THE UNIVERSITY *of* EDINBURGH

This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.
A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.
This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.
The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.
When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.

Analysis and Optimisation of High Throughput Digital Silicon Photomultipliers

Salvatore Gnecchi

Doctor of Philosophy

THE UNIVERSITY OF EDINBURGH

2016

Abstract

Large area detectors for time correlated single photon counting (TCSPC) are nowadays being implemented in CMOS technology to benefit a large variety of applications including positron emission tomography (PET) and 3D laser ranging (LiDAR), exploiting the advanced timing and counting capabilities inside single chips. Single photon avalanche diodes (SPADs) and silicon photomultipliers (SiPMs) represent a great option to realise such detectors thanks to their exceptional timing resolution and the ability to be arranged into arrays. Recently, digital SiPMs (dSiPMs) have been introduced to improve the integration with CMOS technology overcoming limitations on the readout of analogue SiPMs and thus improving the photon resolution of the detector. This work presents a 14 GSamples/s time-to-digital converter (TDC) to improve the throughput of dSiPM sensors commonly limited by the sampling rate of the timing/counting readout circuitry. The converter has been demonstrated on a test chip in 130nm CMOS imaging technology paired with a novel XOR-based 32×32 SPAD array single-channel detector. The overall achieved throughput equals 1 GEvents/s demonstrated in a direct time-of-flight LiDAR experiment. By acquiring a number of photons significantly higher than one per laser pulse, this approach represents the first example in TCSPC of an input rate and conversion rate both higher than the excitation rate. The following part of the work presents a modelling analysis on how to match the achieved high sampling rate / throughput of the single-channel TDC to the performance of a SPAD array. The impact of a selection of dSiPM design parameters, such as photon detection efficiency, dead time and size of the SPAD cell, number of cells per single-channel, digital N-to-1 combining network and channel bandwidth, on the overall sensor throughput and the dynamic range has been characterised thanks to a computational Monte-Carlo simulator and useful equations describing each of the processes in the sensing chain. The *pile-up* effect, i.e. the event-loss causing non-linear distortions on the output signal, has been characterised on each element of the dSiPM and optimisations have been proposed. Event losses in the SPAD cells due to dead time, in the digital combining network due to network dead time and single-channel bandwidth have all been identified, simulated and described by analytical equations. All the results coming from the theoretical analysis have been reproduced in real dSiPM design thanks to a reconfigurable test chip realised in the same 130 nm CMOS imaging technology specifically to validate the proposed theory. The manufactured test chip provides the very first direct comparison between OR-based and XOR-based single-channel dSiPM sensors highlighting the promising timing and counting performance of the newly introduced XOR-based dSiPM. Direct evidence of pile-up distortions and subsequent reduction through design optimisations are demonstrated. A recommended design flow for next generation dSiPMs is proposed at the end of the publication.

Declaration

I hereby declare that the content of this document has been produced entirely by myself with the acknowledgement of the help and contribution of the following designers:

- Prof. Robert Henderson for the design and layout reviewing process
- Dr. Justin Richardson for the structure of the blue-sensitive SPAD structure included in the reconfigurable dSiPM test chip
- Dr. Eric Webster for the structure of the red-sensitive SPAD structure present in the TDC test chip
- Dr. Luca Parmesan for the layout of the SPAD pixels included in the reconfigurable dSiPM test chip and general layout and firmware support in both designed test chips
- Dr. Neale Dutton for the layout of the detector array in the TDC test chip as well as the design of the digital readout of the TDC test chip as well as FPGA firmware for the testing

Prior to this, the content of this document has not been submitted for any professional or academic purposes.

Salvatore Gnecchi

Acknowledgements

First of all, I would like to send my sincere *thank you* to Prof. Robert Henderson for giving me the opportunity to work with his group at the University of Edinburgh where I had the chance for the very first time to challenge myself with real chip design. To all the great people in the group who have helped me during this tough experience. To Dr. Neale Dutton, for his constant mentoring, guiding and personal support; it was very much appreciated, it was a pleasure to work together on the TDC test chip and in general to be part of the same design team. To my Italian fellow, Dr. Luca Parmesan, nothing will beat our non-sense chats, the laughs which made the intense work more enjoyable. To Dr. Laurence Stark, for all the encouragement, his amazing sense of humour and the perfectly on-time "tea o'clock" alarms which made our afternoons more bearable. Thanks also to Tarek Al Abbas for his precious reviews and proof reading, I cannot wait to read all about your project. To ST Microelectronics for funding the project and its amazing team at the design centre in Edinburgh: to my industrial supervisor Dr. Bruce Rae, Dr. Sara Pellegrini, Dr. Lindsay Grant, Dr. Andrew Holmes, Stuart McLeod, Jeff Raynor, Nicky Grant goes my gratitude for making it a great working experience.

To all the near and long-distance support I received throughout these four years: my entire family, my closest friends Cristina Frascà, Azzurra Costa, Nicola Zonno, for their patience and the unconditional support over these many many years. To the new friends that this journey has brought close to me. A particular thank you goes to Georgios Tsiolis, our friendship is one of the most valuable accomplishments of these four years, I cannot thank you enough for making the difference in my life in Edinburgh and beyond! To Georgie Wilde, for being a lovely flatmate as soon as I got to Scotland for the very first time, I am very happy life made our paths cross. To the amazing Dr. Liam Brierley: sharing a flat together for two years has been a blast. We shared the madness of the post-graduate student life and great stories. Thank you for taking care of me as your flatmate. To all the people I had the chance to spend great time together in Edinburgh: George Stefanakis, Lito Kriara, Anastasia Vardaki, Struan Brydone , Dustin Hosseini, Jonathan Freinberg and Martino Sorbaro (thank you for taking care of the actual printing of this thesis) and whoever I might be missing in these few lines. And to the new people I have been living and working with in Cork. My new boss, Dr. Carl Jackson for giving me the opportunity to start my career after this tough experience and for the understanding and support during the thesis writing. To my new friends, for coping with my intense level of stress in the last months, thank you Damiano Tescaro, Nico Mendes, Jamie Cotter, Dr. Feargal Ryan and also a special thank you to Christopher Lyons (thank you for all the kind words, the good food we shared and for just being there).

To each and everyone who believed in me. Thank you.

List of Publications

First Author Publications

[1] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. Pellegrini, L. A. Grant, and R. K. Henderson, "Gphotons/s TCSPC with XOR Digital Silicon Photomultipliers," in *Single Photon Workshop*, 2015.

[2] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. J. McLeod, S. Pellegrini, L. A. Grant, and R. K. Henderson, "A Simulation Model for Digital Silicon Photomultipliers," *IEEE Transactions on Nuclear Science*, vol. 63, no. 3, pp. 1343–1350, 2016.

[3] S. Gnecchi, N. A. W. Dutton, L. Parmesan, R. Bruce, S. Pellegrini, S. J. Mcleod, L. A. Grant, and R. K. Henderson, "Analysis of Photon Detection Efficiency and Dynamic Range in SPAD based Visible Light Receivers," *Journal of lightwave technology*, vol. 8724, no. 1, pp. 1–7, 2016

[4] 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, 2016.

Co-author Publications

[5] O. Almer, N. A. W. Dutton, T. A. Abbas, S. Gnecchi, and R. K. Henderson, "4-PAM Visible Light Communications with a XOR-tree Digital Silicon Photomultiplier," *in IEEE Summer Topicals Meeting Series*, 2015, pp. 2–3.

[6] N. Dutton, J. Vergote, S. Gnecchi, L. Grant, D. Lee, S. Pellegrini, B. Rae, and R. Henderson, "Multiple-event direct to histogram TDC in 65nm FPGA technology," *Ph.D. Research in Microelectronics and Electronics (PRIME), 2014 10th Conference on*, pp. 1–5, 2014.

[7] 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," i*n Solid- State Circuits Conference - (ISSCC), 2015 IEEE International, 2015*, pp. 204–206.

[8] O. Almer, D. Tsonev, N. A. W. Dutton, T. A. Abbas, S. Videv, S. Gnecchi, H. Haas, and R. K. Henderson, "A SPAD-based Visible Light Communications Receiver Employing Higher Order Modulation Schemes," *IEEE Globecom*, 2015.

[9] 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 singlephoton counting and quanta imaging," *IEEE Transactions on Electron Devices*, vol. 63, no. 1, pp. 189–196, 2016.

[10] N. A. Dutton, L. Parmesan, S. Gnecchi, I. Gyongy, N. J. Calder, B. R. Rae, L. A. Grant, and R. K. Henderson, "Oversampled ITOF Imaging Techniques using SPADbased Quanta Image Sensors," *International Image Sensor Workshop*, 2015

Patents

[11] N. A. Dutton, R. K. Henderson, and S. Gnecchi, "TIME TO DIGITAL CONVERTER AND APPLICATIONS THEREOF," 2015.

Contents

Abstract ii

Declaration iii

Acknowledgements iv

List of Publications vi

Contents viii

List of Figures xii

List of Tables xxv

1 Introduction 1

1.1 Single-Photon Timing Applications 4

1.1.1 Time-of-flight 5

1.1.2 Positron emission tomography (PET) 7

1.1.3 Fluorescence lifetime imaging microscopy (FLIM) 9

1.1.4 3D imaging: LiDAR 10

1.2 Time correlated single-photon counting 16

1.2.1 Pile-up distortions 18

1.2.2 Sensor throughput 19

1.3 Research aims 20

1.4 Contribution to knowledge 21

1.5 Thesis structure 22

2 State of the Art of TCSPC Systems 24

2.1 Detectors 26

2.1.1 Photomultiplier Tubes (PMTs) . . . 26
2.1.2 Avalanche photo-diodes (APDs) . . . 27
2.1.3 Single Photon Avalanche Diodes (SPADs) . . . 29
2.1.4 Silicon Photomultipliers (SiPMs) . . . 37
2.1.5 Digital SiPM . . . 42
2.2 Photodetectors and Time-of-Flight 3D LiDAR . . . 50
2.2.1 Photo-current calculation . . . 52
2.2.2 Signal-to-Noise Ratio . . . 53
2.3 Time to Digital Converters (TDCs) . . . 54
2.3.1 Figures of Merit . . . 55
2.3.2 Counters . . . 57
2.3.3 Cyclic Interpolation . . . 58
2.3.4 Vernier Principle . . . 58
2.3.5 Flash TDC . . . 61
2.3.6 TDC Summary . . . 62

3 Modelling of dSiPMs 64

3.1 Overview on Modelling Approaches . . . 65
3.2 Structure of the dSiPM Simulator . . . 67
3.2.1 Generation of stream of photons . . . 68
3.2.2 SPAD cell . . . 69
3.2.3 N-to-1 Network . . . 72
3.3 Simulation results for dSiPMs . . . 76
3.3.1 Photon generation . . . 76
3.3.2 Detection of photon per unit cell . . . 77
3.3.3 N-to-1 Network Simulation . . . 80
3.4 Analytical Modelling for dSiPMs . . . 83
3.4.1 Analytical model for OR-based dSiPMs . . . 83
3.4.2 Analytical model for the XOR-tree . . . 86
3.5 Pile-up Analysis in Time Domain . . . 88

3.5.1 Detector Pile-Up . . . 89
3.5.2 Channel Pile-Up . . . 94
3.6 Summary . . . 102

4 Test Chip Description 105
4.1 Multi Event Time-to-Digital Converter . . . 106
4.1.1 SPAD Array . . . 108
4.1.2 XOR Network . . . 112
4.1.3 METDC . . . 115
4.1.4 Input Serial-to-Parallel interface . . . 122
4.1.5 Overview and Design Achievements . . . 122
4.2 Configurable dSiPM Chip . . . 123
4.2.1 SPAD cell variants . . . 125
4.2.2 Digital N-to-1 Networks . . . 127
4.2.3 On-chip Counters . . . 130
4.2.4 Additional Features . . . 130
4.3 FPGA Firmware . . . 131
4.4 Summary . . . 132

5 Results and Modelling Comparison 133
5.1 Experimental Setup . . . 134
5.2 DSiPM Test Chip . . . 136
5.2.1 Number of Unit cells . . . 138
5.2.2 Unit cell dimension . . . 142
5.2.3 Digital N-to-single channel network . . . 145
5.3 Multi Event Time-to-Digital Converter . . . 149
5.3.1 Code Density Test and Linearity . . . 149
5.3.2 Throughput Test . . . 152
5.3.3 Time Resolution . . . 154
5.4 Pulse Distortion Analysis . . . 157

5.5 Summary 165

6 Conclusions and Outlook **168**

6.1 Summary 168

6.2 Discussion 171

6.3 Future Work 174

Bibliography **178**

Appendices

A Modelling Scripts **198**

A.1 Photo Detection - Active Recharged SPAD cell 198

A.2 Photo Detection - Passive Recharged SPAD cell 199

B Experiments **201**

B.1 Number of Cells Analysis 201

B.2 Combining Technique Analysis 205

Publications **210**

List of Figures

1.0.1 **Beam of light -** Light is represented as both waves and particles. Photons are described by their wavelength and the intensity of the beam is the total sum of flux of photons defined as the energy irradiated over a unit area per second. . 2

1.1.1 **Time-of-Flight system -** The scene is illuminated by a controlled beam and synchronised to a photo-detector which then can record the time elapsed between the emission of the light and the arrival at the detector [12]. 5

1.1.2 **Time-of-Flight -** The measure of the time-of-flight can be done by the direct measurement of the time shift of a received pulsed light (a.1) or from the phase shift of a modulated sine wave (a.2). Alternatively, indirect measurements can be performed by measuring in four time windows the resulting charge at the photo-detector due to the pulsed (b.1) or modulated (b.2) illumination [12]. . . 6

1.1.3 **PET system -** A picture of a general PET scanner. The emission of a positron by a radioactive nuclide is followed by the generation of two high energy photons scattered simultaneously in two opposite directions. A ring of detectors captures the photon pair giving both timing and spatial information on the occurred event [13]. 8

1.1.4 **Time-activity curve -** and example of the obtained data from a scanned plasma after PET tracer injection [14]. 9

1.1.5 **Fluorescence lifetime -** Excited atoms emits fluorescence photon when relaxing back into their low energy configurations. The average timing for this process is in the order of few nanoseconds. The intensity of the emitted fluorescence light decays exponentially as shown. Diagram taken from [15]. . . . 10

1.1.6 **LiDAR system -** A collimated beam of light is sent to the scene. A target diffuses and reflects a fraction of the incident light back onto the sensor. The receiver lens collects the diffused light and concentrates it onto the sensor with additional filtering. A control unit manages the timing sync between transmitter and receiver so that the time elapsed by the illumination to the collection of light can be recorded and the distance extracted. 11

1.1.7 **LiDAR measurement example -** An obtained histogram from a LiDAR experiment in time correlated single photon counting mode using pulsed laser light with a reflective target at a given distance. Results taken from [16]. 15

1.2.1 **Classic TSPC setup -** A time correlated single photon counting system is pictured highlighting its components: a laser source, providing time-controller illumination, a receiver with single-photon sensitivity, and a readout chain consisting of an amplifier, a comparator and a digitiser. The data is collected over many acquisitions and therefore a memory is attached to integrate the results and build up a TCSPC histogram. Original diagram taken from [15]. 16

1.2.2 **TCSPC -** The acquisition of the optical waveform is performed over multiple number of excitations and a histogram is built from single-photon detections [15]. 17

1.2.3 **Pile-up example -** The simulated data of three exponential decays are plotted together with the expected / corrected models for pile-up correction. Depending on the *intensity* of pile-up, i.e. based on the event rates, some reconstructed curves do not match the data due to uncorrected pile-up [17]. 19

1.2.4 **Sensor throughput and pile-up sources -** Each component of a TCSPC system is characterised by a maximum event rate at its output. This is referred to as throughput. Limited throughput causes potential non-linear photon loss at any stage of the sensor. This diagram presents the principal causes of pile-up in TCSPC systems [18]. 20

2.0.1 **TCSPC principle** - The optical waveform is reconstructed by acquiring single photons per each laser pulse. All the time stamped photons are counted onto a histogram that resembles the input waveform. It is essential to minimise the loss of acquired photon in order to avoid non-linearity distortions. 25

2.1.1 **Photomultiplier Tubes** - A PMT consists of a photo-cathode to generate primary photo-electrons which are then multiplied through a series of dynodes to generate a high current at the opposite terminal of the tube. 26

2.1.2 **Avalanche photo-diodes** - APDs in the linear mode of operation are used to detect the light level incident on the sensor by measuring the generated avalanche current after a stage of amplification due to the realative low internat gain. The generated carriers, due to reverse barrier, cause a series of impact ionisations resulting in a macroscopic current through the diode. This current is the amplified by an external circuit which then allows the usage of comparators for photon-counting. 28

2.1.3 **SPADs** - An APD biased above the reverse breakdown voltage, (a.1), is able to trigger a self-sustaining avalanche when a single photon is absorbed. The generated avalanche resulting in the photon current (a.2), must be *quenched* by lowering the current through the diode itself, (a.3). Once the quenching is done, the diode can be restored in its initial state ready for the next photon-detection, (a.4). The quenching and the recharge process can be done by a quenching resistor and a load capacitance, (b). The waveforms for the SPAD note and its digitised output are presented, (c). 30

2.1.4 **The region of operation of a SPAD** - A SPAD can operate in three regions: a DCR dominated region where the response is independent on the incident rate, a linear region where the response is proportional to the incident rate through the PDE, and a saturation region where the registered counts either reach a maximum given by the dead time - in the case of active recharge, or drop down to zero due to the paralysis of the detector - a for passive recharge SPAD. 34

2.1.5 **Silicon Photomultipliers** - A set of $k \times k$ SPAD cells are connected in parallel to create an SiPM. The connection makes all the independent cells contribute to the output current/voltage of the entire detector. Image taken from [19]. . . 37

2.1.6 **SensL SiPM with Fast Output** - Each SPAD is coupled to a dedicated node through a capacitance. The shape of the output current from this node is essentially the derivative of the traditional output allowing narrower pulses to be produced enhancing the photon number resolution. Drawings taken from datasheet [20]. 41

2.1.7 **The digital silicon photomultiplier** - In opposition to conventional analogue SiPMs (a), a dSiPMs combines the digitised output of each SPAD cell through a digital network (b). Additional counting and timing circuitry can be put inside the same chip. 43

2.1.8 **First dSiPM chip** - A microphotograph (a) and a block diagram (b) of the very first fully digital SiPM realised by Frach *et al.* [21]. 44

2.1.9 **OR tree** - Two close-in-time events can be merged together by an OR tree when the inter-arrival time of the events is lower than the individual cell dead time. 45

2.1.10 **OR tree with monostable circuit**- An additional compression stage is introduced to minimise the merge of SPAD pulses during the OR tree combination network. An example timining diagram (a) is shown beside the implemented circuit schematic (b) as shown in [22]. 46

2.1.11 **XOR gate for SPAD combination**- An alternative to the OR gate is suggested to improve the timing combination of SPADs through an XOR gate. Both the circuit (a) and timing diagram (b) are taken from [22]. 46

2.1.12 **A column parallel TDC for dSiPMs** - In opposition to a single shared TDC (a) and an individual TDC per pixel (b), a column parallel TDC (c) is introduced by Mandai *et al.* [23]. 47

2.1.13 **XOR tree** - The toggle and the XOR tree encode all the SPAD events onto both rising and falling edges maintaining the timing information. 48

2.1.14 **Pile-up** - Alrt *et al.* show a comprehensive diagram highlighting all the sources of non-linear distortions in a typical dSiPM sensing chain, from detector to timing circuitry [18]. 49

2.2.1 **Incident photon rate** - Calculated average photon rate on the sensors in a reference LiDAR systems. 51

2.2.2 **LiDAR SNR** - Calculated signal-to-noise ratio for the three detector types in a reference LiDAR system. The high internal gain of SiPMs offers higher SNR at long distances for low reflective targets. 53

2.3.1 **TDC with Cyclic Interpolation** - The fine measurement of the cyclic TDC is done by successively shrinking the width of the input pulse and by counting the number of required cycles for the pulse to vanish. Diagram taken from [24]. . . 59

2.3.2 **Vernier Principle** - By using two oscillators with a slight difference in clock frequency, a fine time interval is expanded into a multiple of the period mismatch of the used clocks (a). The extended delay is measured by the desired coarse architecture. An example is from Porat *et al.* (b) [25]. 60

2.3.3 **Flash TDC** - Sampling the state of a delay line enables faster conversion rates and the expenses of poorer single-shot resolution. Diagrams for tapped delay lines (a) and Vernier delay lines (b) are taken from [26]. 61

3.2.1 **Monte Carlo simulator -** Each probabilistic process occurring in the dSiPM is being simulated in different blocks. Each one of them is presented as an algorithm. The output of each block represents the response in the time domain of the component of the dSiPM. 67

3.2.2 **Generation of stream of photons -** For a selected time window, an average photon rate is distributed in time. 68

3.2.3 **SPAD cell simulation -** Algorithms used to simulated the output of a single SPAD cell with (a) active recharge and (b) passive recharge mechanism. . . . 71

3.2.4 **Output of OR-tree based dSiPMs -** Shown for a no timing compression technique - the digital output of each cell is directly connected to the combining OR-tree (a) and for a pulse-shortened technique - each individual output is compressed by a monostable circuit which enables shorter pulses to be combined by the OR-tree (b) 73

3.2.5 **Output of XOR-tree based dSiPMs -** Each event is now encoded onto rising and falling edges of the digital output of the pixel through a toggle flip-flop. All the edges are propagated onto the final single channel by XOR-gates. 75

3.3.1 **Generation of photons - Test -** For a given set of average number of photons in the time window, photon streams are generated. Statistical analysis show that the probability of generation fits the Poissonian expression (a). An example of the photon stream is provided (b). 77

3.3.2 **Detection of photon in single SPAD cells - Test -** The SPAD simulator has been tested for a variety of incident photon rates and a fixed set of SPAD parameters. The obtained registered even rates match the known charateristics of non-paralysable detector, valid for the case of active recharged SPAD (a) and for paralysable detector for passive recharged diodes (b). 79

3.3.3 **N-to-1 network simulation - Test -** simulated OR tree for a different number of SPAD cells N_{SPADs} and different input photon rates. The collected results are plotted together with paralysable (red) and non-paralysable (blue) models. . . 81

3.3.4 **N-to-1 network simulation - Test -** simulation results for the XOR tree with a different number of SPAD cells N_{SPADs}. Paralysable (red) and non-paralysable (blue) models for comparison. 82

3.4.1 **OR-Tree Model -** An OR-based dSiPM has been modelled with a fixed number of SPAD cells and a decreasing width of the monostable output pulse. This impacts the counting performance of the array in terms of maximum count rate and saturation profile. The model treats the OR-tree as a paralysable detector. 85

3.4.2 **XOR-Tree - Test -** Predicted curves describing the registered count rates of a XOR-based dSiPM with increasing number of SPADs activated in the array. The length of the linear region of operation is reduced for a large dSiPM due to limited throughput of the single-channel bandwidth. The XOR-based dSiPM showcases a flat saturation typical of non-paralysable detectors. 87

3.5.1 **Detector pile-up distortion -** A single SPAD cell distorts an intense square pulse into an exponential decay due to its internal dead time. The output histogram generated by simulations is the digitisation of the optical input. A situation of partial saturation (a) is shown against full saturation due to very intense incident light (b). 91

3.5.2 **Detector pile-up distortion -** The distortion due to detector saturation is usually overcome by attenuating the incident light on the sensor. In this case, an ND filter with coefficient equal to 1.2 is being simulated. 92

3.5.3 **Detector pile-up distortion -** When filters are used to avoid pile-up, the exposure time has to be increased to improve the statistic and consequently the shape of the collected histogram. In this example, the exposure time has been increased to 1 second. 93

3.5.4 **Channel pile-up analysis -** For a crescent number of cells on which the light is spread, OR and XOR are compared. The numbers of cells here simulated are: 2 (b), 4 (c), 10(d) and 50 (e). The left column corresponds to the OR network with a monostable circuit output of 1 ns while the right column shown the respond of an XOR-based dSiPM. The latter network better rejects the distortion of the input waveform (a). 95

3.5.5 **OR-based dSiPM temporal response -** The paralysable nature of the OR-tree network causes pile-up distortion of the input square waveform which translates into an initial decay before reaching the average value due to throughput. Obtained waveforms are simulated for an increasing number of cells: $N_{cells} =$ 5 (a), 50 (b), 200 (c), 500 (d). 98

3.5.6 **Saturated OR-based dSiPM distortions -** When each single cells of the dSiPM enters the saturation region, the input rates of the OR network reduces significantly. This causes distortions due to detector pile-up. However, an additional distortion can be caused by the saturation of the network as well. 99

3.5.7 **XOR-based dSiPM temporal response -** The limitation due to pulse merging is overcome in XOR-based dSiPM. Higher throughputs are achieved and the pulsed waveform is maintained also at high number of cells. Obtained waveforms are simulated for an increasing number of cells: $N_{cells} = 5$ (a), 50 (b), 200 (c), 500 (d). 100

3.5.8 **XOR dSiPM saturation -** The high throughput of the XOR tree enters a saturation region when the event rates hit a maximum given by the bandwidth of the single channel. As results, the registered counts per each bin is lower than the ideal value. Here the simulated data are compared to an updated analytical model taking into account such limitation. 101

3.5.9 **XOR dSiPM saturation -** The high throughput of the XOR tree enters a saturation region when the event rates hit a maximum given by the bandwidth of the single channel. As results, the registered counts per each bin is lower than the ideal value. Here the simulated data are compared to an updated analytical model taking into account such limitation. The plots correspond to $N_{cells} = 10$ (a), 50 (b) and 100 (c). 103

4.1.1 **METDC test chip -** A picture of a high throughput TDC implemented together with a SPAD array combined onto single channel by XOR gates. 107

4.1.2 **METDC diagram -** The main blocks of the METDC test chip are pictured showing the detection chain from the detector array to the main TDC front end as well as the histogram formation directly on chip. [7] 108

4.1.3 **METDC timing diagram -** The VCO generates 33 phases that sample in parallel the return signal coming from the XOR-based dSiPM. A folded XOR ring detects continuously events as variations in the sampled signals. At every clock cycles, these values are shifted over a 8-bit pipeline to extend the range. Every 8 cycles, the pipeline is passed onto ripple-counters effectively forming the histogram on chip. The readout of the counters is done sequentially on a different clock domain and after the exposure has ended. 109

4.1.4 **SPAD cell -** Layout of the unit cell used in the detector of the METDC sensor chip. The space at the top right of the pixel is reserved to the column XOR-cells. 110

4.1.5 **SPAD cell -** Schematic of the SPAD cell including on-chip circuitry: quenching resistor, level shifter/buffers, SRAM for cell enabling (a) and on-chip toggle-cell (b). 111

4.1.6 **XOR H-tree -** The timing path of the output of each SPAD in the column is balanced through an H-tree geometry (a). To improve the fill factor of the array, a flattened H-tree (b) is adopted to ensure the minimum distance among columns. The blue line highlights the equalised signal path. 113

4.1.7 **XOR cells in METDC chip -** The XOR network is divided in two levels. The first one is a column XOR based on a single ended cell (a,b) while the second level, combining the 32 columns, is external to the array and consists of symmetric cells creating both XOR and XNOR signals (c,d). 114

4.1.8 **METDC SPAD array-** Layout of the 32 $\times$ 32 SPAD array with internal column XOR-tree and external XOR-tree for column combination. A second external XOR-tree has been added to provide a direct output signal onto a dedicated pad. 115

4.1.9 **METDC core -** 33 flip-flops sample in parallel the digital output of the XOR combined SPAD output. An XOR ring detects the transitions between consecutive clock phases. The output is connected to ripple counters. [7] 116

4.1.10 **METDC PLL diagram -** The generation of the 33 phases of a clock is done by a voltage controlled oscillator put in a phase locked loop. The PLL is based on a phase detector (PD) comparing the current clock signal with an external reference, which generates UP and DOWN signals to a charge pump (CP). This one generates the necessary voltage for the VCO. 117

4.1.11 **METDC VCO -** A 33 stages differential VCO is included in the METDC chip. The schematic (a) shown how the VCTRL voltage sets the upper voltage of the delay elements. The output nodes are sent to level shifters to be brought back to the full supply swing before being connected to the front-end of the METDC. 118

4.1.12 **METDC sampling flip-flop -** A delay flip-flop is used to sample the input signal per each clock phase. A replica cell is connected in series to reject metastability errors. Both the schematic (a) and layout (b) are here presented. 119

4.1.13 **METDC pipeline timing diagram -** The output of every one of the 33 stages of the XOR ring goes into a pipeline, which is a 8-bit shift register on the correspondent clock phase. For 8 clock periods, the output of the ring is shifted through the pipeline therefore being stored of 8 clock periods. Then, thanks to the 1:8 divided clock cycle, all the stages of the pipeline are passed to the counters, which work on the slower clock to be then counted through the 16-bit ripple counters. The different timing domains are highlighted by a different colour to help visualise the timing in the pipeline system. 120

4.2.1 **Configurable dSiPM picture -** A programmable dSiPM chip allows the selection of different SPAD cell size and number together with the digital N-to-1 network. The picture of the chip is here proposed. 123

4.2.2 **Configurable dSiPM diagram -** The diagram of the sensor is here shown highlighting the main design options for a desired dSiPM configuration. 125

4.2.3 **D2 Array layout -** The circuitry for the 4×4 arrays has been placed outside the light sensitive area. The structure is common for all the arrays **D2-D5**. . . 126

4.2.4 **Configurable dSiPM cell -** A blue sensitive SPAD cell includes a quenching resistor, enabling/disabling transistors and a buffer for level shifting to 1.2 V. . 127

4.2.5 **Configurable dSiPM 4×4 arrays -** The layout of the arrays is shown for 7 µm (a),9 µm (b),13 µm (c),18 µm (d),34 µm (e) 128

4.2.6 **Dedicated 16×16 array in the dSiPM chip -** A larger array is placed in the chip for the deep analysis of XOR-based dSiPMs. The structure priorities the time-balance among elements rather than the uniformity of the light sensitive areas. 129

4.2.7 **Monostable pulse shortening circuit -** The external bias allows the width of the output pulse to be tuned for dedicated analysis on OR-based dSiPMs. Schematic (a) and layout (b) of the cell is presented. 129

5.1.1 **Light level analysis** - The experimental setup consists of a light source, a voltage controlled LED for continuous light (a), or a laser diode for pulse light (b); a PCB to connect the dSiPM chips to an FPGA. All the experiment have been run using an external PC for configuration and acquisition of data. A photograph of the test bench (c) is also provided. 135

5.2.1 **Throughput experiment for XOR-based dSiPMs -** For a high number of activated cells, the registered curves show a limited throughput given by the channel bandwidth. 139

5.2.2 **Efficient number of cells -** The recorded maximum count rates for an XOR-based dSiPM is compared to the analytic model. The latter allows the optimal number of unit cells. 141

5.2.3 **Dynamic range of a dSiPM with increasing number of enabled cells -** The limitation on the channel bandwidth degrades the dynamic range. 142

5.2.4 **Impact of unit cell dimension in fixed dSiPM area -** Photon transfer curves are presented for different configurations to be compared in terms of count rates. From a single big cell to 14 small cell dSiPM, the effect on throughput and dynamic range can be seen all together. 144

5.2.5 **PDE and Dynamic Range dependence on size of unit cell size** - The change in the PDE (a) and dynamic range (b) is measured for different number of unit cells fitted in a total area. 146

5.2.6 **N-to-single channel network -** The OR-tree is directly compared to the XOR-tree with on-chip counters for throughput analysis 147

5.2.7 **N-to-single channel network with off-chip counters -** The throughput of both networks is limited by the bandwidth of the dSiPM pad. 148

5.3.1 **METDC Linearity -** A Code Density Test has been performance to calculate (a) the DNL and (b) the INL of the TDC. 150

5.3.2 **Throughput experiment for the METDC -** The maximum number of converted events has been measured with increasing input signal rate of the TDC. The mean values are plotted (a) together with the deviation from the modelled value (b). 153

5.3.3 **One-event resolution analysis -** Trough a digital delay generator, the minimum variation of the input of the TDC producing a detectable change has been measured. The obtained time-stamps are plotted against the input delays (a) together with the deviation from the ideal line (b) 155

5.3.4 **Depth resolution analysis -** Two events spaced by 1.4 ns are successfully discriminated by the METDC. 156

5.4.1 **Pile-up experimental setup -** The METDC sensor is mounted on a PCB and illuminated with a square laser pulse. To adjust the incident optical power on the detector, filters and diffusers have been used. 158

5.4.2 **Square pulse pile-up analysis -** A low intensity square pulse is distorted by a large single-channel dSiPM with the effect of a reduction of the pulse hight due to single-channel saturation. Waveforms are plotted in absolute values (a) and normalised to the peak value (b) to better appreciate the reduction of pulse height. 159

5.4.3 **Distortion due to single-channel -** the peak and the noise floor of the reconstructed optical waveform is plotted. A high number of cells makes the registered counts deviate from the ideal linear behaviour making the output waveform having a distorted (lower) height. 160

5.4.4 **Detector pile-up experiment -** A saturated SPAD cell distorts a square pulse into an exponential decay waveform. This is used on purpose in a pile-up distortion experiment to analyse how a single-channel dSiPM can distort this shap onto a square pulse. The collected counts (in blue) are compared to the modelled funtion $W_1(t)$ of (5.4.3). 162

5.4.5 **Distortion of an exponential decay waveform -** As soon as the number of cells increases, the dSiPM saturation starts clipping the top of the output waveform distorting its shape from a pure exponential decay curve onto a square wave which therefore does not contain any information on the time constant of the decay. All the data points are plotted together with the proposed model of (5.4.6). 163

5.4.6 **Distortion due to XOR cancellation -** An improved analytical model shown in (5.4.8) describes the reduction of registered counts in an XOR-tree when the input rate exceeds an intrinsic saturation level of the tree here empirically determined. 166

6.2.1 **Proposed design flow** - Highlighted in blue, the input coming from analysis on the optical photon budget and information regarding the CMOS technology including the SPAD performance. The flow chart is intended as starting point of the discussion for future design of dSiPM receivers. 175

List of Tables

2.1 **Commercial PMTs** A comparison among PMTs currently available on the market 27

2.2 **Commercial APDs** Summary of some APDs available on the market 29

2.3 **Review of recent SPAD design** - A selection of works on SPADs is here reported in terms of a selection of figures of merit [27]. 35

2.4 **Commercial aSiPMs** Figures of merit for commercially available analogue SiPMs 38

2.5 **Survey on published dSiPM design** - An overview on a selection of dSiPMs parameters is taken from Mandati *et al.* [28]. 49

2.6 **LiDAR System** - Set of example parameters for comparison between PIN, APD and SiPM devices. 50

2.7 **LiDAR System** - Set of example system parameters for compact systems based on PIN, APD and SiPM devices. 53

2.8 **TDCs** Summary of published TDC architectures for dSiPM readout. 62

3.1 **SPAD cell simulation -** Parameter setup. 78

3.2 **OR-Tree Model -** Parameter Setup. 84

3.3 **XOR-Tree Model -** Parameter Setup. 88

3.4 **Detector Pile-Up Test -** Parameter Setup. 89

3.5 **OR Pile-Up Test** - Parameter Setup 97

4.1 **DSiPM** - Properties and Parameters of unit cells. 125

5.1 **Number of unit cells analysis** - List of parameter values used in the experiment. 139

5.2 **Unit cell dimension** - List of parameter values used in the experiment are here reported. 143

5.3 **Digital N-to-1 network** - List of parameter values used in the experiment. . . 145

5.4 **dSiPM Performance Summary** - Unit cell pitch = 7 µm 149

5.5 **Experimental setup for pulse distortion analysis** - List of parameter values used in the experiment. 158

Chapter 1

Introduction

Cameras are nowadays one of most exciting pieces of technology that has become a big part of everyday life. From phones to accessories, from cars to robots, the realisation of powerful cameras with a lower and lower cost has enabled the interest of many manufacturers to incorporate their many capabilities into an extended variety of devices. Cameras are not just meant to take *pictures*. At least not in general meaning of the term. Light, visible or not, can be exploited in many more ways than just a description of what objects *look* like. In fact, when cameras are paired with specific light sources, they can turn into more exciting devices for which the term of *imaging devices* is preferred.

Light and *photons* have important characteristics that can be exploited to extract more information about the scene in front of the imaging device. Photons can be considered both particles and waves, therefore they can be represented as a propagating radiation with a certain frequency (or wavelength), speed and energy, or they can be represented as particles, see Fig. 1.0.1. The speed of photons, and in general of any electromagnetic radiation, is the commonly called *speed of light* and its measure in the vacuum represents one of the most famous physics constants as it is the upper limit of any observable speed. Its value is accepted as

$$c = 299792458 \text{ m/s} \tag{1.0.1}$$

The energy of a single photon is expressed as:

$$E_p = \frac{hc}{\lambda_p} \tag{1.0.2}$$

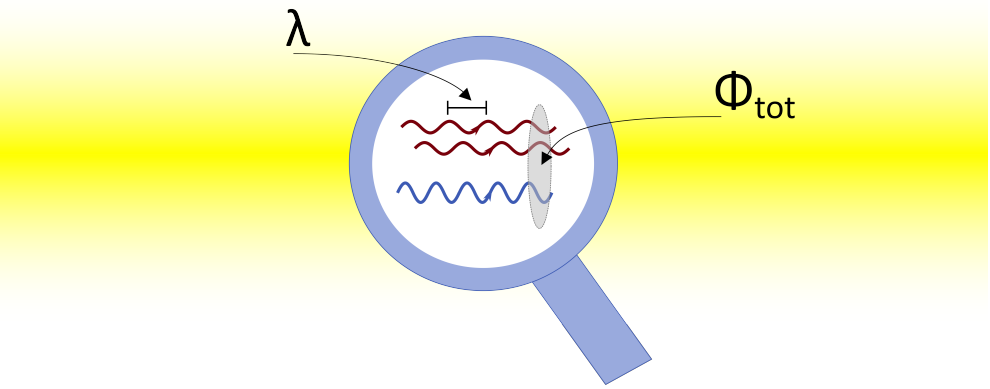


Figure 1.0.1: Beam of light - Light is represented as both waves and particles. Photons are described by their wavelength and the intensity of the beam is the total sum of flux of photons defined as the energy irradiated over a unit area per second.

where λ_p is the wavelength of the photon and h is Planck's constant equal to

$$h = 6.62607004 \cdot 10^{-34}\ \mathrm{m^2 kg/s} \tag{1.0.3}$$

The *intensity* of a beam of light can be expressed therefore by its flux of energy which is defined as the amount of energy travelling over a unit area per second and can be measured as:

$$\mathfrak{I} = \Phi_{tot} = \frac{E_{tot}}{A\Delta t} \qquad \left[\mathrm{W/m^2}\right] \tag{1.0.4}$$

where E_{tot} is the sum of the energy of every single photon in the beam $E_{tot} = \sum_i E_{p,i}$, A is the area of interest and Δt is a fixed amount of time for which the measurement has taken place. When a picture of an object is taken, the camera measures for a certain amount of time, called *exposure time*, the quantity of light that has been *reflected* by the object towards the camera itself. Typically, the light that is reflected comes from what is usually called *ambient light*. This type of light can come from the sun in outdoor environments, or from light bulbs etc. The main characteristic of this type of illumination is that it typically contains a large range of wavelengths. The concept of wavelength is often associated with the human perception of *colour* and therefore the measure of the intensity of the reflected light from an object over a

certain set of wavelengths results in the information of how the object *looks like*. However, what the human eye can perceive is only a small part of the wavelength spectrum and for such reason the light that falls into that range is called **visible light**. The colour information is therefore only one of the many others that can be obtained by the analysis of the reflected light from an object. Moreover, illuminating with different types of light also enables the extraction of additional information thanks to the ability of image sensors to perform both timing and energy measures.

It is worth to distinguish different types of illumination. Light can be composed by photons of a single wavelength, therefore called **monochromatic**, or by photons of all different wavelengths in the electromagnetic spectrum, referred to as **white light**. The intensity of the emitted beam can be **continuous**, i.e. constant in time, or **pulsed** i.e. with a certain time-profile and periodical. Here are some examples:

- **Monochromatic continuous light** is the illumination approximated by laser diodes or sodium lamps which have a narrow bandwidth in the spectral emission to be considered monochromatic.
- **Monochromatic pulsed light** can be obtained by controlling the laser diodes to produce a pulsed output.
- **White continuous light** examples can be daylight and sunlight, halogen light bulbs etc.
- **White pulsed light** can be thought as flashing white lamps.

The continuous light is often referred as *ambient light* even though it can be generated by lamps or LEDs; the terminology comes from the idea that this type of illumination does not provide any *timing information* which is, as explained later in this chapter, essential for many applications. For such reason, a particular attention is focused on monochromatic pulsed light which represents the signal source in most systems.

The aim of this introduction is to present some examples of applications where image sensors are employed to identify important properties about the observed scene. These examples are:

- **LiDAR:** in both consumer and scientific applications it provides distance measures between the sensor and the scene by exploiting a controlled light source and the known speed of light.

- **Positron emission tomography (PET):** used in medical science it allows, through imagining, the observation of metabolic processes in living organisms.

A particular attention is given to the 3D LiDAR ranging, for such reasons, the tests presented later on in this work will be ranging experiments. However, results in particular scenarios offer interesting outlooks also for PET experiments. Both applications often have to deal with very low intensities and therefore require the ability to sense *single photons* to obtain the desired information. The next two sections give a general picture of these two single-photon applications with the aim of highlighting the specific requirements on the sensors.

1.1 Single-Photon Timing Applications

In the early years of the 20^{th} century, Einstein and Planck introduced the concept of the quantisation of the light. For a given wavelength, the energy of a beam of light can be observed only in certain discreet values proportional to the energy of the *quantum*, i.e. the single *photon*. Different light levels are therefore described by the *average* amount of photons travelling through a certain area in a given amount of time as expressed by (1.0.4). High light levels are represented by a high number of photons incident on a surface per unit time while a very low light level can be thought as rare single photons travelling over the given area and time. Due to the quantised nature of light, one cannot think of a low light level just as a very low energy flux, but has to think of it as one single photon over given time. This understanding has significant implications, the most relevant to this research being the necessity of capturing every single photon to extract its energy and its timing information.

The already mentioned applications base their principles of operation on these two properties of single photons as described in the following subsections.

1.1.1 Time-of-flight

The timing information of the transmitted (and reflected) light is often captured to measure the *time-of-flight* (ToF) from the emission to the detection of the light as shown in Fig. 1.1.1. Two examples relying on this principle are:

- **Coincidence time detection**: it is used in scientific applications to determine the spatial and temporal correlation of pairs of photons emitted after the occurrence of a certain metabolic process.
- **Depth measurement** is used in many consumer applications to build a 3D depth map or to measure the distance with a single object in the scene.

The ToF techniques are classified according to the type of the incident illumination on the detector, pulse or continuous wave modulation, and by the calculation method, direct or indirect, see Fig. 1.1.2. Using a pulsed illumination, the received pulse is shifted in time (and eventually attenuated by the target) by an offset equal to the ToF τ_{ToF}. In the continuous modulation approach, the intensity of the light is described by a periodic sinusoidal waveform which, due to the time-of-flight, will be shifted (and attenuated) exhibiting a phase shift φ_{ToF}. These two values can be measured directly or indirectly as explained in the following paragraphs.

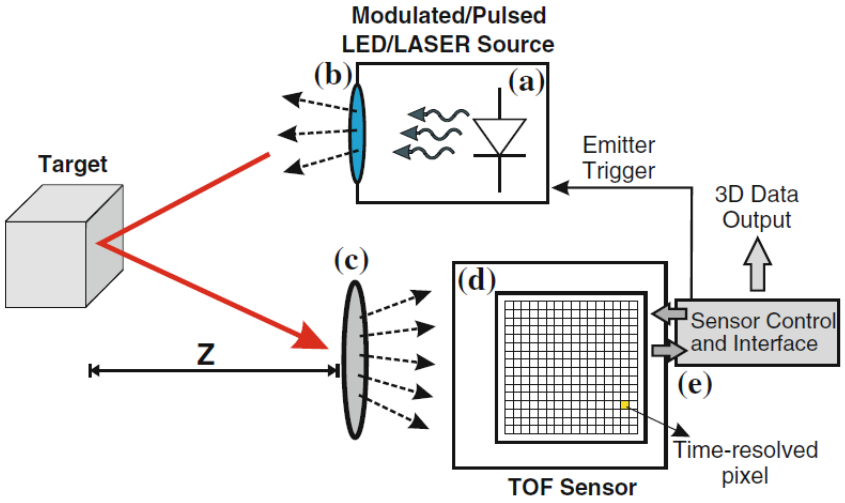


Figure 1.1.1: Time-of-Flight system - The scene is illuminated by a controlled beam and synchronised to a photo-detector which then can record the time elapsed between the emission of the light and the arrival at the detector [12].

Indirect ToF

The indirect measurement of the ToF (iToF) is illustrated for both pulsed and modulated light in Fig. 1.1.2 (b.1) and (b.2) respectively. The figure shows an example of measurements with four time-gating windows equal to half of the period of the incident light pulse T_p or sine wave $1/(2f_m)$. The photo-detector is reset and activated in the four different widows during which the amount of charged produced by the incident light is measured and stored as $C_{1...4}$. From these four values, both the time and phase shift can be measured respectively as:

$$\tau_{ToF} = \frac{T_p}{2}\left(1 + \frac{C_3 - C_1}{C_2 - C_4}\right) \tag{1.1.1}$$

$$\varphi_{ToF} = \arctan\left(\frac{C_3 - C_1}{C_4 - C_2}\right) \tag{1.1.2}$$

In order to extract a correct time-of-flight, it is essential that the shape of the received light, i.e. its timing profile, is preserved by the detection process since the calculations rely on it. Later

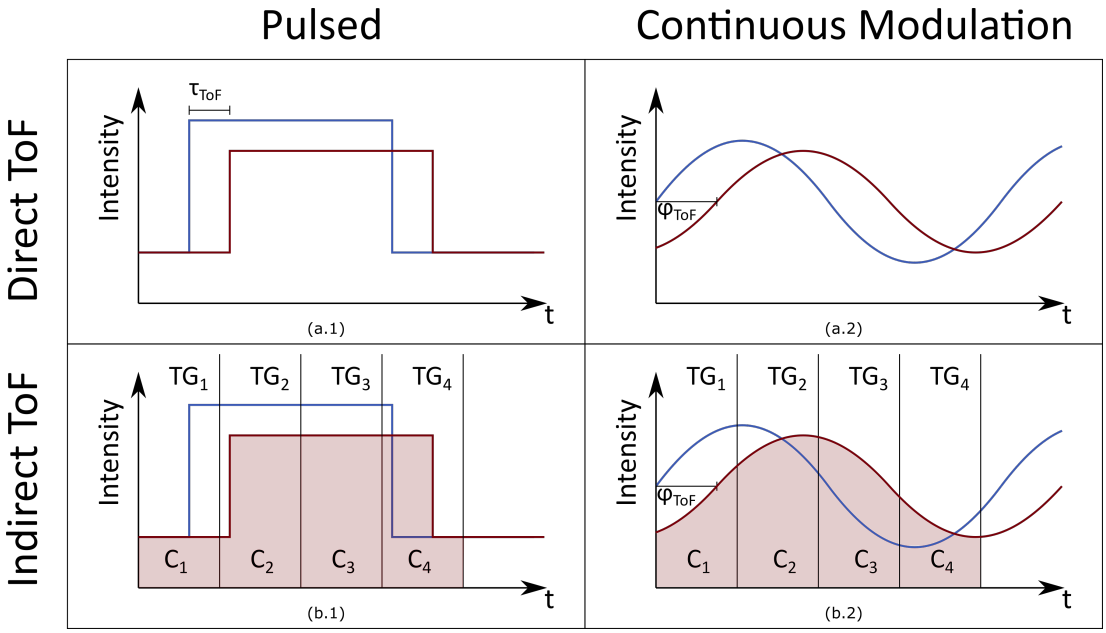


Figure 1.1.2: Time-of-Flight - The measure of the time-of-flight can be done by the direct measurement of the time shift of a received pulsed light (a.1) or from the phase shift of a modulated sine wave (a.2). Alternatively, indirect measurements can be performed by measuring in four time windows the resulting charge at the photo-detector due to the pulsed (b.1) or modulated (b.2) illumination [12].

in this research it will be shown how such assumption translates in strong requirements for the photo-detector and readout system.

Direct ToF

The direct time-of-flight (dToF) measurement, in opposition to iToF, does not measure a charge but it is done directly in the time domain. Instead of measuring a charge in the analogue (amplitude) domain as for iToF, the sensor gives an information in either the time or the phase domain (which can be considered the same). This method does not require time gating and relies only on the detection of a particular event which is typically the crossing of a set threshold by the received signal. This method does not depend on the shape of the received pulse as long as the threshold is exceeded by the signal. This means that direct ToF, especially in pulsed illumination, is less affected by signal distortions (see later sections).

Due to the increasing timing performance of modern photo-detectors, described in Chapter 2, and the relaxed requirements on power consumption, direct ToF based on pulsed illumination is going to be assumed throughout the research.

1.1.2 Positron emission tomography (PET)

Positron emission tomography (PET) is an imaging technique used in medical application for the monitoring of many dynamic metabolic processes in living organisms. Specifically, one of the most popular usage of this technique is cancer diagnosis and staging [13]. Patients affected by tumours must undergo drug treatments in order to contrast the affected cells. This process is not always successful because of the reaction from the body to the different drugs. It is therefore necessary to understand when the drugs are not producing the desired changes in the metabolism in order to better address the drug delivery. The medical information about the affected tissues can be extracted by the observation of physical and chemical processes that happen within the body when stimulated by external agents. For this purpose, a variety of chemicals, called *PET tracers*, are used to be inserted in the tissues and study, with PET scanning, their reaction. The data analysis enables the understanding of the clinical status of the analysed tissue.

A typical PET scanner is pictured in Fig. 1.1.3. The delivered tracer gets into the metabolism of the desired tissue and it is trapped into it until it decays: its nuclide is radioactive and therefore it quickly emits a positron, a process known as *beta-decay*. Almost immediately, such emission is followed by an annihilation process: the positron interacts with an electron, its anti-particle, and therefore the two *annihilate* each other with the liberation of the equivalent energy emitted as two gamma photons. The propagation of these photons, for the law of conservation of momentum, occurs simultaneously in two opposite directions. The detection of such pair of photons and the rate of such events provide the desired information about the metabolism of the examined tissue. The acquired data are collected and presented in *time-activity curves* that are fitted with known models to extract the relevant parameters. An example of such curve is presented in Fig. 1.1.4 reported from the work by Graham [14].

PET scanning requirements

A photo-detector needs to capture all the occurring events. Usually, due to the high energy of the emitted photons, scintillators are employed to "convert" the gamma rays to lower energy photons that can be detected by standard detectors. The requirements for the detectors are:

- **Timing resolution:** in order to discriminate random detections from the true-events, a

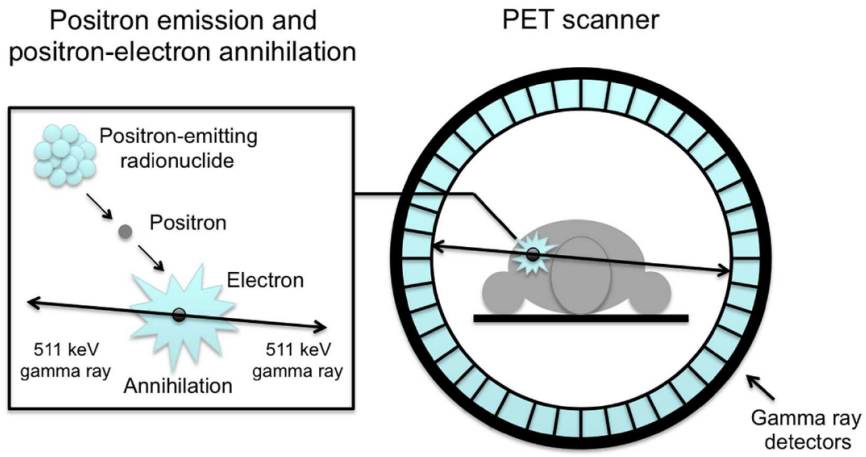


Figure 1.1.3: PET system - A picture of a general PET scanner. The emission of a positron by a radioactive nuclide is followed by the generation of two high energy photons scattered simultaneously in two opposite directions. A ring of detectors captures the photon pair giving both timing and spatial information on the occurred event [13].

very low coincidence time has to be resolved.

- **High dynamic range:** the time activity curves are often described by an intense burst of events at the start of the process, followed by a decay. It is essential for the acquired data not to be distorted by detection artefacts to guarantee the success fit to the known models.
- **Fast operation:** the PET scanning has to occur in a reasonably short time to avoid complications and discomfort to the patients.
- **Compactness:** reduced dimensions and cost of the scanners would improve the popularity of the technique.

1.1.3 Fluorescence lifetime imaging microscopy (FLIM)

A vastly used technique in biology to obtain information about the transfer of energy among cells is the fluorescence lifetime imaging microscopy (FLIM). The physical principle exploited in this technique is based on the fluorescent luminescence of an excited sample. The excitation of atoms can be induced by an external source which in the case of FLIM is often a laser with certain wavelength and power. The sample, once excited, falls back into its lower-energy state

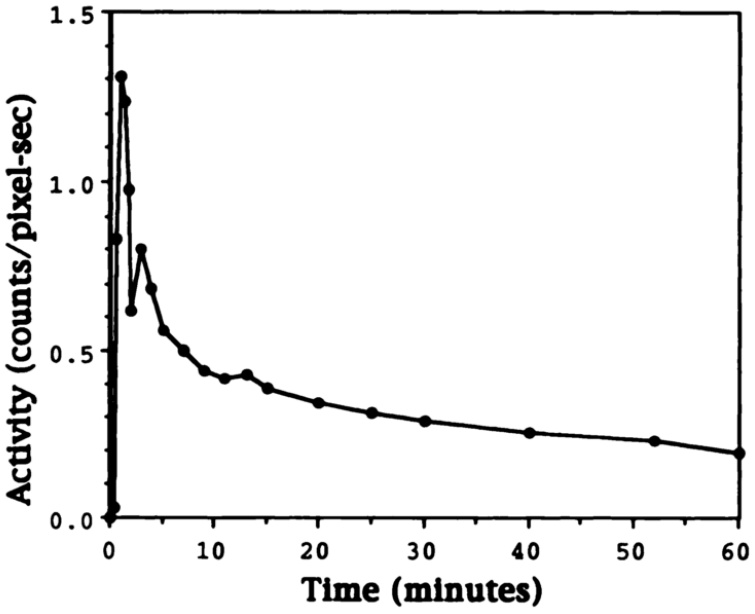


Figure 1.1.4: Time-activity curve - and example of the obtained data from a scanned plasma after PET tracer injection [14].

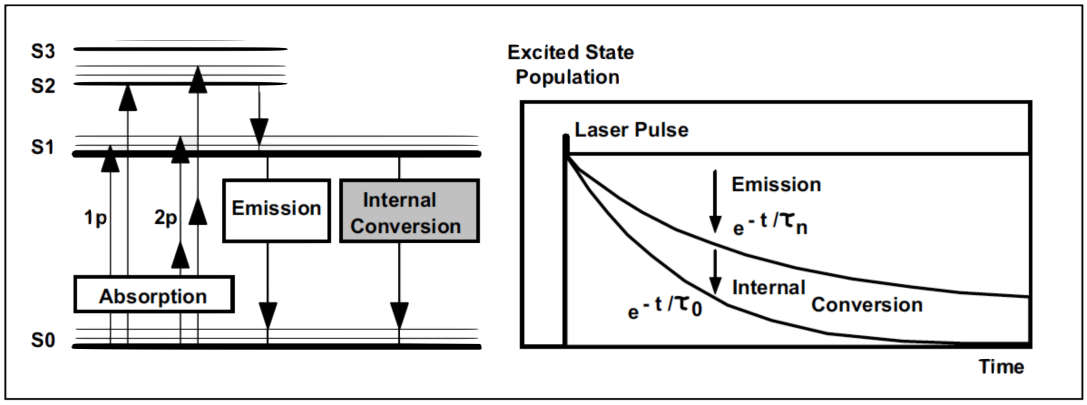


Figure 1.1.5: Fluorescence lifetime - Excited atoms emits fluorescence photon when relaxing back into their low energy configurations. The average timing for this process is in the order of few nanoseconds. The intensity of the emitted fluorescence light decays exponentially as shown. Diagram taken from [15].

by emitting photons of the corresponding energy. The relaxation process is not instantaneous. The average time between the excitation and the relaxation is called *fluorescence lifetime* and provides information on the sample itself. Typically, fluorescence lifetimes are in the orders of a few nanoseconds. Fig. 1.1.5 illustrates the principle of the excitation, left side, and the exponential decay of the intensity of the emitted fluorescent light from the relaxing atoms. The challenge for the sensor is to provide the required dynamic range in order to correctly capture both the intense burst at the start of the process as well as the exponential tail characterised by a low rate of emitted photons. The requirements for FLIM can be considered equivalent to the ones listed for PET expect for the dimension of the sensor given the fact that FLIM typically deals with small samples rather than large object such as human bodies, animals etc.

1.1.4 3D imaging: LiDAR

One of the most interesting physical properties of photons is their speed. The electromagnetic waves associated with the photons travel with a well-known speed which in the vacuum is given by (1.0.1). When a scene is illuminated by the transmitter part of the sensor, photons travel towards objects which can absorb, diffuse or reflect the incident light. The amount of reflected and diffused light travels back to the sensor where it is collected by the receiver optics and focused onto the detector. By measuring the amount of time that light has take to travel from the transmitter to the target and back to the receiver, the sensor can obtain a depth information,

i.e. the distance between the sensor and the object:

$$D = \frac{c \cdot \Delta t}{2} \tag{1.1.3}$$

A general scheme of a LiDAR system is pictured in Fig. 1.1.6 and it is composed of:

1. Controlled laser source
2. Collimator to reduce the beam divergence
3. Target
4. Receiver lens
5. Filter
6. Photodetector
7. Readout and control unit

The laser source emits a beam of light of a certain optical power which is collimated by the transmitter lens in order to reduce the divergence of the beam down to:

$$\theta_B \sim 1 \text{ mrad} \tag{1.1.4}$$

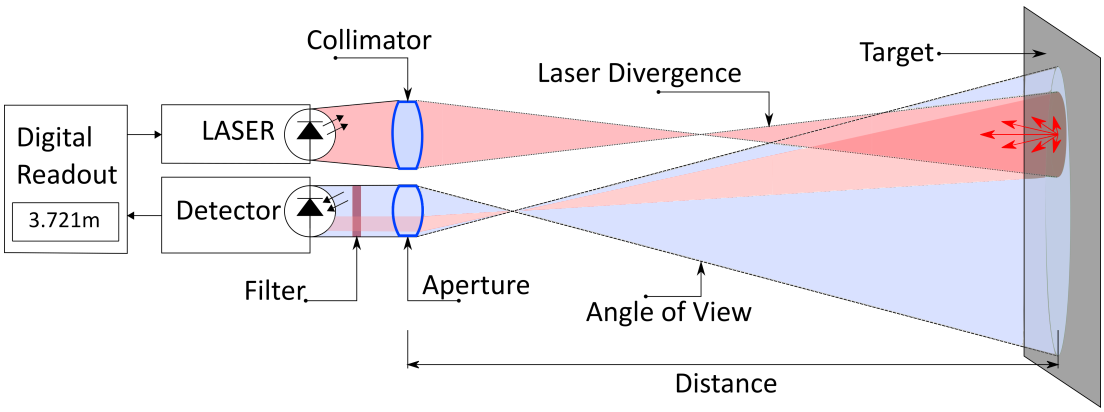


Figure 1.1.6: LiDAR system - A collimated beam of light is sent to the scene. A target diffuses and reflects a fraction of the incident light back onto the sensor. The receiver lens collects the diffused light and concentrates it onto the sensor with additional filtering. A control unit manages the timing sync between transmitter and receiver so that the time elapsed by the illumination to the collection of light can be recorded and the distance extracted.

Ideally, the transmission of the laser beam occurs with 100% efficiency, i.e. no loss due to the absorption or reflections of photons in the medium of propagation. The beam reaches a target which is here characterised by its property of reflecting photons (absorption is not relevant in LiDAR). The reflection can occur in two ways:

- **Specular reflection:** the light is sent back with an angle of reflection equal to the incident angle. For normal incidence, the beam is sent on the same axis, but in opposite direction. No (or minimal) scattering occurs. This type of reflection is observed in mirrors.
- **Diffusive reflection:** it occurs on different levels due to the microscopic non-uniformities of the material resulting in the light being reflected in every directions. This is highly dependant on the property of the material. A useful model is the **Lambertian reflectivity** described by *Lambert's cosine law*. The intensity of the reflected light is constant per unit of solid angle $d\Omega$, which means that the for two different observers, one on the normal line of the surface and one at a given angle θ_{obs} will perceive different intensities:

$$\Im_{obs} d\Omega_{obs} dA_{obs} = \Im d\Omega dA \Rightarrow \tag{1.1.5}$$

$$\Rightarrow \Im_{obs}(\theta_{obs}) = \frac{\Im \cos(\theta_{obs}) d\Omega dA}{\cos(\theta_{obs}) d\Omega_{obs} dA_{obs}} \tag{1.1.6}$$

  For LiDAR system, the angle θ_{obs} is small enough to neglect the cosine term and therefore the intensity of the light at the observer can be estimated by considering the power of the incident light, calculated as the product of the light intensity $\Im$ and the area of the target A_{target}, spread over a spheric surface of diameter equal to the distance d between the reflecting target and the observer. Therefore the relation between incident and observed light intensity can be written as:

$$\Im_{obs} \simeq \frac{\Im \times A_{target}}{4\pi d^2} \tag{1.1.7}$$

The main difference between a specular reflector and a diffusive (or Lambertian) reflector is that the perceived intensity in the former case does not depend on the distance between the observer and the reflector while in the latter case the intensity is inversely proportional to the power of two of the distance. Moreover, the relation expressed by (1.1.7) needs to take into account

the fact the different materials diffuse only a fraction of the incident light. Such property is described by the **reflectivity** η which can be written in percentages from 0 to 100% or fractions from 0 to 1. Note that the reflectivity is a property defined for purely diffusive materials and a hybrid reflecting behaviour can occur and be modelled by thinking of a diffusive target with $\eta > 100\%$ (i.e. with a component of specular reflection). Moreover, since the diffusion depends on the structure of the material, this is also dependant on the incident wavelength:

$$\eta = \eta(\lambda)$$

The receiver consists of a photo-detector and an optical system. The latter is necessary to collect the light and direct it onto the sensor. The two important parameters of the receiver lens are: the aperture, i.e. the physical size of the lens, and the focal length. The **aperture** sets the number of photons, or the power of the received beam, that is directed onto the sensor and is given by the product of the reflected power density $\Im_{obs}$ and the area of the aperture A_{lens}:

$$P_{opt,r} = \Im_{obs} \times A_{lens} = \Im_{obs} \times \pi \frac{D_{lens}^2}{4} \tag{1.1.8}$$

The **focal length** f of the receiver lens sets the *angle of view* θ_R, i.e. the portion of the scene that the sensor is pointing at. This depends on both the focal length of the receiver lens and the size of the detector. Calling L_x and L_y the size of the detector in the horizontal and vertical directions, the angle of view in the corresponding direction $\theta_{R,x}$ and $\theta_{R,y}$ are given by:

$$\theta_{R,x/y} = 2 \arctan\left(\frac{L_{x/y}/2}{f}\right) \tag{1.1.9}$$

At a given distance d, the area of the target that the sensor is looking at depends on the angle

of view and can be written, assuming a rectangular shape (as most sensors are), as:

$$A_{target}(d) = a \times b \tag{1.1.10}$$

$$a = 2 \times d \tan\left(\frac{\theta_{R,x}}{2}\right) \tag{1.1.11}$$

$$b = 2 \times d \tan\left(\frac{\theta_{R,y}}{2}\right) \tag{1.1.12}$$

Combining (1.1.7) and (1.1.10) onto (1.1.8) allows the calculation of the collected optical power at the sensor. Assuming monochromatic light, the photon rate incident at the aperture, hence onto the sensor, is calculated by dividing the optical power by the energy of the single photon (1.0.2).

As shown in Fig. 1.1.6, additional filters can be placed at the receiver. First of all, bandpass filters are used to prevent the detection of wavelength outside the range of interest, which is typically set by the wavelength of the laser. Filters are characterised by a central λ and a bandpass $\Delta\lambda$. The use of this type of filters is essential when LiDAR systems operate with ambient light which has intense components in the spectrum outside the region of interest and therefore would contribute to the optical budget as *incoherent noise*. Ambient light, natural or artificial, is described by its spectral irradiance $\mathfrak{F}$ measured in $\left[\,\mathrm{W/m^2/nm}\right]$ from which the intensity is calculated as:

$$\mathfrak{I}_{amb} = \int_{\lambda-\Delta\lambda/2}^{\lambda+\Delta\lambda/2} \mathfrak{F}(\lambda')d\lambda' \tag{1.1.13}$$

where λ and $\Delta\lambda$ come from the bandpass filter specification.

An example of obtained data in direct pulsed LiDAR experiments is included in Fig. 1.1.7. A noise floor is typically recorded due to both incoherent incident light or dark events within the photo-detector. The reflection of the laser pulse can seen as a peak in the recorded counts at the time-stamp corresponding to the time-of-flight hence the distance of the target. Digital algorithms are in charge of detecting the peak and providing the depth measure.

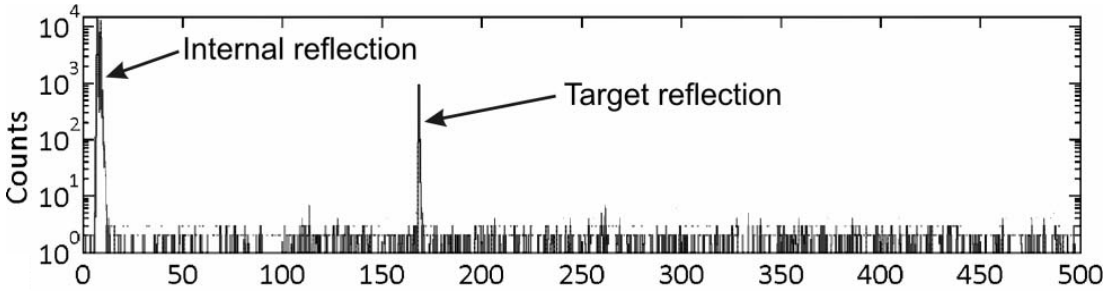


Figure 1.1.7: LiDAR measurement example - An obtained histogram from a LiDAR experiment in time correlated single photon counting mode using pulsed laser light with a reflective target at a given distance. Results taken from [16].

LiDAR requirements

The challenges in LiDAR systems are represented by the following requirements:

- **Laser power**: it can be restricted to power consumption specifications or safety.
- **Target reflectivity**: some materials exhibit reflectivity values as low as 5% or less.
- **Aperture**: in compact systems, the amount of collected photons is restricted by the usage of small aperture lenses.
- **Angle of View**: when the scene is scanned or imaged, a small angle of view is required in order to obtain sufficient spatial resolution in the X-Y plan of the target. This sets a severe restriction to the focal length which gets longer for smaller angles.
- **Single photon efficiency**: the potential low photon rate at the sensor due to the restriction presented above, the detector is required to be sensitive to single photons.
- **Depth resolution**: in order to discriminate small distances in depth, a high timing resolution is required to the sensor.
- **Ambient light rejection**: incoherent light must not interfere with the measurement to prevent false detections of objects.
- **Fast operation**: consumer applications require fast acquisition timing.

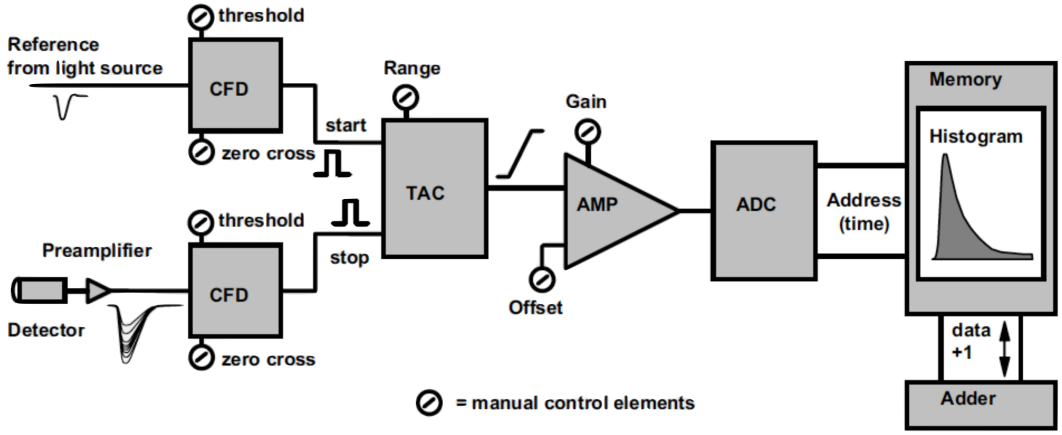


Figure 1.2.1: Classic TSPC setup - A time correlated single photon counting system is pictured highlighting its components: a laser source, providing time-controller illumination, a receiver with single-photon sensitivity, and a readout chain consisting of an amplifier, a comparator and a digitiser. The data is collected over many acquisitions and therefore a memory is attached to integrate the results and build up a TCSPC histogram. Original diagram taken from [15].

1.2 Time correlated single-photon counting

Time Correlated Single-Photon Counting (TCSPC) is one of the most powerful techniques in imaging because it ideally provides timing and counting information on all the single-photon events incident on the sensor. A useful diagram showing the typical components of a classic TCSPC system is shown in Fig. 1.2.1. First of all, the challenge in TCSPC consists of the necessity of detectors to be sensitive to single photons, and therefore not every photo-detector is suitable for such technique. For example, standard PIN photo-diodes and linear avalanche photo-diodes (APDs) are limited by their low internal gain preventing the extraction of single-photon information. Historically, photomultiplier tubes (PMTs) have been employed to demonstrate the TCSPC method and nowadays they are being replaced by solid-state counterparts, such as single-photon avalanche diodes (SPADs) as examined in Chapter 2.

Classic TCSPC has been demonstrated in applications where the incident rate of photons is low enough to have, on average, a number of photons per excitation period *significantly* below one. The idea underneath this operation is that by iterating the time stamping and counting of the rare single-photons, a histogram can be built so that, at the end of the acquisition time, it resembles

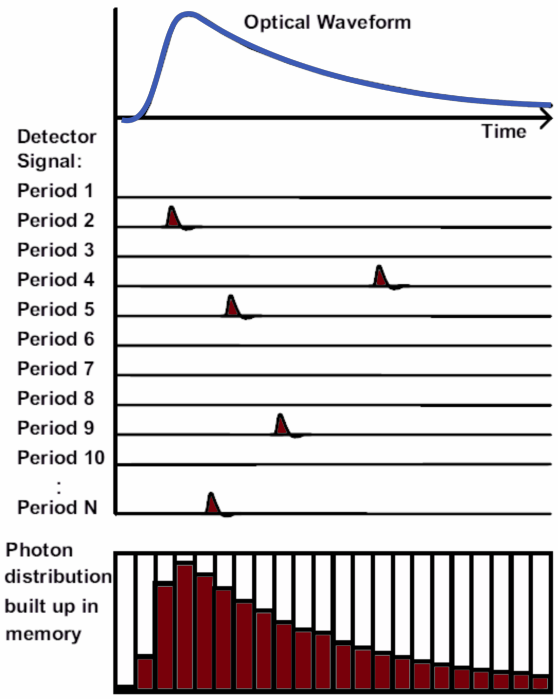


Figure 1.2.2: TCSPC - The acquisition of the optical waveform is performed over multiple number of excitations and a histogram is built from single-photon detections [15].

the time-profile of the incident photon flux, i.e. the intensity of the light, see Fig. 1.2.2, [15]. With this constraint on the incident light level, typically satisfied thanks to the usage of neutral density filters, the requirements on the sensor (including readout) is to detect only one photon, avoiding therefore high dynamic range requirements. Given this operation principle, TCSPC has been initially proposed only for low light level applications with high repetition rates so that the required acquisition time to build the histogram would be acceptable.

However, the requirement on the incident photon rate does not come from the principle itself, which is based on the single-photon sensitivity only, but it comes from the requirement of avoiding non-linear distortions of the shape of the re-constructed optical waveform occurring during the detection process. The cause of such distortions is due to the limited detection and conversion rate: the detection of a photon prevents the record of subsequent photons, therefore lost. This phenomenon is generally called *pile-up*.

1.2.1 Pile-up distortions

An extensive analysis of pile-up distortions can be found in the work by Arlt *et al.* [18] who define pile-up and describe its sources in a TCSPC system. Non-linear distortions are caused by the loss of detected photons in the receiver sensor which make the resulting histogram to be no more proportional to the intensity of the optical waveform. In initial TCSPC systems, this was mainly due to the inability of the electronics to convert the detected photons into time-stamps quickly enough not to lose subsequent events with the result of the detection of only the first photons missing the information of the rest of the events. Typical readout electronics in fact used to detect only one event per excitation period forcing the TCSPC system to work up to a maximum incident photon rate equal to $1 - 20\%$ of the repetition rate of the illumination. When this requirement is not matched, then the probability of having more than a single photon incident on the sensor per excitation pulse becomes high enough to make the loss of photons significant during the creating of the built-up histogram effectively distorting the shape of the curve. In the case of time-activity curves in PET and fluorescence decays in FLIM, exponential curves are distorted by the non-linear loss of photons by altering the timing constant relevant for the measure. Many works have been presented to correct the obtained data through data processing in order to extract the real timing decay from the distorted curves [29–32]. In order to visualise an example of what pile-up looks like for exponential curves Fig. 1.2.3 reports obtained data resulting from low-to-high intensity of the received signal: the exhibited exponential decay timing constant is highly affected by the intensity and therefore, if the condition of operations are not controlled, the extraction can be erroneous.

The loss of photon events is connected to the specification of the system called **dead time**, i.e. the minimal inter-arrival time of events for the sensor to be able to successfully produce two (or more) time-stamps. An ideal TCSPC system would have therefore “zero dead time” and would be able to detect *every single photon*. However, real-life sensors are characterised by dead times which sets the requirements on the maximum incident photon rate on the detector. The concept of dead time is often paired with the *throughput*.

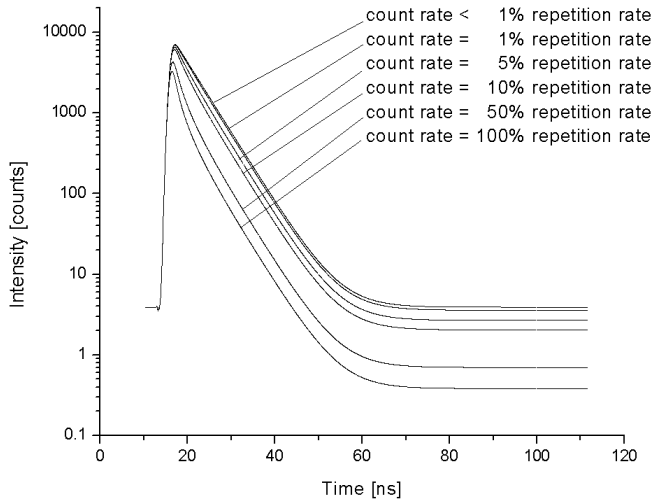


Figure 1.2.3: Pile-up example - The simulated data of three exponential decays are plotted together with the expected / corrected models for pile-up correction. Depending on the *intensity* of pile-up, i.e. based on the event rates, some reconstructed curves do not match the data due to uncorrected pile-up [17].

1.2.2 Sensor throughput

The throughput of a sensor is defined as its maximum event rate. This concept is equivalent to the dead time and therefore can be calculated as its reciprocal and it is often expressed in count rate or events per second. The overall sensor throughput, just as the dead time, depends on the contribution of each of its components, mainly the detector and the readout system. The sensor maximum convertion rate and the incident photon rate defines the amount of missed events per excitation and therefore it is the main contribution to pile-up. Arlt *et al.* have classified the source of pile-up identifying the contribution to the overall sensor throughput from each component of a TCSPC system, see Fig. 1.2.4. Initial realisations of such receivers were dominated by the timing pile-up, i.e. the loss of photons due to limited readout throughput. This, as it will be analysed in the next chapter, has been the consequence of the popular design choice to enhance the timing accuracy of the readout at the expenses of throughput. This has been the main reason for classic TCSPC to the used only in very low incident photon rate applications. The work from Arlt *et al.* has however opened a variety of options to overcome such limitations by optimising the overall throughput of each component of the sensor. This

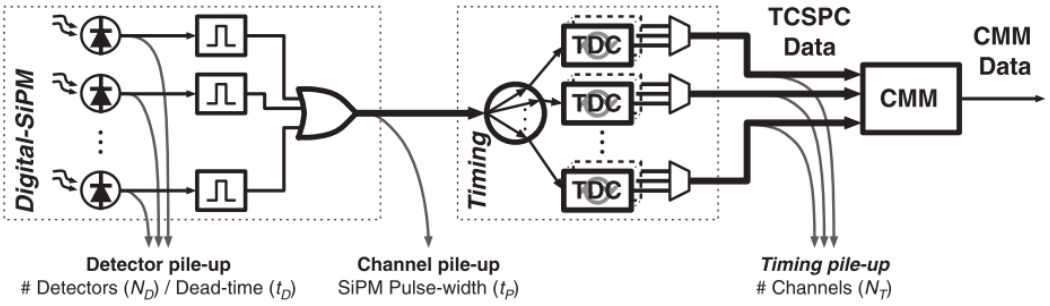


Figure 1.2.4: Sensor throughput and pile-up sources - Each component of a TCSPC system is characterised by a maximum event rate at its output. This is referred to as throughput. Limited throughput causes potential non-linear photon loss at any stage of the sensor. This diagram presents the principal causes of pile-up in TCSPC systems [18].

research takes its starting point from this result.

1.3 Research aims

The object of this research is to demonstrate how high throughput sensors are able to minimise, or cancel when optimised, the distortions due to pile-up without the needs of reducing the incident photon rate to the classic TCSPC requirement. The attention is focused on the optimisation of the throughput of digital silicon photomultipliers (dSiPMs) TCSPC systems. The research aims at the demonstration of a new concept of TCSPC for which a high input event rate can be successfully converted without the need of high illumination rates. The strategy of the research can be summarised in:

1. **Readout throughput -** A high throughput time-to-digital converter demonstrates an improvement of a factor of 10 compared to published architectures overcoming the limitation due to timing pile-up.
2. **Single channel analysis -** A theoretical analysis is run in order to determine the maximum performance achievable by a single-channel TCSPC system for a given technology node. The obtained information is essential to minimise the channel pile-up.
3. **Optimisation of detector array -** An intense analysis on the detector array is presented and matched with test data providing design rules of thumb on how to optimise the architecture of a detector array in order to avoid detector pile-up.

The benefits of overcoming such limitations are:

- Improvement of optical waveform reconstruction - ideal for PET scanning, indirect ToF measurements and low SNR LiDAR systems.
- Decrease of acquisition times due to optimised exploitation of incident photons - required for fast LiDAR systems and PET scanners
- Better performance with lower optical powers by removing the necessity to decrease the laser power to the desired level

A particular attention is given to *single-channel* TCSPC systems which, if optimised, can potentially realise more compact and contained power and cost solutions compared to multi-channel counterparts.

1.4 Contribution to knowledge

The research on digital SiPMs for throughput optimisation has produces the following results:

- **Multi-event TDC architecture -** A 14 GSample/s time-to-digital converter with effective zero conversion dead time has been designed in a 130 nm CMOS Imaging technology in collaboration with ST Microelectronics, funder of the entire project.
- **Monte-Carlo dSiPM simulator -** A high-level, fully configurable dSiPM simulator has been developed in order to characterize, analyse and predict the performance of dSiPMs under a variety of design parameter values. Both timing and counting performances can be predicted by the analysis of the simulator output.
- **Configurable dSiPM chip -** A test chip has been designed and manufactured in the same technology as the TDC chip. The extended reconfigurability of the chip has enabled the validation of the Monte-Carlo simulator.
- **Analytical Models -** An intense analysis of both simulated and collected data has been performed in order to obtain analytical equations describing the detection process in dSiPM sensors. Equations have been presented in both time and frequency (count rate) domain. All the obtained results from the theoretical analysis have been fully validated and backed up with experimental data collected with the usage of the designed test chips.

- **Design flow -** The overall results of the research has been used to propose guidelines on the design of the next generations of dSiPM sensors.

1.5 Thesis structure

The research background, structure and results are presented in this document and divided in the following chapters:

Chapter 1 - An introduction on the applications of single photon detection is provided highlighting the requirements for future sensors. A general overview on the typical challenges in the field are introduced as well as the strategy of the presented research.

Chapter 2 - A survey on current solutions for TCSPC sensor is provided. The variety of available detectors is analysed in terms of requirements and performance focusing on the achieved throughput to keep the review useful for the research purposes. Popular readout architectures have been included in the survey highlighting the need of throughput optimised both readout and detectors. The attention is directed to the choice of single-channel digital SiPMs with flash TDC architecture.

Chapter 3 - The theoretical analysis is presented. The chapter opens with a general overview on available models for SPAD and dSiPM sensors and introducing the level of modelling chosen for the research to match the need of a throughput analysis on an architecture level. The chapter then is divided into two main parts: the description and evaluation of the Monte-Carlo simulator, including detailed algorithms and simulation routines; and the analytical model presented as a consistent set of equations coming from the analysis of the simulated data and later compared to recorded data. Timing and counting performance have been described. The simulator has moreover enabled the ability to simulate pile-up distortions at different levels of the detecting chain.

Chapter 4 - The two manufactured test chips are presented. The first one presented is the demonstration of the high throughput TDC, born as a collaboration with fellow designers and therefore the attention is focused specifically on the core of the TDC design as essential part of the research topic. The design description of the reconfigurable dSiPM chip follows next and it highlights the variety of configuration that the user can set up.

The aim is to present the options included in the test chip which were driven by the modelled results of Chapter 3.

Chapter 5 - The experimental setup is described in this chapter. The list of utilised instrumentation as well as test chip configurations is presented. Results are obtained for both chips in timing and counting domain. The high throughput and high resolution of the TDC sensor is demonstrated in a LiDAR experiment. Moreover, pile-up situations have been recreated for a direct comparison with the output from the Monte-Carlo simulator. The experiments highlight the high performance of newly proposed XOR-based dSiPM paired with a high throughput TDC.

Chapter 6 - A summary of the obtained results and achievements is given in the final chapter of this thesis. A dedicated section offers discussion points for the design of the next generations of dSiPM sensors for TCSPC. A final section on future work conclude the thesis presenting possible extensions of this research for further improvements.

Chapter 2

State of the Art of TCSPC Systems

Time correlated single photon counting (TCSPC) [15, 33] is the most desired technique for low light applications such as fluorescence lifetime imaging microscopy (FLIM) [34, 35], and positron emission tomography (PET) [36–38]. The aim of TCSPC is to reconstruct the temporal shape of the intensity of the incident light on the sensor. To make this possible, every photon has to be detected and recorded. For such reasons, single photon detection capability is strictly necessary to allow the low light levels to be recorded by the detector avoiding distortions on the reconstructed waveform. However, single photon sensors are typically limited in dynamic range and for such reason high bursts of photons are not detected properly by these sensors. The classic way to solve this bottleneck is to exploit the probabilistic processes within the detection of single photons and the ergodicity principle for which the response to a periodic stimulus can be reconstructed by sampling multiple times the response of the detector to each stimulus. In practice, the intensity of the incident light is deliberately lowered down, mostly by filtering, to a level below single photon in order to ensure that per every sample, i.e. laser cycle, the average number of detected photons is below the unity. By running the system (light emitter and detector) for a sufficient amount of time, the single detected photons will eventually recreate the shape of the incident light from the output of the sensor. At each cycle a memory is used to record the time-stamp information, i.e. the time when the photons are detected, and the total number of acquired photons. The result of this process is therefore a time histogram that, if the exposure time of the system is high enough, resembles the optical input, see Fig. 2.0.1. It is evident how this solution is best suitable for applications with a very high repetition rate so that the number of acquired samples is enough to represent a statistical relevant ensemble to reconstruct the desired output.

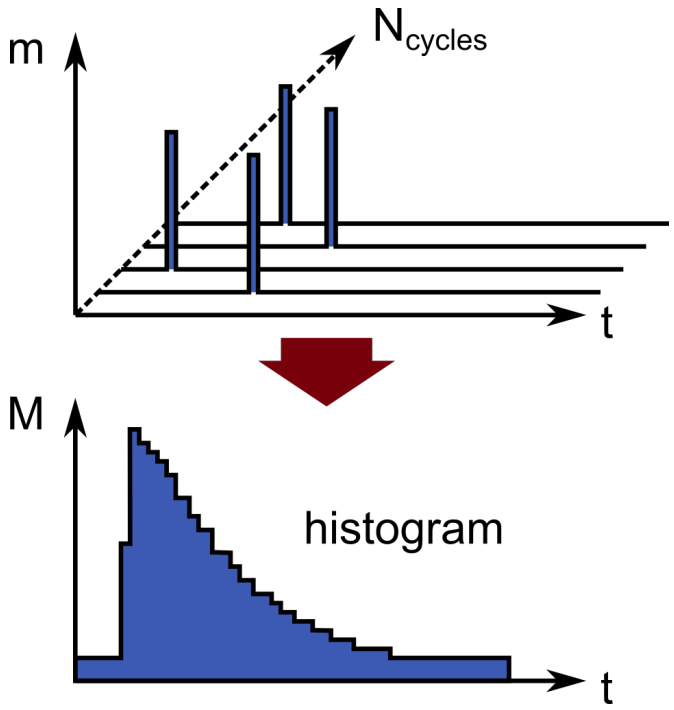


Figure 2.0.1: TCSPC principle - The optical waveform is reconstructed by acquiring single photons per each laser pulse. All the time stamped photons are counted onto a histogram that resembles the input waveform. It is essential to minimise the loss of acquired photon in order to avoid non-linearity distortions.

This chapter aims at presenting a selection of relevant works in the literature to analyse the state-of-the-art of TCSPC systems. The chapter is organised in sections dedicated to the main components of such systems: first of all, the detector, showing the evolution from photomultiplier tubes (PMTs) to solid state CMOS single photon detectors (Geiger APDs or SPADs). Then the chapter continues with an overview of time to digital converter (TDC) architectures proposed for recent TCSPC systems.

The goal of the chapter is to identify the gap in the literature highlighting the limitation of classic TDC introducing therefore the analysis and the solutions presented in the rest of this thesis work. The main attention is given to the open problem of minimisation of non-linearity distortions also called *pile-up* effect [18].

2.1 Detectors

The principle of the TCSPC technique is to detect single photons and build up a time histogram of the incident light through periodic illumination. Three main types of detectors have been employed to fulfil such task in TCSPC systems:

- Photomultipliers tubes (PMTs)
- Avalanche photo-diodes (APDs)
- Single-photon avalanche diodes (SPADs or Geiger-APDs)

It is in the interest of this section to provide an overview on the performance of such devices and to highlight the benefits and limitations in applications where the TCSPC technique is preferred.

2.1.1 Photomultiplier Tubes (PMTs)

The first choice adopted to demonstrate TCSPC was represented by the photomultiplier tubes (PMTs) [39–45]. The invention of PMTs is an object of a controversy between Kubetsky and Zworykin *et al.* who independently during the 1930s have proposed a source of primary photo-electrons and a series of multiplication stages consisting of, respectively, a photo-cathode and multiple dynode stages [46], see Fig. 2.1.1. The major success in the development of PMTs is represented by the high detection efficiency in the infra-red spectrum and low dark currents

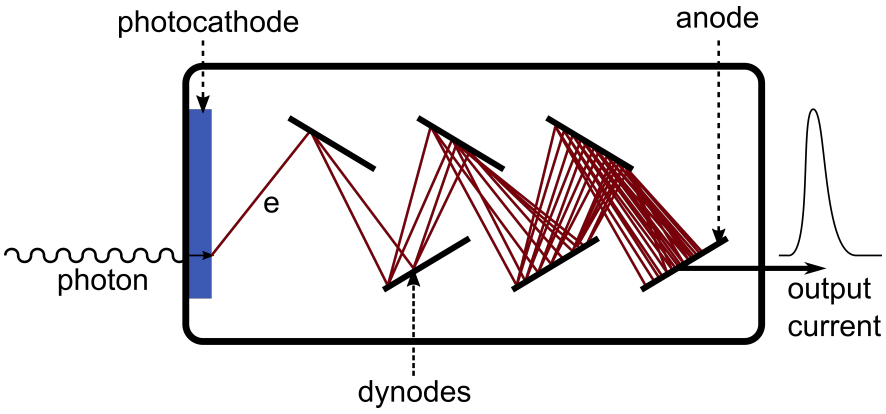


Figure 2.1.1: Photomultiplier Tubes - A PMT consists of a photo-cathode to generate primary photo-electrons which are then multiplied through a series of dynodes to generate a high current at the opposite terminal of the tube.

[40, 47–50]. PMTs represent large area detectors. The active area can reach many cm^2 and in order to reach a high internal gain they need a long dynode chain which makes the size of the entire tubes much larger than the active area itself [19]. The introduction of micro-channel plate PMTs (MCP-PMTs) [51–56] has dramatically reduced the dimensions of PMTs by replacing the multiplication stage of dynodes with ultra-thin conductive glass capillaries. Each of the capillaries is an independent multiplication stage thus realising a two-dimensional electron multiplication. The performance however does not match classic PMTs as shown in Table 2.1.

Even with micro-channel plates, the main limitation is that PMTs require high voltage ($\sim$ kV) and do not provide insensitivity to magnetic fields. Although alternatives to PMTs are constantly investigated and developed, these devices still represent the favourite option in the scientific field where cost, compactness and low voltage operation are not strict requirements. PMTs are still valid options of nuclear science and scientific applications.

Manufacturer	Product	Peak Wavelength λ_p	Sensitivity	Gain	QE at λ_p	Voltage	Dark Current
Hamamatsu [57]	R10601	420 nm	$1.9 \cdot 10^4$ A/W	$2 \cdot 10^5$	28%	1.25 kV	2 nA
Horiba [58]	PDS-3000	260 nm	$7.4 \cdot 10^5$ A/W	$1 \cdot 10^5$	25.4%	1.25 kV	3 nA
Hamamatsu [56]	R10754-07-M16	$350 - 400$ nm	$7.73 \cdot 10^4$ A/W[1]	$1 \cdot 10^6$	25%	$2 - 3.6$ kV	30 nA

1 Calculated value for comparison

Table 2.1: Commercial PMTs
A comparison among PMTs currently available on the market

2.1.2 Avalanche photo-diodes (APDs)

With the evolution of solid state devices, positive intrinsic negative (PIN) diodes [59–63] have been analysed as a possible alternative to photo-cathodes and PMTs. Although insensitive to magnetic fields and operating at lower voltages than PMTs, the main limitation of such devices consist in the lack of internal gain, i.e. a gain equal to one. For such devices, the external stage of amplification is necessary to detect low light levels, see an example in Fig. 2.1.2. However, the requirements on the noise figure of amplifiers becomes too restrictive. As shown by Donati *et al.* [19], to overcome such bottleneck, the photo-diodes have been improved, in dedicated technology, in order to have a higher internal gain. It was in the 60s that photo-diodes have been utilised in a novel way to provide the necessary boost in internal gain. The mechanism behind

this is the avalanche multiplication [64, 65]. Avalanche photo-diodes (APDs) are in fact photo-diodes in reserve bias condition below the breakdown voltage of the junction [65–70]. This allows the detection of photons to cause a so-called avalanche, i.e. an internal multiplication stage where the electron-hole pairs generated through photoelectric effect starts an impact ionisation process through the device resulting in a macroscopic current proportional to the amount of incident light on the diode itself. Because of the relation to the incident photon rate, these diodes are also called *linear-mode* APDs. The ionisation process provides an internal gain related to the ionising power of the carriers and the size of the junction. In optimal designs this can reach the order of one or few hundreds [71, 72]. APDs are nowadays manufactured in dedicated processes, even in silicon technology, and reach high quantum efficiency (QE) values that make them competitive with PMTs [73–75]. The voltage of operation of such devices does not match yet the requirements of low voltage operation, low cost and ease to aggregate into arrays, however, they are still considered as a main alternative to PMTs due to relative low dark currents and no second order effects such as cross talk. However, APDs fail to meet the single-photon detection requirement for the TCSPC technique. Nevertheless, APDs are still used with different techniques in the mentioned applications such as PET [71,76–80] and LiDAR [81,82]. A selection of commercially available silicon APDs have been compared in Table 2.2 showing

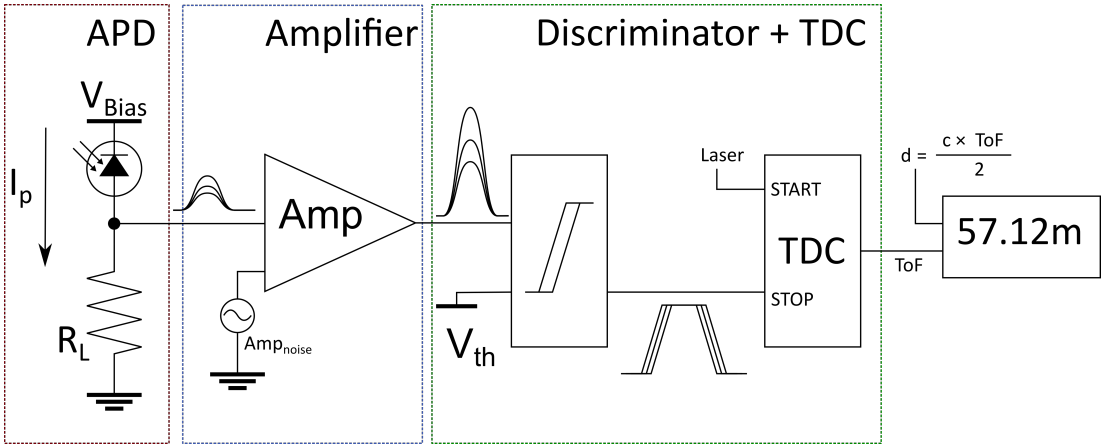


Figure 2.1.2: Avalanche photo-diodes - APDs in the linear mode of operation are used to detect the light level incident on the sensor by measuring the generated avalanche current after a stage of amplification due to the realative low internat gain. The generated carriers, due to reverse barrier, cause a series of impact ionisations resulting in a macroscopic current through the diode. This current is the amplified by an external circuit which then allows the usage of comparators for photon-counting.

Manufacturer	Product	Peak Wavelength λ_p	Sensitivity	Gain	QE at λ_p	Voltage	Dark Current
Excelitas [83]	C30737PH-230-80	800 nm	50 A/W	100	78%[1]	120 – 210 V	0.5 – 1 nA
Hamamatsu [84]	S8664	600 nm	0.24 A/W[2,3]	50	70%[3]	400 – 500 V	0.1 – 1 nA
Hamamatsu [85]	G8931-20[4]	1550 nm	0.9 A/W*	10	7%[1]	40 – 60 V	40 – 60 nA
OSI Optpel. [86]	8-150	800 nm	20 – 50 A/W	40 – 100	75%[2]	150 – 250 V	0.05 – 30 nA

[1] Calculated value for comparison
[2] For unity gain
[3] For $\lambda = 420$ nm
[4] InGaAs technology

Table 2.2: Commercial APDs
Summary of some APDs available on the market

the main characteristics that have been mentioned in the section.

2.1.3 Single Photon Avalanche Diodes (SPADs)

The limited internal gain of avalanche photo-diodes has been solved by operating the devices in the the so-called Geiger mode. Introduced by Cova *et al.* [87] explicitly for TCSPC, single-photon avalanche diodes (SPADs) are APDs operating beyond the breakdown voltage in a state where no carriers are free to contribute to a reverse current. However, due to the strong electric field, even a single-photon can start a self-sustaining avalanche creating therefore an intense reverse current, see Fig. 2.1.3(a) (1-2). The intensity of the generated avalanche photon current is so high that the concept of gain stops being useful since the current is not proportional to the incident light. Moreover, the generation of such current is very rapid in time giving the device great timing performance resulting in *timing jitter* as low as tens of picoseconds. The major difference, however, between a linear mode APD and a Geiger mode APD, is that in the latter, the avalanche needs to be quenched, i.e. the current through the diode must be stopped in order not to damage the device and to make sure more photons can be detected. This is shown in Fig. 2.1.3(a) as process (3) where the current is lowered so that the voltage is below the breakdown value therefore stopping the avalanche. Once the quenching is done, the SPAD can be brought back to its initial condition (4). Different techniques exist for the quenching process [88]. It can be often performed by a resistor R_q, see Fig. 2.1.3(b): the photo-current I_p flows through the diode and the resistor, the latter then sees a voltage drop through its nodes bringing the voltage V_{SPAD} to a much lower value below the breakdown. The avalanche is quenched. At this stage,

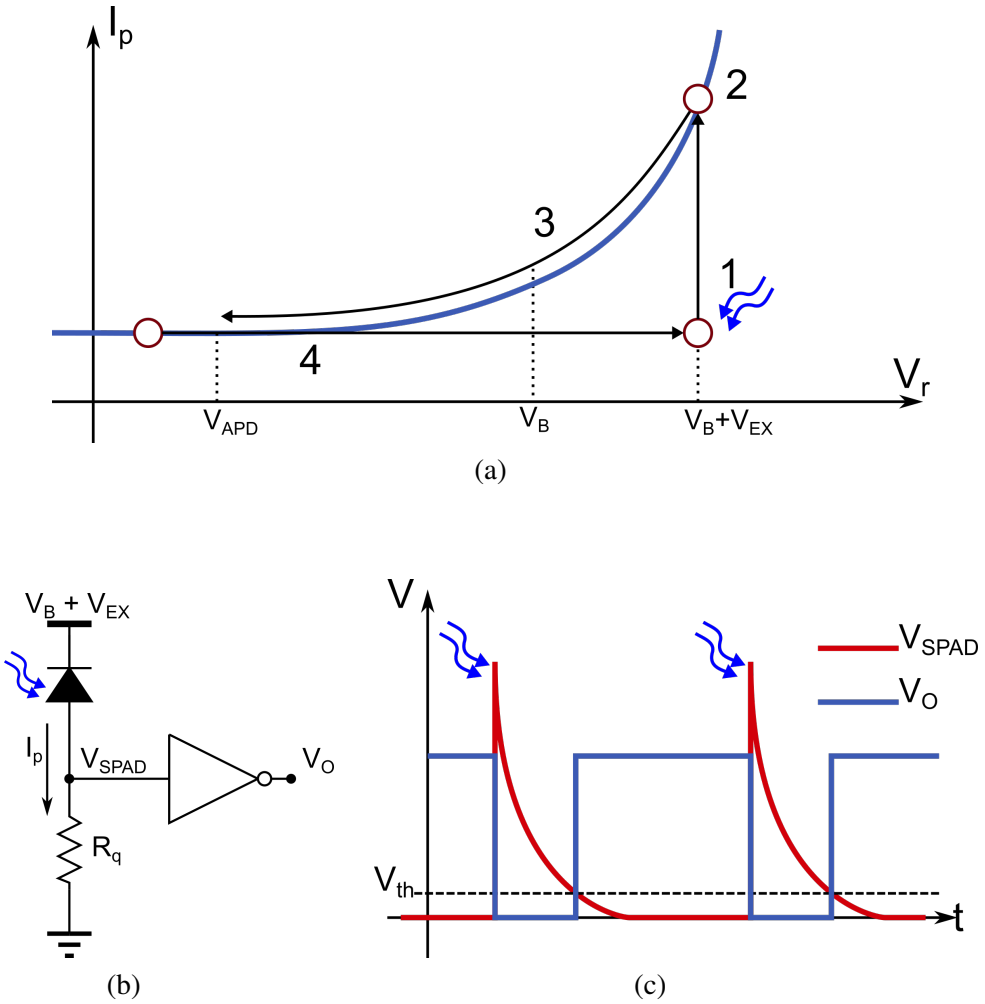


Figure 2.1.3: SPADs - An APD biased above the reverse breakdown voltage, (a.1), is able to trigger a self-sustaining avalanche when a single photon is absorbed. The generated avalanche resulting in the photon current (a.2), must be *quenched* by lowering the current through the diode itself, (a.3). Once the quenching is done, the diode can be restored in its initial state ready for the next photon-detection, (a.4). The quenching and the recharge process can be done by a quenching resistor and a load capacitance, (b). The waveforms for the SPAD note and its digitised output are presented, (c).

the node V_{SPAD} starts recharging to its initial value since no more current is flowing through the diode. This recharge process is a passive R_qC profile where the C is the load capacitance of the cell. Fig. 2.1.3(a) shows how an inverter can be used to create a pulse out of the SPAD signal essentially digitising the node V_{SPAD}, see Fig. 2.1.3(c). The necessity of SPADs to quench the avalanche has two main fundamental consequences that makes their performance intrinsic different from linear mode APDs. First of all, a SPAD cell has to include the quenching resistor and therefore the active area available for the detector is reduced. This is often expressed with a parameter called *fill-factor*[1] FF which is defined as the ratio between the active area of the SPAD cell and the total area:

$$FF = \frac{A_{active}}{A_{cell}} \tag{2.1.1}$$

This parameter is included to extend the concept of quantum efficiency by taking into account the geometrical limitations imposed by the additional required circuitry in the cell. The photon detection efficiency PDE is therefore defined as:

$$PDE = QE \times FF \times P_{avalanche} \tag{2.1.2}$$

where $P_{avalanche}$ is a property of SPADs which represents the probability of a photon triggering the avalanche process. This parameter is a function of the excess voltage V_{EX} and has been modelled by Savuskan *et al.* [89] as:

$$P_{avalanche}(V_{EX}) = 1 - \exp\left(-\frac{V_{EX}}{V_0}\right) \tag{2.1.3}$$

where V_0 is a normalisation factor. Very often the product between QE and $P_{avalanche}$ is referred as photon detection probability PDP. The geometry of the active area and the space occupied by the quenching circuitry impact on the performance of the SPAD especially in the miniaturisation process of the cell where the circuitry area becomes comparable to the active area hence intensively reducing the fill factor and the PDE. Moreover, in advanced SPAD design, other additional circuitry might be included inside the cell further reducing the PDE. This parameter is used to calculate many figures of merit in SPAD performances [27], for such reason the

1. The FF is sometimes referred as geometrical efficiency ε.

impact of the fill factor in the cell design is of fundamental importance.

Another essential parameter that describes the unique performance of a SPAD is the *dead time* or *recovery time* τ_{dead}. As previously mentioned, after the avalanche is immediately quenched, a recharge process is necessary before the excess bias and the *PDE* are sufficiently restored so that a new detection can occur. This means that the average inter-arrival time of the incident photons, defined as the reciprocal of the incident photon rate, has to be longer than the dead time in order to maintain a linear relation between detected and incident photons. This parameter therefore impacts on the maximum detection rate of a SPAD cell. As explained in detail by Eisele et al [90], the detection rate m of a SPAD cell, given an incident photon rate n on the total area must take into account the type of recharge process. Two ways can be distinguished: in the first one, described by Fig. 2.1.3(b), the recharge is done by a simple *RC* process with no additional circuit, which means that the SPAD voltage, and therefore the excess bias, varies in time exponentially during the recharge process. This is referred to as *passive recharge*. Alternatively, additional circuits can be employed so that while the recharge takes place, the SPAD voltage is *fixed* to a value below the breakdown, i.e. the SPAD is *off*, until the recharge is completed and then it is restored to its bias condition. This process is referred to as *active recharge* due to the use of active circuitry. The substantial difference between these two processes is the following: in high incident photon rates, i.e. when the probability of many photons to be incident on the SPAD within the dead time is high, a passive recharge SPAD can be triggered by many photons, although its *PDE* is reduced by the lower bias, and the recharge process has to restart again, effectively extending the dead time window. This behaviour falls into the so-called category of *paralysable detectors*. The relation between the detected photon rate m_{PQ} of such SPAD cells and its incident photon rate n is described by:

$$m_{PQ} = PDE \times n \times e^{-PDE \times n \times \tau_{dead}} \tag{2.1.4}$$

Eq. (2.1.4) allows to calculate the maximum detection rate $m_{PQ,\max}$ and the corresponding maximum input rate $n_{\max}$:

$$\frac{dm_{PQ}}{dn} = 0 \Leftrightarrow n = n_{\max} = \frac{1}{PDE \times \tau_{dead}} \Rightarrow$$
$$m_{PQ,\max} = \frac{1}{e \times \tau_{dead}} \tag{2.1.5}$$

Active recharged SPADs instead, due to the additional circuitry, cannot detect more than one photon in a dead time window regardless of the incident rate. For such reason, these devices can be considered *non-paralysable detectors* and their detected count rate m_{AQ} is given by:

$$m_{AQ} = \frac{PDE \times n}{1 + PDE \times n \times \tau_{dead}} \tag{2.1.6}$$

which gives a maximum count rate of:

$$m_{AQ,\max} = \frac{1}{\tau_{dead}} \tag{2.1.7}$$

In opposition to APDs, where the main limitation to the photon detection is a lower limit due to the inability to produce a current visible after the amplification stage, SPADs have an opposite limitation: they are capable of single photon detection but have an intrinsic upper limit due to their dead time. By plotting the two curves given by (2.1.4) and (2.1.6), normalised by the dead time, see Fig. 2.1.4, three main regions of operation can be identified. The dark count region, where the detection rate does not depend on the incident rate but is set to a minimum value called dark count rate *DCR* due to spontaneous avalanche events triggered by carriers "*generated through Shockley-Read-Hall processes, trap-assisted tunneling (TAT), or Poole-Frenkel emission*" [27,91]. The second region is the linear region of operation where the detection rate is proportional to the incident rate. The extension of this region, which can be defined as the *dynamic range DR* of the SPAD, is given by the ratio between the maximum and the minimum detected rate and can be expressed in decibels:

$$DR(dB) = 20 \cdot \log_{10} \frac{m_{\max}}{DCR} \tag{2.1.8}$$

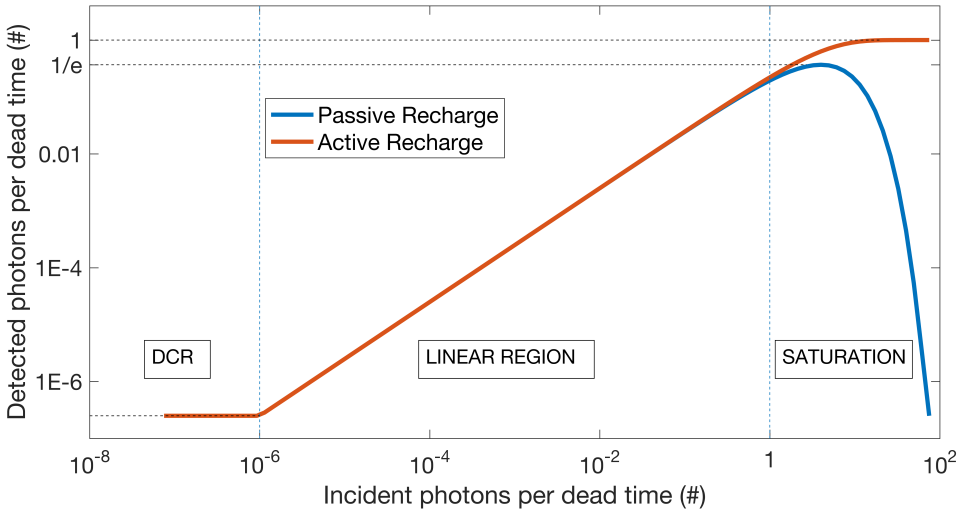


Figure 2.1.4: The region of operation of a SPAD - A SPAD can operate in three regions: a *DCR* dominated region where the response is independent on the incident rate, a linear region where the response is proportional to the incident rate through the *PDE*, and a saturation region where the registered counts either reach a maximum given by the dead time - in the case of active recharge, or drop down to zero due to the paralysis of the detector - a for passive recharge SPAD.

and therefore is greater for low *DCR* and dead times values. The third region represents the saturation of the device: for increasing number of incident photons, the rate of detected events does not increase. As mentioned earlier in the chapter, this region represents the main difference between active and passive recharge devices. In the first case, the detection rate tends asymptotically to a maximum, therefore exhibiting a non-linear behaviour; while in the second case the detection rate tends rapidly to zero due to paralysis of the device itself.

Bronzi *et al.* [27] incorporate into the definition of dynamic range a system level parameter for photon counting application which is the *integration time* (or *acquisition time*) T_{acq} and define the *DR* as the ratio between the maximum counts and the noise affecting the minimum counts, therefore:

$$DR = \frac{m_{\max} \times T_{acq}}{\sqrt{DCR \times T_{acq}}} \tag{2.1.9}$$

Another useful figure of merit for a SPAD performance is the *signal-to-noise ratio SNR* defined as:

$$SNR = \frac{m}{\sqrt{(m + DCR)}} \tag{2.1.10}$$

Work	Peak Wavelength λ_p (nm)	PDE at λ_p (%)	DCR[1] (cps)	Jitter FWHM (ps)	m_{max} (Mcps)	Area (μm^2)	Process
Gulinatti [92]	650	60	550	92	27.6	1,963	Custom
Excelitas [93]	700	70	500	350	35	2,5447	Custom
MPD [94]	550	50	2.5k	35	13	7,854	Custom
IDQ [95]	650	62	150k	400	1	196,3504	Custom
Laser Components [96]	650	73	250	800	12	7,854	Custom
Rochas [97]	470	28	900	60	12.3	38.5	800 nm CMOS
Rochas [98]	470	20	50	50	31.3	32.2	800 nm CMOS
Villa [99]	420	55	155	75	48.1	1,963	350 nm CMOS
Niclass [100]	460	36	750	80	4	78.5	350 nm CMOS
Guerrieri [101]	450	42	4.9k	39	27.6	314	350 nm CMOS
Faramarzpour [102]	470	2.5	60k	163	33.3	78.5	180 nm CMOS
Mandai [103]	600	36	5k	165	0.7	78.5	180 nm CMOS
Niclass [104]	610	64.8	1.6k	161	55.9	625	180 nm CMOS
Niclass [105]	450	34	105	144	10	78.5	130 nm CMOS
Richardson [106]	500	28	60	200	10	50.3	130 nm CMOS
Palubiak [107]	570	2	1k	163	83.3	100	130 nm CMOS
Veerappan [108]	500	27.5	160	140	10	24.6	130 nm CMOS
Webster [109]	560	25	5k	18	66	50.3	130 nm CMOS
Karami [110]	470	16	16k	398	0.6	50.3	90 nm CMOS
Henderson [111]	410	36	250	107	16.6	3.1	90 nm CMOS
Webster [112]	690	38	70	82	66.4	32.2	90 nm CMOS

[1] At room temperature.

Table 2.3: Review of recent SPAD design - A selection of works on SPADs is here reported in terms of a selection of figures of merit [27].

In order to have a good picture of the performance of recently proposed SPAD design, the work from Bronzi *et al.* comes very helpful [27]. Table 2.3 shows a selection of reviewed works highlighting some of the figures of merit introduced in this section. The comparison is very useful to see what the compromises are often made in SPAD design toward one of the FoM rather than the other ones. Some designs are specifically made to enhance the *PDE*, hence a custom process, a large area are appealing to this goal, however they often have to compromise on maximum count rate or timing jitter [93, 95]. Other designs instead focus on the integration with standard CMOS technology at various nodes, this compromises on the *PDE* but they exhibit great timing performances [97, 98, 109, 109].

It is clear how custom SPADs are used to efficiently create large area sensors. When CMOS integration is considered, the trend is almost inverse: the size of the SPAD shrinks down to few

microns squared and the maximum count rates is generally higher than the previous examples [97,98,107,110,111]. The reason behind is most likely to lie underneath the application of such devices. It is reasonable to think that large SPAD detectors are realised to directly compete with APDs and solve their limitation in low light-level applications. For such reasons, the maximum count rate is not particularly limiting since the application requires low counts due to the incident light level. And at the same time, to enhance the performance and be comparable to APDs, a high *PDE* is very appealing. However, large area sensors are not ideal for imaging systems where arrays of detectors are needed in order to resolve in space and not just in time. Although the integration into arrays has been demonstrated also for APDs [77, 113–115], imaging sensors tend to require low area detectors therefore limiting the performance of the APDs due to the consequent lower gain. For such reasons, the design of SPADs as direct competitors of APDs has been focused on the shrinking of the cell size/pitch trying to preserve, as much as the technology allows, the fill factor and therefore the sensitivity of the device. Moreover, the integration with CMOS technology gives significant benefits on the readout of the detectors. The process of reducing the SPAD size also gives additional benefits. First of all, two of the figures of merit of SPADs are related to their size. For examples, the *DCR* is highly dependant on the size of the junction in which the trapped carriers can be released and cause undesired avalanche processes. Also the recharge time of the SPAD node is highly dependant on the size of the load capacitance which again increases with the area of the device. For such reasons it is fair to conclude that the dynamic range of a SPAD increases with smaller sizes at the expenses of the sensitivity.

Due to the appealing feature of the single-photon capability, the SPAD has not been limited to a small area detector but it has always been considered as a potential replacement not only for large area APDs but ultimately also for PMTs. Since the early days, the possibility of aggregating many SPADs together to create a large area detector by exploiting the individual and independent operation of each cell has attracted the attention of designers. The next section is entirely dedicated on the analysis and review of the solutions adopted to create large detectors from SPAD cells, the so-called *silicon photomultipliers* (SiPMs).

2.1.4 Silicon Photomultipliers (SiPMs)

A silicon photomultiplier (SiPM) consists of several SPAD cells combined together to create a large detector with single-photon capability, high gain (Geiger-mode operation) and excellent timing performance inherited by their unit cell properties [116–118]. Thanks to the integration of many cells, such devices are able to reach much higher gains compared to APDs and their progress has made them comparable to the high gains of PMT devices [119–121]. Historically, the first attempt to create such detectors was by integrating a series of SPAD cells (diode and quenching resistor) in parallel on the same silicon substrate so that the bias and output node could be shared by the entire set of cells. This section therefore introduces a SiPM which is now often referred as the *analogue SiPM* (aSiPM) to distinguish them from the later introduced digital SiPM (dSiPM) that will be discussed later on in this chapter. The analogue SiPM is shown in Fig. 2.1.5: each diode has its own quenching resistor to ensure an independent operation, i.e. each cell operates as a single Geiger-mode counter, but the outputs of the cells are summed up together to generate a macroscopic current that therefore is the analogue sum of

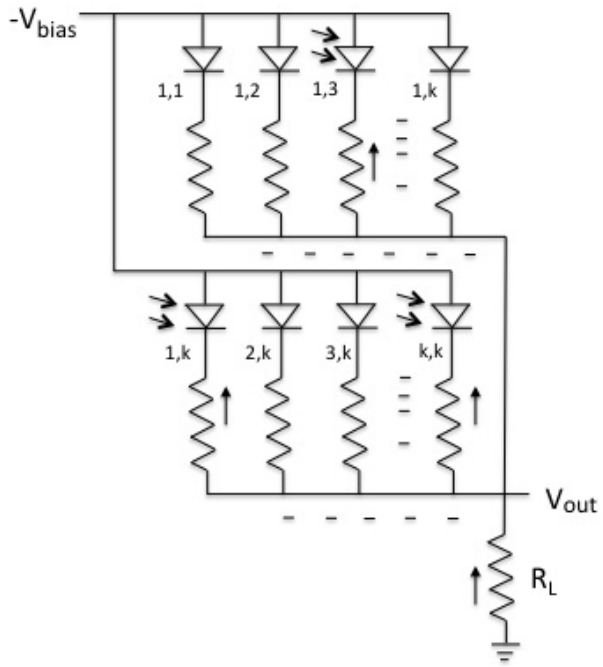


Figure 2.1.5: Silicon Photomultipliers - A set of $k \times k$ SPAD cells are connected in parallel to create an SiPM. The connection makes all the independent cells contribute to the output current/voltage of the entire detector. Image taken from [19].

all the currents in the SPAD cells. For such reasons, the concept of the gain is yet maintained for such devices, even though it is meaningless in SPAD characterisation. The gain of the SiPM can be calculated as the ratio between the generated charge and the electron charge q_e. The generated charge is expressed by the over-voltage bias V_{EX} and the capacitance C_{cell}:

$$G = \frac{C_{cell} \times V_{EX}}{q_e} \tag{2.1.11}$$

Higher gains are achieved by larger SiPM cells thanks to larger capacitance values. To directly compare the SiPM to PMTs and APDs, the concept of *PDE* of the SPAD cells is used to calculate and estimate the sensitivity, or responsivity $\mathfrak{R}$ of the SiPM together with the gain:

$$\mathfrak{R} = PDE \times q_e \times G \times \frac{\lambda}{hc} \qquad \left[\frac{\mathrm{A}}{\mathrm{W}}\right] \tag{2.1.12}$$

Thanks to the figures of merit expressed by (2.1.11) and (2.1.12), currently available SiPMs can be compared against each other and directly to commercially available PMT detectors. Table 2.4 compares a few examples of SiPMs from SensL Technologies, Hamamatsu Photonics and First Sensor. Many works in the literature have compared such devices by measuring their parameters in different conditions and also by evaluating the performance in dedicated applications. Eckert *et al.* [124] have compared SiPMs from Hamamtsu and SensL by describing an experimental setup for *PDE* analysis over a wide spectrum of incident wavelengths and have studied and provided results on the uniformity of the SiPM cells in terms of both sensitivity and gain. Puill *et al.* [125] have instead focused on the timing performance by studying the variation of the jitter over the applied voltage and temperature finding the best value in 150 ps FWHM for all the tested devices. Blosert *et al.* [126] have used analogue SiPMs as readout systems for

Manufacturer	Product	Peak Wavelength λ_p	Cell Pitch	N_{cells}	Gain	PDE at λ_p	FF	Voltage	DCR
SensL [20]	C-Series	420 nm	20 µm	10998	$1 \cdot 10^6$	31%	48%	24.2 V	300 kHz
			35 µm	4774	$3 \cdot 10^6$	41%	64%		
			50 µm	2668	$6 \cdot 10^6$	47%	72%		
Hamamatsu [122]	S13360	450 nm	50 µm	3584	$1.7 \cdot 10^6$	40%	78%	53 V	500 kHz
First Sensor [123]	SiPM-RGB-3S	550 nm	40 µm	5520	$2.7 \cdot 10^6$	32.5%	60%	27 V	< 1 MHz

Table 2.4: Commercial aSiPMs
Figures of merit for commercially available analogue SiPMs

scintillators gamma-ray detectors concluding their high potential as PMTs replacement. The most popular application for SiPMs nowadays is still represented by time-of-flight PET (ToF-PET) for their high performance in low light levels and reduced costs. Works from Powolny *et al.* [127], Seifert *et al.* [128–130], Del Guerra *et al.* [131] have all demonstrated the suitability of such devices in PET applications by optimising the readout electronics to maintain the high timing and energy resolution of the detector.

Another interesting aspect of SiPMs is their counting performance in high photon rates. This aspect differs slightly from the initial applications SiPMs were designed for, i.e. low light level conditions, but their outstanding advantages compared to both PMTs and APDs are pushing their counting performance towards higher regimes of incident photons rates. For such reasons, works by Akiba *et al.* [132] and Eraerds *et al.* [119] represent a very interesting analysis on such devices. Akiba *et al.* have demonstrated the ability of an analogue SiPM by Hamamatsu to reach count rates in the order of 1 GHz thanks to the usage of a readout amplifier and a high-speed oscilloscope by Tektronix. By illuminating the detector with a pulsed source with selectable repetition rate, the output of the detector has been analysed on the oscilloscope verifying the highest repetition rate to be correctly seen at the readout chain. The limitation beyond such repetition rate is identified in the waveform distortion. Eraerds *et al.* had previously run a similar experiment on a different SiPM by idQuantique and have demonstrate a minimum inter-arrival time of incident photons of 2.3 ns corresponding to a maximum input photon rate of 430 MHz. However, this analysis also takes into account the intensity of the pulsed source rather than focusing only on the repetition rate. In fact, it is stated that the efficiency of the SiPM decreases at higher values of laser power. These results were further investigated by Zhang *et al.* [133] who have studied the analogue SiPM as a potential receiver for digital visual light communications (VLC). They identify in the high sensitivity of the SiPM the main advantage against solid-state alternatives of PIN and avalanche photo-diodes but they have also foreseen the limitation of such receivers where ultra high data transmission rates are required. The reported maximum rate is 140 Mbit/s.

During this brief analysis and comparison against PMTs and APDs, it has emerged that the most important limitation of SPADs and SiPMs is represented by their long dead times and

the limited number of cells which compromise the performance of the detector at high light intensities. The number of cells N_{cells} (or N_{SPADs}) in which the active area of the detector is divided into is essential to understand the effect of dead times in SiPMs. As said before, PMTs and APDs do not have upper limits in high photon conditions since there is no internal saturation and the external saturation of the amplifier can be easily reduced in case of high input by lowering the amplification factor. SPADs and SiPMs however suffer of a minimum recovery time before the photons can be detected again due to the high internal gain. Reducing the internal gain of the SiPM by changing the over-voltage, therefore dealing with saturation in the same way of an APD, would potentially work but it significantly compromises the single photon capability which is the key feature of such devices hence it is not considered as an optimal solution. If n is the incident photon rate on the entire detector then the incident photon per cell is calculated as n/N_{cells}. Each cell enters its saturation region when the incident photon rate on the SiPM is n_{sat}:

$$\frac{n_{sat}}{N_{cells}} > \frac{1}{PDE \times \tau_{dead}} \Rightarrow n_{sat} > \frac{N_{cells}}{PDE \times \tau_{dead}} \tag{2.1.13}$$

In general, the number of fired cells is linked to the number of incident photons on the total active area of the SiPM by:

$$N_{fired} = N_{cells} \times \left[1 - \exp\left(-\frac{PDE \times N_{photons}}{N_{cells}}\right)\right] \tag{2.1.14}$$

As stated by Donati *et al.* [19], having a large number of cells cancels the effects of the dead time. However, the author would like to underline that the saturation level according to (2.1.13) depends also on the *PDE* and the dead time of the cells: while the number of cells increases during the process of dividing down the active area of the detector, the *PDE* is reduced due to geometrical limitations and the dead time can vary according due to the change in capacitance of each cell. The way these parameters play against each other is yet to be investigated and it is one of the goals of this work.

Once the dead time limitation has been correctly identified, a lot of efforts have been concentrated on the analysis on how to possibly solve such bottleneck even further than just moving

from SPADs to SiPMs. Obviously, the realisation of single cells with lower dead times is essential in such direction, but a very important aspect has been detected in recent works. In analogue SiPMs, the output of each cell is directly summed up to create the macro-current. With the increasing number of cells firing simultaneously, the changes on the output current are not easy to distinguish, meaning that even if a high-frequency oscilloscope might be able to resolve the identification of subsequent photons, the requirements on the readout can be very demanding. A very interesting approach to tackle this bottleneck has been introduced by Pavlov for SensL Technologies [134]. Looking at the drawings present in one of the datasheets provided by the company, see Fig. 2.1.6, an additional capacitance element can be found per SPAD cell. In this way, the SPAD node is coupled to an additional terminal which SensL refers to as *fast output*. A dedicated work by Dolinsky *et al.* [135] shows a direct comparison of these two readout modes by connecting the terminals to an amplifier and analysing the shape of the output on an oscilloscope. It is highlighted how the smaller capacitance ($\sim 10\%$ of the total C_{SiPM}) put as a load of every SPAD cells significantly reduces the recharge time of the cell therefore giving to the output pulse a much lower width. The immediate benefits

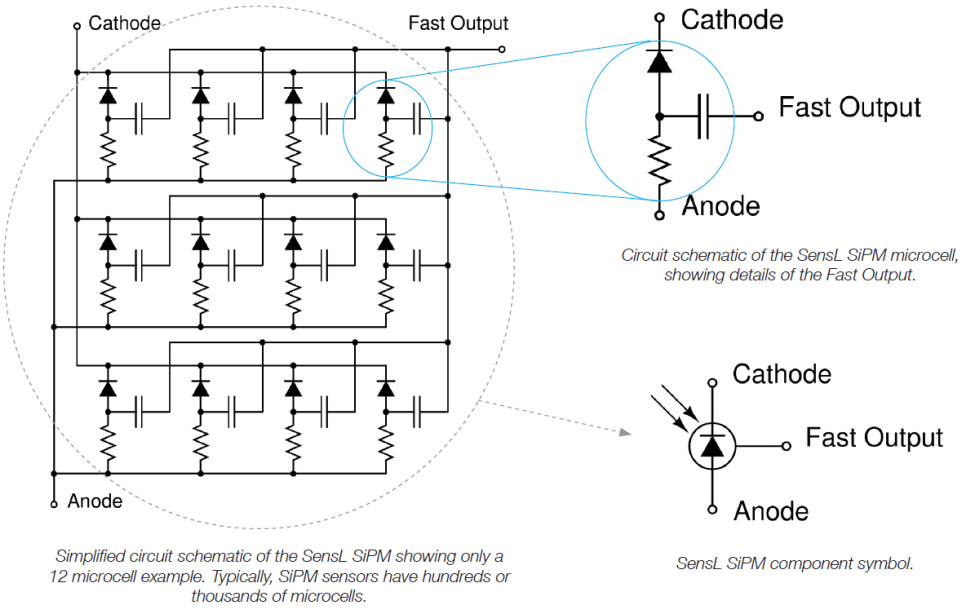


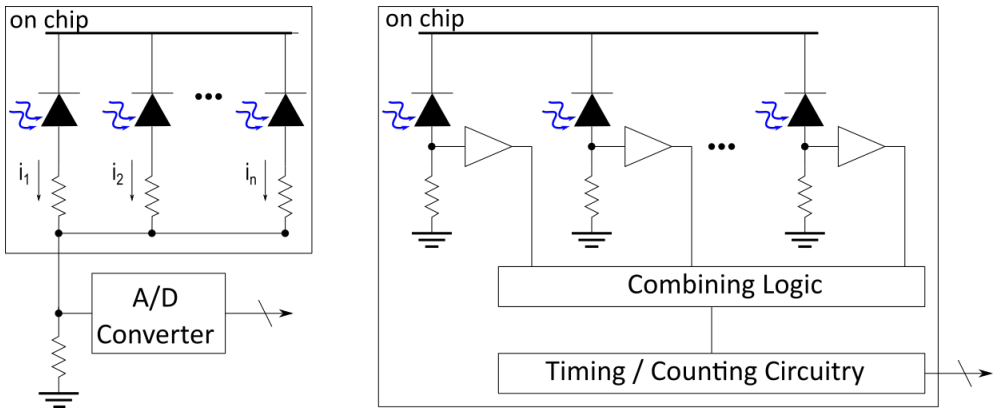
Figure 2.1.6: SensL SiPM with Fast Output - Each SPAD is coupled to a dedicated node through a capacitance. The shape of the output current from this node is essentially the derivative of the traditional output allowing narrower pulses to be produced enhancing the photon number resolution. Drawings taken from datasheet [20].

of such readout mode is to effectively remove the need of additional filters on the SiPM for timing analysis thanks to a much faster pulse. Moreover, additional benefits can be found in high photon rates: while the standard terminal registers an additional, potentially small, variation due to subsequent close in time photons, the newly introduced fast output produces two separate narrow pulses more easily distinguished. To the knowledge of the author, there are no publications showing experimental results in such conditions, however the given published works show great potential in the use of analogue SiPMs in high photon rates.

A more common way to deal with the limited counting abilities of SiPMs is found in the introduction of an alternative approach to the SPAD cells combination: the analogue sum of the SPAD currents is replaced by a digital combination in the so-called *digital silicon photomultipiers* (dSiPMs).

2.1.5 Digital SiPM

In 2009, a new concept for silicon photonmultipliers has been introduced by Frach *et al.* [21,136] for a new SPAD based detector named *digital SiPM* (dSiPM). The idea behind this invention is to overcome the intrinsic limitations of conventional SiPMs due to load capacitances by the exploitation of the digital nature of SPAD cells. As mentioned earlier in the chapter, the output of a SPAD is not a current proportional to the light intensity but a train of pulses with a minimum width set by their dead time and with a frequency proportional to the light intensity when operating outside the saturation region. By connecting all the output terminals together in an analogue fashion, the digital nature of the SPAD is lost and then the digitisation has to be re-introduced outside the SiPM to enable the timing and counting of the recorded events with the given challenges explained in the earlier section. What Frach *et al.* proposed is to enhance the digital nature of a SPAD output and to introduce the digitisation step at the cell level rather than once the (analogue) combination has taken place. In other words, instead of creating an overall current to be then digitised into a train of pulses, the already-digital SPAD output signals can be added up *digitally* to create directly the desired digital waveform to be then passed onto timing and counting circuitry. A simple diagram showing the difference between traditional and digital SiPMs is shown in Fig. 2.1.7. The output of each SPAD is connected to what is shown to be a

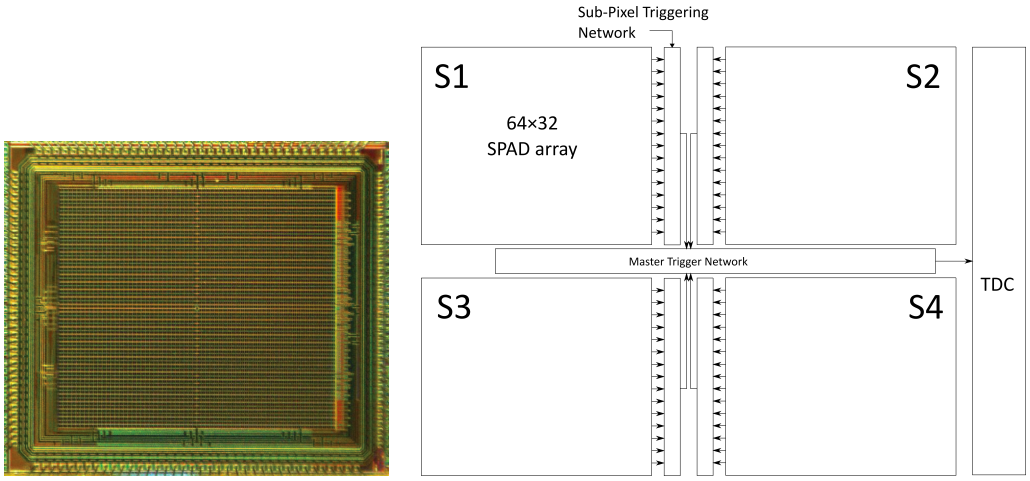


(a) (b)

Figure 2.1.7: The digital silicon photomultiplier - In opposition to conventional analogue SiPMs (a), a dSiPMs combines the digitised output of each SPAD cell through a digital network (b). Additional counting and timing circuitry can be put inside the same chip.

simple buffer/inverter which can be seen as the simplest 1 bit ADC. The circuitry inside each SPAD cell can include more sophisticated elements to improve the performance of the dSiPM.

The biggest assumption underneath this type of SiPM is therefore the CMOS integration. As shown in Table 2.3, the SPAD technology has been very well integrated in standard CMOS processes therefore enabling the idea of dSiPMs. However, the introduction of extra circuitry inside each SPAD makes it more challenging to preserve the fill factor and the *PDE*. The first introduced dSiPM by Frach, see Fig. 2.1.8(b), consists of four sections, called sub-pixels, each consisting of $(64 \times 32 - 1 = 2047)$ SPAD cells. The cell dimensions are reported as 52×20 μm^2 with a fill factor of 50% giving to the dSiPM a peak *PDE* of 30 % at a wavelength of 420 nm. The in-cell electronics inlcude a pair of transistors for active quench and fast recharging and an inverter for pulse shaping. All the four sub-pixels share together one time-to-digital converter put outside the sensitive area of the test chip. The TDC is realised as event driven, i.e. a stop-watch triggered by the detection of the photon event and stopped by the next edge of a synchronous clock. A coarse counter is also inserted for dynamic range extension. One very interesting feature included in the dSiPM is a triggering network which enables the controlling of how the timing and counting readout is activated. All the details of how the network is implemented can be found in the same publication [21], it is worth to underline that, assuming a single-photon threshold, the triggering network is equivalent to a digital OR



(a) (b)

Figure 2.1.8: First dSiPM chip - A microphotograph (a) and a block diagram (b) of the very first fully digital SiPM realised by Frach *et al.* [21].

tree. The idea is to avoid the triggering of the TDC when dark events occur or, in general, to set a desired photon level to analyse light intensity. This is to mimic the discriminator voltage operation of analogue SiPMs and also overcomes the limitation of precise TDCs to process only the one photon per pulse. Note that the average number of incident photons has to be kept below the the saturation level of the dSiPM itself to avoid non-linearities between number of incident photons and detected events described by (2.1.14). The triggering network, in other words, allows to control the system dead time, i.e. to avoid the limitations of the TDC. The input rate of the TDC in fact has to be kept below its maximum throughput in order to avoid the loss of detected events due to readout. Although no explicit information on the TDC throughput is given on the publication, by analysing the reported histograms, the reader can conclude that the readout is able to process a maximum of one event per laser pulse. In other words, this work shows the high potential of fully digitised SiPM devices and the higher ease of the readout system compared to the analogue counter part. The prototype sensor has been commercialised by Philips Digital Photon Counting and has been demonstrated in the detection of scintillation light by Degenhardt *et al.* [136] and in PET applications by Seifert *et al.* [137] while Haemisch *et al.* have directly compared the performance of the new devices to PMTs [49].

An important step in the evolution of dSiPMs has been done by Braga *et al.* with their real-

isation of a mini-dSiPM pixel [22]. Of particular interest is the proposed triggering network, or combining network, from each SPAD cell of the mini-SiPM to the TDC interface. In the original concept of the dSiPM, setting the network to a single/first photon triggering level is effectively equivalent to connect all the SPAD cells to an N_{cells}-to-1 OR tree. This work instead introduces a crucial step during the combination tree based on OR gates from the analysis of the propagation of the digital SPAD pulses. Suppose that two different SPADs in the array detect two photons within a dead time window, two events are generated thanks to the fact that the photons are incident on different diodes, however during the OR tree the events will be *merged* together losing the information of the second-in-time event, see Fig. 2.1.9. To avoid this bottleneck due to the combining network, which later on will be referred to as *channel pile-up* by Arlt *et al.* [18], two solutions have been proposed. The first one, adopted in the test chip presented in the same work [22], consists of the introduction of an intermediate step in the OR-tree: a monostable pulse shaper circuit which compresses in time the pulse at its input to a lower value τ_{MS}. Thanks to the evolution of CMOS technology, this pulse width can be easily set to sub-nano second values thus improving the timing performance compared to the typical > 10 ns dead times of the SPAD cells. Fig. 2.1.10 shows the principle of operation of the OR-tree with a monostable stage implemented in the reported test chip. This principle has been used in other works by Walker *et al.* [138] who improved the fill factor of the SPAD array by optimised geometry.

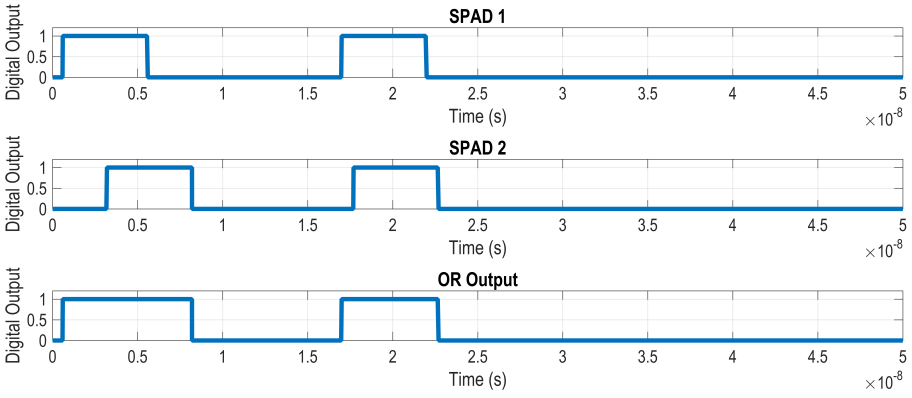
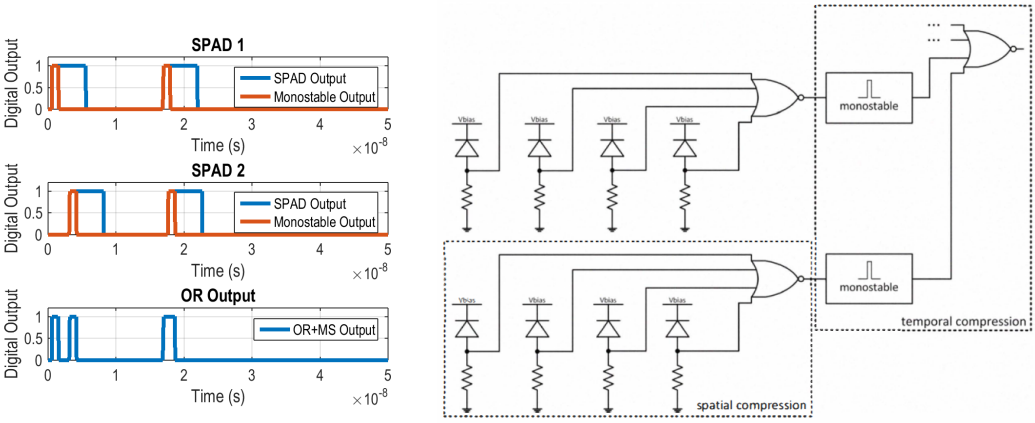


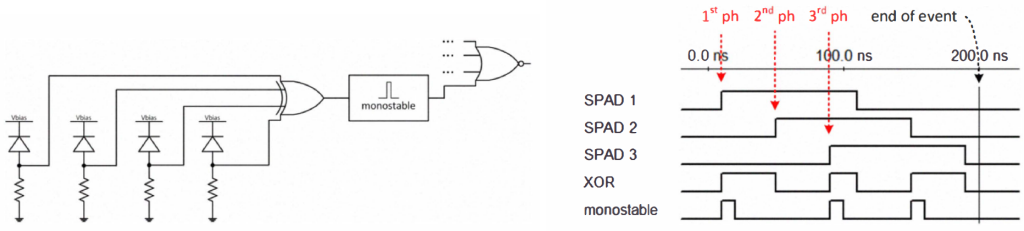
Figure 2.1.9: OR tree - Two close-in-time events can be merged together by an OR tree when the inter-arrival time of the events is lower than the individual cell dead time.



(a) (b)

Figure 2.1.10: OR tree with monostable circuit- An additional compression stage is introduced to minimise the merge of SPAD pulses during the OR tree combination network. An example timining diagram (a) is shown beside the implemented circuit schematic (b) as shown in [22].

Although implementing the monostable solution, Braga *et al.* have also identified another approach which they realise being even more effective at high photon rates. This solution consists of replacing the OR tree with a digital XOR tree, see Fig. 2.1.11. By looking at the timing diagram in Fig. 2.1.11(b) provided in the publication, the reader can conclude two important things. This solution allows to maintain the information that three events have been recorded by the SPAD array, in fact, on the monostable output three rising edges are present that otherwise would have been merged together by an OR gate registering only one event (the first one). Therefore two more events can now be counted. However, looking at the same diagram, the reader can also realise how the timing information on the second and third event



(a) (b)

Figure 2.1.11: XOR gate for SPAD combination- An alternative to the OR gate is suggested to improve the timing combination of SPADs through an XOR gate. Both the circuit (a) and timing diagram (b) are taken from [22].

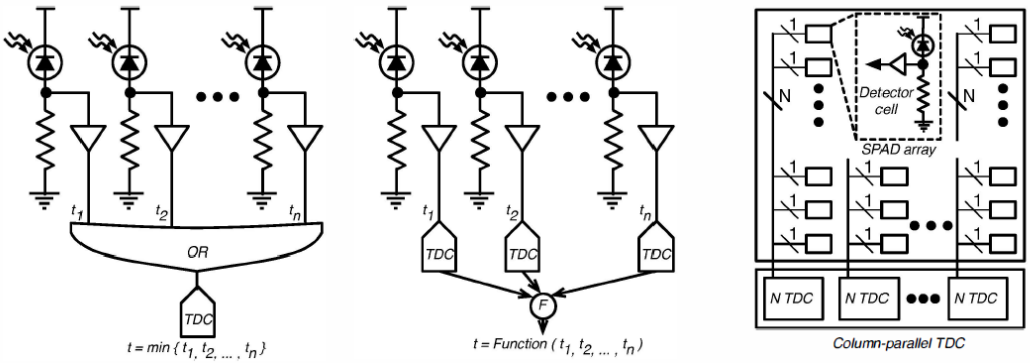


Figure 2.1.12: A column parallel TDC for dSiPMs - In opposition to a single shared TDC (a) and an individual TDC per pixel (b), a column parallel TDC (c) is introduced by Mandai *et al.* [23].

is lost: the position of the later rising edges on the monostable output do not reflect the time where the detection has taken place. This solution has improved the counting performance at the expenses of the timing resolution since the two subsequent timestamps will be affected by a random error.

A fundamentally different approach to the problem of enabling multiple time-stamping at higher photon rates has been adopted by Veerappan *et al.* [108] where the timing of the events occurs directly at the cell level: each cells contains its individual TDC and therefore no combination of events takes place. The time stamps are collected from every single cell in the dSiPM separately. This approach represents the ultimate solution to eliminate any *pile-up* due to combination of events but has the evident drawback of requiring as many TDCs as the number of cells also reducing, in standard technology, the fill factor of the dSiPM itself. The work by Veereppan reaches however a *PDE* of 27.5%. Another solution has been later proposed by Mandai *et al.* who have introduced the concept of a column-parallel TDC in dSiPMs [23]. Fig. 2.1.12 shows the three different approaches seen so far: a common TDC shared by all the cells through an OR tree (a), a dedicated TDC per cell (b) and the newly introduced concept of one TDC per column. This last solution allows to mitigate the channel pile-up giving that the number of cells sharing each TDC is smaller compared to the solution shown by Fig. 2.1.12(a) and at the same time, the TDC can be placed outside the active area of the array therefore having a much less impact on the fill factor and *PDE* of the dSiPM itself. This solution has

been called multi-channel dSiPM or multi-digital SiPM (MD-SiPM). An intensive analysis for this type of digital SiPM can be found in the work by Gundacker *et al.* who, through Monte-Carlo simulations, have analysed and compared the analogue and digital readout mode and have concluded that the performance of the MD-SiPM is highly competitive with the more popular analogue SiPMs [139].

More recently, the idea of exploiting XOR gates rather than OR gates has been however re-proposed by Dutton [140] with a fundamental modification resolving the time-stamping issue seen in the work by Braga. As described in the patent, the proposed solution is based on the replacement of the monostable circuit with a toggle cell which changes its logical output at every low-to-high transition of the SPAD signal therefore encoding the events on both edges rather than just on the rising ones as in the case of the monostable. Each toggled signal in the array is then passed onto an N_{cells}-to-1 XOR tree. The concept is shown in Fig. 2.1.13: by removing the need of creating a pulse per each SPAD detection, the XOR tree can easily propagate each transition onto the end of its output preserving, given an optimal design, the time stamp of each transition. The potential increment of the throughput of the SPAD array has been one of the main reason behind the work described in the later chapters of this work. A general overview on currently published dSiPMs is given in Table 2.5 highlighting the design achievements of each work.

Referring one more time to the work of Arlt *et al.* on the study of pile-up effect, i.e. the loss

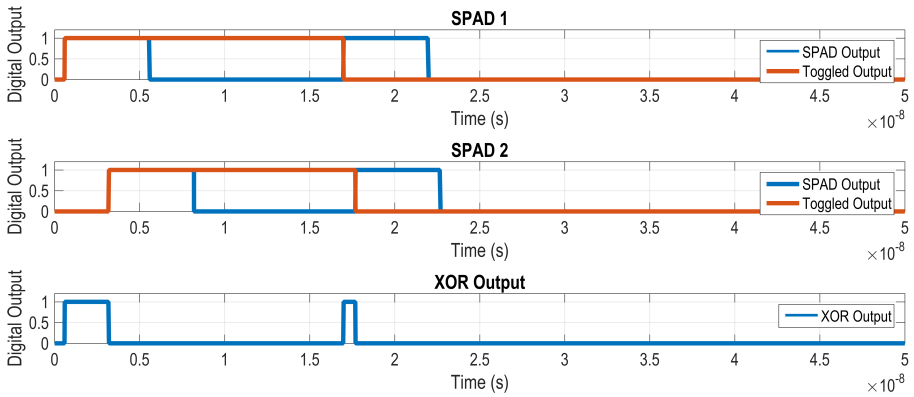


Figure 2.1.13: XOR tree - The toggle and the XOR tree encode all the SPAD events onto both rising and falling edges maintaining the timing information.

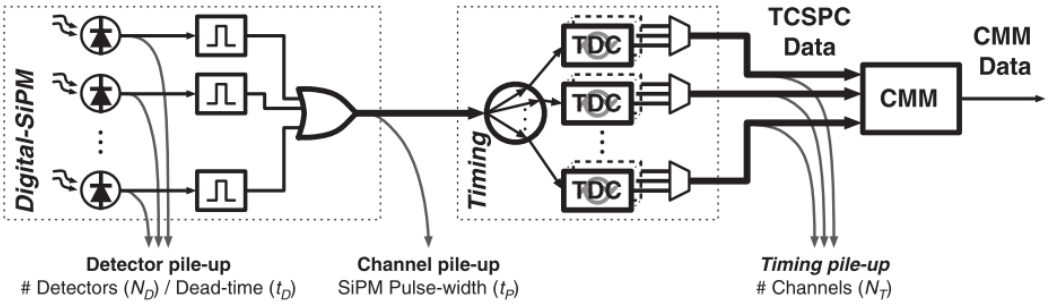


Figure 2.1.14: Pile-up - Alrt *et al.* show a comprehensive diagram highlighting all the sources of non-linear distortions in a typical dSiPM sensing chain, from detector to timing circuitry [18].

of detected photons in the sensing chain [18], see Fig.2.1.14, the reader can notice that so far, in the analysis of the dSiPM devices a main component has yet to be investigated: the time-to-digital converter which is the topic of discussion of the next section.

Work	PDE at 420 nm (%)	DCR[1] (Mcps)	Jitter FWHM (ps)	TDC per SiPM	Triggering Network	Area (mm^2)	Process
Frach [21]	~ 30	2	-	1	OR tree	12.52	180 nm
Veerappan [108]	0.27[2]	-	140	20480	None	51.2	130 nm
Braga [22]	9.33[2]	4.48	325	14	OR tree	4.3	350 nm
Mandai [28]	17.1[2]	10	114	48	Scan	0.624	350 nm

[1] At room temperature.
[2] Calculated as $PDP \times FF$

Table 2.5: Survey on published dSiPM design - An overview on a selection of dSiPMs parameters is taken from Mandati *et al.* [28].

2.2 Photodetectors and Time-of-Flight 3D LiDAR

In time-of-flight applications, linear photodiodes have found great success over PMTs due to their compactness and cost reduction. However, with the evolution of SiPM devices it is natural to compare these devices in terms of ToF performance. To do this, this section presents a useful system example to compare the detectors' output signal-to-noise ratio under the same assumptions in order to highlight the benefits and limitations of each type of devices in the specific example of ToF ranging and LiDAR systems.

As introduced in section 1.1.4 on page 10, a LiDAR system is defined by a series of components which contribute to the incident photon rate on the sensor. Estimating the average rate of incident photons is essential for the estimation of the output current of the detectors. A complete analysis needs also to take into account the noise contribution coming from the amplification stage.

Parameter	Symbol	Value	Description
Wavelength	λ	905 nm	Central wavelength of laser diode
Laser power	P_{laser}	20 W	Peak laser power per pixel
Laser pulse width	T_{pw}	4 ns	Pulse width of outgoing laser pulse
Reflectivity	η	$10-100\%$	Target reflectivity assuming diffusive materials
Aperture	D_{lens}	25.4 mm	Diameter of principal aperture lens
Angle of view	θ_R	0.1°	Angngle of view of the receiver lens of the single-point system
Bandpass	$\Delta\lambda$	50 nm	FWHM of bandpass filter around central wavelength
PIN gain	G_{PIN}	1	PIN diode internal gain
APD gain	G_{APD}	50	Linear APD diode internal gain
SiPM gain	G_{SiPM}	$1 \cdot 10^6$	SiPM internal gain
PIN dark current	$I_{D,PIN}$	10 pA	Thermal contribution to dark current
APD dark current	$I_{D,APD}$	1 nA	-
SiPM dark current	$I_{D,SiPM}$	10 nA	-
PIN excess noise	ENF_{PIN}	1	No amplification noise as its internal gain is equal to 1
APD excess noise	ENF_{APD}	4	Amplification noise in APD devices
SiPM excess noise	ENF_{SiPM}	1.1	Amplification noise in SiPM devices
SiPM size	N_{cells}	1000	Number of SPAD per SiPM device
SiPM dead time	τ_{dead}	20 ns	Dead time of the individual SPAD
PIN efficiency	QE	85%	Quantum efficiency of PIN diodes at 905nm
APD efficiency	QE	55%	Quantum efficiency of silicon APD devices at 905nm
SiPM efficiency	PDE	1%	Photo detection efficiency at 905nm
Amplifier noise	$i_{N,amp}$	$10\ \mathrm{pA}/\sqrt{\mathrm{Hz}}$	Amplifier current noise density referred to the input

Table 2.6: LiDAR System - Set of example parameters for comparison between PIN, APD and SiPM devices.

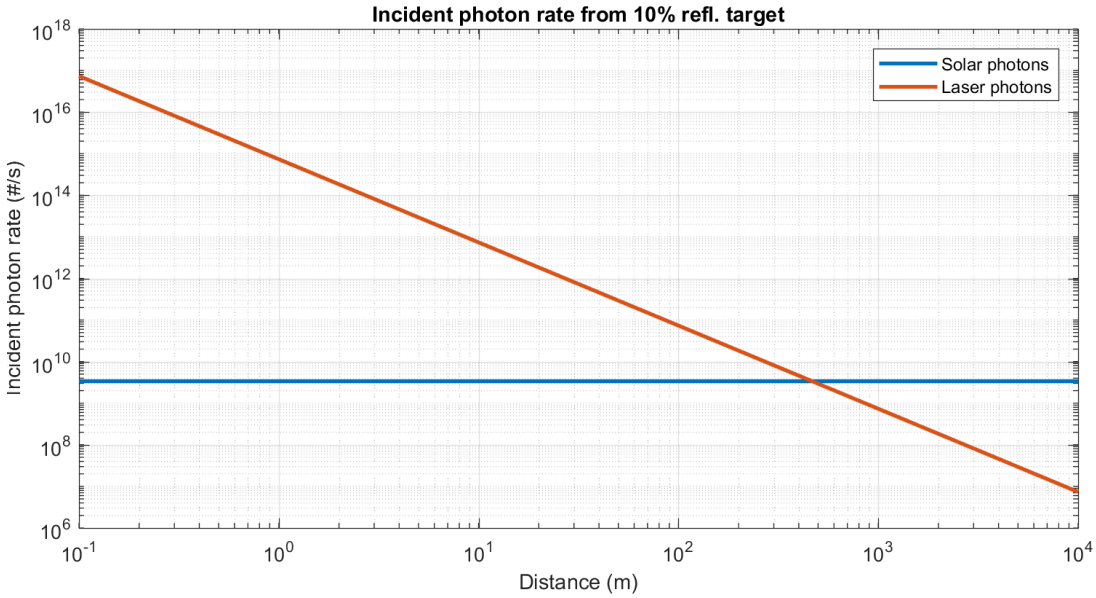


Figure 2.2.1: Incident photon rate - Calculated average photon rate on the sensors in a reference LiDAR systems.

In this section, the example system refers to a compact LiDAR system for outdoors ranging applications. Modern, and future, LiDAR systems aim at high resolution, compact and low power eye-safe systems that can be described by the set of system parameters summarised in Table 2.6. The comparison between linear photo-diodes and Geiger-mode SiPM is done with a series of general assumptions for each device contained in the second part of Tab. 2.6. Note that the same noise figure for the amplification stage is used for a fair comparison between detectors . However, it needs to be noticed that a SiPMs do no require such low-noise amplifiers because the external amplification needed for such device is much lower due to much higher internal gain. However, due to the higher bandwidth of the sensor, a high-speed amplifier is typically preferred to exploit the performance of SiPMs.

Using the equations presented in section 1.1.4, the incident photon rate is calculated and plotted in Fig. 2.2.1 over a range of distance up to 10 kilometres for a worst case scenario of a low reflective target of $\eta = 10\%$. The two curves represent the average signal rate and the uncorrelated noise photons coming from reflected ambient light. These values are here used to estimate the signal to noise ratio (SNR) of the output of each detector.

2.2.1 Photo-current calculation

The intensity of the output photo-current I_p can be calculated from the incident optical power and the sensitivity (or responsivity) of the devices. This can be expressed as the product of the incident photon rate, the elementary charge, the detection efficiency and the internal gain of the sensor:

$$I_p = \Re \times P_{opt} = q_e \times \Phi_{inc} \times G \times PDE(\lambda_p) \tag{2.2.1}$$

For linear photo-diodes and avalanche photo-diodes, the photon detection efficiency is often replaced by the quantum efficiency QE as the geometrical efficiency is approximately 100%.

For the Geiger-mode photo-diodes, SPADs and SiPMs, the detection efficiency has to be considered as a function of the incident photon rate. Due to the mode of operation, the dynamic range of each SPAD is limited by its dead time τ_d and in SiPM configurations also the number of cells N_{cells} contribute to the dynamic range. When calculating the photon current for non-linear diodes, the event rate at the output described by (2.1.4) and (2.1.6) can be used.

Noise level

The noise contribution, in terms of output current, in LiDAR systems is given primarily by:

- Dark current flowing in the detector
- Shot noise of ambient light generated photo-current
- Shot noise of laser (signal) photo-current
- Excess noise factor of the detector
- Amplifier noise - expressed as current density at the amplifier input

while the dark current and the amplifier noise do not depend on the incident light and can be easily characterised, the photo-current generated by uncorrelated ambient light photons depends on the system configuration and light conditions. As per the system parameters, it is useful to take as example a compact LiDAR system, i.e. characterised by a tight angle of view, a small aperture and a narrow NIR bandpass filter to satisfy the general trend of outdoors LiDAR. The chosen parameter values are summarised in table 2.7. All the contributions to the

Parameter	Symbol	Value	Description
Angle of view	$\theta_{R,x/y}$	0.1 deg	Angle of view of a single pixel
Aperture	D_{lens}	25.4 mm	1″ diameter of primary receiver lens
Bandpass	$\Delta\lambda$	50 nm	Filter bandpass around central wavelength expressed as FWHM
Light condition	$\mathfrak{I}_{amb}$	100 klux	Power density associated with ambient light
		38.4 W/m^2	Calculated power density incindent on the target within the filter bandpass

Table 2.7: LiDAR System - Set of example system parameters for compact systems based on PIN, APD and SiPM devices.

noise level σ_{noise} are summed up as follows:

$$\sigma_{noise} = \sqrt{2q_eB \times G\left(I_D + I_{amb} + I_{laser}\right) ENF + i_{N,amp}^2 B} \tag{2.2.2}$$

2.2.2 Signal-to-Noise Ratio

By calculating the output photo-current for each device using the incident photon rates shown in figure 2.2.1 and (2.2.1), and the noise level from (2.2.2), the signal-to-noise ratio of the detectors' output can be calculated and plotted as in Fig. 2.2.2. When the return signal becomes weak, i.e. when the target is further away from the sensor, the internal gain of the SiPM devices

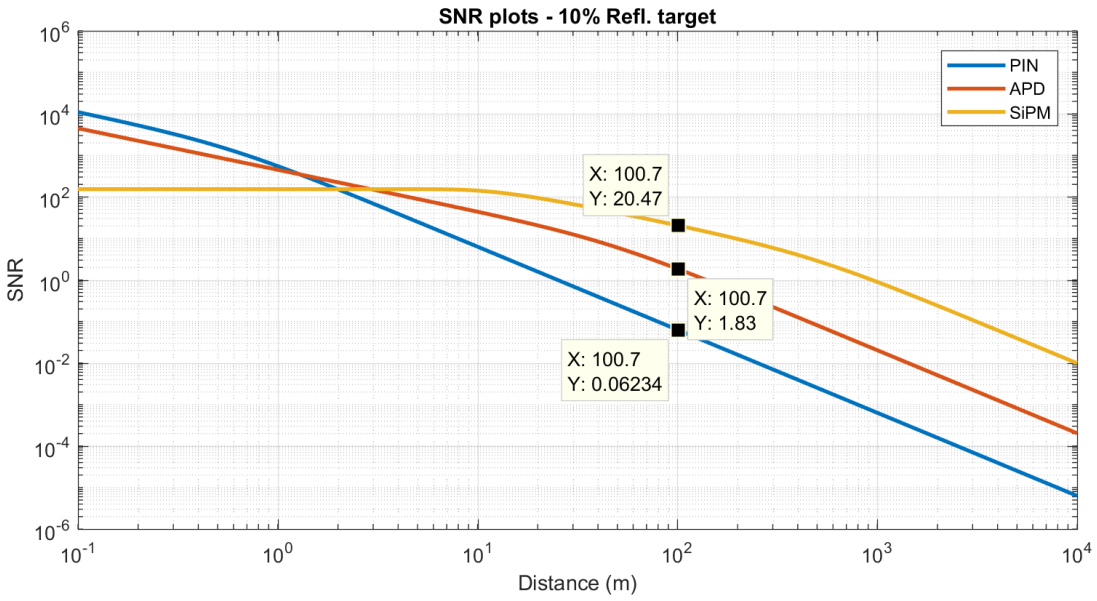


Figure 2.2.2: LiDAR SNR - Calculated signal-to-noise ratio for the three detector types in a reference LiDAR system. The high internal gain of SiPMs offers higher SNR at long distances for low reflective targets.

allows for a higher SNR to be maintained compared to linear photo-diodes. In the graph, the calculated SNR as:

$$SNR = \frac{I_{p,laser}(d)}{\sigma_{noise}(d)} \tag{2.2.3}$$

is plotted over the ranging distance for a diffusive target with a reflective coefficient $\eta = 10\%$. The internal gain of avalanche diodes allows for the limitation set by the amplification noise to be reduced. In particular, the highest internal gain showcased by the SiPM devices gives the system the higher SNR at long distance ranging.

However, the SNR for SiPMs is limited at short distance, or in general in situations where the return photon rate is high. This is due to the inherent limited dynamic range of such detectors. As explained in previous sections, the dead time and limited number of diodes per SiPM limits the maximum detection rate. In the worst cases of very high ambient photons, the SNR is limited by the noise occupancy of the SiPM which prevents the detection of incoming laser photons due to the saturation of the device.

The scope of this work is to investigate the dynamic range of SiPM based systems with a particular attention given to outdoors long distance LiDAR. The understanding of limitations and challenges for future SiPM devices will enable the design of more suitable architectures. Different techniques to investigate and improve the dynamic range of digital SiPMs will be presented and demonstrated in further chapters of this work together with LiDAR experiments to demonstrate the high dynamic range of realised single-channel systems.

2.3 Time to Digital Converters (TDCs)

The time stamping of the output of photo-detectors has initially been implemented by circuits able to produce a signal proportional to the time difference of their input, called time-to-amplitude converter (TAC) [15, 77, 141, 142] followed by analogue to digital converters (ADC). The signal coming from the detector is generally passed to a discriminator, or comparator, (with fixed or constant fraction threshold) which generates one of the two inputs of the TAC. The other input is typically a reference start/stop signal. The output of the TAC is then passed onto the ADC which then produces the correspondent digital value of the registered time stamp.

Recent TCSPC systems however, have replaced the two separate blocks of TAC and ADC with a unique fully digital device which performs directly the conversion from the time difference of its inputs to a digital code overcoming the limitations of the analogue step such as the difficulty to be implemented in IC technology, the higher sensitivity to external conditions and the longer conversion times. The fully digital circuits are referred to as the *time-to-digital converters* (TDC) [143, 144]. A useful overview on TDCs can be found in the work by Kalisz [145]. The aim of this section is to walk the reader through different types of TDCs highlighting a selection of architectures used in time-of-flight sensors based on the photo-detectors described earlier in this chapter.

2.3.1 Figures of Merit

Before entering the review, it is necessary to introduce a selection of parameters and figures of merit useful to evaluate the timing performance of such circuits and the trade-off among different designs. Three properties of TDCs are generally mentioned to describe the timing capability: the accuracy, precision and resolution. Most of these terms come directly from the world of ADCs for their similarities of operations with TDCs. However, it is useful to recall some definitions in order to avoid misinterpretations and to best understand the performance of one TDC in comparison to others. The **accuracy** of a TDC is a concept that describes how well a measure given by TDC relates to the *real value* (or expected value) of the measured quantity [146]. It can be expressed numerically by the *absolute error* between the measure and its real or expected value. The **precision** of the TDC instead is a concept related to the repeatability of the measurement, i.e. it describes the variance among values obtained by the TDC coming from independent measurements under the same conditions. For such reason it is often described as a *standard deviation* of a single measure, or alternatively as a *full-width-half-maximum* (FHWM). The precision and the accuracy should not be mixed: a high precision TDC can suffer of a low accuracy and vice-versa. The **resolution** of a TDC is a concept that describes the minimum variation at the TDC input that can be appreciated at its output resulting therefore in a change of the time stamp. However, when multiple-stamps can be obtained by a single TDC measurement, the resolution can also be expressed as the minimum time-difference between

two input events in order to be successfully distinguished as two separate time stamps. This latter concept can be considered analogue to the resolution of a microscope which is defined as the minimum spatial distance between two points in order for them to be seen (resolved) as two and not merged into one point. For a TDC, the *two-event resolution* is a concept that combines together both the precision and the one-event resolution. For histogramming/binning systems, another design parameter often mention as a figure of merit is the **bin size** (or width): the total dynamic range of the converter is divided in time bins, typically uniformly spaced, to which each measure is placed during the histogram acquisition. This value can be considered more of a design parameter rather than a figure of merit, in fact, the accuracy, precision and resolution can often be expressed as fraction or multiple of the bin width but *per se* it does not express any of these concepts. This is analogue to the concept of the *LSB* in ADCs.

A common way to describe the TDC is represented by the deviation from the ideal linear performance. Two parameters are employed to describe it: the **differential non-linearity** (DNL) and the **integral non-linearity** (INL). Per each time bin, the DNL describes how the actual size of the i-th bin deviates from the ideal *LSB* value:

$$DNL(i) = \frac{OUT(i+1) - OUT(i)}{LSB} - 1 \tag{2.3.1}$$

The INL instead is defined as the deviation of each measured point from the best line fit $m \times i + \text{offset}$ and it is also usually expressed in terms of *LSB*.

$$INL(i) = OUT(i) - m \times i - \text{offset} \tag{2.3.2}$$

As mentioned in the previous section of this chapter, TDCs in TCSPC systems tend to be the main bottleneck in terms of **conversion rate** or throughput. A useful figure of merit is the so-called **converter dead time** $\tau_{dead,TDC}$, i.e. the shortest inter-arrival time of two consecutive events that has to pass in order for the TDC to produce two separate time-stamps corresponding to the two events. The throughput TP is then defined as:

$$TP = \frac{1}{\tau_{dead,TDC}} \tag{2.3.3}$$

The underlying distinction among state-of-the-art TDC architectures is based on two main figures of merit: the precision/resolution and the throughput. Those TDCs which favour the resolution typically compromise on the throughput: the sampling frequency, based on a reference clock signal, is typically low and only one measurement is obtained per clock cycle. The long dead time due to the low frequency is employed by the TDC to improve the timing resolution. On the opposite side, TDC can compromise on the single-shot resolution to benefit the throughput: a high sampling rate, based either on a high clock frequency or on multiple phases of a common clock, generates a higher number of time stamps. However, due to the fast performance, the single-shot resolution does not compare due to the absence of a finer interpolation technique.

The following section analyses four types of implementations by:

- Counters
- Cyclic Interpolation
- Vernier Principle
- Flash TDC

2.3.2 Counters

In its simplest implementation, a TDC is essentially a stop-watch: a start signal enables a counter which is then stopped at the arrival of a second reference signal. The status of the counter n, based on a reference clock of a certain frequency F_{clk}, represents the digital measurement of the time delay between the two events:

$$T_{meas} = n \times T_{clk} = \frac{n}{F_{clk}} \tag{2.3.4}$$

In order to resolve a time interval below the nanosecond scale it easy to see that the required clock frequencies go up to $\sim$ GHz making this solution not suitable for high resolution. This technique is in fact nowadays implemented only as a coarse-timer for dynamic range extension, the reference clock frequency is kept down to few mega-hertz, and therefore the accuracy of such circuits is not high. To achieve a higher performance, finer stages are implemented and

they are described as follows.

2.3.3 Cyclic Interpolation

One of the techniques to increase the TDC resolution down to a few picoseconds or less consists in the *cyclic interpolation*. This approach has been known and used widely in the analogue domain of ADCs while in the time domain it has not been applied directly to TDC until the first introduced by Chen [147]. The proposed architecture is based on the successive pulse-shrinking of an input signal whose initial width is equal to the time to be measured, typically the residual of a coarse stage. By successively shirking the pulse, a counter is set to register how many cycles are required in order for the initial pulse to vanish and, from the status of the counter, the initial time value is determined, see Fig. 2.3.1. The resolution of this type of interpolator is set by the step of the pulse-shrinking per cycle. Since this is used for finer interpolation only, the maximum width of the input pulse set by the length of the pulse-shrinking delay line has to be set to the maximum value of the coarse stage residual. The basic structure of the architecture sets however a big constraint on the conversion rate of the TDC. In fact, the speed of the operation is based on the maximum number of cycles corresponding to the measure of the resolution of the converter. The more accurate the result, the more cycles the TDC needs to measure its input. This is reflected in a relatively poor conversion rate equal to 33 kHz for 57.3 ps resolution [24]. The same architecture has been used to realise a 6 ps resolution TDC in improved CMOS technology at the expenses of a lower conversion rate (not stated in the publication) [148]. Mantyniemi *et al.* have achieved a resolution of 1.22 ps with a cyclic time domain successive approximation (CTDSA) design reducing the number on necessary cycles to detect the *LSB* from 2^N of typical cyclic architectures down to N [149].

2.3.4 Vernier Principle

The Vernier principle [150] uses two clock signals with a slight difference in pulse period to measure fine time intervals. The resolution of TDCs based on this principle is given by:

$$\tau_{res} = T_2 - T_1 \tag{2.3.5}$$

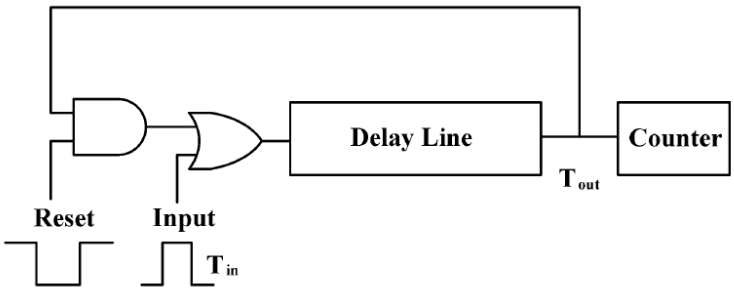


Figure 2.3.1: TDC with Cyclic Interpolation - The fine measurement of the cyclic TDC is done by successively shrinking the width of the input pulse and by counting the number of required cycles for the pulse to vanish. Diagram taken from [24].

where T_2 and T_1 are respectively the periods of the slow and fast clock. The stability of the two clocks is essential and therefore they are often generated in locked loops. This brings a design challenge of avoiding that the two loops to lock to the same frequency/period as observed in very-close oscillators with similar frequencies. The mode of operation is well described by Porat [25], see Fig.2.3.2. The first oscillator is enabled by the start signal while a second clock is activated at the arrival of the stop signal. Simultaneously to the start signal, the output of the Vernier stage is activated. Thanks to the difference in frequencies, the two oscillators will be in phase sync only after a certain amount of cycles n. A phase detection circuit captures the time of occurrence of the sync between the oscillations and triggers the output stage with a stop. The initial time difference τ has been therefore extended to nT_1 which can be now measured with coarse techniques. Once n is measures, the initial τ is calculated as $n \times \tau_{res}$.

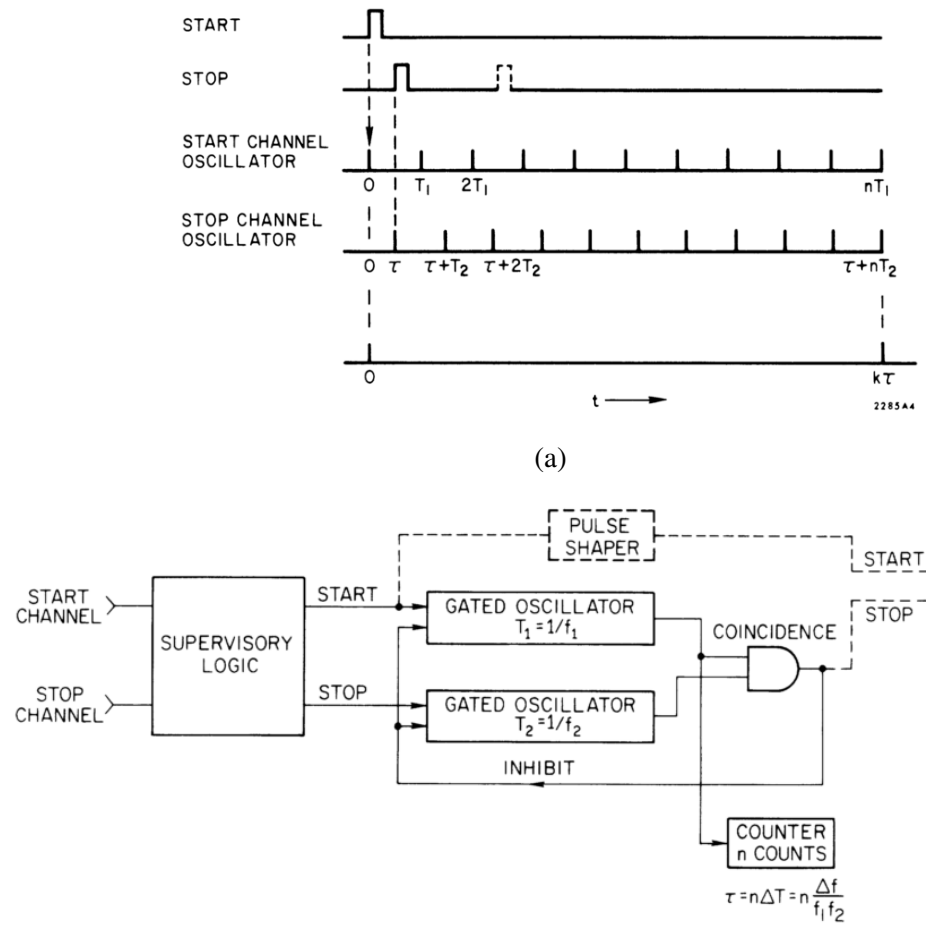


Figure 2.3.2: Vernier Principle - By using two oscillators with a slight difference in clock frequency, a fine time interval is expanded into a multiple of the period mismatch of the used clocks (a). The extended delay is measured by the desired coarse architecture. An example is from Porat *et al.* (b) [25].

2.3.5 Flash TDC

Gray *et al.* [151] investigated a way to digitise waveforms by propagating the data and/or a clock signal through delay lines and by sampling with a flip-flop the output of each delay element of the *tapped delay lines*, see Fig. 2.3.3(a), as an evolution of a concept introduced by Rahkonen *et al.* [152]. The most important advantage of these architectures is that their sampling rate can be higher than the utilised clock frequency. For such reason they are classified as *flash TDC*. It was again Rahkonen *et al.* that have studied how to implement the Vernier method without the need of two oscillators but using *tapped delay lines*, often called for such reason Vernier delay lines, in which the delay element is realised either but a simple gate, reaching high resolution thanks to a minimal gate delay, or using the delay of a flip-flop element reducing the area needed for the entire digitiser [153], see Fig. 2.3.3(b). The limitation of such implementation is represented by the potential mismatch of the delay line. For such reason, these are recommended to be included in closed loops such as delay locked loops (DLLs) or phase locked loops (PLLs) to ensure a uniform time delay throughout the length of the line. Another approach analysed in the same publication is represented by the usage of *pulse shrinking delay lines*. Essentially they operate under the same principle, but instead of having the signals misplaced in time and propagating through the line until phase coincidence, a single input, i.e. a pulse width equal to the time interval to be measured, is propagated to a delay line having a mismatched rise a fall time. As consequence of the mismatch, at each step of the delay line, the width is reduced by the amount of mismatch and therefore a point on the delay line will not produce any output. Although providing a better single-shot resolution, this solution showcase a higher jitter. The Vernier principle has enabled fast converters to be realised in

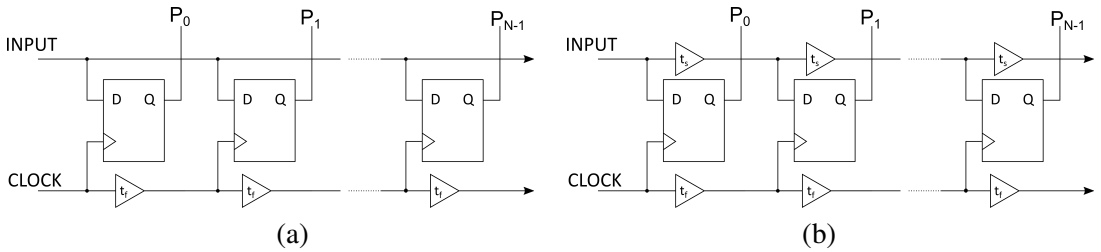


Figure 2.3.3: Flash TDC - Sampling the state of a delay line enables faster conversion rates and the expenses of poorer single-shot resolution. Diagrams for tapped delay lines (a) and Vernier delay lines (b) are taken from [26].

Work	Technique	Resol.	Conv. Rate	Dead Time	LSB	DNL	Tech.	Power
Chen [147]	Cyclic Interpolation	286 ps	33 kHz	-	286 ps	-	800 nm	-
Tisa [160]	Cyclic Interpolation	20 ps	50 kHz	-	20 ps	< 10 ps	800 nm	-
Chen [24]	Cyclic Interpolation	57.3 ps	33 kHz	-	57.3 ps	0.8LSB[3]	350 nm	1.8 mW
Staszewski [161]	Delay Line	-	-	-	$17-21$ ps	$<70\%$LSB	90 nm	1.8 mW
Yousif [162]	Vernier Delay Line	31 ps	-	2 ns	31 ps	<0.7LSB	130 nm	1 mW
Lee [163]	Time Stretching	1.25 ps	-	15 ns	<1.25 ps	-	90 nm	3 mW
Liu [148]	Cyclic Interpolation	6.5 ps	-	-	6.5 ps	15 ps[4]	130 nm	1.8 mW
Mantyniemi [149]	Successive Approx.	3.2 ps	3 kSample/s[1]	200 ns	1.22 ps	-	350 nm	33 mW
Vercesi [155]	2D Vernier Delay Line	4.8 ps	50 MSample/s	-	4.8 ps	$< 1LSB$	65 nm	1.64 mW
Menninga [164]	Vernier Delay Line	9.8 ps	300 MSample/s[1]	-	9.8 ps	<14.7 ps	40 nm [2]	-
Veerappan [108]	Ring Oscillator	-	6 MSample/s[1]	100 ns	55 ps	$<30\%$ LSB	130 nm	3 mW
Markovic [165]	Delay Line	17 ps	3 MSample/s	150 ns	10 ps	$<4\%$ LSB	350 nm	<80 mW
Narku-Tetteh [157]	Vernier Loop	8.125 ps	100 MSample/s	7.5 ns	8.125 ps	$0.64LSB$	180 nm	35 mW
Andersson [156]	Delay Line	5.7 ps	100 MSamples/s	-	5.7 ps	-	90 nm	1.7 mW
Cheng [158]	Ring Oscillator	7.3 ps	2.4 MSamples/s	415 ns	7.3 ps	$1.2LSB$[3]	130 nm	1.2 mW
Lu [159]	2D Vernier Oscillator	2.2 ps	50 MSample/s	>20 ns	2.2 ps	-	90 nm	2.6 mW

[1] Calculated as 1/(Dynamic Range)
[2] FPGA
[3] INL reported
[4] Estimated from graph

Table 2.8: TDCs
Summary of published TDC architectures for dSiPM readout.

recent technologies reaching hundreds of mega samples per second. [26, 154–159].

2.3.6 TDC Summary

In Table 2.8 the reviewed architectures of TDCs are summarised. The fastest designs have been manufactured exploiting the flash principle reaching hundreds of mega events per second on several CMOS technologies [156] including FPGAs [164]. Picoseconds resolution is instead achieved by analogue techniques of time stretching [163] or Vernier oscillators [159] showcasing resolution as low as 1.25 ps with conversion rates under tens of mega events per second. The latter types of TDCs are demonstrated in low photon rate applications where the conversion rate is not required to be high and timing resolution is therefore enhanced by the

chosen interpolative technique.

This research aims at the realisation of a high throughput, high resolution TDC which represents an evolution of the flash TDC design to be demonstrated in low and high photon rate applications. Chapter 4 introduces the newly proposed TDC and the validating results can be found in Chapter 5.

Chapter 3

Modelling of dSiPMs

In this chapter a developed model for a complete dSiPM sensor is documented. The aim is to provide a theoretical tool to analyse different aspects and parameters of dSiPMs. The outcome model, as well as this chapter, is divided in two main parts. Before the presentation of the structure of the model, a brief introduction provides a general overview of currently available SPAD and SiPM models.

The first part of the model is computational and consists of a Monte Carlo simulator developed in MATLAB which emulates the behaviour of each dSiPM cell (SPAD cell) and provides a digital output given a set of parameters resembling the main design choices for dSiPMs. The cells are then combined together in order to replicate the performance of two typical digital networks: OR-tree and XOR-tree. The readout system of the dSiPM is also simulated at high level in order to obtain a complete insight on the performance of the chosen configuration for a dSiPM. This avoids the need of long computational times required by a full CAD simulator (for example, a SPICE modelling approach). The goal is to re-create all the stochastic events that occur in the detection process of a dSiPM: starting from the photon detection in each cell, up to the digital combination from each cell to the single channel output. This part of the model will be presented as a set of computational algorithms shown as flow charts. In this way, the reader will have the general concept behind the simulator which can then be implemented in any computational tool of preferred choice. Nevertheless, a complete set of MATALB scripts is included in the Appendix A of this work and each of them will be recalled throughout this chapter as reference. The simulator has been employed in both throughput and timing analysis.

A second part of the modelling is analytical. From both computational (simulated) results and empirical data obtained by laboratory experiments, a set of theoretical equations has been

derived to describe the performance of dSiPMs. The desire to back-up the simulated results with analytical expression first of all comes from the lack in literature, to the knowledge of the author, of simple formulas that can be implemented to describe and predict the performance of dSiPMs in certain working conditions. Another reason for this analysis is to give the reader a further tool of understanding the physical processes behind the detection of photons in dSiPMs.

Chapter 5 will therefore validate the expressions here found and presented with experimental data. This model has been already briefly presented in published papers [2–4].

3.1 Overview on Modelling Approaches

Many publications in the literature show how the design of SPAD and SiPM sensors is generally preceded by extended modelling. The approach of each model aims at the analysis of certain characteristic of the detector useful to answer design questions and or to satisfy system requirements. Given the large variety of applications of such detectors, many types of models are already available. A general overview is given by Dandin *et al.* [166] who classify the reported models into four different classes according to level of analysis and the outcome.

- **Class 1: Semiconductor level** - The physics underneath the SPAD operation is modelled and simulated in order to obtain information on parameters such as breakdown voltage, gain and noise figures. These models are common for innovative SPAD and APD designs for which these parameters need to be modelled in order to have an early estimation before fabrication [167–171].
- **Class 2: Equivalent circuit level** - The models describe the SPAD as its equivalent circuit together with quenching, recharge and readout circuitry. IV curves are generally derived from such models and they are often implemented in the simulation and testing of the readout system [166, 172–175].
- **Class 3: Stochastic circuit level** - In addition to the equivalent circuit, the stochastic processes are included in this class of models. To this class belong most of the simulators which provide an output close to reality thanks to the simulation of probabilistic effects in the SPAD operation such as detection efficiency, dark counts and after-pulsing. They

are often implemented in hardware description language (HDL) [176–179].

- **Class 4: SPAD array operation** - This class of models describe the operation of the SPADs not just on a single pixel level but as part of an array. Therefore the focus of the analysis is set on array parameters such as cross-talk, breakdown voltage uniformity etc. [170, 180–182]. Models for SiPMs are included in this class.

Classes 1-3 provide significant information on a pixel level giving the designer expectation values for essential parameters for the SPAD behaviour such as breakdown voltage, dead time etc. Class 4 takes into account the arrangement of SPADs into an array, focusing the attention on how the overall signal is affected by the non-isolation of each pixels from one another. However, even for advanced SiPM models [181, 183], most of the modelling is restricted to the description of selected SiPM designs to fully understand and replicate the behaviour in order to predict the operation in general conditions.

The gap in the literature that this work is addressing in this chapter is the lack of a fully reconfigurable SiPM simulator capable of generating the temporal response of an SiPM at various light conditions, continuous and pulsed focusing the analysis on the digital SiPM design. The closest approach to this work is represented by Pulko *et al.* [181] who have set up a Monte Carlo simulator to recreate the detection process, including the probability of the avalanche, efficiency etc., and generate a macroscopic output, i.e. the SiPM current, taking into account the gain, the raise time, and the decay. The proposed simulator has been structured to focus on the digital nature of each SPAD cell of the SiPM, therefore the concept of the gain is abandoned. The sum up of the cell output is done entirely in a digital way. For such reason, the output of this model is not a current signal but it is a digital waveform, a train of pulses with photon events encoded on its rising/falling edges. Included in the simulator, a time-to-digital converter allows the histogram shaping to be generated as well. In this way, the recreated waveform, over a configurable exposure time, can be compared to the initial optical waveform, in case of pulsed light analysis, or, alternatively, the output can be simply scanned to count the detected events in the given time, in case of continuous light. In this way, the performance of the configured SiPM can be directly compared to its input in both time and count rate domain. Thanks to the ability to analyse the counting performance, different SiPM scenarios are

analysed in terms of: maximum count rate, dynamic range and saturation behaviour. The time domain approach is exploited to characterise and simulate pile-up distortions of two examples of pulsed light: a square pulse and an exponential decay. The structure of the simulator is presented in the following section.

3.2 Structure of the dSiPM Simulator

This section is dedicated to the MATLAB dSiPM simulator and presents the algorithms used to reproduce the output of a generic dSiPM. As shown in the general diagram in Fig. 3.2.1, the simulator generates a stream of photons, represented by a discrete sequence in time of ones and zeros corresponding respectively to photons or no photons in the given time stamp. This stream of photons is then served as input to a block which emulates the statistical process of the detection of photons within the dSiPM cells. Each of them is simulated individually and their output is represented as discrete digital vectors where every low-to-high transition represents a detected photon. This is the output of a typical SPAD cell where the photo-diode is paired to an inverter which can be seen as a 1b A/D converter. An N-to-1 network then combines the individual output vectors onto a single channel, i.e. onto a single vector, given the chosen type

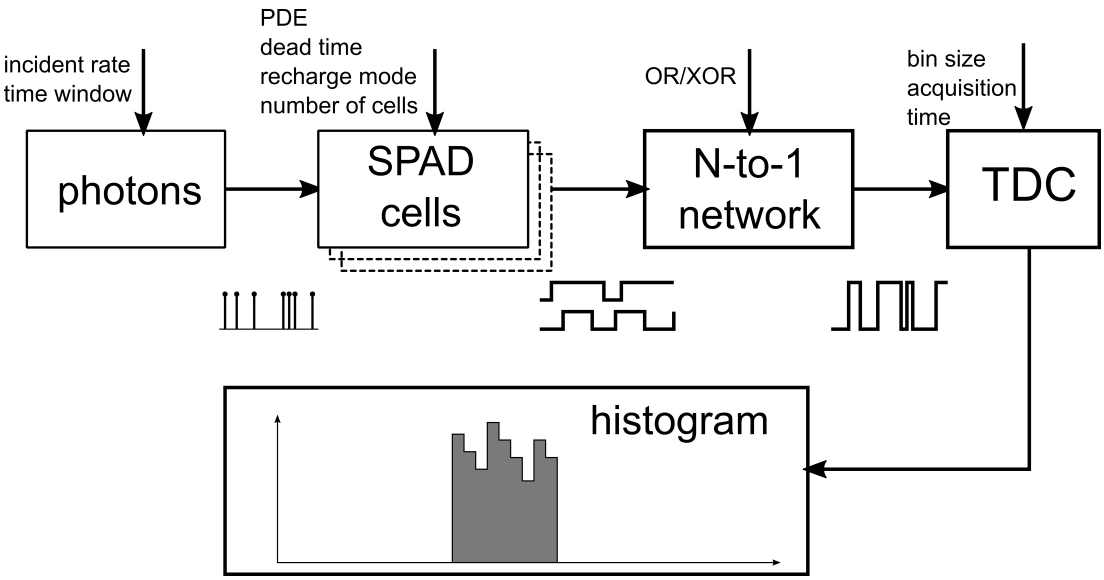


Figure 3.2.1: Monte Carlo simulator - Each probabilistic process occurring in the dSiPM is being simulated in different blocks. Each one of them is presented as an algorithm. The output of each block represents the response in the time domain of the component of the dSiPM.

of network and relative configuration. The replicated output of the detector is then analysed by a high level TDC function able to generate a histogram of the detected events. From the histograms, both counting and timing performance are obtained and analysed respectively in Section 3.3 and 3.5. The following subsections describe each of the blocks implemented in the simulator.

3.2.1 Generation of stream of photons

For a given photon rate PR and a desired time window T, the algorithm used to generate the stream of photons is described by the flowchart in Fig. 3.2.2. The probability $prob$ of having a photon within the time step of the simulation, is estimated by the average number of photons calculated as the product of the photon rate PR and the time step t_{step}. If this is less than 1,

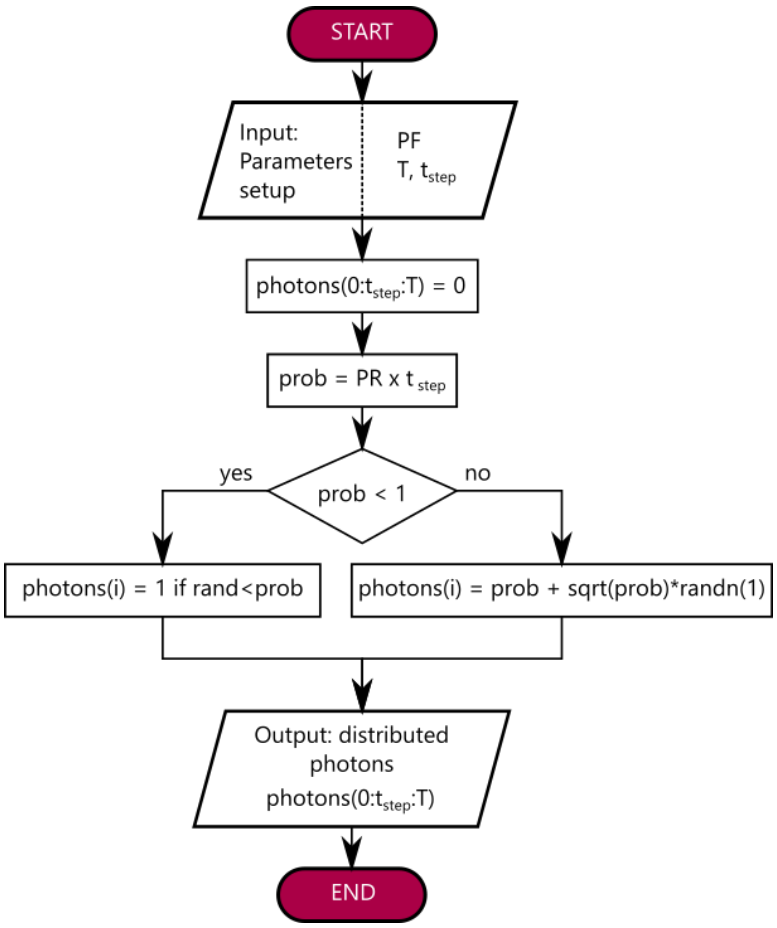


Figure 3.2.2: Generation of stream of photons - For a selected time window, an average photon rate is distributed in time.

then the vector representing the photons is filled with 1s in a Monte Carlo way: per each step, a pseudo-random number is chosen, if this is less than the probability *prob*, then the photon is generated. This process gives an overall average number of photons through the time window equal to $PR \times T$ with a standard deviation given by the square root of the average value.

If the average number of photons per time stamp is grater than 1, then each element of the stream of photons needs to be filled with a random (positive) number with mean value given by $PR \times t_{step}$ and with a standard deviation of $\sqrt{PR \times t_{step}}$.

The output of this function is ready to be used in the detection block of each SPAD cell.

3.2.2 SPAD cell

The core of the simulator is based on the single SPAD cell. The operation of a SPAD is determined by two main parameters: the photon detection efficiency (*PDE*) and the dead time τ_{dead}. The shape of the digital output of such device differs due to the effect of the two cited parameters in relation to the incident photon rate. A mechanism that plays an important role in the shape of the output is the recovery process of the cell. In fact, due to the high internal gain, the avalanche of the diode needs to be quenched in order to prevent the damage of the device itself. The quench of the avalanche and the recharge of the diode can be done actively, with the usage of controlled gates, or passively, with the use of a simple resistor. When active circuits are employed, the detection of further photons after an avalanche is inhibited for a fixed amount of time during which the SPAD is free to recharge without the interference of subsequent photons. As results, the minimum interval between two consecutive detections is fixed by the active circuits. In passive recharge mechanism instead, the SPAD restores its initial conditions through a simple RC recharge process. Although the initial condition can be considered restored only after a certain amount of time, i.e. after the dead time of the cell, the bias of the diode might be sufficient to allow another avalanche effect to occur before the complete recharge. As consequence, at high photon rates, the rate of avalanches gets higher but the output of the SPAD cell does not change because of the incomplete recharge. As results, the inter-arrival time between two consecutive detections depends on the incident photon rate and can exceed the original dead time of the cell. Both mechanisms have been simulated through

individual MATLAB functions described in the following paragraphs.

Active Recharge SPAD cell

The algorithm for active recharge SPAD cells, shown in Fig. 3.2.3(a), has been kept to a high level: per each time stamp, the function scans the input vector and checks if one or more photons are incident on the SPAD. If yes, the occurrence of an avalanche is simulated only if the SPAD bias voltage V_S has been restored to the initial value V_e which represents the excess bias of the diode. In this condition, the probabilistic photon detection process is recreated in a Monte Carlo way: a pseudo-random number is generated between 0 and 1, if this is less than the *PDE* of the SPAD, then the digital output of the SPAD cell is set to 1. This transition represents a detected photon. This is then repeated for all incident photons. Moreover, after a successful detection, the bias voltage of the SPAD is considered 0 (meaning no further detections are possible) and the time stamp τ_{fire} is stored in order to start the dead time window. The time stamp is increased in a loop. Note that when no photons hit the SPAD, the algorithm checks if the SPAD is still in the dead time window and checks if it has completed by comparing the difference between the current time stamp and the last detection to the dead time τ_{dead} fixed by the active circuits. In case of complete recharge, the digital output is reset to 0 and the bias voltage is put back to V_e. In this case, the actual value of V_e is not relevant as it has the only meaning of "on/off". The script used in this work is provided on page 198.

Passive Recharge SPAD cell

For a passive recharged SPAD cell, the algorithm is shown in 3.2.3(b). This needs a different simulation mainly due to the necessity to simulate the RC recharge process of the SPAD terminal and the consequent modulation of the *PDE*. At the start of the process, for simplicity, the time stamp of the last detection τ_{fire} is set to a negative number. At each time stamp, the voltage of the SPAD terminal is calculated as:

$$V_S = V_e \times (1 - \exp(-\Delta t / RC)) \tag{3.2.1}$$

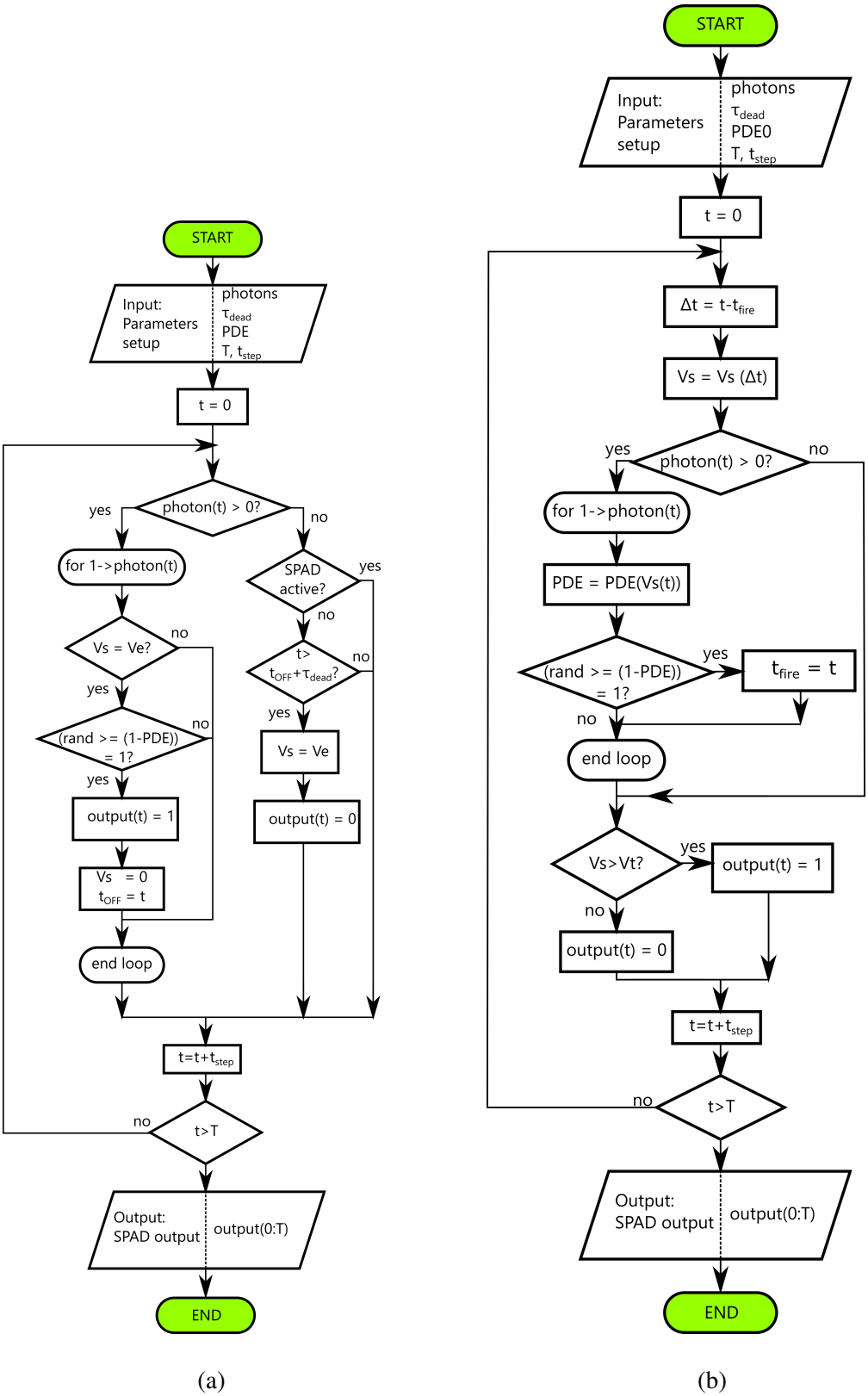


Figure 3.2.3: SPAD cell simulation - Algorithms used to simulated the output of a single SPAD cell with (a) active recharge and (b) passive recharge mechanism.

where $\Delta t = t - \tau_{fire}$. The modulation of the *PDE* according to the value of V_S is modelled as shown by Savuskan *et al.* [89]:

$$PDE = PDE_0 \times (1 - \exp(V_S/V_0)) \tag{3.2.2}$$

where the factor V_0 is inserted for normalisation. The further modification inserted in the algorithm is the following: the digital output of the SPAD cell is set to 1 only if the SPAD voltage exceeds a certain threshold. This is due to the fact that the in-pixel buffer/inverter triggers only if its input exceeds the voltage threshold of the buffer itself. This is here modelled as a parameter V_t. Otherwise, if the voltage V_S is lower than V_t the output is set to 0. In this way, in opposition to the active recharge, there is no need to "actively" close the dead time window but this is naturally set by the value of V_S. The script is provided on page 199.

3.2.3 N-to-1 Network

This section presents the algorithm used to combine the digital output coming from each cell of the dSiPM onto a single channel. Two options are available: OR-tree logic together with a pulse shortening function imitating the operation of an ideal monostable circuit per each cell, and an XOR-tree logic together with a toggle function to emulate a T-type flip-flop per each cell. The two options are individually described in the following paragraphs of this section.

OR-tree Network

To enhance the timing performance of an OR-based dSiPM, pulse compression is performed on each cell output. Because of its dead time, the digital output of each cell has a minimum pulse width of τ_{dead}. The effect of digitally combining each output using OR gates is shown in Fig. 3.2.4(a): combining two close-in-time pulses has the effect of merging them onto a single pulse having a greater width than the individual inputs. The information of subsequent detected photons, contained in the rising edge of the input pulses, is lost through the OR gates propagating only the first detected photon. The loss of photons during the combination onto single channel has been formally given the name of channel pile-up [18]. This example is conceptually similar to the detector pile-up in the case of a passively recharged SPAD where

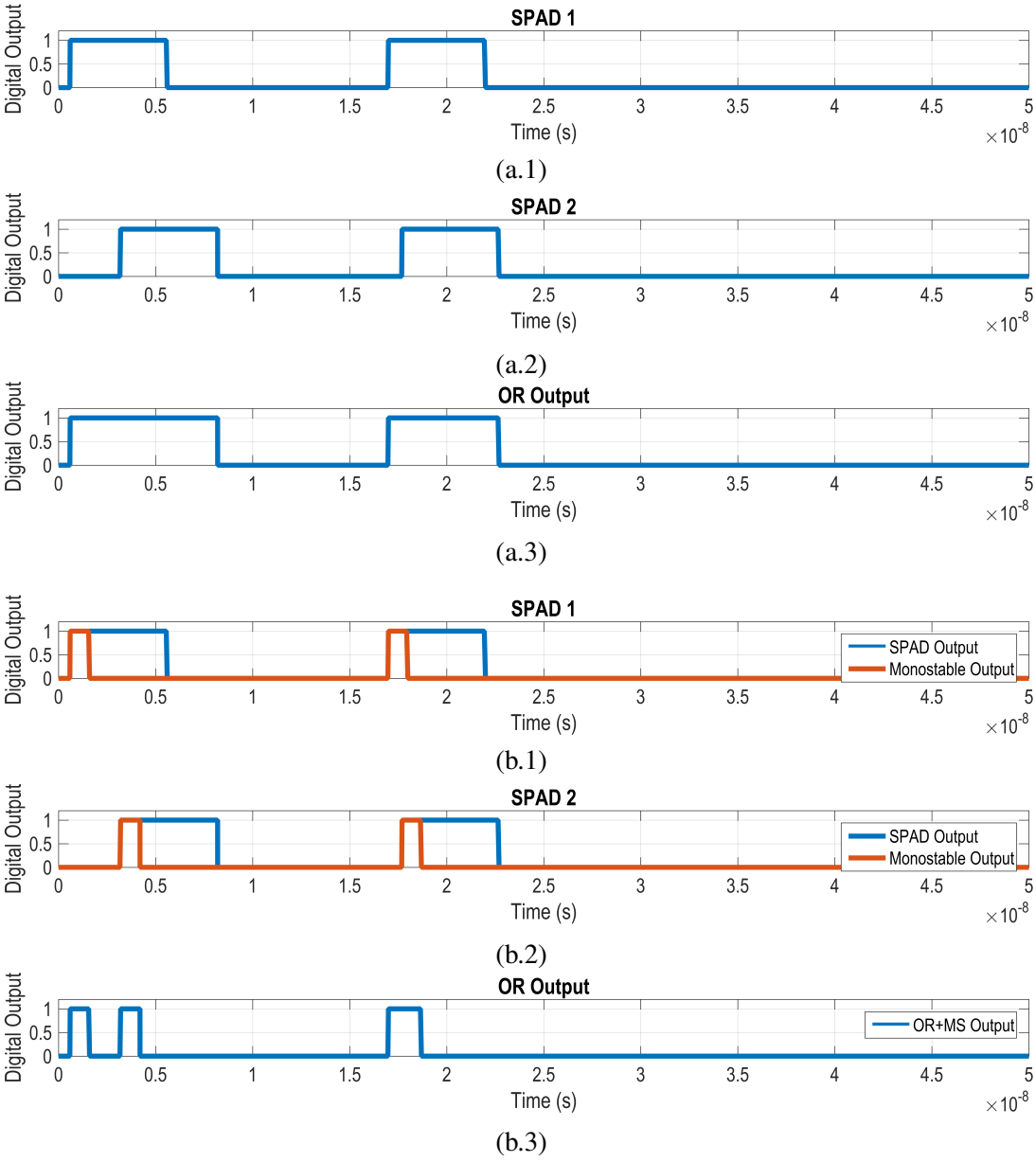


Figure 3.2.4: Output of OR-tree based dSiPMs - Shown for a no timing compression technique - the digital output of each cell is directly connected to the combining OR-tree (a) and for a pulse-shortened technique - each individual output is compressed by a monostable circuit which enables shorter pulses to be combined by the OR-tree (b)

the detection of consecutive photons enlarges the dead time window hence the pulse width of the output.

To minimise this effect, each SPAD output has to be time-compressed. In this way, although each individual SPAD still has its own dead time, the OR-tree does not merge pulses together if sufficiently spaced in time thanks to the shorter pulses at its input. This is shown in Fig. 3.2.4(b). The first two pulses, being generated in two different SPADs, are compressed in time and the OR-tree does not merge them together but propagates the two rising edges thus avoiding pile-up. In the same figure, we also show a case where the time-compression is not successful: if the compressed pulse are not short enough, two pulses might still be merged. It is therefore required to provide small pulses at the input of the OR in order to minimise the channel pile-up and ensure the highest throughput.

In the simulator, such operation is provided by a function which takes as its input the digital output (of the SPAD cell) and the value of the output pulse width. The function makes a simple copy of the input with the exception of forcing a zero output when the input is high for a time longer than the value specified as input.

To perform an OR operation, the binary function present in MATLAB is utilised. This allows only two inputs at each time, therefore to easily perform the OR function without the restrictions on the number of inputs, the OR-tree is simulated in the following way: initially, the output of the OR-tree is set to all zeros. Cell by cell, the output of the monostable is ORed with the current output of the OR-tree. The first cell to be processed is therefore ORed with all zeros, giving as output the monostable output itself, all the rest will be sequentially ORed with the previous output. A more advanced model could simulate the tree in a more detailed way to resemble the implemented circuit. However, it is not in the intent of this analysis to cover such possibilities.

XOR-tree Network

To overcome the limitation on the time compressing circuit, this work proposes an XOR-tree to replace the OR-tree as N-to-1 network. A functional block takes the digital output of the SPAD pixel and simulates a toggle operation: point by point the input is scanned and the output is inverted at every low-to-high transition of the input, see top diagrams in Fig. 3.2.5. Similarly to the case of the simulation of a OR-tree, also in this case the binary MATLAB XOR function is exploited sequentially to XOR all the outputs together onto single channel, see bottom plot in Fig. 3.2.5.

Limitations

To avoid full SPICE simulations, the behaviour of the pulse shortening circuit, the toggle flip-flop and the OR/XOR gates has been highly simplified for computational timing purposes. First of all, the rising and falling times of the gates are set to be equal to the time-step of the simulation, which in the performed analysis is kept in the orders of picoseconds to avoid long computational timings. The two values are therefore supposed to be equal to one another. This

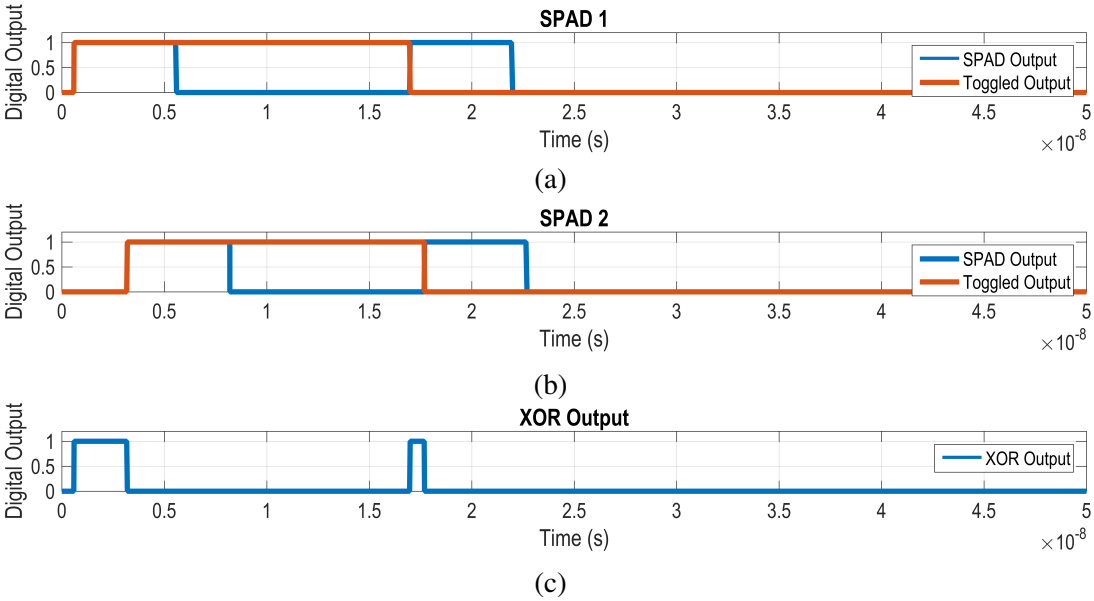


Figure 3.2.5: Output of XOR-tree based dSiPMs - Each event is now encoded onto rising and falling edges of the digital output of the pixel through a toggle flip-flop. All the edges are propagated onto the final single channel by XOR-gates.

is a very strong assumption that in real-life CMOS implementation of such circuits is very often not met. However, for this analysis, this is to be considered a second order limitation that has been left out at this stage. For an improved version of the simulator that wants to correctly emulate timing effects of such circuits, such as jitter propagation, this has to be addressed at the expenses of more intense code and longer computational times.

3.3 Simulation results for dSiPMs

Results from evaluation tests run to analyse the capabilities of the simulator are presented in this section. Each function has been tested with specific simulations to evaluate its outcome. The following subsections provide the description of each test and comments on the obtained results.

3.3.1 Photon generation

The generation of photons has been tested by checking that the distributed events follow the Poissonian statistics. The probability function $P(x = K)$ of having K photons incident of a detector when the average number is L is described by:

$$Prob(x = K) = \frac{e^{-L} \times L^K}{K!} \tag{3.3.1}$$

For a given time window set to 50 ns, the block is called to generate a number of average photons $L = 1, 2, 4, 10$ and 20 corresponding therefore to photon rates $PR = 20, 40, 80, 200$ and 400 MHz. The acquired output stream is summed up to calculate the generated number of photons K. For statistical analysis, the function has been run for 5000 iterations per each value of given photon rate. For each value of L, by normalising the number of occurrence of each number K of generated photons, a histogram is built. Figure 3.3.1 shows the histograms generated for all the values of L and on top of them, the probability function given by (3.3.1) is shown to demonstrate the Poissoninan process simulated by the block.

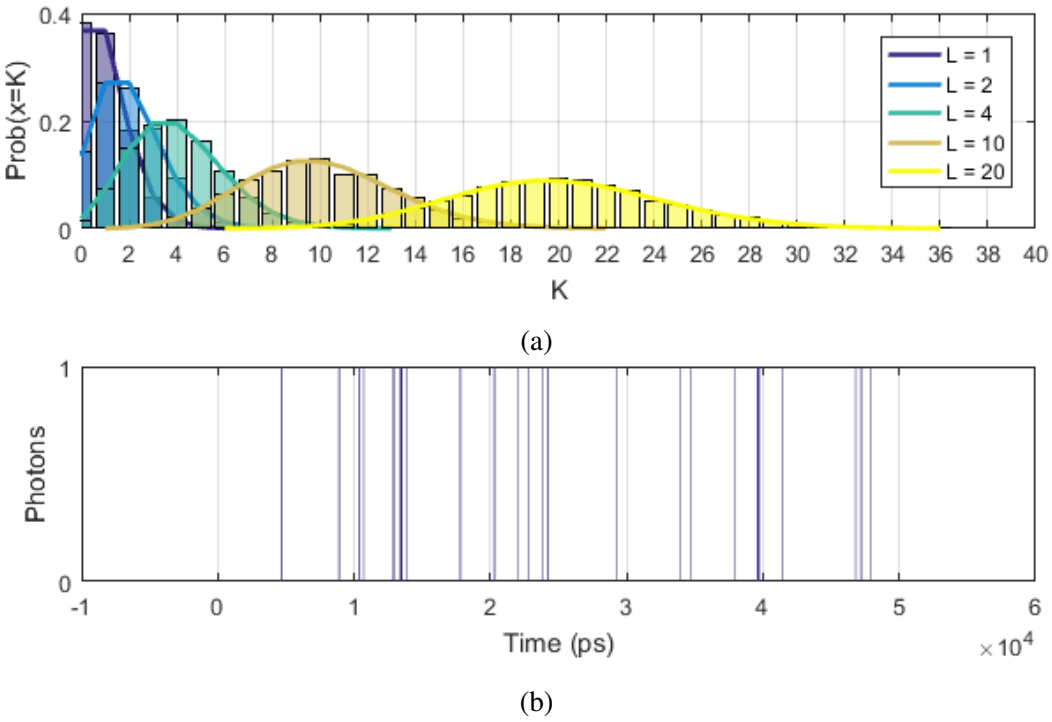
3.3.2 Detection of photon per unit cell

The simplest way to validate the function emulating the detection of photons of each unit cell is to demonstrate the average rate of detected photons. The relation between incident and detected photon rates has been successfully modelled in the literature [90]. SPAD cells are well described by general non-paralysable and paralysable detectors respectively for active and passive recharge cells. The expression used are reported here by (3.3.2) and (3.3.3):

$$m_{AQ} = \frac{n \times PDE}{1 + n \times PDE \times \tau_{dead}} \tag{3.3.2}$$

$$m_{PQ} = n \times PDE \times e^{-n \times PDE \times \tau_{dead}} \tag{3.3.3}$$

where in both equations, n represents the incident photon rates on the total area of the SPAD



(a)

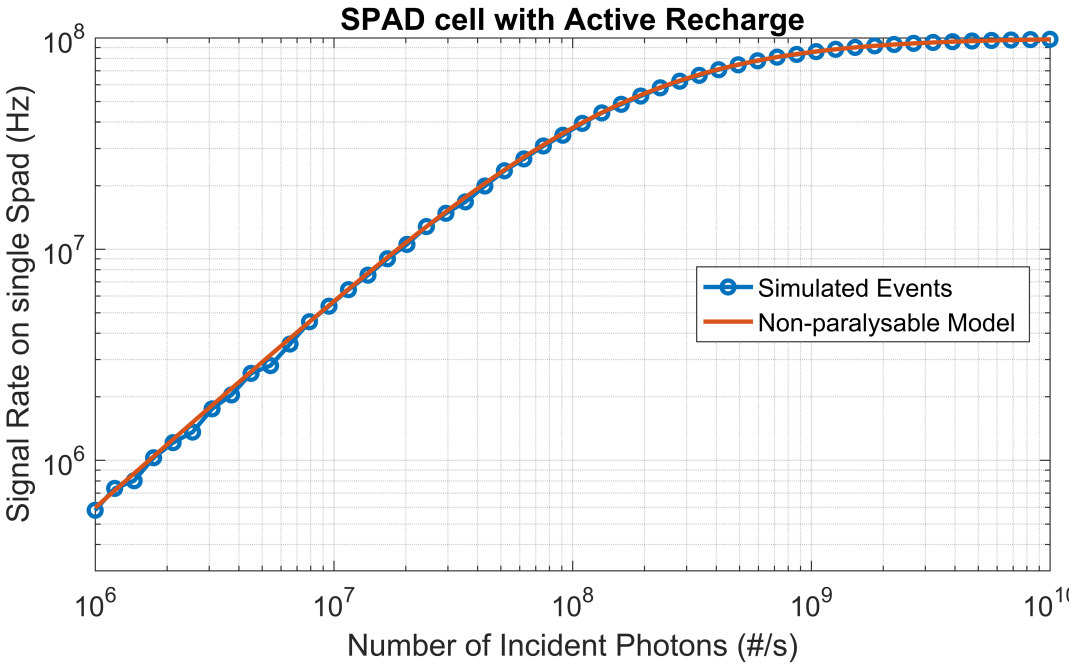
(b)

Figure 3.3.1: Generation of photons - Test - For a given set of average number of photons in the time window, photon streams are generated. Statistical analysis show that the probability of generation fits the Poissonian expression (a). An example of the photon stream is provided (b).

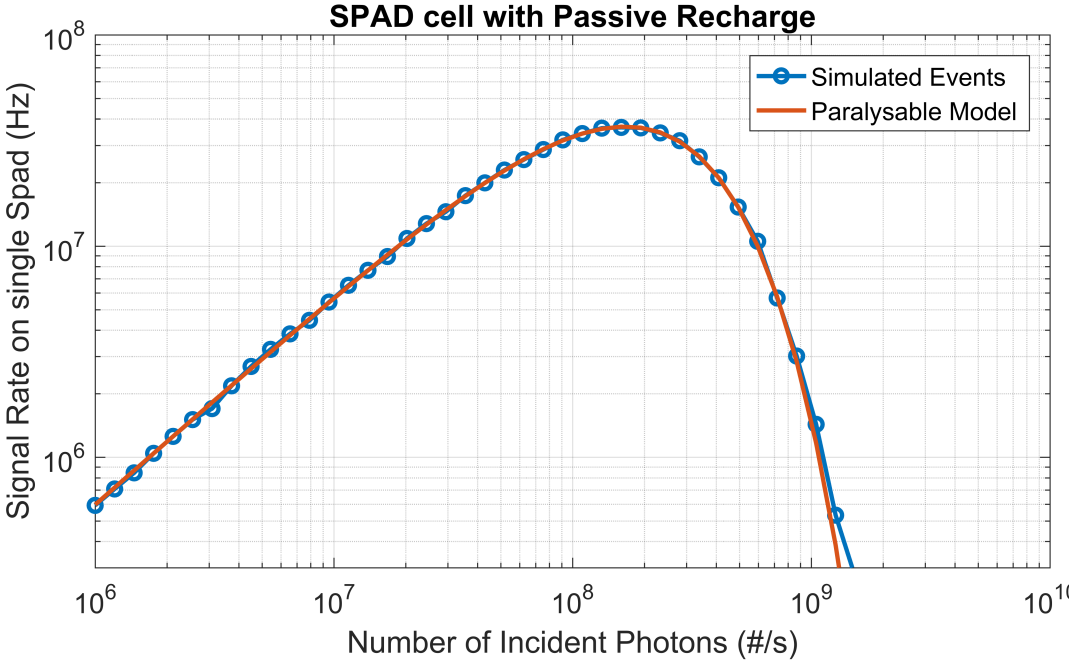
Parameter	Description	Chosen Value
T	Time window	5 µs
t_{step}	Simulation time step	1 ps
PDE	Photon Detection Efficiency	0.6%
τ_{dead}	SPAD dead time	10 ns
V_e	Excess Bias	1.2 V
n	Incident Photon Rate	10^6 to 10^{11} Hz
N_{iter}	Statistical iterations	50

Table 3.1: SPAD cell simulation - Parameter setup.

cell. The two functions emulating the detection of photon in active and passive recharge SPADs have been run with a set of parameters summarised in Table 3.1: a 0.6% PDE and a 10 ns dead time. The script simulated a time window of 5 µs. Once the digital output of the cell has been generated, the vector is scanned and the low-to-high transitions are counted. This number is then divided by the length of the time window in order to obtain a count rate. Since this is a Monte Carlo simulation, the script is launched for a number of iterations equal to 50. From these ensemble of data, an average and standard deviation are obtained. The average collected count rates have been plotted against the input incident photon rates in Fig. 3.3.2. In the case of the non-paralysable detector, i.e. active recharge SPAD, Fig. 3.3.2(a), the blue circles represent the simulated count rates while the red line shows the known model of (3.3.2). The plots demonstrate the efficiency of the simulator to create a digital output for a SPAD cell that resembles the known characteristic of real devices. In the case of active recharge cells, the count rates asymptotically tend to a maximum given by one over the dead time, according to (3.3.2) while for passive recharge SPADs the count rates reach a maximum when the incident photon rates is equal to $1/(PDE \times \tau_{dead})$ and its maximum value is given by $1/(e \times \tau_{dead})$. Note that, at this stage, no second-order effects such as dark counts or after-pulsing have been modelled. The former can be modelled as an added input photon rate while the second can be implemented given a probability of occurrence after each detection. The implementation of such effects is left to future work.



(a)



(b)

Figure 3.3.2: Detection of photon in single SPAD cells - Test - The SPAD simulator has been tested for a variety of incident photon rates and a fixed set of SPAD parameters. The obtained registered even rates match the known charateristics of non-paralysable detector, valid for the case of active recharged SPAD (a) and for paralysable detector for passive recharged diodes (b).

3.3.3 N-to-1 Network Simulation

To evaluate the performance of the block responsible to simulate the N-to-1 network, a fixed set of SPAD cells have been individually simulated, with the same amount of incident photons, and then the digital output have been summed together digitally. This has been done for both an OR-tree paired with a monostable circuit and with an XOR-tree together with a toggle cell. In Fig. 3.3.3 and 3.3.4 the count rates registered by the two flavours of network are presented. The output monostable pulse width has been set to 100 ps as well as the minimum pulse width obtainable by the XOR logic. This has been done to represent the very best case for the OR-tree and compare it to the novel XOR counterpart. The simulation has been repeated for a different number of cells from 1 to 1000. For simplicity, this simulation has been run only for passive recharged SPADs. The reason is that no active recharge SPADs were planned to be manufactured throughout the time-line of this work and therefore the simulated results could not have been validated by experimental data. However, the script can be easily adapted to simulate the opposite case by calling in the script the given function for the active recharged cell. The results from the simulation show how the dSiPM behaves like a paralysable detector (red line) for a low number of activated cells, see top left plot for 10 SPADs cells combined onto single channel for both OR and XOR tree. However, enabling more cells, the recorded count rates starts being higher than the paralysable model making the dSiPM approach the behaviour of a non-paralysable detector (blue line), see top right plots, 100 cells combined. This happens for both XOR and OR tree. The simulator shows a different behaviour between the OR and XOR tree for even higher number of input cells: a stronger non-paralysable characteristic is registered in the XOR-tree compared to the OR-tree, see bottom right plots, 1000 SPADs going into the network. The reason for the higher throughput of the XOR tree can be explained by the ability of the network to encode the events on both edges of the channel enabling more events to be successfully propagated to the counting circuit/block.

The following section aims to back up these initial results in order to better understand the process underlying the sum of the events through different networks while in Chapter 5 the validation of the theoretical findings will be carried by comparison with real data thanks to test chips described in Chapter 4.

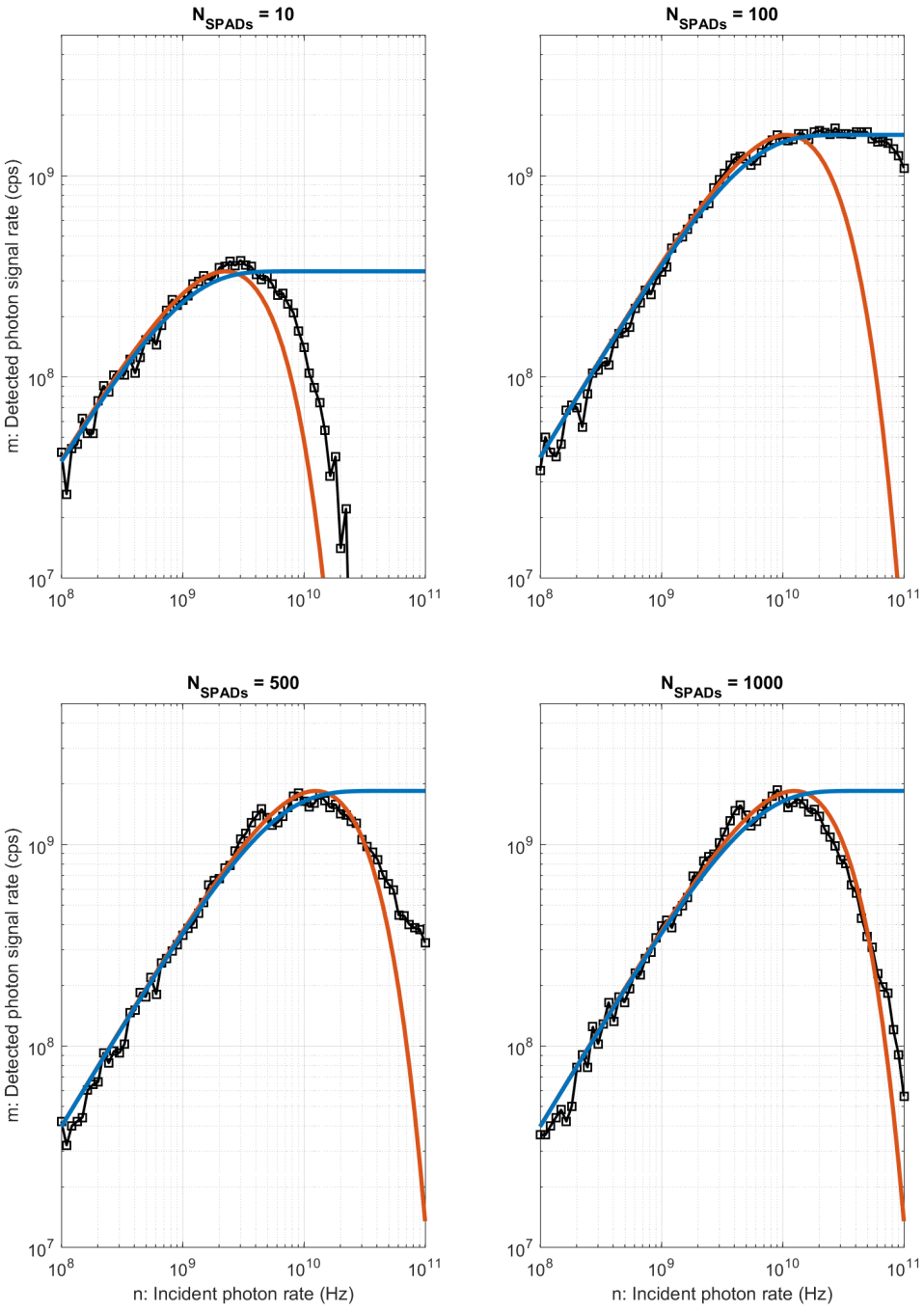


Figure 3.3.3: N-to-1 network simulation - Test - simulated OR tree for a different number of SPAD cells N_{SPADs} and different input photon rates. The collected results are plotted together with paralysable (red) and non-paralysable (blue) models.

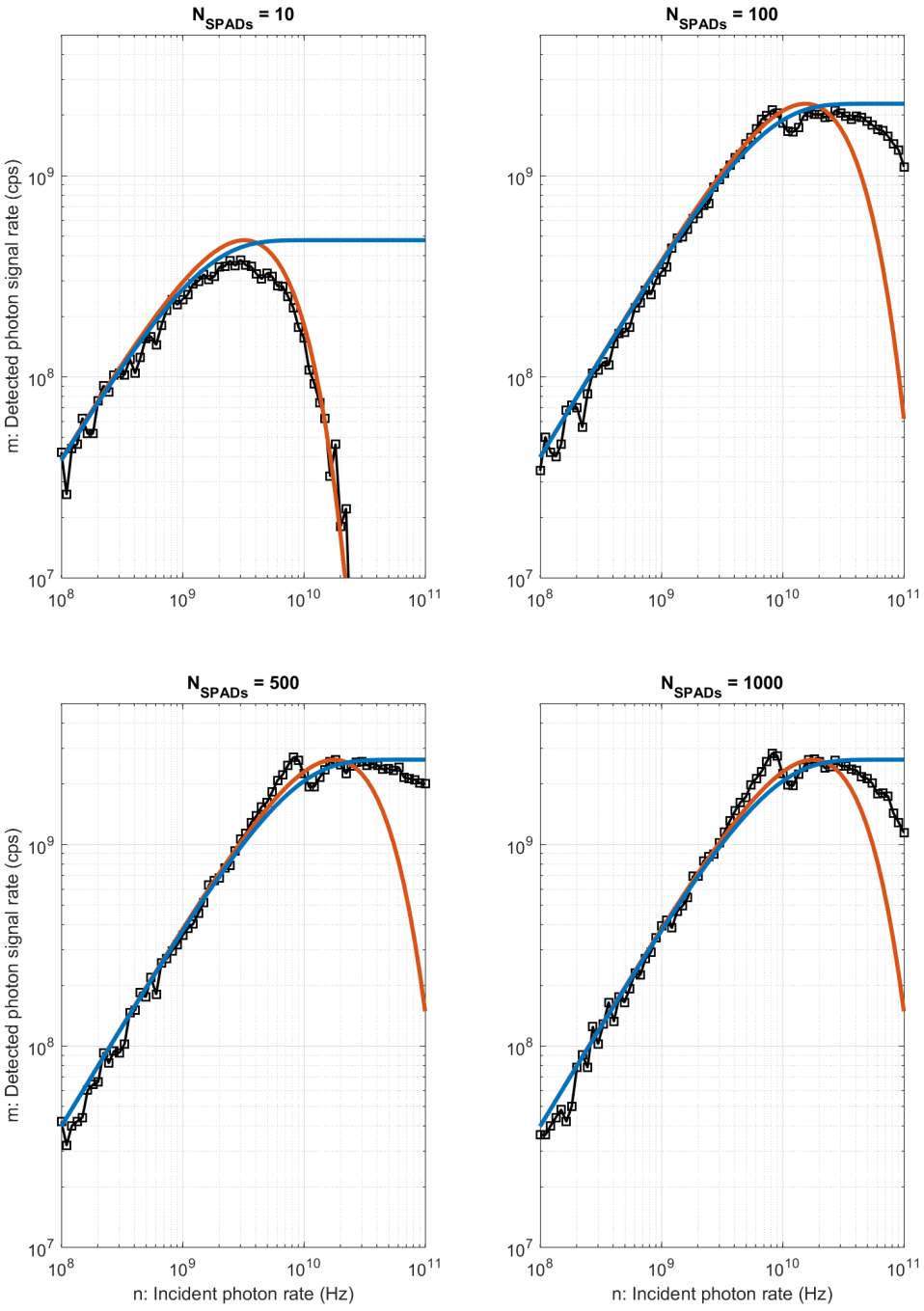


Figure 3.3.4: N-to-1 network simulation - Test - simulation results for the XOR tree with a different number of SPAD cells N_{SPADs}. Paralysable (red) and non-paralysable (blue) models for comparison.

3.4 Analytical Modelling for dSiPMs

In this section, the theoretical analysis on the throughput of a dSiPM is approached in a pure analytical way to be paired to the Monte Carlo analysis described earlier in this chapter. The goal of this section is to understand the processes behind the propagation of events in the considered networks and to provide general equations useful to describe and predict different configurations of dSiPMs. These expressions are similar to the well known characteristic of paralysable and non-paralysable detectors as in (3.3.2) and (3.3.3). As hinted by the Monte Carlo simulator, for increasing number of cells, the dSiPM shows an in-between characteristic that is here to be analysed. The two networks are studied separately starting with the OR-tree.

3.4.1 Analytical model for OR-based dSiPMs

The analysis of the OR-tree is based on the considerations made on Page 72: when two or more of the input cells detect independently two photons, if there are within a time window equal to the cell dead time, the OR gates will merge any subsequent detect photon throughout the tree propagating only the first detected photons, see Fig. 3.2.4(a). This loss of detected photons can be considered analogue to the loss of incident photons in a paralysable detector, i.e. a passive recharged SPAD. Therefore to model the throughput of an OR tree, a similar expression to (3.3.3) is proposed where the input photon rate has to be replaced by the sum of the detected count rates of the individual cells $m_0(n)$:

$$\begin{aligned} m_0(n) &= N_{SPADs} \times m(n) = \\ &= N_{SPADs} \times n \times PDE \times e^{-PDE \times n \times \tau_{dead}} \end{aligned} \tag{3.4.1}$$

The expression then becomes:

$$m_{OR}(n) = m_0(n) \times e^{-m_0(n) \times \tau_{dead}} \tag{3.4.2}$$

Parameter	Description	Chosen Value
PDE	Photon Detection Efficiency	0.25 %
τ_{dead}	SPAD dead time	5 ns
n	Incident Photon Rate	10^2 to 10^{10} Hz
N_{SPADs}	Number of SPADs	16
τ_{MS}	Monostable pulse width	1 to 20 ns

Table 3.2: OR-Tree Model - Parameter Setup.

From this, the maximum detection rate of the OR tree max_{OR} can be calculated and found to be equal to the maximum count rate of a single SPAD, max_{SPAD}:

$$max_{OR} = max_{SPAD} = \frac{1}{e \times \tau_{dead}} \tag{3.4.3}$$

To increase the throughput and effectively take advantage of the multiple available cells, monostable circuits are paired to each individual output before going onto the OR tree. As previously explained, this allows the limitation set by the dead time of the single cell to be overcome, see Fig. 3.2.4(b). In the case of a monostable cell shortening the pulse to a width τ_{MS}, this parameter can replace the dead time in (3.4.1):

$$m_{OR,MS}(n) = m_0(n) \times e^{-m_0(n) \times \tau_{MS}} \tag{3.4.4}$$

The maximum throughput in this case is related to the output of the monostable circuit:

$$max_{OR,MS} = \frac{1}{e \times \tau_{MS}} \tag{3.4.5}$$

The expression for this analytical model has been plotted using the parameters shown in Table 3.2. The choice of these parameters has been dictated by the available test chip used to validate the results later in Chapter 5. The plot is presented in Fig. 3.4.1. Different values of the monostable circuit pulse widths, from 1 to 20ns shown with different colours, have been plotted in order to visualise the effect on the dSiPM. The dotted horizontal lines represent the predicted maximum value for the given pulse width according to (3.4.5). The model not only manages to take into account the different pulse widths of the monostable circuit, but also

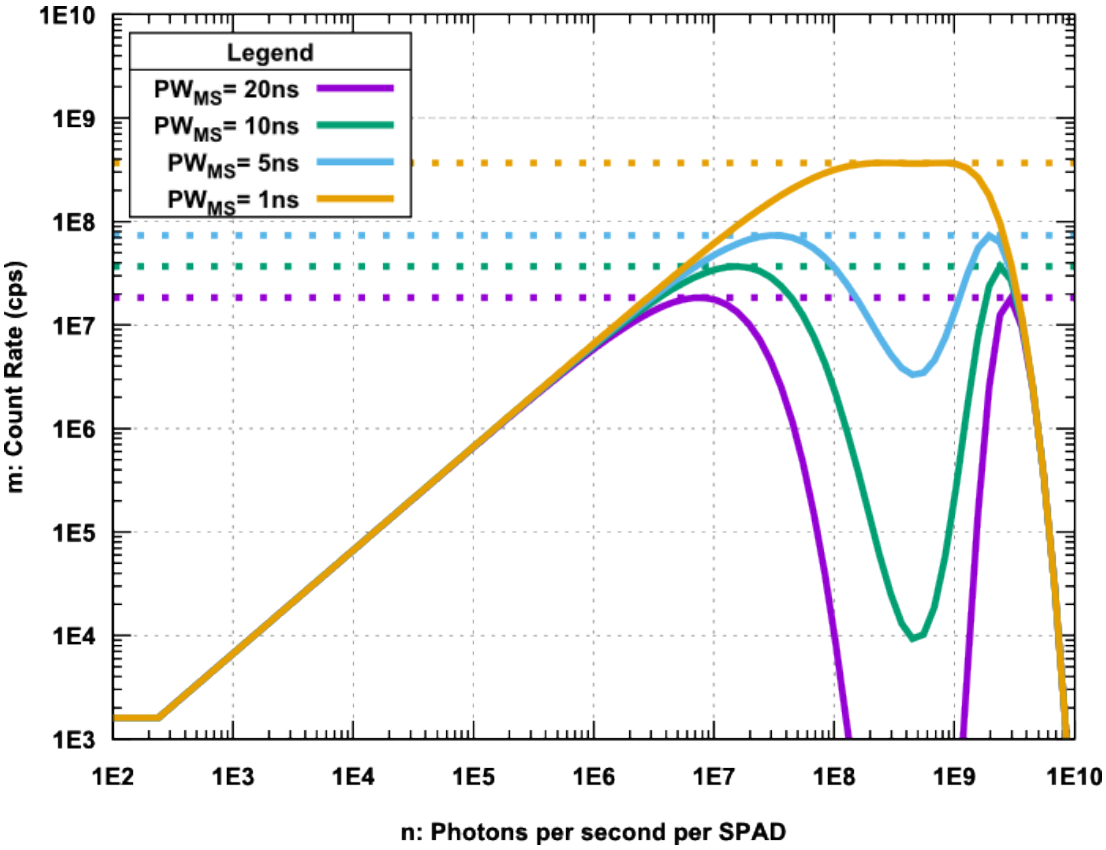


Figure 3.4.1: OR-Tree Model - An OR-based dSiPM has been modelled with a fixed number of SPAD cells and a decreasing width of the monostable output pulse. This impacts the counting performance of the array in terms of maximum count rate and saturation profile. The model treats the OR-tree as a paralysable detector.

shows the semi-paralysis of the dSiPM that was seen by the simulator. It can be noticed that, for the lowest pulse width, 1ns, the dSiPM shows a flat region of saturation, typical of non-paralysable detectors, followed then by the drop in count rates typical of paralysable detectors. This confirms what was seen and presented in the simulation results.

3.4.2 Analytical model for the XOR-tree

For the XOR, the proposed analytical expression for the count rates comes from the intuition that, overcoming the limitation given by the monostable pulse width, the ultimate bottleneck in the dSiPM is given by the bandwidth of the single-channel itself. This might arise from the electrical load of the electrical signal or to the input bandwidth of the counting/timing circuit. In any case, a general parameter f_{BW} is here introduced to express these multiple source of limitation altogether. In opposition to how the OR-tree has been modelled, the XOR-tree count rate is limited in a closer way to the non-paralysable detector: the maximum count rate is given by the bandwidth and any higher rate are simply not recorded, i.e. no paralysis of the detector. This approach is empirical and the proposed expression comes from the analysis of experimental data acquired by the test chip that will be introduced in the next chapters. However, the expression has also been justified by an analogue theory to the non-paralysable detector. The count rate for the XOR is therefore modelled by:

$$m_{XOR} = f_{BW} \times \left(1 - e^{-m_0(n)/f_{BW}}\right) \tag{3.4.6}$$

The maximum is therefore given by

$$max_{XOR} = f_{BW} \tag{3.4.7}$$

To plot the proposed equation, different numbers of enabled input cells have been selected from 1 to 256. Once again, this has been chosen given the test chip available for data comparison. The curves of Fig. 3.4.2, shown in different colours per each number of cells, show how the registered count rates linearly increase with the number of activated inputs. However, due to the limitation of the channel, higher count rates are missed with the effect of registering a

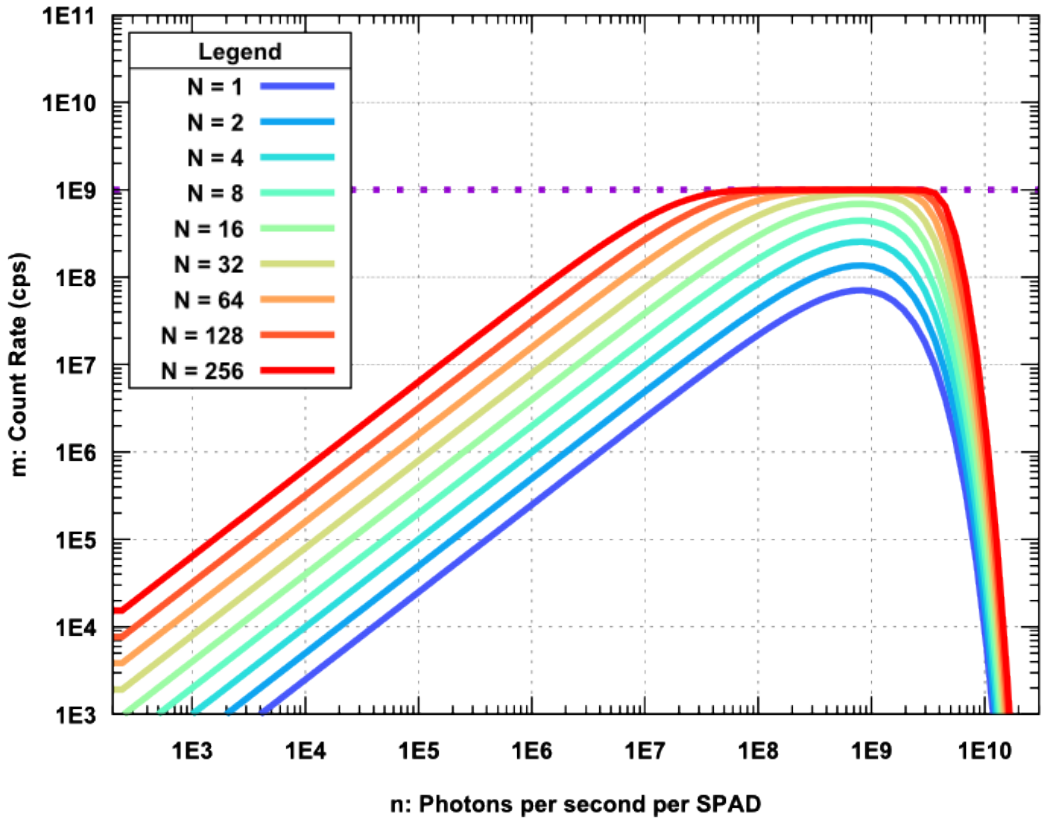


Figure 3.4.2: XOR-Tree - Test - Predicted curves describing the registered count rates of a XOR-based dSiPM with increasing number of SPADs activated in the array. The length of the linear region of operation is reduced for a large dSiPM due to limited throughput of the single-channel bandwidth. The XOR-based dSiPM showcases a flat saturation typical of non-paralysable detectors.

Parameter	Description	Chosen Value
PDE	Photon Detection Efficiency	0.25 %
τ_{dead}	SPAD dead time	5 ns
n	Incident Photon Rate	10^2 to 10^{10} Hz
N_{SPADs}	Number of SPADs	1 to 256

Table 3.3: XOR-Tree Model - Parameter Setup.

throughput only up to a maximum predicted by (3.4.7), reported in the plot as a horizontal dotted line.

Higher throughputs are therefore predicted in favour of the XOR-based dSiPMs compared to OR counterpart thanks to both the simulator and the analytical model. These results will be compared in Chapter 5.

3.5 Pile-up Analysis in Time Domain

The simulator that has been presented earlier in this chapter can reproduce the count rate of dSiPMs but it also has the ability to provide the entire temporal response of the single SPAD cells and the global dSiPMs. This means that the digital output can be scanned not only for throughput analysis but also to generate entire waveforms in the time-domain. Thanks to the photon generation blocks, an optical waveform can be generated by varying in time the photon rate to represent the desired optical shape. For example, a square pulse can be easily reproduced using the block described in the Subsection 3.2.1 on Page 68. This waveform is then used as optical input to the dSiPM simulator and the output is analysed by timing all the events successfully detected by the dSiPM. This is how an ideal TDC with no conversion dead time and with a sampling period equal to the time step of the simulator itself would operate: no additional photon loss is here modelled. The purpose of simulating such a TDC is to visualise the performance of the detector by itself. Additional effects due to non-ideal TDC have been at this point of the analysis intentionally left out.

The reconfigurability of the simulator allows the different types of pile-up sources to be individually analysed. Using the terminology in literature [18], the section is divided in subsections

Parameter	Description	Chosen Value
PDE	Photon Detection Efficiency	0.25 %
τ_{dead}	SPAD dead time	5 ns
n	Incident Signal Photon Rate	7×10^8 Hz
T_{pulse}	Pulse Width	10 ns
PRR	Pulse Repetition Rate	500 kHz
n_{amb}	Ambient Photon Rate	2×10^7 Hz
N_{SPADs}	Number of SPADs	1
t_{exp}	Exposure Time	0.1 s
T_{BIN}	Bin size of ideal TDC	10 ps

Table 3.4: Detector Pile-Up Test - Parameter Setup.

to study the detector pile-up, i.e. the single SPAD cell, and the channel pile-up, i.e. the loss of counts during the propagation through the N-to-1 networks. The latter is done for both OR and XOR under certain assumptions illustrated in the subsections.

3.5.1 Detector Pile-Up

The first simulation represents the distortion of a square pulse by a single SPAD detector. The simulated parameters are summarised in Table 3.4 showing a square pulse of 10 ns with a repetition rate of 500 kHz being sent to the SPAD cell. The intensity of the light is expressed in photon rate. Superimposed to the pulse, which can be considered the signal that has to be encoded, there is uniform light source expressed as constant level of 20 Mcps. This level can be seen as incoherent ambient light but it is also an easy way to model the dark count of the SPAD itself. Given the PDE of the SPAD, here set to 25 %, and the TDC bin size set to 10 ps, the maximum count can be calculated as:

$$C_{max} = n \times PDE \times T_{BIN} \times N_{pulses} \tag{3.5.1}$$

where N_{pulses} is the number of laser pulses acquired per histogram. This is set by the repetition rate of the laser and the acquisition/exposure time:

$$N_{pulses} = t_{exp} \times PRR = 50\,\text{k} \tag{3.5.2}$$

The average registered count per bin can be calculated by the average detection rate expressed in (3.3.3):

$$C_{avg} = m_{PQ} \times T_{BIN} \times N_{pulses} \tag{3.5.3}$$

However, due to the detection mechanism, the first photons hitting the SPAD are being favoured compared to subsequent ones due to the dead time of the single cell. For such reason there is a transition between the maximum registered count, which occurs at the very start of the pulse, and the "steady state" and such transition can be expressed by an exponential decay due to the Poissonian distribution of the probability of the photons in the pulse to the first one to be detected. The detected count versus the acquisition time can be expressed as:

$$C(t) = C_{max} \times e^{-t/\tau_p} \tag{3.5.4}$$

where τ_p is a time constant which depends on the input photon rate by the following relation:

$$\tau_p = \frac{1}{n \times PDE} \tag{3.5.5}$$

The simulated output is presented in Fig. 3.5.1. Two curves have been superimposed to the plot of the histogram: the red line shows the exponential decay expressed by (3.5.4) while the yellow line shows the average count calculated by (3.5.3). It can be seen how the simulation confirms the initial peak of the output waveform due to the Poissonian probability and how, later within the pulse, the shape falls onto the average count. After the pulse, the waveforms returns to the floor level set by the ambient light:

$$C_{floor} = n_{amb} \times PDE \tag{3.5.6}$$

The shape of the pulse has been effectively distorted as shown in the figure. This is one clear example of pile-up distortion due to detector dead time. Typically this is avoided by reducing the intensity of the light incident on the sensor by the usage of neutral density (ND) filters. This optical devices attenuate the average photon flux by a factor that does not depend on the incident wavelength, in opposition of bandpass filters which are utilised to reduce noise

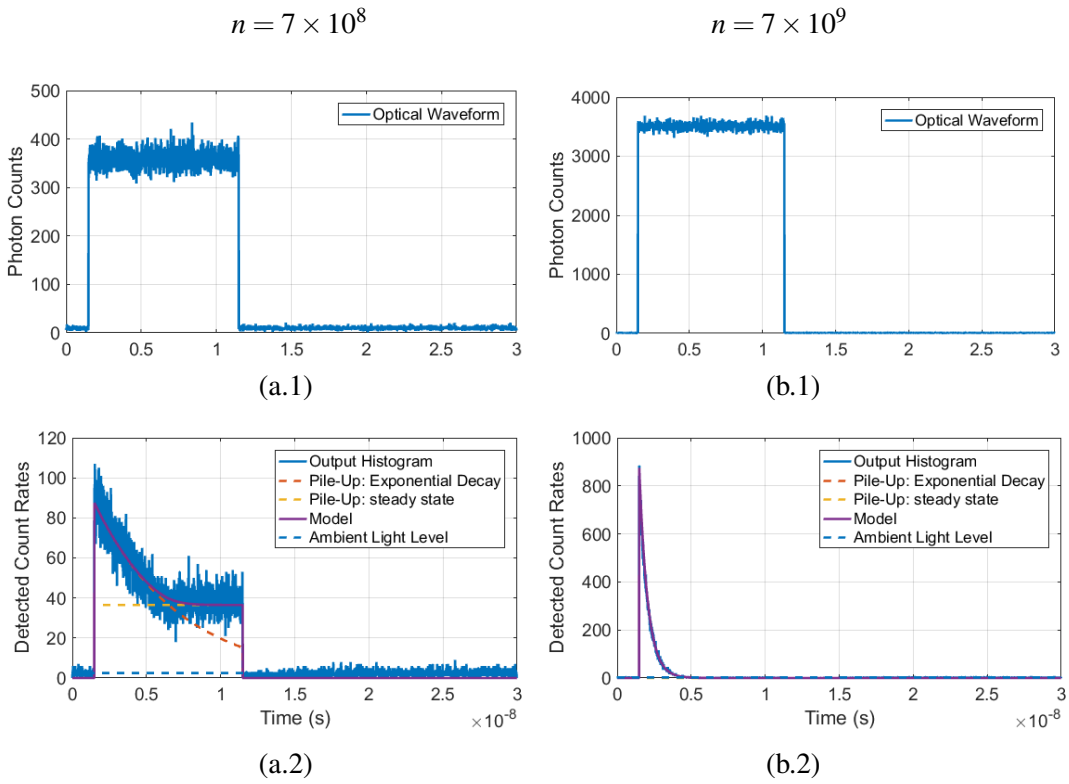


Figure 3.5.1: Detector pile-up distortion - A single SPAD cell distorts an intense square pulse into an exponential decay due to its internal dead time. The output histogram generated by simulations is the digitisation of the optical input. A situation of partial saturation (a) is shown against full saturation due to very intense incident light (b).

coming from wavelengths different than the dominant wavelength of the light source. This is shown in Fig. 3.5.2 with a simulated ND filter coefficient equal to 1.2 which attenuates the light level by a factor of $1 \times 10^{-1.2}$. Although this solution restores the shape of the optical waveform, the quality of the histogram is heavily affected by the lower registered counts due to filtering. A higher exposure time is required to improve the statistics. An example is shown in Fig. 3.5.3 where the exposure time has been increased to 1 s. As can been seen, the shape of the optical waveform is now well preserved but the reduced amount of light significantly slowed down the experiment. This is very crucial not only for the practical duration of experiments but it might exceed the timing requirements of certain applications where fast measurements are needed. Moreover, if the intensity of the optical waveform has a wide peak-to-peak amplitude,

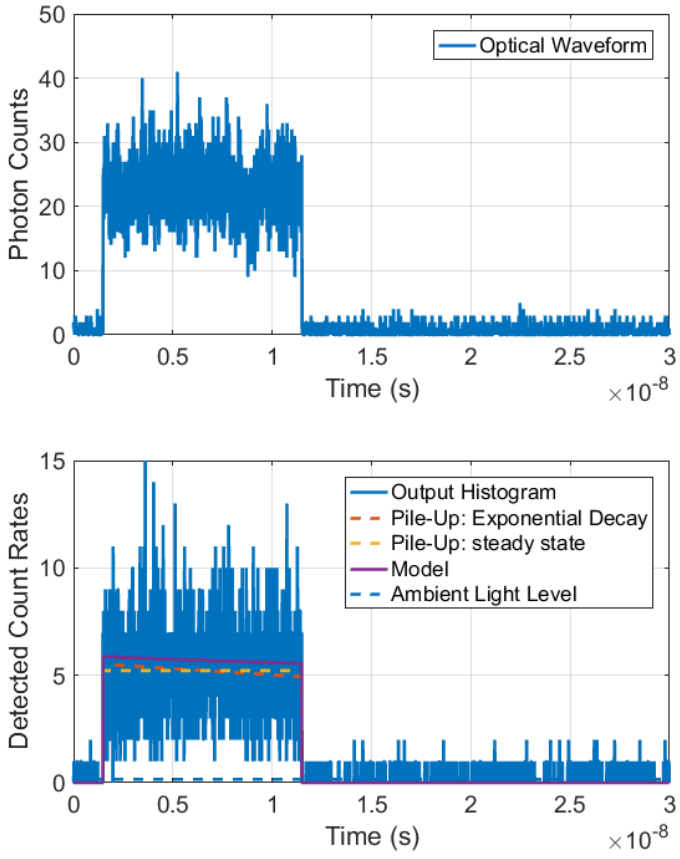


Figure 3.5.2: Detector pile-up distortion - The distortion due to detector saturation is usually overcome by attenuating the incident light on the sensor. In this case, an ND filter with coefficient equal to 1.2 is being simulated.

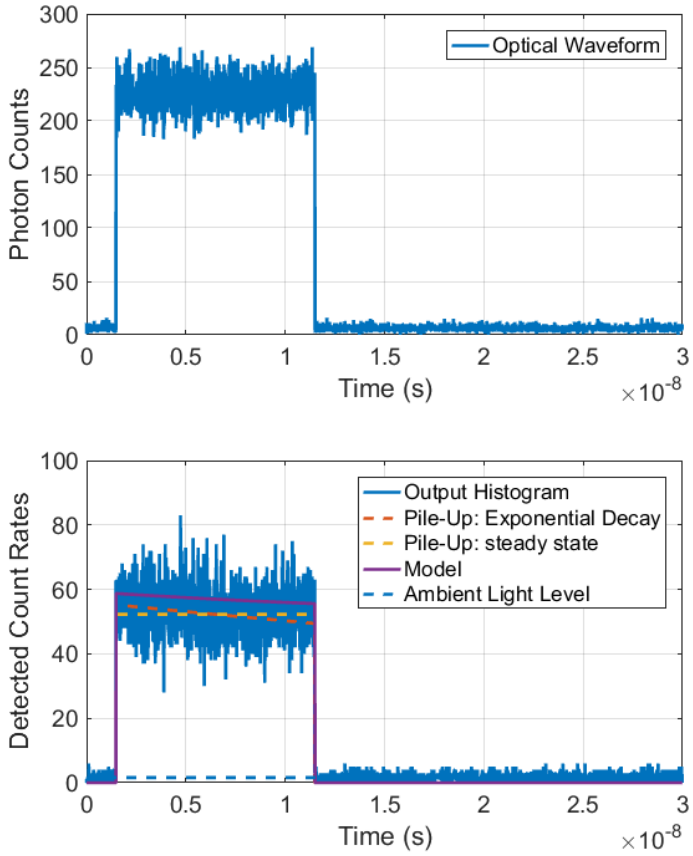


Figure 3.5.3: Detector pile-up distortion - When filters are used to avoid pile-up, the exposure time has to be increased to improve the statistic and consequently the shape of the collected histogram. In this example, the exposure time has been increased to 1 second.

reducing the incident light might allow the detection of the higher range but could compromise the detection of the lower range if that becomes comparable to the dark count rate of the sensor itself.

A more efficient solution to process the high quantity of photons incident at the sensor is to spread the light over a larger number of cells. In other words, instead of using simple SPADs, dSiPMs are preferable.

3.5.2 Channel Pile-Up

The effect of a N_{cells} dSiPM has been simulated for both OR and XOR type of digital N-to-1 network. According to the light intensity and the dead time of the cells, a minimum efficient number of SPADs can be verified to preserve the shape of the square pulse. In order to have more cells available, either smaller SPADs have to be fit into the given area, with the consequent reduction of PDE due to fill factor loss but potential improvement of dead time due to smaller capacitances per diode, or, more easily, lenses need to be used to direct the same amount of light over a larger area. To keep the analysis simple, the latter scenario is simulated so that the PDE and the dead time can be maintained constant during the analysis. In this way, the improvement of the reconstructed waveform can be seen by increasing the number of cells, see Fig. 3.5.4. On the top of the figure, the distribution of the total photon events (square pulse with noise floor) is shown and used as input for all the configurations. Below that, on the left side of the figure, the plots for the OR-based dSiPM are shown with increasing number of cells: $N_{cells} = 2, 4, 10, 50$. Note that in these assumptions, the collected light per each cell is $n_1 = n/N_{cells}$ because the light is supposed to be uniformly distributed over all the dSiPM. This expression has to be used in (3.5.5) to calculate the pile-up time constant of the output of each SPAD cell. On the right side, for direct comparison, the plots for the XOR-based dSiPM are shown. Together with the simulated detected counts, blue line, the analytical model, purple line, is shown as well as its two main components: the exponential decay due to distortion within each cell (dashed red line), and the "steady state" due to average count rates of the dSiPM (dashed yellow line). Because of the higher throughput of the XOR network, the shape due to the SPAD pile-up become quicker less and less dominant in the overall shape of the reconstructed waveform, see Fig. 3.5.4(e). With these tests, the simulator shows a better rejection of detector pile-up in XOR-based dSiPMs compared to OR counterpart.

However, another scenario has yet to be analysed. In fact, dSiPMs may be used not just to increase the dynamic range of the sensor compared to 1-SPAD devices, but they are often used also to realise large area sensors. When more cells are put together in arrays to form dSiPM with this purpose, no additional lenses are used. Therefore if one SPAD cell sees an incident photon rate of n_1 then this does not change when more cells are put together. In fact this is done

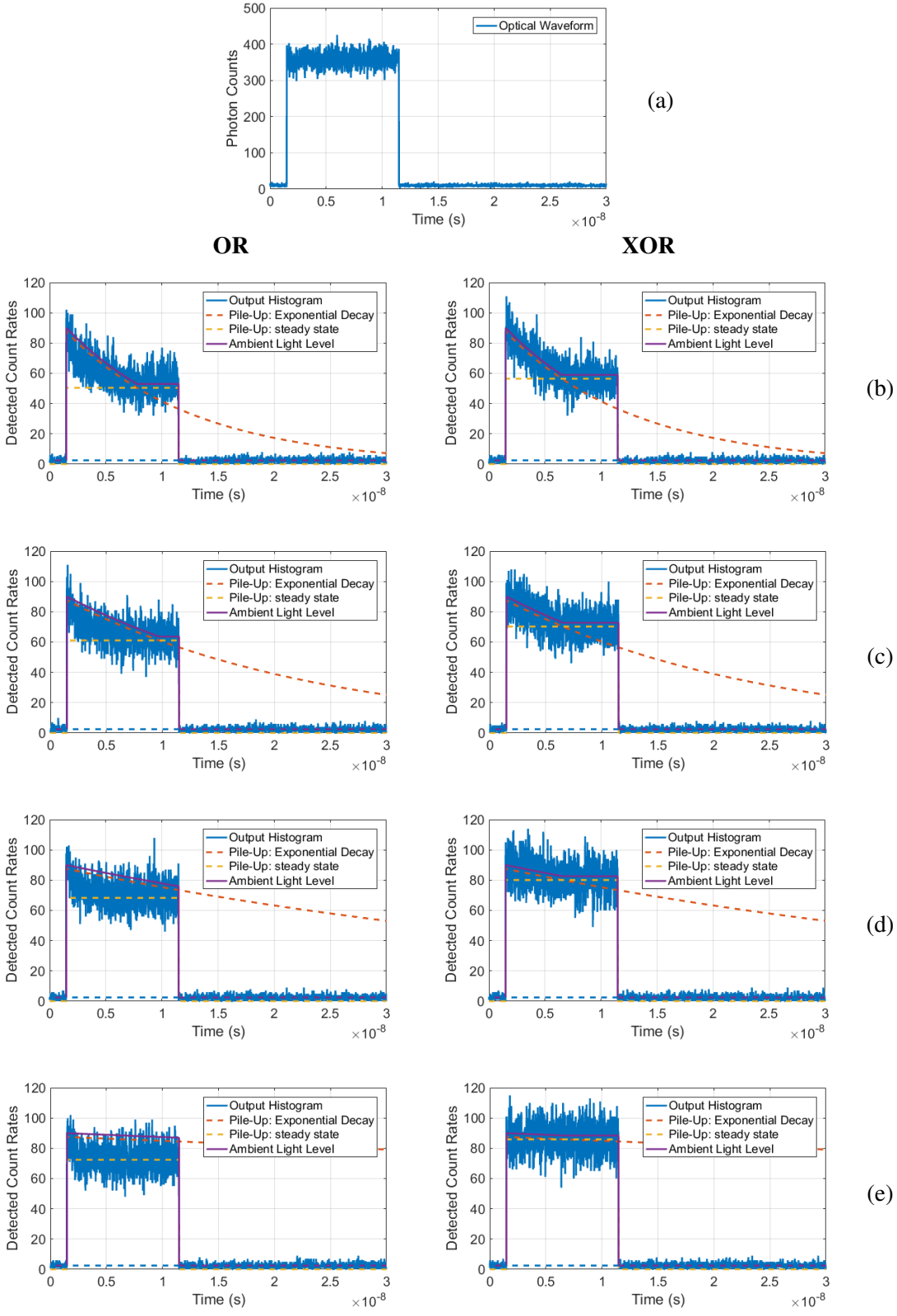


Figure 3.5.4: Channel pile-up analysis - For a crescent number of cells on which the light is spread, OR and XOR are compared. The numbers of cells here simulated are: 2 (b), 4 (c), 10(d) and 50 (e). The left column corresponds to the OR network with a monostable circuit output of 1 ns while the right column shown the respond of an XOR-based dSiPM. The latter network better rejects the distortion of the input waveform (a).

to improve the collected light which linearly increases with the number of cells: $n = N_{cells} \times n_1$. Ideally, when more cells are combined onto a single channel, the final waveform is the sum of the individual cell output with the advantage of reducing the exposure times thanks to the higher collection of light. However, due to the single-channel limitations, when the input rate of detected events to the network is very high, saturation of the single-channel may occur. In this case, the overall detected counts reach the saturation level and do no increase further. With this limitation, the reconstructed waveform is therefore cut at its higher levels to a maximum value given by the limited channel throughput. Let $W_1(t)$ the temporal response of each single cell, then the overall output waveform of the dSiPM is ideally given by $\widetilde{W}_o(t)$ where the $\sim$ stands to underline that this is an ideal output:

$$\widetilde{W}_o(t) = \sum_{i=1}^{N_{cells}} W_{1,i}(t) = N_{cells} \times W_1(t) \tag{3.5.7}$$

To model the distortion of the ideal waveform due to channel saturation, the two cases of OR and XOR are analysed separately.

The OR-tree dSiPM is limited by the output pulse width of the monostable circuits which are connected to each SPAD cell. In fact, at high event rates, large pulses are merged together by OR gates. This translates into the loss of detected events from the cell output to the single channel. As anticipated in the Subsection 3.4.1, the process describing the loss of events in an OR tree is similar to a paralysable detector which in the domain of count rates is described by (3.4.4). In the time domain this translates into an exponential decay with a time constant $\tau_{OR,decay}$ in complete analogy with (3.5.4) which described the decay of another paralysable detector: the single SPAD cell. For that case, the time constant was defined as the reciprocal of the average input event rate. For the OR-tree this is defined in the exact way, therefore:

$$\tau_{OR,decay} = \frac{1}{n \times PDE} = \frac{1}{N_{cells} \times n_1 \times PDE} \tag{3.5.8}$$

The initial shape of the output waveform can be therefore written as:

$$W_o^{OR}(t) = C_{max} \times e^{-t/\tau_{OR,decay}} \tag{3.5.9}$$

where C_{max} is the maximum registered count per TDC bin, analogue to the expression (3.5.1), here given by:

$$C_{max} = n \times T_{BIN} \times N_{pulses} \tag{3.5.10}$$

The "steady state" of the output waveform depends on the average count rates of the OR-tree expressed by (3.4.4). Therefore the registered counts after the initial decay is given by:

$$C_{avg} = m_{OR,MS} \times T_{BIN} \times N_{pulses} \tag{3.5.11}$$

To demonstrate this model, a simulation has been run with the parameters shown in Table 3.5 and the results have been plotted in Fig. 3.5.5. Note that the intensity of the pulsed light and the continuous background have been lowered by the same factor to mimic the effect of neutral density filters. Moreover, the TDC bin size has been increased to 100 ps to increase the speed of the simulation. The simulated count rates show as predicted an initial decay due to saturation of the network caused by the large number of input cells before flatting onto the average value which comes directly from the expression of the throughput of the OR-tree itself. As it can be seen by the comparison between Fig. 3.5.5(a) and (b), which correspond respectively to 5 and 10 cells dSiPM, the increased number of input cells causes the saturation of the OR-tree as a paralysable detector. The distortion of the output waveform is similar to the effect seen in

Parameter	Description	Chosen Value
PDE	Photon Detection Efficiency	0.25 %
τ_{dead}	SPAD dead time	5 ns
n	Incident Signal Photon Rate	1.4×10^7 Hz
T_{pulse}	Pulse Width	10 ns
PRR	Pulse Repetition Rate	500 kHz
n_{amb}	Ambient Photon Rate	4×10^5 Hz
N_{SPADs}	Number of SPADs	5-50-200-500
t_{exp}	Exposure Time	0.1 s
T_{BIN}	Bin size of ideal TDC	100 ps

Table 3.5: OR Pile-Up Test - Parameter Setup

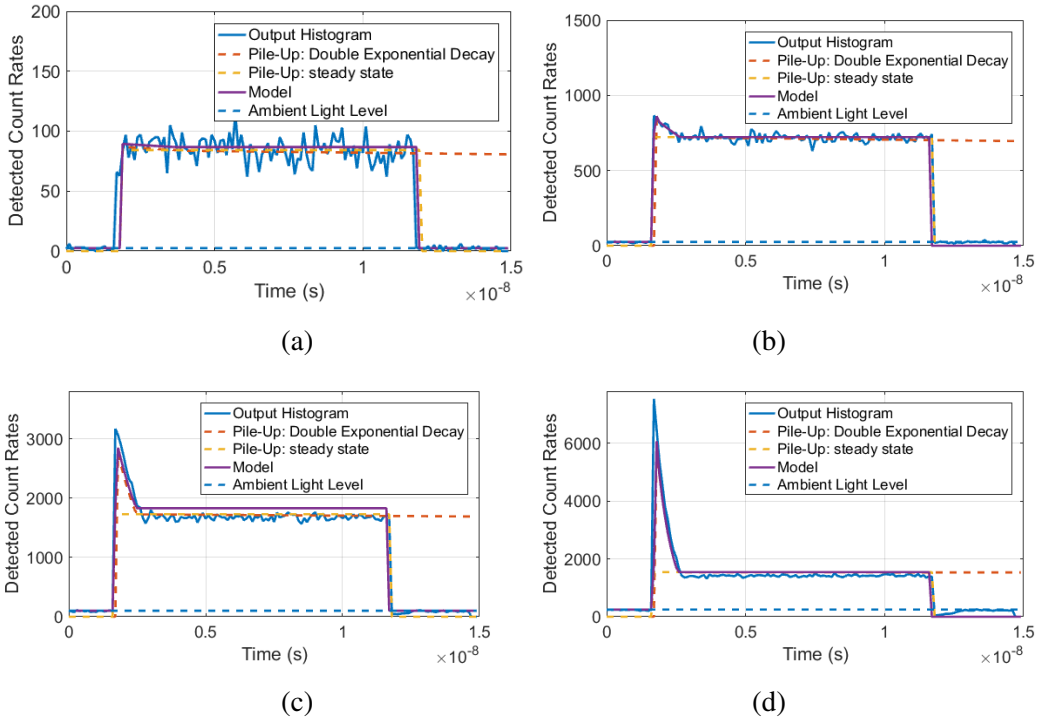


Figure 3.5.5: OR-based dSiPM temporal response - The paralysable nature of the OR-tree network causes pile-up distortion of the input square waveform which translates into an initial decay before reaching the average value due to throughput. Obtained waveforms are simulated for an increasing number of cells: $N_{cells} = 5$ (a), 50 (b), 200 (c), 500 (d).

Fig. 3.5.1. Note that in this case, the saturation of the network has been shown in a particular case where the single cells are not saturated, i.e. the light intensity on each cell does not cause pile-up within the detector but the events are loss only within the N-to-1 channel digital sum of events. When the two processes combine, modelling the two distortions, i.e. exponential decay, might not be immediate since the two are not uncorrelated: a loss in detected photons within the cells results in a less input event rate to the network reducing the pile-up due to the network but increasing the distortions due to single cells. The analysis of such scenario has been left at high level. In Fig. 3.5.6 an alternative scenario is described where the input photon rate has been increased back to the values shown in Table 3.4. Two types of distortions are seen: the earliest one represents the distortion due to detector pile-up and therefore depends only on the input photon rate. The second tail of the output instead is still exponential but described by a different time constant. This constant depends not on the input photon rate but on the input event rate of the sensor. When increasing even further the number of cell, the former decay constant prevails and the distortion can be described by a single process exponential decay.

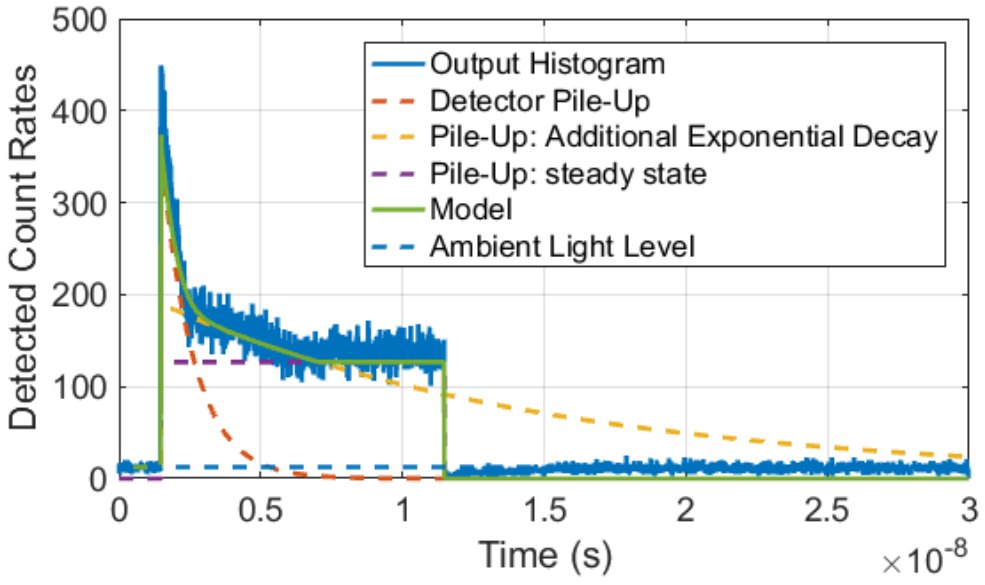


Figure 3.5.6: Saturated OR-based dSiPM distortions - When each single cells of the dSiPM enters the saturation region, the input rates of the OR network reduces significantly. This causes distortions due to detector pile-up. However, an additional distortion can be caused by the saturation of the network as well.

When an XOR network is used to combine SPAD cells digitally, the events are encoded of both rising and falling edges of the signals and no pulses are therefore merged together causing loss in detected events. Therefore, at first stage, the output of an XOR-based dSiPM can be written using (3.5.7). Repeating the simulations done for the OR tree but replacing the network with an XOR, the results shown in Fig. 3.5.7 are obtained. These plots can be directly compared to Fig. 3.5.5 since both refer to the same set of parameters written in Table 3.5. Already at $N_{cells} = 50$, Fig. 3.5(b), the higher throughput of the XOR can be appreciated as there is no evident distortion of the square pulse in opposition of the OR network. Overall, the shape of the waveforms remains a square pulse. However, there is still an artefact in this configuration that needs attention: the simulated counts in 3.5(d) deviate significantly from the ideal model proposed by (3.5.7). The reason is simple: an XOR-based dSiPM is always limited by the bandwidth of the single channel. This can be seen as how fast the counting/timing system is able to sample the detected the photons. As mentioned before in the Subsection 3.4.2, the count rate of an XOR tree can be modelled as (3.4.6). As result of this limited throughput, the

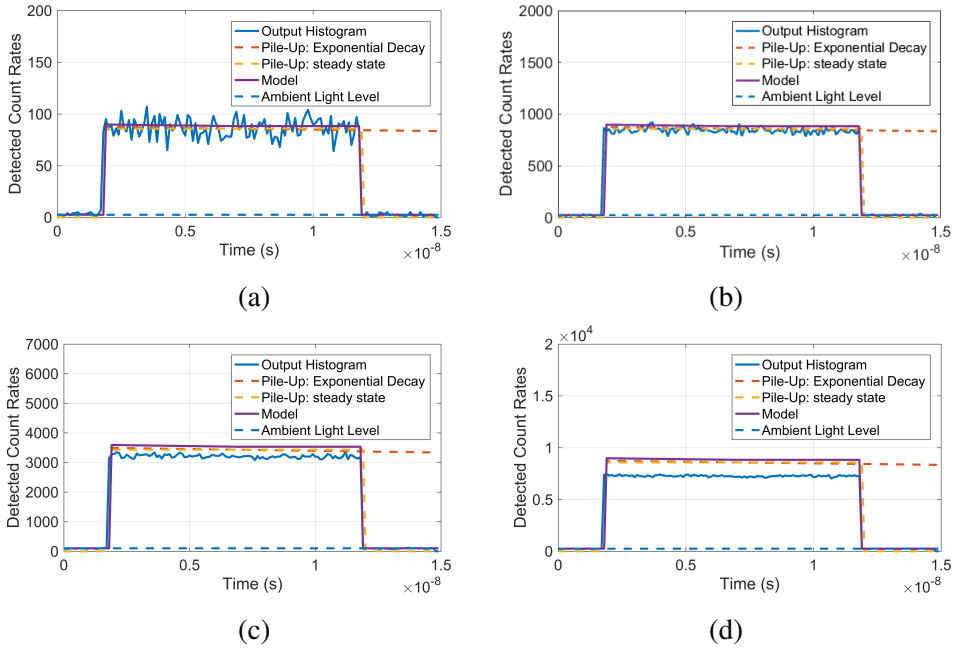


Figure 3.5.7: XOR-based dSiPM temporal response - The limitation due to pulse merging is overcome in XOR-based dSiPM. Higher throughputs are achieved and the pulsed waveform is maintained also at high number of cells. Obtained waveforms are simulated for an increasing number of cells: $N_{cells} = 5$ (a), 50 (b), 200 (c), 500 (d).

registered counts will not exceed a certain maximum. This can be included in the analytical model of the output waveform expression as:

$$W_o(t) = C_{BW,MAX} \times \left(1 - e^{-\widetilde{W}_o(t)/C_{BW,MAX}}\right) \tag{3.5.12}$$

For how the simulator is built, the bandwidth of the channel coincides with simulation time step which, for simplicity, also corresponds to the bin width of the TDC. Referring to Table 3.5, $T_{BIN} = 100$ ps. Therefore, two consecutives events can be simulated and processed within this time, therefore the bandwidth of this simulated channel is:

$$f_{BW,sim} = \frac{2}{100 \text{ ps}} = 20 \text{ GHz} \tag{3.5.13}$$

which is very close to be ideal, but it is sufficient to prove limitations. In Fig. 3.5.8, the updated model is compared to the simulated data only for the worst deviation case of Fig. 3.5.7 (d) $N_{cells} = 500$. The agreement between the analytical model and the simulated data is now highly improved thanks to the consideration on the bandwidth limitation for single channel dSiPMs.

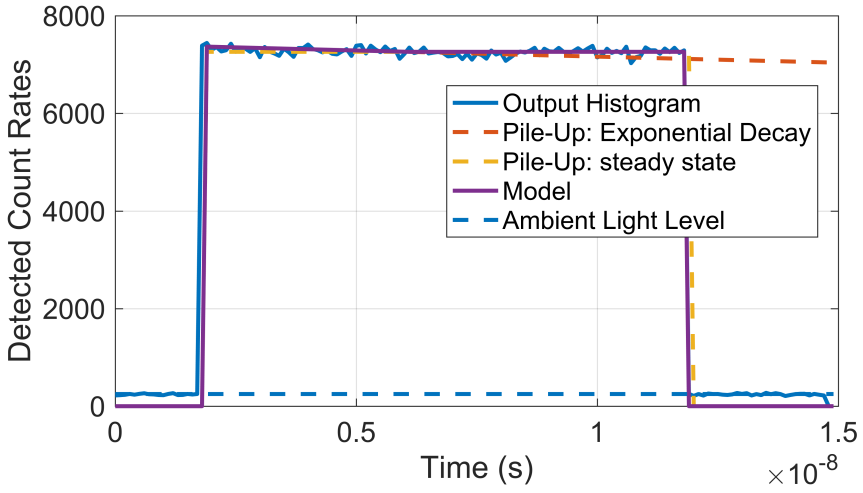


Figure 3.5.8: XOR dSiPM saturation - The high throughput of the XOR tree enters a saturation region when the event rates hit a maximum given by the bandwidth of the single channel. As results, the registered counts per each bin is lower than the ideal value. Here the simulated data are compared to an updated analytical model taking into account such limitation.

For such reason, there is a distortion of the square pulse although not evident: the height of the pulse has been reduced uniformly. In other words, the input square pulse has been distorted into a smaller square pulse.

A way to make this effect more evident is by showing the distortion due to XOR of an exponential shape rather than a square pulse. To do so, the shape of a saturated SPAD cell is utilised, see Fig. 3.5.9. With increasing number of cells, due to high throughput and limited bandwidth, the shape is gradually distorted into a square wave, see Fig.3.5.9(c).

3.6 Summary

In this chapter, a theoretical analysis on the response of dSiPMs has been provided. The analysis consisted in the realisation of a computational simulator to generate the digital output of configurable dSiPMs for any given input optical waveforms. The Monte Carlo approach provides a digital output signal which has been analysed in terms of average detected count rates and also in the time-domain. This has enabled the simulation of pile-up effects. The distortion of a square pulse and an exponential decay is shown and characterised through data analysis and theoretical modelling. The potential sources of pile-up demonstrated through this modelling have been identified in: SPAD cells (detector pile-up), triggering network (channel pile-up). Per each type, the conditions causing the photon loss have been identified: incident photon rate per unit cell, number of cells, digital network type and single channel bandwidth. Rules of thumb to prevent non-linear distortions have been derived from the obtained results. All the simulations have been run for both OR-based and XOR-based dSiPMs and for a selected number of SPAD cells. In all the cases, the higher throughput of the XOR has demonstrated in simulations the ability to reduce distortions due to channel pile-up.

Moreover, the results of the simulations have been paired with an analytical model enabling the description and the prediction of the shape of the output waveform from a reconstructed histogram for a given input. This result is of interest in those cases where pile-up cannot be avoided by a new design of the receiver and techniques of *pile-up corrections* can be developed [29–32].

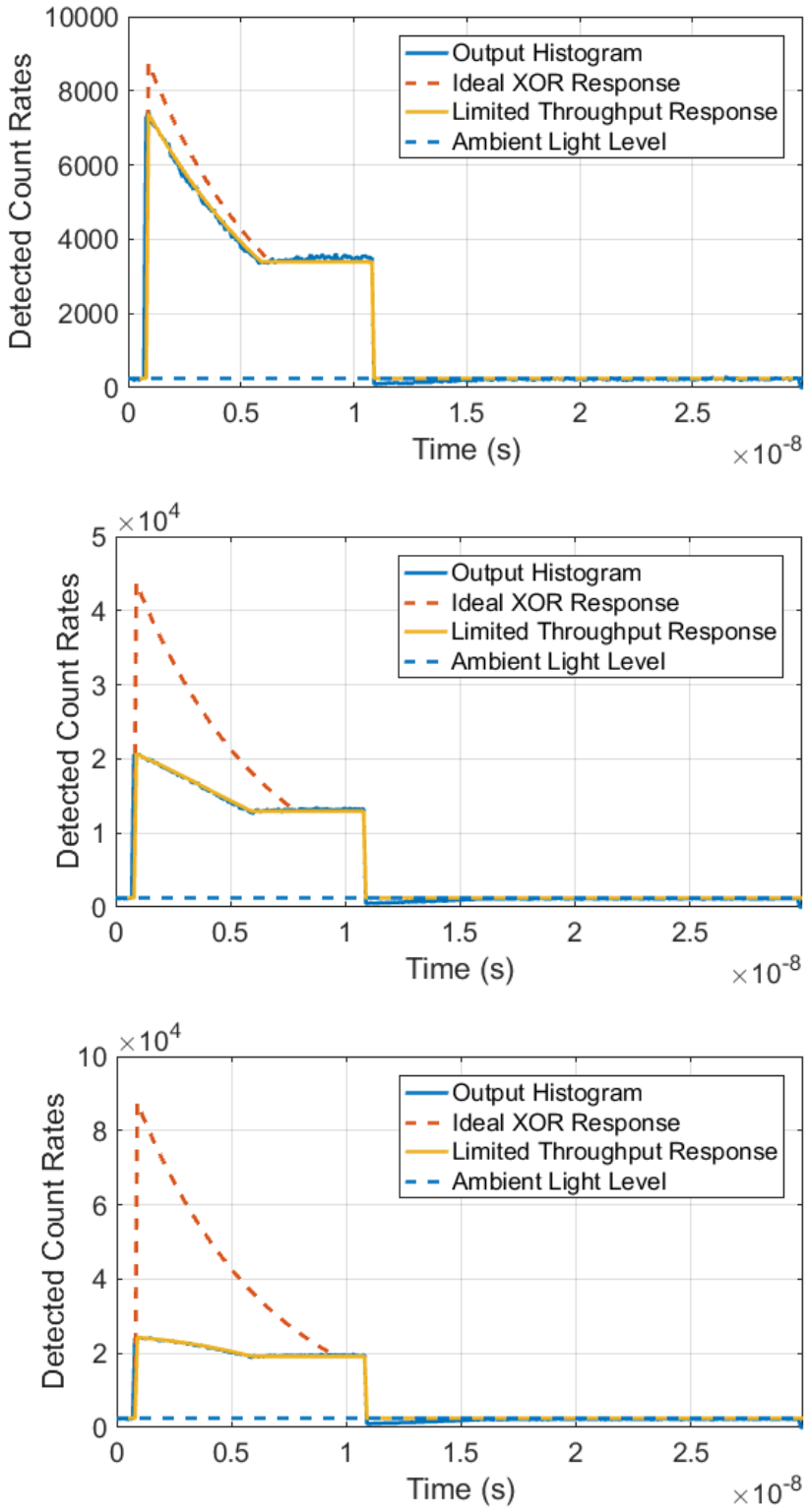


Figure 3.5.9: XOR dSiPM saturation - The high throughput of the XOR tree enters a saturation region when the event rates hit a maximum given by the bandwidth of the single channel. As results, the registered counts per each bin is lower than the ideal value. Here the simulated data are compared to an updated analytical model taking into account such limitation. The plots correspond to $N_{cells} = 10$ (a), 50 (b) and 100 (c).

This modelling analysis will be validated later on in Chapter 5 with a particular attention on the XOR-based dSiPM thanks to two test chips described in details in the following chapter. Extensions of the proposed simulator are discussed in the final Chapter 6 in the future work Section 6.3.

Chapter 4

Test Chip Description

Typical TCSPC systems are limited by the conversion time of the sensor and therefore they are commonly used in applications where the low light intensity does not require to capture more than one photon every laser pulse. In fact, high precision TDC circuits process only one event per laser cycle setting the typical requirement on TCSPC systems. In the long inter-arrival time between the detection of the photo-events, which is required to exceed a laser period, TDCs implement techniques to enhance the temporal resolution down to picosecond level [162, 164, 184] . However, a variety of applications where the precision of the TCSPC technique is required, do not satisfy this strong assumption. For example, PET and FLIM return signals are usually characterised by an intense burst of light at the very start before starting the typical exponential decay and classic TCSPC fails to correctly acquire the initial part of the signal effectively distorting the captured light profile. Another example is offered by outdoor LiDAR systems where the ambient light might prevent the acquisition of the return signal due to its high intensity. All of these applications would benefit from faster conversion times, i.e. higher throughputs of the TDC since they would enable the capture of multiple photons within a given time thus avoiding the so called *pile-up* effect.

This chapter starts with the presentation of a test chip where a novel TDC architecture able to process more than a single photo-event per laser cycle has been implemented. This circuit will be referred throughout the publication as Multi-Event Time-to-Digital Converter (METDC). The proposed architecture has been designed with a sampling frequency of $\sim$ 14 GSample/s and no conversion dead time and therefore enabling multi timestamps to be recorded within the same laser pulse. The test chip consists of a configurable SPAD array combined digitally onto single channel through an XOR network. On-chip counters allow histograms to be built

directly on the chip and streamed via firmware onto the controlling software. This test chip has demonstrated the overcome limit of conversion time in high incident light applications. An initial description and results from this test chip have been published prior to the publication of this thesis [7].

Having solved the bottleneck of the conversion time, the work has been shifted towards the analysis of the optimisation of the detector side of the sensor, i.e. the SPAD array. A high TDC throughput has to be matched with the detector in order to maintain a high sensor throughput. For such reasons, a second test chip has been manufactured in order to have the ability to test different configurations from a variety of parameter values: size of the SPAD cells, number of activated cells, digital N-to-1 network and different channel bandwidth. This configurable dSiPM chip has been the key tool to verify the variety of assumptions and results obtained by the theoretical and computational analysis proposed in Chapter 3. Part of the results have been presented in published articles [3, 4].

Both test chips have been realised in collaboration with STMicroelectronics whose team has provided considerable help in terms of reviewing process in the implementation of the circuit blocks. The STMicroelectronics 0.13 μm 1P4M imaging CMOS technology node is common to the two test chips.

The Chapter is divided in two main sections, one per each test chip, and contain a complete description of the relevant features available used for the purpose of the scope of this work.

4.1 Multi Event Time-to-Digital Converter

The chip here presented, pictured in Fig. 4.1.1, consists of:

1. 32×32 SPAD array
    - SPAD + passive quenching and enabling circuitry (*)
    - In-pixel toggle cell (*)
    - In-pixel column XOR gate
2. Toggle + XOR 1024-to-1 network
3. Multi-Event TDC

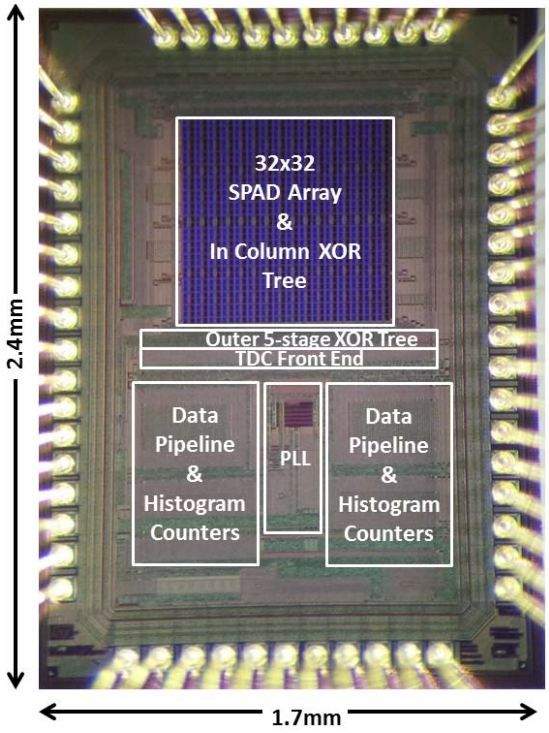


Figure 4.1.1: METDC test chip - A picture of a high throughput TDC implemented together with a SPAD array combined onto single channel by XOR gates.

4. Digital memory for on-chip histogram composition (*)
5. Read-out (*)
6. Serial-to-Parallel input Interface (input SPI) (*)
7. Parallel-to-Serial output Interface (PSI) (*)

The elements highlighted with a (*) represent the contribution of supporting members of the team at the University of Edinburgh who have worked together with the author to the realisation of the test chip. Such contributions are here presented at higher level in order to concentrate the description of the test chip on the core of the METDC design, which the author focussed on. As pictured in Fig. 4.1.2, the output of the SPAD array is combined onto a single channel by two 5-stages XOR-trees, one in-pixel per column, and one external to the array. The single channel is sampled by 33 phases of a clock generated through a PLL and the front end of the TDC detects the events encoded onto every transition of the input signal. A histogram is generated on chip thanks to 264 ripple counters. A timing diagram is presented in Fig. 4.1.3 showing the detection process: the laser clock driver periodically illuminates the scene, the return photons

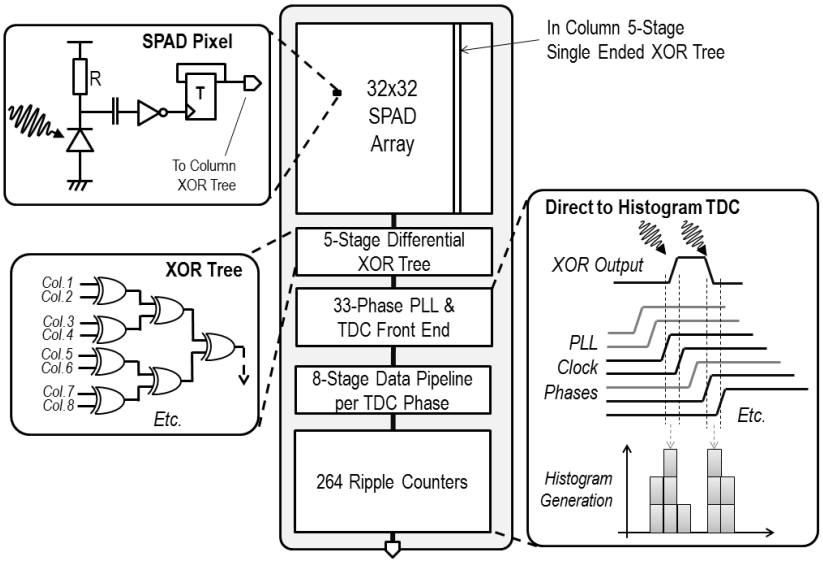


Figure 4.1.2: METDC diagram - The main blocks of the METDC test chip are pictured showing the detection chain from the detector array to the main TDC front end as well as the histogram formation directly on chip. [7]

are captured onto the single channel by the SPAD array and XOR tree where the front end of the TDC samples the signal of 33 phases. Each one of the sampled signal is compared by the XOR ring of the TDC detecting in parallel all the photons. A pipeline then stores these detected photons for 8 clock cycles after which the data is sent to the counters, in sync with the next laser pulse, in order for the histogram to be formed . The next sections describe in details the main components of the sensor.

4.1.1 SPAD Array

The light sensitive part of the sensor consists of SPAD cells of 21 µm $\times$ 21 µm pitch organised in a 32×32 array. The Deep N-well to P-substrate structure of the diode, previously published by Webster *et al.* [112, 185], gives the diode a higher photon detection probability in the infrared spectrum and it is paired with a quenching resistor, an in-pixel 8-transistor buffer, a 12-transistor T-type flip flop cell, a single ended XOR cell and a 6-transistor SRAM for region-of-interest (RoI) selection through input serial interface. The SPAD cells reach a fill factor of 34%, its layout is shown in Fig. 4.1.4 while the schematic is presented in Fig. 4.1.5. An alternative version of the system, realised later on in the project with the assistance of PhD candidate Tarek Al Abbas, sees the SPAD structure being replaced with a more blue-sensitive diode published

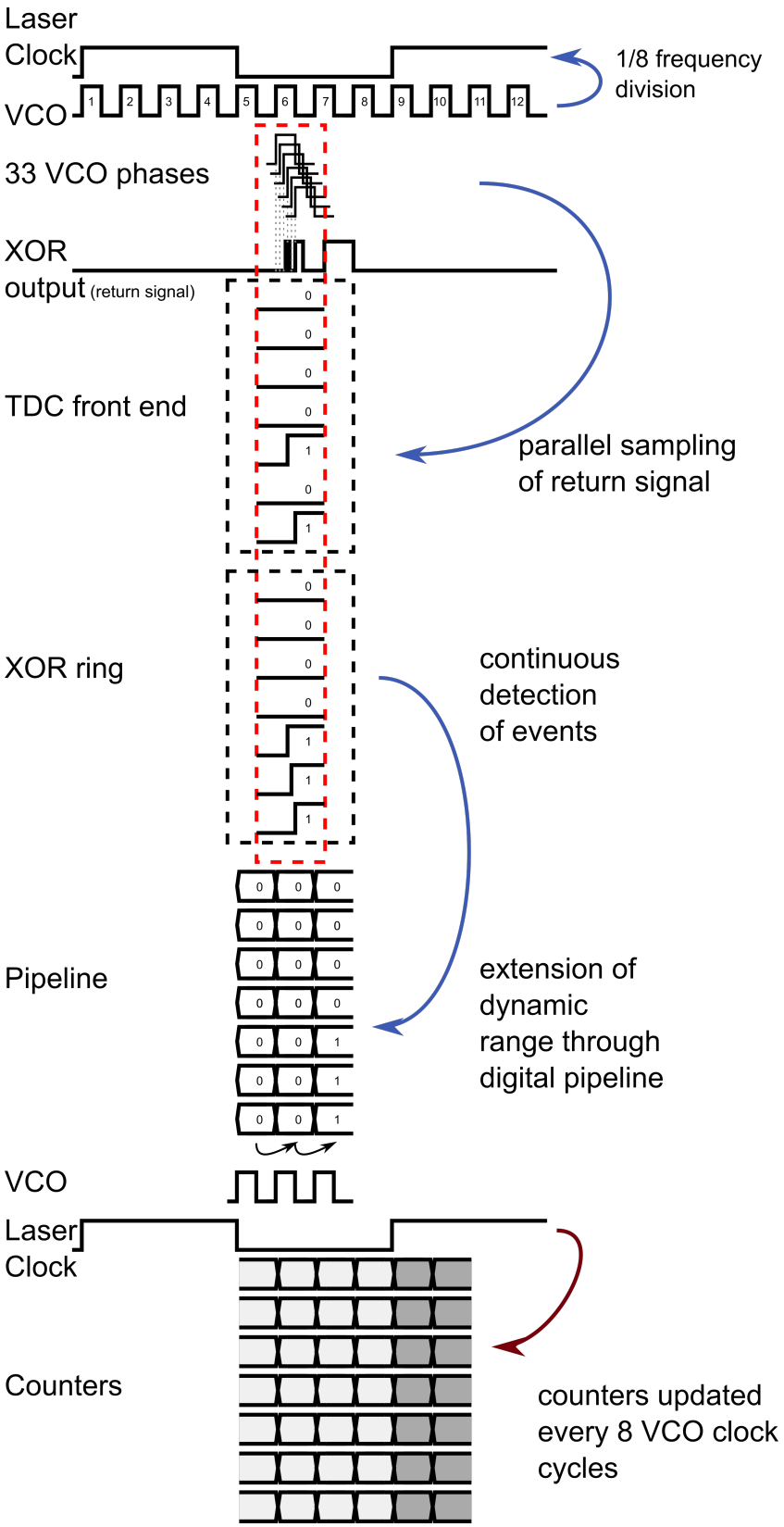


Figure 4.1.3: METDC timing diagram - The VCO generates 33 phases that sample in parallel the return signal coming from the XOR-based dSiPM. A folded XOR ring detects continuously events as variations in the sampled signals. At every clock cycles, these values are shifted over a 8-bit pipeline to extend the range. Every 8 cycles, the pipeline is passed onto ripple-counters effectively forming the histogram on chip. The readout of the counters is done sequentially on a different clock domain and after the exposure has ended.

by Richardson *et al.* [186]. The architecture here presented is common to both test chips.

At every photon detection, the voltage of the cathode, VSPAD in Fig. 4.1.5(a), drops in a fall time of $\sim$ 100 ps, the passive resistor then starts the quenching process restoring the voltage back to its initial value in an RC constant value that determines the so called dead time of the SPAD cell τ_{dead}. The voltage on the cathode is then coupled with a metal fringe capacitor to the input of an internal buffer which also shifts down the voltage to be then connected to standard 1.2 V digital logic. The now-digital pulse then goes into the T-type flip-flop, 4.1.5(b), which encodes every SPAD detection onto both rising ad falling edge. The output of the flip-flop goes into the flatten H-tree column XOR.

Although the SPAD array contains 1024 SPAD cells, a smaller portion of it might be of interest for analysis or performance purposes. Therefore, an S-RAM circuitry has been included on each cell. The input comes from the input serial interface and allow enable/disable signals to

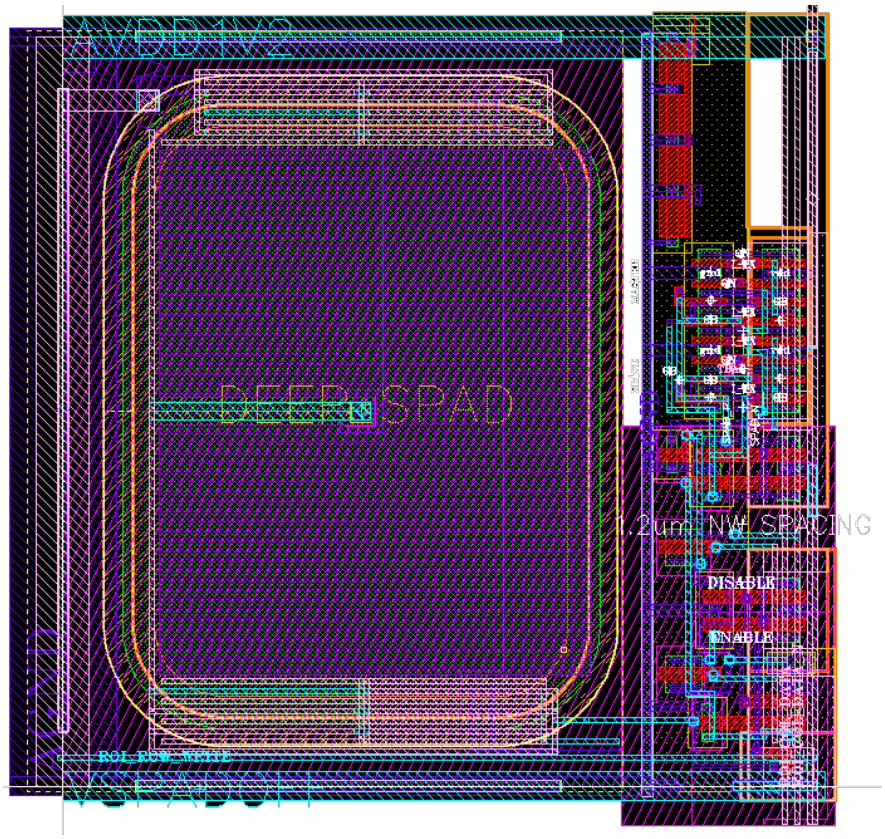


Figure 4.1.4: SPAD cell - Layout of the unit cell used in the detector of the METDC sensor chip. The space at the top right of the pixel is reserved to the column XOR-cells.

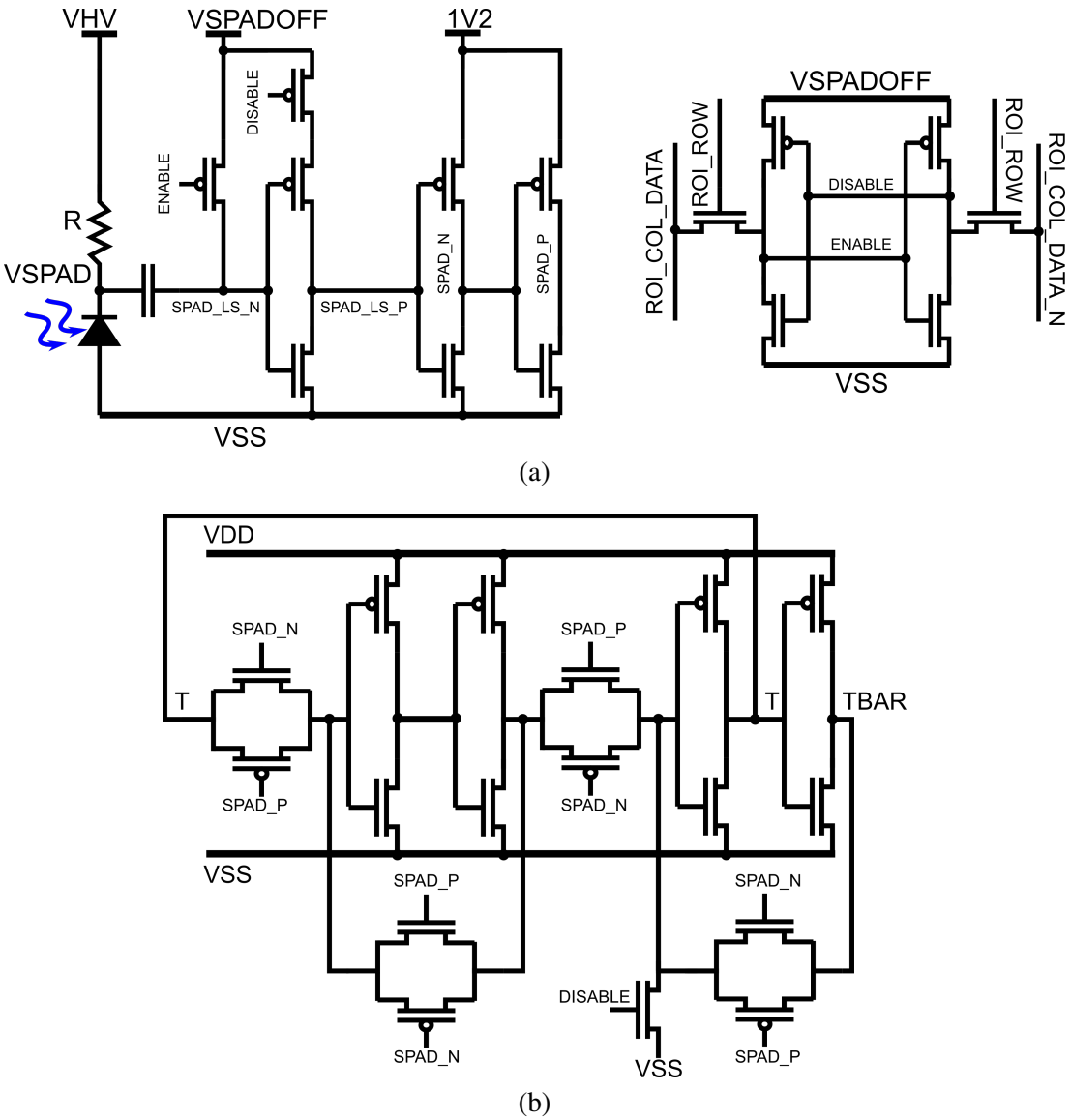


Figure 4.1.5: SPAD cell - Schematic of the SPAD cell including on-chip circuitry: quenching resistor, level shifter/buffers, SRAM for cell enabling (a) and on-chip toggle-cell (b).

be locally generated. At the sensor start-up, the entire array is disabled, then sequentially the RoI is set selecting the row/column address of the desired SPAD(s). Programming the RoI therefore needs certain cycles of the input serial interface and it is generally done at the start of the experiment prior to the exposure. Thanks this programmability, SPAD cells can be inhibited to prevent any change in the SPAD voltages and therefore transitions to be propagated to the XOR tree. This is very critical in order to maintain the current due to incident light to be proportional to the number of activated cells which can be crucial in case of high light levels.

4.1.2 XOR Network

The XOR 1024-to-1 channel network represents one of the innovation proposed in the sensor as this replaces the more common OR counterpart. Each SPAD event is here encoded onto the cell output as both rising and falling edges through a T-type flip-flop connected to the internal buffer. This differs from an OR-based dSiPM where the output of the buffer is time-compressed by monostable circuits and therefore the events are encoded only on rising edges. The structure of the T-type cell has been robustly designed with 12 transistors to avoid instabilities that might cause severe oscillations on the power supplies and the output itself, at the expenses of a lower fill-factor. The output of the T-type cell is then connected to the first stage of the XOR-tree. This is in fact split in two levels. The first one represents a column, in-pixel flatten H-tree for time uniformity, see Fig. 4.1.6, and it is kept single-ended not to affect further the fill-factor of the cell. The array is therefore seen as 32 columns of 32 cells combined by the in-pixel XOR tree. The second level of the tree is placed outside the array. Since this does not affect the fill factor of the array, the cells have been designed with more details to enhance the time-balance among column events. The output of each column XOR is converted to fully differential by an inverter and a latch circuit to align the differential outputs. The external XOR is therefore fully differential. Again, it is designed as a flatten 32-to-1 H-tree with increasing size. This is done to make the tree robust to a high load: the output of the tree is going to see the input capacitance of the METDC core as load.

The full layout of the detector (SPAD array and XOR tree) is shown in Fig. 4.1.8.

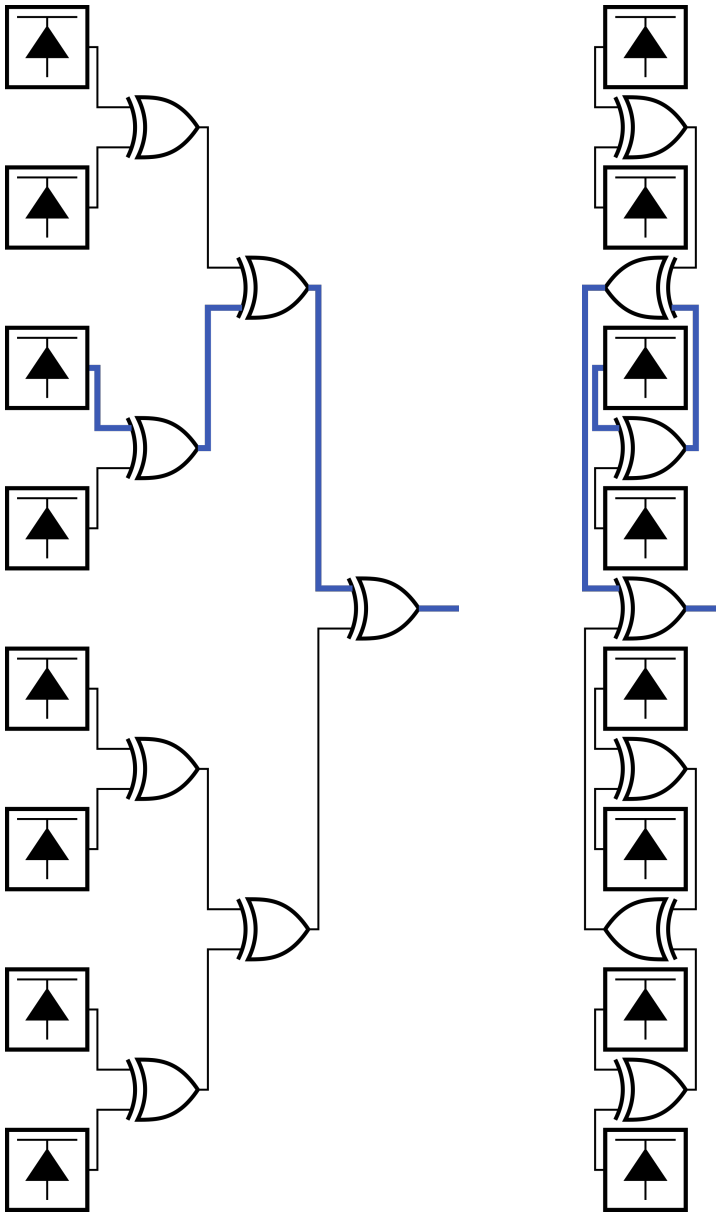


Figure 4.1.6: XOR H-tree - The timing path of the output of each SPAD in the column is balanced through an H-tree geometry (a). To improve the fill factor of the array, a flattened H-tree (b) is adopted to ensure the minimum distance among columns. The blue line highlights the equalised signal path.

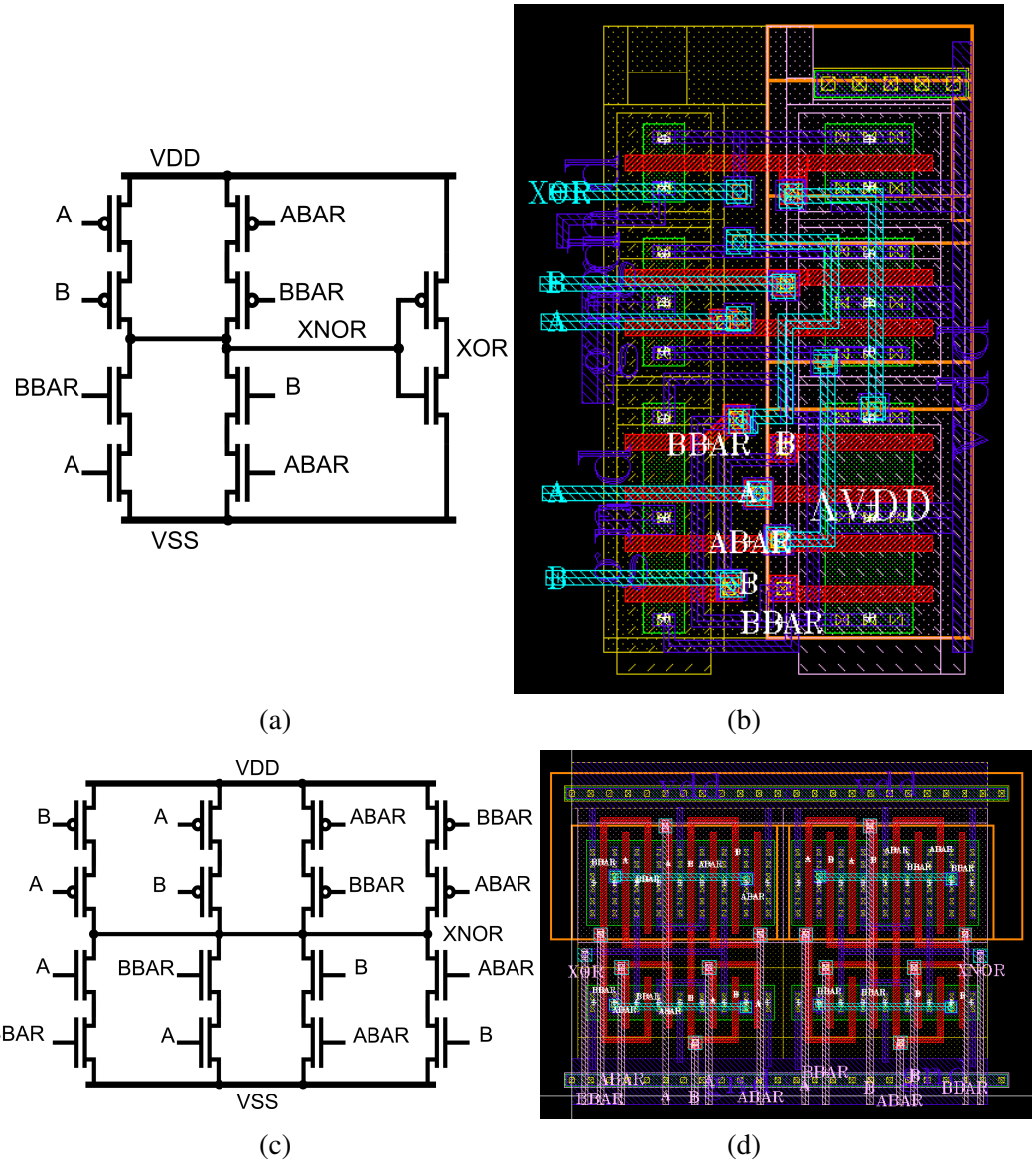


Figure 4.1.7: XOR cells in METDC chip - The XOR network is divided in two levels. The first one is a column XOR based on a single ended cell (a,b) while the second level, combining the 32 columns, is external to the array and consists of symmetric cells creating both XOR and XNOR signals (c,d).

4.1.3 METDC

The Multi-Event TDC, pictured in Fig. 4.1.9, consists of a flash architecture based on a phase locked loop (PLL) where 33 phases of a clock are used to sample in parallel the signal coming from the XOR output. After the sampling, the occurrence of events is detected by an XOR ring which folds the first phase to the last one ensuring continuous sampling and zero conversion dead time. Moreover the readout and histogram generation are also performed in parallel inside the chip itself. The structure of the TDC can be divided into:

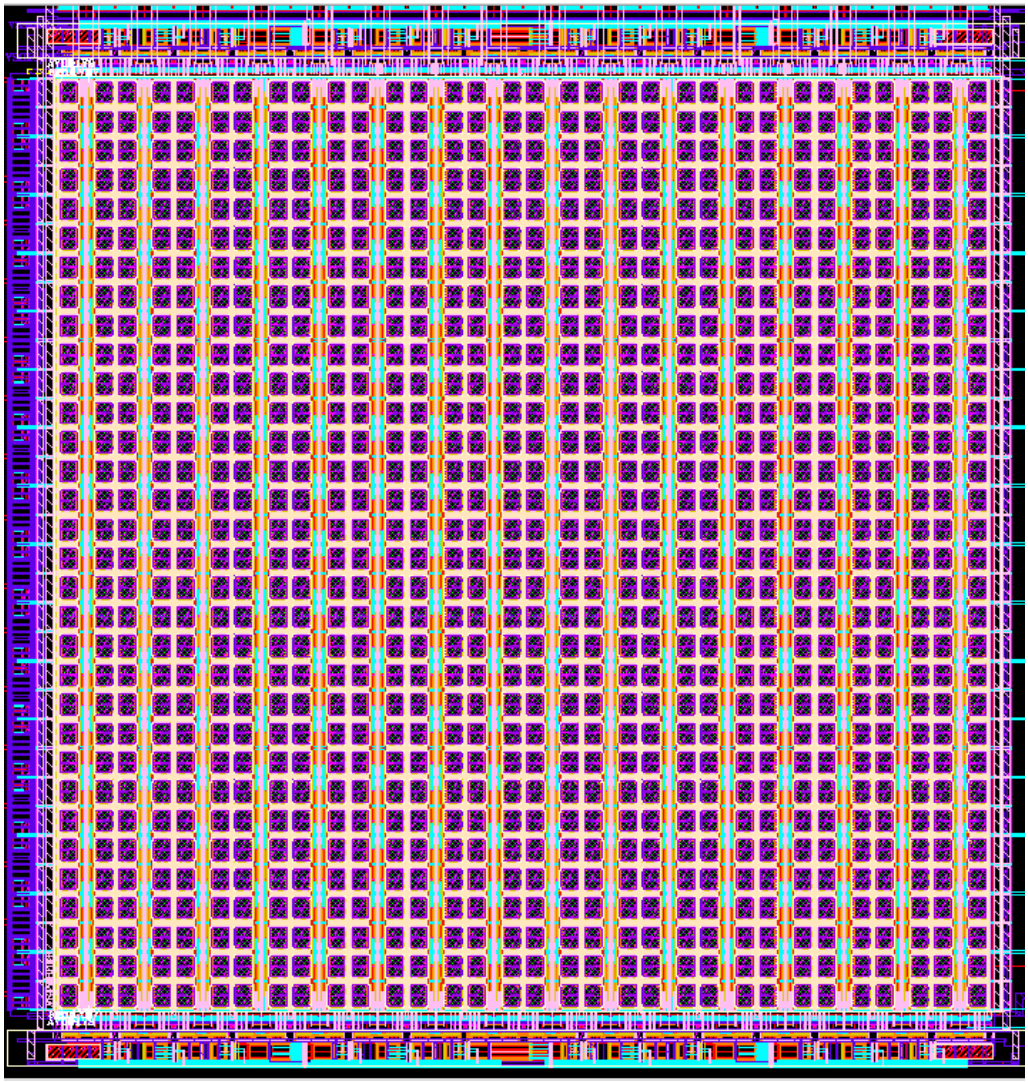


Figure 4.1.8: METDC SPAD array- Layout of the 32 $\times$ 32 SPAD array with internal column XOR-tree and external XOR-tree for column combination. A second external XOR-tree has been added to provide a direct output signal onto a dedicated pad.

- PLL - clock phase generation
- Front-End - sampling
- Digital pipeline - dynamic range extension
- Digital memory - histogram generation

Each of these components are described in the following paragraphs.

Phase Locked Loop

The core of the Phase Locked Loop (PLL), pictured in Fig. 4.1.10, is a differential voltage controlled oscillator (VCO) where 33 current starved differential inverters act as delay elements in a closed oscillating loop. The output of the VCO is a series of 33 clock signals with a frequency given by

$$F_{VCO} = \frac{1}{\tau_{delay} \times 33} \tag{4.1.1}$$

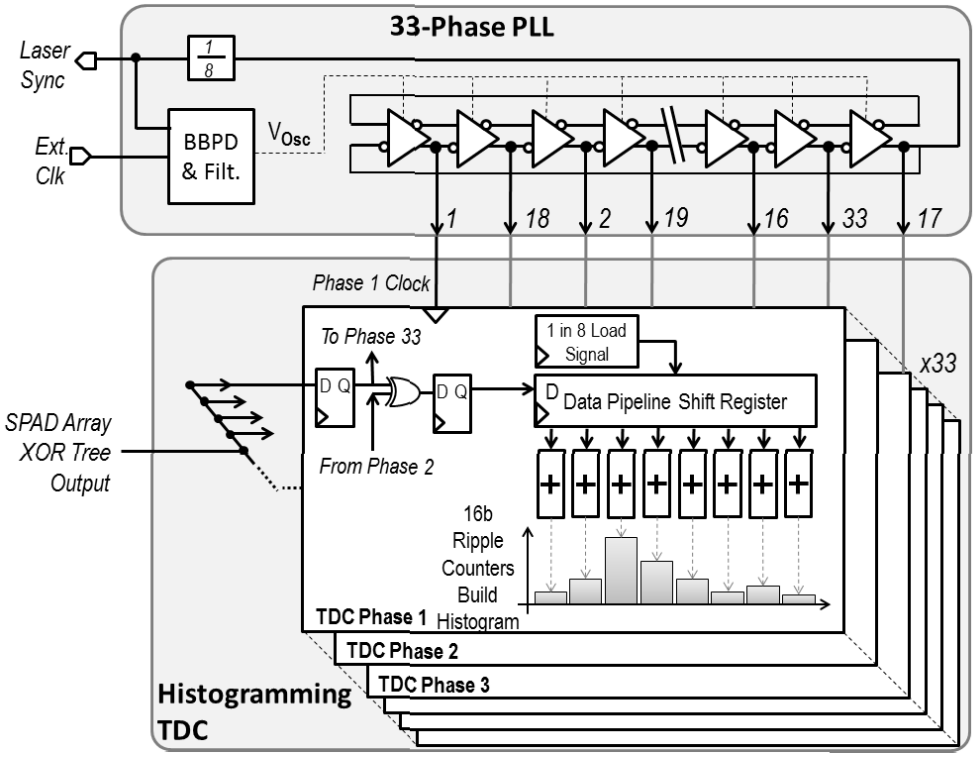


Figure 4.1.9: METDC core - 33 flip-flops sample in parallel the digital output of the XOR combined SPAD output. An XOR ring detects the transitions between consecutive clock phases. The output is connected to ripple counters. [7]

where τ_{delay} is the delay of the inverters in the oscillator. To lock this value, one of the clock phases is sent to a feedback loop in the PLL: the feedback clock signal is compared to an external clock signal, in this case coming from the FPGA, with lower frequency, therefore a programmable frequency divider (set to a 1 : 8 division) is present in the loop. The two clocks are the input of a bang-bang phase detector (BBPD) and a charge pump. Together, these circuits generate the control voltage for the VCO in order to reach the locking condition and therefore set a stable τ_{dealy}. The voltage is filtered through a passive filter to improve the performance of the PLL. The VCO is presented in Fig. 4.1.11(a) where the control voltage sets the current of each starved inverter in the oscillator. Due to the large number of inverters in the oscillators, the PMOS in the design are very large and take most of the layour space, see Fig. 4.1.11(b). Once the 33 phases are generated, the signals are shifted up to the full voltage swing so that they can be used to sample the signal in front-end of the TDC. The clock phases are routed to the flip-flops of the sampling part of the TDC through a balanced H-tree.

Front-End

The sampling is made by a pair of differential D-type flip-flops in sequence for metastability rejection. The architecture for each cell has been chosen to be a high-resolution symmetric cell selected from a literature review [161]. The schematic and the layout are presented respectively in Fig. 4.1.12(a,b) and (c). The sampled waveform output of each flip-flop is XORed with the neighbours output: in this way, a transition in the common input signal between two consecutive

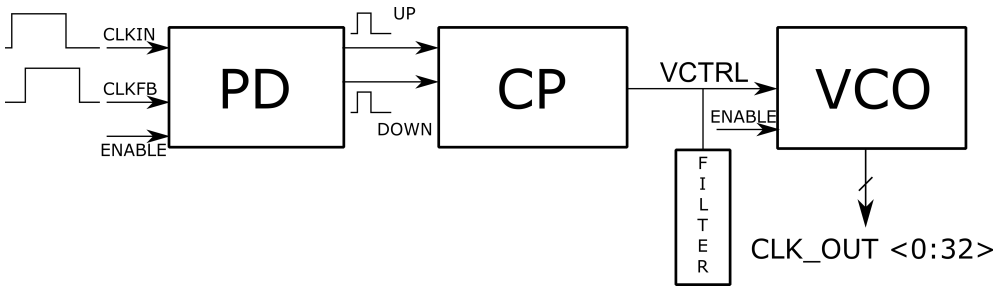
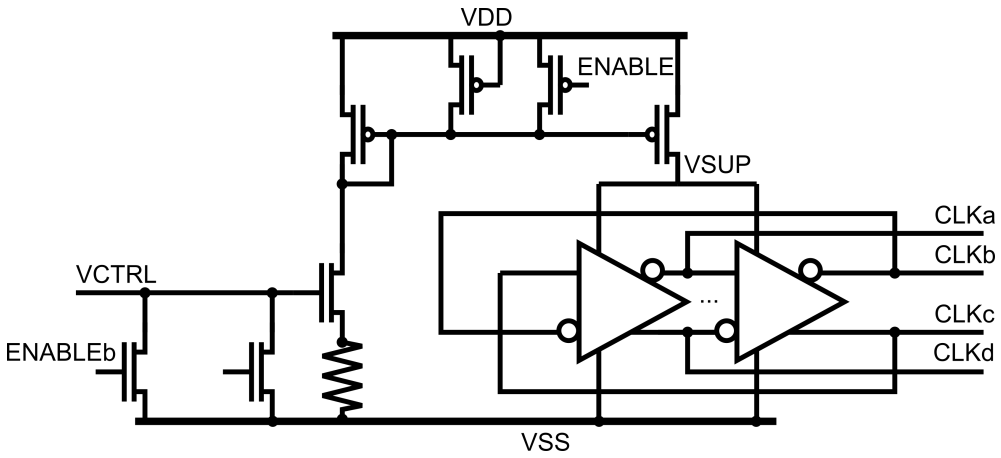
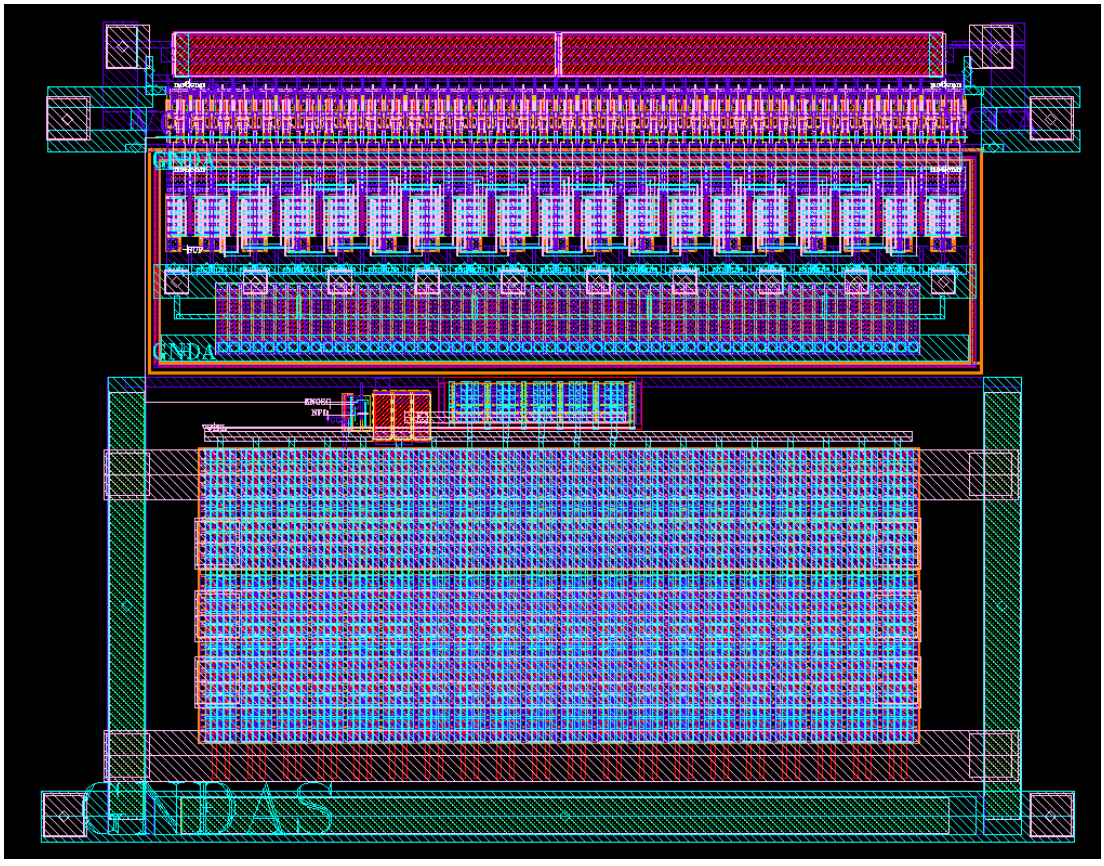


Figure 4.1.10: METDC PLL diagram - The generation of the 33 phases of a clock is done by a voltage controlled oscillator put in a phase locked loop. The PLL is based on a phase detector (PD) comparing the current clock signal with an external reference, which generates UP and DOWN signals to a charge pump (CP). This one generates the necessary voltage for the VCO.



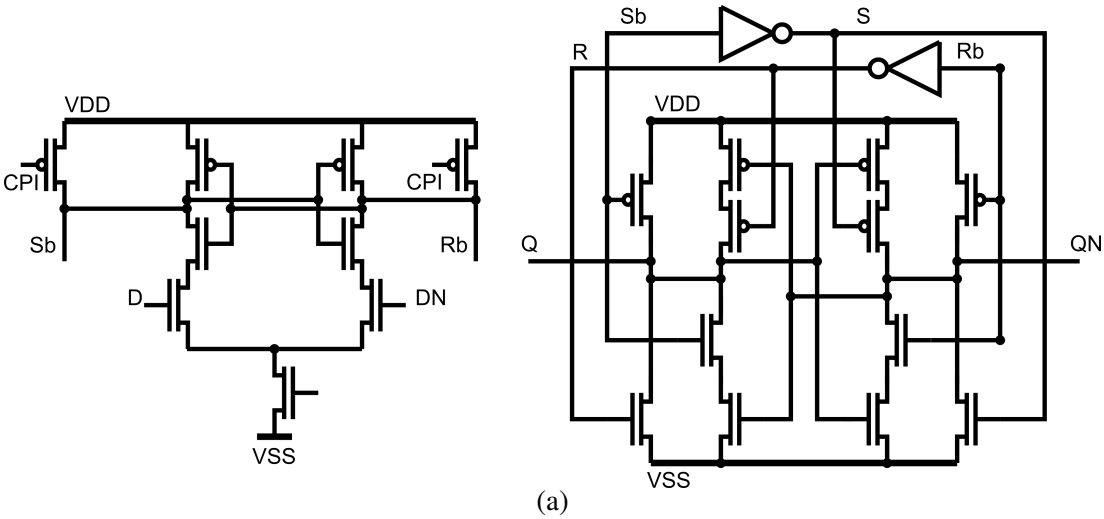
(a)



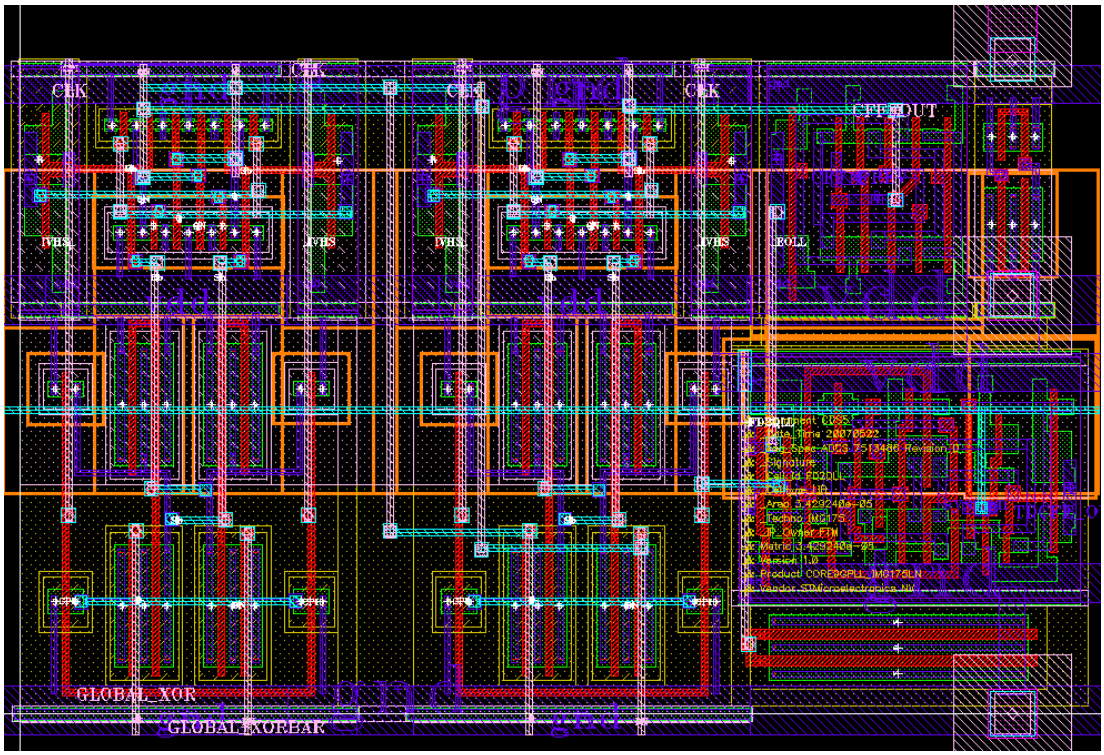
(b)

Figure 4.1.11: METDC VCO - A 33 stages differential VCO is included in the METDC chip. The schematic (a) shown how the VCTRL voltage sets the upper voltage of the delay elements. The output nodes are sent to level shifters to be brought back to the full supply swing before being connected to the front-end of the METDC.

clock phases is detected at the output of the correspondent XOR gate as a digital 1. The XOR ring has been designed with low leakage standard cells since no particular constraints were foreseen by the design. In order to increase the throughput of the converter and ensure a



(a)



(c)

Figure 4.1.12: METDC sampling flip-flop - A delay flip-flop is used to sample the input signal per each clock phase. A replica cell is connected in series to reject metastability errors. Both the schematic (a) and layout (b) are here presented.

zero-dead time and continuous sampling operation, the last of the flip-flops is XORed with the first of the chain creating therefore a folded loop. This sampling mechanism cancels the necessity of clearing the chain, as in typical delay-line based flash architectures. At every clock period, the output of these XOR gates are summed together through ripple counters which have a controllable enable and reset signal. These signals have two dedicated pads which can be controlled by firmware/software. In this way, an exposure/acquisition time can be set by sending a square pulse with a pulse width as wide as the desired exposure time. The counters must be reset before every acquisition and after the readout.

Digital pipeline

The natural dynamic range of the METDC, i.e. the difference between the maximum and the minimum time stamp produced, is equal to the clock period. To generate a long dynamic range, a low frequency clock must be used and therefore, in order to maintain the time-step, i.e. the TDC bin, a longer VCO must be utilised. However, an alternative solution has been here adapted to avoid the need of a longer VCO. To extend the dynamic range by a factor of 8, a digital pipeline has been adopted. As shown in Fig. 4.1.13, each shift register works on the

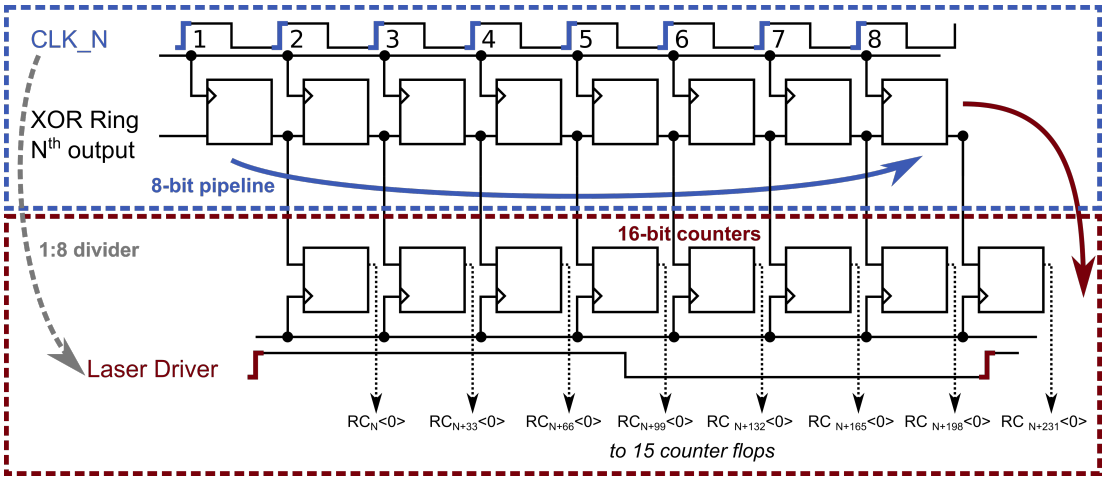


Figure 4.1.13: METDC pipeline timing diagram - The output of every one of the 33 stages of the XOR ring goes into a pipeline, which is a 8-bit shift register on the correspondent clock phase. For 8 clock periods, the output of the ring is shifted through the pipeline therefore being stored of 8 clock periods. Then, thanks to the 1:8 divided clock cycle, all the stages of the pipeline are passed to the counters, which work on the slower clock to be then counted through the 16-bit ripple counters. The different timing domains are highlighted by a different colour to help visualise the timing in the pipeline system.

clock domain of the correspondent *N-th* phase. For 8 cycles the output of the XOR ring goes through the series of flip-flops to be temporary stored in the pipeline. The 33 bins generated at each cycle, by the 33 phases, are therefore digitally stored to create a total of $8 \times 33 = 264$ bins thus extending the dynamic range of the TDC. Every 8 clock cycles, thanks to a divided clock, each and every node of the pipeline is clocked into a series of 16-bit ripple counters. In a general sensor view, the laser pulse can be controlled by the 1:8 frequency divided clock signal to match the dynamic range extension. This clock is generated dividing one of the 33 phases of the VCO output through a frequency divider circuit, see Fig. 4.1.9.

Digital memory

Every 8 clock cycles, the data stored in the pipeline is ready to be passed onto the counters that forms the histogram. These are 16-bit ripple counters designed to represent the bins of the TDC. The flip-flops of the counters have the ability to be enabled through a common signal that it is used on system level to effectively set an exposure time to the sensor. Although the actual detector is free running, the acquisition of the histogram can be set by the counter enable signal. This is created by the FPGA and for pulsed light measurements has to be synchronised with the laser signal. For example, the exposure time can be set as a multiple N_{cycles} of laser cycles and therefore a square pulse with a width of $N_{cycles} \times T_{laser}$ must be generated by the FPGA and sent to the ripple counter enable which has got its own dedicated pad.

The counters represent the most area-intensive part of the read-out (second only to the SPAD array). The creation of the histogram on-chip allows faster acquisitions since it cuts out the need of sending the data outside the chip but it requires a large silicon area to host the ripple counters. The 16-bit size has been chosen in order to prevent saturation for reasonable small acquisition times. The extension of the dynamic range has the unavoidable consequence of requiring 8 times more counters. To avoid the connection of 264×16 bits of every counter per bin, a single set of 16 pads has been used to read the histogram data.

An additional circuitry performs a serial scan of the 264 counters connecting each time the 16 bits to the common output bus.

4.1.4 Input Serial-to-Parallel interface

To reduce the requirements on the number of pads of the chips, a list of digital configuration parameters have been provided to the chip through a serial interface. In this way, the number of required pad has been locked to three: a data, clock and load line. These parameters include:

- Region of Interest for the SPAD array (row and column selection)
- Histogram reset signal
- Histogram read signal
- PLL enable signal
- Frequency divider configuration bits

The interface has been realised with standard cells and the architecture is not here presented. However, intense work from Dr. Neale Dutton and Luca Parmesan has been highly appreciated in the process of layout of such circuitry to ensure a good performance of the sensor.

4.1.5 Overview and Design Achievements

The sensor here presented has demonstrated a high throughput time-to-digital converter allows parallel sampling and readout with on-chip histogram generation. The sampling frequency is set by the minimum time propagation of a differential delay element. For the operation of this converter, the following parameters have been achieved through design: ~ 71 ps bin width, ~ 14 GHz of sampling frequency, ~ 53 MHz locking frequency for the PLL, ~ 18.8 ns of dynamic range. To the date of the publication of this document, this architecture represents the highest throughput TDC realised for imaging. The scope of the rest of this thesis is to prove why this is a fundamental advantage to dSiPM receivers and what the further steps are in the maximisation of the global throughput of the sensor.

From this point of view, the realisation and testing of the METDC test chip has highlighted how the overall throughput of the sensor is no more limited by the conversion rate of the TDC but instead the bottleneck of the receiver is the detector side of the sensor, i.e. the SPAD array connected onto the single channel. The maximum event rate recorded from this test chip, which will be presented in Chapter 5, is much lower than the throughput of the TDC which therefore

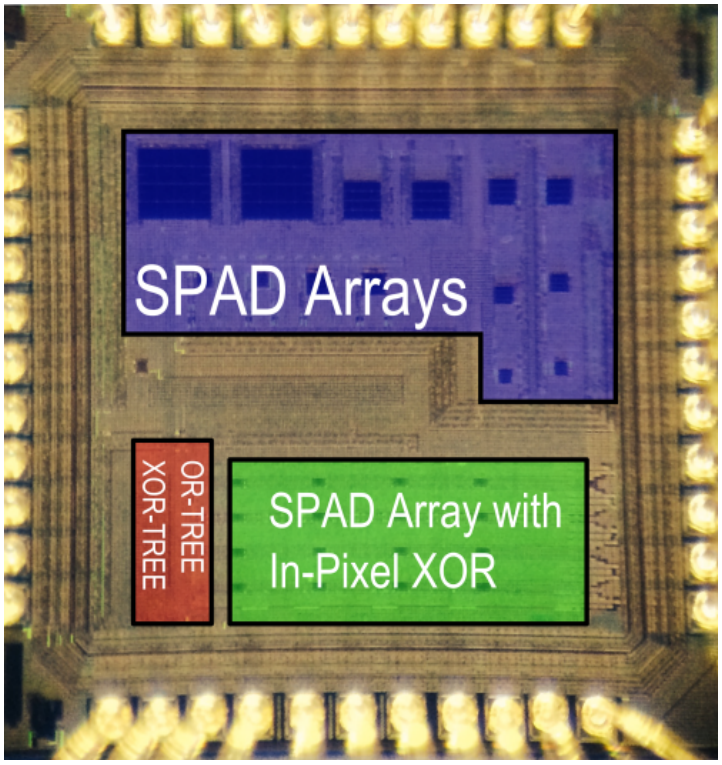


Figure 4.2.1: Configurable dSiPM picture - A programmable dSiPM chip allows the selection of different SPAD cell size and number together with the digital N-to-1 network. The picture of the chip is here proposed.

is a proof of the unmatched detection rate of the SPAD detector with the TDC sampling rate and throughput.

4.2 Configurable dSiPM Chip

After the realisation of the METDC chip and the achievement of the high throughput, the analysis has moved to answer how to ensure a high detection rate from the detector in order to maintain and exploit the overall sensor throughput. In other words, the main questions are: what are the bottlenecks in the SPAD array design; which are the options among different design choices and their impact on the detector performance; and what are the real requirements of a given SPAD array in terms of converter specifications? All these open problems have been approached on a theoretical level thanks to the model described in Chapter 3, however, the analysis has been validated through concrete design thanks to a configurable dSiPM chip, shown in Fig. 4.2.1, that has been specifically designed to validate the theoretical results.

The main requirement for this test chip has been the reconfigurability: a selection of dSiPM parameters are available to the user in order to directly compare, on the same chip, the impact of each one of them. The reconfigurability can be summarised by the following list showing the choice for each desired configuration:

- Programmable SPAD quenching voltage for dead time tuning
- Different SPAD cell size: $7-9-13-18-34$ µm cell pitch
- Programmable number of activated SPAD cells: a 4×4 array per each size plus a dedicated $1-256$ option of the 7 µm pitch array with XOR network
- Selectable OR-tree and XOR-tree external to the arrays - with the exception of integrated XOR-tree in the 16×16 array.
- Voltage controlled monostable circuits for on-chip OR-tree
- On-chip counters of high bandwidth acquisition
- Dedicated pad for direct analysis or external connection to off-chip counters

The choice of the SPAD cell pitch has been chosen in order to keep the work relevant with the SPAD design trend so that both a small pitch (7 µm) and high pitch (34 µm) could be realised on the same chip. Higher pitch values have been left out in order to maintain a reasonable compact chip size. The choice of the array size has been instead driven by the results of the modelling analysis as well as the measurements from the METDC test chip. In fact, knowing the bandwidth limitation for a single channel in the given technology, measured to be $\lesssim 1$ GHz, and the foreseen minimum dead time for the SPADs of ~ 5 ns, the optimal number of SPAD cells in terms of throughput has been calculated to be around 13 and therefore the chip has been populated by 4×4 arrays. An exception has been made by inserting in the chip a 16×16 array with dedicated XOR tree to better analyse the effect of aggregating too many cells and also to better link the results to the METDC chip. A general diagram is provided in Fig. 4.2.2. In the following subsections, each part of the sensor is presented.

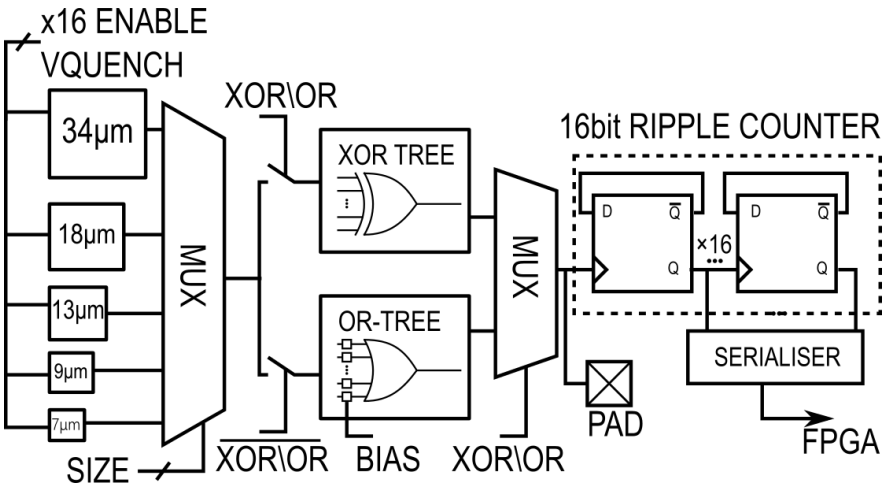


Figure 4.2.2: Configurable dSiPM diagram - The diagram of the sensor is here shown highlighting the main design options for a desired dSiPM configuration.

4.2.1 SPAD cell variants

Five variants of 4×4 SPAD array having different cell pitch, throughout this document, these variants will be referred as in Table 4.1. The structure of the arrays is common to all the variants **D2-D5** while the dedicated array **D1** has been designed specifically for a focussed analysis on XOR-tree dSiPMs. The 4×4 arrays are arranged to have the diodes close to together while the entire circuitry, i.e. enabling transistors, buffers and toggle cell, is placed outside the light sensitive area, this is shown in Fig. 4.2.3 with the example of the 7 µm array. The circuit differs from the one presented in the METDC: the SPAD chosen for this design is a conventional blue-sensitive SPAD [186] which requires a different circuitry. The main difference is that the

dSiPM	Unit cell pitch (µm)	Number of cells	Fill Factor (%)
D1	7	16×16	6.4
D2	9	4×4	18.7
D3	13	4×4	37.4
D4	18.62	4×4	73.6
D5	34.62	4×4	85.4

Table 4.1: DSiPM - Properties and Parameters of unit cells.

resistor for the quenching process is here realised with a transistor biased with an external voltage called VQUENCH which enabled dead time tuning. The schematic of the circuit is presented in Fig. 4.2.4. The architecture of the toggle cells is equivalent to the one presented in Fig. 4.1.5(b). For completeness, the layout of all the arrays is shown in Fig. 4.2.5. The help of Luca Parmesan in generating the SPAD cell layout has been greatly acknowledged during the realisation of the test chip. As can be seen in the general diagram of Fig. 4.2.2, the output of all the SPAD cells is connected to a series of multiplexers which, according to a list of configuration parameters selecting the size of the array and the RoI of the desired 4×4 array, connect the selected output cells onto a 16-wide buffer which is then connected either to the XOR-tree or the OR-tree. The array **D1**, as mentioned earlier in the section, has been designed in a different way: instead of favouring the compactness of the light sensitive part of the array, the speed of the interconnections has been prioritised. Therefore, instead of creating a 16×16 array surrounded by the circuitry, which would have been very challenging from the time-balanced point of view, the array is structured effectively as a 4×4 array of 4×4 arrays shown in Fig. 4.2.3 plus additional circuitry, i.e. enabling circuits and XOR cells (part of a dedicated tree only for the **D1** array). This is shown in Fig. 4.2.6.

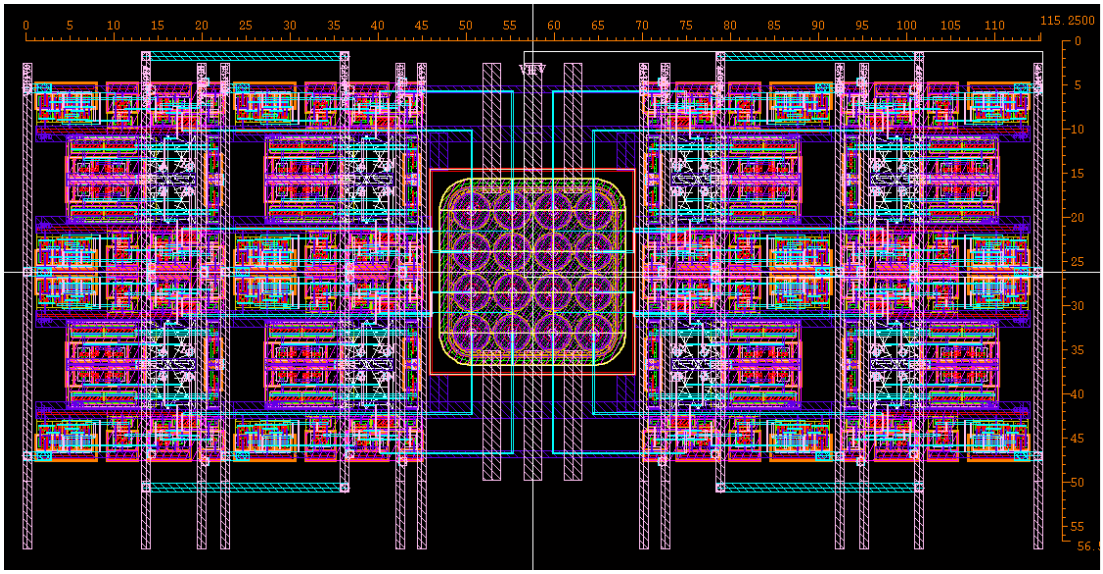


Figure 4.2.3: D2 Array layout - The circuitry for the 4×4 arrays has been placed outside the light sensitive area. The structure is common for all the arrays **D2-D5**.

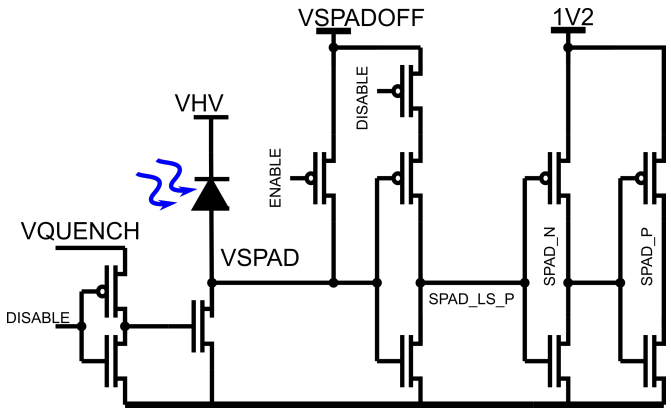


Figure 4.2.4: Configurable dSiPM cell - A blue sensitive SPAD cell includes a quenching resistor, enabling/disabling transistors and a buffer for level shifting to 1.2 V.

4.2.2 Digital N-to-1 Networks

This section describes two separate 16-to-1 networks that have been placed inside the chip for direct comparisons. Both networks, the OR-tree and the XOR-tree, have been designed with standard cells in a H-tree disposition to optimise the time balance among inputs and the layout for the H-tree has been flattened for compactness. The OR-tree in particular has been enhanced by a monostable pulse shortening circuit at its 16 inputs to enable pulse compression of the each SPAD output. The architecture of such cell is presented in both schematic and layout respectively in Fig. 4.2.7(a) and (b). Thanks to a multiplexer, the output bus of the SPAD arrays is connected to one of the two networks. In the case of the XOR, the input of in selected to be the output of the toggle circuit while in the case of the OR-tree, the direct SPAD output goes into the monostable cell before entering the tree. The output of both the networks goes into another multiplexer, controlled by the same bit choosing either of the two, allowing the connection of the desired output onto the on-chip ripple counter and onto the direct output pad for external analysis. The latter feature is essential for any time-domain or waveform analysis since in this test chip there is no TDC included.

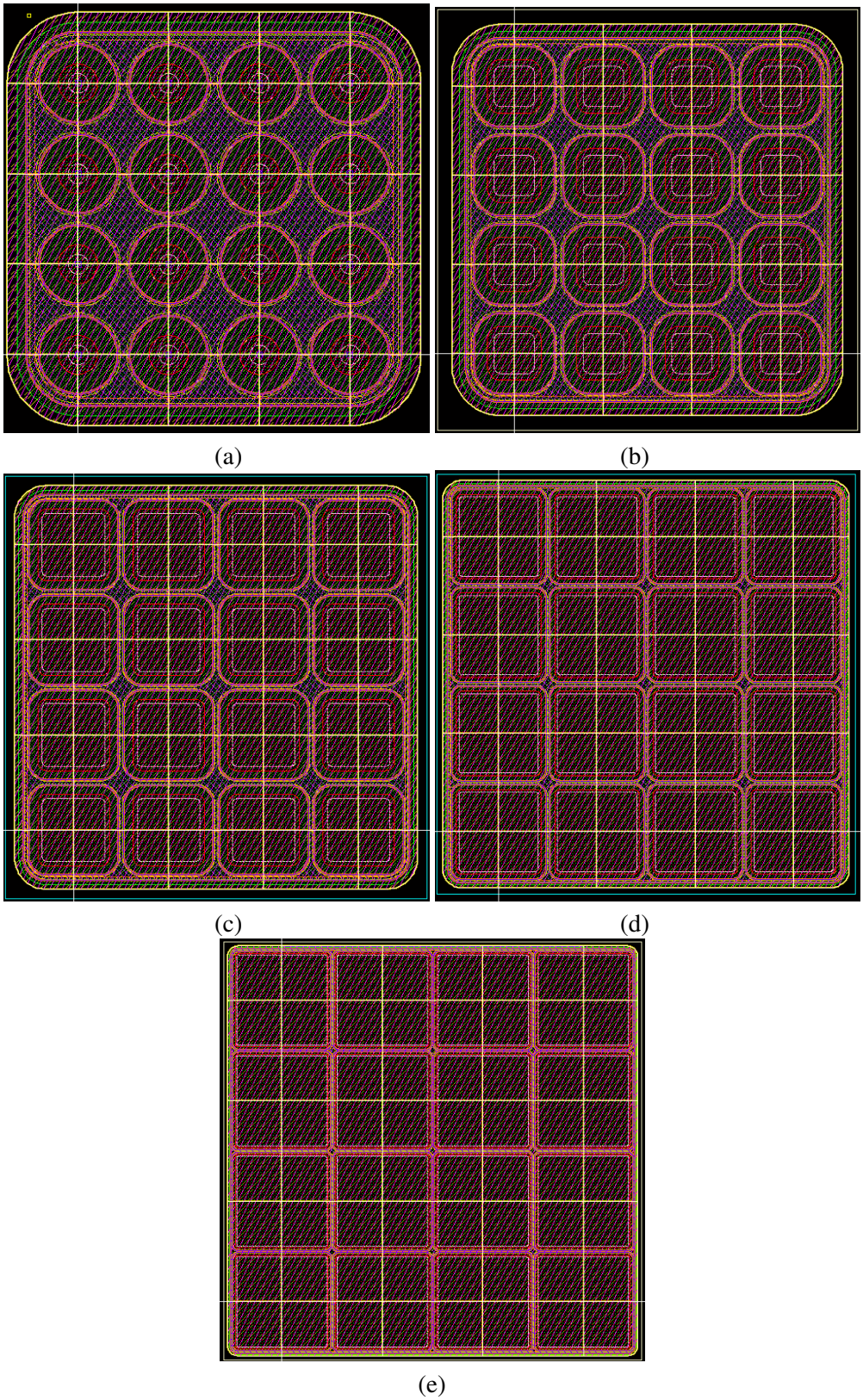


Figure 4.2.5: Configurable dSiPM 4×4 arrays - The layout of the arrays is shown for 7 µm (a),9 µm (b),13 µm (c),18 µm (d),34 µm (e)

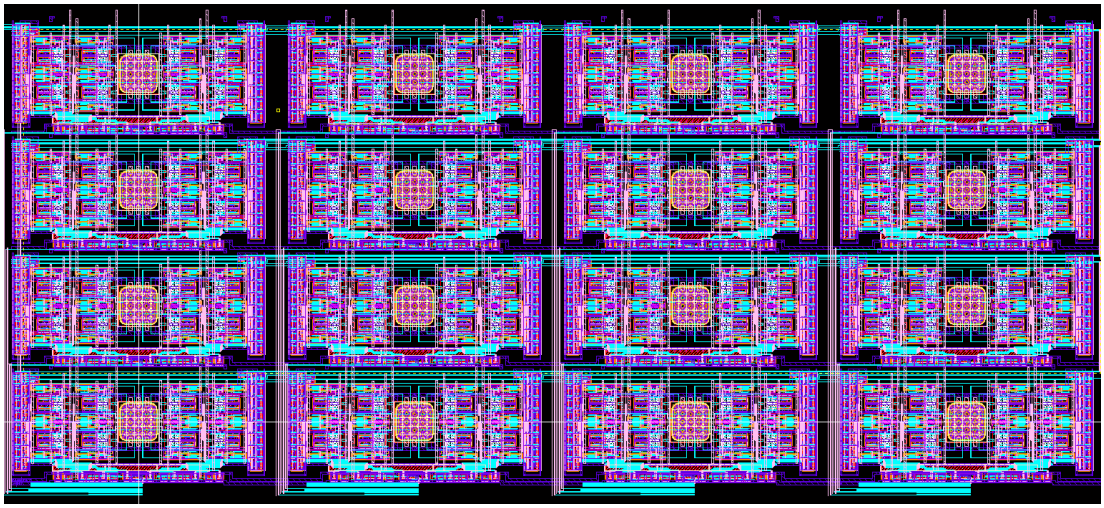


Figure 4.2.6: Dedicated 16×16 array in the dSiPM chip - A larger array is placed in the chip for the deep analysis of XOR-based dSiPMs. The structure priorities the time-balance among elements rather than the uniformity of the light sensitive areas.

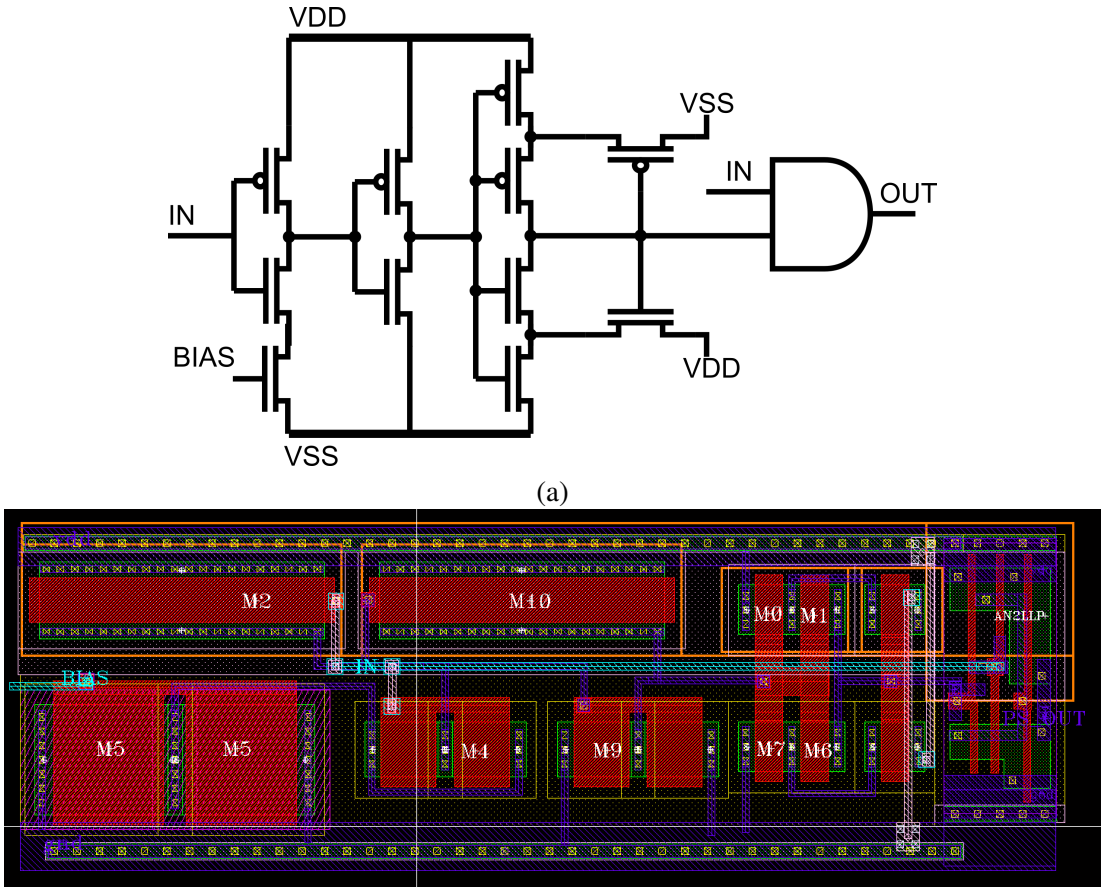


Figure 4.2.7: Monostable pulse shortening circuit - The external bias allows the width of the output pulse to be tuned for dedicated analysis on OR-based dSiPMs. Schematic (a) and layout (b) of the cell is presented.

4.2.3 On-chip Counters

On-chip counters have been included to optimise the record of detected events. The structure of the counters is a standard 16-bit ripple counter architecture. The input of the counters is the output of a multiplexer which selects which of the two networks has been enabled for the acquisition or, in alternative, it can also be connected to the dedicated output of the XOR-combined array **D1** or to the output of a specified SPAD cell (which is also connected to the dedicated otuput pad). The counters have a dedicated enable signal, which allows a fixed acquisition time to be set through an external square pulse with pulse width equal to the desired acquisition time, and also a reset signal to clear the status of the counters before each acquisition takes place. The 16 output bits however are not directly connected to 16 pads but are instead connected to a readout parallel to serial interface described in the next subsection.

The ability of the chip to record directly the number of detected events will be compared in experimental tests, see Chapter 5, to the alternative option of taking the array/network output off-chip and use, for example, FPGA counters.

4.2.4 Additional Features

To avoid a large number of required pads without compromising the features of the test chip, an input serial-to-parallel interface and an output parallel-to-serial interface have been included in the chip to access to the all the control signals needed to set the desired configuration of the dSiPM and to read the output of the counters.

The following signals are brought inside the chip through the input SPI:

- SPAD cell size (3 bits) - for pitch selection
- Toggle selection - for connection of the toggle or direct SPAD output onto the common output bus
- Counter input selection (2 bits) - for: XOR,OR, dedicated XOR or direct SPAD input
- XOR/OR selection - for the connection of the SPAD output bus onto the XOR/OR tree
- SPAD array RoI (16 bits) - for the selection of the region of interest of the selected 4×4 array

- SPAD output (4 bits) - for the selection of a single cell in the array to be connected to the dedicated output pad or counters

The data, clock and load signal for this interface are meant to be generated directly on the FPGA and connected to three pads. A software interface will be used to easily set and change each of the listed parameters.

The parallel-to-serial interface instead is designed to output in sequence each of the 16 bit of the counters. The clock and load signal are, once again, generated by the FPGA and are kept different from the input clock and load signal. To read the serial stream, the FPGA has been paired with a complementary serial-to-parallel interface so that the software interface can receive directly the 16 bits.

4.3 FPGA Firmware

The FPGA firmware used for these test chips consist of purely the generation of the configuration signals needed in the chip. These mainly include:

1. Input serial data - for input SPI, see previous sections for the complete list of parameters included in the SPI
2. Synchronous clock and load signal - for SPI function
3. Generation of clock reference – for the METDC this clock is used to lock the PLL and provide therefore a reference to the laser driver circuit
4. Generation of exposure/acquisition time - a square pulse must be sent to the chip to enable the counters/TDC for a fixed amount of time. This is set in terms of number of clock cycles of the reference FPGA clock
5. Synchronous clock and load signal - for output PSI function, in the case of the configurable dSiPM chip

The FPGA firmware is paired to a MATLAB software which, thanks to dedicated functions available from OpalKelly, allows the control of system parameters directly on software without the need to recompiling the entire FPGA image at every experiment.

4.4 Summary

The test chips here presented have been designed to investigate, in parallel to the modelling analysis, the limitations of dSiPM sensors and the impact of a selection of design choices, on the three main steps in the detection chain.

The first test chip, the METDC, has been design to analyse and prove the benefit of high conversion rates which enable the overcome of typical TCSPC bottleneck given by the loss of recorded events due to converter dead time. The TDC architecture allows parallel sampling and event detection as well as parallel readout with on-chip histogram formation. The architecture enabled multiple events to be recorded within the same laser cycle. The achieved 14 GSamples/s sampling rate of the METDC has been paired with a 32×32 SPAD array reaching 34% fill factor and using the newly proposed XOR combining network for single channel operation.

The second test chip has been designed to have a better and complete look at the detector side of the sensor. Different arrays available on the same chip enable the configuration of many SPAD cell parameters, i.e. dead time, area and number of cell. The dSiPM can be chosen to be OR-based and XOR-based for direct comparison. The impact of the single-channel bandwidth is represented in the test chip by the comparison between on-chip and off-chip counters. The test chip has been used to validate and reproduce the results of the modelling analysis in terms of detector throughput and timing performance.

Chapter 5 presents a series of tests that highlight the capability of the two test chips to provide different scenarios to analysed and characterised. These test chips have represented the perfect opportunity to confirm and explore the results obtain by the theoretical analysis presented in Chapter 3.

Chapter 5

Results and Modelling Comparison

The aim of this chapter is to present the experimental results on the analysis of the throughput of dSiPMs and to characterise the effect on the so-called pile-up distortions in the time domain. Each experiment has been run to confirm the theoretical results found in the Chapter 3. Moreover, a high throughput configuration is highlighted and proposed as possible solution to the pile-up minimisation problem. This configuration, consisting of a novel XOR-based dSiPM, is then paired to the high throughput Multiple-Events Time-to-Digital Converter (METDC) presented in Chapter 4. The combination of these two circuits allows the sensor to achieve an overall throughput in the order or 1 GSamples/s. A selection of the results presented in this chapter has been published in publications [3, 4].

Section 5.1 provides the information on the experimental set-ups that have been used for the realisation of each test. Any additional tool/material is specified in the specific section, where needed. In continuous light analysis, an LED is driven to control the incident light sensor on the dSiPM chip which has been configured accordingly to the goal of the experiment. The goal of such tests has been to analyse the output of the sensor, in terms of number of recorded events for different light levels and for a set of configurations of the dSiPM itself. For pulsed light analysis, a plused laser diode has been used to illuminate the sensor with coherent light to analyse its timing performance. In opposition to the experiments in continuous light, the recorded events are time-stamped and counted. This enables the analysis of the electrical signal that is associated with the optical stimulus of the sensor, i.e. the laser pulse. For all scenarios, the chips are mounted on a common custom PCB populated according to the I/O requirements of the chips. An FPGA is utilised for the configuration and the read out of the sensor.

The performed experiments are divided according to the utilised chip: dSiPM and METDC

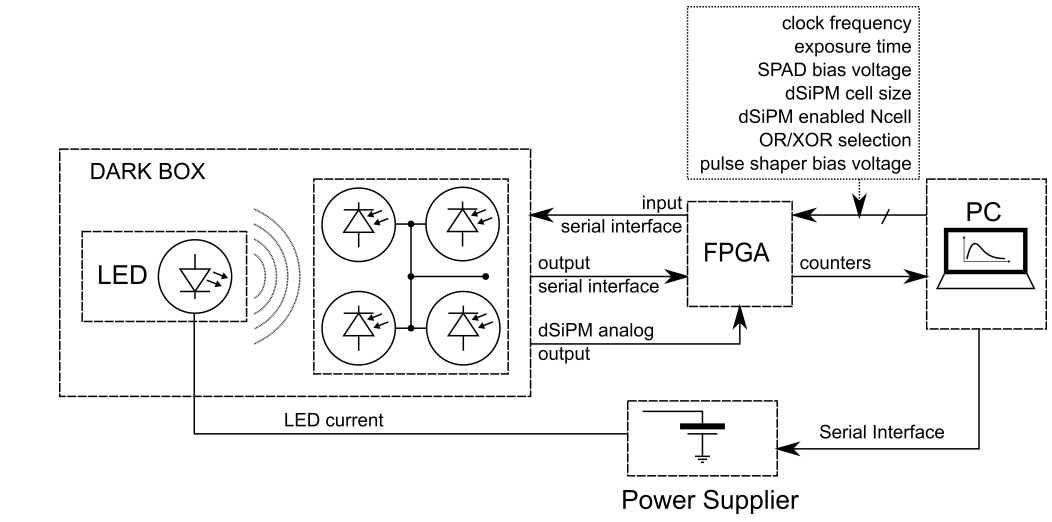
results are given respectively in Section 5.2 and Section 5.3. For each experiment, a flowchart of the corresponding algorithm is given throughout the chapter to help the reader understand and reproduced the reported results.

5.1 Experimental Setup

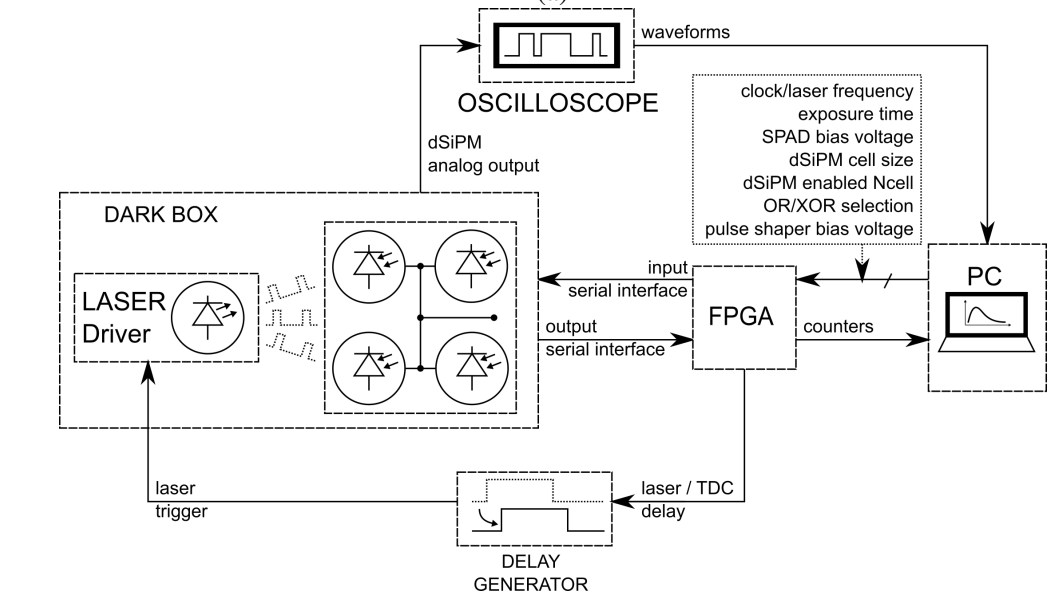
This section intends to describe in details the two experimental set-ups of the experiment run to characterise the performance of the two test chips. All the material utilised are listed and a general description of the algorithm used in the MATLAB software is provided in this section while the complete set of commands is to be found in Appendix B.

The test bench, pictured in Fig. 5.1.1, has been set up for the analysis of the response of dSiPM chips to continuous and pulsed light. The utilised materials are described together with reference to data sheets and details about the configuration such as bias voltages and geometrical set-ups.

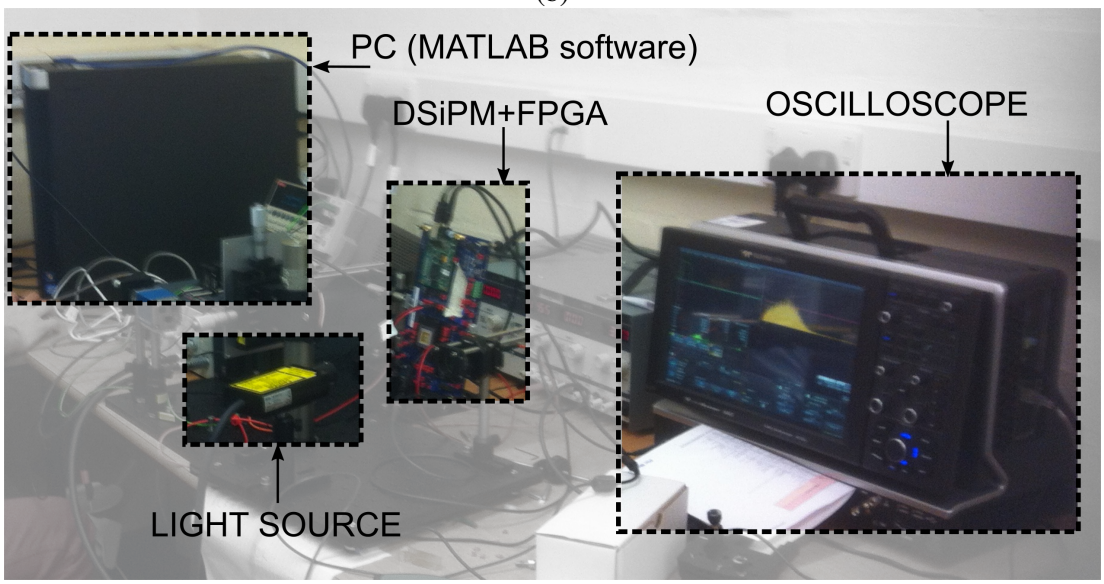
- **Voltage controlled LED** - Thorlab® M470L3 placed in close proximity ($\sim$ 1 cm) to the sensor in order to minimise diffusion of light. The dominant wavelength is chosen to be $\lambda_0 = 470$ nm to match the peak detection efficiency of the photodiodes. The voltage has been set through a digitally controlled power supply. The calibration of the emitted LED optical power versus the bias voltage has been done using a Thorlab power meter PM100D paired with a S130C photodiode from the same manufacturer.
- **Pulsed Laser** - A PicoQuant LHD-D-C-470 pulsed laser head has been paired to a PDL 800-D laser driver. The diode dominant wavelength has been chosen again to be $\lambda_0 =$ 470 nm to match the sensitivity of the dSiPM. The laser has been driven in a short pulse mode of operation with 63 ps FWHM. A square pulse mode of operation, with a width of 5 ns, has been used in the final experiment to analyse the distortion in high intensity mode.
- **Spartan-3 FPGA** - Mounted on Opal Kelly XEM3010, it is used to generate all the timing signals needed in the chip (serial-to-parallel interfaces clock, adjustable exposure time etc.) as well as driving DACs to set bias voltages and internal configuration



(a)



(b)



(c)

Figure 5.1.1: Light level analysis - The experimental setup consists of a light source, a voltage controlled LED for continuous light (a), or a laser diode for pulse light (b); a PCB to connect the dSiPM chips to an FPGA. All the experiment have been run using an external PC for configuration and acquisition of data. A photograph of the test bench (c) is also provided.

parameters as well as to read the output signals (counters and analogue/digital voltages). The interaction between the loaded firmware and the MATLAB® software has been implemented thanks to already-available Opal Kelly functions.

- **LeCroy WaveRunner - 640zi - 40GHz Oscilloscope** - The fast sampling oscilloscope has been used to acquire the analogue and digital waveforms from dedicated output of the chips. The fully digitised waveforms have been recorded and saved in digital files which have been subsequently processed on the PC by MATLAB software.

5.2 DSiPM Test Chip

This section presents the experiments dedicated to characterisation of the configurable dSiPM test chip. The results are presented to validate the analysis on the impact of each configurable parameter to the overall throughput of the sensor. Each key parameter has been swept throughout a possible range. The considered parameters are the following:

- Bias voltage for the quenching resistor: $V_{QUENCH} = 0.1\,\text{V}$ to $3.3\,\text{V}$
- Number of cells (SPAD unit): $N_{cells} = 1$ to 16
- Unit cell dimension: $pitch = 7 - 9 - 13 - 18.6 - 34.6$ µm
- Digital N-to-single channel network: OR - XOR tree
  - Bias voltage for monostable pulse-shortening circuit (OR-tree only):
    $V_{bias} = 0.1\,\text{V}$ to $3.3\,\text{V}$

The analysis on the throughput that is here validated is linked-up to the analysis of pile-up distortion in the time domain in Section 5.4.

The structure of the continuous light analysis experiments can be summarised in the following common steps:

1. Configuration of the dSiPM
   - SPAD bias voltage - V_{SPAD}
   - SPAD excess bias voltage - V_{EX}
   - SPAD quenching voltage, i.e. dead time setup - V_{QUENCH}
   - Unit cell dimension $pitch$

- Enabled cells N_{cells}
- Combining network
- Monostable pulseshaper bias voltage (for OR tree) V_{MS}
- Exposure time T_{exp}

2. Setting the LED voltage, i.e. the light level - V_{LED}
3. Record of acquired counts - M
4. Iteration for all light levels - N_{iter}

The circuit is powered up through the FPGA firmware. The analogue bias voltages, such as SPAD bias voltage, excess bias, quenching voltange etc., are set through DACs available on the PCB which are controlled by the FPGA via firmware code integrated in the software. The internal configuration is loaded up through the on-chip serial interface. The software allows each piece of data to be set individually. This includes the configuration of the dSiPM (unit cell dimension, enabled cells etc.) as well as the exposure time, i.e. the temporal window during which the counters are enabled to record counts. The firmware then generates the serial stream of data which is then fed directly into the chip.

The main loop of the experiments consists in the setting different light levels, i.e. the intensity of the LED which is controlled digitally though a power supply. Once the incident light is set, the counters are enabled for a fixed exposure time T_{exp}. This is controlled directly via software which converts the desired value in terms of FPGA clock cycles. Because of this, the exposure time can only be set as an integer multiple of the FPGA clock period. For such reason, the function also outputs the set exposure time for the user to evaluate the eventual difference the desired exposure and the set value.

Once the exposure is done, it is necessary to acquire the recorded counts M by reading the status of the counters. For what concerns the internal counters, to maintain a limited number of pads in the chip, the readout has been done in serial. The parallel outputs are converted to a serial stream. The conversion back to parallel is done in the FPGA which allows the software to acquire the digits. This significantly limits the performance of the dSiPM as a real-time counter since it introduces a significant readout time. For such reasons, the dSiPM has not been tested for direct communication experiments which require close-to-real-time readout.

Nevertheless, important conclusions on the dSiPM performance are obtained regardless of the specific implementation of the test chip.

The exposure and record of the counts are performed for a number of iterations N_{iter} set by software. In this way, a statistical relevant mean value and a standard deviation, respectively M and σ_M are calculated. From the known exposure time T_{exp}, the count rate m is calculated as:

$$m = \frac{M}{T_{exp}} \tag{5.2.1}$$

$$\sigma_m = \frac{\sigma_M}{T_{exp}}. \tag{5.2.2}$$

Due to the limited size of the on-chip counters, a long exposure time, in high light level conditions, might cause the saturation of the counters hence producing a wrong information on the count rate. On the other opposite, in low light level conditions (or dark), the exposure time needs to be long enough in order to efficiently estimate the count rate. For such reason, the exposure time has been configured considering the set light level to avoid such scenarios. As result of such necessary choice, the standard deviation on the count rate (shown as error bar in the reported graphs) might appear non uniform with the incident light level according to (5.2.2). Details on the chosen exposure time is given in a table for each experiment.

5.2.1 Number of Unit cells

In this section, the analysis is centred on the number of cells composing a single-channel dSiPM. The goal is to determine the optimal number of cells $N_{cells,best}$ to efficiently trade-off the detected photons and the dynamic range of the detector. To maximise the number of detected photons, the bandwidth of the combining network and the the common timing/counting circuit needs to be considered in order to avoid or minimise the detector channel pile-up, i.e. the loss of detected events due to limited channel bandwidth. The proposed model shows the following relation between the optimal number of cells and the single channel bandwidth:

$$N_{cells,best} = e \times \tau_d \times f_{BW} \tag{5.2.3}$$

To validate this result, the 16×16 XOR-combined SPADs were enabled in the dSiPM chip.

Parameter	Description	Value
V_{BIAS}	Common bias voltage of SPADs	14.4V
V_{QUENCH}	Voltage on quenching transistor	0.6 V
V_{EX}	Excess bias	1.2 V
pitch	Unit cell pitch	7 µm
N_{ITER}	Number of statistical iterations	500
$T_{exp,low}$	Exposure time for low incident light	0.6 ms
$T_{exp,high}$	Exposure time for high incident light	3 µs
DSiPM Type	Digital combination technique	XOR
N_{cells}	Number of enabled unit cells	1 to 256

Table 5.1: Number of unit cells analysis - List of parameter values used in the experiment.

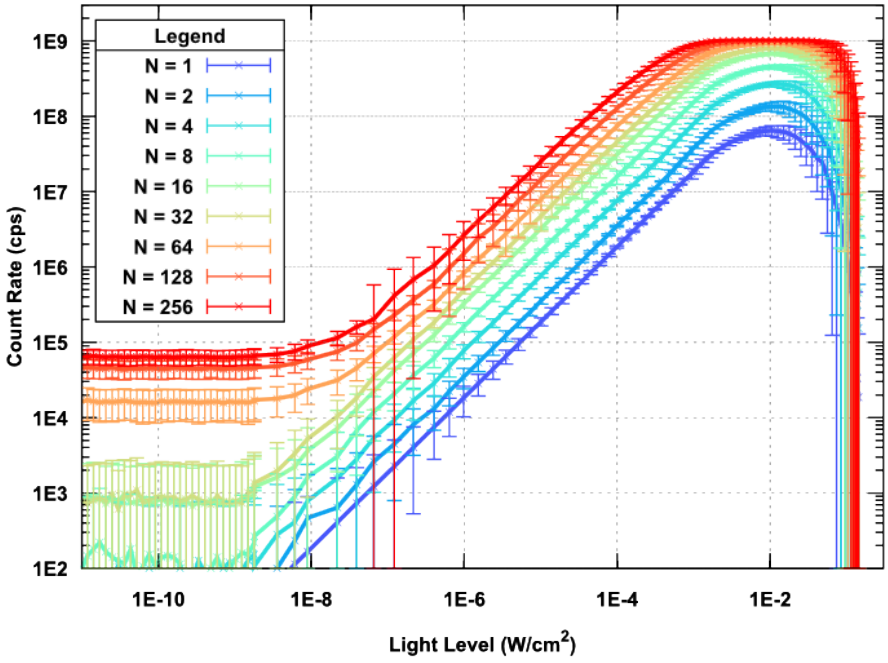


Figure 5.2.1: Throughput experiment for XOR-based dSiPMs - For a high number of activated cells, the registered curves show a limited throughput given by the channel bandwidth.

The general configuration of the dSiPM chip has been summarised in Table 5.1. Light level measurements have been performed increasing the number N_{cells} of activated SPAD cells. Fig. 5.2.1 presents the collected count rates. For $N_{cells} = 1$ the acquired curve coincides to the well-known response of a paralysable detector. Its maximum count rate is related to the dead time of the unit cell. At low light levels, the registered count rate is constant and only due to dark count rate, therefore independent on the light level. With increasing number of enabled cells, the dark count rate of the dSiPM increases as the sum of the single DCR values of the enabled cells (ideally linearly with N_{cells} in the assumption of identical cells). Before reaching the maximum count rate, the dSiPM exhibits a linear region of operation, i.e. the count rate is proportional to the input light level. The slope of the linear region can be seen to be independent on the number N_{cells} as it is connected to the *PDE* of the dSiPM. The maximum count rate is linear with the increasing N_{cells}. However, the bandwidth of the channel sets a limit to the maximum throughput of the dSiPM measured as ~ 1 Gevents/ s, see Fig. 5.2.2. In the graph, the measured maximum count rate for each dSiPM configuration has been compared to the proposed model:

$$m_{\text{dSiPM,max}}(N_{cells}) = f_{BW} \cdot \left[1 - \exp\left(-\frac{m_{\text{cell,max}}(N_{cells})}{f_{BW}}\right)\right] \tag{5.2.4}$$

$$m_{\text{cell,max}}(N_{cells}) = \frac{N_{cells}}{\exp(1) \cdot \tau_d} \tag{5.2.5}$$

Another interesting results that has been here verified is the saturation behaviour of XOR-based dSiPMs: the loss in counts is not abrupt when the number of cells is high (red line in Fig. 5.2.1): this is the proof that the saturation is due to the single channel which sets the limit on the recorded counts in a similar way of a non-paralysable detector, as modelled in Eq. 5.2.4. Each line of Fig. 5.2.1 has been fitted in order to empirically extract the parameters and perform a $\chi-$squared test. From this, the number of the most efficient number of activated cells is extracted and verified to be:

$$N_{cell,best} = e \times \tau_d \times f_{BW} = e \times 5.7 \text{ ns} \times 1.09 \text{ GHz} \simeq 16 \tag{5.2.6}$$

Although for a number of cells greater than $N_{cell,best}$ the maximum count rate increases towards its saturation limit, it is useful to underline the effect on the dynamic range of the dSiPM, i.e.

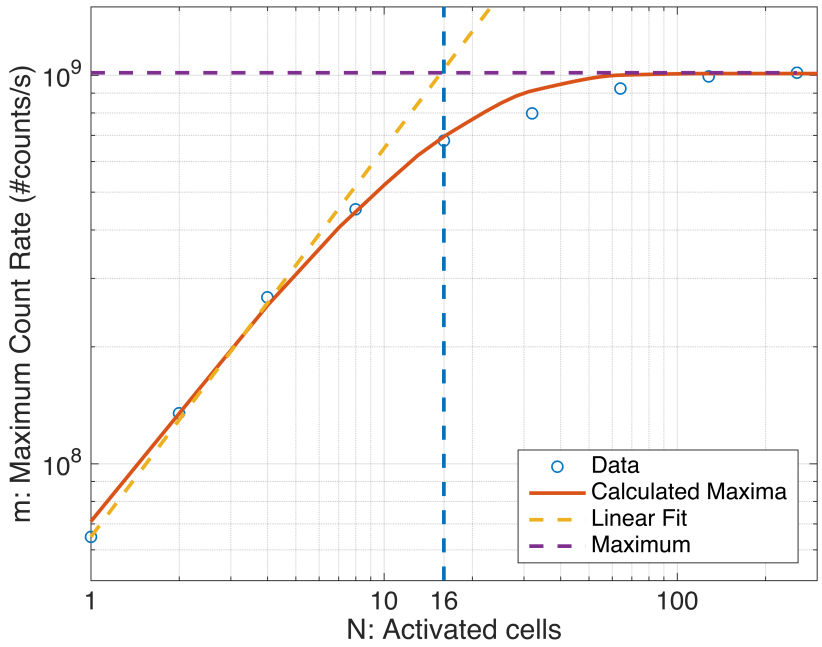


Figure 5.2.2: Efficient number of cells - The recorded maximum count rates for an XOR-based dSiPM is compared to the analytic model. The latter allows the optimal number of unit cells.

the ratio between the maximum count rate and the noise level. In fact, whilst the noise level, i.e. the DCR, increases with the number of cells as:

$$DCR(N_{cells}) = \sum_{i=1}^{N_{cells}} DCR_i \sim N_{cells} \cdot DCR_1 \tag{5.2.7}$$

the maximum count rate deviates from the ideal linear tend due to the saturation effect thus limiting the overall dynamic range. This gives a narrower linear region to the transfer functions shown in Fig. 5.2.1. The dynamic range (DR) calculated as:

$$DR(N_{cells}) = 20 \cdot \log_{10} \frac{m_{\mathrm{dSiPM,max}}(N_{cells})}{DCR(N_{cells})} \tag{5.2.8}$$

has been plotted in Fig. 5.2.3. The error bars presented in the plot have been calculated from the standard deviation of the registered count rates and of the DCR by propagating the error:

$$\sigma_{DR} = \sqrt{\left(\frac{\partial DR}{\partial m} \cdot \sigma_m\right)^2 + \left(\frac{\partial DR}{\partial DCR} \cdot \sigma_{DCR}\right)^2} \tag{5.2.9}$$

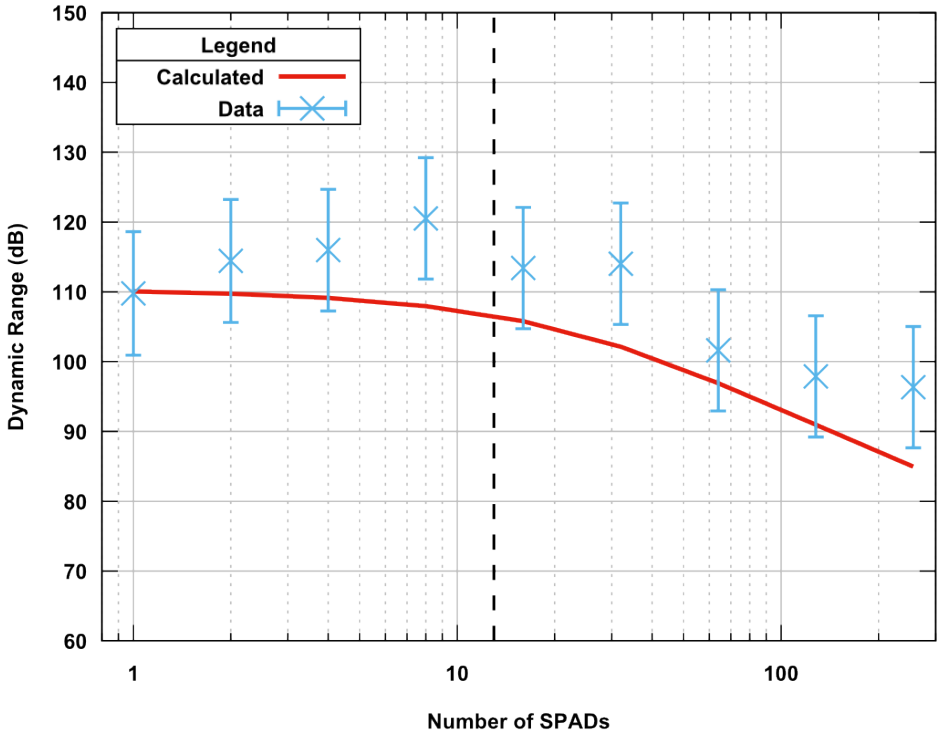


Figure 5.2.3: Dynamic range of a dSiPM with increasing number of enabled cells - The limitation on the channel bandwidth degrades the dynamic range.

This calculation, although being affected by a large standard deviation, provides a qualitative confirmation on the degradation of the dynamic range when too many cells are enabled in a single channel dSiPM.

This analysis confirms that a trade-off between overall throughput and dynamic range in a single-channel dSiPM is necessary since, while the latter increases asymptotically to a maximum given by the channel bandwidth, the latter is significantly limited due to the linear increase of the DCR.

5.2.2 Unit cell dimension

One of the main purposes of SiPMs is to realise large detectors for improving the collection of light, therefore an important choice of parameters is the size of the unit cell. Suppose the available area for an SiPM is A_{total}, then the smaller the unit, the more will fit into the given area. Two main effects fight against each other: the smaller the cell, the smaller the fill factor of the SiPM hence the lower the PDE; on the other side, more cells can be fitted in the area

Parameter	Description	Value
V_{BIAS}	Common bias voltage of SPADs	14.4V
V_{QUENCH}	Voltage on quenching transistor	0.6 V
V_{EX}	Excess bias	1.2 V
N_{ITER}	Number of statistical iterations	100
$T_{exp,low}$	Exposure time for low incident light	0.6 µs
$T_{exp,high}$	Exposure time for high incident light	3 µs
DSiPM Type	Digital combination technique	XOR

Table 5.2: Unit cell dimension - List of parameter values used in the experiment are here reported.

thus reducing the saturation of the sensor. However, it has been already discussed how single channel SiPMs have an intrinsic limitation on the optimal number of cells. In this section, both effects are analysed together.

As described in the Chapter 4, in the configurable dSiPM chip four pitch variants of a 4×4 SPAD array are available. For simplicity, let the total available area be equal to the area of the biggest available SPAD. The configurable chip can realise such SiPM in the following options (see Chapter 4 for array nomenclature):

a.) 1 cell from D5

b.) 3 cells from D4

c.) 7 cells from D3

d.) 14 cells from D2

Light sweep measurements have been performed per each configuration with the same procedure provided on Page 136. The value for each parameter is reported in Table 5.2. The recorded count rates for the XOR dSiPM have been plotted in Fig. 5.2.4. The option of a single large cell, yellow line, is not efficient: this confirms that SiPMs are a necessary solution for large area detectors. In fact, this configuration shows the smallest dynamic range and the narrowest linear region of operation due to the high DCR and the low maximum count rate due to the limited number of cells. However, this configuration shows the highest PDE due to the highest fill factor. For a very short range of incident light level (coinciding with the dSiPM linear region),

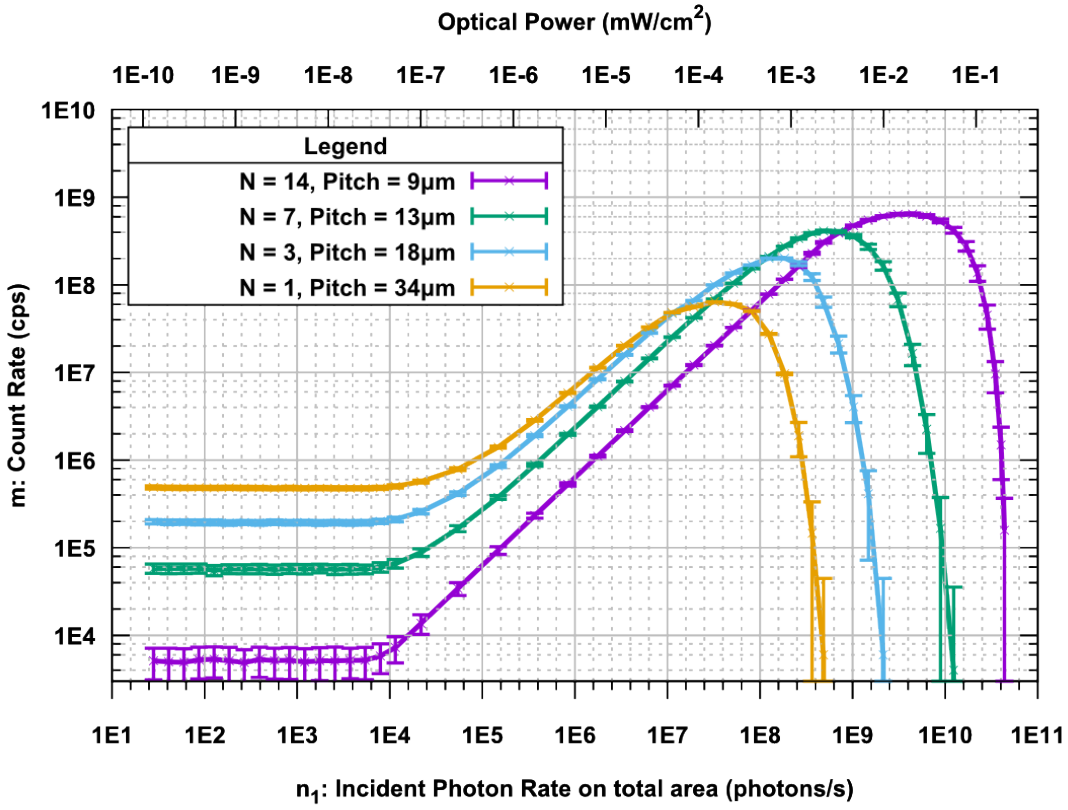


Figure 5.2.4: Impact of unit cell dimension in fixed dSiPM area - Photon transfer curves are presented for different configurations to be compared in terms of count rates. From a single big cell to 14 small cell dSiPM, the effect on throughput and dynamic range can be seen all together.

this configuration has the highest throughput. On the opposite, in the configuration d.), purple line in the plot, the dSiPM shows the highest dynamic range thanks to a lower DCR and a high maximum count rate guaranteed by the high number of cells. Both the DCR and the PDE of such dSiPM are limited by the reduction of the fill factor. The dynamic range and the PDE of all the configurations are plotted in Fig. 5.2.5.

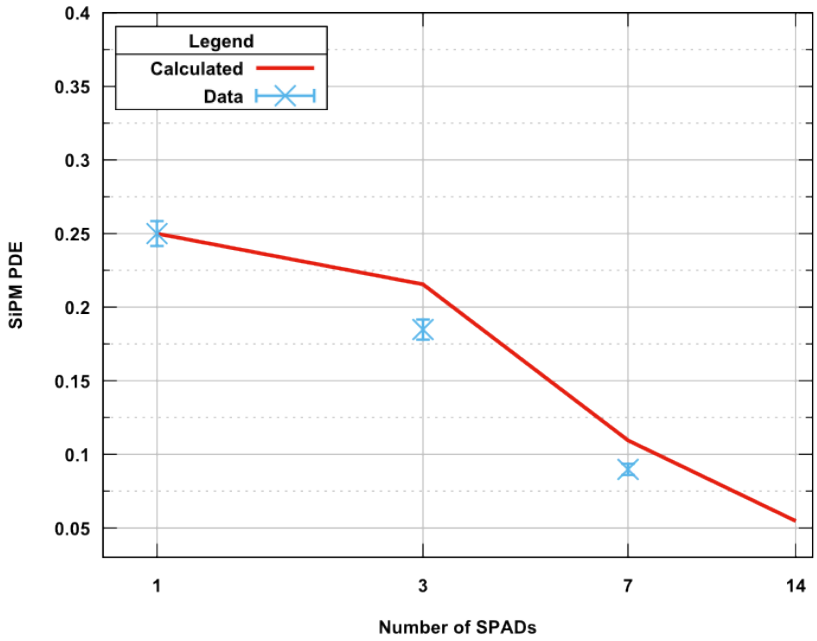
These results can be extended to analogue SiPMs which are intrinsically single channel SiPMs with no temporal compression technique (similar to single OR-tree dSiPM with no monostable circuits). The channel bandwidth in the analogue domain is set but the bandwidth of the readout circuit which typically consists of a transimpedance amplifier followed by a discriminator/comparator block which then is fed to TDC circuits. All the equations here verified are therefore valid for a large variety of SiPMs.

5.2.3 Digital N-to-single channel network

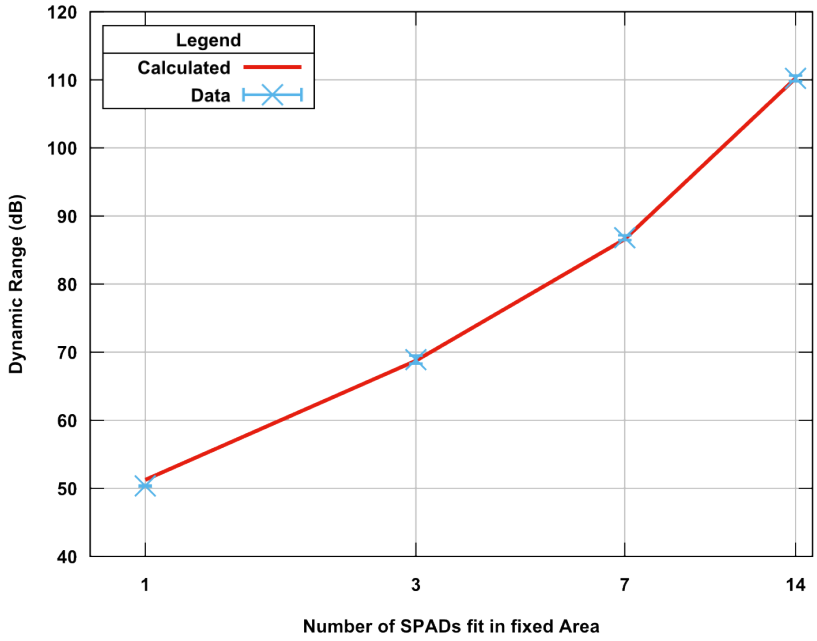
In the Chapter 3, two models were proposed to characterise two main N-to-single channel networks: OR-tree (with monostable pulse shapers) and XOR-tree. To validate such models, a chosen configuration of the SPAD array in the dSiPM chip has been combined onto single channel with the OR-tree as well as as the XOR-tree. The details on such configuration are summarised in Table 5.3. The results presented in Fig. 5.2.6, showing the registered count rates against the incident light level, refer to 16 enabled unit cells of 7 µm pitch (array D2). The throughput has been measured with on-chip counters as explained on Page 138. The red

Parameter	Description	Value
V_{BIAS}	Common bias voltage of SPADs	14.4V
V_{QUENCH}	Voltage on quenching transistor	0.6 V
V_{EX}	Excess bias	1.2 V
V_{MS}	Bias voltage of monostable circuit	$0.35 - 1.2$ V
pitch	SPAD cell pitch	7 µm
N_{ITER}	Number of statistical iterations	100
$T_{exp,low}$	Exposure time for low incident light	0.6 µs
$T_{exp,high}$	Exposure time for high incident light	3 µs
DSiPM Type	Digital combination technique	OR-XOR

Table 5.3: Digital N-to-1 network - List of parameter values used in the experiment.



(a)



(b)

Figure 5.2.5: PDE and Dynamic Range dependence on size of unit cell size - The change in the PDE (a) and dynamic range (b) is measured for different number of unit cells fitted in a total area.

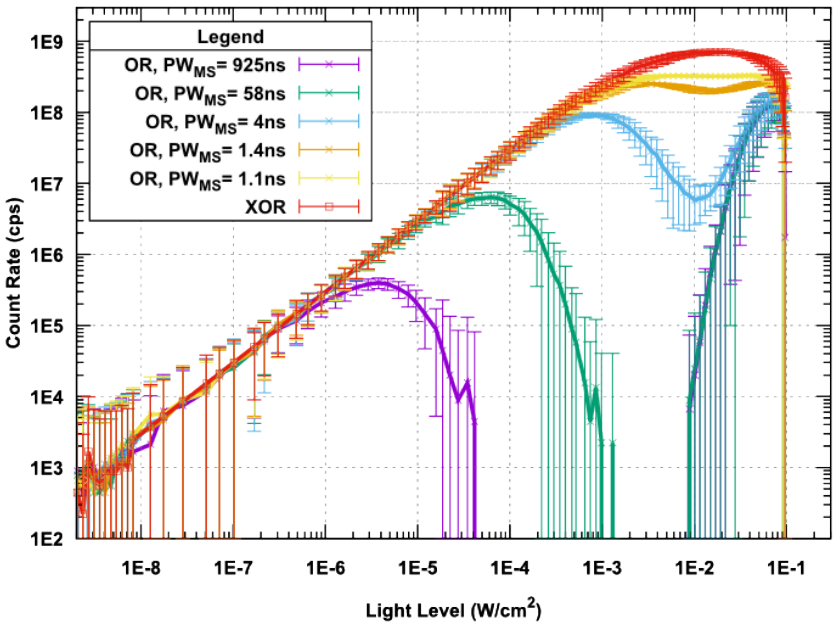


Figure 5.2.6: N-to-single channel network - The OR-tree is directly compared to the XOR-tree with on-chip counters for throughput analysis

line/error bars represents the count rates registered using the XOR-tree while the other coloured curves refer to the OR-tree with different monostable pule-width output. In the latter case, pulse widths larger than the actual cell dead time have been included for academic reasons only, i.e. to highlight the effect of such parameter. The results show highest throughput with XOR dSiPM: the overcome limitation of creating short pulses per each photon event typical of OR dSiPMs allows higher count rates to be registered. To numerically validate the modelled equations (3.4.4) and (3.4.6), a reduced chi-squared test has been run which gave a $\chi^2/\text{DoF} \simeq 1.49$ (worst case) as proof of a good match with the registered data. A further analysis has been run to understand if additional limitations may arise in general dSiPMs. The configurable dSiPM chip has the ability to connect the output of each network through a dedicated pad to off-chip (FPGA) counters. Since the output pad does not have the same high bandwidth of in-chip gates, the addition of such intermediate step in the flow of the signal causes a limitation in the overall dSiPM throughput. For such reason, the experiment described above has been repeated with the difference of using the off-chip counters. The results are presented in Fig. 5.2.7. The bandwidth of the pad and off-chip counters combination $f_{BW,pad}$ limits both dSiPMs in the same way, as

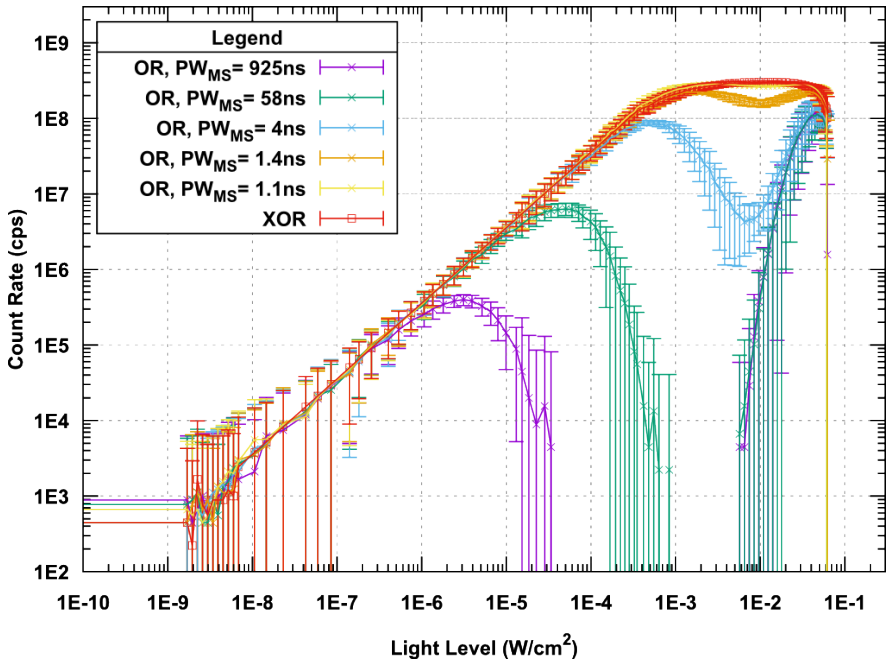


Figure 5.2.7: N-to-single channel network with off-chip counters - The throughput of both networks is limited by the bandwidth of the dSiPM pad.

described by:

$$m_{dSiPM} = \frac{f_{BW,pad}}{1 + m_{XOR/OR}/f_{BW,pad}} \tag{5.2.10}$$

By accurately characterising the bandwidth of channel, it is therefore here provided an efficient way to model the average count rate of a dSiPM. Or, in terms of design choices, the parameters of the dSiPM can be appropriately chosen in order to match the limitation of the available single channel bandwidth (i.e. to match the property of the counting or timing systems as well as technology limitations such as the bandwidth of the electrical path of the signal). A general summary of the performance of the analysed dSiPMs is presented in Table 5.4. It is here reported an optimal configuration with 16 enabled unit cells of 7 µm giving a throughput of $>$ 600 Mcps and a dynamic range $>$ 110 dB.

dSiPM Type	N_{cells}	PW_{MS}	Dynamic Range (dB)	Linearity (R^2)	Throughput
OR	**16**	**1ns**	**107**	**0.991**	**327 Mcps**
		1.4ns	107	0.997	252 Mcps
		4ns	96	0.997	91.7 Mcps
		58ns	74	0.989	6.37 Mcps
		925ns	51	0.979	398 kcps
XOR	1	-	140	1	64.9 Mcps
	2		130	1	136 Mcps
	16		**117**	**0.994**	**683 Mcps**
	128		87	0.987	992 Mcps
	256		83	0.987	998 Mcps

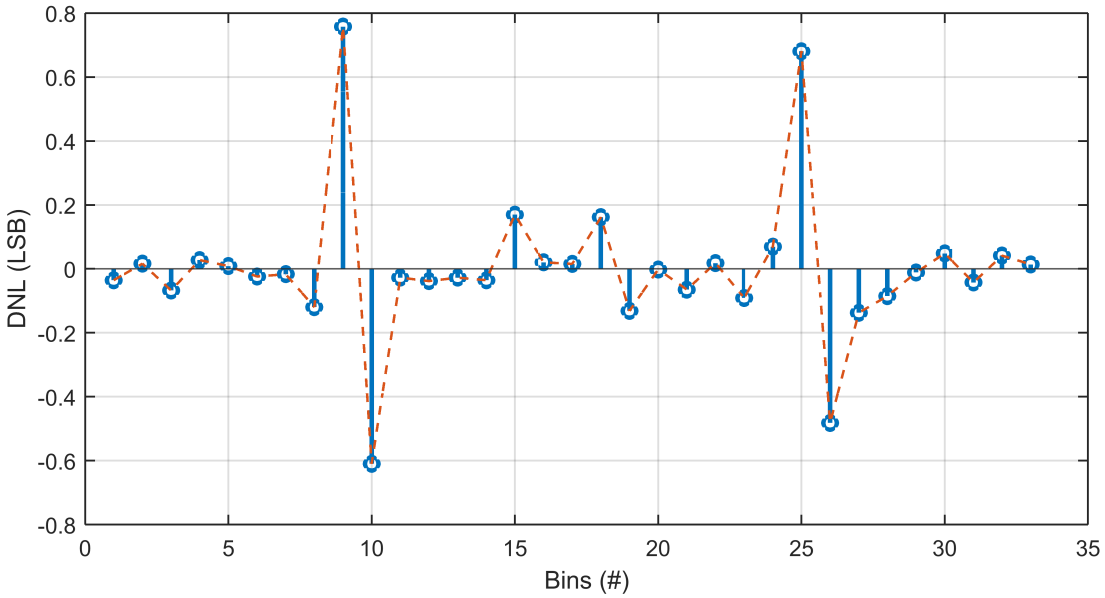
Table 5.4: dSiPM Performance Summary - Unit cell pitch = 7 µm

5.3 Multi Event Time-to-Digital Converter

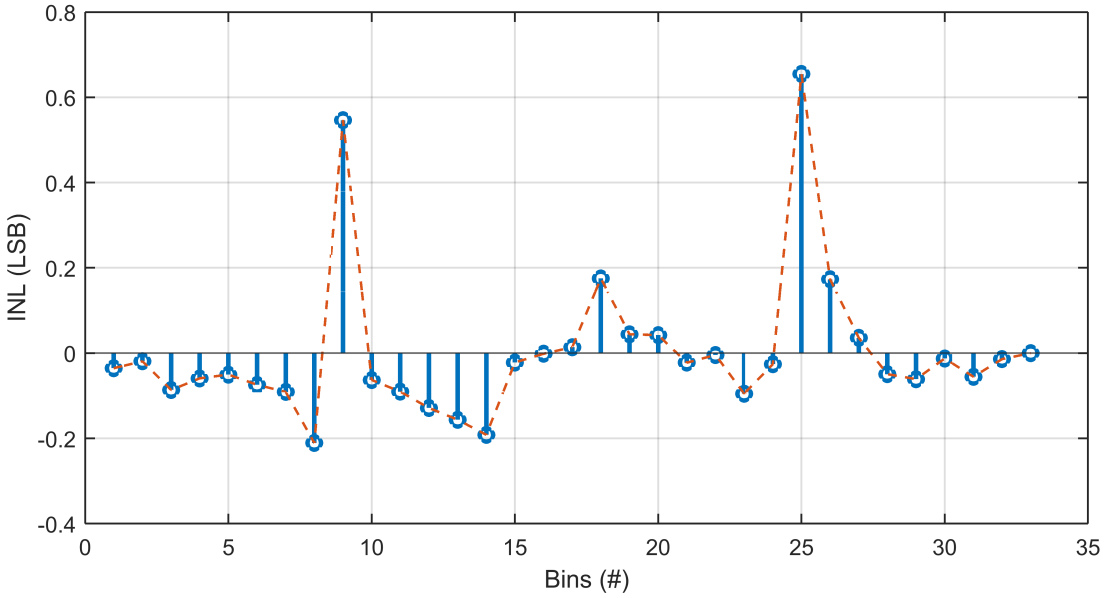
In the previous section the XOR dSiPM has been proven to potentially increase the throughput of single-channel SiPMs. Moreover, it has been highlighted how such improvement needs to be matched by a high bandwidth counting or timing circuit to maintain a high overall sensor throughput. In case of dSiPM employed as timing systems, the presented analysis implies that a high bandwidth - high sampling rate TDC is necessary. The benefits of such systems in advanced TCSPC have been already described in previous chapters while this section focuses to present the experimental results from the characterisation of the Multi Event TDC (METDC) described in section Chapter 4. The following subsections will describe the linearity, the sampling rate and throughput of the TDC as well as the temporal precision, resolution and accuracy.

5.3.1 Code Density Test and Linearity

One of the most relevant proprieties of a TDC is its linearity, i.e. the uniformity of the time bins which ensure a linear relation between the input timing information and the output digital code. This has been tested through a code density test (CDS). The METDC chip has been illuminated with a constant uncorrelated light source (standard office light) and the output histogram of the TDC has been registered. The intensity of the light and the exposure time are not relevant and



(a)



(b)

Figure 5.3.1: METDC Linearity - A Code Density Test has been performance to calculate (a) the DNL and (b) the INL of the TDC.

have not been measured. The fundamental assumptions for this test are the zero time-correlation of the input light and non-saturation of the counters. The latter can be achieved with relatively short exposures or by enabling a limited number of unit cells within the array inside the chip. In such conditions, every bin of the TDC, hence of the output histogram, has the same probability of being filled over time, i.e. the output counts per each been is ideally the same. Due to the TDC non-linearity, in other worse the non-uniform size of the temporal bin, the population of the histogram deviates from the ideal constant shape. Calling μ_{bin} the average registered count of the ideal bin, the differential non-linearity (DNL) is described by the deviation of the average count of the bins $\{p_i\}_{i=1}^{N_{bins}}$ from its ideal value:

$$DNL_i = \frac{p_i - \mu_{bin}}{\mu_{bin}} \tag{5.3.1}$$

The internal PLL of the TDC has been set to lock at a frequency of ~ 56 MHz which is scaled down by an internal $\times 8$ divider so that the 33-stage VCO sets the TDC bin size to a nominal value $\mu_{bin} \simeq 71.4$ ps. In Fig. 5.3.1(a) the *DNL* has been reported as calculated with (5.3.1) against the bin number. The maximum/minimum *DNL* is registered as $DNL = (+0.75/-0.61)$LSB. From the plot, two main peaks can be seen around bins 9-10 and bin 25-26. These are due to the layout of the VCO, see Chapter 4, and correspond to the longest path of the clock distribution which therefore causes such non-uniformity. Overall the TDC shows a strong linearity.

Another typical way to describe this behaviour is the integral non-linearity (INL) which is defined as:

$$INL_i = \sum_{k=1}^{i} DNL_k \tag{5.3.2}$$

This is plotted in Fig. 5.3.1(b) showing a measured maximum/minimum $INL = (+0.65/-0.2)$LSB. These figures of merit can be utilised for post processing analysis to correct the non-linearities distortion of the recorded histograms. In this work, a *DNL* correction has been been applied to the reported histograms. This is done by multiplying the registered counts of each $i-th$ bin with a number equal to $1 + DNL_i$. Where necessary, both uncorrected and corrected histograms are reported in plot legends.

5.3.2 Throughput Test

The METDC has been proposed for its high throughput performance. For such reason, experiments were set up in order to register this figure of merit. The aim is to demonstrate a maximum number of photon events successfully processed by the TDC that is greater than the typical restriction of classic TCSPC system on the number of laser photons per pulse $N_{photons-pulse} << 1$. Moreover, in the next section, the benefits of such sampling rate and throughput are going to be verified in the time domain.

The experiment has been set up as follows: a pulsed laser, synchronised to the chip via an external laser driver controlled directly from the FPGA, illuminates the sensor which is configured with an increasing number N_{cells} of activated cells. The laser pulse repetition rate has been set to $PRR = 56$ MHz and the TDC bin size is set to 71.4 ps. The intensity of the laser power has not been measured as not relevant to the purpose of this experiment. However, sweeping the number of activated cells of the dSiPM allows the input rate of events of the TDC to be increased. For each value of N_{cells}, the histogram is obtained and on software all the bins are summed together in order to calculate the total number of converted events, which is here indicated as $M(N_{cells})$. From these data, the registered throughput TP is calculated as:

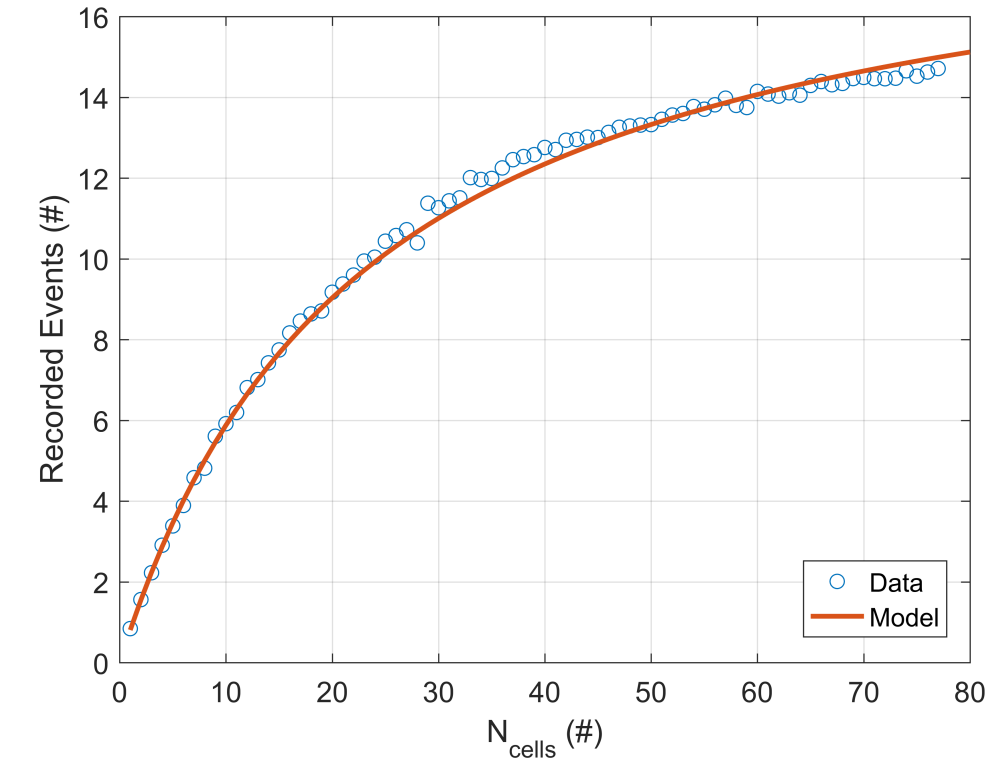
$$TP(N_{cells}) = M(N_{cells}) \times PRR \tag{5.3.3}$$

The collected data are presented in Fig. 5.3.2. The modelled red line is the plot of the following equation:

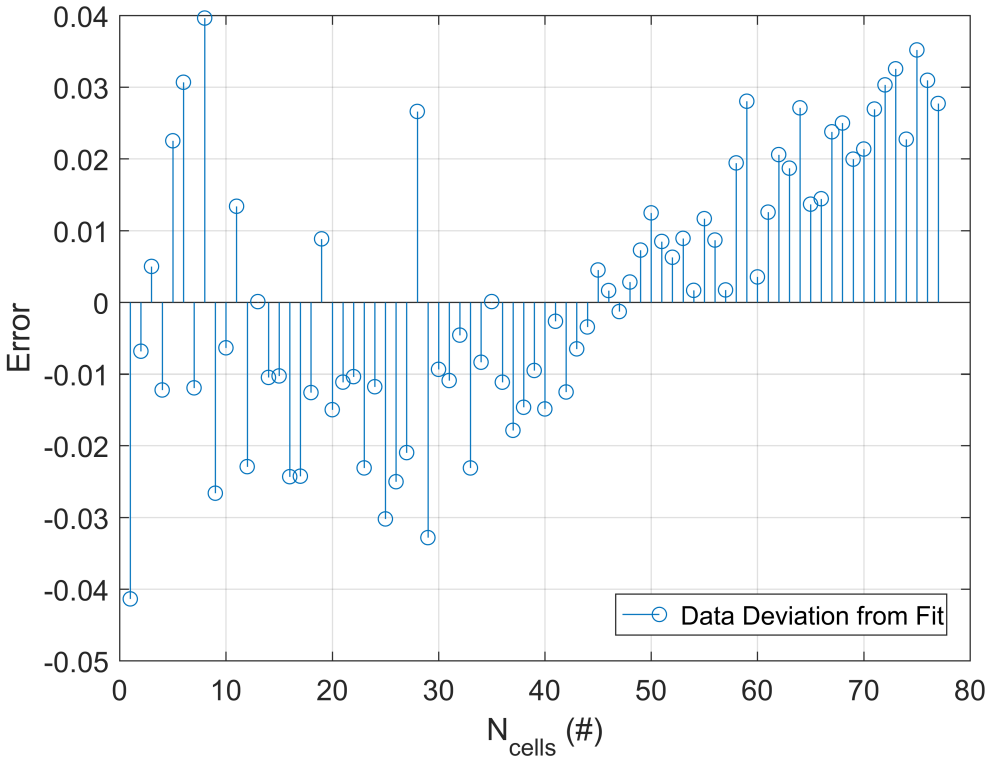
$$M(N_{cells}) = M_{max} \cdot \left[1 - \exp\left(-\frac{N_{cells} \times m_1}{M_{max}}\right)\right] \tag{5.3.4}$$

where m_1 is the detected photons per unit cell, while the maximum recorded numbers of converted event $M_{\max} \sim 16$ has been extracted by a non-linear fit. Moreover, this experimental value matches the relation between the maximum bandwidth of the single channel f_{BW} and the laser pulse repetition rate:

$$M_{\max} = \frac{f_{\mathrm{BW}}}{PRR} = \frac{0.9\ \mathrm{GHz}}{56\ \mathrm{MHz}} \sim 16 \tag{5.3.5}$$



(a)



(b)

Figure 5.3.2: Throughput experiment for the METDC - The maximum number of converted events has been measured with increasing input signal rate of the TDC. The mean values are plotted (a) together with the deviation from the modelled value (b).

From Eq. (5.3.3) and $M_{\max}$, the measured maximum throughput is therefore $TP_{\max} \sim 900$ Mevents/ s for $N_{cells} \sim 80$.

The results from this experiment confirms that the proposed TDC design is able to successfully convert a number of photons per laser pulse of $16 \gg 1$ in opposition to classic TCSPC requirements. This enables high incident photon rates to be correctly digitised by single-photon counting and therefore the pile-up due to conversion dead time is here minimised. The limitations for this prototype dSiPM consist in: the sampling rate of the TDC, which has no additional the dead time, and the bandwidth of the input channel. For this architecture, which is based on a flash topology, the sampling rate F_{TDC} can be written as:

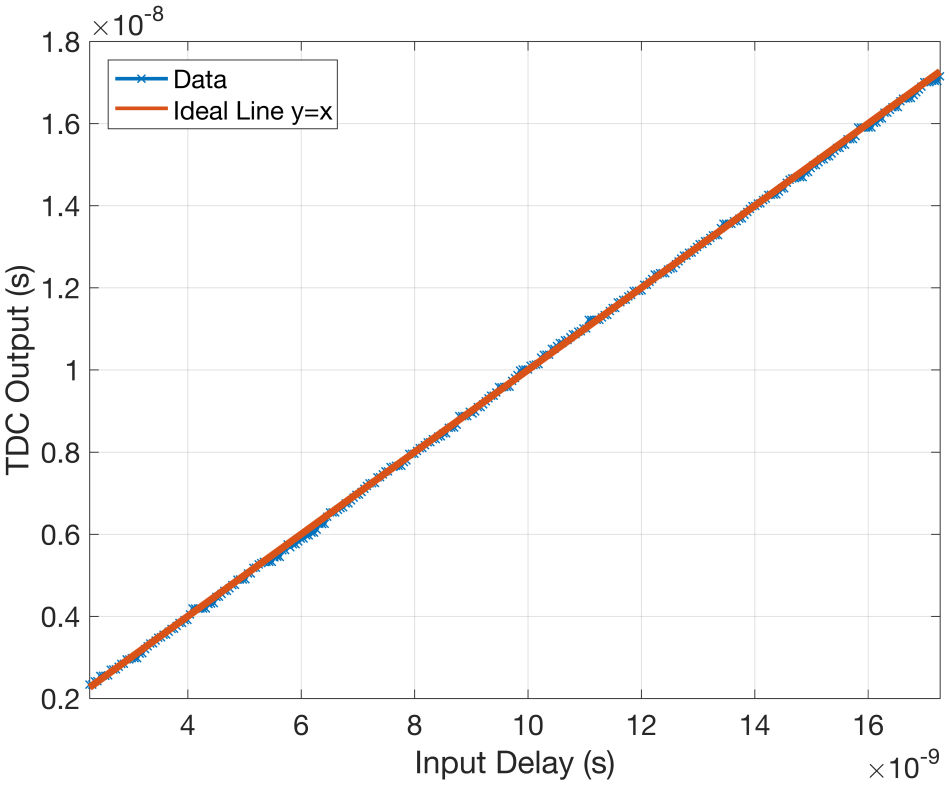
$$F_{TDC} = \frac{1}{T_{bin}} = \frac{1}{71.4 \text{ ps}} = 14 \text{ GSamples/s} \tag{5.3.6}$$

The single channel bandwidth instead is related to the load of the signal itself and therefore has to be evaluated during the design process. Here this has been characterised by the parameter f_{BW} and estimated as ~ 0.9 GHz.

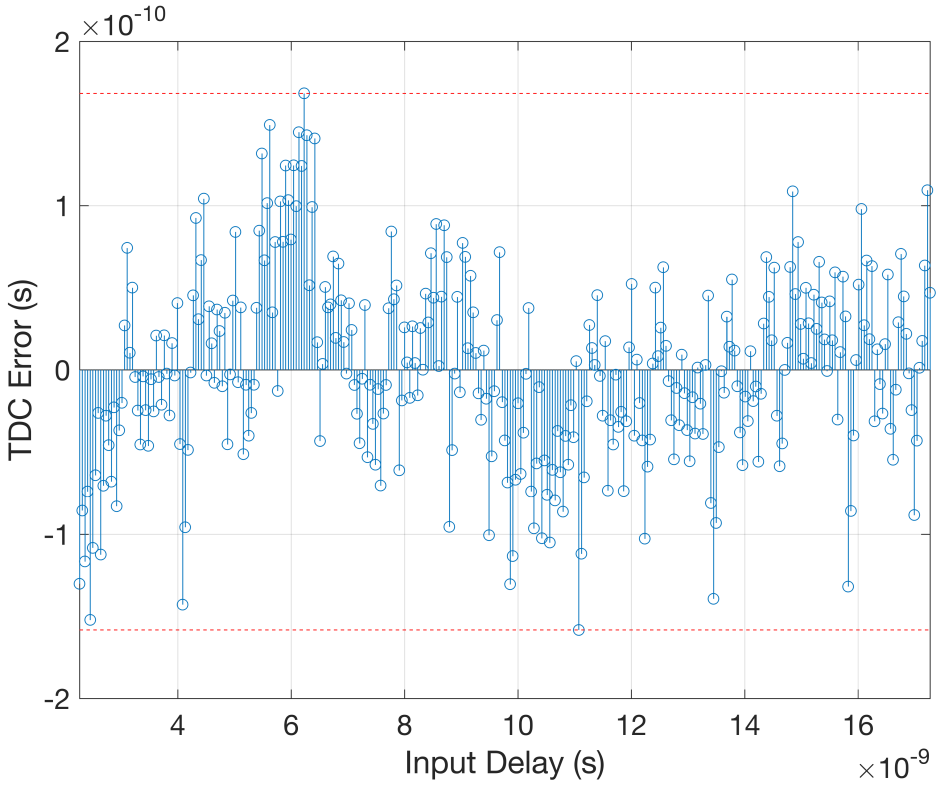
5.3.3 Time Resolution

Another figure of merit of METDC sensor is the time resolution. Typically, the resolution of a system is defined as the minimum change in the input that produces a detectable change in the output. This definition is sufficient for "one-event" TDCs. However, since the proposed architecture is capable of producing multiple time stamps within a laser cycle, the concept of resolution is extended. Just as in optics, especially in displays, the resolution is also defined as the minimum closeness in space (or in ToF, in this case) of *two* separate objects that produces two discernible outputs (time stamps). These two concepts in the case of METDC translate respectively into the minimum change in distance (depth) of a single object in the field of view that is correctly detected, i.e. that produces a relevant variation in the output histogram resulting in a different time stamp; and in the minimum depth difference between *two* objects in the field of view of the sensor that produces *two distinguishable peaks within the same histogram*.

To verify these two aspects of the time resolution, two separate experiments have been run.



(a)



(b)

Figure 5.3.3: One-event resolution analysis - Trough a digital delay generator, the minimum variation of the input of the TDC producing a detectable change has been measured. The obtained time-stamps are plotted against the input delays (a) together with the deviation from the ideal line (b)

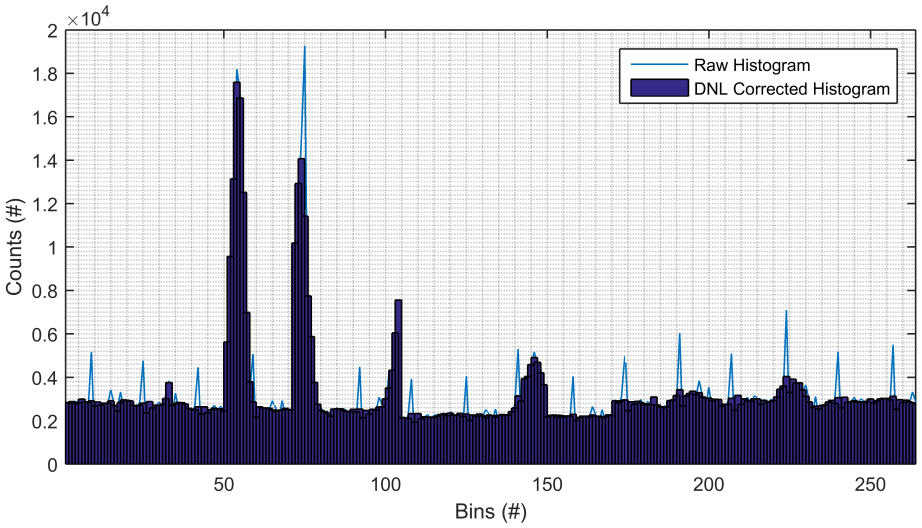


Figure 5.3.4: Depth resolution analysis - Two events spaced by 1.4 ns are successfully discriminated by the METDC.

With the use of a digital delay generator, a variable delay between the laser pulse and the TDC clock has been added to the ToF measurements. This allows the one-event time resolution to be estimated without the need of physically change the distance between the sensor and the target. The sweep has been run by software and has been synchronised to the TDC clock. Per each time delay, the histogram has been acquired from the TDC and a time-stamp has been calculated by a weighted average algorithm after *DNL* corrections on the histogram. Fig. 5.3.3(a) shows a plot of the obtained time-stamps against the input delay set digitally by the software together with the ideal response $y = x$. In the graph below, Fig. 5.3.3(b), the deviation from the ideal line is reported per each bin. The maximum value can be read as the overall resolution of the sensor. This is reported as $\sim$ 168 ps and it must be underlined that this is the resolution of the entire sensor, i.e. includes the timing response of the detector together with the TDC.

The continuous operation and the high sampling frequency of the converter allow close-in-time events to be recorded and discriminated. To demonstrate this statement, two laser pulses were sent to the sensor. This has been done by using two laser heads driven at the same frequency but with a programmable digital delay between the two pulses. This has been done thanks to a delay generator. The two pulses have been put as close in time as 1.4 ns. The collected histogram can be seen in Fig. 5.3.4: the two pulses are distinguishable and the peaks are separated by $\sim$ 10

bins which therefore correspond to 1.4 ns resolved time. The smaller peaks present in the histogram represent optical reflections that have not been filtered. The baseline of the recorded counts represent the uncorrelated ambient light. The presented histogram has been corrected with the DNL results.

This section has presented the timing performance of the METDC paired with the XOR-based SPAD array design. Typical one-event time resolution has been measured as $\sim$ 168 ps with the additional feature of the TDC to be able to digitise multiple laser pulses as close as 1.4 ns thanks to continuous sampling and high throughput. This capability is further exploited in the demonstration of the minimisation of the pulse distortion effect presented in the next section.

5.4 Pulse Distortion Analysis

It is in the interest of this section to demonstrate the effect of pile-up distortions and the minimisation of such phenomenon in optimised configurations for an input optical square pulse. This section determines and analyses the causes of pile-up within the detection and the processing of the photons within the sensor. In two experiments, the XOR-based dSiPM with the METDC sensor has been illuminated with a square laser pulse. The histograms read from the chip are effectively digitising the shape of the detected laser pulse. The sensor has been configured with an increasing number of activated SPAD cells in order to acquire a histogram per each number or activated cells and demonstrate the effect of such parameter on the distortion of a known shape of the input signal. Note that no focusing lenses have been used, therefore, the amount of light incident on each cells is constant and supposed to be equally distributed in space (i.e. the intensity of the laser pulse is assumed to be constant throughout all the dSiPM).

In a first experiment, the intensity of the incident laser has been kept low enough to prevent saturation of the single dSiPM cell. This has been achieved by the usage of diffusers to attenuate the intensity and also by pointing the beam to the dSiPM with a convenient angle to ensure a further attenuated intensity (due to the cosine law of diffusion) suitable for the hypothesis of the experiment, see diagram in Fig. 5.4.1.

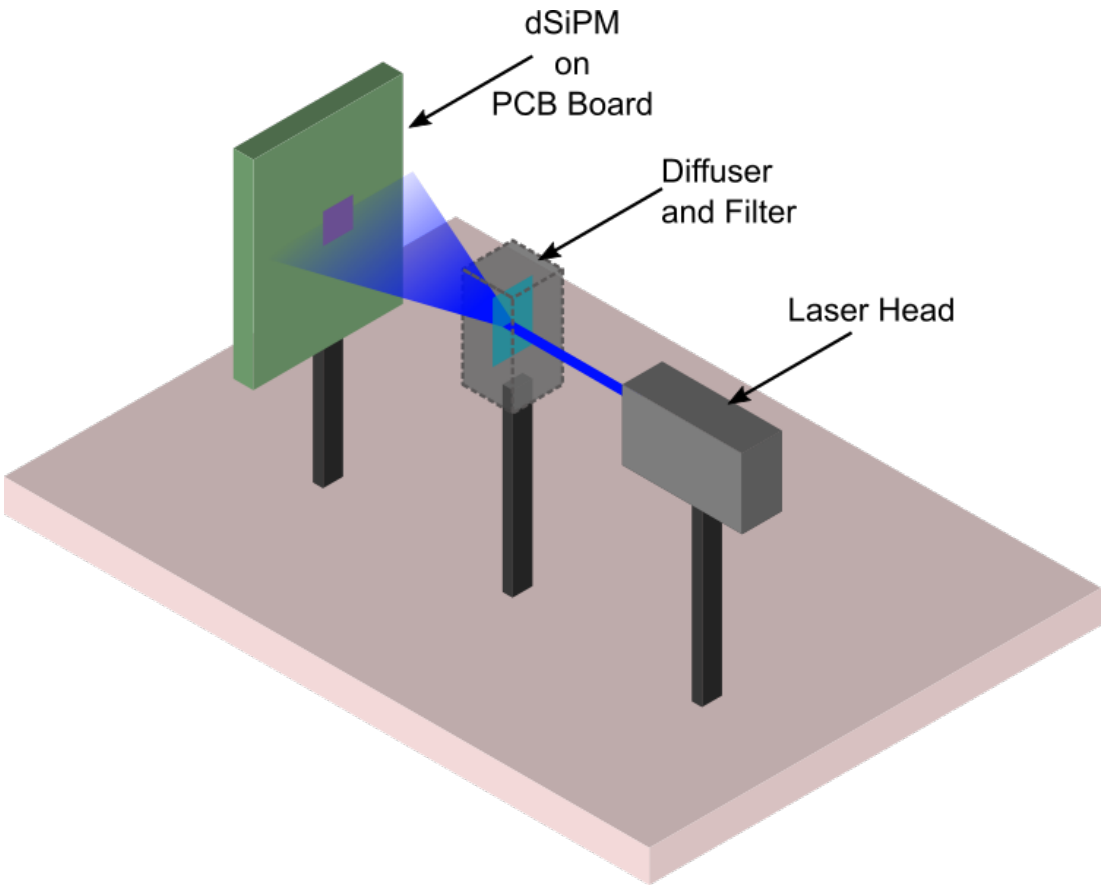
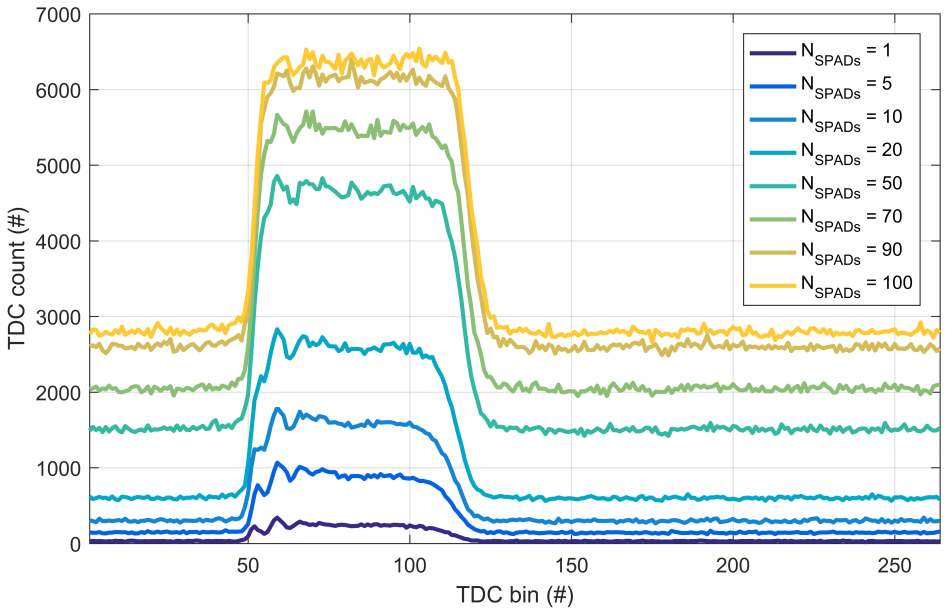


Figure 5.4.1: Pile-up experimental setup - The METDC sensor is mounted on a PCB and illuminated with a square laser pulse. To adjust the incident optical power on the detector, filters and diffusers have been used.

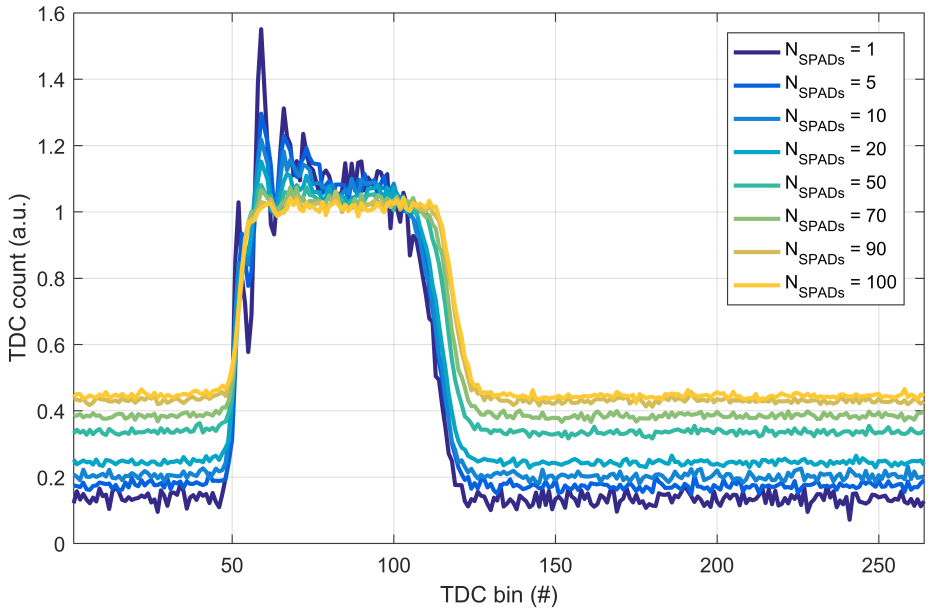
Parameter	Description	Value
V_{BIAS}	Common bias voltage of SPADs	14.4V
V_{QUENCH}	Voltage on quenching transistor	0.6 V
V_{EX}	Excess bias	1.2 V
PRR	Laser pulse repetition rate	53 MHz
T_{PW}	Laser pulse width	5 ns
T_{exp}	Exposure time for low incident light	10 ms
T_{bin}	TDC bin size	72 ps
N_{pulses}	Number of laser cycles per exposure	$PRR \times T_{exp} \simeq 95000$

Table 5.5: Experimental setup for pulse distortion analysis - List of parameter values used in the experiment.

Fig. 5.4.2 reports the recorded digitised waveforms overlapped for a selected number of activated unit cells in the SPAD array. For low number of cells, the dSiPM maintains the shape



(a)



(b)

Figure 5.4.2: Square pulse pile-up analysis - A low intensity square pulse is distorted by a large single-channel dSiPM with the effect of a reduction of the pulse hight due to single-channel saturation. Waveforms are plotted in absolute values (a) and normalised to the peak value (b) to better appreciate the reduction of pulse height.

of the original pulse: the low intensity of the incident light together with the low number of activated cells avoid the pile-up conditions of both the detector itself and of the single channel. Increasing the number of cells allows more events to be digitised exploiting the high sampling frequency of the METDC. The noise floor, set by the combination of DCR of the cells and ambient light, linearly increases with the number of cells. An interesting result consists of height/peak of the pulse. When more cells are activated onto the single channel, at a certain level, the registered counts reach a saturation point beyond which no further elements are recorded resulting in a distorted square pulse having a shorter height, see 5.4.2(b), demonstrating the reduction of dynamic range of the dSiPM. This is a proof of pile-up due to pure single-channel, which was the object of this analysis. In Fig. 5.4.4 the noise floor and the peak of each waveform are plotted against the number of activated cells. The data has been compared directly to the modelled equation:

$$C_{noise/peak}(N_{cells}) = C_{max} \cdot \left[1 - \exp\left(-\frac{N_{cells} \times C_{noise/peak,1}}{C_{max}}\right)\right] \tag{5.4.1}$$

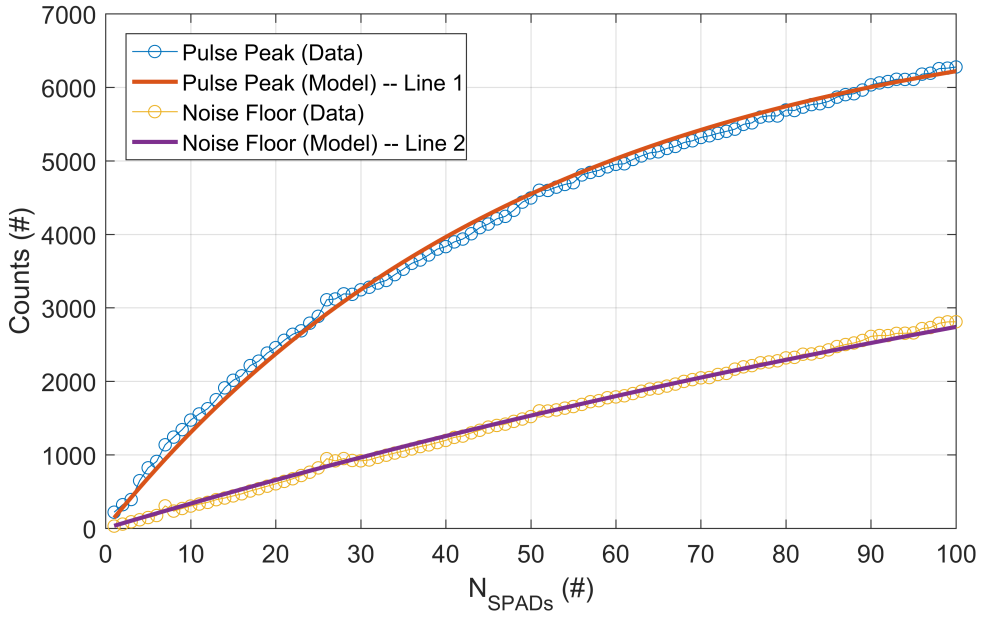


Figure 5.4.3: Distortion due to single-channel - the peak and the noise floor of the reconstructed optical waveform is plotted. A high number of cells makes the registered counts deviate from the ideal linear behaviour making the output waveform having a distorted (lower) height.

where $C_{noise/peak,1}$ is the average recorded count of a single bin when only one cell is enabled and C_{max} is the maximum count per each bin. The latter is directly given by the maximum count rate of the single channel (channel bandwidth f_{BW}), the number of laser pulses $N_{pulses} = PRR \times T_{exp}$ per measurements and the size of the TDC bin:

$$C_{max} = f_{BW} \times T_{bin} \times N_{pulses} \tag{5.4.2}$$

For the plotted curves, the three parameters have been extracted by a fitting algorithm:

$$C_{peak,1} = 144$$

$$C_{noise,1} = 34.5$$

$$C_{max} = 7188.5$$

The deviation of the fit can be read as a deviation from the ideal assumption of identical cell in the dSiPM and in the ideal linearity of the TDC.

Apart from the height, the original square pulse has been distorted into a shorter square pulse. To show something more graphically powerful and also relevant to applications as PET scanning, a second experiment has been run to emulate a decay response from a light source and analyse the pile-up effect in such optical shapes. The distortion of a square pulse onto an exponential decay due to the high light intensity and the dead time of each unit SPAD cell has been exploited to recreate the optical shape. To do so, the power of the incident light has been increased by the use of less aggressive diffusers and by illuminating the detector with zero degree from the laser source. The distorted pulse is shown in Fig. 5.4.4. Although this does not come from an actual exponential decay light source, this waveform can be used for this analysis without the need of fluorescent materials, which made the experiment much easier to build up. Now, using a fit, the decay has been characterised with a single constant decay of $\tau_{decay} = 1.4$ ns. The output waveform $W_1(t)$ of the single cell can be written as:

$$W_1(t) = D_1 \times \exp\left(-\frac{t}{\tau_{decay}}\right) \tag{5.4.3}$$

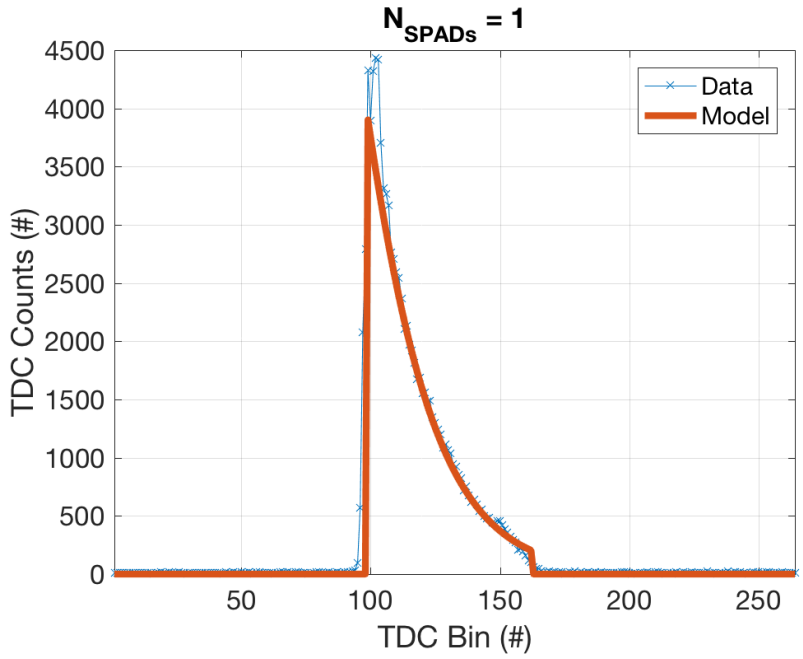


Figure 5.4.4: Detector pile-up experiment - A saturated SPAD cell distorts a square pulse into an exponential decay waveform. This is used on purpose in a pile-up distortion experiment to analyse how a single-channel dSiPM can distort this shap onto a square pulse. The collected counts (in blue) are compared to the modelled funtion $W_1(t)$ of (5.4.3).

The parameter D_1 is given by:

$$D_1 = \Phi_{laser} \times PDE \times T_{bin} \times N_{pulses} \tag{5.4.4}$$

where Φ_{laser} is the incident photon rate on the sensor (in photons per second), PDE. For simplicity, here the parameter has been empirically evaluated as $6.63 \cdot 10^3$ photons per second. The time decay can be estimated by:

$$\tau_{decay} = \frac{1}{\Phi_{laser} \times PDE} \tag{5.4.5}$$

By aggregating more and more SPAD cells, these waveforms together are effectively summed up onto a single channel through the XOR tree. The channel pile-up effect here is more dramatic. In fact, the bins corresponding to the start of the decay pulse, which contain higher counts, reach the saturation limit of the single-channel and therefore the shape of the pulse starts being distorted by cutting the upper part of the pulse. As results, first of all, a simple

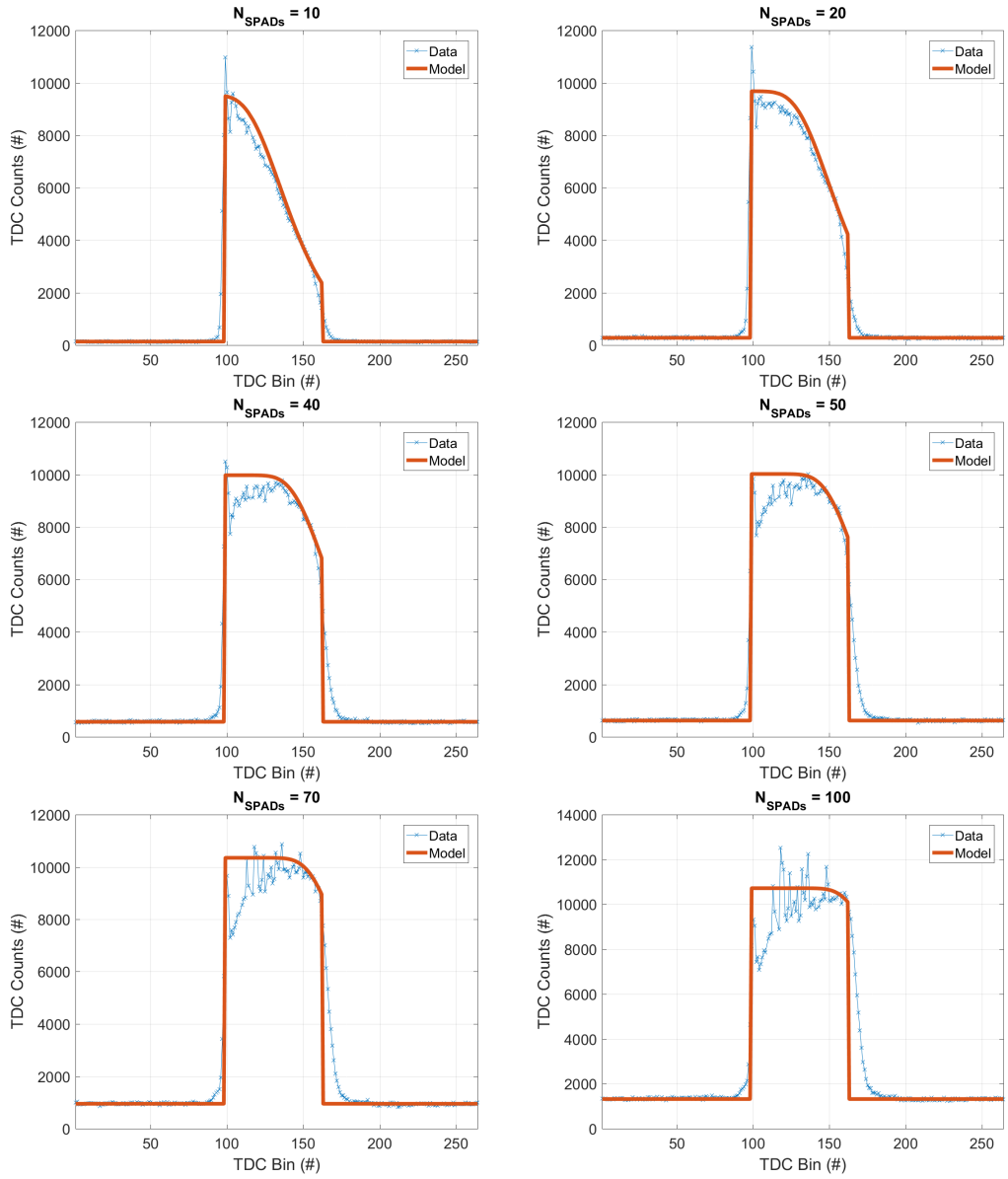


Figure 5.4.5: Distortion of an exponential decay waveform - As soon as the number of cells increases, the dSiPM saturation starts clipping the top of the output waveform distorting its shape from a pure exponential decay curve onto a square wave which therefore does not contain any information on the time constant of the decay. All the data points are plotted together with the proposed model of (5.4.6).

extraction of the time decay constant would fail due to the corrupted pulse shape. With more and more SPADs, the shape is effectively lost and transformed into a square pulse. This is due to the recorded counts from each cells being higher than the capability of the single channel, as results, further events are not timed/counted and therefore the curve does not reach higher values. The resulting curve is cut to a square pulse shape. At this stage, all the decay information is lost, see Fig. 5.4.5. The modelled waveforms $W_{N_{cells}}(t)$, the red line plots, can be calculated as:

$$W_{N_{cells}}(t) = C_{max} \times \left[1 - \exp\left(-\frac{N_{cells} \times W_1(t)}{C_{max}}\right)\right] \tag{5.4.6}$$

where the parameter C_{max} represents the maximum values of successfully timed events due to the channel bandwidth. As noted by the equation itself, this represents the effect due to the limited bandwidth of the timing system and takes as input the individual output of each cells $W_1(t)$ which are here supposed to be identical for simplicity of the mathematical expression.

However, the final plots of the same figure show an additional distortion: the initial part of the pulse, after being cut to the maximum level, starts showing lower recorded events. This saturation effect is not represented in the modelled output of Fig. 5.4.5 and (5.4.6) but the reason of such effect has been already hinted at in Chapter 3. The XOR-tree, although allowing higher throughputs compared to the OR-tree, has the disadvantages of failing when the inputs transition are very close in time: this is due to the inability of the XOR to create two or more consecutive digital transitions with the effect of cancelling all of them. The probability of having close in time input transitions to the XOR tree increases with the number of enable input cells. This is more evident a the beginning of the pulse, where the single cell record a higher number of events (counts). To include this into the model, the Eq. 5.4.6 is multiplied point-by-point by the probability of the input of the XOR-tree to exceed the limit of the XOR itself. This has been done empirically and the chosen expression for the probability of cancellation is given by:

$$p_{XOR}(t) = \exp\left(-\frac{N_{cells} \times W_1(t)}{C_{XOR}}\right) \tag{5.4.7}$$

where C_{XOR} has been extracted empirically to $C_{XOR} = 1 \cdot 10^6$ and represents the maximum

counts handled by the XOR before cancellation. The improved model is summarised by:

$$W^*_{N_{cells}}(t) = W_{N_{cells}}(t) \times p_{XOR}(t) \tag{5.4.8}$$

The improved modelled output has been plotted against the same set of data in Fig. 5.4.6. The closeness between data and the model represents an important validation of the predicted limitation of XOR-based dSiPMs with a high rate of input events due both to single channel limitation and to the intrinsic nature of the digital gate.

5.5 Summary

In this chapter, results coming from the analysis of the throughput of dSiPMs have been presented. The analysis with continuous light analysis and have proven how XOR-based dSiPMs overcome limitations of classic OR-based dSiPMs thanks to the use of toggle cells instead of monostable pulse-shortener circuits. A maximum throughput in the order of 1 Gphotons per second in dSiPM design with 7 μm unit cell pitch have been proven and documented. Moreover, analytical expressions have been introduced to model the throughput and dynamic range of the detectors highlighting each limiting factors in dSiPM design. An optimised XOR-based dSiPM has been paired with the fastest TDC presented in literature at the time of this publication. As results, a demonstration of advanced TCSPC has been here proposed. This does not require the classical limit of $\ll 1$ photons per laser pulse but instead a successful number of converted photons of ~ 16 per laser pulse is here documented. This feature enables faster TCSPC systems allowing more efficient LiDAR detections and would offer a robust dynamic range also useful for rejection of distortion in PET systems. To enrich the analysis, all the experimental results have been paired with analytical expression useful for further modelling. Each contribution of pile-up distortion in a full dSiPM sensor, from the unit cell (detector pile-up) to the N-to-single channel network (channel pile-up) down to the read-out system (counter/converter pile-up) has been successfully characterised. This analytical model allows an easy prediction of the performance of a given design in certain light conditions and dSiPM configuration. Important considerations can then be derived by the usage of the model: for example, the optimisation

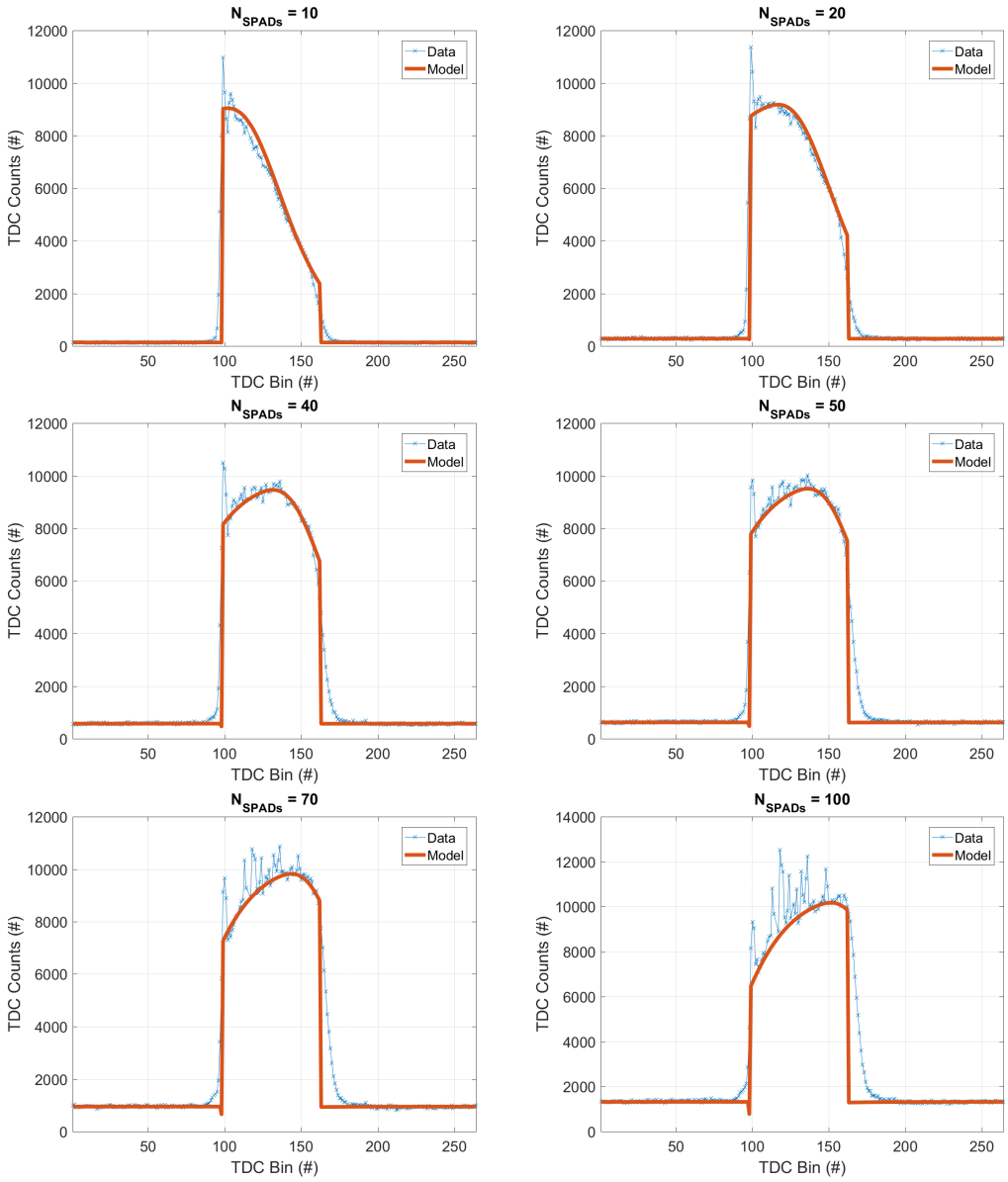


Figure 5.4.6: Distortion due to XOR cancellation - An improved analytical model shown in (5.4.8) describes the reduction of registered counts in an XOR-tree when the input rate exceeds an intrinsic saturation level of the tree here empirically determined.

of the dSiPM for a better exploitation of the dynamic range, the comparison between different technology nodes or also the evaluation of the single-channel versus multi-channel solutions according to the specifications of the application.

Chapter 6

Conclusions and Outlook

The work presented in this thesis document has illustrated an intense analysis on design challenges of the recently introduced digital silicon photomultiplier. The driving goal of the entire study as been to full understand and solve, where possible, all the potential causes of non-linear distortion in a dSiPM sensor. ‘

6.1 Summary

Typical TCSPC systems are limited by the low data throughput of time-to-digital converters needed for high precision and high accuracy measurements. For such reasons, it is very common to reduce the intensity of incident light to avoid the loss of events in the conversion process. This solution represents a bottleneck in applications where the incident light pulse is characterised by intense and fast bursts and slow exponential decays, as in PET systems, but also in applications where a high photon budget is available and yet not exploited to avoid distortions causing a significant slow down in the operation speed of the sensor. Particular attention has been given to 3D LiDAR systems under the assumptions of outdoors operation, i.e. high ambient light, and from low to high reflective objects, i.e. a high range of incoming signal-light. The dynamic range of such devices has been therefore intensely studied.

For such reason, the first optimisation for dSiPM sensors here proposed consists on the realisation of the highest sampling frequency TDC in literature at moment of the publication of this work. The designed chip, referred to as the METDC, is based on a histogramming multiple-event flash architecture with no conversion dead time. The chip has been realised in 130 nm CMOS imaging technology demonstrating a 14 GSamples/s sampling frequency.

This represents a $\times 10$ improvement over the highest throughput solutions in literature. The bin width of the TDC is measured to be as low as 71 ps. The performance of the TDC has been tested in a LiDAR experiment thanks to an on-chip SPAD detector and a controlled laser pulse. The overall throughput of the sensor has been demonstrated to reach the GPhotons/s level. The resolution of system has been characterised by measuring the minimum change in the ToF resolved by the TDC, reported as 168 ps including the effect of the laser, detector and TDC jitter all together. Moreover, to underline the ability of the proposed TDC to detect more than a single time stamp per laser pulse, a two-event resolution has been characterised by measuring the minimum difference in ToF of two separate events (two laser pulses) to be resolved as separate histogram peaks in the TDC output. This is measured to be 1.4 ns which demonstrates the maximum sampling frequency of 14 GSamples/s.

The outstanding performance of the realised converter highlighted the necessity to fully understand how to design the detector part of the sensor to match the single-channel TDC throughput.

Analysing the dynamic range of digital SiPMs has been considered as the key goal that allowed this technology to be brought to more competitive performance against its direct competitor which is represented by linear avalanche photo-diodes, APD.

The approach adopted to answer this question consists of assembling a high level Monte-Carlo simulator of the main parts of the sensing chain: the Poissonian photon generation; the detection process on each SPAD cell; the combination onto single-channel through digital networks; and the timing/counting readout. This has given the ability to re-produce the output of dSiPMs over a selection of different scenarios (light conditions and pulsed illuminations) and varying typical design choices (both on the detector side and on the readout chain) thanks to a series of fully programmable parameters including:

- Indicent light level (incoherent light, DC value)
- Indicent pulsed light level (amplitude, width and frequency of a laser pulse)
- SPAD cell size, *PDE* and dead time
- Number of SPAD cells per dSiPM
- Digital N-to-1 network
  - OR-tree with programmable monostable pulse width

  - XOR-tree
- Readout bandwidth
- TDC bin size

The output of the simulator has been analysed on the time-domain by plotting the obtained waveforms or passing them onto a more sophisticated SPICE simulator. Moreover, the frequency-domain, i.e. the count rate domain, has been studied by scanning the waveforms and counting the detected events over the simulated acquisition time. The entire simulator is here provided as flow-charts (see Appendix for correspondent MATLAB codes) to give the reader the full freedom to implement it in the preferred computational language or simulating tool.

The simulator has been utilised in this work to best analyse the possible scenarios in dSiPM designs when pile-up distortions occur. In particular, results have been presented showing how an incident optical square pulse can be distorted due to the high intensity of the incident photons per SPAD cell, causing the output waveform to exhibit an exponential decay distortion with a time constant that has been linked to the incident rate and the *PDE* of the cell. Moreover, it has shown how a square pulse can also be distorted in amplitude when the number of cells in the dSiPMs is too high for the single-channel readout to match the high detection rate of the detector array. To make the results relevant also for PET and life-time extracting systems, the distortion of an exponential decay has been studied. It has been demonstrated how the limited bandwidth of the single-channel tends to distort the output shape onto a square pulse waveform due to saturation of the single-channel readout: after reaching a maximum in the registered the upper part of the waveform is effectively cut making the waveform appear as a square pulse and inhibiting any decay time constant extraction. The obtained results have been intensively analysed on both timing and counting domain which led to the formulation of useful equations to describe and predict:

- Count rates versus input photon rate: (3.4.6) and (3.4.4)
- Maximum counts: (3.4.7) and (3.4.5)
- Dynamic range: (2.1.8)
- Temporal output waveform for a given input optical shape: Section 3.5

A configurable dSiPM chip consisting of a set of selectable SPAD arrays and two available

combining networks has been specifically designed to validate the modelled results. The chip has been manufactured, as the METDC chip, in 130 nm CMOS imaging technology. On-chip counters have demonstrated a maximum frequency of a single channel in the given technology equal to 1.09 GHz. All the simulations have been recreated in the laboratory thanks to controllable laser sources for both continuous illumination (LED) and for pulsed light illumination (laser). The collected data allowed the verification of the set of equations found in the theoretical analysis and the validation of the quality of the developed dSiPM simulator.

6.2 Discussion

The outcome of the work presented in this thesis can be used to discuss the design of the next generation of single-point dSiPMs sensors. These type of detectors are of major interest for both large area PET scanner and in particular for scanning LiDAR systems. Conclusions can be extended in the analysis of imaging devices, i.e. arrays of single-point sensors.

First of all, this thesis has highlighted the advantages of using, for a given available active area, smaller SPADs aggregated onto a dSiPM type of sensor compared to the realisation of big area SPADs. The loss in PDE due to fill factor is compensated by the potential lower dead time and the higher number of SPAD cells that enhances the overall dynamic range of the detector. The work has demonstrated how the dynamic range is essential to avoid distortions without the need of using strong filters to lower the level of incident light which extends the acquisition time compromising the speed of operation of the sensors and resulting in a significant increase of power consumption.

Even though the thesis has focused its attention onto single-channel dSiPMs, the obtained results can be used to extend the analysis to multi-channel dSiPMs. In fact, the work presented in the previous chapters highlights how essential it is to characterise the available technology which typically sets the bandwidth of a signal-channel and also to consider and specify the throughput and sampling rate of the readout circuitry needed for the application. To avoid channel and timing pile-up, this work has presented a criterion to choose the optimal number of SPADs to be aggregated together onto a single-channel dSiPM. In the given example, a

limited number of 13 SPADs has been identified as the optimal trade-off.

The initial question for a discussion can be the following: what is the most efficient design for a dSiPM sensor to match the requirements of a given application? The author suggests the following process. First of all it is necessary to calculate at system level the amount of minimum and maximum light that the application requires to be successfully detected by the sensor. The minimum value will set a requirement for the noise floor, i.e. the *DCR*, while the maximum light level will set the specification for the overall sensor throughput. An optimal design has to try to achieve the best performance with no additional filtering or diffusing of the light. This is because filters and diffusers attenuate both the noise level and the signal therefore even though they might solve the saturation problem of the detector, the price to pay is a longer acquisition time, not amenable for many reasons (power consumption, speed etc.). Once these specifications are clear to the designer, they must be met taking into account additional parameters such as: available active area, technology node, power constraints and overall costs. The requirements on the noise floor is very important in applications where the *ambient light* can be considered low, such as PET applications. The choice of low-noise SPAD is out of the scope of this work.

The requirement on the upper limit of the incident light level is what this work has investigated. This is due to the inherent limitation of SPAD devices in terms of dynamic range. To improve the overall performance against linear photo-diodes, the upper limit has to be enhanced by adapting the design to the requirements of the application such as outdoor 3D LiDAR. Ideally, the rate of incident photons should be matched with the overall detection rate and therefore it can be used to specify the throughput of the converter. When the conversion rate does not need to be high, then the best choice is to prefer slow but high resolution converters in order to exploit the low photon rate incident on the sensor by accurately resolving time-stamps effectively in one single measurement. The challenge is at high incident photon rates to then combine a high throughput and a high time-resolution taken into account that most converters trade-off high resolution with low throughput. However, in high light conditions, the resolution can be improved by histogramming which naturally benefits of the high conversion rate. Once the chosen TDC architecture has been established, the designer must understand

if the technology allows a single channel to match such conversion rate. This work illustrates how the 14 GSample/s of the METDC enables to oversample the single-channel bandwidth of 1 GHz. This parameter has to be simulated in the given technology: the designer must understand the rate of events that can be encoded on a given electrical node. This work has demonstrated how XOR-based dSiPMs exploit both transitions of the digital signal to encode as many events as possible. In case the technology does not support the required rate of events, the designer has to consider a multi-channel approach. In this case, the number of channels needed can be estimated by the ratio of the required throughput (considered oversampling) and the maximum bandwidth of the single-channel.

The required detection rate per channel (if more than one) needs to be matched by a SPAD design exhibiting the correspondent dead time, see (3.4.4), (3.4.6). The designer has now the task to understand if this required dead time can be met by a single SPAD occupying the entire available active area or if an array of SPADs is required. This information can come either from previous SPAD designs or from technology computer-aided design (TCAD) simulations. In case a large SPAD does not meet the dead time requirement, smaller designs has to be considered. Since the dead time of a SPAD is proportional to its load capacitance and therefore to the area, one can describe the division of the total active area A_{total} into N_{cells} SPADs by:

$$\tau_{dead}(N_{cells}) \propto A_1(N_{cells}) \propto \frac{A_{total}}{N_{cells}} \tag{6.2.1}$$

The achieved maximum detection rate, assuming no saturation of the single-channel, can be extracted by (3.3.1) as:

$$m_{\max} = \frac{N_{cells}}{e \cdot \tau_{dead}(N_{cells})} \tag{6.2.2}$$

The designer can therefore specify the dead time and size of the unit SPAD cell so that the maximum detection rate matches the single-channel bandwidth expressed in events rate. Note that the designer has to keep in mind what can be done in the technology, i.e. has to evaluate what the smallest SPAD dimension in the given process is to ensure a good performance, and if the correspondent minimum dead time is achievable. In case this is not possible, a larger total active area or filters are required to reduce the rate of incident photons on each minimum size

SPAD cell. The proposed design flow is shown as flow chart in Fig. 6.2.1.

6.3 Future Work

The research presented in this thesis can be extended starting from the following considerations. The high throughput of the proposed METDC and its one-shot resolution, i.e. the bin width of the histogramming system, can be expanded implementing additional interpolating techniques. An advanced concept of the METDC is included in the same patent document [11]. The evolution of the architecture sees the introduction of an interpolating stage on the data line, i.e. on the SPAD channel. The patent illustrates an example of such interpolation implemented with the usage of a resistor ring to achieve sub-gate delay enabling finer time-resolution. By preserving the parallel readout through the flip-flop and XOR ring system, the interpolative METDC does not only achieve a lower bin size but also a higher sampling rate and therefore throughput. The factor of improvement is linked to the number of interpolative stages. The implementation of this architecture has not been validated on a design level which therefore represents an interesting chance to demonstrate and validate the concept of the folded flash architecture.

Another point of improvement consists in the SiPM simulator. The model here presented has been kept at a very high level to focus the attention on open questions that the research has been based on such as the number of the cells per dSiPM, digital network etc. However, many parameters have been intentionally left out but that can be included in the simulator to improve the match with a real SiPM output. Some examples include the implementation of **dark events**, i.e. the *DCR* level of each SPAD cell, **after-pulsing**, **cross-talk** and **timing jitter**. The former can be implemented just as the generation of a constant level of photons, since the nature of the dark counts is uniformly random in time, the triggering of each cell can be triggered by introducing the additional input vector that can be called “dark count events”. If implemented as input to the SPAD generation, then the rate has to be divided by the *PDE* of the cell so that the output rate will be as the desired *DCR* level. Alternatively, this can be implemented inside the detection routine of the SPAD cell and add the *DCR* as a probability of event triggered when

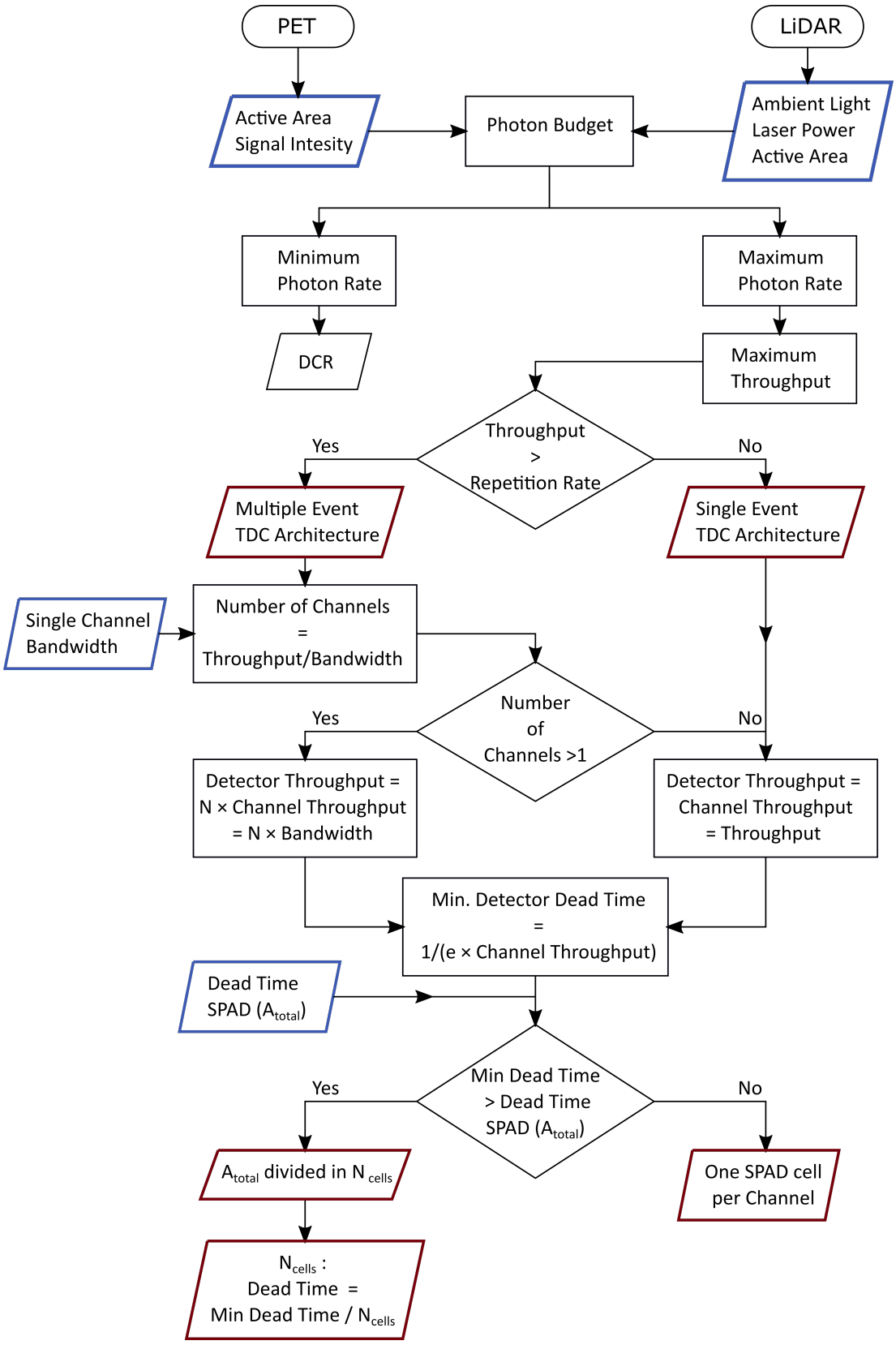


Figure 6.2.1: Proposed design flow - Highlighted in blue, the input coming from analysis on the optical photon budget and information regarding the CMOS technology including the SPAD performance. The flow chart is intended as starting point of the discussion for future design of dSiPM receivers.

no photons are incident on the cell. Both approaches are believed to have the same output. The after-pulsing, being a time-correlated noise, cannot be implemented as in an analogue way as a constant light level and therefore requires to be inserted within the SPAD detection function. As input, this function should have a time-depending probability of occurrence starting from the last photo-detection of that same cell. Cross-talk however requires spatial information to be included in the SiPM routing. The model at this stage does not have any information on the geometrical structure of the SPAD array which is instead essential for the simulation of cross-talk. A solution would be inserting an identification variable to each SPAD cell, an X-Y identifier, that can be used to simulate cross-talk. For example: the time stamps of an adjacent SPAD cells can be used as input to the neighbours cells to enable the generation of cross-talk events. This has a major impact on the level of complexity of the simulator. Last, the timing jitter can be inserted as a function having as input the ideal *no-jitter* output of the SPAD/SiPM (depending on the level of implementation) and the jitter sigma (or FWHM). The aim of this function would be to detect every event-transition and move its position on the output vector by a time-step equivalent to a random-generated number of steps with zero-mean and standard deviation equal to the jitter. As results, the output vector would be a copy of its input but with a random uncertainty on the time detection of the events which is in fact the definition of timing jitter.

In addition to the SiPM output, the simulator can be coupled with a system-level TDC modelling which differs from the ideal TDC assumed in the current model. The utility of incorporating different TDC designs would be to compare or recreate the output of dSiPM sensors with specific readout systems. Moreover, a digital-to-analogue step could be easily implemented to emulate aSiPMs by generating current signals from the rate of detected events with additional input such as: gain, rise time etc.

A further study can also involve a more intense testing of the manufactured test chips. One example would be to investigate different shape of laser pulses for pile-up investigations using the METDC test chip. An alternative set-up could involve a fluorescence lifetime experiment to practically see the distortion due to each of the sources and establish the relation between the extracted decay time with no-pile up and with different levels of sensor saturation. The

model could be recalled to predict and analytically characterise such relation. These results could be analysed in terms of pile-up correction techniques. The obtained information could be used to adjust the configuration of the dSiPM according to the extracted data to minimise the distortions.

Both of the test chips were not analysed in terms of timing jitter. The METDC has been tested in a LiDAR experiment and the overall sensor linearity included the effect of the detector jitter but its characterisation and its dependency on number of SPADs, bias voltage etc. has yet to be done. Moreover, thanks to the second test chip manufactured in the same technology, the jitter performance of thee detector could eventually include further parameters such as SPAD size and digital network. Future test chips might need to be required to improve the signal routing which, in the case of the second test chip, has been severely limited due to the large amount of SPAD array variety and the limited chip dimensions compromising the quality of the timing experiment results. However, the described structure of both chips represent a good point to start the design of improved test chips.

Finally, a natural step would be comparing the results obtained during this research with future technologies and advanced SPAD designs. The evolution of CMOS and the progress in the SPAD design would enable the exploitation of: high *PDE* SPAD cells, lower dead times, a better fill factor (with particular interest in 3D stacking and the use of micro-lenses), higher channel bandwidths, faster gate delay and many more. This research presents a valid method to compare and predict the impact of each of this technology improvement on existing and future dSiPM designs which will help designers understanding the possible trade-offs both at system level and at CMOS design level in order to fulfil the different application requirements.

Bibliography

[1] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. Pellegrini, L. A. Grant, and R. K. Henderson, "Gphotons/s TCSPC with XOR Digital Silicon Photomultipliers," in *Single Photon Workshop*, 2015.

[2] S. Gnecchi, N. A. W. Dutton, L. Parmesan, B. R. Rae, S. J. McLeod, S. Pellegrini, L. A. Grant, and R. K. Henderson, "A Simulation Model for Digital Silicon Photomultipliers," *IEEE Transactions on Nuclear Science*, vol. 63, no. 3, pp. 1343–1350, 2016.

[3] S. Gnecchi, N. A. W. Dutton, L. Parmesan, R. Bruce, S. Pellegrini, S. J. Mcleod, L. A. Grant, and R. K. Henderson, "Analysis of Photon Detection Efficiency and Dynamic Range in SPAD based Visible Light Receivers," *Journal of lightwave technology*, vol. 8724, no. 1, pp. 1–7, 2016.

[4] 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, 2016.

[5] O. Almer, N. A. W. Dutton, T. A. Abbas, S. Gnecchi, and R. K. Henderson, "4-PAM Visible Light Communications with a XOR-tree Digital Silicon Photomultiplier," in *IEEE Summer Topicals Meeting Series*, 2015, pp. 2–3.

[6] N. Dutton, J. Vergote, S. Gnecchi, L. Grant, D. Lee, S. Pellegrini, B. Rae, and R. Henderson, "Multiple-event direct to histogram TDC in 65nm FPGA technology," *Ph.D. Research in Microelectronics and Electronics (PRIME), 2014 10th Conference on*, pp. 1–5, 2014.

[7] 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 *Solid- State Circuits Conference - (ISSCC), 2015 IEEE International*, 2015, pp. 204–206.

[8] O. Almer, D. Tsonev, N. A. W. Dutton, T. A. Abbas, S. Videv, S. Gnecchi, H. Haas, and R. K. Henderson, "A SPAD-based Visible Light Communications Receiver Employing Higher Order Modulation Schemes," *IEEE Globecom*, 2015.

[9] 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, 2016.

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

[11] N. A. Dutton, R. K. Henderson, and S. Gnecchi, "TIME TO DIGITAL CONVERTER AND APPLICATIONS THEREOF," 2015.

[12] F. Remondino and D. Stoppa, *TOF range-imaging cameras*. Springer, 2013.

[13] A. A. M. van der Veldt, E. F. Smit, and A. A. Lammertsma, "Positron Emission Tomography as a Method for Measuring Drug Delivery to Tumors in vivo: The Example of [(11)C]docetaxel." *Frontiers in oncology*, vol. 3, no. August, p. 208, 2013.

[14] M. M. Graham, "Physiologie Smoothing of Blood Time-Activity Curves for PET Data Analysis," *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 38, no. 7, pp. 1161–8, 1997.

[15] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*. Springer Berlin Heidelberg, 2005, vol. 81.

[16] A. McCarthy, R. J. Collins, N. J. Krichel, V. Fernández, A. M. Wallace, and G. S. Buller, "Long-range time-of-flight scanning sensor based on high-speed time-correlated single-photon counting." *Applied optics*, vol. 48, no. 32, pp. 6241–51, 2009. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/19904323

[17] PicoQuant, "Time Correlated Single Photon Counting Wiki." [Online]. Available: http://www.tcspc.com/doku.php/glossary:pile-up{_}effect

[18] 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, pp. 103 105–103 110, oct 2013.

[19] S. Donati and T. Tambosso, "Single-Photon Detectors: From Traditional PMT to Solid-State SPAD-Based Technology," *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 20, no. 6, 2014.

[20] SensL Technologies, "C-Series SiPM Datasheet," 2014. [Online]. Available: http://sensl.com/downloads/ds/DS-MicroCseries.pdf

[21] T. Frach, G. Prescher, C. Degenhardt, R. de Gruyter, A. Schmitz, and R. Ballizany, "The Digital Silicon Photomultiplier - Principle of Operation and Intrinsic Detector Performance," *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, pp. 1959–1965, oct 2009.

[22] L. H. C. Braga, L. Pancheri, L. Gasparini, M. Perenzoni, R. Walker, R. K. Henderson, and D. Stoppa, "A CMOS Mini-SiPM Detector with In-Pixel Data Compression for PET Applications," *IEEE Nuclear Science Symposium Conference Record*, pp. 548–552, 2011.

[23] S. Mandai and E. Charbon, "Multi-channel digital SiPMs: Concept, analysis and implementation," in *IEEE Nuclear Science Symposium Conference Record*, no. 1, 2012, pp. 1840–1844.

[24] C. C. Chen, P. Chen, C. S. Hwang, and W. Chang, "A precise cyclic CMOS time-to-digital converter with low thermal sensitivity," *IEEE Transactions on Nuclear Science*, vol. 52, no. 4, pp. 834–838, 2005.

[25] D. I. Porat, "Review of Sub-Nanosecond Time-Interval Measurements," *IEEE Transactions on Nuclear Science*, vol. NS-20, no. 5, pp. 36–51, 1973.

[26] C. S. Hwang, P. Chen, and H. W. Tsao, "A high-precision time-to-digital converter using a two-level conversion scheme," *IEEE Transactions on Nuclear Science*, vol. 51, no. 4 I, pp. 1349–1352, 2004.

[27] 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, 2016.

[28] S. Mandai and E. Charbon, "A 4 x 4 x 416 digital SiPM array with 192 TDCs for multiple

high-resolution timestamp acquisition," *Journal of Instrumentation*, vol. 8, no. 05, p. P05024, 2013.

[29] P. B. Coates, "The correction for photon 'pile-up' in the measurement of radiative lifetimes," *Journal of Physics E: Scientific Instruments*, vol. 1, no. 8, pp. 878–879, 2002.

[30] J. G. Walker, "Iterative correction for 'pile-up' in single-photon lifetime measurement," *Optics Communications*, vol. 201, no. 4-6, pp. 271–277, 2002.

[31] M. Kafaee and S. Saramad, "Pile-up correction by Genetic Algorithm and Artificial Neural Network," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 607, no. 3, pp. 652–658, 2009.

[32] E. Vicente, J. L. Herraiz, S. España, E. Herranz, M. Desco, J. J. Vaquero, and J. M. Udías, "Deadtime and pile-up correction method based on the singles to coincidences ratio for PET," in *IEEE Nuclear Science Symposium Conference Record*, 2012, pp. 2933–2935.

[33] M. Wahl, "Time-correlated single photon counting," pp. 1–14, 2014. [Online]. Available: http://www.picoquant.com/images/uploads/page/files/7253/technote{_}tcspc.pdf

[34] J. R. Lakowicz, H. Szmacinski, K. Nowaczyk, and M. L. Johnson, "Fluorescence lifetime imaging of calcium using Quin-2," *Cell Calcium*, vol. 13, no. 3, pp. 131–147, 1992.

[35] W. Becker, A. Bergmann, M. A. Hink, K. König, K. Benndorf, and C. Biskup, "Fluorescence Lifetime Imaging by Time-Correlated Single-Photon Counting," *Microscopy Research and Technique*, vol. 63, no. 1, pp. 58–66, 2004.

[36] G. Muehllehner and J. S. Karp, "Positron emission tomography." *Physics in medicine and biology*, vol. 51, no. 13, pp. R117—-37, 2006.

[37] S. Surti, A. Kuhn, M. E. Werner, A. E. Perkins, J. Kolthammer, and J. S. Karp, "Performance of Philips Gemini TF PET/CT scanner with special consideration for its time-of-flight imaging capabilities." *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 48, no. 3, pp. 471–480, 2007.

[38] W. W. Moses, "Recent advances and future advances in time-of-flight PET," in *AIP Conference Proceedings*, vol. 1204, 2009, pp. 119–125.

[39] R. Foord, R. Jones, C. J. Oliver, and E. R. Pike, "The use of photomultiplier tubes for photon counting." *Applied optics*, vol. 8, no. 10, pp. 1975–1989, 1969.

[40] M. B. Das, S. Bose, and R. Bhattacharya, "Single photon response of photomultiplier tubes," *Nuclear Inst. and Methods in Physics Research, A*, vol. 242, no. 1, pp. 156–159, 1985.

[41] S. V. Polyakov, "Photomultiplier Tubes," *Experimental Methods in the Physical Sciences*, vol. 45, no. September, pp. 69–82, 2013.

[42] M. Watanabe, K. Shimizu, T. Omura, N. Sato, M. Takahashi, T. Kosugi, K. Ote, A. Katabe, R. Yamada, T. Yamashita, and E. Tanaka, "A high-throughput whole-body PET scanner using flat panel PS-PMTs," *IEEE Transactions on Nuclear Science*, vol. 51, no. 3 II, pp. 796–800, jun 2004.

[43] S. J. Brice, L. Bugel, J. M. Conrad, B. Fleming, L. Gladstone, E. Hawker, P. Killewald, J. May, S. McKenney, P. Nienaber, B. Roe, V. Sandberg, D. Smith, and M. Wysocki, "Photomultiplier tubes in the MiniBooNE experiment," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 562, no. 1, pp. 97–109, 2006.

[44] Y. Guan, "Photomultiplier tubes in biosensors." pp. 375–387, 2009.

[45] S. V. Polyakov, "Photomultiplier Tubes," *Experimental Methods in the Physical Sciences*, vol. 45, pp. 69–82, 2013.

[46] B. K. Lubsandorzhiev, "On the history of photomultiplier tube invention," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 567, no. 1 SPEC. ISS., pp. 236–238, 2006.

[47] S. Yamamoto, H. Watabe, K. Kato, and J. Hatazawa, "Performance comparison of high quantum efficiency and normal quantum efficiency photomultiplier tubes and position sensitive photomultiplier tubes for high resolution PET and SPECT detectors." *Med phys*, vol. 39, no. 11, pp. 6900–7, 2012.

[48] M. Itaya, T. Inagaki, T. Iwata, G. Y. Lim, H. Oishi, H. Okuno, Y. Tajima, H. Y. Yoshida, and Y. Yoshimura, "Development of a new photomultiplier tube with high sensitivity for a wavelength-shifter fiber readout," *Nuclear Instruments and Methods*

in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 522, no. 3, pp. 477–486, 2004.

[49] Y. Haemisch, T. Frach, C. Degenhardt, and A. Thon, "Fully Digital Arrays of Silicon Photomultipliers (dSiPM) – a Scalable Alternative to Vacuum Photomultiplier Tubes (PMT)," *Physics Procedia*, vol. 37, pp. 1546–1560, 2012.

[50] X.-C. Lei, Y.-K. Heng, S. Qian, J.-K. Xia, S.-L. Liu, Z. Wu, B.-J. Yan, M.-H. Xu, Z. Wang, X.-N. Li, X.-D. Ruan, X.-Z. Wang, Y.-Z. Yang, W.-W. Wang, C. Fang, F.-J. Luo, J.-J. Liang, L.-P. Yang, and B. Yang, "Evaluation of New Large Area PMT with High Quantum Efficiency," *arXiv:1504.03174 [hep-ex, physics:physics]*, vol. 40, no. 2, 2015.

[51] I. Yamazaki, N. Tamai, H. Kume, H. Tsuchiya, and K. Oba, "Microchannel-plate photomultiplier applicability to the time-correlated photon-counting method," *Review of Scientific Instruments*, vol. 56, no. 6, pp. 1187–1194, 1985.

[52] K. Inami, T. Ohshima, Y. Enari, K. Hayasaka, M. Tomita, T. Kubota, S. Yoshino, N. Kishimoto, Y. Akita, and Y. Endo, "A TOF counter of σ˜10 ps with Cherenkov photons and a MCP-PMT," *IEEE Symposium Conference Record Nuclear Science 2004.*, vol. 3, no. C, pp. 1947–1950, 2004.

[53] M. Akatsu, Y. Enari, K. Hayasaka, T. Hokuue, T. Iijima, K. Inami, K. Itoh, Y. Kawakami, N. Kishimoto, T. Kubota, M. Kojima, Y. Kozakai, Y. Kuriyama, T. Matsuishi, Y. Miyabayashi, T. Ohshima, N. Sato, K. Senyo, A. Sugi, S. Tokuda, M. Tomita, H. Yanase, and S. Yoshino, "MCP-PMT timing property for single photons," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 528, no. 3, pp. 763–775, 2004.

[54] Hamamatsu Photonics, "Technical Information MCP assembly," 2006.

[55] J. Va'vra, J. Benitez, D. W. G. S. Leith, G. Mazaheri, B. Ratcliff, J. Schwiening, J. Va'vra, J. Benitez, D. W. G. S. Leith, G. Mazaheri, B. Ratcliff, and J. Schwiening, "A 30 ps timing resolution for single photons with multi-pixel Burle MCP-PMT," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 572, no. 1 SPEC. ISS., pp.

459–462, 2007.

[56] Hamamatsu Photonics, "Microchannel Plate Photomultiplier Tube R10754-07-M16," pp. 7–8, 2016.

[57] ——, "Photomultiplier Tubes R10601, R10601-100," pp. 1–2, 2016.

[58] Horiba, "PDS-3000." [Online]. Available: http://www.obbcorp.com/optical-components/light-detectors/PMT-housings/PMT-Detection-System-PMT.html

[59] K. Yamamoto, Y. Fujii, Y. Kotooka, and T. Katayama, "Highly stable silicon pin photodiode," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 253, no. 3, pp. 542–547, 1987.

[60] M. Kyomasu, "Development of an integrated high speed silicon PIN photodiode sensor," *Electron Devices, IEEE Transactions on*, vol. 42, no. 6, pp. 1093–1099, 1995.

[61] Y. Hosono, Sjafruddin, T. Iguchi, and M. Nakazawa, "Fast neutron detector using PIN-type silicon photodiode," *Nuclear Inst. and Methods in Physics Research, A*, vol. 361, no. 3, pp. 554–557, 1995.

[62] L. Gomez-Rojas, N. J. Gomes, X. Wang, P. A. Davies, and D. Wake, "High performance optical receiver using a PIN photodiode and amplifier for operation in the millimeter-wave region," in *2000 30th European Microwave Conference, EuMC 2000*, 2000.

[63] F. Rêgo and L. Peralta, "Si-PIN photodiode readout for a scintillating optical fiber dosimeter," *Radiation Measurements*, vol. 47, no. 10, pp. 947–950, 2012.

[64] K. Johnson, "Photodiode signal enhancement effect at avalanche breakdown voltage," *ISSCC Digest of Tech. Papers*, pp. 64–65, 1964.

[65] R. J. McIntyre, "Multiplication Noise in Uniform Avalanche Diodes," *IEEE Transactions on Electron Devices*, vol. ED-13, no. 1, pp. 164–168, 1966.

[66] J. R. Biard and W. N. Shaunfield, "A Model of the Avalanche Photodiode," *IEEE Transactions on Electron Devices*, vol. ED-14, no. 5, pp. 233–238, 1967.

[67] R. B. Emmons, "Avalanche-photodiode frequency response," *Journal of Applied Physics*, vol. 38, no. 9, pp. 3705–3714, 1967.

[68] H. Ruegg, "An optimized avalanche photodiode," *IEEE Transactions on Electron Devices*, vol. 14, no. 5, pp. 239–251, 1967.

[69] H. Ando and H. Kanbe, "Effect of avalanche build-up time on avalanche photodiode sensitivity," *IEEE Journal of Quantum Electronics*, vol. 21, no. 3, pp. 251–255, 1985.

[70] K. Yamada, T. Tsuchizawa, T. Watanabe, R. Kou, H. Nishi, H. Shinojima, Y. Ishikawa, K. Wada, and S. Itabashi, "Avalanche-mode operation of a simple vertical p-i-n germanium photodiode coupled with a silicon waveguide," *IEEE International Conference on Group IV Photonics GFP*, vol. 1, pp. 329–331, 2011.

[71] M. C. Maas, D. R. Schaart, D. J. Van Der Laan, H. T. Van Dam, J. Huizenga, J. C. Brouwer, P. Bruyndonckx, C. Lemaître, and C. W. E. Van Eijk, "Signal to noise ratio of APD-based monolithic scintillator detectors for high resolution PET," in *IEEE Transactions on Nuclear Science*, vol. 55, no. 3, 2008, pp. 842–852.

[72] W. Sun, Z. Lu, X. Zheng, J. C. Campbell, S. J. Maddox, H. P. Nair, and S. R. Bank, "High-Gain InAs avalanche photodiodes," *IEEE Journal of Quantum Electronics*, vol. 49, no. 2, pp. 154–161, 2013.

[73] K. Deiters, T. Flügel, D. Renker, H. P. Wirtz, T. Kirn, and P. Denes, "Comparison of photomultiplier and avalanche photodiode readout of PbWO4," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 387, no. 1-2, pp. 211–213, 1997.

[74] I. Britvitch, Y. Musienko, and D. Renker, "Investigation of a photon counting avalanche photodiode from Hamamatsu photonics," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 567, no. 1 SPEC. ISS., pp. 276–280, 2006.

[75] W. G. Lawrence, G. Varadi, G. Entine, E. Podniesinski, and P. K. Wallace, "A comparison of avalanche photodiode and photomultiplier tube detectors for flow cytometry," *Proc. SPIE*, vol. 6859, no. 68590, pp. 68 590M–68 590M–11, 2008.

[76] D. Binkley, B. Puckett, M. Casey, R. Lecomte, and A. Saoudi, "A power-efficient, low-noise, wideband, integrated CMOS preamplifier for LSO/APD PET systems," *IEEE Transactions on Nuclear Science*, vol. 47, no. 3, pp. 810–817, 2000.

[77] R. Grazioso, M. Aykac, M. E. Casey, G. Givens, M. Schmand, R. Grazioso, M. Aykac, M. E. Casey, G. Givens, M. Schmand, R. Graziose, M. Aykac, M. E. Casey, G. Givens,

and M. Schmand, "APD performance in light sharing PET applications," *IEEE Transactions on Nuclear Science*, vol. 52, no. 5 I, pp. 1413–1416, 2005.

[78] R. Grazioso, N. Zhang, J. Corbeil, M. Schmand, R. Ladebeck, M. Vester, G. Schnur, W. Renz, and H. Fischer, "APD-based PET detector for simultaneous PET/MR imaging," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 569, no. 2 SPEC. ISS., pp. 301–305, 2006.

[79] B. J. Pichler, M. S. Judenhofer, C. Catana, J. H. Walton, M. Kneilling, R. E. Nutt, S. B. Siegel, C. D. Claussen, and S. R. Cherry, "Performance test of an LSO-APD detector in a 7-T MRI scanner for simultaneous PET/MRI." *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, vol. 47, no. 4, pp. 639–647, 2006.

[80] S. Shimizu, C. M. Pepin, S. Member, M. Bergeron, and R. Lecomte, "APD-Based PET Detector," *IEEE Transactions on Nuclear Science*, vol. 57, no. 1, pp. 55–62, 2010.

[81] J. M. B. Dias, E. S. R. Fonseca, and D. P. Resendes, "Time and Range Averaging of Lidar Echoes Using APD Based Receivers," *Proc. SPIE*, vol. 3757, pp. 158–166, 1999.

[82] R. Agishev, B. Gross, F. Moshary, A. Gilerson, and S. Ahmed, "Simple approach to predict APD/PMT lidar detector performance under sky background using dimensionless parametrization," *Optics and Lasers in Engineering*, vol. 44, no. 8, pp. 779–796, 2006.

[83] Excelitas Technologies, "C30737CH-300-70," pp. 3–6, 2012. [Online]. Available: http://www.excelitas.com/

[84] Hamamatsu Photonics, "S8664 series datasheet," pp. 3–5, 2005.

[85] ——, "G8931-20 Datasheet."

[86] OSI Optoelectronics, "APD Series 8-150," 2016.

[87] 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.

[88] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes," *IEEE Transactions on Nuclear Science*, vol. 57, no. 6 PART 2, pp. 3815–3826, 2010.

[89] V. Savuskan, I. Brouk, M. Javitt, and Y. Nemirovsky, "An Estimation of Single Photon Avalanche Diode (SPAD) Photon Detection Efficiency (PDE) Nonuniformity," *IEEE Sensors Journal*, vol. 13, no. 5, pp. 1637–1640, may 2013.

[90] A. Eisele and R. Henderson, "185 MHz Count Rate, 139 dB Dynamic Range Single-Photon Avalanche Diode with Active Quenching Circuit in 130nm CMOS Technology," *2011 International Images Sensor Workshop*, pp. 6–8, 2011.

[91] F. Zappa, S. Tisa, A. Tosi, and S. Cova, "Principles and features of single-photon avalanche diode arrays," *Sensors and Actuators, A: Physical*, vol. 140, no. 1, pp. 103–112, 2007.

[92] A. Gulinatti, I. Rech, P. Maccagnani, S. Cova, and M. Ghioni, "New silicon technologies enable high-performance arrays of Single Photon Avalanche Diodes." *Proceedings - Society of Photo-Optical Instrumentation Engineers*, vol. 8727, pp. 1–10, 2013.

[93] Excelitas Technologies, "SPCM-AQRH Datasheet," pp. 1–13, 2016.

[94] Mpd, "PDM Series," pp. 4–5, 2007. [Online]. Available: http://www.micro-photon-devices.com/Docs/Datasheet/PDM.pdf

[95] IDQ, "ID120 Datasheet," 2014. [Online]. Available: http://marketing.idquantique.com/acton/attachment/11868/f-00fd/1/-/-/-/-/ID120Datasheet.pdf

[96] Laser Components, "Single Photon Counting Modules." [Online]. Available: http://www.lasercomponents.com/de-en/photodiodes/avalanche-photodiodes/single-photon-counting-modules/

[97] 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.

[98] A. Rochas, M. Gösch, A. Serov, P. A. Besse, R. S. Popovic, T. Lasser, and R. Rigler, "First fully integrated 2-D array of single-photon detectors in standard CMOS technology," *IEEE Photonics Technology Letters*, vol. 15, no. 7, pp. 963–965, 2003.

[99] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter

and 55% detection efficiency at 420 nm," *Journal of Modern Optics*, vol. 61, no. 2, pp. 102–115, 2014. [Online]. Available: http://dx.doi.org/10.1080/09500340.2013.864425

[100] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in 0.35-μm CMOS and based on an event-driven readout for TCSPC experiments," *Proc. SPIE 6372, Advanced Photon Counting Techniques, 63720S*, pp. 63 720S–63 720S–12, 2006.

[101] F. Guerrieri, S. Tisa, A. Tosi, and F. Zappa, "Two-dimensional SPAD imaging camera for photon counting," *IEEE Photonics Journal*, vol. 2, no. 5, pp. 759–774, 2010.

[102] N. Faramarzpour, M. J. Deen, S. Shirani, and Q. Fang, "Fully integrated single photon avalanche diode detector in standard CMOS 0.18-um technology," *IEEE Transactions on Electron Devices*, vol. 55, no. 3, pp. 760–767, 2008.

[103] 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, p. 5849, 2012.

[104] C. Niclass, H. Matsubara, M. Soga, M. Ohta, M. Ogawa, and T. Yamashita, "A NIR-Sensitivity-Enhanced Single-Photon Avalanche Diode in 0.18µm CMOS," 2015.

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

[106] 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, 2009.

[107] D. Palubiak, M. M. El-Desouki, O. Marinov, M. J. Deen, and Q. Fang, "High-speed, single-photon avalanche-photodiode imager for biomedical applications," *IEEE Sensors Journal*, vol. 11, no. 10, pp. 2401–2412, 2011.

[108] 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 160x128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," in *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, 2011, pp.

312–313.

[109] 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, 2012.

[110] M. A. Karami, M. Gersbach, H.-J. Yoon, and E. Charbon, "A new single-photon avalanche diode in 90nm standard CMOS technology." *Optics express*, vol. 18, no. 21, pp. 22 158–22 166, 2010.

[111] R. K. Henderson, E. A. G. Webster, R. Walker, J. A. Richardson, and L. A. Grant, "A 3x3, 5 um pitch, 3-transistor single photon avalanche diode array with integrated 11V bias generation in 90nm CMOS technology," *Technical Digest - International Electron Devices Meeting, IEDM*, pp. 336–339, 2010.

[112] E. A. G. Webster, J. A. Richardson, L. A. Grant, D. Renshaw, and R. K. Henderson, "A Single-Photon Avalanche Diode in 90-nm CMOS Imaging Technology with 44% Photon Detection Efficiency at 690 nm," *IEEE Electron Device Letters*, vol. 33, no. 5, pp. 694–696, 2012.

[113] A. Biber, P. Seitz, and H. Jackel, "Avalanche photodiode image sensor in standard BiCMOS technology," *IEEE Transactions on Electron Devices*, vol. 47, no. 11, pp. 2241–2243, 2000.

[114] K. S. Shah, R. Farrell, R. Grazioso, R. Myers, and L. Cirignano, "Large-area APDs and monolithic APD arrays," in *IEEE Transactions on Nuclear Science*, vol. 48, no. 6 II, 2001, pp. 2352–2356.

[115] I. C. Sandall, S. Zhang, and C. H. Tan, "Linear array of InAs APDs operating at 2 μm." *Optics express*, vol. 21, no. 22, pp. 25 780–25 787, 2013.

[116] G. Bondarenko, P. Buzhan, B. Dolgoshein, V. Golovin, E. Guschin, A. Ilyin, V. Kaplin, A. Karakash, R. Klanner, V. Pokachalov, E. Popova, and K. Smirnov, "Limited Geiger-mode microcell silicon photodiode: new results," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 442, no. 1, pp. 187–192, 2000.

[117] V. Saveliev and V. Golovin, "Silicon avalanche photodiodes on the base of metal-

resistor-semiconductor (MRS) structures," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 442, no. 1-3, pp. 223–229, mar 2000.

[118] N. Otte, "The Silicon Photomultiplier-A new device for High Energy Physics, Astroparticle Physics, Industrial and Medical Applications," *Proceedings to SNIC symposium (SLAC, Stanford, ...*, no. April, pp. 1–9, 2006.

[119] P. Eraerds, M. Legré, a. Rochas, H. Zbinden, and N. Gisin, "SiPM for fast Photon-Counting and Multiphoton Detection." *Optics express*, vol. 15, no. 22, pp. 14 539–14 549, 2007.

[120] D. J. Herbert, N. D'Ascenzo, N. Belcari, A. Del Guerra, F. Morsani, and V. Saveliev, "Study of SiPM as a potential photodetector for scintillator readout," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 567, no. 1 SPEC. ISS., pp. 356–359, 2006.

[121] D. R. Schaart, E. Charbon, T. Frach, and V. Schulz, "Advances in digital SiPMs and their application in biomedical imaging," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 809, pp. 31–52, 2016.

[122] Hamamatsu Photonics, "MPPC S13360 Datasheet," pp. 1–7, 2016. [Online]. Available: http://www.hamamatsu.com/resources/pdf/ssd/s13360-2050ve{_}etc{_}kapd1053e.pdf

[123] B. R. G. B. Sipms and B. R. G. B. Sipms, "First Sensor," pp. 1–8. [Online]. Available: www.first-sensor.com

[124] P. Eckert, H. C. Schultz-Coulon, W. Shen, R. Stamen, and A. Tadday, "Characterisation studies of silicon photomultipliers," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 620, no. 2-3, pp. 217–226, 2010.

[125] V. Puill, C. Bazin, D. Breton, L. Burmistrov, V. Chaumat, N. Dinu, J. Maalmi, J. F. Vagnucci, and A. Stocchi, "Single photoelectron timing resolution of SiPM as a function of the bias voltage, the wavelength and the temperature," *Nuclear*

Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, vol. 695, pp. 354–358, 2012. [Online]. Available: http://dx.doi.org/10.1016/j.nima.2011.12.039

[126] P. F. Bloser, J. Legere, C. Bancroft, M. L. McConnell, J. M. Ryan, and N. Schwadron, "Scintillator gamma-ray detectors with silicon photomultiplier readouts for high-energy astronomy," *UV, X-Ray, and Gamma-Ray Space Instrumentation for Astronomy XVIII*, vol. 8859, p. 88590A, 2013.

[127] F. Powolny, E. Auffray, S. E. Brunner, E. Garutti, M. Goettlich, H. Hillemanns, P. Jarron, P. Lecoq, T. Meyer, H. C. Schultz-Coulon, W. Shen, M. C. S. Williams, E. Auffray, S. E. Brunner, M. Despeisse, E. Garutti, M. Goettlich, H. Hillemanns, P. Lecoq, T. Meyer, F. Powolny, W. Shen, H. C. Schultz-Coulon, M. C. S. Williams, E. Auffray, S. E. Brunner, E. Garutti, M. Goettlich, H. Hillemanns, P. Jarron, P. Lecoq, T. Meyer, H. C. Schultz-Coulon, W. Shen, M. C. S. Williams, E. Auffray, S. E. Brunner, M. Despeisse, E. Garutti, M. Goettlich, H. Hillemanns, P. Lecoq, T. Meyer, F. Powolny, W. Shen, H. C. Schultz-Coulon, and M. C. S. Williams, "Time based readout of a silicon photomultiplier (SiPM) for time of flight positron emission tomography (TOF-PET)," *2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC)*, vol. 58, no. 3 PART 1, pp. 1212–1219, oct 2009.

[128] S. Seifert, D. R. Schaart, H. T. Van Dam, J. Huizenga, R. Vinke, P. Dendooven, H. Lohner, and F. J. Beekman, "A high bandwidth preamplifier for SiPM-based TOF PET scintillation detectors," in *IEEE Nuclear Science Symposium Conference Record*, 2008, pp. 1616–1619.

[129] S. Seifert, R. Vinke, H. T. Van Dam, H. Lohner, P. Dendooven, F. J. Beekman, and D. R. Schaart, "Ultra precise timing with SiPM-based TOF PET scintillation detectors," *IEEE Nuclear Science Symposium Conference Record*, pp. 2329–2333, 2009.

[130] S. Seifert and D. R. Schaart, "Improving the time resolution of TOF-PET detectors by double-sided readout," *IEEE Transactions on Nuclear Science*, vol. 62, no. 1, pp. 3–11, 2015.

[131] A. Del Guerra, N. Belcari, M. Giuseppina Bisogni, F. Corsi, M. Foresta,

P. Guerra, S. Marcatili, A. Santos, and G. Sportelli, "Silicon Photomultipliers (SiPM) as novel photodetectors for PET," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 648, no. SUPPL. 1, pp. S232–S235, 2011. [Online]. Available: http://dx.doi.org/10.1016/j.nima.2010.11.128

[132] M. Akiba, K. Inagaki, and K. Tsujino, "Photon number resolving SiPM detector with 1 GHz count rate," *Optics Express*, vol. 20, no. 3, p. 2779, 2012.

[133] G. Zhang, C. Yu, C. Zhu, and L. Liu, "Feasibility study of Multi-Pixel Photon Counter serving as the detector in digital optical communications," *Optik*, vol. 124, no. 22, pp. 5781–5786, 2013.

[134] N. Pavlov, "Silicon Photomultiplier and Readout Method," 2013.

[135] S. Dolinsky, G. Fu, and A. Ivan, "Timing resolution performance comparison for fast and standard outputs of SensL SiPM," *IEEE Nuclear Science Symposium Conference Record*, 2013.

[136] C. Degenhardt, G. Prescher, T. Frach, A. Thon, R. De Gruyter, A. Schmitz, and R. Ballizany, "The digital silicon photomultiplier - A novel sensor for the detection of scintillation light," *IEEE Nuclear Science Symposium Conference Record*, pp. 2383–2386, 2009.

[137] S. Seifert, H. T. Van Dam, R. Vinke, P. Dendooven, H. Löhner, F. J. Beekman, and D. R. Schaart, "A Comprehensive Model to Predict the Timing Resolution of SiPM-based Scintillation Detectors: Theory and Experimental Validation," *IEEE Transactions on Nuclear Science*, vol. 59, no. 1 PART 2, pp. 190–204, 2012.

[138] R. J. Walker, E. A. G. Webster, J. Li, N. Massari, and R. K. Henderson, "High Fill Factor Digital Silicon Photomultiplier Structures in 130nm CMOS Imaging Technology," *IEEE Nuclear Science Symposium Conference Record*, pp. 1945–1948, 2012.

[139] S. Gundacker, E. Auffray, P. Jarron, T. Meyer, and P. Lecoq, "On the comparison of analog and digital SiPM readout in terms of expected timing performance," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 787,

pp. 6–11, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.nima.2014.10.020$\
delimiter"026E30F$nhttp://linkinghub.elsevier.com/retrieve/pii/S016890021401167X

[140] N. Dutton, "CIRCUIT FOR COMBINING SIGNALS," 2014.

[141] B. Leskovar, C. C. Lo, P. R. Hartig, and K. Sauer, "Photon counting system for subnanosecond fluorescence lifetime measurements," *Review of Scientific Instruments*, vol. 47, no. 9, pp. 1113–1121, 1976.

[142] E. Raisanen-Ruotsalainen, T. Rahkonen, and J. Kostamovaara, "A high resolution time-to-digital converter based on time-to-voltage interpolation," *Proceedings of the 23rd European Solid-State Circuits Conference*, 1997.

[143] J. Kalisz, M. Pawlowski, and R. Pelka, "Error analysis and design of the Nutt time-interval digitiser with picosecond resolution," *Journal of Physics E: Scientific Instruments*, vol. 20, no. 11, p. 1330, 1987.

[144] S. Henzler, "Time-to-Digital Converters," *Media*, vol. 29, no. 9, pp. 1–3, 2010.

[145] J. Kalisz, "Review of methods for time interval measurements with picosecond resolution," *Metrologia*, vol. 41, no. 1, pp. 17–32, feb 2004.

[146] International Organization for Standardization, "Statistics – Vocabulary and symbols – Part 1: General statistical terms and terms used in probability."

[147] P. Chen, S.-i. Liu, and J. Wu, "A low power high accuracy cmos time-to-digital converter," *Circuits and Systems, 1997. ISCAS '97., Proceedings of 1997 IEEE International Symposium on*, vol. 1, pp. 281–284, 1997.

[148] Y. L. Y. Liu, U. Vollenbruch, Y. C. Y. Chen, C. Wicpalek, L. Maurer, T. Mayer, Z. Boos, and R. Weigel, "A 6ps resolution pulse shrinking Time-to-Digital Converter as phase detector in multi-mode transceiver," *2008 IEEE Radio and Wireless Symposium*, pp. 163–166, 2008.

[149] A. Mantyniemi, T. Rahkonen, and J. Kostamovaara, "A CMOS time-to-digital converter (TDC) based on a cyclic time domain successive approximation interpolation method," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 11, pp. 3067–3078, 2009.

[150] R. D. Barton and M. E. King, "Two Vernier Time-Iinterval Digitizers," *Nuclear Instruments and Methods*, vol. 97, pp. 359–370, 1971.

[151] C. T. Gray, W. Liu, S. Member, W. A. M. V. Noije, T. A. Hughes, and R. K. Cavin, "A sampling technique and its CMOS implementation with 1 Gb/s bandwidth and 25 ps resolution," *IEEE Journal of Solid-State Circuits*, vol. 29, no. 3, pp. 340–349, 1994.

[152] T. Rahkonen, J. Kostamovaara, and S. Saynajakangas, "Time interval measurements using integrated tapped CMOS delay lines," in *Circuits and Systems, 1989., Proceedings of the 32nd Midwest Symposium on*, vol. 1, no. 2, Champaign, IL., 1989, pp. 201–205.

[153] T. E. Rahkonen and J. T. Kostamovaara, "The Use of Stabilized CMOS Delay Lines for the Digitization of Short Time Intervals," *IEEE Journal of Solid-State Circuits*, vol. 28, no. 8, pp. 887–894, 1993.

[154] S. Henzler, S. Koeppe, W. Kamp, H. Mulatz, and D. Schmitt-Landsiedel, "90nm 4.7ps-Resolution 0.7-LSB single-shot precision and 19pJ-per-shot local passive interpolation time-to-digital converter with on-chip characterization," *Digest of Technical Papers - IEEE International Solid-State Circuits Conference*, vol. 51, pp. 548–550, 2008.

[155] L. Vercesi, a. Liscidini, and R. Castello, "Two-Dimensions Vernier Time-to-Digital Converter," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 8, pp. 1504–1512, 2010.

[156] N. U. Andersson and M. Vesterbacka, "A vernier time-to-digital converter with delay latch chain architecture," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 61, no. 10, pp. 773–777, 2014.

[157] N. Narku-tetteh, A. Titriku, and S. Palermo, "A 15b , Sub-10ps Resolution , Low Dead Time , Wide Range Two-Stage TDC," in *2014 IEEE 57th International Midwest Symposium on Circuits and Systems (MWSCAS)*, 2014, pp. 13–16.

[158] Z. Cheng, M. J. Deen, and H. Peng, "A Low-Power Gateable Vernier Ring Oscillator Time-to-Digital Converter for Biomedical Imaging Applications," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 10, no. 2, pp. 445–454, 2016.

[159] P. Lu, Y. Wu, and P. Andreani, "A 2.2ps 2-D Gated-Vernier Time-to-Digital Converter with Digital Calibration," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 7747, no. c, pp. 1–1, 2016.

[160] S. Tisa, A. Lotito, A. Giudice, and F. Zappa, "Monolithic time-to-digital converter with 20ps resolution," in *European Solid-State Circuits Conference*, 2003, pp. 465–468.

[161] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, "1.3 V 20 ps time-to-digital converter for frequency synthesis in 90-nm CMOS," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 53, no. 3, pp. 220–224, mar 2006.

[162] A. S. Yousif and J. W. Haslett, "A Fine Resolution TDC Architecture for Next Generation PET Imaging," *IEEE Transactions on Nuclear Science*, vol. 54, no. 5, pp. 1574–1582, oct 2007.

[163] M. Lee and A. A. Abidi, "A 9b, 1.25ps resolution coarse-fine time-to-digital converter in 90nm CMOS that amplifies a time residue," in *IEEE Symposium on VLSI Circuits, Digest of Technical Papers*, 2007, pp. 168–169.

[164] H. Menninga, C. Favi, M. W. Fishburn, and E. Charbon, "A multi-channel, 10ps resolution, FPGA-based TDC with 300MS/s throughput for open-source PET applications," *IEEE Nuclear Science Symposium Conference Record*, pp. 1515–1522, 2012.

[165] B. Markovic, S. Tisa, F. a. Villa, A. Tosi, and F. Zappa, "A High-Linearity, 17 ps Precision Time-to-Digital Converter Based on a Single-Stage Vernier Delay Loop Fine Interpolation," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 60, no. 3, pp. 557–569, 2013.

[166] M. Dandin, M. H. U. Habib, B. Nouri, P. Abshire, and N. McFarlane, "Characterization of Single-Photon Avalanche Diodes in a 0.5- μm Standard CMOS Process—Part 2: Equivalent Circuit Model and Geiger Mode Readout," *IEEE Sensors Journal*, vol. 16, no. 9, pp. 3075–3083, 2016.

[167] A. Lacaita, S. Cova, A. Spinelli, and F. Zappa, "Photon-assisted avalanche spreading in reach-through photodiodes," *Applied Physics Letters*, vol. 62, no. 6, pp. 606–608, 1993.

[168] S. Wang, F. Ma, X. Li, G. Karve, X. Zheng, and J. C. Campbell, "Analysis of breakdown probabilities in avalanche photodiodes using a history-dependent analytical model," *Applied Physics Letters*, vol. 82, no. 12, pp. 1971–1973, 2003.

[169] Y. Kang, H. X. Lu, Y. H. Lo, D. S. Bethune, and W. P. Risk, "Dark count probability and quantum efficiency of avalanche photodiodes for single-photon detection," *Applied Physics Letters*, vol. 83, no. 14, pp. 2955–2957, 2003.

[170] F. Corsi, A. Dragone, C. Marzocca, A. Del Guerra, P. Delizia, N. Dinu, C. Piemonte,

M. Boscardin, and G. F. Dalla Betta, "Modelling a silicon photomultiplier (SiPM) as a signal source for optimum front-end design," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 572, no. 1 SPEC. ISS., pp. 416–418, 2007.

[171] L. Pancheri, D. Stoppa, and G. F. Dalla Betta, "Characterization and Modeling of Breakdown Probability in Sub-Micrometer CMOS SPADs," *IEEE Journal on Selected Topics in Quantum Electronics*, vol. 20, no. 6, 2014.

[172] A. Dalla Mora, A. Tosi, S. Tisa, and F. Zappa, "Single-Photon Avalanche Diode Model for Circuit Simulations," *IEEE Photonics Technology Letters*, vol. 19, no. 23, pp. 1922–1924, 2007.

[173] R. Mita, G. Palumbo, and P. Fallica, "Accurate model for single-photon avalanche diodes," in *Circuits, Devices & Systems, IET*, vol. 2, no. 2, 2008, pp. 207–212.

[174] F. Zappa, A. Tosi, A. Dalla Mora, and S. Tisa, "SPICE modeling of single photon avalanche diodes," *Sensors and Actuators, A: Physical*, vol. 153, no. 2, pp. 197–204, 2009.

[175] V. Savuskan, M. Javitt, G. Visokolov, I. Brouk, and Y. Nemirovsky, "Selecting single photon avalanche diode (SPAD) passive-quenching resistance: An approach," *IEEE Sensors Journal*, vol. 13, no. 6, pp. 2322–2328, 2013.

[176] G. Giustolisi, R. Mita, and G. Palumbo, "Behavioral modeling of statistical phenomena of single-photon avalanche diodes," *Int. J. Circ. Theor. Appl.*, vol. 40, no. January 2011, pp. 661–679, 2012.

[177] M. M. García, Ó. G. Vinuesa, R. d. R. Fernández, B. P. Verdú, and Á. R. Vázquez, "CMOS SPADs selection, modeling and characterization towards image sensors implementation," in *Electronics, Circuits and Systems (ICECS)*, Seville, 2012, pp. 332–335.

[178] S. Chick, R. Coath, R. Sellahewa, R. Turchetta, T. Leitner, and A. Fenigstein, "Dead Time Compensation in CMOS Single Photon Avalanche Diodes with Active Quenching and External Reset," *IEEE Transactions on Electron Devices*, vol. 61, no. 8, pp. 2725–2731, 2014.

[179] Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen, and H. Peng, "A Comprehensive

and Accurate Analytical SPAD Model for Circuit Simulation," *IEEE Transactions on Electron Devices*, vol. 63, no. 5, pp. 1940–1948, 2016.

[180] H. T. V. Dam, S. Seifert, R. Vinke, D. Dendooven, H. Lohner, F. J. Beekman, and D. R. Schaart, "A Comprehensive Model of the Response of Silicon Photomultipliers," *IEEE Transactions on Nuclear Science*, vol. 57, no. 4, pp. 2254–2266, 2010.

[181] J. Pulko, F. R. Schneider, A. Velroyen, D. Renker, and S. I. Ziegler, "A Monte-Carlo model of a SiPM coupled to a scintillating crystal," *Journal of Instrumentation*, vol. 7, no. 02, pp. P02 009–P02 009, 2012.

[182] A. C. Therrien, B.-L. Berube, S. A. Charlebois, R. Lecomte, R. Fontaine, and J.-F. Pratte, "Modeling of Single Photon Avalanche Diode Array Detectors for PET Applications," *IEEE Transactions on Nuclear Science*, vol. 61, no. 1, pp. 14–22, 2014.

[183] Y. Du and F. Retière, "After-pulsing and cross-talk in multi-pixel photon counters," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 596, no. 3, pp. 396–401, 2008.

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

[185] E. A. G. Webster, J. A. Richardson, L. A. Grant, D. Renshaw, and R. K. Henderson, "An Infra-Red Sensitive , Low Noise , Single- Photon Avalanche Diode in 90nm CMOS," *International Image Sensor Workshop (IISW)*, pp. P17–P20, 2011.

[186] 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, jul 2011.

Appendix A

Modelling Scripts

A.1 Photo Detection - Active Recharged SPAD cell

```
function [ output ] = SPADsDetectAQ(photons,trange,PDE,dead_time)
ve = 4;
vt = 2;
v0=1;

t_end = trange(length(trange));
step = trange(5)-trange(4);

L = length(trange);
time_fire=-t_end;

output = zeros(1,L);
n=1;
for time = trange;
    deltat = time - time_fire;
    if (deltat < dead_time)
        vs = 0;
    else
        vs = ve;
    end
    if photons(n)==1
        if (rand >= (1-PDE) == 1)
            time_fire = time;
        end
```

```
    end
    if vs > vt
        output(n) = 0;
    else
        output(n) = 1;
    end
    n=n+1;
end
end
```

A.2 Photo Detection - Passive Recharged SPAD cell

```
function [output] = SPADDetectPQB( photons,trange,PDE0,dead_time,jitter)

%% SETUP
ve = 4;
vt = 2;
v0=1;
RC = -dead_time/log(1-vt/ve);

t_end = trange(length(trange));
step = trange(5)-trange(4);

L = length(trange);
time_fire= -t_end;

output = zeros(1,L);

n=1;

%% JITTER SPAD can be simulated by adding jitter on the received photons

photons_spad = zeros(1,length(photons));

photons_stamps = find(photons>0);
```

```
for index=1:length(photons_stamps)
    current = photons_stamps(index);
    current_spad = current + round((randn * jitter)/step);
    if (current_spad>1 && current_spad<L)
        photons_spad(current_spad) = photons(photons_stamps(index));
    end
end

%% POINT-WISE DETECTION

for time = trange;
    % lucky SPAD
    deltat = time - time_fire;
    vs = ve*(1-exp(-deltat/RC));
    if photons_spad(n)>0
        for ki=1:photons_spad(n)
            % PHOTON IS DETECTED
            PDE = PDE0*(1-exp(-vs/v0));
            if (rand >= (1-PDE) == 1)
                time_fire = time;
            end
        end
    end
    if vs > vt
        output(n) = 0;
    else
        output(n) = 1;
    end
    n=n+1;
end

end
```

Appendix B

Experiments

B.1 Number of Cells Analysis

```
%%
clc;
SensorStop;
SensorStart;
tic;
% SETUP
VHV_DEEPSPAD = 3.3;
VHV_SHALLOWSPAD = 14.2;
VQUENCH = 0.7;
VSPADOFF = 1.2;
VEB = 1;
V1V2 = 1.2;
V3V3 = 3.3;
%-----------------
% START-UP
% -----------------
s.SetVoltage('VHV_DEEPSPAD',VHV_DEEPSPAD);
s.SetVoltage('VHV_SHALLOWSPAD',VHV_SHALLOWSPAD);
s.SetVoltage('VQUENCH',VQUENCH);
VBIAS = 0;
s.SetVoltage('BIAS',VBIAS);
s.SetVoltage('VSPADOFF',VSPADOFF);
s.SetVoltage('VEB',VEB);
s.SetVoltage('V1V2',V1V2);
```

```
s.SetVoltage('V3V3',V3V3);
wireindata(s.okComms,s.bank,'CHIP_RESET',0);
wireindata(s.okComms,s.bank,'CHIP_RESET',1);
wireindata(s.okComms,s.bank,'CHIP_RESET',0);
%%
wireindata(s.okComms,s.bank,'INPUT_COUNTER_INVERT',0);
% ----------------------------
% --- SERIAL INTERFACE ---
% ----------------------------
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUXB',0);
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUXB_SEL',0);
wireindata(s.okComms,s.bank,'SPI_COUNTER_MUXB',0);
wireindata(s.okComms,s.bank,'SPI_ENABLE_DEEP',0);
wireindata(s.okComms,s.bank,'SPI_ENABLE_SHALLOW',0);
wireindata(s.okComms,s.bank,'SPI_DELAY_GEN_SEL',0);
wireindata(s.okComms,s.bank,'SPI_COUNT_OUT_EN',1);
wireindata(s.okComms,s.bank,'SPI_SEL_DG1',0);
wireindata(s.okComms,s.bank,'SPI_SEL_DG2',0);
wireindata(s.okComms,s.bank,'SPI_SEL_DG3',0);
wireindata(s.okComms,s.bank,'SPI_TEST_INPUT_DELAYGEN',0);
wireindata(s.okComms,s.bank,'SPI_TEST_OUTPUT_DELAYGEN',0);
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUX',0);
wireindata(s.okComms,s.bank,'SPI_DELAYGEN_OUT_EN',0);
wireindata(s.okComms,s.bank,'SPI_SERIAL_DEBUG_OUT',0);
%%
% ----------------------------
% --- ALL FAST SPADs DISABLED ---
% ----------------------------
wireindata(s.okComms,s.bank,'SPI_FAST_ROW',0);
wireindata(s.okComms,s.bank,'SPI_FAST_COLUMN',0);
trigger(s.okComms,s.bank,'SPI_FAST_TRIGGER');
trigger(s.okComms,s.bank,'PROG_CTRL_SR');
Niter = 500;
Light_Steps = 100;
crange = 6*logspace(-9,-1,Light_Steps);
Light_Min = min(crange);
```

```
Light_Max = max(crange);
actual_current = zeros(1,Light_Steps);
column = 1;
wireindata(s.okComms,s.bank,'SPI_SEL_ARRAY',1);
counts1 = zeros(9,Light_Steps);
counts2 = zeros(9,Light_Steps);
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUXB',0);
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUXB_SEL',0);
wireindata(s.okComms,s.bank,'SPI_COUNTER_MUXB',0);
trigger(s.okComms,s.bank,'PROG_CTRL_SR');
pause on;
VQUENCH_RANGE = 0.6;
s.SetVoltage('VQUENCH',VQUENCH);
results = struct('DCR',zeros(1,9),'SigmaDCR',zeros(1,9),'VQUENCH',0,'Counts'...
        ,zeros(9,Light_Steps),'FPGACounts',zeros(9,Light_Steps),'Sigma'...
        ,zeros(9,Light_Steps),'Sigma2',zeros(9,Light_Steps));
for kk = 1:length(VQUENCH_RANGE);
    s.SetVoltage('VQUENCH',VQUENCH_RANGE(kk));
    results(kk).VQUENCH = VQUENCH_RANGE(kk);
    for index = 1:9
        Nspads(index) = 2^(index-1);
        if Nspads(index)<16
            spad_y = 1;
            SPADs = 2^Nspads(index)-1;
        else
            SPADs = 65535;
            spad_y = round(Nspads(index)/16);
        end
        wireindata(s.okComms,s.bank,'SPI_FAST_ROW',0);
        wireindata(s.okComms,s.bank,'SPI_FAST_COLUMN',0);
        trigger(s.okComms,s.bank,'SPI_FAST_TRIGGER');
        trigger(s.okComms,s.bank,'PROG_CTRL_SR');
        for col = 1:spad_y-1
            wireindata(s.okComms,s.bank,'SPI_FAST_ROW',65535);
            wireindata(s.okComms,s.bank,'SPI_FAST_COLUMN',col);
            trigger(s.okComms,s.bank,'SPI_FAST_TRIGGER');
```

```
            trigger(s.okComms,s.bank,'PROG_CTRL_SR');
        end
        wireindata(s.okComms,s.bank,'SPI_FAST_ROW',bin2dec(fliplr(dec2bin(SPADs,16))));
        wireindata(s.okComms,s.bank,'SPI_FAST_COLUMN',spad_y);
        trigger(s.okComms,s.bank,'SPI_FAST_TRIGGER');
        trigger(s.okComms,s.bank,'PROG_CTRL_SR');
        power_meter = PowerSupplySET;
        for l=1:Light_Steps
            actual_current(l) = SetCurrent(power_meter,crange(l));
            if crange(l)*Nspads(index) < 1e-3
                Exposure_Time = 6e-4;
            else
                Exposure_Time = 3e-6;
            end
            cc1 = zeros(1,Niter);
            cc2 = zeros(1,Niter);
            s.SetExposureTime(Exposure_Time*1e6);
            for i=1:Niter
                trigger(s.okComms,s.bank,'EXPOSURE_START');
                pause(Exposure_Time);
                trigger(s.okComms,s.bank,'READ_COUNTER_SR');
                cc1(i) = wireoutdata(s.okComms,s.bank,'RC_OUT');
                cc2(i) = wireoutdata(s.okComms,s.bank,'COUNTERS');
            end
            counts1(index,l) = mean(cc1./Exposure_Time);
            sigma1(index,l) = std(cc1./Exposure_Time);
            counts2(index,l) = mean(cc2./Exposure_Time);
            sigma2(index,l) = std(cc2./Exposure_Time);
        end
        PowerSupplyOFF(power_meter);
        results(kk).Counts(index,:) = counts1(index,:);
        results(kk).Sigma(index,:) = sigma1(index,:);
        results(kk).FPGACounts(index,:) = counts2(index,:);
        results(kk).FPGASigma(index,:) = sigma2(index,:);
    end
end
```

```
experiment_runtime = toc;
day_of_experiment = datestr(now);
day_of_experiment=strrep(day_of_experiment,':','_'); %Replace colon with underscore
day_of_experiment=strrep(day_of_experiment,'-','_');%Replace minus sign with underscore
day_of_experiment=strrep(day_of_experiment,' ','_');%Replace space with underscore
filename = ['\NumSPADsComp\' day_of_experiment];
save(filename);
```

B.2 Combining Technique Analysis

```
%%
clc;
SensorStop;
SensorStart;
tic;
% SETUP
VHV_DEEPSPAD = 3.3;
VHV_SHALLOWSPAD = 14.2;
VQUENCH = 0.7;
VSPADOFF = 1.2;
VEB = 1;
V1V2 = 1.2;
V3V3 = 3.3;
%-----------------
% START-UP
% -----------------
s.SetVoltage('VHV_DEEPSPAD',VHV_DEEPSPAD);
s.SetVoltage('VHV_SHALLOWSPAD',VHV_SHALLOWSPAD);
s.SetVoltage('VQUENCH',VQUENCH);
s.SetVoltage('VSPADOFF',VSPADOFF);
s.SetVoltage('VEB',VEB);
s.SetVoltage('V1V2',V1V2);
s.SetVoltage('V3V3',V3V3);
wireindata(s.okComms,s.bank,'CHIP_RESET',0);
wireindata(s.okComms,s.bank,'CHIP_RESET',1);
```

```
wireindata(s.okComms,s.bank,'CHIP_RESET',0);
%%
wireindata(s.okComms,s.bank,'INPUT_COUNTER_INVERT',0);
% ----------------------------
% --- SERIAL INTERFACE ---
% ----------------------------
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUXB',0);
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUXB_SEL',0);
wireindata(s.okComms,s.bank,'SPI_COUNTER_MUXB',0);
wireindata(s.okComms,s.bank,'SPI_ENABLE_DEEP',0);
wireindata(s.okComms,s.bank,'SPI_ENABLE_SHALLOW',1);
wireindata(s.okComms,s.bank,'SPI_DELAY_GEN_SEL',0);
wireindata(s.okComms,s.bank,'SPI_COUNT_OUT_EN',1);
wireindata(s.okComms,s.bank,'SPI_SEL_DG1',0);
wireindata(s.okComms,s.bank,'SPI_SEL_DG2',0);
wireindata(s.okComms,s.bank,'SPI_SEL_DG3',0);
wireindata(s.okComms,s.bank,'SPI_TEST_INPUT_DELAYGEN',0);
wireindata(s.okComms,s.bank,'SPI_TEST_OUTPUT_DELAYGEN',0);
wireindata(s.okComms,s.bank,'SPI_OUTPUT_MUX',0);
wireindata(s.okComms,s.bank,'SPI_DELAYGEN_OUT_EN',0);
wireindata(s.okComms,s.bank,'SPI_SERIAL_DEBUG_OUT',0);
%%
% ------------------------------
% --- ALL FAST SPADs DISABLED ---
% ------------------------------
wireindata(s.okComms,s.bank,'SPI_FAST_ROW',0);
wireindata(s.okComms,s.bank,'SPI_FAST_COLUMN',0);
trigger(s.okComms,s.bank,'SPI_FAST_TRIGGER');
trigger(s.okComms,s.bank,'PROG_CTRL_SR');
Niter = 100;
Light_Steps = 100;
crange = 6*logspace(-7,-1,Light_Steps);
Light_Min = min(crange);
Light_Max = max(crange);
power_meter = PowerSupplySET;
actual_current = zeros(1,Light_Steps);
```

```
column = 1;
which_array = 0;
% 0 = ENABLE_SPAD_A
% 1 = ENABLE_SPAD_B
if (which_array == 0 || which_array == 1)
    wireindata(s.okComms,s.bank,'SPI_SEL_ARRAY',which_array);
else
    disp('Error: choose array A-0 or B-1');
end
wireindata(s.okComms,s.bank,'SPI_SEL_SIZE',1);
BIAS_POINTS = 10;
counts1 = zeros(BIAS_POINTS+1,Light_Steps);
counts2 = zeros(BIAS_POINTS+1,Light_Steps);
VBIAS_RANGE = linspace(0.35,1.5,BIAS_POINTS);
results = struct('Counts',zeros(BIAS_POINTS+1,Light_Steps),'FPGACounts'...
            ,zeros(BIAS_POINTS+1,Light_Steps),'Sigma',zeros(BIAS_POINTS+1,Light_Steps)...
            ,'Sigma2',zeros(BIAS_POINTS+1,Light_Steps));
pause on;
power_meter = PowerSupplySET;
Nrange = [1 2 4 6 8 10 12 14 16];
for Nspads = Nrange;
    SPADs = 2.^Nspads -1;
    wireindata(s.okComms,s.bank,'SPI_ENABLE_SPAD_A',bin2dec(fliplr(dec2bin(SPADs,16))));
    wireindata(s.okComms,s.bank,'SPI_ENABLE_SPAD_B',bin2dec(fliplr(dec2bin(SPADs,16))));
    trigger(s.okComms,s.bank,'PROG_CTRL_SR');
    trigger(s.okComms,s.bank,'PROG_CTRL_SR');
    trigger(s.okComms,s.bank,'PROG_CTRL_SR');
    for l=1:Light_Steps
        actual_current(l) = SetCurrent(power_meter,crange(l));
        if crange(l)*Nspads < 1e-3
            Exposure_Time = 6e-5;
        else
            Exposure_Time = 3e-6;
        end
        s.SetExposureTime(Exposure_Time*1e6);
        s.SetExposureTime(Exposure_Time*1e6);
```

```
s.SetExposureTime(Exposure_Time*1e6);
for index = 1:BIAS_POINTS+1;
    if index<BIAS_POINTS+1
        which_logic = 0; UNCOMMENT THIS
        %which_logic = 1; % COMMENT THIS
        VBIAS = VBIAS_RANGE(index);
    else
        which_logic = 1;
    end
    s.SetVoltage('BIAS',VBIAS);
    if which_logic == 0 || which_logic == 1
        %--- 0 = OR TREE, No Toggle
        %--- 1 = XOR TREE, Toggle
        wireindata(s.okComms,s.bank,'SPI_NOTOGGLE_SEL',which_logic);
        wireindata(s.okComms,s.bank,'SPI_XOR_OR_SEL',which_logic);
        wireindata(s.okComms,s.bank,'SPI_COUNTER_MUXB',1+2*which_logic);
        wireindata(s.okComms,s.bank,'SPI_COUNTER_MUX',3-2*which_logic);
        trigger(s.okComms,s.bank,'PROG_CTRL_SR');
        %disp('OR TREE');
    else disp('Error');
    end
    cc1 = zeros(1,Niter);
    cc2 = zeros(1,Niter);
    for pp=1:Niter
        trigger(s.okComms,s.bank,'EXPOSURE_START');
        trigger(s.okComms,s.bank,'EXPOSURE_START');
        trigger(s.okComms,s.bank,'EXPOSURE_START');
        trigger(s.okComms,s.bank,'READ_COUNTER_SR');
        cc1(pp) = wireoutdata(s.okComms,s.bank,'RC_OUT');
        cc2(pp) = wireoutdata(s.okComms,s.bank,'COUNTERS');
    end
    counts1(index,l) = mean(cc1./Exposure_Time);
    sigma1(index,l) = std(cc1./Exposure_Time);
    counts2(index,l) = mean(cc2./Exposure_Time);
    sigma2(index,l) = std(cc2./Exposure_Time);
end
```

```
        results(Nspads).Counts(:,l) = counts1(:,l);
        results(Nspads).FPGACounts(:,l) = counts2(:,l);
        results(Nspads).Sigma(:,l) = sigma1(:,l);
        results(Nspads).Sigma2(:,l) = sigma2(:,l);
    end
end
PowerSupplyOFF(power_meter);
experiment_runtime = toc;
day_of_experiment = datestr(now);
day_of_experiment=strrep(day_of_experiment,':','_'); %Replace colon with underscore
day_of_experiment=strrep(day_of_experiment,'-','_');%Replace minus sign with underscore
day_of_experiment=strrep(day_of_experiment,' ','_');%Replace space with underscore
filename = ['\LightSweep\LogicComp\' day_of_experiment];
PowerSupplyOFF(power_meter);
save(filename);
```

Publications