both with a cutoff frequency at the PLL loop bandwidth. Thus, the optimum method for loop bandwidth selection is to minimise the combined contribution of phase noise from both the PFD and the VCO. Unfortunately, loop bandwidth design is often not so straightforward, since for stability reasons, the PLL loop bandwidth should be at most

5-10% of the reference frequency, STOP_{min}, of the PLL. This means that reducing STOP_{min} results in an increased jitter accumulation from the VCO.

In the design of Ocelot, a STOP_{min} of 2.5 MHz is assumed to encompass a wide range of possible applications. This implies a conservative loop bandwidth of approximately 125 kHz. With the loop bandwidth established, it is possible to calculate the jitter added by the VCO. With the RO designed for Piccolo, the simulated phase noise is -81.36 dBc/Hz at 1 MHz offset from the carrier frequency, which translates to an rms jitter of 38 ps, or 89.3 ps FWHM. In relation to the system response for Piccolo, the jitter for contributed by the VCO is almost of the same order. The system response FWHM for Ocelot according to Equation 5.6 would then be *at least* 142 ps. Although the VCO phase noise is likely to dominate the contribution from the PFD due to the low PLL bandwidth, there are several other sources of noise in a PLL which contribute to the overall performance, e.g. reference spurs, power supply noise, substrate noise coupling, etc. Furthermore, a disparity between the simulated and measured phase noise [122] could also result in a poorer system timing performance.

One option to reduce the jitter contribution from the VCO would be to redesign the Piccolo RO for improved jitter performance. This would, however, require a significant increase in area and power consumption [122]. This is highly undesirable due to the compact layout already required to fit into a 28.5 μm column, a large increase in the RO size would be impractical to layout in such a space. Additionally, the added power consumption in the TDC array is a high price to pay since improved phase noise performance is not required in the TDC array. In a single period of the STOP_HF clock, the RO already accumulates only 413 fs rms jitter. Thus an improvement would not be noticeable from this perspective, it is only when the RO is embedded into a PLL with a lower bandwidth that the issue becomes apparent.

For this reason, separate PLLs were designed for PVT compensation of the TDC RO and generation of the STOP_HF clock. In the latter case, a 3-stage VCO [122] was designed with a phase noise of -98.61 dBc/Hz at 1 MHz from the carrier frequency. With a PLL bandwidth of 125 KHz, the rms jitter contribution of the 3-stage VCO is 14 ps.

5.7 Results

The Ocelot sensor is also implemented in the same 180nm CMOS technology as the Piccolo sensor, Chapter 4, and occupies an area of $21.6 \times 10.2\text{mm}^2$. A photomicrograph is shown in Figure 5.10. Ocelot is divided into four quadrants, where each quadrant can function almost completely independently of the remaining three. With 6 TDCs per half-column, the size of

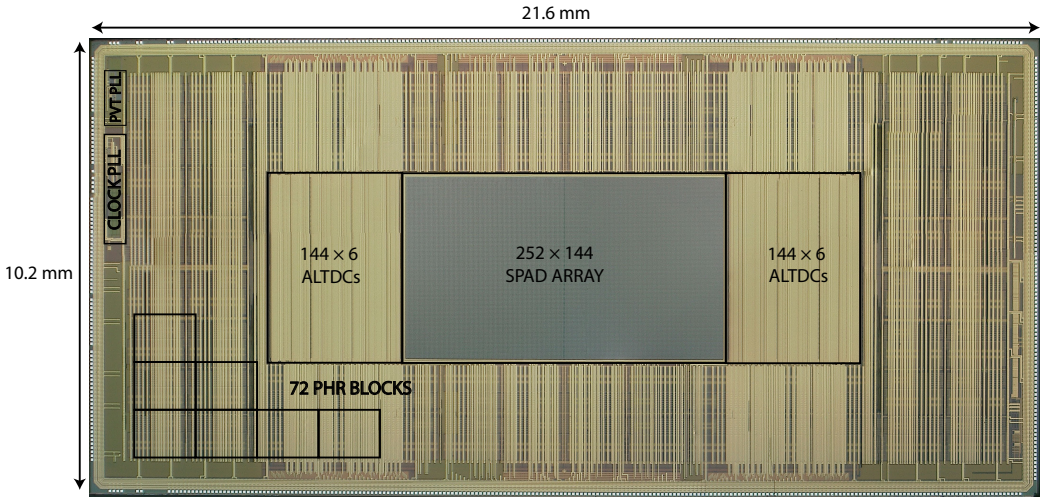
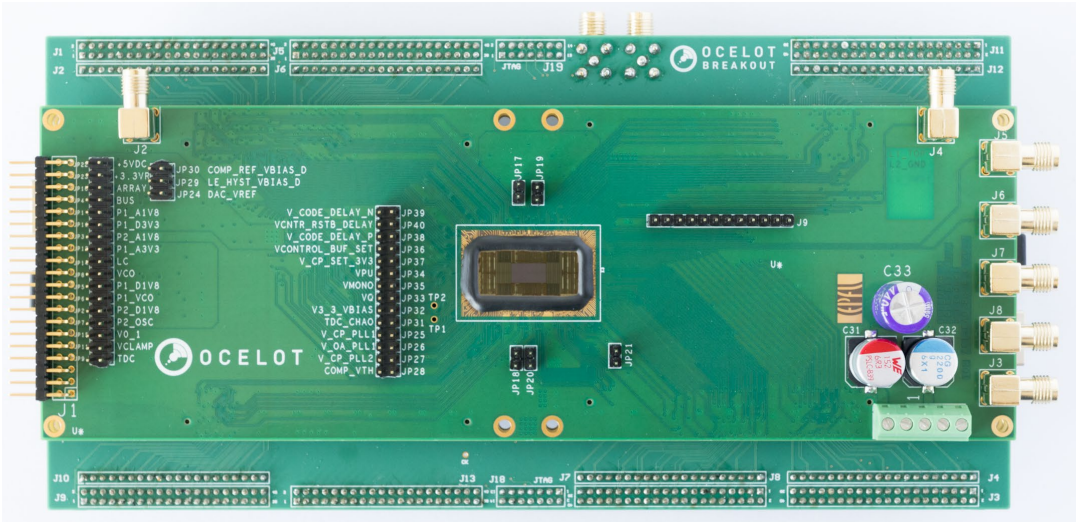


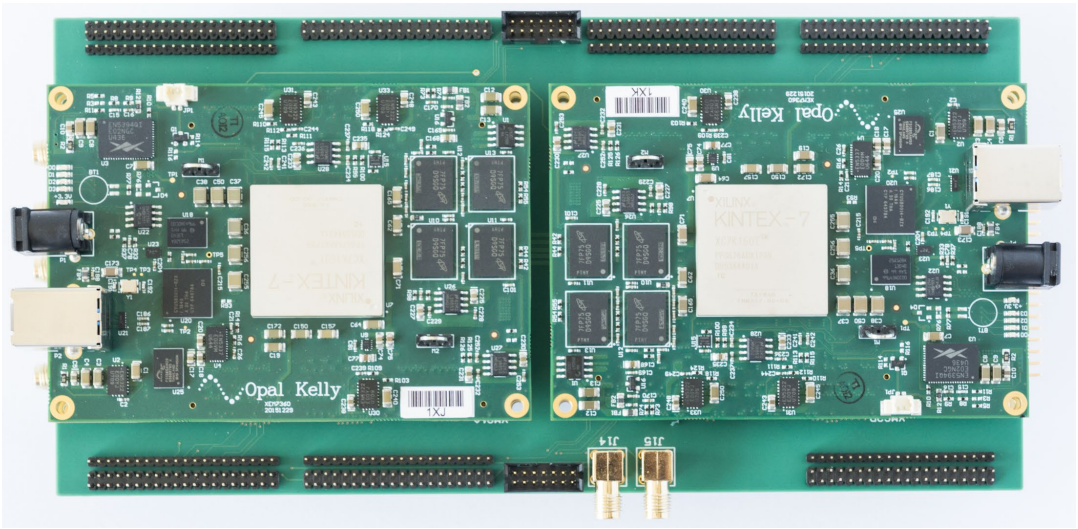
Figure 5.10 – Photomicrograph of Ocelot sensor. ©2018 IEEE.

the pixel array is still larger than the total area of the ALTDC arrays. If the area of the ALTDC is combined with the pixel array, as was calculated for Piccolo, the resulting fill-factor would be 15.8%. In comparison to Piccolo, which had a figure of 9%, the increase in area due to two ALTDC arrays is more than offset by the increase in SPAD array size. Finally, approximately 70% of the core area of the sensor is occupied by the PHR blocks. This is due to the large amount of logic and SRAM required to implement the blocks, and the relatively mature technology. Since the PHR is entirely digital, it would scale very well when manufactured in more advanced technologies.

The Ocelot camera system is shown in Figures 5.11a (front) and 5.11b (back). The complete system comprises of 5 PCBs, a mainboard which includes the sensor and some auxiliary components, a power PCB (not shown), a custom interconnect board, and two integration boards (XEM7360, Opal Kelly, USA) which each include a Kintex-7 FPGA (XC7K160T-1FFG676C, Xilinx, USA) and USB 3.0 interface. The sensor die is glued and bonded directly to the PCB due to the large number of connections between the sensor and PCB. The mainboard contains a number of quad DACs (DAC7554, Texas Instruments) for bias voltage generation, and a fast comparator to convert the laser trigger to a level which is compatible with the 3.3 V I/O pads. A separate power PCB contains a number of voltage regulators (LT3081, Linear Technology) to generate all of the power supplies for the board and sensor. The mainboard is mounted onto an interconnect board which provides some headers for debugging purposes. On the back side of the board, Figure 4.1b, two XEM7360 integration modules mounted onto the reverse of the interconnect board for control and processing of data from Ocelot. Two modules are required



(a) front



(b) back

Figure 5.11 – (a) Front of the Ocelot camera system displaying Ocelot sensor bonded directly to the PCB. The main auxiliary components are included on the same PCB, whilst a separate power PCB containing only voltage regulators supplies the voltages necessary to power the sensor and board. (b) Back of the Ocelot camera system displaying dual XEM7360 integration modules (Opal Kelly, USA), mounted on a custom interconnect board. The system dimensions are 204×100.5 mm^2.

due to their limited number of 3.3 V compatible I/O pads, however, due to the division of the full sensor into quadrants, sensor control and readout are relatively simple.

5.7.1 TDC Nonlinearity

The TDC nonlinearity is measured by illuminating the sensor with uncorrelated light. Providing each TDC receives on average less than 1 photon per cycle then the events should be uniformly distributed in time. As such, by comparing the deviation in the number of photons per bin in relation to the mean number of photons per bin, the DNL can be calculated. This test, which was also performed for the Piccolo TDC, is called a code density test. The DNL is shown at the top of Figure 5.14. Since there are more than 1024 bins in the DNL, some photons

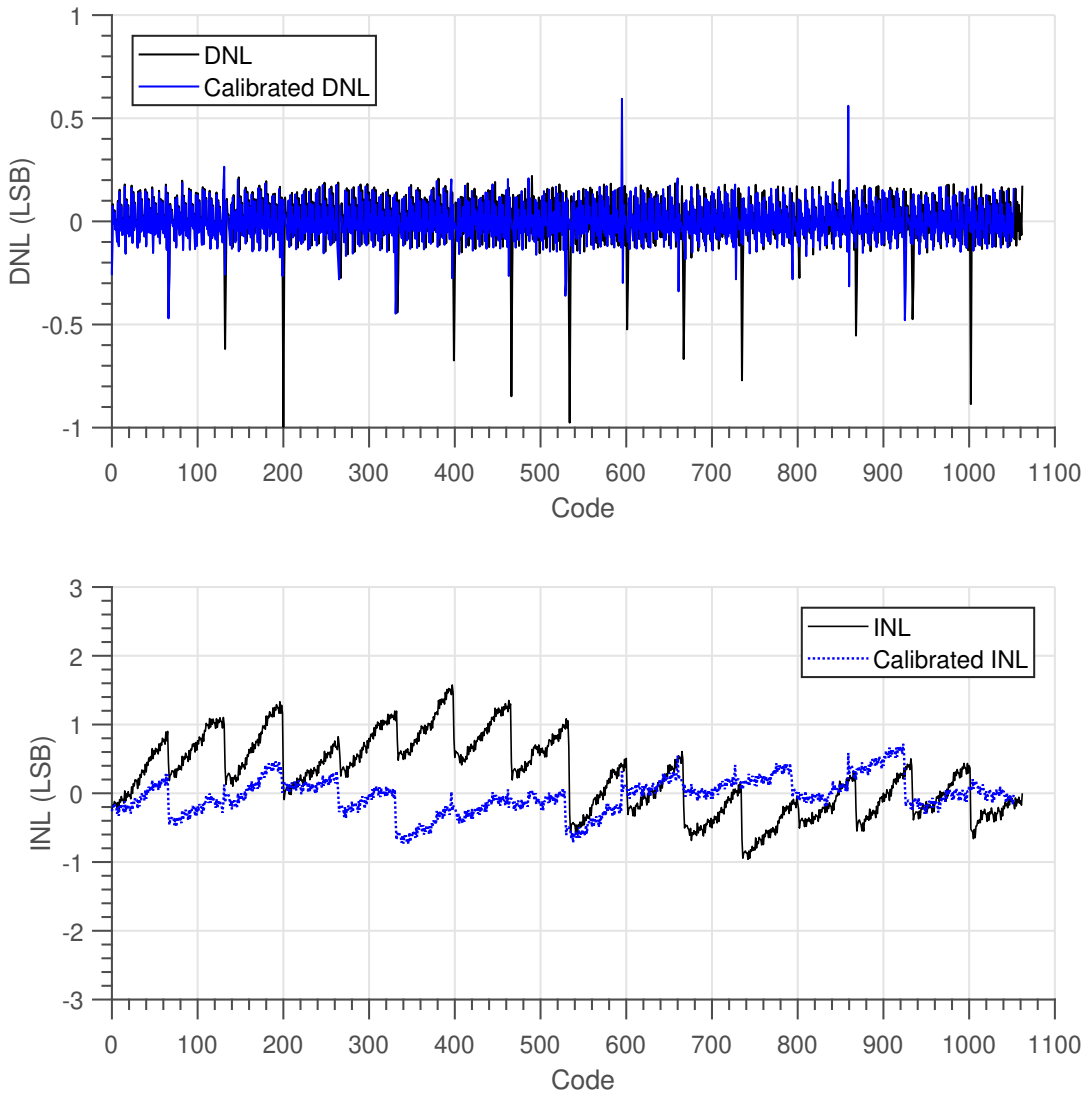


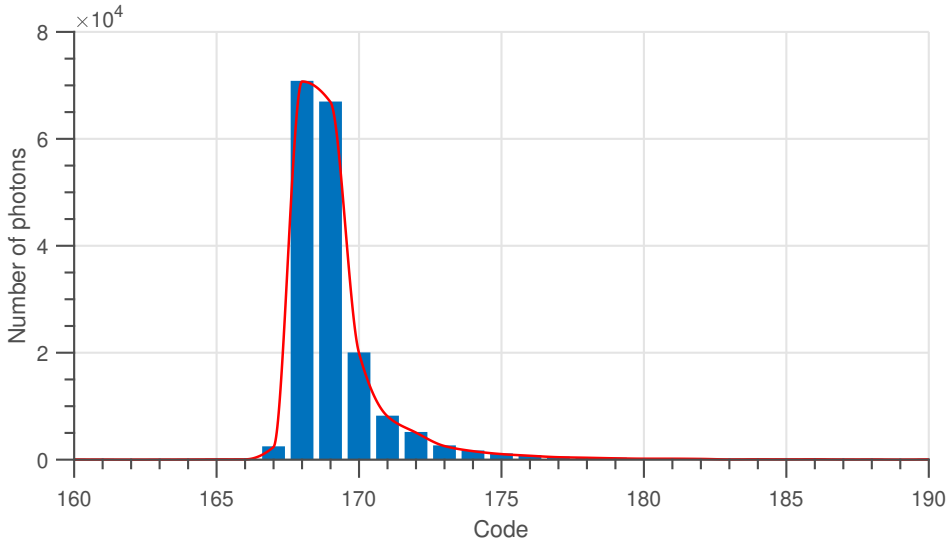
Figure 5.12 – DNL (top) and INL (bottom) of dual-clock TDC. DNL and INL are improved by calibrating for almost empty bins at the transition from the RO to STOP_HF counter. ©2018 IEEE.

have arrived with a code that has a '1' in the most-significant bit of the RO counter. This bit is included in the case that the RO can complete more than 8 cycles during one period of STOP_HF. With only a 3-bit counter, the counter would reset to the minimum value on the 9th cycle and photons arriving over 3 ns apart would land in the same TDC bin. The use of an extra counter bit to capture these photons correctly also means that there are many unused codes, in which no photons will arrive. As such, the TDC response must be reconstructed via a lookup table to place the bins in the correct order and ensure there are no unused codes in the response.

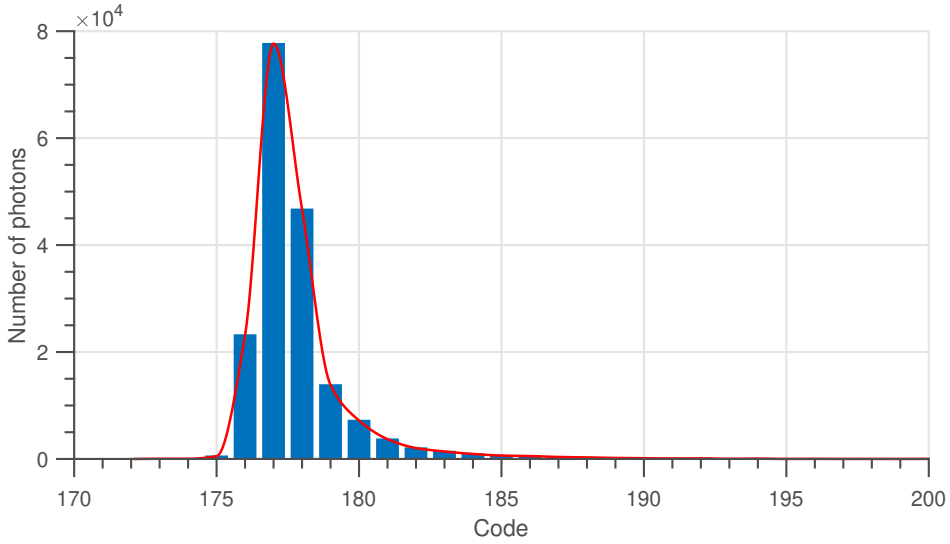
Whilst for the most part, the bins display the same good linearity that was exhibited in the Piccolo sensor, there are now some sharp periodic troughs. This is caused by the transition between the counter for the RO and the counter for STOP_HF. Since the STOP_HF clock exhibits some jitter, the transition will show stochastic variation between cycles. The troughs in the DNL are the TDC bins for which the STOP_HF clock arrived the latest. Thus, they are not due to any design flaw such as inadequate capacitance matching or threshold offsets, rather it is an intrinsic feature of the architecture. Including these troughs into the TDC nonlinearity results in a worst-case DNL of +0.22/-1 LSB and an INL of +2.39/-2.6 LSB. A simple calibration is performed which redistributes photons from the trough regions at 1/2 the median count in the TDC histogram into the closest earlier bin. This calibrated worst case DNL is then +0.6/-0.48 LSB and INL is +0.89/-1.67 LSB.

5.7.2 Timing Response

The timing response of the system is measured by providing the electrical STOP signal from the laser via a fast comparator for minimal jitter. The Ocelot sensor is illuminated by light from a supercontinuum laser (SuperK Extreme, NkT Photonics), where an AOTF narrows the spectrum of the transmitted light to a narrow band around 700 nm. The light intensity is limited such that the measured pixel detects a photon on less than 1% of laser cycles to limit the effects of pile-up. The timing response is measured with the TDC in single-clock mode, Figure 5.13a), where the RO oscillates until the rising edge of STOP, and dual-clock mode, Figure 5.13b), where the RO oscillates until the rising edge of STOP_HF. The system timing responses for the single-clock and dual-clock modes are 110.2 and 98.1 ps, respectively. The decrease in the dual-clock FWHM in comparison to the single-clock can be explained by the position of the peak in the single-clock mode, which appears to land directly between two bins. When the response is interpolated to determine the FWHM, it is wider in comparison to the dual-clock result. Nevertheless, these results validate the design of a dedicated PLL for the



(a) Single-clock mode



(b) Dual-clock mode

Figure 5.13 – Ocelot timing response for one pixel and one TDC in TCSPC mode at 700 nm. a) single-clock TDC mode, 110.2 ps FWHM and b) dual-clock TDC mode, 98.1 ps FWHM.

generation of STOP_HF as the timing performance of the sensor in the dual-clock case has not been adversely affected.

5.7.3 Distance Linearity

The linearity of the Ocelot sensor at ranges of up to 50 m is shown in Figure 5.14. The linearity was measured in dark conditions, employing a 6 × 128 subset of the array, the data from which

is combined to form a single histogram. The light source is a 2 mW 637 nm laser operating at a frequency of 40 MHz. At this laser frequency, the unambiguous range of the system is 3.75 m. As such, a measurement range of up to 50 m is achieved by exploiting prior knowledge of the scene. Despite this requirement, there is nothing in principle which prevents the system being applied to measurements at this distance. The minimum STOP frequency of the Ocelot sensor is 2.5 MHz, which translates to an unambiguous range of 60 m. The linearity measurement up to 50 m at 40 MHz is useful for two reasons. Firstly, it validates the system for use in a direct time-of-flight (ToF) LiDAR system. Secondly, it tests the system over a wide portion of its TDC range, since it must be traversed multiple times to reach 50 m. The nonlinearity results in Figure 5.14 (top) show a maximum nonlinearity of 8.8 cm over the entire 50 m range, results which are in line with the state-of-the-art in SPAD based LiDAR sensors [101]. A worst case precision of 1.4 mm is achieved with approximately 30k photons per histogram.

5.7.4 Flash Image

Time-resolved measurements of the entire sensor operated in flash mode are shown in Figure 5.15. The imaged mannequin is illuminated by a 2 mW 637 nm laser operating at 40 MHz. Due to a limited illumination angle, the measurement was performed in a sequence of 8 exposures, illuminating different sections of the mannequin. The image was acquired in partial histogramming mode, validating the operation of all four quadrants of the array, timing circuitry, peak detection and readout circuitries. Calibration is performed to correct for the insertion of buffer delays between different sections in the half column according to the bus repeater scheme. The reconstructed image is resolved with millimetric detail, demonstrating the utility of the system in precise measurements. It should be noted that it is simpler to acquire and reconstruct ToF data for LiDAR in comparison to NIROT. In particular, in NIROT the time-resolved data is typically corrected for DNL [50] and information from the entire TPSF is exploited. This is in contrast to LiDAR, where the distance can be computed by calculating the mean bin of the acquired TPSF. This explains why in a flash LiDAR measurement, data from the PHR can be used to reconstruct millimetric detailed images. For each pixel, the detected photons have been measured by 6 separate TDCs, each with its own DNL and showing some variation in the LSB. Thus, the DNL is small enough in Ocelot to avoid overly biasing the calculation of the mean. Furthermore, the mean is unaffected by the LSB variations of the 6 TDCs. To make accurate reconstructions in the NIROT measurement whilst employing the PHR readout will require a thorough correction of the data.

Despite the added complexity of employing the PHR in measurements, the results are very

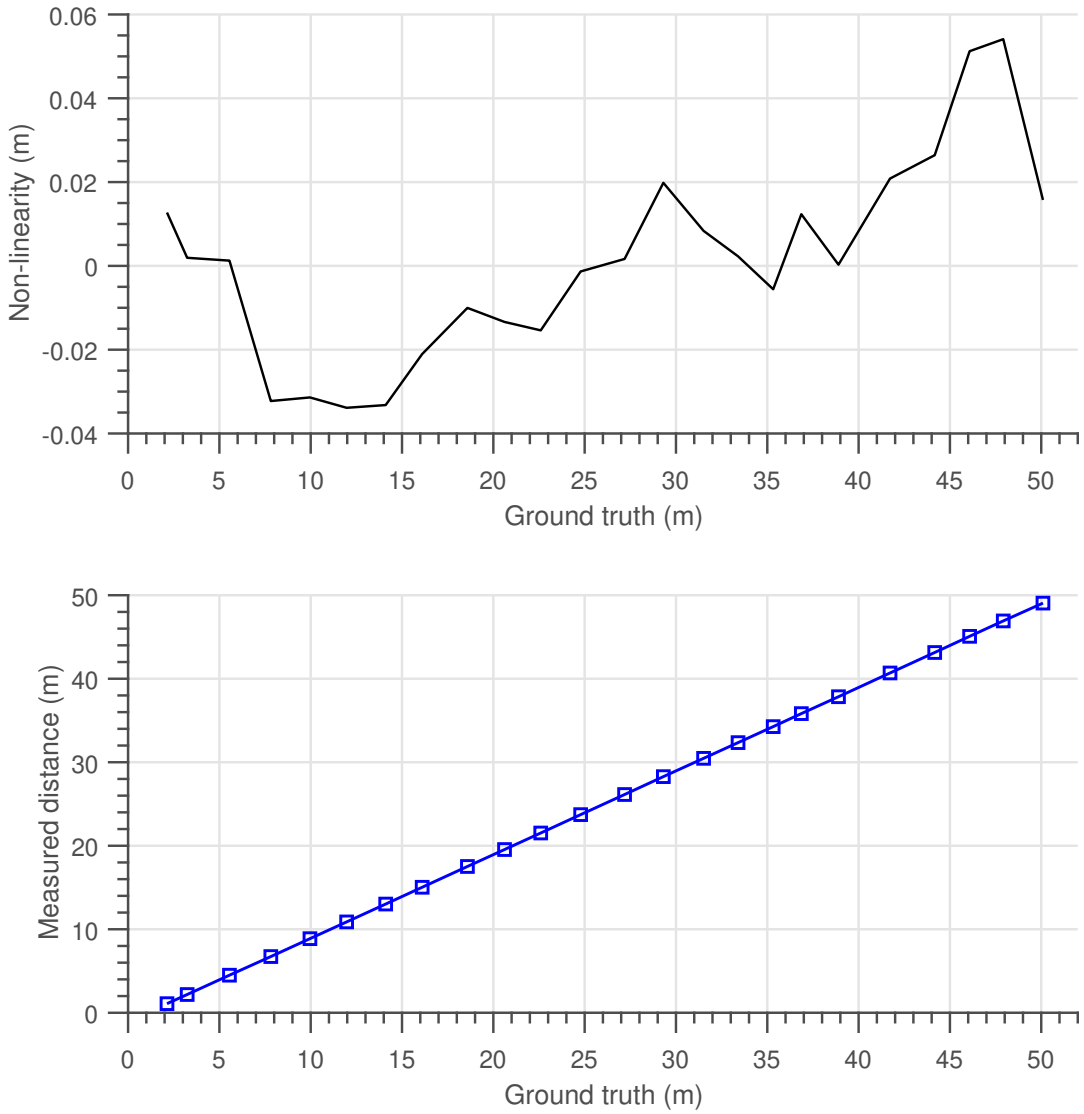


Figure 5.14 – Ranging measurement with non-linearity (top) and the measured distance (bottom) measurements at distances up to 50 m. ©2018 IEEE.

encouraging. The compression factor achieved in this measurement is 14.9, thus a large reduction in off-chip data bandwidth, and thus power consumption in the I/O pads, is achieved. Since the measurement is performed with the slow readout mode, e.g. 60 MHz readout clock, a significant increase in acquisition speed is not yet observed. However, the large compression factor and possibility to operate with a 240 MHz readout clock raises the prospect of high speed time-resolved measurements for the full array.

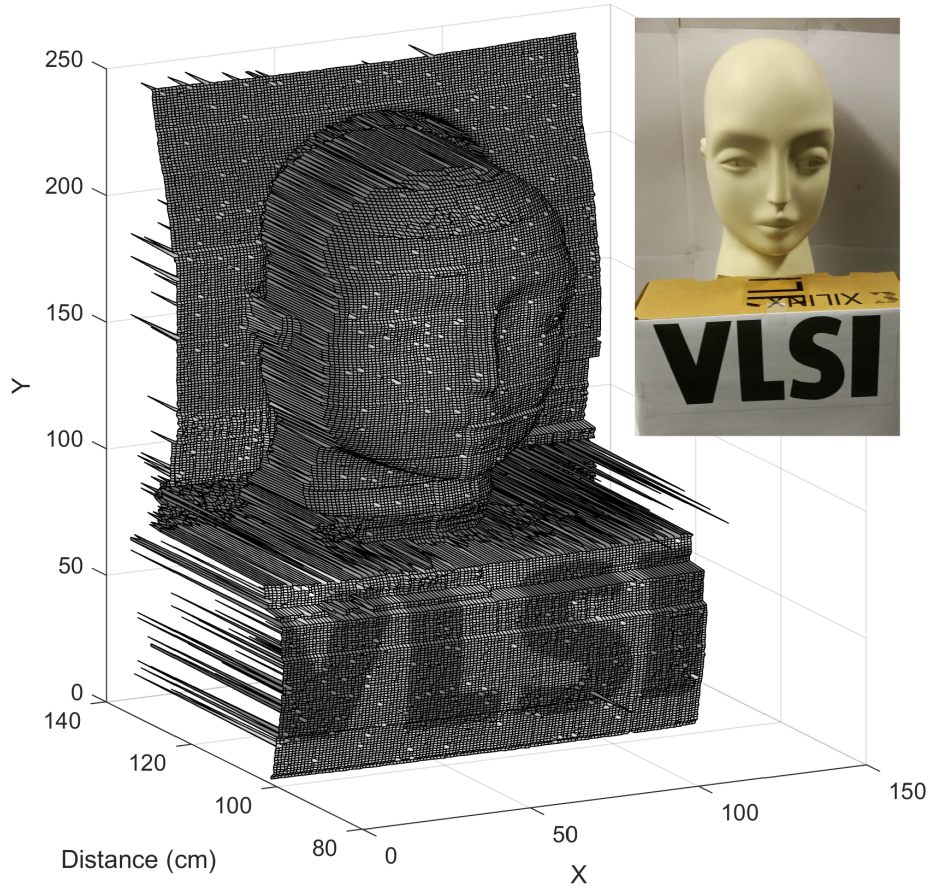


Figure 5.15 – Flash image of mannequin at 1 m distance with 2 mW laser. Image captured in 8 separate exposures due to limited illumination angle.©2018 IEEE.

5.7.5 Power Consumption

The power consumption of the Ocelot sensor was measured during a LiDAR ranging measurement with a global throughput of 156 Mevents/s. At this activity the total power consumption, P_{total} = 2538 mW, with the breakdown between the individual circuit blocks shown in Table 5.1. In comparison to the Piccolo sensor, there is a large increase in the power consumption of the ALTDC array and core circuitry. For the ALTDC array, the TDCs, address latches, and VCOs contribute 63%, 35% and 2%, respectively. Thus, the major increase in power is due to the larger clock distribution networks and the use of higher frequency clocks for the TDC and readout clock. The dynamic component of the TDC is very small, indicating that this component could comfortably scale with greater activity without a major increase in P_{total}. The increase in power consumption for the core is due to the PHR, which has significant digital logic and runs at a 240 MHz clock. Also, it should be noted that whilst there is an increase in

Table 5.1 – Ocelot power consumption

	Power (mW)	Contribution (%)
ALTDC array	838	33
Core	1252	49.3
I/O	198	7.8
Pixel array	74	2.9
PLLs	176	6.9
P_{total}(mW)	**2538**	

the core power consumption, the I/O power consumption has been limited due to the large compression factor achieved.

5.7.6 State-of-the-art Comparison

A comparison of Ocelot against the state-of-the-art in time-resolved SPAD sensor arrays is shown in Table 5.2. A large increase in array size has been achieved against the state-of-the-art, with Ocelot employing almost 80% more pixels than the nearest equivalent competitor [59]. This is achieved whilst maintaining the unprecedented PDE of Piccolo as well as the narrow FWHM of the timing response. Power consumption has risen, however given the large array size increase in comparison to Piccolo this is to be expected. It should also be noted that Piccolo and Ocelot are designed in a 180 nm process where $V_{DD} = 1.8V$, whereas [95] and [59] are designed in 130 nm CMOS, where the supply voltages are 1.5 and 1.2 V, respectively. Since $P_{total} \propto V_{DD}^2$, the P_{total} of Ocelot would reduce significantly if implemented in these technologies.

5.8 Conclusions

The sensor architecture presented here overcomes one of the major challenges in time-resolved SPAD sensor design, namely, achieving both high resolution and a fast acquisition. The application of the dynamic TDC reallocation architecture from Chapter 3, results in a high 28% fill factor in a 28.5 μm pitch for a full 252 × 144 array, an unprecedented result for a sensor with parallel time-resolved measurements from all pixels via TDCs. The reduced area allocated to time-to-digital conversion enables the implementation of an integrated histogramming readout which exploits the structure of ToF data to achieve a high compression of the measured data with a practical allocation of memory. To the best of the author's knowledge, this is the first implementation of integrated histogramming on a per-pixel basis for a full sensor

Table 5.2 – Ocelot state-of-the-art comparison

	[95]	[93]	[59]	Piccolo	**Ocelot**
Array size	64 × 64	32 × 32	160 × 128	32 × 32	**252 × 144**
Fill factor (%)	0.77	3.14	1	28	**28**
PDP at 800 nm (%)	4	6	6	12	**12**
PDE at 800 nm (%)	0.03	0.19	0.06	3.4	**3.4**
DCR (cps/μm^2) at (V_{EB})	30.7 (2.5 V)	0.14 (6 V)	2 (0.73 V)	0.62 (5 V)	**0.62 (5 V)**
TDC LSB (ps)	62.5	312	55	48.8	**48.8**
DNL (LSB)	-2/+2	-0.06/+0.06	-0.3/+0.3	-0.2/+0.2	**+0.6/-0.48**
INL (LSB)	-4/+4	-0.22/+0.22	-2/+2	-0.47/+0.77	**+0.89/-1.67**
Timing FWHM (ps)	200	609	140	110	**110**
Technology (CMOS)	130 nm	0.35μm	130 nm	180 nm	**180 nm**
Die area (mm × mm)	9 × 4.1	9 × 9	12.3 × 11	5 × 2	**21.6 × 10.2**
Data bandwidth (Gbps)	42	1.024	51.2	5.12	**11.52**
TP_{max} (Gevents/s)	2	0.064	0.102	0.220	**0.443**[1]
P_{total} (W)	8.79	0.43(2.8)[2]	0.55	0.31	**2.54**[3]

[1] In TCSPC mode, higher rates possible with PHR dependent on TPSF width.
[2] Gating functionality, where the P_{total} is given for 1 gate (50 gates) per cycle.
[3] Measurement performed with a global throughput of 156 Mevents/s.

array.

The timing circuitry was designed with the proper functioning of the PHR in mind. The dual-clock TDC enables the continued use of a RO based architecture, thus limiting the power consumption, whilst at the same time minimising the error in the code outputs from the half-column TDCs as a result of LSB error accumulation. The clock generation for the sensor employs two PLLs to decouple the requirements of the TDC RO specifications from those of providing a low-jitter clock from a limited bandwidth PLL. The resulting SPAD to TDC signal chain achieves a FWHM almost identical to that of the single-clock TDC case, indicating that negligible jitter has been added by the extra clock.

Applied to LiDAR measurements to demonstrate the sensor operation, a linearity of 8.8 cm is achieved in long distance single-point ranging measurements up to a measurement range of 50 m. Flash measurements at 1 m distance demonstrate millimetric detail with a compression factor of 14.9-to-1. Although this has been demonstrated here with a slow readout, in fast readout mode the PHR will enable NIROT measurements with a much reduced acquisition time in comparison to TCSPC mode. Thus, the sensor is a powerful tool for use in a high resolution NIROT system.

6 Conclusions and Future Work

The goal of this thesis was to develop time-resolved single-photon detector based hardware for a high-resolution clinical NIROT system. This goal was achieved by optimising each stage of the signal chain and with the development of two sensors: Piccolo and Ocelot. In Chapter 2, two new pixel designs were presented for improving the SNR and dynamic range of time-resolved SPAD sensors. The first, was an all-transistor quenching and recharge circuit operating at voltages above the reliability limit of a single transistor for the first time. This circuit is crucial for improving the PDE of SPAD sensors, for both 3D and monolithic technologies. Indeed, the circuit enabled a PDE improvement, in comparison to the single transistor quenching and recharge, of 16% at 800 nm for the Piccolo and Ocelot sensors. Since this configuration is also compatible with active recharge, it raises the prospect of high dynamic range pixels with improved PDE. This could prove very powerful in future NIROT setups, where the dynamic range must be high to simultaneously handle short and distant source-detector separations and the PDE must be maximised for longer distances. The second pixel circuit demonstrated for the first time, quenching and recharge of SPADs with the high-voltage supplied at the anode or cathode terminals, thus it is capable of interfacing with any SPAD. The major advantage of this circuit is that different photodetector tiers may be combined with a single data processing tier in a 3D sensor configuration. Thus application specific photodetector tiers can be designed and fit to a general purpose TCSPC data processing tier.

To improve PDE further in comparison to what was achieved in this thesis, there are a number of avenues which can be pursued. Without custom diffusions or process modifications, the PDE of CMOS SPADs at longer wavelengths, e.g. >700 nm, in FSI technologies is not likely to improve much beyond the current state-of-the-art [67, 82, 91, 109] due to the limitations on junction depth. Thus, one possible avenue of future research would be to design quenching

and recharge circuits which can exceed the range of the cascode quenching and recharge circuit presented in Chapter 2. Another related area, is to verify the reliability of the conventional single transistor quenching and recharge circuit and the cascode circuit at excess bias voltages higher than those of their datasheet values. Both the gate-oxide breakdown [123] and hot carrier degradation [124] are time-dependent, and in SPAD sensors the time that the transistors would spend at voltages exceeding their safe threshold could be very small. This is particularly true for TCSPC sensors where the activity is limited by the output data bandwidth, here the pixel activity can be much less than 1%.

Another route to improving the PDE is with improved SPAD designs in BSI 3D IC technologies. Thus far, only a limited number of designs have been explored [125, 90]. For example, state-of-the-art isolated SPADs [67, 109] would be particularly suitable for a 3D design due to their high PDP in FSI and isolation between adjacent SPADs which is important for low crosstalk. Finally, a further possibility is the exploration of emerging techniques such as light trapping [126] which have been shown to increase the PDE by a factor of 2.5 in the near-infrared.

In Chapter 3, a new TDC sharing architecture for time-resolved SPAD sensors was introduced. By reducing the number of TDCs required in comparison to TDC-per-pixel approaches, the pixel fill factor is greatly increased whilst also simplifying the readout of data from the sensor. When applied to the NIROT application, this results in improved sensitivity, increased photon throughput of the sensor and a corresponding reduction in the measurement time. Thus, the TDC sharing architecture is critical for translating a high resolution NIROT system into clinical use.

The most obvious area for improvement of the sharing architecture in Piccolo, described in detail in Chapter 4, is implementation in static CMOS to minimise the current consumption during events on the bus. Unfortunately, the resulting tree based structure occupies a large area in monolithic sensors, thus any improvement in fill-factor gained by moving the TDCs outside of the pixel array will be eroded. Static CMOS trees will likely see extensive use in 3D IC technologies, where they can be implemented in highly scaled technologies without a large impact on fill-factor.

One of the limiting factors in data throughput is the output data bandwidth. Besides, increasing the bandwidth with more pads or the use of a faster transmission standard, future work for a high speed readout could involve combining the TDC sharing architecture with a parallel datapath such as [95]. Due to the exponential dependence of intensity on source-detector separation, dedicating output pads on a per-column basis can result in significant idle band-

width in columns with very little light. Therefore, by combining all events from a half-sensor or whole sensor into a parallel readout channel, greater dynamic ranges could be accommodated across the sensor as the readout bandwidth is utilized more efficiently.

In Chapter 5, the TDC sharing architecture from Piccolo was extended to a 252 × 144 format. The bus repeater scheme for the collision detection bus allows the area occupation of the column due to the bus to be minimised whilst maintaining a narrow bus dead time. This scheme provides a feasible route to scaling the architecture up to megapixel resolutions. A PHR was implemented to increase the photon throughput without requiring additional output data bandwidth. The PHR readout is, to the best of the author's knowledge, the first time-resolved SPAD sensor which includes on-chip histogramming on a per-pixel basis.

A complication of the PHR is that photons which have been measured by 6 separate TDCs are combined into a single histogram. Therefore, precision measurements where the TPSF must be corrected for DNL and LSB variations will require additional data processing. A future research avenue could involve the implementation of a PHR with a single TDC which is capable of sampling multiple photons per cycle. In this case, the data which is compiled in the SRAM is from a single TDC, and requires no additional data processing.

With the optical and electrical performance of the sensors demonstrated, future work will focus on verification of the sensors as components in complete NIROT systems such as in [110]. This will involve extensive experiments on homogeneous [115] and non-homogeneous [116] phantoms, as well as image reconstructions to investigate the achievable spatial resolution.

In the broader NIROT context, the work contained within this thesis and the suggestions for future work thus far represent a major step forward for high-resolution reflection mode NIROT. Improved PDE and higher throughput architectures will enable the demonstration of clinical tools based on large arrays of SPADs. However, there remains one domain where this work has less impact, that is, increasing the depth sensitivity. The reason for this is that the gains in PDE are fighting an exponential dependence on source-detector separation. Therefore, to achieve contrast at increasing depths, future sensors should combine a high throughput sensor architecture with in-pixel gating of the SPAD array. In the simplest implementation a gated time-resolved sensor could be achieved by adding gating functionality to the Piccolo or Ocelot pixels. The Ocelot architecture is particularly suited to such an application since the PHR already covers a time window of 800 ps (16 × 50 ps bins). Thus, with a gating window of 800 ps, the entire response would be encompassed by the compression. In this mode there is no requirement for peak detection, since the response window can be known in advance. If

a 3D IC technology was available, an array of direct-to-histogram TDCs could be coupled to SPAD macro-pixels or actively recharged SPADs for improved dynamic range. This method has the advantage that the increased dynamic range could enlarge the gating window, which would in turn reduce the acquisition time for measurements. This could enable NIROT image reconstructions at depths of up to 6 cm [34], thus presenting the possibility of applying the method to a much wider range of applications.

The work presented in this thesis opens up a new dimension of NIROT imagers, finally providing an imaging modality capable of quantifying oxygenation of the brains of preterm infants and tumors and thus ultimately achieving substantial progress in the treatment of patients.

Bibliography

[1] M. A. Green, "Self-consistent optical parameters of intrinsic silicon at 300K including temperature coefficients," *Solar Energy Materials and Solar Cells*, vol. 92, no. 11, pp. 1305 – 1310, 2008.

[2] H. Blencowe, S. Cousens, D. Chou, M. Oestergaard, L. Say, A.-B. Moller, M. Kinney, and J. Lawn, "Born Too Soon: The global epidemiology of 15 million preterm births," *Reproductive Health*, vol. 10, no. 1, p. S2, Nov 2013.

[3] G. Greisen, T. Leung, and M. Wolf, "Has the time come to use near-infrared spectroscopy as a routine clinical tool in preterm infants undergoing intensive care?" *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1955, pp. 4440–4451, 2011.

[4] E. A. Hutchinson, C. R. De Luca, L. W. Doyle, G. Roberts, P. J. Anderson, and , "School-age outcomes of extremely preterm or extremely low birth weight children," *Pediatrics*, vol. 131, no. 4, pp. e1053–e1061, 2013.

[5] L. Shalak and J. M. Perlman, "Hypoxic–ischemic brain injury in the term infant-current concepts," *Early Human Development*, vol. 80, no. 2, pp. 125 – 141, 2004.

[6] A. L. Harris, "Hypoxia — a key regulatory factor in tumour growth," *Nature Reviews Cancer*, vol. 2, pp. 38–47, 2002.

[7] A. M. Shannon, D. J. Bouchier-Hayes, C. M. Condron, and D. Toomey, "Tumour hypoxia, chemotherapeutic resistance and hypoxia-related therapies," *Cancer Treatment Reviews*, vol. 29, no. 4, pp. 297 – 307, 2003.

[8] M. Höckel, K. Schlenger, B. Aral, M. Mitze, U. Schäffer, and P. Vaupel, "Association between Tumor Hypoxia and Malignant Progression in Advanced Cancer of the Uterine Cervix," *Cancer Research*, vol. 56, no. 19, pp. 4509–4515, 1996.

[9] M.-C. Kavanagh, A. Sun, Q. Hu, and R. P. Hill, "Comparing Techniques of Measuring Tumor Hypoxia in Different Murine Tumors: Eppendorf pO_2 Histograph, [^{3}H]Misonidazole Binding and Paired Survival Assay," *Radiation Research*, vol. 145, no. 4, pp. 491–500, 1996.

[10] J. M. Weaver and K. J. Liu, "*In vivo* electron paramagnetic resonance oximetry and applications in the brain," *Medical Gas Research*, vol. 7, no. 1, pp. 56–67, 2017.

[11] Siyuan Liu and Sameer J. Shah and Lisa J. Wilmes and John Feiner and Vikram D. Kodibagkar and Michael F. Wendland and Ralph P. Mason and Nola Hylton and Harriet W. Hopf and Mark D Rollins., "Quantitative tissue oxygen measurement in multiple organs using 19F MRI in a rat model," *Magnetic Resonance in Medicine*, vol. 66, no. 6, pp. 1722–1730, 2011.

[12] O. J. Kelada and D. J. Carlson, "Molecular Imaging of Tumor Hypoxia with Positron Emission Tomography," *Radiation Research*, vol. 181, no. 4, pp. 335–349, 2014.

[13] I. N. Fleming, R. Manavaki, P. J. Blower, C. West, K. J. Williams, A. L. Harris, J. Domarkas, S. Lord, C. Baldry, and F. J. Gilbert, "Imaging tumour hypoxia with positron emission tomography," *British Journal Of Cancer*, vol. 112, pp. 238–250, 2015.

[14] K. S. McCommis, T. A. Goldstein, D. R. Abendschein, P. Herrero, B. Misselwitz, R. J. Gropler, and J. Zheng, "Quantification of Regional Myocardial Oxygenation by Magnetic Resonance Imaging," *Circulation: Cardiovascular Imaging*, vol. 3, no. 1, pp. 41–46, 2010.

[15] K. Thorngren-Jerneck, T. Ohlsson, A. Sandell, K. Erlandsson, S.-E. Strand, E. Ryding, and N. W. Svenningsen, "Cerebral Glucose Metabolism Measured by Positron Emission Tomography in Term Newborn Infants with Hypoxic Ischemic Encephalopathy," *Pediatric Research*, vol. 49, pp. 495–501, 2001.

[16] F. Martelli, S. D. Bianco, A. Ismaelli, and G. Zaccanti, *Light Propagation through Biological Tissue and Other Diffusive Media: theory, solutions and software.* Bellingham, Washington: Society of Photo-Optical Instrumentation Engineers (SPIE), 2010.

[17] J. R. Mourant, J. P. Freyer, A. H. Hielscher, A. A. Eick, D. Shen, and T. M. Johnson, "Mechanisms of light scattering from biological cells relevant to noninvasive optical-tissue diagnostics," *Applied Optics*, vol. 37, no. 16, pp. 3586–3593, Jun 1998.

[18] H. Dehghani, M. E. Eames, P. K. Yalavarthy, S. C. Davis, S. Srinivasan, C. M. Carpenter, B. W. Pogue, and K. D. Paulsen, "Near infrared optical tomography using NIRFAST: Algo-

rithm for numerical model and image reconstruction," *Communications in Numerical Methods in Engineering*, vol. 25, no. 6, pp. 711–732, 2008.

[19] F. Scholkmann, S. Kleiser, A. J. Metz, R. Zimmerman, J. M. Pavia, U. Wolf, and M. Wolf, "A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology," *NeuroImage*, vol. 85, pp. 6–27, 2014, celebrating 20 Years of Functional Near Infrared Spectroscopy (fNIRS).

[20] D. T. Delpy and M. Cope, "Quantification in tissue near-infrared spectroscopy," *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, vol. 352, no. 1354, pp. 649–659, 1997.

[21] F. Jobsis, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science*, vol. 198, no. 4323, pp. 1264–1267, 1977.

[22] Y. Zhang, Y. Chen, Y. Yu, X. Xue, V. V. Tuchin, and D. Zhu, "Visible and near-infrared spectroscopy for distinguishing malignant tumor tissue from benign tumor and normal breast tissues *in vitro*," *Journal of Biomedical Optics*, vol. 18, pp. 18 – 18 – 8, 2013.

[23] J. H. Meek, M. Firbank, C. E. Elwell, J. Atkinson, O. Braddick, and J. S. Wyatt, "Regional Hemodynamic Responses to Visual Stimulation in Awake Infants," *Pediatric Research*, vol. 43, pp. 840–843, 1998.

[24] B. R. and P. C.A., "Near-infrared spectroscopy for monitoring muscle oxygenation," *Acta Physiologica Scandinavica*, vol. 168, no. 4, pp. 615–622, 2000.

[25] M. Ferrari and V. Quaresima, "A brief review on the history of human functional near-infrared spectroscopy (fNIRS) development and fields of application," *NeuroImage*, vol. 63, no. 2, pp. 921 – 935, 2012.

[26] T. Hamaoka, K. K. McCully, M. Niwayama, and B. Chance, "The use of muscle near-infrared spectroscopy in sport, health and medical sciences: recent developments," *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, vol. 369, no. 1955, pp. 4591–4604, 2011.

[27] G. Enrico, M. W. W, vandeVen Martin J, F. J. B, M. M. B, and C. Britton, "A novel approach to laser tomography," *Bioimaging*, vol. 1, no. 1, pp. 40–46, 1993.

[28] D. T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, and J. Wyatt, "Estimation of optical pathlength through tissue from direct time of flight measurement," *Physics in Medicine & Biology*, vol. 33, no. 12, p. 1433, 1988.

[29] W. Becker, *The bh TCSPC Handbook*, Berlin, Germany, 2012.

[30] B. Chance, M. Cope, E. Gratton, N. Ramanujam, and B. Tromberg, "Phase measurement of light absorption and scatter in human tissue," *Review of Scientific Instruments*, vol. 69, no. 10, pp. 3457–3481, 1998.

[31] V. Q. Martin Wolf, Marco Ferrari, "Progress of near-infrared spectroscopy and topography for brain and muscle clinical applications," *Journal of Biomedical Optics*, vol. 12, 2007.

[32] A. Torricelli, D. Contini, A. Pifferi, M. Caffini, R. Re, L. Zucchelli, and L. Spinelli, "Time domain functional NIRS imaging for human brain mapping," *NeuroImage*, vol. 85, pp. 28 – 50, 2014, celebrating 20 Years of Functional Near Infrared Spectroscopy (fNIRS).

[33] M. A. Franceschini, K. T. Moesta, S. Fantini, G. Gaida, E. Gratton, H. Jess, W. W. Mantulin, M. Seeber, P. M. Schlag, and M. Kaschke, "Frequency-domain techniques enhance optical mammography: Initial clinical results," *Proceedings of the National Academy of Sciences*, vol. 94, no. 12, pp. 6468–6473, 1997.

[34] A. D. Mora, D. Contini, S. Arridge, F. Martelli, A. Tosi, G. Boso, A. Farina, T. Durduran, E. Martinenghi, A. Torricelli, and A. Pifferi, "Towards next-generation time-domain diffuse optics for extreme depth penetration and sensitivity," *Biomedical Optics Express*, vol. 6, no. 5, pp. 1749–1760, May 2015.

[35] A. Tosi, A. D. Mora, F. Zappa, A. Gulinatti, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli, and R. Cubeddu, "Fast-gated single-photon counting technique widens dynamic range and speeds up acquisition time in time-resolved measurements," *Optics Express*, vol. 19, no. 11, pp. 10 735–10 746, May 2011.

[36] S. R. Arridge and J. C. Schotland, "Optical tomography: forward and inverse problems," *Inverse Problems*, vol. 25, no. 12, p. 123010, 2009.

[37] A. D. Kim, "Transport theory for light propagation in biological tissue," *Journal of the Optical Society of America A*, vol. 21, no. 5, pp. 820–827, May 2004.

[38] C. Zhu and Q. Liu, "Review of Monte Carlo modeling of light transport in tissues," *Journal of Biomedical Optics*, vol. 18, pp. 18 – 18 – 13, 2013.

[39] H. Dehghani, S. Srinivasan, B. W. Pogue, and A. Gibson, "Numerical modelling and image reconstruction in diffuse optical tomography," *Philosophical Transactions of the*

Royal Society of London A: Mathematical, Physical and Engineering Sciences, vol. 367, no. 1900, pp. 3073–3093, 2009.

[40] Y. Yamada and S. Okawa, "Diffuse optical tomography: Present status and its future," *Optical Review*, vol. 21, no. 3, pp. 185–205, May 2014.

[41] J. M. Pavia, "Near-Infrared Optical Tomography with Single-Photon Avalanche Diode Image Sensors," Ph.D. dissertation, École polytechnique fédérale de Lausanne, February 2015.

[42] J. M. Pavia, M. Wolf, and E. Charbon, "Single-Photon Avalanche Diode Imagers Applied to Near-Infrared Imaging," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 291–298, Nov 2014.

[43] G. Bodi and Y. Bérubé-Lauzière, "A new deconvolution technique for time-domain signals in diffuse optical tomography without a priori information," in *Diffuse Optical Imaging II*. Optical Society of America, 2009, p. 7369_14.

[44] M. Diop and K. S. Lawrence, "Deconvolution method for recovering the photon time-of-flight distribution from time-resolved measurements," *Optics Letters*, vol. 37, no. 12, pp. 2358–2360, Jun 2012.

[45] F. E. W. Schmidt, M. E. Fry, E. M. C. Hillman, J. C. Hebden, and D. T. Delpy, "A 32-channel time-resolved instrument for medical optical tomography," *Review of Scientific Instruments*, vol. 71, no. 1, pp. 256–265, 2000.

[46] P. Taroni, A. Pifferi, E. Salvagnini, L. Spinelli, A. Torricelli, and R. Cubeddu, "Seven-wavelength time-resolved optical mammography extending beyond 1000 nm for breast collagen quantification," *Optics Express*, vol. 17, no. 18, pp. 15 932–15 946, Aug 2009.

[47] E. Lapointe, J. Pichette, and Y. Bérubé-Lauzière, "A multi-view time-domain non-contact diffuse optical tomography scanner with dual wavelength detection for intrinsic and fluorescence small animal imaging," *Review of Scientific Instruments*, vol. 83, no. 6, p. 063703, 2012.

[48] J. C. Hebden, A. Gibson, T. Austin, R. M. Yusof, N. Everdell, D. T. Delpy, S. R. Arridge, J. H. Meek, and J. S. Wyatt, "Imaging changes in blood volume and oxygenation in the newborn infant brain using three-dimensional optical tomography," *Physics in Medicine & Biology*, vol. 49, no. 7, p. 1117, 2004.

[49] J. C. Hebden and T. Austin, "Optical tomography of the neonatal brain," *European Radiology*, vol. 17, no. 11, p. 2926, May 2007.

[50] E. Ferocino, E. Martinenghi, A. D. Mora, A. Pifferi, R. Cubeddu, and P. Taroni, "High throughput detection chain for time domain optical mammography," *Biomedical Optics Express*, vol. 9, no. 2, pp. 755–770, Feb 2018.

[51] A. Farina, S. Tagliabue, L. Di Sieno, E. Martinenghi, T. Durduran, S. Arridge, F. Martelli, A. Torricelli, A. Pifferi, and A. Dalla Mora, "Time-Domain Functional Diffuse Optical Tomography System Based on Fiber-Free Silicon Photomultipliers," *Applied Sciences*, vol. 7, no. 12, p. 1235, Nov 2017.

[52] J. J. Selb, D. K. Joseph, and D. A. Boas, "Time-gated optical system for depth-resolved functional brain imaging," *Journal of Biomedical Optics*, vol. 11, pp. 11 – 11 – 13, 2006.

[53] Q. Zhao, L. Spinelli, A. Bassi, G. Valentini, D. Contini, A. Torricelli, R. Cubeddu, G. Zaccanti, F. Martelli, and A. Pifferi, "Functional tomography using a time-gated ICCD camera," *Biomedical Optics Express*, vol. 2, no. 3, pp. 705–716, Mar 2011.

[54] S. Cova, A. Longoni, and A. Andreoni, "Towards picosecond resolution with single-photon avalanche diodes," *Review of Scientific Instruments*, vol. 52, no. 3, pp. 408–412, 1981.

[55] S. Cova, A. Lacaita, M. Ghioni, G. Ripamonti, and T. A. Louis, "20-ps timing resolution with single-photon avalanche diodes," *Review of Scientific Instruments*, vol. 60, no. 6, pp. 1104–1110, 1989.

[56] A. Rochas, M. Gani, B. Furrer, P. A. Besse, R. S. Popovic, G. Ribordy, and N. Gisin, "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology," *Review of Scientific Instruments*, vol. 74, no. 7, pp. 3263–3270, 2003.

[57] C. Niclass, M. Sergio, and E. Charbon, "A Single Photon Avalanche Diode Array Fabricated in Deep-Submicron CMOS Technology," in *Proceedings of the Design Automation Test in Europe Conference*, vol. 1, March 2006, pp. 1–6.

[58] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128 $times$ 128 Single-Photon Image Sensor With Column-Level 10-Bit Time-to-Digital Converter Array," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, Dec 2008.

[59] 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," in *2011 IEEE International Solid-State Circuits Conference*, Feb 2011, pp. 312–314.

[60] M. Alayed and M. Deen, "Time-Resolved Diffuse Optical Spectroscopy and Imaging Using Solid-State Detectors: Characteristics, Present Status, and Research Challenges," *Sensors*, vol. 17, no. 9, p. 2115, Sep 2017.

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

[62] C. Niclass, M. Gersbach, R. Henderson, L. Grant, and E. Charbon, "A Single Photon Avalanche Diode Implemented in 130-nm CMOS Technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 13, no. 4, pp. 863–869, July 2007.

[63] M. Fishburn, "Fundamentals of CMOS Single-Photon Avalanche Diodes," Ph.D. dissertation, Delft University of Technology, September 2012.

[64] E. A. G. Webster and R. K. Henderson, "A TCAD and Spectroscopy Study of Dark Count Mechanisms in Single-Photon Avalanche Diodes," *IEEE Transactions on Electron Devices*, vol. 60, no. 12, pp. 4014–4019, Dec 2013.

[65] E. O. Kane, "Theory of Tunneling," *Journal of Applied Physics*, vol. 32, no. 1, pp. 83–91, 1961.

[66] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, and B. Mamdy, "Industrialised SPAD in 40 nm technology," in *2017 IEEE International Electron Devices Meeting (IEDM)*, Dec 2017, pp. 16.5.1–16.5.4.

[67] C. Veerappan and E. Charbon, "A Low Dark Count p-i-n Diode Based SPAD in CMOS Technology," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 65–71, Jan 2016.

[68] M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti, and M. Ghioni, "Photon-Timing Jitter Dependence on Injection Position in Single-Photon Avalanche Diodes," *IEEE Journal of Quantum Electronics*, vol. 47, no. 2, pp. 151–159, Feb 2011.

[69] I. Rech, A. Ingargiola, R. Spinelli, I. Labanca, S. Marangoni, M. Ghioni, and S. Cova, "Optical crosstalk in single photon avalanche diode arrays: a new complete model," *Optics Express*, vol. 16, no. 12, pp. 8381–8394, Jun 2008.

[70] W. J. Kindt, H. W. van Zeijl, and S. Middelhoek, "Optical Cross Talk in Geiger Mode Avalanche Photodiode Arrays: Modeling, Prevention and Measurement," in *28th European Solid-State Device Research Conference*, Sept 1998, pp. 192–195.

[71] D. Bronzi, S. Tisa, F. Villa, S. Bellisai, A. Tosi, and F. Zappa, "Fast Sensing and Quenching of CMOS SPADs for Minimal Afterpulsing Effects," *IEEE Photonics Technology Letters*, vol. 25, no. 8, pp. 776–779, April 2013.

[72] L. Pancheri, N. Massari, and D. Stoppa, "SPAD Image Sensor With Analog Counting Pixel for Time-Resolved Fluorescence Detection," *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3442–3449, Oct 2013.

[73] N. A. W. Dutton, I. Gyongy, L. Parmesan, S. Gnecchi, N. Calder, B. R. Rae, S. Pellegrini, L. A. Grant, and R. K. Henderson, "A SPAD-Based QVGA Image Sensor for Single-Photon Counting and Quanta Imaging," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 189–196, Jan 2016.

[74] E. Charbon, H. J. Yoon, and Y. Maruyama, "A Geiger mode APD fabricated in standard 65nm CMOS technology," in *2013 IEEE International Electron Devices Meeting*, Dec 2013, pp. 27.5.1–27.5.4.

[75] D. Bronzi, F. Villa, S. Tisa, A. Tosi, and F. Zappa, "SPAD Figures of Merit for Photon-Counting, Photon-Timing, and Imaging Applications: A Review," *IEEE Sensors Journal*, vol. 16, no. 1, pp. 3–12, Jan 2016.

[76] P. Seitz and A. J. P. Theuwissen, *Single Photon Imaging*. Berlin Heidelberg: Springer-Verlag, 2011.

[77] B. Aull, J. Burns, C. Chen, B. Felton, H. Hanson, C. Keast, J. Knecht, A. Loomis, M. Renzi, A. Soares, V. Suntharalingam, K. Warner, D. Wolfson, D. Yost, and D. Young, "Laser Radar Imager Based on 3D Integration of Geiger-Mode Avalanche Photodiodes with Two SOI Timing Circuit Layers," in *2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers*, Feb 2006, pp. 1179–1188.

[78] 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*, vol. 50, no. 10, pp. 2406–2418, Oct 2015.

[79] B. F. Aull, E. K. Duerr, J. P. Frechette, K. A. McIntosh, D. R. Schuette, and R. D. Younger, "Large-Format Geiger-Mode Avalanche Photodiode Arrays and Readout Circuits," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–10, March 2018.

[80] 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," in *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, Feb 2015, pp. 1–3.

[81] C. Niclass and M. Soga, "A miniature actively recharged single-photon detector free of afterpulsing effects with 6ns dead time in a 0.18 μm CMOS technology," in *2010 International Electron Devices Meeting*, Dec 2010, pp. 14.3.1–14.3.4.

[82] 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*, vol. 33, no. 11, pp. 1589–1591, Nov 2012.

[83] C. Veerappan and E. Charbon, "A Low Dark Count p-i-n Diode Based SPAD in CMOS Technology," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 65–71, Jan 2016.

[84] 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*, vol. 24, no. 16, pp. 17 819–17 831, Aug 2016.

[85] S. Lhostis, A. Farcy, E. Deloffre, F. Lorut, S. Mermoz, Y. Henrion, L. Berthier, F. Bailly, D. Scevola, F. Guyader, F. Gigon, C. Besset, S. Pellissier, L. Gay, N. Hotellier, A. L. L. Berrigo, S. Moreau, V. Balan, F. Fournel, A. Jouve, S. Chéramy, M. Arnoux, B. Rebhan, G. A. Maier, and L. Chitu, "Reliable 300 mm Wafer Level Hybrid Bonding for 3D Stacked CMOS Image Sensors," in *2016 IEEE 66th Electronic Components and Technology Conference (ECTC)*, May 2016, pp. 869–876.

[86] 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*, vol. 21, no. 14, pp. 1020–1022, July 2009.

[87] G. P. Singh and R. B. Salem, "High-voltage-tolerant I/O buffers with low-voltage CMOS process," *IEEE Journal of Solid-State Circuits*, vol. 34, no. 11, pp. 1512–1525, Nov 1999.

[88] J. Xiao, A. V. Peterchev, J. Zhang, and S. R. Sanders, "A 4-μA quiescent-current dual-mode digitally controlled buck converter IC for cellular phone applications," *IEEE Journal of Solid-State Circuits*, vol. 39, no. 12, pp. 2342–2348, Dec 2004.

[89] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Scaleable Single-Photon Avalanche Diode Structures in Nanometer CMOS Technology," *IEEE Transactions on Electron Devices*, vol. 58, no. 7, pp. 2028–2035, July 2011.

[90] T. A. Abbas, N. A. W. Dutton, O. Almer, S. Pellegrini, Y. Henrion, and R. K. Henderson, "Backside illuminated SPAD image sensor with 7.83 μm pitch in 3D-stacked CMOS technology," in *2016 IEEE International Electron Devices Meeting (IEDM)*, Dec 2016, pp. 8.1.1–8.1.4.

[91] S. Mandai, M. W. Fishburn, Y. Maruyama, and E. Charbon, "A wide spectral range single-photon avalanche diode fabricated in an advanced 180 nm CMOS technology," *Optics Express*, vol. 20, no. 6, pp. 5849–5857, Mar 2012.

[92] M. Gersbach, Y. Maruyama, R. Trimananda, M. W. Fishburn, D. Stoppa, J. A. Richardson, R. Walker, R. Henderson, and E. Charbon, "A Time-Resolved, Low-Noise Single-Photon Image Sensor Fabricated in Deep-Submicron CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 6, pp. 1394–1407, June 2012.

[93] F. Villa, R. Lussana, D. Bronzi, S. Tisa, A. Tosi, F. Zappa, A. D. 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 of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 364–373, Nov 2014.

[94] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-Pixels Digital Silicon Photomultiplier Direct TOF Sensor With 100-MPhotons/s/pixel Background Rejection and Imaging/Altimeter Mode With 0.14% Precision Up To 6 km for Spacecraft Navigation and Landing," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, Jan 2017.

[95] 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*, vol. 49, no. 4, pp. 867–880, April 2014.

[96] J. Arlt, D. Tyndall, B. R. Rae, D. D.-U. Li, J. A. Richardson, and R. K. Henderson, "A study of pile-up in integrated time-correlated single photon counting systems," *Review of Scientific Instruments*, vol. 84, no. 10, p. 103105, 2013.

[97] C. Niclass, M. Sergio, and E. Charbon, "A CMOS 64 × 48 Single Photon Avalanche Diode Array with Event-Driven Readout," in *2006 Proceedings of the 32nd European Solid-State Circuits Conference*, Sept 2006, pp. 556–559.

[98] D. Tyndall, B. R. Rae, D. D. U. Li, J. Arlt, A. Johnston, J. A. Richardson, and R. K. Henderson, "A High-Throughput Time-Resolved Mini-Silicon Photomultiplier With Embedded Fluorescence Lifetime Estimation in 0.13 μm CMOS," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 6, pp. 562–570, Dec 2012.

[99] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. Pellegrini, S. J. McLeod, L. A. Grant, and R. K. Henderson, "Digital Silicon Photomultipliers With OR/XOR Pulse Combining Techniques," *IEEE Transactions on Electron Devices*, vol. 63, no. 3, pp. 1105–1110, March 2016.

[100] K. Nie, X. Wang, J. Qiao, and J. Xu, "A Full Parallel Event Driven Readout Technique for Area Array SPAD FLIM Image Sensors," *Sensors*, vol. 16, no. 2, 2016.

[101] 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," in *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, Feb 2018, pp. 96–98.

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

[103] 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*, vol. 1, no. 1, 2017.

[104] S. Mandai and E. Charbon, "A 4 × 4 × 416 digital SiPM array with 192 TDCs for multiple high-resolution timestamp acquisition," *Journal of Instrumentation*, vol. 8, no. 05, p. P05024, 2013.

[105] D. Tamborini, B. Markovic, F. Villa, and A. Tosi, "16-Channel Module Based on a Monolithic Array of Single-Photon Detectors and 10-ps Time-to-Digital Converters," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 218–225, Nov 2014.

[106] 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*, vol. 48, no. 2, pp. 559–572, Feb 2013.

[107] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, "A 32 × 32 50ps resolution 10 bit time to digital converter array

in 130nm CMOS for time correlated imaging," in *2009 IEEE Custom Integrated Circuits Conference*, Sept 2009, pp. 77–80.

[108] A. R. Ximenes, P. Padmanabhan, and E. Charbon, "Mutually Coupled Ring Oscillators for Large Array Time-of-Flight Imagers," in *2017 International Image Sensor Workshop*, June 2017.

[109] M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce, and F. Zappa, "Single-Photon Avalanche Diodes in a 0.16μm BCD Technology With Sharp Timing Response and Red-Enhanced Sensitivity," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–9, March 2018.

[110] A. Kalyanov, J. Jiang, S. Lindner, L. Ahnen, A. di Costanzo, J. M. Pavia, S. S. Majos, and M. Wolf, "Time Domain Near-Infrared Optical Tomography with Time-of-Flight SPAD Camera: The New Generation," in *Biophotonics Congress: Biomedical Optics Congress 2018 (Microscopy/Translational/Brain/OTS)*. Optical Society of America, 2018, p. OF4D.5.

[111] L. Ahnen, H. Stachel, S. Kleiser, C. Hagmann, J. Jiang, A. Kalyanov, S. Lindner, M. Wolf, and S. Sanchez, *Development and Validation of a Sensor Prototype for Near-Infrared Imaging of the Newborn Brain*. Cham: Springer International Publishing, 2017, pp. 163–168.

[112] Q. Fang and D. A. Boas, "Monte Carlo Simulation of Photon Migration in 3D Turbid Media Accelerated by Graphics Processing Units," *Optics Express*, vol. 17, no. 22, pp. 20 178–20 190, Oct 2009.

[113] J. Jiang, L. Ahnen, A. Kalyanov, S. Lindner, M. Wolf, and S. S. Majos, *A New Method Based on Graphics Processing Units for Fast Near-Infrared Optical Tomography*. Cham: Springer International Publishing, 2017, pp. 191–197.

[114] J. Jiang, M. Wolf, and S. S. Majos, "Fast reconstruction of optical properties for complex segmentations in near infrared imaging," *Journal of Modern Optics*, vol. 64, no. 7, pp. 732–742, 2017.

[115] A. Pifferi, A. Torricelli, A. Bassi, P. Taroni, R. Cubeddu, H. Wabnitz, D. Grosenick, M. Möller, R. Macdonald, J. Swartling, T. Svensson, S. Andersson-Engels, R. L. P. van Veen, H. J. C. M. Sterenborg, J.-M. Tualle, H. L. Nghiem, S. Avrillier, M. Whelan, and H. Stamm, "Performance assessment of photon migration instruments: the MEDPHOT protocol," *Appl. Opt.*, vol. 44, no. 11, pp. 2104–2114, Apr 2005.

[116] H. Wabnitz, A. Jelzow, M. Mazurenka, O. Steinkellner, R. Macdonald, D. Milej, N. Żołek, M. Kacprzak, P. Sawosz, R. Maniewski, A. Liebert, S. Magazov, J. C. Hebden, F. Martelli, P. D. Ninni, G. Zaccanti, A. Torricelli, D. Contini, R. Re, L. M. Zucchelli, L. Spinelli, R. Cubeddu, and A. Pifferi, "Performance assessment of time-domain optical brain imagers, part 2: nEUROPt protocol," *Journal of Biomedical Optics*, vol. 19, pp. 19 – 19 – 12, 2014.

[117] S. Burri, Y. Maruyama, X. Michalet, F. Regazzoni, C. Bruschini, and E. Charbon, "Architecture and applications of a high resolution gated SPAD image sensor," *Optics Express*, vol. 22, no. 14, pp. 17 573–17 589, Jul 2014.

[118] 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," in *2017 International Image Sensor Workshop*, June 2017.

[119] I. Gyongy, N. Calder, A. Davies, N. A. W. 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*, vol. 65, no. 2, pp. 547–554, Feb 2018.

[120] 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*, vol. 51, no. 1, pp. 155–167, Jan 2016.

[121] S. J. Goldman, *Phase-Locked Loop Engineering Handbook for Integrated Circuits*. Boston, USA: Artech House, 2007.

[122] A. A. Abidi, "Phase Noise and Jitter in CMOS Ring Oscillators," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 8, pp. 1803–1816, Aug 2006.

[123] J. Wu and E. Rosenbaum, "Gate oxide reliability under ESD-like pulse stress," *IEEE Transactions on Electron Devices*, vol. 51, no. 7, pp. 1192–1196, July 2004.

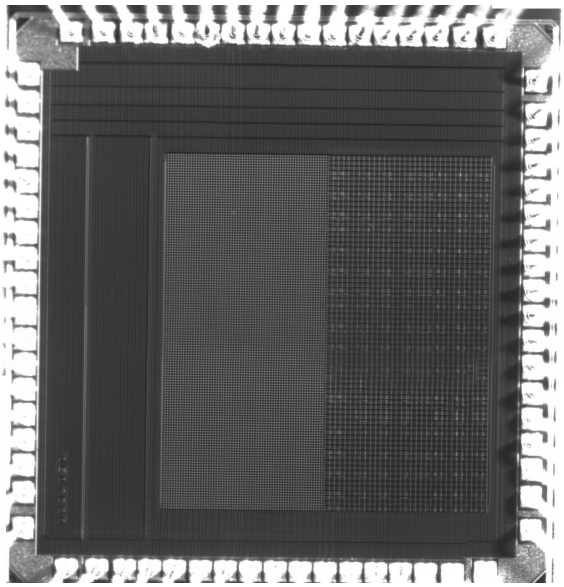
[124] G. V. Groeseneken, "Hot carrier degradation and ESD in submicrometer CMOS technologies: how do they interact?" *IEEE Transactions on Device and Materials Reliability*, vol. 1, no. 1, pp. 23–32, Mar 2001.

[125] M. J. Lee, A. R. Ximenes, P. Padmanabhan, T. J. Wang, K. C. Huang, Y. Yamashita, D. N. Yaung, and E. Charbon, "High-Performance Back-Illuminated Three-Dimensional

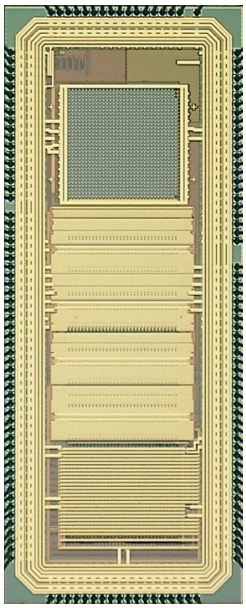
Stacked Single-Photon Avalanche Diode Implemented in 45-nm CMOS Technology," *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 6, pp. 1–9, Nov 2018.

[126] K. Zang, X. Jiang, H. Yijie, 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*, vol. 8, 2017.

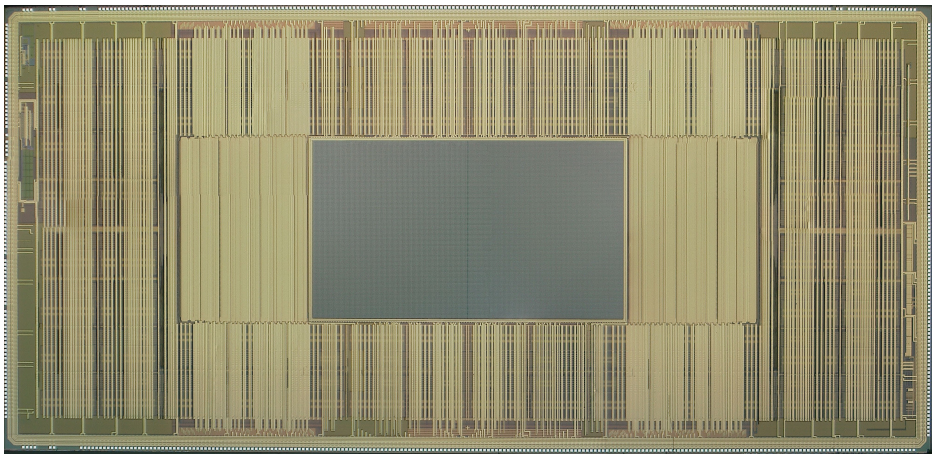
Chip Gallery



(a) BSI 3D IC test sensor, Chapter 2.



(b) Piccolo sensor, Chapter 3.



Ocelot sensor, Chapter 5

List of Publications

International peer-reviewed journal articles

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," in *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547-1550, Nov. 2017.

C. Zhang, **S. Lindner**, I. M. Antolovic, M. Wolf and E. Charbon "A CMOS SPAD Imager with Collision Detection Bus and 128 Dynamic Reallocating TDCs for Single Photon Counting and 3D Time-of-Flight Imaging", [Shared first authorship, submitted to Optics Express]

International conferences

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]

S. Lindner, C. Zhang, I. 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]

A. Kalyanov, J. Jiang, **S. Lindner**, L. Ahnen, A. di Costanzo, J. Pavia, S. Majos, and M. Wolf, "Time Domain Near-Infrared Optical Tomography with Time-of-Flight SPAD Camera: The New Generation," in Biophotonics Congress: Biomedical Optics Congress 2018 (Microscopy-/Translational/Brain/OTS), OSA Technical Digest (Optical Society of America, 2018), paper OF4D.5.

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" in *2018 Symposium on VLSI Circuits*, June 2018. [Shared first authorship]

A. Carimatto, A. Ulku, **S. Lindner**, E. D'Aillon, S. Pellegrini, B. Rae and E. Charbon, "Multipurpose, Fully-Integrated 128 × 128 Event-Driven MD-SiPM with 512 16-bit TDCs with 45 ps LSB and 20 ns Gating,", in *2018 Symposium on VLSI Circuits*, June 2018.

Curriculum Vitae

Scott Lindner
born in 1988 in Chatham, United Kingdom

Professional Experience

2013-2018	**Swiss Federal Institute of Technology, Lausanne (EPFL) & University Hospital Zurich**
	Doctoral assistant
	Time-resolved SPAD sensor design for NIROT
2011-2013	**European Space Agency - Noordwijk, The Netherlands**
	Young Graduate Trainee
	Analog IC design for space applications

Education

2013-2018	**Swiss Federal Institute of Technology, Lausanne (EPFL)**
	Advanced Quantum Architecture Laboratory (AQUA)
	PhD candidate in electrical engineering
2006-2011	**University of Edinburgh - Edinburgh, United Kingdom**
	MEng in electronics and electrical engineering
	International exchange from 2008-2009 at Iowa State University, USA
1999-2006	**Rainham Mark Grammar School - Rainham, Kent, United Kingdom**
	High school, A levels in mathematics, physics and design technology