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.

Miniature High Dynamic Range Time-Resolved CMOS SPAD Image Sensors

Tarek Al Abbas

THE UNIVERSITY
of EDINBURGH

A thesis submitted for the degree of Doctor of Philosophy

THE UNIVERSITY *of* EDINBURGH

2019

To Souha and Ziad,

Tameem and Hani.

With love.

If you think you are beaten, you are.
If you think you dare not, you don't.
If you'd like to win, but think you can't,
It is almost certain you won't.

If you think you'll lose, you're lost.
For out in the world we find
Success begins with a fellow's will;
It's all in the state of mind.

If you think you're outclassed, you are.
You've got to think high to rise,
You've got to be sure of yourself before
You can ever win a prize.

Life's battles, don't always go
To the stronger or faster man;
But soon or late the man who wins
Is the man who thinks he can!

Walter D. Wintle

Supervised by:

Prof. Robert Henderson

School of Engineering,
The University of Edinburgh,
Scottish Microelectronics Centre,
King's Buildings,
Alexander Crum Brown Road,
Edinburgh,
UK

Prof. Ian Underwood

School of Engineering,
The University of Edinburgh,
Scottish Microelectronics Centre,
King's Buildings,
Alexander Crum Brown Road,
Edinburgh,
UK

Prof. Kev Dhaliwal

Edinburgh Medical School,
The University of Edinburgh,
The Queen's Medical Research Institute,
Little France Crescent,
Edinburgh,
UK

Examined by:

Mr. Matteo Perenzoni

Fondazione Bruno Kessler,
Centre for Materials and Microsystems,
Via Sommarive,
Trento,
Italy

Dr. Srinjoy Mitra

School of Engineering,
The University of Edinburgh,
Scottish Microelectronics Centre,
King's Buildings,
Alexander Crum Brown Road,
Edinburgh,
UK

On the 13th of December 2018

Abstract

Since their integration in complementary metal oxide (CMOS) semiconductor technology in 2003, single photon avalanche diodes (SPADs) have inspired a new era of low cost high integration quantum-level image sensors. Their unique feature of discerning single photon detections, their ability to retain temporal information on every collected photon and their amenability to high speed image sensor architectures makes them prime candidates for low light and time-resolved applications.

From the biomedical field of fluorescence lifetime imaging microscopy (FLIM) to extreme physical phenomena such as quantum entanglement, all the way to time of flight (ToF) consumer applications such as gesture recognition and more recently automotive light detection and ranging (LIDAR), huge steps in detector and sensor architectures have been made to address the design challenges of pixel sensitivity and functionality trade-off, scalability and handling of large data rates.

The goal of this research is to explore the hypothesis that given the state of the art CMOS nodes and fabrication technologies, it is possible to design miniature SPAD image sensors for time-resolved applications with a small pixel pitch while maintaining both sensitivity and built-in functionality. Three key approaches are pursued to that purpose: leveraging the innate area reduction of logic gates and finer design rules of advanced CMOS nodes to balance the pixel's fill factor and processing capability, smarter pixel designs with configurable functionality and novel system architectures that lift the processing burden off the pixel array and mediate data flow.

Two pathfinder SPAD image sensors were designed and fabricated: a 96 × 40 planar front side illuminated (FSI) sensor with 66% fill factor at 8.25μm pixel pitch in an industrialised 40nm process and a 128 × 120 3D-stacked backside illuminated (BSI) sensor with 45% fill factor at 7.83μm pixel pitch. Both designs rely on a digital, configurable, 12-bit ripple counter pixel allowing for time-gated shot noise limited photon counting. The FSI sensor was operated as a quanta image sensor (QIS) achieving an extended dynamic range in excess of 100dB, utilising triple exposure windows and in-pixel data compression which reduces data rates by a factor of 3.75×. The stacked sensor is the first demonstration of a wafer scale SPAD imaging array with a 1-to-1 hybrid bond connection. Characterisation results of the detector and sensor performance are presented.

Two other time-resolved 3D-stacked BSI SPAD image sensor architectures are proposed. The first is a fully integrated 5-wire interface system on chip (SoC), with built-in power management and off-focal plane data processing and storage for high dynamic range as well as autonomous video rate operation. Preliminary images and bring-up results of the fabricated 2mm² sensor are shown. The second is a highly configurable design capable of simultaneous multi-bit oversampled imaging and programmable region of interest (ROI) time correlated single photon counting (TCSPC) with on-chip histogram generation. The 6.48μm pitch array has been submitted for fabrication. In-depth design details of both architectures are discussed.

Declaration of Originality

I hereby declare that the research recorded in this thesis (excluding the exceptions stated below) and the thesis itself originated with and was composed by myself.

i. The bandgap, voltage regulator and power-on reset analogue blocks used in the sensor (ENDOCAM) described in Chapter 6 are STMicroelectronics IP blocks and were integrated within the design by myself.
ii. The SRAM digital blocks used in the sensor (ENDOCAM) described in Chapter 6 are STMicroelectronics IP blocks and were integrated within the design by myself.
iii. The digital design of the synthesised micro-control unit and its interface to the SRAM blocks was fully carried out by Dr. Oscar Almer of The University of Edinburgh (now with Semtech Corp.) and integrated within the sensor (ENDOCAM) described in Chapter 6 by myself.
iv. The 1GHz phase locked loop analogue block used in the sensor (CORVETTE) described in Chapter 6 is STMicroelectronics' IP block and was integrated within the design by myself.

Other people who have contributed to this work are acknowledged and / or referenced appropriately within.

This work has not been submitted for any other degree or professional qualification except as specified.

Tarek Al Abbas

Acknowledgements

In August 2013 after finishing my MSc viva at the University of Edinburgh, I was asked if I would like to pursue a PhD. At the time my immediate answer was no as I did not believe I had the knowledge and grit that it takes. Fast forward five years and here I am submitting a thesis for a doctorate degree.

There are many people that I would like acknowledge throughout this journey but there is one person to whom I will always be grateful, a person who saw in me what I did not see in myself. Professor Robert Henderson, thank you for being a tutor, a supervisor and a constant source of support. Your unbridled enthusiasm for new ideas and good engineering is an inspiring trait that not many possess. A lot of the engineer I am, and the engineer I hope to become I owe to you.

Circumstances had it that I spent the four years of this research away from my family; nevertheless both the city and University of Edinburgh, always felt like home. I would like to express my gratitude to all the University students, lectures and staff that helped me throughout this period including my second supervisor Professor Ian Underwood and my colleagues at the Institute for Integrated Micro and Nano Systems. To me Edinburgh will always be a place where dreams come true and I am immensely proud of being a University of Edinburgh graduate.

I am very grateful to have received funding for this research from the EPSRC multi-disciplinary Proteus project. Despite spending most of my time in the realms of engineering away from all the scintillating science produced at the Queens Medical Research Institute (Proteus hub), it was a pleasure being part of the Proteus family and seeing the great progress made over the years. I am particularly thankful to Dr. Anne Moore and Professor Mark Bradley for their trust in what I do and for giving me the absolute freedom to explore and play beyond the objectives of Proteus which accounted for a most enjoyable research experience.

As part of the collaboration between the University and STMicroelectronics imaging division in Edinburgh, I have had access to state of the art CMOS processes from an advanced 40nm node to 3D-stacking, a privilege that many engineers dream of. I would like to acknowledge everyone at ST who gave me a hand during the dozen or so tapeouts that I have been involved in especially Dave Poyner. I am also grateful to Brent Hearn for his characterisation support and fruitful discussions.

It goes without saying that none of these tapeouts would have been possible without the managerial support of Drs. Sara Pellegrini and Bruce Rae. Before starting my PhD I spent a year as an intern at ST under Sara's mentorship, an experience that I cherished and hugely benefited from. At the time, Sara once told me that if you do not ask, you do not get. I bet she regrets doing so now! While Bruce might have had a more logic-driven reasoning behind designing a new chip as opposed to my "just

because we can" excitement-driven one, he always tried to accommodate my ambitious demands for silicon. Thank you both very much.

Two very bright engineers also had a huge impact on the work I have done over the past few years, Drs. Neale Dutton and Oscar Almer. Neale, there is nothing I enjoyed more than passionately discussing ideas with you and I will always be in your debt for all the technical and motivational support. Your friendship is something I hold dearly and it is a joy seeing you and Helena start a new period of your life with baby Fraser. Oscar, your digital wizardry is beyond me and working with you on the ENDOCAM sensor was an absolute pleasure. Getting live stream images 20 minutes after plugging the sensor in for the first time is one the most ecstatic moments of my PhD. Thank you for all your efforts.

I firmly believe that no one makes it on his / her own without the right people around and I have been blessed with having superb colleagues at the CMOS Sensors and Systems group. Dr. Neil Finlayson, I truly admire your patience and your versatile knowledge, thank you for all the long days we spent working on the FlashTDC chip. Dr. Istvan Gyongy you are one of the nicest people I ever met, thank you for being a dear friend and for all the thrilling technical discussions we had. I wish you all the best with your fellowship and I have no doubt you are destined for great achievements.

I cannot say less for Dr. Danial Chitnis. When we first met at UCL in May 2016 I thought to myself that he belongs in Edinburgh. It is incredible how much work we have done together over the past few months but I hope next time we watch the World Cup games we do not have to do it in a dark room through laser safety goggles! My best wishes to you in your academic career, this group will certainly benefit from your expertise.

Drs. Ahmet Erdogan, Aravind Venugopalan and Nick Johnston, PhD colleagues Andras Kufcsak, Andrea Usai, John Kosman and Sarrah Patanwala and ex-group members Drs. Salvatore Gnecchi, Luca Parmesan, Laurence Stark, Nik Krstajic and Richard Walker, I have really enjoyed your company and we have shared a lot of great moments together. I wish you all the best to come, whatever direction you follow. Neil Calder, in July 2017 we had an intense week of tapeout hell for what I call "The Manhattan Project", funny enough this might not be the last time and I am really looking forward to our new career adventure.

As for Francesco Della Rocca and Hanning Mai, I am glad that despite the pressure of work we found the time to become really good friends. Francesco, we have conquered Munros together, had a lot of sushi and did a fair bit of climbing. Thanks for your continuous motivation, support, for all the gossip coffee breaks and fun activities. The all-nighter we pulled back in April 2018 for your chip tapeout will always be a great memory of team work. I cannot wait to see you present it at a top conference. Hanning, thank you for your solidarity over the past few months of writing and for all the football

games we watched and played together over the years. Now that you are writing up I am sure your thesis will be as solid as your defending skills.

Outside of work circles I have had some important people in my life who always kept me distracted, motivated me and were there for the high and low moments. Robert, Jess, Clara and Tom, thanks for being such good friends and for all the weekend breakfast outings. Waseem, thanks for all the deep thoughts, conversations about life and for always being there. Osama and Rikabi our friendship goes back to the days of undergrad and although we live in different parts of the world, it was like you were always here just as the old days; your support meant a lot to me.

Ahmad El Masri, we have been best friends for more than 20 years and fate had it that you would move to London 6 months into my PhD. That is the best thing I could have asked for. Thanks for always keeping me on track and for reminding me of my goal. Thanks for pulling me up when I am down and for taking hours to listen to my concerns. Of course I won't forget all the travelling, fun memories and the crazy nights out. I am very happy with what I have accomplished and even happier that I got to share it with you.

Last but not least I would like to thank my loving mother, caring father and my two amazing brothers for everything. It was hard enough enduring all of this away from you but your unwavering support and encouragement made it feel like a breeze. You always pushed me to follow my dream and to achieve what I desire in life. If I have succeeded then it is because of you. I hope I have done you proud.

Publications

- Chronological List of publications with primary contributions from the author:

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

[2] T. Al Abbas, N. A. W. Dutton, O. Almer, F. M. Della Rocca, S. Pellegrini, B. Rae, D. Golanski and R. K. Henderson, "8.25μm Pitch 66% Fill Factor Global Shared Well SPAD Image Sensor in 40nm CMOS FSI Technology," in *International Image Sensors Workshop*, 2017.

[3] T. Al Abbas, N. A. W. Dutton, O. Almer, N. Finlayson, F. M. D. Rocca and R. Henderson, "A CMOS SPAD Sensor With a Multi-Event Folded Flash Time-to-Digital Converter for Ultra-Fast Optical Transient Capture," in *IEEE Sensors Journal*, vol. 18, no. 8, pp. 3163-3173, 15 April15, 2018.

[4] N. Dutton, T. Al Abbas, I. Gyongy, F. Mattioli Della Rocca, and R. Henderson, "High Dynamic Range Imaging at the Quantum Limit with Single Photon Avalanche Diode-Based Image Sensors," *Sensors*, vol. 18, no. 4, p. 1166, Apr. 2018.

[*] T. Al Abbas, O. Almer, S. W. Hutchings, A. T. Erdogan, I. Gyongy, N. A. W. Dutton and R. K. Henderson, "A 128×120 5-Wire 1.96mm² 40nm/90nm 3D Stacked SPAD Time Resolved Image Sensor SoC for Microendoscopy," in *Symposium on VLSI Circuits*, 2019.

[*] T. Al Abbas, D. Chitnis and R. K. Henderson, "Dual Layer 3D-Stacked High Dynamic Range SPAD Pixel," in *International Image Sensors Workshop*, 2019.

*** The following papers were published / accepted for publication post viva examination and are mentioned here for completeness, therefore they are not referenced directly in the content of this work.**

- Chronological list of publications with secondary contributions from the author:

[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,” *2015 IEEE Summer Topicals Meeting Series (SUM)*, Nassau, 2015, pp. 41-42.

[6] O. Almer, D. Tsonev, N. A. W. Dutton, T. Al Abbas, S. Videv, S. Gnecchi, H. Haas and R. K. Henderson, “A SPAD-Based Visible Light Communications Receiver Employing Higher Order Modulation,” *2015 IEEE Global Communications Conference (GLOBECOM)*, San Diego, CA, 2015, pp. 1-6.

[7] Neil Finlayson, Tarek Al Abbas, Francescopaolo Mattioli Della Rocca, Oscar Almer, Salvatore Gnecchi, Neale A. W. Dutton, Robert K. Henderson, “Hypervelocity time-of-flight characterisation of a 14GS/s histogramming CMOS SPAD sensor,” *Proc. SPIE* 10111, Quantum Sensing and Nano Electronics and Photonics XIV, 101112Z (27 January 2017).

[8] I. Gyongy, T. Al Abbas, N. A. W. Dutton and R. K. Henderson, “Object Tracking and Reconstruction with a Quanta Image Sensor,” in *International Image Sensors Workshop*, 2017.

[9] N. A. W. Dutton, T. Al Abbas, I. Gyongy and R. K. Henderson, “Extending the Dynamic Range of Oversampled Binary SPAD Image Sensors,” in *International Image Sensors Workshop*, 2017.

[10] F. Mattioli Della Rocca, T. A. Abbas, N. A. W. Dutton and R. K. Henderson, “A high dynamic range SPAD pixel for time of flight imaging,” *2017 IEEE SENSORS*, Glasgow, 2017, pp. 1-3.

[*] I. Underwood, H. Mai, T. Al Abbas, I. Gyongy, N. A. W. Dutton and R. K. Henderson, “Invited Paper: Single-Photon-Capable Detector Arrays in CMOS – Exploring a New Tool for Display Metrology,” in *International Conference on Display Technology (ICDT)*, 2018.

[*] I. Gyongy, T. Al Abbas, N. Finlayson, N. Johnston, N. Calder, A. Erdogan, N. A. W. Dutton, R. Walker and R. K. Henderson, “Advances in CMOS SPAD Sensors for LIDAR Applications,” *Proc. SPIE 10799*, Emerging Imaging and Sensing Technologies for Security and Defence III; and Unmanned Sensors, Systems, and Countermeasures, 1079907 (4 October 2018).

[*] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. Al Abbas, N. A. W. Dutton, M. Tyler, S. Chan and J. Leach, “A 256×256 40nm/90nm CMOS 3D-Stacked 120dB Dynamic-Range Reconfigurable Time-Resolved SPAD Imager,” *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2019, pp. 106-108.

[*] J. Kosman, O. Almer, T. Al Abbas, N. A. W. Dutton, R. Walker, S. Videv, K. Moore, H. Haas and R. K. Henderson, “A 500Mb/s -46.1dBm CMOS SPAD Receiver for Laser Diode Visible-Light Communications,” *2019 IEEE International Solid- State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2019, pp. 468-470.

*** The following papers were published / accepted for publication post viva examination and are mentioned here for completeness, therefore they are not referenced directly in the content of this work**.

Acronyms

1-D	One Dimensional
2-D	Two Dimensional
3D-IC	Vertically Integrated Circuit
ADC	Analogue to Digital Converter
AMS	Austria Micro-Systems
APD	Avalanche Photodiode
APS	Active Pixel Sensor
AQAR	Active Quench Active Recharge
AQPR	Active Quench Passive Recharge
AR	Active Recharge
BCD	Bipolar-CMOS-DMOS
BEOL	Back End of the Line
BOX	Buried Oxide
BSI	Backside Illuminated
BTBT	Band to Band Tunnelling
CAPD	Current Assisted Photonic Demodulator
CCD	Charge Coupled Devices
CDS	Correlated Double Sampling
CIS	CMOS Image Sensor
CMOS	Complementary Metal Oxide Silicon
CSS	CMOS Sensors and Systems
CTA	Charge Transfer Amplifier
CW	Continuous Wave
DBI	Direct Bonding Interface
DCR	Dark Count Rate
DFF	D-Type Flip-Flop
DFS	Digital Film Sensor
DLL	Delay Locked Loop
DNW	Deep n-type Well
DOM	Drain Only Modulation
DR	Dynamic Range
DRAM	Dynamic Random Access Memory
dSiPM	Digital Silicon Photomultiplier
DSM	Deep Submicron Technology
DSP	Digital Signal Processing
DTI	Deep Trench Isolation
DToF	Direct Time of Flight
EPFL	Ecole Polytechnique Federale de Lausanne
EPSRC	Engineering and Physical Sciences Research Council
EQE	External Quantum Efficiency
ESD	Electro Static Discharge
FBK	Fondazione Bruno Kessler
FLIM	Fluorescence Lifetime Imaging Microscopy

FPGA	Field Programmable Gate Array
FPN	Fixed Pattern Noise
FRET	Fluorescence Resonance Energy Transfer
FSI	Front Side Illuminated
FW	Full Well
FWHM	Full Width Half Maximum
GM-APD	Geiger Mode Avalanche Photodiode
HB	Hybrid-Bonding
HDR	High Dynamic Range
HFPN	Horizontal Fixed Pattern Noise
IC	Integrated Circuit
ICCD	Intensified Charge Coupled Devices
IEDM	International Electron Devices Meeting
IISW	International Image Sensors Workshop
IO	Input / Output
IoT	Internet of Things
IP	Intellectual Property
IQE	Internal Quantum Efficiency
IRF	Impulse Response Function
ISP	Image Signal Processing
ISSCC	International Solid State Circuits Conference
IToF	Indirect Time of Flight
LED	Light Emitting Diode
LEFM	Lateral Electric Field Modulation
LIDAR	Light Detection and Ranging
LSB	Least Significant Bit
LVDS	Low Voltage Differential Signalling
MC	Monte Carlo
MCP	Micro-Channel Plate
MCU	Micro-Control Unit
MEMS	Micro-Electro Mechanical System
MIM	Metal-Insulator-Metal
MIT	Massachusetts Institute of Technology
MOM	Metal-Oxide-Metal
MPW	Multi-Project Wafer
MSB	Most Significant Bit
MTF	Modulation Transfer Function
MUX	Multiplexer
NIR	Near Infra-Red
NIROT	Near Infra-Red Optical Tomography
NW	n-type Well
OLED	Organic Light Emitting Diode
OVF	Overflow Flag
OVT	OmniVision Technologies
PCB	Printed Circuit Board
PD	Photo-Diode

PDAF	Phase Detection Auto Focus
PDE	Photon Detection Efficiency
PDP	Photon Detection Probability
PEB	Premature Edge Breakdown
PET	Positron Emission Tomography
PLL	Phase Locked Loop
PLS	Parasitic Light Sensitivity
PMD	Photo-Mixing Device
PMT	Photo-Multiplier Tube
POR	Power-On Reset
PPD	Pinned Photodiode
PPS	Passive Pixel Sensor
PQAR	Passive Quench Active Recharge
PQPR	Passive Quench Passive Recharge
PR	Passive Recharge
PRNU	Photo-Response Non-Uniformity
Psub	p-type Substrate
PTC	Photon Transfer Curve
PVT	Process-Voltage-Temperature
PW	p-type Well
QE	Quantum Efficiency
QEM	Quantum Efficiency Demodulator
QIS	Quanta Image Sensor
RLD	Rapid Lifetime Determination
RO	Ring Oscillator
ROI	Region of Interest
ROIC	Readout Integrated Circuit
RTL	Register Transfer Level
RTS	Random Telegraph Signal
sCMOS	Scientific Complementary Metal Oxide Silicon
SCS	Switched Current Source
SLID	Solid Liquid Inter Diffusion
SNR	Signal to Noise Ratio
SoC	System on Chip
SOI	Silicon on Insulator
SPAD	Single Photon Avalanche Diode
SPC	Single Photon Counting
SRAM	Static Random Access Memory
SRH	Shockley-Read-Hall
ST	STMicroelectronics
STI	Shallow Trench Isolation
SWD	Single Wire Debug
TAC	Time to Amplitude Converter
TAR	Temporal Aperture ratio
TCAD	Technology Computer Aided Design
TCSPC	Time Correlated Single Photon Counting

TDC	Time to Digital Converter
ToA	Time of Arrival
ToF	Time of Flight
TSMC	Taiwan Semiconductor Manufacturing Company
TSV	Through Silicon Via
TTS	Time to Saturation
VCSEL	Vertical Cavity Surface Emitting Laser
VLC	Visible Light Communication
WLCSP	Wafer Level Chip Scale Packaging

Contents

Abstract ... 5
Declaration of Originality ... 7
Acknowledgements ... 9
Publications ... 13
Acronyms ... 15
Figures ... 23
Tables ... 33
1. Introduction ... 35
1.1. Proteus Project ... 35
1.2. State of the Art Miniature Endoscopy Cameras ... 36
1.3. The Missing Element: Time-Resolved Capability ... 39
1.4. Time-Resolved Imaging ... 40
1.5. Solid State Time-Resolved Sensors ... 44
1.6. Research Aims ... 49
1.7. Contributions to Knowledge ... 51
1.8. Thesis Outline ... 52
2. CMOS SPAD Devices, Sensors and Technologies ... 53
2.1. The Single Photon Avalanche Diode ... 53
2.1.1. SPAD Operation Principles ... 54
2.1.2. CMOS SPAD Structures ... 57
2.1.3. Exotic SPAD Implementations ... 63
2.1.4. SPAD Performance Parameters ... 64
2.1.5. SPAD Devices State of the Art ... 66
2.1.6. Quench and Recharge Circuits ... 69
2.1.7. Biasing Topologies ... 71
2.2. SPAD Sensor Architectures ... 73
2.2.1. Single Point Sensors ... 73
2.2.2. Line Sensors ... 75
2.2.3. Image Sensors ... 76
2.2.4. SPAD Image Sensor Survey ... 82
2.2.5. Modular Sensors ... 84
2.3. 3D-Stacking Technology and SPADs ... 85
2.3.1. CIS 3D-Stacking Techniques ... 86
2.3.2. Overview of 3D-Stacked SPAD Sensors ... 89
2.4. Summary and Conclusions ... 94
3. High Fill Factor Global Shared Well FSI SPAD Image Sensor ... 95

3.1. SPAD Pixel Layout and Modelling 95
3.1.1. Different Layout Styles 96
3.1.2. SPAD Guard Ring Design Rules 98
3.1.3. Fill Factor and Pixel Pitch Trade-off Modelling 99
3.1.4. Global Well Sharing Feasibility Analysis 104
3.2. 96 × 40 Image Sensor in 40nm 110
3.2.1. System Overview 111
3.2.2. Layout Overview 112
3.2.3. SPAD Trials 118
3.2.4. Pixel Circuit Design 119
3.2.5. Pixel Front End Design 121
3.2.6. Pixel Counter Design 126
3.2.7. Time Gating Logic 129
3.2.8. Electrical Crosstalk on Anode Lines 132
3.3. SPAD and Sensor Characterisation 134
3.3.1. Photon Detection Probability 134
3.3.2. Dark Count Rate 135
3.3.3. Photo Response Non-Uniformity 137
3.3.4. SPAD Array Current Consumption 140
3.3.5. Global Enable (Shutter) Signal Propagation 142
3.3.6. Time Gate Profile Characterisation 143
3.3.7. Time Gate Array Uniformity 144
3.3.8. Time Gate Window Handover 145
3.4. Comparison to State of the Art Sensors 147
3.4.1. Pixel Pitch, Fill Factor and Dynamic Range Trade-offs 147
3.4.2. Time Gate Performance 150
3.5. Summary and Conclusion 150
4. High Dynamic Range Oversampled Binary Image Sensor 153
4.1. Technology Overview 153
4.1.1. The Quanta Image Sensor 153
4.1.2. High Dynamic Range Imaging 156
4.1.3. High Dynamic SPAD Image Sensors 162
4.2. High Dynamic Range SPAD Quanta Image Sensor 164
4.2.1. Challenges of Multi-Exposure QIS 164
4.2.2. Analysis and Measurement Results 167
4.3. HDR QIS and Miniaturisation Discussion 181
4.4. Summary and Conclusions 182
5. A 3D-Stacked SPAD Image Sensor 183

5.1. Chip Overview ... 183
5.1.1. System Architecture ... 183
5.1.2. Pixel Circuitry ... 184
5.1.3. Pixel Layout ... 186
5.1.4. SPAD Trials ... 187
5.2. Characterisation Results ... 190
5.2.1. Breakdown Voltage ... 190
5.2.2. Dark Count Rate ... 191
5.2.3. SPAD Jitter ... 192
5.2.4. Photo-Response Non-Uniformity ... 193
5.2.5. Photon Detection Probability ... 195
5.2.6. Shot Noise Limited Photon Counting ... 197
5.2.7. Intensity Imaging ... 197
5.2.8. Time Gate Profile ... 198
5.2.9. Time-Resolved Imaging ... 200
5.3. Comparison to Other Sensors ... 202
5.3.1. PDP Comparison to FSI SPAD ... 202
5.3.2. Other CMOS BSI 3D-Stacked Sensors ... 204
5.4. Summary and Conclusions ... 206
6. Miniature High Dynamic Range Time-Resolved Sensors ... 209
6.1. System on Chip Design ... 209
6.1.1. Pixel Design ... 211
6.1.2. Array Readout ... 215
6.1.3. Micro-Control Unit and SRAM Memory ... 216
6.1.4. Dynamic Range Enhancement ... 219
6.1.5. Single Pad SoC Data Rate Modelling ... 220
6.1.6. Power Network, Testability and VCSEL Driver ... 228
6.1.7. Gating Logic ... 233
6.1.8. Layout Overview ... 237
6.1.9. Preliminary Bring-up Results ... 240
6.2. Novel Configurable Array Architecture ... 245
6.2.1. Pixel in SPAD Mode ... 245
6.2.2. Pixel in Photodiode Mode ... 247
6.2.3. Region of Interest Operation ... 249
6.2.4. TDC Architecture ... 252
6.2.5. System Level Overview ... 253
6.2.6. Potential Applications ... 255
6.3. Comparison to State of the Art and Discussion ... 256

6.3.1. ENDOCAM System on Chip ... 257
6.3.2. CORVETTE Configurable Architecture ... 261
6.4. Summary and Conclusions ... 261
7. Summary and Conclusions, Future Work and Outlook ... 263
7.1. Thesis Summary and Conclusions ... 263
7.2. Future Work ... 265
7.2.1. ENDOCAM ... 265
7.2.2. CORVETTE ... 266
7.3. Future Outlook ... 267
7.3.1. SPAD Device ... 267
7.3.2. Mixed Signal Pixel ... 268
7.3.3. More Vertical Integration ... 269
7.3.4. Process Additions ... 270
8. Appendices ... 273
8.1. Dual Tier SPAD Pixel ... 273
8.1.1. Pixel Structure ... 273
8.1.2. Characterisation Results ... 275
8.2. Ultra-Miniature QIS Pixel ... 280
8.2.1. Pixel Design ... 280
References ... 283

Figures

Figure 1.1.1. Rendition of the Proteus fibre-based bedside diagnostic tool. 35
Figure 1.2.1. Block diagram of the OV6946 sensor replicated from published datasheet showing the system on chip design with minimal connectivity approach. 38
Figure 1.4.1. Direct time of flight (DToF). 41
Figure 1.4.2. Application examples of time-resolved imaging. (a) Direct time of flight (DToF) 3D imaging. (b) Time correlated single photon counting (TCSPC) fluorescence lifetime imaging microscopy (FLIM). 42
Figure 1.4.3. Application examples of time-resolved imaging. (a) Indirect time of flight (IToF) 3D imaging. (b) Time-gated fluorescence lifetime imaging microscopy with 2-gate rapid lifetime determination (RLD) method. 43
Figure 1.5.1. One-tap gate-based demodulator pixel replicated from [46] showing the substrate potential profile directing photo-generated charge towards the integration node. 44
Figure 1.5.2. Two-tap pinned photodiode pixel presented in [55] for fluorescence lifetime imaging. (a) Pixel layout. (b) Schematic diagram. 45
Figure 1.5.3. Survey of state of the art SPAD sensors pixel pitch as of 2014. Pixels are categorised according to time-resolved technique (TCSPC / Gated) and circuit implementation (Digital / Analogue). 48
Figure 2.1.1. Photodiode gain versus reverse bias voltage. Different regions of operation labelled. 54
Figure 2.1.2. Simple passive quench and recharge SPAD circuit. (a) Schematic. (b) Waveforms. 55
Figure 2.1.3. Five stages of SPAD avalanche cycle. 56
Figure 2.1.4. Illustration of reverse biased p-n junction. (a) Premature edge breakdown. (b) Guard ring structure to ensure uniform planar breakdown. 57
Figure 2.1.5. Variants of the p+ / DNW or NW SPAD with PW physical guard ring. (a) Rochas et al. [60]. (b) Pancheri et al. [94]. (c) Niclass et al. [97]. Multiplication junction highlighted in red. 58
Figure 2.1.6, Variants of the p+ / NW SPAD with PW physical guard ring. (a) Lee et al. utilising SOI process [99]. (b) Veerappan et al. utilising carrier diffusion [100]. Multiplication junction highlighted in red. 59
Figure 2.1.7. Illustration of depletion region borders (dashed red) for a SPAD with PW physical guard ring. (a) Standard device with functional multiplication junction. (b) Multiplication junction disappears when depletion borders or guard ring merge as device scales down. 60
Figure 2.1.8. STI bounded SPADs. (a) Original proposal by Finkelstein et al. [105]. (b) Passivated STI for lower noise contribution proposed by Gersbach et al. [106]. 61
Figure 2.1.9. Virtual guard ring concept. (a) Low doped p-type between spaced NW implants in a DNW [94]. (b) EPI region with blocked PW or NW doping between spaced NW implants with retrograde DNW [109]. Multiplication junction highlighted in red. 62

Figure 2.1.10. University of Edinburgh virtual guard ring SPADs using EPI with no PW or NW doping. (a) PW / DNW SPAD [111]. (b) DNW / Psub SPAD [116]. Multiplication junction highlighted in red. 63
Figure 2.1.11. Concept of SPAD structure in 28nm FDSOI process with CMOS circuits integrated in BOX [123]. Multiplication junction highlighted in red. 64
Figure 2.1.12. Conceptual schematics of different quench and recharge configurations. (a) Passive quench passive recharge (PQPR). (b) Active quench active recharge (AQAR). 69
Figure 2.1.13. Count rate versus illumination level SPAD response for passive recharge configuration (dashed) showing ambiguity due to paralysis and active recharge (solid) configuration showing fixed saturation limit....70
Figure 2.1.14. Biasing topology for non-substrate isolated DNW / Psub SPAD with poly-resistor R_Q and coupling capacitor. (a) Circuit diagram. (b) Waveforms....72
Figure 2.2.1. Single point sensor architecture. 73
Figure 2.2.2. Line sensor architecture....75
Figure 2.2.3. Image sensor architecture. 77
Figure 2.2.4. SPAD image sensor taxonomy....77
Figure 2.2.5. Conceptual schematic of ring oscillator based TDC. 78
Figure 2.2.6. Analogue SPC pixel presented in [65]. (a) Schematic diagram. (b) Timing diagram. 81
Figure 2.2.7. Pixel pitch versus technology node survey for more than thirty reported SPAD image sensors....83
Figure 2.2.8. Fill factor versus technology node survey for more than thirty reported SPAD image sensors....84
Figure 2.2.9. Illustration of the modular sensor architecture presented in [188]. 85
Figure 2.3.1. Generic illustration of TSV 3D-stacking process. (a) Processing and imaging tiers fabricated independently. (b) Imaging tier flipped over processing tier face to face, bonded and thinned from backside. (c) TSVs connectivity established. 87
Figure 2.3.2. Generic illustration of bump bonding 3D-stacking process. (a) Processing and imaging tiers fabricated independently. (b) Bump bonds fabricated over processing tier. (c) Imaging tier flipped over processing tier face to face, bonded and thinned from backside. 88
Figure 2.3.3. Generic illustration of hybrid bonding 3D-stacking process. (a) Processing and imaging tiers fabricated independently. (b) Imaging tier flipped over processing tier face to face, bonded and thinned from backside....89
Figure 2.3.4. Timeline of the key 3D-stacked SPAD published works. Red indicates in-house technologies and grey indicates foundry technologies. 90
Figure 2.3.5. Generic illustration of bridge bonding 3D-stacking process. (a) Imaging tier bonded over processing tier face to face. (b) Imaging tier thinned and etched to expose processing tier metallisation. (c) Metal bridge connecting both tiers is formed....91

Figure 3.1.1. a) Standalone circular SPAD pixel. b) Standalone rectangular SPAD pixel. c) Single strip local well sharing rectangular SPAD pixel. d) Double strip local well sharing rectangular SPAD pixel. e) Global well sharing square SPAD pixel. The pixel electronics and focal plane areas for each of the 2 × 2 arrays is equal thus reflecting the scaling of fill factor with layout style. 97
Figure 3.1.2. Generic SPAD design rules. 99
Figure 3.1.3. Fill factor versus pixel pitch trade-off for standalone SPAD layout pixels. (a) Circular standalone SPAD. (b) Rectangular standalone SPAD. Blue curves assume spacing rule (e) to nearby NMOS circuitry. Red curves assume spacing rule (f) to nearby PMOS circuitry n-well. 101
Figure 3.1.4. Fill factor versus pixel pitch trade-off for local well sharing SPAD layout pixels. (a) Single strip local well sharing SPAD. (b) Double strip local well sharing SPAD. Blue curves assume spacing rule (e) to nearby NMOS circuitry. Red curves assume spacing rule (f) to nearby PMOS circuitry n-well. 101
Figure 3.1.5. Fill factor versus pixel pitch trade-off for global well sharing SPAD layout pixels. Blue curves assume spacing rule (e) to nearby NMOS circuitry. Red curves assume spacing rule (f) to nearby PMOS circuitry n-well. 102
Figure 3.1.6. Fill factor versus pixel pitch trade-off comparison for the five different layout styles. 103
Figure 3.1.7. Fill factor versus resolution given an 800μm × 800μm focal plane area for the five different layout styles. 104
Figure 3.1.8. Generic layout of global well sharing image sensor with M columns and 2 × N rows. 105
Figure 3.1.9. Alternating metals layout strategy. 106
Figure 3.1.10. Projection of metal layers availability in different process nodes. 107
Figure 3.1.11. Projection of minimum metal pitch for different process nodes. 109
Figure 3.1.12. Estimated number of vertical anode connections in a global well sharing layout for different process nodes and SPAD guard ring widths. A comparison against a BSI implementation is also shown where routing can take place over the whole SPAD. 109
Figure 3.2.1. MINIC40 sensor. (a) Block diagram. (b) Chip micrograph. 112
Figure 3.2.2. MINIC40 pixel layout. Green is active area (OD), orange is n-well (NW), red is poly-silicon (PO) and dark blue is metal 1 (MT1). 113
Figure 3.2.3. MINIC40 intra-pixel routing layout. Red is poly-silicon (PO), dark blue is metal 1 (MT1), light blue is metal 2 (MT2) and pink is metal 3 (MT3). Horizontal row signals are annotated. 114
Figure 3.2.4. MINIC40 global vertical routing layout. Red is poly-silicon (PO), dark blue is metal 1 (MT1), green is metal 4 (MT4) and yellow is metal 5 (MT5). Vertical control signals are annotated. 115
Figure 3.2.5. MINIC40 2 × 2 pixel array. (a) With MT2 and MT3 routing visible showing how the pixel arrays by mirroring along the x-axis. (b) With MT4 and MT5 visible with the MT4 / MT5 free channel between the pixels noticeable. Vertical anode connections in MT4 and MT5 would occupy this space. 115

Figure 3.2.6. Global shared well layout of the bottom right corner of the array. Vertical anode MT4 tracks (green) from the SPADs flow in between pixels in the dedicated channel avoiding other horizontal MT4 routing used in-pixel. 116
Figure 3.2.7. (a) MINIC40 horizontal power strapping in MT6 (orange). (b) MINIC40 vertical power strapping in MT7 (grey). No MT6 or MT7 is used over the SPAD array region. 117
Figure 3.2.8. MINIC40 overall IC with four independent trials. 117
Figure 3.2.9. MINIC40 SPAD trials at fixed 8.25μm pitch. Shared NW is in orange, virtual EPI guard ring is in grey and anode connection is in light blue. Encapsulated white area represents the SPAD's active region (a) Reference trial with 3μm guard ring region. (b) Reduced shared NW. (c) Minimum shared NW allowed by process design rules. (d) Aggressive 1.5μm guard ring region trial with minimum shared NW and reduced EPI width. 118
Figure 3.2.10. MINIC40 pixel circuit block diagram with thick oxide transistors MQ, M0 and M1 forming the quench and front end inverter followed by thin oxide 40nm CMOS logic. The front end waveforms and their polarities are indicated in red. The front end output feeds the time gating logic block and a configurable 12-bit ripple counter which outputs its content on a 12-bit column parallel bus. 120
Figure 3.2.11. Thick oxide front end example with SPAD pull-up PMOS and level shifter. Red waveforms show the polarity and signal height at different nodes. 122
Figure 3.2.12. Thick oxide front end example with SPAD pull up PMOS and direct level shifter. Red waveforms show the polarity and signal height at different nodes. 123
Figure 3.2.13. Thick oxide front end example only with direct level shifter. (b) Thick oxide front end example with voltage clamp. Red waveforms show the polarity and signal height at different nodes. 123
Figure 3.2.14. Simulation waveforms for 250 MC runs at typical conditions for the three cases of front end inverter operation. 125
Figure 3.2.15. Histograms for time difference between input rising edge and output falling edge for 250 MC runs at typical conditions for the three cases of front end inverter operation. 126
Figure 3.2.16. Schematic diagram of custom 17 transistor D-type flip-flop. 127
Figure 3.2.17. Layout comparison of standard cell DFF and custom cell DFF. PMOS n-well (NW) is in orange, active area (OD) is in green, poly-silicon (PO) is in red, metal 1 (MT1) is in dark blue and metal 2 (MT2) is in light blue 128
Figure 3.2.18. The two phases of operation of the custom DFF. (a) Hold phase. (b) Sample phase. 129
Figure 3.2.19. Time gating logic schematic diagram. 130
Figure 3.2.20. Gating logic timing diagram. 131
Figure 3.2.21. Top view of extracted column crosstalk simulation. 133
Figure 3.2.22. Extracted column crosstalk simulation waveforms. a) For 1V excess bias case. b) For 2V excess bias case. c) For 3V excess bias case where coupling on horizontal MT5 neighbour lines surpasses the front end inverter 0.55V threshold. 133

Figure 3.3.1. PDP versus wavelength at 1V excess bias for the industrialised 40nm SPAD device replicated from [114] and the imaging 130nm SPAD reported in [65][236]. 134
Figure 3.3.2. Median DCR at room temperature for the four fill-factor SPAD trials implemented in MINIC40. 136
Figure 3.3.3. Cumulative DCR distribution at room temperature and 2V excess bias for the standard 39% and aggressive 66% fill-factor SPAD trials implemented in MINIC40...................................... 137
Figure 3.3.4. Average frame of standard MNIC40 array under fixed high illumination level. Colour scale represents photon counts.. 137
Figure 3.3.5. Illustration of typical passive quenched SPAD response showing the variation in count rate for different dead-times as illumination approaches saturation level. 138
Figure 3.3.6. Average frame of standard MNIC40 array under fixed low illumination level. Colour scale represents photon counts.. 139
Figure 3.3.7. Histogram of pixel photon counts of average frame of MINIC40 standard array showing pixel to pixel variation under different conditions. a) High illumination level. b) Low illumination level. 139
Figure 3.3.8. Higher contrast average frame of standard MNIC40 array under fixed low illumination level. Colour scale represents photon counts... 140
Figure 3.3.9. Zoomed micrograph of MINIC40 showing bottom left corner of SPAD array. 140
Figure 3.3.10. MINIC40 SPAD array current for different trials at different excess bias voltages.... 141
Figure 3.3.11. MINIC40 Enable signal propagation delay sequence across the array. Colour scale represents photon counts.. 143
Figure 3.3.12. MINIC40 time gate profile of a randomly selected pixel showing a record FWHM of 360ps using the edge-to-edge technique. 144
Figure 3.3.13. MINIC40 time gate FWHM map across the array. A clear mismatch between the top and bottom halves of the array is visible.. 144
Figure 3.3.14. MINIC40 time gate FWHM histogram. a) Top half of the array. b) Bottom half of the array. A mismatch of approximately 20ps is seen between the two halves... 145
Figure 3.3.15. MINIC40 handover between two contiguous 20ns time gates generated by the conventional rising-to-falling edge technique. A drop in photon counts is observed at the border between the gates due to mismatch in gate profiles. 146
Figure 3.3.16. MINIC40 handover between two contiguous 20ns time gates generated by the rising-to-rising edge technique. No drop in photon counts is observed at the border between the gates due to improved matching in gate profiles.. 146
Figure 3.4.1. Comparison of fill factor versus pixel pitch for state of the art FSI CMOS SPAD image sensors. 147
Figure 3.4.2. Comparison of fill factor versus single frame photon counting capacity of the art FSI CMOS SPAD image sensor pixels... 148

Figure 4.1.1. Principle of operation of a QIS where multiple binary bit planes are summed to form a single frame or an image 154
Figure 4.1.2. Quanta image sensor (blue) bit density D versus exposure H response (DlogH) along with the corresponding response of an ideal linear CIS (dashed red) for comparison. 155
Figure 4.1.3. Range of illumination levels exhibited in day to day environments 156
Figure 4.1.4. Logarithmic high dynamic range pixel [261]. 157
Figure 4.1.5. Lateral overflow concept from [266]. (a) Circuit diagram. (b) Timing diagram with top waveform representing the barrier control voltage B(t) and the bottom waveform representing the sense diffusion voltage for high light (solid red) and low light (dashed red) conditions. 158
Figure 4.1.6. Basic schematic of time to saturation pixel proposed in [269]. 159
Figure 4.1.7. Basic principle of light-to-frequency pixel reported in [272]. 160
Figure 4.1.8. Dual scan dual output HDR architecture reported in [276] 161
Figure 4.1.9. Dual photodiode mode concept to enhance dynamic range proposed by [284]. 163
Figure 4.2.1. HDR timing diagram comparison for global shutter triple-exposure acquisition and 8-bit depth image specification. (a) For conventional single bit QIS. (b) For MINIC40 with three in-pixel 4-bit counters 166
Figure 4.2.2. Data rates of different HDR architectures and sensor resolutions assuming an effective HDR frame rate of 30fps and 8-bit greyscale images per exposure frame 167
Figure 4.2.3. State of the art survey of maximum single frame photon counting capacity versus pixel pitch for different SPAD sensors grouped by architecture 168
Figure 4.2.4. Photon transfer curve for a single pixel operating in 12-bit linear counting mode showing shot noise limited photon counting. Red line is theoretical shot noise limited response. 169
Figure 4.2.5. Measured normalised intensity (D or 'Bit plane density') to normalized input signal (H) for two sets of integration times for 1 photon threshold (K=1). (a) Exposure ratio of 10 with Short = 100ns, Mid = 1μs, Long = 10μs. (b) Exposure ratio of 2 with Short = 100ns, Mid = 200ns, Long = 400ns. 171
Figure 4.2.6. Measured normalised intensity (D or 'Bit plane density') to normalized input signal (H) for two sets of integration times for 2 photon threshold (K=2). (a) Exposure ratio of 10 with Short = 100ns, Mid = 1μs, Long = 10μs. (b) Exposure ratio of 2 with Short = 100ns, Mid = 200ns, Long = 400ns. 172
Figure 4.2.7. Measured signal, noise and SNR_H responses for 3 exposure settings with exposure ratio of 10 and photon threshold K=1 174
Figure 4.2.8. Measured signal, noise and SNR_H responses for 3 exposure settings with exposure ratio of 10 and photon threshold K=2 176
Figure 4.2.9. Measured signal, noise and SNR_H response for 3 exposure settings and K=1. (a) Exposure ratio of 2. (b) Exposure ratio of 8. 178
Figure 4.2.10. Images captured by 96 × 40 FSI sensor [2]. (a) Sum of 256 fields at 0.1 μs exposure, a Minion figure is visible. (b) Sum of 256 fields at 1.0μs exposure, the Minion is visible but slightly

overexposed while faint letters appear in the background. (c) Sum of 256 fields at 10µs exposure, the Minion is totally overexposed but the letters appear clearer. (d) Linear sum of all the 768 fields from a, b and c to form an HDR image preserving all details. 179

Figure 4.2.11. Images captured by 320 × 240 SPC sensor from [65]. (a) Sum of 1000 fields at 0.1µs exposure, a Minion appears in the lit portion of the scene. (b) Sum of 1000 fields at 1.0µs exposure, the Minion is slightly overexposed but a car figure appears in the dark region of the scene. (c) Sum of 1000 fields at 10µs exposure, Minion is completely overexposed but more detail of the car is apparent. Notice that high DCR pixels appear as white dots. (d) Linear sum of all the 3000 fields from a, b and c to form an HDR image preserving all details. 180

Figure 5.1.1. MINI3D sensor. (a) Block diagram. (b) Chip micrograph showing backside of top tier. 184

Figure 5.1.2. MINI3D pixel circuit block diagram with thick oxide transistors MQ, M0 and M1 forming the quench and front end inverter followed by thin oxide 40nm CMOS logic. The front end waveforms and their polarities are indicated in red. The pixel has a single 1.1V supply (VDD) common to all components. 185

Figure 5.1.3. Cross section of MINI3D backside illuminated 3D-stacked pixel layout showing both top and bottom tiers 186

Figure 5.1.4. MINI3D pixel layout. (a) Layers up to MT3 only. (b) MT4 and MT5 routing frame showing flexible use of these metals compared to MINIC40. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3, green is MT4 and yellow is MT5. Higher metal layers are switched off for clarity. 187

Figure 5.1.5. MINI3D PMOS quench front end with signal polarities shown in red. 188

Figure 5.1.6. Dual SPAD bias concept enabled by full depth DTI. 189

Figure 5.1.7. Layout of 3 × 2 pixels of MINI3D main SPAD trial. 189

Figure 5.2.1. SPAD breakdown voltage characterisation. a) Counts versus VHV sweep for a random pixel with linear fit in red. b) Breakdown voltage distribution across array with mean of 11.7V and 30mV standard deviation with Gaussian fit in red. 190

Figure 5.2.2. Median DCR versus excess bias at room temperature for MINI3D sensor. 191

Figure 5.2.3. Cumulative DCR distribution versus excess bias at room temperature for MINI3D sensor. 192

Figure 5.2.4. SPAD jitter versus excess bias for different wavelengths. 193

Figure 5.2.5. SPAD jitter impulse response functions at different excess bias. a) 443nm. b) 773nm. 193

Figure 5.2.6. MINI3D array uniformity. Mean frame of 5000 captures under fixed illumination. The four darker columns to the right are without metal reflectors. 194

Figure 5.2.7. MINI3D array uniformity. a) Histogram of normalised array response. b) Mean frame with higher contrast colour scale revealing darker edge rows and columns 194

Figure 5.2.8. Leakage effect at edge of MINI3D array due to close proximity of GND substrate contact to the high voltage SPAD shared NW 195

Figure 5.2.9. Photon detection probability of the BSI SPAD with metal reflector at different excess bias voltages. 196
Figure 5.2.10. Photon detection probability of the BSI SPAD with and without metal reflector at 3V excess bias. 196
Figure 5.2.11. Measured photon transfer curve for MINI3D sensor. Blue points show a pixel response while the dashed black line shows a frame response. Inset histograms are for the pixel response at 20ns (left) and (1ms) right exposures with ideal Poisson fit overlaid as red crosses. 197
Figure 5.2.12. MINI3D single shot greyscale image in 12-bit linear counter mode. (a) Raw image. (b) DCR corrected by interpolation. 198
Figure 5.2.13. Time gate profile. (a) 1ns, 4ns and 8ns FWHM time gates for bin 1 of a randomly selected pixel. (b) Distribution of 4ns gate FWHM across the array with 64ps standard deviation... 199
Figure 5.2.14. Indirect time of flight experiment timing diagram. 200
Figure 5.2.15. Indirect time of flight experiment with 4ns laser pulse width. (a) Intensity image of field of view of the sensor showing a mug's handle obstructing a Minion toy. (b) Depth map of the scene distinguishing the two objects at different distances from the sensor. Median filter applied... 201
Figure 5.3.1. PDP comparison of FSI and BSI SPAD at 2V excess bias. 202
Figure 5.3.2. Illustration of a SPAD structure. (a) FSI. (b) BSI. 203
Figure 5.3.3. PDE comparison of FSI and BSI SPAD at 2V excess bias. 204
Figure 5.3.4. PDP comparison of different 3D-stacked BSI SPAD at their highest reported excess bias condition. 205
Figure 6.1.1. ENDOCAM system on chip sensor block diagram. 211
Figure 6.1.2. ENDOCAM pixel circuit diagram. Traces in red indicate signal polarities. 212
Figure 6.1.3. Layout of ENDOCAM 8µm digital pixel. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3 and green is MT4. MT5 to MT7 are not shown for clarity.......... 212
Figure 6.1.4. ENDOCAM pixel different modes of operation and corresponding signal states.......... 213
Figure 6.1.5. ENDOCAM top tier SPAD pixel metallisation structure forming a light reflector and VHV MIM decoupling capacitor cell. a) Shows cross section of the device. b) Shows top view of the MT3 and MT4 metal plates forming the decoupling capacitor alongside a three-dimensional rendition of the cell layout. 214
Figure 6.1.6. Typical row timing diagram in rolling shutter mode for a selected row. 216
Figure 6.1.7. ENDOCAM's synthesised MCU layout occupying a footprint of 225µm × 225µm. Only MT1 to MT5 shown. 217
Figure 6.1.8. ENDOCAM overall system setup. (a) Block diagram. (b) Simplified operation state diagram. 218
Figure 6.1.9. Improvement in dynamic range in dB for a given number of summed frames. 220
Figure 6.1.10. Systems with different data flow models. a) Standard image sensor architecture with single IO pad for serial readout off-chip. b) Proposed image sensor architecture with on-chip frame storage (and summation) with single IO pad for serial readout off-chip.......... 221

Figure 6.1.11. Effective frame rate versus oversampled pixel bit depth for the off-chip and on-chip frame processing architectures assuming a 25MHz system clock frequency. 223
Figure 6.1.12. Required system clock frequency for different in-pixel bit depth counters for an effective frame rate of 30fps and an oversampled bit depth of 16-bits to be achieved given off-chip and on-chip frame store system architectures. 224
Figure 6.1.13. Temporal aperture ratio (TAR) versus chosen oversampled bit depth for ENDOCAM sensor in chained SRAM mode. 225
Figure 6.1.14. ENDOCAM temporal aperture ratio (TAR) versus system clock frequency for oversampling ratios of 2 and 4 in ping-pong SRAM mode. 226
Figure 6.1.15. ENDOCAM effective frame rate (fps) versus system clock frequency in ping-pong SRAM mode. 227
Figure 6.1.16. Simplified block diagram of ENDOCAM power generation network with core blocks, configuration register bank and main signals labelled. 229
Figure 6.1.17. ENDOCAM VQuench MUX settings with 1.1V from regulator as default setting upon start-up. 230
Figure 6.1.18. Start-up simulation of ENDOCAM power generation network with glitch due to capacitive loading on bandgap reference voltage. 231
Figure 6.1.19. Smooth start-up simulation of ENDOCAM power generation network after fixing initial start-up setting to avoid the glitch due to capacitive loading on bandgap reference voltage. 231
Figure 6.1.20. Schematic of inverter based VCSEL driver implemented in ENDOCAM sensor. 232
Figure 6.1.21. Layout of ENDOCAM power generation network with main blocks labelled. 233
Figure 6.1.22. Simplified block diagram of ENDOCAM gating logic. 234
Figure 6.1.23. ENDOCAM gating logic timing diagram for one counter pair. 235
Figure 6.1.24. ENDOCAM ring oscillator timing in self reset mode. 236
Figure 6.1.25. ENDOCAM gate generation logic layout occupying an area of 164μm × 55μm. 237
Figure 6.1.26. ENDOCAM 1.4mm × 1.4mm bottom tier 40nm IC. 238
Figure 6.1.27. ENDOCAM 1.4mm × 1.4mm top tier BSI 90nm IC. 239
Figure 6.1.28. ENDOCAM micrograph showing the backside of the top tier IC. 240
Figure 6.1.29. Live stream image from ENDOCAM sensor in rolling shutter mode at 11fps. The author (right) and colleague Hanning Mai (left). 241
Figure 6.1.30. Data Latch and Row Reset signal propagation paths for ENDOCAM readout. 242
Figure 6.1.31. ENDOCAM images captured with the on-chip time gate generation feature enabled. a) Only time gate for odd columns enabled. b) Both time gates of odd and even columns enabled but ring oscillator switched off resulting in no photons captured. 244
Figure 6.2.1. CORVETTE pixel schematic diagram. 246
Figure 6.2.2. CORVETTE 6.48μm pixel layout. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3 and green is MT4. Higher metal layers are not shown for clarity. The

dashed box encloses the front end transistors from the neighbouring pixel due to the interleaved layout methodology. 247
Figure 6.2.3. Self-reset circuit block of CORVETTE pixel used in photodiode mode. a) Schematic diagram. b) Example timing diagram or generated photodiode reset signals. 248
Figure 6.2.4. Timing diagram of CORVETTE pixel in photodiode mode. 249
Figure 6.2.5. CORVETTE column and horizontal XOR tree structure. 251
Figure 6.2.6. Layout of CORVETTE flash TDC module with on-chip histogramming. 253
Figure 6.2.7. CORVETTE system block diagram. 254
Figure 6.2.8. Layout of CORVETTE 1.7mm × 1.3mm 40nm bottom tier IC. 255
Figure 7.3.1. Illustration of the ideal fully depleted DTI bound SPAD structure. 268
Figure 7.3.2. Mixed signal photon counting pixel with digital front end and LSB counter and analogue MSB counter based on CTA pixel in [65]. 268
Figure 7.3.3. Illustration of vertically integrated modular sensor with mix and match tier options... 270
Figure 8.1.1. Dual-tier pixel schematic. All circuits are implemented in the 40nm bottom tier. Red waveforms show signal polarity, level and profile. 274
Figure 8.1.2. Cross section of the dual-tier pixel. 275
Figure 8.1.3. Dual tier pixel PDP at 3V excess bias. 276
Figure 8.1.4. Dual tier pixel light count rate versus illumination of a white LED 277
Figure 8.1.5. Dual tier pixel count rate versus incident light angle given a fixed illumination level. 278
Figure 8.1.6. Ratio of bottom SPAD counts to top SPAD counts versus incident light angle. 278
Figure 8.1.7. Dual tier pixel SPADs jitter at 2V excess bias and 773nm. 279
Figure 8.2.1. Single-bit self-locking QIS pixel schematic. Red signals show front end waveforms.. 281
Figure 8.2.2. Alternative single-bit self-locking QIS pixel implemented in CORVETTE. (a) Schematic diagram. (b) Pixel layout at 3.24µm pitch. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3, green is MT4 and yellow is MT5. Higher metal layers are not shown for clarity. 282

Tables

Table 1.2.1. Summary of key specifications of some of the most notable commercially available endoscopy cameras 37
Table 2.1.1. SPAD performance summary for some of the recent published FSI SPAD devices 68
Table 2.2.1. Summary of TDC-based FSI SPAD image sensor parameters showing good temporal resolution but low fill factor and large pixel pitch for the exception of [170] implemented in an advanced 40nm node 78
Table 2.2.2. Summary of resource sharing TDC-based FSI SPAD image sensor parameters showing good temporal resolution and improved pixel pitch and fill factor compared to pixel dedicated TDC architectures. 79
Table 2.2.3. Summary analogue SPC SPAD image sensors showing the small pitch and high fill factor but relatively low photon counting capacity per pixel in a single frame capture. 81
Table 2.3.1. Summary of 3D-stacked sensors presented in literature alongside sensors presented in this work. 93
Table 3.2.1. Summary of MINIC40 SPAD trials guard ring dimensions and drawn fill factor. 119
Table 3.2.2. MINIC40 pixel supply and control signals. 121
Table 3.2.3. MINIC40 pixel front end thick oxide transistors size summary 124
Table 3.2.4. Gating logic truth table 132
Table 3.4.1. List of FSI CMOS SPAD image sensor arrays since 2006. Architectures marked in red are intended for photon counting and time-gated operation. 149
Table 3.4.2. Comparison of time-gating techniques and minimum reported time-gate FWHM for different SPAD sensors and a recent pinned photodiode FLIM image sensor. Image sensors are marked in red. 150
Table 4.2.1. Calculated SNR_H and DR_{MAX} from measured data for the case of single, double and triple exposures with a ratio of 10 and K=1. 175
Table 4.2.2. Calculated SNR_H and DR_{MAX} from measured data for the case of single, double and triple exposures with a ratio of 10 and K=2. 176
Table 4.2.3. Measured SNR_H and DR_{MAX} for a three exposures scenario and K=1 with different exposure ratios of 2 (0.1µs, 0.2µs and 0.4µs), 4 (0.1µs, 0.4µs and 1.6µs), 6 (0.1µs, 0.6µs and 3.6µs) and 8 (0.1µs, 0.8µs and 6.4µs) 177
Table 4.3.1. Comparison of data rates of different high dynamic range QIS and in-pixel compression scenarios 181
Table 5.1.1. Summary of trialled SPAD structures on MINI3D. 187
Table 5.3.1. Comparison table of key technology parameters for different all-CMOS 3D-stacked SPAD Sensors. Image sensors are marked in green and the peak PDP wavelength in grey highlighting a trend in BSI SPADs. 205
Table 6.1.1. Summary of target specifications for ENDOCAM SoC sensor. 210

Table 6.3.1. Comparison table of time-resolved image sensors.......... 259

Table 6.3.2. Comparison of temporal response of PPD-based pixels and ENDOCAM's SPAD pixel. 260

Table 6.3.3. Comparison of ENDOCAM to commercially available endoscopy image sensors.......... 261

1. Introduction

This thesis sets out to explore the design of miniature complementary metal oxide silicon (CMOS) single photon avalanche diode (SPAD) image sensors. SPADs offer the unique ability to detect single packets of electromagnetic radiation with picosecond temporal resolution making them the devices of choice for photon counting and time-resolved applications, both features desired in this work. Miniaturisation of the sensor is pursued via three routes: integration in advanced technology nodes and state of the art fabrication processes, smart and optimised pixel designs and novel system architectures that alleviate the performance constraints of such sensors.

1.1. Proteus Project

This work was carried out as part of the Proteus project (https://proteus.ac.uk) funded by the Engineering and Physical Sciences Research Council (EPSRC), UK, under grant number EP/K03197X/1. Proteus is a multi-disciplinary effort bringing together researchers having a wide range of expertise spanning the fields of biology, chemistry, physics and engineering with the shared vision of creating a complete diagnostic tool for intensive care patients with infectious lung diseases.

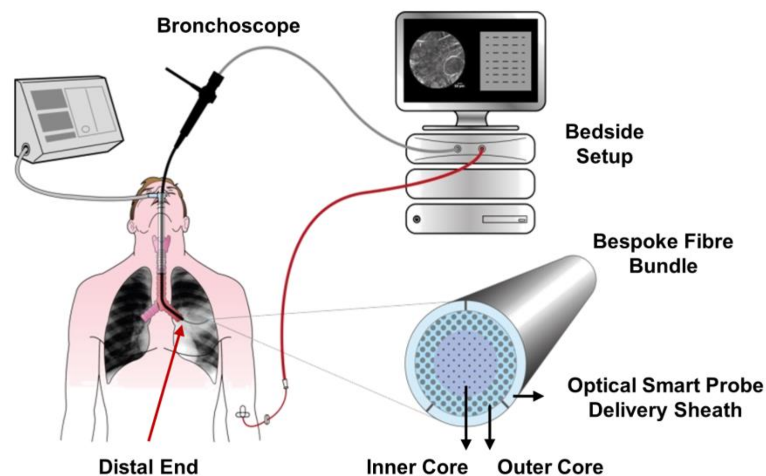


Figure 1.1.1. Rendition of the Proteus fibre-based bedside diagnostic tool.

At the core of Proteus' solution is a fibre-based flexible bronchoscope (Figure 1.1.1) used to deliver specialised biomarkers that tag particular bacteria or biosensors that characterise the lung's environment (pH sensors for example), extract samples for ex-vivo analysis and carry light in and out of the lung for imaging and sensing purposes. Amongst the key elements coupled to the endoscopy system are bespoke CMOS SPAD sensors whose time-resolved capability brings an additional level of sophistication to the extracted data used by clinicians to make life determining decisions.

While the main theme of the Proteus project revolves around a bedside setup which encloses most of the optical and electronic components, this work investigates the alternative of designing a SPAD sensor that could be integrated on the distal end of the endoscope which goes inside the human lung. Miniaturisation is therefore a key requirement. Smaller, lower cost disposable time-resolved endoscopy sensors and the avoidance of optical transmission losses through a fibre by having the sensor in-vivo are the main motives behind this direction.

The challenges associated with such a target application are many - packaging, optics and connectivity to name a few- yet this thesis focuses solely on the silicon design aspects from an integrated circuit (IC) designer point of view.

1.2. State of the Art Miniature Endoscopy Cameras

Endoscopes can serve as means of diagnosis, therapeutic delivery or as invasive surgical tools and can be categorised into three main types of devices. The first is rigid endoscopes which can house several medical instruments within a firm insertion tube [11]. The second is flexible endoscopes which are similar to the rigid ones but with flexible skeletons that allow for manoeuvrability inside the body during a procedure [12]. The third is single-use disposable endoscopy systems that can be in the form of ingestible pills [13], rigid [14] or flexible endoscopes [15].

Rigid and flexible endoscopes are multi-use multi-purpose systems and so include a variety of instruments such as CMOS image sensors and mechanical surgical arms. Hence the module tip can be of several millimetres of diameter although smaller dimensions are preferable. In such modular setup, the image sensor is not necessarily the size restricting element with high resolution and high sensitivity being key requirements.

On the other hand disposable endoscopes, which are an emerging sector in the endoscopy market, are mostly tailored towards a specific application with only necessary instruments integrated for low cost. Ingestible capsules such as PillCam® [16] can still be an inch in length and several millimetres wide due to the optical, power and data transmission components but can benefit from miniaturised image sensors. In contrast, disposable rigid or flexible endoscopes tend to be finer in dimension with the tip measuring down to 1mm in diameter for extended reach into small cavities of the human body and so the image sensor form factor becomes a limitation.

Such application oriented disposable endoscopes are cheaper means of diagnosis, safer and more sterile solutions due to single-use and offer higher patient comfort levels due to smaller form. Yet due to the restricted set of integrated instruments they can gain added value by exploiting recent developments in the field of CMOS sensors, thus justifying the Proteus proposal of a miniature time-resolved chip-on-tip module aimed for disposable endoscopes and bronchoscopy applications in particular.

Miniature cameras for disposable endoscopes are readily available on the market with big players in the CMOS image sensor (CIS) business such as Awaiba; now part of Austria Micro-Systems (AMS), OmniVision Technologies (OVT) and Toshiba leading the way in terms of miniaturisation and wafer level integration. Table 1.2.1 summarises the key specifications of some of the most notable commercially available systems. The applications of these cameras extend beyond medical endoscopy to include security, internet of things (IoT) and industrial inspection due to their unique small form factor.

	Awaiba NanEye [17]	**OmniVision OV6946 [18]**	**OmniVision OV6948 [19]**	**Toshiba IK-CT2 [20]**
Resolution	250 × 250	400 × 400	200 × 200	220 × 220
Pixel Pitch	3μm	1.75μm	1.75μm	n/a
Technology	FSI	0.11μm BSI	0.11μm BSI	BSI
Frame Rate	42 to 55 fps	30 fps	30 fps	60 fps
Full Well	15ke-	n/a	n/a	n/a
Dynamic Range	58 dB	65.8 dB	60.2 dB	n/a
Shutter	Rolling	Rolling	Rolling	n/a
Connection Pins	4	4	4	n/a
Output Interface	LVDS	Analogue	Analogue	n/a
Sensor Dimensions	1mm × 1mm	0.95mm × 0.94mm	0.58mm × 0.58mm	0.7mm × 0.7mm
Endoscopy Module Available	Yes	Yes	Yes	Yes

Table 1.2.1. Summary of key specifications of some of the most notable commercially available endoscopy cameras.

Comparing the different sensors some common trends are observed from which the goals of this research can be derived in order to match or go beyond the state of the art in the field and these include:

1. Spatial resolution close to 200 × 200 pixels.
2. Effective frame rate operation of 30fps.
3. Dynamic range of 60dB in line with typical mainstream CISs.
4. Four wire interface for minimal connectivity overhead.
5. Small sensor footprint of 1mm × 1mm or less.

Another common trend seen when examining the published block diagrams of the aforementioned sensors is the fully integrated system on chip (SoC) approach. This is necessary for fully autonomous operation with minimal peripherals or connectivity requirements with all critical sub-blocks integrated on-chip while avoiding unnecessary processing overhead that could be carried out off-chip. As an example, Figure 1.2.1 shows the block diagram of the OV6946 sensor [18] as seen in the datasheet.

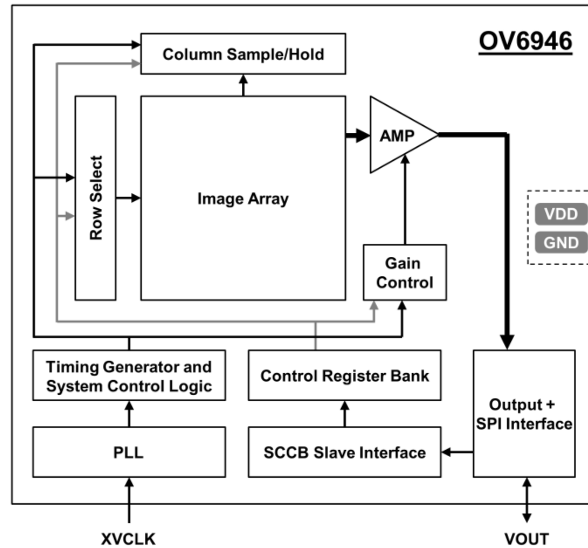


Figure 1.2.1. Block diagram of the OV6946 sensor replicated from published datasheet showing the system on chip design with minimal connectivity approach.

The literature also includes examples of recent developments addressing some of the challenges imposed by miniaturisation. A 10k pixel sensor was presented in [21] where the 3µm pixel design occupies a remarkable footprint of 0.34mm² and utilises a current based analogue to digital converter (ADC) to overcome power supply noise over long cabling distances associated with some endoscopy applications. The work in [22] highlights the challenges of wafer level chip scale packaging (WLCSP) such as loss of peripheral area due to dicing and packaging wall spacers and compatibility with micro-lensing.

Other works focus on optimising the optical efficiency of the focal plane as the pixel area in CISs tends to be rectangular in dimension while the lenses are spherical in shape resulting in aberrations and optical distortion at the edges of the rectangular image. To overcome that, a rectangular IC with an octagonal pixel array was proposed in [23] to match the focal planes of the image sensor and the lens. More recently, a CIS with a unique octagonal layout was demonstrated by [24]. The IC is also sawn in an octagonal form such that it maximises the imaging area when placed on the tip of conventionally cylindrical endoscopes.

On the other hand, other efforts focus on the endoscopy module or hardware system design in which the miniature sensor is a main component. A disposable camera system comprising a CIS, optics and a light emitting diode (LED) was presented in [25] for minimally invasive surgery. Similarly a low power wireless capsule endoscope for fluorescence imaging based on a SPAD sensor was presented in [26][27] including optics, LED source and a battery unit.

As can be seen, the mission of miniaturising a CIS let alone a SPAD-based image sensor is a multi-dimensional problem which depends on the end application, hence the author's decision to limit the scope of this work to the remit of IC design within the available resources.

1.3. The Missing Element: Time-Resolved Capability

Since 1971 when Michael Tompsett proposed the use of charge coupled devices (CCDs) for imaging applications [28], CCDs became the technology of choice for solid state imaging launching a new era of engineering research and development towards high performance silicon-based image sensors. For a full account of the principles of operation and technological developments of CCDs the interested reader is referred to Chapter 4 of the fascinating book "Image Sensors and Signal Processing for Digital Still Cameras" [29].

Two decades later, shortcomings of CCD technology such as high power consumption, limited frame rate and resolution started to surface prompting the industry to turn to CMOS technology as an alternative. The first demonstration of a CMOS camera on chip based on a passive pixel sensor (PPS) was presented by researchers at the University of Edinburgh in 1990 [30] followed by a huge momentum in developing active pixel sensor (APS) CMOS image sensors.

Shortly after, in 1993, Eric Fossum published his famous paper "Active Pixel Sensor: Are CCD's Dinosaurs?" [31] which summarised the various efforts at the time in developing variants of the APS and marked the point in history when CISs were crowned on the throne of electronic imaging over CCDs. The APS relies on a source follower amplifier readout circuit first conceived by Peter Noble in 1968 [32] and later became the predominant configuration due to its superior noise performance over PPS and is now implemented in billions of camera devices.

The pursuit of the optimal CIS took off leveraging the technical advancements pioneered for CCDs, most notably the pinned photodiode (PPD) first invented by Nobukazu Teranishi in 1980 and reported in 1982 [33]. Due to its fully depleted and buried structure the PPD provided better lag and lower dark noise performance and is now at the heart of CIS pixels. An excellent review of this key invention is presented by Fossum et al. in [34].

CISs have come a long way in maturity and it is not possible to capture the history of development and the state of the art in few brief paragraphs, yet the literature provides very informative accounts including Fossum's 1997 paper "CMOS Image Sensors: Electronic Camera-On-A-Chip" [35], Theuwissen's 2008 paper "CMOS image sensors: State-of-the-art" [36] and Fossum's 2013 paper "CAMERA-ON-A-CHIP: TECHNOLOGY TRANSFER FROM SATURN TO YOUR CELLPHONE" [37]. Of course the aforementioned book [29] also provides an excellent technical reference.

Throughout the years, lower noise, higher resolution, higher quantum efficiency (QE) and improved sensitivity were amongst the main drivers of CISs such that we can capture the world around us in a two dimensional image with stunning levels of detail and clarity. Yet we live in a three dimensional space with the third dimension, hereby considered as depth, not being targeted by CISs up until recently, and hence the missing element of time-resolved capability.

In this context time-resolved capability refers to the ability to not only capture photons (i.e. intensity imaging), but also to extract the photon time of arrival (ToA) in order to add a new perspective to the captured image. In a consumer market scenario, this can be exploited as depth information (3D imaging) in applications such as gesture recognition for gaming or light detection and ranging (LIDAR) for autonomous driving. Alternatively in a biomedical scenario, time information provides contrast between biomedical samples that, although spatially or spectrally similar, emit photons having different temporal patterns in applications such as fluorescence lifetime imaging microscopy (FLIM) or fluorescence resonance energy transfer (FRET) [38].

Since all of the endoscopy sensors described in Section 1.2 are based on the CMOS APS technology, they naturally inherit all of the technological advancements and performance of mainstream sensors, and similarly are tailored for capturing two dimensional images within the limits of the application. At the start of this project in September 2014 no such miniature sensor with time-resolved capability existed and so it became the challenge of this work to conceive one.

A time-resolved capable endoscope would be able to identify the presence of pathogens by measuring their auto-fluorescence lifetime against background tissue or by measuring changes in the lifetime of actively introduced biomarkers that bind to targeted molecules [39]. Such specificity in detection would inform treatments prescribed by clinicians and assist in fluorescence-guided surgical oncology for cancerous tissue removal [40]. It also opens new application avenues such as 3D imaging guided keyhole surgery and physical environment characterisation such as pH sensing by Raman probes [41].

1.4. Time-Resolved Imaging

Time-resolved imaging is a broad term encompassing all techniques of extracting temporal information from detected photons and their applications, but for simplicity the discussion in this thesis is limited to 3D imaging (consumer) and FLIM (biomedical) as the most applicable use cases in order to explain concepts since the underpinning principles are similar.

Unlike intensity imaging where photons in the scene are generated by uncorrelated sources such as natural lighting, time-resolved imaging requires a correlated light source synchronised to the integration period of the image sensor in order to extract the temporal information of detected photons with respect to a reference. This principle is known as time of flight (ToF) and is composed of three elements (Figure 1.4.1):

1. Light source and its optics.
2. Light detector and its optics.
3. Timing element.

When a light pulse is sent out by the light source the timing element (stop watch) starts recording time. The light pulse travels for a distance D, hits a target and reflects back towards the detector. Once the detector picks up the return pulse it stops the timing element registering the elapsed period taken by the light pulse to complete a round trip. Thus the time of arrival of the pulse photons is extracted with respect to the reference point defined by the emission of the light source.

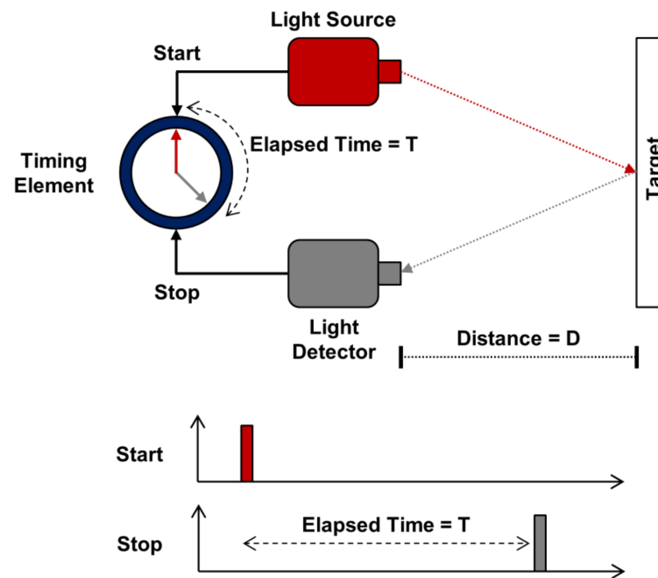


Figure 1.4.1. Direct time of flight (DToF).

To put it in context of 3D imaging and FLIM applications, consider the example of a single pixel observing a target object or a biomedical sample. The light source is pulsed repeatedly with the elapsed time or time stamp (in reference to the laser pulse) of every detected return pulse recorded. After a predefined exposure period, the collected time stamps are sorted into a histogram representing the statistical nature of photon detections where the x-axis represents time and the y-axis represents occurrences.

For 3D imaging, a peak in the histogram will appear at the time bin corresponding to the distance of the target object as depicted in Figure 1.4.2(a). For FLIM (Figure 1.4.2(b)), the histogram would represent a fluorescence decay response. In FLIM the observed sample absorbs the energy of the exciting pulse and undergoes a transition in internal energy states after which it releases a photon having a different (typically longer) wavelength with a time delay with respect to the source. This time delay varies in an exponential fashion reflected in the occurrences of time stamps across the

histogram. A unique lifetime of the photon emission related to the sample and its properties can then be extracted by exponential fitting.

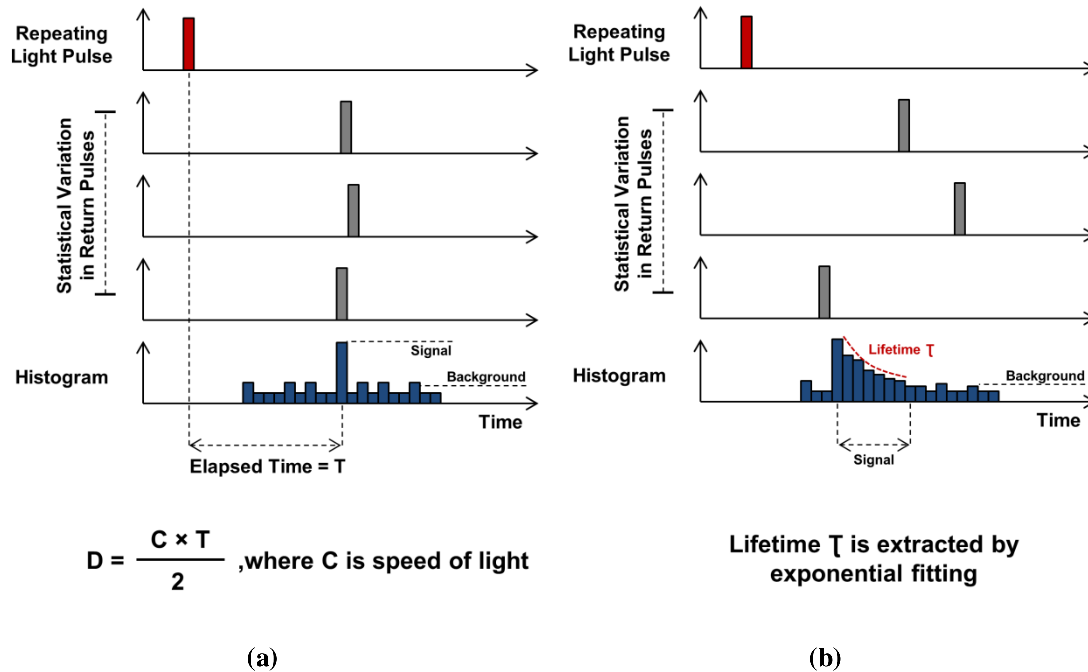


Figure 1.4.2. Application examples of time-resolved imaging. (a) Direct time of flight (DToF) 3D imaging. (b) Time correlated single photon counting (TCSPC) fluorescence lifetime imaging microscopy (FLIM).

Since information is extracted from direct time stamps of every captured photon, this ToF principle is known as direct time of flight (DToF), a term used when referring to ranging applications. In scientific imaging this principle is more commonly known as time correlated single photon counting (TCSPC) and mostly relies on discrete components with photomultiplier tubes (PMTs), custom avalanche photodiodes (APDs) or SPADs as receivers, scientific grade lasers (e.g. Hamamatsu) as light sources and time measurement cards [42] as the timing element coupled to off-line processing.

Another time-resolved measurement principle known as indirect time of flight (IToF) can also be employed to infer distance. IToF differs from DToF in the sense that it does not time stamp individual photons but integrates them within two or more windows of observation (also known as time gates) in reference to the outgoing light pulse. These time gates are coarser in time (typically several nanoseconds) compared to the timing element resolution in DToF (typically tens of picoseconds).

Using the same example of a single pixel observing a target object or a biomedical sample, the light source is repeatedly pulsed with the pixel observing the return pulses within two time gates. For 3D imaging a wider light pulse is used with a first time gate coinciding in time with the light pulse and a second time gate of the same width following contiguously (Figure 1.4.3(a)). The return signal reflected from the target is offset by a period of time corresponding to the target distance. After a

predefined exposure period the accumulated photons in each time gate is proportional to the overlap in time between the time gates and the return pulse. The ratio between the two integrated signals is then used to calculate the distance.

In the case of FLIM, where this principle is referred to as time gating, the two time gates are positioned in time such that they cover the expected fluorescence return duration. The example in Figure 1.4.3(b) assumes two contiguous time gates of equal width but other gating configurations are possible based on the chosen algorithm for lifetime extraction. After several short light pulse repetitions within an exposure period, the accumulated signal in each gate represents the observed portion of the decay and the lifetime can be calculated by rapid lifetime determination (RLD) method [43]. This is the simplest and most circuit efficient approach and will be considered the norm in this work.

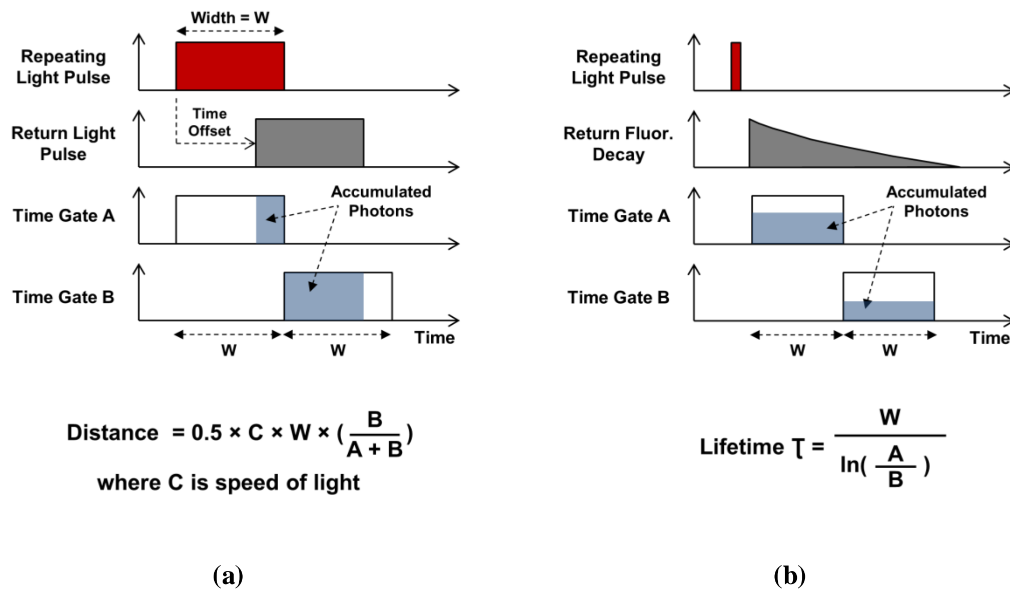


Figure 1.4.3. Application examples of time-resolved imaging. (a) Indirect time of flight (ITоF) 3D imaging. (b) Time-gated fluorescence lifetime imaging microscopy with 2-gate rapid lifetime determination (RLD) method.

Throughout this thesis the terms DToF and TCSPC will be used interchangeably referring to sensors or systems that time stamp received photons. Where time gating operations are being considered, the terms IToF and time-gated will be used to refer to sensors or systems that integrate photons within defined observation windows.

While the ToF examples described above assume a pulsed light source, it is worth noting that the use of continuous wave (CW) light sources is also possible; more commonly with IToF, but omitted from the discussion herein as it is beyond the scope of this introduction. The interested reader is directed to two excellent books "TOF Range-Imaging Cameras" [44] for in-depth detail of ToF principles and

solid state sensors and "Advanced Time-Correlated Single Photon Counting Techniques" [45] for in-depth detail of TCSPC systems and applications.

1.5. Solid State Time-Resolved Sensors

Solid state time-resolved sensors are relatively recent developments in the electronic imaging industry, with the first all solid state array presented by Robert Lange and Peter Seitz in 2001 [46] utilising a demodulating pixel first proposed in 1995 [47] based on CCD principle. Such a pixel belongs to a class of ToF pixels known as photo-mixing devices (PMDs) where the photo-generated charge is steered towards one or more collection nodes in a timely fashion relative to the illumination source.

Figure 1.5.1 replicated from [46] demonstrates the working principle of this one-tap gate-based demodulating pixel. When the pixel is operating within the required observation window, a gradient of voltages is applied to the photo-gates (Middle, Left and Integration in this example) thus creating a potential profile gradient in the substrate that steers photo-generated charges towards the integration node until a readout is performed. Alternatively, by reversing the applied photo-gate voltages the accumulated charge can be disposed of through the dump node. This mechanism of moving charge is similar to that used in CCDs to transport charge through the line.

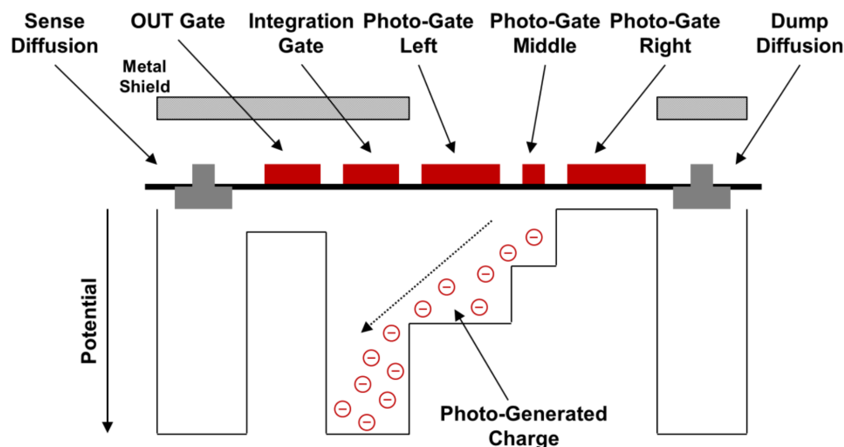


Figure 1.5.1. One-tap gate-based demodulator pixel replicated from [46] showing the substrate potential profile directing photo-generated charge towards the integration node.

A more recent implementation of gate-based demodulators known as quantum efficiency demodulators (QEMs), whereby the charge collection efficiency of regions or pockets of the substrate is modified by a voltage applied to the poly gate covering it, has been demonstrated [48]. Unlike the above gate-based demodulator poly gate voltages are not used to move charge around. The 512×424 10μm pixel is an impressive example of successfully commercialised ToF technology for 3D imaging

with even a more aggressive 1Mpixel 3.5µm pitch backside illuminated (BSI) sensor presented at ISSCC 2018 [49].

Another example of PMDs is a device known as a current assisted photonic demodulator (CAPD) which like the gate-based demodulator uses the voltage difference applied between two substrate nodes to alter the electric field in the substrate and steer the majority carrier current in either direction. The first array of 32 × 32 pixels was demonstrated in 2008 [50] followed by another 160 × 120 array with an optimised pixel of 10µm pitch [51]. More recently a QVGA 10µm BSI pixel has been demonstrated by Sony at VLSI 2017 [52].

The third example of PMDs is based on the pinned photodiode with one or more transfer gates to provide multiple taps. Such an approach demands optimisation of the charge transfer gate such that fast charge transfer from the PPD to the floating diffusion (or storage node) can be achieved with minimal lag. An early demonstration of such a pixel at a 256 × 256 resolution and 7.5µm pitch for FLIM applications was presented by Shoji Kawahito group in 2009 based on a single tap two-stage charge transfer design [53]. A 3D imaging 80 × 60 2-tap 10µm pitch sensor was later presented by Fondazione Bruno Kessler (FBK) at ISSCC 2010 [54].

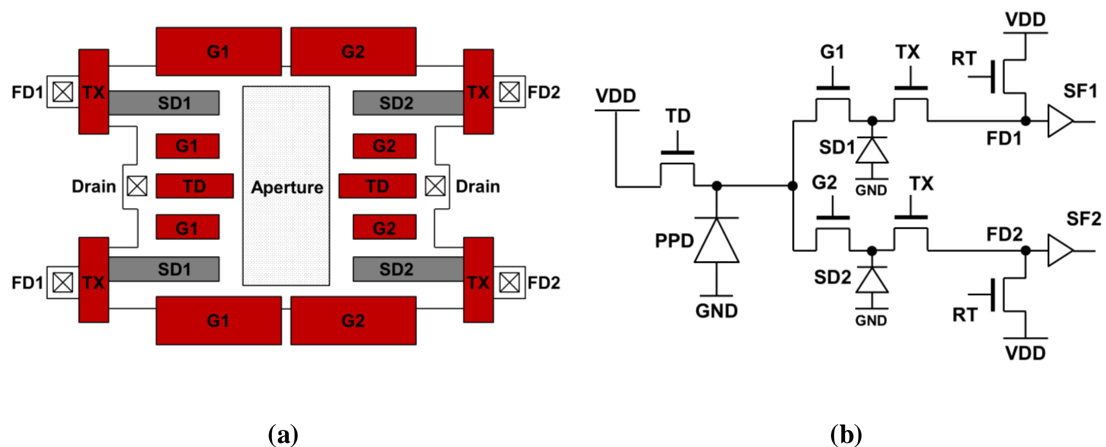


Figure 1.5.2. Two-tap pinned photodiode pixel presented in [55] for fluorescence lifetime imaging. (a) Pixel layout. (b) Schematic diagram.

The Kawahito group presented a variety of pixels based on this technique targeting real time FLIM applications including a symmetric 2-tap design with true correlated double sampling (CDS) achieving a temporal resolution of 10.8ps [55] and a 4-tap design with in-pixel time gate generation logic achieving a time gate width of 0.9ns [56]. The reported intrinsic temporal response of these devices is in the order of few hundred picoseconds at different wavelengths. Figure 1.5.2 shows the layout and schematic of the 2-tap 11.2µm × 5.6µm pixel in [55].

A second class of ToF pixels is the standard photodiode paired with dedicated processing circuitry. This approach was first demonstrated in 2001 [57] for automotive applications where a linear array of

photodiodes is connected to switched capacitor circuits to detect the arrival of the reflected light pulse within two short integration windows. A 50 × 30 imaging array with 82μm pixel pitch was presented later in 2005 [58] demonstrating 3D imaging up to 8m range. This technique will not be discussed further as pixels tend to be large due to the complex analogue processing involved which renders them unsuitable for miniaturisation.

As can be seen from the different examples of solid state time-resolved sensors above, all of them rely on IToF or time gating techniques as opposed to DToF or TCSPC. This is mainly due to the charge integration nature of the detectors in addition to the conventional analogue readout architecture of the APS which naturally lend themselves to IToF. Also, the relatively small pixel pitch of such time-gated pixels due to the simple circuit design makes them amenable to large imaging array formats solidifying the position of IToF as the technique of choice.

As mentioned earlier, TCSPC systems historically relied on discrete components that are bulky, expensive and require dedicated fabrication processes when it comes to APD or SPAD receivers and so a low cost integrated solid state solution is highly desirable. In 2000 and 2003 Pauchard et al. and Rochas et al. demonstrated silicon APDs [59] and SPADs [60] respectively fabricated in standard CMOS technologies paving the way for a new class of ToF sensors: SPADs paired with dedicated processing circuitry.

Due to their high avalanche gain resulting in discernible output voltage pulses per detected photon and the picosecond temporal resolution of this output, SPADs are well poised for TCSPC implementations when coupled with CMOS timing elements be it a time to digital converter (TDC) or a time to analogue converter (TAC).

The first CMOS SPAD array coupled to external TDC cards for 3D imaging was presented by Niclass et al. in 2004 [61] followed by a fully integrated 64 pixel TAC-based linear array in 2005 [62] with the first fully integrated 128 × 128 imaging SPAD array employing column parallel TDCs reported in ISSCC 2008 [63]. These early efforts marked the beginning of solid state TCSPC systems. Apart from TCSPC, SPADs can also be used for IToF by means of counting pulses in the digital or analogue domain with the first SPAD IToF sensor employing an analogue counter presented in 2007 [64] in the form of a 64 pixel linear array.

A detailed review of SPAD devices, architectures and implementations in different technology nodes will be provided in Chapter 2, but it is worth noting that since their emergence, SPAD pixels presented some design challenges that would discount them as candidates for miniature time-resolved image sensors:

1. The SPAD device itself requires a unique layout structure to function correctly imposing limitations on the pixel size and the room for other electronic circuits.
2. In TCSPC implementations, the timing circuitry is complicated and tends to consume a large pixel area whether implemented in the analogue or digital domain.
3. While time-gated circuitry is simpler in principle, the miniaturisation of such pixels is also limited by the footprint of logic gates needed for digital counting especially in older technology nodes, while analogue counters do show promise in this respect but suffer from other performance limitations.

Looking at a survey (Figure 1.5.3) of the plethora of CMOS SPAD sensors and pixels published in literature at the start of this project (Sept. 2014) one can see the state of the art in terms of pixel pitch and time-resolved technique adopted. Overall the landscape is mixed with most of the designs achieving a pixel pitch between 25μm and 100μm whether they implement a TDC / TAC or a counter mainly due to the older technology nodes used and the restrictions of the SPAD size and layout.

Two observations can be made: first, there is a visible trend towards miniaturisation enabled by time-gated analogue pixels which follows in the footsteps of other approaches presented earlier with simple circuitry and more advanced CMOS nodes being a factor. An 8μm pixel presented by Dutton et al. at VLSI 2014 [65] represents the state of the art. Second, and also in 2014, two digital photon counting pixels show signs of miniaturisation when compared to previous similar works with a 35μm pixel implemented in 180nm process [66] and a 25μm pixel implemented in a 3D-stacked technology with custom silicon SPADs on the top tier [67].

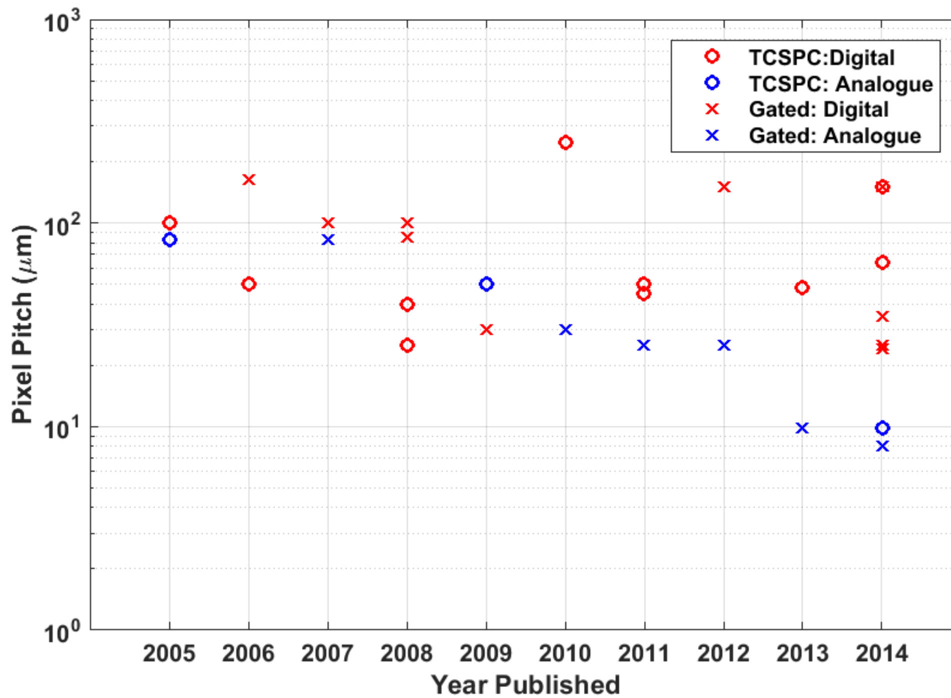


Figure 1.5.3. Survey of state of the art SPAD sensors pixel pitch as of 2014. Pixels are categorised according to time-resolved technique (TCSPC / Gated) and circuit implementation (Digital / Analogue).

The remaining question in the conclusion of this review is why SPADs and no other detectors for a miniature time-resolved sensor? Apart from the natural tendency towards SPADs given the expertise of the CMOS Sensors and Systems (CSS) group at the University of Edinburgh in this field, the following technical justifications are argued.

The first reason relates to timing signals voltage drivers. PMD pixels whether gate-based, PPD-based or CAPD require driving the capacitive load of photo-gates, transfer gates or substrate nodes at high frequencies to well controlled potentials of several volts accounting for significant power consumption by the timing signal drivers.

Similarly SPAD pixels require driving time gates across the array but since the gating can be performed by addressing logic gates instead of device gates or nodes, the capacitive loading is determined by the CMOS node of choice. Thin oxide logic of high density low power nanometre CMOS would amount to smaller switching power figures. Supplying high current loads through few metres of wiring to a miniature endoscopy sensor with limited decoupling capacitance is a serious concern for power supply noise [21].

The second reason relates to the pixel noise performance. All alternative time-resolved pixels rely on an analogue readout chain with a corresponding noise floor limiting their sensitivity. In photon starved biomedical applications such as FLIM high sensitivity is desirable in order to detect weak signals.

SPAD pixels when implemented in the digital domain offer shot noise limited performance and despite their relatively lower photon detection efficiency have been shown to surpass the sensitivity of scientific grade CMOS (sCMOS) in low light conditions and to match that of intensified CCD (ICCD) [68].

The third reason relates to the sensor area overhead. Since all technologies based on charge domain circuits will require analogue to digital conversion, the area and power overhead of the column parallel ADC architecture will be a challenge for miniature sensors. The single ADC channel with multiplexed readout [17] or the single channel multiplexed analogue readout with external ADC [18] configurations of commercially available endoscopy cameras will not suffice as time-resolved imaging requires high speed imaging performance.

Conventional single slope ADCs are more area efficient than other architectures but are limited in speed. Successive approximation or cyclic ADCs are high speed architectures but consume more area and power and so if an analogue pixel is pursued, an optimised area and speed efficient ADC is needed. An area efficient column shared pipelined ADC was proposed by [69] for time-resolved imaging based on PPD yet the ADC block area is significantly larger than the 256×128 pixel array.

Alternatively digital SPAD pixels provide a readily digitised output with no additional area overheads or timing constraints. Thus any area savings due to the absence of ADC circuitry can be utilised for basic on-chip processing and data management. While the survey above suggests that digital SPAD pixels are far from miniaturisation suitability, when implemented in advanced nanometre CMOS technologies; a common theme in this work, new opportunities are possible.

A final argument in favour of SPADs is their intrinsic high temporal resolution commonly between 100-200ps for CMOS implementations with reported figures as low as 27ps [70]. This coupled with digital time gating becomes a powerful tool for observing fast temporal phenomena [71] which is otherwise obscured by the charge transfer speed of other time-resolved pixels necessitating careful process and device optimisation.

1.6. Research Aims

Miniaturisation of SPAD pixels, whether for the purpose of small form factor or high resolution image sensors, is the main but not the only aim of this research project. Maintaining time-resolved capability, high sensitivity or fill factor, high dynamic range and moderate sensor frame rates are also core objectives. Unfortunately, and as is usually the case in engineering, most of these desired properties do not come hand in hand with trade-offs being necessary.

The starting point for this work was the two smallest SPAD pixels reported to date, both at 8 μm pitch and implemented in a front side illuminated (FSI) 130nm imaging technology. The first is an analogue time-gated photon counting pixel [65] and the second is an analogue TAC pixel [72] conceived by

colleagues Drs. Neale Dutton and Luca Parmesan during their PhD research at the CSS group led by Professor Robert Henderson.

While the survey in Figure 1.5.3 above suggests analogue implementations are more suitable for miniaturisation, such pixels do suffer from poor uniformity, readout noise, accumulation noise due to switching behaviour and limited full well (FW) or photon counting capacity [73][74]. Even in advanced CMOS technology nodes thick oxide transistors - normally preferred for analogue design due to better matching, lower leakage current and higher voltage swing - do not scale accordingly offering marginal area advantage.

Up until 2014 all CMOS SPAD sensors were exclusively implemented in FSI technologies with pixel trials in 3D-stacked BSI technologies starting to appear [75]. This meant that the SPAD and circuitry shared the same pixel area and a direct trade-off existed between fill factor and functionality. Optimised circuits and sharing layout techniques [76] are the typical solutions adopted by designers.

As the pixel gets smaller, less room is available for integrated functionality and regardless if a digital or an analogue approach is pursued the photon counting capacity per pixel is limited, thus limiting the dynamic range (DR). If a typical DR of 60dB is to be achieved then a digital counter of 10-bits is needed. Such a pixel was reported in 180nm process achieving 60dB in a 35μm pitch with a limited fill factor of 14.4% [66]. On the other hand, the 26.8% fill factor analogue 8μm pixel in [65] reports a FW of only few hundred photons. A solution to the limited capacity dilemma is oversampling, with the single-bit quanta image sensor (QIS) [77][78] approach being a specific case but this is a discussion reserved for Chapter 4.

Moreover, the effective sensor frame rate is directly impacted by the pixel design as a function of the number of bits per pixel and the oversampling ratio. Traditionally output interface parallelism is the solution with the 10-bit TCSPC pixel, 32×32 sensor reported in [79] achieving 500kfps with 64 output lines and the single-bit time-gated pixel, 512×128 sensor in [80] operating at 156kfps with 128 output lines. Yet for a miniature sensor design, parallelism is an unaffordable luxury as shown in Section 1.2 which adds to the complexity of realising such a sensor.

Finally, the choice between TCSPC and time-gated operation significantly influences the attainable frame rate given a fixed readout scheme. Assuming a 100×100 sensor resolution and a target of 1k photons per pixel for extracting reliable time-resolved measurements, a TCSPC pixel where each photon stamp is represented as a 10-bit code would result in a data rate of 100M-bits per time-resolved frame. Alternatively, a time-gated pixel with a 9-bit counter and two sequential gate settings would result in a data rate of 180k-bits per time-resolved frame. Therefore TCSPC presents a bottleneck in terms of data handling and so a time-gated approach is more favourable for miniature time-resolved sensors.

The research undertaken herein demonstrates that with the aid of advanced CMOS technologies such as a 40nm node and 3D-stacking, it is possible to realise miniature time-gated SPAD image sensors with digital pixels overcoming the noise limitations of analogue counterparts and achieving high fill factor and large counter bit depth. It also demonstrates techniques for improving dynamic range by multi-exposure oversampling while compressing data rates or by on-chip frame processing which is more suited for video rate operation given a single output channel.

1.7. Contributions to Knowledge

Despite the staggering developments over the past 15 years, the field of CMOS SPAD image sensors is still in its early days with many exciting opportunities ahead, and while the efforts described in this thesis are only incremental, the author is very proud and pleased to have had the chance to make his contributions hoping that they would inspire more informed and innovative designs in the future.

Five key strands of work pursued throughout this research are summarised below as the author's contributions to knowledge.

1. Expanded layout design techniques in advanced nanometre CMOS technologies utilising global well sharing of SPAD devices to create miniature high fill factor imaging arrays. A feasibility study of such design and its scalability limitations based on process node parameters is presented alongside a demonstrator IC with an 8.25µm pixel and up to 66% fill factor in a 40nm process [2].
2. Proposed an improved time-gated digital pixel architecture utilising a rising-edge to rising-edge technique enabling sub-nanosecond time gate widths with edge sensitive triggering and lossless photon transition between contiguous observation windows. A demonstrator pixel array is measured achieving a time gate full width half maximum (FWHM) as short as 360ps [2].
3. Demonstrated techniques to enhance the dynamic range of oversampled quanta image sensors through multi-exposure shutters and proposed a novel pixel architecture allowing for in-pixel data rate compression. A demonstrator IC is presented achieving a high dynamic range (HDR) in excess of 100dB with 3.75× data rate compression over conventional QIS implementations [4].
4. Developed a chip architecture enabling miniature high dynamic range endoscopy image sensors. A 1.4mm × 1.4mm demonstrator IC of a fully integrated 3D-stacked SPAD image sensor featuring a 7.83µm time-gated digital pixel is presented. This is the first autonomous 5-wire interface SPAD system on chip with the novel digital architecture allowing for extending the dynamic range by means of on-chip noiseless frame summation and for mediating data transfer achieving video rate operation.

5. Developed a novel configurable chip architecture enabling simultaneous oversampled time-gated image sensor operation with on-demand region of interest (ROI) TCSPC on-chip histogram generation. A 3D-stacked SPAD image sensor design featuring the state of the art 6.48μm digital pixel with dual-mode diode operation (SPAD / photodiode) for extending the dynamic range is presented.

1.8. Thesis Outline

Chapter 2 of this thesis will present a thorough literature review of SPADs, their operation principles and structures in CMOS implementations. The different SPAD sensor architectures will be discussed and the advantages and disadvantages of TCSPC / time-gated and analogue / digital pixels will be assessed in light of miniaturisation. An overview of 3D-stacking technologies for CISs and 3D-stacked SPAD sensors will also be presented.

Chapter 3 will discuss an approach of designing high fill factor miniature sensors in advanced DSM nodes by expanding the SPAD array well sharing layout technique. A 96 × 40 sensor fabricated in a 40nm process will be presented alongside design methodology and characterisation results revealing the scalability and non-uniformity limitations of such a sensor.

Chapter 4 will introduce the quanta image sensor concept which employs a simple 1-bit time-gated pixel that allows for high fill factor designs. Techniques to increase the dynamic range of such small capacity sensors by multi-exposure oversampling will be demonstrated by using the reconfigurable pixel of the sensor described in Chapter 3 which also provides built in data compression capability. The applicability of this binary oversampled imaging paradigm to miniature sensors will be evaluated.

Chapter 5 will discuss a scalable 3D-stacked BSI sensor as an alternative to the planar FSI design described in Chapter 3. SPAD and sensor characterisation results will be presented with focus on the spectral response in comparison to other FSI and 3D-stacked sensors. The implications of the SPAD performance on target applications and the opportunities enabled by 3D-stacking technology will be highlighted.

Taking advantage of the state of the art 3D-stacking technology explored in Chapter 5, Chapter 6 will introduce two novel sensor architectures to realise miniature time-resolved SPAD image sensors. The first architecture is a fully integrated 5-wire system on chip with on-chip frame processing and data management and the second is a highly configurable time-gated array with region of interest TCSPC histogram generation capability. Complete design accounts of both designs will be provided.

Finally, and given the most recent technological developments, Chapter 7 will present the conclusions of this research and will give an outlook to the future of highly integrated intelligent SPAD image sensors.

2. CMOS SPAD Devices, Sensors and Technologies

This chapter presents a literature review of CMOS SPAD devices, sensors and different technology implementations since the realisation of the first SPAD device in standard CMOS in 2003 [60]. It also provides an outlook towards miniaturisation in the light of published works.

The chapter is divided into three main sections. Section one gives a brief historical account of avalanche photo-detectors in the time preceding their realisation in CMOS and introduces the basic operational principles of the SPAD. A review of the different SPAD structures, figures of merit, quenching techniques and biasing topologies is included.

Section two focuses on the different SPAD sensor architectures and their classification with respect to the pixel configuration. Design trade-offs and techniques are discussed. Section three turns attention to the advanced technology step of 3D-stacking and reviews the various implementation strategies adopted for SPAD designs. Finally a conclusion on the overall direction of SPAD development is drawn.

2.1. The Single Photon Avalanche Diode

In the early 1960's, research at Shockley Laboratory into the avalanche multiplication of p-n junctions led to the development of modern day SPAD devices. Although detection of single photons was observed, the main purpose behind the research was to investigate the physics of such devices, and for that a uniform planar avalanche junction was required. To achieve that, a guard ring structure by means of a low doped n-type implant surrounding an n+ into p-type (substrate) junction was proposed by Goetzberger et al. in 1963 [81] which then became one of the most adopted SPAD structures up to this date.

Shortly after, R. H. Haitz was amongst the first to provide an insight into the working principle of avalanche p-n junctions with his work in 1965 explaining the noise mechanisms in such devices including secondary effects such as afterpulsing [82]. This was followed by thorough investigations into avalanche photo-detectors by McIntyre [83] and Webb [84] in the 1970's forming the basis of avalanche photodiodes (APDs). It is worth noting that a SPAD is a specific type of APD as shall be explained in the coming section.

In the early 1980's, Sergio Cova realised the opportunity of using SPAD devices for picosecond resolution photon counting as a solid state replacement for bulky and expensive photomultiplier tubes (PMTs) and micro-channel plates (MCPs) and proposed the concept of active quench circuits to improve SPAD performance [85]. This was followed by decades of device research and development by the Politecnico di Milano group resulting in a variety of SPAD structures. Cova et al. and Ghioni et al. provide an excellent review of this technology in [86] and [87] respectively.

While the above historical snippet is representative, it is difficult to credit all the research efforts that contributed to the field of avalanche photo-detectors. For example in 1985, and in parallel to the work by Cova, McIntyre had also proposed using the previously developed APDs as SPAD devices [88]. Therefore, the author would advise the interested reader to visit two comprehensive references by ex-University of Edinburgh colleagues Drs. Eric Webster [89] and Ed Fisher [90] for more details.

Up until 2002, all SPAD devices were implemented in custom processes which allowed the freedom of optimising doping levels, implant materials and fabrication steps and so the SPADs had to be coupled to external electronic circuitry. This along with the unsuitability of CMOS processes for the requirements of SPAD structures prevented the monolithic integration of large scale detector arrays and processing electronics limiting SPADs to the form of discrete components.

This all changed in 2003 when Rochas et al. presented the first SPAD device in a standard high voltage 0.8μm process [60], launching a new era of highly integrated SPAD sensors for a variety of photon counting and time-resolved applications. That spurred a lot of research by several groups with the developments of the first decade (2003-2013) well captured by Prof. Charbon's review in [91].

As the research carried in this thesis is focused on CMOS design, the following sections will only consider CMOS SPAD devices and sensors.

2.1.1. SPAD Operation Principles

A reverse biased photodiode has three regions of operation: integration (linear), avalanche (proportional) and Geiger (SPAD) as illustrated in Figure 2.1.1. Photodiodes implemented in mainstream image sensors operate in the integration region whereby incident light causes a corresponding amount of electron-hole pairs to be generated in the depletion region contributing to the reverse current. In this case the current is linearly related to the intensity of incident light.

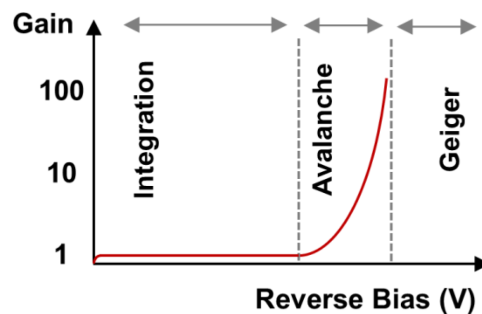


Figure 2.1.1. Photodiode gain versus reverse bias voltage. Different regions of operation labelled.

In the avalanche region on the other hand, the junction is biased near to but below its breakdown voltage V_{BD}. In this mode a single photo-generated electron-hole pair would result in M other

electron-hole pairs being generated where M is a gain factor. This is due to a process known as impact ionisation where the initial electron-hole pair is accelerated by the high electric field of the junction and creates other charge pairs as it collides with the crystal lattice, hence the term avalanche multiplication.

The output avalanche current represents a linear but amplified version of the input incident light intensity with the gain being controlled by the reverse bias voltage. Due to the statistical fluctuation in the gain value such devices are typically limited to gains of few 10s to few 100s in order not to deteriorate the signal to noise ratio. The avalanche process is not self-sustaining and once the light source is switched off the avalanche is halted. Photo-detectors operating in this mode are known as APDs.

Alternatively, a reversed biased p-n junction beyond its V_{BD} (Geiger region) utilises the avalanche multiplication differently. Here the electric field is so intense such that a single detected photon causes a large self-sustaining avalanche multiplication current (very high gain). Due to the internal resistance of the device and the quench element resistance, this large current causes the reverse bias to drop down to the breakdown voltage but unlike an APD, no further photons can be detected unless the device is recovered by means of a recharge circuit.

The resulting output of a photo-detector operating in this region is a distinctive pulse in current (or voltage) representing the detection of a single photon and hence the name Geiger mode APD (GM-APD) or single photon avalanche diode (SPAD). References [92] and [93] provide a detailed description of the physics and principles of operation of APD and SPAD devices.

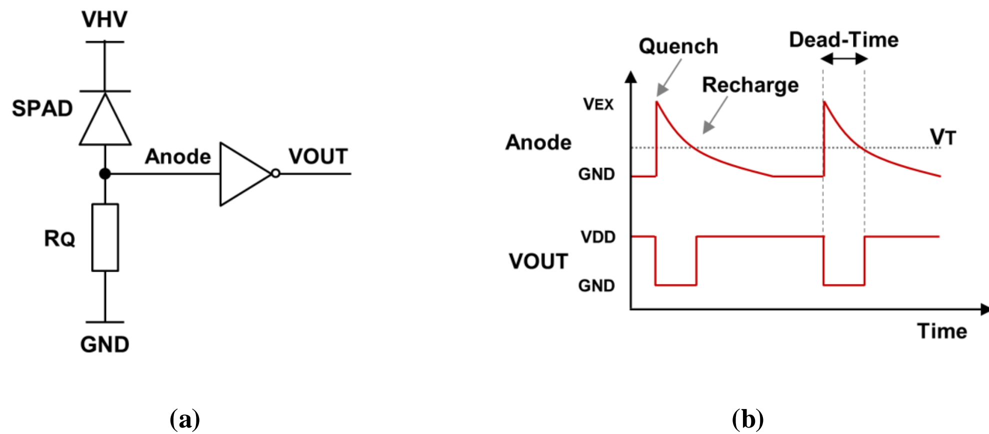


Figure 2.1.2. Simple passive quench and recharge SPAD circuit. (a) Schematic. (b) Waveforms.

Figure 2.1.2(a) shows a simple SPAD circuit with a passive quenching element (R_Q) followed by an inverter logic with threshold V_T. Initially the SPAD is biased to a potential beyond its breakdown voltage known as excess bias or V_{EX} where the high voltage supply VHV is equal to $V_{BD} + V_{EX}$. The

device sits idle in a quasi-stable state until triggered by an event where it undergoes the following transients [91]:

- Seeding: the generation of an electron-hole pair in the depletion region by an absorbed photon. The occurrence of an avalanche breakdown due to impact ionisation at this point depends on the avalanche triggering probability of the device.
- Build-up: upon the triggering of an avalanche, the positive feedback from avalanche multiplication causes the local current density to increase. The flow of this current through the finite charge-space resistance causes the local potential to drop to V_{BD} by negative feedback. This process is microscopic and the current is not observable externally.
- Spreading: the avalanche breakdown spreads across the device junction from the seeding point to the extremities resulting in a macroscopic current observable externally.
- Quenching: the avalanche is stopped or quenched by the rise of potential across the quench element discharging the junction down to V_{BD}. Further incoming photons do not generate a distinctive avalanche at this stage until the device is partially or fully recharged.
- Recharge: the junction is recharged to its initial bias condition above V_{BD} and is ready for further detections. The recharge time is determined by the RC time constant in this example where R is the quench resistance (R_Q) and C is the junction's capacitance.

Figure 2.1.2(b) shows the output waveforms at the moving node (anode in this example) and the digital buffer. The sharp rising edge due to SPAD quenching contains the temporal information of the detected photon. The time taken for the avalanche and recharge is known as the SPAD dead-time and is determined by the quenching circuit configuration. Figure 2.1.3 shows the transient points on the reverse bias I-V curve during a detection cycle.

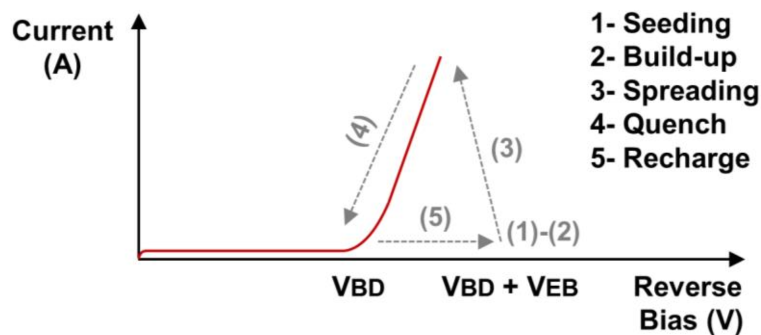
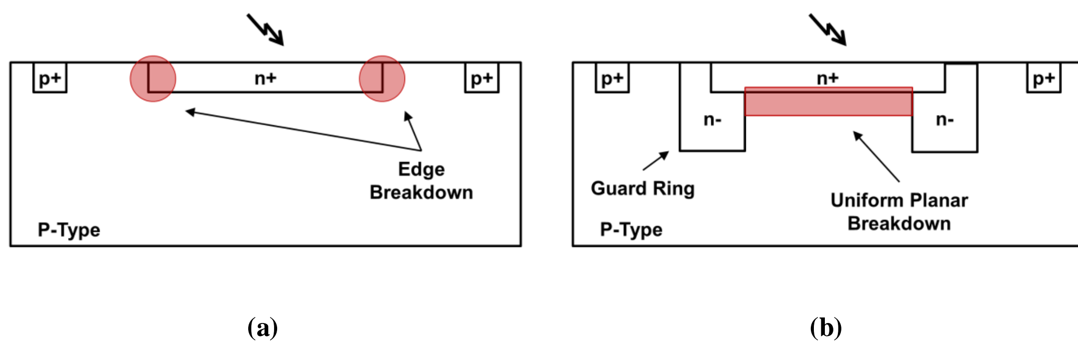


Figure 2.1.3. Five stages of SPAD avalanche cycle.

2.1.2. CMOS SPAD Structures

To ensure a uniform multiplication junction, SPADs require a dedicated physical structure known as the guard ring. In a planar process a p-n junction usually breaks down at the edges (Figure 2.1.4(a)) limiting the multiplication region to small areas and so reducing the efficiency of the device. Moreover, the breakdown near the surface results in high noise due to defects. Alternatively a guard ring such as that proposed by Goetzberger [81] prevents premature edge breakdown (PEB) resulting in a uniform multiplication region and isolates the planar junction (Figure 2.1.4(b)).

The guard ring is formed by a low doped n-type material such that the p-n junction comprised by the higher doped n+ material and the p-type substrate breaks down under reverse bias before the n-type to p-type junction.

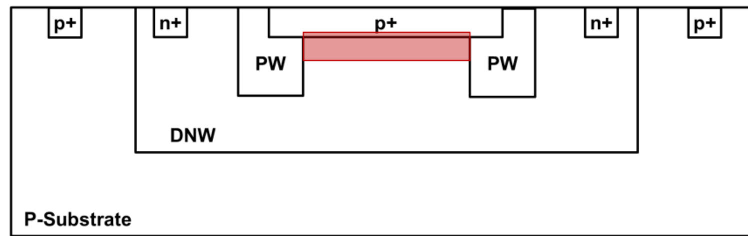


(a) (b)

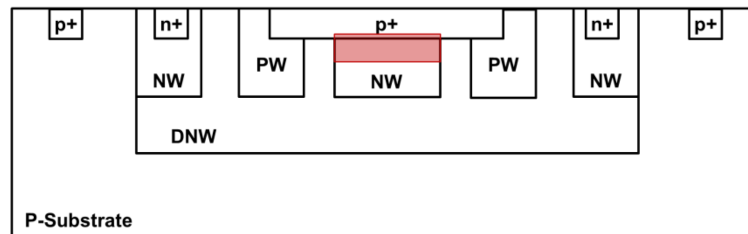
Figure 2.1.4. Illustration of reverse biased p-n junction. (a) Premature edge breakdown. (b) Guard ring structure to ensure uniform planar breakdown.

Many SPAD designs have been presented in the literature and are usually classified based on the guard ring structure. Generally there are two types of guard rings: physical and virtual. Physical guard rings are based on implants or trenches while virtual guard rings are implicit due to the separation between implants of the same material or the absence of doping in specific regions. A summary of the different structures is provided below.

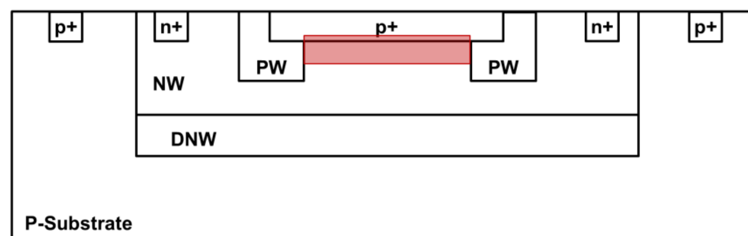
The most common SPAD structure is formed by a p+ diffusion to an n-well (NW) or deep n-well (DNW) junction with a p-well (PW) guard ring (Figure 2.1.5). Such a SPAD was first presented in a 0.8µm high voltage CMOS process [60]. A similar structure was then implemented in a 0.7µm node with the addition of an n-well enrichment implant to increase the junction field through the higher doped NW with respect to the previously used DNW [94]. Similar implementations of the p+ / DNW or p+ / NW junctions were demonstrated in more advanced 0.18µm [95][96] and 0.13µm processes [97][98] all of which are physical guard ring devices.



(a)



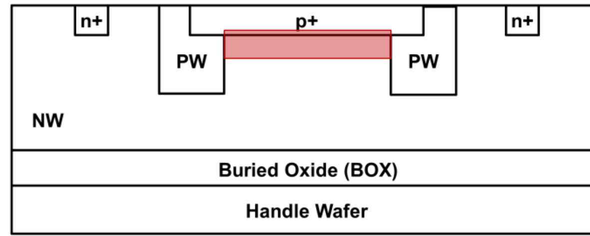
(b)



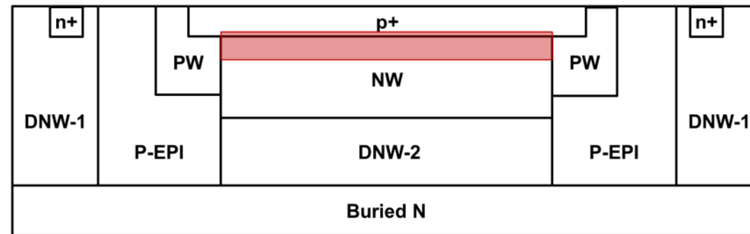
(c)

Figure 2.1.5. Variants of the p+ / DNW or NW SPAD with PW physical guard ring. (a) Rochas et al. [60]. (b) Pancheri et al. [94]. (c) Niclass et al. [97]. Multiplication junction highlighted in red.

Two other notable works utilising the same p+ / NW junction were implemented in a 0.14µm silicon on insulator (SOI) [99] and a 0.18µm [100] processes (Figure 2.1.6). The work in [99] was the first to demonstrate a SPAD in a SOI technology in preparation for backside illumination (BSI) which usually requires the buried oxide (BOX) of an SOI wafer as an etch stop for the backside. The work in [100] utilised a stack of NW implants with gradient doping in order to leverage the carrier diffusion effect for achieving a wide spectral response range.



(a)



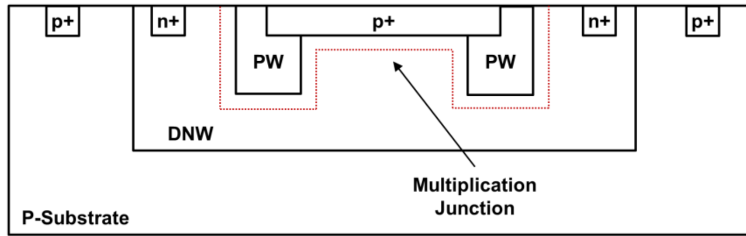
(b)

Figure 2.1.6, Variants of the p+ / NW SPAD with PW physical guard ring. (a) Lee et al. utilising SOI process [99]. (b) Veerappan et al. utilising carrier diffusion [100]. Multiplication junction highlighted in red.

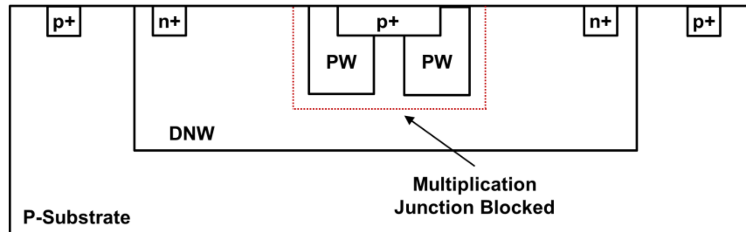
It is worth noting that all of the above structures can be referred to as substrate isolated since the SPAD p-n junction is isolated from the p-type substrate (Psub) by an NW or DNW implant. Alternatively, creating an n+ diffusion into Psub junction results in a non-substrate isolated device (Figure 2.1.4(b)).

A disadvantage of such a structure is its susceptibility to substrate electrical noise but it has the advantage of having the majority of the depletion junction area in p-type material where minority carriers are electrons which are known to have a higher avalanche triggering probability in silicon than holes [101], thus an improved photon detection probability is expected especially for longer wavelengths. Such a SPAD was demonstrated in 0.18µm [102] process and in 90nm [103] and 65nm [104] deep submicron technologies (DSM).

In terms of miniaturisation, the large area required by the implanted guard ring imposes a restriction. Moreover, due to the physical guard ring which exhibits a depletion region once biased, if the multiplication junction is reduced in size too much the depletion regions of the guard ring would merge and cut-off the avalanche junction. Therefore there is a limit to which this structure can scale. This is illustrated in Figure 2.1.7.



(a)

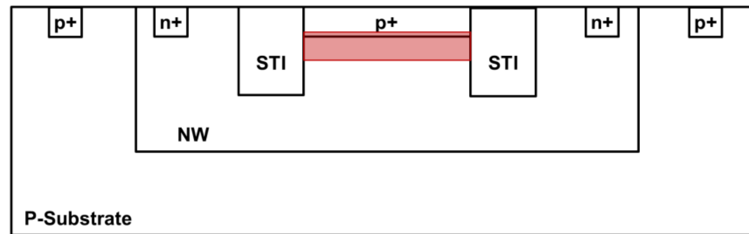


(b)

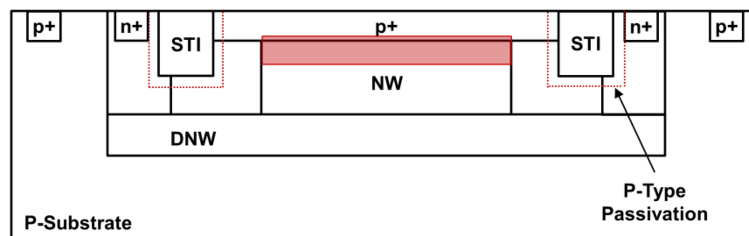
Figure 2.1.7. Illustration of depletion region borders (dashed red) for a SPAD with PW physical guard ring. (a) Standard device with functional multiplication junction. (b) Multiplication junction disappears when depletion borders or guard ring merge as device scales down.

To tackle this issue, Finkelstein et al. proposed using the shallow trench isolation (STI) present in DSM CMOS nodes which is traditionally used to electrically isolate MOS devices as a SPAD guard ring structure in a 0.18μm process [105]. This allows significant reduction of the guard ring area compared to the well-based implant design (Figure 2.1.8(a)).

A major drawback of this design is the high noise level due to interface defects injection into the SPAD's multiplication junction. To mitigate that, a passivation implant in a 0.13μm process was proposed by [106] to reduce the probability of such carriers entering the junction (Figure 2.1.8(b)). Nevertheless, this design is only applicable in processes with passivation options and to SPAD designs with shallow junctions.



(a)

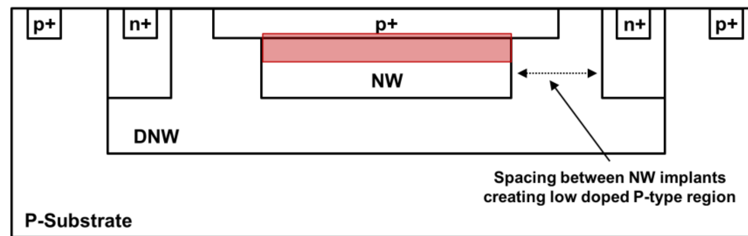


(b)

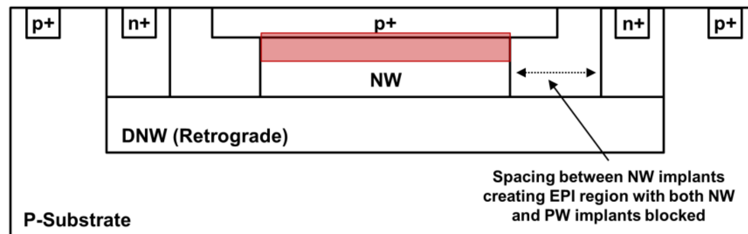
Figure 2.1.8. STI bounded SPADs. (a) Original proposal by Finkelstein et al. [105]. (b) Passivated STI for lower noise contribution proposed by Gersbach et al. [106].

As an alternative to physical guard rings, several virtual guard ring structures were proposed. The first virtual guard ring technique relies on spacing implants of the same material apart such that the separating region results in a low doping profile implicitly forming a guard ring. This was first demonstrated for APDs in CMOS [107] and later for SPAD devices in 0.7μm process by [94] where the separation between two n-wells in a DNW resulted in a lightly doped p-type guard ring. This design was also demonstrated in TSMC's 0.18μm process [108].

The second technique is similar but relies on the retrograde doping of the DNW where the separation between the two n-wells results in p-type epitaxy (EPI) with both PW and NW implants blocked in that region. This was first demonstrated by Henderson et al. in a 0.13μm process [109] and later by [110] in a 0.15μm node. Figure 2.1.9 demonstrates both concepts.



(a)



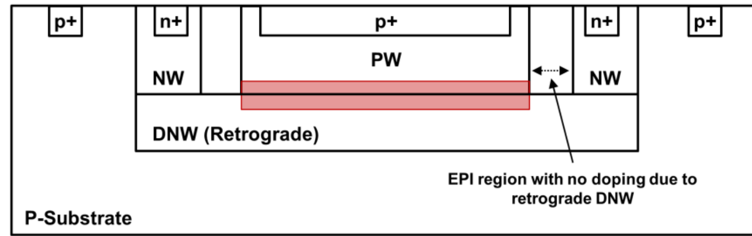
(b)

Figure 2.1.9. Virtual guard ring concept. (a) Low doped p-type between spaced NW implants in a DNW [94]. (b) EPI region with blocked PW or NW doping between spaced NW implants with retrograde DNW [109]. Multiplication junction highlighted in red.

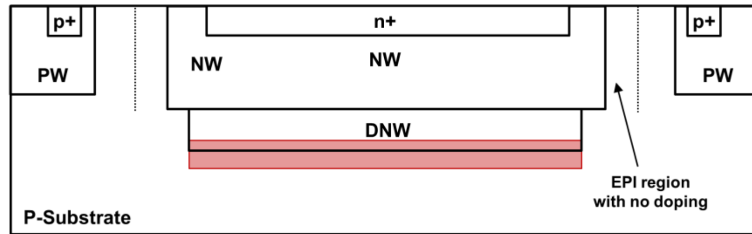
So far all of the presented structures above relied on n+ or p+ diffusion layers resulting in shallow junctions. These junctions have a spectral response geared towards the blue region due to the absorption of shorter wavelengths near the surface. Generally speaking these junctions exhibit high noise due to the high doping of the diffusion implants. Therefore other SPAD structures emerged to tackle these issues.

Based on the retrograde DNW virtual guard ring structure, two SPADs were developed at the University of Edinburgh. The first structure uses a PW to DNW substrate isolated junction providing lower noise and enhanced response at near infra-red (NIR) due to the deeper junction (Figure 2.1.10(a)). The PW / DNW SPAD was demonstrated in 130nm [111] and 90nm [112] CMOS processes. This device was industrialised by STMicroelectronics in 130nm [113] and in an advanced DSM 40nm node [114] used in this work.

The second structure uses the DNW to p-type substrate (Psub) to create an even deeper junction further enhancing the NIR response (Figure 2.1.10(b)). Similarly this device was demonstrated in both 90nm [115] and 130nm [116] technologies. Apart from being non-substrate isolated this structure poses a different challenge related to its biasing topology since the common CMOS substrate is one of its terminals, an issue common to other non-substrate isolated SPADs.



(a)



(b)

Figure 2.1.10. University of Edinburgh virtual guard ring SPADs using EPI with no PW or NW doping. (a) PW / DNW SPAD [111]. (b) DNW / Psub SPAD [116]. Multiplication junction highlighted in red.

As far as miniaturisation is concerned, SPADs with virtual guard rings have great promise with the p+ / NW device reported at 5µm pitch with 2µm active area diameter in 130nm by [117], while the PW / DNW has also been reported at 2µm active area diameter in 130nm [118] and 90nm [112]. More recently the same device was also reported at a remarkable 3µm pitch and 1µm active area diameter in 130nm CMOS at the International Image Sensors Workshop (IISW) 2017 [119]. Therefore, this device is a viable candidate for this work.

2.1.3. Exotic SPAD Implementations

A few other SPAD designs with exotic features or implementations in exotic process nodes are also worth noting. The work by Sun et al. presented a SPAD implemented on a flexible substrate [120]. Such device is attractive for wearable biomedical devices and can be illuminated from both sides. Nevertheless it remains at early stages of development.

In 2015 C. Niclass et al. presented a SPAD device with enhanced NIR response for automotive applications [121]. While implemented in a 0.18µm CMOS, the design incorporated two custom layers to form the SPAD junction. The n-SPAD and p-SPAD layers forming the p-n junction result in a wide deep depletion region. A p+ layer over the n-SPAD layer was used as a pinning layer to reduce surface generated noise. An expanded version of this work was reported in [122].

A very interesting theoretical work by Vignetti et al. in 2016 proposed implementing a SPAD underneath the buried oxide (BOX) of a 28nm FDSOI technology node (Figure 2.1.11) [123]. The PW / DNW device can sit underneath its quenching circuitry which is implemented above the isolating BOX in the PW implant. A potential issue of such a structure is the sensitivity of CMOS circuits to back-gate bias effects especially if the PW is the SPAD's moving node.

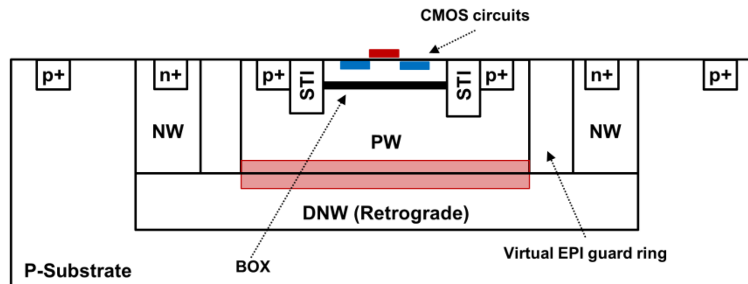


Figure 2.1.11. Concept of SPAD structure in 28nm FDSOI process with CMOS circuits integrated in BOX [123]. Multiplication junction highlighted in red.

In 2017 two SPAD structures were designed and implemented in a 0.16µm bipolar-CMOS-DMOS (BCD) process by Sanzaro et al. [124]. The devices exhibit outstanding noise, spectral response and timing response characteristics. The BCD process intended for high voltage applications is advantageous due to its cleanliness and its modularity which allows for flexibility in implant choice and doping.

Zimmermann et al. presented a circular quad-cell SPAD structure with a thick absorption region for high photon detection probability fabricated in a PIN photodiode 0.35µm CMOS process [125]. The high fill factor device was aimed for near quantum limit fibre optical receivers.

Finally, custom high density and high efficiency silicon photomultiplier SPAD arrays with a cell pitch as small as 5µm have been demonstrated by FBK [126]. The compact arrays employed deep trench isolation to reduce optical and electrical crosstalk between detectors.

2.1.4. SPAD Performance Parameters

The main performance parameters of SPADs are divided into primary and secondary categories. The primary parameters are photon detection probability (PDP), dark count rate (DCR) and temporal response or jitter.

PDP is a measure of the device detection response across the spectra of incident light. It is a product of the SPAD's quantum efficiency (QE) and the SPAD's avalanche triggering probability. The QE is dependent on the internal quantum efficiency (IQE) of silicon in the case of CMOS devices and the

external quantum efficiency (EQE) of the device structure including the metal stack. The avalanche triggering probability on the other hand is dependent on the excess bias V_{EX} above breakdown.

At higher excess bias setting, the electric field in the junction is intensified increasing the avalanche triggering probability. Yet increasing V_{EX} can also have a negative impact on noise level and thus there is an optimal point at which the best signal to noise ratio can be obtained [93][127].

The IQE of the device is dependent on the junction location and the extent of the depletion region with respect to absorption depth of different wavelengths. Therefore many works have investigated extending the depletion region for a wide spectral response [96][102] or specifically for improved NIR response [116][121].

DCR is the rate of spurious counts that are not attributed to photon detections. The main two mechanism of DCR are carrier generation due to thermal or Shockley-Read-Hall (SRH) trap assisted processes and band to band tunnelling (BTBT) [128]. Thermal generation is temperature dependent and can be reduced by cooling the device. BTBT on the other hand is dependent on the electric field in the junction. Hence increasing excess bias would increase noise in tunnelling dominated structures.

It has been observed that some SPAD devices exhibit fluctuations in DCR where the measured value jumps between two low and high states over time. This random telegraph signal (RTS) noise was first reported in [129] and more recently characterised in [130].

BTBT tunnelling in DSM technologies is a common issue given the high doping diffusion implants which result in an intensified electric field. This can be mitigated by utilising lower doping implants (PW instead of p+) [111] or introducing secondary layers to reduce the doping profile (p- over p+) [109]. The impact of lower doping implants needs to be considered as that increases the junction breakdown voltage which can be a challenging problem for on-chip integrated voltage generation circuits.

Careful design of the SPAD active area and guard ring can also reduce DCR. The pinning layer introduced in [122] stops carrier generation centres from the defective surface from reaching the multiplication junction. Similarly the passivation layer used with STI guard rings [106] blocks these carrier generation centres from creating dark events.

SPAD jitter is a measure of the statistical variation from the moment of detecting a photon in the junction to the moment of generating an output pulse. This is measured as an impulse response function (IRF) with respect to laser pulse and usually reported in FWHM. Increasing excess bias improves the statistical nature of avalanche multiplication and so enhances jitter.

The time it takes for a photo-generated carrier created in the device to reach the multiplication junction also influences jitter. If a photon is absorbed in the depletion region, drift times are very small but if a photon is absorbed outside the depletion region, it may diffuse towards the

multiplication junction with a statistically varying time delay. Such effect is observed as a diffusion tail in the measured IRF.

To reduce the diffusion tail, substrate isolated structures are preferred as they block the deeply photo-generated carriers which diffuse towards the relatively shallow junction. This though has a negative impact on PDP at longer wavelengths. For non-substrate isolated SPADs NIR PDP is improved at the cost of jitter diffusion tail. Therefore careful trade-offs should be made when designing a SPAD device.

The secondary performance parameters of a SPAD are afterpulsing and crosstalk. Afterpulsing is similar to DCR as it is a spurious event but it is caused by traps in the junction due to impurities and crystal defects. When a SPAD fires, current flows through the device where a carrier can be trapped in the junction. When the carrier is released it may trigger a secondary firing falsely reporting photon detection.

Afterpulsing can be reduced by a cleaner process, reducing the avalanche current by reducing the junction (i.e. active area) and parasitic capacitances, reducing excess bias or increasing the time the SPAD is off after an initial firing to allow time for these traps to empty. The latter is achieved by active recharge circuits.

Finally, crosstalk can be either electrical or optical. Electrical crosstalk is due to a photon entering through the aperture of one SPAD generating an electron-hole pair that gets detected by a neighbouring device. This is a common issue in CISs and can be resolved by containing the device within isolation implant walls or deep trench isolation (DTI) [131].

Optical crosstalk is more common in avalanche devices as the hot carriers generated during an avalanche can result in a secondary photon emission [132] that could be detected by other pixels. This effect has been reported by several groups for neighbouring devices [133][134] and its dependency of wafer thickness was reported in [135]. Isolation barriers, reducing the charge flow in the device and the spacing of pixels are potential solutions, but this effect is more of concern for miniature arrays when close packing of devices is necessary.

2.1.5. SPAD Devices State of the Art

Table 2.1.1 summarises the performance of some of the most recently published front side illuminated (FSI) SPAD devices to give an indication of the state of the art in terms of various parameters. A thorough review of earlier SPAD works and their figures of merit can be found in [136].

All of the devices are implemented in DSM CMOS nodes which is an expected shift in direction away from older 0.35μm and 0.8μm nodes due to the need for optimising the footprint and performance of the circuitry associated with SPAD pixels and not only focusing on the SPAD device.

SPADs in advanced nodes such as 40nm are emerging with relatively good performance despite the implementation challenges. Of course this requires process optimisation at foundry level, a task suited for industry players such as STMicroelectronics. Toyota which is pursuing SPADs for automotive LIDAR also presented a customised device in 180nm.

As it stands, there is no clear cut winner between physical and virtual guard ring designs but as discussed earlier, as SPAD arrays grow in size for mega-pixel arrays, or when compactness is a key requirement, virtual designs have been shown to scale favourably.

Junction breakdown voltages are reported between 11V and 25V which are not excessively high and can be supplied by on-chip charge pumps [137], a requirement for fully integrated consumer market products [113].

Overall the DCR levels are low ranging from 10's to 100's of counts per second at room temperatures. PDP of most devices is skewed towards the blue region of the spectrum, a response expected for FSI SPADs with relative shallow junctions. In the cases where the device was optimised for NIR response by a deeper junction design PDP peaks in the red region of the spectrum [122]. Peak PDP values at CMOS compatible bias levels are generally between 30% and 40%.

Jitter figures roughly average to ~150ps FWHM with numbers as low as 32ps reported at 820nm. Such performance is acceptable especially that SPAD sensors do not rely on single shot measurements but average over a large number of detections. Afterpulsing probabilities are very low and are less than 1% for all devices at different excess bias voltages and active areas suggesting that the used processes are clean and reliable.

One particular issue relating the cumulative distribution of DCR across SPAD arrays is yet to be solved. Generally publications report a population split of approximately 80% to 20% of low DCR pixels and high DCR pixels (screamers) respectively. This means that in an image sensor ~20% of the pixels are unusable, at least under low light levels, and will appear as random grain noise deteriorating image quality. This problem has to be addressed at foundry level to identify and reduce the source of defects responsible for the screamer pixels.

CMOS SPAD devices have come a long way to meet the different performance demands solidifying their role in time-resolved and photon counting applications. Miniaturisation of SPAD devices and pixels is expected to inherit the same performance with potentially lower DCR and jitter due to smaller device size.

Ref.	[96]	[99][138]	[100]	[121][122]	[139]	[130]	[124]	[113]	[114]
First Author	Veerappan et al.	Lee et al.	Veerappan et al.	Takai et al.	Veerappan et al.	Xu et al.	Sanzaro et al.	Pellegrini et al.	Pellegrini et al.
Institution	TUDelft	TUDelft	TUDelft	Toyota	TUDelft	FBK / Uni. of Trento	Poli. de Milano	STMicroelectronics	STMicroelectronics
Year	2014	2015	2015	2016	2016	2017	2017	2017	2017
Process (nm)	180	140 SOI	180	180	180	150	160 BCD	130 CIS	40
Junction Type	p+ / DNW	p+ / NW	p+ / NW	n-SPAD / p-SPAD	PW-EPI / Burried NW	p+ / NW	PW / DNW	PW / DNW	PW / DNW
Guard Ring Type	PW Implant	PW Implant	PW Implant	Virtual	Virtual	Virtual	Virtual	Virtual	Virtual
Active Area (μm²)	113.1	113.1	113.1	207	113.1	97.12	706.8	50.3	n/a
Breakdown Voltage (V)	23.5	11.3	14.64	20.5	25.46	18.01	25.44	14.2	14.6 (@ 60°C)
Peak PDP (%)	29	25.4	42	62.2	33	31	64	43	33
PDP Wavelength (nm)	480	490	480	610	500	450	480	480	500
PDP @ Excess Bias (V)	3	3	3	5	3	3	5	1.9	1.9
DCR (cps)	22	27.6k	500	<100	25	39	100	100	50
DCR @ Excess Bias (V)	3	3	3	5	3	3	5	1.9	1.9
Jitter FWHM (ps)	90	65	150	161	133	52	32	100	170
Jitter @ Excess Bias (V)	2	3	3	5	3	3	5	3	1.9
Jitter Wavelength (nm)	405	405	637	635	405	468	820	n/a	840
Afterpulsing (%)	0.03	0.1	0.2	0.35	0.08	0.85	<1	0.08	0.1
Afterpulsing @ Excess Bias (V)	2	3	4	5	4	3	5	n/a	1.9

Table 2.1.1. SPAD performance summary for some of the recent published FSI SPAD devices

2.1.6. Quench and Recharge Circuits

There are four different combinations of quench and recharge circuits classified by the circuit operation mechanism:

1. Passive quench passive recharge (PQPR).
2. Passive quench active recharge (PQAR).
3. Active quench passive recharge (AQPR).
4. Active quench active recharge (AQAR).

PQPR rely on a passive element to quench or discharge the SPAD once it avalanches and then to recharge it with an RC delay due to the combination of the quench element resistance and the SPAD's junction capacitance. PQAR circuits on the other hand use a passive element to discharge the SPAD and then begin to passively recharge it. After a delay period an active recharge element is introduced to assist in the recharging process and to quickly recover the detector to its initial condition.

AQPR circuits sense the onset of an avalanche and actively interfere by quenching the SPAD device by forcing the moving node to a voltage value such that the potential across the SPAD is below its breakdown. The SPAD is then recharged back to its initial condition via a passive element. Alternatively, AQAR circuits quench the SPAD in the same way but then introduce a delay also known as hold-off time (τ_{OFF}) after which the SPAD is actively recharged.

For circuits with active elements the SPAD dead-time is determined by the sum of quench, hold-off and recharge times based on the circuit configuration. Figure 2.1.12 conceptually illustrates the two end cases of PQPR and AQAR circuits.

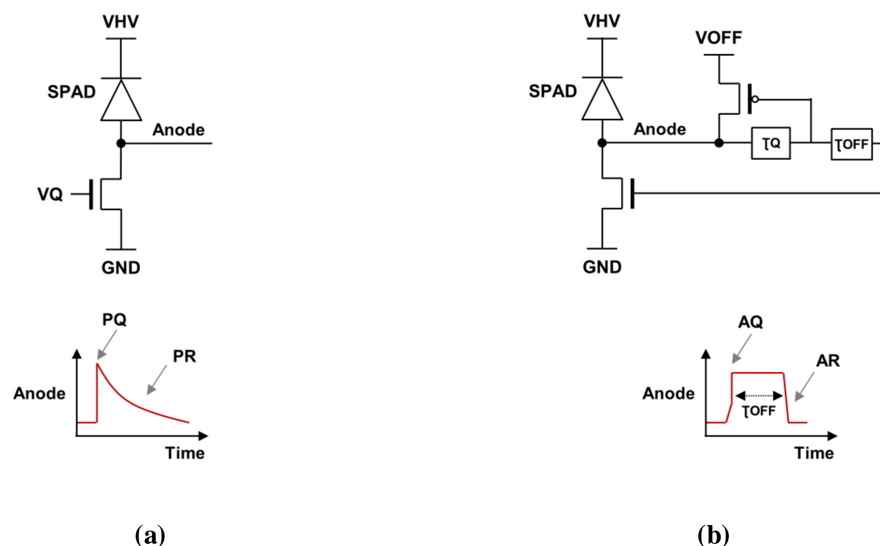


Figure 2.1.12. Conceptual schematics of different quench and recharge configurations. (a) Passive quench passive recharge (PQPR). (b) Active quench active recharge (AQAR).

Active circuits allow for more control over the SPAD behaviour and so are used for two purposes: increasing the intrinsic SPAD dynamic range by reducing the dead-time and so increasing the maximum achievable count rate [140] or reducing after-pulsing effects by allowing the junction enough time to release any trapped charge by extending the hold-off time (i.e. dead-time) of the device [141]. The latter is of particular importance for large area devices that exhibit a large charge flow. These two design objectives do contradict each other.

An advantage of active recharge (AR) circuits over passive recharge (PR) is the elimination of count rate ambiguity at high illumination levels. As shown in Figure 2.1.13, the count rate response versus illumination of a PR circuit would reach a saturation level beyond which the SPAD undergoes paralysis and results in lower count rate at higher illumination. This creates ambiguity as the same rate can be attributed to two illumination levels. Alternatively, AR circuits do not undergo paralysis and saturate at their maximum count rate. The maximum count rates of a PR and AR circuit are [142]:

$$PR_{MaxCountRate} = \frac{1}{e \times \tau_{DeadTime}}$$

$$AR_{MaxCountRate} = \frac{1}{\tau_{DeadTime}} \quad (1)$$

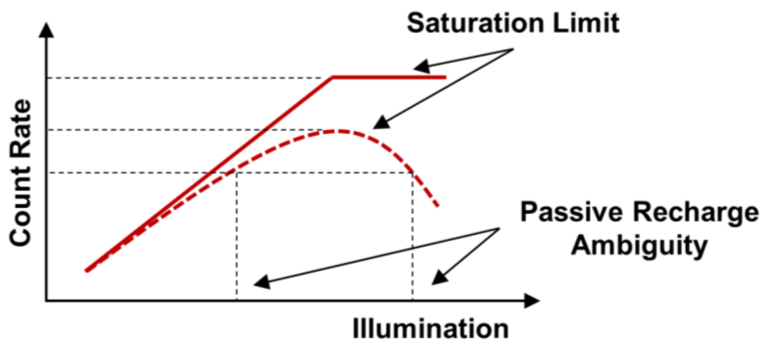


Figure 2.1.13. Count rate versus illumination level SPAD response for passive recharge configuration (dashed) showing ambiguity due to paralysis and active recharge (solid) configuration showing fixed saturation limit.

Several active quench or recharge circuits have been presented some of which are cumbersome and rely on discrete components such as a comparator IC and an FPGA [143] while others are compact CMOS integrated PQAR [144][145] or AQAR [140][146] solutions. An informative review of the evolution of such circuits is provided in reference [147]. Nevertheless, from a miniaturisation

perspective the simple PQPR approach is preferred due to its simplicity and to avoid an additional number of transistors that would compete with the SPAD or time-resolved circuitry for area.

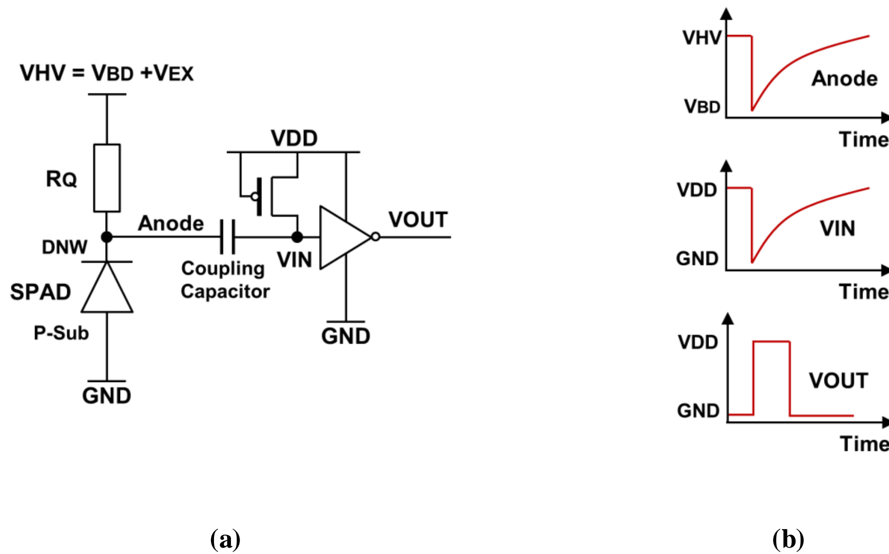
Moreover, as SPAD pixels become smaller the junction capacitance and the parasitics added to the moving node decrease and the noise level decreases allowing the SPAD to reach high intrinsic dynamic range. As for afterpulsing, the data presented in Table 2.1.1 shows that it is very low even in advanced CMOS nodes rendering it of minimal concern and so active measures of reducing it are not necessary.

PQPR can be implemented using a poly-silicon resistor as in [148]. The disadvantage of this configuration is the area needed to realise a typical quench resistance of few 100kΩs. A more efficient approach is to use a PMOS [111] or an NMOS [149] device in linear mode. This simple circuit also has the advantage of controlling the channel resistance by adjusting the MOS gate bias and so control the dead-time.

2.1.7. Biasing Topologies

Since the SPAD is a two terminal p-n junction, one of its nodes would act as the moving node reacting to it avalanching and recovering while the other would act as a bias point to set the device in Geiger mode. There are multiple biasing topologies that can be adapted but are mainly influenced by the chosen polarity of the high voltage (VHV) drive whether negative or positive. This in turn is influenced by the SPAD structure whether it is substrate isolated or not.

For instance and for the typical non-substrate isolated DNW / Psub SPAD in Figure 2.1.10(b), the moving node cannot be the substrate terminal as this is common to other SPAD and circuit devices therefore the DNW terminal is the moving one connected to the quench and recharge circuit.



(a) (b)

Figure 2.1.14. Biasing topology for non-substrate isolated DNW / Psub SPAD with poly-resistor R_Q and coupling capacitor. (a) Circuit diagram. (b) Waveforms.

Biasing the substrate at a constant negative potential to set the SPAD in Geiger mode is also not feasible since this will impact any active circuitry sharing it unless they are isolated in their own wells. Even then care should be taken such that the junction formed by the p-type substrate and the isolating NW does not breakdown. The safest option is to hold the substrate at ground level.

By doing so the required reverse bias potential across the SPAD can only be achieved by applying a positive VHV at the quench element terminal that is not connected to the SPAD's moving node. Any SPAD pulses will result in voltages fluctuating between VHV and VHV – V_{EX}. This is an issue for CMOS compatibility since the quench element cannot be a MOS device with more than 3V to 5V across any of its terminals and the output pulse cannot be interfaced to 3V to 5V logic. To overcome these limitations a passive resistor and a coupling capacitor need to be used along with a pull up leakage PMOS device as in Figure 2.1.14.

For the PW / DNW substrate isolated SPAD in Figure 2.1.10(a) on the other hand either terminal can act as the moving node without interfering with the common substrate. If the PW terminal plays that role then the SPAD bias can be achieved by applying a positive VHV at the DNW terminal while a quench element (resistor or MOS) can be connected to the PW node with the SPAD pulse voltage levels being CMOS compatible.

An added advantage of this scheme is that it allows for well sharing layout techniques first proposed by [94] which is particularly beneficial for closely packing pixels, a requirement for miniaturisation. Also, the option of using a MOS quench is more appealing than a poly-resistor and a coupling capacitor from the point of view of area efficiency although the passive element can be implemented over the SPAD itself at the cost of loss in photon detection efficiency in FSI technologies [150].

In conclusion, SPAD structures that facilitate well sharing for compactness, simple quench element design and avoid the complications of negative drive voltage are preferred for miniaturisation.

2.2. SPAD Sensor Architectures

A variety of SPAD sensors have been presented over the years for photon counting and time-resolved applications with different architectures proposed. In general, SPAD sensors can be categorised into four classes based on their pixel arrangement: single point sensors also known as digital silicon photomultipliers (dSiPMs), line sensors, image sensors and a fourth class which borrows elements from other architectures that will be referred to here as modular sensors.

2.2.1. Single Point Sensors

As the name suggests this class of sensors operates as a single pixel performing single point measurements. In practice the single pixel is a group of individual SPADs with their digital pulses combined together into a single channel output thus losing the spatial resolution. Analogous to the analogue photomultiplier, a group of such SPADs is known as a digital silicon photomultiplier (dSiPM). Figure 2.2.1 shows the single point sensor architecture typically utilised for TCSPC measurements.

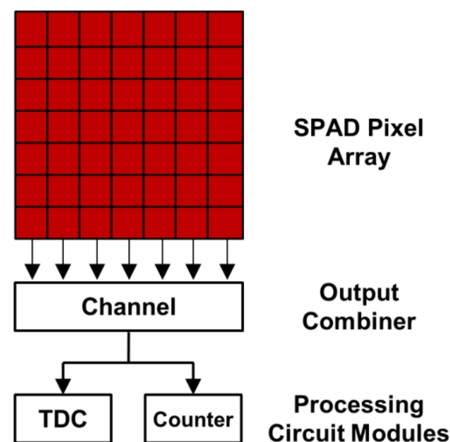


Figure 2.2.1. Single point sensor architecture.

Due to the compromise on spatial resolution where the single channel output can feed external counting or time stamping modules or modules integrated at the edge of the array, SPAD pixels in dSiPMs tend to be minimalistic in circuit design with the SPAD occupying most of the pixel area yielding very high fill factors (50% or higher).

Because of the SPAD dead-time, and given a stream of arriving photons over an active area, a single device would be able to detect the first photon but would miss others arriving within this dead period. This is known as detector pile-up. To overcome this limitation, the active area is divided into multiple SPADs instead such that more devices are available for detection. This also has the advantage of being able to turn off noisy detectors while others remain active as opposed to having a single noisy device. The optimisation of dSiPM performance is beyond the scope of this work and the reader is directed to [151] for detailed analysis of detection efficiency and dynamic range in dSiPM designs.

Single point sensors commonly found in camera phones [152] are used for distance ranging for screen locking or auto-focus assist applications. In the biomedical field, single point sensors are also used in FLIM and PET applications. An early demonstration of such a design for FLIM was presented by Borghetti et al. [153] where the dSiPM comprised an array of 16 × 16 SPADs combined by an OR tree. Photon counting functionality was integrated on-chip but TCSPC measurements were performed externally.

The work by Frach et al. [154] presented a larger design of four dSiPMs each containing 64 × 32 SPADs and sharing a TDC integrated on-chip. Another source of pile-up in single point sensors is the TDC conversion dead-time where the photon that triggers the TDC gets resolved while others arriving within that conversion period are not. To improve the conversion throughput, Tyndall et al. [155] proposed interleaving TDCs to share the load and Mandai et al. [156] took that a step further by proposing a column parallel TDC architecture.

Part of the TDC dead-time is attributed to the data transfer time needed to process the TDC code before a new conversion can take place. While interleaving TDCs can alleviate this limitation, mismatch between the different TDCs can add error to the measurement. To address these issues Dutton et al. presented a 32 × 32 dSiPM with a novel folded flash TDC architecture with a throughput of 14GS/s and direct on-chip histogram generation of TCSPC data [157]. The innovative design also proposed a new method of combining SPAD pulses by means of an XOR tree.

Combining the different SPAD pulses into a single output is also a source of pile-up distortion in dSiPMs as the channel has a limited bandwidth. Combining the pulses by an OR tree as in [153] is limited by the SPAD dead-time where overlapping detections would all be merged into a single pulse by the OR gates. Braga et al. proposed shortening the SPAD pulse before feeding it into the OR tree therefore significantly improving the channel bandwidth as the probability of these short pulses overlapping is decreased [158].

Alternatively the work in [157] opted for feeding the SPAD pulses into a toggle flip-flop thus encoding photon detections on both rising and falling toggle edges and then feeding these edges into an XOR tree. This approach resulted in a 2× improvement in channel bandwidth yet it suffers from

event losses whereby two edges entering the XOR gate within a gate delay would cancel out. A detailed comparison of the OR tree and XOR tree channel architectures is provided in [159].

Single point sensors and dSiPMs are not the topic of this work and therefore will not be discussed further.

2.2.2. Line Sensors

Line sensors are made of a 1-dimensional (1-D) array of pixels operating in parallel with each having its own processing circuitry. The line configuration can be utilised for spectral measurements by spreading the light's spectral components over the array or scanning systems where 2-dimensional (2-D) images can be reconstructed line by line. Figure 2.2.2 illustrates the line sensor architecture.

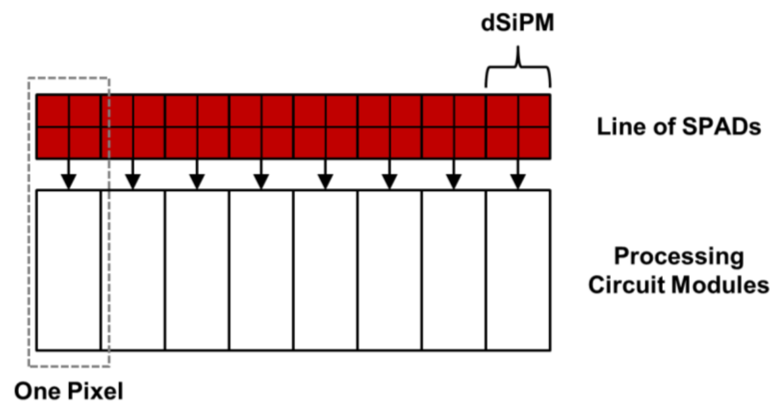


Figure 2.2.2. Line sensor architecture.

Stoppa et al. first presented a 64 SPAD pixel line sensor with an analogue approach implementing a time to digital converter (TAC) for time of flight imaging [62] while Tisa et al. presented a 32 pixel line sensor with a digital counter per pixel for photon counting applications [159/2]. These sensors had one SPAD per pixel coupled to the circuitry.

Due to the absence of scaling restrictions of 2-dimensional arrays, line sensors have the freedom to integrate more detectors and processing electronics per pixel especially in the vertical direction while the horizontal spatial resolution can easily expand. Hence Pancheri et al. presented a 64 pixel line sensor with four SPADs and four parallel counters per pixel for time-gated fluorescence applications [149]. The outputs of the four SPADs are combined forming a dSiPM, an approach commonly adopted by line sensors for improved detection efficiency.

Niclass et al. presented a line sensor coupled to a scanning mechanism for automotive LIDAR applications [160] demonstrating 2-dimensional time of flight depth maps up to 100m range and later

presented a very impressive fully integrated system on chip (SoC) based on two line arrays and a TCSPC digital signal processing (DSP) module [161].

Maruyama et al. presented a large line sensor of 1024 pixels aimed for time-gated laser Raman spectroscopy and laser induced breakdown spectroscopy for space missions [162]. Similarly Nissinen et al. presented several TDC-based [163] and multi-gate [71] line sensors also for Raman spectroscopy.

Krstajic et al. further expanded the dSiPM configuration of the line sensor pixel by adding two banks of different SPAD structures tailored towards the blue and red regions of the spectrum to suit different applications [164] while Erdogan et al. presented a sophisticated 1024 pixel sensor with per-pixel TCSPC histogram generation and time zooming capability [165]. The integration of such complex processing or detector arrangements would not have been possible without the area flexibility of 1-D arrays.

As with the single point sensors, line sensors are not the topic of this work and therefore will not be discussed further but presented here for completeness.

2.2.3. Image Sensors

Similar to CISs, SPAD image sensors are 2-D arrays of pixels with each pixel containing a single SPAD coupled to its dedicated circuitry. Initial SPAD arrays were simple in design and the pixel electronics were minimalistic including only quench and recharge and readout elements. The pixels were sequentially scanned with their pulse outputs routed to external counting or TCSPC modules to create intensity or depth images [61][166].

Eventually SPAD image sensors evolved into more sophisticated systems with fully integrated functionality on per pixel basis and column parallel readout for a wide range of time-resolved applications. Figure 2.2.3 illustrates a typical SPAD image sensor architecture resembling that of a CIS.

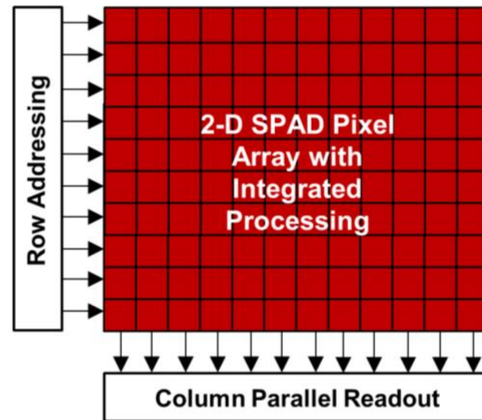


Figure 2.2.3. Image sensor architecture.

Generally speaking such sensors are categorised based on their functionality and pixel circuit implementation. Figure 2.2.4 shows a taxonomy chart of SPAD imager sensors. The two types of functionality embedded in the pixel are time correlated single photon counting (TCSPC) or single photon counting (SPC) which is usually time-gated for time-resolved capability. Both functions can either be achieved through analogue or digital circuits. Digital TDC or analogue TAC circuits are used for TCSPC while analogue or digital counters are used for SPC. There is also a third type of SPC / time-gated pixels relying on single-bit in-pixel memory which can either be analogue or digital.

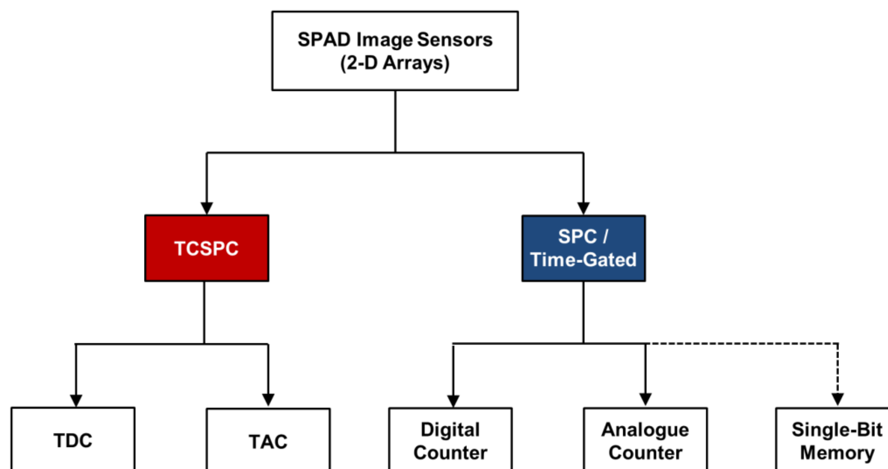


Figure 2.2.4. SPAD image sensor taxonomy.

SPAD image sensors had the lion's share of research and development with many sensors reported in the literature. For TCSPC operation, TDC-based implementations are the most common. Richardson et al. presented a design based on a high speed in-pixel ring oscillator (RO) where the time difference between the start (SPAD trigger) and stop (global clock trigger) signals is measured by the number of

RO oscillations (coarse resolution) and the internal state of the RO stages (fine resolution) once frozen [167].

Several manifestations of this circuit (Figure 2.2.5) were reported due to its high temporal resolution and relative simplicity including a large 160 × 128 array [168] and extended temporal dynamic range design by adding extra bits to the RO coarse counter [169]. The ability to tune the temporal resolution by adjusting the RO speed through a bias voltage is an additional advantage. When implemented in an advanced technology node such as 40nm, the pixel circuitry can occupy less than 100µm² area [170].

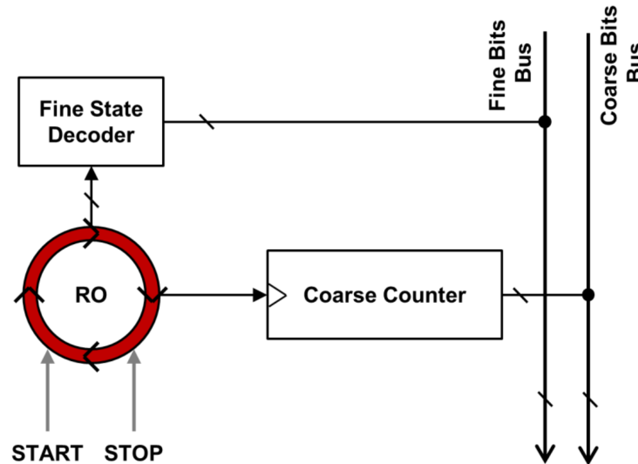


Figure 2.2.5. Conceptual schematic of ring oscillator based TDC.

Another common TDC-based approach was presented by Gersbach et al. [171] where a high speed clock is broadcast globally to the pixel array providing a coarse timing resolution by counting in-pixel the number of clock cycles between the start and stop signals. Higher temporal resolution is obtained by an in-pixel fine interpolator (delay line) that further quantises the coarse clock period into fine intervals. Villa et al. broadcasted 16 phases of the clock globally and sampled them in-pixel [172]. Table 2.2.1 summarises the main parameters of the aforementioned TDC-based sensors.

First Author	Reference	Year	Process (nm)	Pixel Pitch (µm)	Fill Factor (%)	Array Size	TDC Resolution (ps)	TDC Range (ns)	TDC Bits	TDC Type
Richardson et al.	[167]	2009	130	50	1	32 × 32	50	51.2	10	Ring Oscillator
Veerappan et al.	[168]	2011	130	50	1	160 × 128	55	56.3	10	Ring Oscillator
Vornicu et al.	[169]	2014	180	64	3.5	64 × 64	145 - 625	296.9 - 1280	11	Ring Oscillator
Henderson et al.	[170]	2018	40	13	13	192 × 128	33 - 120	135 - 491	12	Ring Oscillator
Gersbach et al.	[171]	2009	130	50	1	32 × 32	119	100	10	Global Clock + In-pixel Interpolator
Villa et al.	[172]	2014	350	150	3.14	32 × 32	312	319.5	10	16 Phases of Global Clock

Table 2.2.1. Summary of TDC-based FSI SPAD image sensor parameters showing good temporal resolution but low fill factor and large pixel pitch for the exception of [170] implemented in an advanced 40nm node.

While they provide high temporal resolution, these sensors suffer from several drawbacks:

1. Power consumption. Broadcasting high speed clocks (few 100MHz) continuously across the whole array is very power hungry and also requires careful management of skews. Similarly for RO designs, when a large number of pixels fire, the high speed ROs (~2GHz) would consume a lot of power and this is light level dependent. Power droops due to poor power gridding would impact the TDC performance.
2. Large pixel pitch and low fill factor. Due to the complex digital processing needed by TCSPC pixels, the pixel area tends to be large with a small proportion of the area dedicated for the SPAD thus resulting in low sensitivity.
3. Data rates. TCSPC measurements require accumulating many time-stamps per pixel before a time-resolved measurement can be extracted. Since each pixel generates a single time-stamp code of several bits (10-bit or more) at a time, the overall data rates required are high.

To tackle the first two issues, several innovative architectures have been proposed. Early on Niclass et al. presented a column parallel architecture where the TDC has been pushed to the column level and shared amongst four column pixels [63]. In this scenario, the first pixel to fire in an enabled row claims ownership of the shared TDC registering its time stamp and address. This has the advantage of minimising the pixel circuitry hence reducing its area and increasing its fill factor while reducing the number of TDCs. A similar design by Schwartz et al. for FLIM uses column level latches that can be triggered by any pixel in the column to store the state of a globally (amongst the column latches) broadcasted TDC code [173].

While this event driven sharing scheme cleverly reuses TDC resources its conversion efficiency can be hindered by the data transfer dead-time and by collisions or contention on the shared line by multiple pixels firing simultaneously especially under higher illumination levels.

This was addressed by Lindner et al. in a sensor aimed at near infra-red optical tomography (NIROT) by dynamically reallocating four TDC's per column of 32 pixels [174]. He later expanded the scheme into an impressive 252 × 144 flash LIDAR array with six TDCs per column and partial on-chip TCSPC histogram generation providing up to 14.9× data rate compression [175]. Table 2.2.2 summarises the main parameters of the resource sharing architectures discussed.

First Author	Reference	Year	Process (nm)	Pixel Pitch (µm)	Fill Factor (%)	Array Size	TDC Resolution (ps)	TDC Type
Niclass et al.	[63]	2008	350	25	6	128 × 18	97	Interpolating
Schwartz et al.	[173]	2008	350	40	0.94	64 × 64	350	DLL Multi-Phase
Lindner et al.	[174]	2017	180	28.5	28	32 × 32	40 - 80	Ring Oscillator
Lindner et al.	[175]	2018	180	28.5	28	252 × 144	48.8	Ring Oscillator

Table 2.2.2. Summary of resource sharing TDC-based FSI SPAD image sensor parameters showing good temporal resolution and improved pixel pitch and fill factor compared to pixel dedicated TDC architectures.

As for the excessive data rates of TCSPC sensors, Field et al. tackled the issue by designing an application specific data-compression data path on-chip to enable continuous FLIM image acquisition [176]. Gasparini et al. proposed a different scheme whereby rows or frames with no sufficient photon content can be skipped to avoid reading out redundant zero data and increase frame rates [177]. This is particularly useful in photon starved environments.

As an alternative to TDCs, analogue time to amplitude converters (TACs) have been proposed. Stoppa et al. demonstrated a 32 × 32 array where a start signal initiates the charging of a capacitor by an in-pixel current source and the stop signal halts it [178]. Knowing the current and capacitor values the photon time of arrival can be calculated from the sampled voltage.

Due to other functionality integrated in the pixel, it had a pitch of 50μm and a fill factor of 1%, yet the 256 × 256 array by Parmesan et al. demonstrated an optimised all NMOS pixel with a globally distributed voltage ramp at an 8μm pitch and 19.63% fill factor [72] showing the advantage of analogue pixels in terms of area.

Despite the wide range of optimisation techniques explored, TCSPC systems are not suitable for miniaturisation as the pixels are large when implemented in digital architectures. This also applies to digital event driven or sharing architectures with large converter circuits. Analogue pixels have the potential to shrink but still require ADC circuits which add to the design overhead. Data rates associated with TCSPC are still restricting and are a serious bottleneck when parallel readout channels are not an option.

As highlighted in Chapter 1, time-gated SPC pixels that integrate photons within an observation window rather than generating a time-stamp per photon significantly reduce the data overhead at the cost of measurement temporal resolution. Digital SPC pixels are straightforward to design with several configurations possible. Niclass et al. presented a dual counter pixel for time of flight applications [179] while Bronzi et al. reported a triple counter pixel for higher gating flexibility [180]. Lee et al. presented a single counter pixel with integrated metal grids on top for angle sensitive measurements [66].

The main disadvantage of digital counting pixels is their relatively large pitch due to the size of logic gates especially in older technology nodes. On the up side, they require no ADC conversion, allow for shot noise limited imaging due to noiseless readout, have relatively high counting capacity which increases exponentially with every added bit and allow for edge sensitive gating where the SPAD's leading edge containing the timing information is what triggers the counter.

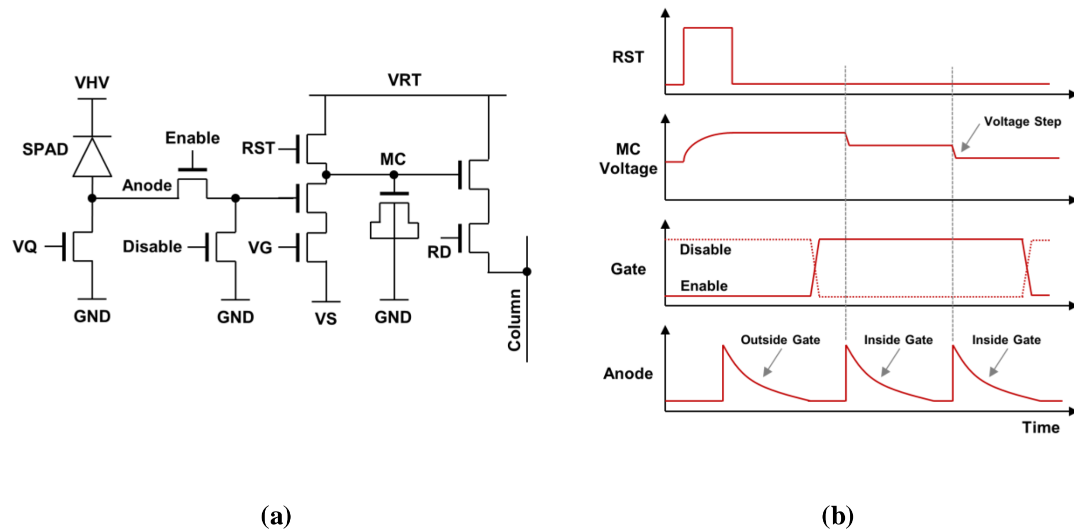
Analogue time-gated SPC image sensors alleviate the area limitation of digital counters and offer a compact solution by counting in the charge domain. By doing so not only is the storage element (MOS or metal-oxide-metal (MOM) capacitor or both) compact but the circuit can be implemented

with a very small number of transistors enabling small pitch and high fill factor. Table 2.2.3 summarises the parameters of the most relevant works.

First Author	Reference	Year	Process (nm)	Pixel Pitch (µm)	Fill Factor (%)	Array Size	Number of Transistors	Counting Capacity in Analogue Mode
Pancheri et al.	[181][183]	2011	350	25	20.8	32 × 32	12T + 2C	150
Dutton et al.	[65][73]	2014	130	8	26.8	320 × 240	8T + 1C	400
Perenzoni et al.	[182][184]	2015	350	15	21	160 × 120	7T + 1C	70

Table 2.2.3. Summary analogue SPC SPAD image sensors showing the small pitch and high fill factor but relatively low photon counting capacity per pixel in a single frame capture.

Although few analogue counting image sensors have been presented, reviewed in [76], they all rely on a common principle which is discharging a capacitor previously reset to an initial value by a controlled amount of charge for every photon detected within a time gate. The uniformity of this charge step from one count to another and from one pixel to another is important in determining the correct number of recorder photons after a readout operation. Figure 2.2.6 shows the pixel schematic and timing diagram of the pixel presented in [65].



(a) **(b)**

Figure 2.2.6. Analogue SPC pixel presented in [65]. (a) Schematic diagram. (b) Timing diagram.

In general analogue pixels tend to suffer from this non-uniformity due to the dependencies of the discharging step on variations in the SPAD pulse width, mismatch of the in-pixel capacitances and mismatch in key transistors in the counter path. The work by Pancheri et al. reported a step size mismatch of 11.6% across the array [183] while the work by Dutton et al. reported on the non-uniformity sources of a pixel that can operate in either switched current source (SCS) or charge transfer amplifier (CTA) mode [73].

In addition to non-uniformity issues, cumulative noise due to the in-pixel switching behaviour [74], temporal noise on supplies and readout noise due to the source follower and column parallel chain

further deteriorate the analogue count. Readout noise can be cancelled out by known correlated double sampling techniques (CDS) [73] but a more novel approach was proposed by Perenzoni et al. where the pixel itself was used as a self-referenced ADC to also cancel out any counter non-linearity [184].

Two other issues arise with the use of analogue counting pixels: first is the limited in-pixel counting capacity which is between few 10s [182] to few 100s photons [65], second is level sensitive time gating which means that if the recharge tail of a SPAD event was present within a time gate it may trigger a false count if the voltage level is above the threshold of the circuit.

The final category of time-gated SPC image sensors is the binary architecture. These sensors utilise a minimalistic single-bit pixel in order to improve the pixel's fill factor and can be implemented in either the digital or analogue domains. A digital static memory design was presented by Maruyama et al. for FLIM in [185] which was then scaled to a large 512 × 128 array [80]. Gyongy et al. presented an optimised 16μm pixel with an analogue dynamic memory and 61% fill factor [186] while Ulku et al. presented a similar circuit design in a 512 × 512 array making it the highest resolution reported for SPAD image sensors [187].

Due to the limited pixel counting capacity, such sensors need to be heavily oversampled which is a different imaging paradigm to conventional sensors that will be introduced and discussed in further detail in Chapter 4.

2.2.4. SPAD Image Sensor Survey

Since developing a miniature image sensor is the main target of this research, it is worth analysing the impact of technology development on the pixel pitch and fill factor. Figures 2.2.7 and 2.2.8 present a survey across more than thirty distinct published image sensors with one SPAD per pixel.

A note to the reader: all the sensors discussed in the sections above are implemented in front side illuminated (FSI) technologies which are the common path for SPAD sensors in general. In recent years the emergence of more advanced BSI 3D-stacking technologies spearheaded by MIT Lincoln Lab has occurred [92]. Section 2.3 will review this development in detail but relevant image sensors were included in the survey for the bigger picture including published work presented in this thesis [1].

Two clear trends emerge in Figure 2.2.7 when analysing the impact of technology choice on pixel pitch. Firstly, digital architectures benefit hugely from the shrink in logic gates allowing for smaller pixels down to 8μm in 40nm CMOS as demonstrated by this work (Chapters 3 and 5). Secondly, analogue designs on the other hand, while generally delivering small pixels, do not scale accordingly with process node as they mostly rely on thick oxide transistors which are relatively large in dimension for better matching and voltage swing. Therefore if all the benefits of a digital pixel are to be reaped in a miniature sensor the cost of an advanced technology node is justifiable.

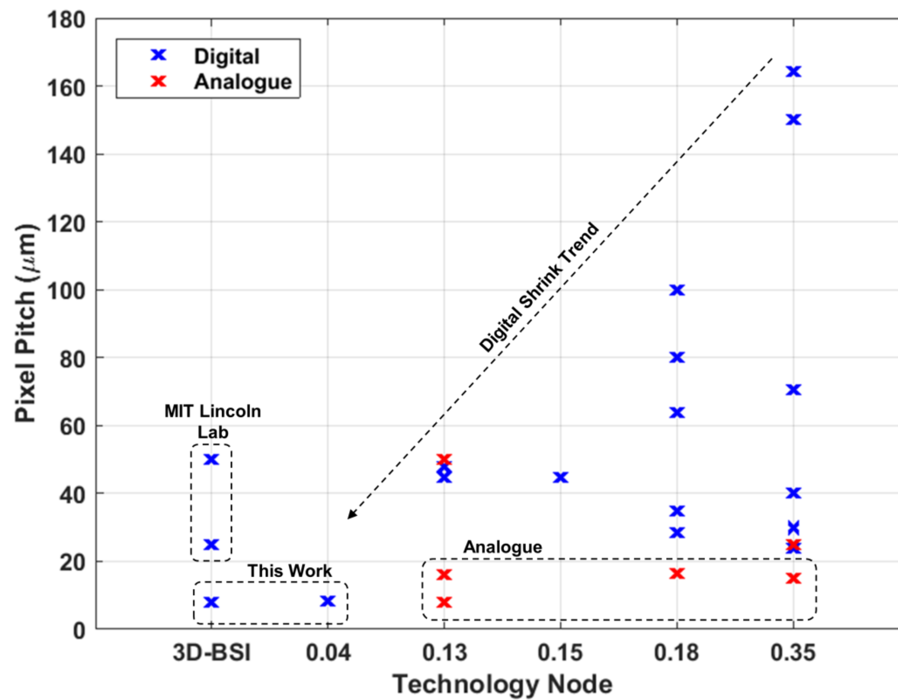


Figure 2.2.7. Pixel pitch versus technology node survey for more than thirty reported SPAD image sensors

Analysing the impact of technology node on fill factor (Figure 2.2.8), it can be seen that based solely on this merit, there is no clear cut winner between analogue or digital designs or even the FSI process choice with the exception of a few works which exploit unique architectures or 3D-stacking.

The single-bit design presented in [186] takes advantage of a highly optimised pixel layout while this work leverages the fine design rules of advanced DSM nodes to expand well sharing layout techniques (Chapter 3). 3D-stacking enables backside illumination (BSI) which allows for scalable and high sensitivity compact SPAD arrays (Chapter 5) with potential fill factor figures reaching 100% as claimed by MIT Lincoln Lab. Therefore, the necessity of using advanced technologies for miniaturisation is further complemented by the gain in pixel sensitivity.

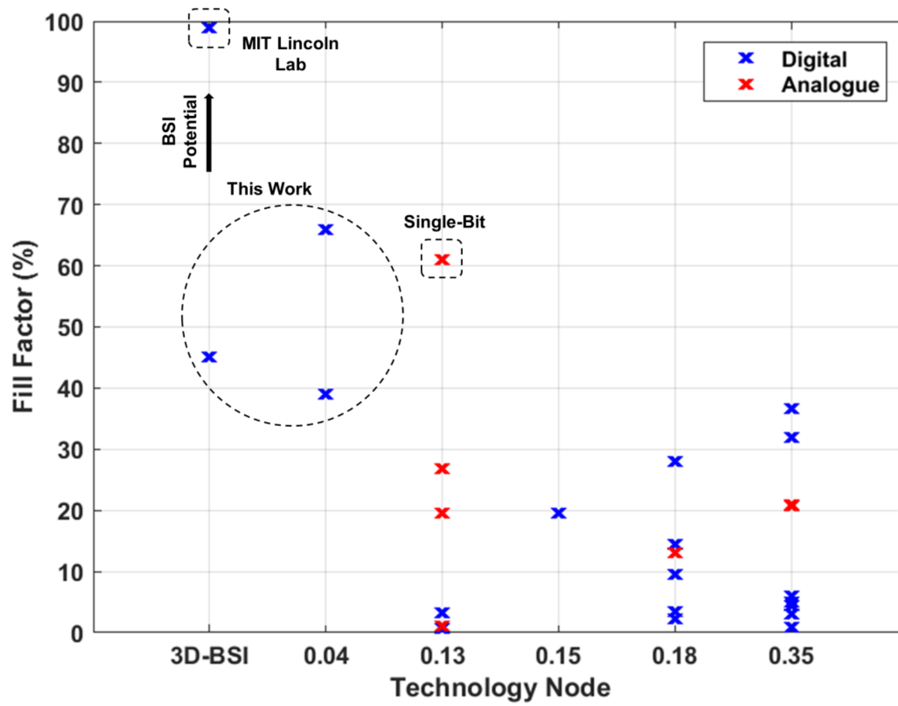


Figure 2.2.8. Fill factor versus technology node survey for more than thirty reported SPAD image sensors.

2.2.5. Modular Sensors

The last class of SPAD sensors is referred to here as modular as these sensors combine features of different architectures in modules that are arrayed to compose the overall system. A clear example is the medical positron emission tomography (PET) sensor presented by Braga and Walker et al. [188][189] which contains an array of 8×16 modules or macro-pixels. Each module in turn is made of four dSiPMs of 12×15 SPADs with their dedicated quench circuits and counter banks. The four dSiPMs then share a common TDC and data management circuit. The architecture of this large scale $5.5\text{mm} \times 10\text{mm}$ sensor is shown in Figure 2.2.9.

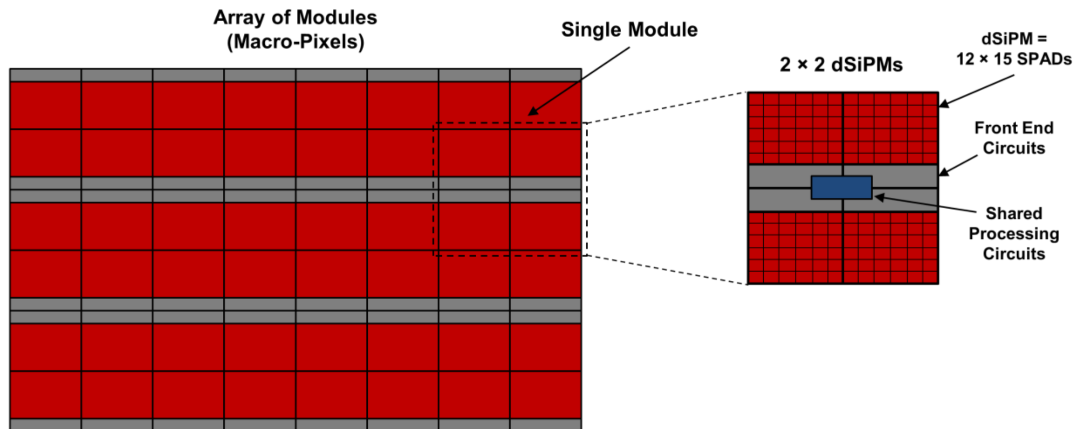


Figure 2.2.9. Illustration of the modular sensor architecture presented in [188].

A similar sensor also intended for PET was presented by Carimatto et al. where the array of 9 × 18 macro pixels each containing a dSiPM of 16 × 26 SPADs with 48 TDCs shared per column [190]. He recently presented a simpler design of four macro-pixels with column parallel TDCs that can operate in frame-based or event-driven modes implemented in the same 40nm process used in this work [191].

Other examples of modular sensors include a 64 × 64 time of flight image sensor for spacecraft navigation and landing by Perenzoni et al [192]. Unlike conventional image sensors with one SPAD per pixel, each pixel is composed of a mini-dSiPM of 2 × 4 SPADs sharing a TDC. The sensor can operate as a short range depth image sensor or as a single point long range altimeter. Akita et al. also reported a 32 × 4 imaging array of macro-pixels for automotive LIDAR [193].

As the architecture of these modular sensors deviates from the standard image sensor model of one SPAD per pixel pursued in this research, they will not be discussed any further in this thesis.

2.3. 3D-Stacking Technology and SPADs

Driven by the demands of high packaging density of transistors, the microelectronics industry started exploring the prospects of vertically integrated ICs (3D-ICs) with demonstrations of stacked layers of active circuitry dating back to the 1980's [194]. In 1987 Nishimura et al. from Mitsubishi Electric Corporation demonstrated one of the first imaging 3D-ICs comprising a 5 × 5 photodiode array fabricated on top of a 2-bit CMOS ADC layer which was fabricated over another layer of CMOS logic [195].

Since then 3D-ICs or 3D-stacking, more commonly used when vertically stacking dies or wafers rather than fabricating multi-layers of devices, have been demonstrated for a variety of applications. In the dynamic random access memory (DRAM) industry 3D-stacking is a key technology to overcoming the memory-processor bus bottleneck [196] improving speeds and reducing power due to

the shorter interconnect paths and lower RC parasitics. Other domains such as micro-electro mechanical systems (MEMS) or silicon photonics benefit from embedded functionality and compactness of hybrid 3D-stacked solutions. An overview of the opportunities and challenges of 3D-ICs is provided in [197].

Similarly CIS sensors would benefit from the improved system performance, reduced form factor and the ability to integrate image signal processing (ISP) on chip not to mention the gain in sensitivity due to backside illumination. Moreover, 3D-stacking enables optimising the process material and fabrication steps of the imaging and processing tiers independently and so leverages the best of both worlds. Such technology offers new opportunities for SPAD sensors.

2.3.1. CIS 3D-Stacking Techniques

There are several 3D-stacking techniques that have been adopted for CMOS image sensors with the most prominent being the use of through silicon vias (TSVs), micro-bumps and face to face hybrid-bonding of wafers.

In 2005 MIT Lincoln Labs demonstrated a 1-Mpix image sensor at an 8μm pixel pitch [198] using a TSV stacking technology with one connection per pixel between the photodiode and its readout circuitry. The details of the in-house developed process are published in [199]. In 2013 Sony announced the first commercial 3D-stacked CIS which used TSVs at the edge of the array to connect the column parallel outputs of the top imaging tier to the ISP bottom tier [200]. Figure 2.3.1 shows an illustration of the TSV 3D-stacking process.

More recently Sony further expanded the TSV process to stack 3-tiers of pixels, 1-Gbit DRAM and processing logic [201]. The impressive sensor utilises a dedicated CIS 90nm process for the imaging tier, 30nm DRAM process and a 40nm logic substrate. The high embedded functionality allows the sensor to capture full-HD resolution images at 960fps overcoming the limitations of output interface speeds [202].

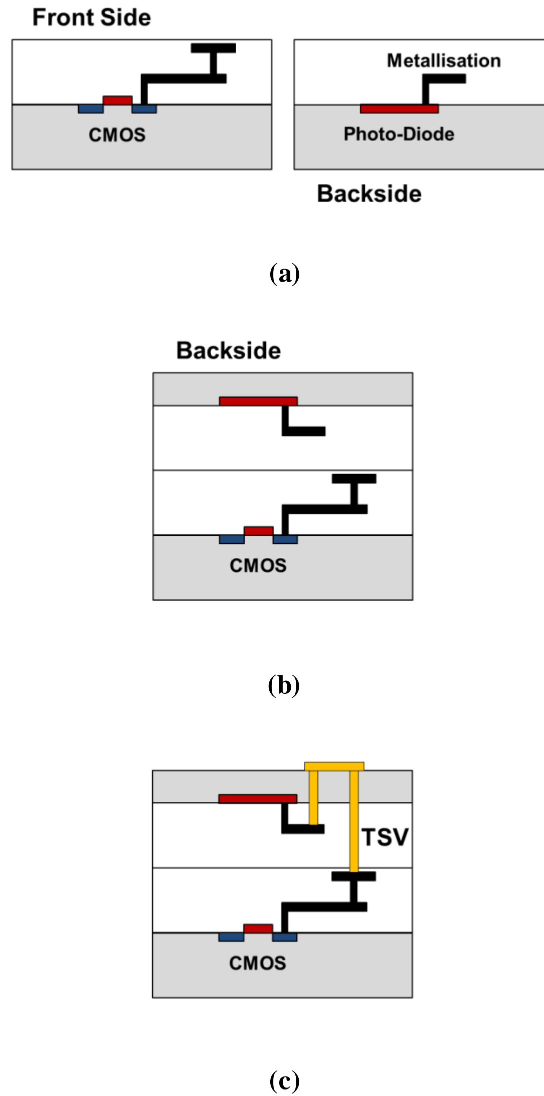


Figure 2.3.1. Generic illustration of TSV 3D-stacking process. (a) Processing and imaging tiers fabricated independently. (b) Imaging tier flipped over processing tier face to face, bonded and thinned from backside. (c) TSVs connectivity established.

Olympus on the other hand adopted a different 3D-stacking technique relying on micro-bumps. In 2013 they demonstrated a global shutter sensor with the photodiode integrated on the top tier and the storage node on the bottom tier to utilise the metallisation between the two wafers as a light shield for improved parasitic light sensitivity (PLS) [203]. The micro-bumps had a pitch of 8.6µm and one connection was shared between four pixels. Figure 2.3.2 shows an illustration of this technique.

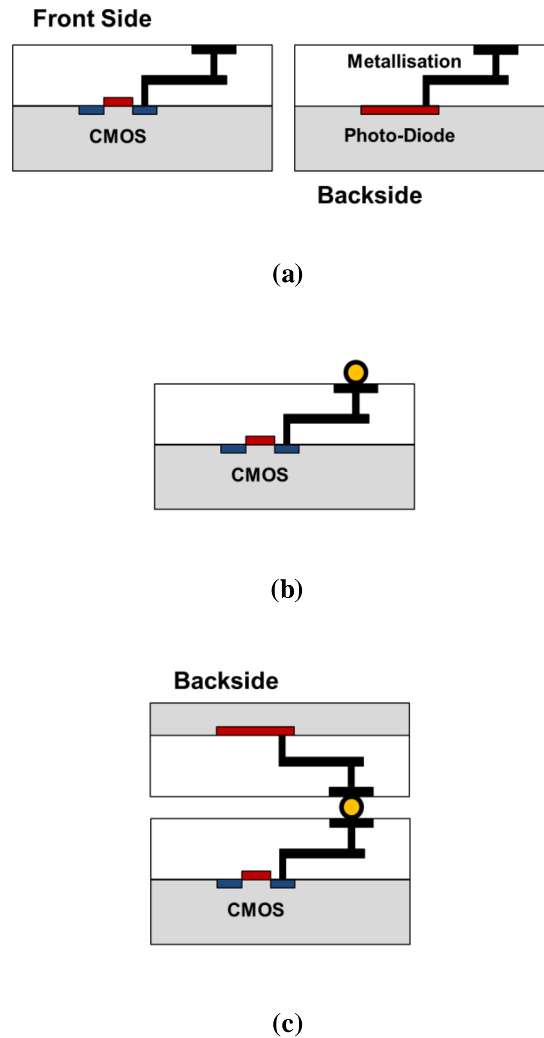


Figure 2.3.2. Generic illustration of bump bonding 3D-stacking process. (a) Processing and imaging tiers fabricated independently. (b) Bump bonds fabricated over processing tier. (c) Imaging tier flipped over processing tier face to face, bonded and thinned from backside.

Another approach to 3D-stacking is the face to face hybrid-bonding of wafers which offers a path towards higher density interconnects compared to TSVs or micro-bumps that are limited by the scalability and reliability of their physical structures. First invented by Ziptronix (now Invensas) under the registered trademark of direct bonding interface (DBI®), this technology has been demonstrated down to 2µm pitch [204].

The term hybrid bonding (HB) refers to the bonding of the stacked wafers at both the oxide-to-oxide interface and the Cu-to-Cu interface of the metallisation interconnect sites fabricated using standard back end of the line (BEOL) damascene process. Other companies have recently demonstrated their own development of this technology including STMicroelectronics / LETI [205] and Sony [206]. Figure 2.3.3 shows an illustration of the HB process.

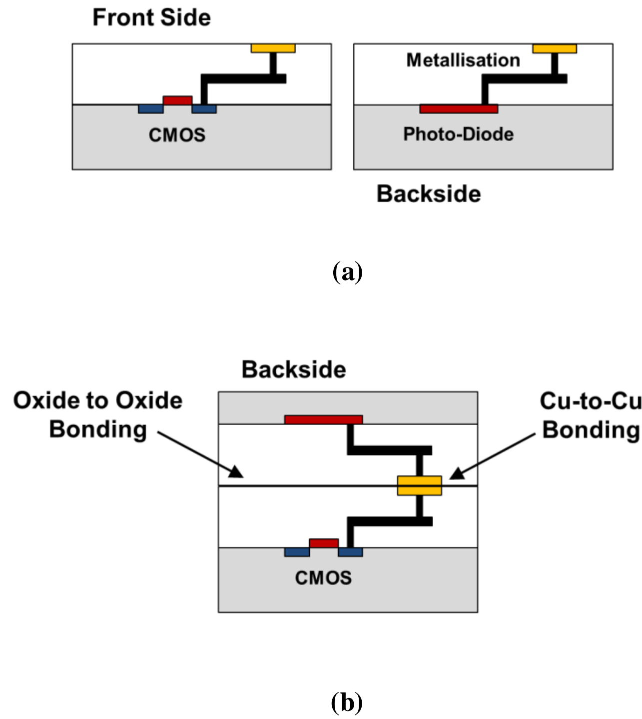


Figure 2.3.3. Generic illustration of hybrid bonding 3D-stacking process. (a) Processing and imaging tiers fabricated independently. (b) Imaging tier flipped over processing tier face to face, bonded and thinned from backside.

Other examples of vertical integration of image sensors although not necessarily 3D-stacking include deposition of exotic material such as organic photoconductive film [207] or quantum film [208] on CMOS for high dynamic range or improved NIR sensitivity. A survey of the 3D-stacking technologies featuring in mainstream cameras can be found in [209][210].

2.3.2. Overview of 3D-Stacked SPAD Sensors

With the advent of 3D-stacking, several groups have seized the opportunity to develop high performance sensors through either in-house developed technologies or through foundry options. Figure 2.3.4 gives a timeline of the key published works.

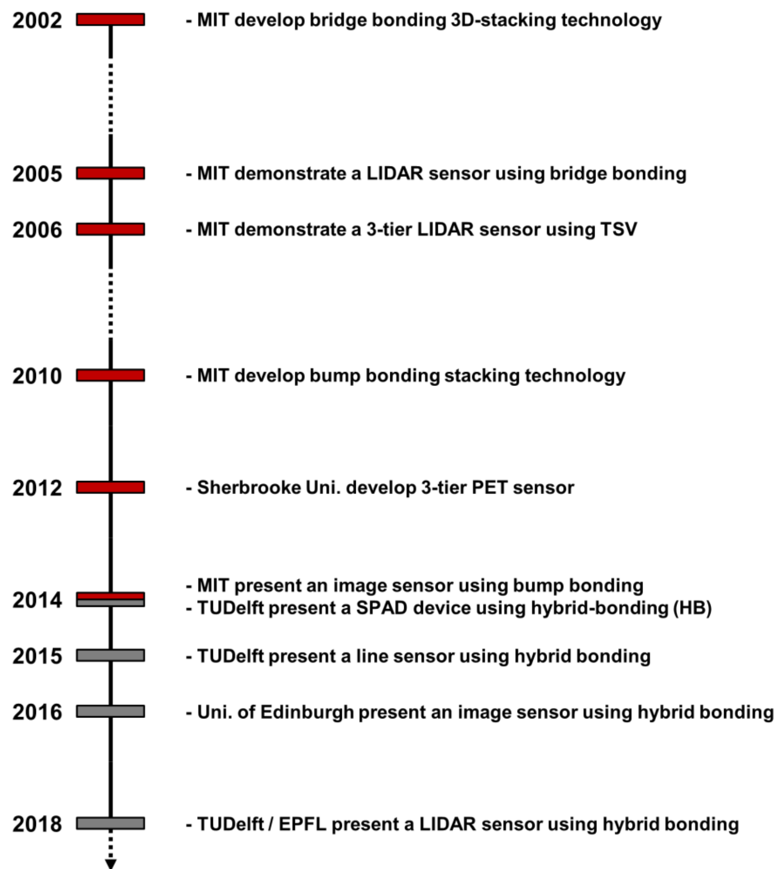


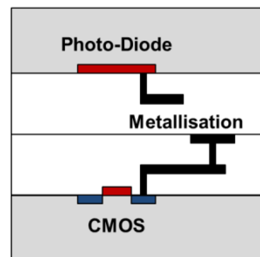
Figure 2.3.4. Timeline of the key 3D-stacked SPAD published works. Red indicates in-house technologies and grey indicates foundry technologies.

3D-stacked SPADs were first pioneered by MIT Lincoln Lab in 2002 where they demonstrated an array of 32 × 32 custom silicon devices stacked over a CMOS readout IC (ROIC) using the bridge bonding technique [92]. The 100µm pitch pixel contained a single backside illuminated SPAD connected to a TDC circuit for LIDAR applications. The sensor was later coupled to a laser chip demonstrating time-resolved depth maps [211].

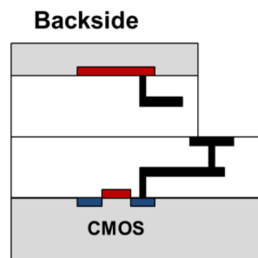
A similar pixel was then presented at ISSCC 2006 at a finer 50µm pitch and a 64 × 64 array resolution using the previously mentioned in-house TSV technique [212]. This sensor was composed of three tiers: custom SPAD in the top imaging BSI tier, 0.35µm CMOS front end circuitry in the second tier and 0.18µm CMOS timing circuitry in the third tier. This was the first demonstration of a 3-tier stacked image sensor.

In 2010 the same group developed an in-house bump bonding technique [213] used later to realise a BSI 256 × 256 custom SPAD array at a pixel pitch of 25µm and a claimed 100% fill factor [67] with a 0.18µm photon counting CMOS ROIC. A complete characterisation account of this sensor was presented in [214] showing yield issues and optical cross-talk between pixels.

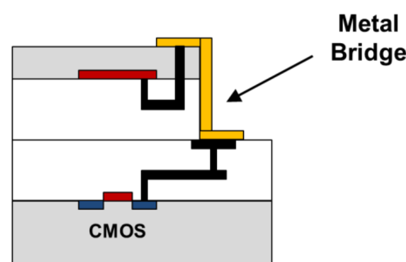
Unlike the bridge bonding technique which requires a portion of the pixel area to form the bridge bonds (Figure 2.3.5), 3D-stacking by TSV and bump bonding allow for high fill factor since the SPAD fully occupies the pixel. A comprehensive review of the MIT technology can be found in [215] and more recently in [216] suggesting next generation sensors will be stacked onto 65nm CMOS.



(a)



(b)



(c)

Figure 2.3.5. Generic illustration of bridge bonding 3D-stacking process. (a) Imaging tier bonded over processing tier face to face. (b) Imaging tier thinned and etched to expose processing tier metallisation. (c) Metal bridge connecting both tiers is formed.

Sherbrooke University in Canada reported in 2012 the development of 3D-stacked SPAD sensors targeting PET applications [217]. Detailed characterisation results of the SPAD device were presented

in [218]. The three tier design contains the processing electronics on the bottom third tier and second contains the front end quench circuitry, both implemented in 0.13µm CMOS.

A top tier of SPADs designed in a high voltage 0.8µm is then stacked on top of the ROICs using an in-house TSV process which results in front side illuminated detectors. Cross section figures suggest the 2nd and 3rd tiers employ face to face hybrid bonding with other publications [219] stating they were fabricated through the Tezzaron foundry known to have licensed the DBI® technology.

Two other groups joined the race for 3D-stacked SPADs in 2014 with Politecnico di Milano and Fraunhofer IMS reporting an attempt to stack a 32 × 32 custom SPAD array onto 0.35µm ROIC using a bump-bonding like technique called solid liquid inter diffusion (SLID) but no results were shown [220].

Prof. Charbon of TUDelft on the other hand reported characterisation results for a test BSI SPAD realised for the first time in a wafer scale hybrid-bonding technology between imaging and ROIC tiers, both fabricated in Tezzaron's 0.13µm CMOS process [75]. The TUDelft team later reported a fully integrated line sensor in the same technology with each top tier pixel containing a dSiPM of 2 × 4 SPADs alongside the front end electronics delivering an output to the bottom tier processing circuitry [221].

At the International Electron Devices Meeting (IEDM) 2016 the University of Edinburgh in collaboration with STMicroelectronics (Chapter 5) presented a 128 × 120 BSI array at state of the art 7.83µm pitch [1] using ST's HB technology with 65nm and 40nm top and bottom tiers respectively. Although the MIT group presented 2-D arrays they were either dedicated for TCSPC or SPC operation while this sensor enabled both SPC intensity and time-gated imaging. EPFL later presented an active quench pixel using the same technology [145].

The most recent development came at IEDM 2017 when EPFL / TUDelft presented a BSI SPAD in TSMC's 3D-stacking technology also with advanced DSM top and bottom tiers [222]. Further results on device design and characterisation were presented in [223] and a modular sensor for LIDAR was recently presented at ISSCC 2018 implemented in the same process [224].

Reference	[92][211]	[212]	[217] also [218][219]	[67] also [214][215]	[220]	[75]	[221]	This Work (1) [1]	[145]	[222][223]	[224]	This Work (2) Unpublished *	This Work (3) Unpublished †
First Author	B. Aull	B. Aull	B. Berube	B. Aull	Y. Zou	E. Charbon	J. Pavia	T. Al Abbas	S. Lindner	M. Lee	A. Ximenes	T. Al Abbas	
Institution	MIT	MIT	Sherbrooke Uni	MIT	Poli. Di Milano	TUDelft	TUDelft	Uni. Of Edinburgh	EPFL	EPFL / TUDelft		Uni. Of Edinburgh	
Year	2002 / 2005	2006	2012	2014	2014	2014	2015	2016	2017	2017 / 2018	2018	2018	
Application	LIDAR	LIDAR	PET	SPC	ToF	Experimental	NIROT	Experimental	Experimental	Experimental	LIDAR	FLIM / Endoscopy	Experimental
Architecture	Image Sensor	Image Sensor	dSiPM	Image Sensor	Image Senor	Test SPAD	Line Sensor	Image Sensor	Test Pixel	Test SPAD	Modular	Image Sensor	
Number of Tiers	2	3	3	2	2	2	2	2	2	2	2	2	2
SPAD	Custom	Custom	Custom	Custom	Custom	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
Illumination	BSI	BSI	FSI	BSI	BSI	BSI	BSI	BSI	BSI	BSI	BSI	BSI	BSI
ROIC	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS	CMOS
Stacking Technology	Bridge Bond	TSV	TSV	Bump Bond	Bump Bond	Hybrid Bond	Hybrid Bond	Hybrid Bond	Hybrid Bond	n/a		Hybrid Bond	Hybrid Bond
Development	In House	In House	In House	In House	In House	Foundry	Foundry	Foundry	Foundry	Foundry	Foundry	Foundry	Foundry
Resolution	32 × 32	64 × 64	22 × 22	256 × 256	32 × 32	1	1 × 400	128 × 120	1	1	256 × 256	128 × 120	128 × 96
* Preliminary silicon results presented in Chapter 6. † Sensor design presented in Chapter 6 pending fabrication.													

Table 2.3.1. Summary of 3D-stacked sensors presented in literature alongside sensors presented in this work.

Table 2.3.1 summarises the main technology parameters of the aforementioned efforts in addition to two other unpublished 3D-stacked sensors presented later in Chapter 6. One main trend appears since 2014 (grey) which is the shift from custom SPADs stacked using in-house technologies to all CMOS foundry implementations predominantly employing hybrid-bonding technology. This highlights the demand for high performance SPAD sensors driven by the industry opening the door to a new era of integration capabilities.

2.4. Summary and Conclusions

CMOS SPAD devices are maturing and are reaching respectable levels of performance in terms of DCR, PDP and jitter even when implemented in deep submicron technologies due to careful junction design and customised processes. Several SPAD structures have been explored in the literature with ones relying on virtual guard rings showing great promise for miniaturisation with active diameters down to 1 μm being presented.

Designing miniature time-resolved image sensors is a challenging task especially if TCSPC operation is to be implemented. Such pixels require complex electronics resulting in large pixel pitch, low fill factor and high data rates. Time-gating could solve some of these issues as the circuits are simpler, data is compressed in-pixel by integration and fill factor can be improved by optimised layout or by resorting to analogue designs. Yet analogue pixels suffer from noise, non-uniformity and low counting capacity.

While digital time-gated pixels also tend to be large, the adoption of advanced CMOS nodes benefits directly from the high integration density reducing the pixel pitch. More advanced technologies such as 3D-stacking allow for more flexibility, greater embedded functionality and higher sensitivity due to backside illumination. The following chapters will present designs and architectures that benefit directly from these technology advancements to achieve a miniature time-resolved image sensor.

3. High Fill Factor Global Shared Well FSI SPAD Image Sensor

A miniature 1mm², 96 × 40 front side illuminated image sensor implemented in an advanced 40nm technology is introduced. By expanding the global shared well layout technique and taking advantage of the fine design rules and high integration density of the process, a 39% to 66% fill factor SPAD array is realised coupled to an 8.25μm time-gated pixel with a 12-bit counter.

The first section of this chapter provides an overview of the different SPAD pixel layout strategies and modelling results for the attainable pixel pitch and fill factor based on specified design rules. This is followed by a feasibility study of the global well sharing technique in context of different process node parameters.

The second section provides a detailed account of the designed 96 × 40 array from system level down to the fine layout specifics. The optimised 12-bit counter pixel is also discussed with focus on area saving and the time-gated mode of operation which utilises an edge-to-edge gate generation technique for fine gate width resolution.

Finally optical and electrical characterisation results are presented highlighting the photo-response non-uniformity shortcoming of the global shared well array and the effectiveness of the time gating technique achieving a time gate FWHM of 360ps. Conclusions and a comparison to the state of the art are also presented.

3.1. SPAD Pixel Layout and Modelling

For the work presented throughout this thesis, several SPAD structures could have been pursued but the author opted for the Richardson p-well (PW) to deep n-well (DNW) junction SPAD [111] as the main choice for the following reasons:

1. The collective know how and experience within the CMOS Sensors and Systems group in implementing such a structure and the familiarity with its design rules since it has been conceived by previous group members. This knowledge reduces the risk factor associated with implementing the SPAD device in a new technology node (i.e. 40nm) for the first time.
2. Portability across different process nodes. Since it is considered a low risk device, it can be implemented in a variety of technologies (FSI and 3D-stacked BSI) which serves well for making comparisons of the same device structure across different platforms including previous implementations within the group in ST's 130nm imaging process [65][72] [167]. Moreover, this structure is proven to be scalable down to 3μm pitch [119] which bodes well with the objective of this work.

3. Taking advantage of process optimisation. Deep sub-micron technologies are known for their poor SPAD performance namely high dark count rate (DCR) due to the high doping concentrations and low breakdown voltage [104]; hence an advanced node such as 40nm needs process optimisation for SPAD implants to yield good performance devices. As part of developing their next generation time of flight (ToF) sensors ST have optimised the PW / DNW SPAD in their 40nm process, so using the same structure would leverage these technology improvements.

Despite the same origin of the SPAD device used in this work and the one commercialised by ST, the author would like to note that any SPAD devices or characterisation results presented are not necessarily identical or representative of ST's foundry offerings. The reader is referred to [114] and [113] for full accounts of ST's industrialised SPADs implemented in 40nm CMOS and 130nm imaging process respectively.

3.1.1. Different Layout Styles

The layout styles of FSI SPAD pixels can be generally classified into three different categories: standalone, local well sharing and global well sharing. In standalone layouts the SPADs are physically separated and share no common sides or implants while they do share electrical connectivity mainly through the high voltage net (VHV). In local well sharing layouts the SPADs share implants at two or more sides as well as electrical connectivity while being physically separated in groupings of rows or columns. In global well sharing the SPAD resources are fully shared on all 4 sides with no physical separation between them.

Five different pixel layout examples are discussed in this chapter for the purpose of making comparisons and these are:

1. Standalone circular SPAD [225].
2. Standalone rectangular SPAD [226].
3. Single strip local well sharing rectangular SPAD [181].
4. Double strip local well sharing rectangular SPAD [65].
5. Global well sharing square SPAD [149].

Figure 3.1.1 shows the layout of a 2×2 array of each of these examples for the same pixel electronics and focal plane areas.

The literature includes other layout techniques and geometries than the five mentioned such as honey-comb global shared well hexagonal SPADs for SiPM pixels [227] and macro-pixel structures where several SPAD devices form one pixel unit which is arrayed to form an image sensor [228]. Such layouts are not discussed further in this section as the focus is SPAD image sensors with one SPAD per pixel.

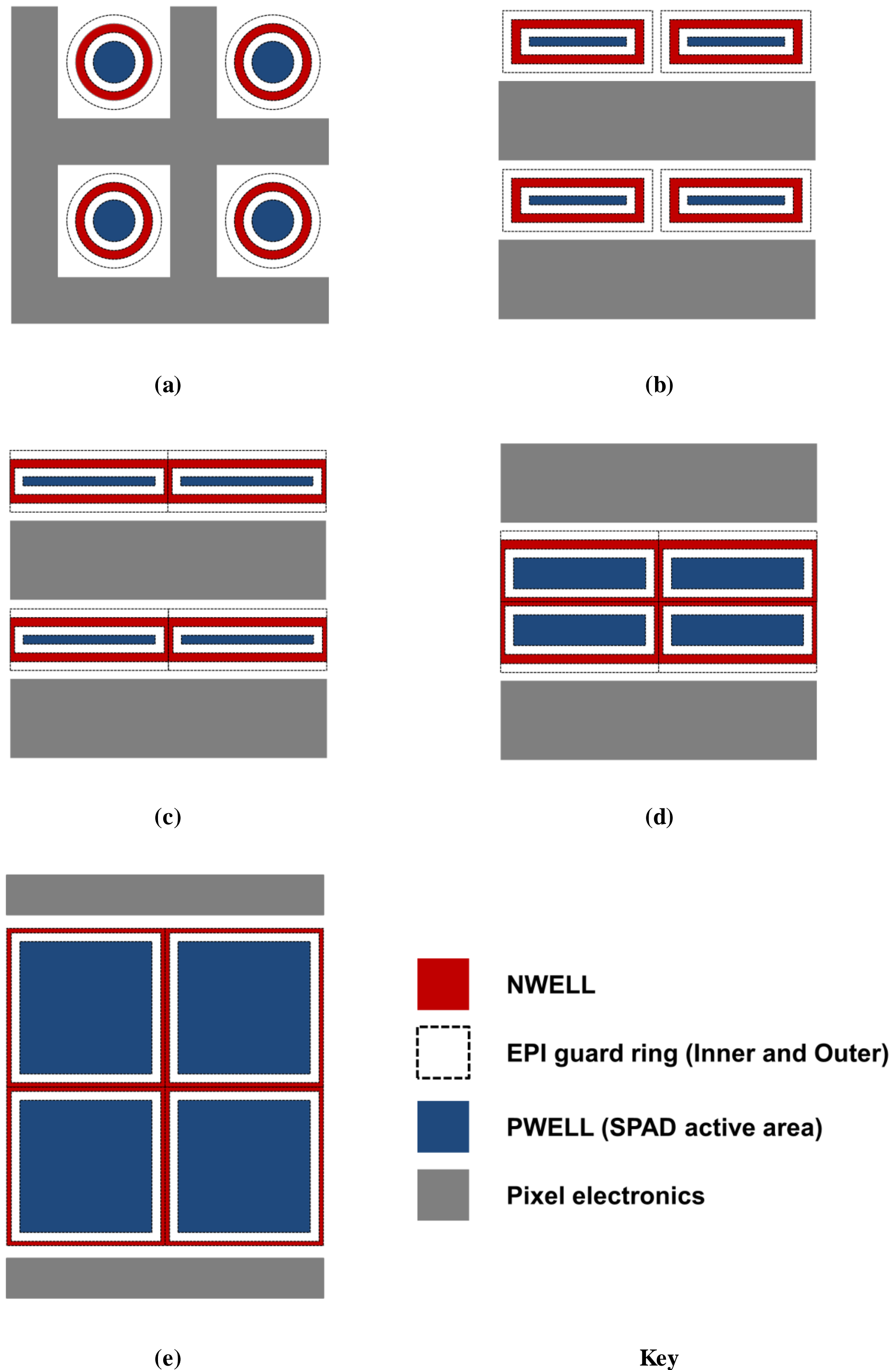


Figure 3.1.1. a) Standalone circular SPAD pixel. b) Standalone rectangular SPAD pixel. c) Single strip local well sharing rectangular SPAD pixel. d) Double strip local well sharing rectangular SPAD pixel. e) Global well sharing square SPAD pixel. The pixel electronics and focal plane areas for each of the 2 × 2 arrays is equal thus reflecting the scaling of fill factor with layout style.

As can be seen from Figure 3.1.1 the chosen layout technique has an immediate impact on the pixel's sensitivity or fill factor. Standalone pixels (Figs. 3.1.1(a) and (b)) have the lowest fill factor as the SPAD device guard ring consumes a big area in addition to the minimum spacing required between the SPAD's edges and nearby circuits or other SPAD devices. Single strip sharing (Fig. 3.1.1(c)) improves the area efficiency of the guard ring by sharing resources between adjacent SPADs and hence improves fill factor. Spacing rules still need to be respected between the SPAD and nearby circuitry along two edges. Double strip sharing takes a step further towards improving fill factor by sharing the SPAD resources along three edges and maintaining the spacing rule along one edge only.

Global well sharing (Fig. 3.1.1(e)) which is the subject of this chapter allows the SPAD to occupy the whole area of the focal plane and to share resources along the four edges of the device and so increase the fill factor even further. While such a layout is not easily scalable (as shall be discussed in Section 3.1.4) the fact that it decouples the circuit area from the SPAD area and eliminates the associated spacing rules between the two pixel components makes it a candidate for high sensitivity image sensor designs.

Other considerations when selecting the layout technique for SPAD pixels are optical crosstalk between SPADs and modulation transfer function (MTF) over the focal plane. Well sharing techniques necessitate packing SPADs in close proximity which increases the probability of crosstalk [68][133]. Certain applications with minimal crosstalk tolerance can benefit from physical isolation between SPADs (Fig. 3.1.1(a)) at the cost of lower fill factor [177]. Symmetrical layouts along both x and y axes (Figs. 3.1.1(a) and (e)) guarantee a uniform MTF along both dimensions while asymmetrical ones (Figs. 3.1.1(b), (c) and (d)) will correspond to a non-uniform spatial sensitivity distribution or MTF. While it has not been reported to be an issue for SPAD image sensors, future deployment of these sensors in scientific or consumer applications might require correction for such effects.

3.1.2. SPAD Guard Ring Design Rules

The SPAD design rules are a main factor in determining the fill factor of the pixel, so before modelling the relationship between fill factor and pixel pitch for the different layout styles these rules have to be established. The following rules are not specific to a process node but are derived from the cumulative experience in designing SPAD devices at the CMOS Sensors and Systems group as generic guidelines. Of course, they are applicable to the adopted SPAD structure in this work. Rules for other SPAD structures relying on physical trench or implantable guard rings can differ.

In reference to Figure 3.1.2 the generic design rules are:

a) Minimum PW (SPAD active area) width or diameter of 1 µm.
b) Minimum inner virtual guard ring (Inner EPI) width of 1 µm.
c) Minimum NW (SPAD cathode implant) width of 1 µm.

d) Minimum outer virtual guard ring (Outer EPI) width of 1µm.

e) Minimum spacing between the SPAD's Outer EPI and nearby active circuitry (NMOS devices) of 1µm.

f) Minimum spacing between the SPAD's NW cathode and nearby active circuitry (PMOS devices) or other nearby SPAD NWs of 3µm. This rule is also known as hot well spacing and varies with process node.

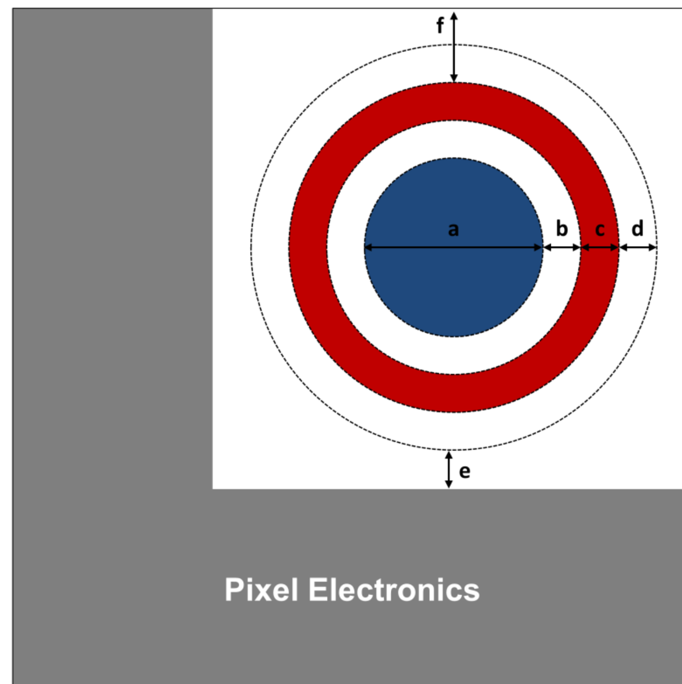


Figure 3.1.2. Generic SPAD design rules.

Design rules (a) and (b) have to be met for each SPAD individually no matter the layout style. For well sharing layouts, whether local or global, rule (c) has to be satisfied between two SPADs sharing a common edge, while rules (d to f) have to be satisfied on edges neighbouring the pixel electronics or other SPADs for local well sharing and standalone layouts respectively. Rules (d to f) also have to be met for global well sharing at the edges of the focal plane but have no impact on pixel fill factor due to the decoupling of the pixel electronics from the photo-devices.

3.1.3. Fill Factor and Pixel Pitch Trade-off Modelling

Given the rules above, the trade-off between pixel pitch and fill factor can now be studied in context of the five aforementioned layout examples (Fig. 3.1.1). Further assumptions are made by the author in the modelling comparison and these include:

1. 50:50 split between pixel area dedicated for the SPAD device and its associated spacing rules and the pixel area dedicated for circuitry.
2. A regular square shaped pixel with equal pitch in x and y dimensions.

The motive behind the first assumption is to give equal weighting to pixel sensitivity and pixel functionality which are both ideally desirable but more often get traded off against one another. Such cases include single photon counting (SPC) pixels [76] where small memory elements are implemented in-pixel with minimal circuitry thus putting more emphasis on fill factor. Event driven architectures [63][174] are another example of minimal circuit electronics implemented in-pixel while pushing the more complex processing elements to the edge of the focal plane. In a way such event driven architectures represent middle ground between local well and global well sharing. More processing intensive time correlated single photon counting (TCSPC) pixels prioritise circuit area over SPAD area giving more importance to in-pixel functionality [167][169][172].

The second assumption derives from the common practice in image sensor design where the pixel is equal in x and y dimensions. Nevertheless the availability of 3D-stacking led to FSI pixel designs that have twice the pitch along the x-axis as that of the y-axis in order to make them 3D-stacking ready. The area of pixel electronics is equal to that of the SPAD but the pixel is rectangular in shape. In such designs columns of pixel electronics and SPADs are interleaved such that the design can be ported to a 3D-stacked implementation by simply replacing the SPAD columns with pixel columns to double the horizontal resolution and by lifting the whole of the SPAD array to the top tier with a global shared well layout [170][191].

Accordingly the area occupied by the SPAD device for each layout style can be calculated as well as pixel pitch and fill factor. Two cases have been considered, NMOS neighbouring circuitry following rule (e) and PMOS neighbouring circuitry reflecting the hot n-well spacing rule (f). This highlights the impact of well spacing rules on minimum achievable pixel pitch and maximum attainable fill factor. For simplicity, perfect rectangular geometries were considered although in practice SPAD devices tend to have rounded corners to reduce the field intensity at sharp edges [153], therefore the fill factor calculations presented will be slightly higher than what would be implemented on silicon.

Figure 3.1.3 shows that with standalone layouts the minimum possible pixel pitch is 12.7μm at 0.48% fill factor while for a large pixel with a pitch of 30μm a fill factor of no more than 17.89% can be obtained.

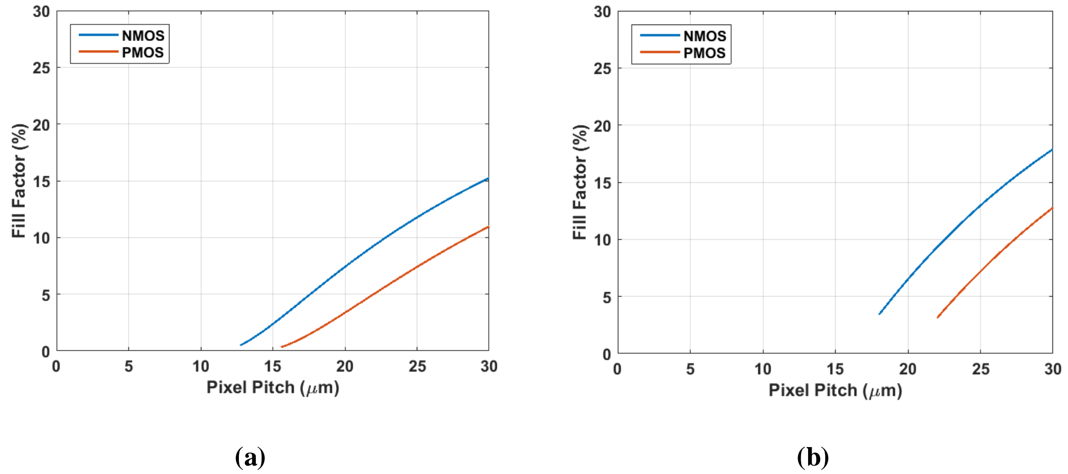


Figure 3.1.3. Fill factor versus pixel pitch trade-off for standalone SPAD layout pixels. (a) Circular standalone SPAD. (b) Rectangular standalone SPAD. Blue curves assume spacing rule (e) to nearby NMOS circuitry. Red curves assume spacing rule (f) to nearby PMOS circuitry n-well.

Figure 3.1.4 shows a similar comparison for local well sharing techniques. It can be seen that while single strip local well sharing does not push towards smaller pixel pitch compared to the rectangular standalone layout, it certainly improves the sensitivity such that the fill factor of a 30μm pixel increases to 21%. Alternatively, the more area efficient double strip local well sharing layout significantly improves fill factor of the same pixel to 28.5% while allowing for a minimum pixel pitch of 13μm but with a fill factor of 5.92% compared to less than 0.5% of a standalone circular counterpart.

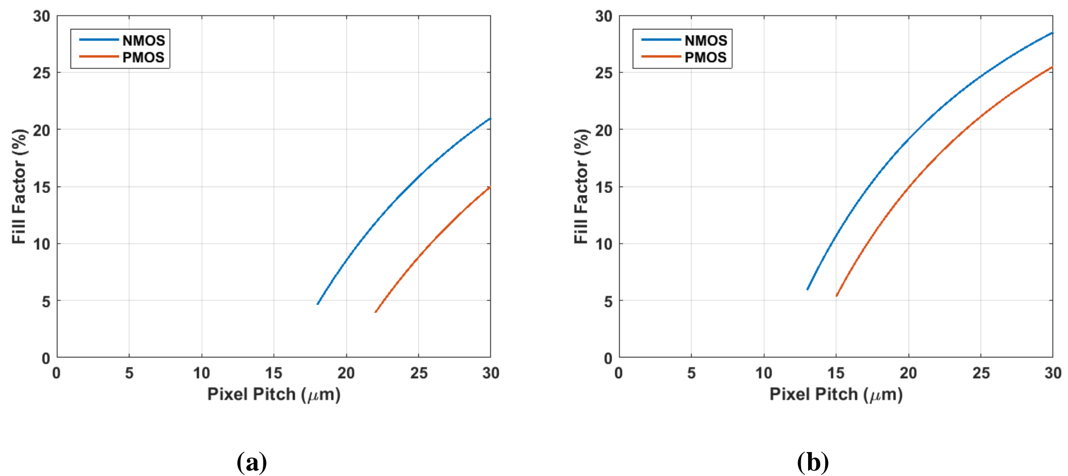


Figure 3.1.4. Fill factor versus pixel pitch trade-off for local well sharing SPAD layout pixels. (a) Single strip local well sharing SPAD. (b) Double strip local well sharing SPAD. Blue curves assume spacing rule (e) to nearby NMOS circuitry. Red curves assume spacing rule (f) to nearby PMOS circuitry n-well.

Global well sharing on the other hand provides the most promising parameters of all layout techniques. Figure 3.1.5 shows that a minimum pixel pitch of 4μm is possible while fill factors higher

than 60% are within reach for average size pixels of 15µm. This of course is expected since the focal plane is fully dedicated to the SPAD devices while all the circuitry is integrated on the edges, and hence the use of NMOS or PMOS devices has no impact on fill factor.

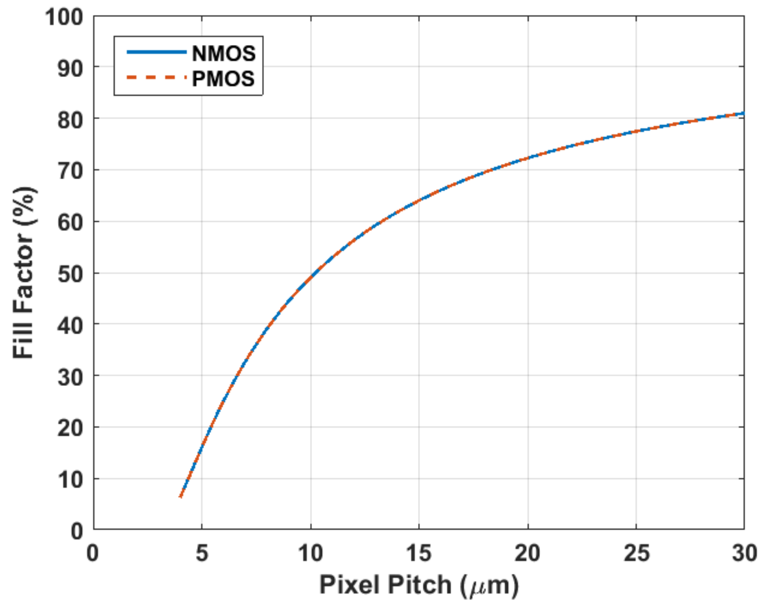


Figure 3.1.5. Fill factor versus pixel pitch trade-off for global well sharing SPAD layout pixels. Blue curves assume spacing rule (e) to nearby NMOS circuitry. Red curves assume spacing rule (f) to nearby PMOS circuitry n-well.

To put things into perspective, Figure 3.1.6 compares the five different layouts for an NMOS circuitry case. It is observed that the standalone circular layout allows for a pixel pitch between 13µm and 19µm when standalone rectangular and single strip local well sharing are not applicable, but from a pitch of 19µm onwards the single strip local well sharing provides higher sensitivity. Double strip local well sharing outperforms the aforementioned layouts but the ultimate winner in terms of minimum pitch and fill factor, is global well sharing.

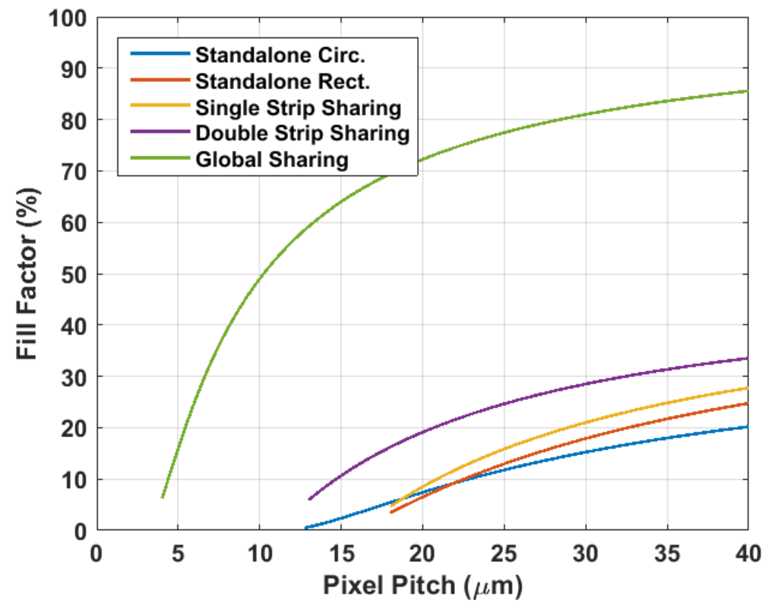


Figure 3.1.6. Fill factor versus pixel pitch trade-off comparison for the five different layout styles.

Since the objective of this analysis is to identify the most efficient miniature pixel design that achieves both high functionality and sensitivity in a restricted area, it is also worth considering the achievable focal plane resolution given any of the layouts above. Figure 3.1.7 provides this analysis assuming a total integrated circuit (IC) area of 1mm × 1mm with an area of 800µm × 800µm dedicated for the pixel array (electronics + photodiode). Here, the global well sharing layout does consume twice the area of other layouts since a pixel is made of two distinctive units of equal pitch, the SPAD and the circuitry. Despite that, global well sharing promises a higher resolution at a higher fill factor than any other layout.

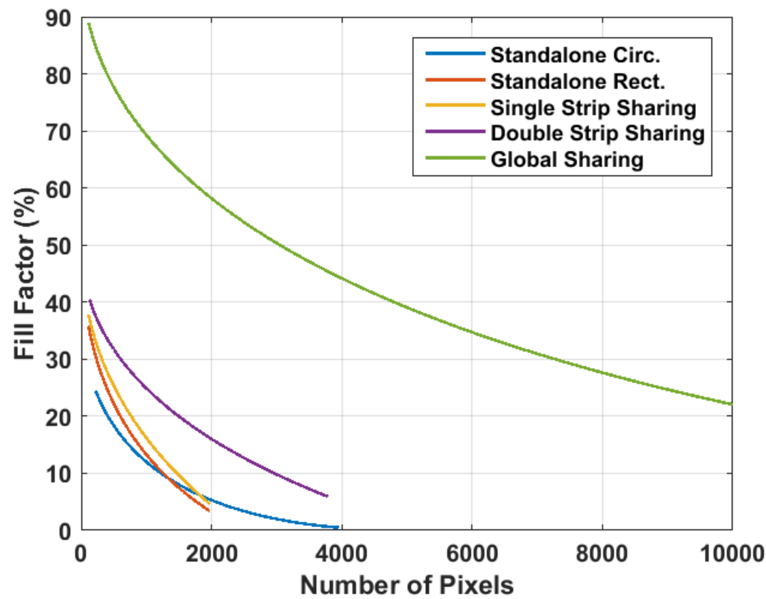


Figure 3.1.7. Fill factor versus resolution given an 800μm × 800μm focal plane area for the five different layout styles.

This analysis concludes by asking the following question:

- If global well sharing allows for smaller pixels, much higher fill factor and higher spatial resolution, why is it that its use has been so far limited to SiPM macro-pixels; whether in linear arrays [149][164][165] or coarse groupings [188][228], and not standard image sensor formats with one SPAD per pixel?

The answer is simply scalability and feasibility of a practical implementation.

3.1.4. Global Well Sharing Feasibility Analysis

A generic overview of global well sharing image sensors is shown in Figure 3.1.8. The imaging array consists of M horizontal pixels by 2 × N vertical pixels. The SPAD anode connections run over the shared guard regions between the SPADs to avoid obstructing incoming light over the active area. To maximise the resolution N rows of SPADs are routed upwards and N rows downwards where the pixel circuits are placed. For each half, the top most SPAD connects to the top most circuit block and bottom most one connects to the bottom most circuits block. This equalises the routing length of each SPAD anode.

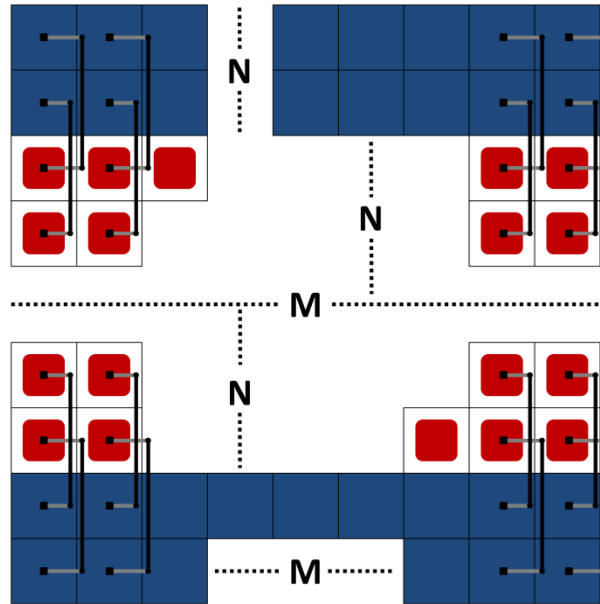


Figure 3.1.8. Generic layout of global well sharing image sensor with M columns and 2 × N rows.

It is immediately noted that such a layout easily scales in the x dimension but has several restrictions on scaling in the y direction dictated by the following:

1. Width of SPAD guard ring region.
2. Minimum metal routing pitch in the process.
3. RC parasitics of long anode connections and associated impact on SPAD dead-time and crosstalk between lines.
4. Availability of metal layers in the process.

The first restriction imposes a trade-off between pixel fill factor and resolution. For a fixed pixel pitch, a larger SPAD guard ring region (i.e. a smaller active area) would allow more anode connections per column which conflicts with the original motive behind global well sharing.

The second restriction is process node dependent. To most efficiently utilise the available SPAD guard ring area a smaller metal pitch is preferred which depends on the design rules of the accessible process node. This pitch or minimum spacing between metal tracks is also influenced by additional design rules of deep sub-micron (DSM) nodes which require a specific number of vertical interconnect access (via) connections to be used when a track length exceeds a certain dimension, and consequently imposes an additional spacing between parallel tracks of the same metal to ensure reliability of such vias.

The third restriction has two direct impacts on the pixel electrical properties. It is commonly known that long parallel tracks with high frequency signals such as SPAD pulses are prone to electrical

coupling meaning that one firing SPAD can crosstalk electrically to its neighbouring pixels through the long anode connections. This crosstalk is also dependent on the SPADs excess bias as in a global well sharing layout the anode tracks carry the raw SPAD pulse with a voltage swing dependent on excess bias. Also, since the anode connections are connected directly to the SPAD's moving node then any track RC parasitic will directly impact the SPAD's dead-time and so its maximum count rate or dynamic range (DR). This is particularly limiting for passively quenched SPADs. Such added parasitics will also influence the SPAD's jitter and after-pulsing as they add up to the SPAD's junction capacitance.

A layout strategy that would mitigate the former two restrictions is to use anode routes alternating between two different metal layers as depicted in Figure 3.1.9. The minimum metal pitch is maintained between interleaving metal layers at different levels of the metal stack such that the pitch is doubled between parallel metals of the same layer. This reduces fringe parasitic capacitance between tracks horizontally and vertically (diagonally) due to increased physical separation. This provides better immunity against crosstalk while satisfying additional spacing rules. This layout strategy will be assumed as the default in the following discussion.

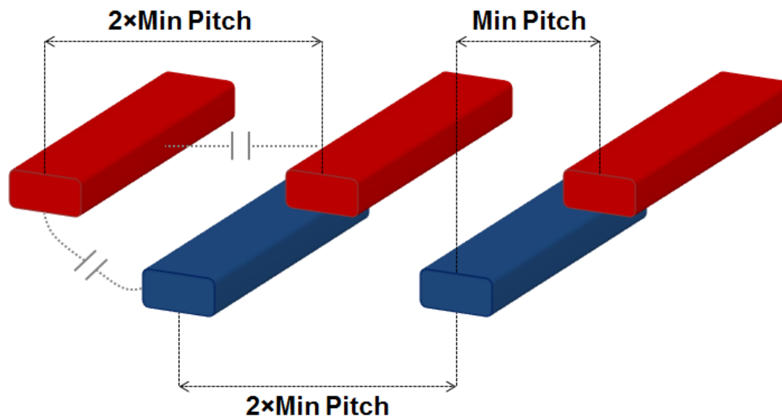


Figure 3.1.9. Alternating metals layout strategy.

This leads to the fourth restriction when using global well sharing layouts which is the availability of metal layers in the process. As shown in Figure 3.1.8 the anode routes not only go over the SPADs guard ring but also over the circuit array which means that at least over certain portions of the pixel connections cannot be made using the same metal layers dedicated for the anodes. Based on the routing layout strategy of Figure 3.1.9 this means two metal layers have to be spared or at least limited in use.

Referring back to the introduction of this chapter, it was stated that the end design presented herein is implemented in a 40nm (45nm with optical shrink) process node so considering this node as a

reference it is possible to project with some approximation the average number of metal layers available in different process nodes. According to Intel's interconnect scalability study presented at the International Electron Devices Meeting in 2016 [229], the average increase in metal layers per technology node step is 0.75 metals/technology. Considering the 40nm process in hand consists of 7 copper metals as a reference a projection is made in Figure 3.1.10.

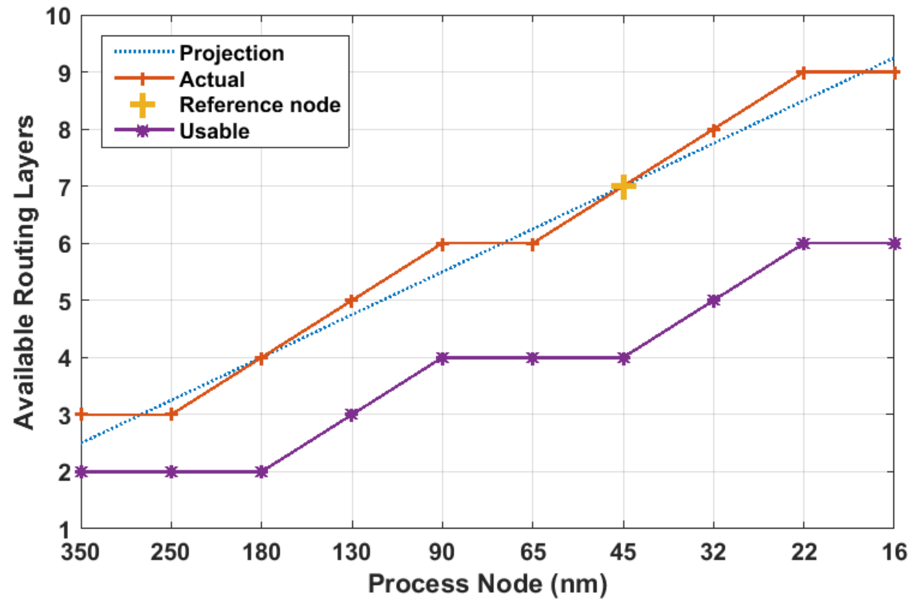


Figure 3.1.10. Projection of metal layers availability in different process nodes.

The dotted blue line is the projection starting from the reference and the red line is the approximate number of layers available after rounding the projected values to the nearest integer. To address the metal layers availability restriction of global well sharing it is not enough to consider the total number of metals available, but the number of metals that are usable. For instance the first metal layer (MT1) usually serves as the chassis or frame of standard cells without being particularly restricted to horizontal or vertical routing orientation and tends to have higher resistivity and so it cannot be used for routing. Higher metals up the stack (MT6 and MT7 in case of the reference 40nm node) are thick layers intended for power routing and have larger minimum width and pitch which renders them unusable for fine routing tasks. Thus the purple line represents the usable routing metals for each node based on the following assumption:

1. MT1 is restricted to standard cells.
2. Process nodes smaller than 65nm are assumed to have two thick metal layers.
3. Process nodes between and including 180nm and 65nm are assumed to have one thick metal layer.
4. Process nodes between and including 350nm and 250nm are assumed to have no thick metal layer.

5. Low resistivity thick aluminium power distribution tracks which might be available across all nodes is omitted as it also does not serve the purpose of fine routing.

Going back to the proposed routing layout strategy of Figure 3.1.9, if two metals are to be limited in use then at least two others are needed (1 horizontal and 1 vertical) for complete connectivity. This sets a target of at least four usable metal layers for global well sharing to be feasible. From Figure 3.1.10 it is concluded that only process nodes of 90nm or less are required for this purpose.

The author would like to point out here that the projection above is based on interconnect scalability of process nodes dedicated for ultra-large scale integration or processors which demand high density integration per unit area, high performance and low power consumption. In the field of imaging on the other hand the main requirement is optimised optical performance and so imaging dedicated nodes will have fewer metal layers than estimated in this analysis. This is intended to reduce the height of the metal stack to improve optical transmission of FSI designs [230] and reduce packaging or module height to suit the demands of consumer markets such as mobile phones.

Since the number of usable metals layers can only be met by advanced DSM nodes this implies that the cost of an FSI global shared well image sensor, albeit a miniature one, will be high as these nodes are expensive. Consequently because of the need for many metal layers, there will also be a cost to the optical performance or quantum efficiency (QE) though this can be compensated for by the big gain in fill factor, by an optimised process where cuts can be made to the passivation layers between metals over the focal plane for better transmission [230] and by micro-lensing to enhance the light collection efficiency [231][232].

While the analysis presented in the previous section has shown that at least mathematically global well sharing looks promising for the development of high resolution image sensors, the practical feasibility of this claim in the light of the restrictions presented has to be analysed further.

Similar to the projection of Figure 3.1.10, and based on Intel's study [229], the minimum metal pitch for each process node can be estimated (Fig. 3.1.11) assuming that metal pitch scales by a factor of 0.7 with every technology node step. The minimum pitch of the reference 40nm process node is 0.14µm.

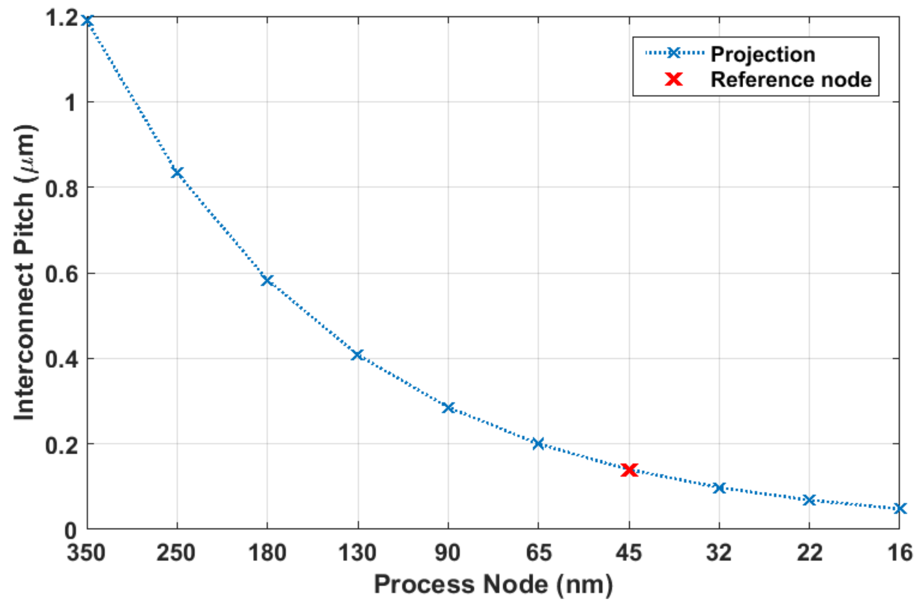


Figure 3.1.11. Projection of minimum metal pitch for different process nodes.

Using the numbers from Figure 3.1.11, the top and bottom routing strategy of Figures 3.1.8 and 3.1.9 and assuming the state of the art 8µm pixel pitch (SPAD and circuit), the total number of vertical anode connections can be estimated given different SPAD guard ring widths (Fig. 3.1.12). A comparison against an 8 µm backside illuminated (BSI) pixel is provided where routing can take place over the whole SPAD.

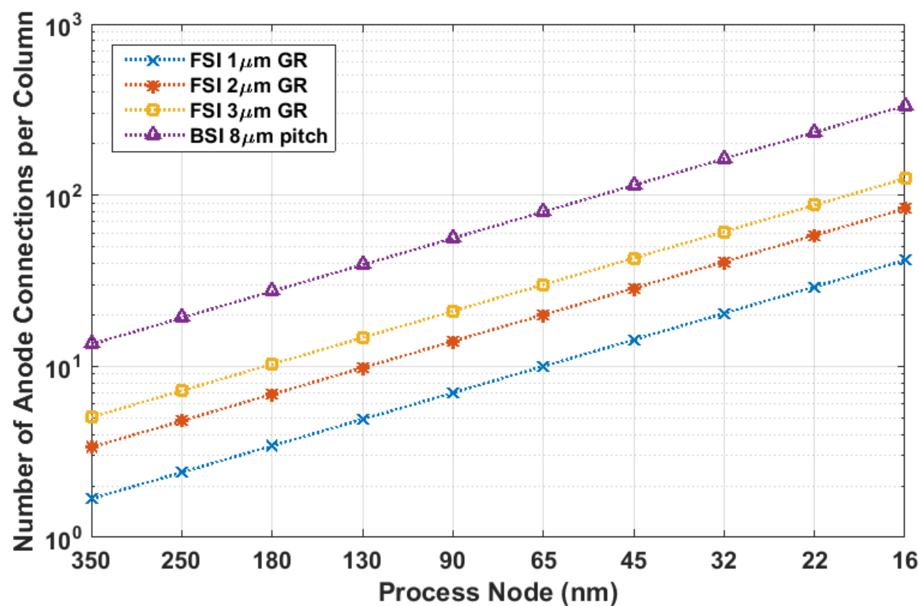


Figure 3.1.12. Estimated number of vertical anode connections in a global well sharing layout for different process nodes and SPAD guard ring widths. A comparison against a BSI implementation is also shown where routing can take place over the whole SPAD.

For a 40nm process, a maximum vertical resolution of 42 SPADs is estimated. This is a relatively low resolution even with such an advanced process node. While pushing towards more aggressive nodes can offer a small increase in vertical resolution (60 SPADs for 32nm node) this brute force solution starts to conflict with the third restriction of global well sharing; long route RC parasitics, let alone the increase in fabrication cost. For an 8μm pixel, an increase in the number of vertical units from 21 to 30 per half array means an increase in the anode route length from 168μm to 240μm with negative impact on crosstalk and SPAD dead-time.

Therefore the conclusion of the feasibility study presented in this chapter is that even with the most advanced process nodes, global well sharing is still limited in scalability and not a practical solution. Yet for applications such as endoscopy with a restricted IC area to start with, it is possible to implement a low resolution high sensitivity global well sharing sensor with careful design to avoid drawbacks such as crosstalk, exaggerated SPAD dead-time and unbalanced routing for uniform temporal performance across pixels.

3.2. 96 × 40 Image Sensor in 40nm

Building on the study above, a global shared well SPAD image sensor was implemented in STMicroelectronics’ front side illuminated 40nm CMOS process. The sensor, which will be referred to as MINIC40, has a resolution of 96(H) × 40(V) at a pixel pitch of 8.25μm. The horizontal resolution was derived from the estimated area allocated for the pixel array (SPAD + circuitry) given a 1mm × 1mm total IC area and taking into account the area required by peripheral circuits and the padring. The vertical resolution stems from the presented modelling for a shared guard ring region of 3μm, or a safe SPAD structure.

The adopted pixel pitch was chosen for the following reasons:

1. Comparison to state of the art since the smallest reported pitch for a SPAD image sensor is 8μm [65] so a pixel with a similar pitch allows for evaluation of how much more functionality can be integrated in-pixel in a 40nm process with the circuits decoupled from the SPAD in a global shared well layout.
2. With the prospects of 3D stacking becoming available in the near future at the time of design, a pixel with such a pitch could be ported with minimal modifications while anticipating a minimum hybrid-bond pitch for the stacking process to be approximately 8μm.
3. Knowing that the pixel layout will rely on standard logic cells and knowing their dimensions, a pixel pitch of ~8μm will allow for integrating at least a 10-bit ripple counter which enables a simple characterisation platform for the newly trialled 40nm process SPADs. The pixel’s vertical pitch has to accommodate an integer multiple of standard cell heights whilst the horizontal pitch should preferably be a multiple of the minimum standard or filler cell width such that pixels abut easily when arrayed.

3.2.1. System Overview

The MINIC40 system has been kept intentionally as simple as possible as the main requirement is to provide an evaluation platform for SPADs in 40nm and to serve as an exploratory sensor for the design capabilities of the new process. This also reflects on the simplicity of the pixel circuit which will be discussed in Section 3.2.4. The simple design also fast tracks the firmware design and bring up process and ensures ease of use.

Figure 3.2.1 shows the block diagram of MINIC40 IC alongside a chip micrograph. The 96×40 SPAD array lies in the centre of the sensor with top and bottom 96×20 pixel circuit blocks. Each half of the array is served by control signals buffers, a single channel serial readout interface and a bank of balanced binary clock trees for distributing the required time gating signals. A row scanner addresses the rows for readout in a rolling fashion by propagating two tokens, one for each half, from the centre of the array outwards. All control signals and voltages are supplied externally through twenty six pads arranged in an L-shape configuration, and critical timing signals are buffered from the source pad in a balanced way to the input stages of the top and bottom clock trees.

While the split in the array pixels necessitates duplicating circuit resources, the area and power consumption trade-offs are not straightforward. Control signals buffers and serial readout interfaces are relatively small blocks and their power consumption is insignificant. Control buffers are mostly static and readout is operated intermittently with other factors such as clock speed and number of Input / Output (IO) pads dominating the power figures.

With clock trees on the other hand, the area consumed is big due to the tree arrangement so for a miniature implementation of a global shared well sensor there is an added area overhead in this respect. As for power consumption, although twice the number of elements is needed to propagate timing signals across the array, the column load driven by each is halved requiring smaller drive strength and so balancing the power consumption. Of course a detailed analysis is needed when designing a final system and no more emphasis was given to this design concern for the purpose of this demonstrator chip.

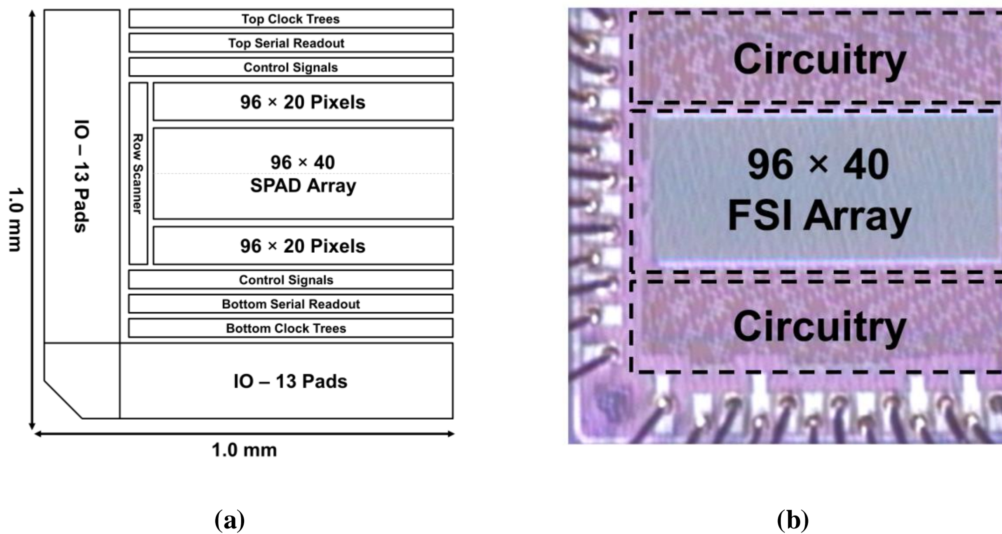


Figure 3.2.1. MINIC40 sensor. (a) Block diagram. (b) Chip micrograph.

3.2.2. Layout Overview

Starting with the building block of the sensor, the pixel, Figure 3.2.2 shows the pixel's layout which is made of six rows of thin oxide standard logic cells and a seventh row shared between standard cells and thick oxide transistors. Every other digital row is mirrored along the x-axis in order to share supply (VDD) and ground (GND) tracks. The vast difference between density and feature size of digital logic and analogue transistors is apparent.

The pixel front end consists of the thick oxide transistors M0, M1 and MQ. MQ acts as a passive quench passive recharge circuit with its resistance being controlled by a bias voltage applied to its gate. This is the smallest quench arrangement possible which is necessary for keeping the pixel size small. Transistors M0 and M1 are an NMOS and a PMOS respectively forming the front end inverter which receives the SPAD's pulse. MD is a thin oxide transistor connected as a decoupling capacitor of approximately 11fF between VDD and GND as the best use of extra space available in the pixel.

Only metal 1 (MT1) connections (dark blue) are shown in Figure 3.2.2 for clarity as the circuit connectivity is quite complex and spans seven metal layers. As mentioned earlier, the use of MT1 as the frame for standard cells renders it unusable for other routing tasks.

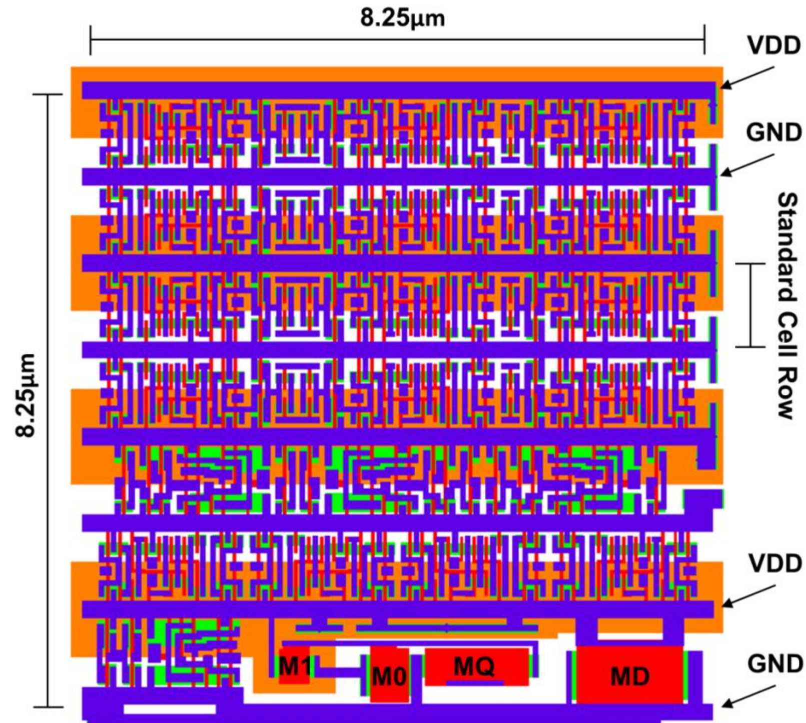


Figure 3.2.2. MINIC40 pixel layout. Green is active area (OD), orange is n-well (NW), red is poly-silicon (PO) and dark blue is metal 1 (MT1).

Figure 3.2.3 on the other hand highlights the complexity of the intra-pixel connectivity through metals 2 and 3 (MT2 and MT3) in light blue and pink respectively. MT2 is used horizontally and MT3 vertically with common row control signals routed in MT2 highlighted. Referring back to the analysis before it is evident that the routing complexity of high density digital logic in deep sub-micron processes requires at least two dedicated metal layers.

Other than the intra-pixel connections, many others remain to connect the pixel as a unit of an imaging array such as column control signals, additional power strapping and column parallel output bus. It is not possible to use MT2 and MT3 for all these purposes (some power strapping can be seen on edge of pixel in Fig. 3.2.3) so additional metal layers have to be utilised. Figure 3.2.4 shows the global vertical grid mainly in metal 5 (MT5) in yellow over the pixel. Some metal 4 (MT4) in green is also used for additional intra-pixel connections to route MT5 signals to lower levels in the pixel.

It is important to note how MT4 and MT5 are confined to a restricted area within the centre of the pixel and how no MT4 or MT5 routing exists on left and right edges. This is absolutely crucial for the global well sharing layout to be achieved.

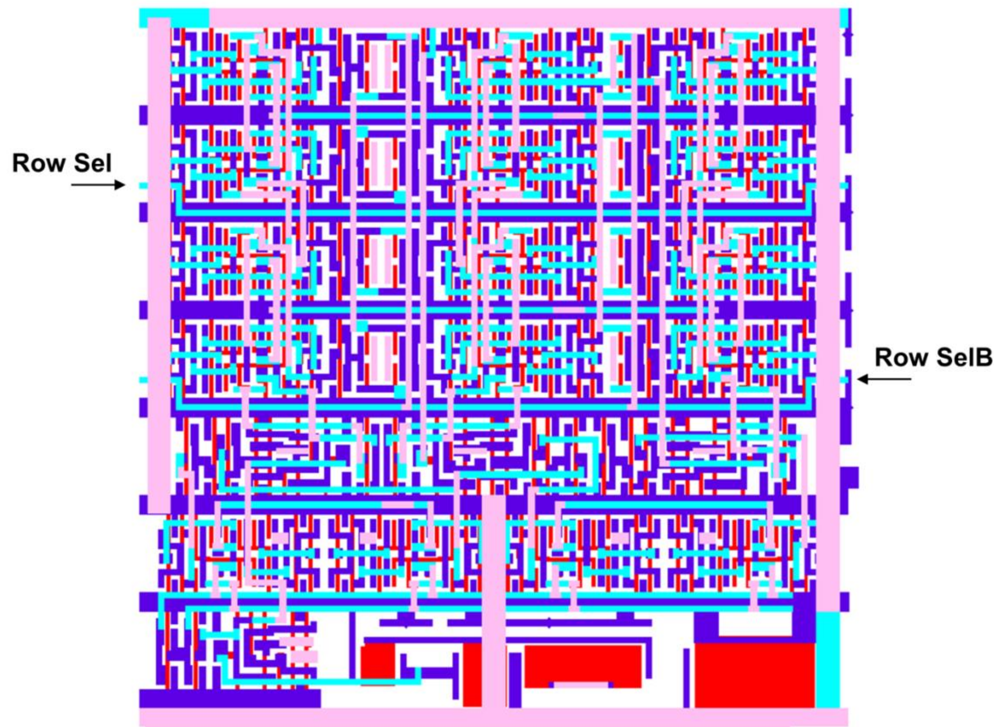


Figure 3.2.3. MINIC40 intra-pixel routing layout. Red is poly-silicon (PO), dark blue is metal 1 (MT1), light blue is metal 2 (MT2) and pink is metal 3 (MT3). Horizontal row signals are annotated.

To form an image sensor, the pixel is simply arrayed horizontally while sharing VDD and GND tracks and row controls with its neighbours. Similarly, it is possible to array it vertically while sharing column signals and the output bus but every other row of pixels has to be mirrored along the x-axis. As seen in Figure 3.2.2 the bottom most track of the pixel is GND while the top most is VDD which requires mirroring rows in order to share resources and avoid shorts. Figure 3.2.5(a) shows a 2×2 pixel array as an example. A consequence of this mirroring is that it introduces a systematic offset in the anode connection length between even and odd rows.

Figure 3.2.5(b) highlights the layout of MT4 and MT5 tracks when the pixel is arrayed. Between every two pixels there is a noticeable channel which is free of any routing in MT4 or MT5 that will be occupied by vertical anode connections coming from the globally shared SPADs in these two metal layers. As concluded in the previous section, a total of at least four usable metals - MT2 to MT5 in this implementation - is a requirement for global well sharing arrays.

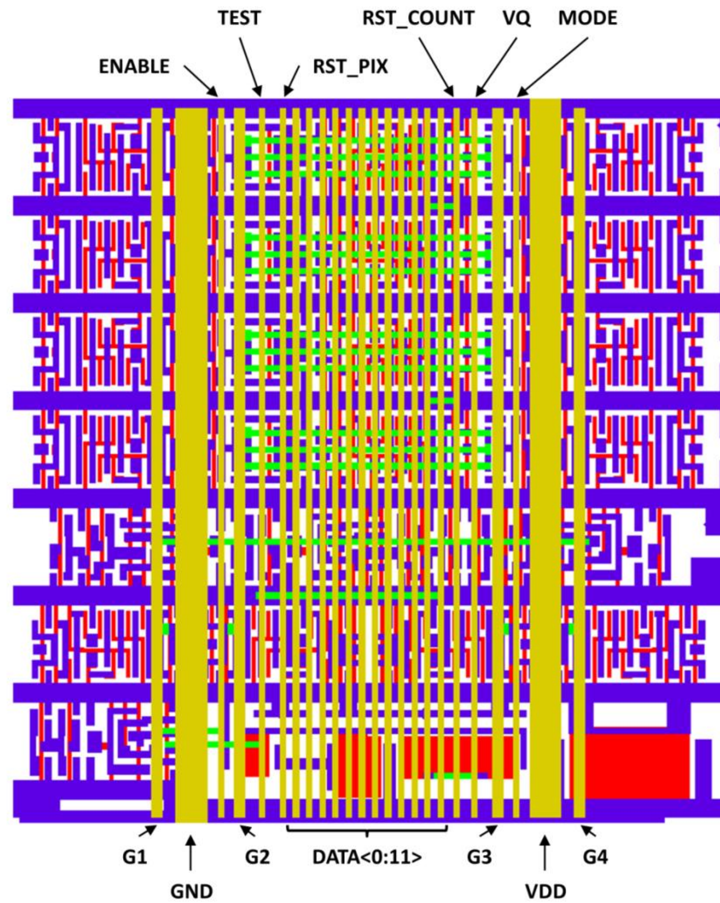
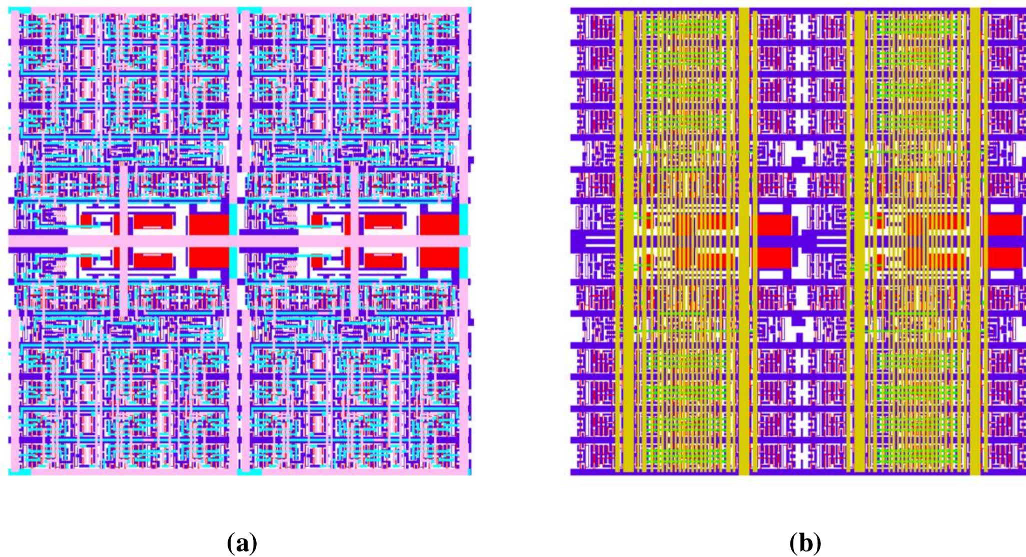


Figure 3.2.4. MINIC40 global vertical routing layout. Red is poly-silicon (PO), dark blue is metal 1 (MT1), green is metal 4 (MT4) and yellow is metal 5 (MT5). Vertical control signals are annotated.



(a) **(b)**

Figure 3.2.5. MINIC40 2 × 2 pixel array. (a) With MT2 and MT3 routing visible showing how the pixel arrays by mirroring along the x-axis. (b) With MT4 and MT5 visible with the MT4 / MT5 free channel between the pixels noticeable. Vertical anode connections in MT4 and MT5 would occupy this space.

The global well sharing design is more clearly depicted in Figure 3.2.6. The anode routes coming from the SPAD array in MT4 are shown for the bottom right corner of the array. The vertical anode routes pass in between pixels through the dedicated channel avoiding other horizontal MT4 intra-pixel connections which were confined to the centre of the layout. Alternating MT5 tracks come down the same way but are switched off in the figure for clarity.

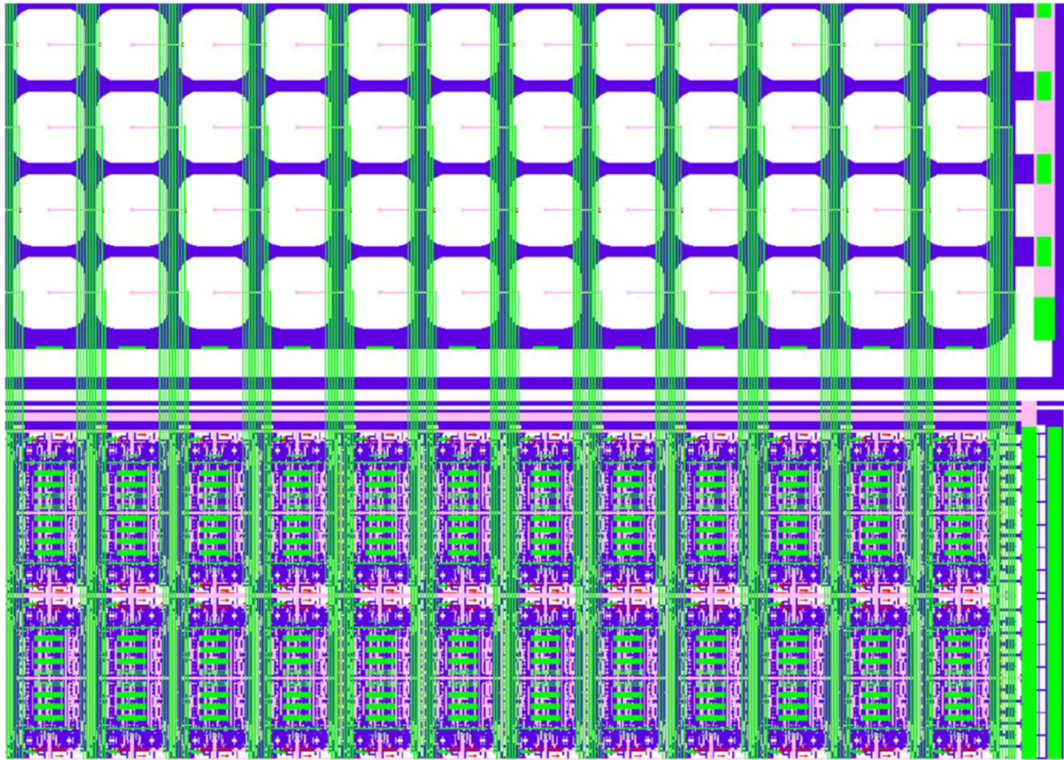
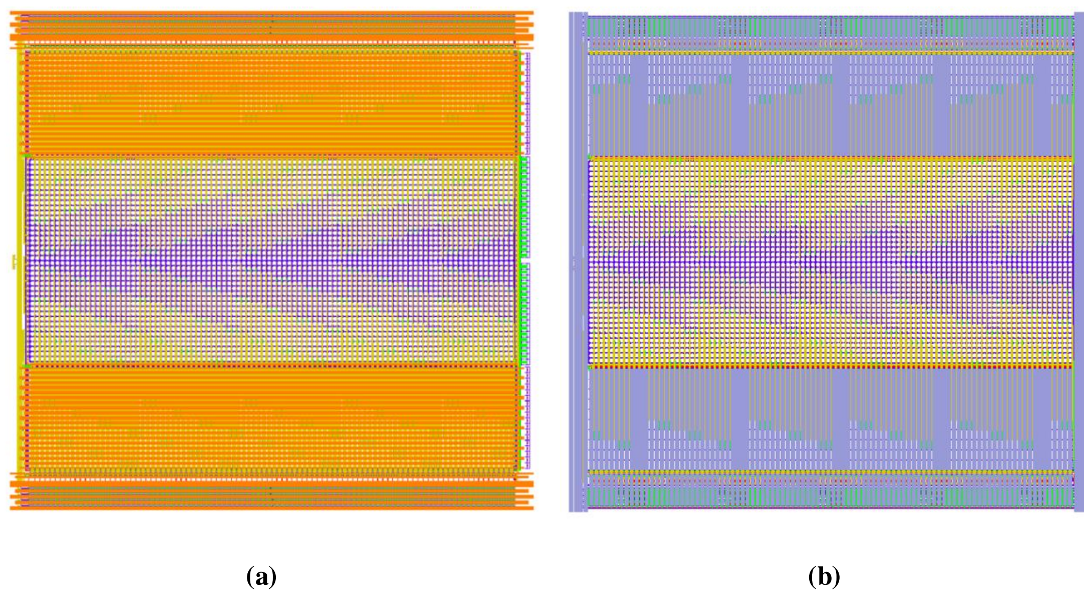


Figure 3.2.6. Global shared well layout of the bottom right corner of the array. Vertical anode MT4 tracks (green) from the SPADs flow in between pixels in the dedicated channel avoiding other horizontal MT4 routing used in-pixel.

Finally, the thick metal layers 6 and 7 (MT6 and MT7) available in the process were used for power routing over the whole image sensor except of course over the SPAD array region. Figure 3.2.7 shows the horizontal power strapping in MT6 and the vertical one in MT7. Figure 3.2.8 shows the overall IC which was fabricated.

The full design is 2mm × 2mm containing four 1mm × 1mm MINIC40 trials within one sealring. The trials are identical except for the SPAD layout and are completely independent and share no electrical connectivity. Every trial was rotated by 90 degrees counter clockwise inside the sealring such that by rotating the packaged IC on the test board immediate access is gained to a new trial without having to reconfigure the voltage supplies or FPGA digital ports.



(a) **(b)**

Figure 3.2.7. (a) MINIC40 horizontal power strapping in MT6 (orange). (b) MINIC40 vertical power strapping in MT7 (grey). No MT6 or MT7 is used over the SPAD array region.

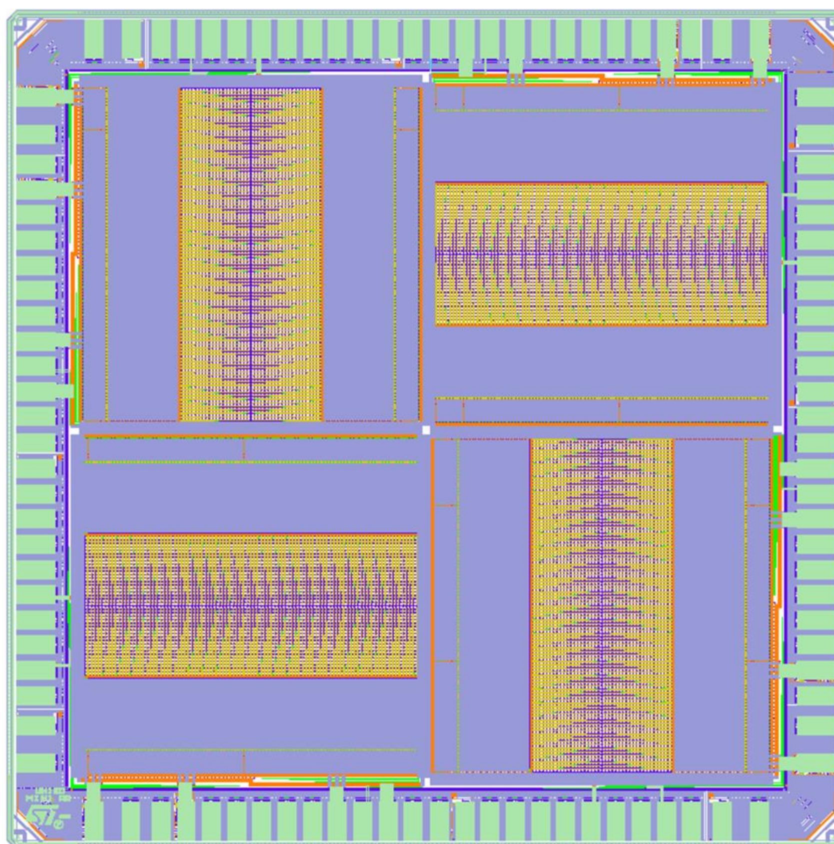


Figure 3.2.8. MINIC40 overall IC with four independent trials.

3.2.3. SPAD Trials

For each of the MINIC40 trials, the SPAD guard ring dimensions (EPI and NW) were slightly varied in order to investigate the limits of the new technology. The overall SPAD structure, layers used and general overlap and spacing rules were all maintained and so was the device pitch. Figure 3.2.9 shows the difference between the four implementations and Table 3.2.1 summarises the guard ring and fill factor figures.

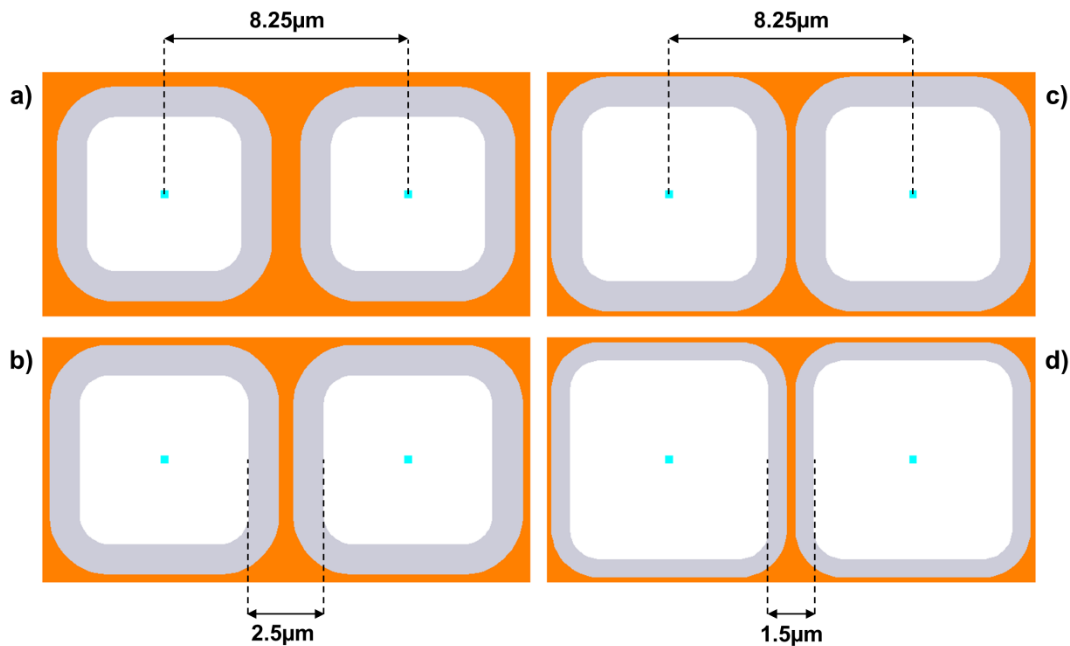


Figure 3.2.9. MINIC40 SPAD trials at fixed 8.25µm pitch. Shared NW is in orange, virtual EPI guard ring is in grey and anode connection is in light blue. Encapsulated white area represents the SPAD's active region (a) Reference trial with 3µm guard ring region. (b) Reduced shared NW. (c) Minimum shared NW allowed by process design rules. (d) Aggressive 1.5µm guard ring region trial with minimum shared NW and reduced EPI width.

Figure 3.2.9(a) shows the reference trial with a 3µm guard ring region (EPI and NW) which is the basis of the MINIC40 sensor design. The shared NW between SPADs is reduced in the next two trials (Figs. 3.2.9(b and c)) with the EPI width maintained as that in the reference trial. The shared NW in the third trial (Fig. 3.2.9(c)) is draw to the minimum allowed by the process design rules. For the last trial (Fig. 3.2.9(d)) the minimum shared NW is used while also reducing the width of the EPI layer. All the experimental trials result in an increased SPAD drawn fill factor albeit the effective pixel fill factor is less than the drawn since anode connections start to obstruct the SPAD's active area.

Trial	Guard Ring Width	SPAD Drawn Fill Factor
1 (Reference)	3μm	39%
2	2.5μm	48%
3	2.3μm	51%
4	1.5μm	66%

Table 3.2.1. Summary of MINIC40 SPAD trials guard ring dimensions and drawn fill factor.

3.2.4. Pixel Circuit Design

Apart from compactness to meet the required pitch of 8.25μm, the pixel circuit was designed with three other objectives in mind:

1. Simplicity. This keeps the design risk low and speeds up the bring-up process.
2. Dynamic range. The pixel should have a reasonable dynamic range in photon counting mode of at least 10 bits (~60dB) for it to function as an image sensor and to act as a characterisation platform for the newly trialled SPAD devices.
3. Time gating capability. This is one of the core reasons behind using SPAD devices and a main project requirement for time-resolved imaging applications.

The first objective was met by adopting an all-digital design relying mostly on standard cells. Digital logic is easy to verify and to reconfigure and standard cells are reliable and tend to function as expected. All digital designs also avoid the complexity and variability of analogue circuits and their associated readout chains.

The second objective was met by integrating a 12-bit digital ripple counter within the 8.25μm pixel thanks to the high integration density of 40nm CMOS. The large counter depth was also possible due to a customised flip-flop design with smaller area footprint. Being digital, the pixel counter suffers from no accumulation noise and no added readout noise as in analogue implementations [74]. Moreover, a digital counter provides a readily digitised output eliminating the need for complex analogue to digital converters (ADCs) which consume power and area.

Similarly the time gating capability was implemented by using logic gates. A time gating approach was preferred over time stamping or time correlated single photon counting (TCSPC) in order to maintain pixel simplicity, leave enough area for counter bits and to avoid high throughput parallel readout to keep up with the high data rates generated by TCSPC systems which conflicts with the miniaturisation target of this work.

In parallel to the MINIC40 design, another project pursued within the group implemented in 40nm CMOS was aimed at realising a TCPSC image sensor based on a ring oscillator design [170] so a time-gated system would provide an alternative approach and avoid duplicating efforts.

The resulting circuit shown in Figure 3.2.10 consists of three main blocks namely:

1. Thick oxide transistors front end and quench.
2. Configurable 12-bit ripple counter.
3. Time gating logic.

The pixel has one voltage supply (VDD), two row control signals, nine column control signals and one bias voltage as labelled in Figures 3.2.3 and 3.2.4. A twelve bit column parallel bus delivers the pixel data to the readout chain. Table 3.2.2 lists all voltage and control signals along with their description.

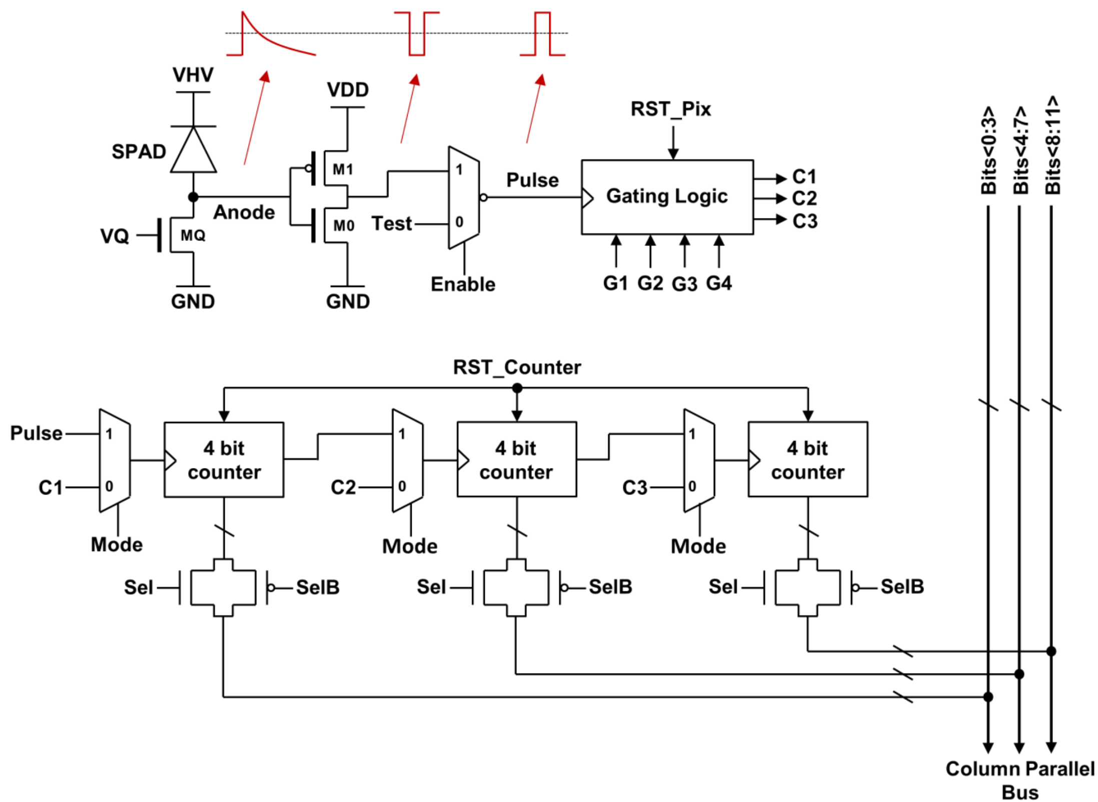


Figure 3.2.10. MINIC40 pixel circuit block diagram with thick oxide transistors MQ, M0 and M1 forming the quench and front end inverter followed by thin oxide 40nm CMOS logic. The front end waveforms and their polarities are indicated in red. The front end output feeds the time gating logic block and a configurable 12-bit ripple counter which outputs its content on a 12-bit column parallel bus.

Signal	Type	Source	Direction	Purpose
GND	Analogue	Global	In / Out	Ground
VDD	Analogue	Global	In / Out	Supply 1.1V
VQ	Analogue	Global	In / Out	Quench transistor (MQ) bias
VHV	Analogue	Global	In / Out	SPAD high voltage supply
BITS<0:11>	Digital	Column	Output	Data from pixel
ENABLE	Digital	Column	Input	Exposure control, active high
G1 to G4	Digital	Column	Input	Time gate generation signals
MODE	Digital	Column	Input	Operation mode, photon counting active high, time gating active low
RST_COUNTER	Digital	Column	Input	Reset signal for counter
RST_PIX	Digital	Column	Input	Reset signal for gating logic
SEL	Digital	Row	Input	Row select for readout
SELB	Digital	Row	Input	Row select bar for readout
TEST	Digital	Column	Input	External test pulse for debug

Table 3.2.2. MINIC40 pixel supply and control signals.

3.2.5. Pixel Front End Design

The pixel front end is what receives the SPAD pulse and conditions it for processing by downstream electronics. Typically three main functions are expected from the front end:

1. Quench and recharge the SPAD device. This can either be achieved asynchronously in a free-running fashion or synchronously with respect to a system clock by charging and discharging (arm / disarm) the SPAD through control signals.
2. Level-shift the SPAD pulse to correct voltage level. The voltage level of the SPAD pulse depends on the excess bias or voltage value above the SPAD's breakdown voltage which can be higher than the voltage rating of the pixel processing circuits, therefore the front end should pass on the SPAD pulse at an acceptable voltage for the backend electronics.
3. Mask the SPAD. This means disabling the SPAD output whether by disabling the SPAD device itself, or by electrically blocking the propagation of SPAD pulses down the circuit chain.

Synchronous quench and recharge operation is usually preferred in implementations that are either limited by the pixel counter bit depth or by in-pixel processing throughput. For example in oversampled SPAD image sensors [80] the pixel contains a 1-bit memory element which indicates whether no photon (logic low) or at least one photon (logic high) has been detected. In such a scenario the first arriving photon within an integration period saturates the pixel and subsequent photon arrivals are undetected. It is possible therefore to synchronously charge / discharge the SPAD by means of control signals once every integration period.

Similarly in TCSPC pixels [212] where the first arriving photon occupies the time to digital converter (TDC) and subsequent photons are unlikely to be processed, a synchronous quench and recharge operation with respect to the TDC timing is feasible.

In this design a free running asynchronous operation is preferred due to the large counter depth where many photons within the integration period can be counted. Also, a free running front end eliminates the need for routing control signals across the array since layout resources are limited.

In the interest of keeping the circuit as compact as possible, an optimum passive quench and recharge was implemented by means of a single thick oxide NMOS transistor with controllable on resistance through an externally supplied bias voltage VQ.

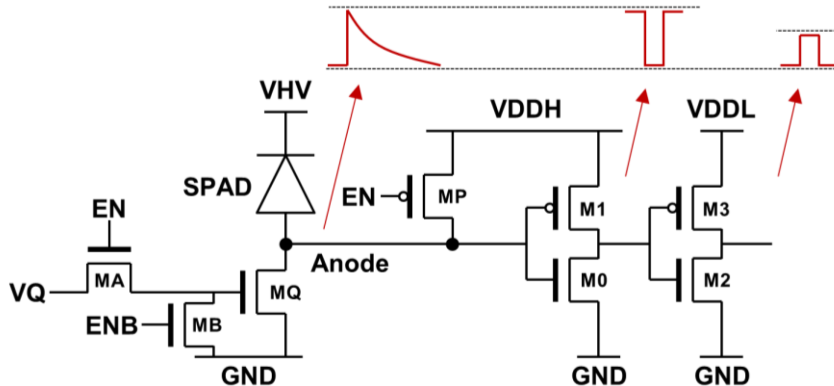


Figure 3.2.11. Thick oxide front end example with SPAD pull-up PMOS and level shifter. Red waveforms show the polarity and signal height at different nodes.

An example of thick oxide front end circuit with both SPAD masking and level shifting functions is shown in Figure 3.2.11. When the pixel is enabled (EN is logic high, ENB is logic low) pull up PMOS MP and NMOS switch MB are disabled while NMOS switch MA connects VQ to the gate of quench transistor MQ. The inverter formed by M0 and M1 receives the anode pulse and is operated at supply of VDDH. VDDH is set to be slightly higher than the SPAD excess bias voltage and sets the inverter threshold to approximately half the anode pulse height.

The second inverter formed by transistors M3 and M4 operates at lower supply VDDL and shifts down the output of the first inverter to the required voltage by following processing electronics. When the pixel is disabled (EN is logic low, ENB is logic high) switch MA cuts off VQ while switch MA shorts the gate of transistor MQ to ground. The pull up transistor MP in turn shorts the SPAD anode to VDDH and since VDDH is slightly higher than the SPAD’s excess bias, the resulting potential across the SPAD is slightly lower than its breakdown voltage turning it off.

There are several issues in the given circuit when it comes to miniature pixel implementations. First of all, there are two different voltage domains in-pixel which means PMOS devices cannot share n-wells and so hot n-well spacing rule has to be obeyed increasing the area footprint. Secondly, there are two sets of thick oxide inverters which are relatively large in size and consume area. The same applies to

switches MA, MB and MP. Finally, since it is possible to selectively enable or disable the pixel, a memory element is required to hold the value of EN and ENB. This memory element, be it a six transistor static random access memory (SRAM) or a latch, needs to be thick oxide to operate with VDDH supply consuming even more area.

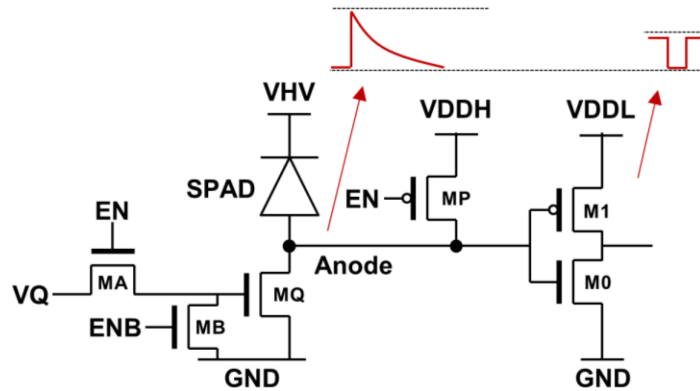
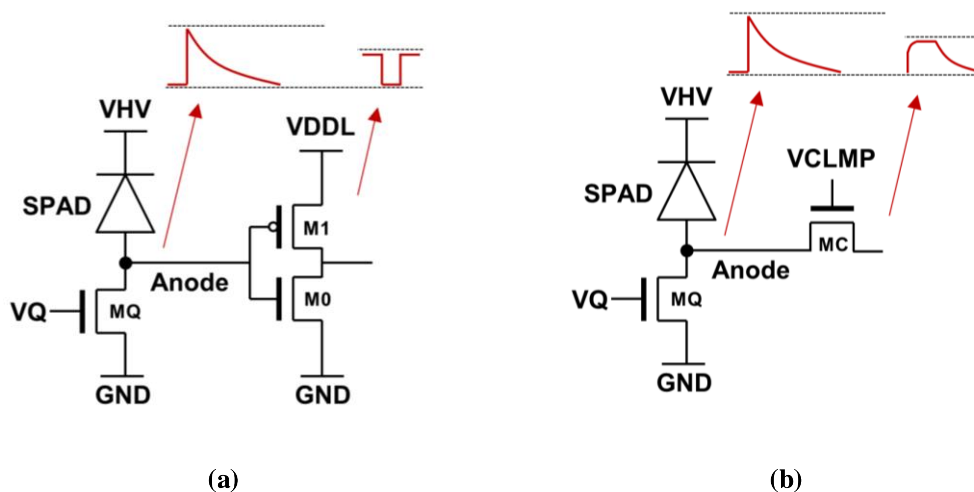


Figure 3.2.12. Thick oxide front end example with SPAD pull up PMOS and direct level shifter. Red waveforms show the polarity and signal height at different nodes.

A more economic implementation would be the circuit shown in Figure 3.2.12. Here only one inverter is needed and it operates at VDDL directly providing a level shifted output. The threshold of the inverter is now fixed at approximately half of VDDL which means that the threshold does not track the SPAD pulse height. While two thick oxide transistors were eliminated the rest of the issues with the circuit in Figure 3.2.11 still persist.



(a) **(b)**

Figure 3.2.13. Thick oxide front end example only with direct level shifter. (b) Thick oxide front end example with voltage clamp. Red waveforms show the polarity and signal height at different nodes.

To fully alleviate the area constraints, the option to turn off the SPAD was fully eliminated as shown in Figure 3.2.13(a) where only a single inverter operating at VDDL is used. This is the simplest configuration but comes at the cost of not being able to disable SPAD devices. Such an option is important to save power, reduce noise on VHV and to prevent crosstalk by turning high DCR SPADs off. Nevertheless it is a necessary trade-off in this scenario.

An even simpler circuit was proposed by [233] (Fig. 3.2.13(b)) whereby an NMOS clamp replaces the inverter and by setting its gate voltage to a bias of VCLMP, all anode pulses higher than VCLMP less a MOS threshold voltage V_T will be skimmed thus controlling the voltage of the signal propagating through. Since it was possible to integrate an inverter in the available area for thick oxide transistors in the pixel (Fig. 3.2.2) while setting VDDL to the same pixel supply VDD and thus sharing the n-well of transistor M1 with thin oxide logic, the circuit in Figure 3.2.13(a) was adopted. This also avoids the need of routing an additional bias (VCLMP) and an additional voltage pad in the IC padring. Thick oxide transistor sizes used are summarised in Table 3.2.3.

Transistor	Width (µm)	Length (µm)
MQ	0.32 (Min)	1.5
M0	0.6	0.55 (Min)
M1	0.32 (Min)	0.44 (Min)

Table 3.2.3. MINIC40 pixel front end thick oxide transistors size summary.

Minimum width device was used for MQ but the length was increased to increase the channel resistance to roughly 200kΩ at 1.1V quench bias. Minimum length devices were used for M0 and M1 with the width of M1 kept at minimum and the width of M0 increased as allowed by area to improve the pull down gain. Since the SPADs rising edge which encodes the timing information translates into a falling edge at the inverter's output, only the NMOS device was optimised.

To evaluate the effect of having a front end inverter with a fixed threshold voltage irrespective of the SPAD pulse height on performance and pixel uniformity, three scenarios for the above dimensions were simulated:

1. Inverter input of 1.1V and supply of 1.1V.
2. Inverter input of 3.0V and supply of 1.1V.
3. Inverter input of 3.0V and supply of 3.0V.

A comparison between the first and second case would show if there is a variation in the low supply inverter depending on its input voltage range, and a comparison between the second and third case would show any variability in the output for a high input voltage and different power supplies. Two hundred and fifty Monte Carlo (MC) runs were simulated at typical conditions.

Figure 3.2.14 shows the simulation waveforms. Both 1.1V and 3.0V inputs have the same slew rate of 25ps/V assuming no added parasitic load to emulate a SPAD event. A small variation in the falling edge of case 1 is observed in comparison to the falling edge of case 2 due the lower overdrive voltage across the NMOS transistor for a 1.1V input. In both cases, the variability in rising edge is large but as mentioned it is of no consequence to this design. No visible difference is observed between the falling edges of case 2 and 3 with the rising edge of case 3 significantly improving due to the large overdrive voltage across the PMOS since the supply increased to 3.0V.

To add context to the comparison, Figure 3.2.15 shows the histograms of time delay between the input rising edge and the output falling edge for all cases. The negligible difference in variability between case 2 and 3 suggests that there is no serious implication on pixel variability or jitter of running a thick oxide inverter at low voltage supply and fixed threshold voltage for different input ranges. The small difference in variability between case 1 and 2 suggests that the size chosen for the NMOS device is reasonable and provides good performance even at low input or overdrive voltages.

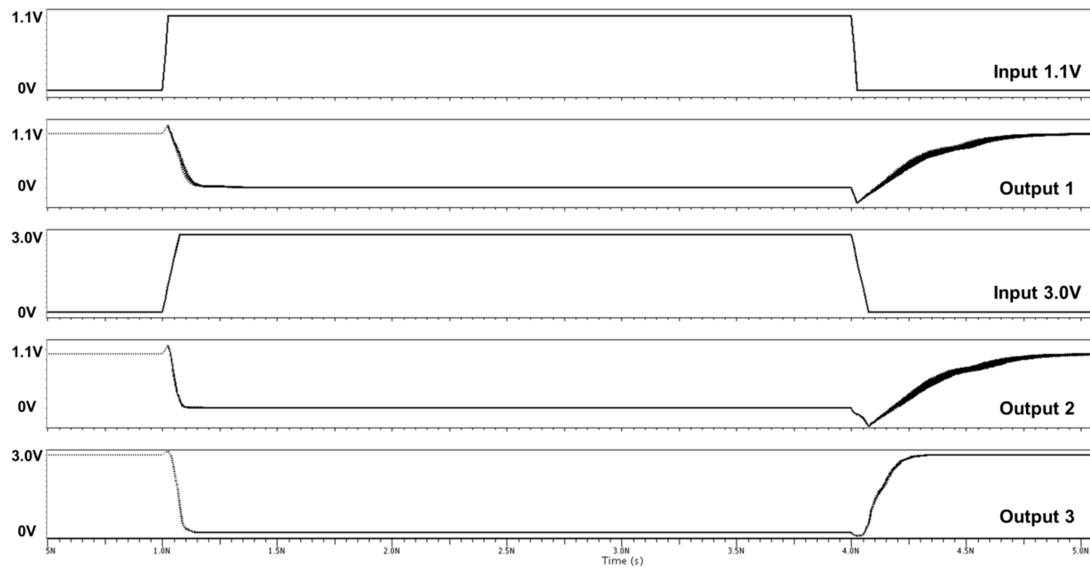


Figure 3.2.14. Simulation waveforms for 250 MC runs at typical conditions for the three cases of front end inverter operation.

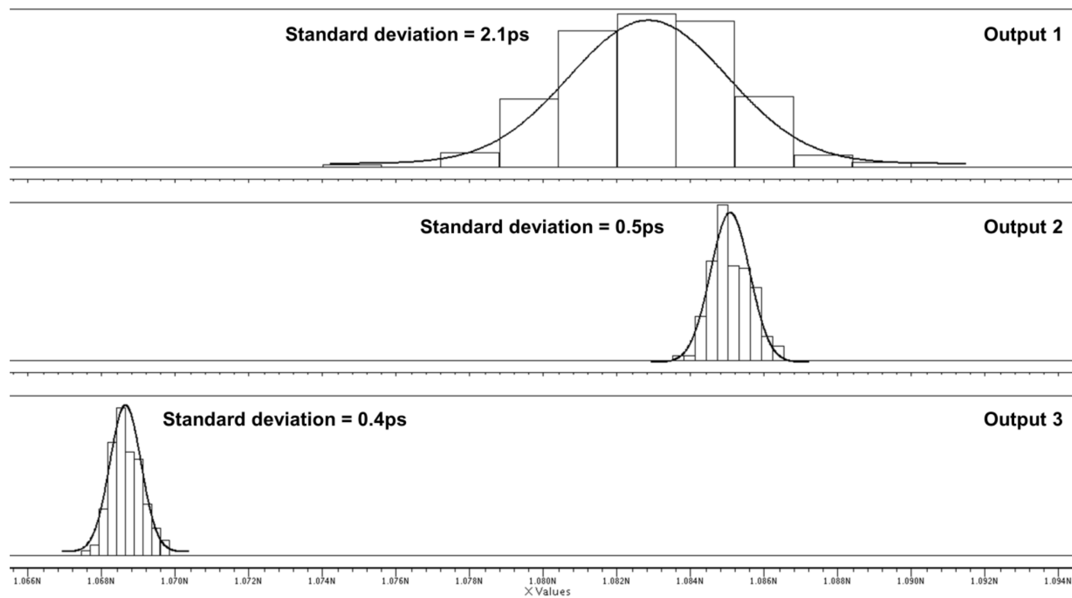


Figure 3.2.15. Histograms for time difference between input rising edge and output falling edge for 250 MC runs at typical conditions for the three cases of front end inverter operation.

3.2.6. Pixel Counter Design

The pixel counter is formed of three banks of 4-bit ripple counters that can either operate as independent memories or a single chained counter of 12-bit depth. In time gating mode the MODE control is set to logic low allowing each of the three counter banks to receive their trigger inputs (C1, C2 and C3) from the time gating logic. This allows the accumulation of counts for three independent yet simultaneous time gates albeit at low bit depth. In photon counting mode, the MODE control is set to logic high chaining the three banks together resulting in a large 12-bit counting capacity with the ENABLE control acting as global exposure signal.

The ripple counters are implemented by cascading thin oxide D-type flip-flops (DFF) but a custom DFF design was used rather than a standard cell. The custom DFF schematic is shown in Figure 3.2.16 and consists of seventeen transistors which make it much smaller than the standard cell. Figure 3.2.17 compares the layout of both DFF designs where the custom DFF layout follows the same frame structure as a standard cell to make it easily tileable, but provides 40% saving in area which translates into more bits per pixel.

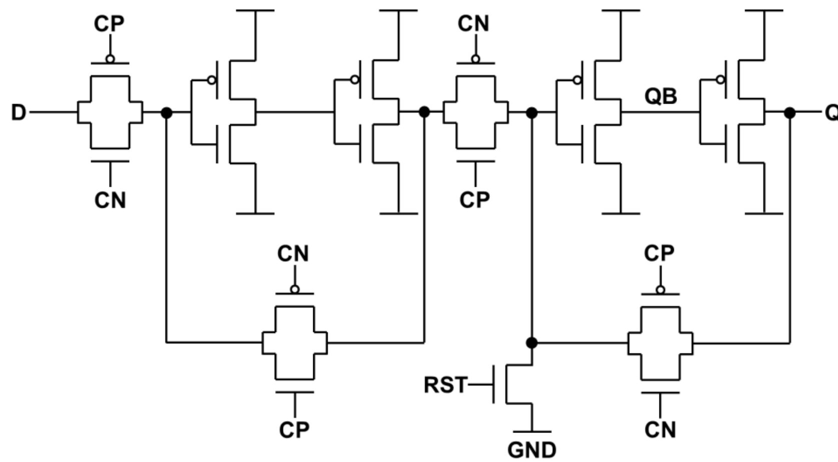


Figure 3.2.16. Schematic diagram of custom 17 transistor D-type flip-flop.

Although the custom flip-flop is advantageous in terms of area, it does suffer from few drawbacks that require careful consideration:

1. It requires differential clock phases (CP and CN) and so few more inverters in the pixel to generate them.
2. Due to its compact layout the transistor sizes are not optimised. All switches are minimum sized devices and the inverter pair driving the outputs (Q and QB) are slightly sized up but limited in drive strength.
3. Since it is stripped down to the bare minimum of devices the output nodes (Q and QB) are also internal nodes of the DFF and are not isolated from loads. This runs the danger of the internal states of the DFF being overwritten when suddenly connected to large loads such as the column parallel bus upon readout. Additional buffers are required for isolation.
4. Reset conditions. Unlike a standard cell where a reset can be applied unconditionally, the custom DFF reset needs to be applied with care to avoid state corruption.

While all of the above concerns are manageable within the pixel, the reset conditions need further thought.

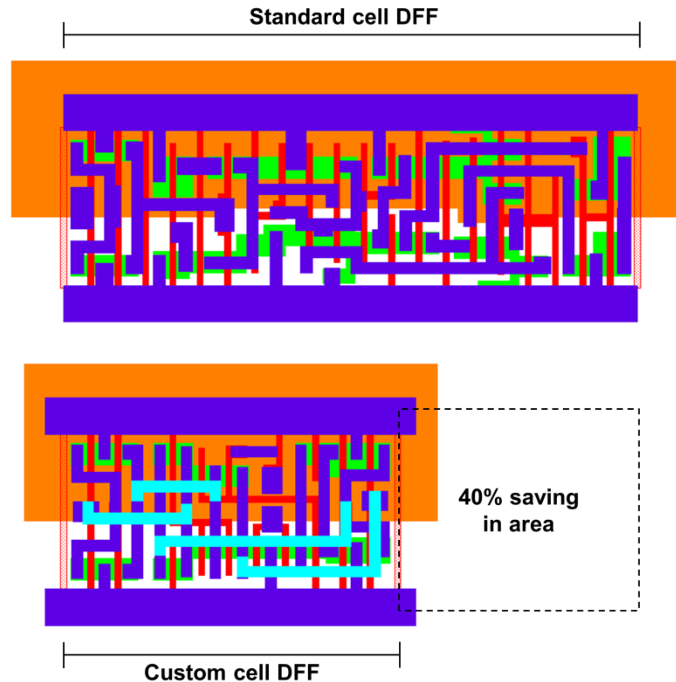
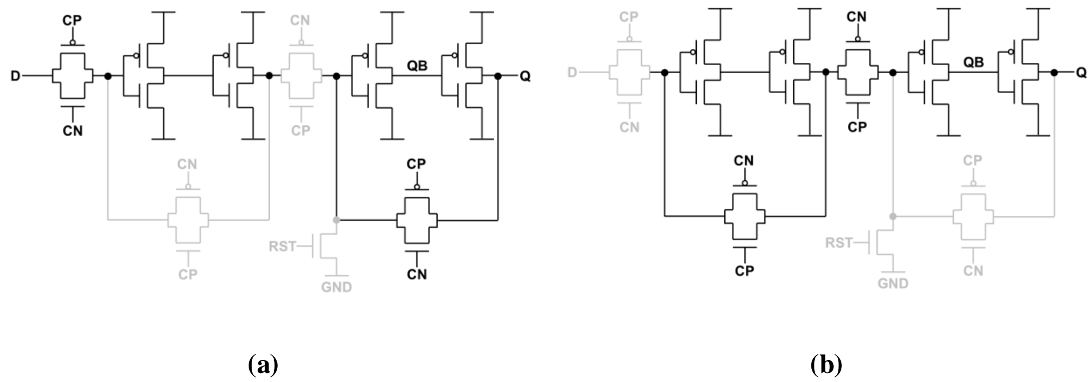


Figure 3.2.17. Layout comparison of standard cell DFF and custom cell DFF. PMOS n-well (NW) is in orange, active area (OD) is in green, poly-silicon (PO) is in red, metal 1 (MT1) is in dark blue and metal 2 (MT2) is in light blue.

The custom DFF is an edge sensitive latch which is an important property for time gating and it operates by employing an input pair and an output pair of inverters. When the clock trigger is low (CP logic low and CN logic high) the input inverter pair are continuously observing the input signal D and are isolated for the output pair which are connected in a self-enforcing loop holding onto the state of output Q. This is referred to as the hold phase (Fig. 3.2.18(a)).

Once the clock trigger goes high (CP logic high and CN logic low) the input inverter pair go into the self-enforced loop state in order to capture the value of input D and simultaneously drive this new value through the output inverter pair which are now released from the self-enforced state. This is referred to as the sample phase (Fig. 3.2.18(b)). After the clock trigger goes back low the DFF reverts to the hold phase and the output inverter loop retains the sampled D value.

Considering the custom DFF operation, a reset could occur while the DFF is in either of the two phases. The preferred phase for a reset is the hold phase as the pair of output inverters are looped and are isolated from the input pair influence whereby a reset would force the output to ground state. Yet it is not always guaranteed that the DFF is in the hold phase especially when several devices are cascaded as a ripple counter.



(a) **(b)**

Figure 3.2.18. The two phases of operation of the custom DFF. (a) Hold phase. (b) Sample phase.

If a reset occurs while the custom DFF is in the sample phase, a couple of points need to be considered. Firstly if the input inverter pair loop is enforcing a logic high value and the reset transistor is activated, there will be a contention between the loop which is trying to force a logic high and the reset which is attempting the opposite, therefore the reset phase should be allowed enough time to succeed.

Secondly in a sample phase the reset transistor is connected to the input pair loop through a transmission gate with high resistance due to the minimum device size which means that under extreme corners the input inverter loop could become sticky and the reset transistor might fail to force logic low. For these two reasons the input inverters were kept deliberately weak and the reset transistor width was increased as much as allowed by layout area.

Finally, if the DFF is allowed to toggle between the two phases while a reset is being applied, there is a chance the DFF might be writing a logic high state just before the reset state ends which results in an incomplete reset and a corrupted state. To avoid dynamic operation while a reset is applied, the ENABLE control has to be low to block SPAD events from triggering the DFF.

3.2.7. Time Gating Logic

The time gating logic consists of four sampling D-type flip-flops followed by a combinational logic decoder as in the schematic diagram of Figure 3.2.19. The principle of operation is as follows:

1. Global time gating signals G1 to G4 are broadcasted across the array.
2. A SPAD event samples the states for G1 to G4 on the DFFs.
3. The combinational logic decides whether C1, C2 or C3 is to go high.
4. The corresponding counters of signals C1 to C3 are incremented by one accordingly.
5. The gating logic is reset by the periodic RST_PIX control.
6. Steps 1 to 5 are repeated.

While basic in principle the adopted approach has two important features. First of all the sampling of the gating signals by the SPAD's rising edge makes it completely edge sensitive and independent of the SPAD pulse dead-time. This provides better timing accuracy and avoids false counts due to convoluting the SPAD pulse with a time gate pulse as in analogue approaches [73][183][184] or level sensitive digital implementations.

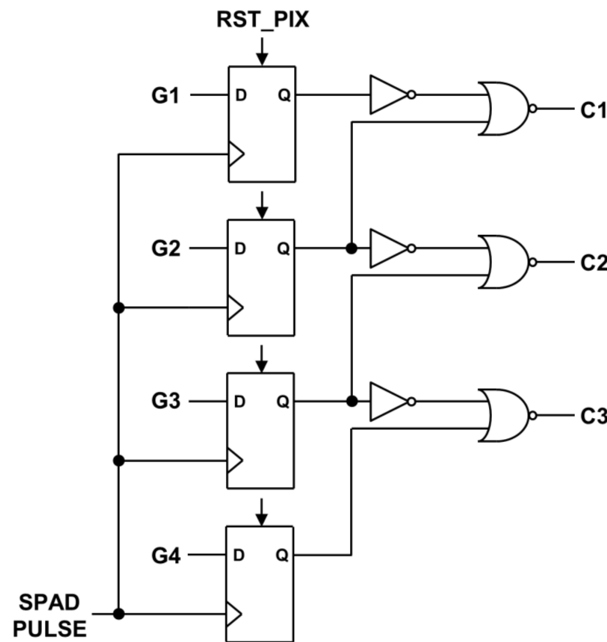


Figure 3.2.19. Time gating logic schematic diagram.

Second of all the time gate width in this case is determined by the time difference between rising edges of signals G1 to G4 rather than conventional pulses. The timing diagram in Figure 3.2.20 better explains this concept. Usually propagating time gate pulses through an IC and across an array introduces errors in the time gate width due to mismatches in rising and falling times of signal drivers and due to slewing of signals as they traverse long signal lines. Such effects can also cause contiguous time gates to overlap or even separate in time creating a dead coverage zone.

To avoid this problem each time gate is defined by two rising edges which maintain the time difference between them even as they slew. This technique not only preserves the signal integrity but also allows for propagating very small (sub-nanosecond) time gates across an array. A similar design has been published previously in linear arrays [71][162] and a modelling of the topic was presented in [234][235].

The combinational decoder decides which of the three counter banks is to be incremented. Only one of the C1, C2 and C3 signals can go high at any particular instant based on the states of G1 to G4. If a SPAD event occurs during the second time window T2 for instance (between G2 and G3), C2 goes

high while C1 and C3 remain low. If a following photon occurs during time window T3 then C2 will go low causing no additional counts in the second counter and C3 will go high. The sampling DFFs are reset at the end of every cycle.

If simultaneous SPAD events occur within the same time window, say T2, the first one will cause C2 to go high incrementing the corresponding counter by one, yet the second SPAD event will be missed since the state of the sampled DFFs does not change and so C2 does not retrigger. Therefore the gating logic can only process one event per time window per cycle. Table 3.2.4 summarises the combinational logic truth table.

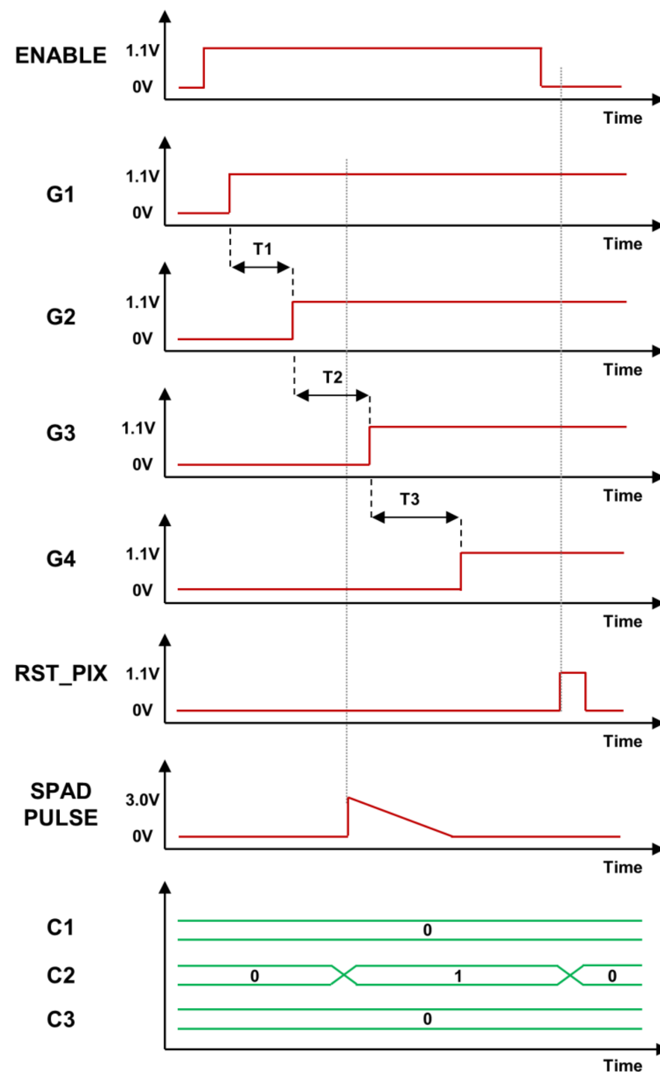


Figure 3.2.20. Gating logic timing diagram.

G1	G2	G3	G4	C1	C2	C3
0	0	0	0	0	0	0
1	0	0	0	1	0	0
1	1	0	0	0	1	0
1	1	1	0	0	0	1
1	1	1	1	0	0	0

Table 3.2.4. Gating logic truth table.

3.2.8. Electrical Crosstalk on Anode Lines

As mentioned earlier in this chapter, a danger of long anode routing in global shared well layout is the potential of electrical crosstalk between lines which result in false counts in neighbouring pixels. To investigate this issue for the implemented layout design, a test column of twenty pixels (SPAD + electronics) was simulated.

Anode<9> which is a central anode route on MT5 was fed with a pulse to mimic a SPAD firing and the other nineteen anode lines were monitored. The typical Cc extracted simulation was repeated for three different SPAD pulse heights of 1V, 2V and 3V emulating different excess bias conditions. Figure 3.2.21 gives a top view of the simulated layout and Figure 3.2.22 shows the simulation waveforms.

Crosstalk between the lines is clearly visible with four immediate neighbours showing strong coupling. These are the two horizontal MT5 neighbours Anode<7> and Anode<11> and the diagonal neighbours Anode<8> and Anode<10> on the lower MT4 lines. The dependence of coupling on excess bias is evident with the coupled signals on the horizontal MT5 neighbours crossing the 0.55V front end inverter threshold at 3V excess bias but for a very short duration insufficient to trigger the counter.

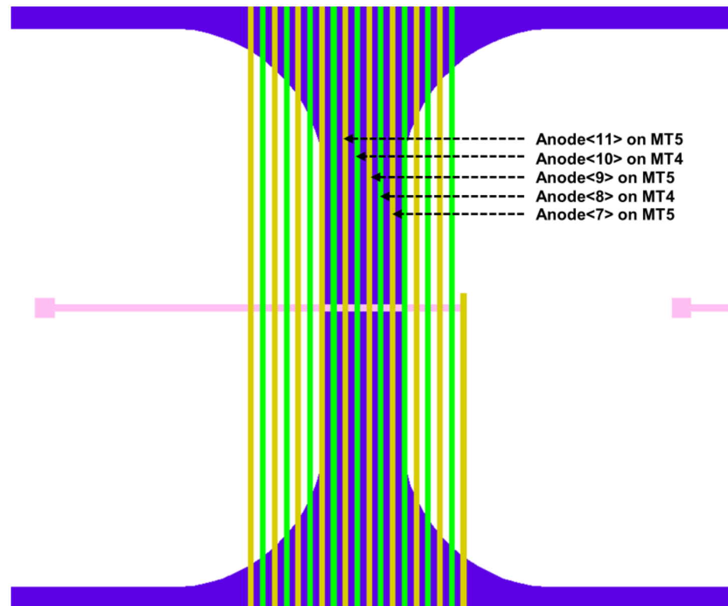


Figure 3.2.21. Top view of extracted column crosstalk simulation.

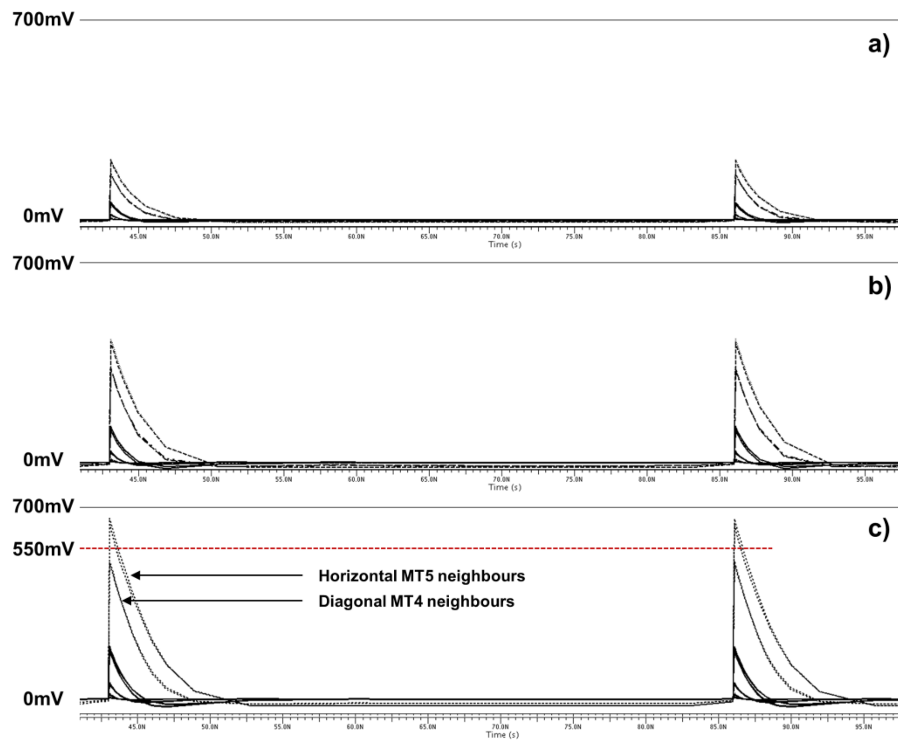


Figure 3.2.22. Extracted column crosstalk simulation waveforms. a) For 1V excess bias case. b) For 2V excess bias case. c) For 3V excess bias case where coupling on horizontal MT5 neighbour lines surpasses the front end inverter 0.55V threshold.

3.3. SPAD and Sensor Characterisation

This section presents optical and electrical characterisation results of the MINIC40 array. Full characterisation results of the SPAD device are not presented as apart from the device size and minor structural differences, the layers and junction of the implemented device are identical to STMicroelectronics' industrialised device where a full characterisation account has been presented at IEDM 2017 [114]. Thus the focus of the results herein is the specifics of the MINC40 global shared well array.

In reference to [114], the 40nm SPADs were found to have an optical crosstalk of 2% to direct neighbours in a shared well layout, afterpulsing probability of 0.1% and a jitter of 140ps which is similar to other reported CMOS devices. All these figures are claimed for 1V excess bias setting.

3.3.1. Photon Detection Probability

Figure 3.3.1 shows the PDP versus wavelength plot of the 40nm SPAD device replicated from [114] in comparison to the PDP curve of the 8μm pixel presented in [65][236] and implemented in an imaging 130nm process.

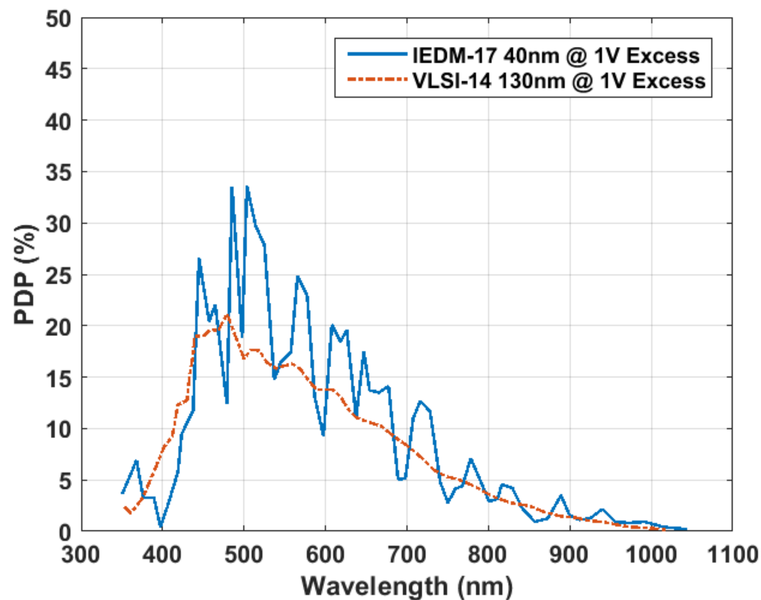


Figure 3.3.1. PDP versus wavelength at 1V excess bias for the industrialised 40nm SPAD device replicated from [114] and the imaging 130nm SPAD reported in [65][236].

While both front side illuminated devices are measured at 1V excess bias, few observations can be made. The PDP profile of both SPADs is similar in trend with a peak around the 500nm region which is expected as both have the same junction structure (PW to DNW). Albeit sharp oscillations in the

response of the 40nm device is observed due to the Fabry Perot interference patterns of the much higher metal stack (7-metal layers in 40nm), while the 130nm imaging process has a much limited and optimised stack. These oscillations can be problematic as small deviations in wavelength result in big changes in detection probability.

Moreover, the 40nm device reaches a peak PDP value of 34% compared to the 21% of the 130nm device which is curious given the expected lower transmission efficiency of the higher metal stack. This discrepancy is due to the fact that the reported 1V excess bias figure in [114] refers to 1V beyond VHV0 and not the SPAD's breakdown voltage, where VHV0 is the voltage at which SPAD pulses start to appear and is dependent on the front end circuit supply. Therefore the 1V excess bias relative to VHV0 in practice corresponds to 1.9V excess bias relative to V_{BD} given the mentioned conditions in [114].

For the reported results in this Chapter and the remainder of this work, the stated excess bias setting is always relative to V_{BD} unless specified otherwise.

3.3.2. Dark Count Rate

The median DCR of the four implemented SPAD trials at room temperature and different excess bias voltages is shown in Figure 3.3.2. At 1V excess bias, all trials exhibit a DCR below 50cps with it increasing exponentially with excess bias indicating tunnelling is the dominant source of DCR. As the guard ring design is pushed further towards aggressive dimensions, the active area of the SPAD increases and so does the measured DCR albeit with a non-linear dependency on area as previously reported by [237]. A much larger increase in DCR is seen for the step from 51% to 66% fill factor when compared to the step from 39% to 51% fill factor suggesting the onset of edge breakdown for the most aggressive design.

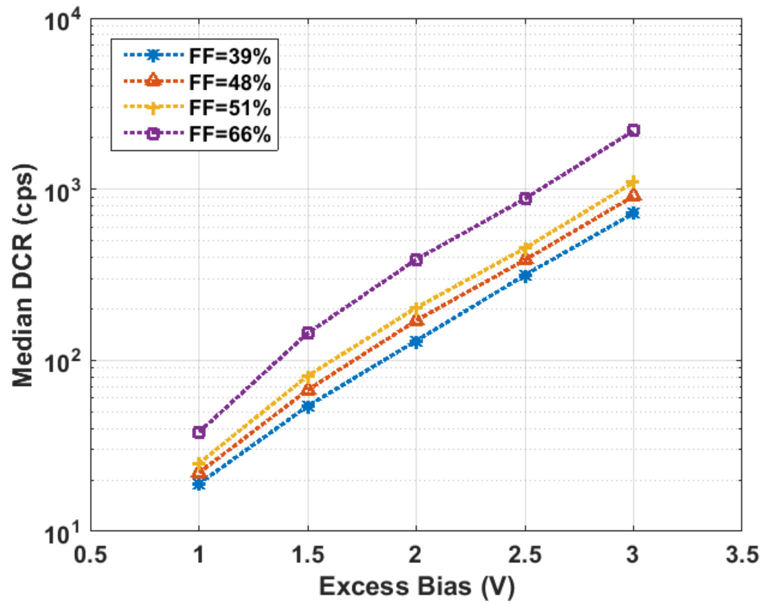


Figure 3.3.2. Median DCR at room temperature for the four fill-factor SPAD trials implemented in MINIC40.

Figure 3.3.3 shows the cumulative distribution of DCR amongst the 3840 SPADs in the array at room temperature and for 2V excess bias for the standard (39% fill factor) and the most aggressive (66% fill factor) trials. Both trials show the same upward sloping trend suggesting a wide distribution of DCR around the median value with a visible tail of high DCR or screamer SPADs which exhibit a different DCR defect. These curves are indicative of the cleanliness of the fabrication process and if a 1kcps benchmark is to be used, 87% and 74% of the standard and aggressive trails respectively meet the specification. From an image sensor perspective where one SPAD is used per pixel this defect rate needs improving and does not compare well to mainstream image sensor standards.

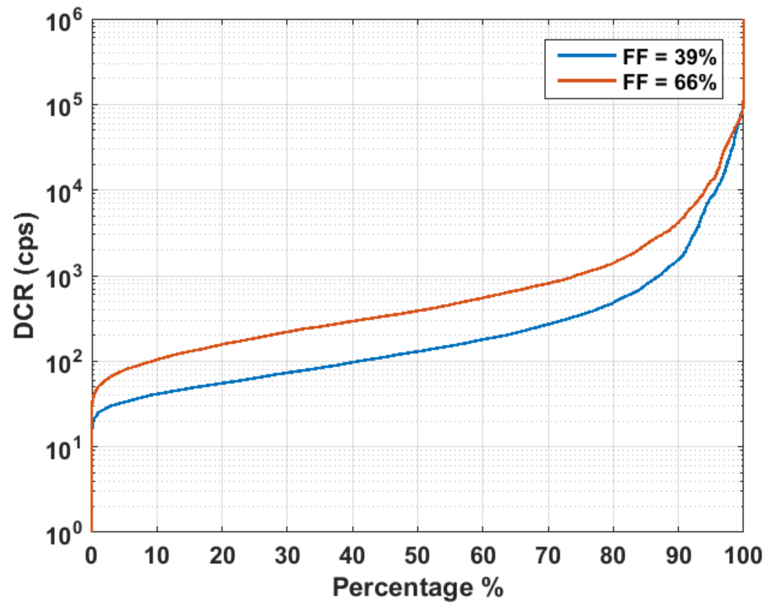


Figure 3.3.3. Cumulative DCR distribution at room temperature and 2V excess bias for the standard 39% and aggressive 66% fill-factor SPAD trials implemented in MINIC40.

3.3.3. Photo Response Non-Uniformity

Photo response non-uniformity (PRNU) is one of the most important figures for image sensors as it reflects the pixel to pixel output variation with respect to constant input photon flux. To evaluate PRNU of the MINIC40 array, the standard 39% fill factor trial at room temperature and 2V excess bias was used and a thousand frames of 100µs exposures under fixed illumination where averaged to cancel out any temporal variation. Figure 3.3.4 shows the averaged frame with a mean photon count of 2.3k equating to a mean SPAD count rate of 23Mcps.

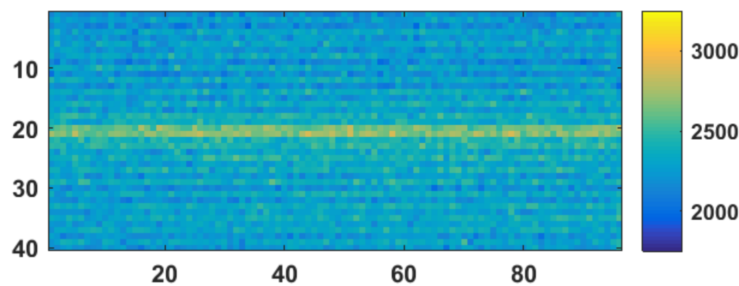


Figure 3.3.4. Average frame of standard MNIC40 array under fixed high illumination level. Colour scale represents photon counts.

Two prominent features jump out to the reader: horizontal fixed patter noise (HFPN) represented in alternating row count levels and a centre strap of higher counts represented in rows 20 and 21. These

two distinct features can be linked directly to the global shared well layout strategy. The alternating row pattern relates to the vertical anode routing from the SPAD array to the circuit array as half the SPADs are routed in M4 while the other is routed in M5 in an alternating fashion as seen in Figure 3.2.21 and thus experience different parasitics contributing towards variations in SPAD dead-time.

Rows 20 and 21 on the other hand are a special case as they are the top most SPADs of each half array which means that their anode route is an edge track running almost entirely over the guard ring region and has one neighbouring anode track as opposed to two and so lower parasitic capacitance (i.e. shorter SPAD dead-time). The same is not true for rows 1 and 40 despite their anode tracks also having one neighbouring track as they run over the SPAD array, they are routed almost entirely over the circuit array and so have other control or supply tracks on the other side.

While variations in dead-time can be small, they become more apparent under high illumination levels such as this case where the SPADs are in the nonlinear region and tending towards the saturation limit as depicted in Figure 3.3.5.

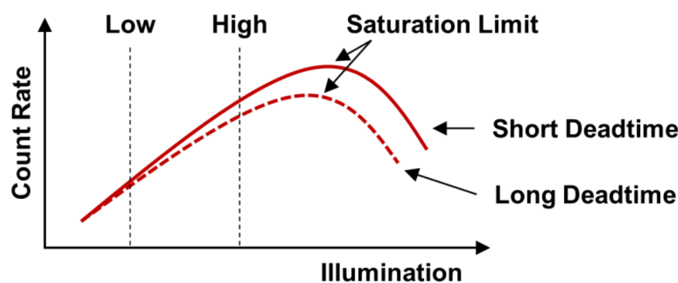


Figure 3.3.5. Illustration of typical passive quenched SPAD response showing the variation in count rate for different dead-times as illumination approaches saturation level.

The same experiment was repeated under a lower illumination level with the average frame of one thousand captures shown in Figure 3.3.6. The exposure was extended to 750µs such that the average photon count is 2.16k which is similar to the photon count in the previous experiment so the same colour scale could be used for comparison. The average SPAD count rate is 2.9Mcps which is almost a decade lower than the previous case. Comparing Figures 3.3.4 and 3.3.6, the HFPN is almost not visible with the exception of the two centre rows.

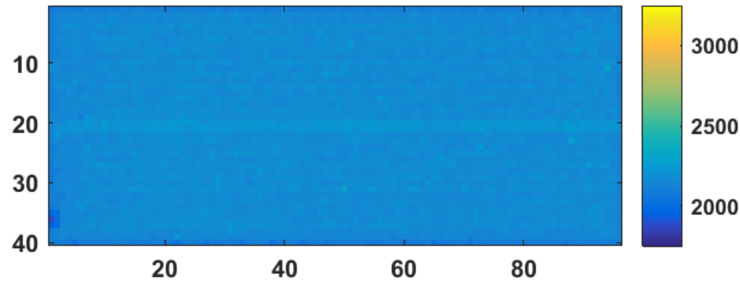


Figure 3.3.6. Average frame of standard MNIC40 array under fixed low illumination level. Colour scale represents photon counts.

Figure 3.3.7 shows the histogram of the recorded photon counts of the average frames for both low and high illumination levels where the distribution at lower illumination is clearly tighter. PRNU was calculated as the standard deviation divided by the mean and equated to 1.3% and 5.8% under low and high illuminations respectively. This measurement shows that while the trialled global well sharing technique allows for miniature high fill factor image sensor arrays, careful consideration of the effect of PRNU given the expected illumination levels per application should be taken into account.

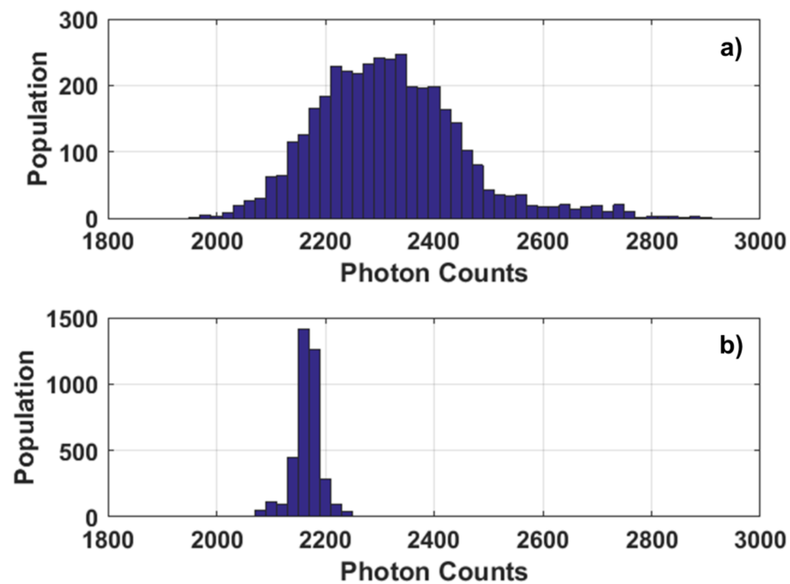


Figure 3.3.7. Histogram of pixel photon counts of average frame of MINIC40 standard array showing pixel to pixel variation under different conditions. a) High illumination level. b) Low illumination level.

Upon closer examination of the low illumination level average frame by adjusting the colour scale other image non-uniformities become visible. Figure 3.3.8 shows the higher contrast frame. It can be seen that apart from the aforementioned artefacts, the peripheral pixels around the edges of the array seem to exhibit lower photon counts while a handful of pixels in the bottom left corner are even darker.

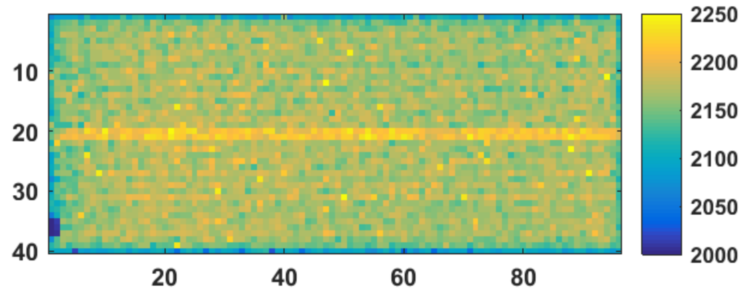


Figure 3.3.8. Higher contrast average frame of standard MNIC40 array under fixed low illumination level. Colour scale represents photon counts.

Likewise, these effects are linked to the array layout. The fainter edge rows and columns are attributed to shading effects due to the metal routing and density dummy patterns surrounding the SPAD array which sits in the centre of the IC. These metallisation patterns up to MT7 and including the top most aluminium dummies form a several micrometres high box-like structure encapsulating the SPAD array which would cast a shadow over its edges.

As for the handful of darker pixels, this is due to a fault in the dummy generation procedure which inserted few dummy structures around these pixels causing partial light blockage. The micrograph image in Figure 3.3.9 better conveys the message.

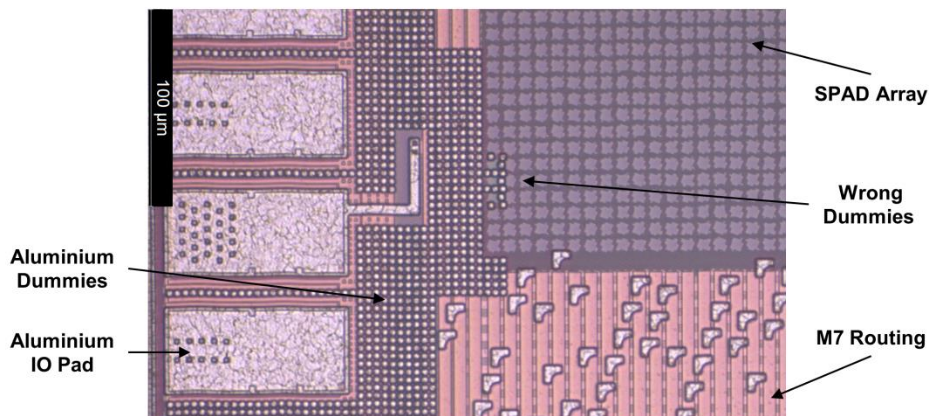


Figure 3.3.9. Zoomed micrograph of MINIC40 showing bottom left corner of SPAD array.

3.3.4. SPAD Array Current Consumption

A crude experiment to get a feel for the SPAD array power consumption was carried out for the different trials. The high voltage (VHV) was supplied by an external Keysight E3647A power supply such that the current drawn can be recorded and the SPADs were allowed to free run. An LED was used to illuminate the array at discrete illumination levels. Figure 3.3.10 shows the plot for array current versus illumination.

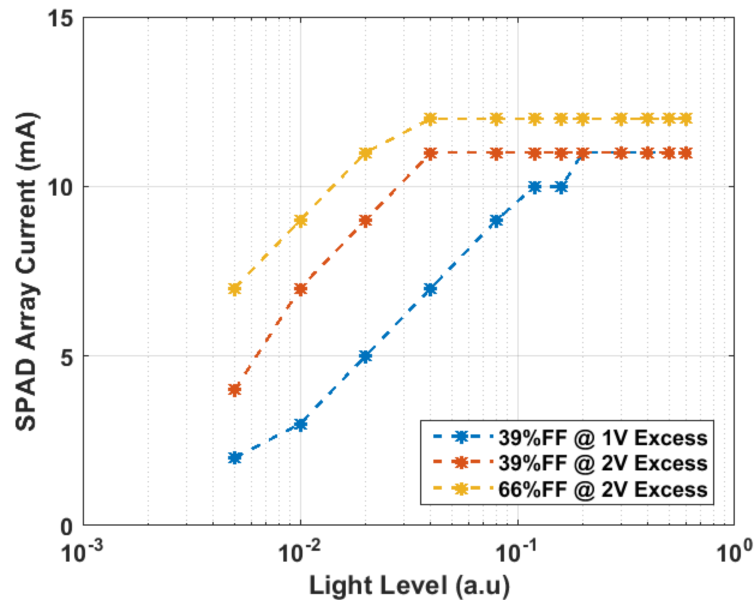


Figure 3.3.10. MINIC40 SPAD array current for different trials at different excess bias voltages.

The quantisation in the recorded array current is due to the 1mA resolution of the used power supply. As expected, and for the standard 39% fill factor trial, increasing the excess bias increases the array power consumption as the junction field intensifies and the avalanching consumes more energy. Also, for the same 2V excess bias, switching from the standard trial to the aggressive 66% fill factor trial increases the power consumption further as the active area of the SPAD increases and so does the junction capacitance. For both trials the current drawn by the 3840 SPADs easily exceeds 5mA at moderate to high illumination levels.

While this graph does not draw any surprising conclusions it does bring up two points concerning the scalability of SPAD imaging arrays:

1. There is a power consumption cost to increasing the fill factor by design which can become a problem for high resolution SPAD arrays. This begs the question of whether it is more viable to restrict the device drawn fill factor if it is high enough to start with (i.e. 39% in this case) and gain more detection efficiency by means of micro-lensing [238] for the purpose of saving power as SPAD arrays scale. In other words, there might be a break point between costs of fabrication versus cost of operation.
2. The measured current figures reflect the case of free running SPAD pixels which makes sense when the pixel has a large counting capacity (12-bits for MINC40), but when the pixel is limited to a single bit memory such as [80], it makes more sense to synchronously arm and disarm the SPAD as oversampled frames are captured since only one event can be registered in any given exposure and so allowing the SPADs to free run is wasteful in terms of power.

While this might seem the case for small arrays, the balance becomes more subtle for large resolution sensors as the power needed to propagate the pixel controls at very high frame rates needs to be weighed against the power of free running SPADs given an expected illumination level.

3.3.5. Global Enable (Shutter) Signal Propagation

As in the case of optical characterisation, the standard 39% fill factor array was used at 2V excess bias for all electrical characterisation measurements.

The first signal to be characterised was the global exposure control referred to as Enable in Figure 3.2.10. The Enable signal allows the SPAD pulses to reach the time gating logic or the 12-bit counter based on the selected mode of operation. To capture the signal behaviour, a 443nm Hamamatsu PLP10 pulsed laser with quoted electrical jitter of 53ps was delayed in time with respect to the Enable signal using a Stanford DG645 delay box. At each time point, several frames were captured and summed. A sequence of such frames is shown in Figure 3.3.11 showing the propagation of the Enable signal falling edge (hence the drop in counts) as the laser pulse drops outside the exposure window.

The sequence clearly reveals an overall propagation delay across the array of at least 1ns and in blocks of 12 columns. This column pattern is due to the re-buffering of the Enable signal every 12 columns as it traverses horizontally before being driven up the columns. This delay is acceptable as the time gating function does not depend on the Enable signal but rather uses a dedicated set of controls which are distributed via a balanced clock tree to ensure timing uniformity across the array.

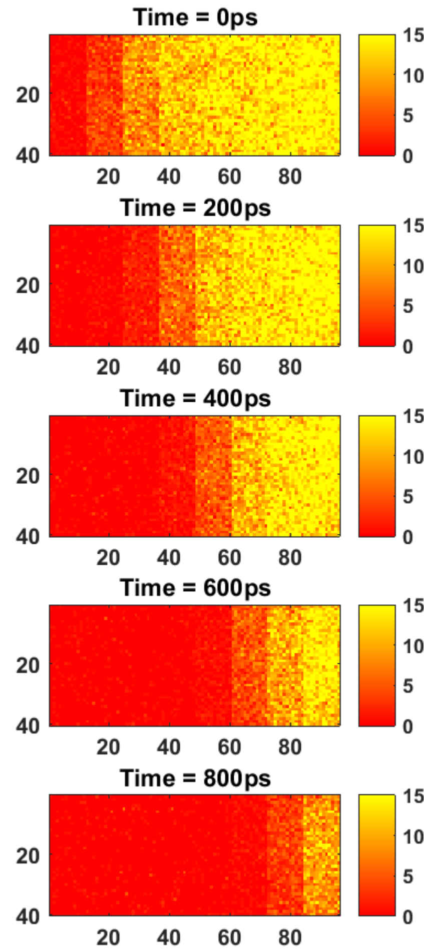


Figure 3.3.11. MINIC40 Enable signal propagation delay sequence across the array. Colour scale represents photon counts.

3.3.6. Time Gate Profile Characterisation

To characterise the time gating performance the Stanford delay box was used to generate and feed rising edges to the chip as illustrated in the timing diagram of Figure 3.2.20. The Hamamatsu laser was then swept in steps of 25ps across the gate and counts were recorded over many frames in order to reconstruct the resulting gate profile.

Figure 3.3.12 shows the profile of the minimum measured time gate using the edge-to-edge technique of a randomly selected pixel. The shape of the time gate is a convolution of the laser jitter, SPAD jitter and the gate generation signals. A record average time gate FWHM of 360ps for a SPAD image sensor 2× improvement over reported state of the art [182] was obtained with 31ps standard deviation across the array.

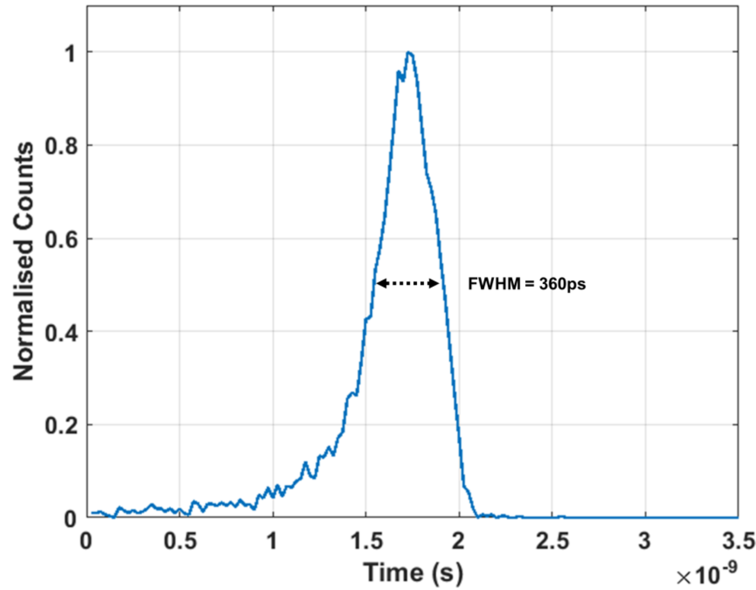


Figure 3.3.12. MINIC40 time gate profile of a randomly selected pixel showing a record FWHM of 360ps using the edge-to-edge technique.

3.3.7. Time Gate Array Uniformity

To better assess the quality of the time gating signals across the array a map of measured FWHM values is shown in Figure 3.3.13. A clear split can be seen at the centre of the map which is not surprising given the layout of the IC as half the pixel circuitry sits on either sides of the global shared well SPAD array. This necessitates duplicating the gating controls and routing them to separately to each half.

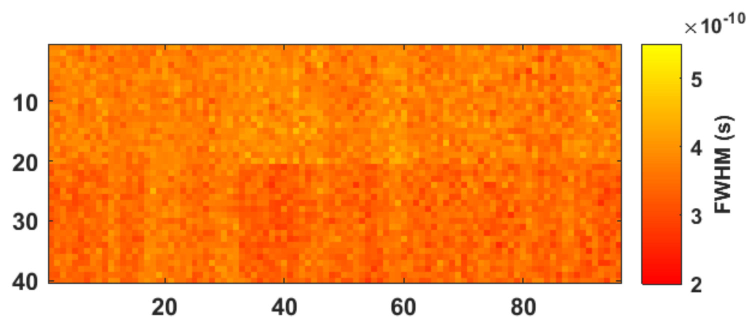


Figure 3.3.13. MINIC40 time gate FWHM map across the array. A clear mismatch between the top and bottom halves of the array is visible.

Figure 3.3.14 shows separate histograms of the time gate FWHM for the top and bottom halves of the array which clearly highlights a difference of approximately 20ps due to the propagation delay mismatch of the edges forming the gates from the source to each of the two halves. The variation

within each half is attributed to the mismatch in the binary tree drivers distributing the signals across the columns.

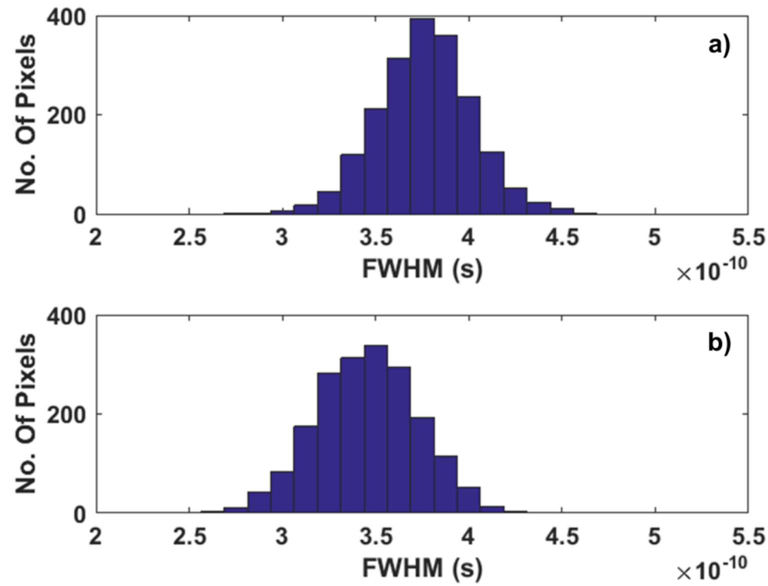


Figure 3.3.14. MINIC40 time gate FWHM histogram. a) Top half of the array. b) Bottom half of the array. A mismatch of approximately 20ps is seen between the two halves.

3.3.8. Time Gate Window Handover

Finally, the effectiveness of the rising edge to rising edge gating technique in terms of photon efficiency is characterised through two sets of measurements. In the first set two contiguous 20ns wide time gates where generated for bins 1 and 2 and the laser was swept across in steps of 100ps while the counts have been recorded. Figure 3.3.15 shows the obtained bin profiles over a region of 9 × 9 pixels at the centre of the array. Spatial oversampling was necessary to improve signal level.

Although the two time gates are contiguous to start with, the mismatch in propagation delay and rise and fall edge times causes them to separate which is evident in the observed dip in counts. This shows the deficiency in the handover region between the gates.

In the second experiment, the same contiguous time gates are generated using the edge-to-edge technique and the same laser sweep was repeated. Figure 3.3.16 shows the resulting handover response overcoming the photon count dip due to improved matching of the gate profiles for the same 9 × 9 pixels.

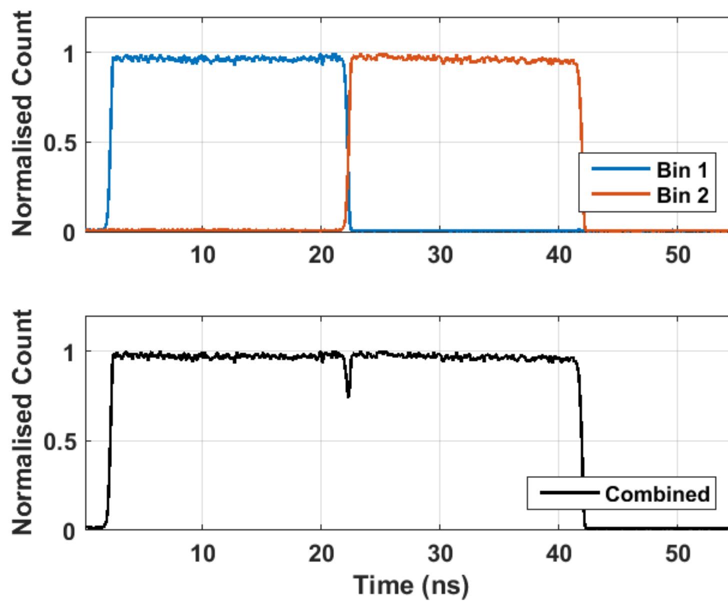


Figure 3.3.15. MINIC40 handover between two contiguous 20ns time gates generated by the conventional rising-to-falling edge technique. A drop in photon counts is observed at the border between the gates due to mismatch in gate profiles.

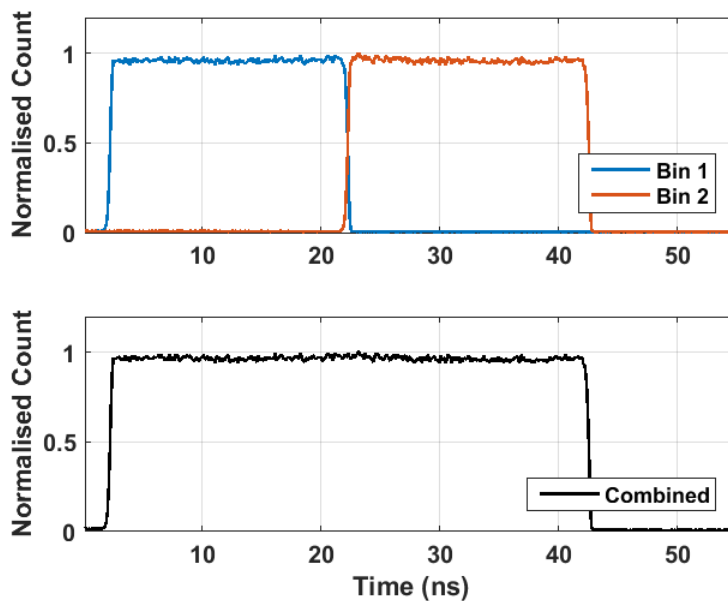


Figure 3.3.16. MINIC40 handover between two contiguous 20ns time gates generated by the rising-to-rising edge technique. No drop in photon counts is observed at the border between the gates due to improved matching in gate profiles.

3.4. Comparison to State of the Art Sensors

To evaluate the validity of the global shared well approach from miniaturisation perspective and the time-resolved capability of the designed pixel, the MINIC40 sensor is compared to the state of the art CMOS SPAD image sensors.

3.4.1. Pixel Pitch, Fill Factor and Dynamic Range Trade-offs

Table 3.4.1 lists the relevant published front side illuminated imaging arrays implemented in various technology nodes since 2006 and classifies them by circuit type and architecture. Sensors that are specifically designed for photon counting and time-gated operation are marked in red to distinguish them from pixels aimed for TCSPC. For pixels with unequal pitch in x and y directions, a normalised pitch is calculated as the square root of the pixel area.

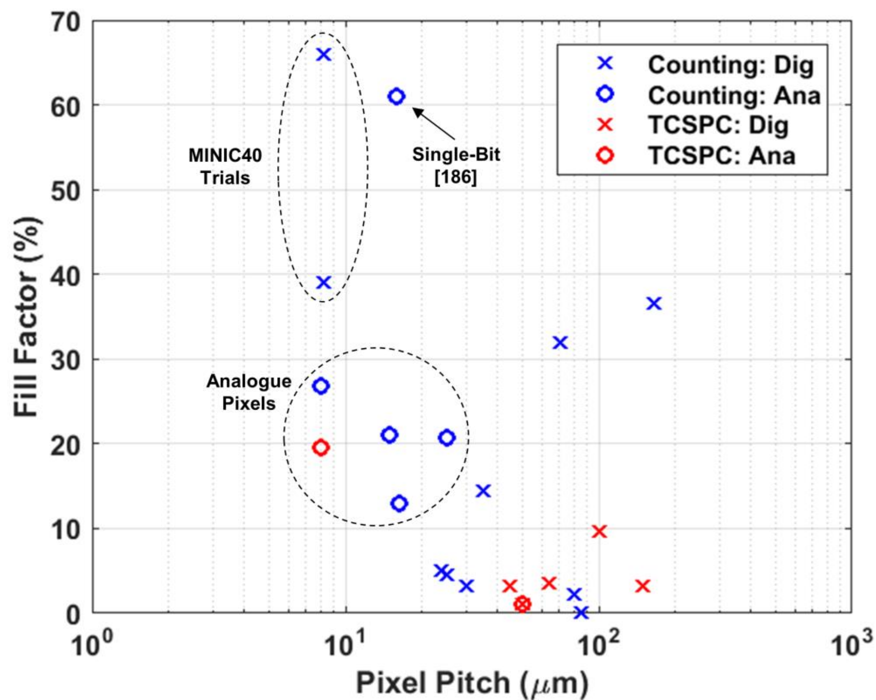


Figure 3.4.1. Comparison of fill factor versus pixel pitch for state of the art FSI CMOS SPAD image sensors.

Figure 3.4.1 highlights the fill factor and pixel pitch trade-off for the listed sensors. It is clear that analogue designs give the best compromise with the two 8μm pixels in [65] and [72] leading with 26.8% and 19.6% fill factors respectively. The MINC40 trials maintain a similar pitch while boosting the fill factor beyond other FSI implementations up to 66%, a virtue of decoupling the SPAD array from associated circuitry.

One other pixel stands out which is the 61% fill factor at 16μm pitch reported at IEDM 2016 [186]. This has been achieved through an optimised SPAD and circuit design albeit at the cost of single-bit counter depth which leads to the next discussion.

Figure 3.4.2. Comparison of fill factor versus single frame photon counting capacity of the art FSI CMOS SPAD image sensor pixels.

Figure 3.4.2 highlights the trade-off between pixel sensitivity and integrated functionality, manifested here as the pixel photon counting capacity within a single frame capture. One interesting class of pixels emerges represented by the single-bit memory designs which minimises the counter bit-depth to the minimum in an attempt to boost fill factor. Such pixels require oversampled operation in order to recover dynamic range. One particular pixel stands out in terms of high photon counting capacity and fill factor by Mosconi et al. [226] but with prohibitively large dimensions unsuited for miniaturisation.

Once again, it is clear that MINIC40 can achieve both sensitivity and dynamic range beyond other FSI sensors due to the decoupling of SPAD and circuitry, optimised circuit design and more specifically the aggressive high density 40nm node.

Ref.	First Author	Institute	Year First Published	CMOS Node (µm)	Techno.	PitchX (µm)	PitchY (µm)	Pitch (µm) [Sqrt of Area]	FillFactor (%)	Resolution	Architecture	Pixel Counter Depth per Single Frame (Photons)	Circuit Type
MINIC40 [2]	Al Abbas et al.	Uni. Of Edinburgh	2017	0.04	FSI	8.25	8.25	8.25	39 66†	96 × 40	Counting	4096	Dig
[226]	Mosconi et al.	FBK	2006	0.35	FSI	180	150	164.32	36.6	7 × 2	Counting	131072	Dig
[63]	Niclass et al.	EPFL	2008	0.35	FSI	85	85	85	0.05	60 × 48	Counting	256	Dig
[167]	Richardson et al.	Uni. Of Edinburgh	2009	0.13	FSI	50	50	50	1	32 × 32	TDC	128	Dig
[178]	Stoppa et al.	FBK	2009	0.13	FSI	50	50	50	1	32 × 32	TAC	64	Ana
[171]	Gersbach et al.	EPFL	2009	0.13	FSI	50	50	50	1	32 × 32	TDC	64	Dig
[239]	Carrara et al.	EPFL	2009	0.35	FSI	30	30	30	3.14	32 × 32	Counting	1	Dig
[181][183]	Pancheri et al.	FBK	2011	0.35	FSI	25	25	25	20.8	32 × 32	Counting	150	Ana
[240]	Walker et al.	Uni. Of Edinburgh	2011	0.13	FSI	44.65	44.65	44.65	3.17	128 × 96	TDC	64	Dig
[168]	Veerappan et al.	TUDelft	2011	0.13	FSI	50	50	50	1	160 × 128	TDC	128	Dig
[185]	Maruyama et al.	TUDelft	2011	0.35	FSI	25	25	25	4.5	128 × 128	Counting	1	Dig
[180]	Bronzi et al.	Poli. di Milano	2014	0.35	FSI	150	150	150	3.14	64 × 32	Counting	512	Dig
[65][73]	Dutton et al.	Uni. Of Edinburgh	2014	0.13	FSI	8	8	8	26.8	320 × 240	Counting	1* 400	Ana
[169]	Vornicu et al.	Uni. Of Seville	2014	0.18	FSI	64	64	64	3.5	64 × 64	TDC	256	Dig
[66]	Lee et al.	Cornell Uni.	2014	0.18	FSI	35	35	35	14.4	72 × 60	Counting	1024	Dig
[80]	Burri et al.	EPFL	2014	0.35	FSI	24	24	24	5	512 × 128	Counting	1	Dig
[172]	Villa et al.	Poli. di Milano	2014	0.35	FSI	150	150	150	3.14	32 × 32	TDC	64	Dig
[182][184]	Perenzoni et al.	FBK	2015	0.35	FSI	15	15	15	21	160 × 120	Counting	70 256*	Ana Ana
[72]	Parmesan et al.	Uni. Of Edinburgh	2015	0.13	FSI	8	8	8	19.6	256 × 256	TAC	1	Ana
[186]	Gyongy et al.	Uni. Of Edinburgh	2016	0.13	FSI	16	16	16	61	256 × 256	Counting	1	Ana
[241]	Portaluppi et al.	Poli. di Milano	2017	0.18	FSI	100	100	100	9.6	32 × 32	TDC	32	Dig
[187]	Ulku et al.	EPFL	2017	0.18	FSI	16.38	16.38	16.38	13	512 × 512	Counting	1	Ana
[242]	Ruokamo et al.	Uni. Of Oulu	2017	0.35	FSI	50	100	70.71	32	80 × 25	Counting	1	Dig
[243]	Ouh et al.	Oregon State Uni.	2018	0.18	FSI	80	80	80	2.2	8 × 8	Counting	16384	Dig

† Aggressive layout trial. * Digital readout mode.

Table 3.4.1. List of FSI CMOS SPAD image sensor arrays since 2006. Architectures marked in red are intended for photon counting and time-gated operation.

3.4.2. Time Gate Performance

Since time gating is core to time-resolved imaging operation of the proposed sensor, the time gate quality is assessed in comparison to other works in terms of uniformity across the array and minimum time-gate width achievable. The time gate width is important as it defines the sensor's ability to observe fast temporal phenomena such as nanosecond fluorescence lifetimes of biomedical samples.

Table 3.4.2 lists relevant time-gated sensors with different architectures including line sensors and test pixels. Standard gating refers to the conventional rising-to-falling gate profile while TDC-based refers to designs relying on multiple clock phases locked to a delay locked loop (DLL) or utilising an external TDC card. Edge-to-edge refers to the time gating technique described in this work.

Generally line sensors are capable of fine gate width due to the small count of vertical pixels which for an image sensor require optimised layout paths and drivers. Sub-nanosecond gates are possible for image sensors enabled by the edge-to-edge technique which minimises the driver rising to falling time mismatch due to single edge polarity and reduces gate profile distortion due to propagation delay skews across column lines. Recent implementations of such technique were reported for SPAD [182][242] and pinned photodiode (PPD) [56] imaging arrays with MINIC40 achieving a minimum time gate down to 360ps FWHM and 31ps standard deviation comparable to the reported sensors.

Ref.	First Author	Year First Published	Architecture	Sensor Type	Resolution	Gating Technique	No. of Bins	Minimum Time Gate Width (ns)	Array Uniformity σ (ps)
[2]	Al Abbas et al.	2017	Image Sensor	SPAD	96 × 40	Edge-to-Edge	3	0.36	31
[149]	Pancheri et al.	2009	Line Sensor	SPAD	64	Standard	4	0.8	20
[244]	Nissinen et al.	2011	Test Pixel	SPAD	1	TDC-Based	1	0.3	na
[181][183]	Pancheri et al.	2011/2013	Image Sensor	SPAD	32 × 32	Standard	1	1.1	na
[245]	Pancheri et al.	2013	Test Pixel	SPAD	1	Standard	1	0.53	na
[163][246]	Nissinen et al.	2013/2014	Line Sensor	SPAD	128	TDC-Based	8	0.08	10
[162][247]	Maruyama et al.	2013/2014	Line Sensor	SPAD	1024	Edge-to-Edge	1	0.7	120
[80]	Burri et al.	2014	Image Sensor	SPAD	512 × 128	Standard	1	4	134*
[71]	Nissinen	2015	Line Sensor	SPAD	128	Edge-to-Edge	4	0.08	17.5
[182][184]	Perenzoni et al.	2015/2016	Image Sensor	SPAD	160 × 120	Edge-to-Edge	1	0.75	80.2†
[186]	Gyongy et al.	2016	Image Sensor	SPAD	256 × 256	Standard	1	4	na
[242]	Ruokamo et al.	2017	Image Sensor	SPAD	80 × 25	Edge-to-Edge	1	0.5	na
[56][248]	kawahito	2017/2018	Image Sensor	PPD	128 × 128	Edge-to-Edge	4	0.8	50⁺
* Calculated from reported FWHM									
† Reported for FWHM 1.4ns gate									
⁺ Reporetd RMS variation across bins									

Table 3.4.2. Comparison of time-gating techniques and minimum reported time-gate FWHM for different SPAD sensors and a recent pinned photodiode FLIM image sensor. Image sensors are marked in red.

3.5. Summary and Conclusion

A miniature 96 × 40 FSI SPAD image sensor with high fill factor by expanding the global well sharing layout technique previously restricted to small SiPM arrays [188] or line sensor pixels [164] is presented. This is enabled by the fine design rules and the 7-metal stack of an advanced 40nm process.

Aside from improved sensitivity when compared to other works, the sensor's 8.25μm pitch pixel allows for multi-bin time-gated operation with time gate FWHM as short as 360ps and a high dynamic

range counter bit depth of 12-bits. This is a result of decoupling the circuits from the SPAD array and the high logic integration density.

Nevertheless, such sensor architecture suffers from two drawbacks:

1. Scalability. Due to process and geometry limitations the array size can only be scaled horizontally while the number of column pixels is constrained. Even if applicable to miniature sensor formats, the resulting low aspect ratio field of view is inconvenient.
2. Photo-response non uniformity. The extension of the shared well layout relies heavily on long SPAD anode connections alternating between available metal layers. Due to line parasitics and unbalanced routing, that results in variations in SPAD dead-time contributing to a non-uniform response across the array.

An alternative to maintaining high sensitivity and uniformity in FSI implementations is to opt for a minimalistic design employing a single-bit pixel counter as shown by the literature survey. Such pixels have a regular layout structure that emphasises sensitivity over functionality. The limited dynamic range of these designs is compensated for by oversampling with time-resolved [78] and high speed tracking [8] applications being demonstrated. Extending the dynamic range of such simplistic pixels even further is the topic of discussion of the next chapter (Ch.4).

Another solution to maintain both benefits of high sensitivity and integrated functionality of miniature pixels allowed by global well sharing is to opt for more advanced CMOS processes such as 3D-stacking. This has the added benefits of improving uniformity due to the direct 1-to-1 connection between the SPAD and its pixel circuits and scalability due to the tileable pixel unit. This will be the topic of discussion of Chapter 5.

4. High Dynamic Range Oversampled Binary Image Sensor

This chapter explores techniques of extending the dynamic range of oversampled binary image sensors, also known as quanta image sensors (QISs). Based on the MINIC40 pixel presented in Chapter 3, the configurable counter is operated with triple exposure settings resulting in a dynamic range in excess of 100dB. The latching front-end gating circuit permits partial in-pixel binary frame summation resulting in a data rate compression of 3.75× compared to typical QIS operation.

The chapter is split into two sections. The first provides an overview of the QIS concept and high dynamic range (HDR) CMOS circuits shedding the light on their applicability to SPAD designs. Published high dynamic range SPAD pixel implementations are also reviewed. The second presents detailed analysis and measurement results of the multi-exposure HDR technique with conclusions drawn regarding its suitability for miniature sensor architectures.

4.1. Technology Overview

As concluded in the previous chapter, one direction for miniaturising SPAD pixels while maintaining high sensitivity is to resort to single-bit designs. These binary pixels require oversampling to form an image with dynamic range content and are different in operation principle to mainstream integrating CISs where the industry has undergone a lot of development in order to boost DR. Therefore both the oversampled binary imaging paradigm and the conventional HDR techniques are summarised.

4.1.1. The Quanta Image Sensor

In 2005 Eric Fossum proposed a new paradigm in image capture known as the digital film sensor (DFS) [249], latter called the quanta image sensor (QIS) [77], inspired by the statistical principle behind photographic film. Unlike the standard image sensors model where each pixel produces an output linearly proportional to the amount of incident light usually within a single frame or sample, the QIS model differs in two aspects:

1. Each QIS pixel, also known as a jot, produces a binary value whereby a logic 1 means at least one or more photons have been captured and a logic 0 means no photons have been captured. Thus if a single QIS frame is to be displayed it will look like an apparently random scatter of 1's and 0's. This binary frame is therefore called a bit plane or a field.
2. Since the sparse content of a bit plane is not sufficient to provide an image with context or bit depth, a content rich image is formed by capturing and summing a number of bit planes or fields to form a single regular frame. In this sense the QIS is an architecture that relies on oversampling (temporally and / or spatially) to create the final image product. For example, if

a standard image sensor can give an 8-bit grey scale image in a single frame capture, a temporally oversampled QIS would need to capture and sum 255 fields (i.e. forming one frame) to result in the same image bit depth.

Figure 4.1.1 demonstrates the bit planes and oversampling QIS principles.

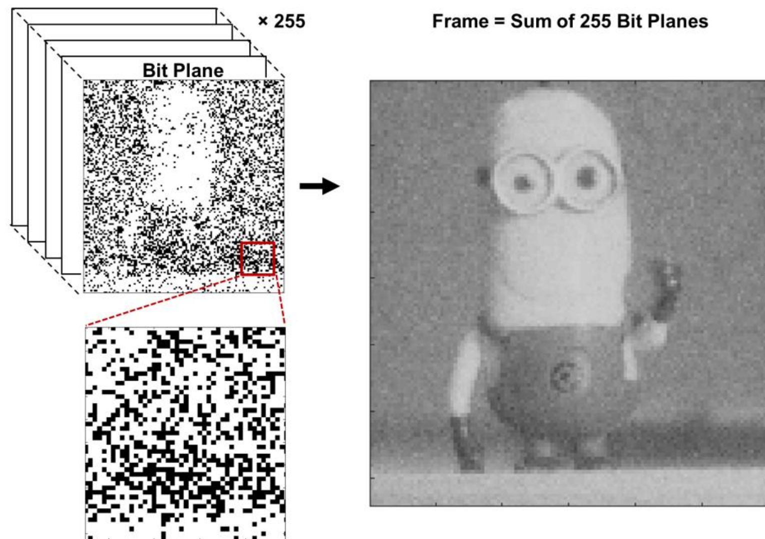


Figure 4.1.1. Principle of operation of a QIS where multiple binary bit planes are summed to form a single frame or an image.

The response of a QIS sensor (blue) is modelled in Figure 4.1.2 (see Section 4.2.2 for equations) with the x-axis representing exposure H and the y-axis representing bit density D. D is defined as the ratio of the number of pixels that detect one or more photons (i.e. binary value of 1) to the total number of pixels in the field at every exposure setting, hence the term bit density. The response of an ideal linear CIS is also modelled (red) for comparison.

A couple of interesting features are observed. The first is that the QIS response has an 'S' shape curve resembling that of photographic film first reported in [250]. This is due to the Poisson arrival statistics of photons where D is the probability of at least one photon arriving at the detector and H is the average number of arriving photons within a time interval. The QIS binary pixels in this context resemble the discrete silver-halide atoms of photographic film.

The second feature is that for low exposures, the response of a QIS follows that of a conventional CIS meaning it is linear in that region. While for high exposures, the QIS sensor deviates from the linear response reaching its saturation limit later than a conventional CIS. Therefore it is easy to see that; despite the QIS pixel having a single-bit photon counting capacity, it naturally exhibits a higher dynamic range than a conventional CIS due to the photo-response compression. It is possible to extend the dynamic range of a QIS even further as shall be demonstrated in Section 4.2.

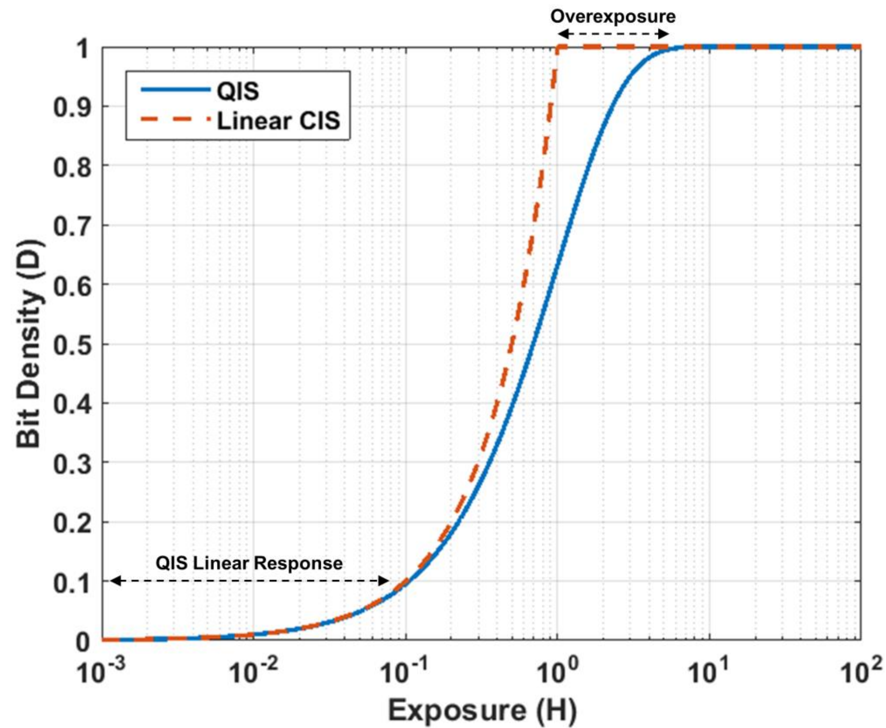


Figure 4.1.2. Quanta image sensor (blue) bit density D versus exposure H response (DlogH) along with the corresponding response of an ideal linear CIS (dashed red) for comparison.

Since the inception of the QIS concept, a lot of research has gone into achieving such sensors mainly through designing and optimising single photon counting pixels. In general there are two approaches towards this goal represented by:

1. High conversion gain low readout noise active pixel sensor (APS) [251][252] jots or pixels with small full well capacity and readout noise below 0.3e- which is the requirement for single photon counting as modelled by N. Teranishi [253].
2. Avalanche based pixels primarily on SPAD designs which leverage the high avalanche gain to render readout noise sources negligible and thus provide single photon counting capability.

It is perceived that the more favourable approach to implement a QIS is the jot APS as it leverages readily available CMOS image sensor (CIS) processes with minimal modifications [254], the higher quantum efficiency of photodiodes compared to SPADs, advances in low noise and low power readout circuits [255] and most importantly scalability as current jot pixels achieve a pitch of 1.1µm [256] in line with mainstream camera pixels which paves the way towards very high resolution sensors.

On the other hand, SPAD pixels do require an optimised process to achieve good performance parameters, have lower detection efficiency, consume more power and are limited in scalability with

the smallest reported SPAD-only pixel at 3.0μm [119]. Such limitations would eventually prohibit the design of large format SPAD QIS arrays.

Despite that, SPAD QISs do have the advantages of temporal resolution and time gating which are beneficial in niche applications such as FLIM [257] and super resolution microscopy [258]. Moreover, the high speed of SPAD QISs lends itself to other applications such as Brownian motion particle tracking [68], molecule localisation and blinking characterisation [259] and object tracking and image reconstruction of high speed scenes [260].

Such properties, combined with the fact that single-bit pixels are an attractive approach for miniature high sensitivity SPAD arrays, makes SPAD QIS architectures merit the effort of investigating their dynamic range behaviour.

4.1.2. High Dynamic Range Imaging

The concept of extending the dynamic range of mainstream image sensors goes back a long way to the starting days of CMOS image sensors and has been explored thoroughly by many groups. The early motive behind such research was to create a sensor that is capable of imaging the widely varying environment around us which ranges from very dark starlight conditions to very bright sunlight. Figure 4.1.3 highlights such dynamic range spanning more than seven decades of illumination or more than 140dB.

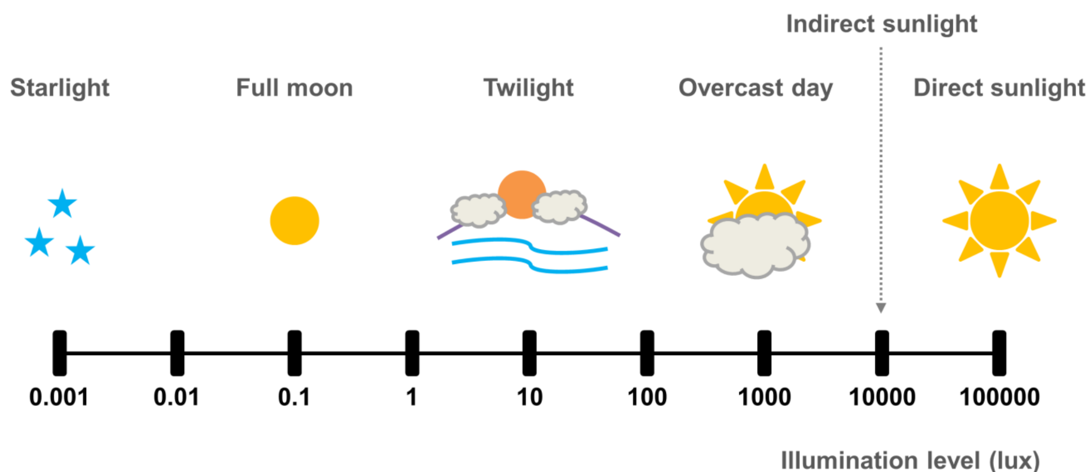


Figure 4.1.3. Range of illumination levels exhibited in day to day environments.

Other than the appeal of capturing high quality images for the high volume consumer market, computer vision has been another motive behind high dynamic range (HDR) imaging development, specifically automotive applications for driver assistant systems and more recently autonomous driving cars. Such applications are required to understand and interpret the world with the user's safety at stake and so need to reliably see beyond the limits of the human eye.

Generally speaking the different approaches of extending the dynamic range of image sensors explored in literature include but are not restricted to:

1. Non-linear or logarithmic pixels.
2. Well adjustment techniques.
3. Time to saturation (TTS) pixels.
4. Light to frequency converters.
5. Multi-exposure or multiple sampling architectures.

In this context dynamic range is defined as:

$$DR = 20 \times \log\left(\frac{S}{N}\right) \tag{1}$$

Where S is the saturation signal level and N is the input referred readout noise floor. There are two options for improving the DR of a sensor, one can either extend the maximum signal level to accommodate the higher end of light intensities, or reduce the noise floor to extend the DR towards the lower end of the scale. While in practical solutions a combination of both options is adopted, the five solutions above focus mainly on the former one.

One of the early adaptations of a non-linear pixel is depicted in Figure 4.1.4 [261]. Transistor M0 is operated in weak inversion whereby its gate-source voltage is logarithmically dependent on incident light levels and so provides a high dynamic range transfer function. Transistors M1 and M2 form the source follower and readout select switch respectively. The advantages of such a pixel are its simplicity and small number of components but it does suffer from high fixed pattern noise (FPN) due to the V_T mismatch of M0, loss of contrast due to the logarithmic compression and complicated post processing due to the ADC's linear quantisation of the pixel's logarithmic response.

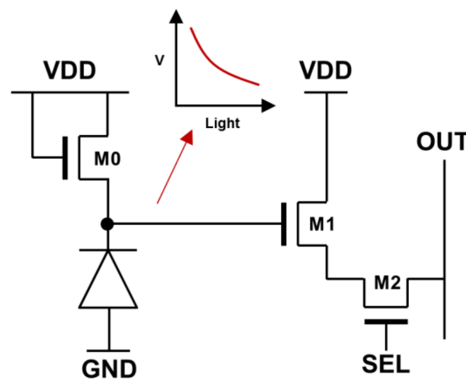
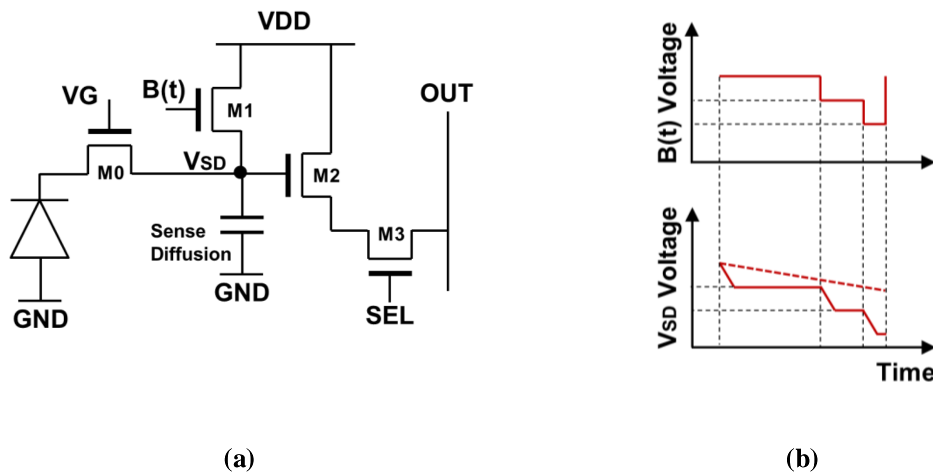


Figure 4.1.4. Logarithmic high dynamic range pixel [261].

Different variants of this concept were proposed including a combined lin-log architecture where both responses can be obtained alternatively [262], a logarithmic pixel with on-chip calibration for improved FPN [263] and an all included lin-log pixel with built in calibration mechanism [264]. Other non-linear pixel architectures which do not rely on a MOS device in weak inversion were proposed such as [265] where by modulating the gate voltage of a MOS switch in between the photodiode and the source follower gate node a logarithmic HDR response is effectively obtained.

Well adjustment techniques are among the common HDR solutions with the work by Decker et. al. [266] being amongst the most prominent examples. Figure 4.1.5(a) shows the corresponding circuit diagram with M0 acting as a charge spill control (not crucial to the HDR operation), M2 and M3 are the source follower and readout switch respectively while M1 is the most critical component known as the lateral overflow transistor.



(a) (b)

Figure 4.1.5. Lateral overflow concept from [266]. (a) Circuit diagram. (b) Timing diagram with top waveform representing the barrier control voltage B(t) and the bottom waveform representing the sense diffusion voltage for high light (solid red) and low light (dashed red) conditions.

The gate voltage of M1 is controlled by a time varying function B(t) (decreasing voltage) which creates a potential barrier (increasing barrier) between the sense diffusion and the lateral overflow drain which influences the integrated charge in the photodiode depending on incident light levels. If the amount of incident light is high there will be durations within an integration where the photodiode cannot integrate further and is said to be barrier limited until the barrier potential is modified. This is illustrated as the red solid V_{SD} line in Figure 4.1.5(b). Alternatively if the incident light is low the change in barrier potential has no influence on the photodiode charge integration and it is said to be free flowing as in the red dashed V_{SD} line in Figure 4.1.5(b).

The advantage of this technique is that it efficiently reuses the pixel resources to compress the high signal response. On the down side it does require tight control of the voltage and time steps of the barrier potential increasing the system complexity. Other well adjustment techniques include the use

of a switched lateral overflow integration capacitor as in [267] or increasing the pixel floating diffusion capacitance by borrowing the floating diffusion of a neighbouring row pixel via in-pixel switches trading off sensitivity for dynamic range [268].

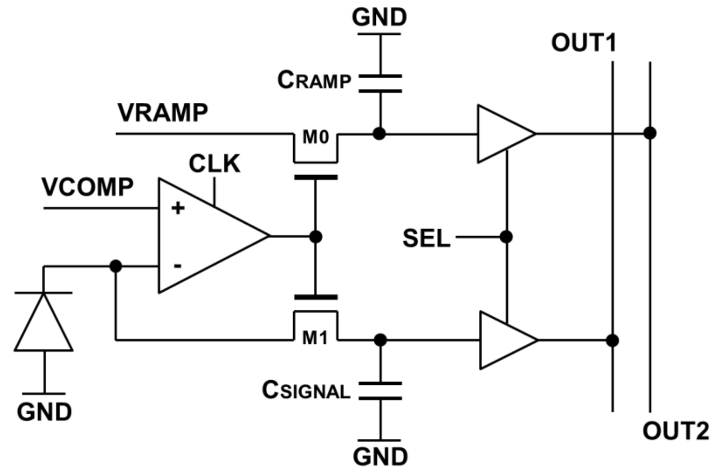


Figure 4.1.6. Basic schematic of time to saturation pixel proposed in [269].

While the above two approaches offer compact pixel designs for improved dynamic range the challenge of resolving FPN artefacts and the processing of compressed signals lead to the development of a new class of HDR pixels known as time to saturation (TTS). The concept of TTS pixels is that the integrated signal is continuously compared to a threshold voltage that if met, the pixel halts its integration to avoid saturation and simultaneously a secondary time varying voltage waveform (usually a ramp) is sampled in-pixel to indicate the time at which the pixel reached the threshold limit.

An early example of such a pixel is demonstrated in Figure 4.1.6 [269] where an in-pixel comparator compares the integrated photodiode signal to VCOMP at specific time intervals within an integration period by a clock trigger CLK. If the integrated signal passes VCOMP then the signal is sampled onto CSIGNAL and VRAMP onto CRAMP via switches M0 and M1, else integration continues with the subsequent clock samples happening at progressively increasing time intervals until the condition is met. VRAMP is a stepped voltage signal which encodes the CLK sample number and so the total time interval taken by the pixel to reach the threshold. Two outputs are available from the pixel, OUT1 which is the integrated signal value even if the pixel never reaches its threshold limit for low light conditions, and OUT2 which is the sampled ramp voltage.

Other implementations of TTS pixels include an asynchronous design where two ramp voltages with different waveform shapes are sampled in-pixel to provide good resolution and linearity with the option to read the integrated signal value [270] and a similar design by the same group with an on-

chip voltage waveform generator that is programmable to adapt to the range of illumination levels exhibited in the scene [271]. Another advantage of TTS architectures is that they allow for adaptive integration time on per-pixel basis but at the expense of large pixel pitch and analogue circuit non-uniformity which limits their scalability to high resolutions.

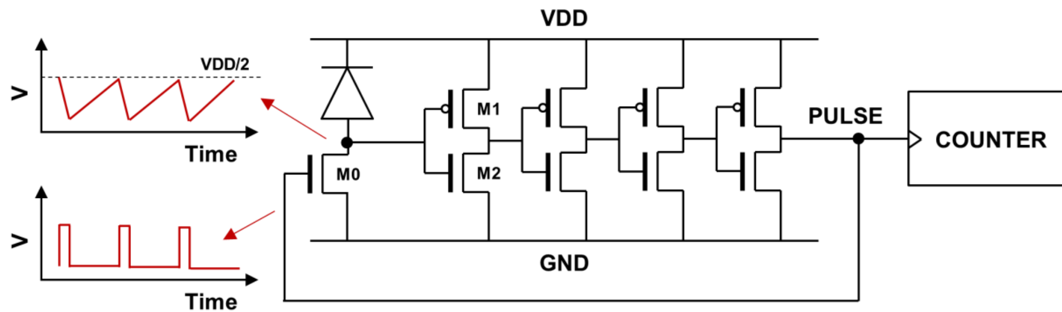


Figure 4.1.7. Basic principle of light-to-frequency pixel reported in [272].

Another class of HDR pixels is light-to-frequency converters. The operation principle of these pixels is simple whereby the photodiode is left to integrate up to a threshold voltage at which point a pulse is generated. This pulse has two purposes, first is to increment a counter, and second to reset the photodiode such that it starts integrating again. The number of pulses counted within an exposure period corresponds to a frequency; hence the name light-to-frequency converter, which is proportional to the level of incident light.

One of the first demonstrations of this technique was reported in [272] and depicted in Figure 4.1.7. The photodiode integrates charge until it hits the threshold level of the front end inverter formed by M0 and M1 which triggers an edge through a cascade of inverter elements. These delay elements are required to generate a feedback pulse long enough to reset the photodiode and trigger the counter.

A modified architecture of the same concept was reported in [273] where charge is integrated on a feedback capacitor of an integrator instead of the photodiode itself in order to maintain a constant bias across the photodiode and avoid light dependent characteristics. A more sophisticated version of this design was later reported in [274] where a hybrid analogue-digital pixel counts pulses and also allows for readout of residual analogue signal to provide both coarse and fine conversions.

Similar to the TTS pixels, light-to-frequency designs tend to be large in area but allow for low power implementations and no parasitic light sensitivity due to the in-pixel digital conversion and storage.

The final HDR technique reviewed in this section utilises multi-exposure or multiple sampling architectures. There are many manifestations of this method but the basic principle relies on capturing multiple image samples; two in the simplest case, at different exposure settings; namely short and long, and then combining them to form the final image. The simplest implementation of multi-

exposure image sensors was reported in 1997 by two different groups, Olympus [275] and JPL / Photobit [276].

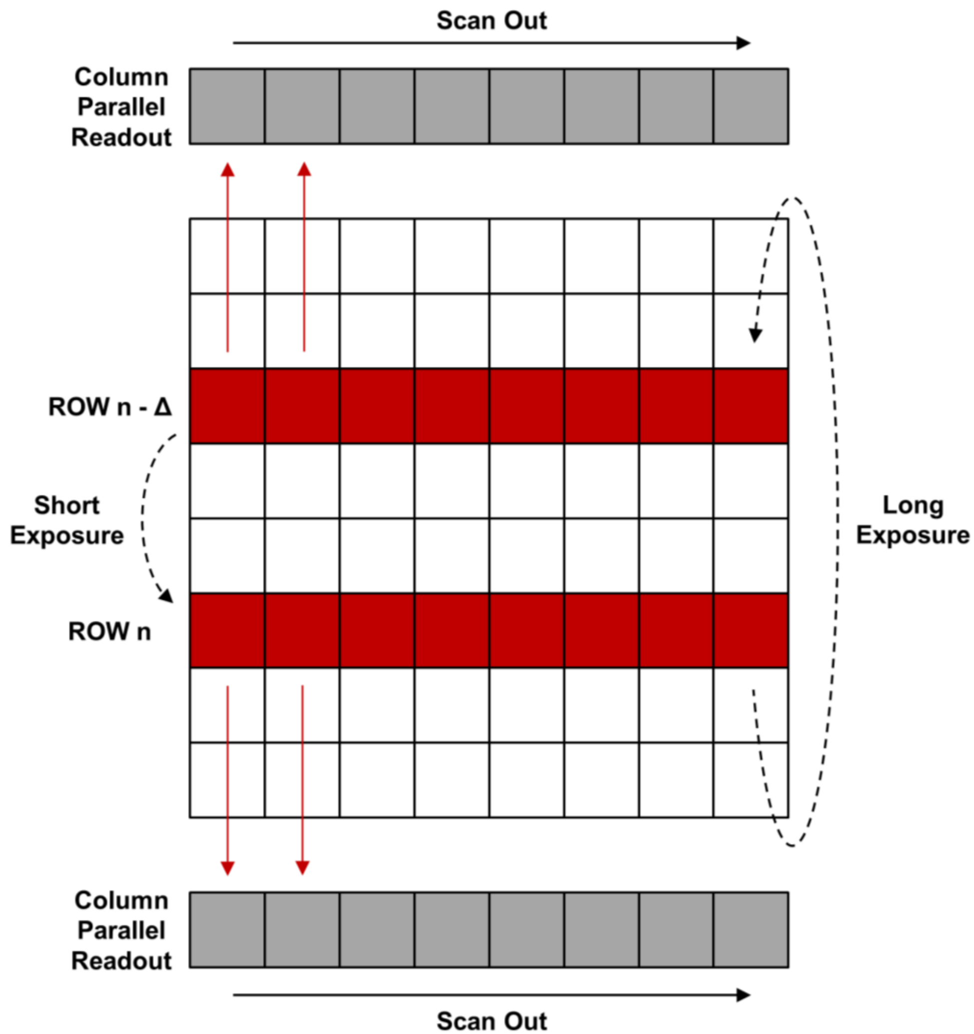


Figure 4.1.8. Dual scan dual output HDR architecture reported in [276].

The Olympus approach performs two scans of the image field where the first scan represents the short exposure and the second represents the long exposure. After the short exposure the pixel values are read out non-destructively while the conventional reset is applied after the read of the long exposure values. The JPL / Photobit approach differs slightly by having dual shutter lines and dual readout columns where the short and long exposure times are defined by the time difference between the two shutter blades as demonstrated in Figure 4.1.8 and the pixel values for each exposure are read through independent signal chains simultaneously.

Some of the other innovative multiple sampling architectures include a pixel-level single bit ADC where bit planes of multiple comparisons of the pixel value against a reference signal take place at exponentially increasing time intervals resulting in a floating point conversion of the pixel value

[277]. The work in [278] presents another approach where multiple samples are captured at different integration times and gain settings and the optimum settings based on saturation prevention criteria for each pixel are digitally encoded in an off-chip memory alongside the pixel digitized value for post processing. Overall the multi-exposure or multiple sampling techniques offer better linearity and signal to noise ratio (SNR) than other HDR approaches [279] but have the downside of requiring frame store memory, high speed readouts and vulnerability to motion artefacts.

The field of HDR is vast and includes many other techniques that were omitted from this brief review ranging from sensor architectures such as down-sampling and split-diode [280] to novel ADC architectures with non-linear slope characteristics [281] or adaptive variable resolution [282]. The interested reader is referred back to the extensive literature on this fascinating topic.

4.1.3. High Dynamic SPAD Image Sensors

Many of the HDR techniques discussed above do not directly apply to SPAD sensors due to the fundamental difference of operation between integrating and avalanching photodiodes, yet the circuit techniques can inspire novel analogue photon counting pixels such as logarithmic counters [283] and some of the concepts can be readily applied such as multi-exposure sampling.

SPAD devices can intrinsically offer a wide dynamic range response in excess of 100dB by virtue of low dark count rate noise floor and high maximum count rate due to short device dead-time [140] but it is difficult to capture this response as it requires a very large pixel photon counting capacity or full well. In digital architectures this translates to a large bit depth counter with negative impact on pixel pitch and fill factor, and in analogue architectures no practical demonstration has exceeded a counting capacity beyond a few hundred photons [184] while overcoming pixel variability and noise constraints.

Recent digital implementations in advanced CMOS processes [10] leverage miniature logic gates to fully capture this intrinsic DR at a pixel size of 20μm × 40μm in an FSI implementation which would easily port to a 20μm × 20μm equivalent in 3D-stacked BSI one but at the cost of expensive technology nodes.

A more novel approach to increase the DR of an avalanche sensor was presented by Panasonic [284] with a fully customised BSI device which can switch between integrating photodiode mode (PD) and avalanche photodiode mode (APD) by adjusting a global reverse bias voltage. When operated in integrating photodiode mode the conventional 3.8μm 4-transistor APS pixel gives an output voltage linearly related to incident light levels offering a conventional DR of 60dB.

When operated in avalanche photodiode mode the sensor becomes a single photon quanta device offering an additional 40dB of DR at lower light levels with a large output voltage swing interpreted as a binary signal. This sensor is also the first demonstration of a mega-pixel APD device and further

results were presented in [285]. The downside of this approach is the high voltage switching time of 1ms required which limits the sensor to an HDR frame rate of 15fps (15fps in PD mode + 15fps in APD mode). Figure 4.1.9 demonstrates the idea behind combining the photodiode modes to enhance DR.

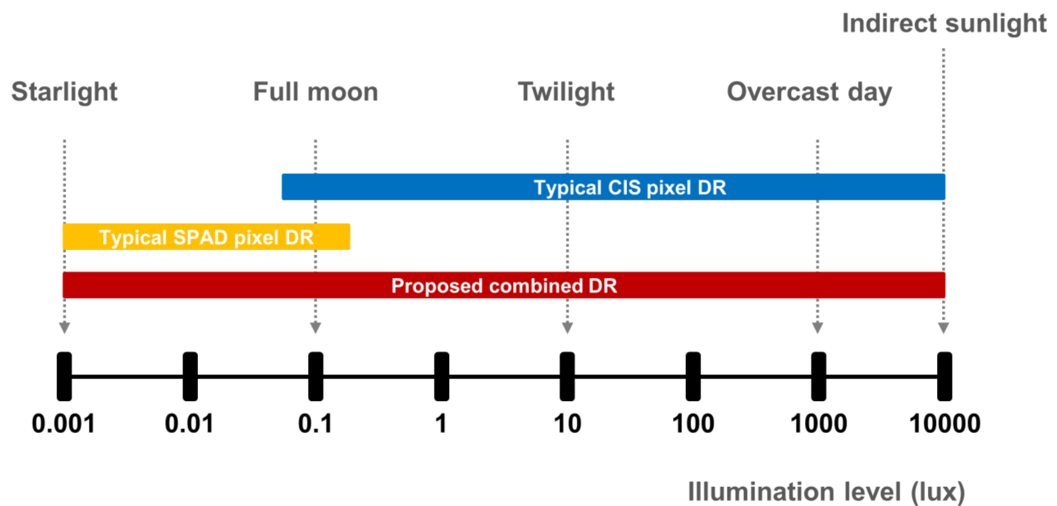


Figure 4.1.9. Dual photodiode mode concept to enhance dynamic range proposed by [284].

More recently another way of extending the DR of SPAD sensors by operating devices in both PD and SPAD modes was presented by [243][286] where a pixel inspired by the light-to-frequency HDR technique provides a readily digitised HDR response. The pixel comprises an integrator, a comparator, a quench and recharge circuit (QRC) and a digital counter. In PD mode the integrator accumulates charge until the comparator threshold is met at which point a trigger is issued resetting the integrator and incrementing the counter. This corresponds to the light-to-frequency conversion circuit and is used for higher illumination levels.

At low light levels the diode is operated in SPAD mode to improve sensitivity and the SPAD pulses trigger the counter to provide a corresponding count rate. By alternating the device between both modes an HDR response was achieved. In comparison to the Panasonic approach, the disadvantage of this implementation is the large pixel pitch (80μm) needed in an 180nm process to integrate both analogue components and digital counters.

4.2. High Dynamic Range SPAD Quanta Image Sensor

Without using large bit depth counters or alternating between device modes of operation, it is also possible to obtain an HDR response from an oversampled SPAD quanta image sensor by combining image fields at different exposure settings into a final HDR image.

In 2013, Eric Fossum presented the theory of single and multi-bit QIS devices alongside the modelling of extended DR by the aforementioned principle [287]. The following section presents the first demonstration of high dynamic range imaging at the quantum limit in a time and data efficient way using the MINIC40 SPAD sensor [4].

4.2.1. Challenges of Multi-Exposure QIS

While in principle it is easy to acquire oversampled QIS image fields at different exposure settings to form an HDR image this presents two system challenges:

1. For an assumed 30fps operation where each frame is a summation of 255 bit planes or binary fields for an 8-bit greyscale image, the sensor has to have a field rate of 7650 fields per second (fips). So for the case of three exposure settings the sensor's frame rate will be cut by a factor of three to 10fps in HDR mode or if the 30fps rate is to be maintained the sensor has to operate at a field rate of 22950fips. Therefore, multi-exposure HDR mode presents a system trade-off between effective frame rate and sensor speed and power consumption.
2. As a consequence of oversampling and the potential need to acquire triple the number of fields as in the example above, the sensor will have to produce large amounts of data in quick bursts which scales with the image spatial resolution. For the above example and at a nominal HDR 30fps operation for a 128 $\times$ 128 resolution, the data rate reached is 376.01Mbps, or 1.50Gbps for a 256 $\times$ 256 resolution. This presents a system challenge in terms of data handling and readout output channels power consumption.

To alleviate these system challenges, the MINIC40 sensor offers two unique features:

1. The ability to capture three image fields with different exposure times in parallel. When operated in time gated mode the in-pixel counter splits into three independent counters where each can be triggered at a different time window or exposure setting. This overcomes the effective frame rate trade-off since multi-exposure settings are acquired simultaneously. This does not solve the data rate challenge however. For the same 128 $\times$ 128 7650fips sensor the number of bits per second still equates to 376.01Mbps since now each field image is a combination of three simultaneous bit planes. On the other hand, being able to acquire the different exposure fields in parallel improves motion induced artefacts in the final synthesised image since the fields are not acquired over different slices in time.

2. To address the data rate issue, each 4-bit counter allows for summing 15 fields (i.e. sub-frame) at each exposure before a readout operation is required, so instead of reading out 15 binary values per sub-frame, these can be represented by an in-pixel 4-bit sum providing a data compression ratio of 3.75×. The same 128 × 128 sensor now has a reduced data rate of 100.27Mbps for the same triple-exposure 8-bit depth specification. The downside of the improvement in data rates is loss of raw field data at high field rates which would be useful for tracking fast moving objects [8] or imaging high speed phenomena [288].

Figure 4.2.1 shows a comparison of the HDR frame timing for a conventional 1-bit pixel QIS device and MINIC40 with three 4-bit counters per pixel. A global shutter triple-exposure acquisition and rolling readout operation is assumed for an 8-bit depth final image specification.

For conventional operation, an HDR frame is formed of three consecutive frames of short, mid and long exposures. Each of these frames is a sum of 255 (8-bit) binary image fields where the single bit pixels are exposed, readout and then reset.

On the other hand a MINIC40 HDR frame is composed of 17 HDR sub-frames each comprising an exposure sequence, readout and pixel counters reset. During an exposure sequence, 15 repetitions of binary fields are acquired for three back-to-back exposure settings with a reset to pixel latching front end after every repetition. The consequence of this HDR timing is that the multi-exposures are captured contiguously rather than over three distinct slices in time and the in-pixel partial field summation to create sub-frames results in data compression.

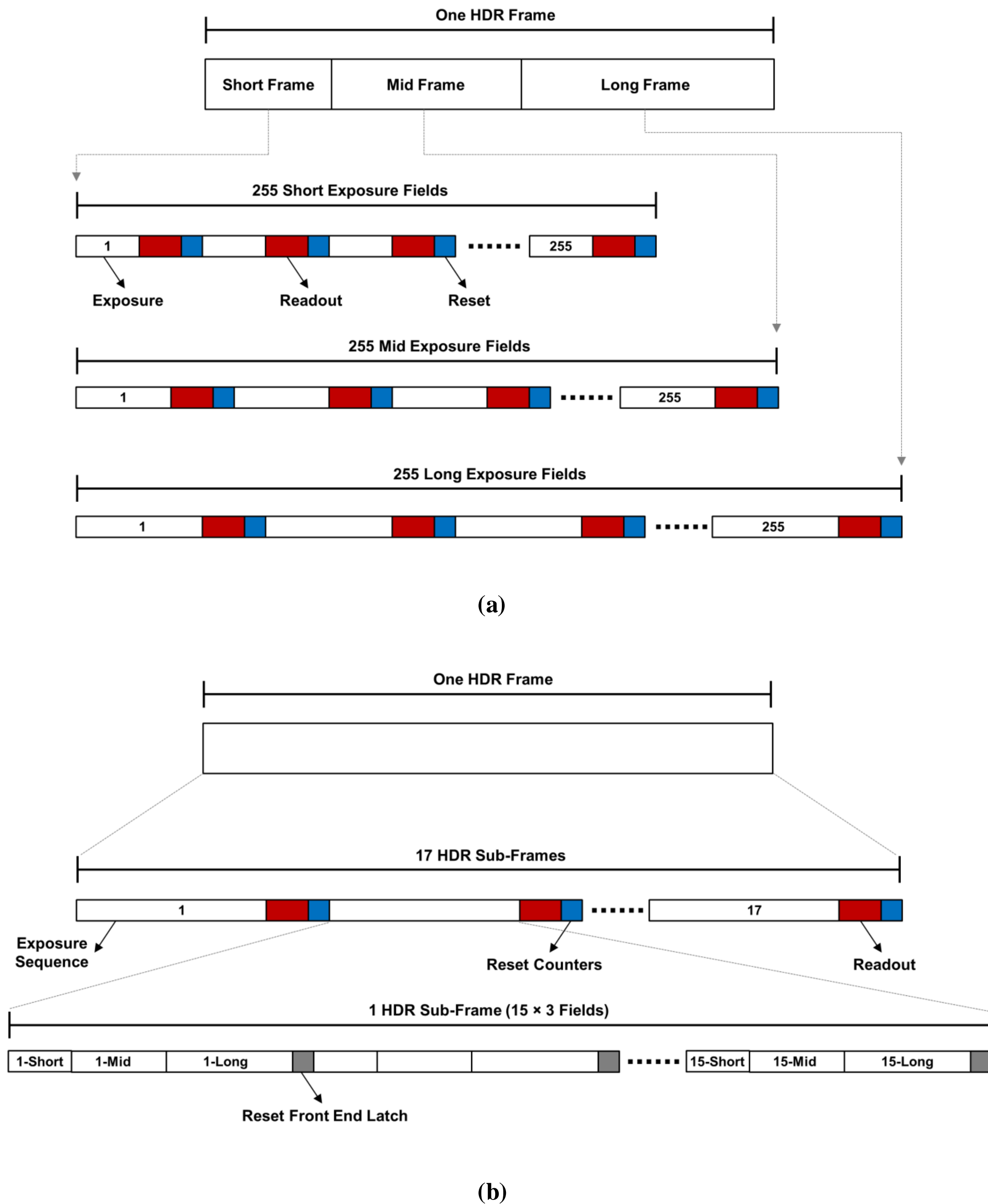


Figure 4.2.1. HDR timing diagram comparison for global shutter triple-exposure acquisition and 8-bit depth image specification. (a) For conventional single bit QIS. (b) For MINIC40 with three in-pixel 4-bit counters.

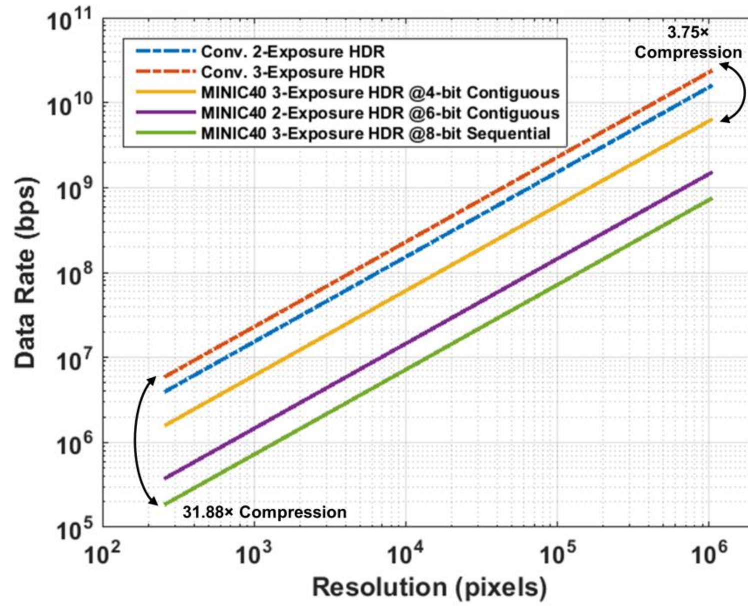


Figure 4.2.2. Data rates of different HDR architectures and sensor resolutions assuming an effective HDR frame rate of 30fps and 8-bit greyscale images per exposure frame.

Figure 4.2.2 shows a comparison between the required data rates for different HDR architectures and sensor resolutions assuming an effective HDR frame rate of 30fps and an 8-bit greyscale image depth per exposure frame. An expected increase in data rates is seen for the conventional QIS architecture when performing HDR imaging with three exposures (dashed red) up from two (dashed blue) while for the same 3-exposure setting MINIC40 (yellow) offers a 3.75× improvement.

Further reduction in data rates could be achieved if the same 12-bit MINIC40 counter is repurposed for 2-exposure captures at 6-bit depth each (purple). This of course comes at the cost of degraded HDR performance due to the use of 2-exposure settings instead of three.

An even bigger compression ratio greater than 30× with 3-exposure HDR mode can be reached if the MINIC40 in-pixel counter can be operated as a single 8-bit one (green), albeit in this case at the cost of sequential HDR frame captures similar to the conventional architecture as opposed to the contiguous one which is more prone to motion artefacts.

4.2.2. Analysis and Measurement Results

Before examining the high dynamic range properties of MINIC40 as a quanta image sensor it is worth taking a look at the state of the art SPAD image sensors in terms of pixel pitch and photon counting (or FW) capacity to gain an insight into their performance limits. Figure 4.2.3 provides a survey of relevant published works in 2D-array formats spanning analogue, digital, 3D-stacked digital and APD sensors.

All sensors with a single-bit pixel are considered QIS devices. As for the others, the relatively larger FW allows them to operate in linear integration mode where the DR can be calculated by Equation 1. With the exception of two very large pixels (circled), The MINIC40 sensor in 12-bit linear mode stands above the crowd with a remarkable 4095 photons capacity at 8.25μm pitch. For any given integration time and a single frame capture, the maximum achievable DR for this pixel assuming a single dark event noise floor is 72.2dB.

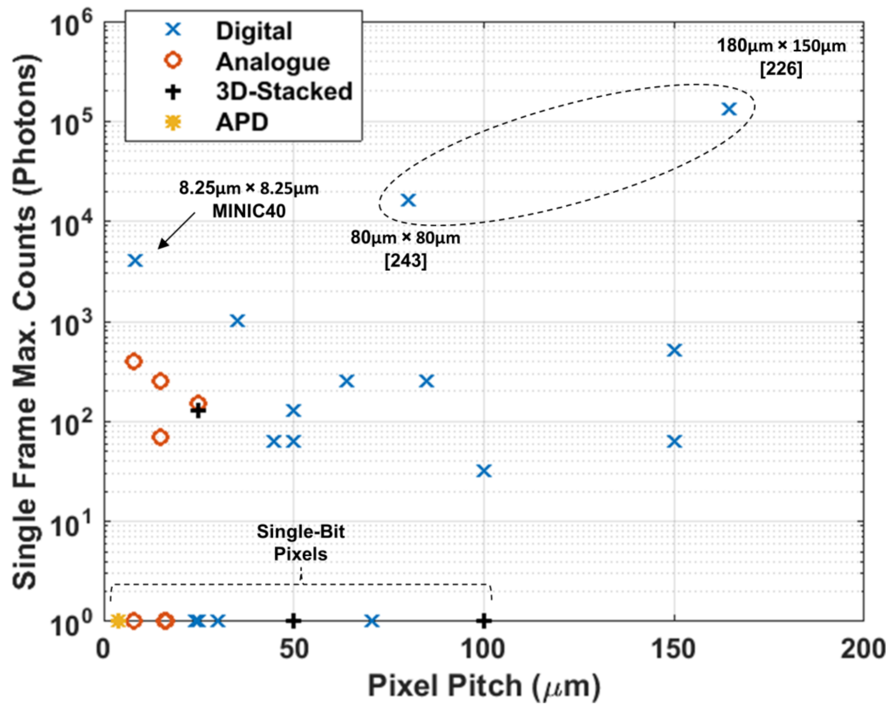


Figure 4.2.3. State of the art survey of maximum single frame photon counting capacity versus pixel pitch for different SPAD sensors grouped by architecture.

Figure 4.2.4 illustrates the photon transfer curve (PTC) of a single pixel in linear counting mode to confirm that the photon counting mechanism of the SPADs and the image sensor is entirely shot-noise limited. The red-line is a model of shot-noise limited single photon counting and there is minimal deviation of measured results from the ideal model. The measurement was limited to a signal level right before the pixel counter saturation limit to avoid roll-over since no overflow mechanism was implemented.

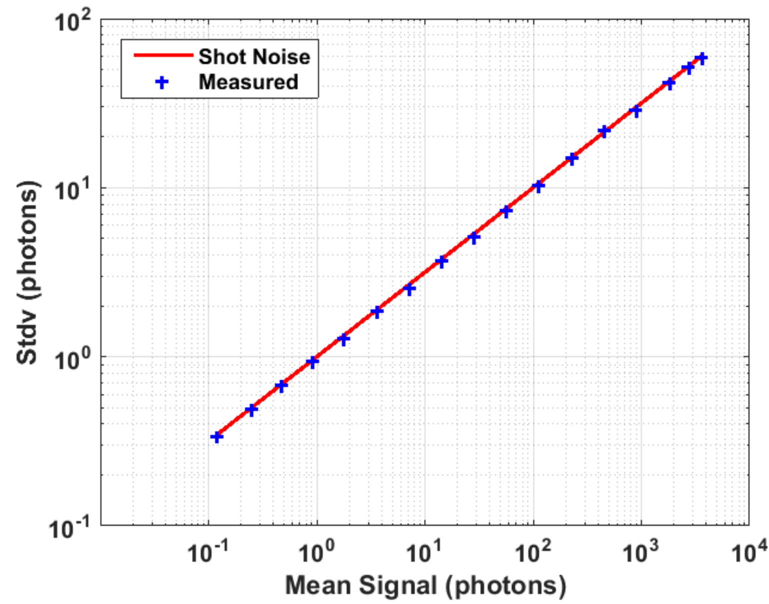


Figure 4.2.4. Photon transfer curve for a single pixel operating in 12-bit linear counting mode showing shot noise limited photon counting. Red line is theoretical shot noise limited response.

To demonstrate the sensor's quanta response the current through an LED source was swept while data was captured at a variety of exposure settings. For each light point, a total of 50 bit planes or fields of 96 × 40 pixels were spatially and temporally combined to result in a total of 192000 ensembles M. For the purpose of speeding up the measurement all of the 96 × 40 pixels were spatially summed to contribute towards the total number of ensembles, while in a practical QIS use case a smaller subset of pixels or jots (8 × 8 for example [256]) would be spatially summed to represent one image element. The bit density D vs the input signal H curves were produced by dividing the total number of counts at each light point by M.

Figure 4.2.5 shows the measured QIS response for a photon threshold K of 1 where a pixel is assigned a binary value of 0 for no photons detected and a binary value of 1 for one or more photons detected. This binary assignment is performed by the in-pixel gating and counting logic depicted in Figure 3.2.19.

Two scenarios were explored where 3 different exposures of ratios of 10 (0.1µs, 1µs and 10µs) and ratios of 2 (0.1µs, 0.2µs and 0.4µs) were used. The x-axis is normalised such that an input signal of H=1 yields a bit density D=0.63 for the shortest exposure setting of 0.1µs. This is known as the full exposure point as defined by [287]. The 0.1µs exposure setting is chosen as the reference as it is the common setting across all measurements to follow. The modelled QIS response for this exposure is shown as the dashed red line where D is defined as:

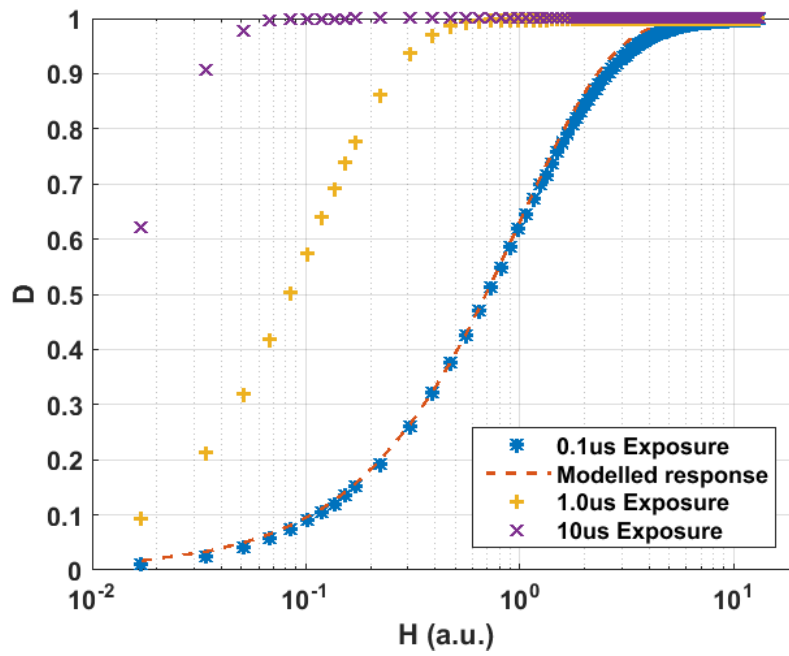
$$D = 1 - e^{-H} \quad (2)$$

The measured data exhibits some deviation from the ideal model which could be attributed to illumination non-uniformity, photo-response non-uniformity and temporal variations as measurements were acquired over hours which would all contribute to the error in the spatio-temporally oversampled data. Moreover the light source used was characterised and found to exhibit non-linearity at higher input currents (i.e. higher output optical power).

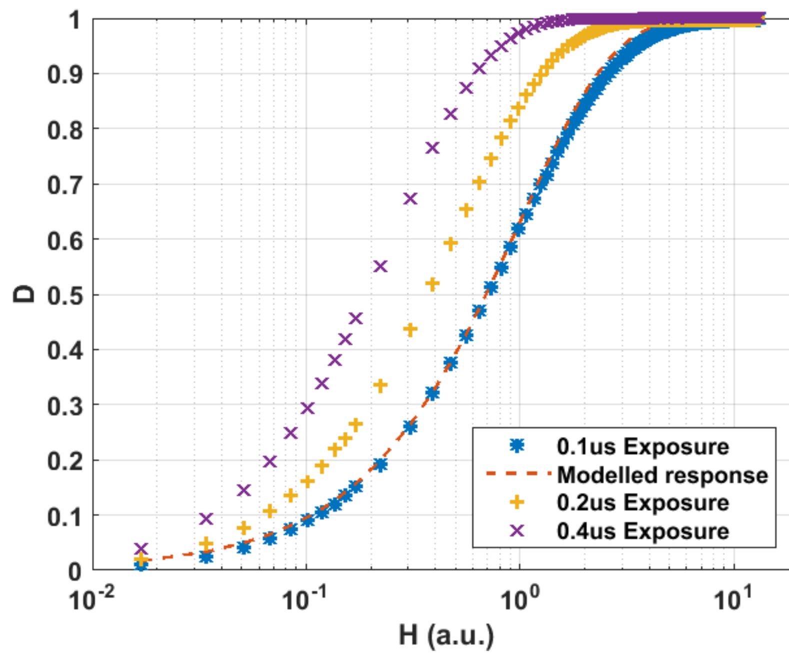
Nevertheless, the measured data offers a qualitative insight into QIS behaviour. As can be seen from the results of the longest exposure setting of 10µs, it was not possible to reach low bit density values due to the limitations in the illumination source used. The author opted for not combining data acquired by using different neutral density filters to avoid adding in more error.

The measurement was repeated for an emulated photon threshold of K=2 (pixel assigned a binary value of 0 for no photons or one photon detected and binary value of 1 for two or more photons detected) by using linear counting mode (12 bit) and three sequential exposures. This emulation is necessary due to the latching single bit (K=1) front end in HDR mode. 50 single frames (no on-chip summation) were captured for each exposure setting where each pixel exhibits photon counts between 0 and 4095.

By post processing the captured intensity frames the pixel values were re-assigned to transform the frame into a binary bit-plane or field. In the future an improved pixel design with multi-photon triggering could achieve the variable K threshold in-pixel. This variable threshold adjusts the non-linear intensity to exposure characteristic which is an interesting property of the QIS. The same exposure ratio settings were used and DlogH curves are shown in Figure 4.2.6.

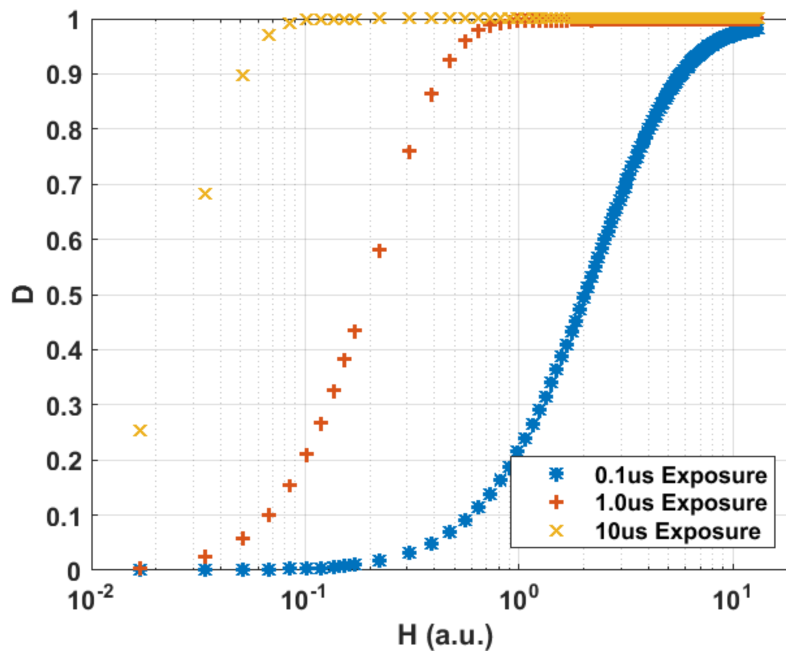


(a) Exposure ratio of 10, K=1

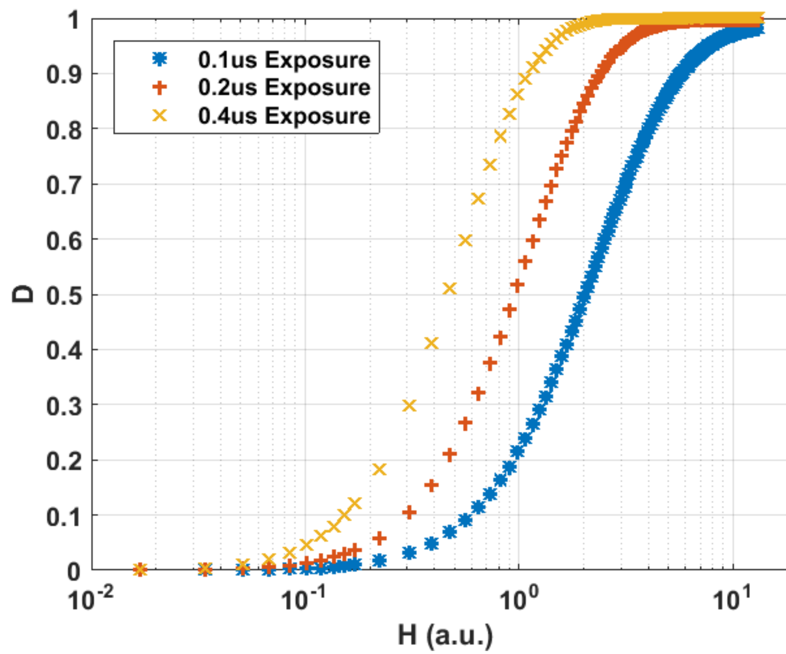


(b) Exposure ratio of 2, K=1

Figure 4.2.5. Measured normalised intensity (D or 'Bit plane density') to normalized input signal (H) for two sets of integration times for 1 photon threshold (K=1). (a) Exposure ratio of 10 with Short = 100ns, Mid = 1µs, Long = 10µs. (b) Exposure ratio of 2 with Short = 100ns, Mid = 200ns, Long = 400ns.



(a) Exposure ratio of 10, K=2



(b) Exposure ratio of 2, K=2

Figure 4.2.6. Measured normalised intensity (D or 'Bit plane density') to normalized input signal (H) for two sets of integration times for 2 photon threshold (K=2). (a) Exposure ratio of 10 with Short = 100ns, Mid = 1µs, Long = 10µs. (b) Exposure ratio of 2 with Short = 100ns, Mid = 200ns, Long = 400ns.

To evaluate the dynamic range (DR) and signal-to-noise ratio (SNR) of the quanta image sensor, and following from the theory presented in [287], DR is hereby defined as:

$$DR = 20 \times \log(\frac{H_m}{H_n}) \tag{3}$$

Where H_m is the H value at which the measured signal reaches 99% of its saturation limit and H_n is the H value equivalent to the noise level (read + dark). Since the digital sensor used here has no read noise as shown in (Figure 4.2.4), the only contribution to H_n is from the dark count rate (DCR) of the SPADs. For all measurements the SPADs were biased at 2V excess voltage for which the median DCR is ~150cps at room temperature [2]. Using Equation 2, and taking D to be 150cps $\times$ 0.1µs, the equivalent H_n is calculated to be 1.5e-5. This value was used for all DR calculations in this work while H_m was estimated from the measured signal.

It is worth noting that the number of ensembles M has an effect on DR as the minimum observable signal is one photon per M ensembles (or 1 / M), so for the maximum DR (DR_{MAX}) to be achieved it is necessary that the used number of ensembles is greater than the noise floor equivalent (i.e. M > [1 / D(H_n)]), else the DR will be limited by the ability to observe a signal. Since M of 192000 used in the presented measurements satisfies this condition, all DR figures reported herein represent DR_{MAX} which might not be achievable in a practical QIS scenario.

For SNR calculations an alternative exposure referred SNR or SNR_H definition was proposed by [287]. The objective of this definition is to project the SNR as measured in the y-axis (bit density D or voltage referred) onto the input x-axis (H). The reason behind this is that the voltage referred SNR will result in an artificial increase due to the compression of noise by the QIS response and so SNR_H is a more meaningful measure. SNR_H is defined as:

$$SNR_H = \frac{H}{\sigma_H} \tag{4}$$

Where σ_H is defined as:

$$\sigma_H = \sigma_{\text{Tot}} \times \frac{dH}{dM_{Tot}} \tag{5}$$

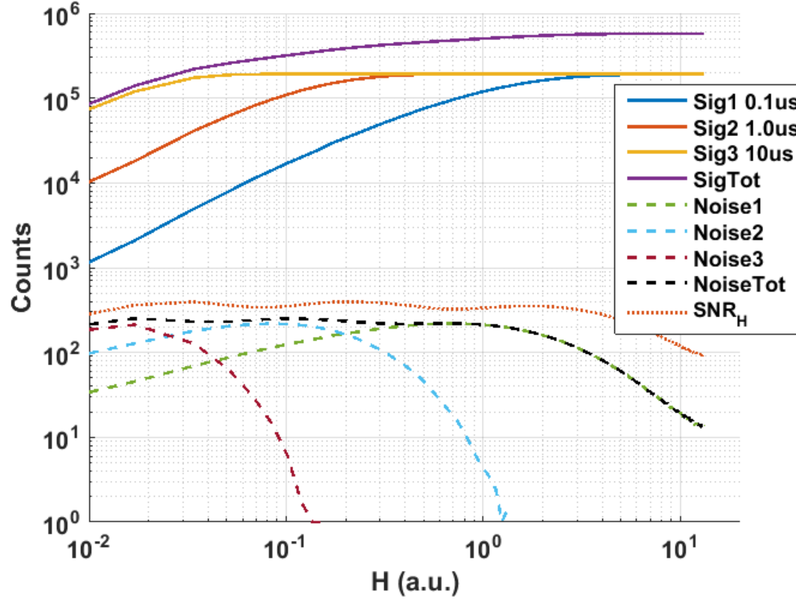


Figure 4.2.7. Measured signal, noise and SNR_H responses for 3 exposure settings with exposure ratio of 10 and photon threshold K=1.

Figure 4.2.7 shows the cumulative QIS signal response and SNR_H for photon threshold K=1 and three different exposures with a ratio of 10 (0.1μs, 1μs and 10μs). Sig1, Sig2 and Sig3 are the counts M1, M2 and M3 of the three corresponding exposures. SigTot (or M_{Tot}) is the linear summation of the counts of the three responses:

$$SigTot = Sig1 + Sig2 + Sig3 = M_{Tot} = M_1 + M_2 + M_3 \quad (6)$$

Noise1 is the standard deviation of Sig1 and under the assumption of Poisson statistics is given by [287]:

$$Noise1 = \sigma_1 = \sqrt{\frac{(M - M_1) \times M_1}{M}} \quad (7)$$

Where M is 192000 (50 fields × 96 × 40 pixels). Noise2 and Noise3 are defined similarly and NoiseTot is the total noise of the cumulative response and is defined as:

$$NoiseTot = \sigma_{Tot} = \sqrt{(\sigma_1)^2 + (\sigma_2)^2 + (\sigma_3)^2} \quad (8)$$

Hence it is possible to calculate SNR_H for the measured data from the above equations. While it is not possible to observe the rise of SNR_H at low H values due to the measurement setup limitations and the fact that the long 10μs exposure response masks the response from the shorter exposures at these low H values, it is interesting to see how SNR_H peaks forming a plateau region with very smooth

transitions or ripples when data from different exposures are summed as opposed to the dips in SNR observed in conventional image sensors.

Using the equations above, SNR_H and DR have been calculated for the cases of single, double and triple exposures with a ratio of 10 showing how DR increases from ~70dB to more than a 100dB in this example (Table 4.2.1).

Exposure	SNR_H (dB)	DR_{MAX} (dB)
10µs	50.5	72
10µs + 1µs	51.8	90
10µs + 1µs + 0.1µs	52	109

Table 4.2.1. Calculated SNR_H and DR_{MAX} from measured data for the case of single, double and triple exposures with a ratio of 10 and K=1.

The same analysis was repeated for the measurements of the same exposure settings with a photon threshold of K=2 to see the effect of photon threshold on SNR_H and DR in the case of multi-photon single-bit pixels. The signal and noise plots are shown in Figure 4.2.8 and SNR_H and DR are summarised in Table 4.2.2. It is observed that while the DR increases slightly above that of K=1 this comes at the expense of more pronounced ripples or variation in SNR_H at the plateau region when combining the three exposures. The measured SNR_H variation in this example was ~2dB.

The increase in DR is attributed to the fact that the QIS response for K=2 (Fig. 4.2.6(a)) is shifted to the right with respect to the response for K=1 (Fig. 4.2.5(a)) moving the 99% saturation point further while the lower end of the response is still dominated by the noise floor. Moreover, the K=2 response exhibits a steeper slope compared to that of K=1 which reflects on the transition between the three exposure settings and hence higher variation in SNR_H.

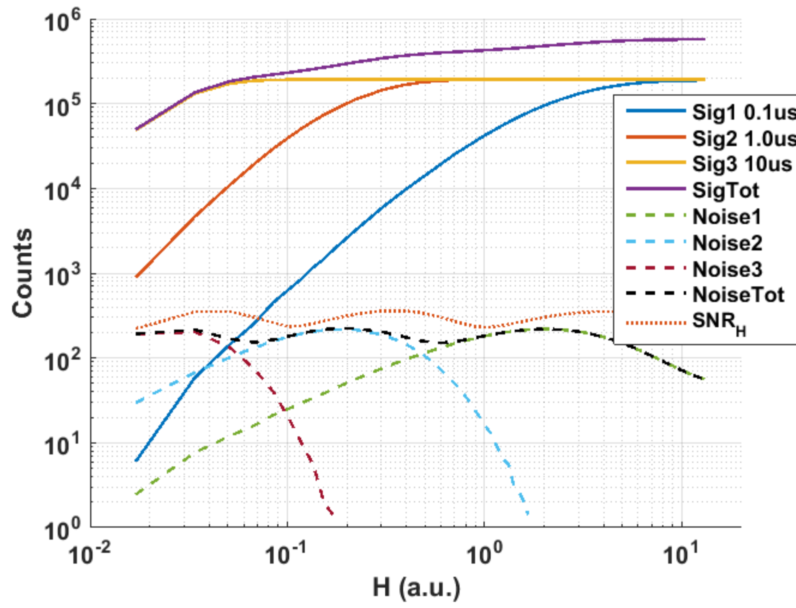


Figure 4.2.8. Measured signal, noise and SNR_H responses for 3 exposure settings with exposure ratio of 10 and photon threshold K=2.

Exposure	SNRH (dB)	DRMAX (dB)
10µs	50.7	75
10µs + 1µs	50.9	92.7
10µs + 1µs + 0.1µs	51.1	115.8

Table 4.2.2. Calculated SNR_H and DR_{MAX} from measured data for the case of single, double and triple exposures with a ratio of 10 and K=2.

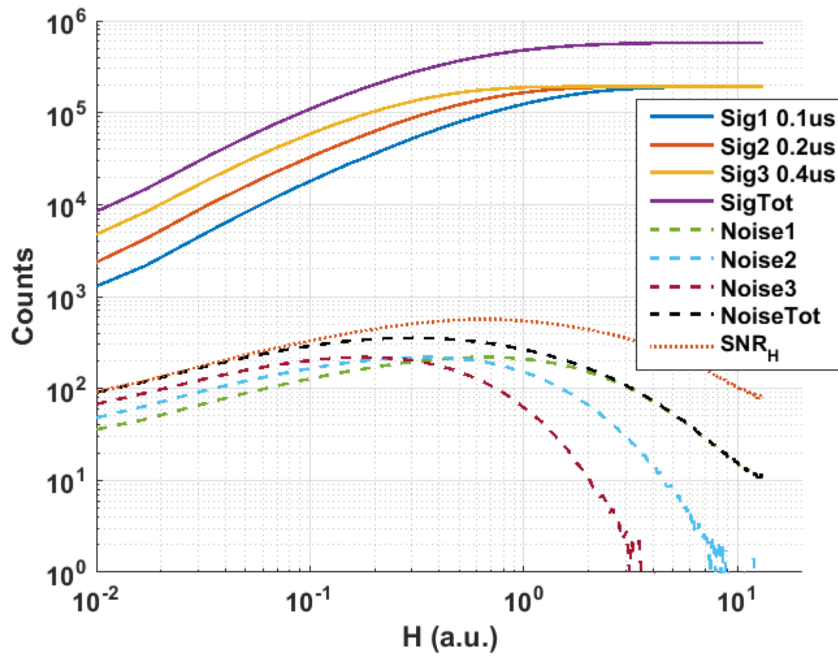
Another factor that has been investigated is the effect of the exposure ratio on SNR_H and DR. For that, the same measurements as above were repeated for K=1 and exposure ratios of 2 (0.1µs, 0.2µs and 0.4µs), 4 (0.1µs, 0.4µs and 1.6µs), 6 (0.1µs, 0.6µs and 3.6µs) and 8 (0.1µs, 0.8µs and 6.4µs). The 0.1µs exposure setting is the common factor across all experiments. The measured SNR_H and DR for all cases are summarised in Table 4.2.3 and Figure 4.2.9 shows the SNR_H curves for exposure ratios of 2 and 8.

It is observed that while SNR_H slightly decreases as the exposure ratio increases, DR is unaffected. This suggests that the DR extension is dominated by the shortest exposure setting which in this example was the common 0.1µs. Of course this holds true due to the fact that the minimum observable signal is dominated by the noise floor as a very large number of ensembles has been used as explained previously. For a smaller number of ensembles the minimum detectable signal will be determined by the longest exposure setting and hence influence the achievable DR. In a rolling shutter sensor the shortest exposure would be dominated by line time and in a global shutter sensor it is down to signal drivers and acceptable temporal aperture ratio.

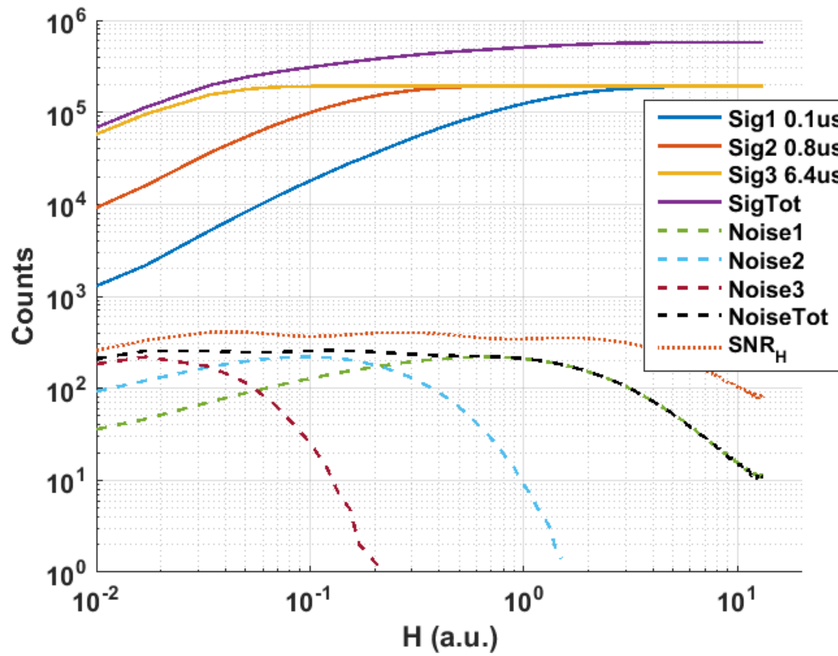
The SNR_H peak is higher for smaller exposure ratios because as can be seen from Equations 4 and 5, SNR_H is dependent on the rate of change in the total signal which is higher for short exposure ratios as the individual responses are close to each other and add up together more rapidly (i.e. dM_{Tot} / dH is higher for shorter exposure ratios). On the other hand, for longer exposure ratios the individual responses are spaced apart resulting in a slower rate of change in the total signal as they are summed together.

Ratio	SNR_H (dB)	DR_{MAX} (dB)
2	55	108
4	53.5	108
6	52.7	108
8	52.2	108
10	52	109

Table 4.2.3. Measured SNR_H and DR_{MAX} for a three exposures scenario and K=1 with different exposure ratios of 2 (0.1µs, 0.2µs and 0.4µs), 4 (0.1µs, 0.4µs and 1.6µs), 6 (0.1µs, 0.6µs and 3.6µs) and 8 (0.1µs, 0.8µs and 6.4µs).



(a) Exposure ratio of 2, K=1

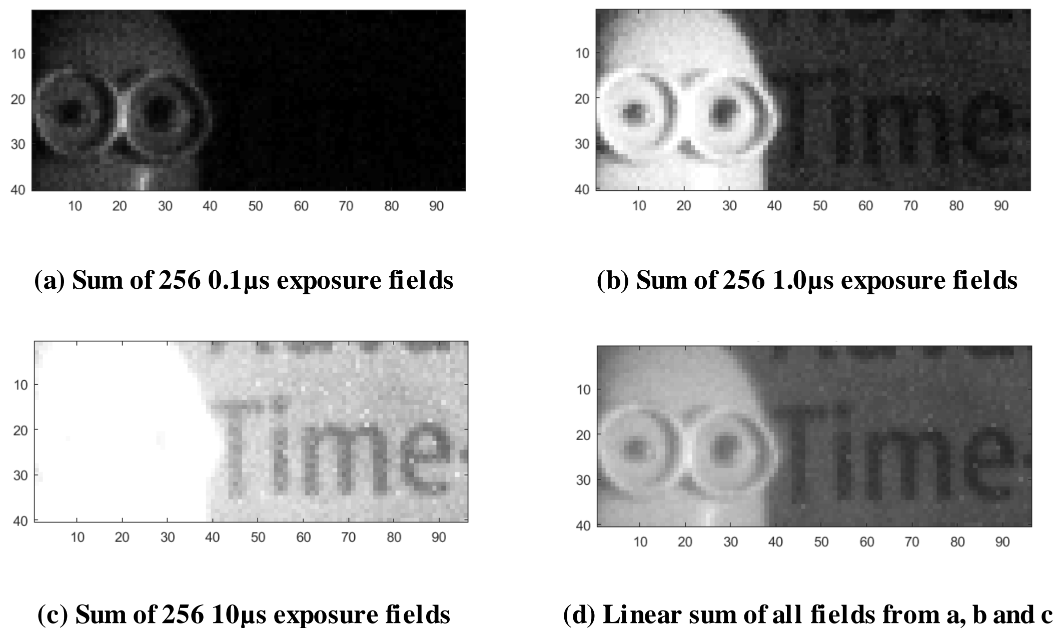


(b) Exposure ratio of 8, K=1

Figure 4.2.9. Measured signal, noise and SNR_H response for 3 exposure settings and K=1. (a) Exposure ratio of 2. (b) Exposure ratio of 8.

The presented results show how the dynamic range of a single frame triple-exposure sensor can be increased which is also an improvement over previous work by the group [9] which required two frames to capture the three sub-exposures for the dynamic range extension. While other QIS sensors can attain similar DR performance, the partial in-pixel field summation providing 3.75× data compression and the ability to capture multiple exposure settings simultaneously significantly reduces readout requirements and offers better immunity against motion artefacts as compared to other works.

The 96 × 40 sensor is used to capture a high dynamic range scene as a demonstration of HDR QIS in operation in Figure 4.2.10. To demonstrate this proof of principle further, Figure 4.2.11 shows images captured by the 320 × 240 SPC imager from [65] which has higher resolution, wider field of view and lower DCR. Both sensors were operated with a photon threshold of K=1 and different exposures were acquired sequentially as only static scenes were imaged.



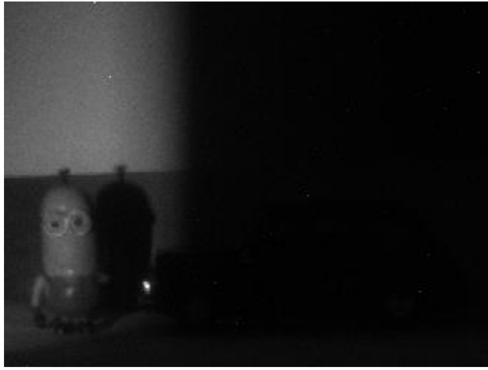
(a) Sum of 256 0.1µs exposure fields

(b) Sum of 256 1.0µs exposure fields

(c) Sum of 256 10µs exposure fields

(d) Linear sum of all fields from a, b and c

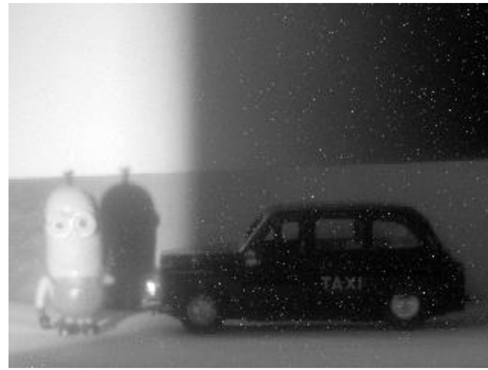
Figure 4.2.10. Images captured by 96 × 40 FSI sensor [2]. (a) Sum of 256 fields at 0.1µs exposure, a Minion figure is visible. (b) Sum of 256 fields at 1.0µs exposure, the Minion is visible but slightly overexposed while faint letters appear in the background. (c) Sum of 256 fields at 10µs exposure, the Minion is totally overexposed but the letters appear clearer. (d) Linear sum of all the 768 fields from a, b and c to form an HDR image preserving all details.



(a) Sum of 1000 0.1µs exposure fields

(b) Sum of 1000 1.0µs exposure fields

(c) Sum of 1000 10µs exposure fields



(d) Linear sum of all fields from a, b and c

Figure 4.2.11. Images captured by 320 × 240 SPC sensor from [65]. (a) Sum of 1000 fields at 0.1µs exposure, a Minion appears in the lit portion of the scene. (b) Sum of 1000 fields at 1.0µs exposure, the Minion is slightly overexposed but a car figure appears in the dark region of the scene. (c) Sum of 1000 fields at 10µs exposure, Minion is completely overexposed but more detail of the car is apparent. Notice that high DCR pixels appear as white dots. (d) Linear sum of all the 3000 fields from a, b and c to form an HDR image preserving all details.

The example given in Figure 4.2.10 allows for a brief benchmarking of HDR QIS performance. The presented analysis in this work shows that for the given 96 × 40 sensor it is possible to achieve a maximum dynamic range (DR_{MAX}) of 108dB. Yet the DR of the example in Figure 4.2.10 is limited by the number of ensembles (M=256) rather than the noise floor, so the effective DR ($DR_{Effective}$) is limited by the minimum observable signal (bit density D = 1 / 256 for each exposure). To calculate $DR_{Effective}$ the equivalent H value for this minimum signal can be calculated from Equation 2, and using that as the denominator in Equation 3 results in an effective DR of 99.6dB for a three exposure (0.1µs, 1µs and 10µs) scenario showing the effect of M on achievable DR.

4.3. HDR QIS and Miniaturisation Discussion

Although the MINIC40 pixel allows for extending the dynamic range of QISs via simultaneous triple-exposure acquisitions while providing in-pixel data rate compression, the applicability of this concept to miniature sensors needs to be examined.

Table 4.3.1 provides a comparison between different HDR QIS and in-pixel compression scenarios to estimate the numbers of data readout channels required. A sensor resolution of 128 × 128 and an effective frame rate of 30fps at an oversampling ratio of 255 (8-bit) are assumed. The reference case represents a conventional QIS sensor with a single exposure (no HDR) and a single-bit pixel.

	Reference	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6
No. of Pixels (128 × 128)	16384	16384	16384	16384	16384	16384	16384
No. of HDR Exposures	1	3	3	3	3	3	3
Oversampling In-Pixel	No	No	Yes	Yes	No	Yes	Yes
Exposures In-Pixel	1	1	1	1	3	3	3
Counter Depth Per Exposure	1	1	4	8	1	4	8
Summed Fields In-Pixel	1	1	15	255	1	15	255
Pixel Data Output Width (bits)	1	1	4	8	3	12	24
Effective Frame Rate (fps)	30	30	30	30	30	30	30
Oversampling Ratio	255	255	255	255	255	255	255
Sensor Field Rate (fips)	7650	22950	1530	90	7650	510	30
Sensor Data Rate (Mbps)	125.3	376.0	100.3	11.8	376.0	100.3	11.8
Pads Required at 25MHz Readout	6	16	5	1	16	5	1
Pads Required at 100MHz Readout	2	4	2	1	4	2	1
In-Pixel Compression Ratio	1.00	1.00	3.75	31.88	1.00	3.75	31.88
Multi-Exposures In-Pixel	No	No	No	No	Yes	Yes	Yes

Table 4.3.1. Comparison of data rates of different high dynamic range QIS and in-pixel compression scenarios.

It can be seen that the only scenarios where a single output channel would suffice (red) are cases 3 and 6 where an 8-bit in-pixel counter is required for a data compression ratio of 31.88. Regardless if the triple HDR exposures are performed simultaneously or sequentially the sensor data rate is constant. The impact of that though reflects directly on the pixel pitch.

Implementing an 8-bit counter in-pixel for case 3 contradicts the motive behind simplistic single-bit pixels to gain high fill factor at a small pitch. Case 6 exaggerates that further by requiring three 8-bit counters. If such a pixel is to be realised the dynamic range obtained by utilising all counter bits in linear mode would exceed the HDR extension benefit of a triple-exposure QIS.

Moreover, even the MINIC40 pixel architecture does not meet the single output channel specification. At its best case assuming a 100MHz clock, two IO pads are required (grey). Considering the reference case where no HDR QIS is performed, a similar conclusion can be drawn. While QISs can offer extended dynamic range, high speed imaging and time-resolved capability with SPADs, it is unsuitable for miniature sensor architectures due to the high data rates imposed by oversampling.

4.4. Summary and Conclusions

A technique of extending the dynamic range of quanta image sensors through triple-exposure samples is demonstrated using the MINIC40 sensor. The configurable pixel counter allows for simultaneous HDR bit-planes acquisition for better motion artefact immunity while the latching circuit front end allows for partial bit-plane summation in-pixel resulting in 3.75× data rate compression.

For the MINIC40 sensor parameters, a dynamic range in excess of 100dB is reported. In oversampled QIS operation the dynamic range upper limit is determined by the shortest possible exposure setting which pushes the DlogH curve response towards higher signal levels. The lower end of the range is dependent on the number of oversampled ensembles.

If the minimum observable signal is higher than the sensor's noise floor, determined by the SPAD DCR for the shot noise limited MINIC40 pixel, then that minimum signal defines the attainable DR ($DR_{Effective}$). Else if a large number of ensembles is possible, then the sensor's noise floor determines the DR (DR_{MAX}). This will have consequences on the oversampling ratio required.

Despite the benefits of QISs, their associated data rates make them incompatible with miniaturisation goals. A single data output channel cannot deliver the necessary bandwidth of a single-bit pixel and to relax data rates multi-bit pixels are needed which would result in a large pixel design.

A solution to this trade-off, whether for QIS or linear mode counting pixels, is to push the oversampling off the focal plane with the aid of on-chip processing. This also has the benefit of creating a mediating medium between the array acquisition and the sensor output channel. Such concept is at the core of the proposed miniature sensor demonstrated in Chapter 6 and has been hinted at recently by the MIT group in [216].

5. A 3D-Stacked SPAD Image Sensor

Pursued since 2002 [92] for SPAD pixels, the idea of vertically integrating the photo-detector and the corresponding electronics is crucial for improving the performance of such sensors. Apart from optimising each of these components in a dedicated fabrication process, the decoupling of the two allows for higher levels of integrated functionality and increased sensitivity while enabling miniaturisation and innovative system architectures.

This chapter presents the first SPAD image sensor realised in an industrial wafer-scale hybrid-bonding technology [1]. Based on a revised version of the MINIC40 configurable 12-bit counter design (Chapter 3), the time-gated pixel achieves a 45% fill factor at a 7.83μm pitch with a 1-to-1 hybrid bond connection between the imaging specific 65nm top tier SPAD and the 40nm bottom tier processing circuitry. The author was privileged to have access to this state of the art technology through the research collaboration agreement between the University of Edinburgh and STMicroelectronics.

An overview of the system and pixel design and layout of the MINI3D sensor is given in the first section with characterisation results presented in section two. Comparisons to previous FSI SPAD sensors and other 3D-stacked works from the literature are made in section three with final remarks and conclusions stated at the end.

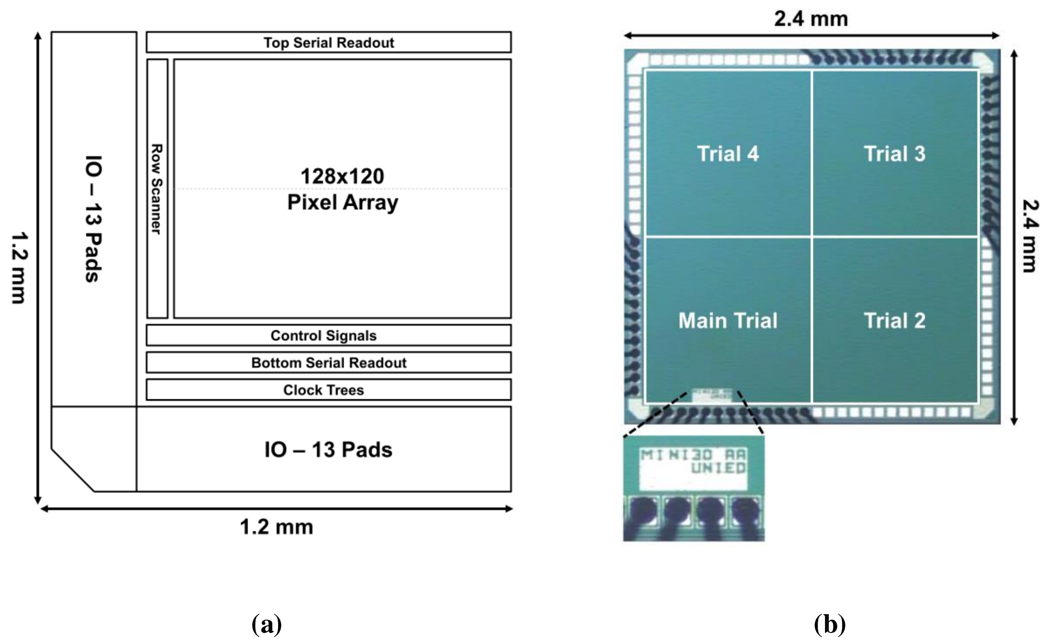
5.1. Chip Overview

The MINI3D sensor is similar to MINIC40 in many respects for the exception of a few minor details that will be highlighted in the sections below. Both employ the exact IO configurations, include four independent array trials and generate 12 bits of data per pixel. Therefore, the same bring-up platform and firmware can be used for characterisation while accounting for the difference in resolution. Unlike the 96×40 MINIC40 sensor, MINI3D has 128×120 pixels.

5.1.1. System Architecture

Figure 5.1.1 shows a block diagram of the basic system architecture alongside a micrograph of the backside of the top tier chip. All dynamic control (exposure and reset) and gating signals are driven from the bottom of the array through balanced binary clock trees while static controls are driven through standard column buffers.

Serial readout is performed through two output pads, one per half array, with rolling blades starting from the centre of the array outwards. This limits the maximum frame rate to 500fps with a 50MHz readout clock and 155μs global shutter period. The overall die area is 2.4mm × 2.4mm with each independent trial measuring 1.2mm × 1.2mm.



(a) (b)

Figure 5.1.1. MINI3D sensor. (a) Block diagram. (b) Chip micrograph showing backside of top tier.

5.1.2. Pixel Circuitry

The 12-bit counter pixel was modified from that of MINIC40 such that in time-gated mode it splits into two 6-bit counter banks. This decision was made to improve the counting depth of each time gate to 63 photons since two time gates are enough to perform time-resolved measurements be it indirect time of flight depth imaging [64] or time-gated fluorescence lifetime imaging [289]. Figure 5.1.2 shows the pixel circuit diagram.

The same minimalistic thick oxide front end (MQ, M0 and M1) was adopted with the ability to bypass the SPAD pulses by the Enable control and to inject electrical test pulses through the Test signal. In normal operation, Test is held low while Enable acts as a global shutter. The global Mode signal chains the two 6-bit ripple counters in imaging mode for 12-bit photon counting capacity or splits them for time-gated mode.

Conventional column parallel readout is enabled through row select signals making the counter bits available on the column bus. After the whole array has been readout the global RST_Counter signal resets all pixel bits to zero.

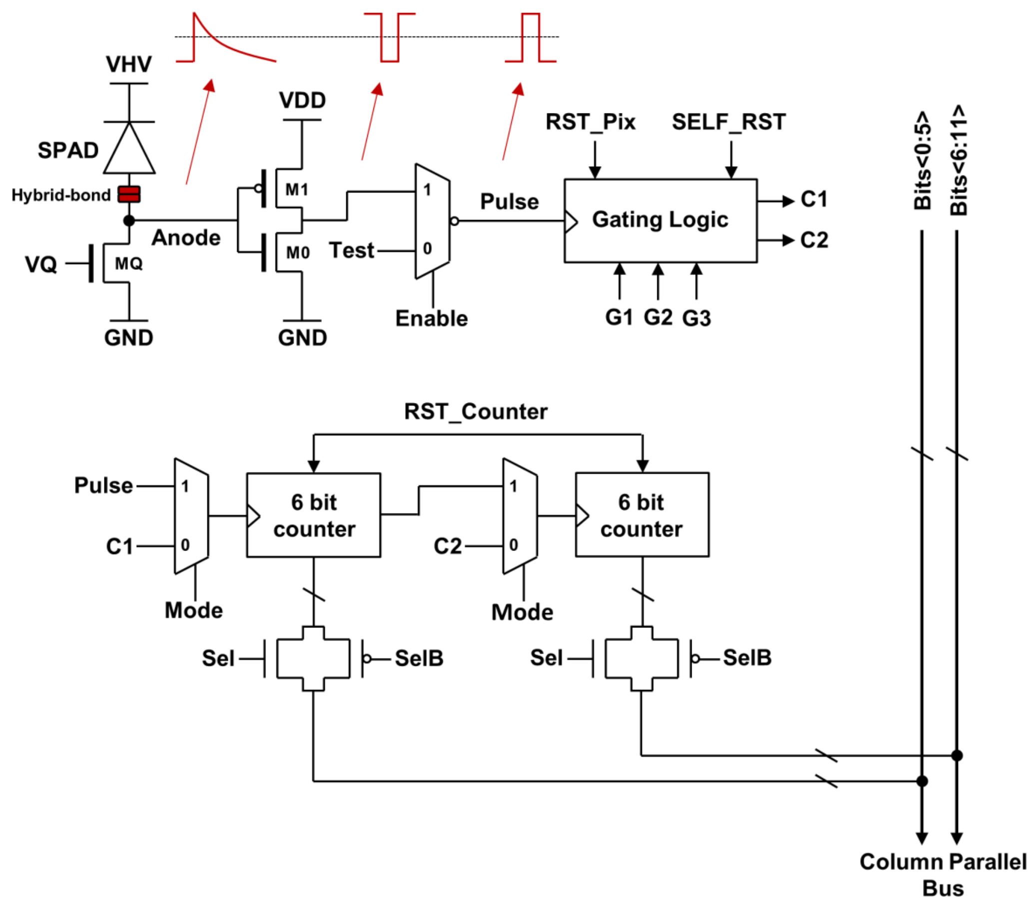


Figure 5.1.2. MINI3D pixel circuit block diagram with thick oxide transistors MQ, M0 and M1 forming the quench and front end inverter followed by thin oxide 40nm CMOS logic. The front end waveforms and their polarities are indicated in red. The pixel has a single 1.1V supply (VDD) common to all components.

Time gating operation is similar to that of MINIC40 relying on rising edge to rising edge technique. Gate control signals G1, G2 and G3 define the observation windows of the two bins. RST_Pix signal is used to reset the latching front end of the gating logic to allow photon detections within subsequent time gates.

An additional self-reset feature is added to the gating front end to automatically reset the sampling latches once a photon is detected to allow multiple photon detections within the same time gate, an improvement over the MINIC40 pixel (Chapter 3.2.7). Thus, the two 6-bit counters can operate in linear counting mode if needed. This feature is activated via the global SELF_RST control.

5.1.3. Pixel Layout

The main difference between MINI3D and MINIC40, or any other FSI sensor, is the fact that the SPAD and circuitry are separated and vertically integrated on top of each other. Figure 5.1.3 shows a cross section of the pixel layout demonstrating the 3D-stacking concept.

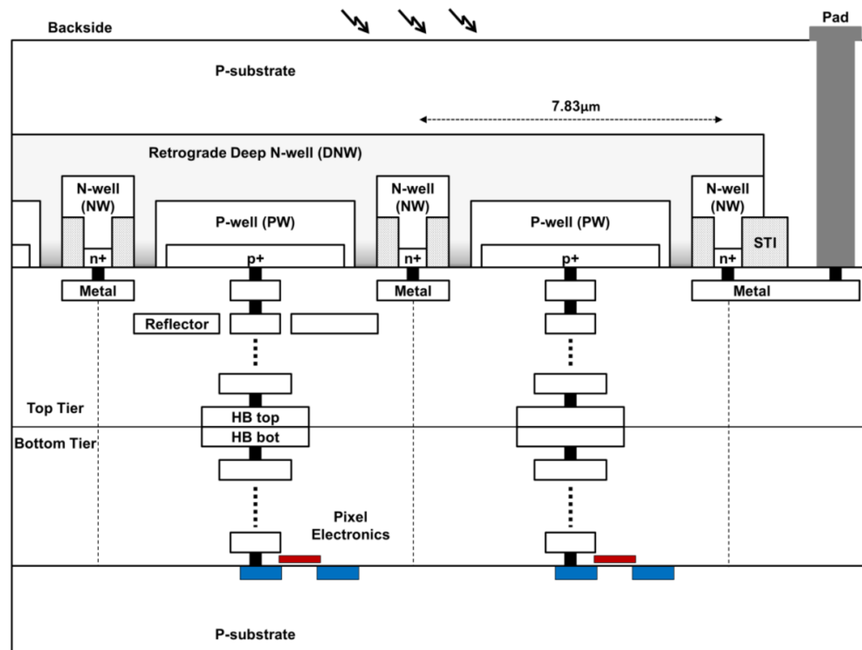


Figure 5.1.3. Cross section of MINI3D backside illuminated 3D-stacked pixel layout showing both top and bottom tiers.

The top tier of MINI3D is implemented in an imaging 65nm process that is not optimised for SPAD performance hosting an array of 128 × 120 global shared well detectors. No active circuitry was included on the top tier with all processing electronics integrated in the 40nm bottom tier process.

The anode of each SPAD is connected vertically through a via stack from the lowest metal layer on the top tier to the lowest metal layer on the bottom one with a 1-to-1 hybrid bond (HB) site in between [205]. It is clear how this configuration overcomes the limits of MINIC40 where no long anode routes with complex layout are needed and the single pixel unit easily scales in both x and y directions.

It is worth noting that due to 3D-stacking, the top tier is flipped upside down with light entering from the thinned backside of the die, hence this is a backside illuminated sensor. The thickness of the backside stack is undisclosed to the author by STMicroelectronics. The hybrid-bond sites have a size of 4µm × 4µm and connectivity to the outer world is achieved through backside etched aluminium pads as depicted for the SPAD high voltage supply in the Figure 5.1.3 (not to scale). Pad connections to the bottom tier would extend all the way through the vertical metal stack of both tiers including multiple HB sites (not shown).

The optimised pixel layout achieves a pitch of 7.83μm (for both SPAD and circuitry) making it the smallest reported for a SPAD image sensor. Figure 5.1.4(a) shows the high density pixel layout in 40nm with only metals 1 (MT1) to 3 (MT3) switched on for clarity. The thick oxide transistors (M0, M1 and MQ) are clearly identifiable as well as the small decoupling capacitor MD. Figure 5.1.4(b) on the other hand shows the routing frame in metals 4 (MT4) and 5 (MT5) which in contrast to the MINIC40 pixel is not restricted over particular regions allowing for a more flexible layout scheme. Thick metal 6 and metal 7 power straps (not shown) fully cover the pixel area which also helps in blocking any potential circuitry photo-emission from reaching the top tier SPADs [290].

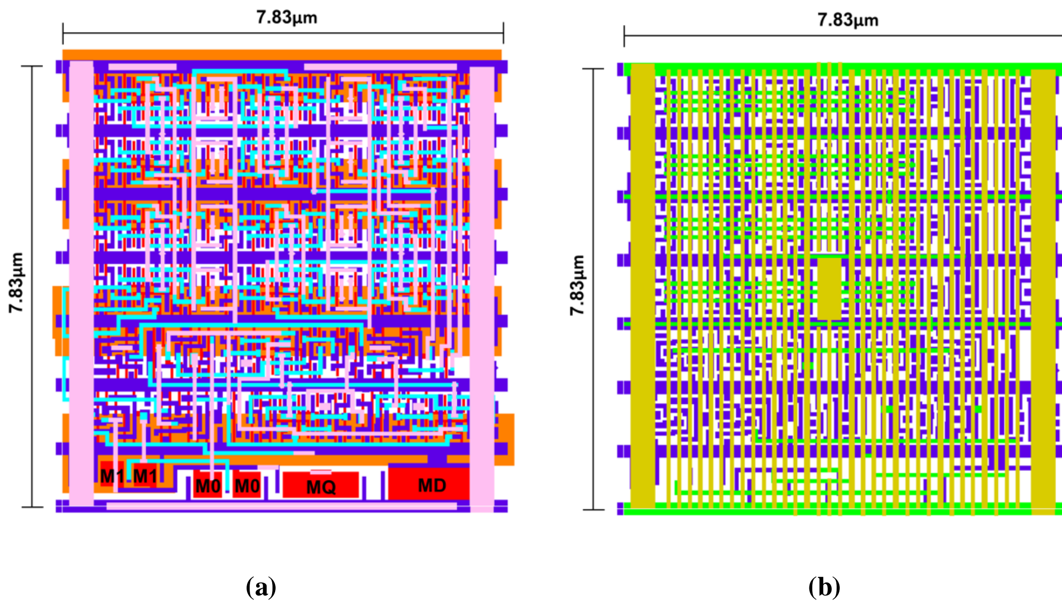


Figure 5.1.4. MINI3D pixel layout. (a) Layers up to MT3 only. (b) MT4 and MT5 routing frame showing flexible use of these metals compared to MINIC40. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3, green is MT4 and yellow is MT5. Higher metal layers are switched off for clarity.

5.1.4. SPAD Trials

The four implemented variants of the chip were used to trial different SPAD structures to test and compare their characteristics in a 3D-stacked BSI technology. Table 5.1.1 summarises the trial types with an independent imaging array of 128 × 120 pixels of each integrated on the top tier.

Trial	SPAD Junction	Substrate Isolated	DTI Isolated	Quench Transistor	VHV Polarity	Working
1 (Main)	PW / DNW	Yes	No	NMOS	Positive	Yes
2	n+ / Psub	No	No	PMOS	Negative	No
3	DNW / Psub	No	No	PMOS	Negative	No
4	DNW / Psub	No	Yes	PMOS	Negative	No

Table 5.1.1. Summary of trialled SPAD structures on MINI3D.

The main trial is the standard PW / DNW SPAD used in this work which is fully functional and has been used for all characterisation results in the rest of this chapter. The shared well device has a drawn fill factor of 45%. While this is a BSI implementation, the actual fill factor is unlikely to be higher than the drawn estimate since this is a substrate isolated device.

Unfortunately the other three trials failed with no signs of functionality. These trials were designed experimentally without any knowledge of the junction implants or substrate doping as this is information is STMicroelectronics confidential. Without access to such information or technology computer aided design (TCAD) simulations it was not possible to debug the reason of failure. A common attribute to all these structures is the substrate being one side of the multiplication junction.

The n+ / Psub trial was intended for having a non-substrate isolated junction to see its effect on photon detection probability (PDP). Moreover, in a BSI implementation, the multiplication junction would be deeper in silicon relative to the reference PW / DNW SPAD which would have made for an interesting comparison for near infra-red (NIR) PDP.

On the other hand, the DNW / Psub trial [116] was intended for the opposite purpose, as the deep NW is closer to the backside surface and would have been interesting to see if it enhances the PDP at the blue region of the spectrum while also being a non-substrate isolated device.

For both these devices, the front end quench circuit was modified to a PMOS quench to take into account the change in expected SPAD pulse polarity since a negative VHV bias can be applied to the substrate of the top tier which is independent of the bottom tier die and has no active circuitry integrated in it. Figure 5.1.5 shows the modified pixel front end.

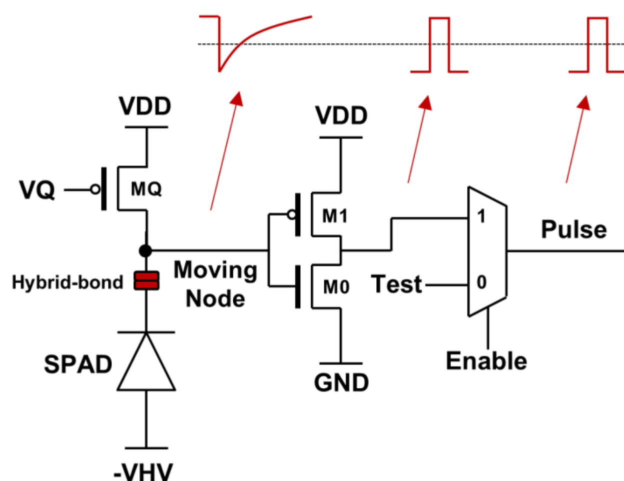


Figure 5.1.5. MINI3D PMOS quench front end with signal polarities shown in red.

The final trial was the same DNW / Psub array but with full depth deep trench isolation (DTI) walls surrounding the SPADs. DTI is commonly used in CISs for optical and electrical isolation of pixels

[131] and a comparison between the two DNW / Psub SPADs with and without DTI would have been a great result. BSI SPAD pixels with DTI walls have already been attempted by [220].

Another interesting feature of full depth DTI is that it allows for different bias conditions for non-substrate isolated and non-well sharing structures since the SPAD structure is encapsulated within the DTI walls. This can be achieved by having multiple VHV bias lines for varying pixel sensitivities or even diode modes of operation. Figure 5.1.6 demonstrates the concept.

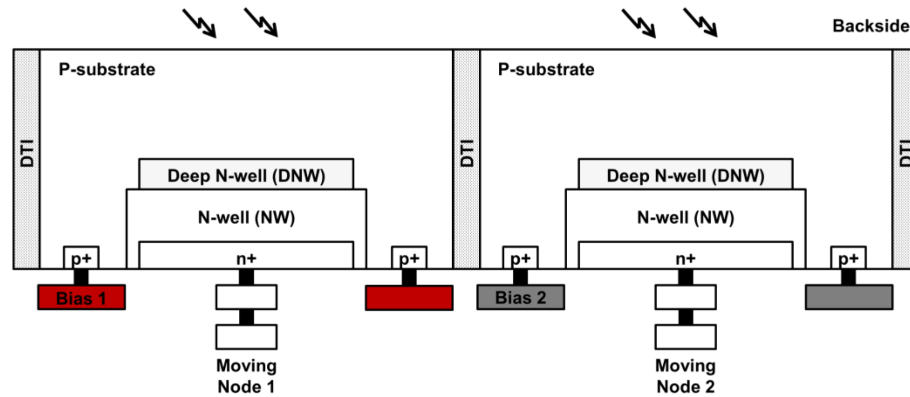


Figure 5.1.6. Dual SPAD bias concept enabled by full depth DTI.

Finally and for all the arrays, a floating plate of metal 2 was drawn on top of the SPADs such that it acts as a reflector for light coming from the backside (see Figure 5.1.3). 124 columns of each array implemented the MT2 reflector while the last 4 columns of the right side were excluded for comparison. Figure 5.1.7 shows the layout of 3 × 2 pixels of the main SPAD trial.

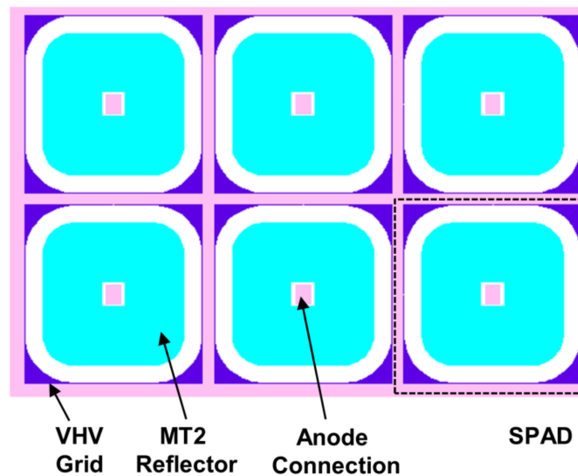


Figure 5.1.7. Layout of 3 × 2 pixels of MINI3D main SPAD trial.

5.2. Characterisation Results

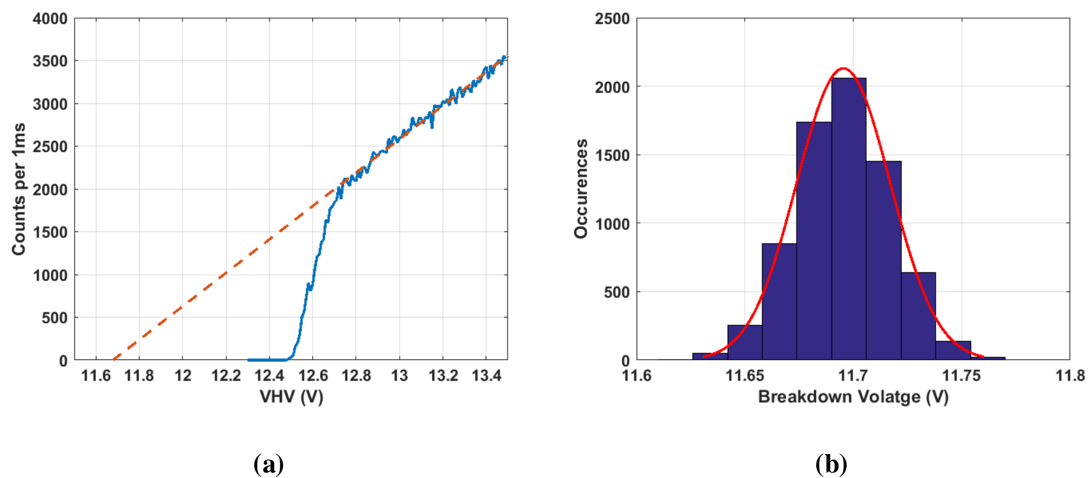
Characterisation results of the SPAD (PW / DNW) and sensor measurements are presented in this section with focus on optical performance such as PRNU and PDP. Images acquired with MINI3D in photon counting and time-gated modes are also included.

5.2.1. Breakdown Voltage

To measure the breakdown voltage of the SPAD the light count rate technique was used [130]. The Sensor was uniformly illuminated with a white light source and the excess bias voltage was swept in steps of 10mV while recording the counts of all pixels within a 1ms interval.

Initially no counts are registered as the SPAD excess bias is not enough to generate pulses exceeding the front end inverter threshold. As the excess bias surpasses the threshold events start registering in the counter with linear dependence on VHV. A line is fitted to the counts in that region and extrapolated to extract the x-axis zero crossing point. This point corresponds to the SPAD breakdown voltage.

The fitting and extrapolation was done for 100 × 100 pixels in the centre of the array to avoid edge effects. Figure 5.2.1(a) shows the sweep for a randomly selected pixel with the fitted line overlaid while figure 5.2.1(b) shows the distribution of the extracted breakdown voltage across the array with a mean value of 11.7V and a standard deviation of 30mV.



(a) **(b)**

Figure 5.2.1. SPAD breakdown voltage characterisation. a) Counts versus VHV sweep for a random pixel with linear fit in red. b) Breakdown voltage distribution across array with mean of 11.7V and 30mV standard deviation with Gaussian fit in red.

5.2.2. Dark Count Rate

Dark count rate was measured at room temperature for different excess bias voltages. Figure 5.2.2 shows the median DCR values which fit an exponential trend suggesting that tunnelling is the dominant DCR mechanism. The sensor exhibits a median DCR less than 200cps and less than 11kcps at 1V and 3V excess bias settings respectively.

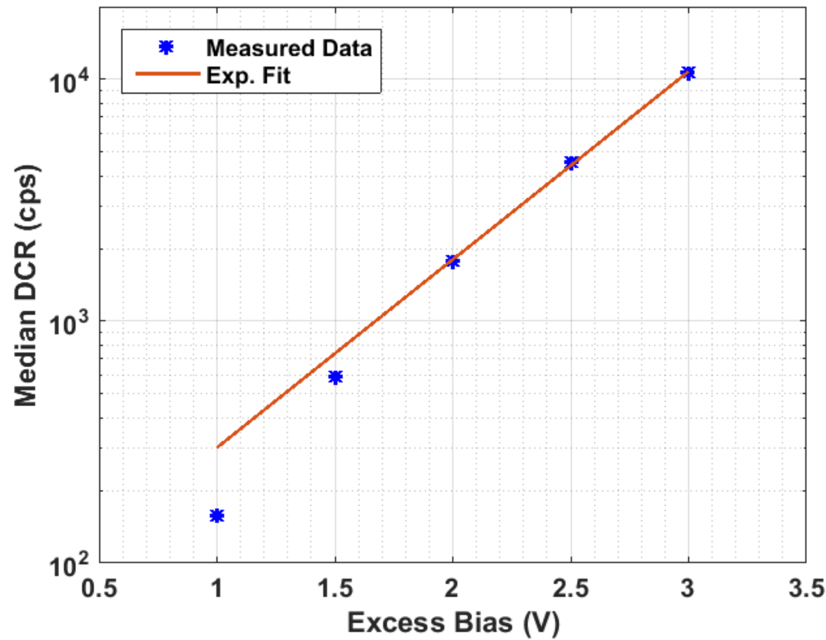


Figure 5.2.2. Median DCR versus excess bias at room temperature for MINI3D sensor.

Figure 5.2.3 shows the cumulative DCR distribution across the array at different excess bias voltages. Under all conditions, roughly 80% of the pixels reveal a relatively tight lower DCR distribution as evident by the flatness of the cumulative curve. The other 20% of pixels exhibit a different defect showing as a secondary distribution (tail) of higher DCR.

Approximately 86% of pixels report a DCR below 1kcps at 1V excess bias and 89% report a DCR below 10kcps at 2V excess bias. Similar to the MINIC40 SPAD, the yield may not be suitable for some scientific imaging applications and needs process improvement.

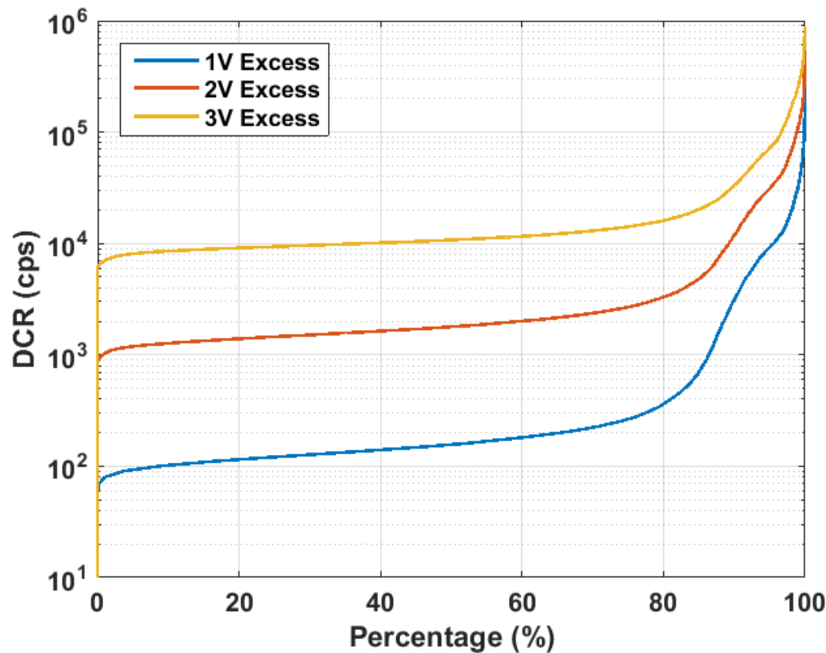


Figure 5.2.3. Cumulative DCR distribution versus excess bias at room temperature for MINI3D sensor.

5.2.3. SPAD Jitter

The output pulse from an edge pixel in the leftmost column of the array is accessible through a test IO pad and was used to characterise the SPAD temporal response or jitter. A LeCroy WaveRunner 4GHz oscilloscope was used for recording the SPAD pulse time with respect to the electrical synchronisation signal of a Hamamatsu PLLP10 laser driver.

Two lasers of 443nm and 773nm wavelengths were used with quoted electrical jitter of 53ps and 56ps respectively. Neutral density filters were used to ensure that the SPAD does not fire more than five times per hundred laser repetitions to avoid pile-up conditions and 30k hits were recorded per measurement. Figure 5.2.4 shows the SPAD jitter for different excess bias conditions without correcting for the laser or circuitry contributions. Figure 5.2.5 shows the measured impulse response function (IRF).

The measured jitter values are thought to be higher than the intrinsic jitter of the SPAD device due to the non-optimal electrical signal chain from the edge pixel output to the chip output pad. Moreover, since it is not possible to isolate the test pixel by turning off the rest of the array, noise on VHV as the rest of the array is simultaneously firing would add error to the measurement.

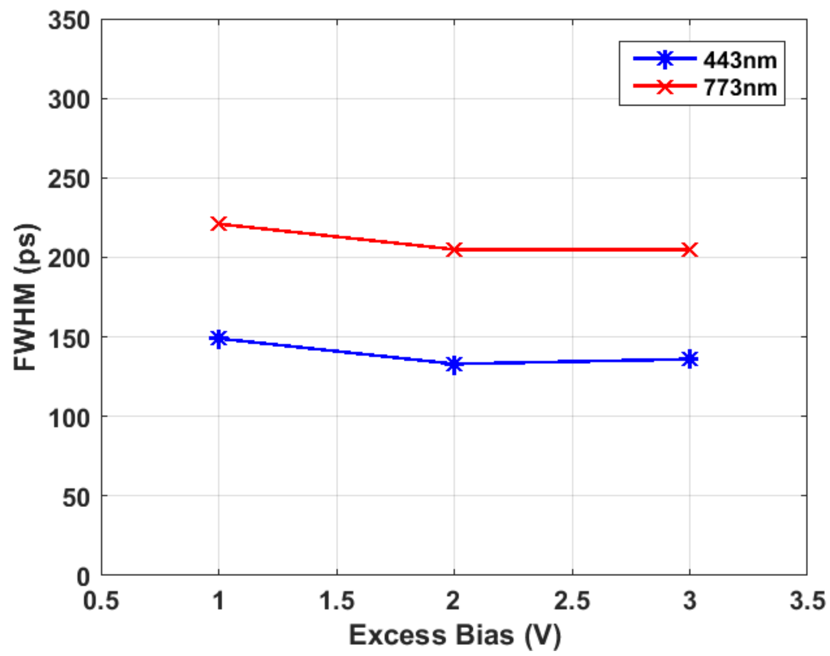
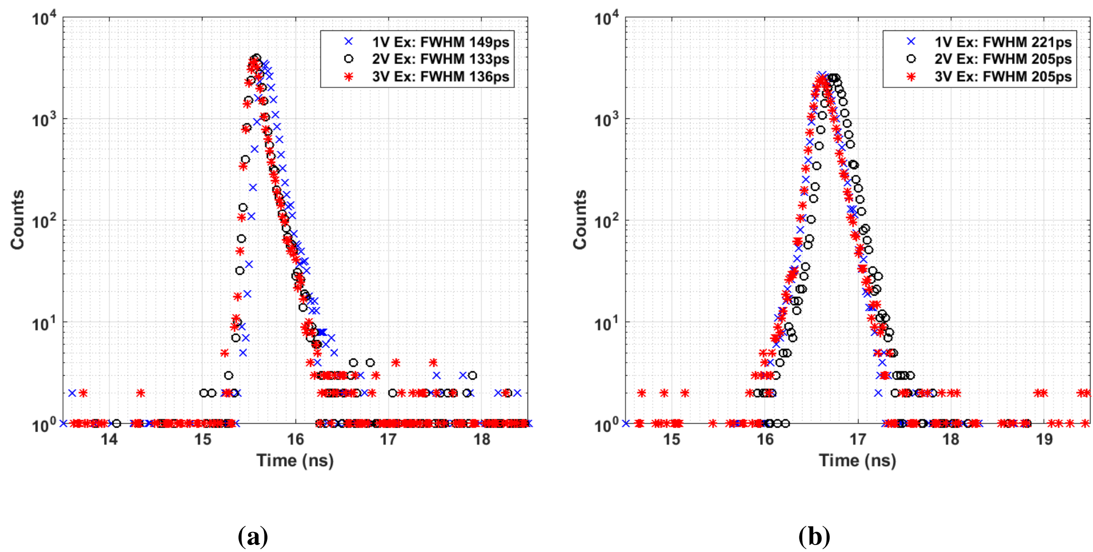


Figure 5.2.4. SPAD jitter versus excess bias for different wavelengths.



(a) **(b)**

Figure 5.2.5. SPAD jitter impulse response functions at different excess bias. a) 443nm. b) 773nm.

5.2.4. Photo-Response Non-Uniformity

To evaluate the uniformity of the imaging array five thousand frames were captured under fixed illumination level filling up approximately 25% of the pixel's photon counting capacity. An excess

bias of 2V was applied. The mean frame is shown in Figure 5.2.6. It can be seen that the four rightmost columns without the metal reflector exhibit lower counts and were excluded from the PRNU calculation. PRNU was calculated as the standard deviation divided by the mean and equated to less than 2% - similar to other CMOS arrays [291] - without excluding the high DCR pixels.

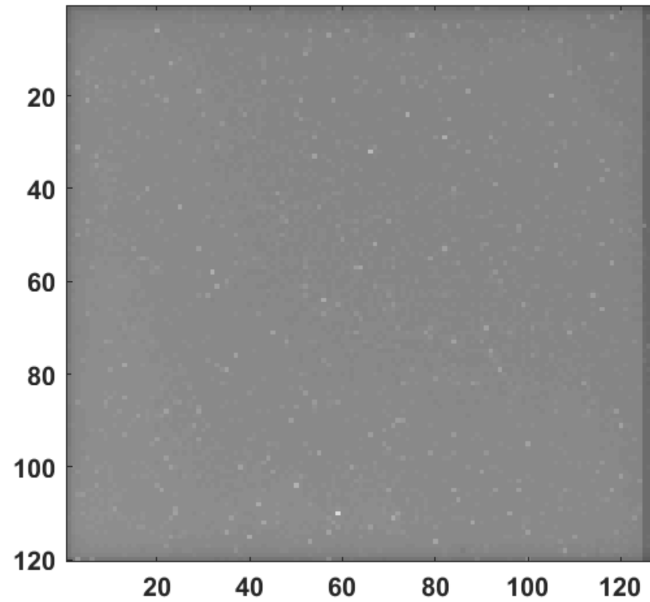


Figure 5.2.6. MINI3D array uniformity. Mean frame of 5000 captures under fixed illumination. The four darker columns to the right are without metal reflectors.

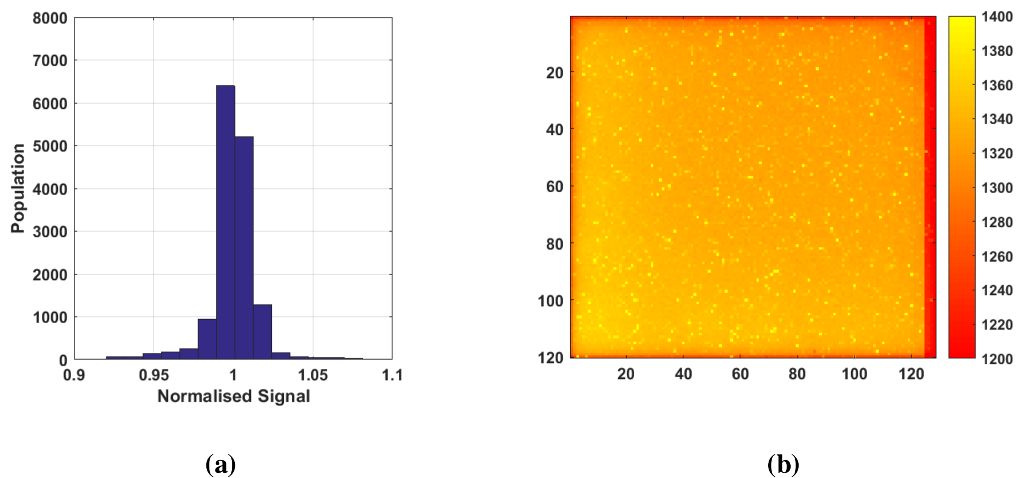


Figure 5.2.7. MINI3D array uniformity. a) Histogram of normalised array response. b) Mean frame with higher contrast colour scale revealing darker edge rows and columns.

The normalised array response is shown in the histogram in Figure 5.2.7(a). The tail on the right is attributed to the unremoved high DCR pixels while an unexpected tail on the left is visible. Upon closer examination of the average array frame, shown in higher contrast in Figure 5.2.7(b), it can be seen that the edge rows and columns (apart from the reflector-free ones) also show lower counts.

Unlike MINIC40, this cannot be attributed to shading effects as this is a BSI array with no metal stack surrounding the SPAD from the direction of impinging light. Referring back to the layout it was found that an error was made by placing a ground substrate contact ring around the SPAD array in close proximity (1µm) to the edge of high voltage NW. It is thought that when the NW is reverse biased with a high potential that a leakage current would flow through the nearby ground connection reducing the sensitivity of the edge pixels (Figure 5.2.8).

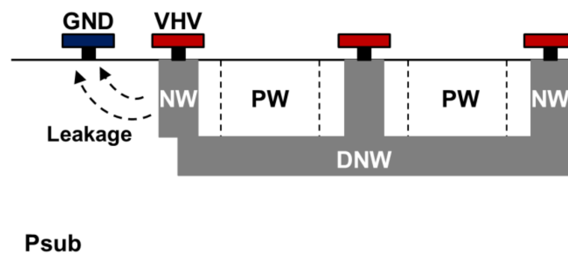


Figure 5.2.8. Leakage effect at edge of MINI3D array due to close proximity of GND substrate contact to the high voltage SPAD shared NW.

5.2.5. Photon Detection Probability

The photon detection probability was measured at room temperature for different excess bias voltages. Figure 5.2.9 shows the mean PDP response of the 124 × 120 pixels with metal reflectors. A peak PDP of 27.5% is reached at 640nm and 3V excess bias.

Compared to the PDP response of the same SPAD structure in an FSI implementation two differences can be seen. First the blue region of the spectrum is heavily attenuated due to the absorption of the short wavelengths at the surface of the backside away from the reach of the deep substrate isolated multiplication junction. Second, this relatively deeper position of the junction with respect to surface of impinging light shifts the peak of the PDP towards the red region of the spectrum due to the deeper penetration of longer wavelengths in silicon.

Figure 5.2.10 compares the PDP of the main array to the PDP of the four columns without the MT2 reflector to evaluate its efficiency. The reflector has a small effect on PDP with a maximum gain of 6% in PDP at 820 nm. It is thought that with optimised material and layer thickness engineering, this gain can be shifted towards the wavelength of interest.

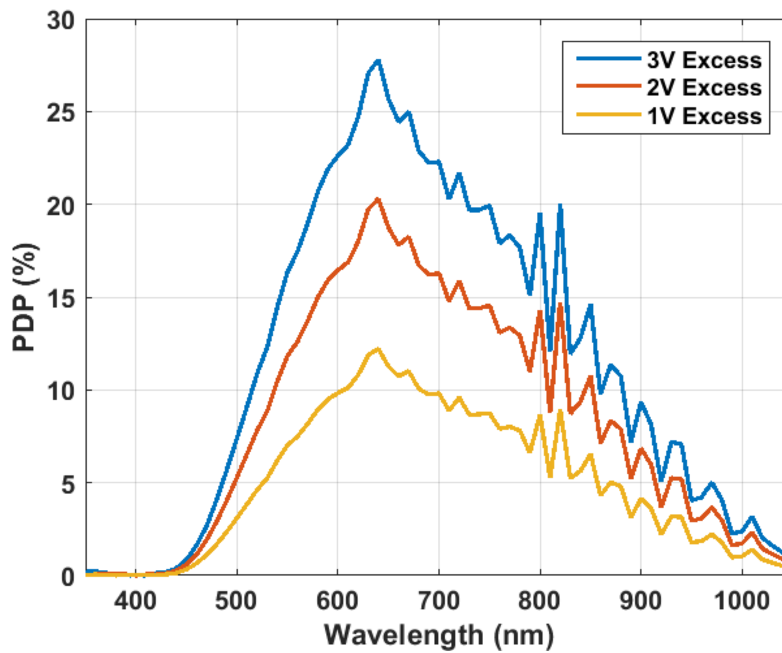


Figure 5.2.9. Photon detection probability of the BSI SPAD with metal reflector at different excess bias voltages.

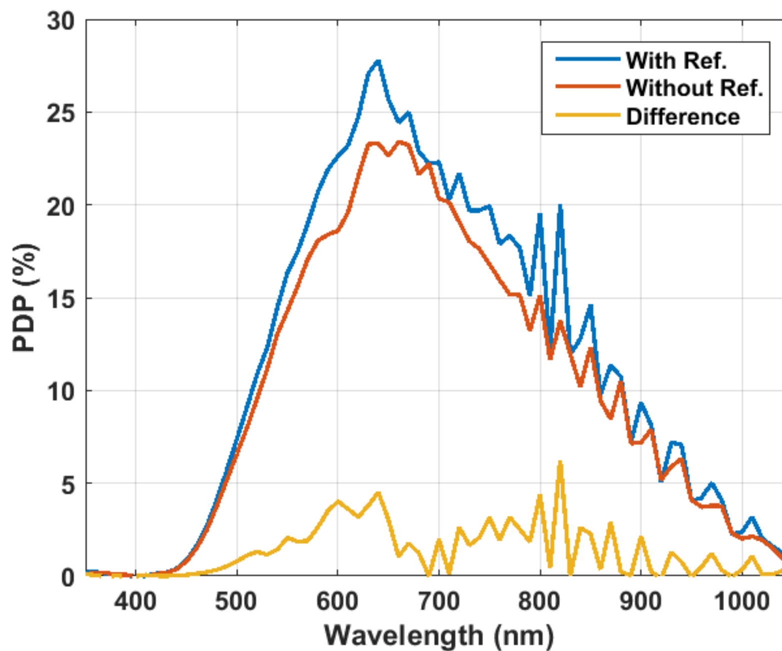


Figure 5.2.10. Photon detection probability of the BSI SPAD with and without metal reflector at 3V excess bias.

5.2.6. Shot Noise Limited Photon Counting

To demonstrate the noiseless photon counting capability of the digital pixel, the sensor's exposure was swept from 20ns to 1ms under constant illumination with five thousand frames captured at each setting. Figure 5.2.11 shows the photon transfer curve (PTC) of the sensor.

For a randomly selected pixel, the mean signal value and standard deviation were calculated across the five thousand frames at each exposure setting (blue) with the response matching the ideal Poisson statistics (red). Similarly, the mean signal value and standard deviation were calculated for all pixels of a randomly selected frame at each exposure setting (black) with the frame response deviating from the ideal behaviour at higher exposures due to the contribution of PRNU.

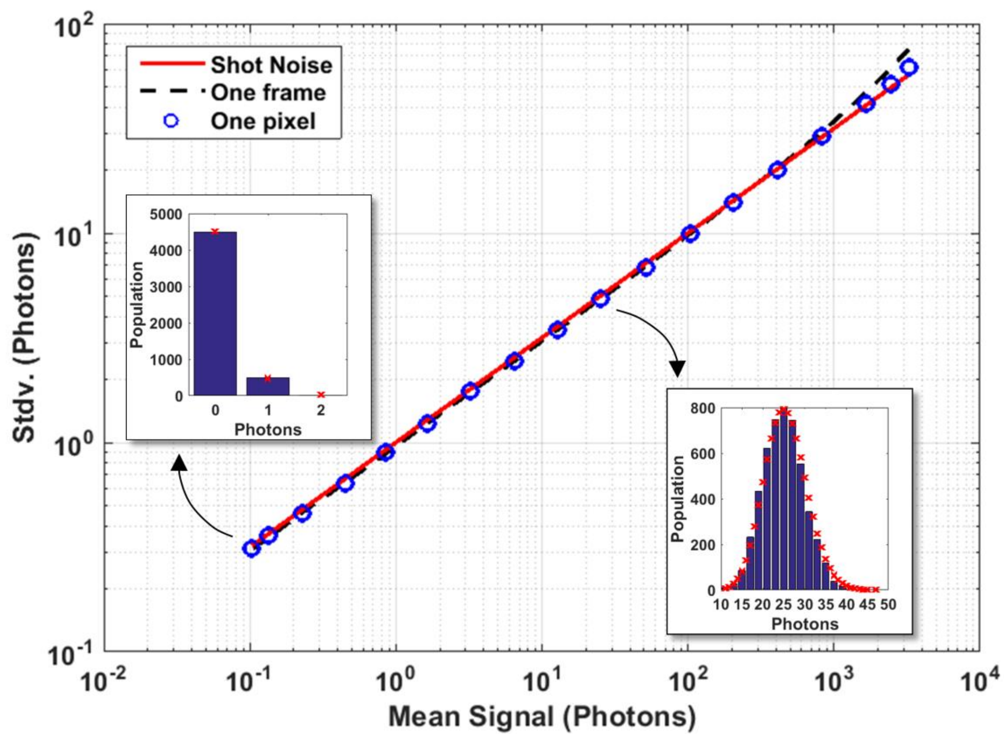
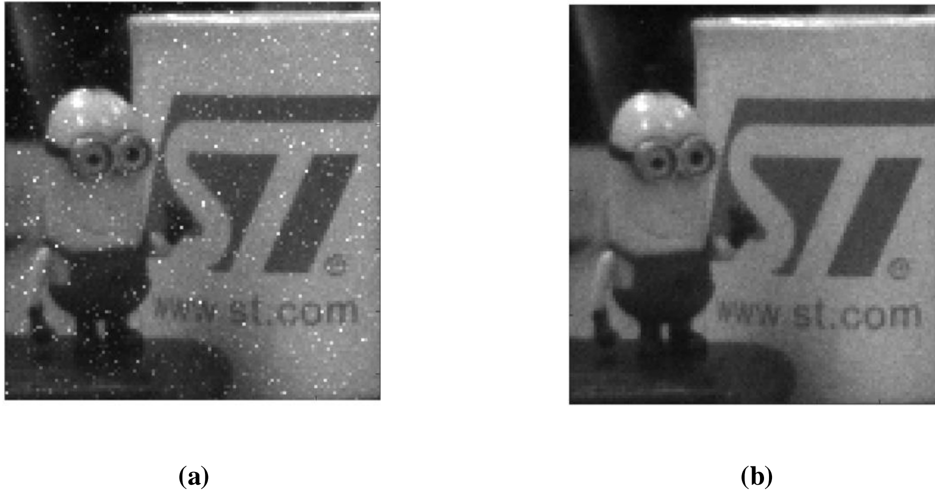


Figure 5.2.11. Measured photon transfer curve for MINI3D sensor. Blue points show a pixel response while the dashed black line shows a frame response. Inset histograms are for the pixel response at 20ns (left) and (1ms) right exposures with ideal Poisson fit overlaid as red crosses.

5.2.7. Intensity Imaging

Figure 5.2.12 shows a single shot greyscale intensity image captured at 2V excess bias in 12-bit linear counter mode. The wider field of view of the scalable sensor is obvious compared to MINIC40 (see Figure 4.2.10). Defective high DCR pixels appear as white dots across the array but can be compensated for by post processing.

Not only does the pixel allow for imaging at the shot noise limit, but it also experiences no parasitic light sensitivity (PLS) due to the digital nature of stored counts which does not degrade the image quality in global shutter mode.

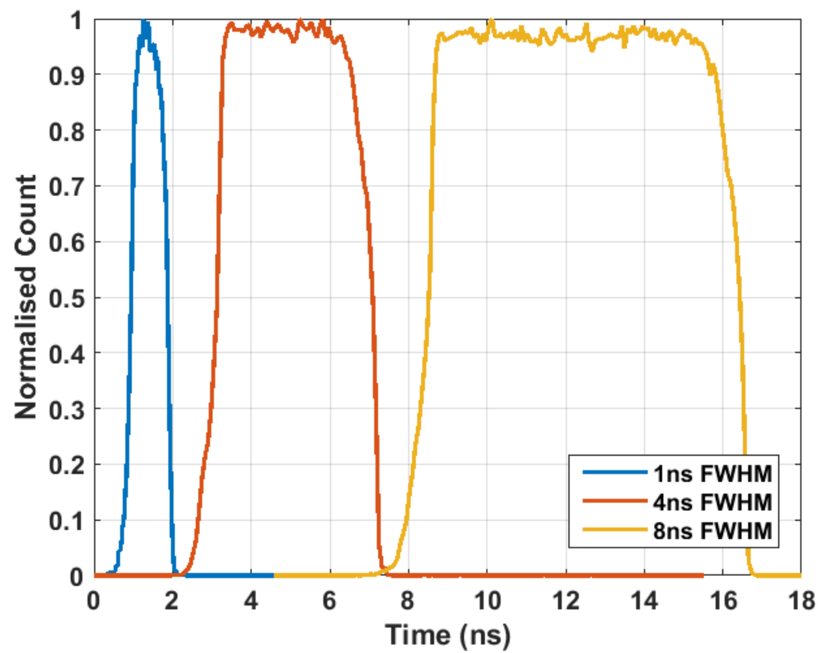


(a) (b)

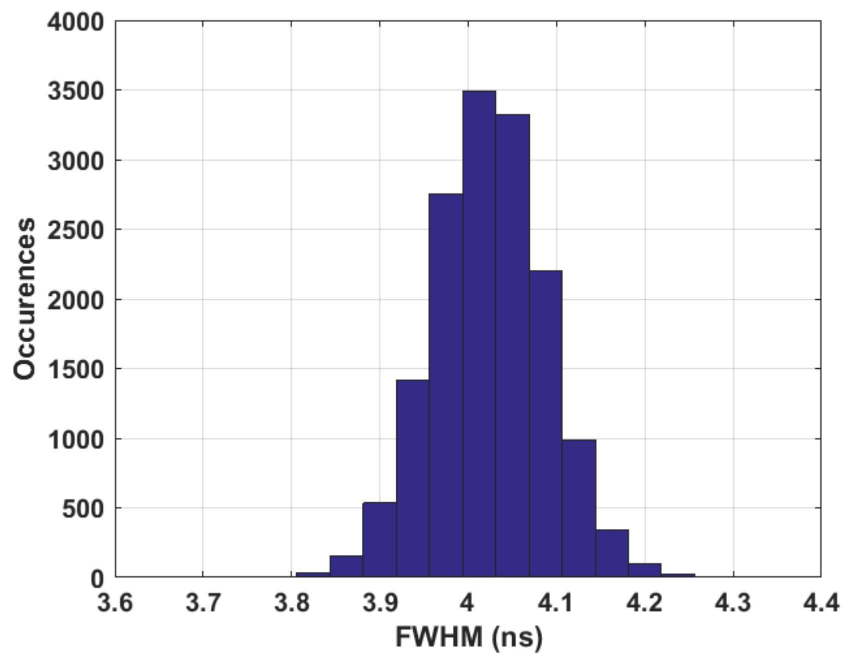
Figure 5.2.12. MINI3D single shot greyscale image in 12-bit linear counter mode. (a) Raw image. (b) DCR corrected by interpolation.

5.2.8. Time Gate Profile

The time gate profile was characterised by sweeping the Hamamatsu laser by steps of 25ps using a Stanford delay generator box (DG645) while recording the photon counts over many repetitions. Figure 5.2.13(a) shows the time gate of bin 1 for a randomly selected pixel of 1ns, 4ns and 8ns FWHM. The standard deviation of the 4ns gate width was calculated to be 64ps across the whole array with the histogram shown in Figure 5.2.13(b).



(a)



(b)

Figure 5.2.13. Time gate profile. (a) 1ns, 4ns and 8ns FWHM time gates for bin 1 of a randomly selected pixel. (b) Distribution of 4ns gate FWHM across the array with 64ps standard deviation.

5.2.9. Time-Resolved Imaging

To demonstrate the time-resolved capability of the sensor, an indirect time of flight experiment was conducted using a PicoQuant 840nm pulsed laser with 4ns pulse width. The 4ns pulse translates to a measurement dynamic range of 60cm. The timing diagram is shown in Figure 5.2.14 where distance is calculated as:

$$Distance = 0.5 \times C \times PW \times \frac{C2}{C2 + C1} \quad (1)$$

Where C is speed of light in free space, PW is the laser pulse width and C1 and C2 are the accumulated counts in time gates 1 and 2 respectively.

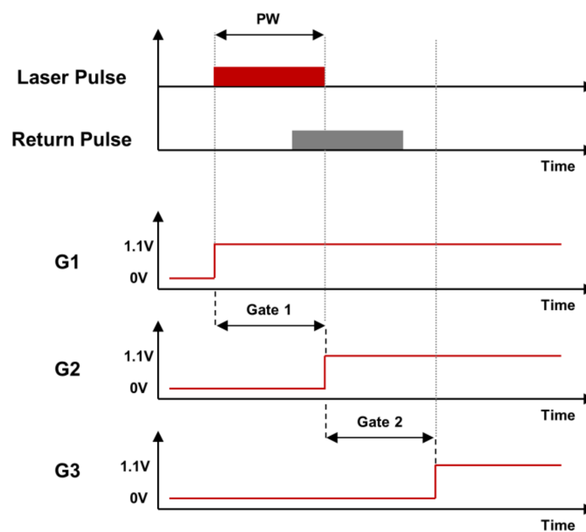
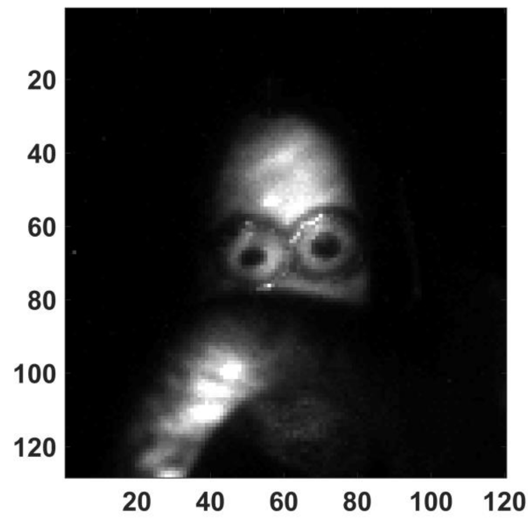
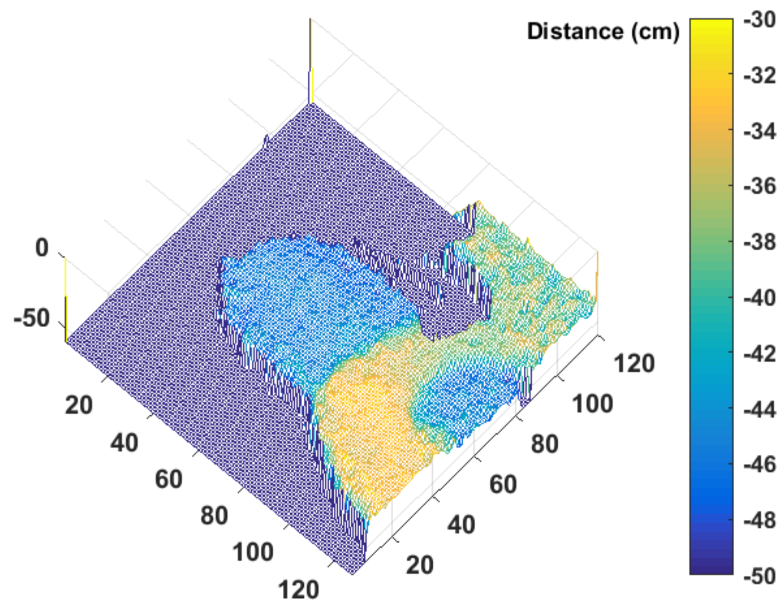


Figure 5.2.14. Indirect time of flight experiment timing diagram.

A mug was placed at ~30cm away from the sensor with a minion toy standing ~15cm behind the mug's handle. Figure 5.2.15 shows an intensity image of the sensor's field of view and a depth map of the scene constructed after acquiring a thousand frames.



(a)



(b)

Figure 5.2.15. Indirect time of flight experiment with 4ns laser pulse width. (a) Intensity image of field of view of the sensor showing a mug's handle obstructing a Minion toy. (b) Depth map of the scene distinguishing the two objects at different distances from the sensor. Median filter applied.

5.3. Comparison to Other Sensors

To understand the implications of implementing SPADs in 3D-stacked technologies and to evaluate the scope for new architectural possibilities, a comparison of the MINI3D is made to other CMOS SPAD sensors.

5.3.1. PDP Comparison to FSI SPAD

Since the 8µm SPC Imager pixel [65][236] reported in a 130nm imaging FSI technology is the reference of comparison to the work of this thesis as it represents the state of the art in terms of miniaturisation, and since both MIN3D and SPC Imager use the same PW / DNW SPAD structure, a comparison of their photon detection probability and efficiency (PDE) is of interest.

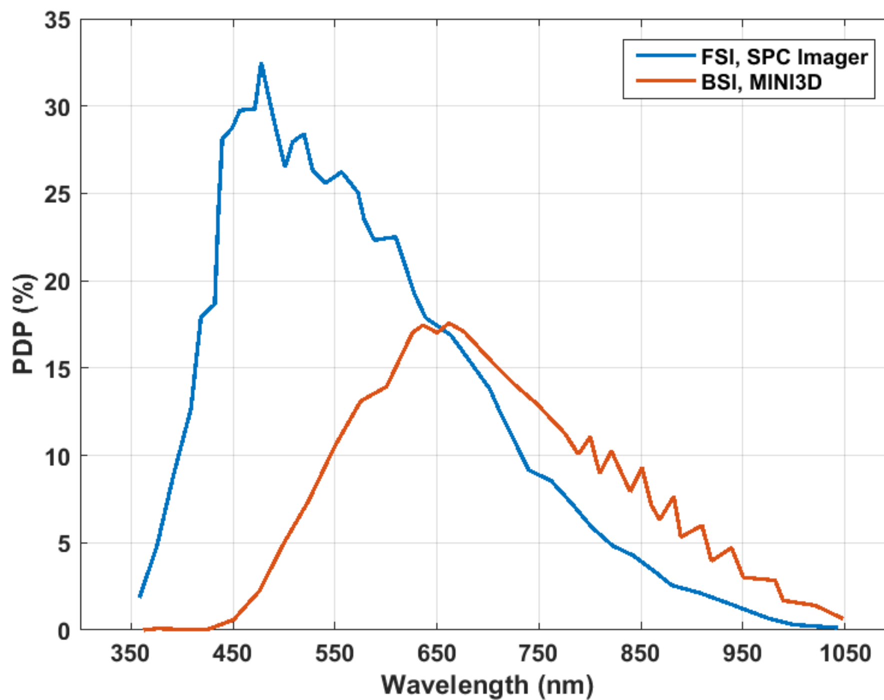


Figure 5.3.1. PDP comparison of FSI and BSI SPAD at 2V excess bias.

Figure 5.3.1 shows the plot of both sensor's PDPs at 2V excess bias. Three important observations are made. First, and as stated in section 5.2.5, the PDP response of the BSI is shifted towards longer wavelengths since the position of the multiplication junction is now deeper into the silicon with respect to the surface of impinging light (Figure 5.3.2).

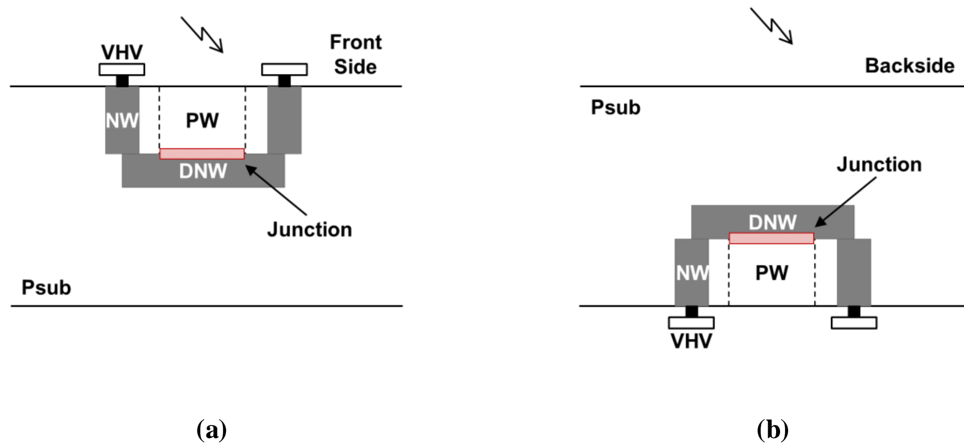


Figure 5.3.2. Illustration of a SPAD structure. (a) FSI. (b) BSI.

Second, and as a consequence of the junction location, a higher PDP for the BSI SPAD at the NIR range is measured. On the other hand a severe cut-off in the blue range is measured due to the absorption of the short wavelengths at the backside surface away from the substrate isolated junction as explained earlier. Finally, the FSI SPAD reports a significantly higher peak PDP (>30%) compared to the 17% of the BSI one which questions the validity of 3D-stacking for improving sensitivity.

Yet the PDP comparison is only a reflection of the SPAD device performance outside the context of a pixel. For a fairer comparison the photon PDE of the two 8µm pixels is shown in Figure 5.3.3. PDE is defined as the SPAD PDP multiplied by the pixel's fill factor. Here the benefit of 3D-stacking becomes much clearer as the high fill factor of MINI3D offsets the high PDP of SPC Imager resulting in similar sensitivity.

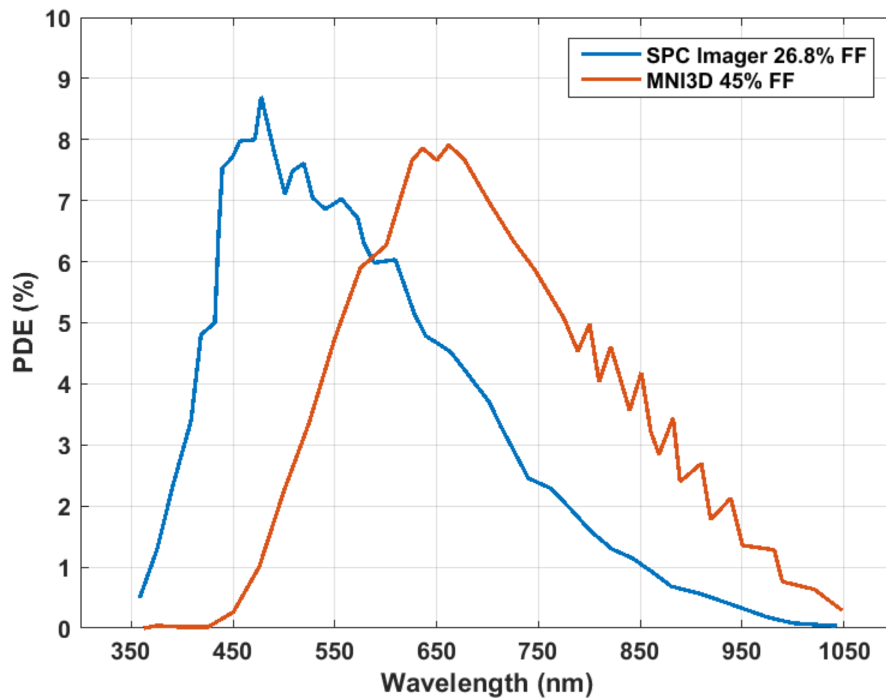


Figure 5.3.3. PDE comparison of FSI and BSI SPAD at 2V excess bias.

Not only is the sensitivity maintained, but the MINI3D pixel offers a much higher photon counting capacity, two simultaneous time-gated bins in-pixel and noise free digital readout and storage all at the same state of the art pixel pitch of ~8μm. Therefore 3D-stacking technology is deemed necessary for miniature high dynamic range time-resolved sensor designs.

A final remark on the difference in PDP of the FSI and BSI SPADs is their suitability for different applications. Since the FSI design has a higher response in the blue region this makes it more suited for biomedical applications such as positron emission tomography (PET) [217] where emission is in that region of the spectrum.

Alternatively, the enhanced NIR response of the BSI SPAD makes it more suitable for consumer applications such as depth imaging which relies on NIR emitters [152]. Other applications such as FLIM can leverage the cut-off of the BSI SPAD to suppress background tissue fluorescence in green against an expected signal emission in red.

5.3.2. Other CMOS BSI 3D-Stacked Sensors

Over the last few years several other all CMOS SPAD devices or sensors have been reported in 3D-stacked BSI technologies, yet in terms of PDP they all seem to exhibit the same blue cut-off and red shift trends. Figure 5.3.4 shows the reported responses on the same plot. The SPAD junction type and the thickness of the backside stack determine the PDP at different wavelengths.

The work in [222] claims an improved blue response due to a thinner optimised backside stack while this work [1] claims an improved NIR response due to the deeper junction location compared the same SPAD in a FSI process.

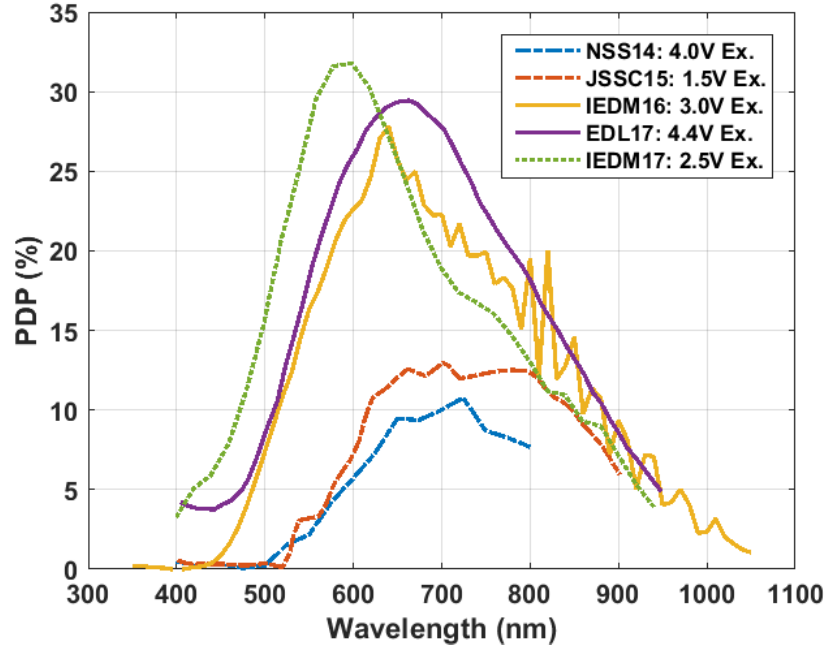


Figure 5.3.4. PDP comparison of different 3D-stacked BSI SPAD at their highest reported excess bias condition.

Table 5.3.1 provides a summary of key technology parameters of the discussed sensors. Apart from the process optimisation of each tier separately, the decoupling of the detector array from the processing electronics allows for enhanced architectural possibilities.

Reference	**[75]**	**[221]**	**This Work [1]**	**[145]**	**[222][223]**	**[224]**
Publication	IEEE NSS/MIC	IEEE JSSC	IEEE IEDM	IEEE EDL	IEEE IEDM/JSTQE	IEEE ISSCC
First Author	Charbon et al.	Pavia et al.	Al Abbas et al.	Lindner et al.	Lee et al.	Ximenes et al.
Year Published	2014	2015	2016	2017	2017	2018
Architecture	Test SPAD	Line Sensor	Image Sensor	Test Pixel	Test SPAD	Image Sensor
Pixel Configuration	Standalone	2 × 4 SiPM + TDC	Counter + Dual Time Gates	Cascoded Passive Quench Active Recharge	Standalone	2 × 8 × 8 SiPM + TDC and Processing Unit
SPAD Junction	P+/NW	NLDD/PW	PW/DNW	PW/DNW	P+/DNW	
Guard Ring Type	PW Implant	NW Implant	Virtual	Virtual	PW Implant	
Substrate Isolated	Yes	No	Yes	Yes	Yes	
DTI Isolated	No	No	No	No	No	
Top Tier Process	130nm	130nm	65nm CIS	65nm CIS	45nm CIS	
Top Tier Integration Level	SPAD Only	SPAD + Minimal Circuitry	SPAD Only	SPAD Only	SPAD Only	
Bottom Tier Process	130nm	130nm	40nm	40nm	65nm	
BSI	Yes	Yes	Yes	Yes	Yes	
Backside Thickness (µm)	5	4.2	n/a	n/a	<3	
Wafer-Scale Bonding	Yes	Yes	Yes	Yes	Yes	
Foundry	Tezzaron	Tezzaron	STMicroelectronics	STMicroelectronics	TSMC	
Bonding Technology	Cu-to-Cu	Cu-to-Cu	Hybrid-Bonding	Hybrid-Bonding	n/a	
Reported Bond Pitch (µm)	n/a	4	7.83	18.36	n/a	
SPAD Active Area (µm²)	28.3	28	27.6	250.7	122.7	
Drawn Fill Factor (%)	n/a	23.3	45	74.4	34.7	
Peak PDP (%)	11	13	27.5	29.5	31.8	
Peak PDP Wavelength (nm)	725	700	640	660	600	
Excess Bias (V)	4	1.5	3	4.4	2.5	

Table 5.3.1. Comparison table of key technology parameters for different all-CMOS 3D-stacked SPAD Sensors. Image sensors are marked in green and the peak PDP wavelength in grey highlighting a trend in BSI SPADs.

The image sensor presented in this work exploits the high integration density of a 40nm process to boost the pixel dynamic range while maintaining time-resolved capability in a small pitch allowing for miniature sensor designs. The fine bonding pitch of the process also contributes towards such venture.

On the other hand, the work in [145] exploits the pixel area differently by using cascaded thick oxide transistors to achieve a passive quench active recharge circuit allowing for a SPAD excess bias up to 4.4V without exceeding the voltage operation limits of the MOS devices. This has the benefit of boosting PDP without having the added circuit area detract from the pixel fill factor. The SPAD reported in [145] was realised in the same STMicroelectronics process in this work and was designed by the author for colleagues at EPFL.

Alternatively, the work in [224] reports a modular pixel comprising a $2 \times 8 \times 8$ block of SPADs sharing the same TDC and digital processing and communications unit. This design leverages the high integration density bottom tier to cleverly share resources in an otherwise computationally intensive TCSPC architecture intended for LIDAR imaging. The module is tileable and can be extended to achieve higher sensor resolutions.

5.4. Summary and Conclusions

The first SPAD image sensor array in a wafer-scale 3D-stacked technology with a 1-to-1 hybrid bond connection between the top tier SPAD and the bottom tier circuitry was presented. The 7.83µm pitch global shared well SPAD array in the top 65nm CIS process achieves a fill factor of 45% and is coupled to a 12-bit digital time-gated pixel implemented in a high density 40nm bottom process.

Characterisation results of the backside illuminated array were also presented with the highlight being its shift in photon detection probability in comparison to a front side illuminated device towards the red region of the spectrum due to the deeper junction. Noiseless single photon counting and time-gated depth imaging results were demonstrated.

In conclusion, 3D-stacking is a key technology for delivering high performance SPAD sensors. The decoupling of the detector from the circuitry allows for optimising each component separately without compromising the other's performance. It also allows for high fill factor pixels due to most or all of the circuitry being implemented on the bottom tier.

Yet depending on the application, innovative SPAD structures are needed to deliver the necessary PDP at the required wavelength with careful engineering of the backside thickness. Other techniques used in CIS can also be used such as DTI for immunity against electrical and optical crosstalk.

As a result of the dedicated bottom tier process, advanced CMOS nodes such as 40nm allow for high integration density increasing the pixel built-in functionality and reducing power consumption. This

development also paves the way for smart sensors utilising configurable pixels that reuse resources efficiently or on-chip digital processing for computational or data management purposes.

The next chapter will present two architectures enabled by 3D-stacking technology, a fully integrated system on chip and a reconfigurable array for miniature time-resolved image sensors.

6. Miniature High Dynamic Range Time-Resolved Sensors

Having explored the opportunities and challenges of designing miniature time-resolved SPAD arrays using the state of the art CMOS technology nodes and fabrication techniques, this chapter looks into novel pixel and system architectures that address the challenges of pixel pitch, embedded functionality, dynamic range and ability to handle high data rates. Two 3D-stacked BSI SPAD image sensor designs are presented.

The first sensor, named ENDOCAM, is a full implementation of a system on chip (SoC) with internal power generation network, programmable micro-control unit (MCU) for autonomous operation, off-focal plane data storage and processing and a 5-wire interface. System level modelling and design considerations are presented alongside simulations and preliminary bring-up results of a fabricated chip proving the functionality of various blocks.

The second sensor, named CORVETTE, presents a novel configurable array architecture at a record 6.48μm pixel pitch. Oversampled time-gated imaging, time correlated single photon counting (TCSPC) on-chip histogramming of regions of interest (ROI) and dual photodiode / SPAD modes of operation are supported. The various application fields of the sensor are discussed alongside a design overview of a chip submitted for manufacturing.

Finally a comparison to published works is made and conclusions are drawn regarding the suitability of the proposed architectures for miniature time-resolved sensors.

6.1. System on Chip Design

The target application behind the design of the SoC sensor is time-resolved biomedical imaging, more specifically endoscopy applications. Since in theory such a sensor will be used in-vivo there are several requirements that need to be met:

1. Small form factor. The overall form factor of any module aimed at such domain has to be maintained as small as possible in order for such sensors to access the smallest cavities in the human body. Therefore it is of importance that the sensor itself is miniaturised.
2. Minimum connectivity. In practice image sensors need a large number of interface connections for power, data delivery and control, but in an applied biomedical setup, it is not conceivable to have tens or hundreds of wires in an endoscopic solution both from miniaturisation and reliability points of view. Hence, the sensor has to follow the industry standard of only a handful of electrical connections.

As a consequence of these requirements, some system design challenges have to be addressed:

1. Pixel pitch and integrated functionality. For the sensor to be miniaturised the pixel pitch has to be small in order to achieve a reasonable resolution within the restricted area. Doing that restricts how much can be processed and stored within the pixel necessitating some form of oversampling in order to acquire enough signal or to extend the dynamic range. In-turn, oversampling means that large amounts of data have to be handled in the process which conflicts with the second system requirement of minimum number of connections, therefore an architecture with data management ability is sought after.
2. Power generation and distribution. Since the number of connections is limited, the system should be able to generate and regulate the necessary voltages internally through an on-chip power network. This might seem to contradict the miniaturisation requirement since more sub-systems are integrated on-chip either increasing the overall area footprint or reducing the available area for the imaging array. Yet this is not entirely true as the number of interface wires is minimised resulting in an overall IC area balance.

Taking that into account, a complete SoC SPAD image sensor was designed and fabricated in STMicroelectronics' 3D-stacked BSI technology with the aim to fulfil the specifications listed in Table 6.1.1 as a compromise between the mainstream endoscopy camera specifications reviewed in Chapter 1 and the capabilities of the CMOS technology in hand.

Parameter	**Specification**	**Comment**
Area	$< 2mm^2$	For 2mm diagonal endoscope tip
Pixel Pitch	$< 10\mu m$	For miniature array in-line with state of the art
Resolution	100×100	Acceptable field of view
Time-Resolved Technique	Time-Gated	Trading-off TCSPC precision for in-pixel bandwidth
IO Connections	< 5	In-line with industry standard
Effective Frame rate	> 10fps	For real-time video output

Table 6.1.1. Summary of target specifications for ENDOCAM SoC sensor.

Figure 6.1.1 shows the ENDOCAM SoC architecture comprising:

1. 128×120 SPAD pixel array with peripheral addressing and readout blocks.
2. Micro-control unit (MCU).
3. Dual 32.768 kilobyte SRAM memory blocks.
4. Ring oscillator (RO) based gate generator and distribution clock trees.
5. Power network with power-on reset (POR).
6. Vertical cavity surface emitting laser (VCSEL) driver.
7. Five wire IO interface (VDD, GND, VHV, CLK and bidirectional DATA)

This block diagram represents the bottom tier chip implemented in 40nm CMOS while the top tier (not shown) contains a corresponding array of 128 × 120 global shared well SPADs with one SPAD per pixel connected to the bottom tier via a 1-to-1 hybrid-bond site.

The peripheral area of the top tier chip was used to implement few independent test structures and banks of MOM capacitors in the available metal layers used as decoupling for VHV and VDD. No active circuits were implemented in the top chip in order to keep the option of replacing it with another process in case of a re-spin with minimal design effort. The top tier process is a 90nm BSI CIS technology and the backside thickness is undisclosed to the author as it is ST confidential.

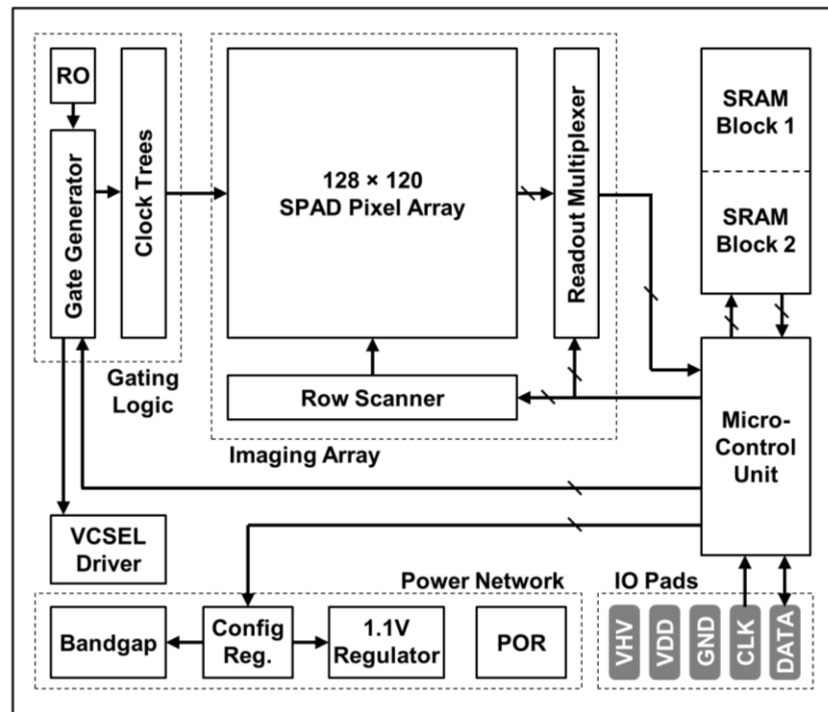


Figure 6.1.1. ENDOCAM system on chip sensor block diagram.

6.1.1. Pixel Design

Similar to the MINIC40 and MINI3D pixels, the ENDOCAM design is fully digital due to the compactness in a 40nm node, readily digitised output and shot noise limited photon counting ability when compared to analogue approaches. The schematic and layout of the 8μm pixel are shown in Figures 6.1.2 and 6.1.3.

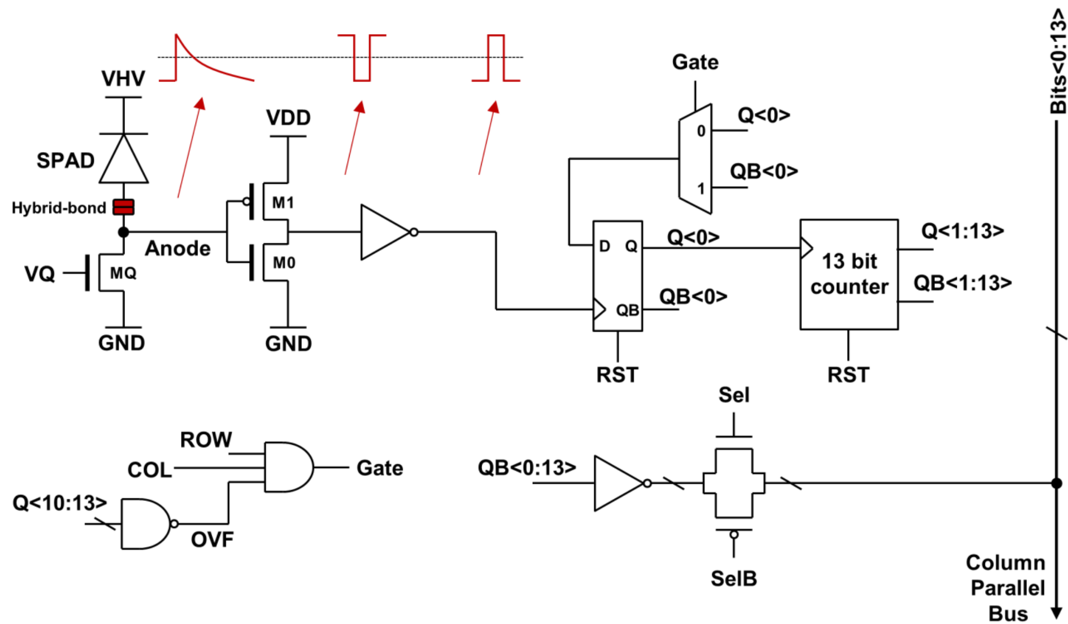


Figure 6.1.2. ENDOCAM pixel circuit diagram. Traces in red indicate signal polarities.

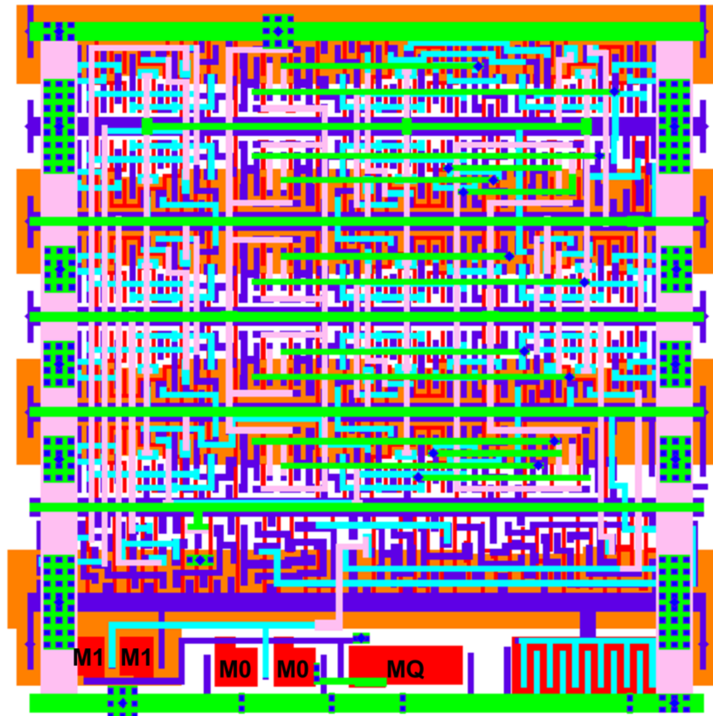


Figure 6.1.3. Layout of ENDOCAM 8μm digital pixel. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3 and green is MT4. MT5 to MT7 are not shown for clarity.

The front end is made of three thick oxide transistors MQ, M0 and M1. MQ is the passive quench passive recharge transistor and M0 and M1 form an inverter operating at 1.1V for direct level shifting. This is followed by a 14-bit ripple counter with the LSB looped back through a MUX to implement edge sensitive gating.

A triple input; row, column and counter overflow, AND gate is used to implement the gating function where the counter overflow is an active low signal based on the four MSBs going high yielding a maximum photon counting capacity of 15360 events.

All of the logic is implemented using 1.1V thin oxide low power standard cells for the exception of the ripple counter bits Q<1:13> where an optimised custom DFF similar to that discussed in Section 3.2.6 was used for area saving. Conventional column parallel readout of the buffered counter bits is implemented through transmission gates driven by row select signals.

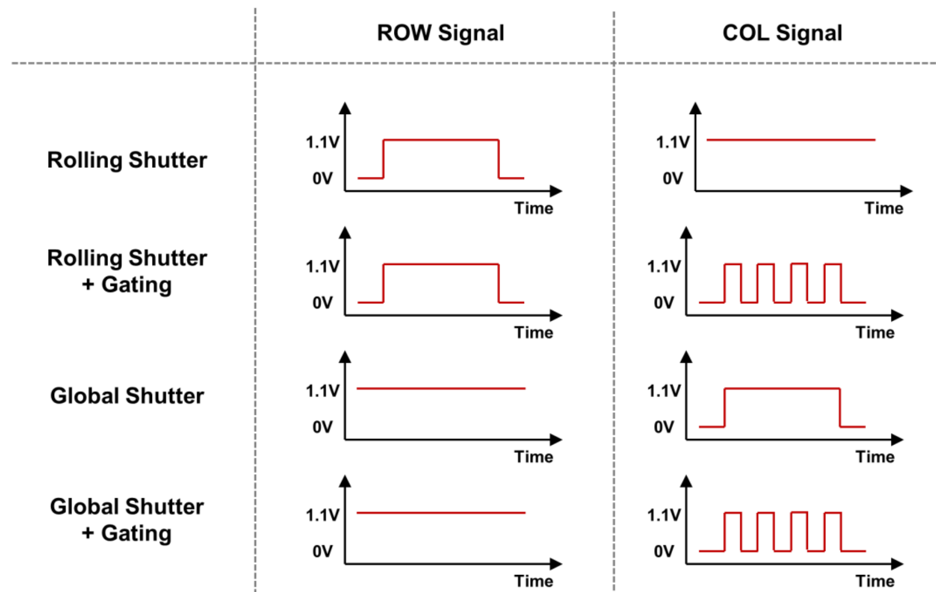


Figure 6.1.4. ENDOCAM pixel different modes of operation and corresponding signal states.

The pixel gating logic allows the sensor to operate in four different ways as shown in Figure 6.1.4 based on the state of the ROW and COL signals. For rolling shutter operation the COL signal is globally held high and the exposure period is defined by the ROW signal which is high throughout the rolling process until the row is selected for readout. Alternatively in time gated mode and while a rolling shutter operation is carried, the COL signal can be globally pulsed to generate intermittent time gates within the rolling exposure period.

For global shutter operation the ROW signal for all rows is globally held high and the COL signal is what determines the exposure period. It can either be set high for a predefined time interval or pulsed to generate time gates. Rolling readout follows in a similar fashion to conventional image sensors.

ROW signals are generated by the row scanner block in the imaging array and COL signals are generated by the gating logic block as shall be explained in Section 6.1.7 where both blocks are driven by the MCU. In all modes of operation, and if the pixel reaches its saturation limit, the active low overflow flag OVF blocks the counter from receiving further SPAD pulses to prevent roll-over.

As for the SPAD device implemented in this chip, it is the same PW / DNW SPAD [111] described in Chapters 3 and 5. The SPAD was designed with conservative guard ring parameters and has a drawn fill factor of 45% at 8μm pitch.

Similar to the array implemented in MINI3D, a metal plate at the back of the device was used as a reflector although in this case only to satisfy metal density rules by having a uniform shape rather than dummy metal fill patterns that vary from pixel to another. In addition, the backside metals were used to create a metal-insulator-metal (MIM) capacitor grid that scales with the array size and functions as a locally distributed VHV decoupling capacitor.

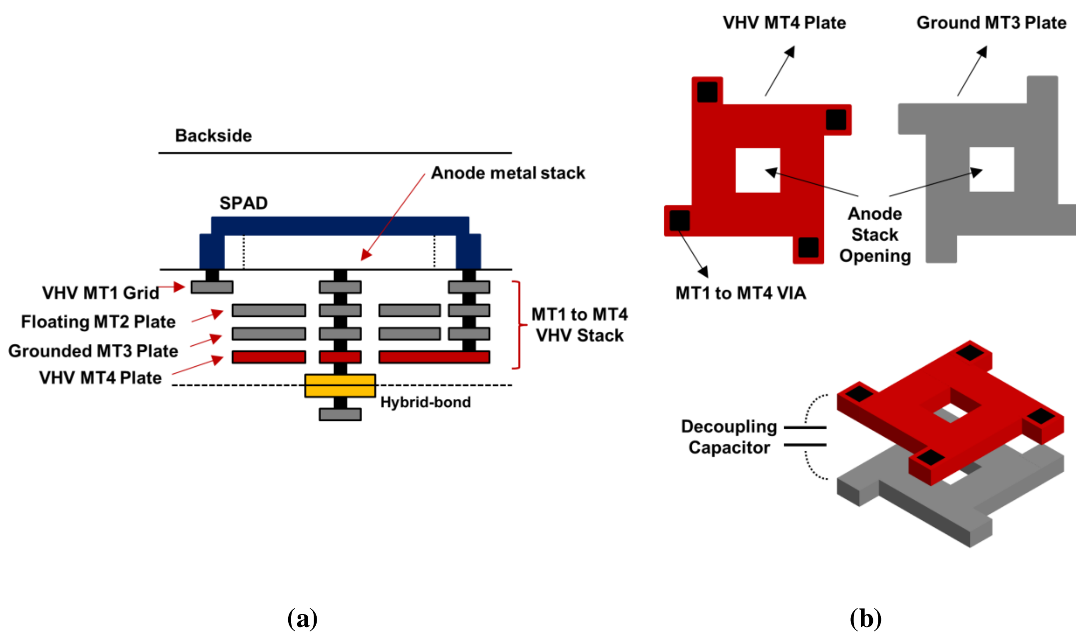


Figure 6.1.5. ENDOCAM top tier SPAD pixel metallisation structure forming a light reflector and VHV MIM decoupling capacitor cell. a) Shows cross section of the device. b) Shows top view of the MT3 and MT4 metal plates forming the decoupling capacitor alongside a three-dimensional rendition of the cell layout.

Figure 6.1.5(a) illustrates the concept showing a cross section of the BSI device while Figure 6.1.5(b) shows a top view of the scalable and tileable structure as the pixel is arrayed. The MT3 plate is

connected to ground while the MT4 one is connected to VHV and has MT1 to MT4 vias to connect it to the MT1 VHV grid of the shared well SPADs.

This configuration not only reduces the resistance of the VHV distribution grid but also creates a local decoupling capacitor as opposed to the global one placed at the edge of the array. This serves the purpose of avoiding VHV droop and maintaining good uniformity at high activity levels across the array. The calculated capacitance per pixel is 4fF which adds up to more than 60pF for the given resolution.

Such capacitance if fabricated reliably can act as a flywheel capacitor for a VHV charge pump if implemented on top tier for example or serve to mitigate supply noise when VHV is supplied over long wires in endoscopy setups. In this design a single IO pad was dedicated for supplying VHV to the top tier SPAD array from an external source.

6.1.2. Array Readout

The readout of the imaging array is carried out by two sub-blocks, the row scanner and readout multiplexer. The row scanner is a standard shift register with a token inserted at the top of the chain and shifted down to select rows for readout. It is also possible to position the token in a dummy location where no rows are selected for readout and so global shutter operation is possible as shown in Figure 6.1.4. When a row is selected for readout three operations take place: LATCH, RESET and READOUT.

During the LATCH phase the column bus data of the selected row is allowed some time to settle and is then latched into a temporary register memory. At this point it is possible to reset the pixel counters by the RESET phase. This is followed by the READOUT phase where 28 bits of two consecutive pixels from the temporary register are funnelled through the readout multiplexer by dialling a 6-bit code to address the 64 pairs of pixels (i.e. 128 columns).

All readout control signals and MUX addresses are issued by the MCU state machine which also uses the same 6-bit code to address the corresponding pixel memory allocation in the SRAM blocks. Figure 6.1.6 shows a typical timing diagram of a single selected row in rolling shutter operation.

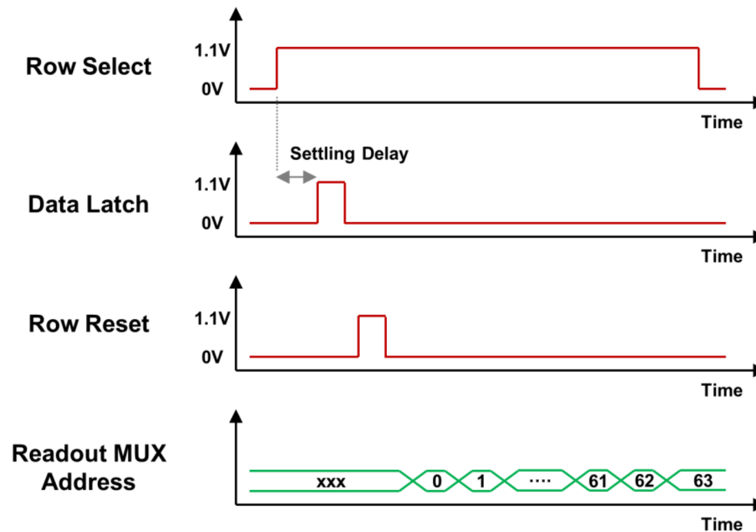


Figure 6.1.6. Typical row timing diagram in rolling shutter mode for a selected row.

6.1.3. Micro-Control Unit and SRAM Memory

The two SRAM memory banks and their controller are standard STMicroelectronics IP blocks which were automatically generated to specification by ST internal tools. As part of the digital design stage carried out by Dr. Almer, these SRAM memories were integrated into the MCU block functionality.

The MCU was designed to perform four main tasks:

1. Control the pixel array readout timing based on selected mode of operation.
2. Manage the IO interface to read out frames or read in register settings as instruction sets.
3. Act as a configuration block by hosting a wide range of programmable register settings that configure the various modes of operation and the settings of all other sub-systems such as the gating logic and power management.
4. Manage the flow of data from the imaging array to the SRAM memories based on the selected mode of operation. It is also responsible for performing simple arithmetic operations such as frame summation or comparison to predefined threshold values.

Figure 6.1.7 shows the layout of the MCU excluding the two SRAM banks. The footprint area is 225µm × 225µm housing roughly 330k logic gates with approximately 33% occupancy density with the rest filled by decoupling cells. It is easy to see how in such an advanced CMOS node it is possible to design very compact low power processing elements opening the door to future smart SPAD CMOS image sensors with built in capabilities such as fluorescence lifetime estimation [292], depth extraction [293], object tracking [260] and much more.

Such capability would significantly reduce the data rates off chip as only the final product is read out and will also reduce the overall system complexity and cost by pushing more components on chip. This is evident in emerging opportunities in the mainstream CIS field enabled by 3D-stacking such as [202], [294] and [295].

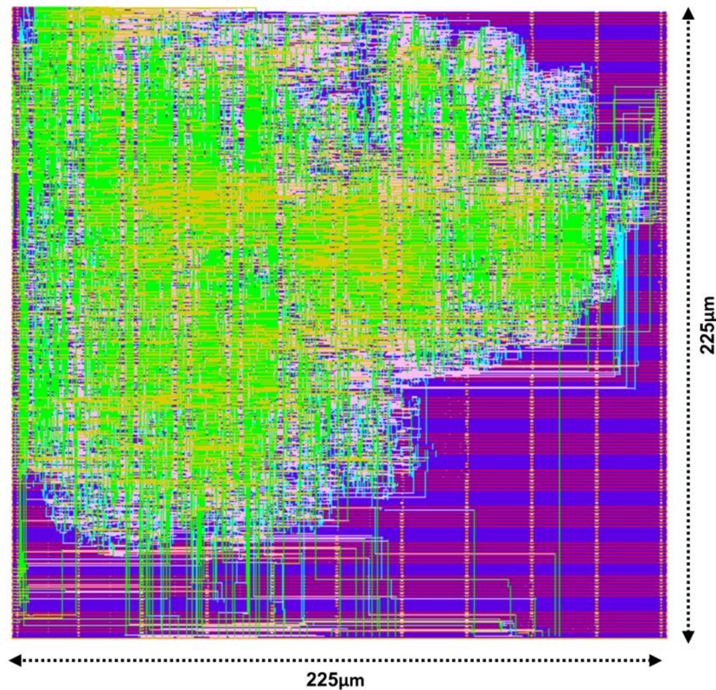


Figure 6.1.7. ENDOCAM's synthesised MCU layout occupying a footprint of 225µm × 225µm. Only MT1 to MT5 shown.

The MCU uses only two IO pads to interface with external world, a CLK pad providing an external system clock for the SoC and a bi-directional data pad configured by the MCU as needed to relay frame information to the outer world or to read in register settings as instruction sets and configuration bits. Digital synthesis tools show the design is capable of running at up to 50MHz without violating timing constraints mainly dominated by SRAM access times but in practice the speed will be limited to more conservative values.

The bi-directional data pad was chosen to have a current specification of 8mA higher than standard IOs in anticipation of large wire loads of an endoscopy system. This load will also determine the system clock speed in a practical scenario.

A single wire debug (SWD) protocol was devised for communication between the chip and the outer world, which is a firmware implemented on an Opal Kelly XEM6310 board interfaced with a MATLAB environment. The sensor initially starts up with initial conditions and is in listening (idle) mode where a set of registers can be configured from MATLAB in a series of 32-bit words. This causes the sensor to trigger capture and readout actions based on the given settings and the transferred

data is received and displayed by MATLAB through a USB3 interface from the Opal Kelly board. The sensor then falls into listening mode again where another handshaking routine triggers it into action.

The result is a continuous acquisition with programmable settings albeit with effective frame rates dependent on the MATLAB interface efficiency. This firmware and software design was fully carried out by Dr. Almer in parallel to the MCU design as part of a co-design and verification approach. Figure 6.1.8 shows a simplified block and state diagrams of the ENDOCAM SoC sensor.

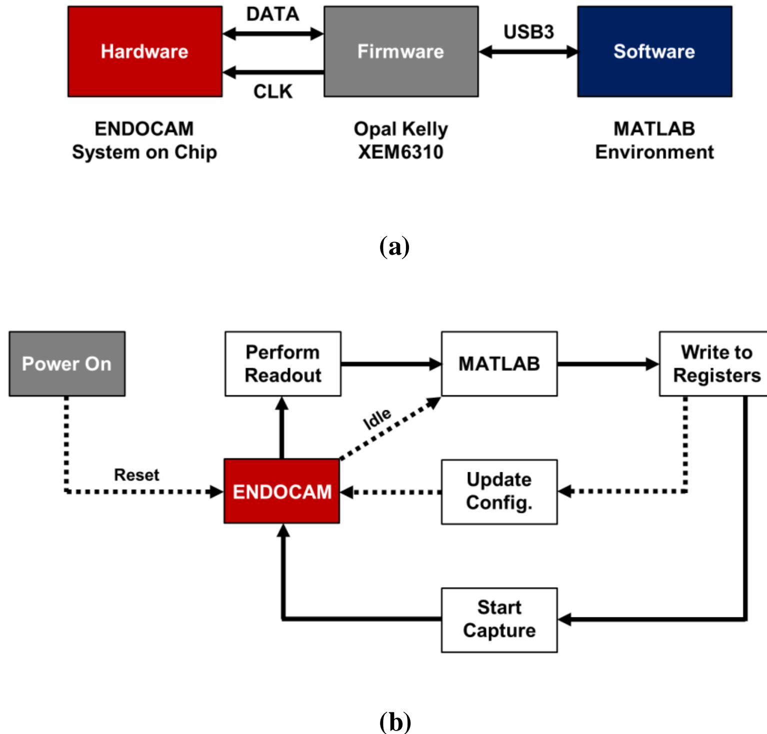


Figure 6.1.8. ENDOCAM overall system setup. (a) Block diagram. (b) Simplified operation state diagram.

Aside from array readout routines, data transfer and configuration setup, the most important feature of the MCU is its data flow and frame manipulation functionality. This feature is at the heart of the SoC design and is aimed at:

1. Improving the sensor's dynamic range.
2. Mediating data flow to reduce data rates and improve frame rates given the single pad output which is necessary for a miniature system.

6.1.4. Dynamic Range Enhancement

As discussed previously, DR is defined as:

$$DR = 20 \times \log\left(\frac{S}{N}\right) \tag{1}$$

Where S is the maximum signal level and N is the minimum signal level observed by the pixel. For the ENDOCAM sensor the maximum signal is 15360 photons as defined by the in-pixel counter limit. Since it has been shown that digital SPAD counters exhibit shot noise limited performance, then the only noise source is the DCR of the SPAD which in value is dependent on exposure time and readout noise is negligible.

Yet since the mean DCR value of the SPAD could be measured, then it can be corrected for by subtraction and so the minimum observable signal is not the DCR value itself but rather its shot noise component (i.e. square root of DCR). This assumption is carried forward in the following DR analysis.

So if for a single frame dynamic range is defined by:

$$DR_{SingleFrame} = 20 \times \log\left(\frac{S}{\sqrt{DCR}}\right) \tag{2}$$

Then when summing M frames together and given that uncorrelated noise sources add in quadrature, dynamic range can be expressed as:

$$DR_{MultipleFrame} = 20 \times \log\left(\frac{M \times S}{\sqrt{M} \times \sqrt{DCR}}\right) \tag{3}$$

And so the improvement in DR is given by:

$$DR_{Improvement} = 20 \times \log\left(\frac{M}{\sqrt{M}}\right) = 20 \times \log(\sqrt{M}) \tag{4}$$

So given the discrete nature of DCR and assuming a minimum integration period which corresponds to a mean DCR value of 1cps, the improvement in DR over the intrinsic pixel DR of 83.7dB is shown in Figure 6.1.9.

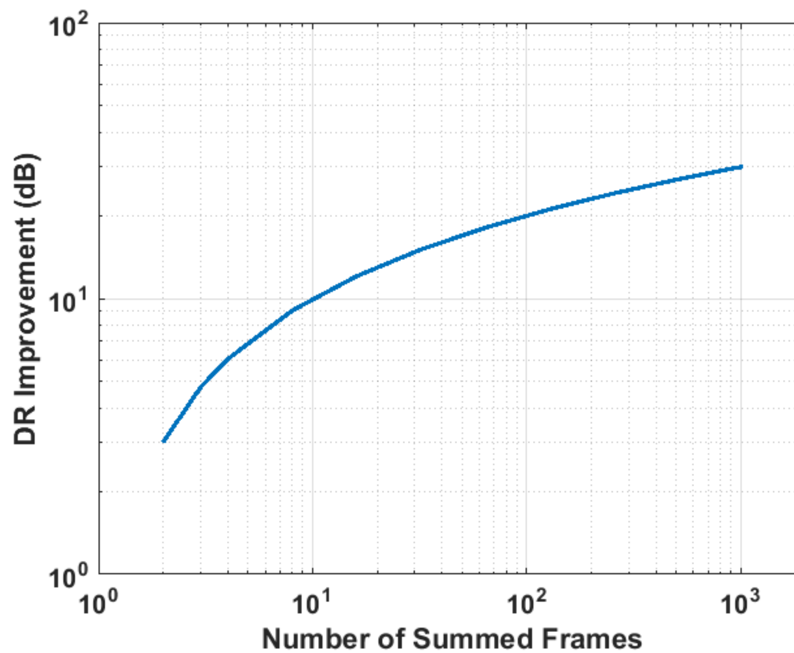
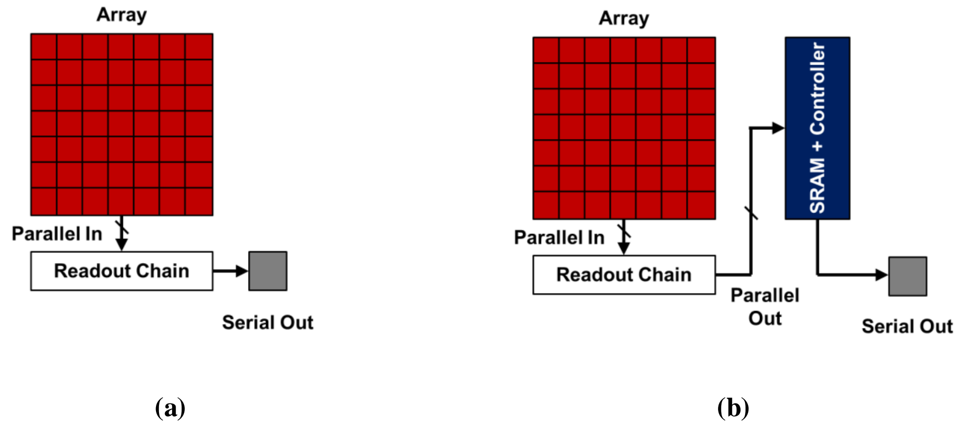


Figure 6.1.9. Improvement in dynamic range in dB for a given number of summed frames.

Of course it is possible to obtain the same improvement by summing frames off-chip, yet this becomes a challenging task for a single output pad interface if a reasonable frame rate is to be maintained, and so an optimal solution between on-chip frame summation and off chip readout is needed.

6.1.5. Single Pad SoC Data Rate Modelling

To explain the necessity of on-chip frame processing or summation, two system examples are considered as depicted in Figure 6.1.10. The first is a standard image sensor with a single pad serial readout to the outer world, the second is an image sensor with parallel data transfer to an on-chip SRAM memory from which the stored data is readout serially through a single output pad.



(a) (b)

Figure 6.1.10. Systems with different data flow models. a) Standard image sensor architecture with single IO pad for serial readout off-chip. b) Proposed image sensor architecture with on-chip frame storage (and summation) with single IO pad for serial readout off-chip.

For the presented modelling results the following parameters and operation specifications are considered based on the ENDOCAM SoC design unless mentioned otherwise:

1. Array consists of 128 columns.
2. Array consists of 120 rows.
3. Each pixel has a 14-bit value.
4. Back to back rolling shutter operation.
5. Single output pad serial readout off-chip.

For both system models the following readout operation parameters are considered:

1. One clock cycle needed for row select.
2. Five clock cycles needed for column bus settling.
3. One clock cycle needed for latching the data into the readout chain.
4. Five clock cycles are needed to issue row reset and other array control signals.

For the standard image sensor serial readout the following operation is considered:

1. A Col × Bit number of clock cycles are needed to stream out a single row content where Col is the number of columns (i.e. 128) and Bit is number of bits per pixel (i.e. 14).

For the data flow of the proposed ENDOCAM system model a different set of readout mechanisms are taken into account:

1. Two pixel values are transferred from the array readout chain to the SRAM block at a time (i.e. 28-bits). This evaluates to 64 clock cycles to transfer a full row content corresponding to 64 readout MUX addresses issued by the MCU.

2. When a pair of pixels is transferred the previous data in their corresponding SRAM locations is fetched by the MCU. The previous data and new data values are added and the resulting sum is written back into the SRAM. Hence an additional clock cycle is needed for the write operation accumulating to 64 clock cycles per row.
3. When frames are accumulated on-chip back to back rolling shutter operation is assumed and a single readout operation of the oversampled frame takes place at the end.
4. When performing a serial readout off-chip the whole 32-bit memory space of each pixel in the SRAM is streamed out regardless of the occupancy of the MSB bits. This is a constraint in the designed system since the SRAM generation tool only allows for specific power of two memory dimensions and so on this occasion a number of unnecessary bits have to be read out. This inefficiency has been factored into the model nevertheless.

Figure 6.1.11 compares the performance of the two sensor architectures in terms of effective frame rate and oversampled pixel bit depth. The effective frame rate refers to the final frame rate output after oversampling (i.e. summing native frames) which for the standard image sensor architecture is referred to hereafter as off-chip frame store since oversampling happens externally and for the ENDOCAM architecture as on-chip frame store as oversampling is carried out internally.

The oversampled bit depth is related to the number of native frames summed to form a single oversampled one. For no oversampling each pixel is represented by a 14-bit number while for an oversampling ratio of 2 for example, each pixel value becomes a sum of two 14-bit numbers equating to 15-bits and so on.

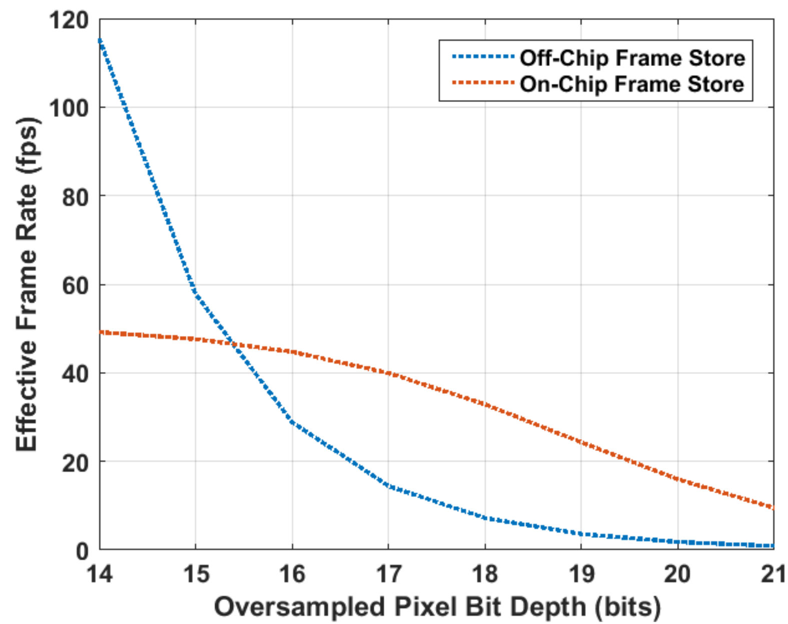


Figure 6.1.11. Effective frame rate versus oversampled pixel bit depth for the off-chip and on-chip frame processing architectures assuming a 25MHz system clock frequency.

The comparison was done assuming a moderate 25MHz system clock frequency. The system clock frequency would influence the frame rate figure but the overall trend and crossing point of both curves still holds true. It can be seen that for the native bit depth of 14-bits the standard system model is more efficient mainly due to the fact that for the on-chip frame store model, data has to be moved into the SRAM first and then readout with some redundant bits as explained earlier.

As more frames are summed together it becomes clear that performing the summation on-chip is the winning case because a single readout operation is needed for an oversampled frame as opposed to a readout operation per native frame in the standard model. Hence the effective frame rate of the standard model drops by the number of frames summed while the ENDOCAM model drops at a much slower rate as the time spent acquiring and summing frames exceeds the readout time allowing for up to 10fps operation even at an oversampled bit depth of 21-bits.

To validate the choice of the pixel counter native bit depth two target specifications were assumed: an effective frame rate of 30fps corresponding to the typical video rates of image sensors and an oversampled bit depth of 16-bits. Modelling results of both the standard and ENDOCAM systems are presented in Figure 6.1.12 showing what system clock frequencies would be required if the mentioned specifications were to be met given different in-pixel counter bit depths to start with.

The 16-bit oversampled bit depth target is chosen since for a 6-bit in pixel counter oversampling up to 16-bits yields a DR of 66dB (36dB native + 30dB improvement by oversampling) which is considered

a typical DR of an image sensor while for a counter with a native bit depth of 16-bits and no oversampling a DR in excess of 90dB can be achieved representing HDR performance. In-pixel counter bit depths below 6-bits were not considered as the pixel response tends away from the linear case towards statistical QIS behaviour which was covered in Chapter 4.

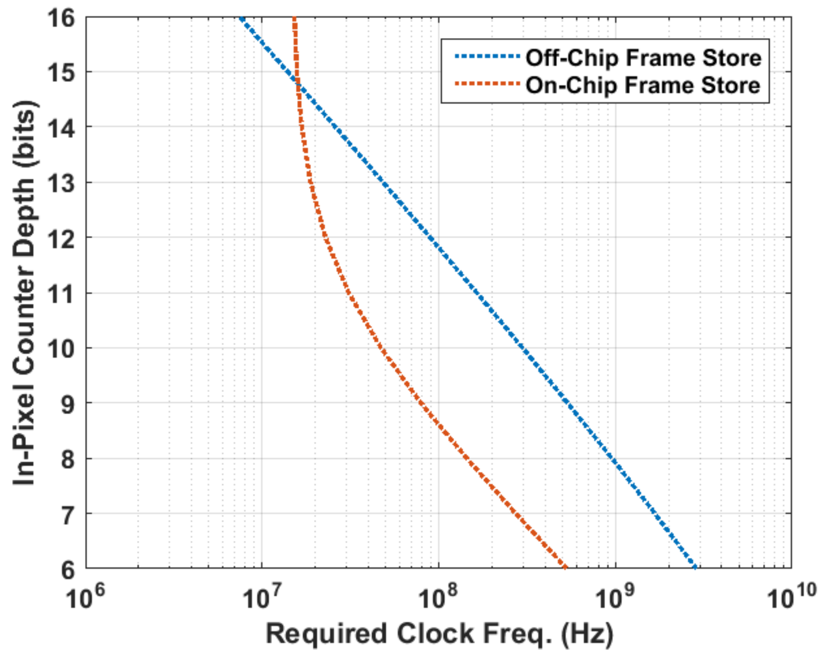


Figure 6.1.12. Required system clock frequency for different in-pixel bit depth counters for an effective frame rate of 30fps and an oversampled bit depth of 16-bits to be achieved given off-chip and on-chip frame store system architectures.

Modelling results show that for the off-chip frame store architecture and for small in-pixel counter bit depths unrealistic system clock frequencies in the GHz region are necessary to meet the assumed specifications and therefore a large in-pixel bit depth is crucial. This of course would immediately impact the pixel pitch and achievable array resolution in a given area.

On the other hand the on-chip frame store architecture requires more relaxed clock frequencies but taking into account the SRAM access times and corresponding timing constraints, operation beyond 50MHz is not feasible. This dictates an in-pixel counter bit depth of 10-bits or more. Beyond a bit depth of 13-bits the required system clock frequency does not reduce further and so the chosen in-pixel bit depth of 14-bits is optimal. Given the optimised flip-flop design used for the counter and the minimum hybrid-bond stacking pitch of 7.2μm for the process in hand a pixel pitch of 8μm was achieved.

Another important system parameter to consider is the temporal aperture ratio (TAR) which is a measure of the shutter efficiency or how much time the sensor spends observing the scene and how much time it is oblivious mainly due to readout dead-time. Figure 6.1.13 demonstrates this in context

of chosen oversampled bit depths for the case of both SRAM blocks operating as a single unit (i.e. chained) with 32-bit memory space per pixel.

It is seen that for the native bit depth of 14-bits the TAR is low and the system is inefficient while a high TAR of ~80% can be achieved when higher oversampling bit depths are chosen. This though comes at the expense of effective frame as the sensor spends more time acquiring than outputting frames. TAR is found to be independent of the system clock frequency as at any chosen oversampled bit depth the number of clock cycles needed to acquire and those required to readout is fixed and only the clock period changes.

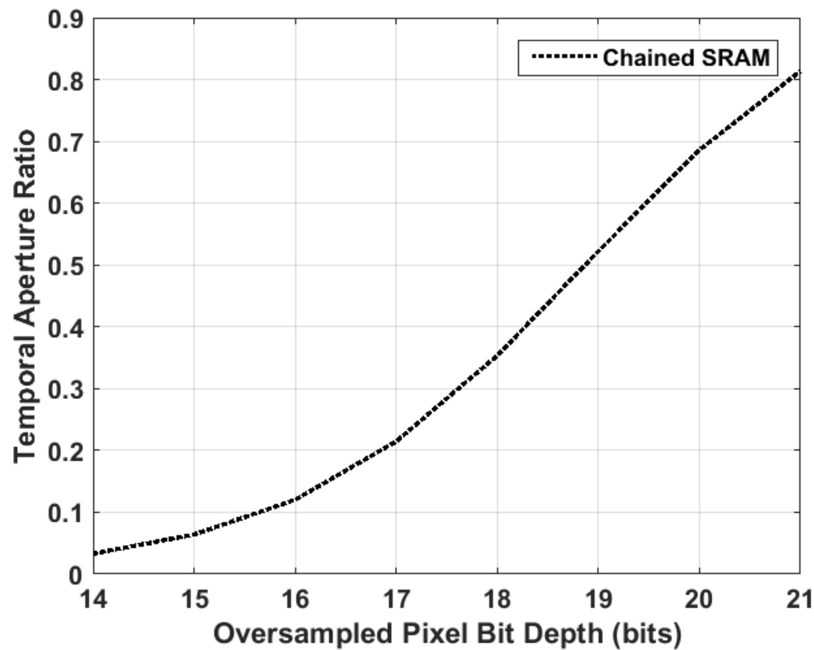


Figure 6.1.13. Temporal aperture ratio (TAR) versus chosen oversampled bit depth for ENDOCAM sensor in chained SRAM mode.

Alternatively, it is possible to operate the sensor in ping-pong SRAM mode where one SRAM bank is readout while frames are accumulated in the other. In this case each 32-bit memory space now represents two independent 16-bit values for two pixels. This makes the readout way more efficient as all 32-bit values are used and there are no redundancies which is partially a motive behind this mode of operation. The main motive though is improving the TAR albeit at the cost of limited oversampling bit depth options with an oversampling ratio of 4 (i.e. 16-bits) being the maximum.

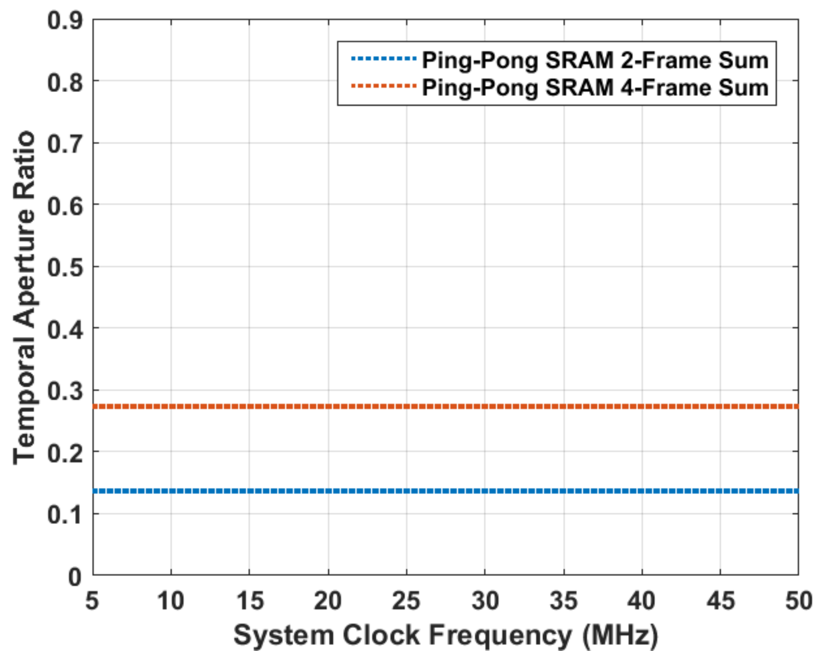


Figure 6.1.14. ENDOCAM temporal aperture ratio (TAR) versus system clock frequency for oversampling ratios of 2 and 4 in ping-pong SRAM mode.

Figure 6.1.14 shows how regardless of the system frequency TAR is fixed for oversampling ratios of 2 and 4. As mentioned before this has to do with the fixed number of clock cycles needed for acquisition and readout in each case but it does point to the fact that the TAR is dominated by the readout time.

If the frame acquisition time was greater than the readout time of the SRAM bank then TAR would be 1 as the sensor will be continuously acquiring as it ping-pongs between the two memory banks. Yet in this case the acquisition is completed in one SRAM block before the second finishes readout meaning that some dead-time is required before switching roles. Nevertheless for both the 2-frame and 4-frame summation cases a TAR of approximately double that of the chained SRAM mode is achieved and so demonstrating an improvement in system efficiency.

Changing the system frequency though has an impact on the effective frame rate as shown in Figure 6.1.15 for the ENDOCAM system in an ideal scenario. Since back to back rolling shutter operation is assumed, and because the line time is directly related to the clock frequency, higher clock speeds translate to shorter exposure times.

If a shorter exposure period than the shutter time is needed then the global COL signal (Fig. 6.1.4) can be used to implement an exposure period within the frame time. Of course that would reduce the TAR even further and so the modelling presented does not consider such cases and avoids any global shutter operation as the best case scenario is always achieved in rolling shutter mode.

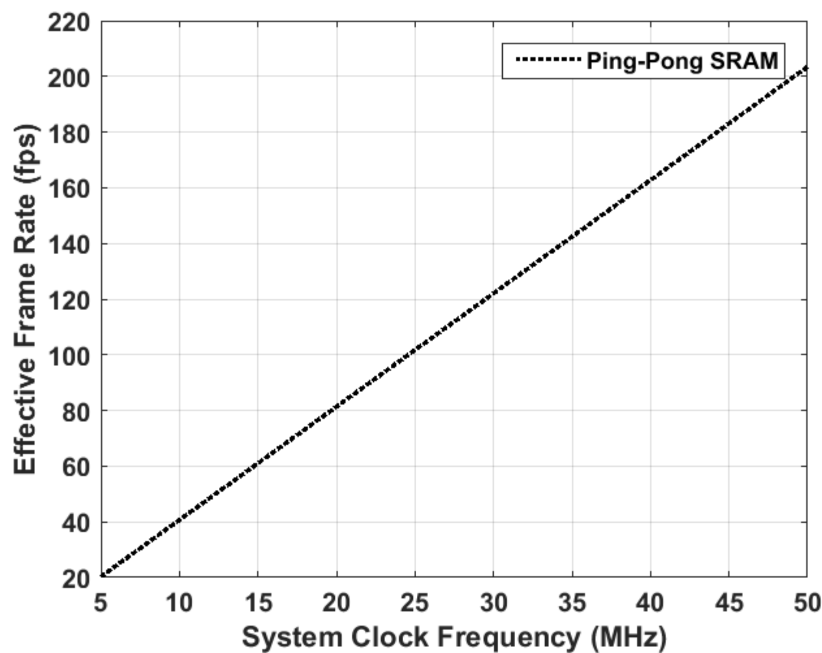


Figure 6.1.15. ENDOCAM effective frame rate (fps) versus system clock frequency in ping-pong SRAM mode.

The effective frame rate is the same for any oversampling ratio (2 frame sum or 4 frame sum) as it is determined by the readout time of the SRAM which dictates how often the switch between the memory banks takes place. In practice such frame rates cannot be reached as the model does not take into account the time required to write instructions to the chip. This is a repetitive procedure which would eventually consume some system time.

Moreover, due to the handshaking and data transfer procedure between MATLAB, the Opal Kelly and the sensor there are some time delays that are unaccounted for. Such an overall system needs careful optimisation and it is beyond the scope of this work which focuses on the silicon design.

In conclusion it is evident that for a miniature sensor design with limited data output capability on-chip frame summation or processing is necessary to reach acceptable dynamic range, effective frame rate (i.e. off-chip data rate) and moderate system clock frequencies. The modelling and design presented are specific to the ENDOCAM sensor but it serves as a proof of principle of such architectures.

The design parameters are more complex than discussed specially when factoring in off-focal plane memories. In this case the SRAM offered a bit density approximately 20× higher than the in-pixel counter (~4.2bits/μm² compared to ~0.22bits/μm²) but the scalability of such design to a higher resolution image sensor is not trivial. For example, for a restricted area the pixel pitch has to drop resulting in a smaller in-pixel bit depth which can be compensated for by oversampling. But that

would dictate a specific system clock frequency and a larger SRAM array which might not be as density efficient as it grows in size due to necessary additional control blocks. Also a bigger SRAM block would impose stricter access timing constraints which would counteract the increase in system clock frequency.

Overall such SoC design has a multi-variable complexity and requires deep understanding of the target application and specifications for optimal design. Yet what is demonstrated in this work is that with the help of advanced CMOS nodes and the advent of 3D-stacking technology the door is open to an endless stream of innovative architectures.

6.1.6. Power Network, Testability and VCSEL Driver

While the innovative aspect of ENDOCAM is in the system design, it is still necessary for the sensor to operate autonomously and with minimum connections to the outer world, which makes the power generation network embedded on-chip crucial to the SoC concept and often overlooked as in essence it is a collective of analogue IP blocks.

The power generation network requires two supply inputs, typically 2.8V and ground, both supplied through a VDD and a GND IO pads. The core blocks making up the network are:

1. Bandgap with multiple reference voltages (0.4V, 0.9V and 1.1V).
2. Voltage regulator for generating the core logic 1.1V supply with up to 20mA current supply.
3. Power-on reset (POR) for initialising the sensor upon start-up.

These three components were provided by STMicroelectronics as standard IP blocks in their 40nm CMOS process and the selected them based on voltage domain, current supply and area requirements. These blocks were instantiated and configured by the author with the assistance of Dr. Neale Dutton. The network was built into a single layout block which also integrates other components such as:

1. Configuration registers addressable by the MCU.
2. Analogue test MUX connected to an analogue test IO pad for debugging purposes.
3. Digital test MUX connected to a digital test IO pad for debugging purposes.
4. Configurable VQuench MUX for different VQuench settings.
5. VCSEL driver circuit for a complete SoC design proof of principle.

All these extra circuits were designed by the author and integrated into the power generation network. The two analogue and digital test pads are part of the SoC IO but are not necessary for its operation and so do not count towards the 5-wire interface. Figure 6.1.16 shows a simplified block diagram of the power generation network with the core blocks, configuration register bank and main signals labelled.

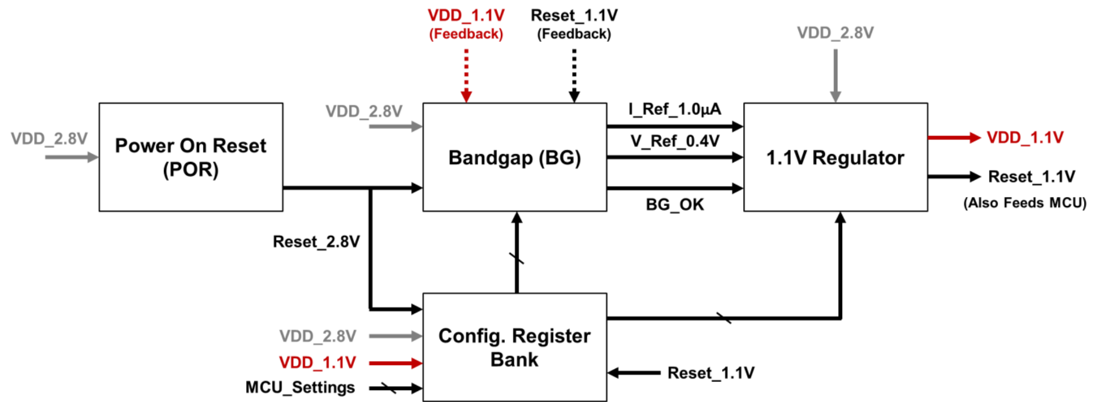


Figure 6.1.16. Simplified block diagram of ENDOCAM power generation network with core blocks, configuration register bank and main signals labelled.

The configuration registers ensure that the system starts with the correct initial conditions and allow reconfiguring several options in the ST IP blocks such as voltage and current trimming values of the bandgap to account for process variability. The initial conditions also guarantee that all test ports are grounded to avoid unnecessary loading upon start-up or conflict of signals.

The analogue test MUX is used to observe internal node voltages such as the regulator output and it is possible to externally drive in static bias voltages such as VQuench as a backup plan. Similarly the digital test MUX allows observing some of the array addressing signals one at a time for debugging reasons and also allows driving in an external digital signal through the bi-directional digital pad such as an external time gate if needed. Care was taken to ensure that all registers were addressed and supplied by the correct voltage domain (2.8V or 1.1V) and level shifters were used when necessary.

The VQuench MUX gives some flexibility in setting the dead-time of the SPADs and so their intrinsic dynamic range as they are implemented in this top tier technology for the first time. Initially the VQuench MUX selects the 1.1V of the regulator upon start-up but it is possible to switch to 1.1V or 0.9V from the bandgap references as a quieter source or even drive in externally a voltage higher than 1.1V. A 2pF capacitor was implemented in layout at the output node of the MUX to filter any noise due to the SPAD activity from feeding back into the bandgap references as shown in Figure 6.1.17.

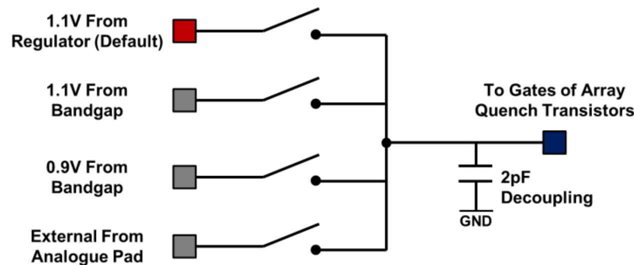


Figure 6.1.17. ENDOCAM VQuench MUX settings with 1.1V from regulator as default setting upon start-up.

Despite being seemingly trivial great care should be taken when determining the start-up conditions of the power generation network to guarantee that voltages do settle correctly and no instability arises causing a system failure. Start-up verification simulations were carried out and one potential issue was spotted. Initially the VQuench MUX was meant to take in 1.1V from the bandgap reference as default value but that was changed later to 1.1V from the regulator instead.

When the 2.8V supply is ramped up the power-on reset (POR) 2.8V reset signal is held low for a period of time and only after it is released (i.e. high state) does the bandgap block start functioning. After some start up delay the bandgap reference voltages are generated and once they settle it releases a BG_OK flag indicating that it is ready for operation.

Following that the 1.1V regulator starts up. At this instance there is no 1.1V supply on chip and all the configuration registers responsible for controlling the test multiplexers are off and so are the MUX switches. The moment the regulator reaches its steady state the 1.1V supply is generated and as soon as it reaches ~600mV of value the digital circuitry wakes up at initial conditions since the 1.1V reset signal from the regulator is still active (i.e. low).

Consequently the VQuench MUX is activated and the 1.1V reference voltage of the bandgap suddenly sees a very large capacitive load in the form of the 2pF decoupling capacitor and the gate capacitance of 15360 long quench transistors in the imaging array. As the bandgap is designed to supply very stable static voltages and not to drive big loads, the reference voltages dip as a reaction. This dip is sensed by the bandgap as a disturbance in the system and so sets the BG_OK flag low causing the regulator to shut down.

Here the 1.1V supply goes off but the VQuench MUX keeps the reference 1.1V connected to the load through an analogue switch powered by the stable 2.8V supply. Eventually the bandgap reference voltage settles back to its 1.1V value and the BG_OK flag is set high causing the regulator to start up correctly this time. Figure 6.1.18 shows a simulation of this condition.

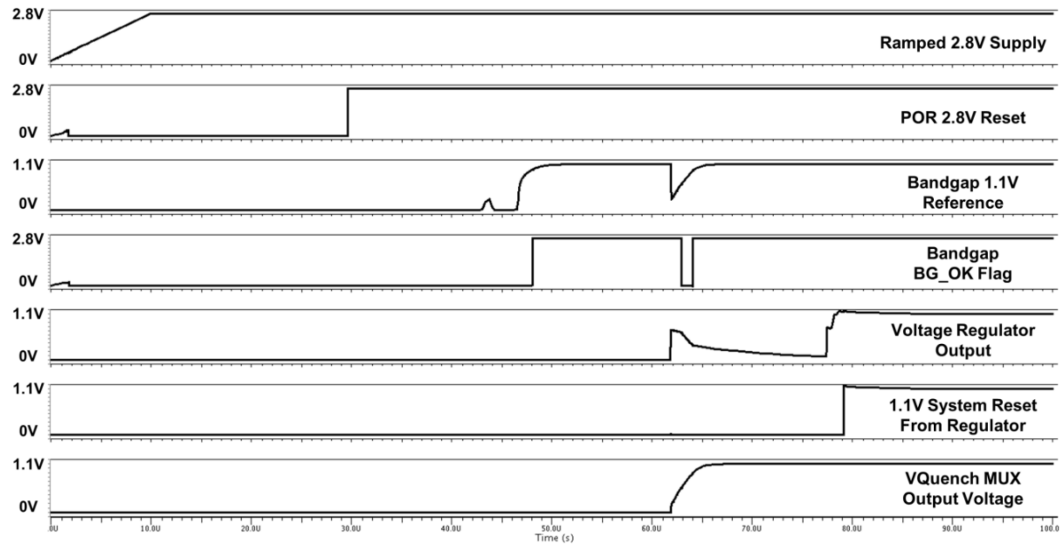


Figure 6.1.18. Start-up simulation of ENDOCAM power generation network with glitch due to capacitive loading on bandgap reference voltage.

As a fix the initial conditions of the VQuench MUX where changed to start with 1.1V for the regulator which is meant to drive loads and after the SoC starts up correctly, the quench voltage can be seamlessly configured to switch to the much quieter 1.1V from the bandgap without causing a dip as the load has already been pre-charged. This ensures a smooth start up and avoids unexpected failures. Figure 6.1.19 shows a simulation of the final design start up sequence.

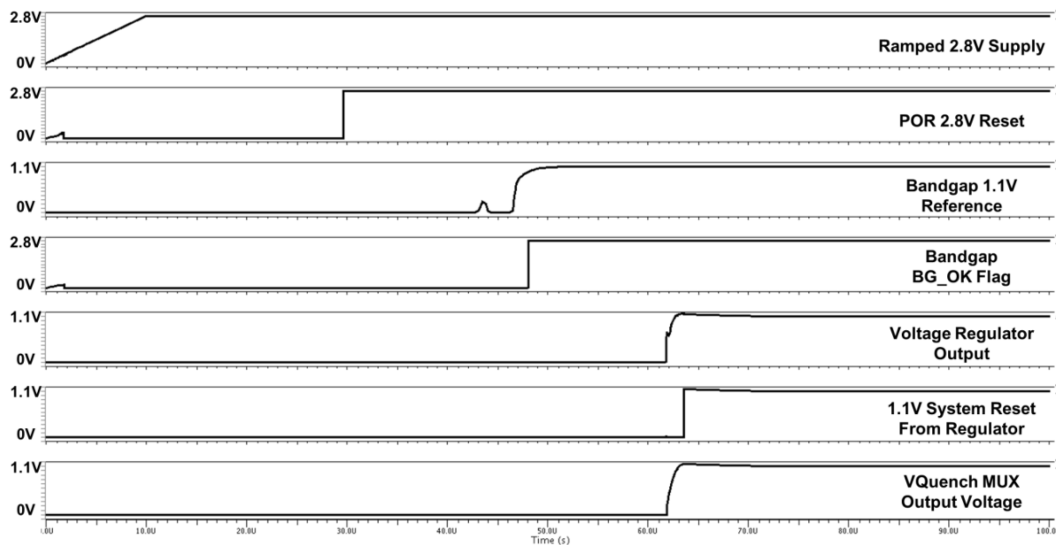


Figure 6.1.19. Smooth start-up simulation of ENDOCAM power generation network after fixing initial start-up setting to avoid the glitch due to capacitive loading on bandgap reference voltage.

A vertical cavity surface emitting laser (VCSEL) driver circuit was also integrated as part of ENDOCAM power network sub-system. This block was a nice to have option and not necessary for

operation but it was added for complete proof of principle of a fully integrated time-resolved SPAD image sensor SoC.

A basic inverter based design was adopted for pulsed illumination and is composed of a cascade of custom inverter cells with increasing size (i.e. drive strength) to ensure minimal propagation delay and sharp switching characteristics when driving capacitive loads. Figure 6.1.20 shows a circuit diagram of the VCSEL driver stages.

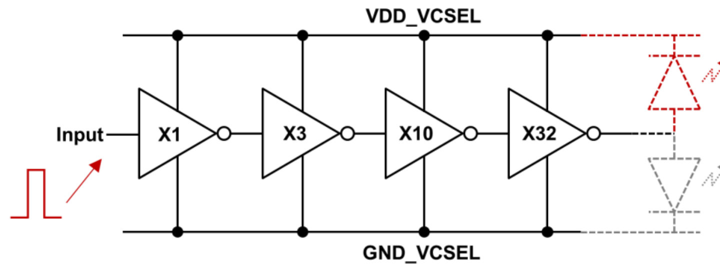


Figure 6.1.20. Schematic of inverter based VCSEL driver implemented in ENDOCAM sensor.

The VCSEL driver was isolated in its own n-well such that it has its own VDD_VCSEL and GND_VCSEL. The two supplies and the driver's output node were connected directly to passive IO pads on the top tier through multiple hybrid-bond sites and no active IO's were integrated on the bottom tier in order to save area since all connections are either to source or drain terminals and not to gates which are highly vulnerable to electro-static discharge (ESD) damage.

The transistor sizes of the final driving stage were chosen such that currents starting from 24mA and up to 88mA can be supplied by adjusting the VCSEL supply between 1.8V and 3.3V. Available area on the top tier was used to implement MOM decoupling capacitors for the VCSEL supply to better cope with large current transients.

Several configuration bits from the MCU make it possible to enable and disable the driver, select its input driving signal polarity to allow for any VCSEL configuration (dotted line in Figure 6.1.20), use it in DC on / off mode if to drive a light source in an endoscopy setup, to take in a pulsed input from the gating logic such that it operates synchronously with the time gates in time-gated mode or even to accept an external pulse through the digital test pad for testing purposes.

An overhead of integrating the VCSEL driver is the need for an additional VCSEL supply due to the high currents needed and the fixed voltage domains on-chip (i.e. 2.8V and 1.1V). Moreover, integrating an additional on-chip voltage regulator for that purpose is not feasible due to strict area constraints. If the sensor is to be bonded in the same package as the VCSEL, the external GND line can be shared if it has low enough resistance while the VCSEL driver's output node can be wire

bonded to the VCSEL node directly, but an independent line for VDD_VCSEL is necessary making the system a 6-wire solution.

Figure 6.1.21 shows the layout of the power generation block with all blocks labelled for the exception of the analogue and digital test multiplexers which sit closer to the IO pads (not shown).

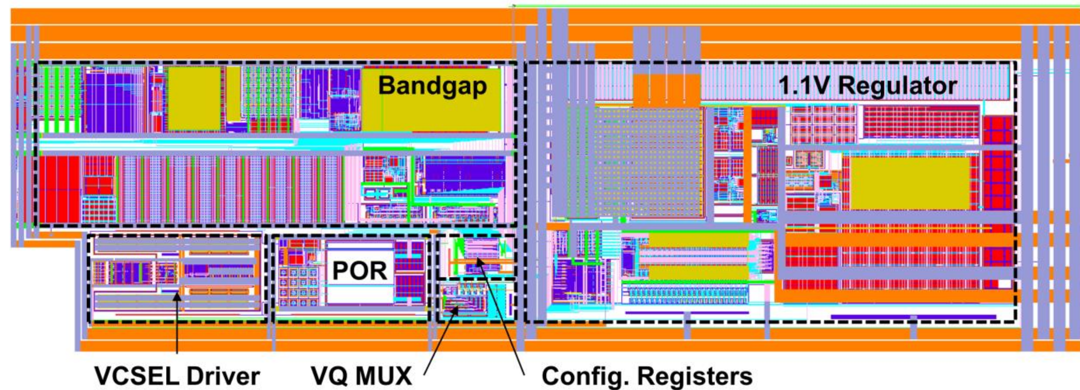


Figure 6.1.21. Layout of ENDOCAM power generation network with main blocks labelled.

6.1.7. Gating Logic

Time-resolved imaging capability is a key motive behind the ENDOCAM SoC and so the sensor should be capable of delivering such functionality. As with the other designs presented in this work time-gating approach was adopted as it is much more area efficient from the miniaturisation point of view.

It is possible to extract temporal information by simply collecting photons in two time gates and so approximate parameters such as fluorescence lifetime by rapid lifetime determination (RLD) [289] or and distance or depth by indirect time of flight methods [64]. Of course this trades-off the accuracy of the statistical approach of histogramming time stamps of captured photons for pixel simplicity and compressed data rates.

In order to achieve this goal, on-chip time gate generation logic was implemented based on a fully digital custom design handcrafted in the analogue flow with the following objectives in mind:

1. Compactness for minimal area overhead.
2. Programmable for flexible and adaptive operation.
3. Sub nanosecond temporal resolution to allow for nanosecond width time gates.

The design takes advantage of the compactness and high speed of the 40nm logic and is based around a ring oscillator (RO) and programmable ripple counters to produce three time gates, GATE_A, GATE_B and GATE_VCSEL to supply the even columns of the array, odd columns of the array and the VCSEL driver respectively. Figure 6.1.22 shows a simplified block diagram of the gating logic.

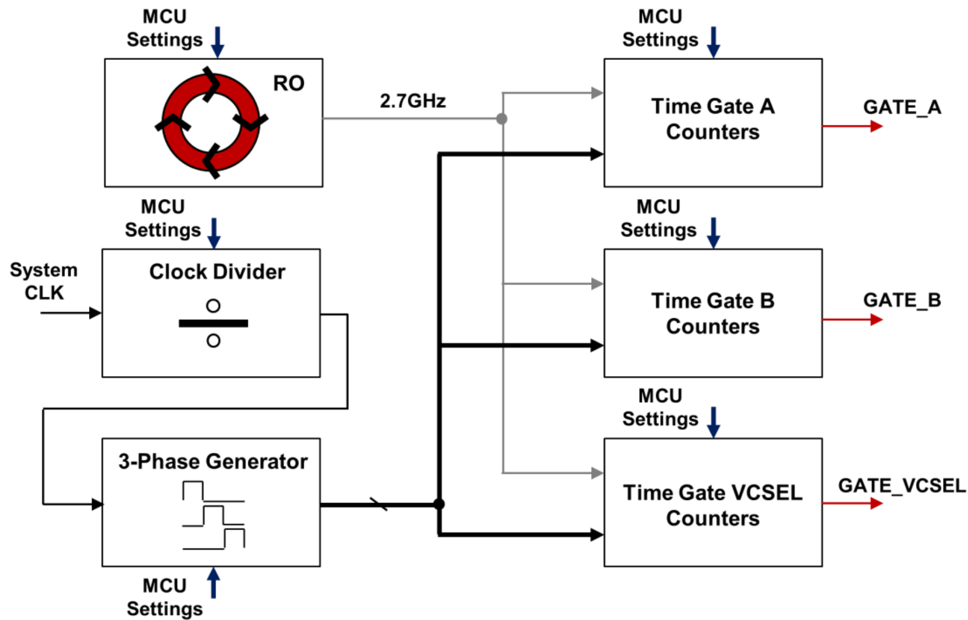


Figure 6.1.22. Simplified block diagram of ENDOCAM gating logic.

The basic principle of the gate generation logic is simple. Each time gate counter block implements two 12-bit counters that operate in three phases where the phases are derived from the system clock. During the first phase both counters are set to logic high and in the second phase each of the counters is set to a programmable 11-bit value by selectively resetting particular bits to logic low. The twelfth bit is always reset to logic low in this phase. During the third phase, the high speed RO clock is enabled and causes both counters to count down until they roll back over to all logic high, which is a different point in time for each of the counters based on their programmed values.

At this instance, once the twelfth bit goes high for any of the counters, a rising edge is generated. The twelfth bit is used instead of the zero-crossing count to avoid meta-stability due to different bit settling times when comparing the 12 counter bits and because a counter's MSB stays high for much longer than a single RO clock cycle allowing sufficient time for signal propagation delays.

The two rising edges generated by the pair of counters feed an AND gate which generates a corresponding time gate. The absolute value of the programmable count of each counter defines the time offset of the generated time gate in number of RO clock cycles with respect to the system clock phase and the count difference between the counters defines the time gate width in number of RO clock cycles. Figure 6.1.23 shows a timing diagram for the three phase operation to better convey the concept.

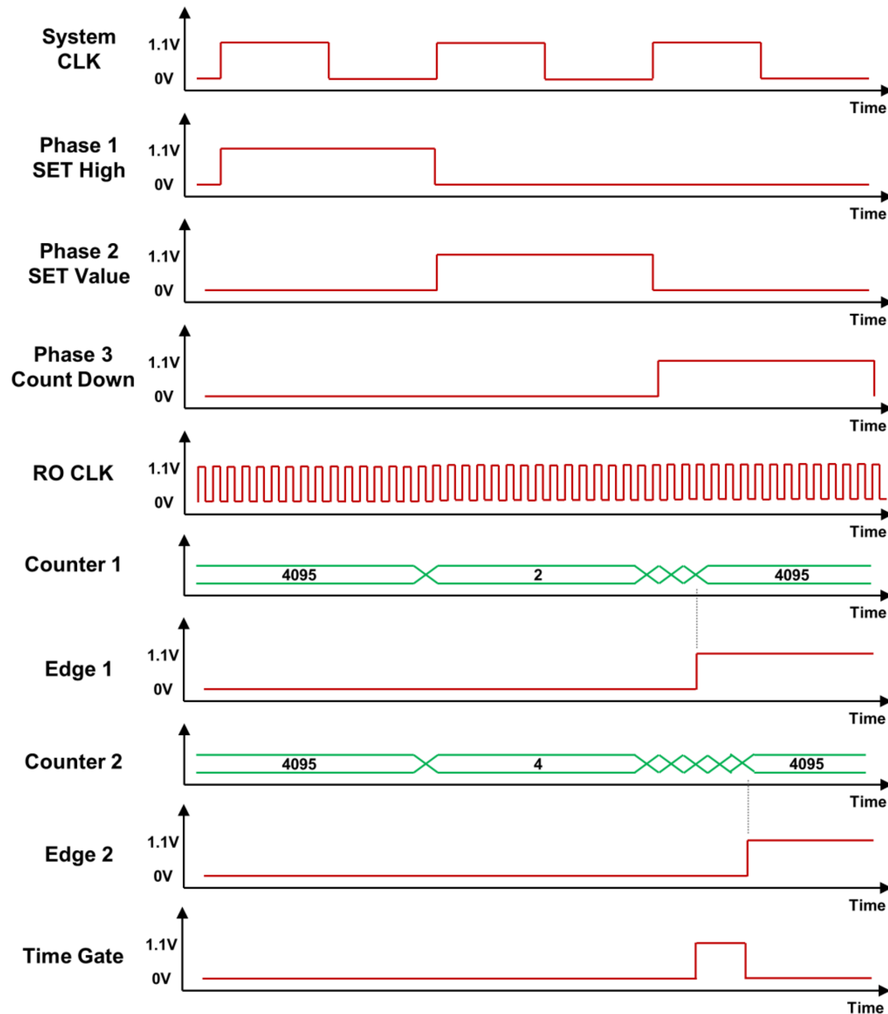


Figure 6.1.23. ENDOCAM gating logic timing diagram for one counter pair.

Since the counter pair requires a three phase operation during which only one of them generates a time gate, the overall design is not very efficient. To overcome that, two additional counter pairs that take in the same MCU settings are implement with the difference being that the order of the phases they use is alternated such that at any system clock phase one of them is being set to all logic high, the second is set to the programmable count value while the third is generating a time gate by the described mechanism. This guarantees continuous time gate generation with every system clock cycle.

This composite counter block is then instantiated in the design three times for generating GATE_A, GATE_B and GATE_VCSEL (Fig. 6.1.22) with each having their independent programmable MCU counter settings and so the different time gates can either overlap or be displaced or delayed in time as necessary.

The temporal dynamic range of the time gate generation logic is dependent of the system clock frequency. An 11-bit value was used to ensure that as many RO clock cycles as required can be

counted within a system clock period. It is also possible to divide the system clock frequency and use the divided version to drive the gating logic in order to increase the temporal coverage or dynamic range. On the other hand, the temporal resolution is defined by the RO clock period which is 370ps from the extracted simulations under typical conditions of the 2.7GHz 4-differential-stage RO.

The ring oscillator also has configurable settings controlled by the MCU allowing for turning it off to save power when time gating is not used. It is also possible to switch the power supply of the RO between the internal 1.1V from the regulator or an external source through an analogue switch. The external source can be provided through an additional passive pad placed on the top tier which is not necessary for the SoC operation and so does not count towards the 5-wire interface but was implemented as a back-up plan and for testing purposes.

Another important programmable feature is the ability to start and stop the RO continuously throughout operation with every system clock cycle. Since no process-voltage-temperature (PVT) compensation mechanism is integrated on-chip through a phase locked loop (PLL) or a delay locked loop (DLL) due to area constraints, allowing the RO to free run for long periods of time would result in accumulation of non-linearity and so integrated jitter [296]. In an attempt to avoid that, this feature stops, resets and restarts the RO with every system clock cycle. By design it is ensured that this operation consumes minimal time such that the temporal dynamic range is not reduced. Figure 6.1.24 illustrates the concept.

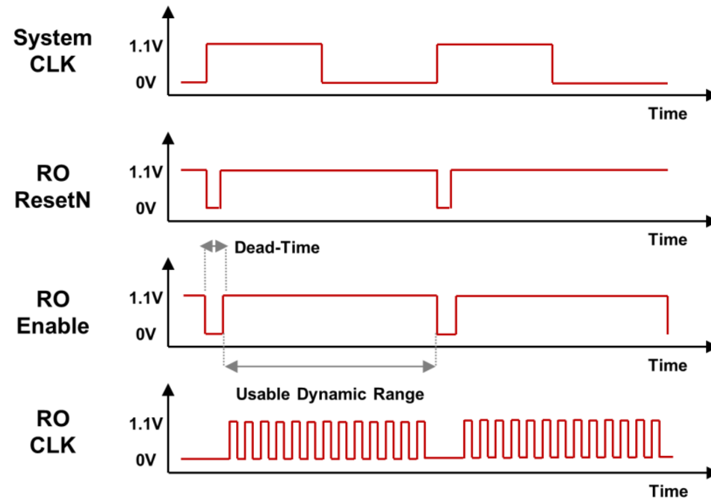


Figure 6.1.24. ENDOCAM ring oscillator timing in self reset mode.

Moreover, if the RO is allowed to free run, and if the pulsed light source is triggered externally the phase offset between the RO clock and the system clock would mean that any defined time gate would dither in time by a maximum of one RO clock cycle period resulting in poor time gate uniformity. Resetting the phase relationship between the two clocks avoids such anomaly.

If the pulsed light source is driven by the VCSEL driver (through GATE_VCSEL) then this will not be an issue as what matters is the relative position of the time gate signals which are all derived from the same block. Alternatively, it is possible to route GATE_VCSEL out of the chip through the digital test MUX / pad to use it as a master synchronisation signal.

Finally, and as a secondary part of the gate generation block, two sets of balanced binary clock trees were integrated in order to drive the time gates through the imaging array odd and even columns. This allows for space division multiplexing of time gates A and B at the cost of effective image resolution.

The design was iteratively refined by resizing metal tracks and driver strengths such that at least a 1ns time gate can be propagated to all pixels with simulated skew across the array less than 150ps worst case scenario. In this case a standard rectangular time gate was opted for instead of the edge-to-edge technique described in Chapter 3 in order to simplify the gating logic design and to more efficiently use in-pixel resources for counter bits.

The clock tree inputs are also configurable by the MCU such that they can be used for global shutter mode (see Fig. 6.1.4), driven by the same time gate in parallel for time division multiplexing of time gates or to be driven by an external time gate through the digital test MUX / pad for characterisation purposes.

The layout of the gate generation block is shown in Figure 6.1.25 demonstrating the small footprint of such design occupying 164μm × 55μm excluding the backup VDD_OSC IO pad and the clock trees for time gate distribution (not shown).

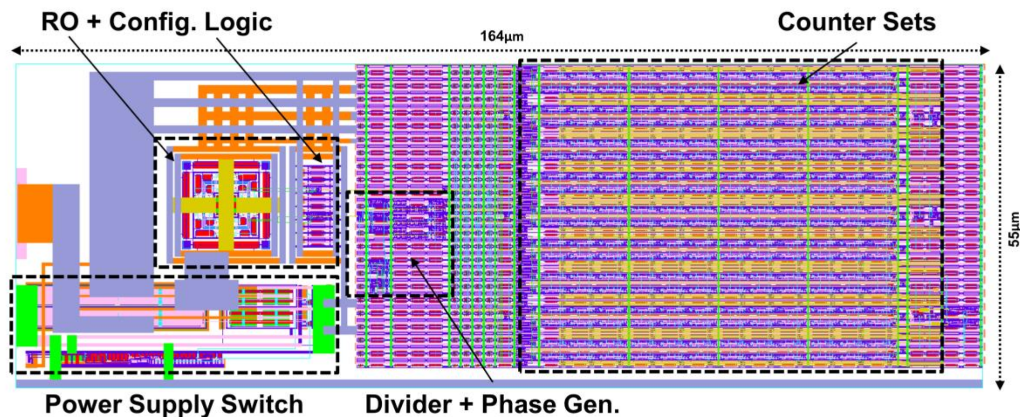


Figure 6.1.25. ENDOCAM gate generation logic layout occupying an area of 164μm × 55μm.

6.1.8. Layout Overview

The design process of ENDOCAM exceeded seven months of continuous work by the author with Dr. Almer working full time on the digital design for half of that period. Apart from technical prowess, communication between the designers is key to progress as the system is very complex and crucial

details can easily be missed or misinterpreted specially when converging the analogue and digital design flows into one implementation. Needed configuration bits, LSB and MSB bit orders, timing delays, signal sequences, reset signals polarity and voltage domains are all examples of things that can go amiss to name a few.

Layout planning is another area where attention to detail is of great importance. The floor plan of the chip was repeatedly thought out throughout the design process as different sub-blocks took shape. It is easier to regenerate layouts in the automated digital flow and the MCU unit was iteratively updated to meet area, pin location, metal layer accessibility, power strapping and substrate tapping requirements. Also, the vertical layout stack of the MCU and SRAM units had implications on timing constraints due to long signal tracks which were dealt with by the synthesis tools mainly by optimised buffering.

The layout of any IC, let alone one of this complexity, is an art form and demands a lot of meticulous work. Nevertheless and despite the ratio of design time spent doing it, it is not possible to easily describe in comparison to circuits but the bigger picture is what tells it all. Figures 6.1.26 and 6.1.27 show the final layout of the top and bottom tier chips of ENDOCAM with the main sub-systems labelled. Figure 6.1.28 shows a micrograph of the top tier backside of the fabricated stacked sensor.

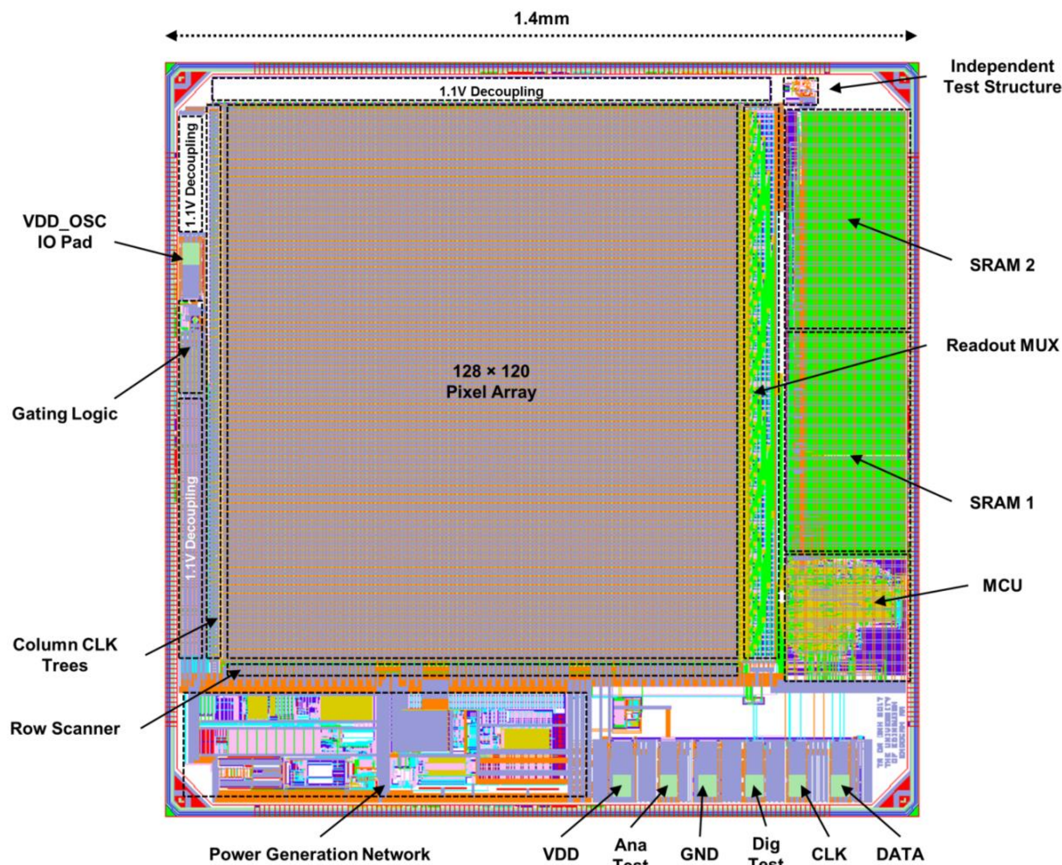


Figure 6.1.26. ENDOCAM 1.4mm × 1.4mm bottom tier 40nm IC.

Since some free area was available on the top tier IC, Prof. Henderson and the author took the chance to implement some SPAD test structures which were wired out to passive pads on the top tier. These structures are completely independent form the ENDOCAM SoC and have no impact on its operation. One particular trial consisting of a dual tier pixel visible in to top right corner of the array is described in Appendix 8.1.

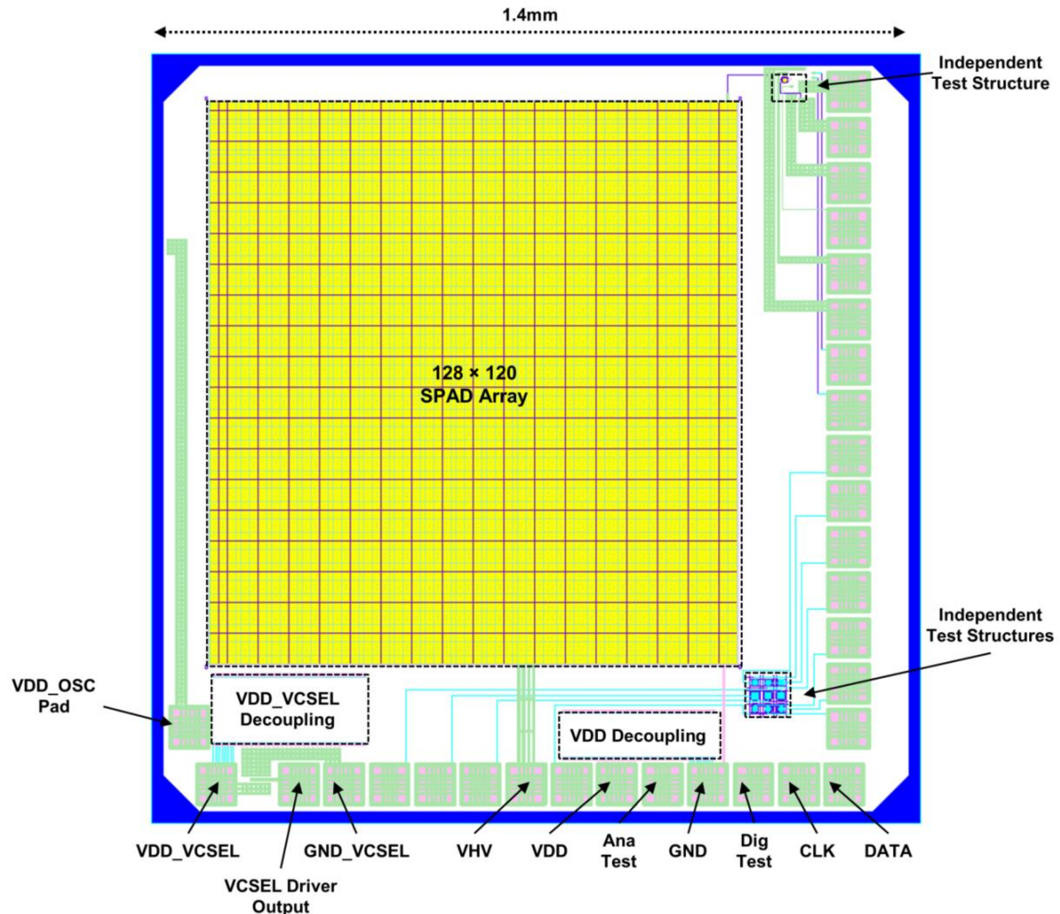


Figure 6.1.27. ENDOCAM 1.4mm × 1.4mm top tier BSI 90nm IC.

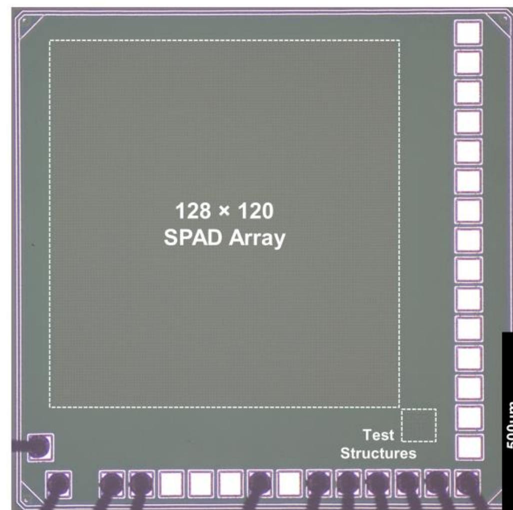


Figure 6.1.28. ENDOCAM micrograph showing the backside of the top tier IC.

6.1.9. Preliminary Bring-up Results

Unfortunate for the author, the chip fabrication and delivery turn around since submission was approximately eight months due to STMicroelectronics' multi-project wafer (MPW) scheduling which meant that there was hardly any chance of characterising the chip at the time of writing particularly with the PhD deadline approaching.

This section demonstrates preliminary plug and play results of the sensor showing that it is mostly functional. This is still considered an encouraging sign given the scale of engineering effort that went into the design and given the fact that it is expected to feed follow up projects within the CSS group where colleagues will be testing and utilising it further.

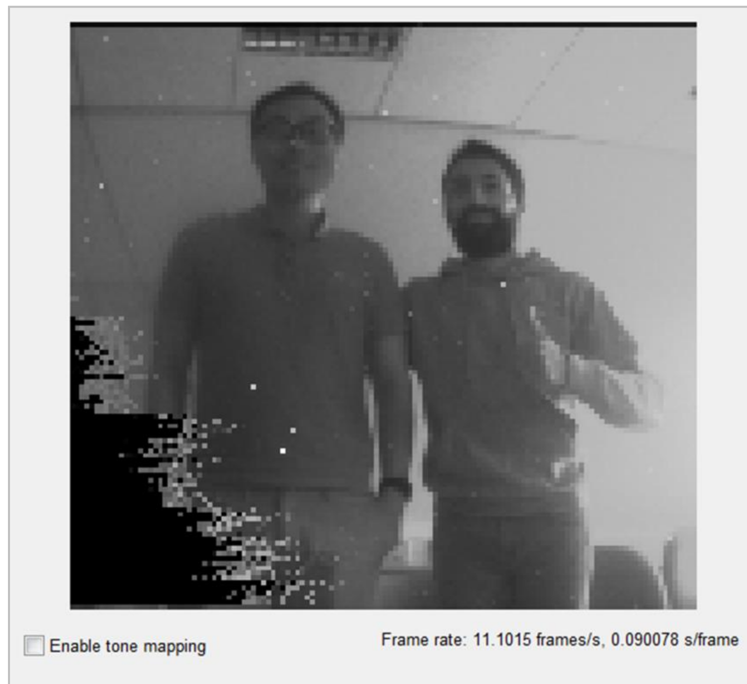


Figure 6.1.29. Live stream image from ENDOCAM sensor in rolling shutter mode at 11fps. The author (right) and colleague Hanning Mai (left).

Being able to observe such images immediately confirms the functionality of several key SoC blocks including:

1. Power generation network.
2. IO interface and bi-directional data pad.
3. Synthesised MCU and SRAM blocks.
4. Pixel array and SPAD array.
5. Imaging array addressing and readout circuits.

Despite the positive signs of life of the ENDOCAM sensor it is also obvious that there are some issues in the streamed images (Figure 6.1.29). The bottom left corner of the array seems to be reading out mostly zero values with visible banding of 20 rows at a time which fades away with higher row numbers. It is immediately known to the author the cause of this issue which is primarily a timing violation in the array addressing controls.

As demonstrated earlier in Figure 6.1.6, and once a row has been selected for readout, two row control signals are issued by the readout state machine in the MCU, Data Latch and Row Reset. Ideally Data Latch should be issued after allowing some settling time of pixel data on the column parallel bus before latching it into a temporary storage register until all bits are multiplexed and written into their corresponding SRAM locations.

Following the latch operation it is safe to reset the pixels in the selected row before releasing the row for subsequent integration. Yet, and from the dead zone of pixels observed in Figure 6.1.29 it appears that there is a race condition between the Data Latch and Row Reset signals.

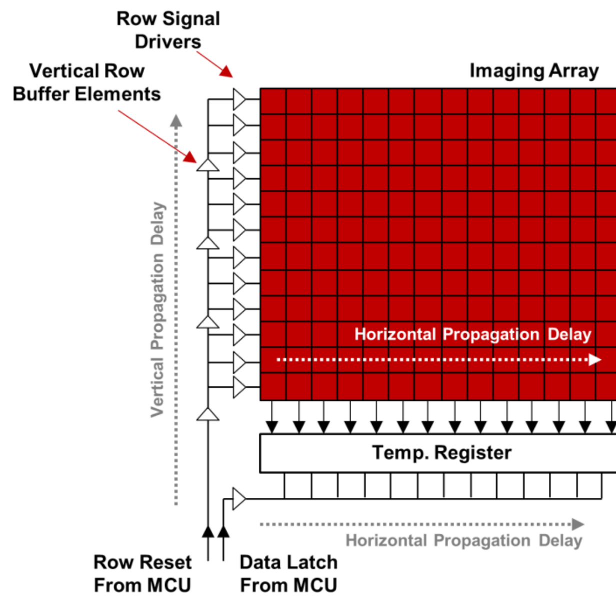


Figure 6.1.30. Data Latch and Row Reset signal propagation paths for ENDOCAM readout.

Both Data Latch and Row Reset signals traverse from the MCU to the bottom left corner of the imaging array where the Data Latch signals propagates horizontally across the register chain with some delay while the Row Reset traverses vertically and then horizontally with different delays depending on layout parasitic RCs and driver strengths. The issue observed can be described as follows:

1. Both Data Latch and Row Reset are issued simultaneously with no time delay between them and race towards the imaging array.
2. The Row Reset signal reaches the beginning of the first few rows earlier than the Data Latch signal resetting the first few pixels to zero and that is what is latched into the column register. As the Row Reset signal traverses horizontally across the selected row it becomes slower than the Data Latch signal and so the further away pixel values are correctly latched into the column register before the reset happens.
3. As higher rows are selected the Row Reset signal experiences further vertical propagation delay causing it to reach fewer pixels at the beginning of the row earlier than the Data Latch signal.

4. This is evident in the decreasing number of corrupt pixel values towards the beginning of the row in bands of 20 rows corresponding to the interval at which the Row Reset signal is re-buffered vertically in layout.
5. Eventually and for the top most rows of the array the vertical delays of Row Select signal are sufficient to overcome this race condition.
6. The groups of sparkling pixel values in between the zero (i.e. corrupt) and valid pixel values is due to the reset and latch happening simultaneously and so a random pixel value is registered mid transition.

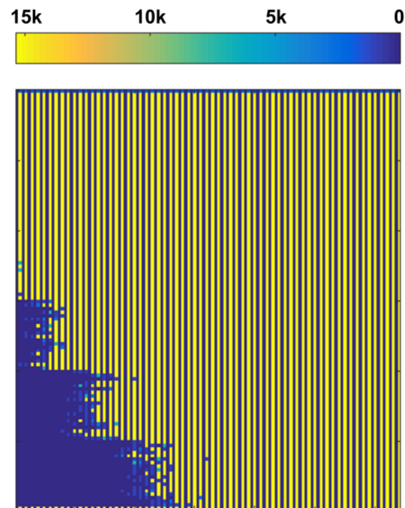
Figure 6.1.30 illustrates the propagation delays associated with both Row Reset and Data Latch signals. During the design stage the designers considered the possibility of such scenario happening and so Dr. Almer implemented a set of configurable registers to select a delay in number of system clock cycles between these and other signals issued by the MCU readout state machine. Unfortunately at the time of writing, and out of all configuration registers these settings seem to have no effect on the captured images.

The issue remains to be fully uncovered and colleagues at the CSS group are looking into the matter as part of the follow up efforts to identify if this is a silicon bug in register transfer level (RTL) or a problem in configuring the delay registers. In the meanwhile a temporary solution would be to reduce the sensor resolution to a cropped 90×90 region for proof of principle in practical applications.

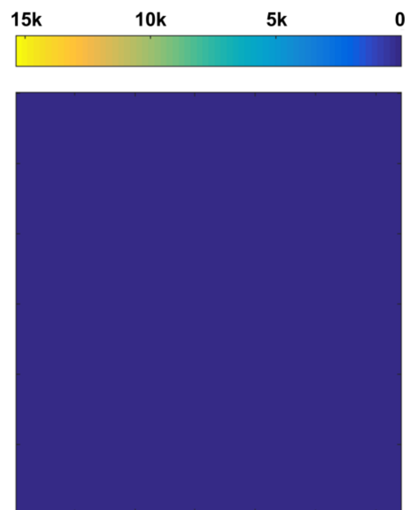
Another important block that has proven to be functional is the time gating logic. Figure 6.1.31 shows images captured under uncorrelated room ambient conditions with time gating features enabled. Figure 6.1.31(a) shows only odd columns enabled by setting the counters of GATE_B while GATE_A was switched off by keeping its counters at reset value.

Figure 6.1.31(b) shows an image with no photons captured despite both time gates of odd and even columns being configured due the ring oscillator being switched off by the MCU. While this is not a detailed characterisation of the time gating performance it shows that the following sub-blocks are functional:

1. Ring oscillator.
2. Time gate generation logic.
3. Clock trees.



(a)



(b)

Figure 6.1.31. ENDOCAM images captured with the on-chip time gate generation feature enabled. a) Only time gate for odd columns enabled. b) Both time gates of odd and even columns enabled but ring oscillator switched off resulting in no photons captured.

Finally, few other features were briefly tested through the analogue and digital test multiplexers and pads by probing signals using an oscilloscope such as:

1. Ability to program the width of time gates (GATE_VCSEL signal probed).
2. Ability to program the voltage regulator output (between 0.9V and 1.2V).
3. Ability to program VQuench MUX settings and to drive it from an external supply.
4. Ability to probe internal digital control signals and configuration bits issued by MCU.

6.2. Novel Configurable Array Architecture

Having presented a complete SoC architecture tailored for miniature time-resolved imaging arrays in the previous section, this section presents an alternative highly configurable architecture based around a miniature smart pixel facilitating multiple modes of operation.

As a side project during this PhD the author had the pleasure of working on the FlashTDC chip [157] (designed by former CSS group members Drs. Neale Dutton and Salvatore Gnecchi) including two tape-outs with fixes and modifications to original design, using it in visible light communications (VLC) application [5][6] and full optical and electrical bench characterisation [3][7].

The FlashTDC sensor boasts an impressive folded TDC architecture with direct on chip histogram generation and a throughput exceeding 10GS/s. Yet this optimised single point architecture aimed at time of flight applications clearly highlights the demanding silicon resources of full TCSPC systems and their restricted scalability beyond multi-point dimensions to image sensor formats.

Nevertheless, and as has been the trend in several architectures presented in literature, sharing of resources in a modular fashion between pixels [224] or at column level [175] allows for scaling TCSPC designs. For the presented sensor, hereby known as CORVETTE, a time-gated imaging array with configurable region of interest TCSPC operation inspired by the FlashTDC architecture is proposed. The key elements of this design include:

1. State of the art 3D-stacked BSI pixel with 6.48µm pitch.
2. High frame rate multi-bit QIS operation with time gating.
3. Dual diode mode for SPAD and photodiode operation for high dynamic range.
4. Region of interest TCSPC on-chip histogramming.

6.2.1. Pixel in SPAD Mode

The CORVETTE pixel is also a fully digital design in order to leverage the high integration density of the 40nm thin oxide logic for implementing several functions in-pixel. Figure 6.2.1 shows the pixel schematic diagram. Apart from the two thick oxide front end transistors MQ and MC all other devices are thin oxide and the whole pixel runs off a common 1.1V supply.

When operated in SPAD mode by setting VHV to reverse bias voltage beyond the SPAD's breakdown, transistor MQ acts as the usual quench and recharge device with tuneable gate voltage VQ and transistor MC acts as a clamp switch similar to [233] used to truncate any SPAD pulses above 1.1V. To keep signal levels compatibly with the digital logic bias voltage VCLMP is set such that VCLMP minus an NMOS threshold V_T is equal to 1.1V. Red waveforms in Figure 6.2.1 illustrate the front end signal profiles and polarities.

Figure 6.2.1. CORVETTE pixel schematic diagram.

SPAD pulses trigger a 3-bit ripple counter which is activated by a gate control logic identical to that implemented in ENDOCAM where ROW and COL signals define the exposure or time gate characteristics. Overflow control is also implemented in order to freeze the counter once it reached a value of six counts (i.e. binary value of 110). Conventional column parallel readout is implemented with row select signals Sel and SelB making the counter bits available on the column bus.

The intentionally small bit depth of the pixel counter is meant for highly oversampled QIS operation and the chip readout consists of eight parallel output channels allowing for frame rates in excess of 10kfps for the 128 × 96 array resolution. Having a multi bit counter allows for exploring processing algorithms where the pixel photon threshold K can be adaptively selected between 1 to 6 photons on per pixel basis through moving window algorithms such as [288].

For this tape-out the minimum allowed hybrid-bond pitch by process was reduced to approximately 6μm and so the layout of the pixel was optimised to a pitch of 6.48μm surpassing the state of the art figures for SPAD image sensor pixels at the time of writing. Figure 6.2.2 shows a layout snapshot of the pixel layout design.

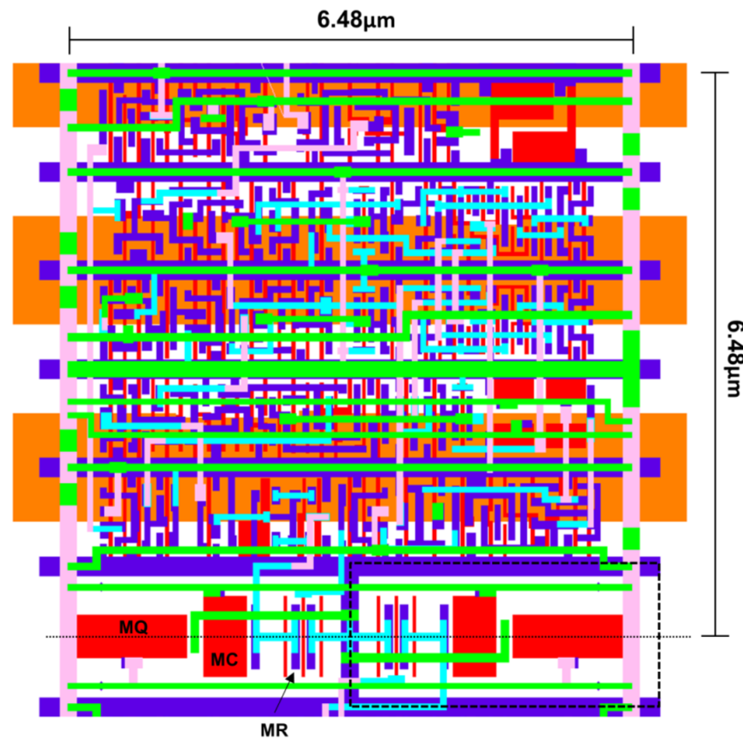


Figure 6.2.2. CORVETTE 6.48µm pixel layout. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3 and green is MT4. Higher metal layers are not shown for clarity. The dashed box encloses the front end transistors from the neighbouring pixel due to the interleaved layout methodology.

6.2.2. Pixel in Photodiode Mode

When operated in photodiode mode (PD), the idea is to bring VHV down to approximately 3V while globally switching the quench transistor MQ off by biasing VQ at 0V. In this mode additional circuit components are needed such as thin oxide reset transistor MR and the self-reset block in Figure 6.2.1. Initially the anode node would be set to zero and as light impinges on the reverse biased diode it integrates charge until the anode voltage exceeds the front end Schmitt inverter threshold causing the counter to trigger.

The motive behind operating the diode in PD mode is to increase dynamic range by merging the outputs of the two modes as first proposed by [284]. Due to the digital circuit design it is not possible to integrate charge in the traditional 3T-pixel way and so an alternative light to frequency approach is proposed. The same digital counters can count the number of cycles the photodiode integrates up to a threshold within an exposure time. For that to work the photodiode needs to be reset once it reach the threshold value and this feature is enabled by the self-reset block.

Dr. Istvan Gyongy, Prof. Robert Henderson and the author have conceived the idea in the early design stages of the pixel and it only came to their attention later after the submission of the design for

fabrication that such an idea was already implemented and presented by [243][286] and so they fully acknowledge that Ouh et.al. were the first to propose and explore such an approach.

Nevertheless, this implementation differs in two main aspects: first the compact implementation using digital logic and no analogue integrators or comparators achieving a remarkable 6.48μm pixel pitch, and second combining both PD integration and high frame oversampling (analogous to QIS) in an attempt to further increase the dynamic range.

The validity of this approach is to be proven but it is thought an extended HDR response similar to that of QIS can be obtained where the photon threshold K in this case does not represent the detection of a discrete number of photons but rather the number of integration cycles of the photodiode.

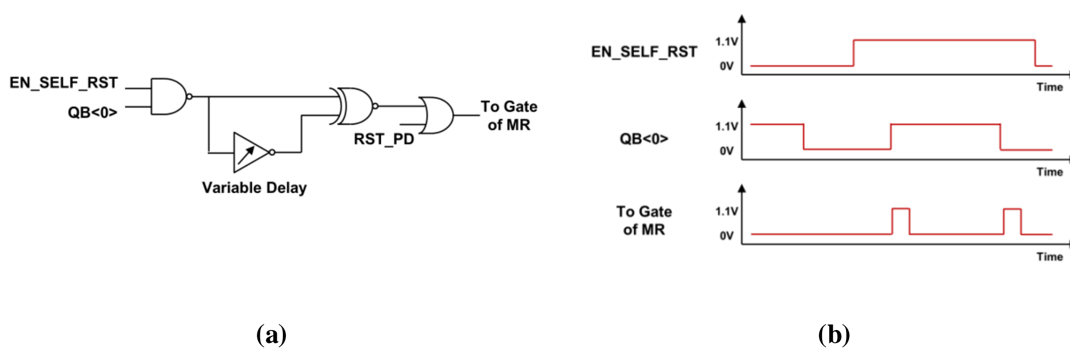


Figure 6.2.3. Self-reset circuit block of CORVETTE pixel used in photodiode mode. a) Schematic diagram. b) Example timing diagram or generated photodiode reset signals.

Figure 6.2.3 shows the schematic diagram of the self-reset circuit needed in PD mode alongside an example timing diagram. To detect that the diode integrated to the set threshold, the state transition of any of the digital nodes in the chain can be used and here the LSB of the counter (QB<0>) was selected to ensure that the count has been registered before resetting the diode. But since this is a toggle bit then events can be represented by either a high to low or low to high transitions and so the self-reset block accounts for that. A global EN_SELF_RST setting is used to enable the self-reset block and a voltage controlled variable delay inverter is used to tune the reset pulse duration.

Both NMOS and PMOS devices are tuneable to create equal reset pulses for both transition cases. Moreover, a global RST_PD signal is used to force reset the photodiode once the VHV and VQ voltages have been set to guarantee a zero anode voltage at start-up conditions.

To optimise the pixel performance in PD mode a Schmitt inverter front end was designed for two reasons. First, to have sharp switching characteristics to avoid a situation where the photodiode integrates slowly causing a near threshold input level that sets the inverter mid-rail with free flowing current and to be less susceptible to input noise. Second, to alter the inverter threshold by transistor

sizing such that it is increased from the typical half supply 0.55V to 0.68V gaining an additional 130mV of integration dynamic range in PD mode.

Moreover Monte Carlo simulations suggest that the Schmitt inverter has a 10mV tighter standard deviation DC characteristics compared to a standard inverter which translates to better pixel to pixel variability. The reset transistor MR was deliberately placed after the thick oxide clamp transistor MC so it can be a thin oxide device driven by 1.1V logic and its width was increased as allowed by layout to improve its pull-down strength.

The MC transistor was also chosen as wide as possible to reduce its on-resistance such that the photodiode resets to a value as close to 0V as possible. Its length was left to the minimum allowed by process sacrificing V_T matching across pixels albeit with no significant impact on the front end circuit behaviour as all switching characteristics are dominated by the lower bound Schmitt inverter threshold which is below the 1.1V output of the clamp switch.

Figure 6.2.4 illustrates the timing of the CORVETTE pixel in PD mode.

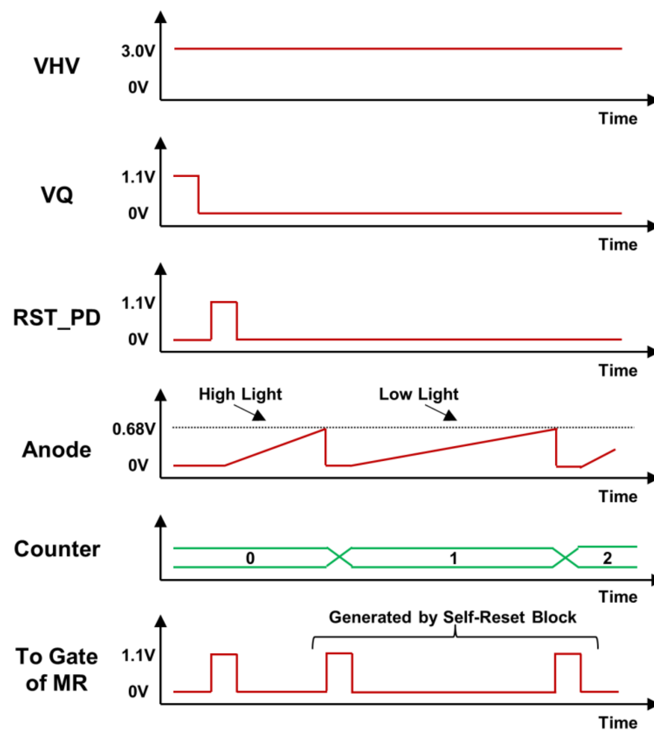


Figure 6.2.4. Timing diagram of CORVETTE pixel in photodiode mode.

6.2.3. Region of Interest Operation

As mentioned earlier, the CORVETTE image sensor is designed to operate in oversampled mode where two dimensional images are generated at 10kfps or higher. It is possible to boost the frame rate

further by selectively enabling rows of interest when performing the readout, a feature commonly found in other sensors designed in the CSS group such as the SPCImager [65] and the MegaFrame32 [167].

Time gating capability is also implemented with time gates being generated by the time difference between rising edges of external signals input to the chip. Like ENDOCAM separate time gates for odd and even columns can be configured. Balanced column drivers broadcast the time gates to the array.

As the sensor operates as an imager, a unique programmable region on interest (ROI) feature is proposed. Since time gating allows for detecting fluorescence of biomedical samples with enough contrast or for estimating distance by means of indirect time of flight while trading off TCSPC accuracy for pixel simplicity, the proposed feature allows for performing single-point on-demand TCSPC with on-chip histogramming for particular regions of interest efficiently reusing resources and lifting the processing burden off the pixel array.

Referring back to the CORVETTE pixel schematics in Figure 6.2.1, the region of interest operation is enabled by programming a configuration bit into a 1-bit latch memory in-pixel. Programing the memories happens sequentially by writing a bit pattern into column wise registers and latching the settings into the row of pixels by pulsing CLK_Bit high. Writing logic high into the memory sets the in-pixel signal EN_XOR high.

At the pixel level, EN_XOR has three main functions:

1. Enables the SPAD pulses represented by the toggling of the counter LSB (Q<0>) to propagate through the column wise XOR tree used to combine pulse streams from all ROI enabled pixels in the column into a single output.
2. Overwrites any time gate or exposure windows governed by the gate generation logic in order to keep the counter (mainly LSB) flowing without interruption.
3. Block any reset signals from reaching the counter flip-flops in order to keep the counter (mainly LSB) flowing without interruption.

It is worth noting at this stage that the effect of the ROI operation is to select any pattern of pixel across the array to operate in a dSiPM like mode where they all act as a single point sensor within the imaging array at the cost of loss of spatial resolution. Drawing a comparison to the FlashTDC architecture whose array of SPADs operates as a single point by default, the ROI operation of CORVETTE allows for selectively moving the single point (or a group of points) across the focal plane.

The necessity of overwriting the reset, exposure or time gate signals for the ROI pixels stems from the fact that the rest of the imaging array continues to operate regularly outputting valid frames while the ROI pixels work as a dSiPM in parallel.

Also similar to the FlashTDC, the column XOR gates are embedded within the pixel and the routing is balanced in order to match propagation delays from all locations. A standard cell XOR gate with intermediate drive strength to compromise on area was used and simulations show a good matching between rising and falling edge times as photons are encoded in the toggling of the counter LSB. This combination of toggle and XOR pulse combining architecture utilises dual data rate encoding independent of the SPAD pulse dead-time to achieve 2× higher bandwidth under uncorrelated light conditions [159] compared to an OR tree one.

While there is a risk of event cancellation when using an XOR tree when two events arrive within a gate delay to the XOR (which is likely in correlated light conditions) resulting in non-linearity, this architecture was selected as it lends itself to the availability of a toggle signal in-pixel and avoids the overhead of designing a pulse shortener needed for an OR tree architecture. Figure 6.2.5 shows the XOR tree arrangement for the 128×96 imaging array where the column XOR trees feed a second horizontal XOR tree at the top of the array combining all column streams into a single output.

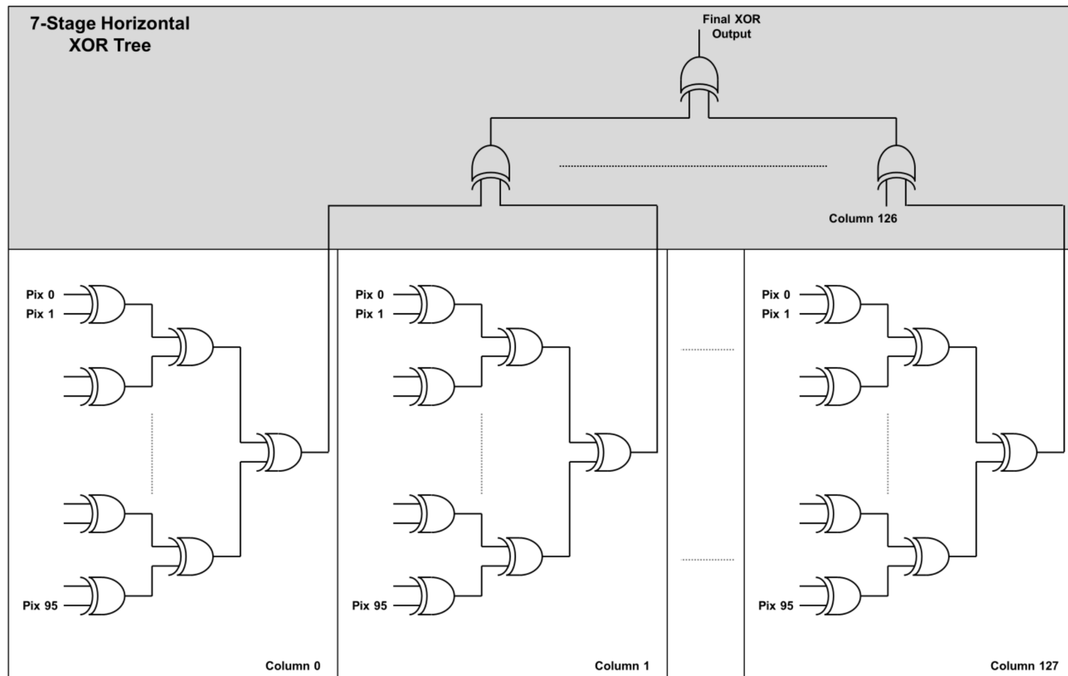


Figure 6.2.5. CORVETTE column and horizontal XOR tree structure.

6.2.4. TDC Architecture

The final XOR tree output feeds into an on-chip histogramming dual mode flash TDC module consisting of 128 bins with a 16-bit depth and an overflow control mechanism. The direct to histogram TDC architecture is based on [157] but differs from the original design in the way the front end edges which sample the XOR output and define the bin width are generated. This implementation offers two modes of operation:

1. Medium range where the front end sampling edges are generated by shifting a token through a 128 element shift register as described in [297][298] at 1GHz resulting in a bin width of 1ns and a temporal dynamic range of 128ns. This mode trades off temporal resolution for dynamic range and is suitable for applications such as mid-range LIDAR covering approximately a distance range of 19m. The 1GHz clock is generated on-chip by a PLL circuit provided by STMicroelectronics as an IP block which is configured by the author as required.
2. Short range where the front end sampling edges are generated by propagating a clock through 128 delay elements providing a bin width of ~170ps (from extracted simulations) and a temporal dynamic range of ~21ns. For design simplicity the delay line was built of standard delay cells from the clock logic library and was not locked to a feedback mechanism such as a delay locked loop (DLL) to control the delays across PVT corners, yet a calibration mechanism was implemented where a clock edge is fed in and can be probed out at the end of line in order to measure the total delay span. Such a temporal dynamic range is suitable for short distance ranging (~3m) and FLIM applications.

A subtle difference between the two modes is that in shift register mode the TDC has no dead-time as seen in [3] and continuously operates by regenerating a token every 128 clock cycles achieving a high throughput. While in delay line mode, the input clock period has to be greater than the expected dynamic range of 21ns to avoid aliasing and so there exists a TDC dead-time between the time of generating a sampling edge from the last delay element and the time a new clock edge is fed into the line. Nevertheless the direct to histogram function is maintained and repetition rates up to 40MHz are permitted which is in line with most lasers repetition frequencies.

Several other control features such as clock dividers, probing points and master / slave configurations are also implemented adding more flexibility to the design. For example, the 1GHz PLL clock can be divided to extend the temporal dynamic range further. Figure 6.2.6 shows the compact layout of the histogramming TDC module which in 40nm process occupies an overall area of 620μm × 126μm.

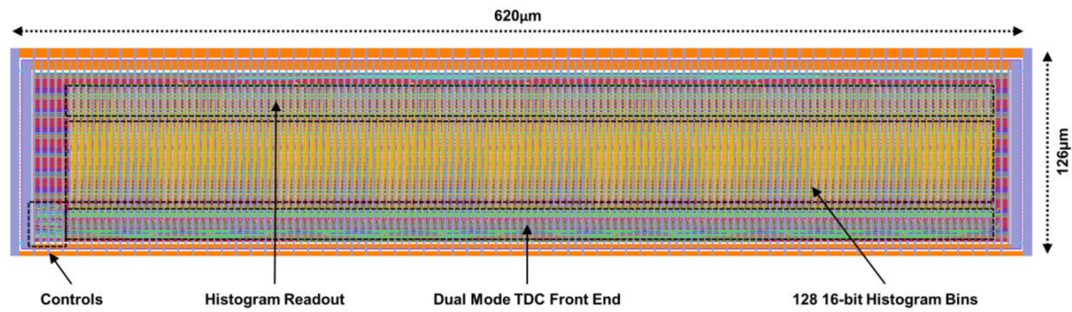


Figure 6.2.6. Layout of CORVETTE flash TDC module with on-chip histogramming.

6.2.5. System Level Overview

Other than the TDC module the XOR output feeds into a high speed counter that can be sampled at a rate of 10MS/s or higher which is useful in particular applications. The counter uses the 3-phase structure described for ENDOCAM's gate generation logic to ensure continuous sampling and counts both rising and falling edges of the XOR toggle signal.

The output channels of the chip can be configured for different readout modes. By default the imaging array frames are readout out but when a histogram has been accumulated or if the counter sampling mode is required the imaging array readout is interrupted and other data is multiplexed out. Several shift registers are used to configure the different modes and setting of the chip including readout rows of interest, dSiPM ROI and TDC configurations. Figure 6.2.7 shows a block diagram of the system architecture.

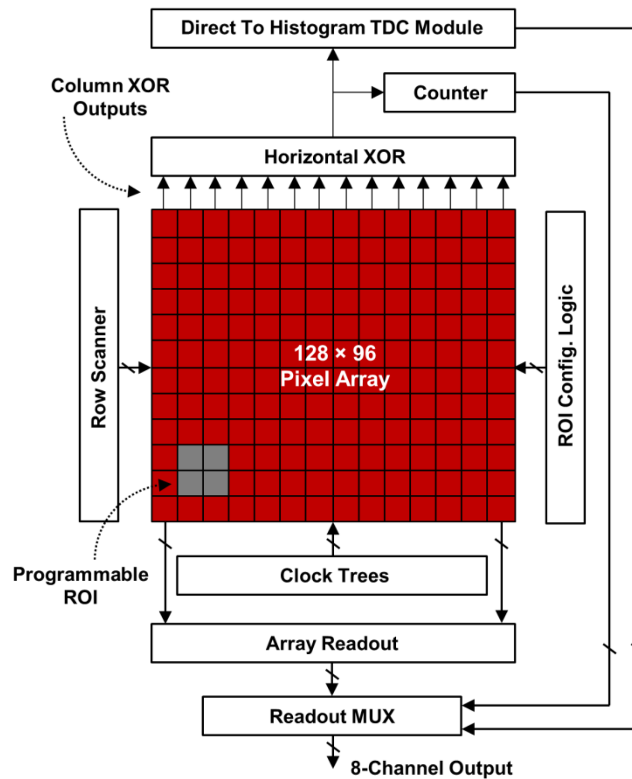


Figure 6.2.7. CORVETTE system block diagram.

Figure 6.2.8 shows the layout of the bottom tier 40nm IC with main block labelled. The top tier is not shown as it only contains a global well sharing array of 128 × 96 SPADs at 6.48μm pitch with a drawn fill factor of 35% and other small test arrays of the same device. The same PW / DNW SPAD structure and guard ring dimensions of ENDOCAM were used.

An experimental ultra-miniature 3.24μm single-bit pixel 32 × 30 QIS array was implemented in the top right corner of the chip as a look ahead towards megapixel time-resolved SPAD arrays. The pixel design is described in Appendix 8.2.

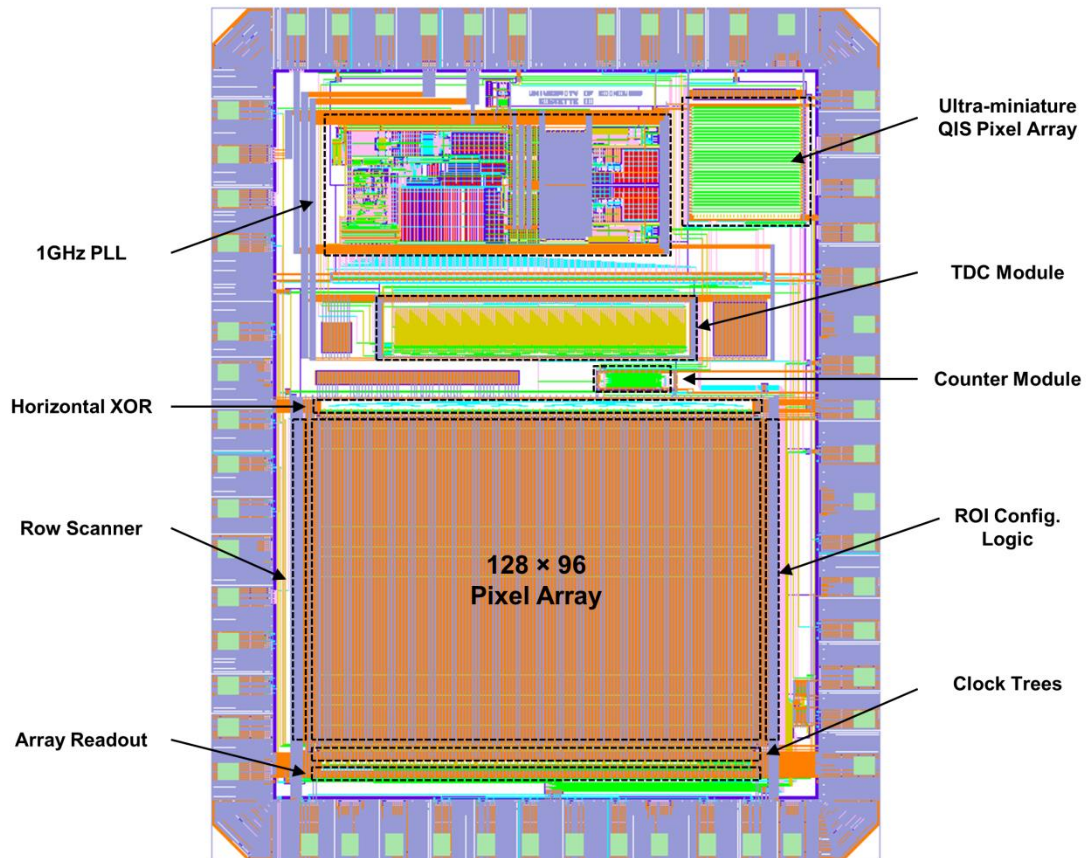


Figure 6.2.8. Layout of CORVETTE 1.7mm × 1.3mm 40nm bottom tier IC.

6.2.6. Potential Applications

The flexibility of the proposed design makes it suitable for a variety of applications such as but not restricted to the following as envisaged by the author:

1. On-demand TCSPC FLIM measurements. As the main target of this work is miniature time-resolved image sensors for biomedical applications, in-vivo FLIM for diagnostics is a relevant application. The sensor operated in time-gated imaging mode can generate 2D images at acceptable frame rates with enough contrast in temporal data to identify samples. Once the clinician identifies a region of interest in the scene, on-demand TCSPC data can be generated to more accurately measure lifetimes or to separate multi-exponential decays.
2. Target tracking and depth imaging. In ranging applications, it is possible to operate the sensor in QIS mode to generate 2-D images at high frame rates in which targets or objects of interest can be identified by on the fly algorithms and then LIDAR measurements can be performed for the regions of interest. The region can be adaptively allocated as the imager keeps acquiring new frames and both short and midrange direct time of flight (ToF) can be

performed by adapting the TDC mode. Another modality is to operate in indirect time of flight mode by time gating and perform on-demand direct ToF measurements when required.

3. Visible light communication [5][6]. The sensor can operate in imaging mode and once a target or more are identified, the ROI feature is enabled and VLC can be performed on per transmitter (user in a room for example) basis. Such a concept has been demonstrated for CISs [299]. Moreover, it is possible to track the different transmitters and adapt the ROI for better performance. Yet more importantly, it is possible to perform ToF measurements on ROI basis to identify the speed and trajectory of the transmitter and consequently adapt the sensor settings (i.e. VHV and VQuench) for best SNR. In a more ambitious scenario where the receiver (based on a CORVETTE sensor) can also communicate such information to the transmitter so it can adapt its transmission power accordingly.
4. LED characterisation. Recently, work by soon to be Dr. Hanning Mai at the University of Edinburgh has shown the ability of high speed SPAD image and point sensors in capturing transients of organic light emitting diode (OLED) micro displays [300]. In that work the SPC imager was used for imaging small arrays of OLEDs while the FlashTDC in counter mode was used for higher sampling rates albeit at a single point resolution. This meant that new alignment of the optical setup was needed for each set of measurements. The CORVETTE sensor offers a 2-in-1 solution with both features available resulting in a more efficient characterisation platform.
5. Low power or security applications. As pointed out by Dr. Neale Dutton, a low power mode can be implemented by operating a programmable pattern of pixels in ROI mode and using the counter to continuously sample the number of count in the scene. As soon as the number of counts changes above or below predefined thresholds, the sensor wakes up and operates in full imaging mode. Subsequently ToF measurements can be performed as explained above. This is similar in principle to dynamic vision albeit at an array level rather than per-pixel level. The concept of using thresholds has been demonstrated in the context of improving sensor frame rates by skipping rows or frames when not enough signal is detected [177]. To improve the bandwidth of the XOR tree in low power mode at high ambient conditions an additional feature was implemented where the toggle signal from the counter MSB is routed into the XOR tree instead of the LSB by means of a global Bit_SEL setting (see Figure 6.2.1). This provides an attenuation factor of 8 for the final XOR output rate.

6.3. Comparison to State of the Art and Discussion

Since ENDOCAM is the main focus of this research, a comparison to the state of the art image sensors aimed for biomedical imaging and endoscopy applications is made followed by a discussion of the CORVETTE configurable architecture and its potential use for miniature sensors.

6.3.1. ENDOCAM System on Chip

Table 6.3.1 lists the main pixel and system parameters for relevant SPAD and pinned photodiode (PPD) time-resolved image sensors with ones designed specifically for biomedical FLIM highlighted in grey. Only time-gated SPAD sensors are considered since TCSPC systems are discounted for miniaturisation purposes.

Assuming a minimum acceptable field of view of 100 × 100 pixels, an extrapolation of the imaging array area occupied by each of the sensors is made. Setting a target of 1mm² for a miniature design only the 8µm pixels of ENDOCAM, MIN3D and SPC Imager meet the criteria (green) which is not surprising given that these are the smallest reported for SPAD image sensors.

Comparing the SPAD pixels on basis of single frame photon counting capacity, the ENDOCAM sensor exceeds any other design with its 14-bit digital counter. Other sensors fall short either due to the limited integration capacity of analogue storage or the single-bit approach which requires heavy oversampling, an option deemed unsuitable for miniaturisation due to the data rate bottleneck as concluded in Chapter 4.

On the other hand, several PPD-based pixels meet the area constraints (green) due to the small pixel pitch of the simple APS configuration while offering a full well of more than 2ke-. Such pixels are pioneered by Shizuoka University led by professor Shoji Kawahito and are mainly designed in a 0.11µm CIS process.

While initially an attractive option for miniature time-resolved sensors, PPD-based designs do have disadvantages. Given the charge integrating APS pixels the analogue readout chain is susceptible to noise and analogue to digital conversion is necessary which adds a large area and power overhead to the design.

Moreover, current sensors report average frames rates of ~20fps which would be reduced when oversampling frames to accumulate enough signal or when acquiring multiple time gates. The latter is particularly true for single tap designs. A solution is to acquire multiple time gates simultaneously by 2-tap [301] or 4-tap [56] designs which come at the cost of increased pixel pitch which is why (for the exception of [55]) only single-tap designs meet the area requirement. The work in [55] provides a compromise at the cost of reduced FW compared to other PPD pixels meaning frame oversampling is still needed which will impact effective time-resolved frame rates.

As for digital SPAD sensors, they suffer from no readout noise and provide readily quantised output with the 8µm ENDOCAM pixel matching the average pitch of PPD-designs with even higher photon counting capacity. Nevertheless this is delivered with a single time gate also necessitating frame oversampling. Splitting the digital counter into two or more simultaneous bins as explored in

MINIC40 would drastically reduce the counting capacity per bin and demand even more oversampling.

Thus, the ENDOCAM sensor can overcome the oversampling of a single time gate versus frame rate trade-off by the on-chip frame summation architecture. For acquiring two time gates which is necessary for constructing a time-resolved frame space division multiplexing is possible by alternating time gates as explained in Section 6.1.7 at the cost of spatial resolution or by time division multiplexing by operating in SRAM ping-pong mode at the cost of effective frame rate. Once the frame acquisition interface between Matlab and the IC is optimised, time-resolved frame rates of 10fps should be feasible.

In light of the above discussion, one important observation is to be noted. All the presented sensors are not customised for autonomous operation needed by miniature sensors and so the pixel and system parameters would change once engineered for that purpose. All reported sensors need more than 25 connections to the outer world, in fact 25 is an understatement.

As it stands, the only sensor capable of delivering video rate time-resolved frames at small pixel pitch, high counting / integration capacity, high fill factor and a handful of interface pads is the ENDOCAM system on chip design (highlighted in red).

Other time-resolved CMOS sensors designed for FLIM such as the differential photodiode with in-pixel trans-conductance amplifier for improved IRF [302], the row level phase shift measurement by zero-crossing and phase-to-time conversion via a TDC [303], the CAPD with fast charge transport time [304] and the in-pixel capacitive trans-impedance amplifier (CTIA) for multi-cycle integration [305] were all omitted due their large pixel pitch of 50μm, 50μm, 30μm and 60μm respectively.

Ref.	First Author	Institute	Year First Published	CMOS Node (µm)	Techno.	Pitch (µm) [Sqrt of Area]	Native FillFactor (%)	Resolution	Pixel Capacity per Single Frame (Photons / e-)	Circuit Type	Area of 100 × 100 array (mm²)	Native Frame Rate (fps)	SoC + Processing	IO Interface	Simultaneous Time Gates / Taps	Single Time Gate / Tap Capacity (photons / e-)
SPAD Image Sensors [Ana: Analogue Counter, Dig: Digital Counter]																
[179]	Niclass et al.	EPFL	2008	0.35	FSI	85	0.05	60 × 48	256	Dig	72.25	46.3k	No	>25	2	256
[239]	Carrara et al.	EPFL	2009	0.35	FSI	30	3.14	32 × 32	1	Dig	9.00	376k	No	>25	1	1
[181][183]	Pancheri et al.	FBK	2011	0.35	FSI	25	20.8	32 × 32	150	Ana	6.25	180	No	>25	1	150
[185]	Maruyama et al.	TUDelft	2011	0.35	FSI	25	4.5	128 × 128	1	Dig	6.25	2.4k	No	>25	1	1
[180]	Bronzi et al.	Poli. di Milano	2014	0.35	FSI	150	3.14	64 × 32	512	Dig	225.00	100k	No	>25	2	512
[65][73]	Dutton et al.	Uni. Of Edinburgh	2014	0.13	FSI	8	26.8	320 × 240	1 (Digital Mode)	Ana	**0.64**	16k	No	>25	1	1
									400		**0.64**	7	No		1	400
[66]	Lee et al.	Cornell Uni.	2014	0.18	FSI	35	14.4	72 × 60	1024	Dig	12.25	n/a	No	>25	1	1024
[80]	Burri et al.	EPFL	2014	0.35	FSI	24	5	512 × 128	1	Dig	5.76	156k	No		1	1
[182][184]	Perenzoni et al.	FBK	2015	0.35	FSI	15	21	160 × 120	70	Ana	2.25	231	No	>25	1	70
									256 (Digital Mode)	Ana	2.25	486	No	>25	1	256
[186]	Gyongy et al.	Uni. Of Edinburgh	2016	0.13	FSI	16	61	256 × 256	1	Ana	2.56	100k	No	>25	1	1
[187]	Ulku et al.	EPFL	2017	0.18	FSI	16.38	13	512 × 512	1	Ana	2.68	156k	No		1	1
64	Ruokamo et al.	Uni. Of Oulu	2017	0.35	FSI	70.71	32	80 × 25	1	Dig	50.00	1000k	No	>25	1	1
[1]	Al Abbas et al.	Uni. Of Edinburgh	2016	0.065 CIS / 0.04	3D-Stacked BSI	7.83	45	128 × 120	4096	Dig	**0.61**	500	No	>25	2	64
[2]	Al Abbas et al.	Uni. Of Edinburgh	2017	0.04	FSI	8.25	39	96 × 40	4096	Dig	Not Possible	2k	No	>25	3	16
							66				Not Possible				3	16
ENDOCAM	Al Abbas et al.	Uni. Of Edinburgh	Unpublished	0.090 CIS / 0.04	3D-Stacked BSI	8	45	128 × 120	15360	Dig	**0.64**	>10	**Yes**	**5**	1	15360
PPD Active Pixel Image Sensors [DOM: Draining Only Modulation, LEFM: Lateral Electric Field Modulation]																
[53]	Yoon et al.	Shizuoka University	2009	0.18 CIS	FSI	7.5	n/a	256 × 256	n/a	LEFM	**0.56**	30	No	>25	1	n/a
[306]	Li et al.	Shizuoka University	2012	0.18 CIS	FSI	7.5	4.6	256 × 256	3800	DOM	**0.56**	15	No	>25	1	3800
[55][307]	Seo et al.	Shizuoka University	2015	0.11 CIS	FSI	7.9	16.7 (µLens)	512 × 256	2700	LEFM	**0.62**	12	No	>25	2	2700
[308]	Li et al.	Shizuoka University	2016	0.11 CIS	FSI	5.6	5.6	512 × 310	n/a	LEFM	**0.31**	n/a	No	>25	1	n/a
[69]	Chen et al.	Dartmouth College	2017	0.18 CIS	FSI	5	n/a	256 × 128	n/a	LEFM	**0.25**	n/a	No	>25	1	n/a
[301]	Seo et al.	Shizuoka University	2017	0.11 CIS	FSI	11.2	12.3	256 × 128	6930	LEFM	1.25	20	No	>25	2	6930
[56][248]	Seo et al.	Shizuoka University	2017	0.11 CIS	FSI	22.4	3.5 / 26.3 (µLens)	128 × 128	5500	LEFM	5.02	45	No	>25	4	5500

Table 6.3.1. Comparison table of time-resolved image sensors.

To assert the validity of SPADs as the detectors of choice for miniature time-resolved image sensors, Table 6.3.2 compares the intrinsic temporal response of the mentioned PPD-based pixels against the ENDOCAM SPAD. Other than readout noise and integration capacity, charge modulation pixels tend to be limited by their temporal response which would determine the shortest lifetimes or optical phenomena that could be observed.

Ref.	First Author	Institute	Year First Published	CMOS Node (µm)	Intrinsic Response (ps)	Wavelength (nm)	Readout Noise (e-)	Pixel Capacity per Single Frame (Photons / e-)	Device Type
[53]	Yoon et al.	Shizuoka University	2009	0.18 CIS	700	374	2.6	n/a	LEFM
[306]	Li et al.	Shizuoka University	2012	0.18 CIS	2000	374	2	3800	DOM
[55][307]	Seo et al.	Shizuoka University	2015	0.11 CIS	180	374	1.75	2700	LEFM
					220	472			
					250	635			
					370	851			
[308]	Li et al.	Shizuoka University	2016	0.11 CIS	150	374	n/a	n/a	LEFM
[69]	Chen et al.	Dartmouth College	2017	0.18 CIS	n/a	n/a	2.52	n/a	LEFM
[301]	Seo et al.	Shizuoka University	2017	0.11 CIS	460	635	1.2	6930	LEFM
[56][248]	Seo et al.	Shizuoka University	2017	0.11 CIS	170	472	0.85 (@ 30fps)	5500	LEFM
ENDOCAM	Al Abbas et al.	Uni. Of Edinburgh	Unpublished	3D-Stacked BSI	<100ps *	500 to 900	Negligible †	15360	SPAD

* Unpublished work. Measured jitter FWHM across spectral range using Fianium supercontuum laser for a standalone SPAD test pixel.
† Not taking into account other SPAD noise sources such as DCR, afterpulsing or crosstalk.

Table 6.3.2. Comparison of temporal response of PPD-based pixels and ENDOCAM's SPAD pixel.

While recent designs optimise the potential between the PPD and storage node for faster transport time in lateral electric field modulation (LEFM) pixels, this is very much dependent on the wavelength of the observed signal. A response of 150ps was reported in [308] at 374 nm, but a more comprehensive overview is provided in [307] showing how the response deteriorates to 370ps at 851nm due to the diffusion of photo-generated charge from deeper regions of silicon. Drain only modulation (DOM) pixels experience even worse charge transport times of 2ns at 374nm.

On the other hand, jitter spectroscopy measurements of a standalone ENDOCAM SPAD test pixel were carried out by Dr. Danial Chitnis at the University of Edinburgh showed a jitter response of less than 100ps FWHM across the 500nm to 900nm spectral range (to be published). This high temporal precision coupled with sub-nanosecond edge triggered time gating provides high performance time-resolved capability.

Finally, and to put the ENDOCAM SoC in the perspective of application, Table 6.3.3 gives a comparison against the commercially available endoscopy cameras mentioned in Chapter 1. While ENDOCAM has a larger area of 2mm² and a smaller number of pixels due to the bigger 8µm pitch, it is capable of video rate operation, provides equivalent full well capacity, higher dynamic range due to noiseless readout which can be boosted further by on-chip frame summation and can operate in both rolling and global shutter modes due to parasitic light insensitive digital storage.

More importantly, and in keeping with the aim of this research, ENDOCAM is the only miniature image sensor capable of time-resolved imaging with a fully integrated SoC 5-wire interface design.

	Awaiba NanEye [17]	OmniVision OV6946 [18]	OmniVision OV6948 [19]	Toshiba IK-CT2 [20]	ENDOCAM
Resolution	250 × 250	400 × 400	200 × 200	220 × 220	128 × 120
Pixel Pitch	3µm	1.75µm	1.75µm	n/a	8µm
Technology	FSI	0.11µm BSI	0.11µm BSI	BSI	3D-Stacked BSI
Frame Rate	42 to 55 fps	30 fps	30 fps	60 fps	>10 fps
Full Well	15ke-	n/a	n/a	n/a	15.36k photons
Dynamic Range	58 dB	65.8 dB	60.2 dB	n/a	83.7 dB
Shutter	Rolling	Rolling	Rolling	n/a	Rolling / Global
Connection Pins	4	4	4	n/a	5
Output Interface	LVDS	Analogue	Analogue	n/a	Digital
Sensor Dimensions	1mm × 1mm	0.95mm × 0.94mm	0.58mm × 0.58mm	0.7mm × 0.7mm	1.4mm × 1.4mm
Intensity Imaging	Yes	Yes	Yes	Yes	Yes
Time-Resolved Imaging	No	No	No	No	Yes

Table 6.3.3. Comparison of ENDOCAM to commercially available endoscopy image sensors.

6.3.2. CORVETTE Configurable Architecture

The highly configurable CORVETTE array pushes the boundaries of pixel miniaturisation beyond 8µm while delivering several interesting modes of operation including TCSPC on-chip histogram generation. But due to its oversampled nature it is not suitable yet for full miniature SoC implementations.

Other opportunities for the CORVETTE sensor lie in the additional intelligence that could be integrated on-chip, specifically in the bottom processing tier. Object recognition and tracking algorithms can automate the sensor's ROI operation and only deliver meaningful extracted data to the user. In biomedical imaging context, lifetime estimation algorithms can be integrated on-chip which could then determine the TCSPC ROI operation based on classification criteria.

Other technological advancements such as 3-tier 3D-stacking [201] would expand its capabilities by having a dedicated memory tier for data management and processing. Alternatively, a 3-tier design such as in [212] can open the door to modular processing [224] or cluster readout [256][295] where the top tier houses the SPADs, the second tier houses a shared TDC and the third tier houses a shared histogram generation or data extraction unit. Such modular design can reduce data rates or operate in different modes as necessary.

6.4. Summary and Conclusions

Two novel architectures for realising miniature SPAD 3D-stacked pixels and sensors were presented. The first encompasses an 8µm 14-bit depth pixel with on-chip data processing allowing for extending the dynamic range by noiseless frame summation and for achieving time-resolved imaging at video rates with a single output connection. The 2mm² 5-wire interface programmable sensor is the first fully integrated system on chip SPAD imaging array.

The second is a highly configurable array featuring state of the art 6.48µm pixel that can operate in oversampled time-gated mode while simultaneously performing region of interest TCSPC

measurements with on-chip histogram generation. Such sensor has a large scope of applications such as LIDAR, FLIM and VLC.

Both architectures are strictly permissible by 3D-stacking technology which is the key to realising miniature or configurable arrays with off-focal plane processing lifting the design constraints of the pixel. Consequently this results in smaller, more sensitivity and higher functionality pixels. Apart from optimised optical performance of top tier and lower power of dedicated bottom tier processes, 3D-stacking enables smarter designs with embedded intelligence.

7. Summary and Conclusions, Future Work and Outlook

This chapter summarises the work described in this thesis outlining the key objectives, the research directions followed to achieve them and the conclusions reached in retrospect. It also lists the main future tasks needed to complete this work such that it yields a high impact value. Finally it gives an outlook to future detector, pixel, system and process developments that would significantly improve the described sensors.

7.1. Thesis Summary and Conclusions

Over the past decade SPAD sensors have shown great promise in biomedical applications such as fluorescence resonance energy transfer (FRET), fluorescence lifetime imaging microscopy (FLIM) and super resolution microscopy. Their single photon sensitivity, high temporal resolution, integration in CMOS and amenability to high speed imaging architectures make them prospective replacements to other technologies such as PMTs and CCDs. Reference [309] provides an excellent review of the main SPAD sensors and the results they have achieved in the field of bio-photonics.

Therefore, this thesis sets out to explore the possibility of designing miniature time-resolved SPAD image sensors aimed for disposable endoscopes such that these systems not only deliver the typical intensity images, but also enable new in-vivo imaging modalities such as FLIM in order to provide clinicians with more informative diagnostic tools. Such an application requires the sensor to have a small form factor, minimal connectivity, conventional dynamic range and video rate output.

The literature survey shows that there are several challenges associated with designing time-resolved SPAD sensors including the detector integration in CMOS, pixel sensitivity and functionality trade-off due to the sophisticated embedded processing and the high data rates associated with complex time-resolved systems [310]. These challenges are further emphasised in the context of a miniature sensor where the silicon area is restricted and the data bandwidth is limited. While smart layout, pixel and system architectures can be exploited, the advent of advanced CMOS technologies such as 3D-stacking opens the door to new design possibilities and higher built-in intelligence [311].

To address these challenges, a miniature FSI sensor (MINIC40) in an advanced 40nm node was designed. This 1mm² sensor leveraged the high integration density and fine design rules of the advanced process to expand the global well sharing SPAD layout technique into a 96×40 pixel array achieving high fill factors up to 66% at a pitch of 8.25µm. The 12-bit digital pixel allowed for shot noise limited photon counting and parallel multi-gate time-resolved capability.

Nevertheless such a sensor suffered from two drawbacks. First, and this was supported by a feasibility study, global well sharing is an impractical approach due to its scalability limitation along the y-axis owing to routing complexity, crosstalk potential and added parasitics to the SPAD moving node. This restricts the attainable field of view and the situation does not significantly improve by adopting even more advanced CMOS nodes. Second, characterisation results showed the negative impact of this layout configuration on photo-response non-uniformity (PRNU) further detracting from its appeal.

While the MINIC40 sensor attempted to achieve high fill factor (i.e. sensitivity) and photon counting capacity (i.e. dynamic range) simultaneously, an alternative single-bit memory approach has been explored in the literature which trades-off in-pixel counting capacity for high sensitivity, while still incorporating a normally scalable pixel configuration. In these oversampled binary sensors, off-chip summation is necessary to reconstruct a greyscale image resulting in high data rates.

The concept of expanding the DR of such single-bit architectures was explored further using the aforementioned FSI sensor which enabled parallel triple-exposure settings achieving a DR in excess of 100dB and partial in-pixel bit plane summation providing 3.75× compression in data rates. Yet it was concluded that this model was not applicable to miniature sensor designs for two reasons. First, the associated data rates of oversampling especially for triple exposures are incompatible with the limited bandwidth of the single data channel of a miniature sensor. Second, if the in-pixel data compression scheme is to be employed, it would require multi-bit counters negating the motive behind single-bit pixels in the first place.

Consequently, another sensor (MINI3D) was designed implementing a similar 12-bit counter pixel but in a 3D-stacked technology where the SPAD and the processing electronics are integrated in separate tiers. With a direct 1-to-1 hybrid-bond connection between each SPAD and its circuitry in the 128 × 120 array, the scalability restrictions of MINIC40 were circumvented and characterisation results showed less than 2% PRNU.

Characterisation results of the 45% fill factor top tier backside illuminated SPAD revealed a change in the PDP response of the device compared to its FSI counterpart. Due to the flipped substrate isolated structure the spectral response in the blue region is heavily attenuated and the overall profile exhibited a shift towards the red region enhancing the detection efficiency at NIR wavelengths. This has implications on the target application where FLIM emission at short wavelengths will not be detected.

Despite that, 3D-stacking allows for embedding more functionality on-chip while maintaining a small sensor form factor due to the fact that the SPAD array is completely lifted to the top tier allowing for a small pixel pitch of 7.83µm in this case. Therefore for a resolution of 128 × 120 and a target IC area of 2mm², the imaging array would occupy roughly 1mm² leaving an equivalent area for system level processing capability.

This opportunity motivated the design of a third sensor (ENDOCAM) which is a fully integrated system on chip architecture with embedded power management circuits, time gate generation logic and a digitally synthesised micro-control unit. The programmable 5-wire interface 2mm² sensor utilises two SRAM memory banks to noiselessly sum frames on-chip enhancing the DR and mediating data rates. Preliminary bring-up results show the sensor is functional and streaming live images at 11fps under nominal operating conditions. Compared to other works in the literature, this is the first demonstration of a complete SoC SPAD image sensor at such a miniature dimension. Compared to commercially available endoscopy CISs, this is the only sensor which enables time-resolved capability.

Finally an exploratory highly configurable architecture (CORVETTE) which is capable of simultaneous time-gated oversampled multi-bit imaging and programmable region of interest TCSPC on-chip histogram generation was proposed. The 3D-stacked BSI 128 × 96 sensor comprises a state of the art 6.48μm pixel with dual photodiode and SPAD modes of operation for enhanced DR. While not directly applicable for miniature sensors this architecture offers the opportunity to explore a wide range of applications such as FLIM, VLC, short to mid-range LIDAR and bit plane processing algorithms.

In conclusion, 3D-stacking is a key technological step to realising miniature highly intelligent sensors with many built-in capabilities that would otherwise prove to be difficult in standard planar implementations. Apart from enabling high fill factor scalable arrays, 3D-stacking allows for optimising each of the tiers for best performance were an imaging specific process can be used for the top IC and a low power high density logic process can be used for the bottom one.

7.2. Future Work

The sensors conceived in this thesis require a lot of work to be done for them to be applicable in real endoscopy applications. The ENDOCAM sensor has shown promising signs of functionality but is in its early days of testing. The CORVETTE sensor has been submitted for fabrication and it will be months before it captures its first images.

7.2.1. ENDOCAM

Four main strands of work need to be pursued for ENDOCAM to reach its full potential. First, the sensor requires in-depth debugging of its different modes of operation. Initial images show a dead zone of pixels attributed to timing violations. While delay register settings have been integrated on-chip, they are currently ineffective and have no influence on the array timing. This feature has to be simulated in Verilog to understand if it is an implementation bug by design or a synthesis bug by the digital tools. Similarly global shutter mode timing and exposure control needs to be investigated.

Second, complete characterisation of the sensor's imaging capabilities needs to be done to demonstrate the DR enhancement although initial images are produced at a default setting of 2-frame summation on-chip. The time gate generation logic also needs to be tested to characterise the minimum possible time gate FWHM, non-uniformity and propagation skews across the array. The same applies to the SPAD array to measure its DCR, PDP, uniformity and crosstalk.

Third, the MATLAB interface needs to be optimised such that it does not limit the sensor's output frame rate. Due to the handshaking and configuration routines between the sensor and the MATLAB interface, any delays or dead-times in the process would influence the speed of acquisition. Additionally, the current firmware driving the sensor operates at either 12.5MHz or 25MHz systems clocks (images have been streamed at both settings), and digital synthesis tools suggest the compiled control unit can operate up to 50MHz. Any gains in system clock frequency would benefit the acquired frame rates.

Finally, the biggest achievement of this sensor is its miniature form and so integrating it on a mini printed circuit board (PCB), packaging it, assembling it on an endoscope tip, wiring it up, coupling it to a synchronised light source and demonstrating it in time-resolved mode with real-time frames would be the ultimate pinnacle. Such a project would require the efforts of experienced hardware and application engineers.

7.2.2. CORVETTE

Similar to the ENDOCAM sensor, the CORVETTE sensor would require complete characterisation but the more interesting goal is to demonstrate it in various applications. For biomedical imaging, acquiring time-gated frames to construct a wide field time FLIM image and then comparing the estimated lifetime accuracy to that acquired through ROI TCSPC mode would prove the validity of this architecture. More importantly, the ability to resolve multi-exponential decays in contrast to an average lifetime approximation is particularly powerful.

Developing bit plane processing algorithms with adaptive bit depth similar to [288] with the added feature of adapting the pixel photon threshold number K for improved dynamic range (see Section 4.2) due to the multi-bit (3-bit) pixel output would be a future research direction. This can be coupled with other object tracking and high speed image reconstruction algorithms as in [8].

Such processing capability will benefit the sensor in VLC or LIDAR modes where image processing algorithms can automatically reconfigure the sensor's ROI as necessary. In VLC mode the transmitter can be identified and its depth information can be used to adapt the receiver's (CORVETTE) detector bias settings for the best SNR. In LIDAR mode, tracking moving objects across the field of view and reconfiguring the sensor's ROI to measure their distance to the sensor as opposed to the whole scene would also be interesting. Linking this configurable sensor to smart image processing algorithms is also a future research direction.

7.3. Future Outlook

Based on the lessons learnt from the sensors designed throughout this research, an outlook on future developments in terms of detector, pixel, system architectures and process is projected. While mainly focused on the miniaturisation of time-resolved SPAD sensors these developments are also applicable to other end uses.

7.3.1. SPAD Device

As discussed in Sections 5.3.1 and 5.3.2, the PDP of all the 3D-stacked CMOS SPADs is cut-off in the blue region of the spectrum mainly due to the substrate isolated multiplication junctions, therefore a future SPAD device should be designed to deliver a wide range response. This can be achieved by a non-substrate isolated SPAD structure and by optimising the top tier backside thickness.

In order not to reduce the PDP in the NIR region, the junction of such a SPAD needs to be as deep as possible with respect to the backside surface, yet if this is the case short wavelength photons absorbed near the surface will have to diffuse towards the multiplication junction. This will have a negative impact on jitter performance. To overcome that, the SPAD has to be fully depleted such that any absorbed photon no matter the wavelength will exhibit drift transients that do not suffer from the diffusion timing uncertainty.

From a miniaturisation point of view, the SPAD should preferably be biased with a positive VHV such that the generated pulses are compatible with CMOS levels without the need for passive quench resistors or coupling capacitances, although these components can be easily integrated over the BSI active area on the top tier. Also, an ideal SPAD would be contained within passivated DTI walls for low DCR and to prevent optical and electrical crosstalk from densely packed neighbouring pixels.

The DTI walls though might create a problem for a small fully depleted device where the high voltage contacts that bias the substrate material are located between the SPAD's guard ring and the DTI walls. The depletion of the junction might pinch-off the substrate bias points and cause the SPAD to malfunction. Hence the substrate should preferably be biased by backside contacts connected to a backside metallisation grid. Figure 7.3.1 illustrates the envisioned ideal SPAD.

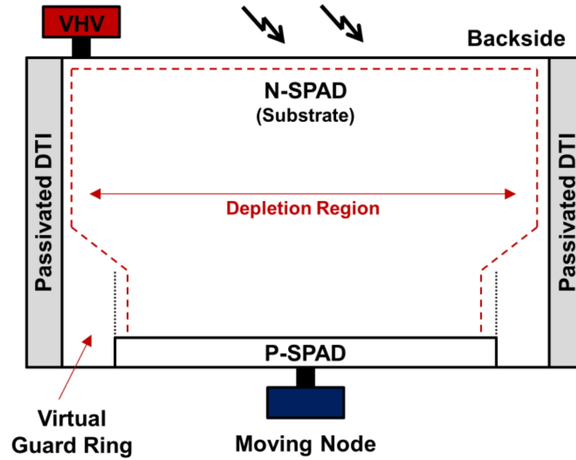


Figure 7.3.1. Illustration of the ideal fully depleted DTI bound SPAD structure.

7.3.2. Mixed Signal Pixel

In order to shrink 3D-stacked pixels further below 8µm, two options are available, either resorting to a more advanced bottom tier node with higher integration density such that a reasonable counter bit-depth can be obtained or resorting to the compact storage in the charge domain (i.e. an analogue pixel). As discussed in Section 2.2.3, analogue counting pixels suffer from non-uniformity, accumulation and readout noise and inaccurate level sensitive gating.

To overcome these limitations, a mixed signal pixel with a digital front end and a digital counter for LSBs and an analogue MSB storage (Figure 7.3.2) can solve most of these issues. Such a pixel can benefit from edge sensitive gating and noiseless LSB counting. An experimental layout trial in 40nm showed that such a pixel is possible at ~6.5µm pitch.

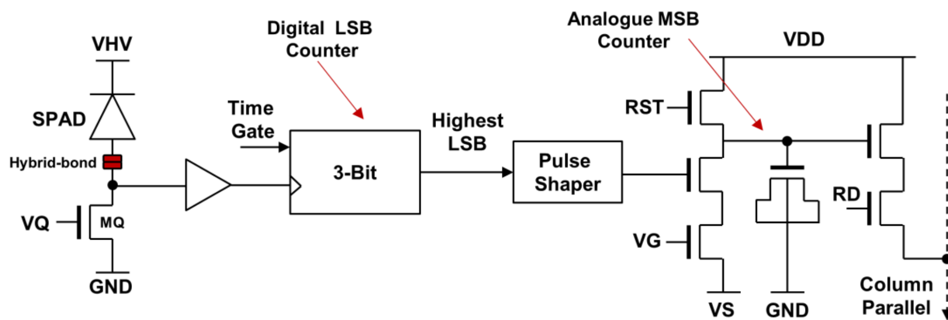


Figure 7.3.2. Mixed signal photon counting pixel with digital front end and LSB counter and analogue MSB counter based on CTA pixel in [65].

Assuming a 3-bit digital counter, whenever 8 photons are captured the counter rolls over back to zeros while triggering an analogue MSB count. The duration of this trigger can be controlled by a pulse shaper circuit and thus the analogue counter is decremented with a controlled signal in width and voltage height independent from the SPAD pulses which improves the step uniformity. Moreover, the MSB counter is triggered once every 8 photons thus reducing the accumulation noise per photon due to the less frequent switching.

Assuming an output voltage swing for the analogue MSB counter of 1.5V and a step size of 15mV, then an MSB capacity of 100 counts can be registered. Yet this corresponds to 8× photons per analogue count and thus a photon counting capacity of approximately 9.5-bits. The digital LSBs can also be read out independently in oversampled digital mode and the analogue counter supply and readout circuitry can be switched off for power saving.

Nevertheless some accumulation and readout noise contributions are inevitable which means that an error in the MSB conversion results in an error of 8 photons in count. For that, clever self-referenced ADC architectures such as that described in [184] are necessary and possible due to the flexible digital front end which allows for triggering the counter from a row / column signal.

To allow for DR extension by oversampling, faster conversion speeds can be obtained by implementing cluster parallel ADCs [312][313] as opposed to the conventional column parallel architecture. Although for SPAD image sensors, and due to the 1-to-1 relationship between the top tier detector and the bottom tier circuitry, this can be a challenging in layout and so another processing tier (3rd tier) as in [202][212] might be necessary especially if the sensor form factor is to be kept small.

7.3.3. More Vertical Integration

To maintain a small form factor and high performance, more vertical integration is anticipated for realising intelligent SPAD image sensors. Due to the size of the individual detectors (or SiPMs) and the complex processing / storage circuitry, SPAD sensors rely on direct connectivity between the different pixel components as opposed to CISs where only the column parallel output needs to be connected to the bottom processing element [200] or where a group of pixels can share an ADC block [295] since the charge storage takes place in the top imaging tier itself.

This implies that if any additional system level processing is to be implemented, it either has to occupy additional area on the bottom tier which is not matched to top tier diodes (ENDOCAM) or has to be integrated on another dedicated tier altogether. The work by MIT in [212] was the first demonstration of a 3-tier SPAD sensor splitting the design into SPAD, pixel front end and pixel backend tiers. Sony recently demonstrated the first commercially available 3-tier CIS with dedicated logic and memory tiers [202].

Building the system in a vertically integrated modular fashion which enables mix and match choices between different functionality and pixel properties would be a future design methodology for realising flexible and computationally powerful sensors.

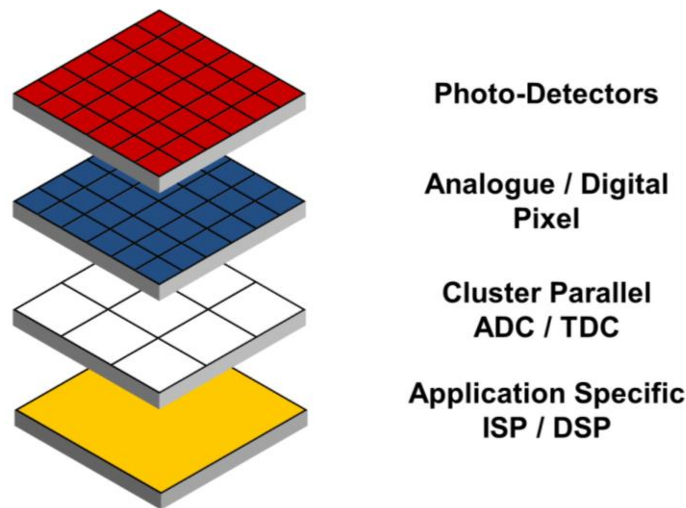


Figure 7.3.3. Illustration of vertically integrated modular sensor with mix and match tier options.

For the 4-tier example in Figure 7.3.3, the top tier which contains only SPAD devices may be implemented in CMOS or non-CMOS processes. The pixel tier underneath can be switched between analogue or digital front ends depending on the application. The third tier could break the 1-to-1 dependency and implement cluster parallel ADC / TDC architectures while the bottom most tier can be alternated between ISP (object recognition) / DSP (machine learning) capability. Such a design can easily be adopted at the wafer level assembly stage.

The more processing done on-chip, the less data has to be transferred to the outer world and so miniature systems with limited bandwidth can benefit from such a development. Other readout standards such as low voltage differential signalling (LVDS) can also be integrated instead of the conventional IO interfaces now that the additional area overhead of reference voltage and clock generators can be vertically accommodated. Moreover, architectures like CORVETTE can benefit from the integration of the 1GHz PLL block on a different tier further reducing the sensor's area.

7.3.4. Process Additions

Two process additions would be very beneficial for miniature sensors: TSV bump connectivity through the backside of the bottom tier IC and high capacitance MOM structures on the top tier IC.

As it stands, a sensor such as ENDOCAM is connected to the outer word through backside pads on the top tier which means that it has to be wire bonded to a mini PCB and then packaged increasing the

footprint of the overall module. If TSV bump connectivity through the backside of the bottom tier is available, then the sensor can be directly bump bonded onto a PCB of equivalent footprint resulting in an overall small form factor module.

Also, in an image sensor such as ENDOCAM, the top tier SPAD array occupies an area equivalent to the corresponding pixel array on the bottom tier but the rest of the area overlaying other circuits such as the SRAM block is unutilised. Ideally this area can implement active circuitry such as a charge pump but future top tiers that are SPAD specific might not be suitable for active devices. Moreover, implementing active circuitry might not be possible due to the reduction in layer and metal masks in order to cut fabrication costs.

Despite that there will be at least one metal layer available on the top tier for routing purposes and so the spare area can be filled with decoupling capacitors for any voltage supply. Thus, future processes can benefit from dedicated MOM structures that allow for reliable high capacitance to maximise the decoupling efficiency and durability at high voltages (i.e. SPAD VHV). A similar thing can be said about MIM capacitors if two or more metal layers are available (See Figure 6.1.5).

8. Appendices

This supplementary chapter introduces a couple of pixel ideas that the author thought are intriguing. Free silicon area available on the ENDOCAM and CORVETTE chips was used to integrate basic versions of these pixels so they could be investigated and learned from.

8.1. Dual Tier SPAD Pixel

Inspired by the two-tier SPAD pixel for particle detection presented by Pancheri et al. [314], a test two-tier pixel was implemented in the top right corner of the ENDOCAM sensor. The pixel was conceived and designed by the author and Prof. Henderson and characterised by Dr. Danial Chitnis. Unlike the work in [314] where the bottom SPADs are shielded to avoid optical crosstalk and the top tier backside is very thick (280µm) blocking all incoming visible light, the bottom SPADs in the presented pixel are not shielded and the backside thickness of the top tier is only few microns making it the first demonstration of a two-tier imaging SPAD pixel.

Dual tier image sensors have been proposed before where Panasonic demonstrated such a sensor for both visible (top tier pixels) and NIR (bottom tier pixels) imaging in [315], while the theoretical work in [316] showed the possibility of using two tier pixels to enhance low light colour reproduction without the use of conventional colour filters. The purposes behind the implemented pixel described in this section are:

1. To investigate the PDP of the bottom tier SPAD to see if it improves at NIR wavelengths due to its deeper location down the stack.
2. To improve dynamic range by switching between the more sensitive top SPAD and the less sensitive bottom one.
3. To investigate interesting concepts such as angular response and vertical spatial coincidence detection.

8.1.1. Pixel Structure

Figure 8.1.1 shows the schematic diagram of the two-tier pixel trial which is completely independent from the ENDOCAM sensor and has its own supplies and controls. The front end circuitry is all implemented in the bottom 40nm tier with the standard passive quench and recharge transistor followed by a voltage controlled pulse shortener.

A MUX is used to select either the pulse of the top SPAD, bottom SPAD, an AND operation of both or a ground state. The output of the MUX feeds into a toggle flip-flop which drives a level shifter that relays the shifted toggle output to an optimised output buffer capable of driving an external circuit or an oscilloscope probe.

The toggle signal was used as the front end pulses are too short (<1ns) to propagate through the output chain. An FPGA with positive and negative edge counters was used for all photon counting measurements while a LeCroy WavePro oscilloscope was used for time correlated experiments.

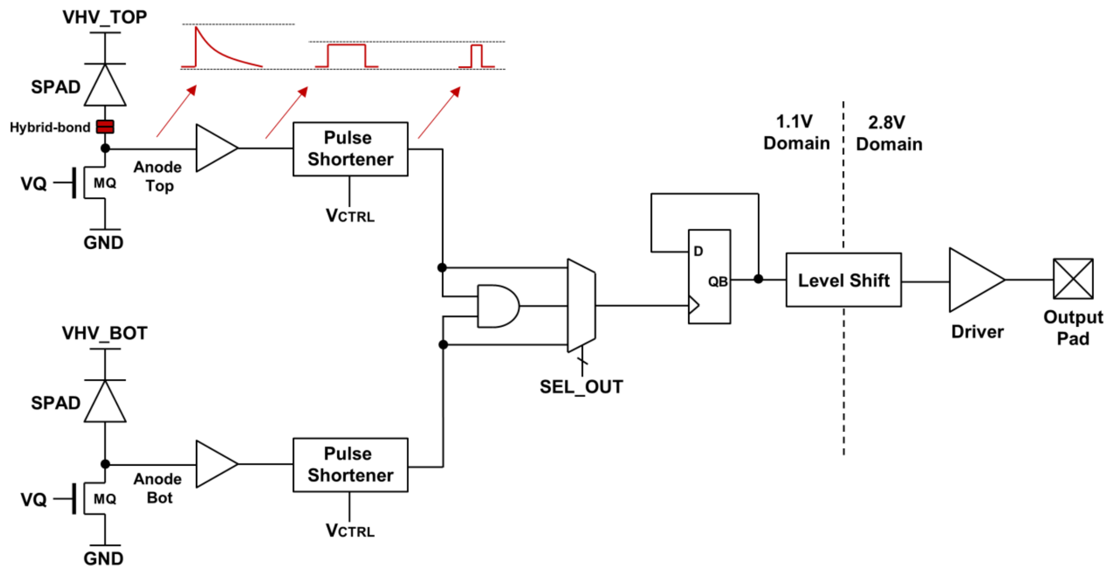


Figure 8.1.1. Dual-tier pixel schematic. All circuits are implemented in the 40nm bottom tier. Red waveforms show signal polarity, level and profile.

Figure 8.1.2 illustrates the layout configuration of the pixel. The top SPAD in the 90nm imaging process is a standalone version of ENDOCAM's 28.8μm² active area PW / DNW structure which makes the characterisation results representative of the main array expected performance. A SPAD with similar dimensions and structure was implemented directly underneath in the bottom tier but without rounded corners to avoid breaking design rules as a standard 40nm process was used and not the SPAD specific one presented in [114]. Therefore, the bottom tier SPAD parameters (breakdown voltage, DCR, etc.) are not representative of ST's industrial offering. Nevertheless it is acceptable for proof of concept.

Care was taken to ensure that the bottom SPAD is not obstructed by any metallisation due to the top tier dummies or hybrid-bond sites or due to the dummy patterns of the bottom tier. In fact, metal dummies were utilised to create isolation walls around the structure in an attempt to limit the incoming light only to that entering the top SPAD's aperture. The top SPAD moving node (anode) was connected to the bottom tier circuitry through a single hybrid-bond connection.

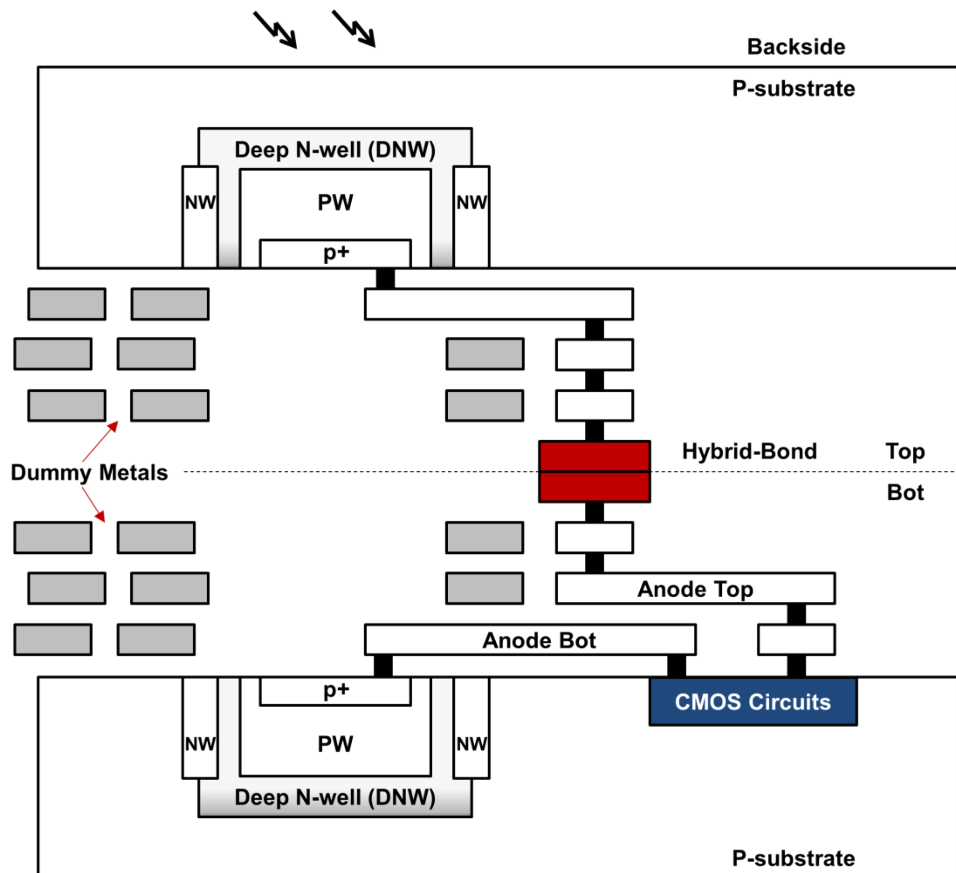


Figure 8.1.2. Cross section of the dual-tier pixel.

8.1.2. Characterisation Results

Figure 8.1.3 shows the PDP response of the two-tier pixel at 3V excess bias for both SPADs. The response of the top BSI SPAD is similar to that of MINI3D (see Fig. 5.2.9) in the sense that it exhibits a cut-off in the blue region and a shift towards the red region of the spectrum. The peak PDP is ~28% at 615nm. The difference in the overall profile could be due to differences in the backside thickness or the backside material stack. Such information is not disclosed to the author. Fabry Perot oscillations are more identifiable in this case due to the fine 2nm sweep of the PDP setup.

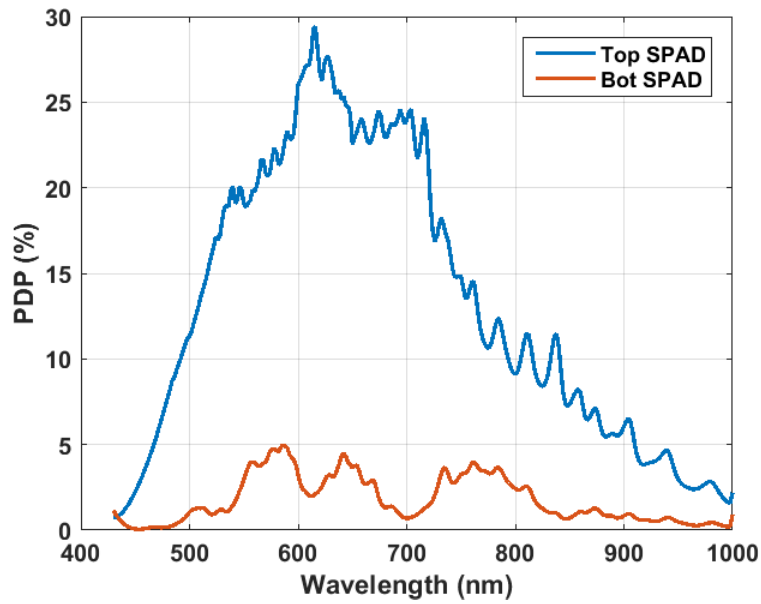


Figure 8.1.3. Dual tier pixel PDP at 3V excess bias.

The PDP response of the bottom SPAD on the other hand is heavily attenuated compared to that of the FSI 40nm SPAD in [114] (see Fig. 3.3.1). This is in a way expected due to the non-optimised device and due to the added stack of the top tier silicon thickness and metallisation. It is thought that the response can be improved by engineering the metal stack such that it creates a light pipe to enhance the detection efficiency [230][316].

Nevertheless this attenuation can be useful for extending the dynamic range in imaging mode. Figure 8.1.4 shows the light count rate versus illumination from a white LED for both SPADs at 2V excess bias. It is obvious that the bottom SPAD results in lower count rates than the much more sensitive top SPAD and so reaches saturation at higher light levels.

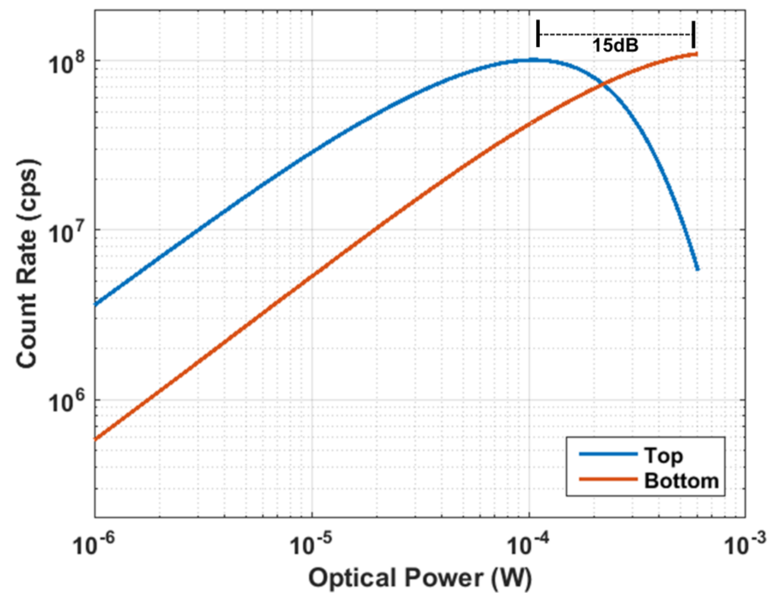


Figure 8.1.4. Dual tier pixel light count rate versus illumination of a white LED.

The top SPAD can be initially used as the main detector until it nears its saturation limit at which point the pixel can be switched to the bottom SPAD. The measured DR extension under the given bias conditions is ~15dB. This can also be useful in ranging applications for example when the target is very close and highly reflective which might saturate the BSI top SPAD and distort time of flight measurements.

An interesting feature of the two-tier pixel is the difference between the angular responses of both SPADs. Since the top SPAD is backside illuminated and not isolated by trenches, it has a wide field of view and is not expected to experience a big change in count rate with the angle of incident light. The bottom SPAD on the other hand has a more restricted field of view emphasised by the barrel like dummy walls surrounding it making it more sensitive to the light incidence angle. This is confirmed by the count rate measurement under fixed illumination level and varying incident light angle in Figure 8.1.5. Due to the symmetrical SPAD structure the angular response was measured across one axis from 0 to 90 degrees.

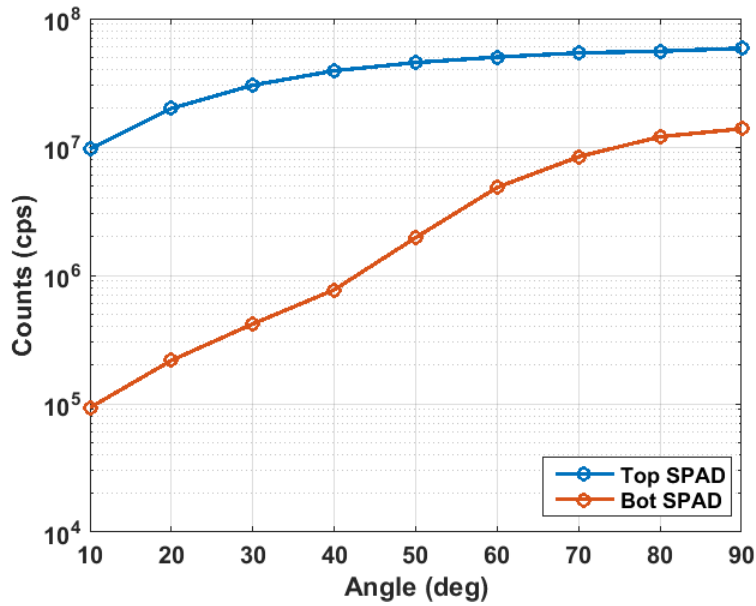


Figure 8.1.5. Dual tier pixel count rate versus incident light angle given a fixed illumination level.

Figure 8.1.6 shows the ratio of counts between the bottom SPAD and top SPAD at different angles clearly demonstrating that such a pixel can be used for angular measurements. Angle sensitive SPADs have been demonstrated in [66] but by using metallisation grids over the SPAD which reduces the photon detection efficiency.

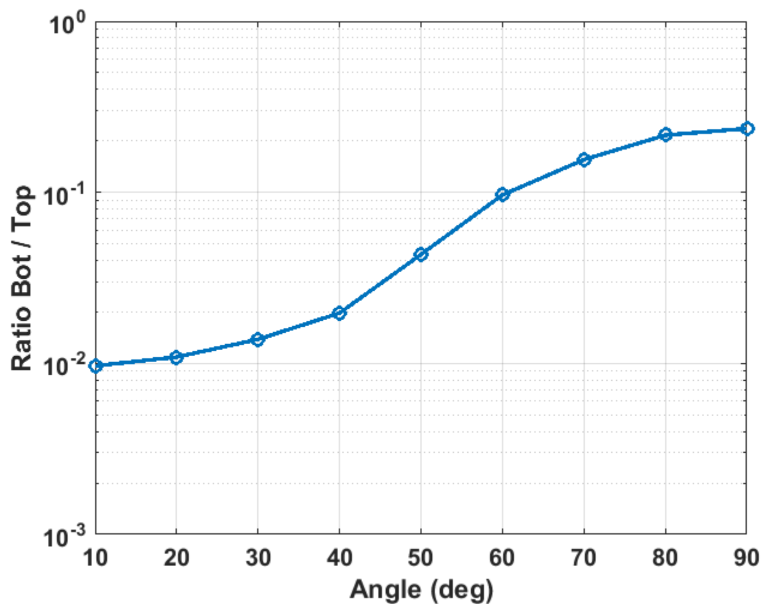


Figure 8.1.6. Ratio of bottom SPAD counts to top SPAD counts versus incident light angle.

The jitter of both SPADs was measured at 2V excess bias and 773nm using a Hamamatsu PLP10 laser head with quoted electrical jitter of 56ps. Both SPADs show a very low jitter FWHM of approximately 70ps (Figure 8.1.7) without correcting for the laser or circuit contributions. The optimised output path and buffer contributed towards better measurements than those reported for MIINI3D SPADs in Figure 5.2.4.

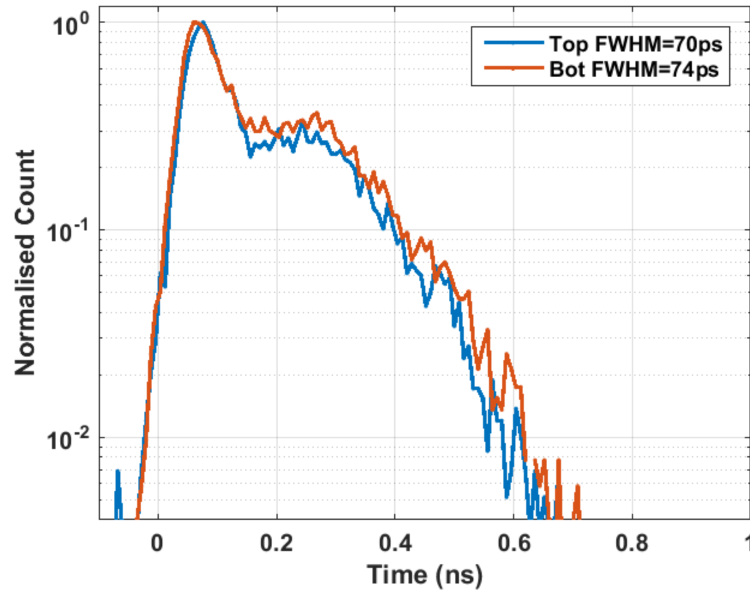


Figure 8.1.7. Dual tier pixel SPADs jitter at 2V excess bias and 773nm.

Finally, although an AND gate arrangement similar to [314] was implemented to investigate the applicability of the two-tier pixel to coincidence detection for LIDAR applications by spatially correlating vertical SPADs as opposed to the conventional planar approach [317], the trial was unsuccessful for several reasons.

First of all the maximum count rate of each SPAD of more than 100Mcps suggests that their dead-times are in the order of 3.5ns which is further reduced to few hundred picoseconds by the pulse shortener resulting in very low coincidence probabilities. This is further deteriorated by the big gap in PDP across all wavelengths between the two devices; hence it was difficult to replicate practical scenarios with meaningful coincidence measurements. This remains under investigation by Dr. Chitnis and it is thought that with better dead-time control and enhanced bottom SPAD PDP such an arrangement might be useful.

8.2. Ultra-Miniature QIS Pixel

Apart from realising miniature sensors for applications such as endoscopy, small pixels are also a key driver towards megapixel resolution avalanche-based image sensors. With Panasonic presenting the first megapixel APD image sensor [284], it is easy to see how the completely rethought APS 3.8μm BSI pixel made that possible. On the other hand, the highest reported resolution for a SPAD image sensor is a 512 × 512 FSI array with a much larger pixel pitch of 16.38μm [187]. Both sensors implement a simple single-bit analogue circuit for photon counting QIS operation.

While the Panasonic approach is elegant as it also allows the megapixel sensor to operate as a conventional CIS by adjusting the diode bias voltage, the SPAD sensor is merely a brute force attempt towards a high resolution array. A major drawback of this large 9.5mm × 9.6mm design is the high power consumption of the 256 parallel output lines needed to achieve a high oversampling ratio. Another drawback of the large area is the distribution and skew of timing signals which is difficult to maintain across such lengths.

Therefore to realise practical megapixel SPAD arrays two innovation directions are needed. First, the sensor needs to implement on-chip data handling and processing capability to avoid transmission of excessive data rates. This can be tackled by architectures such as ENDOCAM or the stacking of processing logic and DRAM as in Sony's high frame rate CIS [202].

Second, the pixel needs to shrink starting with the SPAD itself. Recently at the International Image Sensors Workshop 2017, Prof. Henderson presented a functional 3μm SPAD demonstrating that with better device and technology engineering, it is possible to realise high resolution arrays at such pitch [119]. In the light of the 3D-stacking trend and the ever increasing interconnect density between tiers, he concluded his talk by asking a key research question: *how to realise photon counting and timing pixel electronic matrices in the bottom tier at these small pitches?*

In response to that call, an experimental 3D-stacked array of ultra-miniature 3.24μm QIS pixels was implemented in the top right corner of the CORVETTE sensor (see Figure 6.2.8).

8.2.1. Pixel Design

The schematic of the proposed single-bit self-locking pixel is shown in Figure 8.2.1. Similar to the CORVETTE pixel, a two transistor thick oxide front end conditions the SPAD pulse followed by all-digital 1.1V standard cell logic. At the core of the pixel is a single D-type flip-flop which is initially reset to zero (QN=1) such that incoming SPAD pulses can clock it. If a SPAD pulse samples a high time gate signal defined by the Row and Col controls (as in ENDOCAM and CORVETTE), QN falls low and so locks the pixel state and prevents further SPAD pulses from getting through. Conventional column parallel readout is enabled through a row driven tri-state inverter.

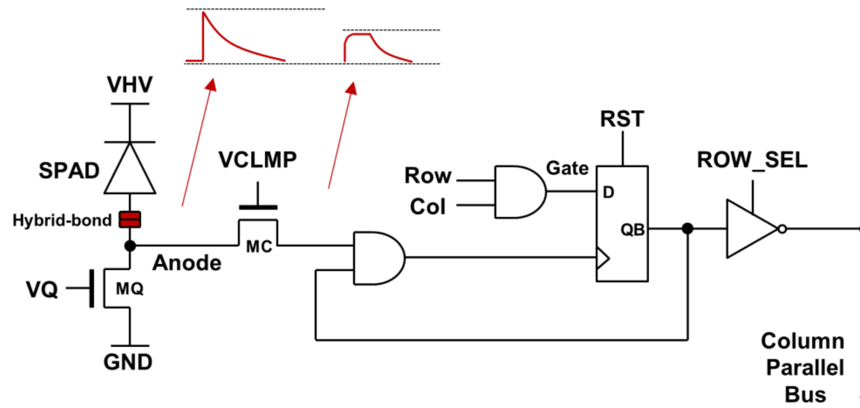
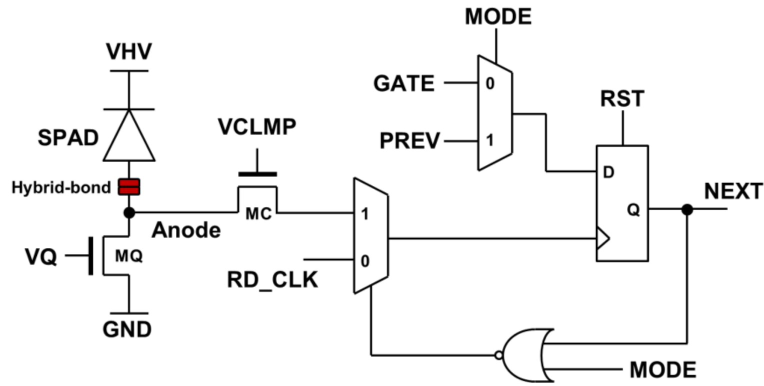


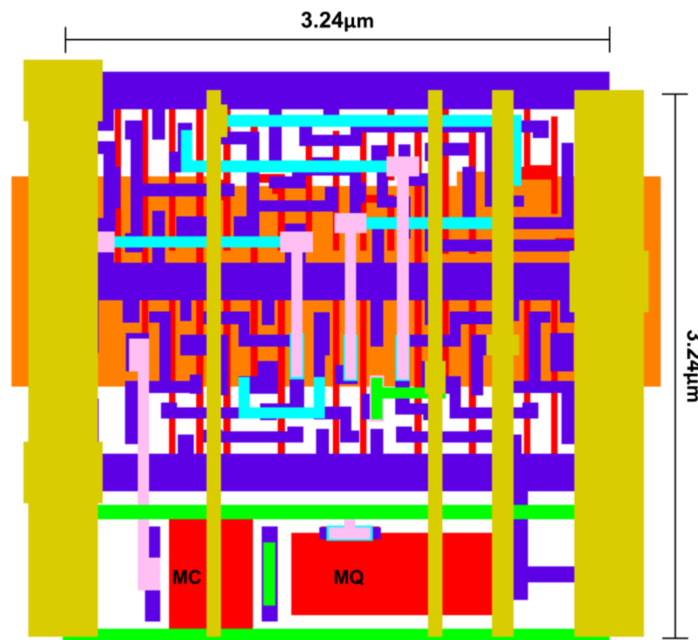
Figure 8.2.1. Single-bit self-locking QIS pixel schematic. Red signals show front end waveforms.

Apart from the miniature layout pitch of 3.24μm which for a megapixel array yields a moderate array size of approximately 3.5mm × 3.5mm with peripheral circuitry, there are several other advantages of this design. The smaller array dimensions make it easier to drive timing signals with manageable skew and less parasitic loading across the array. Moreover, as only thin oxide logic is used, there is the added benefit of lower power consumption of the array drivers due to the smaller capacitive load and lower supply voltage. Unlike analogue pixels, this digital pixel has no parasitic light sensitivity (PLS) or memory leakage issues and offers edge sensitive time gating.

Due to the limited number of IO pads on the CORVETTE chip, a slightly modified version of the self-locking pixel was implemented on silicon. Figure 8.2.2 shows the schematic and layout of the alternative design which also has a pitch of 3.24μm. The main difference is that the AND gates for the gating logic and front end locking are now replaced by multiplexers controlled by a MODE signal. When MODE is low the pixel operates as described above but when MODE is high all the D-type flip-flops in the array chain to form a single shift register that is clocked by RD_CLK. This arrangement requires minimal array controls and a single output pad which is enough to prove the pixel operation even at low frame rates.



(a)



(b)

Figure 8.2.2. Alternative single-bit self-locking QIS pixel implemented in CORVETTE. (a) Schematic diagram. (b) Pixel layout at 3.24µm pitch. Orange is NW, red is PO, dark blue is MT1, light blue is MT2, pink is MT3, green is MT4 and yellow is MT5. Higher metal layers are not shown for clarity.

A 64 × 60 array of this pixel was implemented but because the minimum 3D-stacking hybrid-bond pitch is 6.48µm, only one in every 2 × 2 pixels is connected to a SPAD on the top tier with the other three acting as dummies. Therefore, the effective pixel array resolution is 32 × 30. The same shared well 6.48µm CORVETTE SPAD was implemented on the top tier but future improvements in 3D-stacking hybrid-bond pitch would allow for implementing one 3.24µm SPAD per 3.24µm circuit.

References

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

[2] T. Al Abbas, N. A. W. Dutton, O. Almer, F. M. Della Rocca, S. Pellegrini, B. Rae, D. Golanski and R. K. Henderson, “8.25µm Pitch 66% Fill Factor Global Shared Well SPAD Image Sensor in 40nm CMOS FSI Technology,” in International Image Sensors Workshop, 2017.

[3] T. Al Abbas, N. A. W. Dutton, O. Almer, N. Finlayson, F. M. D. Rocca and R. Henderson, “A CMOS SPAD Sensor With a Multi-Event Folded Flash Time-to-Digital Converter for Ultra-Fast Optical Transient Capture,” in IEEE Sensors Journal, vol. 18, no. 8, pp. 3163-3173, 15 April15, 2018.

[4] N. Dutton, T. Al Abbas, I. Gyongy, F. Mattioli Della Rocca, and R. Henderson, “High Dynamic Range Imaging at the Quantum Limit with Single Photon Avalanche Diode-Based Image Sensors,” Sensors, vol. 18, no. 4, p. 1166, Apr. 2018.

[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,” 2015 IEEE Summer Topicals Meeting Series (SUM), Nassau, 2015, pp. 41-42.

[6] O. Almer, D. Tsonev, N. A. W. Dutton, T. Al Abbas, S. Videv, S. Gnecchi, H. Haas and R. K. Henderson, “A SPAD-Based Visible Light Communications Receiver Employing Higher Order Modulation,” 2015 IEEE Global Communications Conference (GLOBECOM), San Diego, CA, 2015, pp. 1-6.

[7] Neil Finlayson, Tarek Al Abbas, Francescopaolo Mattioli Della Rocca, Oscar Almer, Salvatore Gnecchi, Neale A. W. Dutton, Robert K. Henderson, “Hypervelocity time-of-flight characterisation of a 14GS/s histogramming CMOS SPAD sensor,” Proc. SPIE 10111, Quantum Sensing and Nano Electronics and Photonics XIV, 101112Z (27 January 2017).

[8] I. Gyongy, T. Al Abbas, N. A. W. Dutton and R. K. Henderson, “Object Tracking and Reconstruction with a Quanta Image Sensor,” in International Image Sensors Workshop, 2017.

[9] N. A. W. Dutton, T. Al Abbas, I. Gyongy and R. K. Henderson, “Extending the Dynamic Range of Oversampled Binary SPAD Image Sensors,” in International Image Sensors Workshop, 2017.

[10] F. Mattioli Della Rocca, T. A. Abbas, N. A. W. Dutton and R. K. Henderson, “A high dynamic range SPAD pixel for time of flight imaging,” 2017 IEEE SENSORS, Glasgow, 2017, pp. 1-3.

[11] "Richard Wolf TEXAS Bronchoscope," 2018. [Online]. Available: https://www.richard-wolf.com/en/disciplines/pneumology-thoracic-surgery/texas-bronchoscope/. [Accessed: 18-Aug-2018].

[12] "Olympus BF-190 Bronchoscopes," 2018. [Online]. Available: http://medical.olympusamerica.com/products/bf-190-bronchoscopes. [Accessed: 18-Aug-2018].

[13] "CapsoVision CapsoCamPlus," 2018. [Online]. Available: http://www.capsovision.com/products/capsocam-plus. [Accessed: 18-Aug-2018].

[14] "Integrated endoscopy nuvis single-use rigid endoscope," 2015. [Online]. Available: http://www.iescope.com/nuvis-single-use-arthroscope/. [Accessed: 18-Aug-2018].

[15] "Boston Scientific LithoVue single-use digital flexible ureteroscope," 2018. [Online]. Available: http://www.bostonscientific.com/en-US/products/Ureteroscopes/LithoVue.html. [Accessed: 18-Aug-2018].

[16] "PillCam smart bowel capsule," 2015. [Online]. Available: https://www.pillcamcrohns.com/how-pillcam-works. [Accessed: 18-Aug-2018].

[17] "Awaiba NanEye," 2018. [Online]. Available: http://www.awaiba.com/product/naneye/. [Accessed 24-July-2018].

[18] "OmniVision OV6946 Image Sensor," 2018. [Online]. Available: https://www.ovt.com/sensors/OV6946. [Accessed 24-July-2018].

[19] "OmniVision OV6948 Image Sensor," 2018. [Online]. Available: https://www.ovt.com/sensors/OV6948. [Accessed 24-July-2018].

[20] "Toshiba IK-CT2 Camera," 2018. [Online]. Available: http://www.toshibacameras.com/products/prod_detail_ikct2.jsp. [Accessed 24-July-2018].

[21] M. Wany, S. Voltz, F. Gasparand L. Chen, "Ultrasmall digital image sensor for endoscopic applications," in International Image Sensors Workshop, 2009.

[22] B. Wolfs, C. Esquenet and W. Iandolo, "400×400 pixel image sensor for endoscopy in 1.7mm² CSP package," in International Image Sensors Workshop, 2009.

[23] S. Yoshizaki, A. Serb, Y. Liu and T. G. Constandinou, "Octagonal CMOs image sensor with strobed RGB LED illumination for wireless capsule endoscopy," 2014 IEEE International Symposium on Circuits and Systems (ISCAS), Melbourne VIC, 2014, pp. 1857-1860.

[24] M. Wany, P. Santos, E. Reis, A. Andrade, R. Sousa and L. Sousa, "Octagonal CMOS Image Sensor for Endoscopic Applications," Electronic Imaging, Image Sensors and Imaging Systems 2017, pp. 46-51(6).

[25] D. Covi, C. Cavallotti, M. Vatteroni, L. Clementel, P. Valdastri, A. Menciassi, P. Dario and A. Sartori, "Miniaturized digital camera system for disposable endoscopic applications," Sensors and Actuators A: Physical, Volume 162, Issue 2, 2010, Pages 291-296.

[26] M. A. Al-Rawhani, D. Chitnis, J. Beeley, S. Collins and D. R. S. Cumming, "Design and Implementation of a Wireless Capsule Suitable for Autofluorescence Intensity Detection in Biological Tissues," in IEEE Transactions on Biomedical Engineering, vol. 60, no. 1, pp. 55-62, Jan. 2013.

[27] M. Rawhani, J. Beeley and D. Cumming, "Wireless fluorescence capsule for endoscopy using single photon-based detection," in Scientific Reports 5, Article number: 18591 (2015).

[28] M. F. Tompsett et al., "Charge-coupled imaging devices: Experimental results," in IEEE Transactions on Electron Devices, vol. 18, no. 11, pp. 992-996, Nov. 1971.

[29] J. Nakamura, Ed., Image Sensors and Signal Processing for Digital Still Cameras, First. CRC Press, 2006.

[30] D. Renshaw, P. B. Denyer, G. Wang and M. Lu, "ASIC vision," IEEE Proceedings of the Custom Integrated Circuits Conference, Boston, MA, USA, 1990, pp. 7.3/1-7.3/4.

[31] E. R. Fossum, “Active pixel sensors: are CCDs dinosaurs?,” in IS&T/SPIE’s Symposium on Electronic Imaging: Science and Technology, 1993, pp. 2–14.

[32] P. J. W. Noble, “Self-scanned silicon image detector arrays,” IEEE Trans. Electron Devices, vol. 15, no. 4, pp. 202–209, Apr. 1968.

[33] N. Teranishi, A. Kohono, Y. Ishihara, E. Oda and K. Arai, "No image lag photodiode structure in the interline CCD image sensor," 1982 International Electron Devices Meeting, San Francisco, CA, USA, 1982, pp. 324-327.

[34] E. R. Fossum and D. B. Hondongwa, "A Review of the Pinned Photodiode for CCD and CMOS Image Sensors," in IEEE Journal of the Electron Devices Society, vol. 2, no. 3, pp. 33-43, May 2014.

[35] E. R. Fossum, "CMOS image sensors: electronic camera-on-a-chip," in IEEE Transactions on Electron Devices, vol. 44, no. 10, pp. 1689-1698, Oct. 1997.

[36] A. Theuwissen, "CMOS image sensors: State-of-the-art," in Solid-State Electronics, Volume 52, Issue 9, Pages 1401-1406, 2008.

[37] E. Fossum, "CAMERA-ON-A-CHIP: TECHNOLOGY TRANSFER FROM SATURN TO YOUR CELL PHONE," in Technology & Innovation. 15 (3): 197–209, December 2013.

[38] W. Becker, “Fluorescence lifetime imaging – techniques and applications,” Journal of Microscopy, 247: 119–136. 2012.

[39] L. Marcu, "Fluorescence Lifetime Techniques in Medical Applications," in Ann Biomed Eng., 40(2):304-31, Feb. 2012.

[40] H. A. R. Homulle, F. Powolny, P. L. Stegehuis, J. Dijkstra, D.-U. Li, K. Homicsko, D. Rimoldi, K. Muehlethaler, J. O. Prior, R. Sinisi, E. Dubikovskaya, E. Charbon, and C. Bruschini, "Compact solid-state CMOS single-photon detector array for in vivo NIR fluorescence lifetime oncology measurements," Biomed. Opt. Express 7, 1797-1814 (2016).

[41] K. Ehrlich, A. Kufcsák, S. McAughtrie, H. Fleming, N. Krstajic, C. J. Campbell, R. K. Henderson, K. Dhaliwal, R. R. Thomson, and M. G. Tanner, "pH sensing through a single optical fibre using SERS and CMOS SPAD line arrays," Opt. Express 25, 30976-30986 (2017).

[42] " Becker and Hickl TCSPC Module card SPC-130-EMN," 2018. [Online]. Available: https://www.becker-hickl.com/spc-130-em.htm. [Accessed: 19-Aug-2018].

[43] R. M. Ballew and J. N. Demas, “An error analysis of the rapid lifetime determination method for the evaluation of single exponential decays,” Ana. Chem., vol. 61, pp. 30-33, 1989.

[44] F. Remondino and D. Stoppa, Eds., TOF Range-Imaging Cameras, 1st ed. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.

[45] W. Becker, Advanced Time-Correlated Single Photon Counting Techniques. Springer, 2005.

[46] R. Lange and P. Seitz, "Solid-state time-of-flight range camera," in IEEE Journal of Quantum Electronics, vol. 37, no. 3, pp. 390-397, March 2001.

[47] T. Spirig, P. Seitz, O. Vietze, and F. Heitger, “The lock-in CCD-two-dimensional synchronous detection of light,” IEEE J. Quantum Electron., vol. 31, no. 9, pp. 1705–1708, 1995.

[48] A. Payne et al., "A 512×424 CMOS 3D Time-of-Flight image sensor with multi-frequency photo-demodulation up to 130MHz and 2GS/s ADC," 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, 2014, pp. 134-135.

[49] C. S. Bamji et al., "1Mpixel 65nm BSI 320MHz demodulated TOF Image sensor with 3μm global shutter pixels and analog binning," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, 2018, pp. 94-96.

[50] W. van der Tempel, R. Grootjans, D. Van Nieuwenhove and M. Kuijk, "A 1k-pixel 3D CMOS sensor," SENSORS, 2008 IEEE, Lecce, 2008, pp. 1000-1003.

[51] L. Pancheri, D. Stoppa, N. Massari, M. Malfatti, L. Gonzo, Q. D. Hossain and G. Dalla Betta, "A 120x160 pixel CMOS range image sensor based on current assisted photonic demodulators," Proc. SPIE 7726, Optical Sensing and Detection, 772615 (13 May 2010).

[52] Y. Kato et al., "320×240 Back-illuminated 10µm CAPD pixels for high speed modulation Time-of-Flight CMOS image sensor," 2017 Symposium on VLSI Circuits, Kyoto, 2017, pp. C288-C289.

[53] H. Yoon, S. Itoh and S. Kawahito, "A CMOS Image Sensor With In-Pixel Two-Stage Charge Transfer for Fluorescence Lifetime Imaging," in IEEE Transactions on Electron Devices, vol. 56, no. 2, pp. 214-221, Feb. 2009.

[54] D. Stoppa, N. Massari, L. Pancheri, M. Malfatti, M. Perenzoni and L. Gonzo, "An 80×60 range image sensor based on 10µm 50MHz lock-in pixels in 0.18µm CMOS," 2010 IEEE International Solid-State Circuits Conference - (ISSCC), San Francisco, CA, 2010, pp. 406-407.

[55] M. Seo et al., "A 10.8ps-time-resolution 256×512 image sensor with 2-Tap true-CDS lock-in pixels for fluorescence lifetime imaging," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, San Francisco, CA, 2015, pp. 1-3.

[56] M. Seo et al., "A programmable sub-nanosecond time-gated 4-tap lock-in pixel CMOS image sensor for real-time fluorescence lifetime imaging microscopy," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2017, pp. 70-71.

[57] R. Jeremias, W. Brockherde, G. Doemens, B. Hosticka, L. Listl and P. Mengel, "A CMOS photosensor array for 3D imaging using pulsed laser," 2001 IEEE International Solid-State Circuits Conference. Digest of Technical Papers. ISSCC (Cat. No.01CH37177), San Francisco, CA, USA, 2001, pp. 252-253.

[58] D. Stoppa, L. Viarani, A. Simoni, L. Gonzo, M. Malfatti and G. Pedretti, "A 50x30-pixel CMOS Sensor for TOF-based Real Time 3D Imaging," in International Image Sensors Workshop, 2005.

[59] A. Pauchard, A. Rochas, Z. Randjelovic, P. A. Besse and R. S. Popovic, "Ultraviolet avalanche photodiode in CMOS technology," International Electron Devices Meeting 2000. Technical Digest. IEDM (Cat. No.00CH37138), San Francisco, CA, USA, 2000, pp. 709-712.

[60] A. Rochas, M. Gani, B. Furrer, P. A. Besse, and R. S. Popovic, "Single photon detector fabricated in a complementary metal–oxide–semiconductor high-voltage technology," in Review of Scientific Instruments 74, 3263 (2003).

[61] C. L. Niclass, A. Rochas, P. A. Besse and E. Charbon, "A CMOS single photon avalanche diode array for 3D imaging," 2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519), San Francisco, CA, 2004, pp. 120-517 Vol.1.

[62] D. Stoppa, L. Pancheri, M. Scandiuzzo, M. Malfatti, G. Pedretti and L. Gonzo, "A single-photon-avalanche-diode 3D imager," Proceedings of the 31st European Solid-State Circuits Conference, 2005. ESSCIRC 2005., Grenoble, France, 2005, pp. 487-490.

[63] C. Niclass, C. Favi, T. Kluter, M. Gersbach and E. Charbon, "A 128×128 Single-Photon Imager with on-Chip Column-Level 10b Time-to-Digital Converter Array Capable of 97ps Resolution," 2008 IEEE International Solid-State Circuits Conference - Digest of Technical Papers, San Francisco, CA, 2008, pp. 44-594.

[64] D. Stoppa, L. Pancheri, M. Scandiuzzo, L. Gonzo, G. Dalla Betta and A. Simoni, "A CMOS 3-D Imager Based on Single Photon Avalanche Diode," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 54, no. 1, pp. 4-12, Jan. 2007.

[65] N. A. W. Dutton, L. Parmesan, A. J. Holmes, L. A. Grant and R. K. Henderson, "320×240 oversampled digital single photon counting image sensor," 2014 Symposium on VLSI Circuits Digest of Technical Papers, Honolulu, HI, 2014, pp. 1-2.

[66] C. Lee, B. Johnson and A. Molnar, "An on-chip 72×60 angle-sensitive single photon image sensor array for lens-less time-resolved 3-D fluorescence lifetime imaging," 2014 Symposium on VLSI Circuits Digest of Technical Papers, Honolulu, HI, 2014, pp. 1-2.

[67] B. F. Aull, D. R. Schuette, D. J. Young, D. M. Craig, B. J. Felton and K. Warner, "Photon counting imagers based on high-fill-factor silicon geiger-mode avalanche photodiode arrays," 2014 IEEE Photonics Conference, San Diego, CA, 2014, pp. 166-167.

[68] I. Gyongy et al., "A 256×256 , 100-kfps, 61% Fill-Factor SPAD Image Sensor for Time-Resolved Microscopy Applications," in IEEE Transactions on Electron Devices, vol. 65, no. 2, pp. 547-554, Feb. 2018.

[69] S. Chen and E. R. Fossum, "A time-resolved CMOS image sensor with high conversion-gain pixels and pipelined ADCs," 2017 IEEE 60th International Midwest Symposium on Circuits and Systems (MWSCAS), Boston, MA, 2017, pp. 1-4.

[70] M. J. Hsu, H. Finkelstein and S. C. Esener, "A CMOS STI-Bound Single-Photon Avalanche Diode With 27-ps Timing Resolution and a Reduced Diffusion Tail," in IEEE Electron Device Letters, vol. 30, no. 6, pp. 641-643, June 2009.

[71] I. Nissinen, J. Nissinen, P. Keränen, A. Länsman, J. Holma and J. Kostamovaara, "A 2 × (4) × 128 Multitime-Gated SPAD Line Detector for Pulsed Raman Spectroscopy," in IEEE Sensors Journal, vol. 15, no. 3, pp. 1358-1365, March 2015.

[72] L. Parmesan, N. Dutton, N. Calder, N. Krstajic, A. Holmes, L. Grant and R. Henderson, "A 256x256 SPAD array with in-pixel Time to Amplitude Conversion for Fluorescence Lifetime Imaging Microscopy," in International Image Sensors Workshop, 2015

[73] N. A. W. Dutton et al., "A SPAD-Based QVGA Image Sensor for Single-Photon Counting and Quanta Imaging," in IEEE Transactions on Electron Devices, vol. 63, no. 1, pp. 189-196, Jan. 2016.

[74] N. A. W. Dutton, I. Gyongy, L. Parmesan, and R. K. Henderson, "Single Photon Counting Performance and Noise Analysis of CMOS SPAD-based Image Sensors," Sensors (Basel)., no. Special Issue, pp. 1–17, 2016.

[75] E. Charbon, M. Scandini, J. Mata Pavia and M. Wolf, "A dual backside-illuminated 800-cell multi-channel digital SiPM with 100 TDCs in 130nm 3D IC technology," 2014 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), Seattle, WA, 2014, pp. 1-4.

[76] M. Perenzoni, L. Pancheri and D. Stoppa, "Compact SPAD-Based Pixel Architectures for Time-Resolved Image Sensors," Sensors (Basel)., 2016.

[77] E. R. Fossum, "The Quanta Image Sensor (QIS): Concepts and Challenges," in Imaging and Applied Optics, OSA Technical Digest (CD) (Optical Society of America, 2011).

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

[79] N. Krstajić, S. Poland, J. Levitt, R. Walker, A. Erdogan, S. Ameer-Beg, and R. K. Henderson, "0.5 billion events per second time correlated single photon counting using CMOS SPAD arrays," Opt. Lett. 40, 4305-4308 (2015)

[80] S. Burri, Y. Maruyama, X. Michalet, F. Regazzoni, C. Bruschini, and E. Charbon, "Architecture and applications of a high resolution gated SPAD image sensor," Opt. Express 22, 17573-17589 (2014).

[81] A. Goetzberger, B. McDonald, R. H. Haitz, and R. M. Scarlett, "Avalanche Effects in Silicon p-n Junctions. II. Structurally Perfect Junctions," Physical Review, vol. 34, no. 6 pp. 1591-1600, June 1963.

[82] R. H. Haitz, "Mechanisms Contributing to the noise pulse rate of avalanche diodes," Journal of Applied Physics, vol. 36, no. 10 pp. 3123-3131, October 1965.

[83] R. J. Mc Intyre, "The Distribution of Gains in Uniformly Multiplying Avalanche Photodiodes: Theory," IEEE Trans. Electron Devices ED-19, 703-713 (1972).

[84] P.P.Webb, R. J. McIntyre, J.Conradi , "Properties of Avalanche Photodiodes" , RCA Review 35, 234-278 (1974).

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

[86] S. Cova, Ma. Ghioni, F. Zappa, I. Rech, A. Gulinatti, "A view on progress of silicon single-photon avalanche diodes and quenching circuits," Proc. SPIE 6372, Advanced Photon Counting Techniques, 63720I (25 October 2006);

[87] M. Ghioni, A. Gulinatti, I. Rech, F. Zappa and S. Cova, "Progress in Silicon Single-Photon Avalanche Diodes," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 13, no. 4, pp. 852-862, July-aug. 2007.

[88] R. J. McIntyre, "Recent developments in silicon avalanche photodiodes," Measurement, vol. 3, no. 4 pp. 146-152, 1985.

[89] E. A. G. Webster, "Single-Photon Avalanche Diode Theory, Simulation, and High-Performance CMOS Integration," The University of Edinburgh, 2013.

[90] E. Fisher., Principles and Early Historical Development of Silicon Avalanche and Geiger-Mode Photodiodes, 1st ed. , Open Access Book Chapter, 2018.

[91] E. Charbon, "Single-photon imaging in complementary metal oxide semiconductor processes," Phil Trans R Soc A, vol. 372, no. 2012, Feb. 2014.

[92] B. F. Aull, A. H. Loomis, D. J. Young, R. M. Heinrichs, B. J. Felton, P. J. Daniels, and Deborah J. Landers, "Geiger-Mode Avalanche Photodiodes for Three Dimensional Imaging," Lincoln Lab. J., vol. 13, no. 2, pp. 335–350, 2002.

[93] F. Zappa, S. Tisa, A. Tosi and S. Cova, "Principles and features of single-photon avalanche diode arrays," Sensors and Actuators A: Physical, Volume 140, Issue 1, October 2007.

[94] L. Pancheri and D. Stoppa, "Low-Noise CMOS single-photon avalanche diodes with 32 ns dead time," ESSDERC 2007 - 37th European Solid State Device Research Conference, Munich, 2007, pp. 362-365.

[95] M. A. Marwick and A. G. Andreou, "Fabrication and Testing of Single Photon Avalanche Detectors in the TSMC 0.18μm CMOS Technology," 2007 41st Annual Conference on Information Sciences and Systems, Baltimore, MD, 2007, pp. 741-744.

[96] C. Veerappan and E. Charbon, "A Substrate Isolated CMOS SPAD Enabling Wide Spectral Response and Low Electrical Crosstalk," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 20, no. 6, pp. 299-305, Nov.-Dec. 2014, Art no. 3801507.

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

[98] R. M. Field, J. Lary, J. Cohn, L. Paninski, K. L. Shepard, "A low-noise, single-photon avalanche diode in standard 0.13μm complementary metaloxide-semiconductor process," Applied Physics Letters, vol. 97, no. 211111, pp. 1-3, November 2010.

[99] M. Lee, P. Sun and E. Charbon, "Characterization of Single-Photon Avalanche Diodes in Standard 140-nm SOI CMOS Technology," in International Image Sensors Workshop, 2015.

[100] C. Veerappan and E. Charbon, "CMOS SPAD Based on Photo-Carrier Diffusion Achieving PDP >40% From 440 to 580 nm at 4 V Excess Bias," in IEEE Photonics Technology Letters, vol. 27, no. 23, pp. 2445-2448, 1 Dec.1, 2015.

[101] W. O. Oldham, R. R. Samuelson, and P. Antognetti, "Triggering phenomena in avalanche diodes," IEEE Trans. Electron Devices, vol. ED-19, no. 9, pp. 1056–1060, Sep. 1972.

[102] S. Mandai, M. W. Fishburn, Y. Maruyama, and E. Charbon, "A wide spectral range single-photon avalanche diode fabricated in an advanced 180 nm CMOS technology," Optics Express, vol. 20, no. 6 pp. 5849-5857, 27 February 2012.

[103] M. Karami, H. Yoon and E. Charbon,"Single-Photon Avalanche Diodes in sub-100nm Standard CMOS Technologies," in International Image Sensors Workshop, 2011.

[104] E. Charbon, H. Yoon and Y. Maruyama, "A Geiger mode APD fabricated in standard 65nm CMOS technology," 2013 IEEE International Electron Devices Meeting, Washington, DC, 2013, pp. 27.5.1-27.5.4.

[105] H. Finkelstein, M. J. Hsu and S. C. Esener, "STI-Bounded Single-Photon Avalanche Diode in a Deep-Submicrometer CMOS Technology," in IEEE Electron Device Letters, vol. 27, no. 11, pp. 887-889, Nov. 2006.

[106] M. Gersbach, C. Niclass, E. Charbon, J. Richardson, R. Henderson and L. Grant, "A single photon detector implemented in a 130nm CMOS imaging process," ESSDERC 2008 - 38th European Solid-State Device Research Conference, Edinburgh, 2008, pp. 270-273.

[107] A. Rochas, A. R. Pauchard, P.-A. Besse, D. Pantic, Z. Prijic, and R. S. Popovic, "Low-noise silicon avalanche photodiodes fabricated in conventional CMOS technologies," IEEE Trans. Electron Devices, vol. 49, no. 3, pp. 387–394, Mar. 2002.

[108] M. A. Marwick, A. G. Andreou, "Single photon avalanche photodetector with integrated quenching fabricated in TSMC 0.18um 1.8V CMOS process," Electronics Letters, vol. 44, no. 10 8 May 2008.

[109] R. Henderson, J. Richardson and L. Grant,"Reduction of Band-to-band Tunneling in Deep-submicron CMOS Single Photon Avalanche Photodiodes," in International Image Sensors Workshop,2009

[110] L. Pancheri, G. Dalla Betta, L. H. Campos Braga, H. Xu and D. Stoppa, "A single-photon avalanche diode test chip in 150nm CMOS technology," 2014 International Conference on Microelectronic Test Structures (ICMTS), Udine, 2014, pp. 161-164.

[111] J. A. Richardson, L. Grant and R. Henderson, "Low Dark Count Single-Photon Avalanche Diode Structure Compatible With Standard Nanometer Scale CMOS Technology," in IEEE PTL, vol. 21, no. 14, pp. 1020-1022, July 2009.

[112] E. Webster, J. Richardson, L. Grant and R. Henderson, "Single-Photon Avalanche Diodes in 90nm CMOS imaging technology with sub-1Hz Median Dark Count Rate," in International Image Sensors Workshop, 2011

[113] S. Pellegrini and B. Rae, "Fully industrialised single photon avalanche diodes," Proceedings of Advanced Photon Counting Techniques XI, Volume 10212, 2017

[114] S. Pellegrini et al., "Industrialised SPAD in 40 nm technology," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 16.5.1-16.5.4.

[115] 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," in IEEE Electron Device Letters, vol. 33, no. 5, pp. 694-696, May 2012.

[116] E. A. G. Webster, L. A. Grant and R. K. Henderson, "A High-Performance Single-Photon Avalanche Diode in 130-nm CMOS Imaging Technology," in IEEE Electron Device Letters, vol. 33, no. 11, pp. 1589-1591, Nov. 2012.

[117] R. K. Henderson, E. A. G. Webster, R. Walker, J. A. Richardson and L. A. Grant, "A 3×3, 5µm pitch, 3-transistor single photon avalanche diode array with integrated 11V bias generation in 90nm CMOS technology," 2010 International Electron Devices Meeting, San Francisco, CA, 2010, pp. 14.2.1-14.2.4.

[118] J. A. Richardson, L. A. Grant, E. A. G. Webster and R. K. Henderson, "A 2um diameter, 9hz dark count, single photon avalanche diode in 130nm cmos technology," 2010 Proceedings of the European Solid State Device Research Conference, Sevilla, 2010, pp. 257-260.

[119] Z. You, L. Parmesan, S. Pellegrini and R. Henderson, "3µm Pitch, 1µm Active Diameter SPAD Arrays in 130nm CMOS Imaging Technology," in International Image Sensors Workshop, 2017.

[120] P. Sun, E. Charbon and R. Ishihara," A Flexible 32×32 SPAD Image Sensor with Integrated Microlenses," in International Image Sensors Workshop, 2015.

[121] 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," in International Image Sensors Workshop, 2015.

[122] I. Takai, H. Matsubara, M. Soga, M. Ohta, M. Ogawa and T. Yamashita, "Single-Photon Avalanche Diode with Enhanced NIR-Sensitivity for Automotive LIDAR Systems," Sensors (Basel), March 2016.

[123] M. M. Vignetti, F. Calmon, P. Lesieur, F. Dubois, T. Graziosi and A. Savoy-Navarro, "A novel 3D pixel concept for Geiger-mode detection in SOI technology," 2016 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS), Vienna, 2016, pp. 166-169.

[124] M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce and F. Zappa, "Single-Photon Avalanche Diodes in a 0.16 µm BCD Technology With Sharp Timing Response and Red-Enhanced Sensitivity," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 24, no. 2, pp. 1-9, March-April 2018, Art no. 3801209.

[125] H. Zimmermann, B. Steindl, M. Hofbauer and R. Enne, "Integrated fiber optical receiver reducing the gap to the quantum limit," in Scientific Reports 7, Article number: 2652, 2017.

[126] F. Acerbi et al., "High Efficiency, Ultra-High-Density Silicon Photomultipliers," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 24, no. 2, pp. 1-8, March-April 2018, Art no. 3800608.

[127] D. Chitnis and S. Collins, “A flexible compact readout circuit for SPAD arrays,” in SPIE NanoScience + Engineering, 2010.

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

[129] M. Karami, C. Niclass and E. Charbon, "Random Telegraph Signal in Single-Photon Avalanche Diodes," in International Image Sensors Workshop, 2009

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

[131] A. Tournier, F. Leverd, L. Favennec, C. Perrot, L. Pinzelli, M. Gatefait, N. Cherault, D. Jeanjean, J.-P. Carrere, F. Hirigoyen, L. Grant and F. Roy, "Pixel-to-Pixel isolation by Deep Trench technology: Application to CMOS Image Sensor," in International Image Sensors Workshop, 2011.

[132] A. L. Lacaita, F. Zappa, S. Bigliardi and M. Manfredi, "On the bremsstrahlung origin of hot-carrier-induced photons in silicon devices," in IEEE Transactions on Electron Devices, vol. 40, no. 3, pp. 577-582, March 1993.

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

[134] H. Xu, L. Pancheri, L. Braga, G. Della Betta and D. Stoppa, "Crosstalk characterization of single-photon avalanche diode (SPAD) arrays in CMOS 150nm technology," in Procedia Engineering 87, 2014.

[135] A. Ficorella et al., "Crosstalk mapping in CMOS SPAD arrays," 2016 46th European Solid-State Device Research Conference (ESSDERC), Lausanne, 2016, pp. 101-104.

[136] 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," in IEEE Sensors Journal, vol. 16, no. 1, pp. 3-12, Jan.1, 2016.

[137] S. Mandai and E. Charbon, "Stabilizing sensitivity in large single-photon image sensors with an integrated 3.3-to-25V all-digital charge pump," in International Image Sensors Workshop, 2013.

[138] M. Lee, P. Sun and E. Charbon, "A first single-photon avalanche diode fabricated in standard SOI CMOS technology with a full characterization of the device," OSA, Vol.23, No.10, May 2015.

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

[140] A. Eisele, R. Henderson, B. Schmidtke, T. Funk, L. Grant, J. Richardson and W. Freude, "185 MHz Count Rate, 139 dB Dynamic Range Single-Photon Avalanche Diode with Active Quenching Circuit in 130 nm CMOS Technology," in International Image Sensors Workshop, 2011.

[141] 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," J. Mod. Opt., vol. 61, no. 2, pp. 102–115, Jan. 2014.

[142] S. Gnecchi et al., "A Simulation Model for Digital Silicon Photomultipliers," in IEEE Transactions on Nuclear Science, vol. 63, no. 3, pp. 1343-1350, June 2016.

[143] J. Liu et al., "Fast Active-Quenching Circuit for Free-Running InGaAs(P)/InP Single-Photon Avalanche Diodes," in IEEE Journal of Quantum Electronics, vol. 52, no. 10, pp. 1-6, Oct. 2016, Art no. 4000306.

[144] H. Finkelstein, M. J. Hsu, S. Zlatanovic, and S. Esener, "Performance trade-offs in single-photon avalanche diode miniaturization.," Rev. Sci. Instrum., vol. 78, no. 10, p. 103103, 2007.

[145] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf and E. Charbon, "A High-PDE, Backside-Illuminated SPAD in 65/40-nm 3D IC CMOS Pixel With Cascoded Passive Quenching and Active Recharge," in IEEE Electron Device Letters, vol. 38, no. 11, pp. 1547-1550, Nov. 2017.

[146] J. Richardson, R. Henderson and D. Renshaw, "Dynamic Quenching for Single Photon Avalanche Diode Arrays," in International Image Sensors Workshop, 2007.

[147] A. Gallivanoni, I. Rech and M. Ghioni, "Progress in Quenching Circuits for Single Photon Avalanche Diodes," in IEEE Transactions on Nuclear Science, vol. 57, no. 6, pp. 3815-3826, Dec. 2010.

[148] E. Webster, J. Richardson, L. Grant, D. Renshaw and R. Henderson,"An Infra-Red Sensitive, Low Noise, Single-Photon Avalanche Diode in 90nm CMOS," in International Image Sensors Workshop, 2011.

[149] L. Pancheri and D. Stoppa, "A SPAD-based pixel linear array for high-speed time-gated fluorescence lifetime imaging," 2009 Proceedings of ESSCIRC, Athens, 2009, pp. 428-431.

[150] E. A. G. Webster, R. J. Walker, R. K. Henderson, and L. A. Grant, “A silicon photomultiplier with >30% detection efficiency from 450–750nm and 11.6µm pitch NMOS-only pixel with 21.6% fill factor in 130nm CMOS,” in 2012 Proceedings of the European Solid-State Device Research Conference (ESSDERC), 2012, pp. 238–241.

[151] S. Gnecchi et al., "Analysis of Photon Detection Efficiency and Dynamic Range in SPAD-Based Visible Light Receivers," in Journal of Lightwave Technology, vol. 34, no. 11, pp. 2774-2781, 1 June1, 2016.

[152] "STMicroelectronics Proximity Sensors," 2018. [Online]. Available: https://www.st.com/en/imaging-and-photonics-solutions/proximity-sensors.html?querycriteria=productId=SC1934. [Accessed: 27-Aug-2018].

[153] F. Borghetti, D. Mosconi, L. Pancheri and D. Stoppa, "A CMOS Single-Photon Avalanche Diode Sensor for Fluorescence Lifetime Imaging," in International Image Sensors Workshop, 2007.

[154] 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,” in 2009 IEEE Nuclear Science Symposium Conference Record (NSS/MIC), 2009, pp. 1959–1965.

[155] D. Tyndall, B. Rae, D. Li, J. Richardson, J. Arlt and R. Henderson, "A 100Mphoton/s time-resolved mini-silicon photomultiplier with on-chip fluorescence lifetime estimation in 0.13µm CMOS imaging technology," 2012 IEEE International Solid-State Circuits Conference, San Francisco, CA, 2012, pp. 122-124.

[156] S. Mandai, V. Jain and E. Charbon, "A 780×800 µm² Multichannel Digital Silicon Photomultiplier With Column-Parallel Time-to-Digital Converter and Basic Characterization," in IEEE Transactions on Nuclear Science, vol. 61, no. 1, pp. 44-52, Feb. 2014.

[157] N. A. W. Dutton et al., "A time-correlated single-photon-counting sensor with 14GS/S histogramming time-to-digital converter," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, San Francisco, CA, 2015, pp. 1-3.

[158] L. H. C. Braga et al., "A CMOS mini-SiPM detector with in-pixel data compression for PET applications," 2011 IEEE Nuclear Science Symposium Conference Record, Valencia, 2011, pp. 548-552.

[159] S. Gnecchi et al., "Digital Silicon Photomultipliers With OR/XOR Pulse Combining Techniques," in IEEE Transactions on Electron Devices, vol. 63, no. 3, pp. 1105-1110, March 2016.

[159/2] S. Tisa, F. Guerrieri, A. Tosi and F. Zappa, "100 kframe/s 8 bit monolithic single-photon imagers," ESSDERC 2008 - 38th European Solid-State Device Research Conference, Edinburgh, 2008, pp. 274-277.

[160] C. Niclass, M. Soga, H. Matsubara and S. Kato, "A 100m-range 10-frame/s 340×96-pixel time-of-flight depth sensor in 0.18µm CMOS," 2011 Proceedings of the ESSCIRC (ESSCIRC), Helsinki, 2011, pp. 107-110.

[161] C. Niclass, M. Soga, H. Matsubara, M. Ogawa and M. Kagami, "A 0.18µm CMOS SoC for a 100-m-Range 10-Frame/s 200×96-Pixel Time-of-Flight Depth Sensor," in IEEE Journal of Solid-State Circuits, vol. 49, no. 1, pp. 315-330, Jan. 2014.

[162] Y. Maruyama, J. Blacksberg and E. Charbon, "A 1024×8 700ps time-gated SPAD line sensor for laser raman spectroscopy and LIBS in space and rover-based planetary exploration," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, 2013, pp. 110-111.

[163] I. Nissinen, J. Nissinen and J. Kostamovaara, "A time-gated 4×128 SPAD array with a 512 channel flash 80 ps-TDC for pulsed Raman spectroscopy," 2013 European Conference on Circuit Theory and Design (ECCTD), Dresden, 2013, pp. 1-4.

[164] N. Krstajić et al., "A 256 × 8 SPAD line sensor for time resolved fluorescence and raman sensing," ESSCIRC 2014 - 40th European Solid State Circuits Conference (ESSCIRC), Venice Lido, 2014, pp. 143-146.

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

[166] C. Niclass and E. Charbon, "A single photon detector array with 64/spl times/64 resolution and millimetric depth accuracy for 3D imaging," ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005., San Francisco, CA, 2005, pp. 364-604 Vol. 1.

[167] J. Richardson et al., "A 32×32 50ps resolution 10 bit time to digital converter array in 130nm CMOS for time correlated imaging," 2009 IEEE Custom Integrated Circuits Conference, Rome, 2009, pp. 77-80.

[168] C. Veerappan et al., "A 160×128 single-photon image sensor with on-pixel 55ps 10b time-to-digital converter," 2011 IEEE International Solid-State Circuits Conference, San Francisco, CA, 2011, pp. 312-314.

[169] I. Vornicu, R. Carmona-Galán and A. Rodríguez-Vázquez, "A CMOS 0.18μm 64×64 single photon image sensor with in-pixel 11b time-to-digital converter," 2014 International Semiconductor Conference (CAS), Sinaia, 2014, pp. 131-134.

[170] R. Henderson, N. Johnston, H. Chen, D. Li, G. Hungerford, R. Hirsch, P. Yip and D. McLoskey, "A 192 x 128 Time Correlated Single Photon Counting Imager in 40nm CMOS Technology," in ESSCIRC, 2018.

[171] M. Gersbach et al., "A parallel 32×32 time-to-digital converter array fabricated in a 130 nm imaging CMOS technology," 2009 Proceedings of ESSCIRC, Athens, 2009, pp. 196-199.

[172] F. Villa et al., "CMOS Imager With 1024 SPADs and TDCs for Single-Photon Timing and 3-D Time-of-Flight," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 20, no. 6, pp. 364-373, Nov.-Dec. 2014, Art no. 3804810.

[173] D. E. Schwartz, E. Charbon and K. L. Shepard, "A Single-Photon Avalanche Diode Array for Fluorescence Lifetime Imaging Microscopy," in IEEE Journal of Solid-State Circuits, vol. 43, no. 11, pp. 2546-2557, Nov. 2008.

[174] S. Lindner, C. Zhang, I. Antolovic, J. Pavia, M. Wolf and E. Charbon, "Column-Parallel Dynamic TDC Reallocation in SPAD Sensor Module Fabricated in 180nm CMOS for Near Infrared Optical Tomography," in International Image Sensors Workshop, 2017.

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

[176] R. M. Field, S. Realov and K. L. Shepard, "A 100 fps, Time-Correlated Single-Photon-Counting-Based Fluorescence-Lifetime Imager in 130 nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 49, no. 4, pp. 867-880, April 2014.

[177] L. Gasparini et al., "A 32×32-pixel time-resolved single-photon image sensor with 44.64μm pitch and 19.48% fill-factor with onchip row/frame skipping features reaching 800kHz observation rate for quantum physics applications," in 2018 IEEE International Solid-State Circuits Conference (ISSCC), pp. 98–100, 2018.

[178] D. Stoppa et al., "A 32x32-pixel array with in-pixel photon counting and arrival time measurement in the analog domain," 2009 Proceedings of ESSCIRC, Athens, 2009, pp. 204-207.

[179] C. Niclass, C. Favi, T. Kluter, F. Monnier and E. Charbon, "Single-photon synchronous detection," ESSCIRC 2008 - 34th European Solid-State Circuits Conference, Edinburgh, 2008, pp. 114-117.

[180] D. Bronzi et al., "100 000 Frames/s 64 × 32 Single-Photon Detector Array for 2-D Imaging and 3-D Ranging," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 20, no. 6, pp. 354-363, Nov.-Dec. 2014, Art no. 3804310.

[181] L. Pancheri, N. Massari, F. Borghetti and D. Stoppa, "A 32x32 SPAD Pixel Array with Nanosecond Gating and Analog Readout," in International Image Sensors Workshop, 2011.

[182] M. Perenzoni, N. Massari, D. Perenzoni, L. Gasparini and D. Stoppa, "A 160×120-pixel analog-counting single-photon imager with Sub-ns time-gating and self-referenced column-parallel A/D conversion for fluorescence lifetime imaging," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, San Francisco, CA, 2015, pp. 1-3.

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

[184] M. Perenzoni, N. Massari, D. Perenzoni, L. Gasparini and D. Stoppa, "A 160×120 Pixel Analog-Counting Single-Photon Imager With Time-Gating and Self-Referenced Column-Parallel A/D Conversion for Fluorescence Lifetime Imaging," in IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 155-167, Jan. 2016.

[185] Y. Maruyama and E. Charbon, "An all-digital, time-gated 128X128 spad array for on-chip, filter-less fluorescence detection," 2011 16th International Solid-State Sensors, Actuators and Microsystems Conference, Beijing, 2011, pp. 1180-1183.

[186] I. Gyongy et al., "256×256, 100kfps, 61% Fill-factor time-resolved SPAD image sensor for microscopy applications," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 8.2.1-8.2.4.

[187] A. Ulku, C. Bruschini, X. Michalet, S. Weiss and E. Charbon, "A 512×512 SPAD Image Sensor with Built-In Gating for Phasor Based Real-Time siFLIM," in International Image Sensors Workshop, 2017.

[188] L. H. C. Braga et al., "An 8×16-pixel 92kSPAD time-resolved sensor with on-pixel 64ps 12b TDC and 100MS/s real-time energy histogramming in 0.13µm CIS technology for PET/MRI applications," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, 2013, pp. 486-487.

[189] R. Walker, L. Braga, A. Erdogan, L. Gasparini, L. Grant, R. Henderson, N. Massari, M. Perenzoni and D. Stoppa," A 92k SPAD Time-Resolved Sensor in 0.13µm CIS Technology for PET/MRI Applications," in International Image Sensors Workshop, 2013.

[190] A. Carimatto et al., "A 67,392-SPAD PVTB-compensated multi-channel digital SiPM with 432 column-parallel 48ps 17b TDCs for endoscopic time-of-flight PET," 2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers, San Francisco, CA, 2015, pp. 1-3.

[191] A. Carimatto, A. Ulku, S. Lindner, E. Aillon, B. Rae, S. Pellegrini and E. Charbon, "A. Carimatto et al., "Multipurpose, fully-integrated 128×128 event-driven MD-SiPM with 512 16-bit TDCs with45ps LSB and 20ns gating," in VLSI, 2018.

[192] M. Perenzoni, D. Perenzoni and D. Stoppa, "A 64×64-pixel digital silicon photomultiplier direct ToF sensor with 100Mphotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6km for spacecraft navigation and landing," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2016, pp. 118-119.

[193] H. Akita, I. Takai, K. Azuma, T. Hata and N. Ozaki, "An imager using 2-D single-photon avalanche diode array in 0.18-µm CMOS for automotive LIDAR application," 2017 Symposium on VLSI Circuits, Kyoto, 2017, pp. C290-C291.

[194] Y. Akasaka and T. Nishimura, "Concept and basic technologies for 3-D IC structure," 1986 International Electron Devices Meeting, Los Angeles, CA, USA, 1986, pp. 488-491.

[195] T. Nishimura et al., "Three dimensional IC for high performance image signal processor," 1987 International Electron Devices Meeting, Washington, DC, USA, 1987, pp. 111-114.

[196] M. Koyanagi, H. Kurino, Kang Wook Lee, K. Sakuma, N. Miyakawa and H. Itani, "Future system-on-silicon LSI chips," in IEEE Micro, vol. 18, no. 4, pp. 17-22, July-Aug. 1998.

[197] J. Michailos et al., "New challenges and opportunities for 3D integrations," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 8.5.1-8.5.4.

[198] V. Suntharalingam et al., "Megapixel CMOS image sensor fabricated in three-dimensional integrated circuit technology," ISSCC. 2005 IEEE International Digest of Technical Papers. Solid-State Circuits Conference, 2005., San Francisco, CA, 2005, pp. 356-357 Vol. 1.

[199] J. A. Burns et al., "A wafer-scale 3-D circuit integration technology," in IEEE Transactions on Electron Devices, vol. 53, no. 10, pp. 2507-2516, Oct. 2006.

[200] S. Sukegawa et al., "A 1/4-inch 8Mpixel back-illuminated stacked CMOS image sensor," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, 2013, pp. 484-485.

[201] H. Tsugawa et al., "Pixel/DRAM/logic 3-layer stacked CMOS image sensor technology," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 3.2.1-3.2.4.

[202] T. Haruta et al., "A 1/2.3inch 20Mpixel 3-layer stacked CMOS Image Sensor with DRAM," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2017, pp. 76-77.

[203] J. Aoki et al., "A rolling-shutter distortion-free 3D stacked image sensor with −160dB parasitic light sensitivity in-pixel storage node," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, 2013, pp. 482-483.

[204] Liang Wang et al., "Direct Bond Interconnect (DBI®) for fine-pitch bonding in 3D and 2.5D integrated circuits," 2017 Pan Pacific Microelectronics Symposium (Pan Pacific), Kauai, HI, 2017, pp. 1-6.

[205] S. Lhostis et al., "Reliable 300 mm Wafer Level Hybrid Bonding for 3D Stacked CMOS Image Sensors," 2016 IEEE 66th Electronic Components and Technology Conference (ECTC), Las Vegas, NV, 2016, pp. 869-876.

[206] Y. Kagawa et al., "Novel stacked CMOS image sensor with advanced Cu2Cu hybrid bonding," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 8.4.1-8.4.4.

[207] K. Nishimura et al., "An over 120dB simultaneous-capture wide-dynamic-range 1.6e− ultra-low-reset-noise organic-photoconductive-film CMOS image sensor," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2016, pp. 110-111.

[208] L. Barrow, N. Bock, A. Bouvier, D. Clocchiatti, J. Feng, N. Kolli, A. Pattantyus, V. Sharma, T. Shu and E. Mandelli, "A QuantumFilm based QuadVGA 1.5µm pixel image sensor with over 40% QE at 940 nm for actively illuminated applications," in International Image Sensor Workshop, 2017.

[209] R. Fontaine, "The State of the Art of Mainstream CMOS Image Sensors," in International Image Sensors Workshop, 2015.

[210] R. Fontaine, "A Survey of Enabling Technologies in Successful Consumer Digital Imaging Products," in International Image Sensors Workshop, 2017.

[211] B. Aull, "3D Imaging with Geiger-mode Avalanche Photodiodes," in Optics and Photonics News, May 2005.

[212] B. Aull et al., "Laser Radar Imager Based on 3D Integration of Geiger-Mode Avalanche Photodiodes with Two SOI Timing Circuit Layers," 2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers, San Francisco, CA, 2006, pp. 1179-1188.

[213] Schuette, Daniel R. et al. "Hybridization process for backilluminated silicon Geiger-mode avalanche photodiode arrays." Advanced Photon Counting Techniques IV. Ed. Mark A. Itzler & Joe C. Campbell. Orlando, Florida, USA: SPIE, 2010.

[214] B. F. Aull, D. R. Schuette, D. J. Young, D. M. Craig, B. J. Felton, and K. Warner, "A Study of Crosstalk in a 256x256 Photon Counting Imager Based on Silicon Geiger-Mode Avalanche Photodiodes," IEEE Sens. J., vol. 15, no. 4, pp. 2123–2132, Apr. 2015.

[215] B. Aull, "Geiger-Mode Avalanche Photodiode Arrays Integrated to All-Digital CMOS Circuits," Sensors (Basel), 2016.

[216] B. F. Aull, E. K. Duerr, J. P. Frechette, K. A. McIntosh, D. R. Schuette and R. D. Younger, "Large-Format Geiger-Mode Avalanche Photodiode Arrays and Readout Circuits," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 24, no. 2, pp. 1-10, March-April 2018, Art no. 3800510.

[217] B. Bérubé et al., "Development of a single photon avalanche diode (SPAD) array in high voltage CMOS 0.8 µm dedicated to a 3D integrated circuit (3DIC)," 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), Anaheim, CA, 2012, pp. 1835-1839.

[218] B. Bérubé et al., "Implementation Study of Single Photon Avalanche Diodes (SPAD) in 0.8µm HV CMOS Technology," in IEEE Transactions on Nuclear Science, vol. 62, no. 3, pp. 710-718, June 2015.

[219] M. -. Tétrault et al., "Real-Time Discrete SPAD Array Readout Architecture for Time of Flight PET," in IEEE Transactions on Nuclear Science, vol. 62, no. 3, pp. 1077-1082, June 2015.

[220] Yu Zou, D. Bronzi, F. Villa and S. Weyers, "Backside illuminated wafer-to-wafer bonding single photon avalanche diode array," 2014 10th Conference on Ph.D. Research in Microelectronics and Electronics (PRIME), Grenoble, 2014, pp. 1-4.

[221] J. M. Pavia, M. Scandini, S. Lindner, M. Wolf and E. Charbon, "A 1 × 400 Backside-Illuminated SPAD Sensor With 49.7 ps Resolution, 30 pJ/Sample TDCs Fabricated in 3D CMOS Technology for Near-Infrared Optical Tomography," in IEEE Journal of Solid-State Circuits, vol. 50, no. 10, pp. 2406-2418, Oct. 2015.

[222] M. -. Lee et al., "A back-illuminated 3D-stacked single-photon avalanche diode in 45nm CMOS technology," 2017 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2017, pp. 16.6.1-16.6.4.

[223] M. Lee et al., "High-Performance Back-Illuminated Three-Dimensional Stacked Single-Photon Avalanche Diode Implemented in 45-nm CMOS Technology," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 24, no. 6, pp. 1-9, Nov.-Dec. 2018, Art no. 3801809.

[224] A. R. Ximenes, P. Padmanabhan, M. Lee, Y. Yamashita, D. N. Yaung and E. Charbon, "A 256×256 45/65nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6dB interference suppression," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, 2018, pp. 96-98.

[225] A. Rochas et al., "First fully integrated 2-D array of single-photon detectors in standard CMOS technology," in IEEE Photonics Technology Letters, vol. 15, no. 7, pp. 963-965, July 2003.

[226] D. Mosconi, D. Stoppa, L. Pancheri, L. Gonzo and A. Simoni, "CMOS Single-Photon Avalanche Diode Array for Time-Resolved Fluorescence Detection," 2006 Proceedings of the 32nd European Solid-State Circuits Conference, Montreux, 2006, pp. 564-567.

[227] 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," 2012 IEEE Nuclear Science Symposium and Medical Imaging Conference Record (NSS/MIC), Anaheim, CA, 2012, pp. 1945-1948.

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

[229] R. Brain, "Interconnect scaling: Challenges and opportunities," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 9.3.1-9.3.4.

[230] N. Teranishi, H. Watanabe, T. Ueda and N. Sengoku, "Evolution of optical structure in image sensors," 2012 International Electron Devices Meeting, San Francisco, CA, 2012, pp. 24.1.1-24.1.4.

[231] I. Gyongy, A. Davies, B. Gallinet, N.A.W. Dutton, R. Duncan, C. Rickman, R. Henderson, and P. Dalgarno, "Cylindrical microlensing for enhanced collection efficiency of small pixel SPAD arrays in single-molecule localisation microscopy," Opt. Express 26, 2280-2291 (2018)

[232] J. M. Pavia, M. Wolf, and E. Charbon, "Measurement and modeling of microlenses fabricated on single-photon avalanche diode arrays for fill factor recovery," Opt. Express 22(4), 4202–4213 (2014).

[233] E. Panina, G. Dalla Betta, L. Pancheri and D. Stoppa, "Design of CMOS Gated Analog Readout Circuits for SPAD Pixel Arrays," PRIME 2012; 8th Conference on Ph.D. Research in Microelectronics & Electronics, Aachen, Germany, 2012, pp. 1-4.

[234] O. Maciu, W. Uhring, J. Le Normand, J. Kammerer, F. Dadouche, N. Dumas, “Sub-nanosecond Gating of Large CMOS Imagers”, Sensors & Transducers Journal, IFSA Publishing, pp. 41-49, Vol. 193, No 10, October 2015.

[235] O. Maciu et al., "Sub-nanosecond gated photon counting for high spatial resolution CMOS imagers," 2016 14th IEEE International New Circuits and Systems Conference (NEWCAS), Vancouver, BC, 2016, pp. 1-4.

[236] N. A. W. Dutton, “A CMOS SPAD-based Image Sensor for Single Photon Counting and Time of Flight Imaging,” The University of Edinburgh, 2015.

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

[238] S. Cheng, C. Chang, K. Lin, C. Huang, L. Tseng, H. Yang, K. Wu and J. Hseih, "Lens Solution for Intensity Enhancement in Large-Pixel Single-Photon Avalanche Diodes," in International Image Sensors Workshop, 2017.

[239] L. Carrara, C. Niclass, N. Scheidegger, H. Shea and E. Charbon, "A gamma, x-ray and high energy proton radiation-tolerant CIS for space applications," 2009 IEEE International Solid-State Circuits Conference - Digest of Technical Papers, San Francisco, CA, 2009, pp. 40-41,41a.

[240] R. J. Walker, J. A. Richardson and R. K. Henderson, "A 128×96 pixel event-driven phase-domain ΔΣ-based fully digital 3D camera in 0.13μm CMOS imaging technology," 2011 IEEE International Solid-State Circuits Conference, San Francisco, CA, 2011, pp. 410-412.

[241] D. Portaluppi, E. Conca and F. Villa, "32 × 32 CMOS SPAD Imager for Gated Imaging, Photon Timing, and Photon Coincidence," in IEEE Journal of Selected Topics in Quantum Electronics, vol. 24, no. 2, pp. 1-6, March-April 2018, Art no. 3800706.

[242] H. Ruokamo, L. Hallman, H. Rapakko and J. Kostamovaara, "An 80 × 25 pixel CMOS single-photon range image sensor with a flexible on-chip time gating topology for solid state 3D scanning," ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference, Leuven, 2017, pp. 59-62.

[243] H. Ouh and M. L. Johnston, "Dual-mode, in-pixel linear and single-photon avalanche diode readout for low-light dynamic range extension in photodetector arrays," 2018 IEEE Custom Integrated Circuits Conference (CICC), San Diego, CA, 2018, pp. 1-4.

[244] I. Nissinen et al., "A sub-ns time-gated CMOS single photon avalanche diode detector for Raman spectroscopy," 2011 Proceedings of the European Solid-State Device Research Conference (ESSDERC), Helsinki, 2011, pp. 375-378.

[245] L. Pancheri, E. Panina, G. Dalla Betta, L. Gasparini and D. Stoppa, "Compact analog counting SPAD pixel with 1.9% PRNU and 530ps time gating," 2013 Proceedings of the ESSCIRC (ESSCIRC), Bucharest, 2013, pp. 295-298.

[246] I. Nissinen, J. Nissinen, J. Holma and J. Kostamovaara, "A TDC-based 4×128 CMOS SPAD array for time-gated Raman spectroscopy," ESSCIRC 2014 - 40th European Solid State Circuits Conference (ESSCIRC), Venice Lido, 2014, pp. 139-142.

[247] Y. Maruyama, J. Blacksberg and E. Charbon, "A 1024×8, 700-ps Time-Gated SPAD Line Sensor for Planetary Surface Exploration With Laser Raman Spectroscopy and LIBS," in IEEE Journal of Solid-State Circuits, vol. 49, no. 1, pp. 179-189, Jan. 2014.

[248] M. Seo, Y. Shirakawa, Y. Kawata, K. Kagawa, K. Yasutomi and S. Kawahito, "A Time-Resolved Four-Tap Lock-In Pixel CMOS Image Sensor for Real-Time Fluorescence Lifetime Imaging Microscopy," in IEEE Journal of Solid-State Circuits, vol. 53, no. 8, pp. 2319-2330, Aug. 2018.

[249] E. R. Fossum, “Gigapixel Digital Film Sensor (DFS) Proposal,” in Nanospace Manipulation of Photons and Electrons for Nanovision Systems, Proceedings of., 2005.

[250] F. Hurter and V. C. Driffield, “Photo-chemical investigations and a new method of the sensitiveness of photographic plates,” J. Soc. Chem. Ind.455–469 (1890).

[251] J. Ma and E. R. Fossum, "Quanta Image Sensor Jot With Sub 0.3e- r.m.s. Read Noise and Photon Counting Capability," in IEEE Electron Device Letters, vol. 36, no. 9, pp. 926-928, Sept. 2015.

[252] M. Seo, S. Kawahito, K. Kagawa and K. Yasutomi, "A 0.27e-rms Read Noise 220-µV/e-Conversion Gain Reset-Gate-Less CMOS Image Sensor With 0.11-µm CIS Process," in IEEE Electron Device Letters, vol. 36, no. 12, pp. 1344-1347, Dec. 2015.

[253] N. Teranishi, "Required Conditions for Photon-Counting Image Sensors," in IEEE Transactions on Electron Devices, vol. 59, no. 8, pp. 2199-2205, Aug. 2012.

[254] J. Ma, S. Masoodian, D. Starkey and E. Fossum, "Photon-number-resolving megapixel image sensor at room temperature without avalanche gain ," in OSA Optica, 2017.

[255] S. Masoodian, A. Rao, J. Ma, K. Odame and E. R. Fossum, "A 2.5 pJ/b Binary Image Sensor as a Pathfinder for Quanta Image Sensors," in IEEE Transactions on Electron Devices, vol. 63, no. 1, pp. 100-105, Jan. 2016.

[256] S. Masoodian, J. Ma, D. Starkey, Y. Yamashita and E. Fossum, "A 1Mjot 1040fps 0.22e-rms Stacked BSI Quanta Image Sensor with Cluster-Parallel Readout," in International Image Sensors Workshop, 2017.

[257] I. Antolovic, S. Burri, R. Hoebe, Y. Maruyama, C. Bruschini and E. Charbon, "Photon-Counting Arrays for Time-Resolved Imaging," Sensors (Basel), 2016.

[258] I. Antolovic, S. Burri, C. Bruschini, R. Hoebe and E. Charbon, "SPAD imagers for super resolution localization microscopy enable analysis of fast fluorophore blinking," Scientific Reports 7, Article number: 44108, March 2017.

[259] I. Gyongy, A. Davies, N. Dutton, R. Duncan, C. Rickman, R, Henderson and P. Dalgarno, "Smart-aggregation imaging for single molecule localisation with SPAD cameras," Scientific Reports 6, Article number: 37349, November 2016.

[260] I. Gyongy, N. Dutton and R. Henderson, "Single-Photon Tracking for High-Speed Vision," in Sensors (Basel), 2018.

[261] N. Ricquier and B. Dierickx, "Pixel structure with logarithmic response for intelligent and flexible imager architectures," ESSDERC '92: 22nd European Solid State Device Research conference, Leuven, Belgium, 1992, pp. 631-634.

[262] N. Tu, R. Hornsey and S. G. Ingram, "CMOS active pixel image sensor with combined linear and logarithmic mode operation," Conference Proceedings. IEEE Canadian Conference on Electrical and Computer Engineering (Cat. No.98TH8341), Waterloo, Ontario, Canada, 1998, pp. 754-757 vol.2.

[263] S. Kavadias, B. Dierickx, D. Scheffer, A. Alaerts, D. Uwaerts and J. Bogaerts, "A logarithmic response CMOS image sensor with on-chip calibration," in IEEE Journal of Solid-State Circuits, vol. 35, no. 8, pp. 1146-1152, Aug. 2000.

[264] G. G. Storm, J. E. D. Hurwitz, D. Renshaw, K. M. Findlater, R. K. Henderson and M. D. Purcell, "Combined linear-logarithmic CMOS image sensor," 2004 IEEE International Solid-State Circuits Conference (IEEE Cat. No.04CH37519), San Francisco, CA, 2004, pp. 116-517 Vol.1.

[265] H. Cheng, B. Choubey and S. Collins, "An Integrating Wide Dynamic-Range Image Sensor With a Logarithmic Response," in IEEE Transactions on Electron Devices, vol. 56, no. 11, pp. 2423-2428, Nov. 2009.

[266] S. Decker, R. McGrath, K. Brehmer and C. Sodini, "A 256×256 CMOS imaging array with wide dynamic range pixels and column-parallel digital output," 1998 IEEE International Solid-State Circuits Conference. Digest of Technical Papers, ISSCC. First Edition (Cat. No.98CH36156), San Francisco, CA, USA, 1998, pp. 176-177.

[267] N. Akahane, R. Ryuzaki, S. Adachi, K. Mizobuchi and S. Sugawa, "A 200dB Dynamic Range Iris-less CMOS Image Sensor with Lateral Overflow Integration Capacitor using Hybrid Voltage and Current Readout

Operation," 2006 IEEE International Solid State Circuits Conference - Digest of Technical Papers, San Francisco, CA, 2006, pp. 1161-1170.

[268] T. Lule, C. Mandier, A. Glais, G. Roffet, R. Monteith, B. Deschamps and D. Herault, "High Performance 1.3MPix HDR Automotive Image Sensor," in International Image Sensors Workshop, 2015.

[269] T. Lule, H. Keller, M. Wagner and M. Bohm, "100,000 pixel 120 dB imager in TFA-technology," 1999 Symposium on VLSI Circuits. Digest of Papers (IEEE Cat. No.99CH36326), Kyoto, Japan, 1999, pp. 133-136.

[270] D. Stoppa, A. Simoni, L. Gonzo, M. Gottardi and G. -. Dalla Betta, "Novel CMOS image sensor with a 132-dB dynamic range," in IEEE Journal of Solid-State Circuits, vol. 37, no. 12, pp. 1846-1852, Dec. 2002.

[271] D. Stoppa, M. Vatteroni, D. Covi, A. Baschirotto, A. Sartori and A. Simoni, "A 120-dB Dynamic Range CMOS Image Sensor With Programmable Power Responsivity," in IEEE Journal of Solid-State Circuits, vol. 42, no. 7, pp. 1555-1563, July 2007.

[272] W. Yang, "A wide-dynamic-range, low-power photosensor array," Proceedings of IEEE International Solid-State Circuits Conference - ISSCC '94, San Francisco, CA, USA, 1994, pp. 230-231.

[273] J. Raynor and P. Seitz, "A linear array of photodetectors with wide dynamic range and near photon quantum-noise limit," Sensors and Actuators A: Physical, June 1997.

[274] J. Raynor, A. Scott, C. Holyoake and D. Reay, "A single-exposure linear HDR 17-bit hybrid 50µm analogue-digital pixel in 90nm BSI," in International Image Sensors Workshop, 2015.

[275] T. Nakamura and K. Saitoh, "Recent Progress of CMD Imaging," in IEEE Workshop on Charge-Coupled Devices, 1997.

[276] O. Yadid-Pecht and E. R. Fossum, "Wide intrascene dynamic range CMOS APS using dual sampling," in IEEE Transactions on Electron Devices, vol. 44, no. 10, pp. 1721-1723, Oct. 1997.

[277] D. X. D. Yang, A. E. Gamal, B. Fowler and H. Tian, "A 640×512 CMOS image sensor with ultrawide dynamic range floating-point pixel-level ADC," in IEEE Journal of Solid-State Circuits, vol. 34, no. 12, pp. 1821-1834, Dec. 1999.

[278] B. J. Hosticka et al., "CMOS imaging for automotive applications," in IEEE Transactions on Electron Devices, vol. 50, no. 1, pp. 173-183, Jan. 2003.

[279] B. Fowler, "High dynamic range image sensor architectures," Proc. SPIE 7876, Digital Photography VII, 787602 (24 January 2011).

[280] J. Solhusvik, J. Kuang, Z. Lin, S. Manabe, J. Lyu and H. Rhodes, "A comparison of high dynamic range CIS technologies for automotive applications," in International Image Sensors Workshop, 2013.

[281] M. Sasaki, M. Mase, S. Kawahito and Y. Tadokoro, "A Wide-Dynamic-Range CMOS Image Sensor Based on Multiple Short Exposure-Time Readout With Multiple-Resolution Column-Parallel ADC," in IEEE Sensors Journal, vol. 7, no. 1, pp. 151-158, Jan. 2007.

[282] M. Seo et al., "A Low-Noise High Intrascene Dynamic Range CMOS Image Sensor With a 13 to 19b Variable-Resolution Column-Parallel Folding-Integration/Cyclic ADC," in IEEE Journal of Solid-State Circuits, vol. 47, no. 1, pp. 272-283, Jan. 2012.

[283] D. Chitnis and S. Collins, “Compact readout circuits for SPAD arrays,” in Proceedings of 2010 IEEE International Symposium on Circuits and Systems, 2010, pp. 357–360.

[284] M. Mori et al., "A 1280×720 single-photon-detecting image sensor with 100dB dynamic range using a sensitivity-boosting technique," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2016, pp. 120-121.

[285] M. Mori et al., "An APD-CMOS image sensor toward high sensitivity and wide dynamic range," 2016 IEEE International Electron Devices Meeting (IEDM), San Francisco, CA, 2016, pp. 8.3.1-8.3.4.

[286] H. Ouh, S. Sengupta, S. Bose and M. L. Johnston, "Dual-mode, enhanced dynamic range CMOS optical sensor for biomedical applications," 2017 IEEE Biomedical Circuits and Systems Conference (BioCAS), Turin, 2017, pp. 1-4.

[287] E. R. Fossum, "Modeling the Performance of Single-Bit and Multi-Bit Quanta Image Sensors," in IEEE Journal of the Electron Devices Society, vol. 1, no. 9, pp. 166-174, Sept. 2013.

[288] I. Gyongy, N. Dutton, L. Parmesan, A. Davies, R. Saleeb, R. Duncan, C. Rickman, P. Dalgarno and R. Henderson, "Bit-plane Processing Techniques for Low-Light, High Speed Imaging with a SPAD-based QIS," in International Image Sensors Workshop, 2015.

[289] D. Li, S. Ameer-Beg, J. Arlt, D. Tyndall, R. Walker, D. Matthews, V. Visitkul, J. Richardson and R. Henderson, "Time-Domain Fluorescence Lifetime Imaging Techniques Suitable for Solid-State Imaging Sensor Arrays," in Sensors (Basel), 2012.

[290] C. Chao, C. Chang, M. Mhala, P. Chou, H. Tu, S. Yeh, K. Chou, C. Liu and F. Hsueh, "Detection and Shielding of Photon Emission in Stacked CIS," in International Image Sensors Workshop, 2015

[291] I. M. Antolovic, S. Burri, C. Bruschini, R. Hoebe and E. Charbon, "Nonuniformity Analysis of a 65-kpixel CMOS SPAD Imager," in IEEE Transactions on Electron Devices, vol. 63, no. 1, pp. 57-64, Jan. 2016.

[292] D. Li et al., "FPGA implementation of a video-rate fluorescence lifetime imaging system with a 32×32 CMOS single-photon avalanche diode array," 2009 IEEE International Symposium on Circuits and Systems, Taipei, 2009, pp. 3082-3085.

[293] C. S. Bamji et al., "A 0.13 μm CMOS System-on-Chip for a 512 × 424 Time-of-Flight Image Sensor With Multi-Frequency Photo-Demodulation up to 130 MHz and 2 GS/s ADC," in IEEE Journal of Solid-State Circuits, vol. 50, no. 1, pp. 303-319, Jan. 2015.

[294] T. Yamazaki et al., "A 1ms high-speed vision chip with 3D-stacked 140GOPS column-parallel PEs for spatio-temporal image processing," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2017, pp. 82-83.

[295] T. Takahashi et al., "A 4.1Mpix 280fps stacked CMOS image sensor with array-parallel ADC architecture for region control," 2017 Symposium on VLSI Circuits, Kyoto, 2017, pp. C244-C245.

[296] M. Perenzoni, L. Gasparini and D. Stoppa, "Design and Characterization of a 43.2-ps and PVT-Resilient TDC for Single-Photon Imaging Arrays," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 65, no. 4, pp. 411-415, April 2018.

[297] T. Iwata, T. Taga and T. Mizuno, "FPGA-based photon-counting phase-modulation fluorometer and a brief comparison with that operated in a pulsed-excitation mode," The Optical Society of Japan, 2017.

[298] T. Iwata and T. Mizuno, "High-speed, FPGA-based photon-counting fluorometer with high data-gathering efficiency," Measurement Science and Technology, Volume 28, Number 7, June 2017.

[299] C. Lin and C. Hsieh, "A Dual-Mode CMOS Imager for Free-Space Optical Communication with Signal Light Source Tracking and Background Cancellation," in International Image Sensors Workshop, 2015.

[300] Mai, H., Gyongy, I., Dutton, N. A. W., Henderson, R. K., and Underwood, I. (2018) Characterization of electronic displays using CMOS single-photon avalanche diode image sensors. Jnl Soc Info Display, 26: 255–261.

[301] M. Seo and S. Kawahito, "A 7 ke-SD-FWC 1.2 e-RMS Temporal Random Noise 128×256 Time-Resolved CMOS Image Sensor With Two In-Pixel SDs for Biomedical Applications," in IEEE Transactions on Biomedical Circuits and Systems, vol. 11, no. 6, pp. 1335-1343, Dec. 2017.

[302] T. D. Huang, S. Sorgenfrei, P. Gong, R. Levicky and K. L. Shepard, "A 0.18μm CMOS Array Sensor for Integrated Time-Resolved Fluorescence Detection," in IEEE Journal of Solid-State Circuits, vol. 44, no. 5, pp. 1644-1654, May 2009.

[303] J. Guo and S. Sonkusale, "A CMOS imager with digital phase readout for fluorescence lifetime imaging," 2011 Proceedings of the ESSCIRC (ESSCIRC), Helsinki, 2011, pp. 115-118.

[304] H. Ingelberts and M. Kuijk, "High-speed gated CMOS detector for fluorescence lifetime microscopy extending to near-infrared wavelengths," 2015 IEEE SENSORS, Busan, 2015, pp. 1-4.

[305] G. Fu and S. Sonkusale, "CMOS sensor for dual fluorescence intensity and lifetime sensing using multicycle charge modulation," 2017 IEEE Custom Integrated Circuits Conference (CICC), Austin, TX, 2017, pp. 1-4.

[306] Z. Li et al., "A Time-Resolved CMOS Image Sensor With Draining-Only Modulation Pixels for Fluorescence Lifetime Imaging," in IEEE Transactions on Electron Devices, vol. 59, no. 10, pp. 2715-2722, Oct. 2012.

[307] M. Seo et al., "A 10 ps Time-Resolution CMOS Image Sensor With Two-Tap True-CDS Lock-In Pixels for Fluorescence Lifetime Imaging," in IEEE Journal of Solid-State Circuits, vol. 51, no. 1, pp. 141-154, Jan. 2016.

[308] Z. Li, M. Seo, K. Kagawa, K. Yasutomi and S. Kawahito, "CMOS image sensor with lateral electric field modulation pixels for fluorescence lifetime imaging with sub-nanosecond time response," Japanese Journal of Applied Physics, 2016.

[309] C. Bruschini, H. Homulle and E. Charbon, "Ten years of biophotonics single-photon SPAD imager applications - retrospective and outlook," Proc. of SPIE, Vol. 10069, 2017.

[310] M. Perenzoni, M. Moreno-García, L. Gasparini and N. Massari, "Time-resolved single-photon detectors: Integration challenges in CMOS technologies," 2018 International Conference on IC Design & Technology (ICICDT), Otranto, 2018, pp. 197-200.

[311] E. Charbon, C. Bruschini and M. Lee, "3D-Stacked CMOS SPAD Image Sensors: Technology and Applications," IEEE International Conference on Electronics, Circuits and Systems (ICECS), 2018.

[312] T. Takahashi et al., "A Stacked CMOS Image Sensor With Array-Parallel ADC Architecture," in IEEE Journal of Solid-State Circuits, vol. 53, no. 4, pp. 1061-1070, April 2018.

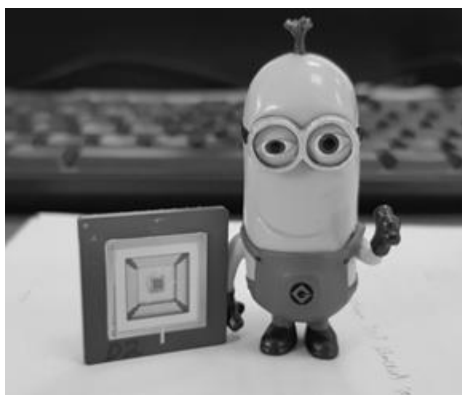
[313] S. Masoodian, J. Ma, D. Starkey, T. Wang, Y. Yamashita and E. Fossum, "Room Temperature 1040fps, 1 Megapixel Photon-Counting Image Sensor with 1.1um Pixel Pitch," Proc. of SPIE, Vol 10212, 2017.

[314] L. Pancheri et al., "First Demonstration of a Two-Tier Pixelated Avalanche Sensor for Charged Particle Detection," in IEEE Journal of the Electron Devices Society, vol. 5, no. 5, pp. 404-410, Sept. 2017.

[315] Y. Takemoto et al., "Multi-storied photodiode CMOS image sensor for multiband imaging with 3D technology," 2015 IEEE International Electron Devices Meeting (IEDM), Washington, DC, 2015, pp. 30.1.1-30.1.4.

[316] L. Anzagiraand E. Fossum, "Two Layer Image Sensor Pixel Concept for Enhanced Low Light Color Imaging," in International Image Sensors Workshop, 2015.

[317] C. Niclass, M. Soga, H. Matsubara, S. Kato and M. Kagami, "A 100-m Range 10-Frame/s 340×96-Pixel Time-of-Flight Depth Sensor in 0.18µm CMOS," in IEEE Journal of Solid-State Circuits, vol. 48, no. 2, pp. 559-572, Feb. 2013.



Lena embracing a MINI3D sensor after a long day in the lab.

A special Minion repeatedly featured in recent publications by the author and the CSS group prompting the joke of calling him Lena. Throughout 4 years of research, the yellow fellow was subjected to extensive hours of cold and severely bright environments, ridiculous laser power levels and even got involved in few toy car accidents for experimental imaging purposes. Despite the harsh treatment, Lena never complained and always kept a smile on his face, therefore his integral role in the completion of this work is dearly acknowledged.

First draft submitted on the 31st of August 2018

Final draft Submitted on the 30th of May 2019