High-Speed Wide-Field Time-Correlated Single-Photon Counting Fluorescence Lifetime Imaging Microscopy

Ryan Michael Field

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2014

Abstract

High-Speed Wide-Field Time-Correlated Single-Photon Counting Fluorescence Lifetime Imaging Microscopy

Ryan Michael Field

Fluorescence microscopy is a powerful imaging technique used in the biological sciences to identify labeled components of a sample with specificity. This is usually accomplished through labeling with fluorescent dyes, isolating these dyes by their spectral signatures with optical filters, and recording the intensity of the fluorescent response. Although these techniques are widely used, fluorescence intensity images can be negatively affected by a variety of factors that impact the fluorescence intensity. Fluorescence lifetime imaging microscopy (FLIM) is an imaging technique that is relatively immune to intensity fluctuations and also provides the unique ability to directly monitor the microenvironment surrounding a fluorophore.

Despite the benefits associated with FLIM, the applications to which it is applied are fairly limited due to long image acquisition times and high cost of traditional hardware. Recent advances in complementary metal-oxide-semiconductor (CMOS) single-photon avalanche diodes (SPADs) have enabled the design of low-cost imaging arrays that are capable of recording lifetime images with acquisition times greater than one order of magnitude faster than existing systems. However, these SPAD arrays have yet to realize the full po-

tential of the technology due to limitations in their ability to handle the vast amount of data generated during the commonly used time-correlated single-photon counting (TCSPC) lifetime imaging technique.

This thesis presents the design, implementation, characterization, and demonstration of a high speed FLIM imaging system. The components of this design include a CMOS imager chip in a standard 0.13μm technology containing a custom CMOS SPAD, a 64-by-64 array of these SPADs, pixel control circuitry, independent time-to-digital converters (TDCs), a FLIM specific datapath, and high bandwidth output buffers. In addition to the CMOS imaging array, a complete system was designed and implemented using a printed circuit board (PCB) for capturing data from the imager, creating histograms for the photon arrival data using field-programmable gate arrays, and transferring the data to a computer using a cabled PCIe interface. Finally, software is used to communicate between the imaging system and a computer.

The dark count rate of the SPAD was measured to be only 231 Hz at room temperature while maintaining a photon detection probability of up to 30%. TDCs included on the array have a 62.5 ps resolution and a 64 ns range, which is suitable for measuring the lifetime of most biological fluorophores. Additionally, the on-chip datapath was designed to handle continuous data transfers at rates capable of supporting TCSPC-based lifetime imaging at 100 frames per second. The system level implementation also provides sufficient data throughput for transferring up to 750 frames per second from the imaging system to a computer.

The lifetime imaging system was characterized using standard techniques for evaluating SPAD performance and an electrical delay signal for measuring the TDC performance.

This thesis concludes with a demonstration of TCSPC-FLIM imaging at 100 frames per second – the fastest 64-by-64 TCSPC FLIM that has been demonstrated. This system overcomes some of the limitations of existing FLIM systems and has the potential to enable new application domains in dynamic FLIM imaging.

Contents

List of Figures **v**

List of Tables **x**

List of Acronyms **xi**

Acknowledgments **xiii**

Chapter 1 Introduction **1**

Chapter 2 Fluorescence Lifetime Imaging Microscopy **5**

2.1 Fluorescence 5

2.1.1 Fluorescence Lifetime 9

2.2 Fluorescence Lifetime Measurements 12

2.2.1 Frequency-Domain Lifetime Measurements 12

2.2.2 Time-Domain Lifetime Measurements 17

2.3 System Requirements 22

Chapter 3 Single-Photon Detectors **24**

3.1 Photomultiplier Tubes 24

3.2 Avalanche Photodiodes 25

3.3 CMOS Single-Photon Avalanche Diodes 28

3.4 Optimal Biasing of SPADS for FLIM 31

3.4.1 Non-homogeneous Poisson Process 32
3.4.2 Probability of detecting a true positive event 32
3.4.3 Probability of recording a true negative event 34
3.4.4 Probability of detecting an arriving photon 35
3.4.5 Figure-of-Merit 37
3.4.6 Comparison to Simulated FLIM Data 37
3.5 CMOS SPAD Implementation 38
3.5.1 SPAD Design 39
3.5.2 SPAD Characterization 42
3.6 Summary 45

Chapter 4 Wide Field Fluorescence Lifetime Imager Integrated Circuit Design 47

4.1 Imager Design 49
4.1.1 Imager Architecture Overview 49
4.1.2 Pixel Circuitry 51
4.1.3 Time-to-Digital Converters 57
4.1.4 FLIM Datapath 79
4.1.5 Output Buffers 85
4.1.6 Phase-Locked Loop 87
4.1.7 Imager Control 89
4.1.8 Additional Circuits 93
4.2 Integrated Circuit Characterization 93
4.2.1 Test Pixel Measurements 95
4.2.2 Pixel Control Measurements 99
4.2.3 SPAD Array Measurements 100
4.2.4 PLL Measurements 102
4.2.5 TDC Measurements 102

4.2.6 Impulse Response of SPAD and TDC . . . 107
4.2.7 LVDS Buffer Measurements . . . 111
4.2.8 Preliminary Images . . . 112
4.2.9 FIFO Controller Bug . . . 114
4.2.10 Column Counter Bug . . . 114
4.2.11 Power Consumption . . . 114
4.3 Summary . . . 115

Chapter 5 Fluorescence Lifetime Imaging Microscopy System Design 117

5.1 System Overview . . . 117
5.2 IC Packaging . . . 118
5.3 FPGAs . . . 122
5.3.1 FPGA Characteristics . . . 123
5.3.2 FPGA Architecture . . . 125
5.4 Printed Circuit Board . . . 137
5.4.1 Power Conversion and Distribution . . . 138
5.4.2 Auxiliary Circuits . . . 139
5.4.3 Liquid Cooling . . . 140
5.5 Software . . . 142
5.5.1 Linux Kernel Module & Device Driver . . . 142
5.5.2 C Application Programming Interface . . . 146
5.5.3 Graphical User Interface . . . 147
5.5.4 Lifetime Extraction . . . 148
5.6 Imaging Results . . . 150
5.6.1 Array-Wide Dark Count Rate . . . 150
5.6.2 Lifetime Measurement Setup . . . 153
5.6.3 Lifetime Imaging Results . . . 155
5.7 Summary . . . 162

Chapter 6 Conclusions **163**

6.1 Summary of Contributions 163

6.2 Future Work 164

Bibliography **167**

Appendix A Single-Photon Avalanche Diode Test Chips **178**

A.1 First SPAD Test Chip - 0.13 μm IBM CMOS 178

A.2 Second SPAD Test Chip - 0.35 μm AMS CMOS 180

A.3 Third SPAD Test Chip - 0.13 μm IBM CMOS 180

A.4 Fourth SPAD Test Chip - 0.35 μm AMS CMOS 182

A.5 Fifth SPAD Test Chip - 0.13 μm IBM CMOS 183

A.6 FLIM Array Test Sites 184

A.7 JFET Test Chip - 0.18 μm IBM CMOS 185

List of Figures

2.1 Jablonski diagram of fluorescence transitions. 6

2.2 Absorption and emission spectra for FITC. 8

2.3 Illustration of a common epi-fluorescent microscope. 9

2.4 Illustration of frequency-domain FLIM signal - short lifetime. 13

2.5 Illustration of frequency-domain FLIM signal - long lifetime. 14

2.6 Illustration of time-resolved fluorescence decay measurements. 17

2.7 Illustration of gated-integration lifetime measurement. 18

2.8 Illustration of TCSPC lifetime measurement. 19

3.1 I-V curve for a avalanche photodiode. 26

3.2 SPAD passive quench and reset. 27

3.3 Illustration of the structure of a SPAD device. 28

3.4 Simulation of minimum number of laser repetitions. 39

3.5 Plot of FOM corresponding to simulated results 40

3.6 Illustrated cross-section of CMOS SPAD 41

3.7 Simulated cross-section of SPAD using TCAD software. 42

3.8 A photograph of the SPAD used in this work. 42

3.9 Current-voltage relationship for the fabricated SPAD 43

3.10 Measured dark count rate of SPAD 44

3.11 Photon detection probability of the SPAD 45

3.12 Instrument response function of the SPAD 46

4.1 Block diagram for CMOS imager chip. 48
4.2 Die photograph of FLIM IC. 50
4.3 SPAD pixel layout 51
4.4 Schematic of pixel circuitry. 52
4.5 Timing diagram for pixel event and reset. 56
4.6 Tapped delay-line. 59
4.7 Time-to-digital converter architecture. 61
4.8 Overview of delay-locked loop. 63
4.9 Schematics of alternate delay elements. 64
4.10 Tuning curves for alternative delay elements. 65
4.11 Delay element for DLL and tuning curve. 67
4.12 Schematic of level shifter used in DLL. 68
4.13 Extracted delay tuning curve. 69
4.14 Complementary clock generator schematic. 70
4.15 Complementary clock generator simulated results. 71
4.16 Phase detector schematic. 72
4.17 Phase detector performance simulation. 73
4.18 Calibrated charge pump schematic. 75
4.19 Schematic of linear regulator used in each of the 32 DLLs. 77
4.20 Linear regulator simulation. 78
4.21 Datapath block diagram. 80
4.22 First stage of the datapath. 81
4.23 Overview of data flow on IC. 82
4.24 Datapath stage 1 schematic 83
4.25 LVDS output buffer schematic. 86
4.26 Simulated data eye for LVDS buffer. 87

4.27 Overview of PLL. 88
4.28 Pixel controller timing diagram. 90
4.29 Datapath stage one timing diagram. 91
4.30 Stage two, three, and four controller timing diagram 92
4.31 Printed circuit board for initial testing 95
4.32 Test pixel output events. 96
4.33 Test pixel maximum count rate. 97
4.34 Afterpulsing Probability plots. 98
4.35 Afterpulsing inter-spike interval plot. 99
4.36 Pixel on/off control. 100
4.37 Dark Count Rate for one quadrant of the array. 101
4.38 Sensitivity demonstration for one quadrant of the array. 102
4.39 Charge pump mismatch characteristics. 103
4.40 Charge pump calibration. 104
4.41 TDC Fine Offset. 105
4.42 DNL and INL Measurements. 108
4.43 TDC counter stop synchronizer. 109
4.44 TDC counter stop synchronizer waveforms 109
4.45 DNL using code density method 110
4.46 Impulse response of SPAD and TDC combined 110
4.47 LVDS output buffer data eye. 112
4.48 Photo of fluorescein dye. 113
4.49 Preliminary FLIM image. 115

5.1 Photograph of custom IC package 121
5.2 BGA package artwork layers. 122
5.3 Photograph of package warping 123
5.4 Temperature profile for package solder reflow 124

5.5 Photograph of eliminated package warping 125
5.6 Block diagram of the FPGA architecture 127
5.7 Diagram of the data demultiplexers on the FPGA 129
5.8 Flow diagram of the duplicate data detection block 130
5.9 Data binner finite state machine diagram 132
5.10 Topology of PCIe system 133
5.11 PCIe interface layers from base specification 134
5.12 Block diagram of FPGA PCIe controller 135
5.13 Block diagram of the raw data acquisition system 138
5.14 Photograph of the final system PCB 139
5.15 Photograph of the cooling system 141
5.16 Plot of DCR versus time with cooling system 142
5.17 Main GUI window 148
5.18 Configuration subwindow of GUI 149
5.19 Dark count rate for entire array 150
5.20 Partial array configuration showing gradient in failing rows 151
5.21 Demonstration of improved row distribution with TDCs disabled 152
5.22 Row hit distribution with controller error 153
5.23 Optical setup for array testing 154
5.24 Location of dye sample during imaging 155
5.25 Lifetime image covering the entire SPAD array 156
5.26 Lifetime image with side masked. 157
5.27 Lifetime image with corner masked. 158
5.28 Sixteen frame acquisition at 100 fps 159
5.29 Eight frame acquisition at 100 fps with blocking 160

A.1 Cross-section of SPADs in IBM test chip 1 179
A.2 I-V characteristics of SPADs in IBM test chip 1 180

A.3 I-V characteristics of SPADs in the first AMS test chip 181
A.4 Layout of SPADs in IBM test chip 2 182
A.5 I-V characteristics of SPADs in IBM test chip 2 183
A.6 I-V characteristics of SPADs in IBM test chip 3 184

List of Tables

2.1 Table of resolution-limited pixel sizes for camera imaging 10

2.2 Table of maximum frame rate for published work. 23

4.1 Characteristics of the IBM 0.13 μm technology. 48

4.2 Threshold voltages for available devices. 54

4.3 Subset of charge pump calibration codes. 74

4.4 Power consumption of imaging IC. 115

5.1 Xilinx Virtex-6 LX130T characteristics 126

5.2 Table of maximum payload sizes 136

5.3 Table of BAR offsets for PCIe Controller 145

5.4 Summary of IC characteristics 161

List of Acronyms

ADC	Analog-to-Digital Converter
APD	Avalanche Photodiode
BAR	Base Address Register
CCD	Charge-Coupled Device
CFD	Constant Fraction Discriminator
CMOS	Complementary Metal-Oxide-Semiconductor
CPU	Central Processing Unit
DCR	Dark Count Rate
DLL	Delay-Locked Loop
DMA	Direct Memory Access
DNL	Differential Non-Linearity
DW	Double Word
FET	Field Effect Transistor
FF	Flip-Flop
FIFO	First-In, First-Out
FITC	Fluorescein Isothiocyanate
FLIM	Fluorescence Lifetime Imaging Microscopy
FOM	Figure-of-Merit
FPGA	Field-Programmable Gate Array
FPS	Frames per Second
FRET	Fluorescence (or Förster) Resonance Energy Transfer
IC	Integrated Circuit
ICCD	Intensified Charge-Coupled Device
INL	Integral Non-Linearity

LOCOS	Local Oxidation of Silicon
LSB	Least Significant Bit
LUT	Look-up Table
LVDS	Low-Voltage Differential Signaling
MOS	Metal-Oxide-Semiconductor
MPS	Maximum Payload Size
MSB	Most Significant Bit
N/PFET	N/P-type Field Effect Transistor
N/PMOS	N/P-type Metal-Oxide-Semiconductor
OS	Operating System
PCB	Printed Circuit Board
PCIe	Peripheral Component Interconnect Express
PCH	Platform Controller Hub
PDP	Photon Detection Probability
PLL	Phase-Locked Loop
PMT	Photomultiplier Tube
QW	Quad Word
SFF	Scan Flip-Flop
SPAD	Single-Photon Avalanche Diode
STI	Shallow Trench Isolation
TCSPC	Time-Correlated Single-Photon Counting
TDC	Time-to-Digital Converter

Acknowledgments

I would like to take a moment to thank all of the mentors, colleagues, friends, and family that have supported me throughout the duration of my thesis work.

First, I would like to thank my advisor, Ken Shepard, for encouraging me to come to Columbia for my graduate studies and then providing the support, resources, freedom, and encouragement for me to succeed. Ken's endless enthusiasm for engineering, entrepreneurial insights, and commitment to his students at Columbia and beyond is inspirational. I learned an enormous amount from Ken technically and otherwise and I appreciate his support of the many educational outreach activities I pursued during my studies.

I would also like to thank the professors who served on my committed and took the time to review my work and provide useful feedback: Charles Zukowski, John Kymissis, Liam Paninski, and Elizabeth Hillman. In particular, I'd like to thank Liam for the extended statistical discussions that helped me to develop a deeper understanding of the underlying mathematics.

Additionally, an enormous amount of gratitude should be directed toward the many students I had the opportunity to work with at Columbia. Simeon Realov is not only a great friend but also helped with me in taping out my imager integrated circuit and designed the PLL on the chip. His encouragement and voluntary support through numerous all-night

sessions to meet the submission deadline were more than anyone could expect and I cannot express enough thanks. Matthew Johnston provided technical, business, and life support and was an enthusiastic partner in teaching pursuits and continues to be great friend outside of Columbia. I appreciate his continued friendship and support. I'd also like to thank: Mike Lekas for always asking critical questions and his friendship away from the lab. Eyal Aklimi who provided invaluable help machining parts while making it seem easy. Jared Roseman for many technical discussions and long hours solving server problems with me. Jacob Rosenstein, David Tsai, and Tarun Chari for their interest in my work and help in the lab. Steven Warren and Scott Trocchia for stepping in as the next generation of system administrators for the Shepard Lab computing infrastructure.

I'd also like to thank David Schwartz for introducing me to this work. Sebastian Sorgenfrei, Inanc Meric, Omar Ahmed, and Peter Levine for their mentoring and guidance as I started work on this dissertation.

I'd also like to wish Dan Fleischer and David Gidony luck as they continue building new devices within the constraints of CMOS technologies and improve upon the work done in this thesis.

Cecilia Townsend, Robert Kolbas, Ginger Yu, Maysam Ghovanloo, Erik Heijne, and Randall Victora were all instrumental in helping me realize the joy of engineering research early in my education and for that I'm grateful.

Ria Miranda also deserves recognition for all of her outstanding administrative support over the years. She had a talent for getting things done and I could always count on her when in a bind.

Finally, I'd like to thank my family. They were all infinitely supportive through the

many years I spent working on this dissertation. In particular, I'd like to thank my wonderful wife, Lauren, who, while simultaneously pursuing her Ph.D., was my biggest supporter and ideal companion. She was the perfect distraction when I needed one, a patient reviewer, and is an excellent engineer.

Chapter 1

Introduction

Observations of fluorescence were first recorded during the middle of the nineteenth century by Sir George Stokes, who observed that a glass of tonic water emitted light of a different color than that which was incident on the liquid. This color translation is now known as the Stokes shift and serves as the basis for fluorescence microscopy – an enabling tool for many areas of biological research. Nearly a century after Stokes' observation, advances in the conjugation of fluorescent molecules (or fluorophores) to antibodies allowed fluorescence to be utilized for microscopic imaging, providing specific labeling capability and improved signal-to-background ratios over previous microscopic staining methods [1]. These advantages have made fluorescence microscopy widely useful for biological applications. A few notable applications are the study of neuronal activity using calcium-sensitive fluorescent dyes [2], intracellular monitoring using fluorescent timers [3], membrane dynamics with fluorescence recovery after photobleaching (FRAP) [4], and biological interactions on small scales by recording the co-localization of molecules through ratiometric fluorescence resonance energy transfer (FRET) [5].

Most fluorescence microscopy is performed using an epi-fluorescent microscope where the excitation source and imaging objective lens are positioned on the same side of the sample. Consequently, the excitation light and emitted fluorescence share part of the optical

path through the microscope and a set of filters, which are chosen based on the Stokes shift of the fluorophore being imaged, are required to produce high contrast images. This set includes an excitation filter, which constrains the bandwidth of the excitation light, a dichroic mirror, which separates the excitation light from the fluorescence emission, and an emission filter that is used to block any unwanted intrinsic fluorescence or excitation light that is transmitted by the dichroic mirror. After passing through this filter set, the emission light forms an image on a wide-field sensor or the observer's retina.

There are a number of factors that can present a challenge to capturing high quality fluorescence images. These include: fluctuations in fluorophore concentration within the sample, light source limitations and filter requirements for excitation, cross-talk when imaging multiple fluorophores with multiple detection channels, fluorophore photobleaching, background fluorescence and autofluorescence, detector sensitivity limits, sample preparation requirements, and optical resolution limits [6, 7]. Various techniques can be used to address some of these challenges. Confocal laser scanning microscopy (CLSM) and two-photon laser scanning microscopy (TPLSM) [8, 9, 10, 11] combine specialized lasers and extremely sensitive detectors (such as photomultiplier tubes (PMTs) or avalanche photodiodes (APDs)) to improve background rejection and allow for imaging into thick samples. Super-resolution imaging techniques, including stimulated-emission-depletion (STED) microscopy [12], stochastic optical reconstruction microscopy (STORM) [13], and photoactivated localization microscopy (PALM) [14] continue to push the resolution limits of fluorescence microscopy.

Fluorescence lifetime imaging microscopy (FLIM) is another fluorescence microscopy technique that addresses a subset of the challenges mentioned above. Image contrast in FLIM is based on the temporal response of a fluorophore to an excitation source [15]. Fluorescence lifetimes can be measured either in the frequency domain, using a modulated continuous wave (CW) excitation source, or in the time domain, using a pulsed laser excitation source and measuring the fluorescence intensity decay over time. In the time domain, the intensity

decay is typically measured using either a gated-integration technique or time-correlated single-photon counting (TCSPC). The detector used can either be a wide-field camera [16] or a PMT/APD [17], depending on the measurement approach used. Lifetime imaging is able to address some of the limitations of fluorescence microscopy that are associated with fluorophore intensity fluctuation, background, autofluorescence [18], and cross-talk from multiple dyes. In addition, FLIM can provide a direct measure of the local microenvironment through monitoring shifts in the measured fluorescence lifetime, which are associated with changes in properties like pH [19], charge concentration [20], metabolic state [21, 22], or the presence of other molecules through fluorescence resonance energy transfer (FRET) interactions [23]. FLIM has also found application outside of the biological sciences in the detection of counterfeit currency [24].

The benefits of FLIM have traditionally come at the cost of image acquisition speed. Commercial TCSPC systems can take tens of seconds to capture a single image and frequency domain techniques have achieved imaging speeds of a few hertz but with reduced lifetime measurement accuracy relative to TCSPC [25]. As a result applications leveraging TCSPC FLIM have been limited to recording images of static samples. Recent work has leveraged integrated arrays of complementary metal-oxide-semiconductor (CMOS) single-photon avalanche diodes (SPADs) and time-to-digital converters (TDCs) to create parallelized TCSPC imaging systems [26, 27, 28, 29, 30, 31]. Although improved imaging speeds have been demonstrated in some of these designs, the parallel acquisition channels generate output data rates that limit the achievable frame rate. With improvements in TCSPC hardware, it could be possible to achieve the accuracy of TCSPC with the integration speed of wide-field gated-integration measurements, which would allow lifetime imaging of dynamic samples. This would unlock an entirely new range of applications to which FLIM could be applied, such as calcium transient imaging, metabolic-based diagnostic imaging, and real-time monitoring of molecular interactions.

This thesis presents the design, characterization, and demonstration of a 64-by-64

SPAD array with integrated TDCs designed in a standard 0.13μm CMOS process and an optimized system for high-speed FLIM applications. The key components of this system include a custom CMOS SPAD, a novel active quench and reset circuit, delay-locked loop based TDCs with fine calibration control, a FLIM-specific compressing datapath, low-voltage differential signaling (LVDS) output banks, field-programmable gate arrays (FPGAs) for histogramming data, a cabled second generation peripheral component interconnect express (PCIe) interface for sustained rapid data transfer, and a direct memory access (DMA) approach for buffering data to a computer.

This thesis is organized into 6 chapters and includes one appendix. Chapter 2 provides background on fluorescence and details of lifetime imaging techniques. Chapter 3 provides background information on SPADs and describes the custom CMOS SPAD developed in this work. Characterization data for the SPAD is also presented. Chapter 4 discusses the design of the imaging integrated circuit (IC), which includes the SPADs, TDCs, and a custom datapath. Measurements to demonstrate the features and capabilities of the IC are included along with preliminary imaging results. Chapter 5 describes the system-level requirements for high-speed FLIM and presents the design of hardware and software for achieving fast imaging performance. Chapter 6 provides summarizing remarks on the original contributions presented and direction for future work in this field.

Chapter 2

Fluorescence Lifetime Imaging Microscopy

This chapter aims to provide a general foundation in fluorescence within the context of microscopy in order to better understand the constraints required of a lifetime imaging system. The following sections present an overview of fluorescence and fluorescence lifetimes, fluorescence lifetime measurement techniques, time-correlated single-photon counting, and system requirements for high-speed FLIM.

2.1 Fluorescence

Fluorescence is the emission of light from a molecule that was excited by a photon when the electron from its excited singlet state returns to the ground state. Fluorescent molecules, or fluorophores, are abundant in nature (chlorophyll is fluorescent) but can also be synthesized, including the manipulation of proteins to produce fluorescent structures [32] or through quantum confinement in semiconductors as in the case of quantum dots (Q-dots) [33]. The two properties of fluorophores that most fluorescent microscopes rely on for imaging are the wavelength of light that can cause an electron to transition from the ground state to an excited state and the wavelength of light that is emitted when that electron returns to the

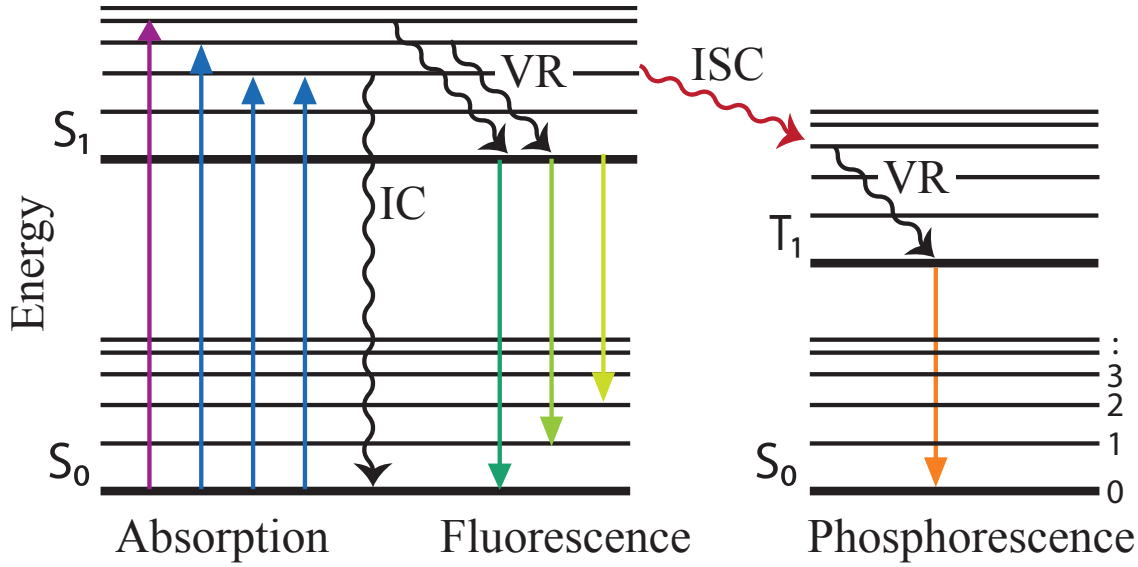

Figure 2.1: A Jablonski diagram illustrates the energy transitions that can occur within a fluorescent molecule. S_0 is the ground state, S_1 is an excited singlet state, and T_1 is a triplet state. Each state is comprised of several energy levels that form an energy band. Depending on the energy difference between the levels, the energy of absorbed and emitted light can vary, creating a spectrum for both absorption (excitation) and emission. After an electron is excited, it can return to the ground state through a number of pathways, including vibrational relaxation (VR), inter-system crossing (ISC), internal conversion (IC), or by emitting a photon.

ground state. A common representation of these transitions is a Jablonski diagram, shown in Figure 2.1. In the Jablonski diagram photon absorption or emission is indicated by a straight colored arrow and the photon energy associated with these transition wavelengths is given by $E = h\nu = hc/\lambda$, where h is Planck's constant. Throughout this chapter, E and λ are used interchangeably. For the visible portion of the electromagnetic spectrum, the photon energy ranges from around 1.6eV to 3 eV. This is also an important parameter in later sections during the discussion of semiconductor-based photodetectors.

In Figure 2.1, the fluorescent molecule is excited by a short wavelength light (purple/blue arrows). The likelihood that an incident photon will excite an electron in the molecule is given by its absorption coefficient. The direct conversion of energy to a photon as the electron transitions from the the singlet state, S_1, to the ground state, S_0, is the fluorescence emission processes. Also shown are vibrational relaxation and internal conver-

sion processes, during which the electron loses energy as heat, and inter-system crossing to a triplet state, T_1, where the electron decay to S_0 is classically forbidden. The probability that a fluorophore will produce a photon when an excited electron returns to the ground state is its quantum yield.

Because each energy state is comprised of several energy levels between which the electron can transition, a range of wavelengths can be used to excite an electron. Similarly, a spectrum of emission wavelengths is associated with each fluorescent molecule. An example absorption and emission spectrum is presented in Figure 2.2. In the Jablonski diagram, the emission energy must always be lower than the excitation energy. This difference in wavelengths can also be seen in the fluorescein isothiocyanate (FITC) spectrum of Figure 2.2 where the peak of the emission spectrum is at a longer wavelength (lower energy) than that peak of the absorption spectrum. The separation between these peaks is known as the Stokes shift.

For epi-fluorescence microscopy, an optical microscope is fitted with a filter cube, which consists of a combination of two filters and a dichroic mirror. The filter cube is designed to provide optimal isolation of the emission spectrum from the broadband excitation source (commonly a xenon-arc or mercury-vapor lamp). An example filter cube for FITC is drawn with the spectra in Figure 2.2 and a schematic of a common epi-fluorescent microscope is shown in Figure 2.3. The first filter is used to filter a narrow wavelength band from the excitation light. Then the dichroic mirror reflects the excitation light at 90 degrees toward the sample and transmits the emitted fluorescence from the sample. The final emission filter is used to further remove excitation light that was not previously filtered and to eliminate non-specific background fluorescence. The resulting light is focused onto an imaging device, where the image is observed.

Fluorescence microscopy can be performed with a wide-field camera, like a charge-coupled device (CCD) or complementary metal-oxide-semiconductor (CMOS) image sensor. For camera systems, the spatial resolution is limited by the Nyquist criterion, whereby the

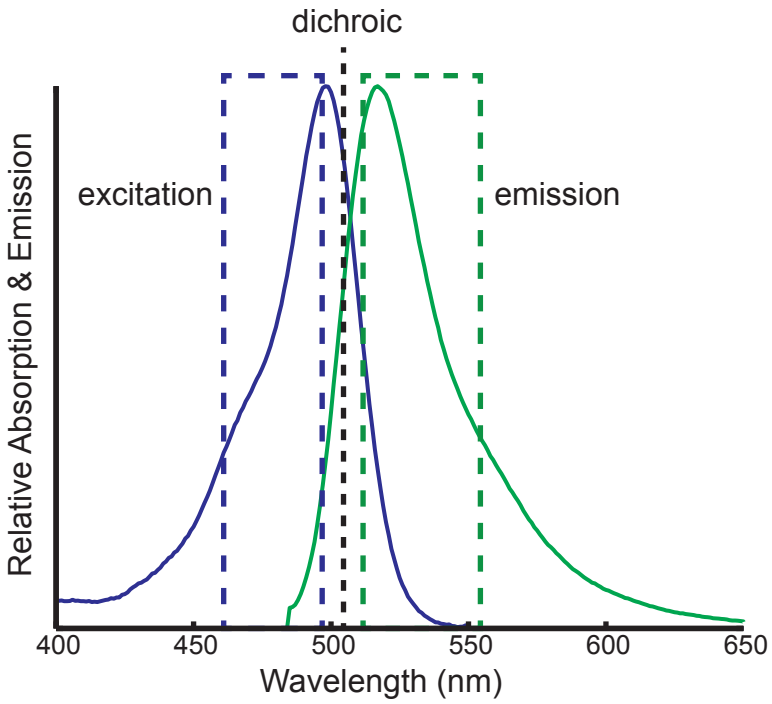

Figure 2.2: The absorption (blue curve) and emission (green curve) spectra for FITC. Overlaid on the spectra are the dichroic mirror and excitation and emission band-pass filters, which are represented by dashed boxes and lines. The filters and mirror are drawn as if they are ideal with perfect cut-offs. However, actual mirrors and filters will have a sloped transition between wavelengths that are transmitted and reflected. The filter cube data was obtained from the Edmund Optics FITC filter cube specifications.

diffraction limited spot projected on the sensor should be sampled by at least two pixels on each axis [11]. The diffraction limited spot is given by the Abbe limit:

$$r = \frac{\lambda}{2 \cdot NA} \tag{2.1}$$

Where NA is the numerical aperture (the numerical aperture is also given by $NA = n \cdot sin(\theta)$ where n is the index of refraction and θ is the angle of incidence). The projection of this diffraction limited spot is obtained by multiplying the Abbe limited resolution by the magnification of the objective lens in use. Table 2.1 gives a list of the maximum allowed pixel sizes for diffraction limited imaging across the visible part of the electromagnetic spectrum with common objective lenses.

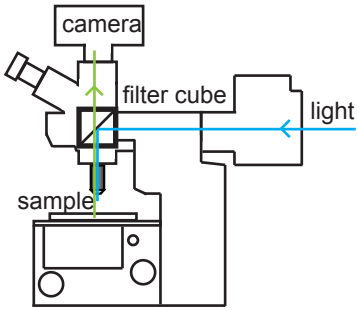

Figure 2.3: An illustration of a typical epi-fluorescent microscope with details of the internal optics omitted. A broadband light source is filtered before reaching the dichroic mirror where the excitation light is directed at the sample through the objective. The emission light is collected by the objective and passes through the dichroic mirror before reaching the emission filter and either being reflected to the eyepieces or transmitted to a camera. The camera is mounted to the microscope using a standard C-mount adapter.

While the filtering process previously described is relatively straightforward for imaging with a single fluorophore, imaging with two or more fluorophores poses a challenge. Multi-fluorophore images often suffer from problems of bleed-through or cross-talk where fluorescence from one fluorophore is transmitted through the emission filter of another fluorophore [7]. Additionally, contrast in fluorescence microscopy is given by the relative intensity of the background as compared to the feature of interest. As a result, quantitative fluorescence microscopy is complicated by fluctuations in fluorescence intensity. These fluctuations can be the result of effects including: fluorophore concentration gradients, variation in background signal, photobleaching, non-uniform excitation, and fluorophore cross-talk [6]. Because so many factors influence the fluorescence intensity, identifying genuine fluctuations due to an event of interest is challenging.

2.1.1 Fluorescence Lifetime

In addition to the fluorescence intensity, information about the fluorophore can be gained from the rates associated with the electron transitions described in Figure 2.1. The initial vibrational relaxation process typically occurs on the order of picoseconds. Once the excited

Table 2.1: A table listing maximum allowed pixel sizes to meet the Nyquist sampling requirement for the visible spectrum with common objective lenses. The lenses used for this comparison are from the Nikon CFI Plan Fluor series of objectives and assume that no additional magnification occurs between the camera and objective.

Magnification	N.A.	Wavelength	Resolution	Projected Size	Pixel Size
4x	0.13	450 nm	1.73 μm	6.92 μm	3.46 μm
		750 nm	2.88 μm	11.54 μm	5.77 μm
10x	0.30	450 nm	0.75 μm	7.50 μm	3.75 μm
		750 nm	1.25 μm	12.5 μm	6.25 μm
20x	0.50	450 nm	0.45 μm	9.00 μm	4.50 μm
		750 nm	0.75 μm	15.0 μm	7.50 μm
40x	0.75	450 nm	0.30 μm	12.0 μm	6.00 μm
		750 nm	0.50 μm	20.0 μm	10.0 μm
60x	0.85	450 nm	0.265 μm	15.88 μm	7.94 μm
		750 nm	0.44 μm	26.47 μm	13.23 μm
100x	0.90	450 nm	0.25 μm	25 μm	12.5 μm
		750 nm	0.416 μm	41.6 μm	20.8 μm

electron reaches the lowest energy level in S_1, it is in a metastable state and the time before it decays to S_0 and emits a photon is typically a few nanoseconds. If an electron undergoes a spin conversion to reach the triple state, T_1, then it can remain in that excited state for milliseconds or longer. Photon emission from the triplet state to the ground singlet state is called phosphorescence and typically occurs at longer wavelengths than the fluorescence [18].

Through precise temporal measurements, it is possible to determine the average amount of time that a fluorophore is excited before emitting a photon. This average emission time is the fluorescence lifetime of the molecule and its measurement is the focus of this thesis. Most fluorescent molecules exhibit a mono-exponential decay when measured independently. However, many applications of fluorescence lifetime will inherently have fluorescent species other than that of interest present. As a result, a typical intensity decay for a population of fluorophores will consist of a multi-exponential decay of the form:

$$I(t) = I_0 + A_0 e^{-t/\tau_0} + A_1 e^{-t/\tau_1} + \cdots \tag{2.2}$$

A fluorophore's lifetime is an intrinsic property of the molecule, but it can be affected by external factors that influence its emission rate constant. Consequently, lifetime measurements can be used to detect changes in the microenvironment surrounding a fluorophore. This property makes fluorescence lifetime a powerful method for direct *in vivo* sensing of changes in pH due to interactions with hydrogen ions [34], transients of intracellular Ca^{2+} concentration [35], gradients of intercellular viscosity [36], fluctuations in temperature [37], and presence of other macromolecules [38].

The ability of FLIM to image interactions between the fluorophore and other molecules is of particular interest for *in vivo* studies. A powerful technique that leverages FLIM is fluorescence resonance energy transfer (FRET) [39]. In FLIM-FRET measurements, the lifetime of the donor fluorophore is shortened due to the presence of a quenching molecule within the Förster radius. This quenching molecule must be selected such that there is a considerable overlap between the emission spectrum of the donor and the absorption spectrum of the quencher. Despite the requirement that the donor and acceptor have overlapping spectra, an intermediate photon is not produced. Rather, the excited electron from the donor is directly transferred to the acceptor through a dipole-dipole interaction. Because the mechanism of electron transfer relies on dipole interactions, the Förster radius is only a few nanometers and FRET is a common tool for detecting binding of molecules [38].

Another instance in which FLIM can identify binding between molecules is in the observation of *in vivo* metabolic states. There are two metabolic coenzymes, reduced nicotinamide adenine dinucleotide (NADH) and flavin adenine dinucleotide (FAD), that are intrinsically fluorescent and exhibit changes in their lifetime when in a protein-bound versus unbound state [40]. The ratio of bound versus unbound coenzymes provides information about the favored metabolic pathway of a cell. Through leveraging fluorescence lifetime information associated with these molecules, lifetime imaging has the potential to provide

early identification of cancers. This is possible since lifetime imaging can distinguish between metabolic states and the favored metabolic pathway is known to shift from oxidative phosphorylation to glycolysis during tumor progression [41]. This shift is known as the Warburg effect and will result in increased FAD lifetimes and shorter NADH lifetimes [42]. A series of diagrams showing the metabolic pathways and the roles of NADH and FAD can be found in Chapter 2 of reference [43].

From these few examples, it is clear that fluorescence lifetime imaging is a powerful technique. However, all of these applications depend on the ability to precisely measure combinations lifetimes of only a few nanoseconds. The following sections discuss the methods for measuring fluorescence lifetimes and the challenges associated with lifetime measurements.

2.2 Fluorescence Lifetime Measurements

Fluorescence lifetime imaging microscopy can be performed by measuring the temporal response of a sample in either the time domain or the frequency domain. There are trade-offs for performing the measurements using either domain and multiple approaches exist within each broad classification. In the following sections, the fundamental theory and methods for measuring lifetimes with each approach is presented. For the engineering work in this thesis, the time-domain approach was taken.

2.2.1 Frequency-Domain Lifetime Measurements

Frequency-Domain lifetime measurements are most often performed using an intensity modulated light source and measuring temporal response of the fluorescence signal. The amplitude and phase shift of the measured fluorescence signal relative to the modulated source will depend on the lifetime of the fluorophore and the period of the modulated light source. Figure 2.4 illustrates how the fluorophore lifetime results in a phase shift and amplitude modulation at the output signal. The measured fluorescence signal can be thought of as a linear

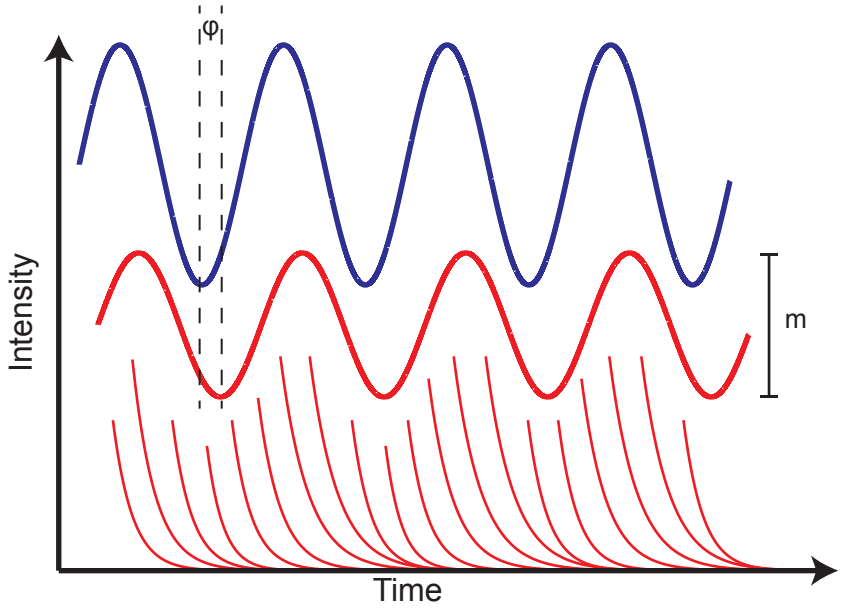

Figure 2.4: An illustration showing how the amplitude and phase shift of the frequency-domain FLIM output signal depend on the lifetime of the fluorophore. Fluorophores with short lifetimes have small phase offsets and larger amplitudes due to their faster response and corresponding ability to closely follow the excitation waveform. (blue) excitation source and (red) emission signals.

combination of intensity decays that are continuously excited with the initial intensity of the decay proportional to the instantaneous amplitude of the excitation light.

If the excitation modulation period is decreased while measuring the same fluorophore as Figure 2.4, the output phase shift would increase and the amplitude modulation of the signal would decrease, as seen in Figure 2.5. Instead, if the lifetime of the fluorophore were longer than in Figure 2.4 but the excitation modulation frequency kept the same, the result would follow the relative trend of Figure 2.5. As the period of the modulated excitation source becomes short relative to the lifetime of the fluorophore, the amplitude modulation of the output signal approaches zero as the long fluorescence decays become averaged.

The fluorescence lifetime is related to the measured phase shift and modulation amplitude by [18]:

$$\tan \phi_\omega = \omega \cdot \tau \tag{2.3}$$

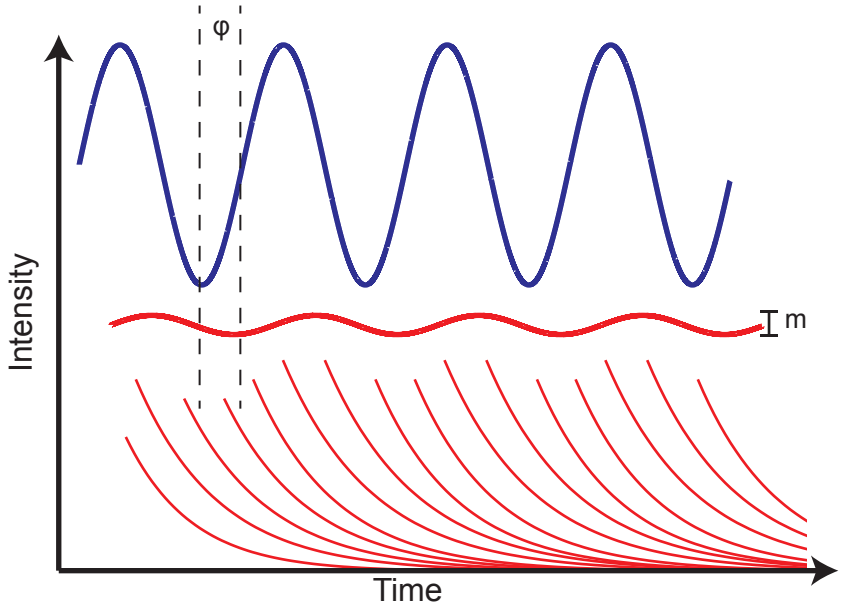

Figure 2.5: An illustration showing how the amplitude and phase shift of the frequency-domain FLIM output signal depend on the lifetime of the fluorophore. Fluorophores with long lifetimes have large phase offsets and smaller amplitudes compared with short lifetime fluorophores excited with the same modulation frequency. (blue) excitation source and (red) emission signals.

$$m_\omega = \sqrt{1 + \omega^2 \tau^2} \tag{2.4}$$

Because the phase shift and amplitude modulation are relative measures data is typically collected at several excitation modulation frequencies and used to extract the lifetime. The frequency response of a fluorophore is a plot of the phase shift and amplitude modulation for a range of excitation modulation frequencies such that the output phase shift is swept from 0° to 90°. The range of modulation frequencies that is appropriate for a given fluorophore depends on its lifetime. Examples of frequency responses of fluorophores can be found in Chapter 5 of reference [18]. In the instance where multiple fluorophores are present in a sample and are spatially co-localized, the lifetime extracted from the modulation amplitude and phase shift will not be the same. As a result, complex lifetime extractions are required and limited to apparent lifetime values, which depend on the measurement system and weighting factors during the lifetime extraction process[18]. This inability to defini-

tively resolve multiple fluorescent species in a sample is one of the major drawbacks to using frequency-domain FLIM.

Frequency-domain FLIM can be performed using either wide-field or point detection (used with laser scanning). Commonly, wide-field frequency-domain methods use light-emitting diodes (LEDs), which are directly modulatable and cheaply available. Continuous wave (CW) lasers can also be used and are commonly applied when excitation wavelengths in the ultraviolet range are required. Another limiting factor of frequency-domain FLIM is photobleaching of the fluorophores due to the continuous excitation of the molecules during measurements. Laser scanning systems can also be used to perform point-by-point detection using a confocal arrangement, which will reduce the extent of photobleaching but require long image acquisition times.

The measured output signal in a frequency-domain FLIM laser scanning system exactly matches the waveforms outlined in the discussion above. A photomultiplier tube (PMT) with sufficient bandwidth to match the modulated frequency can be used to record a waveform similar to that in Figure 2.4. As the laser is rastered over each pixel in the image, phase and amplitude information are recorded and this process is then repeated for several modulation frequencies. This rastering process limits the speed at which lifetime images can be recorded but the measurement process is straightforward.

The faster wide-field detection method is commonly performed using an intensified charge-coupled device (ICCD) sensor. The intensifier has a modulated gain at or near the same frequency as the excitation source. When the sample is illuminated, the modulated fluorescent signal impinges on the intensifier, which then amplifies the optical signal through use of a phosphor. The modulated fluorescence signal at an individual pixel in the image will have the form of:

$$I(t) = I_0 \cdot [1 + m_F \cdot \sin(\omega t + \phi_F)] \tag{2.5}$$

where I_0 is the average fluorescence intensity, m_F is the modulation amplitude, and ϕ_F is the phase offset for the modulation frequency ω. The response time of the phosphor is on the

order of milliseconds, which limits the bandwidth of the measured signal to less than 1 kHz. Additionally, the phosphor degrades the spatial resolution. Further, CCDs are often limited to 100 frames per second or less, depending on the intensity of the light at the sensor. This reduces the sampling rate of the signal to only 100 Hz. Through the gain modulation of the intensifier, the high frequency emission signal is mixed to a lower frequency that falls within the bandwidth of the intensifier and CCD [15]. The gain follows a function similar to the fluorescence signal of the form:

$$G(t) = G_0 \cdot [1 + m_D \cdot \sin(\omega t + \phi_D)] \tag{2.6}$$

where G_0 is the average gain and m_D and ϕ_D are the modulation amplitude and phase offset applied to the detector. The product of equations 2.5 and 2.6 give the resulting signal after the intensifier (including the low-pass filtering of the phosphor):

$$S(\phi_D) = S_0 \cdot \left[1 + \frac{1}{2} m_F m_D \cdot \cos(\phi_F - \phi_D)\right] \tag{2.7}$$

As a result of the gain modulation, the intensity at each pixel now depends on the modulation and phase offset parameters, allowing differences in lifetime to be easily identified. Additional lifetime information can be obtained by sweeping the modulation frequency of the detector ϕ_D. A lifetime measurement with fewer than 10 modulation steps will require approximately 1 second to acquire an image [23]. Techniques have recently been explored to use higher harmonics present in pulsed measurements to obtain multiple modulation frequency components simultaneously and increase imaging speed [44].

Although these techniques allow for easy differentiation between lifetimes of fluorophore species in a sample, they are limited in the ability to provide precise lifetime values. Recent work has also been conducted to use non-fitting methods for frequency-domain FLIM where phasor plots are constructed from the measured data that allow for differentiation between regions in an image without measuring exact lifetimes [45]. The fastest

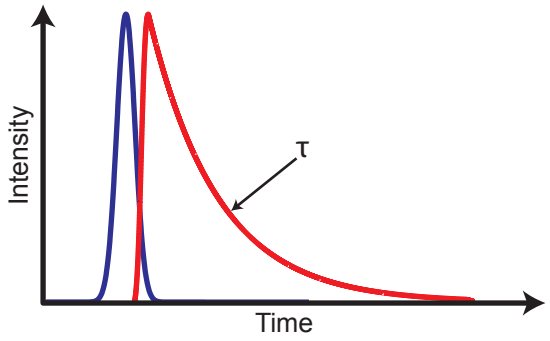

Figure 2.6: With time-domain measurements, the intensity decay rate is measured directly. A short pulse of laser light (blue) excites a population of fluorophores which then exhibit the intensity decay (red).

frequency-domain FLIM measurements have achieved up to 55 frames per second (fps) [46].

2.2.2 Time-Domain Lifetime Measurements

The alternative to measuring fluorescence lifetimes in the frequency domain is to directly measure the intensity decay using time-domain techniques. The two primary time-domain techniques are the time-gated integration method and time-correlated single-photon counting (TCSPC). These techniques differ greatly from the frequency domain in the measurement requirements and the type of data that are recorded. Most notably, all time-domain measurements require the use of a sub-nanosecond pulsed laser. The pulsed laser excites a population of fluorophores in the sample and the intensity decay of that population is the measured, as seen in Figure 2.6. Because typical biologically relevant fluorescent lifetimes are on the order of several nanoseconds, accurate time measurement of the decay can be challenging. The following sections present the details for each of the time-domain approaches to lifetime imaging.

Time-Gated Integration Method

The simplest time-domain fluorescence lifetime measurement uses two gated photon integration windows after the laser pulse to accumulate the signal from the fluorescence decay as

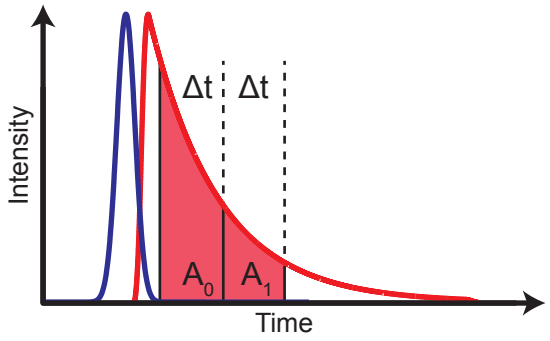

Figure 2.7: The fluorescence decay is integrated during two equally sized time bins. The lifetime is determined using Equation 2.8 with the measured photocurrent recorded during the two windows. More accurate lifetime measurements can be made by integrating over shorter windows and increasing the number of total windows.

shown in Figure 2.7 [47]. Time-gated measurements enable rapid lifetime estimation, and a lifetime with two measurement windows of equal width can be approximated by:

$$\tau = \frac{-\Delta t}{\ln\left(A_1/A_0\right)} \tag{2.8}$$

Where Δt is the window width and A_0 and A_1 are the integrated photocurrents during the first and second time intervals, respectively. This method results in an averaged lifetime value and produces results quickly but with poor accuracy. It is impossible to extract multiple lifetime components while using only two integration windows. However, the gated-integration mode also allows for wide-field imaging using ICCDs with the gating signal applied to the intensifier. Applications where measuring an average lifetime value or a shift in lifetime are all that is required will be well suited for this mode of lifetime imaging. The fastest demonstrated FLIM frame rate using this method is 100 fps [48].

The accuracy of this time-gated approach can be improved by increasing the number of integration windows and adjusting their widths to optimize the windows such that the integrated values are approximately the same throughout the decay [49]. Through these methods, the gated-integration technique offers improved lifetime determination at the cost of system complexity, imaging speed, and photon efficiency.

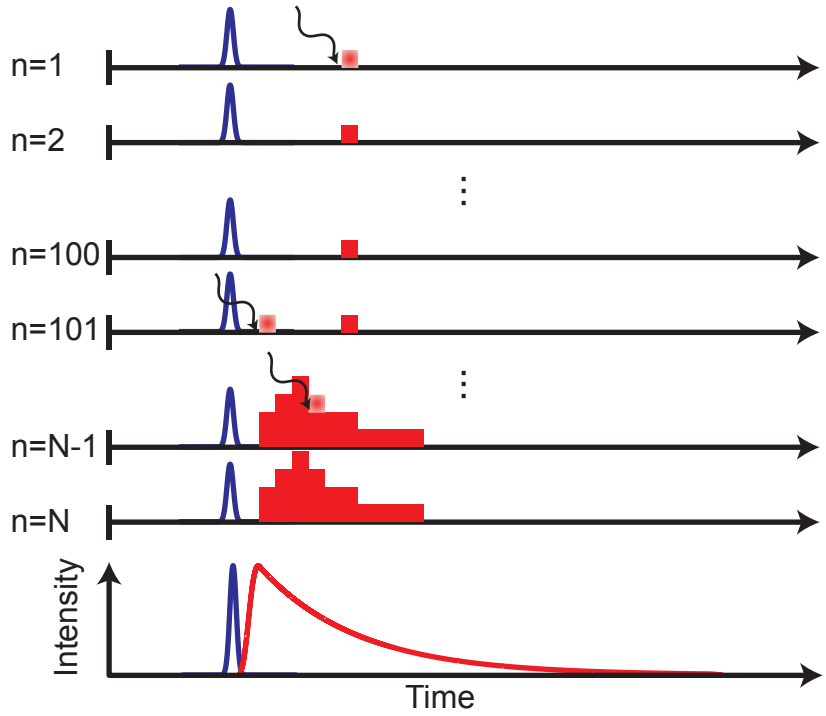

Figure 2.8: In time-correlated single-photon counting, at most one photon is detected per laser pulse. In this figure, N laser pulses are shown (represented by the blue spikes) and photon detection events are indicated by the curved arrows. On average 1% of laser pulses should result in a photon being recorded in order to avoid pulse pile-up. As the laser is repeatedly pulsed, the arrival times of the photons are binned to form a histogram. With sufficient laser repetitions and proper imaging conditions, the resulting histogram will closely match the intensity decay of the fluorophore.

Time-Correlated Single-Photon Counting

Another method for time-domain lifetime measurements is time-correlated single-photon counting (TCSPC). This method leverages single-photon sensitive detectors to measure the time at which individual photons are emitted from the population of fluorophores and extracts the lifetime from the statistical distribution of these events [50]. The details of this process are presented in Figure 2.8. Again, a short laser pulse is used to excite the fluorophores in the sample being measured. The excited electrons in the fluorophores will return to the ground state with an average time given by its lifetime, τ, as discussed above. Each individual excitation event and corresponding photon emission will result in a photon arrival

time that is distributed as a non-homogeneous Poisson process with the rate parameter given by:

$$\lambda(t) = ae^{-t/\tau} \tag{2.9}$$

where λ is the Poisson rate parameter, a is the intensity of the emission signal, and τ is the lifetime of the fluorophore. Each time the laser is pulsed, a single photon emission is detected and the time between when the laser was pulsed and the photon was detected is recorded. By repeating thousands of laser pulses, a sufficient number of photon arrival times can be collected to reconstruct a histogram that matches the distribution given by the non-homogeneous Poisson process. Once the histogram of arrival times has been collected, an exponential decay can be fit to the histogram from which the lifetime can be estimated. For each pixel in the image, thousands of laser repetitions are accumulated, which leads to long image acquisition times. Typical fluorescence lifetime images using TCSPC can be acquired in under 30 seconds, but acquisition times can reach up to 30 minutes for detailed decay measurements [51].

A key constraint on lifetime measurements using TCSPC is that the fluorescence intensity must be relatively low. In order for the distribution of photon arrival times to accurately represent the fluorescence intensity decay, the fluorescence intensity should be sufficiently dim such that photons are detected for only 1% of laser repetitions. If the fluorescence intensity is too bright, the distribution of photon arrival times will become skewed toward short arrival times, resulting in an effect known as pulse pile-up [52].

The necessary hardware for recording TCSPC lifetime data is a single-photon sensitive detector, such as a photomultiplier tube (PMT) or avalanche photodiode (APD), a time-to-digital converter (TDC) for recording the photon arrival time, and a pulsed laser [53]. Most commercial systems use a single PMT with one TDC channel. Photon arrival times are accumulated as a pulsed laser is rastered over the entire sample. Both the detector and the TDC require a short period of inactivity following a photon event to reset. During this period they are unable to detect other photons. This period is called the dead time and

is a limiting factor in TCSPC measurements. A typical PMT dead time is 10-50ns while a typical TDC dead time is upwards of 100ns. The TDC dead time limits the maximum count rate to less than 10 MHz, which, combined with the point-by-point laser scanning, restricts the maximum imaging rate of commercial TCSPC systems. As an example, for a 20 MHz laser repetition rate, a fluorescence intensity tuned for a 1% detection rate, and an ideal scanning and detection system, it will take 250 μs, on average, to acquire a minimum of 500 photon detections for each pixel in the image. Acquiring a 64-by-64 pixel image, therefore, requires at least 1s.

Combining commercial systems into highly parallel imaging systems is impractical due to the size of the components and cost. Recent developments in complementary metal-oxide-semiconductor (CMOS) single-photon avalanche diodes (SPADs) have led to the opportunity to combine thousands of single-photon detectors with TDCs for parallelized measurements [54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66]. CMOS technology nodes as advanced as 90nm have successfully been used to design CMOS SPADS [64]. These CMOS SPADs provide a number of capabilities that could be leveraged for performing high speed fluorescence lifetime imaging. Because of the close integration of the SPAD and control circuitry, dead times can be reduced to negligible durations. Additionally, the economics of CMOS scaling allow for thousands of SPADs and TDCs to be combined on a single silicon chip, providing many parallel detection channels without the high costs associated with the currently available commercial systems. A major concern with CMOS SPADs is high noise levels due to either high dark count rates [67] or after-pulsing [68], both of which will be discussed in detail in Chapter 3. Additionally, the wavelengths at which CMOS SPADs are most sensitive often fall below 525nm due to constraints of the CMOS processes, which will also be discussed in Chapter 3.

These CMOS SPADs have been used in the design of SPAD arrays ranging in size from 3-by-3 [69] to 160-by-128 [31] and many sizes between [27, 70, 71, 72, 73, 30, 74, 75, 76, 77, 78]. In addition to FLIM, these SPAD arrays have been used for a number of

applications, including 3-D imaging [70], protein microarrays [79], gamma ray detection [80], and automatic laser alignment [81].

Although some of these SPAD and TDC arrays have been directly applied to TCSPC imaging systems [26, 27, 28, 29, 30, 31], and they have demonstrated improved imaging acquisition times through the parallel nature of the imagers, the overall image acquisition rates have been limited by off-chip data transfer rates. Using a similar example as above, for a laser repetition rate of 20 MHz and a 64-by-64 array of pixels (this requires 12 bits of position data to tag each photon arrival event with its location) with 10-bit timing resolution, the required data rate reaches 1.8 Tbps. While event-driven readout approaches have reduced these data rates, previous SPAD array systems have still been limited in the number of parallel channels, frame rate, or number of continuously acquired frames [82]. Table 2.2 lists published SPAD arrays with integrated TDCs and the theoretical data-bandwidth limited maximum frame rate. The actual frame rate achieved by each of these devices is often omitted from publication.

The objective of this thesis is to design a FLIM optimized SPAD array consisting of 64-by-64 pixels with independent TDCs that is capable of processing the large amount of data required for parallel TCSPC at high image acquisition rates. Additional emphasis is placed on the system-level design for processing and storing the lifetime data.

2.3 System Requirements

In order acquire lifetime images at high frame rates, a number of requirements must be met. First, a low-noise CMOS SPAD is necessary for detecting single-photon emission from excited fluorophores with a low average photon detection rate. An effective control circuit is needed to ensure that this SPAD is biased at the optimum operating point and is capable of detecting photon events with every laser pulse (i.e. the dead time is less than the laser pulse period). Additionally, a TDC with resolution below 100ps is desired for accurately

Table 2.2: A table showing the maximum theoretical frame rate for previously reported SPAD arrays with integrated TDCs. This assumes an event-driven readout scheme, which requires that each TDC data must also be tagged with the pixel location from which it was generated. This theoretical maximum assumes that 1000 photon events are needed to extract the lifetime and that photon event data is output on every possible I/O clock cycle.

Pixels	Bandwidth (Mbps)	Time Bits	Position Bits	Frame Rate (fps)	Ref.
1024	1	10	10	0.0488	[30]
1024	10240	10	10	500	[83]
1024	5120	10	10	250	[72]
4096	0.073	37	12	0.00036	[27]
16384	7680	10	14	19.5	[71]
20480	51200	10	15	100	[31]
4096	42000	10	10	466	This Work

determining biologically relevant fluorescence lifetimes, which are on the order of a few nanoseconds. In order to address the 1.8 Tbps data requirement, an event-driven readout technique is necessary along with high-speed output buffers capable of driving at least 16 Gbps off of the imaging chip. At the system level, a technique for pre-processing the lifetime data to reduce the required data bandwidth between the imaging system and a computer for lifetime processing should be designed to reduce the datarate by at least a factor of 10. The following chapters present each of these components in the context of high frame rate fluorescence lifetime imaging.

Chapter 3

Single-Photon Detectors

Single-photon sensitive detectors are essential to time-correlated single-photon counting. Commonly used single-photon detectors are photomultiplier tubes (PMTs) and avalanche photodiodes (APDs). Each of these devices have varying levels sensitivity, response time, noise, spectral response, and dead time. For time-correlated single-photon counting applications, high sensitivity is the most critical parameter while low noise and fast response time are also highly important. The wavelengths over which the device is most sensitive is an application-specific parameter and the dead time only becomes a factor when high-speed measurements are desired. The mechanism by which each class of detector operates and the key detector properties are discussed briefly below.

3.1 Photomultiplier Tubes

Photomultiplier tubes are high-gain photodetectors that produce a current in response to incident photons. A PMT relies on the photoelectric effect to generate electrons from photons as they collide with the photocathode. These electrons are then accelerated in an electric field toward a series of electrodes (dynodes) of sequentially increasing potential. When the initial electrons collide with the first dynode, electrons are freed from that electrode with more electrons being released than are incident on the dynode due to secondary emission. This now

larger population of electrons is then accelerated to the next, higher potential dynode. With each subsequent dynode interaction more electrons are produced, which results in the high photocurrent gain and fast temporal response. After the last dynode, the resulting electrons are collected by the anode and output from the PMT to a transimpedance amplifier. The photocathode and dynodes of a PMT are enclosed in a vacuum tube to prevent the electrons from ionizing particles in the air.

Because the chain of electron multiplication interactions begins through the photoelectric effect at the photocathode, the noise of the PMT is dictated by the photocathode material and is typically low. The photocathode material also determines the range of wavelengths to which the PMT will respond. The dead time, which is the period of time after the amplification cascade starts that it takes for the PMT to be able to detect another photon, is generally tens of nanoseconds. Although PMTs offer high gain and low noise, they require high voltages ($\sim$1000V) to accelerate the electrons through the tube and are bulky, fragile, and expensive. These devices are well suited for laser scanning imaging but are not appropriate for combining to form an imaging array.

3.2 Avalanche Photodiodes

Avalanche photodiodes (APDs) are the semiconductor equivalent of a PMT and can be made of silicon (Si), germanium (Ge), indium gallium arsenide (InGaAs), or gallium nitride (GaN) depending on the desired wavelength range over which the device should be sensitive. An APD is a diode that is reverse biased such that a high electric field is generated across the depletion region of the device. When a photon with sufficient energy is incident on the detector, it produces an electron-hole pair. This electron-hole pair will be accelerated in the electric field and collide with other atoms in the silicon lattice, generating additional electron-hole pairs through avalanche multiplication. APDs are typically less sensitive than PMTs and have higher noise. However, they offer a similar temporal response, often have

Figure 3.1: The I-V characteristic for a typical avalanche photodiode. The reverse bias breakdown voltage is indicated by the vertical dashed line.

broader wavelength sensitivity ranges, are more durable, cheaper, and have higher immunity to electromagnetic interference than PMTs.

A current-voltage (I-V) curve for a typical discrete APD is presented in Figure 3.1. The reverse bias breakdown voltage, V_{br}, is labeled with a vertical dashed line where the negative current begins to exponentially increase. In order for an APD to provide single-photon sensitivity, the reverse bias across the photodiode must exceed its breakdown voltage such that the electric field across the depletion region of the device is strong enough to sustain an avalanche indefinitely through the constant regeneration of impact ionized electron-hole pairs.

In order to stop the avalanche current, the voltage across the APD must be reduced below its breakdown voltage through a process known as quenching. This technique allows for the generation of a large photocurrent pulse in response to a single photon. This mode of operation is often called Geiger mode due to the commonality with radiation detectors. When an APD is biased and quenched in Geiger mode, it is referred to as a single-photon avalanche diode (SPAD). The most straightforward approach to quenching a SPAD is to place a resistor in series with the SPAD (see Figure 3.2a), which will produce voltage drop when an avalanche current begins flowing. During an avalanche, this voltage drop will reduce the SPAD bias voltage and stop the avalanche, resulting in a short voltage pulse

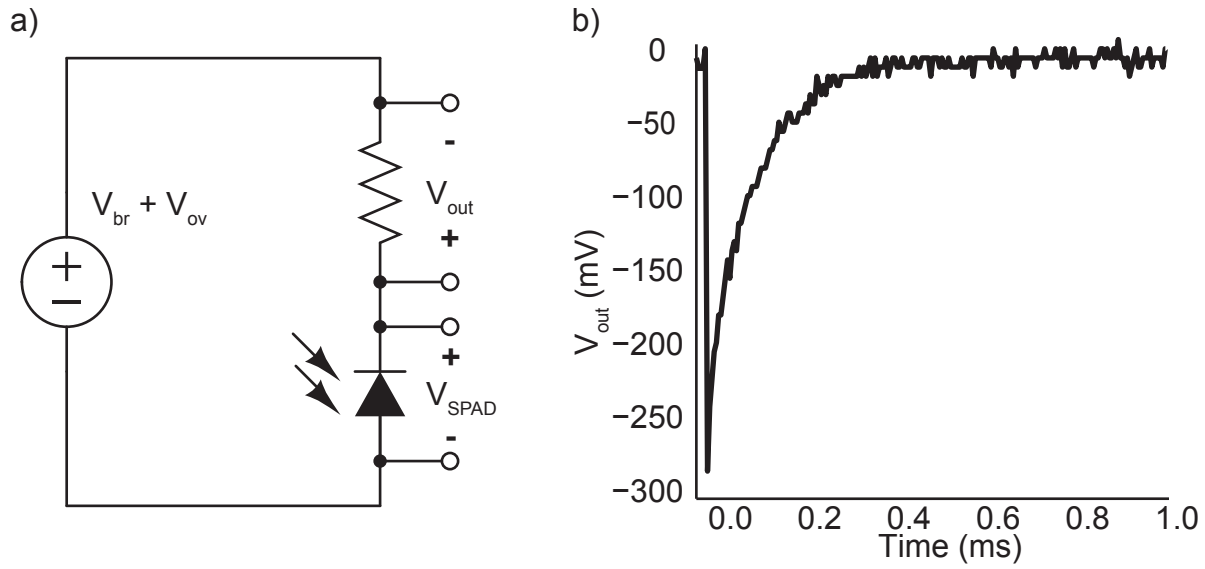

Figure 3.2: Passive quench and reset of SPAD using a large resistor. a) Schematic showing the passive quenching circuit. b) In the measured waveform, V_{ov} was 275 mV and the quench/reset resistor was 1 MΩ. The ground probe of the oscilloscope was the highest potential in the measurement, hence the inversion of the output voltage, V_{out}.

across the resistor that indicates that an avalanche has occurred (Figure 3.2b). The size of the quenching resistor will determine the magnitude of the avalanche current that will be reached before the SPAD is turned off. Consequently, this resistor also sets the duration of the avalanche. Ideally, one would like a small current and short avalanche duration to minimize the total charge involved in the avalanche, which will reduce the potential for trapped charge and the likelihood of re-triggering the device, an effect called afterpulsing. Figure 3.2b shows a typical oscilloscope trace for an avalanche event and the quenching process. Additionally, a slow R-C limited recharge process occurs through the quenching resistance as the SPAD resets.

SPADs are typically designed with special features to limit noise events, such as smaller active areas and a lightly-doped guard ring surrounding the structure [84]. The smaller active area reduces the likelihood of a thermal noise event and reduces the magnitude of the avalanche current while the guard ring protects against premature breakdown at the periphery of the device due to the high electric field near the edges. Additionally, the P-N

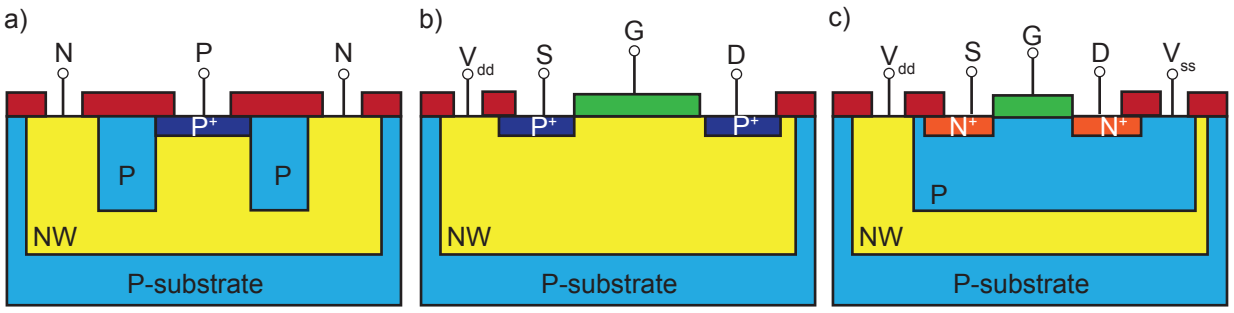

Figure 3.3: a) Cross-sectional view of an idealized planar SPAD structure. b) A cross-sectional view of a standard PFET and c) a cross-sectional view of a standard isolated (or triple-well) NFET. The FET devices are idealized and do not include any of the complex implants that are used in modern CMOS processes.

junction used for a SPAD is typically asymmetrically doped such that the depletion region is mostly on one side of the junction. The depletion width is a tuning parameter for optimizing between sensitivity to incoming photons and the photocurrent temporal response [85]. An illustration of a SPAD is presented in Figure 3.3a. Similar to PMTs these devices also require high voltages and commercial versions are bulky, which ultimately prevents their incorporation into an array-based imaging system.

3.3 CMOS Single-Photon Avalanche Diodes

Because silicon has a bandgap of 1.11 eV, it is capable of absorbing light at wavelengths below 1.1 μm, which covers the visible range (1.6 - 3 eV) and includes the wavelengths of interest for most fluorescence applications. Additionally, advances in semiconductor manufacturing have enabled tremendous technological advancement in the last half century through reducing the minimum feature size of devices used to make a wide range of integrated circuits for use in computing, communications, imaging, and many other applications [86]. The current state-of-the-art integrated circuits (ICs) are complementary metal-oxide semiconductor (CMOS) devices with minimum features sizes as small as 14nm. By designing SPADs in a CMOS technology, the advances in semiconductor manufacturing can be leveraged to make large arrays of SPADs for an array-based FLIM system. Additionally, with SPADs designed in

CMOS technologies, the TDCs and data processing circuitry can be integrated onto the same chip as the photodetectors, which allows for accurate time measurement and efficient data handling.

The first work on CMOS compatible SPADs was published in 1994 and demonstrated that planar SPADs could be successfully integrated with CMOS circuitry [87]. Since this initial work, a number of CMOS SPADs have been published with the aim of enabling large arrays of devices [73, 88, 89, 90, 91]. CMOS SPADs have been designed in both specialized imaging processes and standard CMOS processes. In standard CMOS technologies, the diffusion implants for n-type and p-type metal-oxide-semiconductor field effect transistors (MOSFETs) are repurposed for designing the SPAD. As a result, diffusions used for the p-type MOSFET (PFET) and isolated n-type MOSFET (NFET) (illustrated in Figure 3.3b and 3.3c respectively) are used. CMOS SPADs almost always consist of a heavily doped P^+ anode and a lightly doped n-type cathode (n-type well from PFET device) combined with a lightly doped p-type guard ring (p-type well from the isolated NFET device). The key performance metrics for a SPAD are its photon detection probability (PDP) and peak wavelength sensitivity, its dark count rate (DCR), and its impulse response function (IRF).

The PDP is primarily a function of the overvoltage, V_{ov}, which is the bias beyond V_{br} applied to the device. Additionally, the wavelength of peak sensitivity is determined by the depth of the depletion region for the device. As light enters the silicon structure, short wavelengths are absorbed at shallow depths and the photons can be absorbed before reaching the multiplication region. Conversely, long wavelengths are absorbed deep in the silicon and can be absorbed too far from the detector to be sensed. Photon interactions deep in the silicon can diffuse into the active region but will lead to a degradation of the IRF due to the relatively long diffusion times. Photon absorption follows the Beer-Lambert law:

$$I(x) = I_0 \cdot e^{-\alpha x} \tag{3.1}$$

where x is the depth into the material, I_0 is the light intensity at the surface, and α is the absorption coefficient. In silicon, the average absorption depth is approximately 100nm for 400nm wavelength light and is nearly 10μm for 800nm light [92]. Because the PFET source/drain implant is used for the anode of the SPAD, the junction depth will be relatively shallow (100-200nm) and will result in peak sensitivity at short wavelengths. SPADs implemented in more advanced technology nodes will have shallower multiplication regions and a peak wavelength that is blue-shifted. Recent work has explored using backside illumination for SPAD structures formed by the p-type substrate and the n-type well implant used for PFET devices. This work has demonstrated a peak wavelength sensitivity of 650 nm without degradation of the IRF [93].

The DCR has a first order dependence on the CMOS process and geometry of the device. Early CMOS SPADs were designed in technology nodes with minimum feature sizes ranging from 0.8μm to 0.35μm [54, 56, 60, 73]. These processes all used local oxidation of silicon (LOCOS) to isolate devices from one another. This isolation technique uses a wet oxidation step to selectively grow an oxide between devices. This oxide growth is generally clean and leaves few interface traps at the silicon/silicon dioxide boundary. The CMOS SPADs in these technologies typically exhibit low DCRs due to this clean interface. SPADs have also been designed in more recent CMOS technology nodes (180nm to 90nm) but must contend with the increase in DCR that can be caused by the defect-rich interface between the silicon and the shallow trench isolation (STI) used for separating devices. These interfacial defects can trap charge near the multiplication region that can lead to noise events when the trapped charge gains enough thermal energy to free itself from the trap and diffuse into the multiplication region. While some devices have suffered with DCRs in the MHz range [59, 67], others have found ways to mitigate the impact of STI through techniques like hydrogen passivation [61]. Another technique for reducing the STI induced DCR is to spatially separate the STI interface from the multiplication region [65].

The IRF also depends primarily on the CMOS technology and structure of the device.

It is a function of the physical path taken by the generated charge carriers and depends on the implant profile used to make the SPAD. For an unmodified CMOS process, the designer has no control over these parameters and the IRF is constrained by the CMOS technology.

In addition to the factors discussed above, the PDP, DCR, and IRF are all functions of the applied overvoltage, V_{ov}, to the SPAD. As the overvoltage increases, the PDP will increase, the DCR will increase and the IRF will become narrower. The trade-off between PDP and DCR has often been obfuscated in the literature and publications sometimes cite peak PDP and minimum DCR at different operating points. Depending on the application, the optimal bias point for the SPAD may differ. In this work, a FLIM specific figure-of-merit (FOM) was established to provide guidance when selecting the bias point for a SPAD and to make comparisons between devices more objective.

3.4 Optimal Biasing of SPADS for FLIM

As described in Section 2.1.1, the duration of a fluorophore's metastable state following a pulsed excitation determines its fluorescence lifetime. The time at which an individual molecule will emit a photon and return to its ground state follows a non-homogeneous Poisson process. In TCSPC the individual photon arrival times are recorded from which a statistical distribution and the lifetime are extracted. The performance of a SPAD can be evaluated specifically for FLIM by considering the statistical properties of the fluorescence emission. Additionally, this same analysis can be used to determine the optimal bias point of the device such that an image can be formed with the minimum number of laser repetitions. The discussion that follows presents the theoretical background for the FLIM-specific FOM and provides simulated results as verification.

3.4.1 Non-homogeneous Poisson Process

The non-homogeneous Poisson process describing photon emission from a fluorophore gives a photon emission rate of:

$$\lambda(t) = ae^{-t/\tau} \tag{3.2}$$

where a corresponds to the intensity of the emission, which is typically limited by adjusting the excitation laser power such that only 1% of measurement windows contain hits, as constrained by the pulse pile-up requirement [52]. τ is the lifetime of the fluorophore to be measured. The mean value of this rate during a measurement window of length T is:

$$\mu = \int_0^T \lambda(t)\, dt = a\tau\left(1 - e^{-T/\tau}\right) \tag{3.3}$$

The probability that a photon incident on the SPAD will be detected is determined by the measured PDP. Further, the measured noise of the detector follows a homogeneous Poisson process with a rate parameter equal to the DCR.

3.4.2 Probability of detecting a true positive event

The probability of detecting a true positive event is the probability that the detector indicates that there was an event and a photon is responsible. This can be described as:

$$P(\text{ Photon Detected } \mid \text{ Hit Recorded }) = \frac{P(\text{ Photon Detected } \cap \text{ Hit Recorded })}{P(\text{ Hit Recorded })} \tag{3.4}$$

The two probabilities on the right-hand side of this equation are written as:

$$\begin{aligned} P(\text{ Photon Detected } \cap \text{ Hit Recorded }) &= P(\text{ Photon Detected }) \\ &= P(\text{ Detecting Photon } \mid \geq 1 \text{ Photon Arrives }) \cdot P(\geq 1 \text{ Photon Arrives }) \\ &= \sum_{k=1}^{\infty} P(\text{ Detecting a Photon } \mid k \text{ Photons Arrive }) \cdot P(k \text{ Photons Arrive }) \\ &= \sum_{k=1}^{\infty} [1 - P(\text{ No Detections } \mid k \text{ Photons Arrive })] \cdot P(k \text{ Photons Arrive }) \end{aligned} \tag{3.5}$$

$$\begin{aligned} P(\text{ Hit Recorded }) &= P(\text{ Photon Detected } \cup \text{ Dark Count }) \\ &= P(\text{ Photon Detected }) + (1 - P(\text{ No Dark Counts })) \end{aligned} \tag{3.6}$$

These probabilities are defined by the Poisson processes and probability theory as:

$$P(k \text{ Photons Arrive }) = \frac{e^{-\mu}\mu^k}{k!} \tag{3.7}$$

$$P(\text{ No Detections } \mid k \text{ Photons Arrive }) = (1 - \text{PDP})^k \tag{3.8}$$

$$P(\text{ No Dark Counts }) = e^{-\text{DCR}\cdot T} \tag{3.9}$$

Combining Equations 3.7, 3.8, 3.9 with Equations 3.5 and 3.6 gives the expressions:

$$
\begin{aligned}
&P(\text{ Photon Detected } \cap \text{ Hit Recorded }) \\
&\qquad = \sum_{k=1}^{\infty} \left[1 - (1 - \text{PDP})^k\right] \cdot \frac{e^{-\mu}\mu^k}{k!} \\
&\qquad = \sum_{k=1}^{\infty} \frac{e^{-\mu}\mu^k}{k!} - \sum_{k=1}^{\infty} (1 - \text{PDP})^k \cdot \frac{e^{-\mu}\mu^k}{k!} \\
&\qquad = \left(1 - e^{-\mu}\right) - \frac{e^{-\mu}}{e^{-\mu\cdot(1-\text{PDP})}} \sum_{k=1}^{\infty} \frac{e^{-\mu\cdot(1-\text{PDP})} \left(\mu \cdot (1 - \text{PDP})\right)^k}{k!} \\
&\qquad = \left(1 - e^{-\mu}\right) - e^{-\mu\cdot\text{PDP}} \left(1 - e^{-\mu\cdot(1-\text{PDP})}\right) \qquad (3.10)
\end{aligned}
$$

$$
\begin{aligned}
P(\text{ Hit Recorded }) &= P(\text{ Photon Detected }) + (1 - (\text{ No Dark Counts })) \\
&= \left(1 - e^{-\mu}\right) - e^{-\mu\cdot\text{PDP}} \left(1 - e^{-\mu\cdot(1-\text{PDP})}\right) + \left(1 - e^{-\text{DCR}\cdot T}\right) \qquad (3.11)
\end{aligned}
$$

Finally, substituting Equations 3.10 and 3.11 into Equation 3.4, the final probability of a true positive detection is:

$$
\begin{aligned}
&P(\text{ Photon Detected } \mid \text{ Hit Recorded }) = \\
&\qquad \frac{(1 - e^{-\mu}) - e^{-\mu\cdot\text{PDP}} \left(1 - e^{-\mu\cdot(1-\text{PDP})}\right)}{(1 - e^{-\mu}) - e^{-\mu\cdot\text{PDP}} \left(1 - e^{-\mu\cdot(1-\text{PDP})}\right) + \left(1 - e^{-\text{DCR}\cdot T}\right)} \qquad (3.12)
\end{aligned}
$$

3.4.3 Probability of recording a true negative event

The probability of recording a true negative event is the probability that no events were recorded when there were no photons incident on the detector. This probability is described

as:

$$\begin{aligned} P(\text{ Recording a Miss } &\mid \text{ No Photons Arrive }) \\ &= \frac{P(\text{ Recording a Miss } \cap \text{ No Photons Arrive })}{P(\text{ No Photons Arrive })} \\ &= \frac{P(\text{ No Dark Counts }) \cdot P(\text{ No Photons Arrive })}{P(\text{ No Photons Arrive })} \\ &= P(\text{ No Dark Counts }) \end{aligned}$$

If the dark count rate is considered as a homogeneous Poisson process, then this probability can be simply written as:

$$P(\text{ Recording a Miss } \mid \text{ No Photons Arrive }) = e^{-\text{DCR}\cdot T} \tag{3.13}$$

3.4.4 Probability of detecting an arriving photon

The final probability to consider is the probability of detecting an arriving photon. This is the probability that a hit that was caused by a photon arrival is recorded in the device and that no dark counts have occurred. This is a measure of the sensitivity of the device.

$$\begin{aligned} P(\text{ Hit Recorded } &\mid \geq 1 \text{ Photon Arrives }) \\ &= \frac{P(\text{ Hit Recorded} \cap \geq 1 \text{ Photon Arrives })}{P(\geq 1 \text{ Photon Arrives })} \end{aligned} \tag{3.14}$$

The components of this probability are given by:

$$
\begin{aligned}
&P\,(\ \text{ Hit Recorded} \cap \geq 1 \text{ Photon Arrives }) \\
&\quad = \sum_{k=1}^{\infty} P\,(\ \text{Hit Recorded } \cap k \text{ Photons Arrive }) \\
&\quad = \sum_{k=1}^{\infty} P\,(\ \text{Hit Recorded } \mid k \text{ Photons Arrive }) \cdot P\,(k \text{ Photons Arrive }) \\
&\quad = \sum_{k=1}^{\infty} [1 - P\,(\ \text{No Hit Recorded } \mid k \text{ Photons Arrive })] \cdot P\,(k \text{ Photons Arrive }) \\
&\quad = \sum_{k=1}^{\infty} [1 - P\,(\ \text{No Detections } \cap \text{ No Dark Counts } \mid k \text{ Photons Arrive })] \\
&\qquad\qquad \cdot P\,(k \text{ Photons Arrive }) \\
&\quad = \sum_{k=1}^{\infty} [1 - P\,(\ \text{No Detections } \mid k \text{ Photons Arrive }) \cdot P\,(\ \text{No Dark Counts })] \\
&\qquad\qquad \cdot P\,(k \text{ Photons Arrive }) \qquad (3.15)
\end{aligned}
$$

$$
P\,(\ \geq 1 \text{ Photon Arrives}) = 1 - P\,(\ \text{No Photons Arrive }) \qquad (3.16)
$$

Combining Equations 3.7, 3.8, and 3.9 with Equations 3.15 and 3.16 gives the expression for the probability of detecting an incident photon:

$$
\begin{aligned}
&P\,(\ \text{ Hit Recorded } \mid \geq 1 \text{ Photon Arrives }) = \\
&\quad = \frac{1}{1 - e^{-\mu}} \cdot \sum_{k=1}^{\infty} \left[1 - (1 - \text{PDP})^k \cdot e^{-\text{DCR}\cdot T}\right] \cdot \frac{e^{-\mu}\mu^k}{k!} \\
&\quad = \frac{1}{1 - e^{-\mu}} \cdot \left[\sum_{k=1}^{\infty} \frac{e^{-\mu}\mu^k}{k!} - e^{-\text{DCR}\cdot T} \sum_{k=1}^{\infty} (1 - \text{PDP})^k \cdot \frac{e^{-\mu}\mu^k}{k!}\right] \\
&\quad = \frac{1}{1 - e^{-\mu}} \cdot \left[\sum_{k=1}^{\infty} \frac{e^{-\mu}\mu^k}{k!} - \frac{e^{-(\text{DCR}\cdot T + \mu)}}{e^{-\mu\cdot(1-\text{PDP})}} \cdot \sum_{k=1}^{\infty} \frac{e^{-\mu\cdot(1-\text{PDP})}\left(\mu \cdot (1 - \text{PDP})\right)^k}{k!}\right] \\
&\quad = \frac{1}{1 - e^{-\mu}} \cdot \left[\left(1 - e^{-\mu}\right) - \frac{e^{-(\text{DCR}\cdot T + \mu)}}{e^{-\mu\cdot(1-\text{PDP})}} \left(1 - e^{-\mu\cdot(1-\text{PDP})}\right)\right] \\
&\quad = 1 + \frac{e^{-(\text{DCR}\cdot T + \mu)} - e^{-(\text{DCR}\cdot T + \mu\cdot\text{PDP})}}{1 - e^{-\mu}} \qquad (3.17)
\end{aligned}
$$

3.4.5 Figure-of-Merit

From these probabilities, a figure-of-merit (FOM) is developed for a SPAD used for FLIM. This FOM is the product of the true positive probability, the true negative probability, and the probability of detecting an incident photon. This optimizes for a device that accurately records events while maintaining sensitivity to incident photons. This combined figure-of-merit is:

$$FOM = e^{-\mathrm{DCR}\cdot T}\cdot\left[\frac{(1-e^{-\mu})-e^{-\mu\cdot\mathrm{PDP}}\left(1-e^{-\mu\cdot(1-\mathrm{PDP})}\right)}{(1-e^{-\mu})-e^{-\mu\cdot\mathrm{PDP}}\left(1-e^{-\mu\cdot(1-\mathrm{PDP})}\right)+(1-e^{-\mathrm{DCR}\cdot T})}\right] \cdot\left[1+\frac{e^{-(\mathrm{DCR}\cdot T+\mu)}-e^{-(\mathrm{DCR}\cdot T+\mu\cdot\mathrm{PDP})}}{1-e^{-\mu}}\right] \tag{3.18}$$

By assuming that $\mu << 1$ and $\mathrm{DCR}\cdot T << 1$, Equation 3.18 reduces to:

$$FOM = PDP\cdot\left[\frac{\mu\cdot\mathrm{PDP}}{\mu\cdot\mathrm{PDP}+\mathrm{DCR}\cdot T}\right] \tag{3.19}$$

This FOM provides a simple guideline for biasing and comparing SPADs. Recently, others have published FOMs using a similar approach that focuses on maximizing the signal-to-noise ratio as a general optimization goal without considering the specific application of FLIM [94].

3.4.6 Comparison to Simulated FLIM Data

In order to evaluate the chosen FOM, a simulated FLIM experiment is used to determine the minimum number of laser repetitions needed to extract a lifetime, within a chosen accuracy, for several SPAD operating points and fluorescence intensity levels. The time-rescaling method [95] is used to efficiently generate photon arrival times using MATLAB. As previously described, the DCR was considered as a homogeneous Poisson process and the PDP as a scaling factor for the incident photon rate. Lifetime and system parameters were chosen to match [65] ($\tau = 3$ ns, $T = 20$ ns). Additionally, a timing resolution of 50 ps was

assumed, and the intensity, a, was chosen such that 1% or less of the measurement windows contain hits.

Each simulated experiment consisted of N laser repetitions and used a non-linear least squares method to fit the monoexponential decay. For each value of N, the experiment was repeated 300 times and the standard deviation was used as a measure of the error for the set of N experiments. A successfully extracted lifetime was defined as a set of 300 repetitions whose standard deviation was within 10% of the known true lifetime (3 ns for these parameters). Once this threshold has been reached, the experiment is stopped and N is considered to be the minimum number of laser repetitions required to accurately determine the lifetime for that set of parameters. The results of these simulations are shown in Fig. 3.4. For comparison, the probability-based FOM derived in the previous sections is plotted in Fig. 3.5 for the same fluorescence intensity values as Fig. 3.4. The minima in both Fig. 3.4(a) and Fig. 3.4(b) correspond with the maximum FOM for the respective intensities in Fig. 3.5. This serves as evidence supporting the FOM as a method for optimally choosing the SPAD bias point such that the number of laser repetitions (and, consequently, image acquisition time) can be minimized.

With a FLIM optimized FOM, it is now possible to evaluate, compare, and optimize a SPAD for use in TCSPC FLIM measurements. The following section describes the CMOS SPAD used in this thesis.

3.5 CMOS SPAD Implementation

The details of the SPAD used for the imaging array design are included in this section. In addition to this SPAD, there were four other test chips that were fabricated to evaluate SPADs in 0.35 μm Austrian Microsystems (AMS) and 0.13μm International Business Machines (IBM) processes. The details of these test chips are provided in Appendix A and the result of these test chips was a low-noise SPAD developed in both 0.35μm AMS and

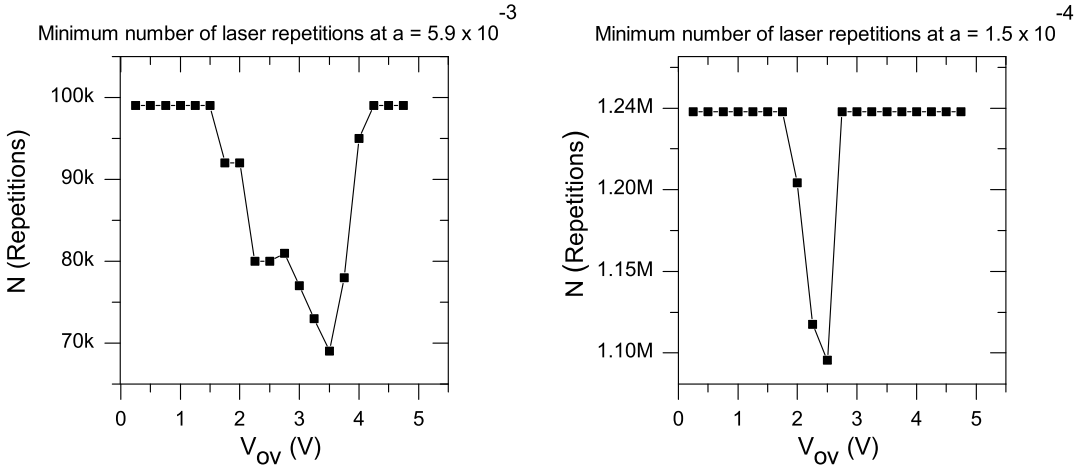

Figure 3.4: Simulated results for the minimum number of laser repetitions, N, necessary in order to achieve a standard deviation of 0.3 ns for 300 repeated fittings of the monoexponential decay at each bias point. a) An intensity given by $a = 5.9 \times 10^{-3}$. b) An intensity given by $a = 1.5 \times 10^{-4}$.

0.13μm IBM CMOS technologies. The 0.13μm design was used for this work so that the faster transistors could be leveraged for better timing precision in the TDCs and improved data handling techniques. The 0.35μm AMS design was never incorporated into an array but was leveraged in a quantitative polymerase chain reaction (qPCR) on CMOS chip design [96].

3.5.1 SPAD Design

A custom low-noise SPAD was designed in an 0.13μm IBM radio frequency (RF) process for integration into a large array. As mentioned in Section 3.3, a combination of the implants that are traditionally used for NFETs and PFETs must be adapted in order to generate the isolated well structure illustrated in Figure 3.3. Additionally, the multiplication region of the device should be physically isolated from the STI to prevent charge traps at the interface with the STI from generating a high DCR.

A diagram showing the cross section and the corresponding CMOS design mask layers

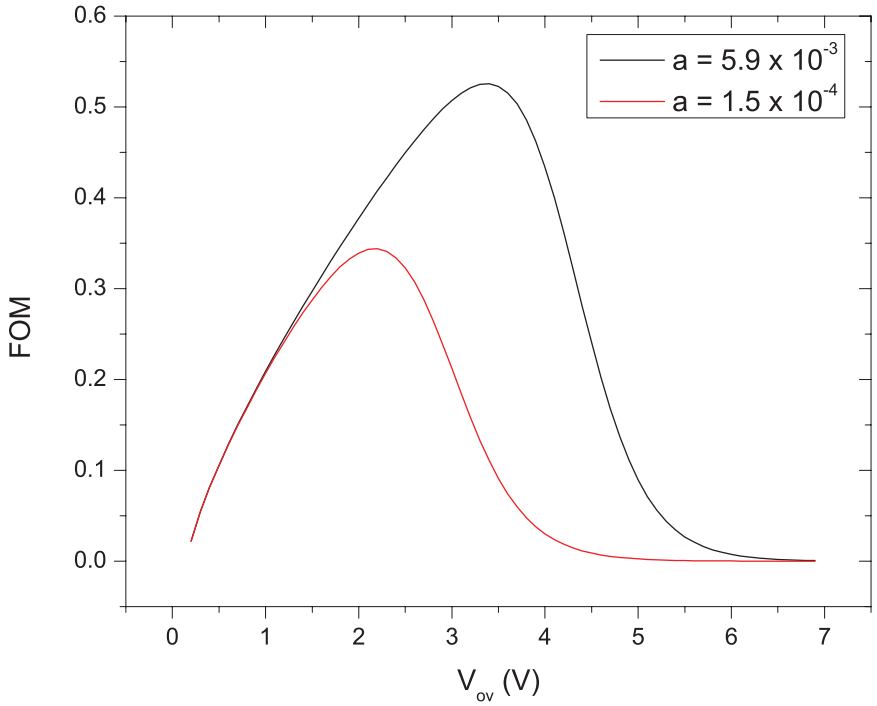

Figure 3.5: Plot of the FOM given by Equation 3.18 at two different intensities that correspond with the simulated results in Figure 3.4.

is presented in Figure 3.6. Commonly in standard STI CMOS processes, the only mechanism for blocking STI is with the use of the "active" mask, labelled RX in Figure 3.6. This mask defines that region of silicon as either an NFET or a PFET and combines with the BP design mask to define the regions of silicon that will be doped with a P^+ implant. Because CMOS processes are complementary, the N^+ implant mask in this technology is a processed mask and is defined as any area covered by RX and not BP. To separate the STI from the multiplication region, it is desirable to have RX that is doped with neither the P^+ nor the N^+ implants. The BN mask drawn in Figure 3.6 is a development mask (not typically used in CMOS designs) that blocks the N^+ implant in regions covered by BN that are adjacent to BN covering BP. This creates a region surrounding the main active area of the device that is neither P^+ nor N^+ between the STI and the multiplication region. The NW mask is used for the NW implant that serves as the cathode of the SPAD while the PI mask produces the p-type guard ring surrounding the active region of the SPAD and a deep n-type implant for joining the two NW regions below the guard ring to prevent shorting between the guard

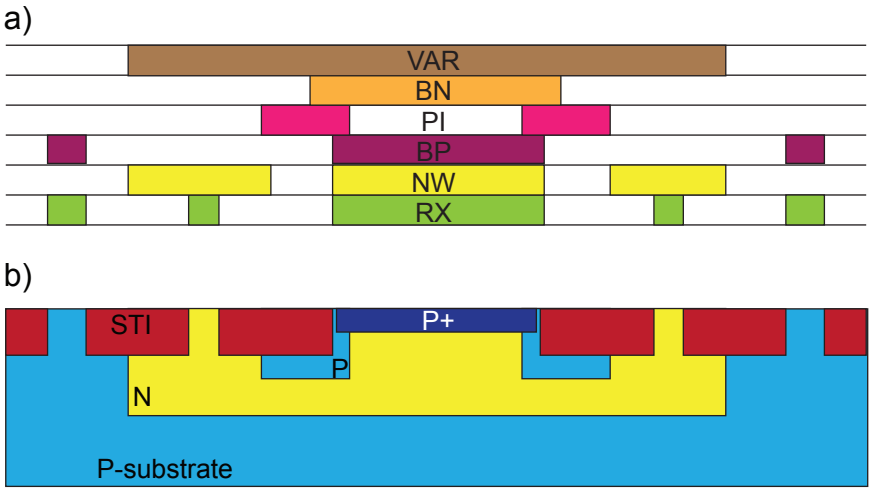

Figure 3.6: a) The mask layers used in the CMOS process design environment to fabricate the SPAD. b) The anticipated SPAD cross-section that would result from a fabrication run using the drawn masks.

ring and substrate, which is also p-type.

These drawn masks define the key features for the CMOS SPAD. However, CMOS processes are optimized for the design of high quality transistors and a number of masks are generated from the design masks that are used to enhance the properties of short-channel MOSFETS. These masks include features like lightly-doped drain implants, source and drain halo implants, and self-aligned silicide (salicide) regions. These additional implants can extend beyond the drawn dimensions of the P^+ and N^+ regions and lead to unexpected shunting across the diode. The salicide is used to make low resistance contacts to the MOSFET and is grown wherever the RX mask is drawn. If the SPAD is designed with adjacent n-type and p-type regions, this salicide can short them together. In order to prevent these enhancement features from negatively impacting the SPAD, additional masks are drawn to block these layers. The VAR mask is traditionally used for making variable capacitors (varactors) where the additional MOSFET implants are not needed. This mask blocks the lightly-doped drain and halo implants. Additionally, the OP mask (not drawn here) can be used to block the salicide formation. This layer is traditionally used to increase the resistance for on-chip polysilicon resistors.

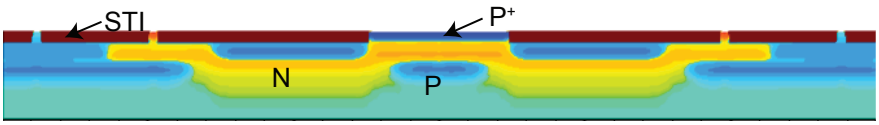

Figure 3.7: Simulated cross-section of SPAD using TCAD software.

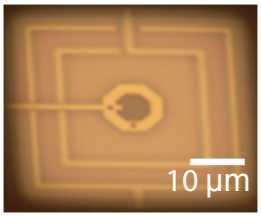

Figure 3.8: A photograph of the SPAD used in this work.

The full process recipe for this technology was used to simulate the implants, annealing stages, masks, and etch processes for the entire 0.13μm process. The technology computer aided design (TCAD) software Sentaurus was used to simulate the structure outlined in Figure 3.6. The results of this simulation are shown in Figure 3.7. This device was then fabricated and I-V, PDP, DCR, IRF, and afterpulsing measurements were taken for the device. Although the simulated STI is close to the multiplication region, this device had the lowest noise out of any of the structures that were tested in this technology, which was unexpected. One possible explanation for this is an incomplete understanding of the technology flow or a systematic mask alteration that increased the STI separation. While the process recipe was made available, the exact computational algorithms for producing the masks were not. Additionally, post-computational processing masks were never obtained from the foundry for comparison between the modelled and the fabricated masks.

3.5.2 SPAD Characterization

Figure 3.8 is a photograph of the fabricated SPAD. The SPAD was designed with an octagonal active area in order to meet design rules for the CMOS process and has a diagonal dimension measuring 5μm. An additional mask layer was used to remove the polyimide over

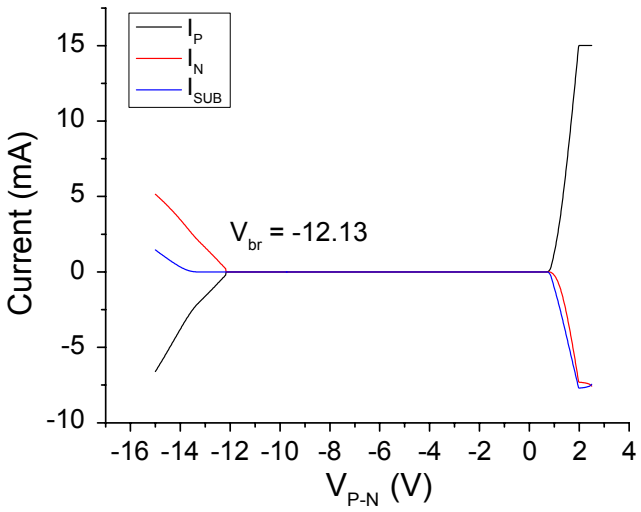

Figure 3.9: The current-voltage relationship for the fabricated SPAD. The reverse bias breakdown voltage, V_{br} is -12.13V. There is a substrate current that begins flowing when the reverse bias voltage reaches -13.7V.

the device so that active region of the device will only be covered by the transparent silicon dioxide. The I-V characteristic for this device is presented in Figure 3.9 and the device exhibits a reverse bias breakdown voltage, V_{br}, of -12.13V. Measurements of the SPAD DCR, PDP, IRF, and afterpulsing were taken using an off-chip passive quenching resistor (as shown in Figure 3.2) with a measured value of 423 kΩ.

The DCR was characterized by covering the SPAD test setup with a blackout sheet and recording the average number of counts over a period of 10 seconds using an Agilent 53132A universal counter. The overvoltage, V_{ov}, was swept over a range of 0.25V to 1.50V in steps of 0.25V. A plot of the DCR and a best fit line showing an exponential dependence on V_{ov} is presented in Figure 3.10. The DCR is as low as 6 Hz at 0.25V and reaches only 231 Hz at 1.5V, measured at room temperature.

Measurements of the PDP were performed using the same passively quenched configuration described above. This passive quenching technique will limit the measured PDP due to the long reset time that results from resetting the SPAD through the large quenching resistance. To measure PDP, a xenon arc lamp is used with a Spectral Products CM110 monochromator to produce a narrow-band light source. Additionally, the output of the

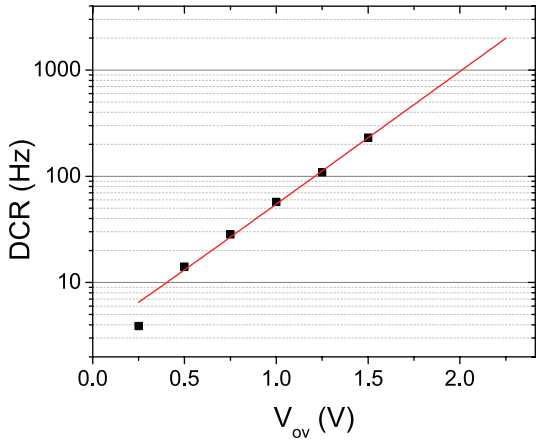

Figure 3.10: The measured dark count rate reaches only 231 Hz at a V_{ov} of 1.5V. Beyond 1.5V, the substrate current seen in the I-V curve causes SPAD failure.

monochromator is passed through a 2-inch integrating sphere to produce a uniform illumination over the SPAD. A Thorlabs PM130D power meter is used to measure the incident light intensity on the SPAD, from which the number of incident photons per second can be calculated. The photon flux was calculated using:

$$\text{Photons/s/m}^2 = \frac{P \cdot \lambda}{A \cdot h \cdot c} \tag{3.20}$$

where P is the power measured by the power meter, λ is the wavelength of light, A is the area of the power meter detector, h is Planck's constant, and c is the speed of light. For these measurements, the photon flux was approximately 1×10^{15} Photons/s/m^2 for all wavelengths. The PDP versus V_{ov} for the visible spectrum is plotted in Figure 3.11a. Additionally, the peak PDP value versus V_{ov} is plotted in Figure 3.11b. The peak PDP is approximately linear with V_{ov} whereas the DCR increase exponentially, which leads to the complexity associated with determining an optimal operating point and comparing SPADs that was addressed in Section 3.4. It should also be noted that the peak sensitivity for this SPAD occurs at 425nm, which is consistent with the shallow junction depths expected in the 0.13μm CMOS technology. This peak sensitivity is well suited for measuring fluorophores like NADH that have emission peaks near 450nm.

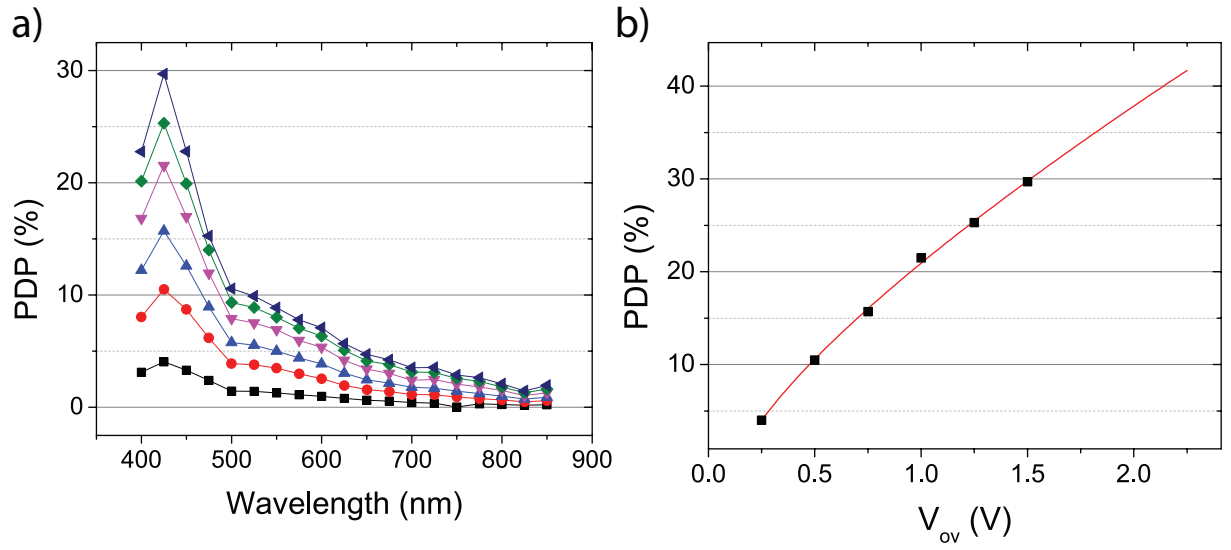

Figure 3.11: a.) The SPAD PDP for the visible wavelengths. V_{ov} is increased from 0.25V (bottom black curve) to 1.5V (top blue curve) in steps of 0.25V. b) A plot showing the almost linear relationship between the PDP and V_{ov}. This linear relationship is extrapolated to 2.0V but will eventually saturate, limited by the reset time of the SPAD.

The impulse response function of the SPAD was measured using a pulsed 408nm laser with a specified full width at half maximum (FWHM) pulse width of 45 ps. An HP 8118A pattern generator produced a trigger signal that was used to trigger the laser and an oscilloscope. The SPAD response was measured using a Tektronix TDS7404 oscilloscope with a sampling rate of 20 Gsps. The SPAD was biased with $V_{ov} = 1.5$V and the resulting IRF of the system (laser, SPAD, and oscilloscope) was 198ps and is plotted in Figure 3.12.

The afterpulsing probability for this SPAD was measured but the long R-C reset time of the passive quenching resistor resulted in a negligible afterpulsing probability. A detailed analysis of the afterpulsing probability using an active quench and reset circuit can be found in Section 4.2.

3.6 Summary

The device presented in this chapter is a low-noise SPAD that is sensitive to visible wavelengths and is well suited for inclusion in an imaging array. The standard 0.13 μm CMOS

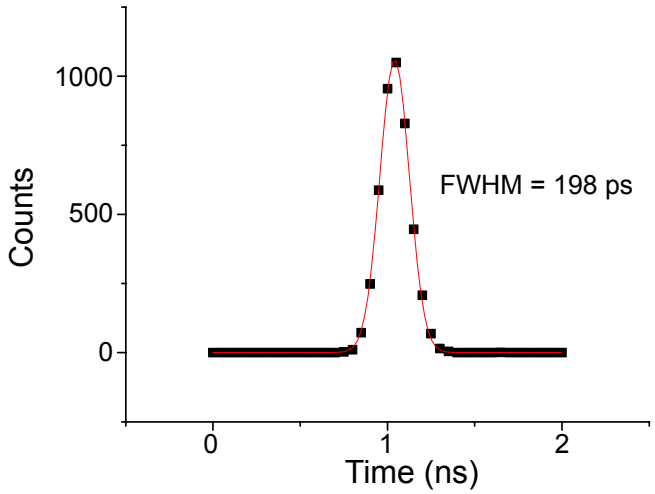

Figure 3.12: The instrument response function for the SPAD. The FWHM of 198ps includes the instrument response of the laser and the oscilloscope that was used for the measurement.

technology node in which it was designed includes compact and fast devices that can enable high speed TCSPC FLIM data acquisition. The following chapters discuss the integration of this SPAD into a high speed imaging platform.

Chapter 4

Wide Field Fluorescence Lifetime Imager Integrated Circuit Design

This chapter presents the design and characterization of an imager integrated circuit (IC) that was optimized for wide field fluorescence lifetime imaging. The imager consists of an array of SPADs that was designed for time-correlated single-photon counting (TCSPC) based fluorescence lifetime imaging at high frame rates. It can be used as an attachment to the camera mount (c-mount) port of a standard upright microscope, as an active imaging substrate on which biological samples can be directly placed, or as a multi-hit point detector for laser scanning applications.

This array consists of the single-photon avalanche diodes (SPADs) that were developed in standard IBM 0.13 μm CMOS technology and presented in Chapter 3. A few key characteristics of this CMOS process are presented in Table 4.1. The triple well feature of this technology is used to isolate all circuits from the substrate in order to minimize noise coupled through the substrate due to SPAD events. In addition, each pixel incorporates an active quench and reset circuit and has an independent time-to-digital converter (TDC) channel. The output from the TDCs passes through a datapath that was optimized for TCSPC data before it is buffered off-chip.

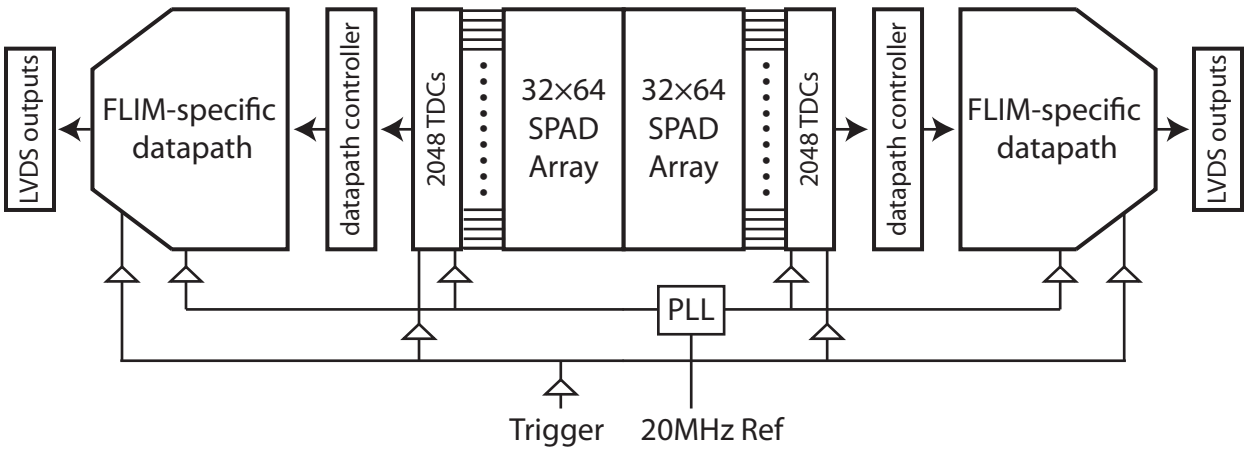

Figure 4.1: A block diagram showing the major functional components of the FLIM IC.

Table 4.1: Characteristics of the IBM 0.13 μm technology.

Characteristic	Value	Units
Core V_{dd}	1.2 - 1.6	V
I/O V_{dd}	2.5 - 2.7	V
Thick FET V_{dd}	3.3 - 3.6	V
Thin Metals	6	
Thick Metals	2	
FO1	18	ps
FO4	43	ps
β-Ratio FO1	2.4	
β-Ratio FO4	3.0	

The IC consists of a 64-by-64 array of pixels divided into quadrants of 32-by-32 pixels. A total of 32 delay-locked loops (DLLs) generate fine timing phases for each TDC, 16 first-in-first-out (FIFO) buffers synchronize data from the datapath to four banks of 21 low-voltage differential signaling (LVDS) output buffers. This design provides a timing resolution of 62.5 ps with a measurement window of up to 64 ns. The output bandwidth of the datapath and LVDS buffers supports lifetime image acquisition at up to 100 frames per second. A block diagram of the overall architecture is presented in Figure 4.1.

4.1 Imager Design

This imaging array was designed and fabricated using a standard IBM 0.13 μm CMOS process. By integrating the previously presented low-noise SPADs in this technology, this imager is able to leverage faster circuits and achieve an overall imaging frame rate that is faster than any previously demonstrated TCSPC-based system. In particular, this technology has a fan-out-of-one (FO1) of 18 ps and a fan-out-of-four (FO4) of 43 ps. These circuit performance metrics establish limits on the maximum on-chip clock frequency and the TDC precision. The timing targets for this design were a 1 GHz clock and a TDC precision of 62.5 ps. Additionally, this 0.13 μm process includes mask layers for triple well implants. This allows for complete isolation of all circuits from the substrate, which is directly coupled to the SPADs. This isolation can be used to avoid substrate noise associated with SPAD events.

4.1.1 Imager Architecture Overview

A die photograph of the imager IC with all of the major functional blocks highlighted is shown in Figure 4.2. The IC measures 9.1 mm x 4.2 mm and consists of approximately 10 million transistors. In the center of the IC is the 64-by-64 array of SPADS. Each SPAD has an octagonal active area with a 5 μm diagonal, resulting in an active area of 17.68 μm^2 per device. Due to the control circuitry within each pixel and the required spacing between diffusions in the SPAD structure, the pixel-to-pixel pitch is 48 μm. This corresponds to a total array imaging area of 9.43 mm^2 and an overall fill factor of only 0.77%. The details of the SPADs were presented in Section 3.3 and the circuits included in each of the pixels are detailed in 4.1.2, below. An image showing the layout of a single pixel within the array and associated control circuitry is presented in Figure 4.3.

The two time-to-digital converter blocks each contain 2048 TDCs that are driven by 16 delay-locked loops (DLLs). The specifics of the TDC blocks are provided in Section 4.1.3.

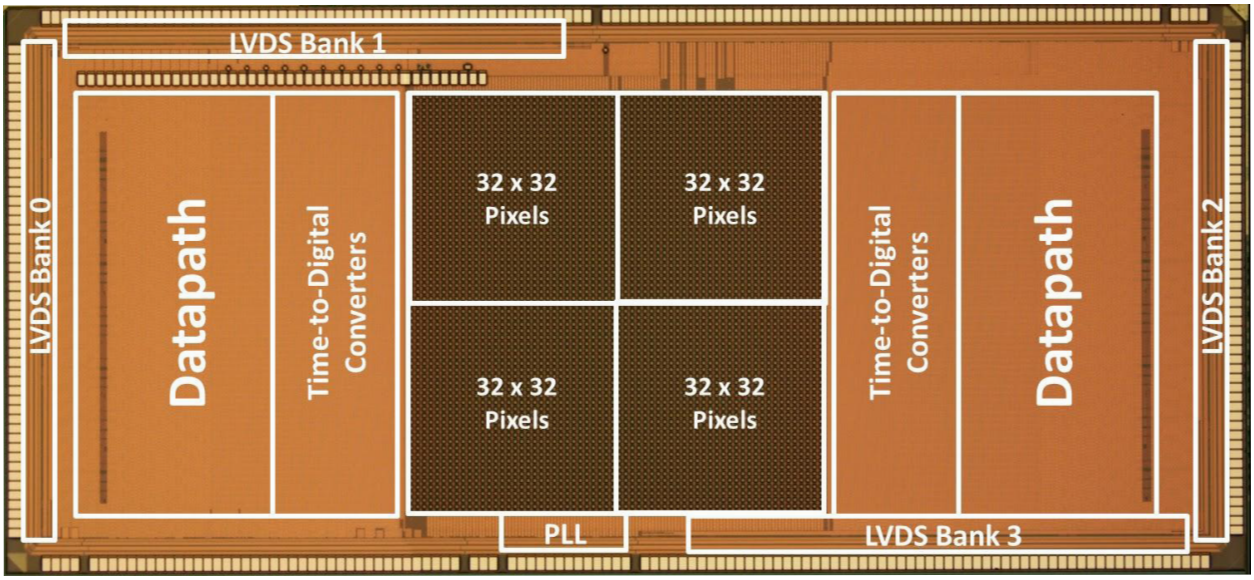

Figure 4.2: Die photograph showing the major functional blocks of the FLIM imaging IC from Figure 4.1.

The datapath circuitry has been designed specifically for processing FLIM data as it moves from the TDCs' registers to the output buffers at the periphery. This datapath is crucial in reducing the output data bandwidth required by ensuring that only data from pixels with SPAD events are transmitted off-chip. Section 4.1.4 provides a detailed description of the datapath circuitry.

In order to continuously transmit large volumes of data, the imager has been designed with four banks of 21 LVDS output buffers. These LVDS buffers have been designed to operate at up to 500 MHz, providing a total data bandwidth of 42 Gbps. The design considerations for the LVDS drivers are found in Section 4.1.5.

In addition to the primary FLIM-related circuits, a phase-locked loop was included for synchronizing the global clock network to the laser used for FLIM experiments and a global scan-chain is used for configuring the circuits in the array. The considerations for designing the PLL are presented in Section 4.1.6. A brief overview of the control circuits is also presented in Section 4.1.7.

All simulations presented in this section were performed using Spectre models in the Cadence Analog Design Environment, unless stated otherwise.

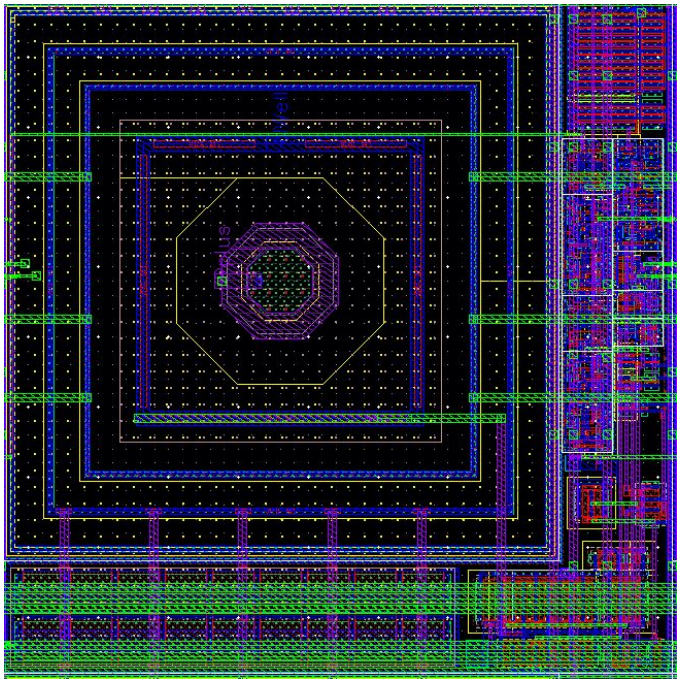

Figure 4.3: The layout of a single pixel in the array, including SPAD and the pixel control circuit of Fig. 4.4. The pixel and circuitry occupy an area that is 48 μm $\times$ 48 μm, of which a considerable amount is white space due to the conservative SPAD structure and guard rings used.

4.1.2 Pixel Circuitry

In this section the details of the pixel control circuits are presented. At a minimum, each pixel must have the ability to detect an avalanche when the SPAD is triggered, quench the SPAD avalanche, and reset the SPAD following the quench. In addition, overall imaging performance can be improved by minimizing noise events at each SPAD. Two particular areas of concern are reducing afterpulsing and disabling pixels with unusually high DCR levels. Figure 4.4 shows the complete circuit schematic for the SPAD control circuit. In addition to drawn circuits, two flip-flops for selecting between the calibration or disabled pixel modes that are part of the global scan chain and not shown here. Local decoupling capacitors for the pixel output buffers are also included in every pixel. The flip-flops are D-type flip-flops from an Artisan standard cell library and the capacitors are comprised of a combination of metal-insulator-metal (MIM) capacitors and thin oxide NMOS capacitors (NCAPs) in order to maximize the amount of capacitance per area in each pixel.

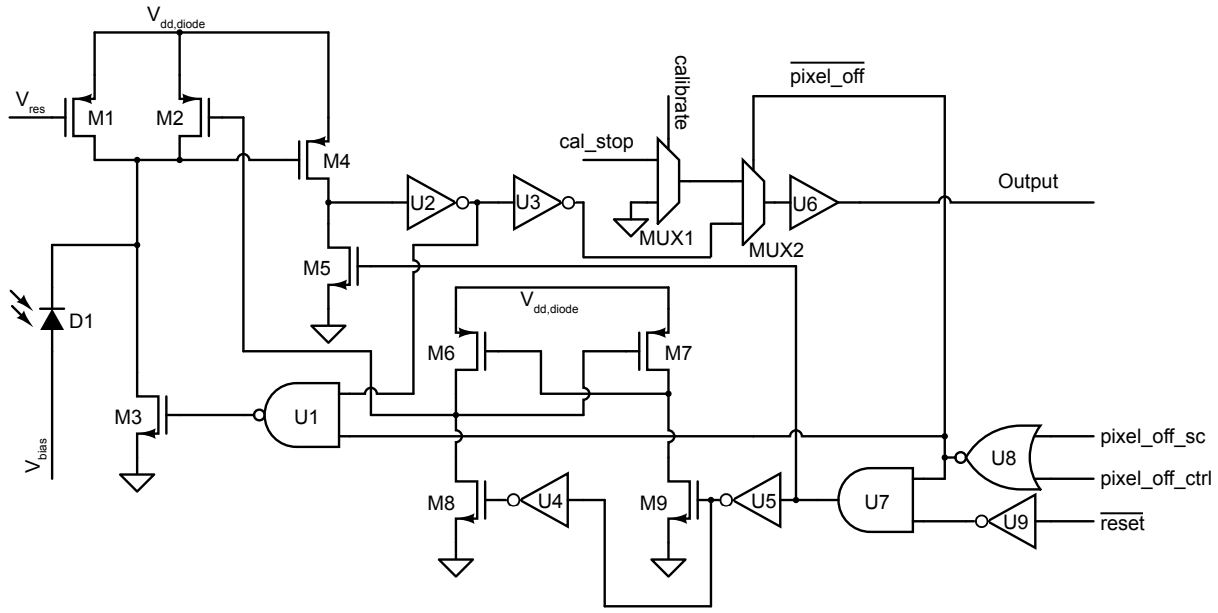

Figure 4.4: Detailed circuit schematic of the SPAD sense, quench, reset, and control. Transistors M1-M9 and inverter U2 are designed using thick oxide devices.

Event Detection

When a photon arrives at the detector and produces an avalanche, the avalanche must be detected and communicated to the TDCs so that the arrival time can be recorded. There are a number of approaches to achieving this task.

In traditional TCSPC setups where PMTs are used, event detection is typically performed using a constant fraction discriminator (CFD). Because the pulse height from the PMT will vary between arrival events, setting a reliable fixed threshold can be challenging. A CFD is used to detect when a fast pulse crosses a certain fraction of the pulse height, which is independent of the actual height of an individual pulse [97]. A common CFD circuit will generate an inverted and attenuated version of the input pulse as well as a time delayed version, then add them together to create a bipolar signal that can be detected using a zero-crossing detector [98]. This type of design would result in a significant overhead when designing and array of detectors, each requiring independent level sensors.

With SPAD-based systems, pulse detection can be simplified. Every time an event triggers the detector, a voltage drop equal to the overvoltage, V_{ov}, can be observed across

the quenching device. If V_{ov} is greater than one transistor threshold voltage, V_t, a single transistor can be used to detect the pulse. In the IBM 0.13 μm technology, the approximate threshold voltages available for each of the different types of field effect transistors (FETs) are shown in Table 4.2. In order to have sufficient sensitivity for making lifetime measurements, the SPADs are generally biased with a V_{ov} greater than 0.5 V. Consequently, event detection can be performed using any of the FET devices available in this technology.

In Figure 4.4, the PFET M4 is used for event detection. This PFET is a thick oxide 3.3V device, which has an allowable tolerance for gate to source voltages of up to 3.6 V. This allows the detection of any events for SPADs with V_{ov} between 0.35 V and 3.6 V. V_{ov} for this circuit is determined by $V_{dd,diode} - V_{diode,p} - V_{br}$. Special care must be taken to ensure that all of the P-N junction diffusions are properly reverse-biased. Typically, the substrate potential is held at 0 V. This places a constraint on the minimum possible voltage applied to the cathode of the SPAD, which must be less than the forward bias potential of the P-N junction, $-V_f$. Thus, the voltages $V_{dd,diode}$ and $V_{diode,p}$ should be chosen such that the cathode will never drop below this threshold during an avalanche. Typical safe operating voltages for a V_{ov} of 1.0 V with these SPADs and a substrate voltage of 0 V are $V_{dd,diode} = 2.5$ V and $V_{diode,p} = -10.10$ V.

The inverter U2 is connected to the core IC power supply and level-shifts the output from the event detection PFET, with a supply of $V_{dd,diode}$, to the core logic voltage for the rest of the IC (typically 1.5 V). Following a second inverter, U3, are two multiplexers for selecting between the SPAD output, an electrical calibration input, or a constant 0 V signal for turning off the output. The output from the multiplexers is then buffered to drive a $\sim$3.5 mm long wire connecting the pixel output to the stop signal of the TDC. Special care was taken when performing the physical design of these output buffers and long wires due to the close spacing required to transmit stop signals from all 4096 pixels. In particular, top level metal was used in order to minimize resistance of the long wires, and cross-talk between them should be minimized due to the asynchronous nature of photon stop times throughout

Table 4.2: Approximate threshold voltages for the various types of field effect transistors available in the IBM 0.13 μm technology. All V_t values are $V_{t,sat}$ for the minimum length & width device. PFET values are given as the absolute value.

NFET Type	V_t (V)	PFET Type	V_t (V)
nfet	0.300	pfet	0.350
lpnfet	0.475	lppfet	0.455
lvtnfet	0.150	lvtpfet	0.275
nfet25	0.400	pfet25	0.450
nfet33	0.375	pfet33	0.325

the array.

SPAD Quenching

Following the discussion in Section 3.3, when a photon event triggers a SPAD avalanche, the avalanche current must be stopped so that the device can be reset and used in subsequent detection windows. In order to stop the current, the voltage across the device must be reduced to below its breakdown voltage, V_{br}, in a process called quenching.

The most straightforward method for quenching a SPAD is to use a resistor in series with the device as was shown in Figure 3.2a of Section 3.2. The need for low current and a short avalanche duration led to the selection of a large quenching resistance. In this design, a PFET device, M1 in Figure 4.4, is used as the quenching resistor. A tunable voltage, V_{res}, is applied to the gate of the PFET, which allows the drain to source resistance, R_{ds}, of the device to be adjusted. The resistance of this PFET is 400 kΩ when V_{gs} is 500 mV and can be tuned from 10 kΩ to several MΩ as V_{gs} approaches the threshold voltage. Non-linearity in the resistance of the PFET is inconsequential for this application since the gate of the PFET is not actively modulated and the exact resistance does not matter.

SPAD Resetting

After the SPAD has been quenched, the device must be reset before it can be used to detect another event. Because the quenching resistor is placed in series with the SPAD, the SPAD

will begin to passively reset by recharging the cathode of the SPAD to $V_{ov} + V_{br}$ through this series resistance. Since a large resistor is typically used for quenching the device, the R-C time constant for resetting the device through this resistance can be on the order of several microseconds. An example waveform that shows the quench and reset phases of the device used in this work is plotted in Figure 3.2b. Here, a 1 MΩ resistor is used to passively quench and reset the device. The measured device was wire-bonded to a package, placed in a socket, and measured through connections on a PCB. Thus, a significant parasitic capacitance contributes to the overall R-C delay in this measurement, in addition to the large reset resistor.

Instead of waiting for this passive R-C recharge process, the reset time can be reduced by using an active reset mode where a low resistance path is switched on during the reset phase. It is sometimes also desirable to control the timing of the reset process so that afterpulsing is minimized. With an active reset mode, precise control of the reset device is possible and techniques for reducing afterpulsing can be implemented. Discrete active quenching and reset circuits were first demonstrated in previous work by Nightingale [99]. Integrated active quench and reset circuits were later demonstrated by Zappa et. al. [100]. The implementation of active quenching and reset circuits can vary widely in complexity. The schematic of Figure 4.4 shows the circuit implementation used in this work. A wide channel PFET device, M2, with a resistance of 1 kΩ to 400 Ω, depending on $V_{dd,diode}$, is used to reset the device in approximately 500 ps. The SPAD capacitance is estimated to be less than 20 fF based on its area and the junction capacitance for a PFET in this technology. In addition, the NFET M3 is used to hold the bias across the SPAD below breakdown and can be used to prevent the SPAD from resetting or to disable the SPAD completely. M2 and M3 are independently controlled to minimize the probability for after-pulsing.

The primary reset path used during normal operation for imaging is shown in Figure 4.4. In this path, when an avalanche event occurs, the output of inverter U2 is fed back into a NAND gate, U1, which enables M3 and holds the SPAD bias below V_{br}. NFET M3 stays

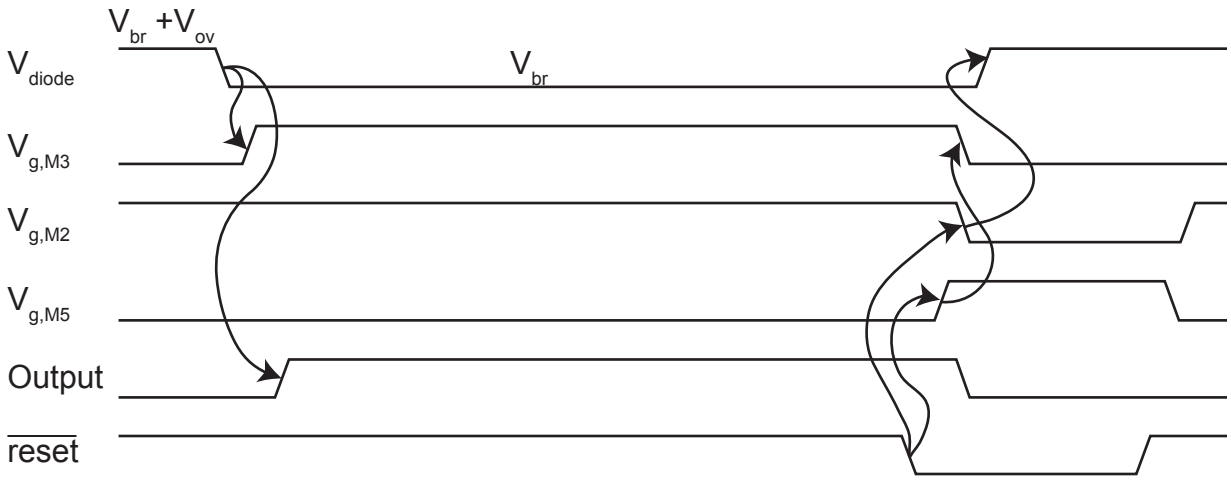

Figure 4.5: Timing diagram for pixel event and reset.

enabled until the $\overline{\text{reset}}$ signal is asserted. When the $\overline{\text{reset}}$ signal is asserted, it passes through an AND gate, which allows for the option to disable the pixel, and a level-shifter that brings the $\overline{\text{reset}}$ signal up to the $V_{dd,diode}$ supply level. While the pixel is being reset, device M2 is enabled and charges the cathode of the SPAD. In order to avoid a fight between M2 pulling up and M3 pulling down, the $\overline{\text{reset}}$ signal is also passed to device M5. When M5 turns on, it pulls the input to U2 low, and causes U1 to pull down, turning off M3. The timing diagram of this event detection and reset cycle is presented in Figure 4.5.

The $\overline{\text{reset}}$ signal is generated by a controller in the datapath that is timed relative to the laser trigger signal such that a reset event will occur immediately after a laser pulse. With the 1 ns reset time and the electrical delay of the trigger signal through the IC, the SPAD is gated such that the laser pulse intensity (FWHM 20 ps) will diminish before the SPAD is reset. This allows for time-gated operation without the need for optically filtering the excitation source from the fluorescence emission.

Calibration and Disablement

Two additional capabilities included in each pixel are the ability to trigger the pixel output buffer using an electrical calibration signal and the ability to disable the pixel. The electrical

calibration path is used to calibrate the TDCs using a known delay. This measurement will be discussed in more detail in Section 4.1.3.

There are two signals that can be used to turn off the pixel. The pixel_off_sc signal is a scan chain bit that can be used to completely disable the pixel during all measurements. By using this control signal to disable abnormally noisy pixels, data bandwidth that would have been consumed by these noise events will be freed for pixels that are contributing meaningful data. The pixel_off_ctrl signal comes from a datapath controller and is used to disable the pixel at the end of a measurement window, if no events have occurred. The details of the controller that asserts this signal are presented in Section 4.1.7. This controller has a programmable timer that allows the user to specify how long to activate the SPAD. This provides the capability to set a measurement window that is appropriate for the lifetimes under observation, leading to a reduction of the $\mathrm{DCR} \cdot T$ term in Equation 3.18 and resulting in better optimized performance for FLIM measurements. A similar technique for defining the measurement window has recently been used to achieve extremely low noise levels in CMOS SPADs for measurements where the photon arrival time is tightly constrained, like 3-D imaging [101].

4.1.3 Time-to-Digital Converters

The time-to-digital converter (TDC) is used to measure the arrival time of the first photon detected during each measurement window. The design targets for this TDC were a 62.5 ps resolution with a 64 ns range. These parameters will provide sufficient resolution and range for most biological fluorophores, which can have lifetimes ranging from hundreds of picoseconds to several nanoseconds.

There are a number of TDC design architectures, all with varying levels of precision, power consumption, area, and complexity. TDCs are becoming an important circuit block in modern CMOS designs as they are increasingly used as phase detectors in all-digital PLLs for frequency generation in advanced technology nodes [102, 103]. Additionally, continued

research into TDCs is being driven by applications that require time-of-flight measurements like laser range finding [104], 3-D imaging [105], and positron emission tomography [106].

Early TDC designs leveraged analog-to-digital converters (ADCs) to perform the time-to-digital conversion. These ADC based designs require two steps: first is the conversion of the time interval to a voltage, second is the translation of that voltage into a digital value. Typically the time-to-voltage conversion is done by using the time interval being measured to control the width of a digital pulse. This pulse is then integrated onto a capacitor, which translates the pulse width into an analog voltage. After the time interval has been converted to a voltage, any suitable ADC architecture can be used to finish the time-to-digital conversion [102]. More complex analog approaches for TDCs can use sample-and-hold methods like those used in work by Napolitano et. al. [107]. Although these architectures are straightforward, they suffer from limitations associated with analog circuits, such as poor linearity, limited bandwidth, quantization noise, limited dynamic range, and calibration requirements as discussed by Henzler [102]. Additionally, as technology nodes advance, the analog circuit performance tends to deteriorate.

Naturally, TDC architectures have been developed that use only digital circuits for the conversion process. The simplest all-digital TDC is a synchronous counter, but a counter-based TDC has a timing resolution that is limited by the maximum clock rate of the counter. However, counter-based TDCs that utilize quadrature clocks and multiple counters can achieve a factor of four increase in precision [108]. Finer timing resolution is possible through the use of a tapped delay-line architecture. TDCs based on this approach utilize the quantized delay through a logic gate to precisely measure the time period of interest [109]. In a tapped delay-line, the start signal is launched into a series of delay elements and the stop signal freezes the state of each delay element into a set of flip-flops, resulting in a thermometer code that represents the measured time interval. Figure 4.6 illustrates the principle of operation for a tapped delay-line. The measured time interval is determined by the location of the binary 1-0 boundary between buffer stages.

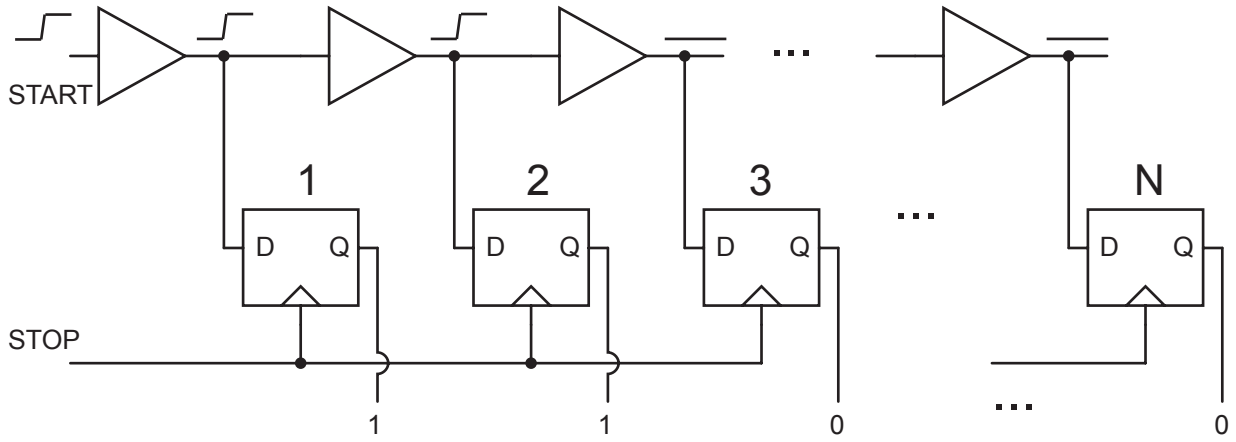

Figure 4.6: An example of a tapped delay-line using buffers. The start signal is launched into the buffer chain and the stop signal latches the output of all buffers to record a thermometer encoded digital value of the time interval. In this example, the start signal only propagated through two of the N total buffer stages corresponding to a time interval of $2 \cdot T_{buf}$.

The best precision possible with this style of TDC is limited by the gate delay of an inverter for the technology that is used for implementation. There are two major limitations with this TDC architecture – the delay line requires calibration to determine the post-fabrication gate delay for the TDC, and the range of the TDC is limited by the number of delay stages used. As the number of stages (and corresponding TDC area) increases, process variations between individual delay stages will combine to increase the timing uncertainty. Consequently, as the number of delay stages increases, the non-linearity of the TDC also increases as a factor of $\sqrt{N}$ [102].

A delay-line based architecture can be modified to provide an extended range without paying a penalty in noise, non-linearity, or variability through the use of a looped architecture. In this approach, a small number of delay elements is used such that the output from the last element in the delay-line is connected to a counter and is fed back to the input of the delay-line. The counter is incremented every time the start pulse completely traverses the delay-line, providing coarse timing information. Although this approach addresses the range and non-linearity issues related to long delay-lines, the resolution is still limited to an

inverter delay and the delay-line must be calibrated with a known timing interval. Through the use of feedback, a looped-delay architecture can be locked to a reference clock, forming a delay-locked loop (DLL). In a DLL the stage delay is calibrated to the reference clock, providing absolute timing measurements.

The approaches discussed thus far have been limited by the gate delay of the technology in which the the TDC is implemented. There are a number of techniques for achieving sub-gate delay precision from TDCs. These approaches include Vernier delay lines, local passive interpolation, and pulse shrinking techniques. The techniques used to achieve sub-gate delay resolution typically require large areas, extra power, and long conversion times [102]. Because the SPADs used in this work were developed in an IBM 0.13 μm process with an FO4 of only 43 ps, sub-gate delay techniques were deemed unnecessary for this application.

The previous techniques highlight the trade-offs associated with TDC architectures. Namely, precision comes at the cost of larger area designs that consume more power and take longer to perform the conversion. The primary criteria for the TDC used in this design were a conversion time less than 50 ns, precision better than 100 ps, low susceptibility to process, voltage, and temperature (PVT) variations, and the ability to easily provide 4096 individual measurement channels. By reusing some or all of the TDC components in multiple pixels, area and power constraints on the TDCs can be relaxed. Other approaches have successfully implemented per-pixel TDCs using compact ring oscillator based converters, such as the work by Richardson et. al. [72].

The TDC selected for this work is based on a DLL architecture with a synchronous counter. This architecture was chosen because of its well defined precision and dynamic range, its fast conversion speed, and the ease with which it could be shared among many pixels in the array. In this design, the DLL and counter outputs are distributed to groups of 128 pixels. A block-level schematic showing the overview of the DLL used for the TDCs is shown in Figure 4.7.

The DLL generates the fine timing information by generating 16 precisely aligned

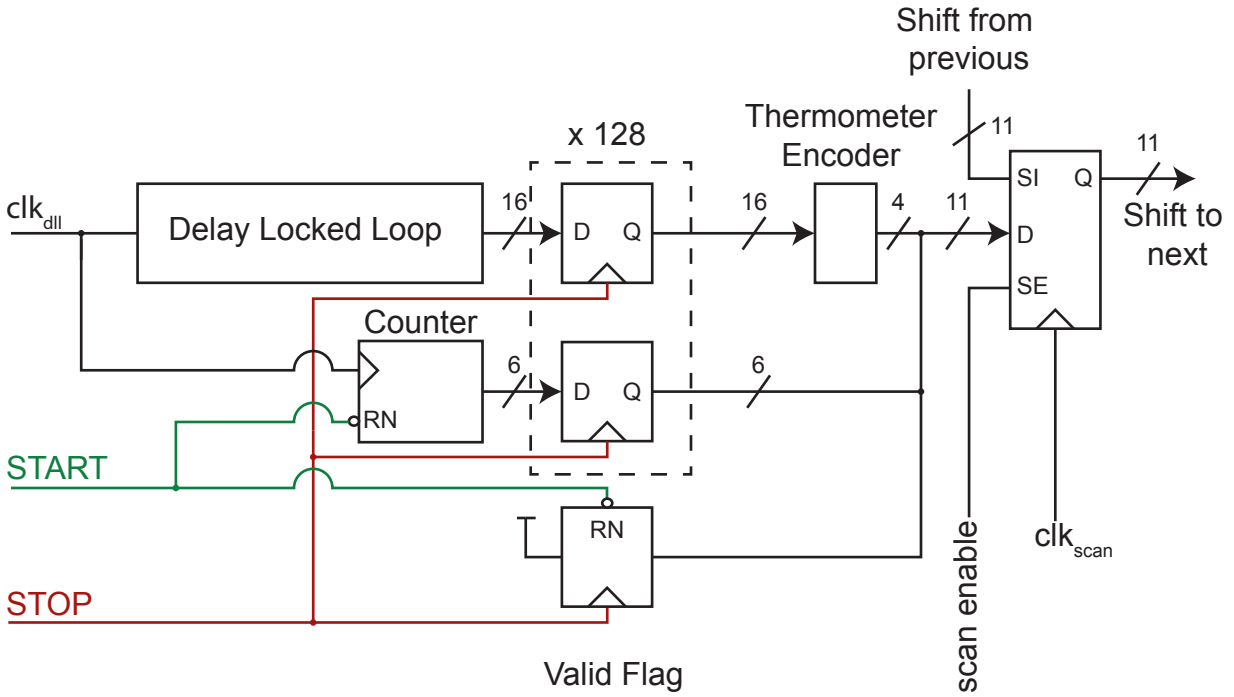

Figure 4.7: A high-level overview of the time-to-digital converter used in this work is presented. A delay-locked loop subdivides the reference clock (clk_{dll}) into 16 evenly spaced phases. clk_{dll} also increments a coarse counter. The phase and counter outputs are buffered to flip-flops to be used in TDCs for groups of 128 pixels. The thermometer encoder converts the 16-bit thermometer code into a 4 bit value, which, along with 6-bits from the counter and a valid data flag, is clocked into a chain of shift flip-flops at the end of each measurement window. clk_{scan} is a gated version of $clk_{datapath}$, which is controlled by a datapath controller and is discussed in detail in Section 4.1.4.

phases from a 1 GHz global clock (62.5ps timing resolution). The DLL is comprised of several functional blocks as seen in Figure 4.8. The main phase-generating mechanism is the voltage controlled delay line (VCDL) in the center of the loop. Two output phases from the VCDL are compared by the phase detector, which then instructs the charge pump to either increase or decrease the reference voltage, V_{ctrl}, for the linear regulator. The linear regulator closes the feedback loop by controlling the supply voltage of the stages in the VCDL, which in turn adjusts the stage delay.

As discussed previously, there are a number of potential sources of error in TDC designs. Considerations for minimizing error in the TDC are presented alongside of the discussion of the relevant sub-blocks that follow. One concern not discussed below is the

potential for timing skew due to trace length mismatch in the physical routing of the output phases to each pixel's flip-flop bank. Additionally, STOP signal skew within each flip-flop bank can result in timing errors. In this design, the metal traces used to connect the sixteen phases to the flip-flop banks were length matched so that each buffer would see a similar capacitive load and have comparable R-C delay. Additionally, a local H-tree was used for the STOP signal that is driven from the pixel's output buffer and simultaneously latches the state of the phases at all of the pixel FFs. This ensures that any timing skew is minimized and that the phases latched by the flip-flops after distribution are accurate representations of the phases at the output of the VCDL.

Voltage Controlled Delay Line

A voltage controlled delay line (VCDL) is a chain of devices whose delay can be adjusted through a control voltage. These devices could be any digital gate but inverters and buffers are the most commonly used stages for VCDLs. The voltage control can be achieved through a number of mechanisms including variable loading, current starving, or supply regulating [110, 111, 112]. An example schematic of the current starving and variable loading techniques is shown in Figure 4.9.

Current starved NFETs and PFETs can be used to control the current available to the delay stage by controlling the gate voltage of the starving device, which will adjust the charging or discharging resistance and change the delay. Using either an NFET or PFET starving device will cause one of the edges in the delay line to slow, which can affect timing performance. For instance, if a current starving NFET is used to adjust the delay, then the pull-down strength of the delay element will become weaker as the control voltage approaches 0 V. This results in a slow falling edge and a timing asymmetry between the $t_{L\rightarrow H}$ and $t_{H\rightarrow L}$ transitions within the pulse. If delay stages are comprised of an inverter with a current starved NFET and a second stage inverter is used to generate a monotonic thermometer code, this delay asymmetry propagated through a long chain of elements could result in

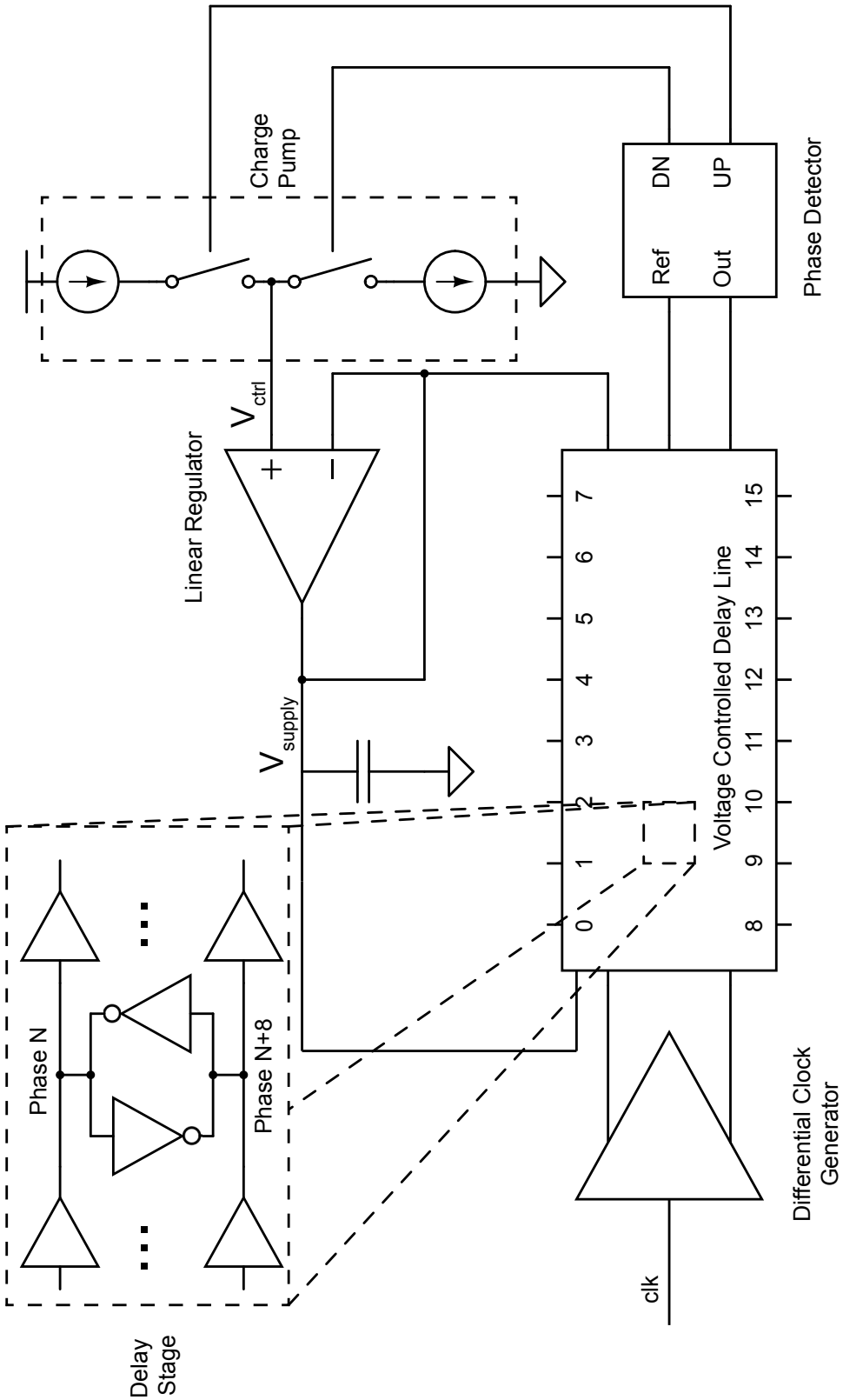

Figure 4.8: Overview of delay-locked loop.

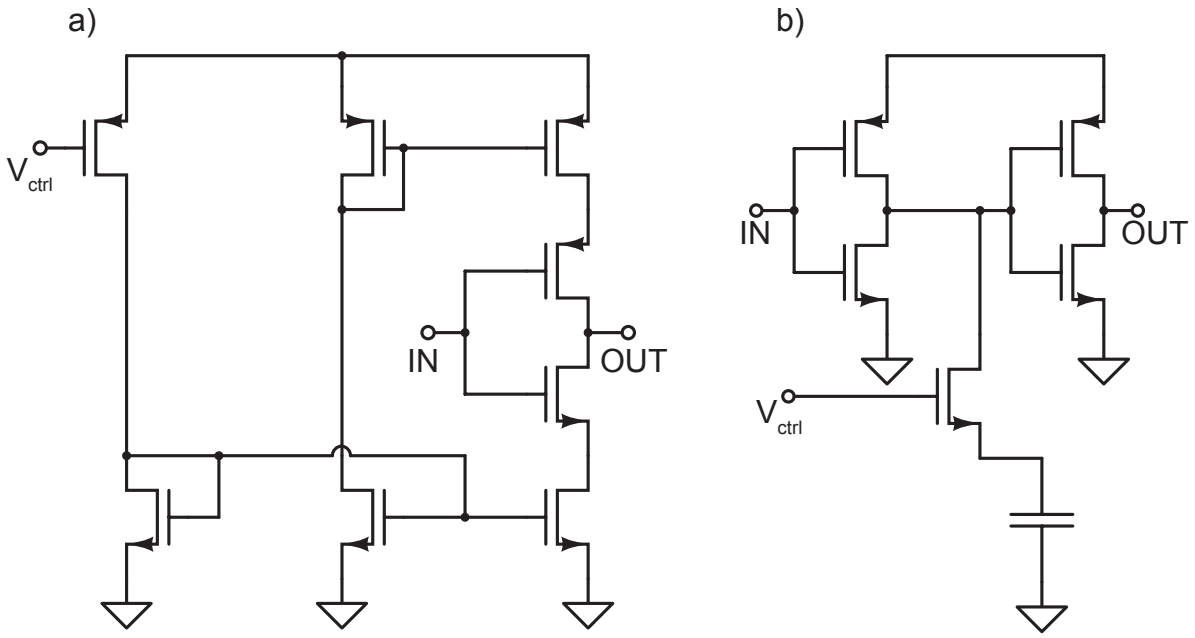

Figure 4.9: a) Schematic of a current starved delay element and b) a schematic of a variable load delay element.

significant duty cycle increase. Conversely for a similar architecture with current starved PFETs, the duty cycle would decrease as the signal propagates. By using both PFETs and NFETs to starve the inverter symmetrically, these duty cycle changes can be reduced. Another drawback to using current starving devices to control the delay is that the tuning range is highly non-linear and often narrow. A simulation of the tuning curve for the delay element in Figure 4.9a that uses both NFET and PFET current starving devices is shown in Figure 4.10a. A significant advantage to using current starving header or footer devices is that the control node is a high-impedance input and will draw little current.

Delay elements based on varying the loading at the output of an inverter can improve the linearity of the tuning curve, but such designs require capacitors that can consume significant area and the designs are much more sensitive to PVT variations. An example tuning curve for a typical variable load stage design is shown in Figure 4.10b. In this simulation, the capacitive load was approximately 10 fF.

In the design of the TDC for this FLIM imaging IC, supply regulated buffers were

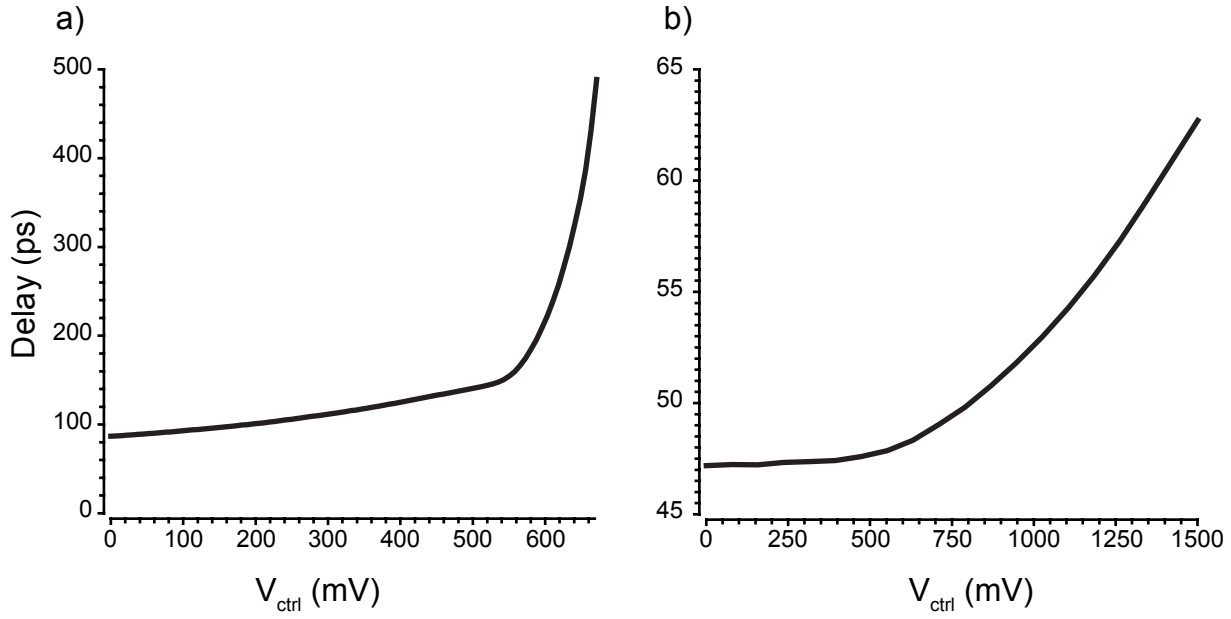

Figure 4.10: Tuning curves for current starved and variable capacitive loading delay elements. a) Current starved, b) variable load.

chosen as the delay element. These buffers are arranged in a differential chain with weakly cross-coupled inverters to form the delay line. Because the inverter propagation delay in this technology is well below the specified 62.5 ps delay, a buffer stage can be used as the fundamental delay element and still meet the timing resolution requirement.

There are a number of advantages to using buffers over inverters for the delay elements. First, a buffer generates more uniform delays than an inverter since each delay will have a pull-up and pull-down component. This results in increased immunity to PVT variations compared to an inverter-based element. Additionally, the thermometer code with an inverter-based design would be a 'pseudothermometer code' of alternating 1's and 0's. In such a code the delay time is determined by the location in the delay line with either two consecutive 1's or two consecutive 0's, and detecting this pattern requires more complex logic for measuring the time interval. This coding scheme also makes inverter-based designs vulnerable to non-linearities resulting from the asymmetric setup times of the flip-flops used to capture the event time.

By using a differential delay chain, the number of delay stages is reduced by a factor of 2 since each stage produces two outputs exactly 180 degrees out of phase. This helps to improve both the integral non-linearity (INL) and the differential non-linearity (DNL), which increase proportionally to the square-root of the number of series delay stages [102]. Additionally, the weakly cross-coupled inverters reduce the impact of power supply noise and local process variations while also aligning the complementary phases. A section of the delay line showing details of consecutive buffer stages with the cross-coupled inverters is shown in Figure 4.11a. A simulated plot of the tunability of one of these buffer delay stages is also shown in Figure 4.11b. The tuning for the supply regulated delay stage is much more linear than the current starved stage and has a wider tuning range than the variable load delay stage shown in Figure 4.10. This schematic level simulation shows a minimum delay of 31 ps at a supply voltage of 1.6 V. In addition to the 8 delay stages that comprise the VCDL core, an extra stage is added to the beginning of the VCDL to condition the complementary input clocks before entering the VCDL and also to the end of the VCDL to ensure uniform loading for all stages. Thus, each VCDL consists of a total of 10 differential buffer stages with cross-coupled inverters, of which 8 stages are used to generate the 16 phases separated by 62.5 ps.

Sizing and layout of the VCDL must be done carefully in order to minimize non-linearities and avoid degrading the minimum achievable delay of the stage with parasitic capacitances. The individual inverters within the buffered delay cell were sized such that the first inverter is slightly smaller than the second. This spreads the required stage gain between the two inverters, taking into account the loads of the next stage, level shifter, and cross-coupled inverters. The result is a maximum individual stage gain of less than 1.4, which allows for individual stage delays that approach the technology limit. The cross-coupled inverters are sized to be one-third of the size of the buffer output, which avoids overloading the buffer output. The NFET and PFET devices in the cross-coupled inverters are skewed to have a β-ratio of 2.5, which aligns the rising and falling $V_{dd}/2$ crossing point

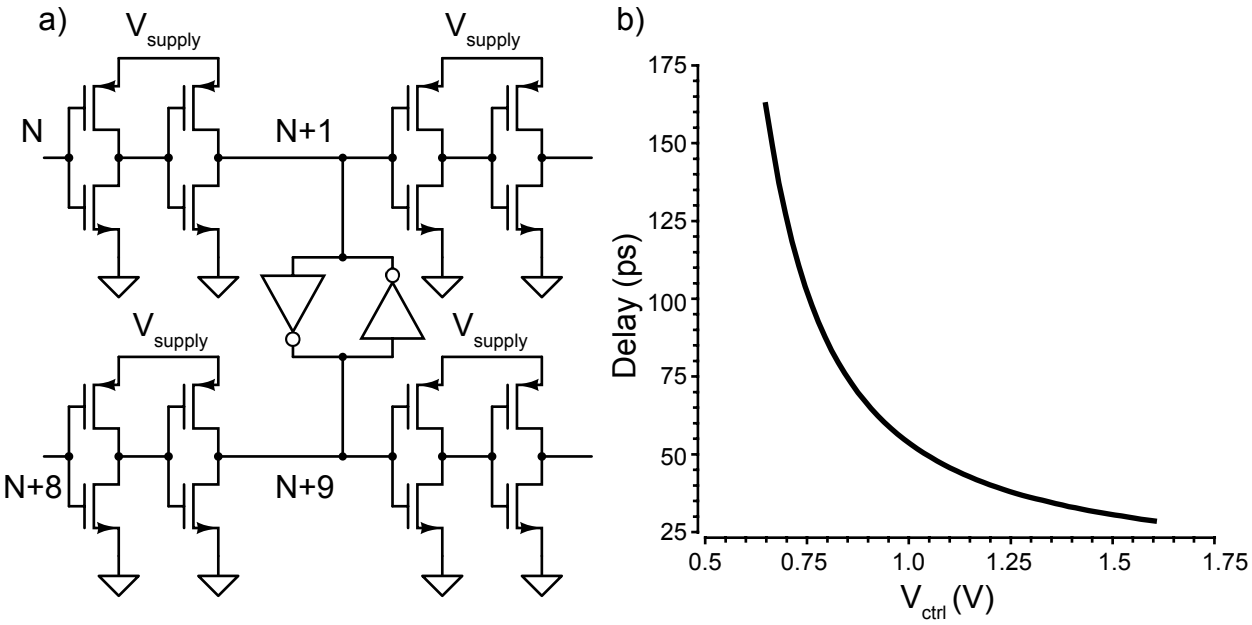

Figure 4.11: a) Schematic of the delay element used in this DLL design. The cross-coupled inverters also use the regulated supply, V_{supply}. b) Simulated tuning curve for one delay stage showing a minimum single-stage delay (TDC resolution) of 31 ps at the maximum allowed core supply voltage of 1.6.

at the typical process corner. Each phase output must pass through a level shifter in order to bring the regulated output signals up to the core supply voltage. The schematic for the level shifter design is shown in Figure 4.12. In this circuit, the NFETs M1 and M2 must be skewed larger than usual so that they can overcome the cross-coupled PFETs during a switching event.

The level shifter drives an output buffer and either one of the phase detector inputs or an equivalently sized dummy load to ensure comparable loading of each stage. Local decoupling capacitors and power and ground traces were placed between the complementary delay stages to provide shielding and minimize the performance degradation due to simultaneously switching nodes in opposite directions. Post layout simulation was performed on an R-C-CC extracted model and the design was capable of a locking to an input clock with a range of 357 MHz to 1.38 GHz, corresponding to an individual stage delay ranging from 45 ps to 175 ps as shown in Figure 4.13.

Special care was taken when designing the input clock that feeds the delay line to

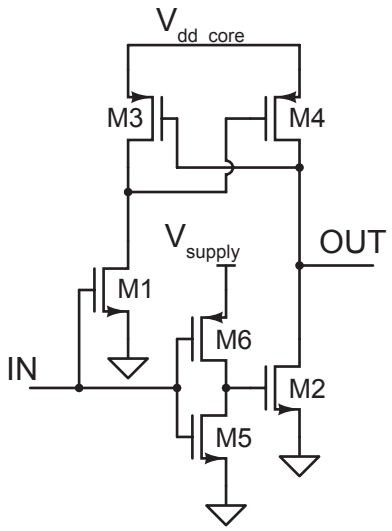

Figure 4.12: Schematic of level shifter used in DLL.

ensure that the complementary clock edges are well-aligned with a 50% duty cycle. A single-ended 1 GHz clock is distributed globally throughout the chip. At the input of each DLL is a simple circuit to generate a complementary differential clock from this single-ended clock. Any variation from a 50% duty cycle on the incoming single-ended clock will result in a phase error between the two halves of the VCDL when the DLL is locked. This phase error will result in an error in the temporal separation between phases 7 and 8 equal to twice the phase error. Additionally, the complementary phases of the clock must be well aligned such that the $V_{dd}/2$ crossing point for both phases are coincident.

A PLL is used to generate the single-ended 1 GHz TDC input clock (clk_{dll}) from the 20 MHz trigger signal of the laser. An overview of the PLL is provided in Section 4.1.6. The 1 GHz clock generated from the PLL has a well controlled 50% duty cycle. This clock is then distributed using clock-specific buffers that have been sized for maintaining this 50% duty cycle across the chip. Additionally, the PLL can be configured to output a 2 GHz clock and each DLL complementary clock generator is preceded by a divide-by-two stage that can be switched into the clock path using a configuration bit in the scan chain. By dividing the clock down to 1 GHz at the input to the DLL, a 50% duty cycle will be ensured. In this design, 2 GHz is an aggressive clock speed and the clock distribution network was not robust

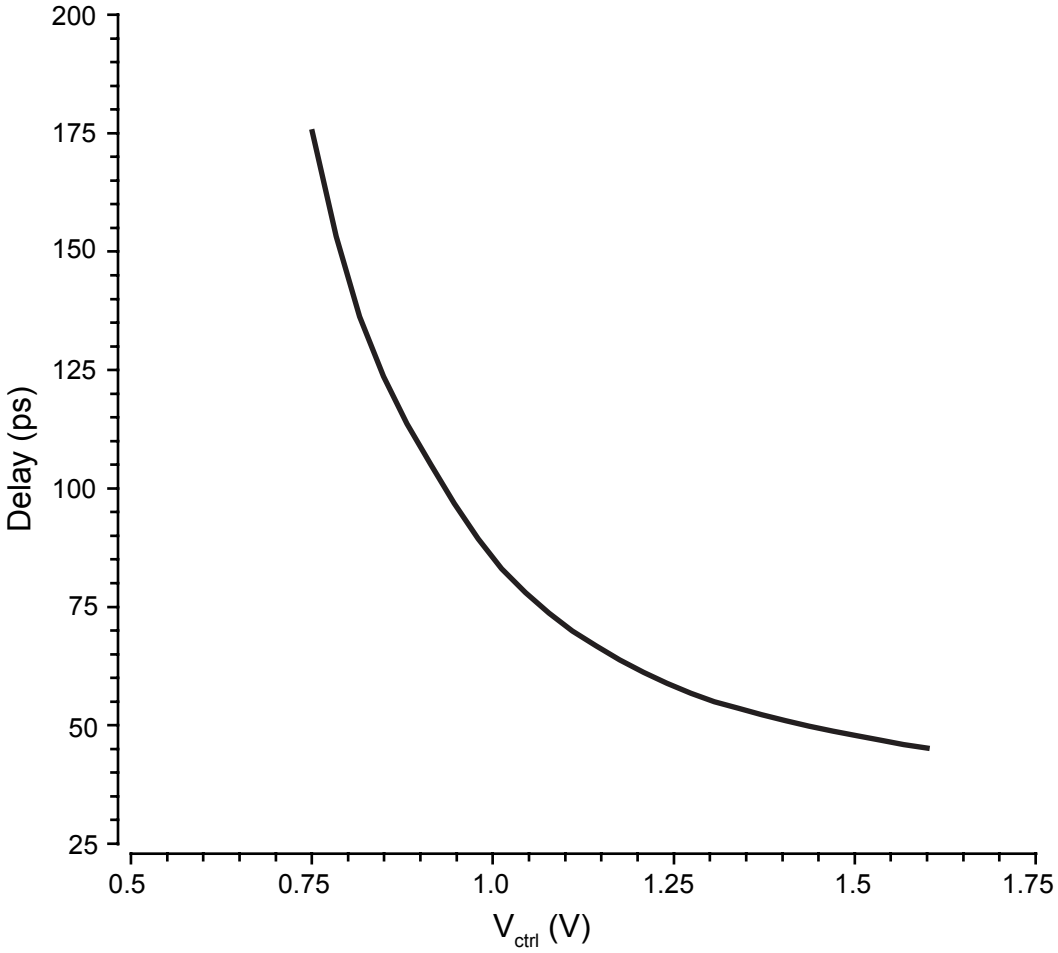

Figure 4.13: Plot of the delay element tuning curve after post-layout extraction and simulation.

enough to deliver a 2 GHz clock across the chip to each DLL input. Consequently, the 1 GHz single-ended clock was distributed throughout the chip to the input of each DLL.

To generate the complementary clocks from the single-ended clocks, a novel pass-gate circuit was designed to closely align the crossing points of the complementary edges. This design is presented in Figure 4.14 and is compact (only 6 transistors), generates well aligned complementary edges, and is relatively tolerant to PVT variations. This complementary clock generator uses transmission gates that are controlled by the opposite phase to gate each clock. On the rising edge of clk_{in}, the falling edge of the inverted clock ($\overline{clk_{in}}$) will be delayed by one inverter delay. The input buffer driving clk_{in} is designed such that its rise

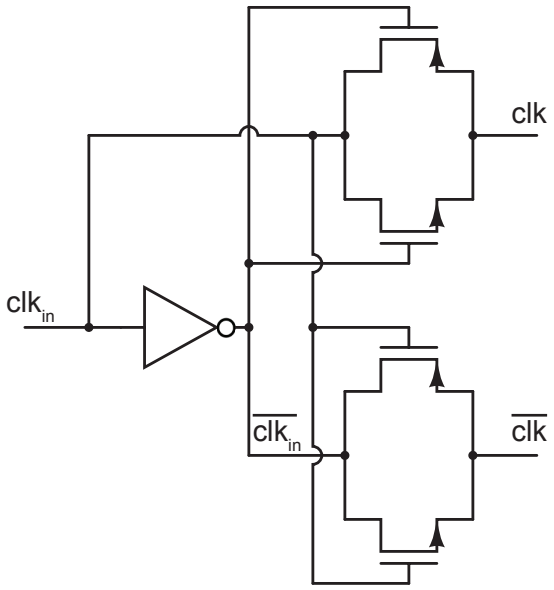

Figure 4.14: A simple complementary clock generator.

and fall times are greater than this inverter delay. Consequently, clk will be in the middle of its low-to-high transition when $\overline{\text{clk}_{\text{in}}}$ reaches the pass gate control inputs for the clk gate output. The resistance of the clk pass-gate increases as the pass-gate is switched during the clk transition, slowing it down. $\overline{\text{clk}_{\text{in}}}$ is able to pass through its transmission gate with a constant low resistance because clk has already switched these transmission gates before $\overline{\text{clk}_{\text{in}}}$ arrives. This results in well aligned transitions for both clk and $\overline{\text{clk}}$ despite the inversion needed to generate $\overline{\text{clk}}$. A simulation showing the typical output from this complementary clock generator for both typical and skewed corners is shown in Figure 4.15.

Phase Detector

A phase detector is used to set the control voltage, V_{ctrl}, of the charge pump by generating up (UP) and down (DN) pulses with relative widths that are proportional to the phase delay between its inputs. For this DLL design, phase 1 and the loading buffer output (following phase 16) are used as the reference and output clocks. Ideally, when the delay line is locked to the reference clock, the average value of these UP and DN signals will be equal. In order to allow for the smallest possible capacitance on V_{ctrl} with a reasonable voltage ripple,

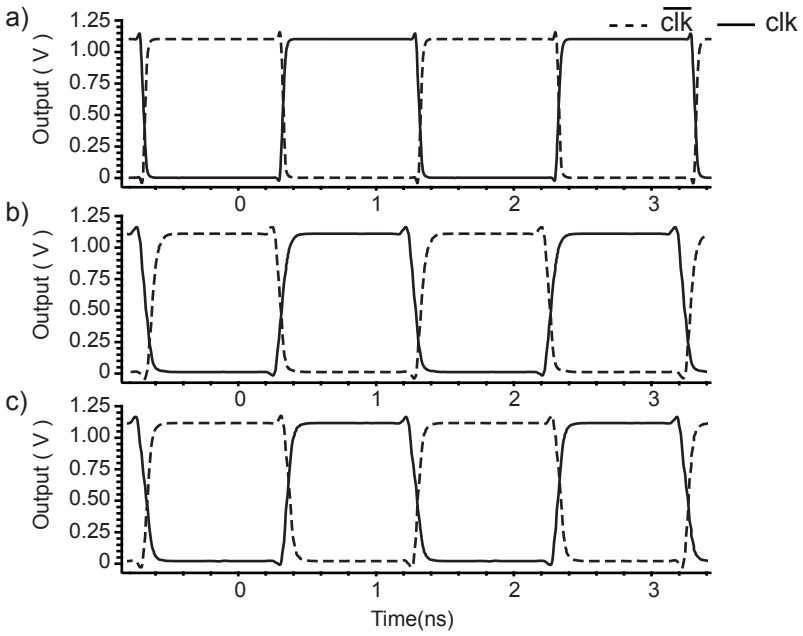

Figure 4.15: Simulated results showing the well aligned edges of clk and $\overline{\text{clk}}$ using the complementary clock generator presented in Figure 4.14. Simulation results showing complementary clock edge alignment for nominal devices (a), skewed fast NFET, slow PFET devices (b), and skewed slow NFET, fast PFET devices (c). The input clock frequency is 1 GHz.

the UP and DN signals should be designed to overlap. There are a number of different approaches used for designing phase detectors, these include NAND based designs [113], S-R latch designs [114], precharged inverters [111], and XOR based designs [115]. In this design, a precharged phase detector similar to the work done in reference [116] is used because of its compact size, simplicity, small capacitive load, and zero offset at zero phase difference. Each input to the phase detector goes through a divide-by-two stage, which reduces the load on the VCDL outputs to only the clock input of a flip-flop. The divide-by-two also serves to reduce the frequency at which the charge pump updates. Additionally, when the input phases are aligned, the UP and DN control signals are of equal width and overlapping. This results in reduced ripple on the V_{ctrl} voltage, which allows for a smaller capacitance on V_{ctrl} and also results in better regulation and phase control within the VCDL. A schematic of the phase detector is presented in Figure 4.16. Simulated results showing the average difference between UP and DN pulses are presented in Figure 4.17a. Additionally, simulated waveforms

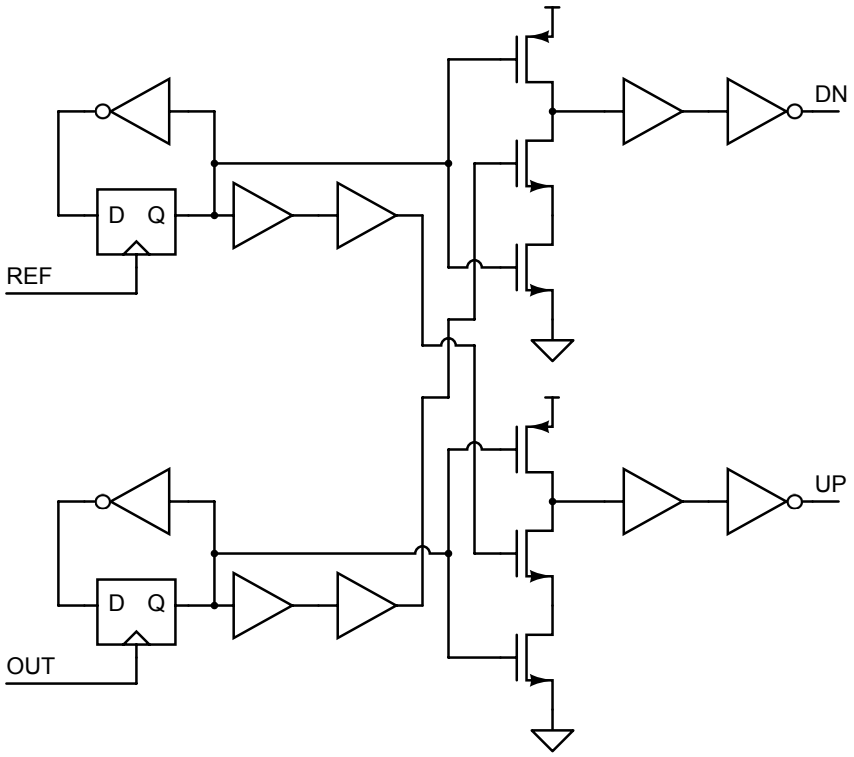

Figure 4.16: Schematic drawing of the precharged phase detector used in the DLL. When the REF and OUT clocks are aligned, the Up and DN pulses output by the phase detector are equal.

showing the up and down pulses when the output clock leads the reference clock by 200 ps and when they are both perfectly aligned are shown in Figure 4.17b.

Charge Pump

Because the phase detector generates equally sized UP and DN pulses when the input phases are aligned, any mismatch in the up and down currents of the charge pump will result in a static phase offset in the loop. This offset is due to unequal amounts of charge being delivered to the V_{ctrl} capacitor for equally sized UP and DN control signals. PVT variations can cause differences in the relative strength of NFETs and PFETs, so a calibrated charge pump was designed to eliminate potential differences between NFET and PFET strength. Additionally, the charge pump calibration control was designed such that the total combined width of the current mirror NFETs that set the average UP and DN currents can be adjusted

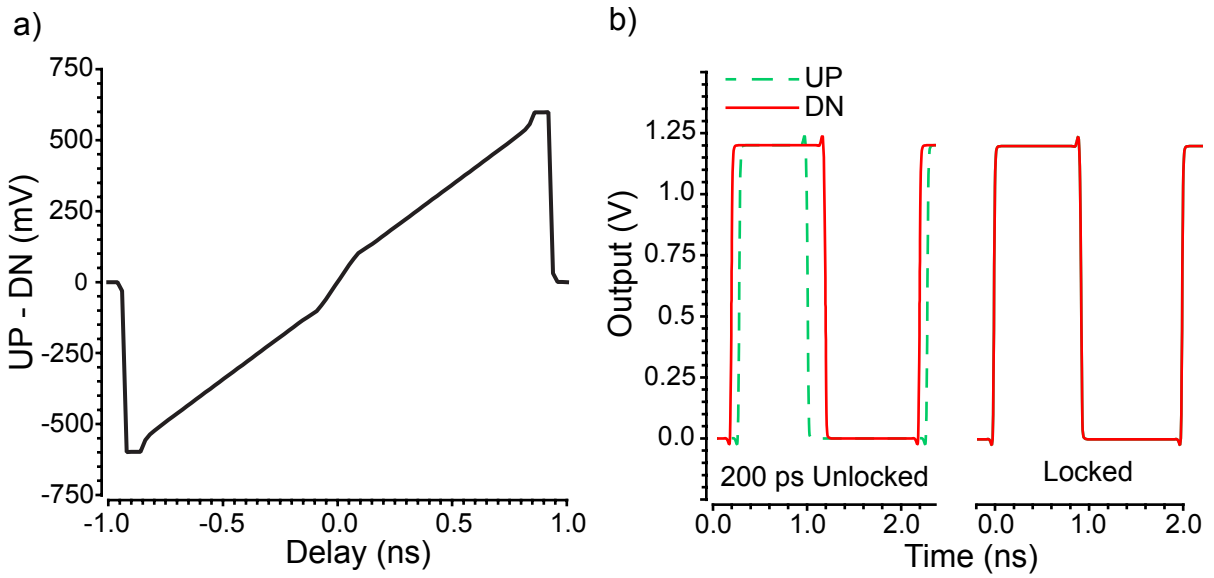

Figure 4.17: Simulated results showing a) the linear response of the phase detector to phase offsets at the input and b) the UP and DN phases during lock condition. In (b), the unlocked condition has the OUT phase 200 ps ahead of the REF phase. The ideal locked condition has both UP and DN outputs perfectly overlapping.

in increments of 10 nm. This corresponds to current steps of approximately 3 nA despite using minimum length devices . A subset of the complete coding scheme is shown in Table 4.3. Binary-weighted calibration codes are a common method for calibrating charge pumps [27, 117], but this limits the trimming precision to the minimum width device of the technology. Consequently, long channel length devices are necessary to make fine adjustments to the UP/DN currents. With the calibration coding scheme chosen here, precise current matching is achieved to compensate for slight differences in drive strength due to PVT variations using minimum device length transistors. A schematic of the charge pump is presented in Figure 4.18.

Ideally, the charge pump should consist of two switched current sources, as diagrammed in Figure 4.8. As such, the topology chosen for the charge pump was a switched, low headroom, self-biased, cascoded current mirror. This architecture provides a high output resistance and closely matches an ideal current source. Both the UP and DN switches were implemented with NFETs to minimize variability due to mismatch in PFET and NFET

Table 4.3: Subset of calibration codes and combinations demonstrating the width tuning capabilities of the calibrated charge pump. The codes continuously increase in increments of 10 nm from 0 to 630 nm differences. Additional width differentials beyond 630 nm can be generated from these codes with the maximum calibration difference at 2470 nm. The bits in the code correspond to the following device widths in order from MSB to LSB: 520 nm, 440 nm, 400 nm, 380 nm, 370 nm, 360 nm. The minimum 2.7 V device width is 360 nm.

Code 1	Code 2	$W_{total,1}$ (nm)	$W_{total,2}$ (nm)	Tuning Diff. (nm)	Current Diff. (nA)
000000	000000	0	0	0	0
000001	000010	360	370	10	3
000001	000100	360	380	20	6
000010	001000	370	400	30	9
000001	001000	360	400	40	12
000011	001100	730	780	50	15
000100	010000	380	440	60	19
...	...	...	...	...	...
000101	111000	740	1360	620	156
000011	111000	730	1360	630	159

fabrication. The calibration control is implemented with six additional NFET devices in parallel with each of the switches (M24 - M35). This allows for an adjustment of the average current for fine tuning. Replica biasing devices (M39 - M50) were also used to ensure that the UP and DN adjustment currents are well matched. Finally, the charge pump uses the 2.5 V I/O devices in this technology to allow for a supply up to 2.7 V. This ensures that the charge pump voltage can span the entire operating supply range of the VCDL, 0.75 V to 1.6 V, despite the headroom consumed by the cascode devices. An off-chip reference current is used to provide the bias voltage, V_{bias}, for the current mirrors.

Linear Regulator

The charge pump output sets the control voltage for the linear regulator, which must then regulate the supply of the VCDL and provide sufficient current for operation at 1 GHz. The linear regulator used in this design is a two-stage op-amp configured in a unity-gain arrangement as shown in Figure 4.19. The output stage of the op-amp is a cascoded current

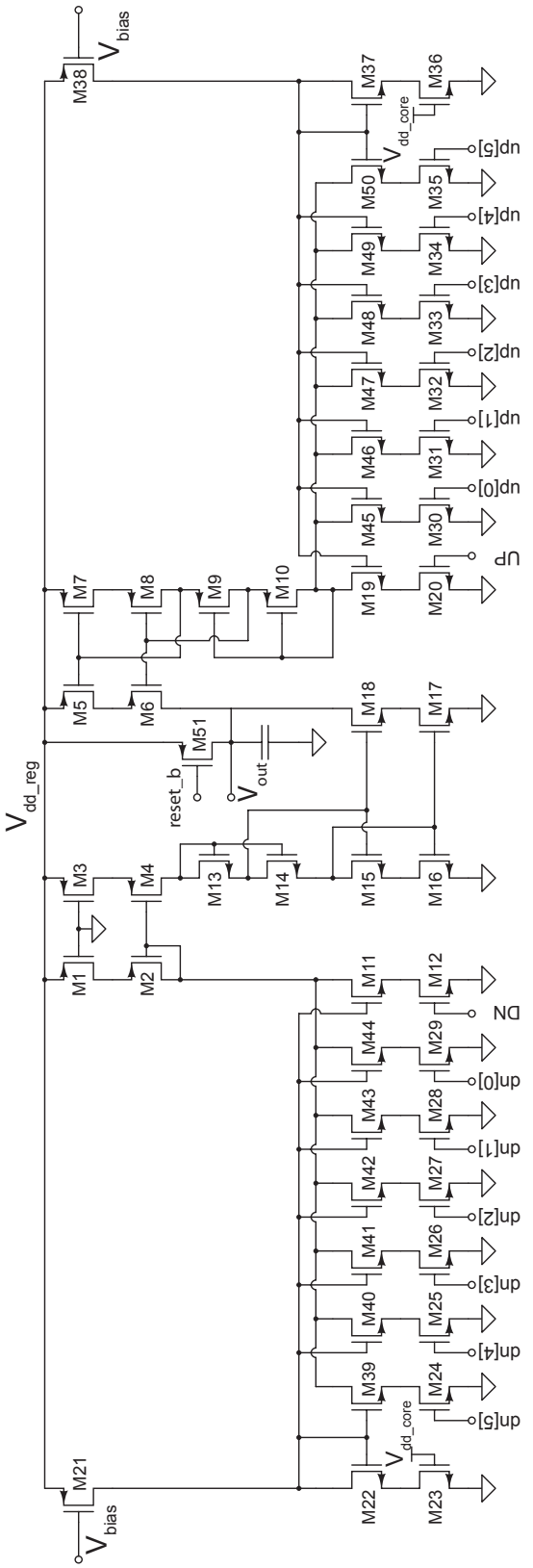

Figure 4.18: Calibrated charge pump schematic.

mirror, which provides additional open-loop gain and improves power supply noise rejection. The expected load of the VCDL operating at 1 GHz is simulated to be 1.7 mA. The regulator is designed to source up to 6 mA and uses the same 2.5 V I/O devices and 2.7 V supply as the charge pump. By consuming additional headroom in the design, the cascoded current mirror also helps to restrict the output voltage level of the regulator to no more than 1.6 V, which is the maximum rated supply for the FETs used in the VCDL. The bias current for the differential input pair is generated from a current mirror biased by V_{ctrl}; this regulator design is similar to the design in [111]. A simulation showing the voltage tracking of the regulator and the output limiting behavior is shown in Figure 4.20. Additionally, the open loop gain of the op-amp is 18 dB, the phase margin is 88° and the power supply rejection ratio is -44 dB. The amplifier is output compensated at the output by a 16.7 pF local decoupling capacitor on the regulated supply. This capacitance is implemented as a n-type MOS capacitor.

Phase Distribution

Each phase from the DLL must be distributed to 128 flip-flop banks to provide timing information for the TDCs. This distribution is performed using a set of 16 phase buffers for each DLL. Each phase buffer is driven by the output from the level shifter and must drive a load of roughly 1.2 pF that consists of gate and interconnect capacitance. Four inverters are used with a stage gain of 4.2, which provides sufficient drive strength while adding minimal jitter to the phase outputs.

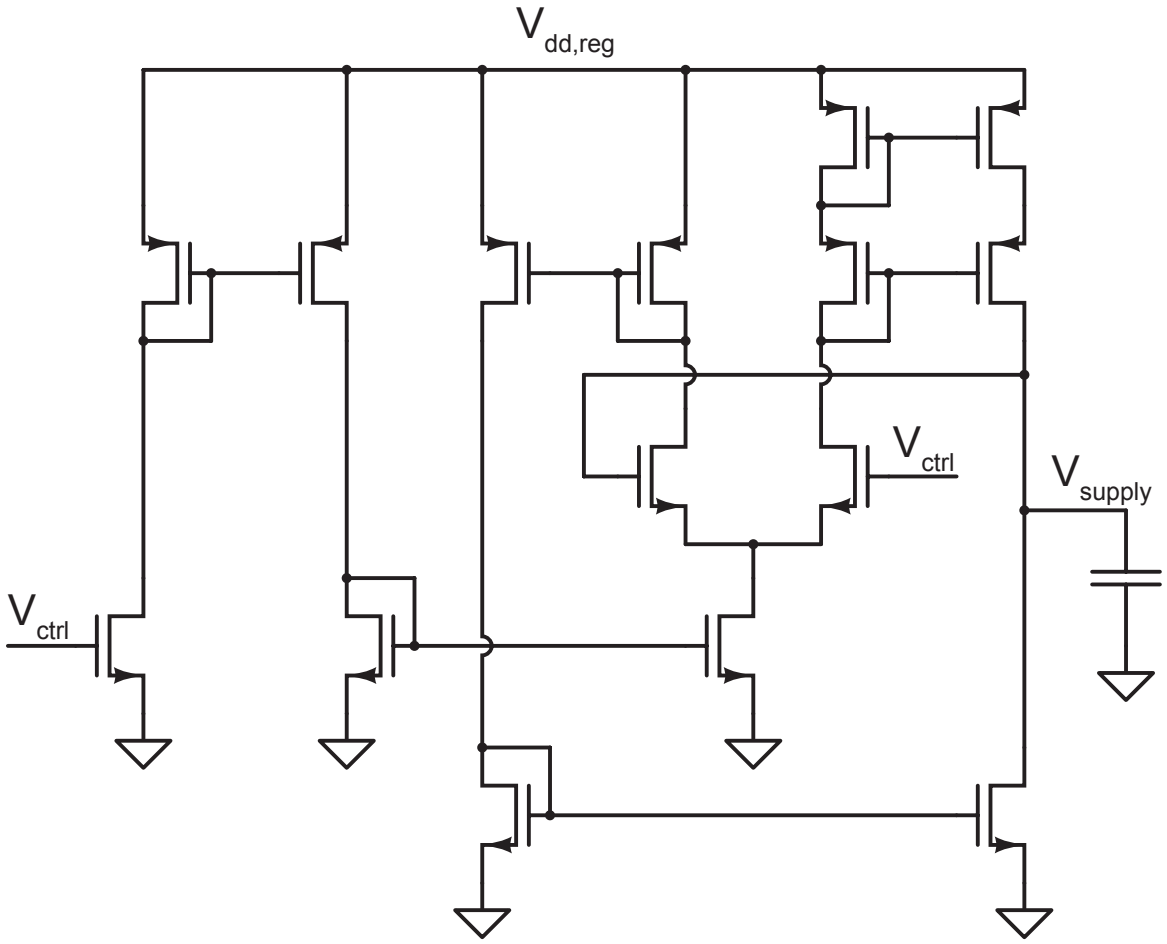

Figure 4.19: Schematic of linear regulator used in each of the 32 DLLs.

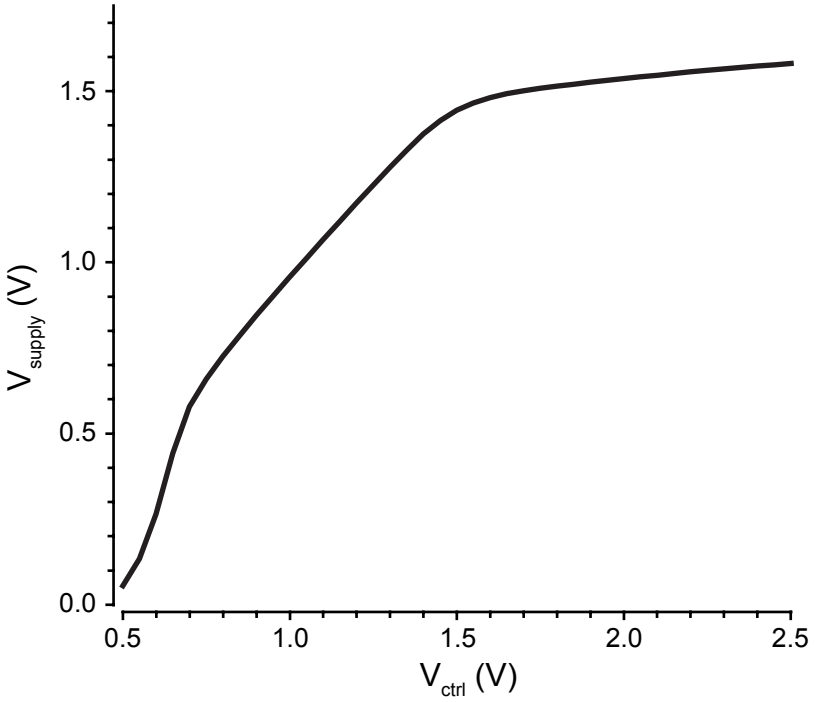

Figure 4.20: Simulation results showing the output tracking of the regulator circuit with a 400 Ω load. Between 0.7 V and 1.5 V, which is the primary operating range of the VCDL, the regulator tracks the control voltage well.

4.1.4 FLIM Datapath

Perhaps the most important consideration in this IC design was how to manage the vast amount of data generated during TCSPC-based FLIM. For each laser repetition, every pixel in the array could possibly record a photon arrival. With an array of 4096 pixels where each requires 12 bits of position information and 10-bit timing precision, while using a laser repetition rate of 20 MHz, this array would generate 1.8 Tb of data per second. However, as discussed in Section 2.2.2, TCSPC FLIM places a constraint on each pixel such that the hit rate should not exceed 1-2% of the laser repetition rate per pixel in order to avoid pulse-pileup. As a result, of the 1.8 Tbps of data, only 18-36 Gbps of meaningful data (i.e. photon hits) will be generated. An efficient event-driven data management technique is used to transfer the data from the TDC flip-flops to the output buffers at the periphery of the chip while discarding all data that does not contain a valid photon arrival time. This datapath is designed to leverage the statistical nature of the photon arrival times during TCSPC FLIM such that the fewest number of data storage elements are used for an average overflow failure rate lower than 1 in 1,000,000,000 laser repetitions. This limits data overflow failures to no more than once every five seconds when operating with a $f_{\text{laser,trigger}}$ of 20 MHz. An overview of one of the 16 identical sections of this datapath is shown in Figure 4.21.

In Figure 4.21, 8 rows of 32 pixels (8 half rows of the array) are read into the datapath in parallel from the left. The valid data bit, which indicates a photon was detected at the pixel and is shown with the TDC structure in Figure 4.7, is used by the datapath to separate the photon arrival events from the data without a photon arrival. This data separation occurs fluidly as the data progresses from left to right through the datapath. The total data input into this 8-row section of the datapath block is 81 Gbps (5 bits position, 10 bits timing, 1 bit valid, 256 pixels, and a laser repetition rate of 20 MHz) and the data rate out of the datapath is less than 8.5 Gbps, depending on the number of photons detected following an excitation pulse from the laser.

TDC data enters the datapath through the scan flip-flops (SFFs) shown in Figure

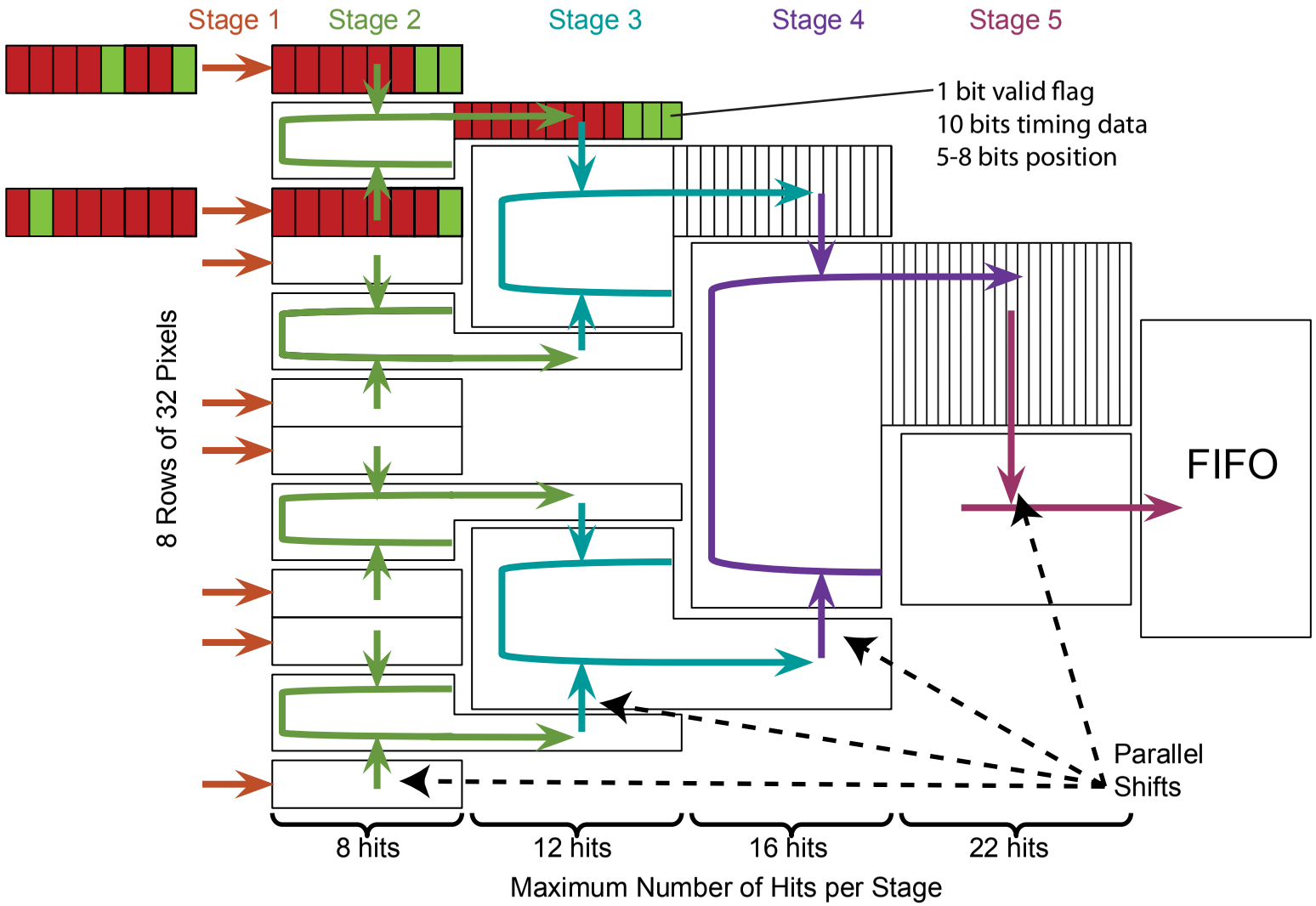

Figure 4.21: Diagram showing the movement of data through the datapath. Each of these blocks is repeated sixteen times on the imager chip. An example showing how data is compressed in each stage is shown for the top two rows. Incoming valid data (green) initially has non-event data (red) between it. Each incoming data consists of 10-bits of timing information, 5-bits position information, and one-bit valid flag. As data enters the datapath, the non-event data is discarded, resulting in the two valid data packets in the top row finishing in the rightmost flip-flops. In the second stage, the two rows of data are shifted in parallel into the the U-shaped shift chain and then shifted clockwise with the non-event data being discarded once again. In this figure, all horizontal arrows represent serial data shifts while vertical arrows indicate a parallel data shift.

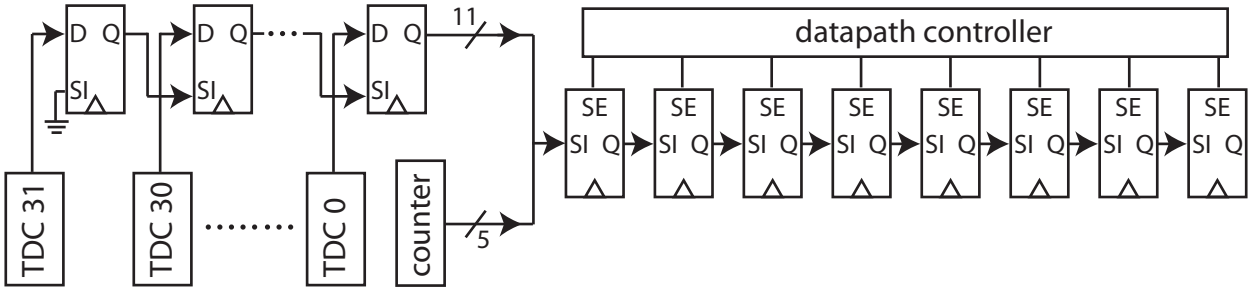

Figure 4.22: The details of the first datapath stage. A counter keeps track of the pixel from which the photon data originated. The timing data and valid flag comprise the 11 bits that are combined with the 5 position bits from the counter. The datapath controller uses the valid bit to lock valid photon arrival time data in the flip-flops from right to left.

4.22, which are the same SFFs that are shown in 4.7. These SFFs are D-type flip-flops with a multiplexer at the input that selects the scan input when scan enable (SE) is asserted. As data is shifted into the datapath, a position counter increments every clock cycle and its value is used to record the position in the row from where the data originated. This approach reduces the number of flip-flops needed for each pixel in the TDC, as location information does not initially have to be stored with the arrival time.

The number of data storage elements for each stage in the datapath was chosen such that the probability of more photons arriving than can be stored in each stage is less than 1 in 1,000,000,000 for a 1% hit rate. The equation that describes the probability of N photon events occurring in P pixels with an average event rate of μ, is given by 4.1.

$$P\,(\ N \text{ hits out of } P \text{ pixels}) = \mu^N\,(1-\mu)^{P-N}\left(\frac{P!}{N!\,(P-N)!}\right) \tag{4.1}$$

For TCSPC measurements, μ is the average photon arrival rate and should be approximately 1%. Equation 4.1 will be used in the detailed discussion of the datapath that follows.

Datapath Flow

The following is a detailed description of how the data is transported through the datapath for the right two quadrants of the pixel array. For the left two quadrants, this same operations

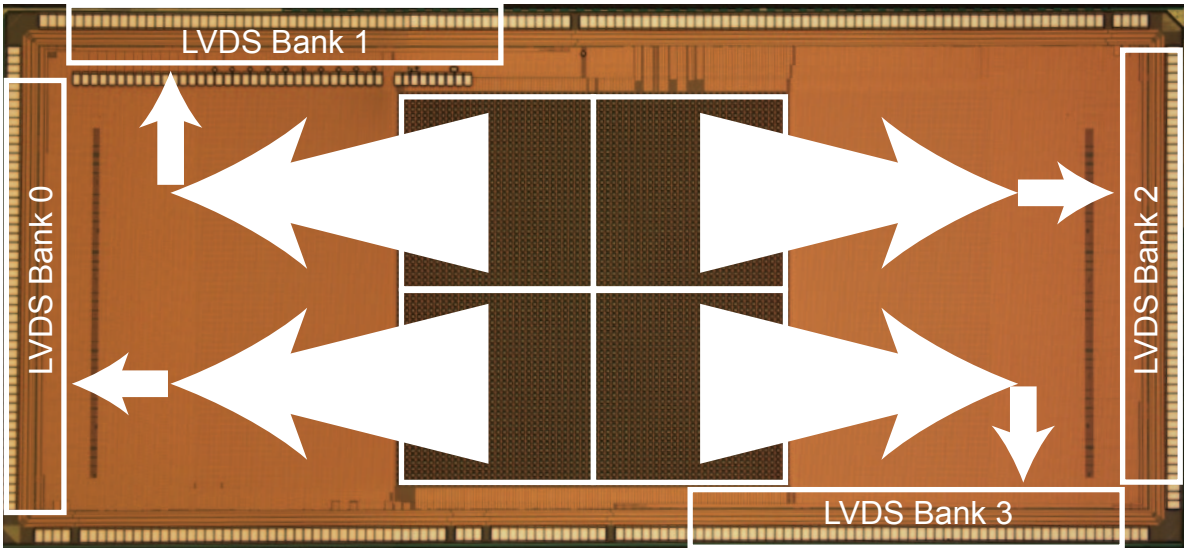

Figure 4.23: A high level overview of the data moving from the pixel array to the periphery. The tapered arrows indicate data compression.

occur but follow a right to left progression. A top-level view of the flow of data from the pixel array to the output buffers is shown in Figure 4.23. Each stage of the datapath is separated by a parallel shift operation, which allows the datapath to be segmented such that the required operations can be performed within $f_{\text{clk,datapath}}/f_{\text{laser,trigger}}$ cycles. In Figure 4.21, the vertical arrows represent these parallel shift operations while the horizontal arrows are the serial data compression operations. Data is continuously processed by the datapath such that new measurements can be recorded with every laser pulse. The clock frequencies used in designing this datapath are a 1 GHz datapath clock ($f_{\text{clk,datapath}}$) and a 20 MHz laser trigger signal ($f_{\text{laser,trigger}}$). This provides 50 cycles between laser pulses to move the data from one stage to the next.

As the data moves into the first stage of the datapath (Figure 4.24), a thermometer controller recognizes event data from the valid bit and de-asserts the scan enable control signals in order from right to left as valid data reaches the rightmost available SFFs. At the end of the first stage cycle, all of the valid data for the half row will be packed in the rightmost storage elements with no empty pixel data between them. Example data patterns are shown in Figure 4.21. The process of moving data out of the TDC flip-flops and packing

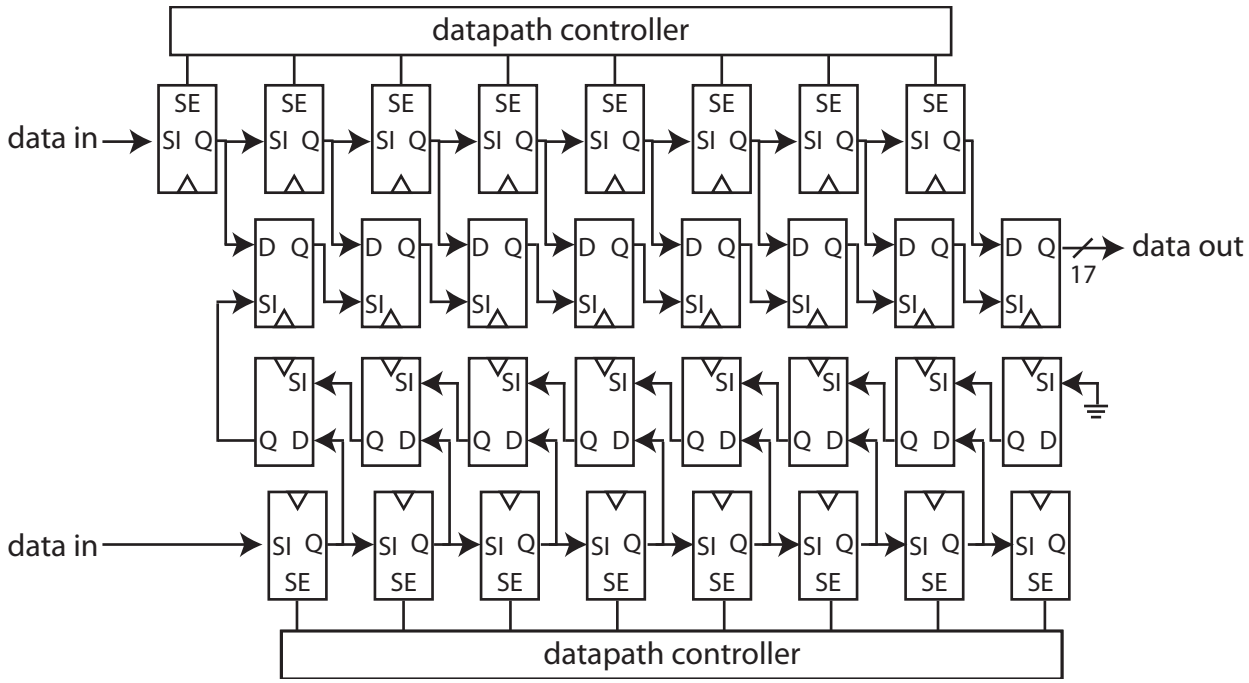

Figure 4.24: The first stage of the datapath captures data shifted out of two rows of the TDCs and collects all valid pixel events. The datapath controller checks the valid bit of the incoming data and organizes the data into the rightmost available flip-flops. After the data input shift is complete, the next laser pulse triggers a parallel shift operation into the central flip-flops that are connected in a U-shape. During the next measurement window, the data in these flip-flops will be scanned into the next stage of the datapath. The number of data bits increases by one to 17-bits, with the 17th bit representing the input row of the data.

it into the first stage takes a minimum of 32 clock cycles. This first stage consists of SFFs for holding 8 events per row, resulting in a stage one overflow probability of 1 in 44,000,000,000 laser repetitions, based on equation 4.1.

After the data has been packed into the right side of stage one, a parallel shift operation occurs where the output data from two adjacent rows are combined and then serially shifted to the second datapath stage. Valid data from these adjacent rows are compressed together in the same manner as in stage one. When the data undergoes the parallel shift operation, an additional bit is added to the data word to indicate whether it came from the top or bottom row of those that were combined. The top row is indicated by a 0 and the bottom by a 1. This additional bit brings the total word length to 17 bits at the input to the second stage.

The second stage of the datapath can store up to 12 words of data. The 12 available storage locations result in a second stage overflow failure probability of 1 in 22,000,000,000 laser repetitions. The scanned data is locked in place by a thermometer controller similar to that of stage one, which will stack all of the valid data at the rightmost positions of these 12 SFFs. The shift process in stage two requires a minimum of 16 clock cycles to complete, and could require up to 28 cycles in the worst case (worst case is when the bottom row has only one pixel, requiring 8 cycles to move through the bottom storage elements, 8 cycles to move through the top, and 12 cycles to reach the rightmost storage element at the end of stage two).

The third and fourth datapath stages operate similarly to the second stage and each stage adds an additional bit to the data word – increasing the data word length to a total of 19 bits by the end of the fourth stage. In each stage, a parallel shift operation first moves the data into a set of SFFs for the stage and then a controller is used to lock valid data in the rightmost storage elements. These stages are sized to handle 16 and 22 photon arrivals, limiting the overflow probabilities to 1 in 10,000,000,000 and 1 in 5,000,000,000 laser repetitions, respectively. Stages three and four require a minimum of 24 cycles and 32 cycles and in the worst case scenario these stages could require 40 and 54 cycles, respectively. Although the worst case scenario for stage 4 exceeds the number of cycles available, it is permissible as the only consequence will be that there could be up to 4 empty spaces separating the valid data from the rightmost storage element. This does not cause a problem because those empty spaces will not have the valid flag asserted and will be ignored by subsequent stages.

In stage five of the datapath, a parallel shift operation moves the 22 words stored at the end of stage four into a new set of SFFs. These SFFs are primarily used to relax timing requirements on the input to the FIFO. The 22 data words in this final set of SFFs are shifted at half of the datapath frequency into the FIFO (500 MHz using the parameters described above), which requires a total of 44 cycles to complete.

The FIFO stage is 32 words deep, which provides some buffering room in the event that the output controller cannot read all of the data contained in the FIFO between the trigger signals. Four FIFOs are grouped together and controlled by a single FIFO coordinator. The FIFO coordinator cycles through each of the four FIFOs to check for valid data in a round-robin fashion. If valid data is found in one of the FIFOs, the coordinator reads from this FIFO, adds two additional position bits to the data word (total length of 21 bits), and then sends the data to the output buffers for transmission off chip. Each FIFO coordinator block controls the data output for one quadrant (1024 pixels) of the array. On average, this block will handle 11 valid data words per laser trigger. The datapath has been designed such that local bursts of data can be handled without missing events. As such, the FIFO coordinator can process up to 25 words per trigger period, and the four FIFOs can store up to 88 words per trigger period. At the described frequencies, each datapath can support output data rates of up to 10.5 Gbps.

4.1.5 Output Buffers

Each of the FIFO coordinator controllers outputs data to a bank of low voltage differential signaling (LVDS) output drivers. Each bank of LVDS drivers is 21 bits wide and is designed to operate at 500 MHz. There are a total of four LVDS banks on the IC, which results in a total output data bandwidth of 42 Gbps.

LVDS is a current-mode signaling standard and was chosen for this application because of its many advantages over common voltage-mode signaling standards, like HSTL and SSTL. Among these are reduced AC supply noise, sharper transient response, and lower power consumption [112, 118]. These properties are crucial, because 84 output buffers are necessary to meet the data bandwidth requirements of the IC. A schematic of the LVDS buffers used is shown in Figure 4.25 and are designed to be compliant with the TIA/EIA 644-A standard [119]. The LVDS buffer is designed to drive a 100 Ω differential load over a 100 Ω differential PCB trace at up to 1 GHz.

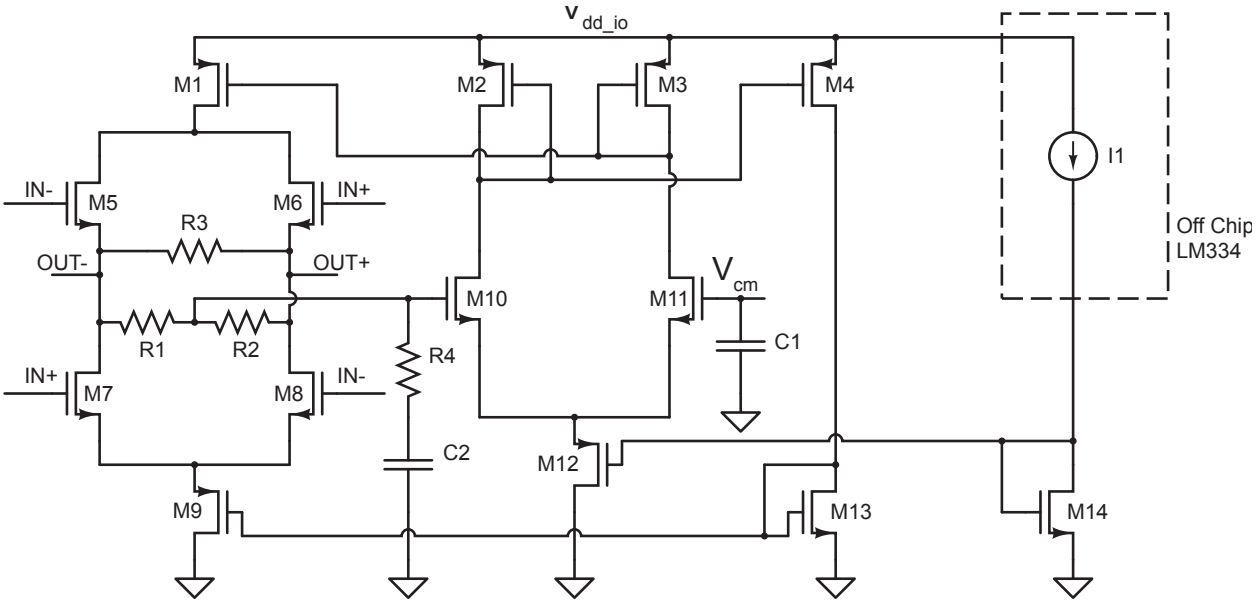

Figure 4.25: Schematic diagram of the LVDS output buffer used. Four banks of 22 of these LVDS buffers (21 data bits and 1 clock signal per bank) are used to drive the TCSPC data off chip.

In Figure 4.25, the main current steering devices are the NFETs M5-M8. Resistor R3 is a source termination resistor that is nominally sized at 198 Ω to reduce reflections. A trade-off between signal integrity, area, and power was made in choosing not to exactly match the 100 Ω differential trace impedance. The common mode feedback sense is performed by resistors R1 and R2, which have a designed value of 131 kΩ. Common-mode control is performed by the common-mode amplifier, which compares the sensed common-mode voltage (between R1 and R2) with the reference input, V_{cm}. The amplifier adjusts the common-mode by controlling the header/footer current mirrors M1 and M9, respectively. This feedback loop is stabilized by R4 and C2, which combine to provide a zero in the frequency response. The simulated common-mode gain is 24 dB with a phase margin of 66°. Capacitor C1 is used to decouple the input common-mode voltage, which is set off-chip. The reference current, I1, is also set off-chip and should be equal to 10 μA per buffer.

These LVDS buffers are meant to drive the inputs of Xilinx Virtex-6 FPGAs, which will be discussed in detail in Section 5.3. Simulations using IBIS models for the Virtex-6 inputs and a transmission line model for the PCB traces were performed. The LVDS buffer

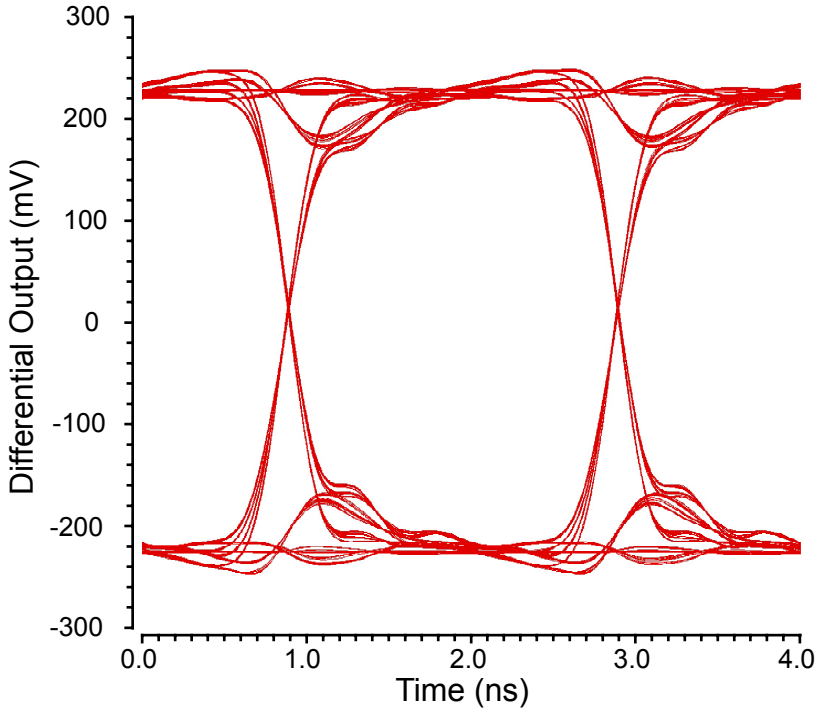

Figure 4.26: Simulation results showing the data eye for the designed LVDS buffer.

was designed with a target differential voltage swing of 500 mV. In simulation with a random bit stream, the LVDS buffer had an eye width of 1.921 ns and height of 384 mV at a 500 MHz frequency, which is shown in Figure 4.26. Analog layout techniques for addressing polysilicon gradients and implant shadowing [120] were used to ensure close matching between all current mirrors and differential switches.

4.1.6 Phase-Locked Loop

The phase-locked loop (PLL) in this design is used for synchronizing the on-chip clocks to the trigger signal from the laser. The PLL uses a charge pump architecture based on the design in [121]. A block-level diagram of the PLL is presented in Figure 4.27. The divide-by-two on the reference clock input is optional and allows the PLL to generate either a 1 GHz or 2 GHz output from the same 20 MHz input. The voltage controlled oscillator (VCO) is a current starved inverter ring oscillator design. The output from this VCO passes through a level shifter and then to three divide-by-two stages. One is a programmable divide-by-two in the

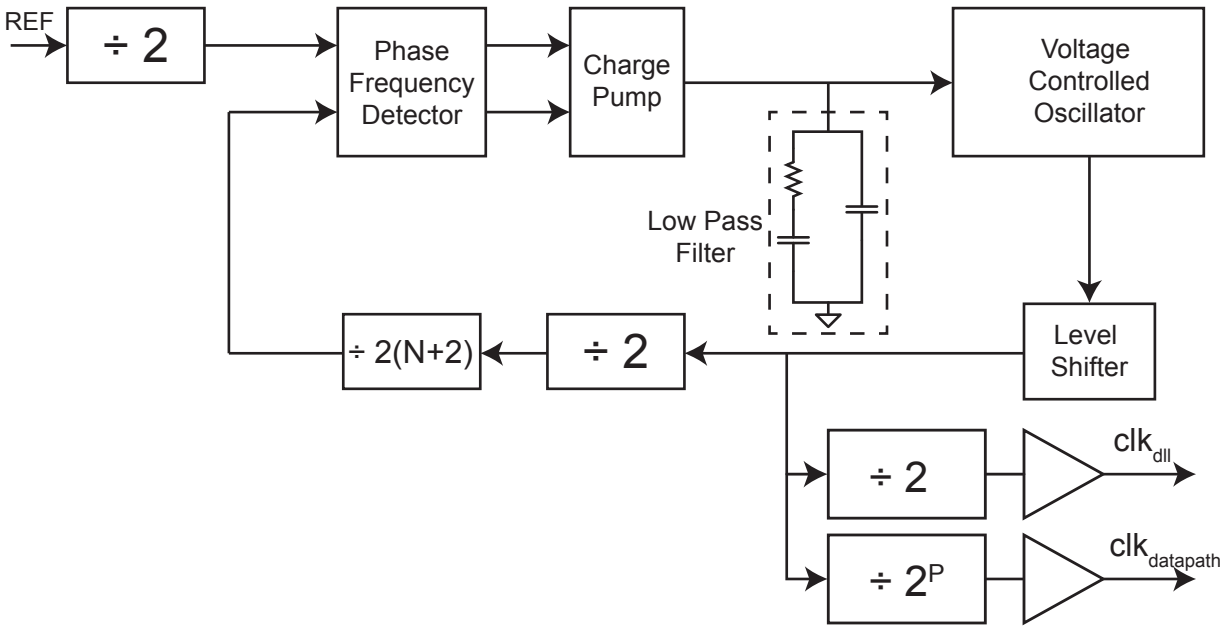

Figure 4.27: Overview of PLL.

clk_{dll} output path, another is a divide-by 2^P, in the $\text{clk}_{\text{datapath}}$ output path, and the last is a fixed divide-by-two that feeds a divide-by $2\,(N+2)$ in the feedback path. The divide-by-two in the clk_{dll} path can be used to produce a 1 GHz TDC clock when the PLL is locked at 2 GHz, as discussed in the complementary clock generation portion of Section 4.1.3. This will produce a clock output with a 50% duty cycle that is distributed to all of the DLLs. The $\text{clk}_{\text{datapath}}$ division parameter, P, can be set to 0, 1, or 2 allowing for a datapath clock of $f_{\text{clk,dll}}$, $f_{\text{clk,dll}}/2$, or $f_{\text{clk,dll}}/4$. The feedback division parameter, N, can range from 0 to 31 and the VCO is designed to operate between 700 MHz and 2 GHz. All of the division factors for the PLL are configured using a PLL-specific scan chain. The PLL was designed by Simeon Realov specifically for this IC.

4.1.7 Imager Control

Throughout this imaging IC, there are a number of controllers used to control the pixels and coordinate the movement of data through the chip. This section provides an overview of these controllers and describes their role in the circuits discussed above.

A global scan chain 8,635 bits long allows for configuration of the measurement window, control settings for each pixel (described in Figure 4.4), and charge pump calibration codes shown in Figure 4.18. The scan chain is designed such that the data and clock propagate linearly in opposite directions throughout the chip, eliminating the possibility of short path errors due to clock skew.

Each of the controllers was designed in Verilog HDL and generated using commercial synthesis and place & route tools. Except for the thermometer controllers in each datapath stage and the FIFO controller used in each datapath block, each controller is instantiated twice on the chip. Each instance controls one half of the array (right or left) and the controller output signals are buffered to all associated pixels or datapath blocks.

Pixel Array Controller

Each pixel in the array is carefully controlled to optimize performance for FLIM applications. This controller is used to manage the pixel_off_ctrl and $\overline{\text{reset}}$ signals described in Figure 4.4. The pixel controller is configured through the scan chain with a programmable number of cycles to wait before sending the pixel_off_ctrl signal to all of the pixels. Additionally, this controller receives the trigger signal from the laser and buffers an inverted version to the $\overline{\text{reset}}$ input of each pixel. A timing diagram showing the behavior of this controller is presented in Figure 4.28.

Controller for Datapath Stage One

The first stage of the datapath, which is described in Section 4.1.4, uses two controllers to coordinate data movement through this stage. The first controller is the thermometer

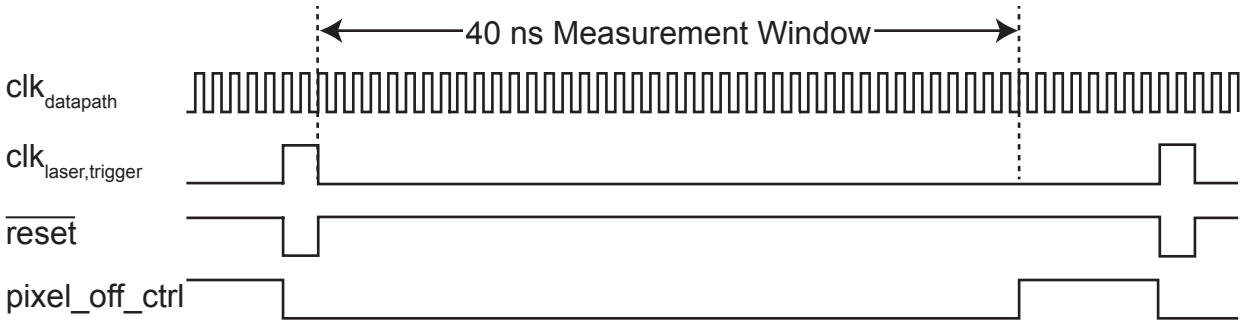

Figure 4.28: A timing diagram that demonstrates the operation of the pixel controller. $clk_{datapath}$ is operating at 1 GHz and $clk_{laser,trigger}$ is at 20 MHz. The parameter that sets the measurement window has been set to 40 cycles.

controller, which is used to lock valid data into sequential SFF-based storage elements. The second controller is used to assert the scan enable, reset, and scan clock signals that coordinate the data flow through stage one. The inputs to this controller are the laser trigger signal and the number of cycles to wait after the laser trigger before starting to scan data into the stage one SFFs. After the programmed number of datapath clock cycles have passed following a laser trigger event, the scan clock is toggled once to lock the pixel TDC data into the SFFs from Figure 4.7. Additionally, the pixel position counter, which tracks the pixel position within the row, is reset. On the following datapath clock cycle, the scan enable signal is asserted and TDC data is scanned out of the pixel TDC SFFs and into the stage one SFFs. A timing diagram showing the relationship of these signals is presented in Figure 4.29. After a total of 41 scan clock cycles (enough to capture the initial data and scan the data from the 32nd pixel in the row to the right most SFF location in stage one), clk_{scan} is disabled and scan enable is de-asserted. The scan enable signal is also passed to the stage two controller to indicate to the next stage when the scan process has finished. This entire process is repeated following every laser trigger signal.

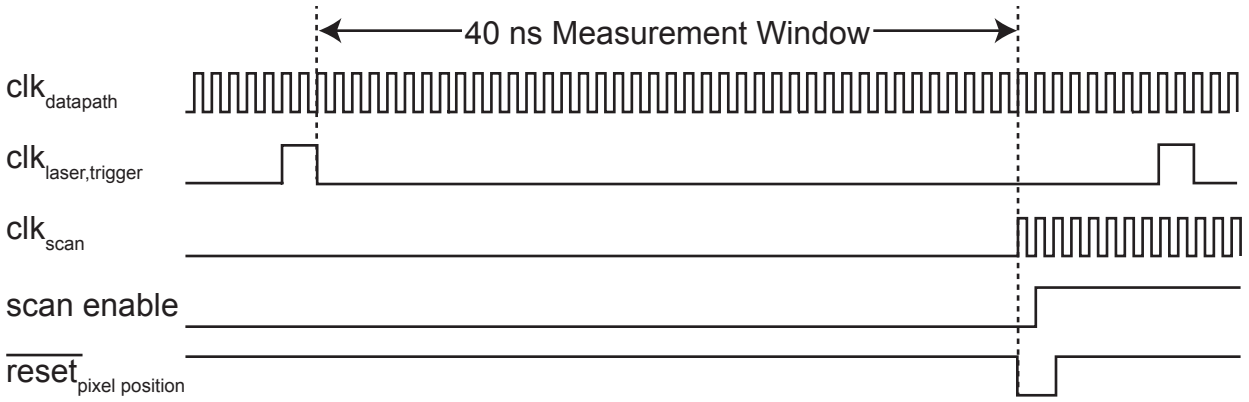

Figure 4.29: A timing diagram that demonstrates the operation of the stage one datapath controller. $clk_{datapath}$ is operating at 1 GHz and $clk_{laser,trigger}$ is at 20 MHz. The number of cycles parameter has been set to 40. This parameter sets the measurement window and determines the waiting period after a trigger edge before the data scanning process begins.

Controllers for Stages Two, Three, and Four of the Datapath

The controllers for stages two, three, and four are identical in their operation. All three stages have a thermometer controller for locking valid data. In addition, they each have a controller that synchronizes the data movement between stages. These controllers differ only in the number of shift operations that each performs due to differences in the number of storage elements available in each stage. In general, they all receive a scan enable signal from the previous stage and output scan enable and scan clock signals for the current stage. A $\overline{reset}$ signal for the previous stage is also asserted by this controller to clear the data from the previous stage after the parallel shift has occurred. This ensures that no data is cleared from the previous stage until the current stage has performed the parallel shift operation. An example timing diagram for stage two is presented in Figure 4.30. In this figure, the clk_{scan} is toggled once following the high-low transition of the scan enable signal from the previous stage and before the scan enable signal of the current stage. This clock cycle performs the parallel shift operation between stages one and two, as shown by the vertical arrows in Figure 4.21. Following this clock cycle, the scan enable signal for stage two is asserted and the U-shaped serial scan operation is performed. For stage two, this will last for 16 clock cycles,

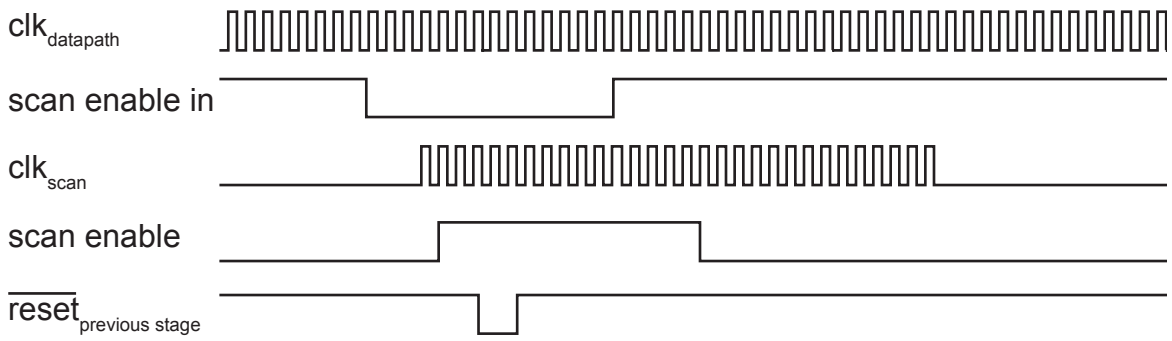

Figure 4.30: A timing diagram that shows the typical timing operations for the scan enable and scan clock signals used in the datapath stage two, three, and four controllers. The scan enable input is from the previous stage datapath controller and the reset signal is transmitted back to the previous stage once the data has been scanned in.

at which point the scan enable signal is de-asserted. However, the scan clock continues for an additional 12 cycles to ensure that the data is completely compressed to the right most SFFs in stage two. The reset signal for stage one is asserted after the first three scan clock cycles occur in stage two. The reset is only applied to the valid bit FF to reduce the load driven by the reset buffer.

FIFO Controller

The final controller in the datapath is the controller at the interface between stage four and the FIFO. In addition to performing scan operations similar to the previous stages, this controller also implements the FIFO by using a two-port synchronous register file generated with a memory generator from Artisan. This controller generates the addresses, enables, and clock signals that are used to access the register file as a FIFO. A half-rate clock is generated from the datapath clock to use for shifting data into the FIFO, which has a maximum operating frequency of 790 MHz.

4.1.8 Additional Circuits

In addition to the primary circuits used for fluorescence lifetime imaging, there are two circuits that were included in this design for test and calibration.

Calibration Buffer

A dedicated input pad is used to drive a buffer that is connected to the cal_stop multiplexer input in Figure 4.4 for a subset of pixels in the array. This cal_stop signal can be used to electronically trigger the pixel output and record a known time interval in the TDC. This calibration signal is connected to a total of 64 pixels in the middle two columns of the array using an H-tree. This translates to two calibration paths per DLL. The primary purpose of this calibration path is to calibrate the charge pump using the digital control bits of the charge pump (Figure 4.18) for each DLL on the chip.

Test Pixel

Outside of the main 64-by-64 pixel array there is an individual pixel with the full pixel control circuitry of Figure 4.4 that has its $\overline{\text{reset}}$, pixel_off_ctrl, and output signals connected directly to I/O pads. Additionally, the cal_stop signal discussed above is also connected to this test pixel. This test pixel can be used to independently verify the functionality of the pixel control circuit and SPAD.

4.2 Integrated Circuit Characterization

In this section, some of the circuits described in Section 4.1 are characterized. Because this IC was designed as a high-performance imaging system, there are relatively few access points for obtaining direct measurements. Where possible, circuit performance will be inferred from output data. Only data related to the circuit blocks described in Section 4.1 will be presented here. SPAD specific data was presented in Section 3.3 and full imaging array data

and system performance will be presented after the discussion of the complete system level camera design in Chapter 5.

A printed circuit board (PCB) was designed specifically for these initial IC characterization tests. This PCB was configured such that the two LVDS output banks on the left side of the chip were connected to Virtex-6 field-programmable gate arrays (FPGAs) and the two output banks on the right side of the chip were connected to low capacitance logic analyzer footprints and SMA connectors. This combination of connections allowed for electrical characterization of the LVDS interface, relative timing analysis, and initial testing of the Verilog hardware description language (HDL) code used to program the FPGAs. Fast comparators were included to enable synchronization of the IC to the laser trigger signals and SMA connectors were used to connect to the test pixel and the cal_stop signal. An Opal Kelly XEM6010-1500 was used to provide control signals to the IC and to serially read data from each of the Virtex-6 devices over a USB interface. The serial interface between the Virtex-6 devices and the XEM6010 created a data bottleneck that limited image acquisition to two frames of one quadrant of the imaging IC at at time. A photograph of this PCB is produced in Figure 4.31.

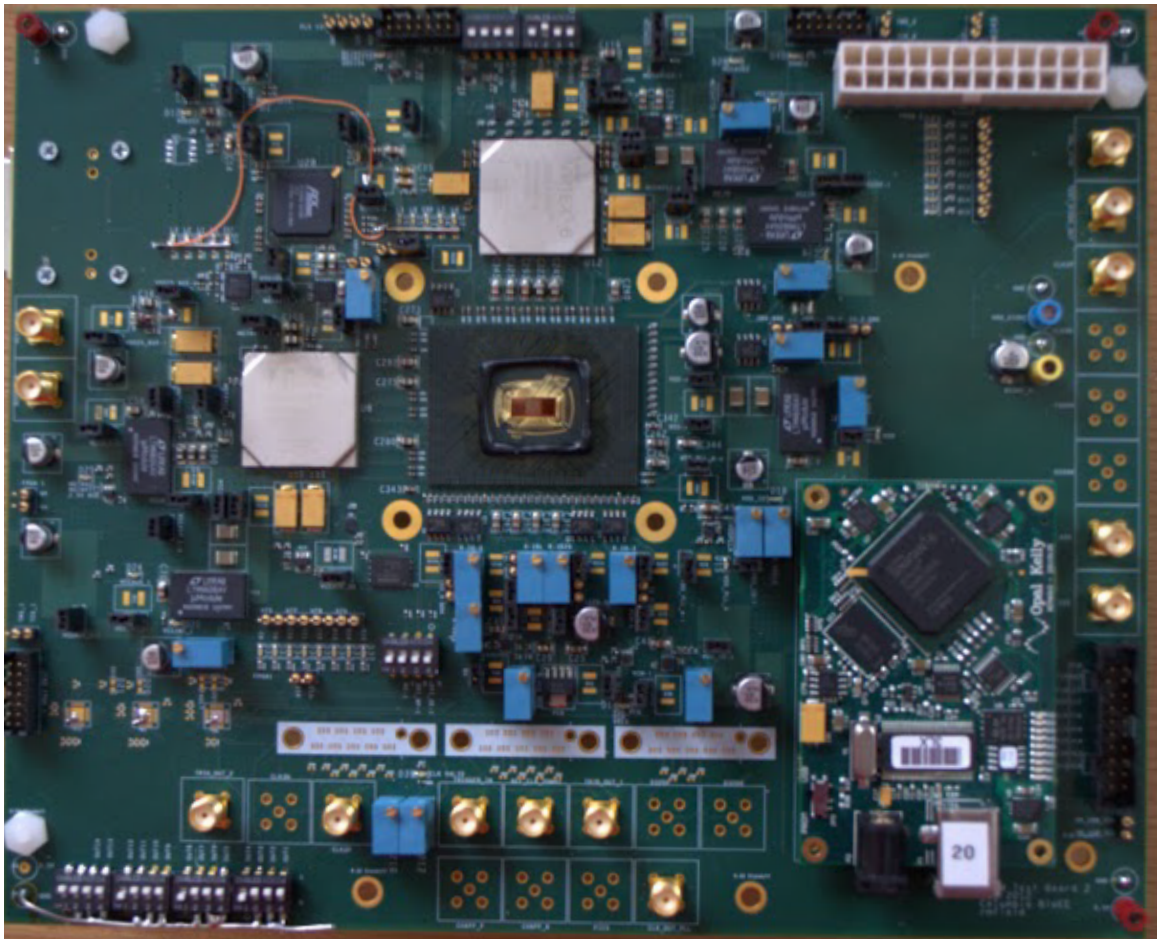

Figure 4.31: The PCB used for initial testing of the imaging IC.

4.2.1 Test Pixel Measurements

In order to verify the operation of the pixel control circuitry from Figure 4.4 and the SPAD in the complete imaging IC design, tests were first performed using the isolated test pixel. As described above, this test pixel is identical to a single pixel in the array. Measurements were taken using a 20 MHz $\overline{\text{reset}}$ signal, similar to what the pixel array controller (Figure 4.28) is expected to generate from the laser trigger pulse. $V_{dd,diode}$ was set to 3.3 V and correct pixel operation was observed with V_{bias} values ranging from -9.40 V to -12.40 V, which corresponds to an overvoltage, V_{ov} range of approximately 0.5 V to 3.5 V. Representative waveforms showing three different photon arrival times at the detector are presented in Figure 4.32.

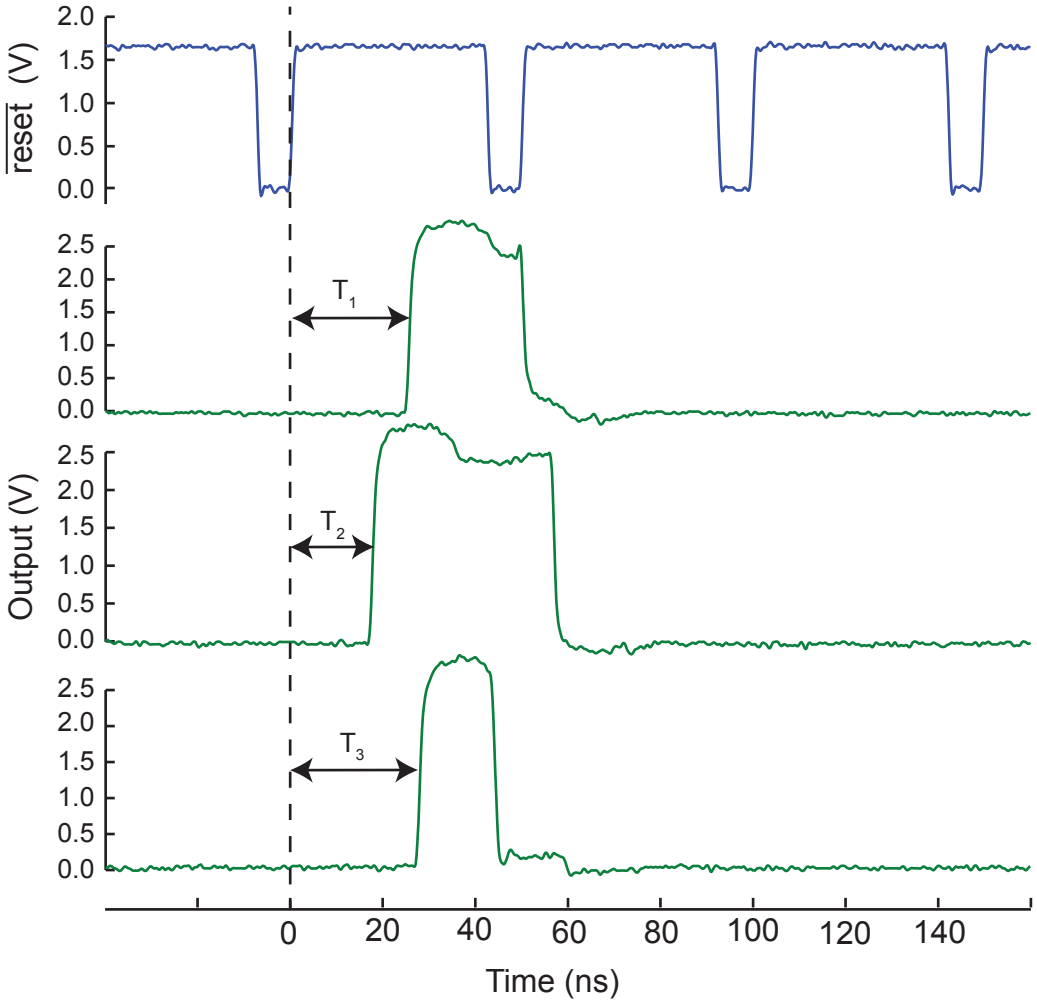

Figure 4.32: Oscilloscope traces showing the timing behavior of the test pixel. The top trace shows the $\overline{\text{reset}}$ signal for the pixel. The bottom three traces show different event times (T_1 through T_3.

With this configuration, similar dark count rates and sensitivity to the results presented in Section 3.3 were observed. However, due to the improved temporal response of the active quench and reset circuit, described in Section 4.1.2, higher count rates are possible. With high count rates, afterpulsing could become problematic due to short reset times that are insufficient for allowing charge carriers to vacate traps near the multiplication region after an avalanche event and before the SPAD is recharged [122].

The maximum count rate for this device was evaluated using a bright, uncorrelated white light source. The test pixel was biased with $V_{dd,diode}$ = 3.3 V and V_{bias} =-12.4 V.

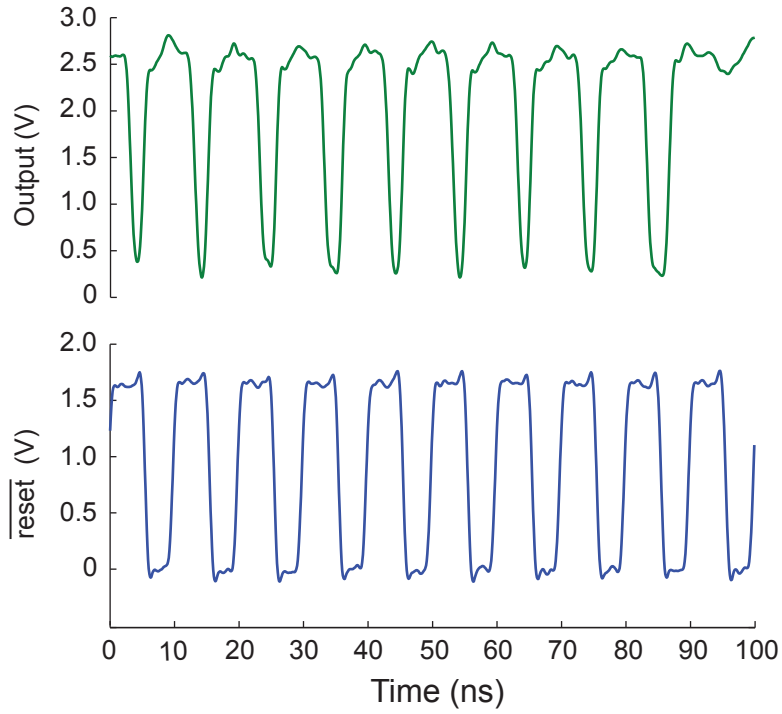

Figure 4.33: Oscilloscope traces showing the pixel output signal for a 100 MHz reset rate and a total count rate of 89.2 MHz.

The $\overline{\text{reset}}$ signal used a frequency of 100 MHz. Under these conditions, the pixel dead time, quench time, and reset time sum to 10 ns. The maximum count rate observed was 89.2 MHz and a representative signal trace demonstrating the high count rate performance is shown in Figure 4.33.

The afterpulsing for the pixel using the active quench and reset circuitry was evaluated using uncorrelated white light for both the standard 20 MHz reset rate and the 100 MHz high frequency counting reset rate. Afterpulsing probability can be measured by recording signal traces of pixel output pulses then computing the autocorrelation of the traces and scaling by the average number of uncorrelated events [123, 66]. The results of the afterpulsing measurements are shown in Figure 4.34. To record this data, 3710 signal traces of 4000 ns were acquired with 800 ps precision using Tektronix TDS7404 oscilloscope. The autocorrelation plots are generated by sweeping the value of k, the lag, in the equation 4.2. N is the total number of samples used in the calculation and x is a discrete signal of pulse arrival times.

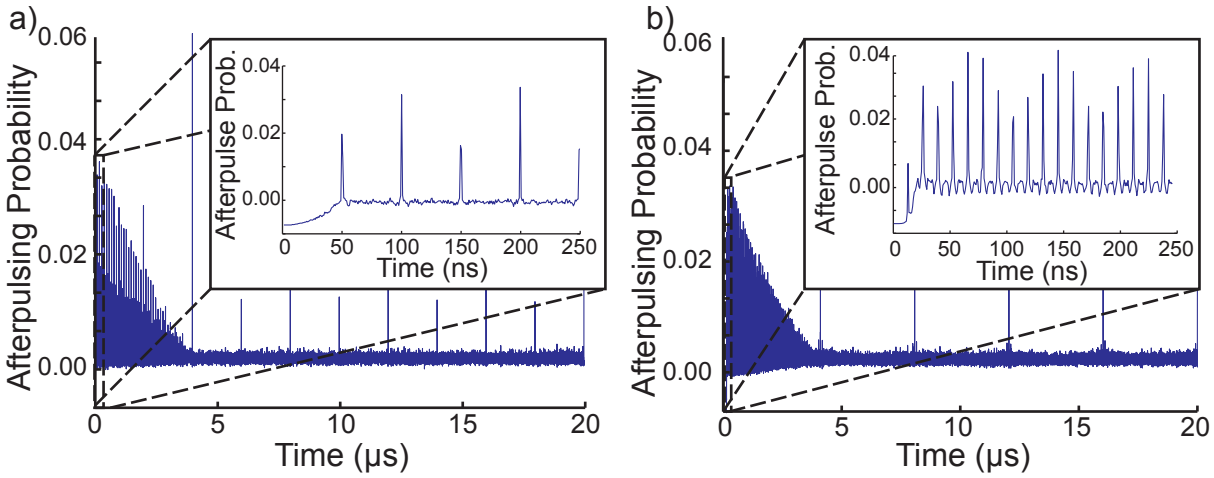

Figure 4.34: Plots showing afterpulsing probabilities at a) 20 MHz reset rate and b) 100 MHz reset rate. Both measurements show periodic spikes in the afterpulsing probability plots at multiples of the reset frequency. For both measurements, V_{ov} was set at 3.5 V.

$$R(k) = \frac{1}{N-k} \sum_{n=1}^{N-k} \frac{x(n) \cdot x(n+k)}{\left(\frac{1}{N} \sum_{n=1}^{N} x(n)\right)^2} \tag{4.2}$$

The results from the afterpulsing probability measurements show that the likelihood of afterpulsing is close to zero for both the 20 MHz and the 100 MHz reset rates with a relatively high V_{ov} of 3.5V. The periodic spikes in the measurement are a result of the synchronous SPAD reset signal, which leads to correlated events.

Additionally, a histogram of the inter-spike interval (ISI) times can be used to determine if there is afterpulsing. In a detector with afterpulsing, the histogram of the ISI times will show a bi-exponential decay. The short decay time is a consequence of afterpulsing and the long decay rate is related to the uncorrelated light source[66]. As seen in Figure 4.35, the ISI histogram is monoexponential and provides further evidence that no afterpulsing occurs with this detector.

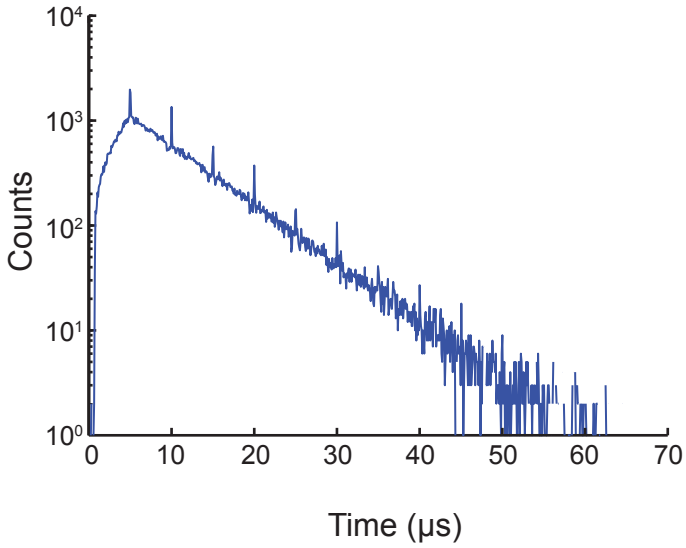

Figure 4.35: Semi-log plot showing a histogram of the inter-spike intervals measured with a SPAD V_{ov} of 3.5V and a reset rate of 20 MHz. The monoexponential decay indicates that no afterpulsing is present. Spikes in the histogram are observed at multiples of the reset frequency.

4.2.2 Pixel Control Measurements

As mentioned in Section 4.1.2, each SPAD in the array can be individually turned on or off. This allows for a particularly noisy SPAD to be completely silenced and conserves data bandwidth that would have otherwise been consumed transmitting the noise-filled data. Figure 4.36 shows a region in one quadrant of the array with a group of four noisy pixels which have selectively been turned off.

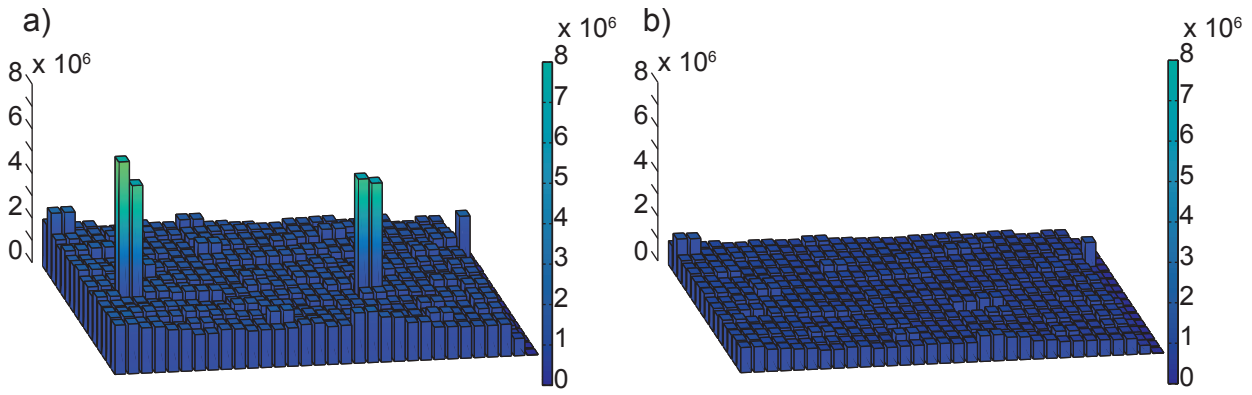

Figure 4.36: Demonstration of the capability to control individual pixels within the array. Intensity data (total number of hits per pixel) is shown for 1/8th of the total array with a total of four noisy pixels as seen in (a). These four noisy pixels have been disabled in (b).

4.2.3 SPAD Array Measurements

The DCR and sensitivity of the two quadrants of the pixel array that were connected to FPGAs were evaluated. The DCR was comparable to the measurements made on individual SPADs in Section 3.3. A representative image for one of the quadrants is shown in Figure 4.37. The left side of Figure 4.37 is closest to the periphery and the gradual decrease in DCR from left to right across this quadrant of the array is likely due to a reduced V_{ov} near the center of the array, which is the result of a voltage drop across the insufficiently sized power distribution network within the array. The average DCR for this measurement was 302 Hz and the V_{ov} for these measurements was 1.25 V.

Additionally, the global sensitivity of this quadrant of the array was evaluated using uncorrelated white light. Figure 4.38 shows recorded images with the overhead lights on and a neutral density (ND) filter covering the pixel array. The counts at each pixel are summed to form an intensity image. In Figure 4.38a, an ND 2.5 filter is placed over the array and the even illumination from the uncorrelated light source can be seen. In Figure 4.38b, an additional ND 0.9 filter is placed over the array, creating an effective attenuation of ND 3.4. The number of integrated counts per pixel is reduced uniformly throughout the array. The total count rate drops by the appropriate proportion ($1/10^{0.9}$) from around 4 MHz to about

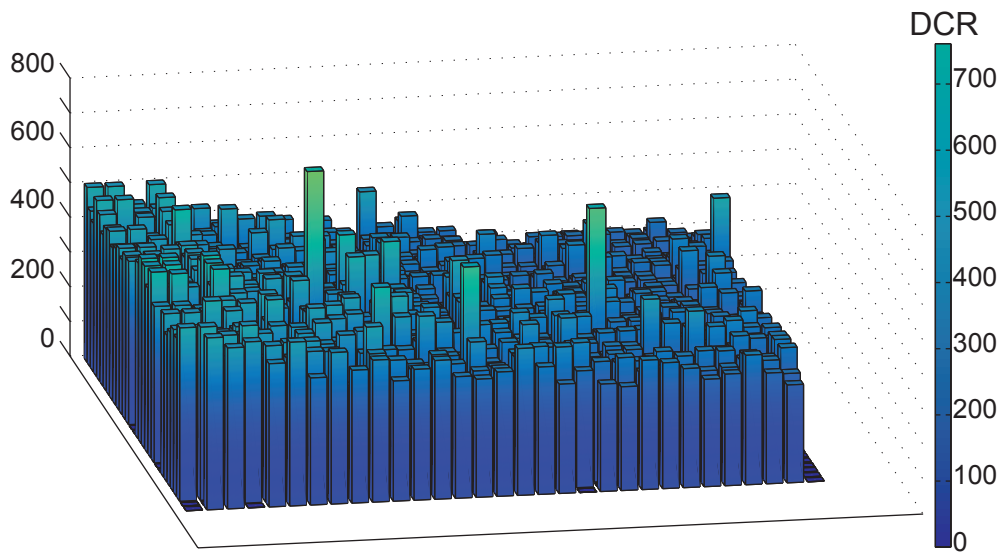

Figure 4.37: The DCR at room temperature measured for the top left quadrant of the array with $V_{ov} = 1.25V$.

0.5 MHz.

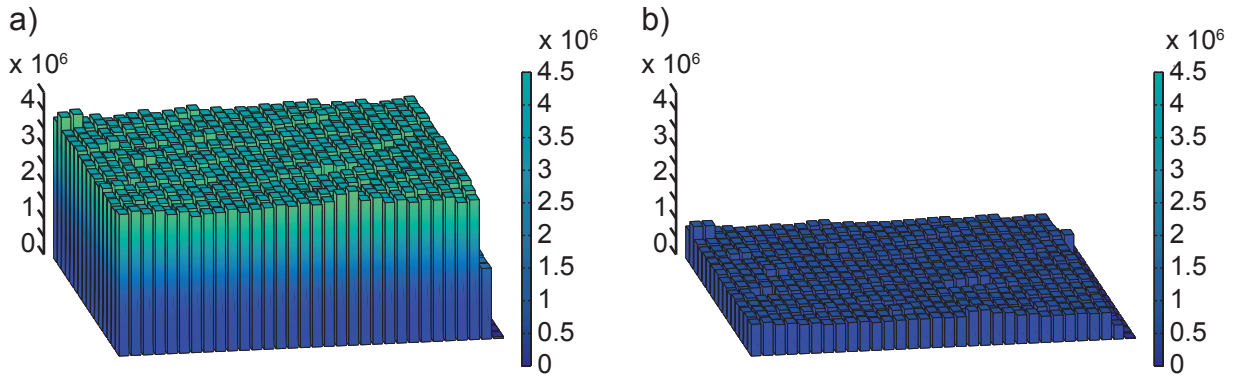

Figure 4.38: Intensity measurements that show the sensitivity of the array to light. a) An ND 2.5 filter is placed over the array and the counts at each pixel are summed. b) An additional ND 0.9 filter is placed over the array, which causes a drop in count rate by a factor of $10^{-0.9}$.

4.2.4 PLL Measurements

The PLL was designed to operate at up to 2 GHz with programmability for distributing a wide range of clocks to the DLL (for time of arrival measurements) and the datapath (for moving data off chip). The PLL could not be analyzed independently because of its tight integration into the system. The PLL outputs a 1/256th frequency reference clock that monitors the output frequency. This signal indicated that the PLL could successfully lock at up to 2 GHz, however, the on chip clock distribution was unable to support frequencies above 1.1 GHz.

4.2.5 TDC Measurements

The following TDC characterization measurements were taken using the cal_stop input to the pixel circuit that is described in Section 4.1.8. In this mode, a 400 kHz reference signal is generated using an Agilent 81130A pulse generator. This reference signal is input into the trigger port of a Stanford Research Systems DG535 digital delay generator. The AB output of the DG535 is connected to the trigger input on the IC and the CD output of the DG535 is connected to cal_stop. Delays were generated in steps of 6 ps by sweeping the 'D' delay value of the DG535.

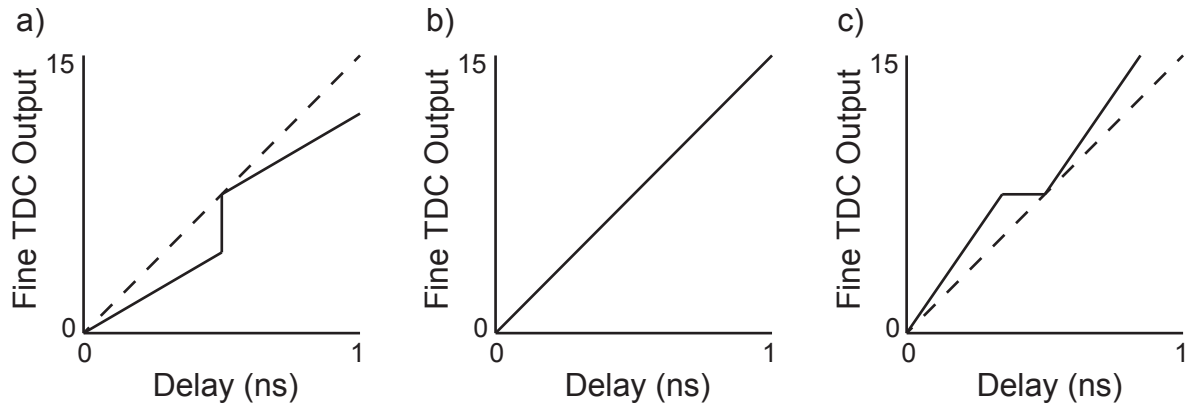

Figure 4.39: Diagrams showing the different charge pump mismatch states. a) The DN current is stronger than the UP current causing the VCDL to run slowly and the TDC output to lag behind the input delay. This results in a vertical jump in the fine TDC transfer curve. b) When the UP and DN currents are equal, the TDC output linearly tracks the delay input. c) In the case when the UP current is stronger than the DN current, the VCDL runs fast and the TDC output leads the delay input. This causes a horizontal plateau in the fine TDC transfer curve.

Charge Pump Calibration

Ideally, the phase detector and charge pump will set the VCDL control voltage such that the sixteen phases are perfectly aligned with the 1 GHz clock. When the loop is locked, the net change in charge on the V_{ctrl} capacitor will be zero. If there are imperfections in the phase detector or the charge pump, this zero net charge condition can occur with a static phase error. The most common cause of such a static phase error is a difference in strength between the charge pump PFETs that provide the UP current and the NFETs that provide the DN current. As discussed in Section 4.1.3, a calibrated charge pump was designed in order to compensate for these differences.

Each charge pump is manually calibrated by setting 12 control bits (6 bits each for UP and DN) in the global scan chain. Measurements of the fine TDC value can be used to determine if the currents are mismatched and which of them is stronger. A diagram highlighting the indicators of UP/DN current mismatches is presented in Figure 4.39.

Measurement results showing the adjustment of the fine TDC transfer characteristic through changing the charge pump calibration bits are shown in Figure 4.40. In Figure

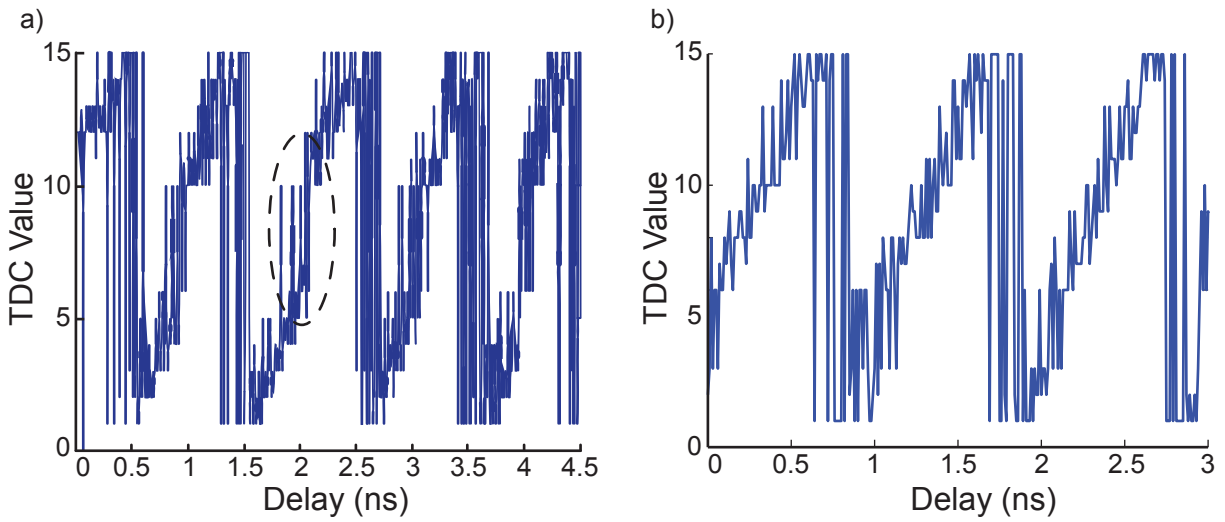

Figure 4.40: a) TDC output when the UP and DN charge pump calibration codes are 010000 and 010001, respectively. The dashed oval highlights the region of the transfer characteristic that indicates that the VCDL is running slowly (the DN is stronger than the UP current). b) The TDC output after charge pump adjustments are made. The UP calibration code is 000101 and the DN calibration code is 101000. In this hardware, the UP devices are stronger than the DN devices.

4.40a the DN calibration bits have been set to 010001 and the UP calibration bits to 010000. This creates a mismatch with the total DN width 360 nm larger than the combined UP device width in the current mirrors of the charge pump. As a result, a stronger DN current and a characteristic vertical jump in the fine TDC transfer curve are observed, as shown schematically in Figure 4.39a. In Figure 4.40b the DN calibration bits are set to 101000 and the UP bits are set to 000101, which leads to the DN current mirror device being effectively 180 nm larger than the UP device. This perfectly cancels the drive strength difference between the UP and DN currents. Consequently, the vertical jump in the fine TDC transfer characteristic is no longer present and corresponds with the schematic representation of Figure 4.39b.

TDC Fine Offset

Due to the time delay between when the rising edge of the reference clock enters the VCDL and when it clocks the coarse TDC counter input, the fine TDC values output from the

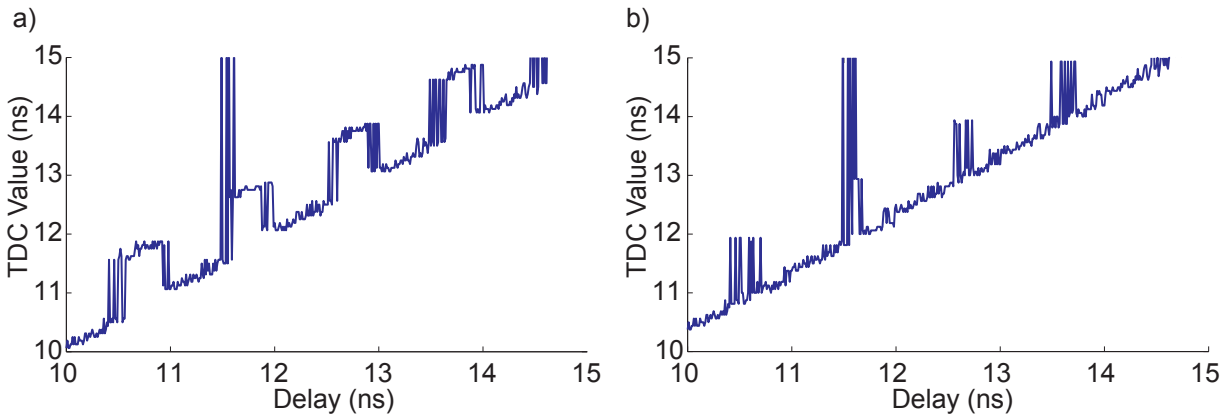

Figure 4.41: a) Raw data recorded from the TDC in calibration mode. The fine values are at a minimum in the middle of a coarse step. b) An offset of 5 has been added to the fine TDC values of the same raw data and a much more linear transfer curve is the result.

VCDL are offset relative to the coarse TDC steps. This offset is shown in Figure 4.41a. In order to correct for this, a constant value of 5 is added to all fine TDC values with any carry out bits discarded (i.e. 14 (1110) + 5 (0101) = 19 (1 0011) becomes 3 (0011)). This operation results in the proper alignment of the fine TDC values with the coarse TDC steps, as presented in Figure 4.41b.

Non-Linearities

The two measures of non-linearity in a TDC are the differential non-linearity (DNL) and the integral non-linearity (INL). The DNL is an indication of the amount that each LSB step differs from an ideal step. The INL is a measure of the overall non-linearity of the converter and can be evaluated by the slope of a best-fit line or by considering the values of the two endpoints of the transfer function.

The non-linearities were measured using the electrical calibration path described in Section 4.1.2. A known delay was generated and distributed across the test PCB to the IC where the cal_stop buffer distributes the signal to the pixels used for calibration. This delay value is swept over the entire range of the TDC in steps of 10 ps in order to evaluate the linearity for all possible time intervals between 0 and 64 ns. The results of these delay

sweeps are shown in Figure 4.42.

In Figure 4.42a, the overall linearity of the TDC tracks well with the input time interval. However, the jitter on the electrical stop signal used to define the delay is on the order of 200 ps, which is more than three times the LSB of the TDC. As a result, accurate determination of the TDC linearity is difficult. The fine resolution TDC data shown in Figure 4.42d shows the roughly 3 LSB jitter in the measurement. This jitter was confirmed using a real time oscilloscope. In the transfer curve in Figure 4.42a, periodic large spikes can be observed. By separating the fine and coarse components of the TDC value in Figures 4.42d and 4.42e, it is clear that these spikes in the transfer curve are contributions to the overall response made by the coarse counter. This undesirable contribution is due to the asynchronous stop signal that is used to lock the coarse counter value into flip-flops, as shown in Figure 4.7. Because the stop signal is not synchronized with the counter clock, if the counter value transitions within the setup and hold times of the flip-flop, the latched value will be non-deterministic. This error could be corrected by simply synchronizing the stop signal for the coarse counter with the counter clock. An example circuit to accomplish this synchronization that is based on a pre-charged domino logic stage is shown in Figure 4.43.

A timing diagram showing how the circuit in Figure 4.43 can be used to synchronize the coarse counter stop signal is presented in Figure 4.44. In this figure, the circuit is precharged by the $\overline{\text{reset}}$ input. When the asynchronous stop signal arrives from the pixel output, it will discharge the bottom NFET. When the clk input is driven high, the middle NFET will discharge and trigger the count stop signal. The buffers on the clk signal are used to generate a delay such that the count stop signal will latch the counter state at a minimum of Δt after rising edge of the clock. By ensuring that Δt is large enough to allow the counter value to settle before the counter stop triggers, the flip-flop metastability window can be avoided and the non-linearities associated with the coarse counter metastability in Figure 4.42 would be eliminated.

An alternative technique for quantifying the DNL is the code-density measurement. In this approach, the time of uncorrelated dark count events is recorded for many cycles and the results plotted in a histogram. Because these events are uncorrelated, an ideal converter should have a uniform distribution of counts in the histogram. Non-linearity can be quantified by normalizing the histogram to the average histogram bin value [124]. The result of this normalization is the DNL in LSB. Through eliminating jitter in the measurement electronics, this technique results in a measured jitter of less than 2.27 LSB over the lowest 128 code values. Only the lowest 128 code values could be recorded due to the histogram size limitation (128 bins) of the FPGA HDL. The data was recorded with the minimum bin width of 62.5 ps, which corresponds to the range of 0 – 8 ns. A figure plotting the code-density derived DNL is shown in 4.45.

4.2.6 Impulse Response of SPAD and TDC

The impulse response function (IRF) of the SPAD and TDC combined was evaluated by repeatedly triggering the SPADs using a 500 nm band from a Fianium supercontinuum laser with pulse width of 10 ps. The outputs of the TDC were collected and formed into histograms, which resulted in a distribution with a peak width of only two LSB, or 125 ps 4.46.

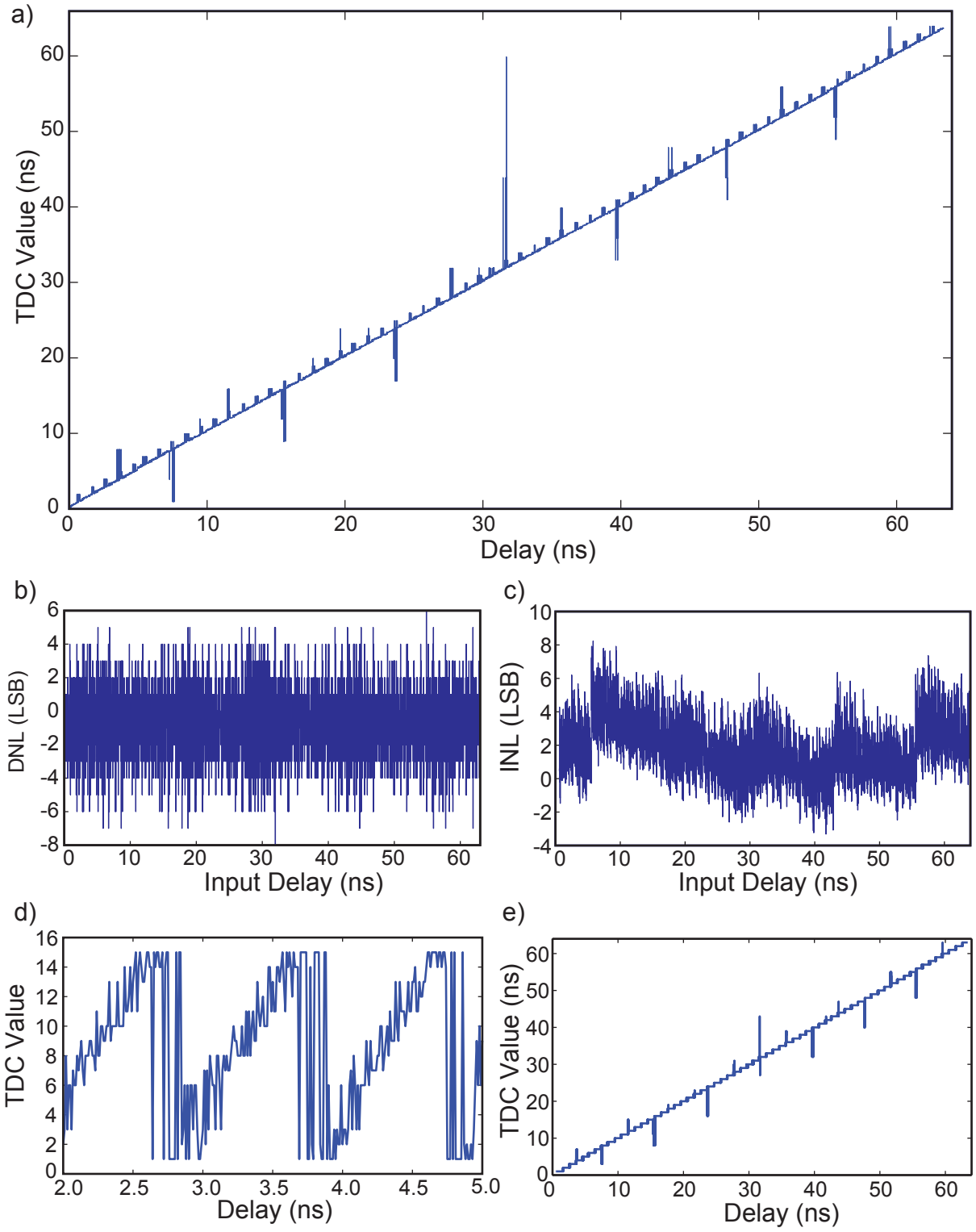

Figure 4.42: Plots showing the a) transfer curve of the TDC, b) DNL of the TDC and c) INL of the TDC. The transfer curves of fine and coarse components of the TDC value are presented in d) fine and e) coarse.

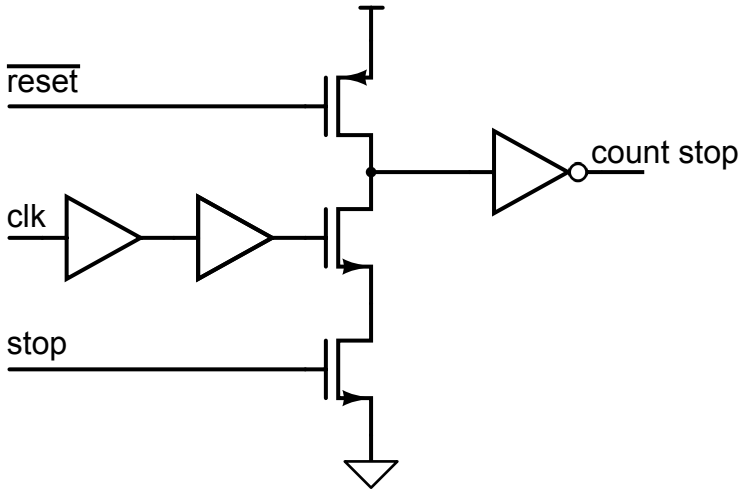

Figure 4.43: TDC counter stop synchronizer.

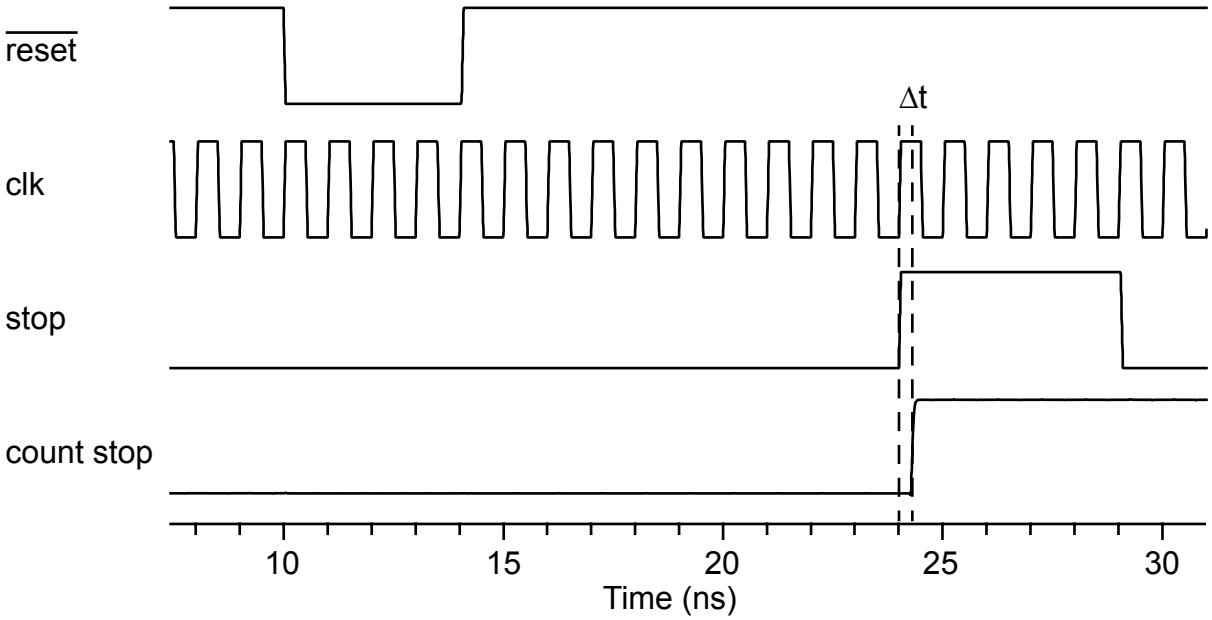

Figure 4.44: TDC counter stop synchronizer waveforms

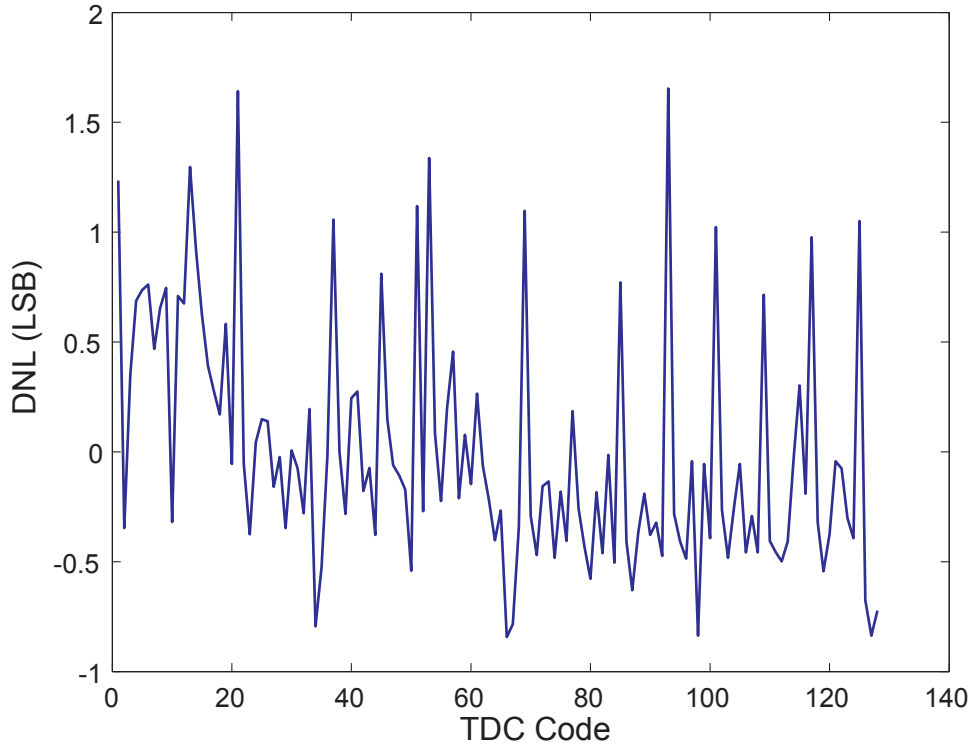

Figure 4.45: Plot of the DNL acquired through the code density technique for the lowest 128 TDC values.

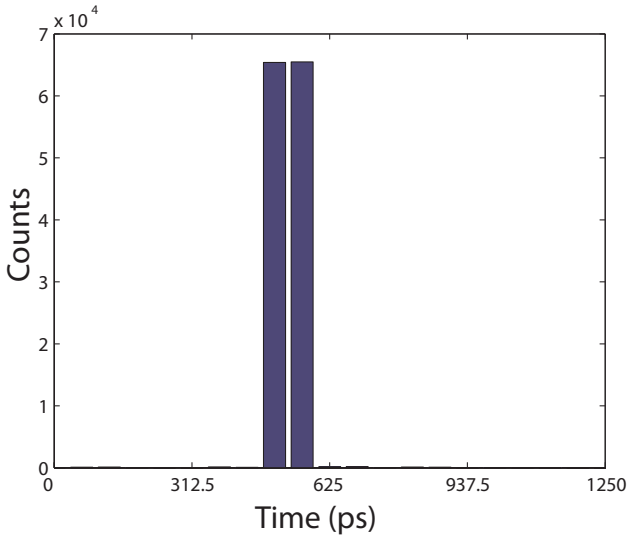

Figure 4.46: Figure showing the impulse response recorded using the minimum bin width (62.5 ps). The impulse response spans two histogram bins for a width of 125 ps. This includes contributions from both the SPAD and the on-chip TDC.

4.2.7 LVDS Buffer Measurements

A representative data eye measured from an on-chip LVDS output buffer driving a 100 Ω differential load over a 100 Ω PCB trace is shown in Figure 4.47. This measurement was taken at an output frequency of 500 MHz from bit 20 of the output bus. Because of the round-robin pattern used by the FIFO coordinator, bit 20 will change regularly with a period twice that of the data clock. This measurement was taken with an Agilent 86100 wide bandwidth oscilloscope.

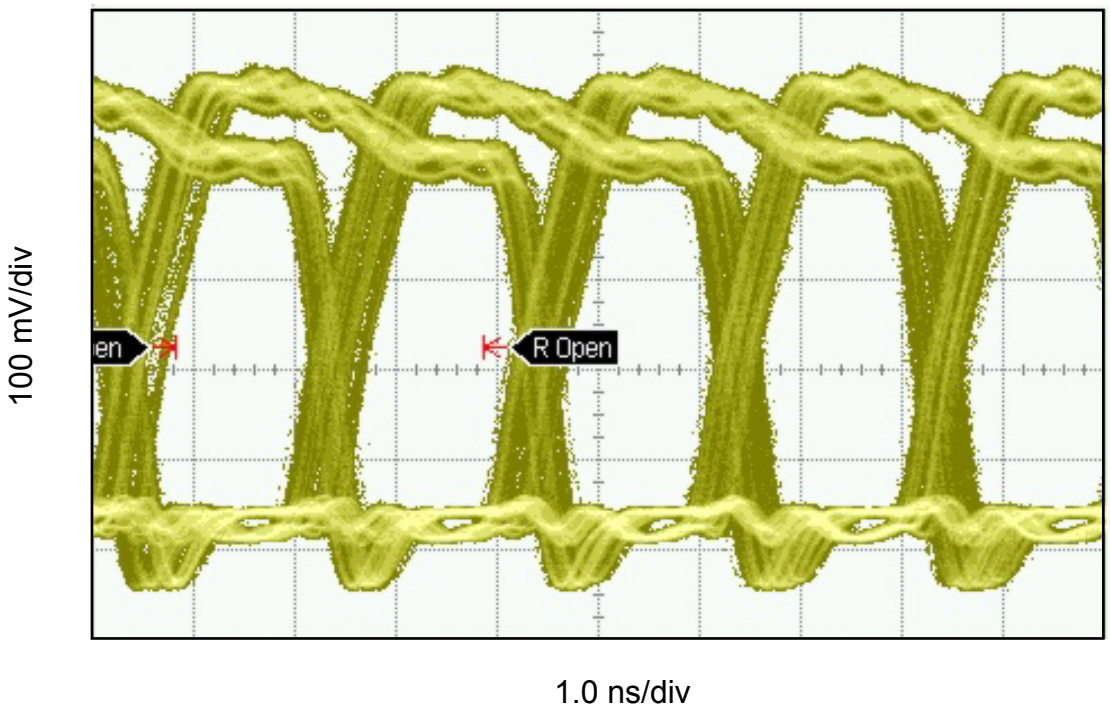

Figure 4.47: A representative data eye for bit 20 of the LVDS output bus.

4.2.8 Preliminary Images

The characterization data presented in the previous sections demonstrate that the main components of the imaging IC function properly. In order to demonstrate that the FLIM specific datapath transfers the data as designed, it is necessary to acquire an image that records lifetime data. Due to the previously discussed system level limitations of the system characterization PCB, image acquisition is limited to a single frame. Figure 4.49 shows a representative FLIM image acquired using the imaging IC. To capture this image, a Petri dish was placed over the array and a drop of 0.5 mM fluorescein dye in pH 7.4 phosphate buffered saline (PBS) was pipetted into the dish over the active region. A Fianium SC-450pp supercontinuum laser with acousto-optic tunable filter (AOTF) was used to excite the spot of fluorescein dye with 488 nm light. A picture of the drop of fluorescein dye being excited with the laser is shown in Figure 4.48. The Fianium includes a pulse-picker allowing the 20 MHz repetition rate to be divided down by integer factors. In these experiments the laser was

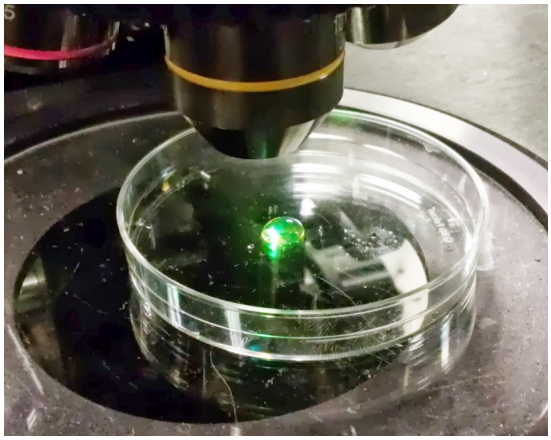

Figure 4.48: Photo of fluorescein dye.

pulse-picked to a repetition frequency of 1 MHz. The IC was configured with a V_{ov} of 1.25 V, a DLL reference frequency of 1 GHz, a datapath frequency of 250 MHz, calibration bits for DN = 40 and UP = 5, and a measurement window of 32 ns. The FPGAs were configured to bin the arrival time data into histograms with bin widths of 250 ps, which is appropriate for the measured 125 ps impulse response of the SPAD and TDC combination. An image captured during these experiments is shown in Figure 4.49. In 4.49a, an intensity map showing the total number of photons collected per pixel is shown. Figure 4.49b shows the corresponding lifetime image with an inset of a representative fluorescence decay measured at one pixel. Missing pixels or disabled pixels in this image were filled by using interpolation. The expected fluorescence lifetime for fluorescein dye is around 4-5 ns according to [125], which is consistent with the values measured in Figure 4.49b.

From these FLIM images, the independence of fluorescence lifetime with respect to fluorescence intensity can readily be observed. Regions with fewer than 10,000 total photon counts measure the same lifetime value as those with more than 25,000. Additionally, this image demonstrates that accurate lifetime imaging is possible despite the TDC error from the metastability in the coarse counter.

4.2.9 FIFO Controller Bug

Through the initial imaging tests, a minor bug in the FIFO coordinator controller was identified. After all of the data that was written to a FIFO has been transferred, the last data word will remain on the output port of the FIFO. Because the data is written to the FIFO with the valid bit set high, all data coming out of the FIFO and being passed to the LVDS buffer banks will have the valid bit asserted. This creates a bug when the last data has been written but the valid bit has not been cleared. Consequently, the last data word out of the FIFO will be repeated with the valid bit asserted until new data is written to the FIFO. In order to use the array to acquire FLIM images, a duplicate data detection block is included in the FPGA HDL. This duplicate data detection circuit identifies data from the same FIFO output that does not change from cycle to cycle. The duplicate detection scheme could easily be moved on-chip or the FIFO coordinator controller could be modified to de-assert the valid bit after the data is transmitted to the LVDS bank in future implementations.

4.2.10 Column Counter Bug

There is timing error for the 5-bit counter used to track the pixel location within the half-row at which the photon event occurred (Section 4.1.7). There is a full clock period delay between when the data starts shifting and the counter first increments. This causes pixels from both columns 0 and 1 (the two columns closest to the edge of the chip) to record a position of column 0. The impact of this bug can be limited by disabling column 0 for both half-rows such that array is effectively reduced to a 62-by-64 array.

4.2.11 Power Consumption

The test PCB used for these measurements included sense resistors and test points for measuring the current consumed by the IC. Measurements for the current consumption were taken while all four quadrants of the chip were enabled and all TDCs were running. Table

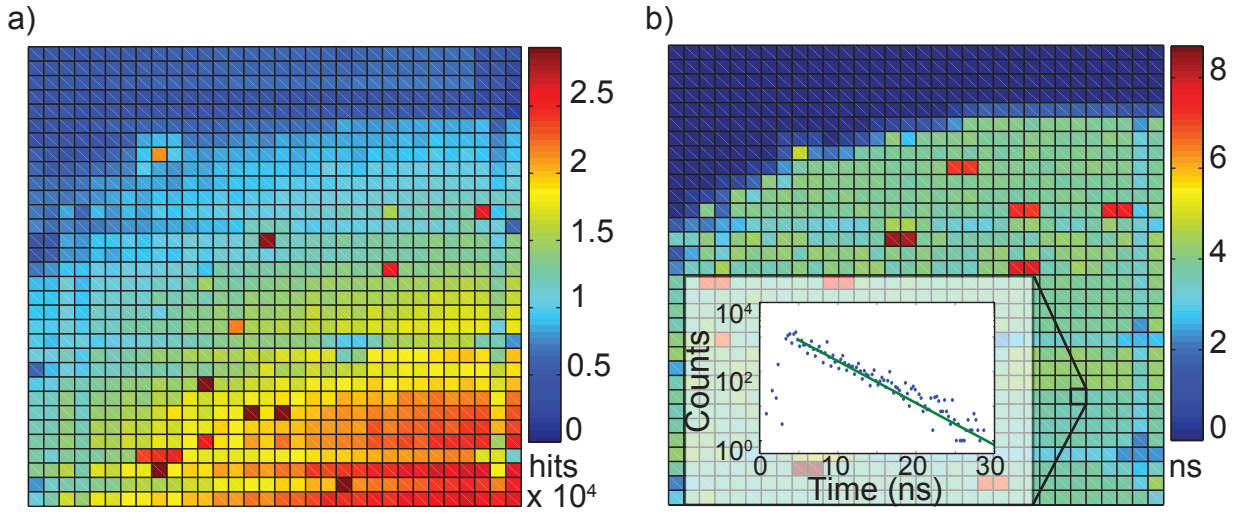

Figure 4.49: a) an intensity-based image that is formed by summing the total number of photon counts at each pixel. b) lifetime image of fluorescein dye.

Table 4.4: Power consumption of imaging IC.

Supply Name	Voltage (V)	Power (W)
V_{dd}	1.6	6.10
$V_{dd,IO}$	2.5	2.15
$V_{dd,reg}$	2.5	0.125
$V_{pll,analog}$	1.6	0.06
$V_{diode,p}$	-11	0.35
Total Power:		8.79

4.4 lists the supply voltages and the power drawn from each. The total measured power drawn by the IC is 8.79 W. Both Vdd,diode and Vdd are dependent on the incident photon flux and the values presented in Table 4.4 are average values for a typical photon flux.

4.3 Summary

In this chapter an overview of the FLIM specific imaging IC was presented along with some initial tests that demonstrate the capabilities of this IC. Although some imaging data has been presented, the full capabilities of the IC have not been tested due to limits on the data bandwidth available on the PCB described in the section above. These initial tests show

that the complete system is functional. More in-depth tests, including images acquired with the whole array, will be presented in Chapter 5.6

Chapter 5

Fluorescence Lifetime Imaging Microscopy System Design

This chapter presents system level considerations for the high speed fluorescence lifetime imaging camera. Although a considerable effort was made to reduce the amount of data that is transmitted off chip, as described in Chapter 4, the imaging IC can still produce up to 42 Gbps of data. Consequently, this camera system is designed around managing this data and further compressing it before transferring it to a computer for lifetime extraction. Previous SPAD arrays have been limited by the data bandwidth available for transmitting TCSPC data, resulting in restrictions on the number of frames that can be acquired to bursts of a few hundred frames [126] or to small regions of the array [26]. Everything in the system described here has been designed such that a continuous stream of full images can be transferred to a computer, processed, and saved.

5.1 System Overview

The complete imaging system consists of a single FLIM imager IC that connects with four Xilinx Virtex-6 FPGAs. These FPGAs receive the data output from each quadrant of the IC and create histograms with 128 bins for each pixel. Through the binning operation, the total

amount of data per frame is reduced to 2 Mb per quadrant. Each of the Virtex-6 FPGAs contains a high-speed transceiver that can be configured as a 2nd generation Peripheral Component Interconnect Express (PCIe) interface. This PCIe interface provides enough data bandwidth (up to approximately 4Gbps per lane after packet overhead) to transmit up to 2000 frames per second. The four PCIe transceivers connect to a PCIe switch that combines them into a single four lane (x4) interface. This x4 interface then connects directly with a computer over a cabled PCIe interface. Because the data transfer happens over PCIe, a direct memory access (DMA) approach is employed whereby the Virtex-6 FPGAs can write directly to the system memory of the computer without any involvement from the central processing unit (CPU). Once the data has reached system memory, the CPU can access it and perform computations on it or save it to disk.

The infrastructure to enable this system along with the details for each of the subsystems and power delivery are presented in the following sections. Additionally, considerations for software design are discussed. Imaging results using the complete array are presented in Section 5.6.

5.2 IC Packaging

The IC described in Chapter 4 has a number of characteristics that required a custom packaging solution. The primary considerations were the number of differential I/O pairs required, the need to dissipate power through the back of the die, and the physical size of the IC.

There are a total of 22 differential output buffers in each of the four LVDS output banks for 176 total package pins that must be routed differentially. Additionally, there are a total of 22 single ended signals that are used for controlling the IC and another 16 for supplying current biases. In total there are 214 signal pins. Back-of-the-envelope calculations for the number of package pins that should be used to ensure low resistance paths for all

of the voltage supplies result in a minimum of 135 total power pins. With a minimum requirement of 349 package pins, a grid array package is a suitable choice.

The expected power dissipation for the IC was calculated by combining simulation results for small blocks with coarse approximations for the larger ones that could not be simulated in their entirety. The TDCs and output buffers could be easily segmented and simulated to estimate their dynamic power consumption. From simulation, it was estimated that the TDCs would draw 1.5 A from a 1.5 V supply (2.25 W) and the output buffers would consume 0.5 A from a 2.5 V supply (1.25 W). The power used by each pixel during a firing event is negligible in the context of overall power consumption. The datapath is the largest load and is difficult to simulate because of its complexity and size. The dynamic power consumption for this large portion of the chip was estimated by considering the total area occupied by the datapath and determining the number of NAND2 standard cells that would fill the same space. The total power is then approximated by taking the average power consumption of the NAND2 cell and making assumptions about the activity of the datapath and the overall utilization of the area.

The total datapath area is 17,238,410 μm^2 and a single NAND2 device in the standard cell library used in this design has an area of 11.52 μm^2. Assuming a 40% fill ratio, this leads to an effective total equivalent number of gates of approximately 900,000. The NAND2 gate used here has a AC power consumption of approximately 15 nW/MHz. For a 1 GHz datapath frequency this results in an estimated power consumption of 9 W for the datapath. Combining each estimate for power consumption yields a total power of 12.5 W and a predicted power density of 33 W/cm^2. This power density is on par with typical microprocessor designs [127]. However, unlike microprocessors, the IC in this design must have its top surface exposed such that the photodetectors can receive incoming photons. As a result, efforts to heat sink the IC must involve thermal conduction through the back side of the IC and its package.

Some CMOS imaging ICs use ceramic lead-less chip carriers (LCCs) packages, but

these are limited in the number of pins and routing layers that can be used [128]. In order to accommodate the large number of pins required for this design, a dielectric laminate ball grid array (BGA) package with a copper core was designed to provide a sufficient number of I/O connections and to allow for heat transfer through the back-side of the package. A bismaleimide triazine (BT) resin is used as the insulating dielectric for the BGA and 6 layers of metal provide power distribution and signal routing capabilities in a total package thickness of 31 mils. A large number of plated through hole (PTH) vias are placed in the center of the package, where the copper core is, in order to promote heat transfer between each of the six layers in the copper core. These vias are connected directly to ground balls of the BGA so that heat can be transferred from the via directly to the PCB.

Each of the differential output signals from the IC is routed through the package over a differential signal trace with a controlled impedance of 100 Ω. Additionally, four power rings were used to connect each of the supplies needed for the IC. In order from the innermost ring to the outermost, these supplies are ground, core supply, I/O supply, and TDC regulator supply. The second power ring was split to separate the right and left I/O supplies. The BGA balls are placed at a 1 mm pitch and a total of 489 balls were used. A photograph showing the top and bottom of the package is presented in Figure 5.1. Each of the features described above is visible in these photographs.

Package Warping

Despite efforts to design a robust package, there is a problem with warping during the solder reflow process. It is likely that this problem is due to the relatively thin package (0.031 inches) combined with an asymmetry in the metal layers. Of the six metal layers used, 3 were dedicated to signal routing and the other 3 were used as power planes. The layers were arranged as: 1.) signal, 2.) power, 3.) signal, 4.) power, 5.) power, 6.) signal. As the package is heated and the metals expand, warping will occur if there is an asymmetric force generated by the thermal expansion during the reflow process. With the chosen layer

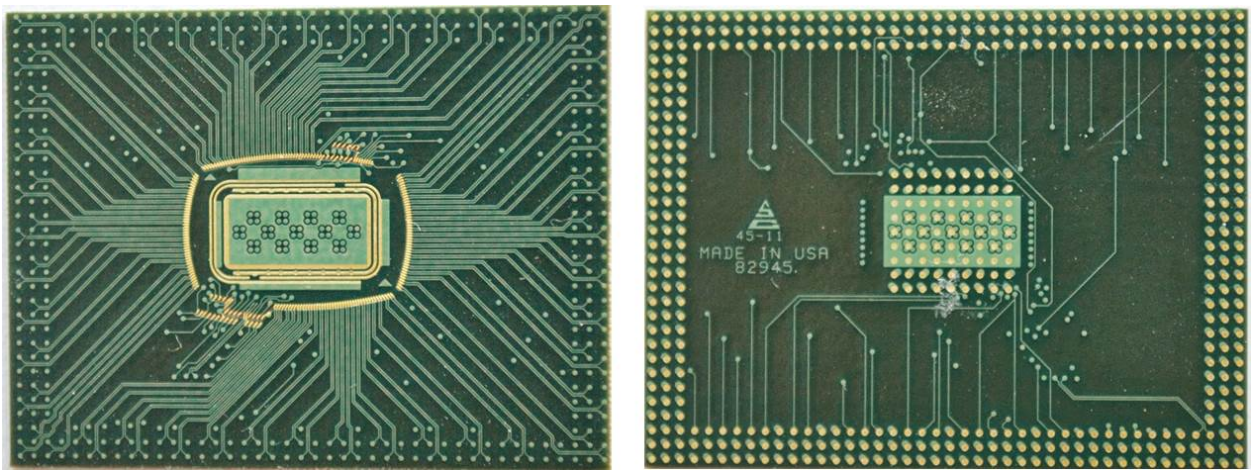

Figure 5.1: A photograph showing the top (left) and bottom (right) of the custom package designed for the imaging IC.

assignment, there will be more metal close to the bottom of the package and warping will occur where the signal routing near the top is sparse. Figure 5.2 shows the artwork for the six package layers that highlights the asymmetry in the manufactured metal distribution. Figure 5.3 is a photograph demonstrating the package warping under normal solder reflow conditions.

Aside from improving the symmetry of the metal distribution within the package, other approaches can be used to mitigate package warping during the reflow soldering process. One such approach is to create a thick dielectric frame that is made of the same material as the insulating layers and will provide additional stiffness to the package during reflow while allowing for a thin substrate at the center where heat transfer will occur during normal IC operation [129]. Another option is to use a metal structure that is designed to reduce package warping similar to what was done in [130]. A third is to use a controlled temperature profile for the reflow soldering process that uses slow ramp rates and multiple soak stages in order to allow more gradual heat expansion. A suitable profile is presented in figure 5.4 and the resulting improvement in package warping using this profile is shown in figure 5.5.

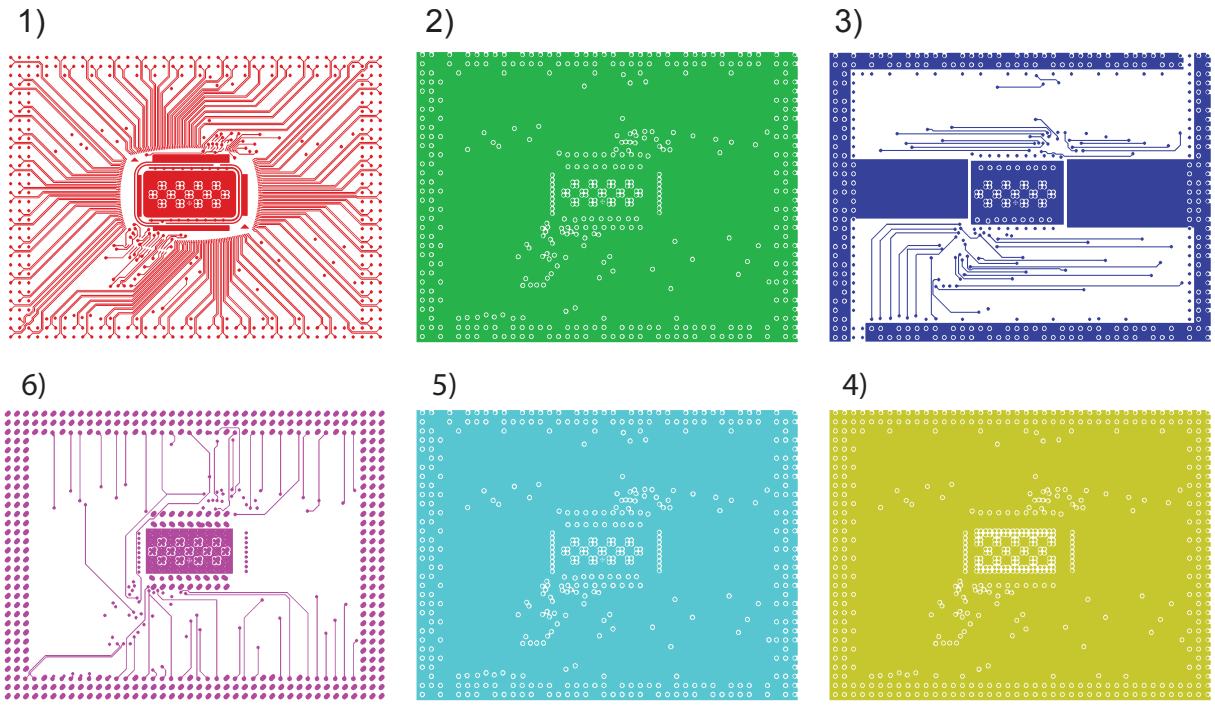

Figure 5.2: The artwork used to make the BGA package showing the uneven metal density distribution between the 6 layers.

5.3 FPGAs

The custom package described above allows for board-level connections to be made with the custom IC. These connections enable the IC to be powered and the outputs to be captured by additional chips using the LVDS standard. Four field-programmable gate arrays (FPGAs) are used to configure, control, and capture the raw arrival time data from the IC (10 bits location, 10 bits timing information, and 1 valid bit). The FPGAs then generate per-pixel histograms from the data and the lifetime is extracted for each pixel to form an image. Additionally, the FPGAs are used to transfer data to a computer over the PCIe bus. This section presents an overview of the FPGA features leveraged in this design as well as a description of the architecture and hardware description language (HDL) code used to properly configure the device.

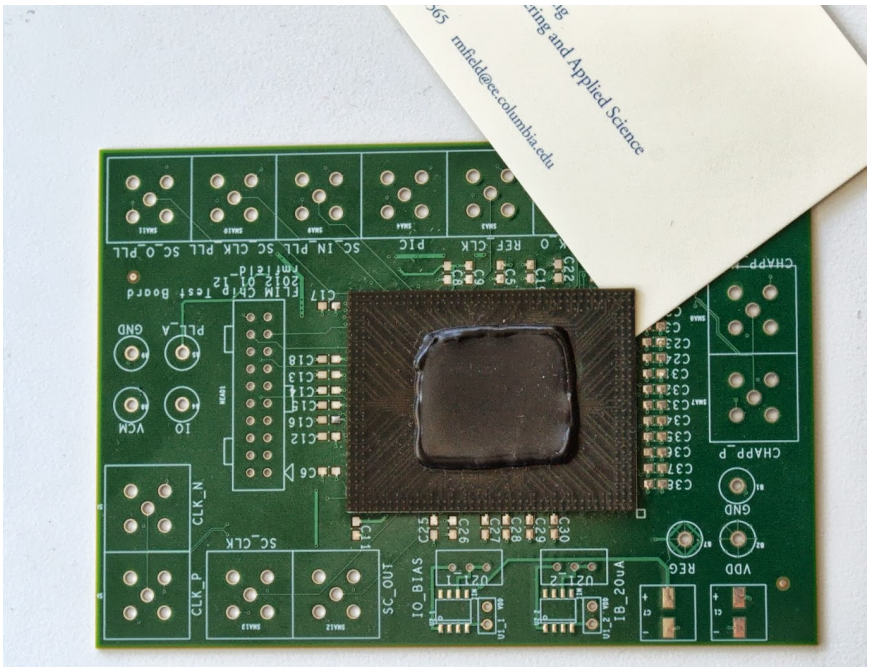

Figure 5.3: Under normal solder reflow conditions, the custom package will warp enough to slide a business card between the PCB and package after soldering.

5.3.1 FPGA Characteristics

An FPGA is a programmable logic device that can be reconfigured to implement arbitrary logic functions. FPGAs are made of thousands of look-up tables (LUTs) that are used to implement logic and typically accept four to six inputs and can produce one or more outputs depending on the manufacturer's architecture and the device version. The LUT output(s) are commonly connected to a combination of multiplexers (MUXes), fixed logic gates, registers, and a switched interconnect matrix. A typical FPGA architecture will combine multiple LUTs with registers and other fixed logic (MUXes and inverters typically) to form a logic slice or module [1]. The switched interconnect matrix allows for programmable routing between slices, which enables arbitrarily complex logic functions to be created. The performance of an FPGA (in terms of maximum operating frequency) is a function of the complexity of the logic that is implemented between two registers. A single LUT has an intrinsic delay, which is

[1]The most common FPGA manufacturers are Xilinx and Altera. Xilinx calls the combination of LUTs and registers a slice while Altera refers to them as adaptive logic modules. In this work, Xilinx FPGAs were used and the slice terminology will be adopted from this point forward.

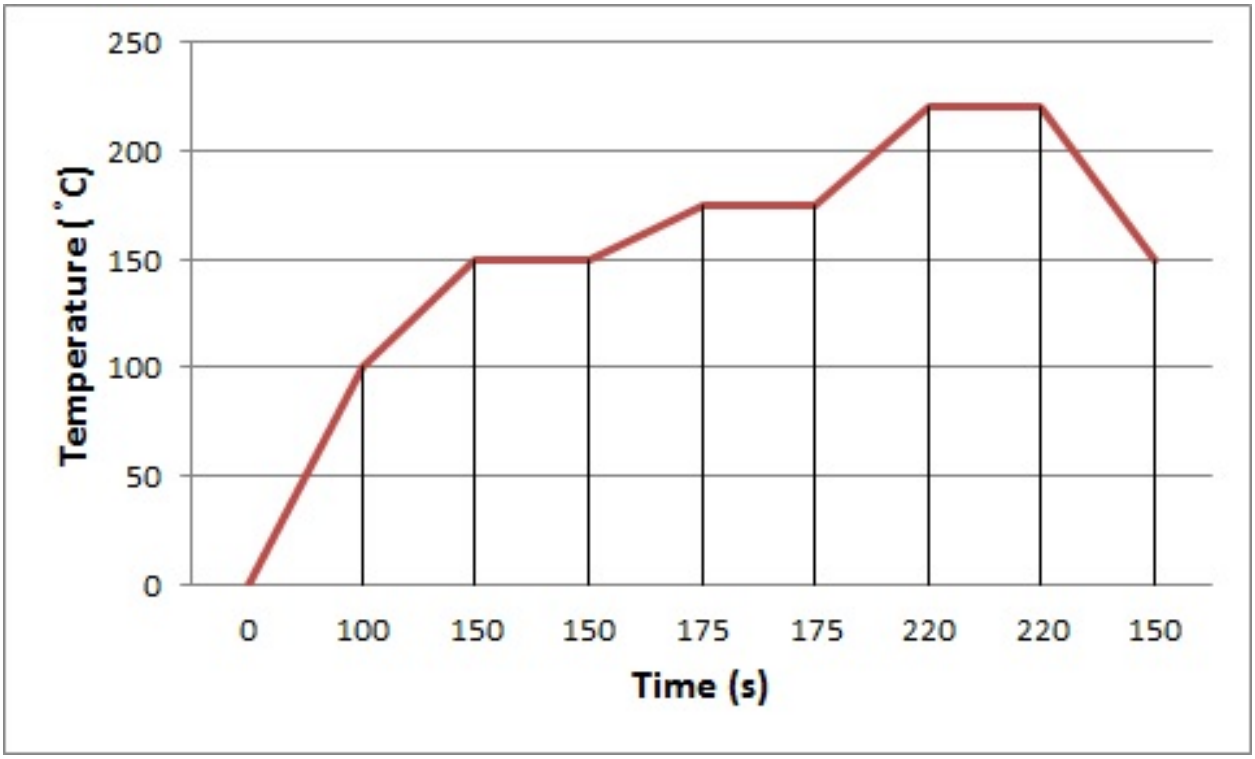

Figure 5.4: A modified temperature profile to reduce the warping during solder reflow of the BGA package.

determined by the technology used to fabricate the chosen FPGA. If more LUTs are required to implement the logic function than are available within a single slice, an additional delay will be incurred for routing signals through the switched interconnect matrix to adjacent slices. Additionally, a more significant performance penalty will result if the logic function requires two or more stages of LUTs to compute. In order to achieve the maximum possible performance with the FPGA, the design must be pipelined such that the worst case delay is limited to a single LUT stage delay and only the switched interconnect delay between the previous register and the LUT input. However, if a large number of slices are utilized (>50%), the interconnect matrix can become a significant source of delay as connected LUTs will more likely be physically separated by a greater distance on the FPGA. In addition to delay considerations, further design specific performance considerations will be discussed in the following sections.

As mentioned in section 4.1.5, the imaging IC has four LVDS output banks that

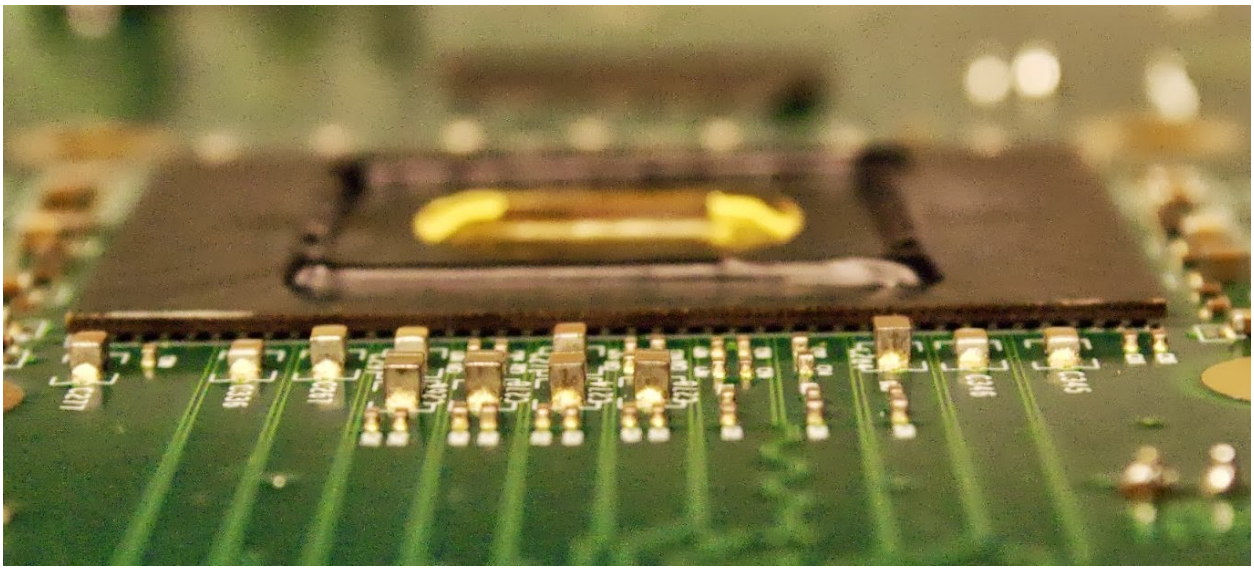

Figure 5.5: By performing the reflow soldering process with the modified temperature profile of Figure 5.4, the package warping is reduced such that all pads on the BGA are successfully soldered to the board.

transmit a clock and 21 data bits using an TIA/EIA-644-A standard compliant LVDS output buffer. Each of the four output banks on the IC connects to a separate FPGA, which minimizes the data handling requirement for each device and makes signal routing on the PCB easier. In this design, Xilinx Virtex-6 FPGAs are used to receive the data. A relatively wide range of Virtex-6 devices are available that differ in their maximum operating frequency, number of I/O, available look-up tables, registers, amount of block RAM, and high-speed transceivers. The device chosen for this design was a XC6VLX130T-3 and its characteristics are listed in table 5.1. This was the minimum size device that could meet the architecture requirements and timing constraints of this system, which will be discussed in more detail below.

5.3.2 FPGA Architecture

The primary purpose of the FPGAs is to capture the raw photon arrival time data from the imaging IC and transmit this data to a computer for analysis. Because the IC outputs up to 42 Gbps of data, which is beyond the data transfer limits of a typical computer, the FPGAs

Table 5.1: Detailed characteristics of the Xilinx Virtex-6 XC6VLX130T device used in this work. A speed grade -3 (fastest version available) device was required to meet the timing performance of the full design, which is reflected in timing parameters listed below. Each slice contains four 6-input, 2-output LUTs and 8 flip-flops in Virtex-6 devices. Additionally, the total available block RAM is divided into 18 kb blocks that are distributed throughout the FPGA.

Characteristic	Value	Units
Slices	20,000	
LUTs	80,000	
Flip-Flops	160,000	
Block RAM	9,504 528	kb 18 kb Blocks
Available I/O	240	Pins
Maximum Clock Frequency	800	MHz
PCIe Transceivers	2	

must perform some processing on the data to reduce the overall data bandwidth requirement. The FPGAs form 128 bin histograms of the arrival times for each pixel in the array, which reduces the overall data requirement to only 8 Mb per frame. For the designed maximum frame rate of 100 frames per second, this results in a data transfer rate requirement of 800 Mbps. This datarate can easily be handled using a PCIe 1.0 interface [131] between the Virtex-6 devices and a computer. The following paragraphs provide details of the FPGA architecture and blocks used to achieve these operations.

FPGA Data Input Interface

A block-level description of the FPGA architecture is presented in Figure 5.6. The data from the imaging IC is received by the Virtex-6 FPGAs using LVDS primitives available on the FPGA. The 21 differential data bits are initially captured in dedicated registers in the I/O blocks. The differential clock input is used to clock these registers and is buffered to drive a mixed-mode clock manager (MMCM) block, which then generates additional clock frequencies for use within the FPGA. Specifically, the MMCM replicates the incoming clock and then generates a half-frequency clock. The outputs from the I/O registers are directed

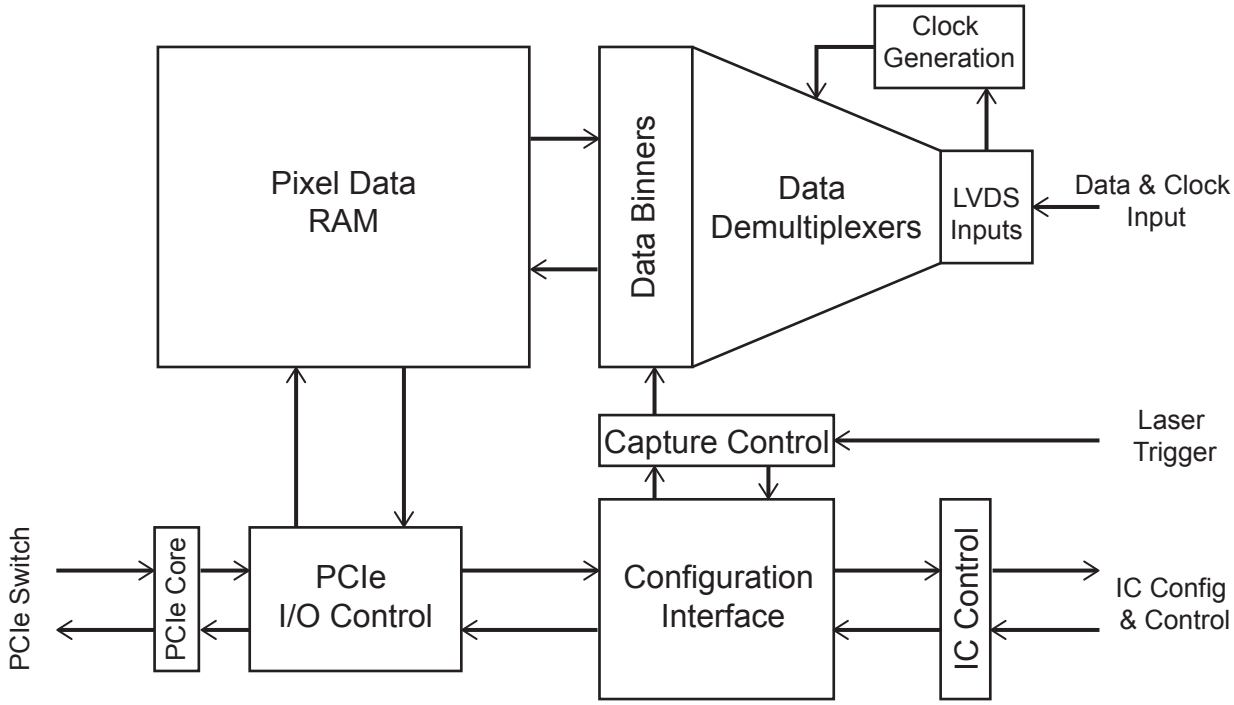

Figure 5.6: A block diagram showing the main features of the FPGA design. Data from the imaging IC enters from the right side and passes through the data demultiplexers before being binned into per-pixel histograms that are stored in the pixel data RAM. The left side interface is with the computer over a cabled PCIe interface. All configuration and control signals are set using the PCIe interface and then passed to the appropriate block within the FPGA or to the IC configuration and control interface.

into a datapath that sorts the incoming data by pixel. After sorting the data, a histogram of photon arrival times is generated for each pixel. In addition, the FPGA serves as the main control interface between the user interface and the imaging IC.

Pixel Data Sorting

The incoming photon data consists of 20 bits of position and timing information along with a single valid bit and the interface clock. The order of pixel data is not known *a priori* due to the statistical nature of whether a pixel will record a photon and the on-chip method for discarding data for pixels in which no hits occurred, as described in Section 4.1.4. As such, the FPGA datapath must process each incoming data word and determine which pixel the

timing information belongs to. Further, this sorting process must be executed at the same speed as the maximum I/O clock from the IC, which is 500 MHz.

In order to sort and bin the data under these constraints, the sorting process is divided into a number of sequential demultiplexing stages such that each stage can be completed using no more than a single FPGA slice per bit. This constraint ensures that the interconnect delay through the FPGA switch matrix is minimized. Since a slice contains four 6-input, 2-output LUTs, the logic synthesis and mapping tool can pack two bits into a single slice when it is beneficial for timing closure. Each demultiplexing stage consists of one or more 1-to-4 word demultiplexers that use the two current MSBs of the data to route the input word to the appropriate output. In this design, there are 1024 unique pixels processed per FPGA and the complete demultiplexing process is accomplished with 5 stages of 1-to-4 demultiplexers. The final stage consists of 256 demultiplexers, as seen in Figure 5.7.

Because of the FIFO controller bug that was mentioned in Section 4.2.8, duplicate copies of the same data will be presented to the FPGA with valid bits asserted whenever there is not enough unique data to fully utilize the output capacity of the imaging IC. As a result of this bug, it is necessary to detect duplicate data on the FPGA in order to avoid counting the same data multiple times. A simple duplicate data detector is implemented in the second demultiplexing stage of the pixel data sorting block. A flow diagram of the duplicate elimination scheme is shown in Figure 5.8. This duplicate removal block was placed at the second demultiplexing stage because of the well defined pattern in the initial 2 MSBs. Since the FIFO coordinator block (Section 4.1.4) uses a round-robin scheme for reading from each of the four FIFOs on the IC and the two MSBs indicate from which of the four FIFOs the data originated, these MSBs will follow a pattern of 00, 01, 10, 11, 00, 01, ... Consequently, each of the four inputs to the second demultiplexing stage will only change once every four clock cycles. This allows for removal of the duplicate data during the extra clock cycles. Additionally, since the duplicate data is a result of a bug in the FIFO coordinator block, which generates the two MSBs, this is also the earliest point at which

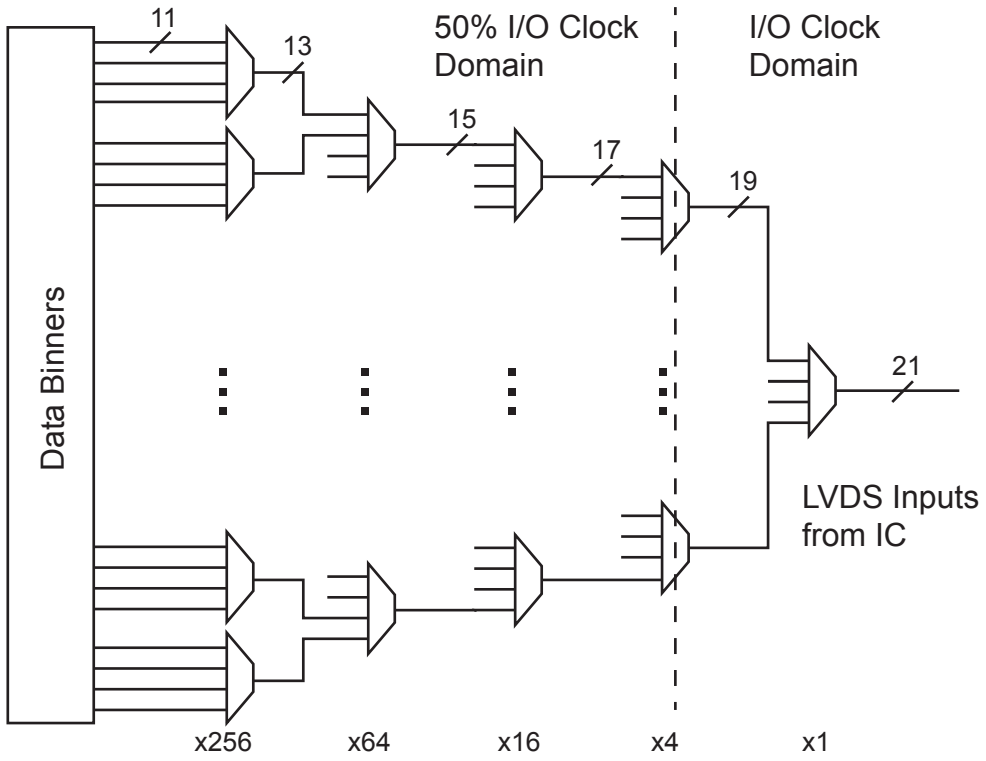

Figure 5.7: A diagram showing the structure of the data demultiplexers used to sort the incoming data by pixel on the FPGAs. Five stages of 1-to-4 demultiplexers are used to separate the incoming data before it reaches the data binning block. The vertical dashed line shows the dividing point between the full I/O clock speed domain and the half-rate I/O clock domain.

duplicate data can be detected using the method presented here.

Because none of the demultiplexer inputs from this stage forward will change more frequently than once every four clock cycles, the data is transitioned to the half-rate clock domain generated by the MMCM discussed above. This relaxes the timing requirements on the FPGA design and makes timing closure for the design easier despite operating with an input frequency of 500 MHz and a total logic utilization over 50%. Additionally, the lower clock rate will reduce the dynamic power consumption for that portion of the design by nearly a factor of two.

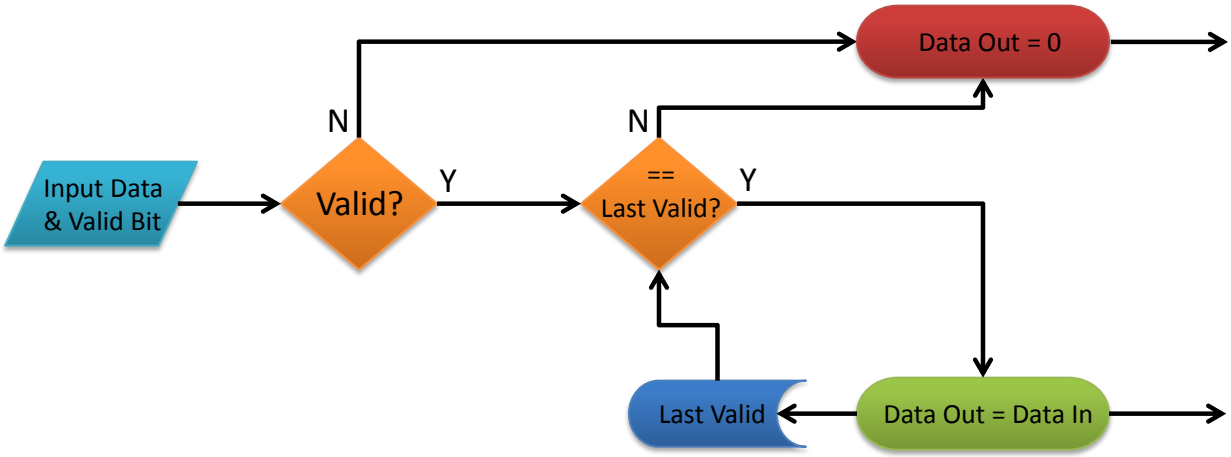

Figure 5.8: The duplicate data detection block is part of the second demultiplexing stage. Its purpose is to eliminate duplicate copies of data that were presented as valid due to a bug in the FIFO coordinator controller on the IC. This flowchart shows the decision making process for the duplicate detection. The block first checks for a valid bit assertion and then compares the incoming data to the last valid data. If the data matches, the output is set to all 0's and the duplicate data is prevented from propagating past this point. If the data is different from the last valid data, then the last valid data register is updated and the input data is passed to the output. The demultiplexing operation that is also performed at this stage is not included in this diagram.

Per-Pixel Histogram Generation

Once the data has passed through the datapath, it will be sorted by pixel position and will consist of only 10 bits of timing information and 1 valid bit. A FSM is used to check if the current data is valid and, if so, increment the appropriate timing bin in the histogram. Because each pixel can have no more than one hit per laser repetition, there will be at least 12 clock cycles between unique timing data for any pixel in design (20 MHz laser repetition rate and a 500 MHz input clock that is divided by 2, as mentioned in the above paragraph). As such, two adjacent pixels can share a single FSM and the FSM is required to process both of them within 12 clock cycles.

A diagram showing the operation of the histogram generating FSM is presented in Figure 5.9. Before reaching the cyclic part of the FSM, incoming data passes through a bin width selection block. This block is globally configured through software and can be

adjusted to select any 7 consecutive bits from the incoming data. These 7 bits are used for a portion of the bin address in the FSM and correspond to bin widths between 62.5 ps (lowest 7 TDC bits) and 500 ps (highest 7 TDC bits). The RAM is split into four address regions such that each dual-ported RAM holds a histogram for two pixels in consecutive frames (four histograms per block RAM in total). The remaining address bits are the frame (either 0 or 1) and whether the pixel is even (0) or odd (1). The 18kb RAM is configured with 16-bit words and is 512 words deep. 1n Figure 5.6, the two ports of the pixel data RAM can be seen with the bottom port connected to the PCIe I/O controller for transmitting data to the computer while the right port is simultaneously recording histogram data. This structure allows for continuous simultaneous recording and reading of the FLIM histogram data for each pixel while ensuring that no address is accessed from both ports at the same time, causing a read/write collision.

Image Capture Controller

A programmable image capture controller is designed to enable precise timing of the image acquisition. This controller receives a programmable number of laser repetitions that it uses to coordinate frame capture. In addition, it receives status signals from the PCIe output controller and resets the pixel histograms after they have been read. The capture process is started by a trigger signal originating from the controlling computer, which allows the user to control image acquisition through software.

PCIe Interface

Data transfer from the block RAM, where the pixel histogram data are stored, to the computer for further processing occurs over a cabled PCIe interface. An overview of the PCIe topology is shown in Figure 5.10. The Intel Xeon E31225 microprocessor has a integrated memory controller and its associated platform controller hub (PCH) can support up to 8 PCIe root complexes, with each capable of operating a single lane (x1) second generation

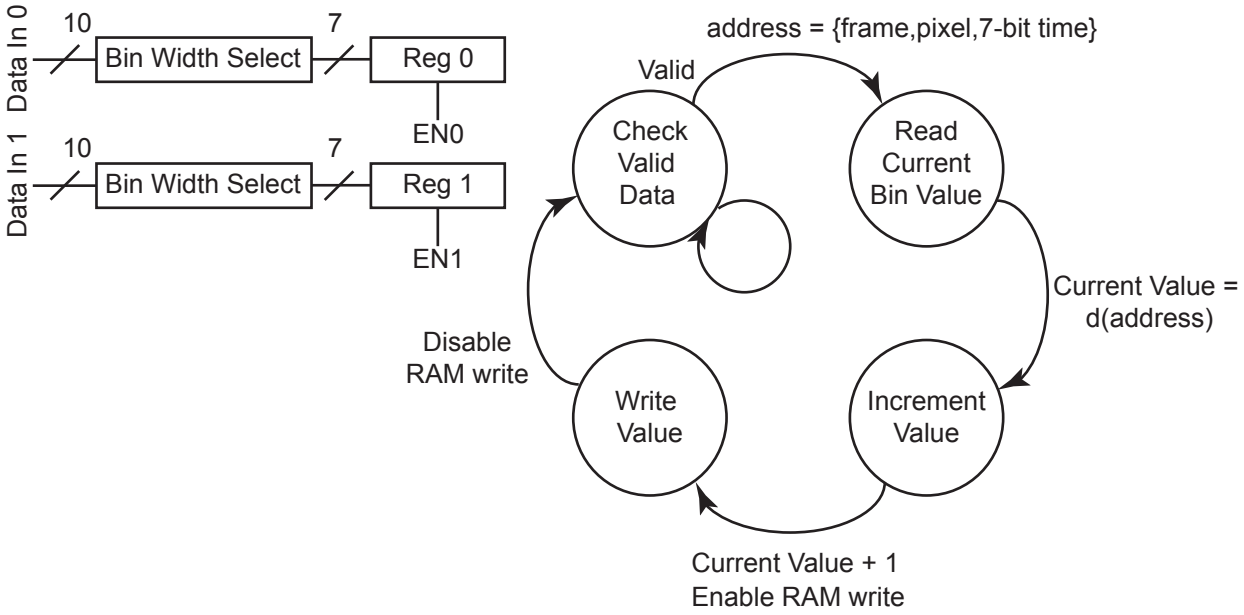

Figure 5.9: The finite state machine for the data binner is outlined. The FSM initializes to the check valid data state where it alternates between checking register 0 and 1 for valid data. If valid input data is found in either register, a cascade of events occurs whereby an address to the pixel RAM is generated, the bin value at that address is read, the value is incremented by one, and the new value is written back to the address. The process takes only four clock cycles. The enable signals after the bin width select block allow incoming valid data to be locked if the FSM is in the middle of a cycle when valid data arrives.

PCIe link having a 5 Gbps bandwidth. In this case, four of these root complexes are combined to make a four lane (x4) PCIe link between the PCH and the CPU. The PCH allows for direct memory access (DMA) operations whereby PCIe endpoints can directly access system memory without requiring intervention from the CPU. The other port of the root complex in the PCH is connected to the PCIe switch, which aggregates the data from the four x1 PCIe endpoints on the FPGAs into a single x4 link for the computer.

The Xilinx PCIe endpoints are implemented using a PCIe endpoint core for Virtex-6 FPGAs available from Xilinx. This core handles the physical layer (interface configuration, speed & width negotiation, 8b/10b encoding [132]), data link layer (error checking and packet retransmission), and provides ports for the transaction layer (header and data). A diagram showing the different layers to the PCIe protocol is reproduced from the PCIe Base

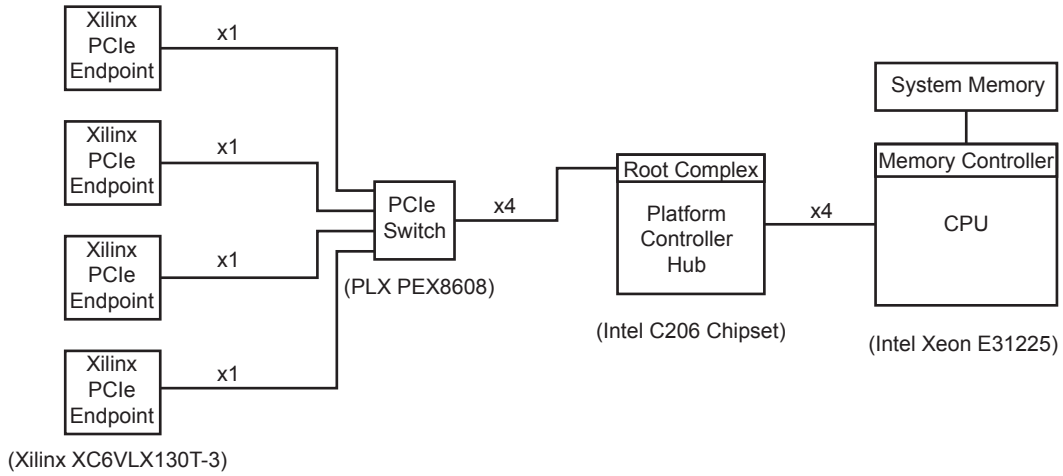

Figure 5.10: A diagram showing the topology of the PCIe system and the specific components used.

Specification [131] in Figure 5.11. The user must design the controllers to support sending and receiving packets through the transaction layer interface port of the Xilinx PCIe core. This includes responding to requests from the root complex, supplying packet headers and data payloads for transmission, receiving data packets, and initiating interrupt signals. A diagram showing the HDL blocks on the FPGAs that are responsible for accomplishing these tasks is presented in Figure 5.12.

RAM-to-FIFO Controller and Pixel RAM Reader The first step in preparing the data for transmission over the PCIe interface is reading the histogram for each pixel of the previously acquired frame and arranging them into 32-bit double words (DWs). When the user defined number of laser repetitions per frame have been completed, the capture controller initiates a readout sequence in the RAM-to-FIFO and DMA transmission control blocks. The pixel RAM reader module receives a pixel number and bin number from the RAM-to-FIFO module and responds with the corresponding 16-bit bin value for the addressed pixel and frame that was previously acquired. An early implementation of this block attached the pixel number and bin number to each 16-bit bin value but the final version omits this information since a strict data ordering scheme is used in the readout. This omission

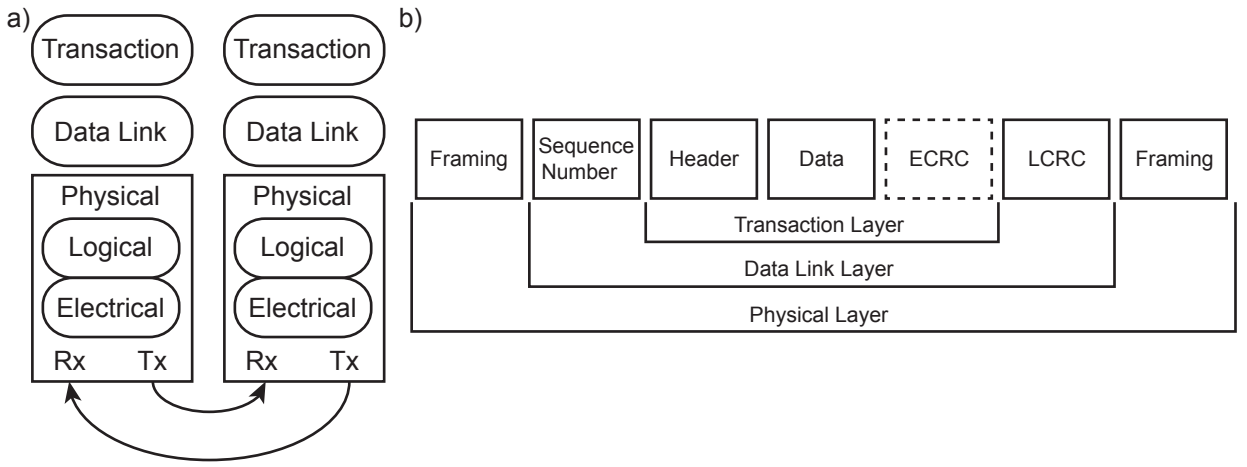

Figure 5.11: a) A block-level perspective of the PCIe interface layers. b) The pieces of a PCIe packet and the layer within the protocol that processes each of them. With the Xilinx PCIe core, the designer is responsible for designing the controller that processes the transaction layer headers. In this design, no error checking is performed and the ECRC portion of the header is skipped. These diagrams were reproduced from the PCI-SIG standard [131]

allows for the minimum transmission of data needed to reconstruct the histogram at each pixel after receipt by the computer, which amounts to 8Mb per frame. The RAM-to-FIFO module coordinates closely with the DMA transmission control block to place packets with address and data information into the output FIFO for transmission over the PCIe interface.

DMA Transmission Controller The DMA transmission controller generates transaction layer headers (Figure 5.11) for the write operation and organizes the 32-bit DWs into packets of DWs. The packet size is determined by the lowest maximum payload size (MPS) supported by all of the PCIe devices in the datapath (See Figure 5.10 for the PCIe topology). Table 5.2 shows each component in the PCIe datapath and its corresponding MPS. In this work, the MPS is limited to 32 DWs (128 bytes) by the Intel chipset and processor used in the receiving desktop computer.

The DMA transmission controller organizes a 3 DW packet header along with the 32 DW payload into eighteen 64-bit quad words (QWs) that matches the transaction interface input bus width for the Xilinx PCIe endpoint and pushes these QWs into the output FIFO.

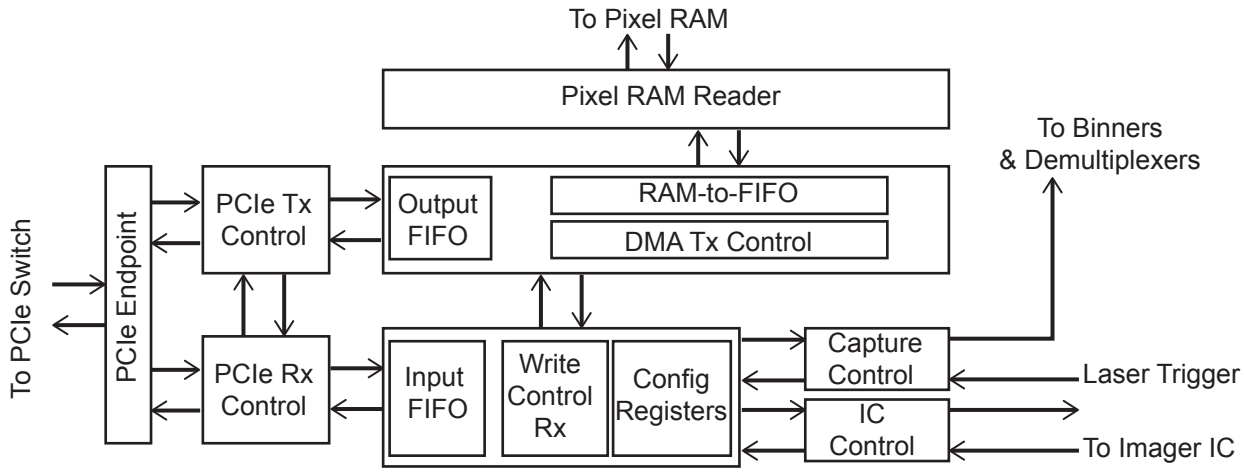

Figure 5.12: This diagram shows the sub-blocks that comprise the PCIe I/O controller. The configuration interface from Figure 5.6 is integrated into the Write Control Rx and Configuration Registers. The capture control and IC control blocks from Figure 5.6 are reproduced here to provide context.

If the PCIe transmission interface is idle, the PCIe transmission controller will start to send data to the computer when the FIFO has been filled with at least 50% of a complete packet (9 QW, including headers and data). The DMA controller will continue to fill the output FIFO until the entire frame of histograms has been read from the Pixel Data RAM. If the FIFO becomes full, then the DMA controller will pause until data can be written over the PCIe interface and additional space is freed in the FIFO. Because of the overhead required for packet transmission, the output FIFO will always fill faster than it can be emptied by the PCIe transmission controller. Due to the number of block RAM units required for all pixels in this design, nearly all of the available block RAMs are allocated for pixel histograms and only eight 18kb block RAMs could be used for the output FIFO. Consequently, the depth of the FIFO was limited to only 2048 words. While this is sufficient for maintaining system performance for 100 frames per second, further optimization could come from increasing this buffer size on a larger FPGA.

An entire image worth of data consists of 65536 DWs requiring 2048 PCIe packets to be transmitted with a total of 36864 writes to the output FIFO. Once the frame has been completely pushed into the FIFO and the PCIe transmission controller has read all of the

Table 5.2: A table of the maximum payload size supported by each PCIe device in the dataptah.

Device	Maximum Payload Size (DWs)
Xilinx PCIe Endpoint	256
PLX PEX8608	512
Intel C206 Chipset	32
Intel Xeon E31225	32

words from the FIFO, the frame transmission is complete and the controllers idle until the next readout initiation signal is received from the capture controller.

Configuration Interface

In an effort to minimize interfaces between the computer and imaging system, all controls that were handled by the Opal Kelly device used on the PCB in Section 4.2 have been migrated to the same Virtex-6 FPGAs that are used to capture data from the IC. This allows for complete control of the imaging system over a single cabled PCIe interface. This control is accomplished by allocating a unique base address register (BAR) for each of the four PCIe endpoints such that the computer recognizes the devices as regions of memory within its system. By writing to specially defined memory regions on the PCIe endpoints, commands (e.g. starting an image capture) or configuration data (e.g. IC scan chain bits) can be sent to the FPGAs. Configuration registers on the FPGA store the bits for programming the scan chain for the PLL and the imaging array. In addition, they hold values for the programmable number of laser repetitions per frame, the bin width to use in histogramming, and the fine data offset (Section 4.2.5). Additional control signals sent over the PCIe interface allow for measuring the PLL frequency, resetting the DLLs and FIFOs, initiating configuration scans, and controlling data acquisition.

Although all four FPGAs were designed with identical system architectures, only the FPGAs corresponding to the top and bottom sides of the IC were used for control and configuration since these were closest to the I/O connections for the control interface of the

imager IC.

FPGA Raw Data Acquisition

For verification and calibration, it is sometimes beneficial to have the ability to directly record the raw binary data coming from the imaging IC. In particular, the TDC calibration presented in section 4.2.5 required access to raw binary data. In order to accommodate this requirement, an FPGA implementation was designed to capture the raw data stream from the FLIM IC and store it in the FPGA's block RAM. This design uses the same LVDS input blocks as the FLIM histogramming HDL, but instead of sorting the data and generating per-pixel histograms, the raw data is written directly to a FIFO that is 65536 words deep. A specialized FSM then transfers the data recorded in this FIFO to the PCIe output FIFO where it then undergoes the same handling as the image data, discussed above. This allows for high-speed acquisition and transmission of the calibration data so that TDC calibration information can be acquired quickly. A block-level overview of the raw data acquisition system is presented in figure 5.13.

5.4 Printed Circuit Board

In Section 4.2.5, the test and debug PCB was presented. This PCB used two of the Virtex-6 FPGAs for data acquisition and a separate Opal Kelly XEM6010 USB-to-FPGA daughter card for controlling the system. It also was designed to only utilize half of the imaging array and image acquisition time was not a design consideration. As a result of these loose requirements, power consumption was not a constraint on the system and linear regulators could be easily used. The complete test and debug system consumed just over 23W from a 5.5V supply.

The PCB for the final system had more demanding constraints. It needed to support four FPGAs, the PCIe switch, the full IC, and several auxiliary circuits. This system required

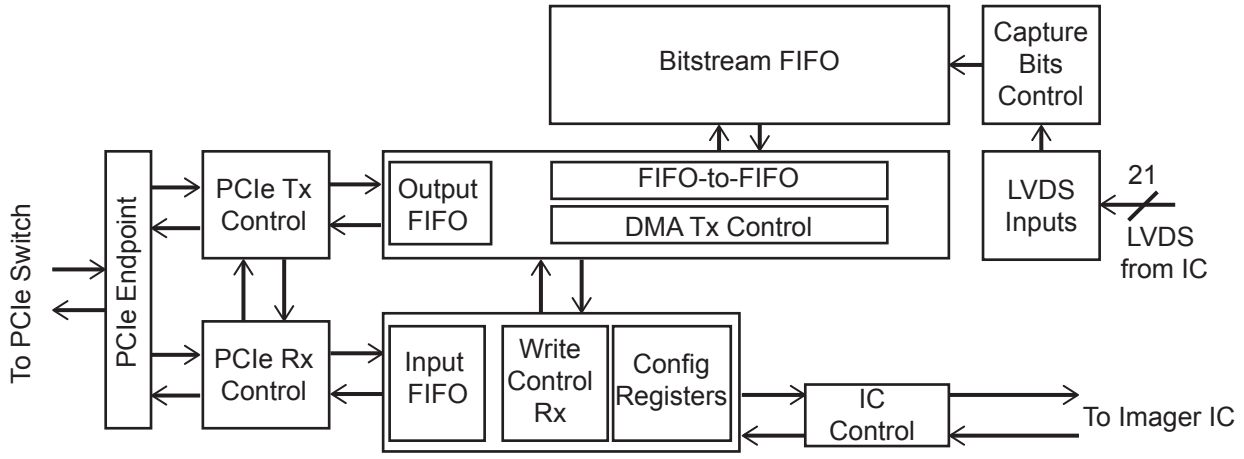

Figure 5.13: The block diagram for the raw data acquisition system is shown. A 21-bit data stream (plus the clock) enters the FPGA through the LVDS inputs. This data is immediately passed to a block that controls the capture of the bitstream. When the bitstream FIFO is filled, a modified output controller (FIFO-to-FIFO) works in concert with the same DMA transmission controller that is used in the imaging system. The remainder of this system is identical to the imaging system.

careful matching of traces and routing delays across the board in addition to an efficient power conversion and distribution network. A photograph of the system PCB is shown in Figure 5.14.

5.4.1 Power Conversion and Distribution

The maximum power estimate for the entire system running at full speed with complete utilization of the FPGAs was 150W. Consequently, a 24V, 7.2A main power supply was chosen to power the system. High efficiency switching regulators were designed to minimize power dissipation in the conversion from the 24V main supply to the supplies needed for each chip in the system, which ranged from -15V to 3.3V. In total 10 different voltage levels needed to be generated from the 24V supply and, because some of these supplies required extra filtering for analog circuits or tunability, a total of 20 regulators were designed for this PCB. While imaging at 100 fps, the entire system consumed only 26.4W from a metered 20V supply. The actual power consumption is much less than the budgeted 150W due to

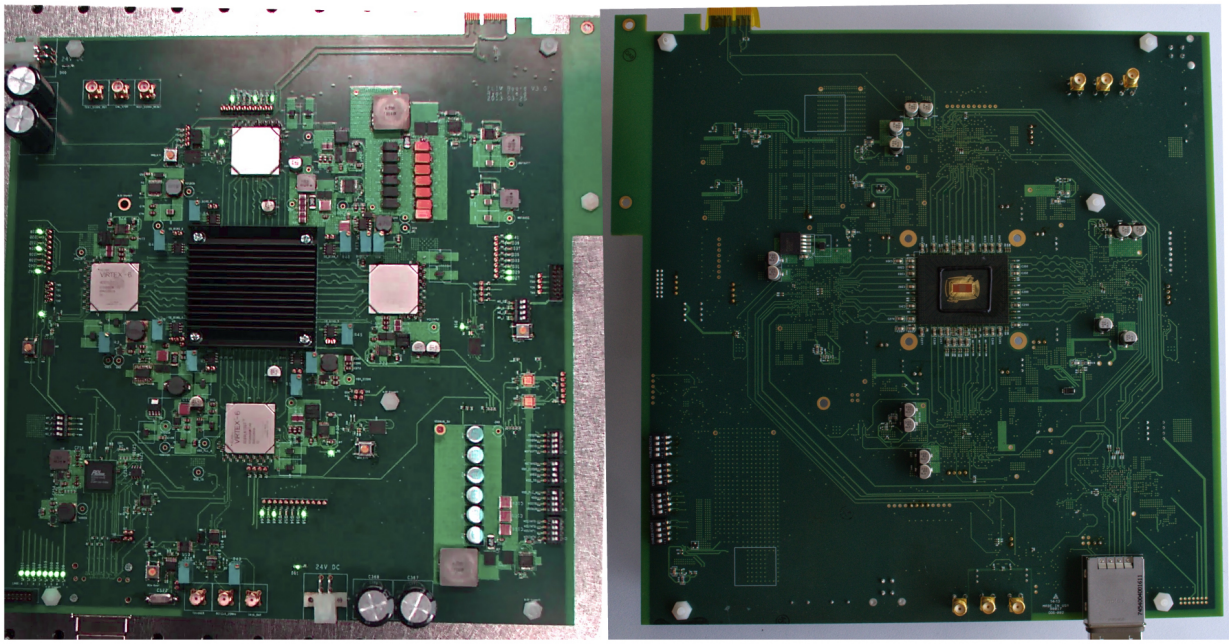

Figure 5.14: A photograph of the final system PCB. The left side shows the top of the PCB and the right side shows the bottom. The imager IC is mounted in the middle of the bottom of the PCB. The cabled PCIe connector is visible in the bottom right corner. Each of the four FPGAs are visible in the top view.

partial utilization of the available FPGA resources.

5.4.2 Auxiliary Circuits

In addition to the main imager IC, FPGAs, and PCIe switch that were previously discussed, there are a few auxiliary circuits on the PCB that enable the imaging system. The first is a high-speed comparator circuit for amplifying the laser trigger signal. The signal that is output from the laser must be 50Ω terminated and has an amplitude of 100mV and pulse width of only 5ns. The comparator circuit increases the amplitude of this pulse to 3.3V and then distributes it to each of the four FPGAs and the imaging IC using a 1:6 clock buffer chip. The 6th output is connected to an SMA connector for monitoring.

The PCIe interface requires that a common 100 MHz differential reference clock be distributed to all PCIe devices in a system. This reference clock is transmitted from the computer over the PCIe cable and is buffered to each of the FPGAs and the PCIe switch.

A 2:5 LVDS clock buffer is used to distribute this clock and proper termination is included to meet the signaling requirements of each device (LVDS for FPGAs, CML for PEX8608). Additionally, the second buffer input is used as a backup whereby a 100 MHz clock is generated using a 25 MHz crystal reference and a 4x clock generator. This capability was added to allow debugging of the PCIe switch in case of challenges bringing up the PCIe switch interface. The I2C interface of the PEX8608 cannot be accessed until a reset signal has occurred while the reference clock is active. Similarly, a manual reset button (active low) was added to the PCIe interface and combined with the reset signal from the PCIe cable using an AND2 gate. The final PCIe related auxiliary circuit was a complete backup interface. A card-edge connector was designed onto the PCB such that it could be plugged into any x1 PCIe header of a motherboard and would allow access to one of the FPGAs. This interface was never tested since the cabled interface worked successfully.

A complete JTAG chain linking together the PEX8608 and the four FPGAs was included. However, the PEX8608 JTAG interface prevented the Xilinx Impact programming software from accessing the FPGAs. Zero-ohm resistors and headers were included in the chain allowing the devices to be separated and manually re-chained. Ultimately, the four FPGAs were manually connected into a JTAG chain of their own, which worked flawlessly.

The final auxiliary circuit is a power supply sequencing circuit. This circuit was designed to automatically turn on each of the 20 voltage supplies sequentially. This circuit included manual override switches and was never tested since it was not essential to the principle mode of operation of the system.

5.4.3 Liquid Cooling

The imager draws a total of 8.79 W when running at full-speed and is liquid-cooled to avoid degraded performance of the SPADs due to heating [133]. To achieve this cooling, a custom BGA package with a copper core is directly soldered to the PCB, which also has a copper core. A copper heat exchanger is placed in tight contact with the copper region of the PCB

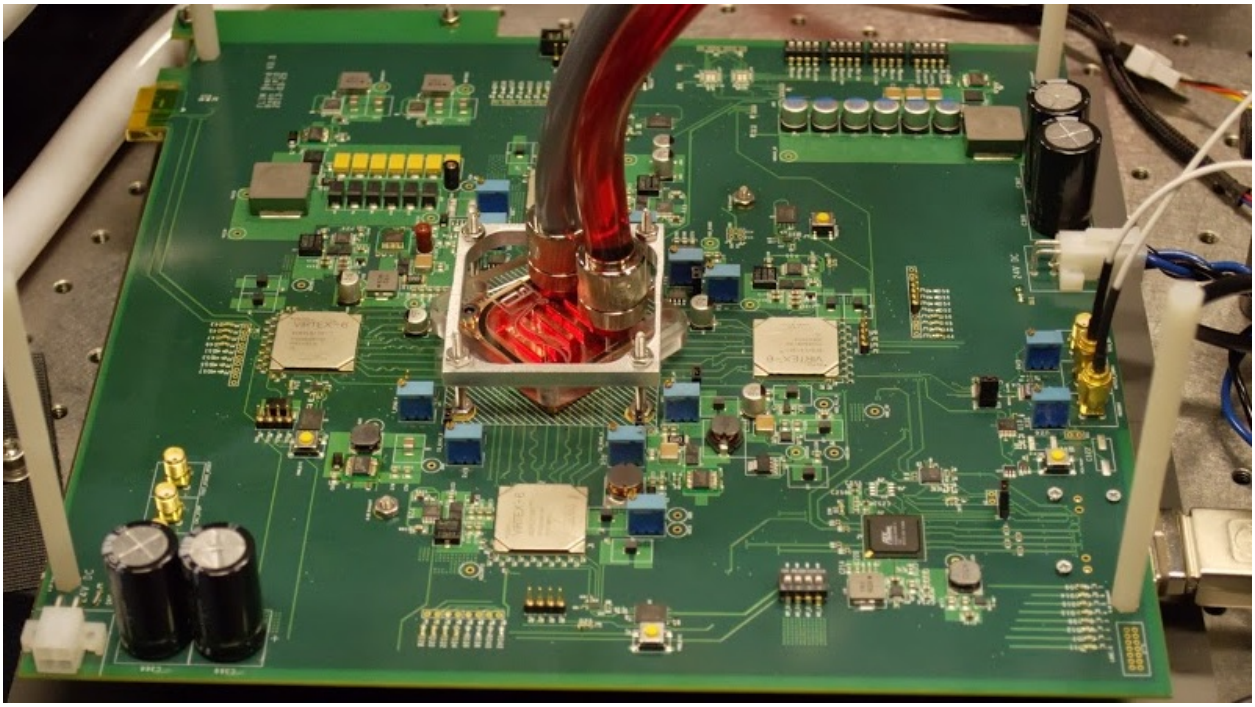

Figure 5.15: A photograph of the final system with the liquid cooling system attached. The copper water block for cooling the imager IC from the back side.

and a coolant is pumped through the heat exchanger. The liquid cooling system was chosen over a fan-based system to minimize vibrations during imaging. The average dark count rate for all pixels in the array with cooling is 544 Hz, which is an improvement from 1036 Hz without cooling. Figure 5.16 shows a plot of the average DCR throughout the array versus time. In this measurement, the full array was enabled with all TDCs running. Initially, no cooling was used and after six minutes the cooling system is turned on. From this plot, it is clear that the cooling system has a significant effect on the DCR and that the DCR is stable over time with cooling. The overvoltage, V_{ov} for this measurement is 2.5V. A photograph of the cooling system attached to the PCB is shown in Figure 5.15.

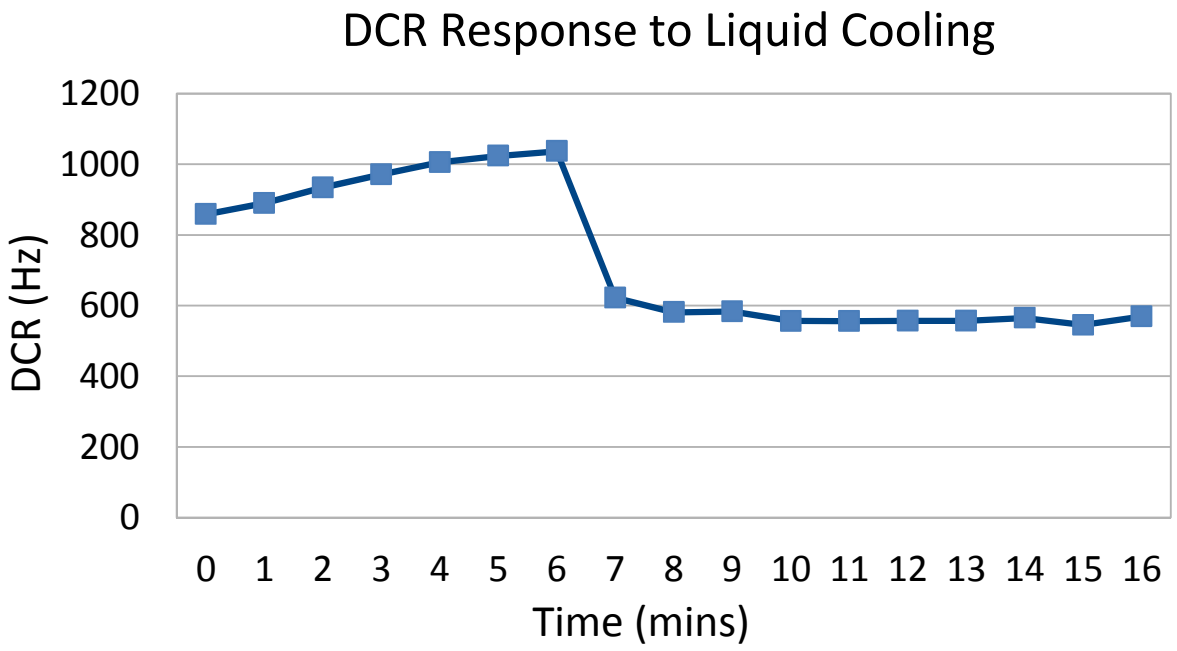

Figure 5.16: A plot of the average pixel DCR versus time while all features of the imager IC are running. Initially, the cooling system is turned off and the DCR begins to rise with the temperature. At t=6 minutes the cooling system is turned on and the DCR immediately drops and remains stable around 600 Hz for a V_{ov} of 2.5V.

5.5 Software

In order to control the complete system, the final essential component is software for communicating with the device, recording data, and processing the lifetime. Each of these topics is covered in this section, beginning with the device driver.

5.5.1 Linux Kernel Module & Device Driver

A Linux kernel module is a piece of software that can dynamically be added to the Linux kernel. The Linux kernel is a low-level program that handles communication between application software, the CPU, memory, and peripherals. All of the work in this thesis was designed for use with the Ubuntu Linux distribution, specifically version 12.04, which is a long term support release. Since the FLIM imaging system has a custom hardware interface

with the computer, it is necessary to design a module that will allow for software running in the operating system (OS) on the computer to communicate with the imaging system and to handle exchanges between the system, memory, and the CPU. This system requires only a subset of the available operations that can be performed between a device and the kernel. At a minimum, the kernel must allow read and write operations to the imaging system. This is enabled within the kernel by mapping each of the Virtex-6 PCIe endpoints to virtual memory address regions within the computer. All configuration and control operations can then be carried out using memory write actions to the FPGAs and status monitoring can be done through memory read actions. The following discussion is directed at the design choices made in building a kernel module for this system and not the specifics of how to write a kernel driver. For details on writing kernel drivers, see reference [134].

Device Mapping When a computer boots, the BIOS initializes the PCIe devices and the kernel maps the memory regions of the device into its memory address space [134]. During the boot process, all of the PCIe interfaces undergo a negotiation process during which settings like the maximum payload size, link width, and data transfer rate are determined. These settings are written to the endpoints' configuration registers and also recorded in the kernel. Also stored in the configuration space are read-only properties like the vendor and device identification values, which can later be read by the computer to identify specific devices. Each of the PCIe endpoints on the Virtex-6 FPGAs is configured to have one base address register (BAR), which is configured with the starting address for a virtual memory region that is allocated to it during boot time. The endpoint requests a certain size virtual memory space within which it will be able to process I/O operations to its addresses. This virtual memory address space for the Virtex-6 PCIe endpoints was chosen to have a size of 16 MB, which provides enough memory addresses for configuration registers and control signals in addition to providing the flexibility to read all of the devices' block RAM (an operation that would only be necessary during debugging). Every command and configuration register

is accessed by the computer through combining a fixed offset with the address stored in the BAR. A table of a subset of the offset values for the operations supported by the driver and FPGA HDL are presented in Table 5.3. Control signals are mapped to individual bits within a configuration register.

After the computer boots and the memory addresses are allocated, the kernel module is loaded. During the loading process, the kernel reads the vendor and device identification values from the configuration space for each of the PCIe devices. Whenever one of the imaging system devices is recognized, a data structure in the driver is filled with important information (BAR address, IRQ address, capabilities, etc.). Most importantly, the kernel allocates a 4 MByte block of contiguous low memory for each FPGA. This block of memory will be used by the FPGA when it is ready to transfer data to the computer.

Driver I/O Control Once the kernel maps the device to a memory region, it is available for direct read and write access through the driver. Such a device is called a memory-mapped device and I/O to the device can be handled through kernel headers that provide 'ioread' and 'iowrite' functions. These functions allow for seamlessly handling any pointer dereferencing between the I/O memory space of the PCIe endpoints and CPU in order to perform read or write functions. All data transfers to or from the Virtex-6 devices occur through these two functions.

Direct Memory Access The process used to transfer data is a direct memory access (DMA) routine whereby the CPU allocates dedicated memory for each FPGA and the FPGA is free to modify this memory without the CPU's involvement. Ideally, the FPGA would send an interrupt signal to the CPU to notify it when the data transfer is complete and it is safe for the CPU to retrieve the data from memory and further process it. In order to avoid the complexities of interrupt handling, this work uses a simpler approach for DMA data transfer, with the consequence that only 16 consecutive frames can be captured in the memory space that is allocated for each FPGA. Increasing the size of the memory allocation

Table 5.3: A table listing a subset of the BAR offsets and corresponding command or configuration register. Each of these are DW aligned byte address offsets and the two least significant 0's are omitted from the table. For instance, 0xC0001 is actually the 32-bit address offset 0000 0000 0011 0000 0000 0000 0000 0100

Register	**BAR Offset**	**Bit #**	**Purpose**
Capture Start	0xC0001	4	Starts capture of set of frames
DMA Start Address	0xC0003	0-31	System memory address for writing output data
Interface Reset	0xC0004	0	Resets PCIe interface controllers
FPGA Clock Enable	0xC0004	1	Enables clock distribution on FPGA
Bin Resolution	0xC0004	2-3	Sets bin width of histogram
Reset RAM Side	0xC0004	4-5	Resets the RAM for frame 0 (bit 4) or frame 1 (bit 5)
PCIe Write Timer	0xC0005	0-31	Stores number of clock cycles to complete PCIe write
LED Outputs	0xC0006	0-7	Control LEDs for debug/status indicators
Fine TDC Offset	0xC0008	0-3	Programmable fine TDC value offset
Frame Count	0xC0009	0-31	Programmable number of laser triggers per frame
PLL Configuration	0xC000A	0-30	Configuration values and control for PLL scan chain
FLIM Configuration	0xC000B	0-31	Control for FLIM Scan Chain
FLIM Config Data	0xC04xx	0-15	Write FLIM config data to RAM using lower 10 bits for addressing
PLL Clock Speed	0xC000C	0-31	Output from frequency counter measuring 1/256th of PLL frequency
Number of Frames	0xC000D	0-31	Number of consecutive frames to capture following a capture start signal

for each FPGA cannot be done with the current approach due to limitations on the low memory accessible by the kernel. Instead of using interrupts, the image recording and data transfer process follows these steps:

1. The CPU sends the capture start signal to all of the FPGAs using the dedicated memory address listed in Table 5.3.

2. The CPU sleeps for an amount of time proportional to the number of frames recorded to allow the capture to finish (1 to 16 consecutive frames recorded).

3. The CPU reads the memory where the four FPGAs wrote the image data.

4. The image histogram data is written to disk and saved for additional processing.

With updates to the kernel module and some changes to the FPGA HDL to leverage interrupts, this system should be able to record an arbitrary number of frames at full imaging speed.

5.5.2 C Application Programming Interface

The discussion leading to this point has focused on the low-level hardware and kernel driver design for the FLIM imaging system. In order to build user-friendly program that can interact with the FPGAs to control the system and record data, a programming interface must be developed with functions that can be accessed through a user interface. This application programming interface (API) was written in C and contains functions for sending the commands/configuration information listed in Table 5.3. C was chosen as the language for its speed and compatibility for interfacing with hardware.

An important feature of the API is in how the software handles the frame capture sequence. There are four FPGAs that all need to be synchronized so that all four quadrants of the imager are recording data synchronously and simultaneously. This simultaneity is handled through the use of POSIX threads (pthreads) [135]. These pthreads allow for each

FPGA control to be handled in separate threads of the CPU so that they can be run in parallel. After the initial pthreads are created (one for each FPGA), they prepare the FPGAs to capture data (configuring the number of laser repetitions, number of frames, etc.). It is not certain how long it will take each thread to perform the initialization tasks when they are launched, so a pthread barrier is used at which the threads pause their operation until all four of them have finished the initialization. Once they have all reached the barrier, they are released at which point they send the start signal for the FLIM acquisition routine to the FPGAs. Synchronizing the FPGAs in software is not ideal because of the inability to precisely time execution. Consequently, a better approach to synchronizing the FPGAs' data acquisition routines would be to send a single signal from software to one of the FPGAs and have it generate an electrical trigger signal that is distributed across the PCB over equal length traces that would then start the frame acquisition on all four FPGAs simultaneously. This synchronization approach was not included in the PCB design but should be added in future implementations. In this work, monitoring the CPU time when each thread is released from the pthread barrier shows that the image acquisition start signals are typically sent to all four FPGAs within 100μs. This timing precision is sufficient since the frame acquisition time is designed to be no less than 10ms.

5.5.3 Graphical User Interface

In order to make configuration and control of the system more efficient, a graphical user interface (GUI) was developed. The GUI was written using GTK+3 libraries for C. The GUI provides graphical elements to control all of the functions of the imaging system. In its current form, it does not perform lifetime extraction or data display, but these functions could be added. A screenshot showing the main GUI window and the control buttons is shown in Figure 5.17. The configuration button opens a subwindow that allows the user to easily set the bias currents for each TDC and individually enable/disable the pixels in the array (Figure 5.18).

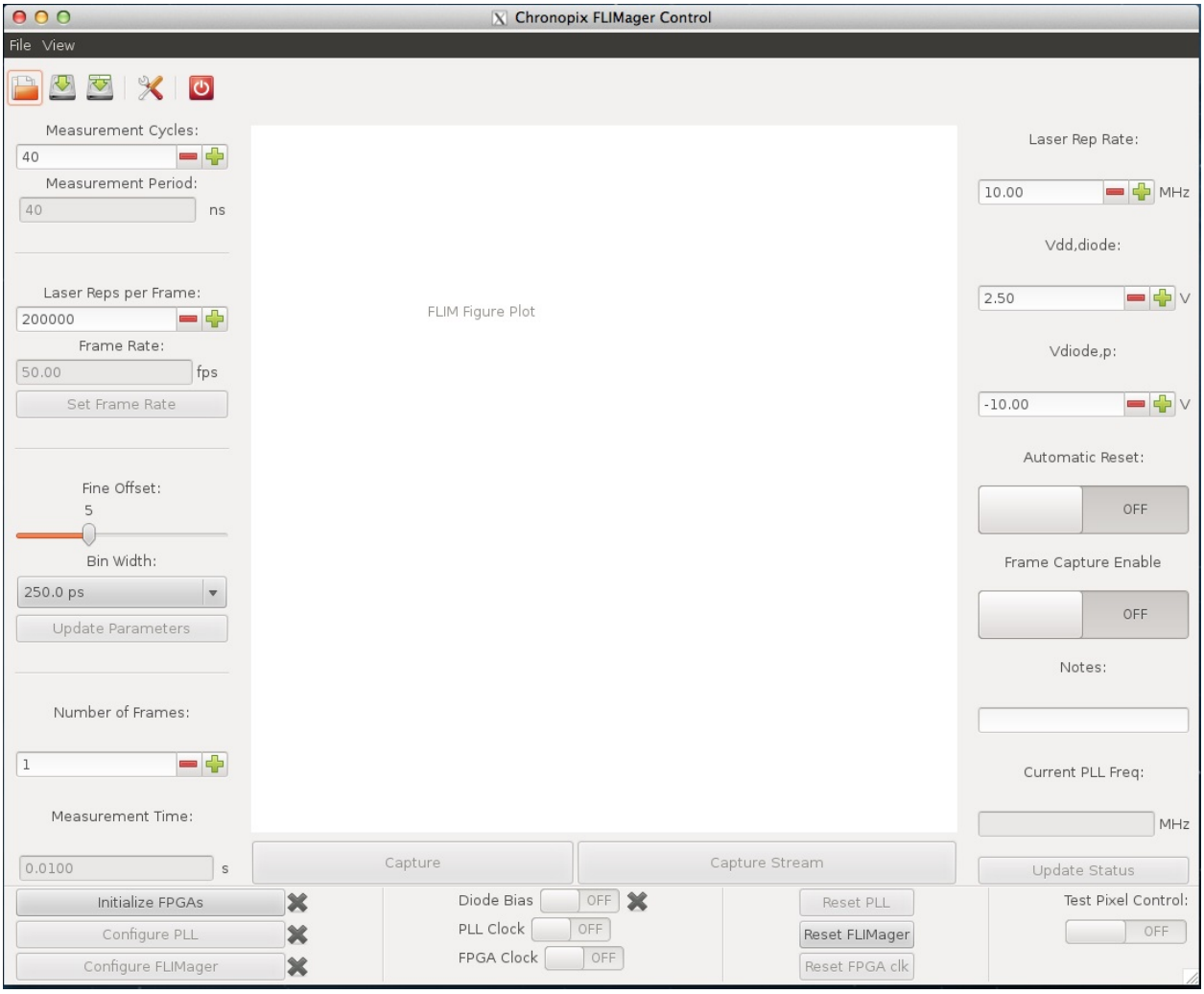

Figure 5.17: A screenshot of the main GUI window showing the primary configuration settings and control buttons.

5.5.4 Lifetime Extraction

The final software component for the imaging system is the algorithm for lifetime extraction. All lifetime extraction was done using MATLAB and the 'exp1' fit algorithm. Measurement results for fluorescein dye are presented in Section 5.6. Since all images contained a single fluorescent species, they will have a single exponential decay and this fitting algorithm is sufficient for all of the measurements taken. The Parallel Computing Toolbox in MATLAB was used to compute the lifetime for multiple pixels at the same time. Additionally, a significant amount of processing time is saved by setting a threshold for the number of

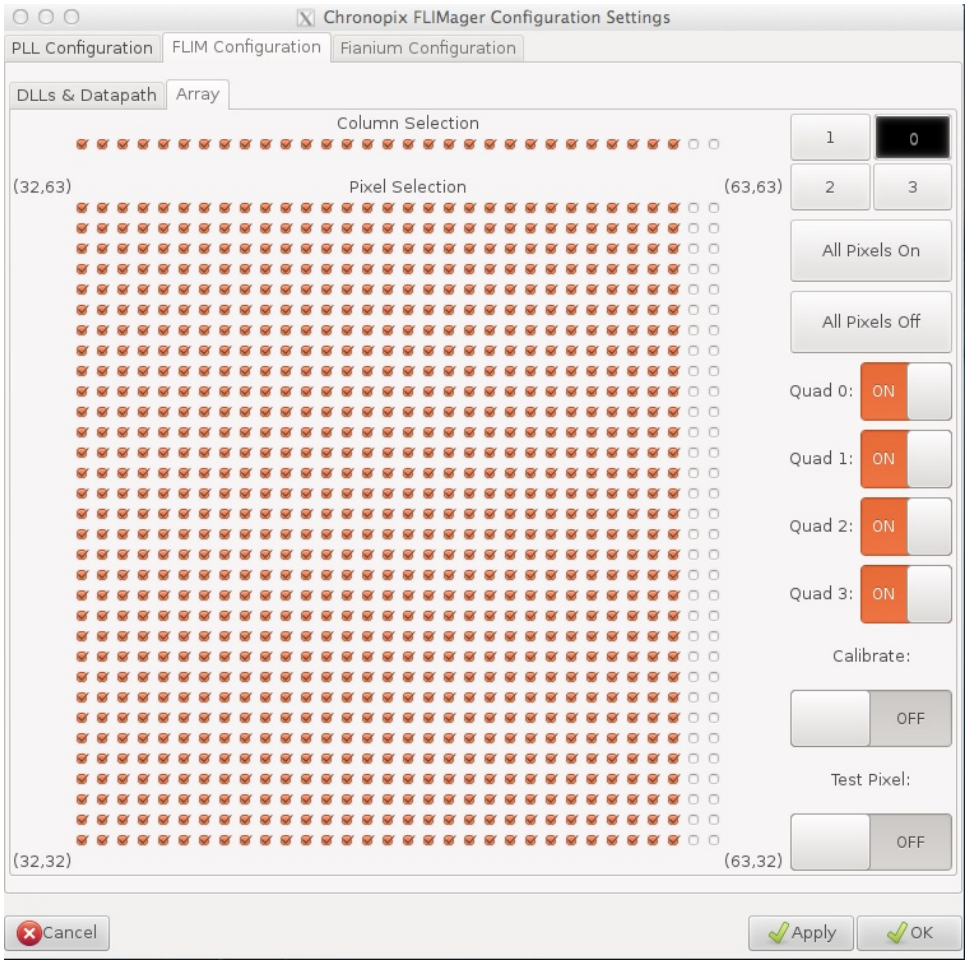

Figure 5.18: One tab of the configuration subwindow that shows the pixel enable switches for every pixel in the top right quadrant. In this example, columns 62 and 63 are turned off completely but all others remain on. Arbitrary patterns of pixels can be enabled using this window.

hits that a pixel must see in order for the fitting algorithm to attempt lifetime extraction. This threshold is set relative to the number of laser repetitions that were used to form the frame and should be chosen such that pixels with mostly noise events are omitted from the fitting process. With the use of this toolbox and thresholding technique, the average lifetime extraction for one complete 64-by-64 frame of data takes 35 seconds.

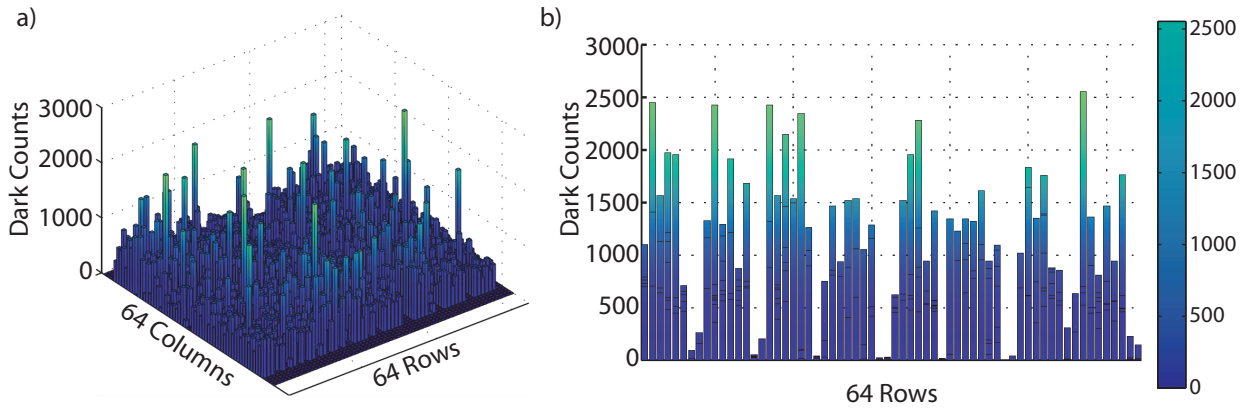

Figure 5.19: a) Angled view showing the distribution of the dark count rate across the array. b) cross-sectional view of the ends of the rows showing the distribution of noise events between rows.

5.6 Imaging Results

In this section results demonstrating the full imaging capabilities are presented. Preliminary images and measurements highlighting the features and basic functionality of the array were presented in Section 4.2.

5.6.1 Array-Wide Dark Count Rate

The dark count rate for all four quadrants operating simultaneously is presented in Figure 5.19. In Figure 5.19a, an angled view of the array is presented from which the distribution of the DCR throughout the array can be seen. The V_{ov} was set to 2.5V and the average DCR in this measurement is 304 Hz with a standard deviation of 114 Hz. Figure 5.19b shows a cross-sectional view showing the distribution between rows in the array. From this figure, it is clear that there is a periodic distribution with some rows showing much lower dark counts. Because there are some hits in these rows, it is likely that this pattern is the result of a layout-dependent effect like a voltage-droop problem in the datapath. This could be due to an insufficiently sized power distribution grid or an undersized reset buffer causing a timing violation.

Support for this failure mode is shown in Figure 5.20. In this figure, the imaging array

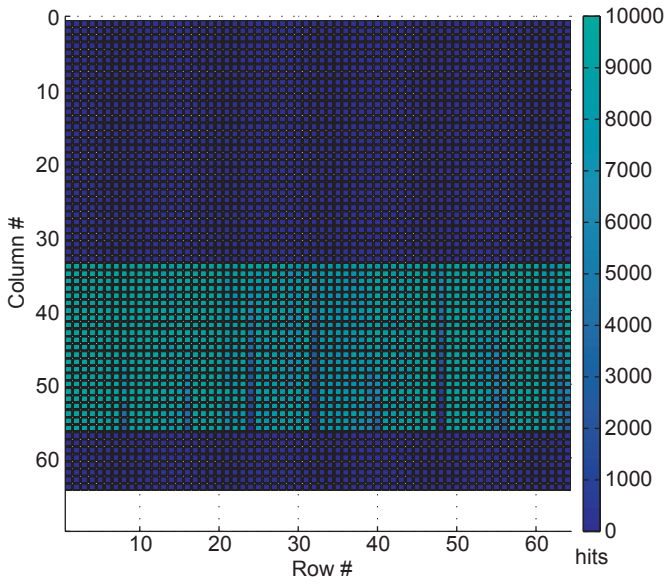

Figure 5.20: The array has been configured with all of the TDCs operating but with only 24 of the columns enabled. A gradient is observed from the center of the array to the periphery in the rows with low hit counts. The maximum color scale has been adjusted in this image to increase the contrast in the rows with gradients.

has been configured with all of the TDCs running but only 24 of the 64 columns enabled and a uniform uncorrelated white light source applied. The rows that were almost entirely empty in 5.19b now show a gradient starting from the center of the array to the periphery. This gradient is opposite to what would be expected if there were a voltage droop problem within the array, which suggests that the voltage droop occurs within the datapath or one of its controllers.

An additional measurement to show that this error is due to a power-related issue is presented in Figure 5.21. In this figure, half of the pixels in the array are enabled and the TDCs in the half of the array with disabled pixels are turned off. The TDCs are turned off by skewing the charge pump currents such that the control voltage for the VCDL goes to zero. This is accomplished by setting the UP current calibration value to 0 and the DN value to 63. The results from this measurement show that the performance of affected rows is partially recovered.

An alternative hypothesis for this problem is that there is a systematic error in one

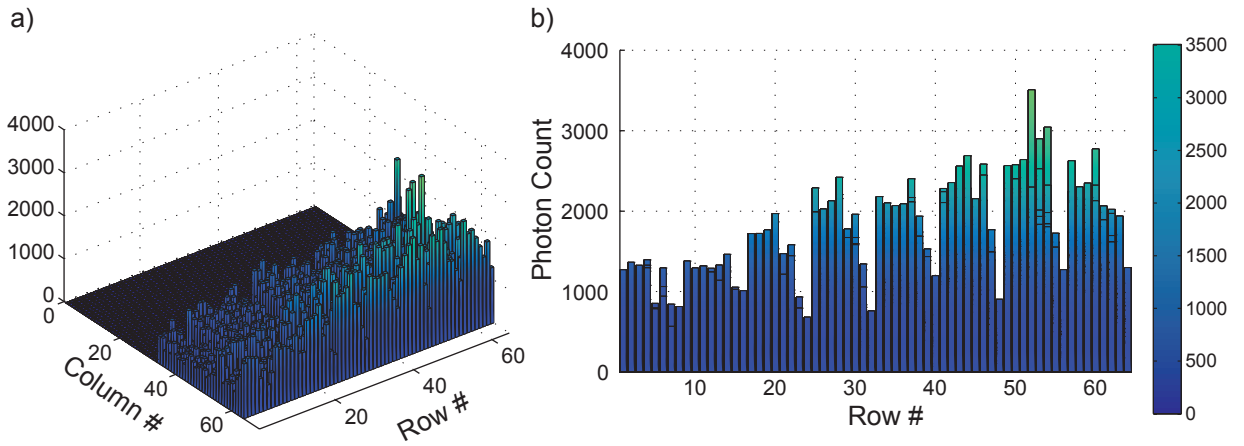

Figure 5.21: This figure shows the result of enabling half of the pixel array and turning off half of the TDCs while measuring bulk fluorescein emission with the excitation source aligned to the top of the array, closest to row 63. The distribution of hits between the rows is more uniform than before.

of the datapath controllers. In Figure 5.22, the array was improperly configured so that the controller will operate twice as fast as it should and overwrite one row entirely with zeros. This demonstrates that a logic failure in the controller would be more deterministic than the data presented in 5.19b and, thus, a failure due to a logic controller error is unlikely.

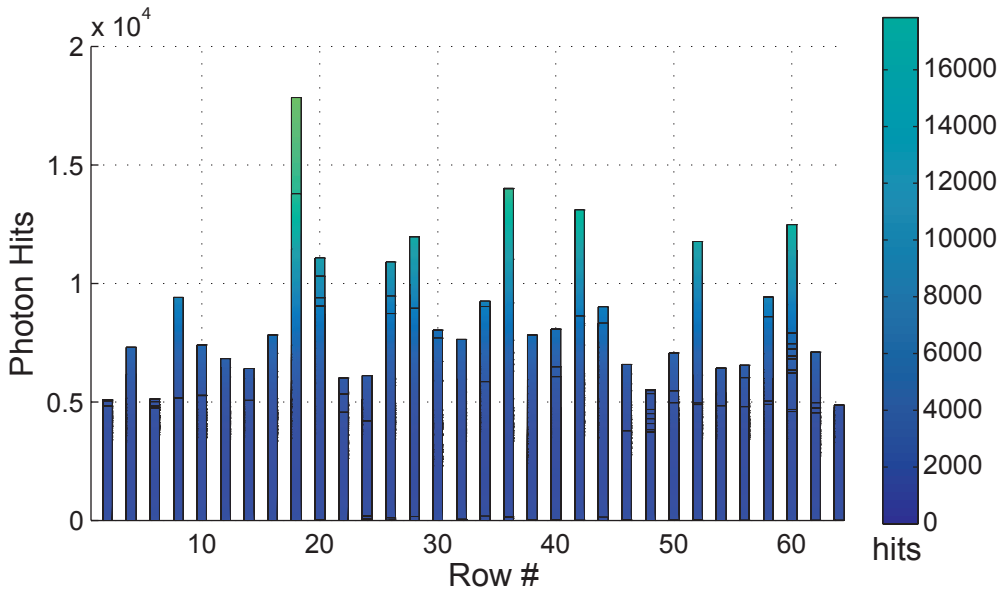

Figure 5.22: The controller for the datapath was intentionally miscalibrated to have too many shift operations from the array to the datapath, relative to the laser repetition rate. With this controller error, the output is deterministic and every other row is completely empty. This is representative of the type of behavior that is expected for a logical error in one of the datapath controllers.

5.6.2 Lifetime Measurement Setup

Optical Setup Tests of the complete lifetime imaging system were performed with the optical setup shown in Figure 5.23. The goal of these measurements was to demonstrate the array imaging performance using a known lifetime reference dye. In the system, a Fianium supercontinuum pulsed light source performs the excitation of a 0.5mM fluorescein solution in pH 7.4 phosphate buffered saline. The broadband light source is first incident on a cold mirror, which reflects only the visible light and transmits the infrared (IR) wavelengths generated by the nonlinear fiber in the Fianium. A second filter is selected from a filter wheel to select a narrow band from the visible spectrum of the Fianium. The beam is slightly divergent so a pair of lenses is used to collimate the beam before the mirrors that direct the pulsed light to the sample.

The sample is a dish of fluorescein dye placed directly above the imaging array with a 550-nm high-pass filter between the array and dish. For some measurements, a ceramic

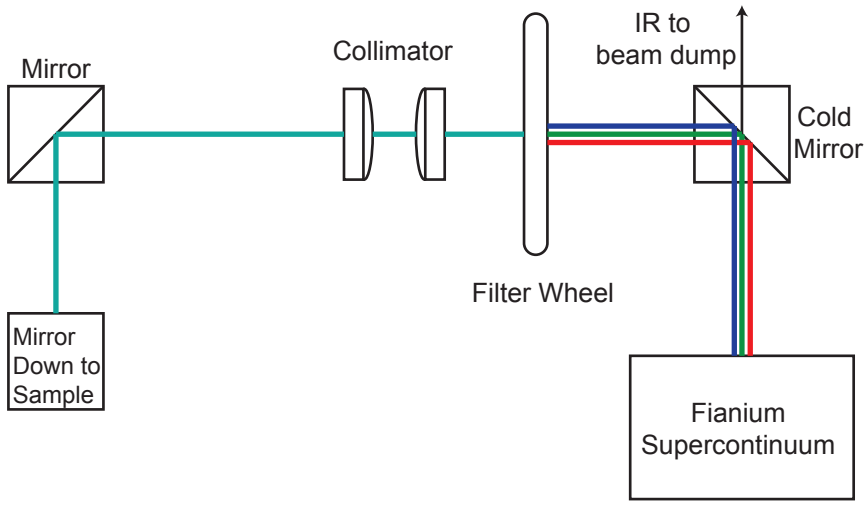

Figure 5.23: The cold mirror separates the IR and visible spectra of the Fianium supercontinuum source. The filter wheel has band-pass filters for 488nm, 500nm, and 520nm. The 520nm filter was selected for these measurements. The collimating lenses are used to correct for a slight beam divergence from the laser and two final mirrors are used to direct the beam into the page where the sample would sit above the imaging IC.

plate was placed between the filter and the imaging array to provide a masked region for producing contrast. A photograph of the IC, ceramic plate, high-pass filter, and fluorescein is show in figure 5.24.

The imaging system was built assuming that the Fianium laser would have an exact 20 MHz reference. However, the supercontinuum source is generated from a non-linear fiber and the intrinsic repetition rate is a function of the length of fiber used to generate the broad spectrum. The manufacturer allows for some tolerance in the fiber length and the repetition rate may not be exactly 20 MHz. For our supercontinuum the repetition rate is 19.4 MHz and the pulse-picker module included with it can produce integer fractions of the 19.4 MHz repetition rate. Additionally, the delay of the electrical trigger signal from the laser output to the IC input was too long. This delay caused the lifetime measurement to start at the end of the intensity decay curve. To compensate for this, the laser trigger signal was routed through an Agilent 8110A that allowed an additional delay to be inserted into the trigger signal. This additional delay allows the user to adjust the start time so that the measurement begins after the laser is pulsed and the intensity decay is still near its maximum.

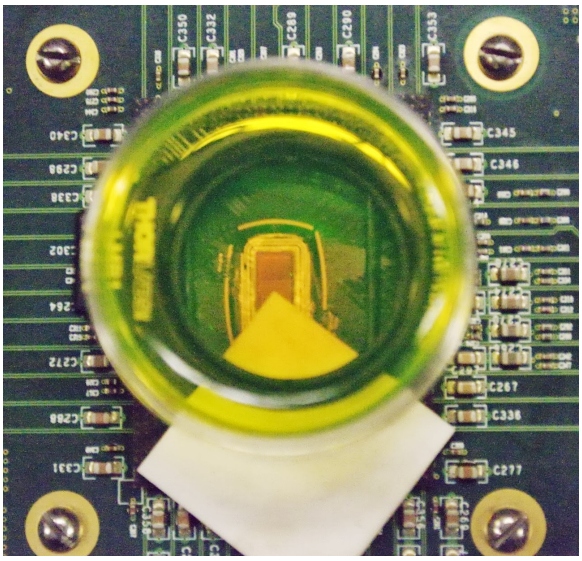

Figure 5.24: The dish containing fluorescein dye is placed directly over the imaging IC. A ceramic plate is used to provide contrast for an image.

5.6.3 Lifetime Imaging Results

In order to demonstrate the capabilities of the lifetime imaging system using only bulk fluorescein dye (lifetime 4-5ns [125]), several geometric arrangements of the ceramic plate were used, each of which is presented below.

Whole Array Fluorescein Image The first images were collected without using the ceramic plate to mask any portion of the array. The purpose of this arrangement is to demonstrate that almost every pixel in the array can be used for simultaneously recording lifetime data. Figure 5.25 shows an intensity image (5.25a) and a corresponding lifetime image (5.25b). In this measurement, the pulse-picked laser repetition rate was set to 3.24 MHz, the TDC resolution was 62.5 ps, the data was binned into 250 ps bins, and the datapath clock was running at 250 MHz. The image acquisition time was 20 ms per frame (50 fps). In this figure the outer four columns on each side were disabled to reduce the overall power consumption and allow imaging across the whole array simultaneously.

The second imaging configuration demonstrates that the imaging system properly localizes fluorescence to the pixels nearest the signal. In this configuration, the ceramic

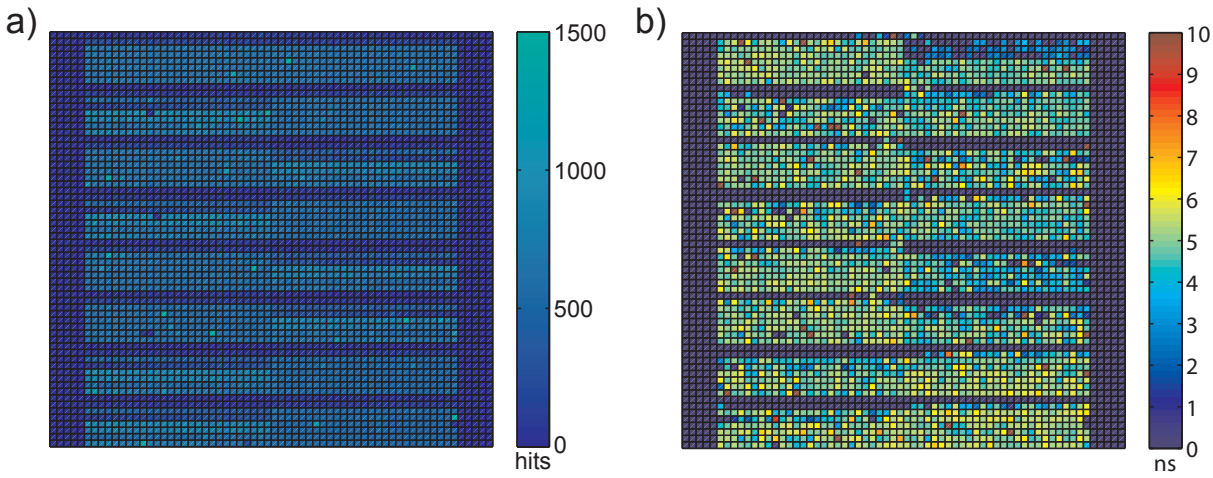

Figure 5.25: a) Intensity image, b) Lifetime image. The whole array was active during this measurement, except for the outer four columns on both sides.

plate is placed across the photosensitive region of the imager and blocks the fluorescence signal from many of the detectors. Figure 5.26a presents the intensity image and Figure 5.26b shows the measured lifetime image. In this measurement, the laser repetition rate was 3.24 MHz, the TDC resolution was 62.5 ps, the data was binned into 250 ps bins, and the datapath clock was running at 250 MHz. The image acquisition time was 100 ms per frame (10 fps).

A third configuration shows one corner of the ceramic plate covering the imaging array. In this case, images were acquired at 50 fps. Again, all of the timing parameters were the same as the previous image but the laser pulse repetition rate was 4.9 MHz. The intensity image is shown in Figure 5.27a and the lifetime image and representative decay are presented in Figure 5.27b. In this measurement, the entire array was enabled because the large masked region lowered the pixel activity across the array and reduced average power consumption. Figure 5.27c is a plot of the lifetime histogram for a single pixel that includes the monoexponential fit.

Finally, the sequential imaging performance of the system is demonstrated by recording 16 consecutive images at a rate of 100 fps. The imaging system parameters were: laser repetition rate of 4.9 MHz, TDC resolution of 62.5ps, bin width of 250ps, datapath frequency

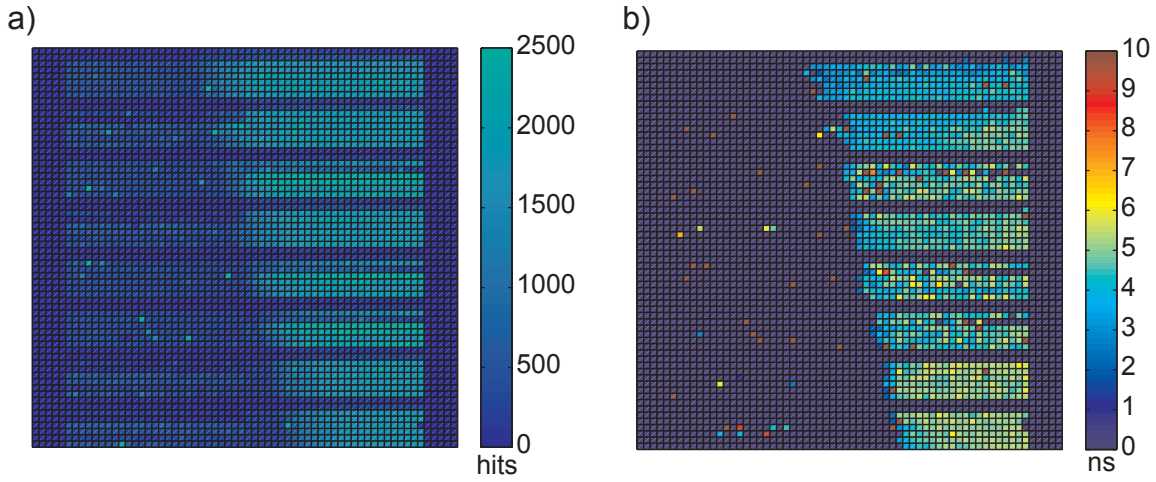

Figure 5.26: a) Intensity image, b) Lifetime image showing the well resolved mask. The outer four columns were disabled in this measurement.

of 250 MHz, and acquisition time of 10 ms. A panel showing these frames for the ceramic plate corner is presented in Figure 5.28

An additional sequence of lifetime data was captured at 100fps while a shutter was abruptly closed in the path of the laser, blocking the fluorescence signal during acquisition. While this sequence was not tightly controlled and cannot be used to quantify the imaging speed of the system, it does demonstrate the ability of the FLIM system to detect rapid changes in the lifetime. Figure 5.29 contains the panel showing this sequence of images.

Table 5.4 provides a summary of the FLIM imaging system performance.

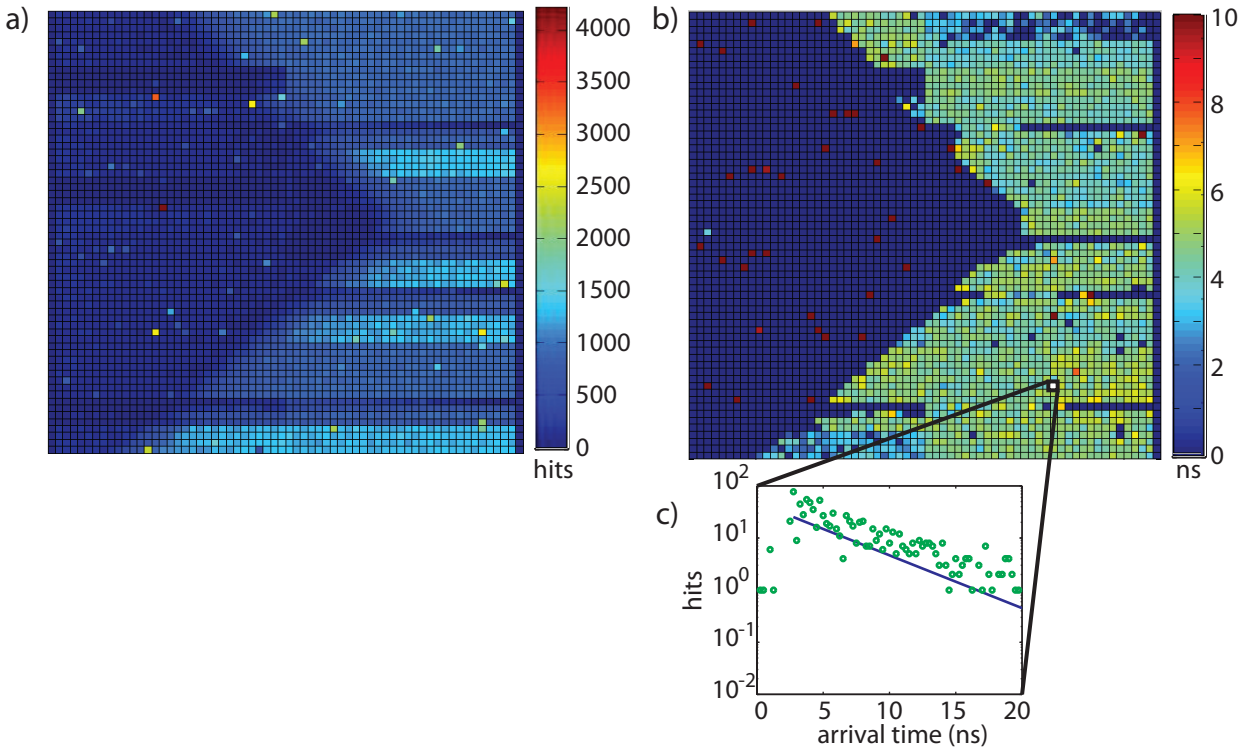

Figure 5.27: a) Intensity image, b) Lifetime image showing the well resolved mask. c.) A representative lifetime decay from pixel (49,10) in the image.

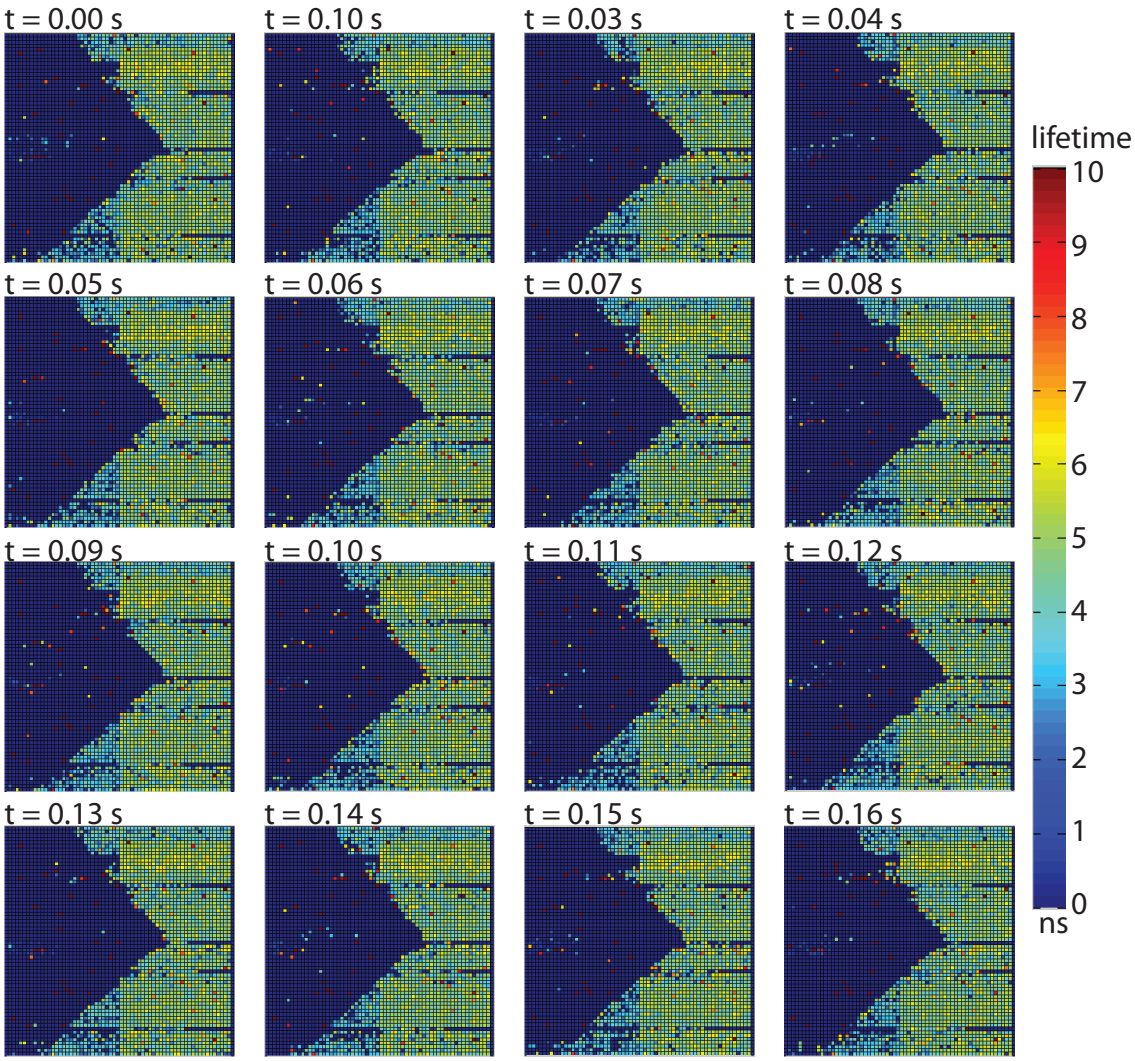

Figure 5.28: Sixteen consecutive frames captured using the lifetime imaging system with an acquisition time of only 10 ms per frame. The lifetime uniformity of Figure 5.27 is slightly better than that of these frames due to the shortened acquisition time.

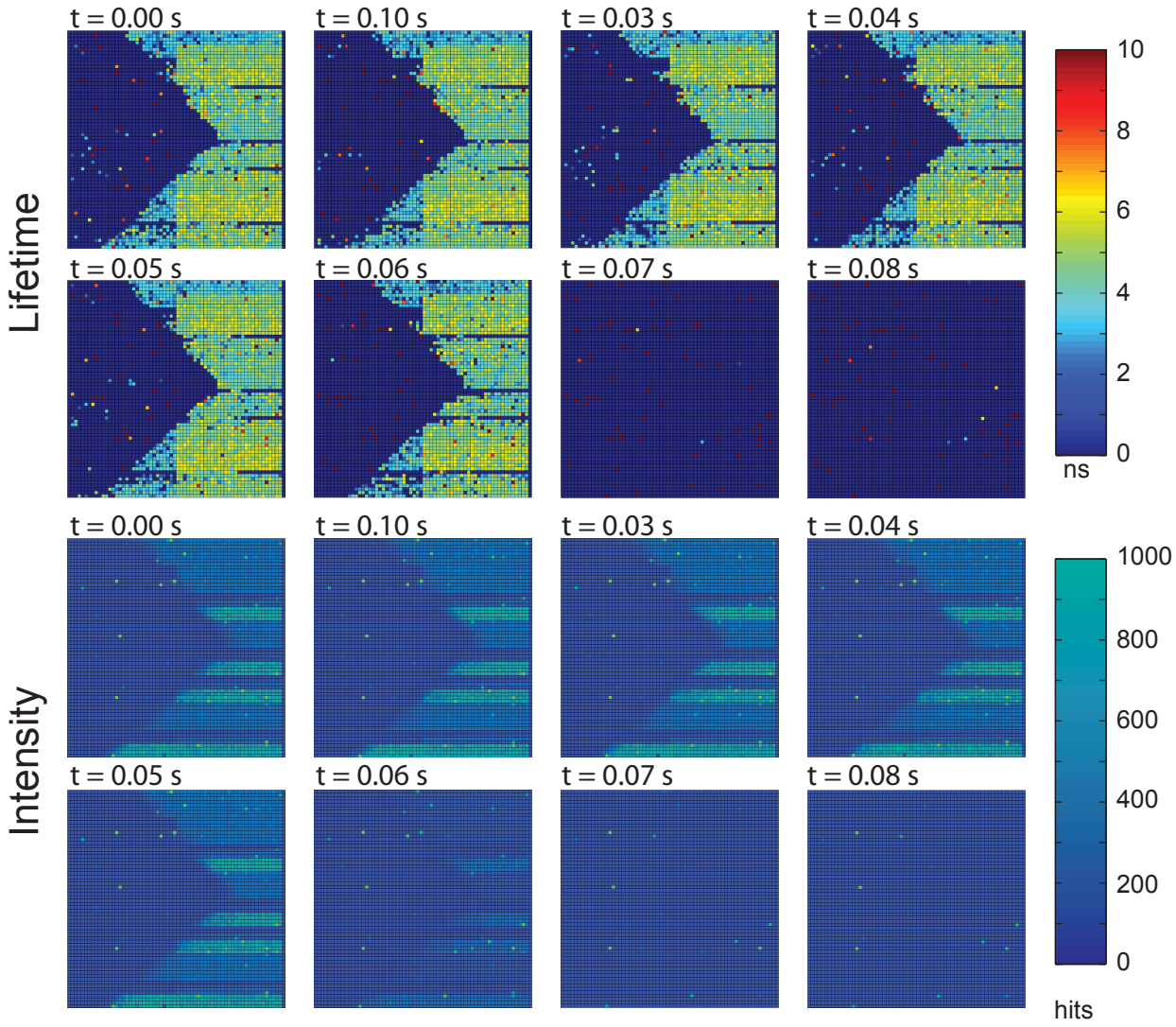

Figure 5.29: Eight consecutive frames captured using the lifetime imaging system and an image acquisition time of 10 ms per frame. A shutter was abruptly closed in the laser path after 6 frames.

Table 5.4: Summary of IC characteristics

Characteristic	Value	Unit
Peak Photon Detection Probability (PDP)	30 [65]	%
Peak PDP Wavelength	425 [65]	nm
Average DCR @ V_{ov}=2.5 and Room Temp.	544	Hz
TDC Resolution	62.5	ps
TDC Range	64	ns
TDC DNL	<4	LSB
TDC INL	<8	LSB
IC Output Bandwidth	42	Gbps
System Output Bandwidth	6.325	Gbps
Maximum Data Limited Frame Rate	466	fps
Maximum Attained FLIM Frame Rate	100	fps
Average Power IC	8.79	W
Average Power System	26.4	W

5.7 Summary

In this chapter, the complete high-speed FLIM system is presented. The combination of the FLIM-specific imaging IC, custom FPGA processing, PCIe interfaces, kernel drivers, and software allows for full-array imaging rates of up to 100 frames per second. While there are some problems with the imaging IC that limit its utility for wide-field imaging experiments, this complete system demonstrates the fastest TCSPC frame rate achieved to-date. Further modifications to the system firmware and driver can enable continuous recording of high-speed FLIM images. Additionally, a revision of the imager IC with improved power distribution and corrections of the bugs outlined above would improve overall image quality.

Chapter 6

Conclusions

6.1 Summary of Contributions

This thesis includes the complete design and characterization of a high frame rate fluorescence lifetime imaging microscopy system based on a CMOS SPAD array. The 64-by-64 SPAD array reached imaging speeds of up to 100 frames per second for time-correlated single photon counting lifetime measurements. The significant contributions emerging from this work include:

- The lowest noise room temperature single-photon avalanche diode in a standard 0.13μm CMOS technology.
- A novel SPAD active quench and reset pixel design that enables high count rate imaging with negligible afterpulsing.
- A complementary clock generator with process variation resiliency for generating well aligned $V_{dd}/2$ crossing points.
- A precisely controlled charge pump calibration circuit with differential-based tuning.
- The design of a fluorescence lifetime imaging microscopy optimized datapath based on statistical optimization.

- The first reported CMOS SPAD array to demonstrate time-correlated single-photon counting FLIM at 100 fps.

- The first demonstration of a PCIe-based direct memory access camera system to computer interface.

This work has led to the following peer-reviewed scientific publications:

- R. M. Field and K. L. Shepard, "A 100-fps Fluorescence Lifetime Imager in Standard 0.13μm CMOS," *2013 Symposium on VLSI Circuits*

- R. M. Field, J. Lary, J. Cohn, L. Paninski, and K. L. Shepard, "A low-noise, single-photon avalanche diode in standard 0.13μm complementary metal-oxide-semiconductor process," *Applied Physics Letters*, 97, 211111 (2010).

- 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," To appear in *IEEE Journal of Solid-State Circuits* April 2014.

6.2 Future Work

While this work has advanced the state-of-the-art in TCSPC imaging, there are areas where this work could be improved or adapted to provide further contributions to the field. These include:

- Improve lifetime accuracy performance of high frame rate imaging system through correcting the TDC coarse counter metastability flaw and considering alternative low-power approaches for time-to-digital conversion.

- Explore options like 3-D chip stacking to combine low-noise detectors in LOCOS processes, or custom SPAD processes with cutting edge CMOS technologies.

- Adapt wide-field image sensor for use as a multi-hit point detector to replace PMTs and APDs in laser scanning TCSPC imaging.

- Develop new applications that leverage the high frame rate capabilities of the camera system. These could include *in vivo* metabolic-based diagnostic imaging, dynamic FRET experiments, and real-time calcium concentration monitoring.

- Improve fitting algorithms and possibly adapt for use with parallel architectures, like GPUs.

Some specific thoughts about the most prominent current challenges associated with this work concern the power consumption and the data handling. These two factors are currently limiting the performance of the FLIM imaging system presented in this thesis.

Improving the data handling should focus on the kernel driver development and is a tractable problem. The kernel memory can be subdivided and double buffering could be used with interrupt driven communication to synchronize DMA transfers from the FPGAs to the computer. Other possibilities include creating memory pools that combine regions of higher memory to allow the FPGAs to access a larger space or performing DMA directly to a GPU DRAM, which could also be leveraged for faster lifetime extraction. The challenge with a direct to GPU implementation is ensuring that there is sufficient throughput in the lifetime extraction algorithm to prevent buffer overruns. Currently the CUDA environment from Nvidia supports remote direct memory access (RDMA), which looks promising for the suggested approach.

Secondly, the power consumption was a limiting factor for the SPADs (thermal degradation) and the data transfer circuits (IR droop causing reduced counts at some pixels) in this thesis. The power consumption on the imager IC is dominated by the datapath and TDCs neither of which were designed with power efficiency in mind. The datapath consumes 48% of the area of the IC and must be aggressively clocked to maintain the required data rate. Low power design techniques (lower power supply with low Vt transistors) could

help with the dynamic power consumption at the cost of higher static power consumption (leakage). Dynamic power consumption for MOS devices is given by:

$$P_{\text{dyn}} = CV^2 f \tag{6.1}$$

where C is the average load capacitance that is switched every clock cycle, V is the logic supply level and f is the frequency of operation. Reducing the datapath supply from 1.5V to 1.0V would lead to a 44% reduction in dynamic power consumption.

Scaling the design to an advanced technology node could also help to reduce the dynamic power consumption. This would come primarily due to the higher clock frequencies that gates in advanced technologies can support. With faster transistors, more logic or data shift operations could be performed between the fixed 20 MHz laser repetition rate. This could allow for a reduction in the number of stages needed in the datapath, which would greatly reduce the average load capacitance and lead to improved dynamic power consumption. Scaling to an advanced node would likely result in lower average load capacitance, lower supply voltage, and higher frequency. Realistic scaling numbers for this could be 2× increase in frequency, 50% reduction in capacitance, and 33% reduction in supply voltage. This would also result in a 44% reduction in dynamic power consumption.

The other block that is a major power consumer is the TDC block. There are a number of techniques that could be used to lower power consumption by the TDCs. Using the reverse-start-top technique [31] would enable an event-driven TDC mechanism whereby no power is consumed unless there is first a SPAD event. TDCs are an active area of research and low powered architectures are constantly being explored.

Bibliography

[1] A. H. Coons, "Localization of Antigen in Tissue Cells: II. Improvements in a Method for the Detection of Antigen by Means of Fluorescent Antibody," *Journal of Experimental Medicine*, vol. 91, pp. 1–13, Dec. 1949.

[2] R. Kerr, V. Lev-Ram, G. Baird, P. Vincent, R. Y. Tsien, and W. R. Schafer, "Optical Imaging of Calcium Transients in Neurons and Pharyngeal Muscle of C. elegans," *Neuron*, vol. 26, pp. 583–594, June 2000.

[3] F. V. Subach, O. M. Subach, I. S. Gundorov, K. S. Morozova, K. D. Piatkevich, A. M. Cuervo, and V. V. Verkhusha, "Monomeric fluorescent timers that change color from blue to red report on cellular trafficking.," *Nature Chemical Biology*, vol. 5, no. 2, pp. 118–26, 2009.

[4] J. Yguerabide, J. a. Schmidt, and E. E. Yguerabide, "Lateral mobility in membranes as detected by fluorescence recovery after photobleaching.," *Biophysical Journal*, vol. 40, pp. 69–75, Oct. 1982.

[5] T. Förster, "Energy migration and fluorescence. 1946.," *Journal of Biomedical Optics*, vol. 17, p. 011002, Jan. 2012.

[6] J. C. Waters, "Accuracy and precision in quantitative fluorescence microscopy.," *The Journal of Cell Biology*, vol. 185, pp. 1135–48, June 2009.

[7] J. W. Lichtman and J.-a. Conchello, "Fluorescence microscopy.," *Nature Methods*, vol. 2, pp. 910–9, Dec. 2005.

[8] W. B. Amos, J. G. White, and M. Fordham, "Use of confocal imaging in the study of biological structures.," *Applied Optics*, vol. 26, pp. 3239–43, Aug. 1987.

[9] W. Denk, J. Strickler, and W. Webb, "Two-photon laser scanning fluorescence microscopy," *Science*, vol. 248, pp. 73–76, Apr. 1990.

[10] P. T. So, C. Y. Dong, B. R. Masters, and K. M. Berland, "Two-photon excitation fluorescence microscopy.," *Annual Review of Biomedical Engineering*, vol. 2, pp. 399–429, Jan. 2000.

[11] D. B. Murphy and M. W. Davidson, *Fundamentals of Light Microscopy and Electronic Imaging*. Hoboken, NJ, USA: John Wiley & Sons, Inc., Sept. 2012.

[12] S. W. Hell and J. Wichmann, "Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy.," *Optics Letters*, vol. 19, pp. 780–2, June 1994.

[13] M. J. Rust, M. Bates, and X. Zhuang, "Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (STORM).," *Nature Methods*, vol. 3, pp. 793–5, Oct. 2006.

[14] S. T. Hess, T. P. K. Girirajan, and M. D. Mason, "Ultra-high resolution imaging by fluorescence photoactivation localization microscopy.," *Biophysical Journal*, vol. 91, pp. 4258–72, Dec. 2006.

[15] J. R. Lakowicz and K. W. Berndt, "Lifetime-selective fluorescence imaging using an rf phase-sensitive camera," *Review of Scientific Instruments*, vol. 62, no. 7, p. 1727, 1991.

[16] J. R. Lakowicz, H. Szmacinski, K. Nowaczyk, K. W. Berndt, and M. Johnson, "Fluorescence lifetime imaging.," *Analytical Biochemistry*, vol. 202, pp. 316–30, May 1992.

[17] X. F. Wang, A. Periasamy, B. Herman, and D. M. Coleman, "Fluorescence Lifetime Imaging Microscopy (FLIM): Instrumentation and Applications," *Critical Reviews in Analytical Chemistry*, vol. 23, pp. 369–395, Jan. 1992.

[18] J. R. Lakowicz, *Principles of Fluorescence Spectroscopy*. Boston, MA: Springer US, 2006.

[19] R. Sanders, A. Draaijer, H. Gerritsen, P. Houpt, and Y. Levine, "Quantitative pH Imaging in Cells Using Confocal Fluorescence Lifetime Imaging Microscopy," *Analytical Biochemistry*, vol. 227, no. 2, pp. 302–308, 1995.

[20] J. R. Lakowicz, H. Szmacinski, K. Nowaczyk, W. J. Lederer, M. S. Kirby, and M. L. Johnson, "Fluorescence lifetime imaging of intracellular calcium in COS cells using Quin-2.," *Cell Calcium*, vol. 15, pp. 7–27, Jan. 1994.

[21] J. R. Lakowicz, H. Szmacinski, K. Nowaczyk, and M. L. Johnson, "Fluorescence lifetime imaging of free and protein-bound NADH.," *Proceedings of the National Academy of Sciences*, vol. 89, pp. 1271–5, Feb. 1992.

[22] V. V. Ghukasyan and F.-j. Kao, "Monitoring Cellular Metabolism with Fluorescence Lifetime of Reduced Nicotinamide Adenine Dinucleotide," *The Journal of Physical Chemistry C*, vol. 113, pp. 11532–11540, July 2009.

[23] T. W. J. Gadella, ed., *FRET and FLIM Techniques*. 2011.

[24] T. H. Chia and M. J. Levene, "Detection of counterfeit U.S. paper money using intrinsic fluorescence lifetime.," *Optics Express*, vol. 17, pp. 22054–61, Nov. 2009.

[25] K. Suhling, P. M. W. French, and D. Phillips, "Time-resolved fluorescence microscopy.," *Photochemical and Photobiological Sciences*, vol. 4, pp. 13–22, Jan. 2005.

[26] D.-U. Li, R. Walker, J. Richardson, B. Rae, A. Buts, D. Renshaw, and R. Henderson, "FPGA Implementation of a Video-rate Fluorescence Lifetime Imaging System with a 32x32 CMOS Single-Photon Avalanche Diode Array," in *IEEE International Symposium on Circuits and Systems*, pp. 3082–3085, IEEE, May 2009.

[27] D. E. Schwartz, E. Charbon, and K. L. Shepard, "A Single-Photon Avalanche Diode Array for Fluorescence Lifetime Imaging Microscopy," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 2546–2557, Nov. 2008.

[28] L. Pancheri and D. Stoppa, "A SPAD-based Pixel Linear Array for High-Speed Time-Gated Fluorescence Lifetime Imaging," in *Proceedings of ESSCIRC*, pp. 428–431, IEEE, Sept. 2009.

[29] E. Charbon, "Highly Sensitive Arrays of Nano-sized Single-Photon Avalanche Diodes for Industrial and Bio Imaging," in *Proceedings of Nano-net*, pp. 161–168, Springer, 2009.

[30] M. Gersbach, Y. Maruyama, E. Labonne, J. Richardson, R. Walker, L. Grant, R. Henderson, F. Borghetti, D. Stoppa, and E. Charbon, "A Parallel 32x32 Time-To-Digital Converter Array Fabricated in a 130 nm Imaging CMOS Technology," in *Proceedings of ESSCIRC*, pp. 196–199, IEEE, Sept. 2009.

[31] C. Veerappan, J. Richardson, R. Walker, D.-U. Li, M. W. Fishburn, Y. Maruyama, D. Stoppa, F. Borghetti, M. Gersbach, R. K. Henderson, and E. Charbon, "A 160x128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter," in *IEEE International Solid-State Circuits Conference*, pp. 312–314, IEEE, Feb. 2011.

[32] M. Chalfie, Y. Tu, G. Euskirchen, W. Ward, and D. Prasher, "Green fluorescent protein as a marker for gene expression," *Science*, vol. 263, pp. 802–805, Feb. 1994.

[33] A. P. Alivisatos, "Semiconductor Clusters, Nanocrystals, and Quantum Dots," *Science*, vol. 271, pp. 933–937, Feb. 1996.

[34] C. Y. Dong, T. French, P. T. So, C. Buehler, K. M. Berland, and E. Gratton, "Fluorescence-lifetime imaging techniques for microscopy.," in *Methods in Cell Biology*, vol. 72, pp. 431–64, Jan. 2003.

[35] J. R. Lakowicz, H. Szmacinski, and M. L. Johnson, "Calcium imaging using fluorescence lifetimes and long-wavelength probes," *Journal of Fluorescence*, vol. 2, pp. 47–62, Mar. 1992.

[36] M. K. Kuimova, G. Yahioglu, J. a. Levitt, and K. Suhling, "Molecular rotor measures viscosity of live cells via fluorescence lifetime imaging.," *Journal of the American Chemical Society*, vol. 130, pp. 6672–3, May 2008.

[37] M. Sumitani and N. Nakashima, "Temperature dependence of fluorescence lifetimes of trans-stilbene," *Chemical Physics Letters*, vol. 5, no. 1, pp. 183–185, 1977.

[38] P. Bastiaens and A. Squire, "Fluorescence lifetime imaging microscopy: spatial resolution of biochemical processes in the cell," *Trends in Cell Biology*, vol. 9, no. 2, pp. 48–52, 1999.

[39] H. Wallrabe and A. Periasamy, "Imaging protein molecules using FRET and FLIM microscopy.," *Current Opinion in Biotechnology*, vol. 16, pp. 19–27, Feb. 2005.

[40] M. C. Skala, K. M. Riching, A. Gendron-Fitzpatrick, J. Eickhoff, K. W. Eliceiri, J. G. White, and N. Ramanujam, "In vivo multiphoton microscopy of NADH and FAD redox states, fluorescence lifetimes, and cellular morphology in precancerous epithelia.," *Proceedings of the National Academy of Sciences*, vol. 104, pp. 19494–9, Dec. 2007.

[41] R. a. Gatenby and R. J. Gillies, "Why do cancers have high aerobic glycolysis?," *Nature Reviews. Cancer*, vol. 4, pp. 891–9, Nov. 2004.

[42] R. Cicchi and F. S. Pavone, "Non-linear fluorescence lifetime imaging of biological tissues.," *Analytical and Bioanalytical Chemistry*, vol. 400, pp. 2687–97, July 2011.

[43] B. Alberts, *Molecular Biology of the Cell*. New York: Garland Science, 4th ed., 2002.

[44] A. Squire, P. J. Verveer, and P. I. Bastiaens, "Multiple frequency fluorescence lifetime imaging microscopy.," *Journal of Microscopy*, vol. 197, pp. 136–49, Feb. 2000.

[45] M. A. Digman, V. R. Caiolfa, M. Zamai, and E. Gratton, "The phasor approach to fluorescence lifetime imaging analysis.," *Biophysical Journal*, vol. 94, no. 2, pp. L14–6, 2008.

[46] O. Holub, M. Seufferheld, C. Gohlke, and R. Clegg, "Fluorescence lifetime imaging (FLI) in real-time-a new technique in photosynthesis research," *Photosynthetica*, vol. 38, no. 4, pp. 581–599, 2000.

[47] R. Ballew and J. Demas, "An error analysis of the rapid lifetime determination method for the evaluation of single exponential decays," *Analytical Chemistry*, vol. 61, pp. 30–33, Jan. 1989.

[48] A. V. Agronskaia, L. Tertoolen, and H. C. Gerritsen, "High frame rate fluorescence lifetime imaging," *Journal of Physics D: Applied Physics*, vol. 36, pp. 1655–1662, 2003.

[49] C. J. de Grauw and H. C. Gerritsen, "Multiple Time-Gate Module for Fluorescence Lifetime Imaging," *Applied Spectroscopy*, vol. 55, pp. 670–678, June 2001.

[50] W. Becker, A. Bergmann, K. Koenig, and U. Tirlapur, "Picosecond fluorescence lifetime microscopy by TCSPC imaging," in *Proceedings of SPIE Vol. 4262, Multiphoton Microscopy in the Biomedical Sciences* (A. Periasamy and P. T. C. So, eds.), vol. 4262, pp. 414–419, Apr. 2001.

[51] W. Becker, A. Bergmann, M. a. Hink, K. König, K. Benndorf, and C. Biskup, "Fluorescence lifetime imaging by time-correlated single-photon counting.," *Microscopy Research and Technique*, vol. 63, pp. 58–66, Jan. 2004.

[52] C. Harris and B. Selinger, "Single-Photon Decay Spectroscopy. II The Pile-up Problem," *Australian Journal of Chemistry*, vol. 32, pp. 2111–2129, 1979.

[53] C. Chang, D. Sud, and M. Mycek, "Fluorescence Lifetime Imaging Microscopy," *Methods in Cell Biology*, vol. 81, no. 06, pp. 495–524, 2007.

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

[55] A. Vilà, A. Arbat, E. Vilella, and A. Dieguez, "Geiger-Mode Avalanche Photodiodes in Standard CMOS Technologies," in *Photodetectors*, pp. 175–204, 2004.

[56] S. Tisa, F. Zappa, and I. Labanca, "On-chip detection and counting of single-photons," in *IEEE International Electron Device Meeting*, vol. 00, pp. 815–818, 2005.

[57] H. Finkelstein, M. J. Hsu, and S. Esener, "An ultrafast Geiger-mode single-photon avalanche diode in 0.18-μm CMOS technology," in *Proceedings of SPIE* (W. Becker, ed.), vol. 6372, pp. 63720W–63720W–10, Spie, Oct. 2006.

[58] M. Marwick and A. Andreou, "Fabrication and Testing of Single Photon Avalanche Detectors in the TSMC 0.18 um CMOS Technology," in *Conference on Information Sciences and Systems*, pp. 741–744, 2007.

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

[60] L. Pancheri and D. Stoppa, "Low-Noise CMOS single-photon avalanche diodes with 32 ns dead time," in *IEEE ESSDERC*, pp. 362–365, Sept. 2007.

[61] M. Gersbach, J. Richardson, E. Mazaleyrat, S. Hardillier, C. Niclass, R. Henderson, L. Grant, and E. Charbon, "A low-noise single-photon detector implemented in a 130nm CMOS imaging process," *IEEE Journal of Solid-State Circuits*, vol. 53, pp. 803–808, July 2009.

[62] B. Nouri, M. Dandin, and P. Abshire, "Characterization of single-photon avalanche diodes in standard CMOS," in *IEEE Sensors*, pp. 1889–1892, 2009.

[63] J. A. Richardson, L. A. Grant, and R. K. Henderson, "Low Dark Count Single-Photon Avalanche Diode Structure Compatible With Standard Nanometer Scale CMOS Technology," *IEEE Photonics Technology Letters*, vol. 21, pp. 1020–1022, July 2009.

[64] M. A. Karami, M. Gersbach, and E. Charbon, "A new single-photon avalanche diode in 90nm standard CMOS technology," *Imaging*, vol. 7780, pp. 77801F–77801F–6, 2010.

[65] R. M. Field, J. Lary, J. Cohn, L. Paninski, and K. L. Shepard, "A low-noise, single-photon avalanche diode in standard 0.13 μm complementary metal-oxide-semiconductor process," *Applied Physics Letters*, vol. 97, no. 21, p. 211111, 2010.

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

[67] H. Finkelstein, M. Hsu, S. Zlatanovic, and S. Esener, "Performance trade-offs in single-photon avalanche diode miniaturization.," *The Rev. of Sci. Inst.*, vol. 78, no. 10, p. 103103, 2007.

[68] D. Contini, A. Dalla Mora, L. Di Sieno, R. Cubeddu, A. Tosi, G. Boso, and A. Pifferi, "Memory effect in gated single-photon avalanche diodes: a limiting noise contribution similar to afterpulsing," in *Proceedings of SPIE Vol. 8619, Physics and Simulation of Optoelectronic Devices XXI* (B. Witzigmann, M. Osinski, F. Henneberger, and Y. Arakawa, eds.), vol. 8619, pp. 86191L–86191L–9, Mar. 2013.

[69] R. K. Henderson, E. A. G. Webster, R. Walker, J. A. Richardson, and L. A. Grant, "A 3x3 , 5μm pitch , 3-transistor Single Photon Avalanche Diode Array with Integrated 11V Bias Generation in 90nm CMOS Technology," in *IEEE International Electron Device Meeting*, pp. 336–339, 2010.

[70] D. Stoppa, L. Pancheri, M. Scandiuzzo, L. Gonzo, G.-f. D. Betta, S. Member, and A. Simoni, "A CMOS 3-D imager based on single photon avalanche diode," *IEEE Transactions on Circuits and Systems I*, vol. 54, no. 1, pp. 4–12, 2007.

[71] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128 x 128 Single-Photon Image Sensor With Column-Level 10-Bit Time-to-Digital Converter Array," *IEEE Journal of Solid-State Circuits*, vol. 43, pp. 2977–2989, Dec. 2008.

[72] J. Richardson, R. Walker, L. Grant, D. Stoppa, F. Borghetti, E. Charbon, M. Gersbach, and R. K. Henderson, "A 32x32 50ps Resolution 10 bit Time to Digital Converter Array in 130nm CMOS for Time Correlated Imaging," in *IEEE Custom Integrated Circuits Conference*, no. 029217, pp. 77–80, IEEE, Sept. 2009.

[73] D. Stoppa, D. Mosconi, L. Pancheri, and L. Gonzo, "Single-Photon Avalanche Diode CMOS Sensor for Time-Resolved Fluorescence Measurements," *IEEE Sensors Journal*, vol. 9, pp. 1084–1090, Sept. 2009.

[74] D.-u. Li, J. Arlt, J. Richardson, R. Walker, A. Buts, D. Stoppa, E. Charbon, and R. Henderson, "Real-time fluorescence lifetime imaging system with a 32 x 32 0.13um CMOS low dark-count single-photon avalanche diode array," *Optics Express*, vol. 18, p. 10257, May 2010.

[75] Y. Maruyama and E. Charbon, "An all-digital, time-gated 128X128 spad array for on-chip, filter-less fluorescence detection," in *2011 16th International Solid-State Sensors, Actuators and Microsystems Conference*, pp. 1180–1183, IEEE, June 2011.

[76] S. Isaak, S. Bull, M. C. Pitter, I. Harrison, A. M. Hashim, and V. K. Arora, "Fully Integrated Linear Single Photon Avalanche Diode (SPAD) Array with Parallel Readout Circuit in a Standard 180 nm CMOS Process," *Physics*, vol. 180, pp. 175–180, 2011.

[77] 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 0.13um CMOS imaging technology," in *IEEE International Solid-State Circuits Conference*, pp. 122–124, IEEE, Feb. 2012.

[78] R. M. Field and K. L. Shepard, "A 100-fps Fluorescence Lifetime Imager in Standard 0.13-um CMOS," in *Symposium on VLSI Circuits*, (Kyoto), pp. 10–11, 2013.

[79] S. Marangoni, I. Rech, M. Ghioni, P. Maccagnani, M. Chiari, M. Cretich, F. Damin, G. Di Carlo, and S. Cova, "A 6 x 8 photon-counting array detector system for fast and sensitive analysis of protein microarrays," *Sensors and Actuators B: Chemical*, vol. 149, pp. 420–426, Aug. 2010.

[80] M. Fishburn and E. Charbon, "System Tradeoffs in Gamma-Ray Detection Utilizing SPAD Arrays and Scintillators," *IEEE Transactions on Nuclear Science*, vol. 57, no. 5, p. 2549, 2010.

[81] D. Tyndall, R. Walker, K. Nguyen, R. Galland, J. Gao, I. Wang, M. Kloster, A. Delon, and R. Henderson, "Automatic laser alignment for multifocal microscopy using a LCOS SLM and a 32 x 32 pixel CMOS SPAD array," *Proceedings of SPIE Vol. 8086, Advanced Microscopy Techniques*, vol. 8086, no. 0, pp. 80860S–80860S–6, 2011.

[82] C. Niclass, M. Sergio, and E. Charbon, "A single photon avalanche diode array fabricated in 0.35-μm CMOS and based on an event-driven readout for TCSPC experiments," in *Proceedings of SPIE Vol. 6372, Advanced Photon Counting Techniques* (W. Becker, ed.), vol. 6372, pp. 63720S–63720S–12, Oct. 2006.

[83] M. Gersbach, Y. Maruyama, R. Trimananda, M. W. Fishburn, D. Stoppa, J. A. Richardson, R. Walker, R. Henderson, and E. Charbon, "A Time-Resolved, Low-Noise Single-Photon Image Sensor Fabricated in Deep-Submicron CMOS Technology," *IEEE Journal of Solid-State Circuits*, vol. 47, pp. 1394–1407, June 2012.

[84] L. Li and L. Davis, "Single photon avalanche diode for single molecule detection," *Review of Scientific Instruments*, vol. 37388, no. January, pp. 1524–1529, 1993.

[85] B. Streetman and S. Banerjee, *Solid State Electronic Devices*. fifth ed., 2000.

[86] G. Moore, "Cramming more components onto integrated circuits," *Electronics*, vol. 38, no. 8, 1965.

[87] A. Mathewson, R. Duane, and G. T. Wrixon, "CMOS Compatible APD Arrays," in *International Society for Optics and Photonics*, vol. 2209, pp. 378–387, 1994.

[88] A. Rochas, A. Pauchard, P.-a. Besse, D. Pantic, Z. Prijic, and R. Popovic, "Low-noise silicon avalanche photodiodes fabricated in conventional CMOS technologies," *IEEE Transactions on Electron Devices*, vol. 49, pp. 387–394, Mar. 2002.

[89] S. Tisa, F. Zappa, A. Tosi, and S. Cova, "Electronics for single photon avalanche diode arrays," *Sensors and Actuators A: Physical*, vol. 140, no. 1, pp. 113–122, 2007.

[90] E. Charbon, "Towards large scale CMOS single-photon detector arrays for lab-on-chip applications," *J. Phys. D*, vol. 41, no. 9, p. 094010, 2008.

[91] D. E. Schwartz, P. Gong, and K. L. Shepard, "Time-resolved Förster-resonance-energy-transfer DNA assay on an active CMOS microarray.," *Biosensors and Bioelectronics*, vol. 24, pp. 383–90, Nov. 2008.

[92] P. Seitz and A. J. Theuwissen, eds., *Single-Photon Imaging*, vol. 160 of *Springer Series in Optical Sciences*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.

[93] E. a. G. Webster, J. a. Richardson, L. a. Grant, D. Renshaw, and R. K. Henderson, "A Single-Photon Avalanche Diode in 90-nm CMOS Imaging Technology With 44% Photon Detection Efficiency at 690 nm," *IEEE Electron Device Letters*, vol. 33, pp. 694–696, May 2012.

[94] D. Bronzi, F. Villa, S. Bellisai, S. Tisa, G. Ripamonti, and A. Tosi, "Figures of merit for CMOS SPADs and arrays," in *Proceedings of SPIE Vol. 8773, Photon Counting Applications IV; and Quantum Optics and Quantum Information Transfer and Processing* (R. Sobolewski and J. Fiurásek, eds.), vol. 8773, pp. 877304–877304–7, May 2013.

[95] E. N. Brown, R. Barbieri, V. Ventura, R. E. Kass, and L. M. Frank, "The time-rescaling theorem and its application to neural spike train data analysis.," *Neural Computation*, vol. 14, pp. 325–46, Feb. 2002.

[96] H. Norian, I. Kymissis, and K. Shepard, "Integrated CMOS quantitative polymerase chain reaction lab-on-chip," in *Symposium on VLSI Circuits*, (Kyoto), pp. 220–221, 2013.

[97] D. Gedcke and W. McDonald, "A constant fraction of pulse height trigger for optimum time resolution," *Nuclear Instruments and Methods*, vol. 55, pp. 377–380, Jan. 1967.

[98] T. J. Paulus, "Timing Electronics and Fast Timing Methods with Scintillation Detectors," *IEEE Transactions on Nuclear Science*, vol. 32, no. 3, pp. 1242–1249, 1985.

[99] N. S. Nightingale, "A new silicon avalanche photodiode photon counting detector module for astronomy," *Experimental Astronomy*, vol. 1, no. 6, pp. 407–422, 1990.

[100] F. Zappa, A. Lotito, A. Giudice, S. Cova, and M. Ghioni, "Monolithic Active-Quenching and Active-Reset Circuit for Single-Photon Avalanche Detectors," *IEEE Journal of Solid-State Circuits*, vol. 38, pp. 1298–1301, July 2003.

[101] E. Vilella and a. Diéguez, "A gated single-photon avalanche diode array fabricated in a conventional CMOS process for triggered systems," *Sensors and Actuators A: Physical*, vol. 186, pp. 163–168, Oct. 2012.

[102] S. Henzler, *Time-to-Digital Converters*. Springer Series in Advanced Microelectronics, Dordrecht: Springer Netherlands, 2010.

[103] R. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, P. Balsara, W. Center, T. Inc, and T. Dallas, "1.3 V 20 ps time-to-digital converter for frequency synthesis in 90-nm CMOS," *IEEE Transactions on Circuits and Systems II*, vol. 53, no. 3, pp. 220–224, 2006.

[104] A. Mantyniemi, T. Rahkonen, and J. Kostamovaara, "A CMOS Time-to-Digital Converter (TDC) Based On a Cyclic Time Domain Successive Approximation Interpolation Method," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 11, pp. 3067–3078, 2009.

[105] C. Niclass, A. Rochas, P.-A. Besse, and E. Charbon, "Design and Characterization of a CMOS 3-D Image Sensor Based on Single Photon Avalanche Diodes," *IEEE Journal of Solid-State Circuits*, vol. 40, no. 9, pp. 1847–1854, 2005.

[106] A. S. Yousif and J. W. Haslett, "A Fine Resolution TDC Architecture for Next Generation PET Imaging," *IEEE Transactions on Nuclear Science*, vol. 54, pp. 1574–1582, Oct. 2007.

[107] P. Napolitano, F. Alimenti, and P. Carbone, "A Novel Sample-and-Hold-Based Time-to-Digital Converter Architecture," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, pp. 1019–1026, May 2010.

[108] M. Fries and J. Williams, "High-precision TDC in an FPGA using a 192 MHz quadrature clock," *Nuclear Science Symposium Conference Record*, vol. 1, pp. 580–584, 2002.

[109] T. Rahkonen and J. Kostamovaara, "The use of stabilized CMOS delay lines for the digitization of short time intervals," *IEEE International Solid-State Circuits Conference*, vol. 28, no. 8, 1993.

[110] M. Johnson and E. Hudson, "A variable delay line PLL for CPU-coprocessor synchronization," *IEEE Journal of Solid-State Circuits*, vol. 23, no. 5, pp. 1218–1223, 1988.

[111] S. Sidiropoulos and M. Horowitz, "Adaptive bandwidth DLLs and PLLs using regulated supply CMOS buffers," in *Symposium on VLSI Circuits*, pp. 124–127, Ieee, 2000.

[112] W. J. Dally and J. Poulton, *Digial Systems Engineering*, vol. 17. Aug. 1985.

[113] W. Lee, J. Cho, and S. Lee, "A high speed and low power phase-frequency detector and charge-pump," *Design*, pp. 0–3, 1999.

[114] J. Christiansen, "An integrated high resolution CMOS timing generator based on an array of delay locked loops," *IEEE Journal of Solid-State Circuits*, vol. 31, pp. 952–957, July 1996.

[115] S. Soliman, F. Yuan, and K. Raahemifar, "An overview of design techniques for CMOS phase detectors," in *IEEE International Symposium on Circuits and Systems*, pp. V–457–V–460, IEEE, 2002.

[116] H. Johansson, "A simple precharged CMOS phase frequency detector," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 2, pp. 295–299, 1998.

[117] C. Liang, S. Chen, and S. Liu, "A Digital Calibration Technique for Charge Pumps in Phase-Locked Systems," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 390–398, 2008.

[118] N. H. E. Weste and D. Harris, *CMOS VLSI Design*. Pearson, 3rd ed., 2005.

[119] TIA/EIA-644-A, "Electrical Characteristics of Low Voltage Differential Signaling (LVDS) Interface Circuits," 2001.

[120] B. Razavi, *Design of Analog CMOS Integrated Cicruits*. McGraw-Hill, 2001.

[121] D. Fischette, "Practical Phase-Locked Loop Design," *ISSCC Tutorial*, 2004.

[122] S. Cova, S. Member, A. Lacaita, and G. Ripamonti, "Trapping Phenomena in Avalanche Photodiodes on Nanosecond Scale," *IEEE Electron Device Letters*, vol. 12, no. 12, pp. 685–687, 1991.

[123] J. C. Jackson, D. Phelan, A. P. Morrison, R. M. Redfern, and A. Mathewson, "Characterization of Geiger Mode Avalanche Photodiodes for Fluorescence Decay Measurements," in *Proceedings of SPIE Vol. 4650, Photodetector Materials and Devices VII*, vol. 4650, pp. 55–66, May 2002.

[124] F. Villa, B. Markovic, S. Bellisai, D. Bronzi, a. Tosi, F. Zappa, S. Tisa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "SPAD Smart Pixel for Time-of-Flight and Time-Correlated Single-Photon Counting Measurements," *IEEE Photonics Journal*, vol. 4, pp. 795–804, June 2012.

[125] D. Magde, G. E. Rojas, and P. G. Seybold, "Solvent Dependence of the Fluorescence Lifetimes of Xanthene Dyes," *Photochemistry and Photobiology*, vol. 70, no. 5, p. 737, 1999.

[126] D. D.-U. Li, J. Arlt, D. Tyndall, R. Walker, J. Richardson, D. Stoppa, E. Charbon, and R. K. Henderson, "Video-rate fluorescence lifetime imaging camera with CMOS single-photon avalanche diode arrays and high-speed imaging algorithm," *Journal of Biomedical Optics*, vol. 16, no. 9, p. 096012, 2011.

[127] A. Danowitz, K. Kelley, J. Mao, J. P. Stevenson, and M. Horowitz, "CPU DB," *Communications of the ACM*, vol. 55, p. 55, Apr. 2012.

[128] P. Vu, B. Fowler, C. Liu, S. Mims, P. Bartkovjak, H. Do, W. Li, J. Appelbaum, and A. Lopez, "High-dynamic-range 4-Mpixel CMOS image sensor for scientific applications," in *Proceedings of SPIE Vol. 8298, Sensors, Cameras, and Systems for Industrial and Scientific Applications XIII* (R. Widenhorn, V. Nguyen, and A. Dupret, eds.), vol. 8298, pp. 82980D–82980D–10, Feb. 2012.

[129] P. J. Brown and A. L. Chan, "IC Package Stiffener with Beam," 2011.

[130] K. Chun Hung Lin, K. H. I Zeng Lee, K. Su Tao, and T. Kun-Ching Chen, "Ball Grid Array Semiconductor Package with Improved Strength and Electric Performance and Methodr Makingthe Same," 2003.

[131] PCI-SIG, "PCI Express Base Specification Revision 1.1," 2005.

[132] A. X. Widmer and P. A. Franaszek, "A DC-Balanced, Partitioned-Block, 8B/10B Transmission Code," *IBM Journal of Research and Development*, vol. 27, pp. 440–451, Sept. 1983.

[133] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Applied Optics*, 1996.

[134] J. Corbet, A. Rubini, and G. Kroah-Hartman, *Linux Device Drivers*. 3rd ed., 2005.

[135] D. R. Butenhof, *Programming with POSIX Threads*. Addison-Wesley professional computing series, Addison-Wesley, 1997.

Appendix A

Single-Photon Avalanche Diode Test Chips

During the development of the 0.13μm CMOS SPAD used in this work, a number of test designs were fabricated. A brief summary of these designs and the knowledge gained from each are presented here. In total, 5 SPAD test chips were fabricated in both 0.13μm IBM and 0.35μm AMS CMOS processes. The result of this work was a low-noise SPAD in each of these technologies. In addition to the SPAD designs, the techniques learned in developing these unique structures were leveraged to assist in the design of junction field effect transistor (JFET) structures in 0.18μm IBM CMOS with the hope of achieving lower flicker noise levels than those of MOSFETs of a comparable size. Each of these efforts is described in the following sections.

A.1 First SPAD Test Chip - 0.13 μm IBM CMOS

The devices on this test chip were designed with the help of David Schwartz. The approach used in this design was to create the maximum possible separation between the STI and the active region of the device. Similar to the design in Section 3.5, the BN mask was used to block the n-type implant in the active RX region. The OP mask was used to block the

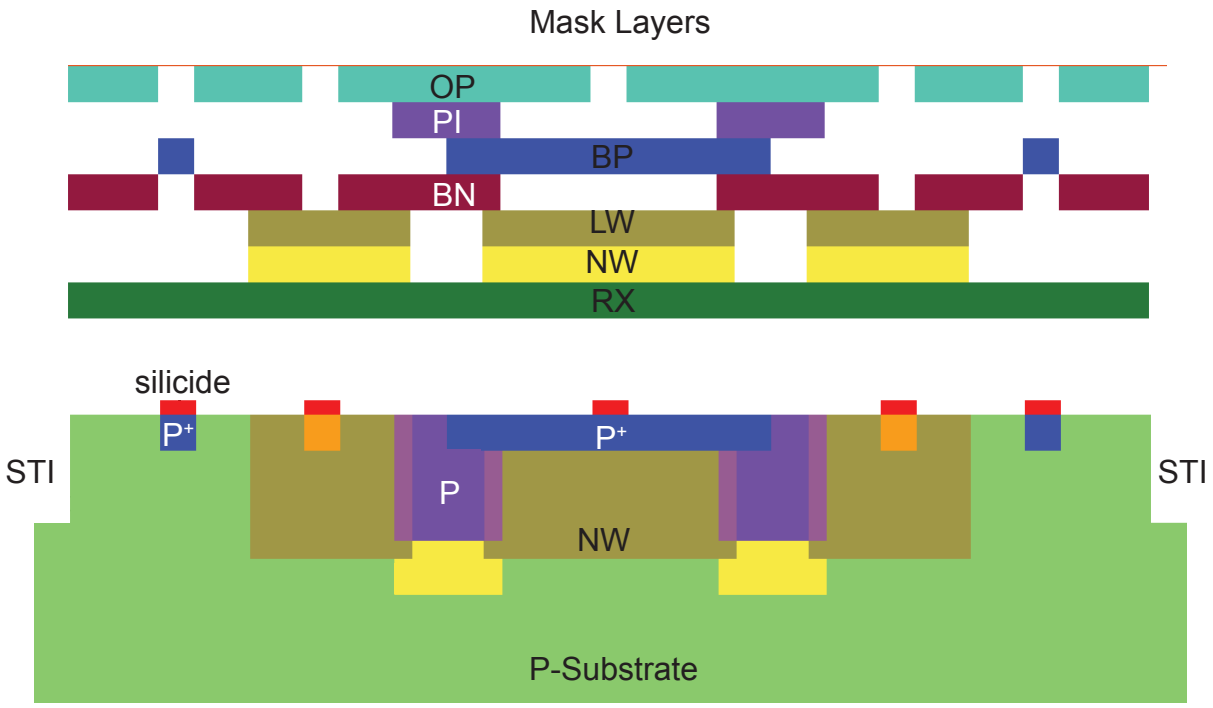

Figure A.1: The mask layers and expected cross-section for the first test chip. The LW layer was used in this design to lower the doping in the N-well.

salicide between the n-type and p-type regions. A diagram of the mask layers used for this design and the expected cross-section are shown in Figure A.1. Several variations based on this structure were designed into the test chip.

A representative I-V characteristic for diodes on this test chip is plotted in Figure A.2. All of the diodes demonstrated only a weak rectifying behavior indicative of a shunt path in parallel with the diode. This shunt path was ultimately isolated to a path between the P^+ connection and the substrate. The major problem that was identified from this test chip was that an additional VAR layer must be used to block the implants for enhancing MOSFET performance that are associated with every RX region.

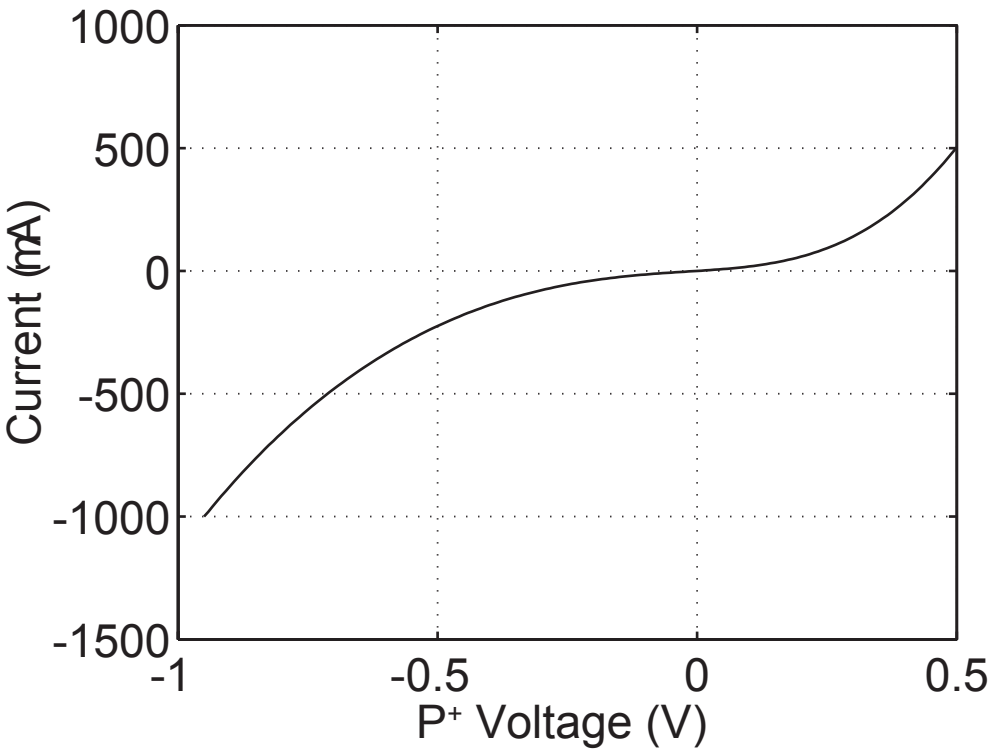

Figure A.2: The I-V characteristic is indicative of a shunt path across the diode.

A.2 Second SPAD Test Chip - 0.35 μm AMS CMOS

Following the first failed test chip in the IBM 0.13μm process, a test chip in an AMS 0.35μm high-voltage process was designed because it was well known that low-noise SPADs could be developed in this technology. The SPAD design was based on the one used in David Schwartz's thesis, which was provided by Edoardo Charbon. The goal of this design was to correct what appeared to be a punch-through problem due to insufficient spacing between the N-well and substrate contacts. However, these devices showed odd I-V characteristics (Figure A.3) and high levels of DCR.

A.3 Third SPAD Test Chip - 0.13 μm IBM CMOS

For the second SPAD test chip in the IBM process, the VAR layer was included over all devices in order to prevent halo and lightly-doped drain implants. Again, a number of

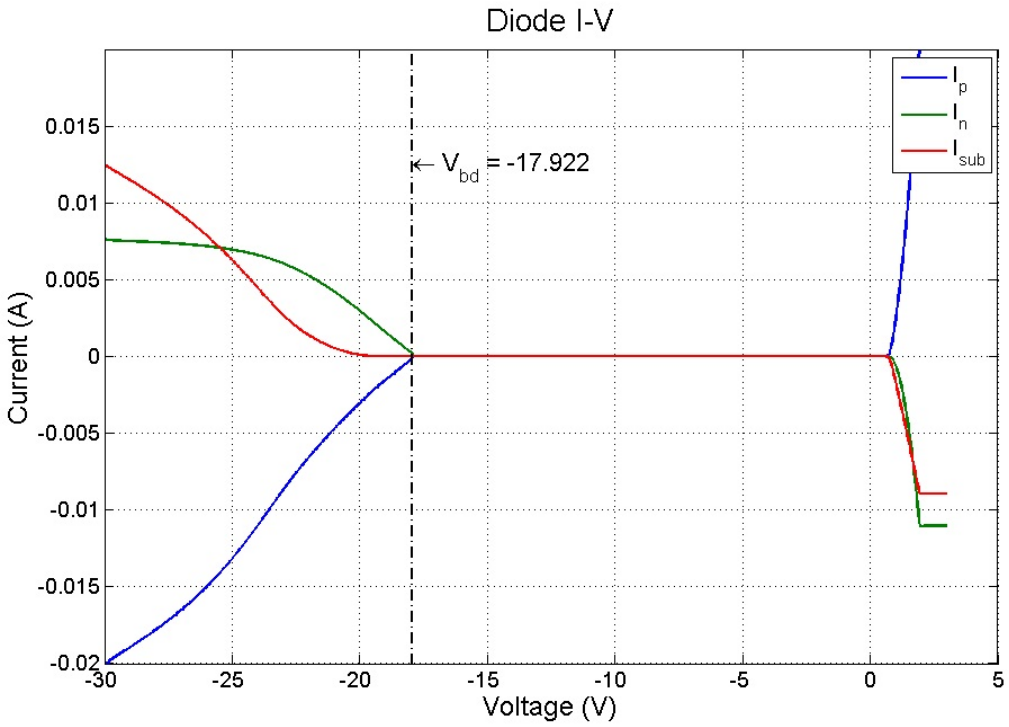

Figure A.3: The I-V characteristic shows a reasonable breakdown voltage but exhibits an odd behavior in the substrate current. These devices also had DCR levels that were greater than 100 kHz.

geometries and layer combinations were tried with the goal of designing a SPAD without STI near the multiplication region. An image of the chip layout showing the 74 different variants that were designed can be found in Figure A.4.

A representative I-V curve for a SPAD from this test chip is plotted in Figure A.5. Again, the diode shows weakly rectifying behavior. The non-linear response of this device is better than the first IBM test chip but the characteristics of these devices are still undesirable. It was postulated from these measurements that there is a surface shunt path that is connecting the terminals of the device other than the salicide, which should be blocked by the OP mask.

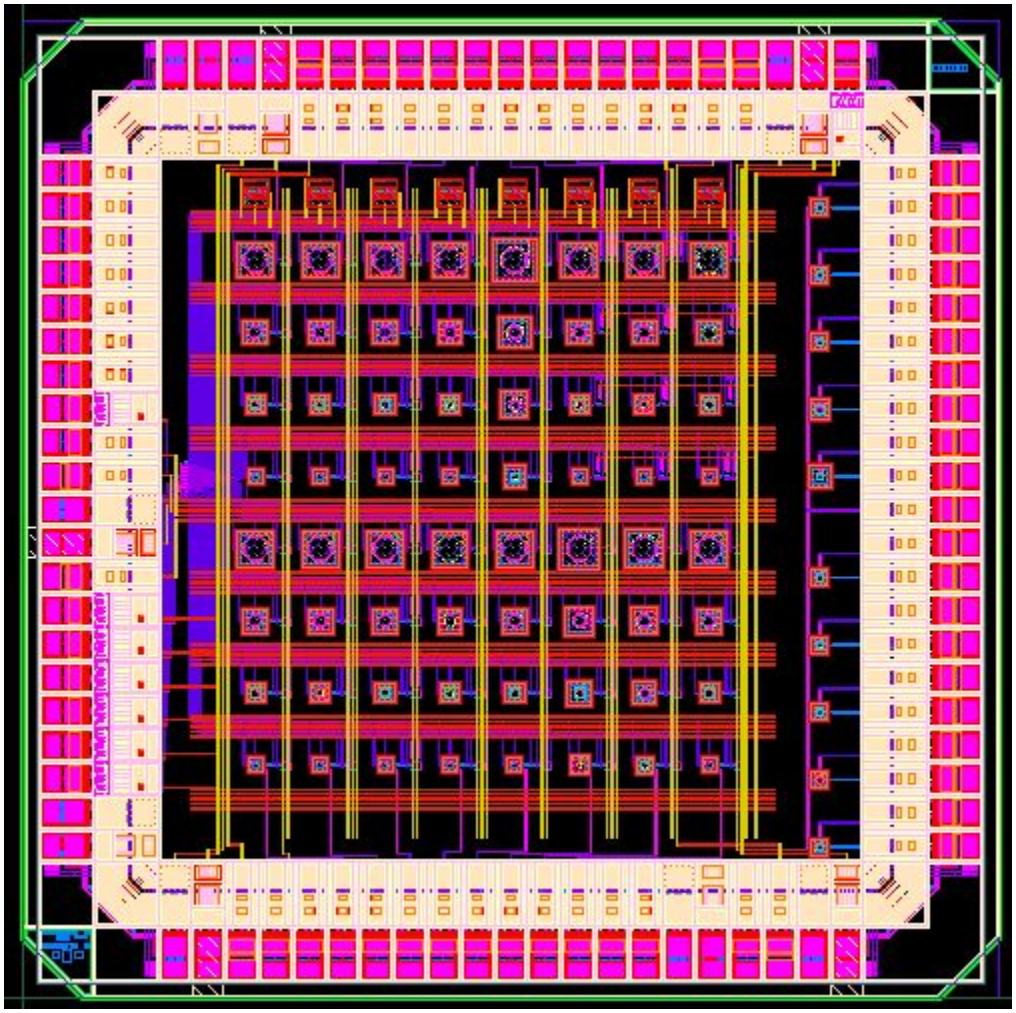

Figure A.4: Chip layout for the second test chip.

A.4 Fourth SPAD Test Chip - 0.35 μm AMS CMOS

With the improved understanding of the additional implants associated with the "active" mask layers that was gained from the IBM test chips, it was discovered that the FIMP layer in the AMS technology plays an important role in blocking implants. Consequently, a new set of devices was fabricated with almost all of them showing favorable I-V characteristics. The success of these devices coincided with the success of the fifth IBM test chip and, consequently, focus on the 0.35μm SPADs was dropped. However, Haig Norian saw an opportunity to incorporate these devices into his quantitative polymerase chain reaction IC. Details of the device performance can be found in reference [96].

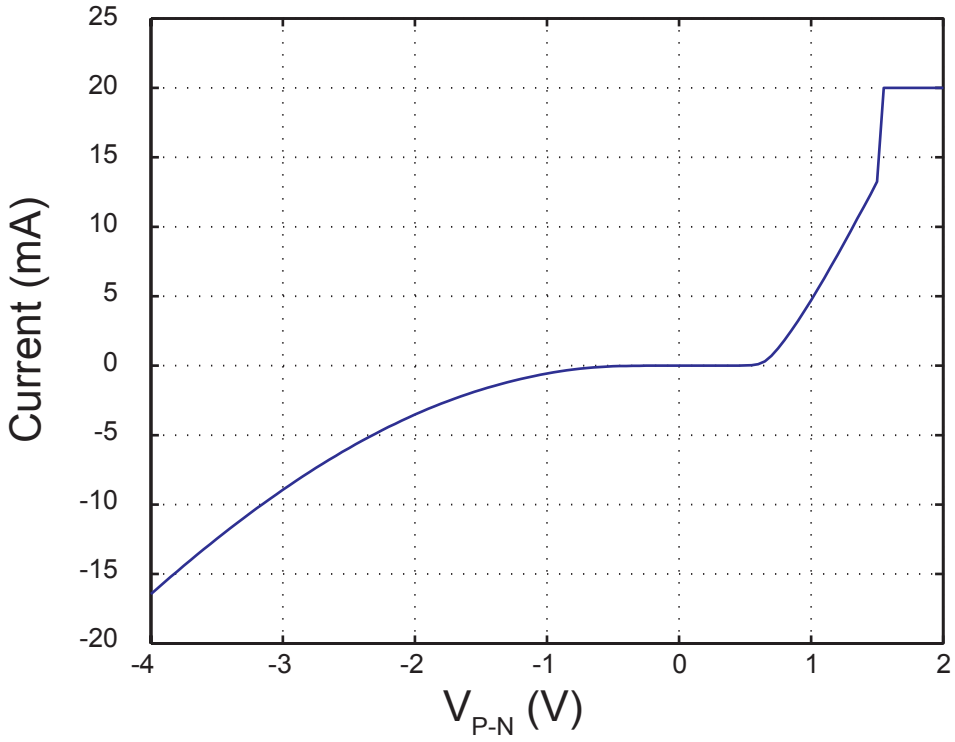

Figure A.5: Representative I-V characteristic for a SPAD from the second IBM test chip.

A.5 Fifth SPAD Test Chip - 0.13 μm IBM CMOS

In the third IBM test chip, a number of techniques were tried to isolate the terminals of the diode and disrupt the shunt path observed in previous test chips. These included using poly traces (with the corresponding gate oxide as a blocking layer), adding OP layers, and separating the RX so that the STI can also be used to separate the sides of the P-N junction. In addition, several devices using the IBM true triple well (T3) process were designed. Of the several varieties of devices that were fabricated ten varieties successfully behaved as diodes. The successful designs included a T3 isolated SPAD, a Poly isolated SPAD, a SPAD with STI touching the multiplication region, and a SPAD with STI in the guard ring close to the multiplication region.

After identifying the SPADs with rectifying I-V characteristics, they were each evaluated for their DCR performance. The SPAD that was presented in Chapter 3 had significantly lower DCR than any of the other devices. Surprisingly, this device had STI in

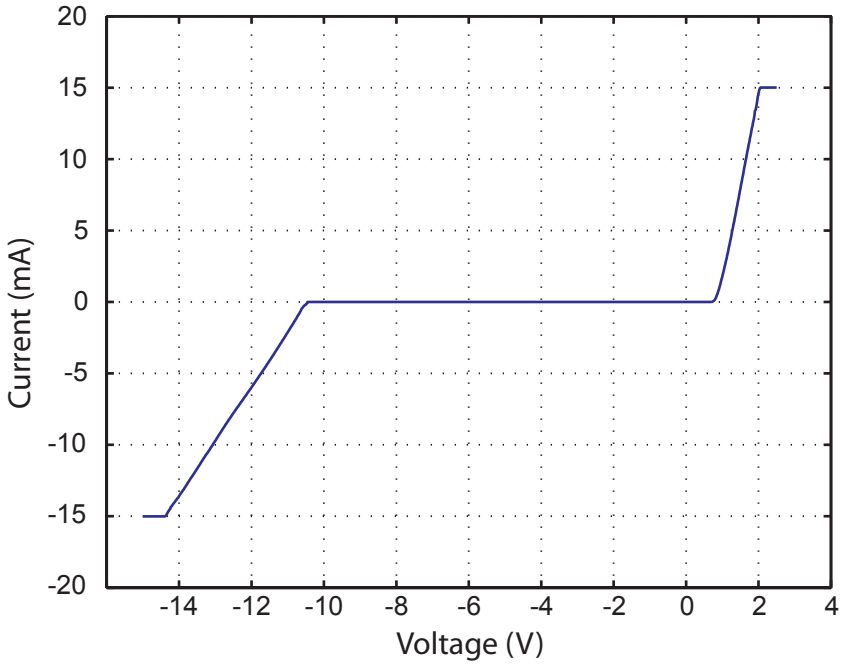

Figure A.6: Representative I-V characteristic for a SPAD from the third IBM test chip.

the guard ring near the multiplication region. This device exhibited outstanding noise and sensitivity performance and variants of this geometry were included on the imager IC presented in Chapter 4 with probing pads to evaluate spacing requirements for future designs. These development devices were tested (and passed) for the expected I-V characteristics but were not characterized for noise or sensitivity. Among these devices are geometries with tighter spacing between the SPAD and surrounding guard rings. These would be of particular interest for future SPAD array designs where density or fill-factor would be an important constraint.

A.6 FLIM Array Test Sites

In addition to the SPAD test sites on the imager IC that were mentioned in Appendix A.5, there were also junction field effect transistor (JFET) test structures that were designed using the true triple well (T3) implants of the technology. I-V characterization of these devices revealed that there was either a shunt path across the device or that the channel was too wide

to pinch off. The I-V results followed the trend of an ohmic device. Additional device test structures were placed on the January 2012 version of the imager. These structures included metal waveguide structures that were designed to route single photons to the SPADs and were designed by Matt Trusheim but never tested.

A.7 JFET Test Chip - 0.18 μm IBM CMOS

Building on the experience gained in adapting CMOS design masks for unconventional structures, an attempt was made to modify existing JFET structures in a 0.18μm technology to have improved flicker noise. The approach taken was to remove as much STI from the channel as possible to eliminate trapping sites that could contribute to low frequency noise. The methodology for removing the STI in this process is similar to that used for the SPADs. Active RX area is drawn to prevent introduction of STI. The OP and poly masks are used to prevent conduction paths across the surface of the silicon due to salicide formation. Combining these techniques will allow for separate P and N type regions within the same contiguous RX area. Additionally, a BP2ND layer is available in the 0.18μm technology that can explicitly block both the N^+ and P^+ implants. Dan Fleischer headed the fabrication and testing of these devices. Early measurement results indicate that these JFETs have reduced flicker noise as compared to similarly sized MOSFET devices.