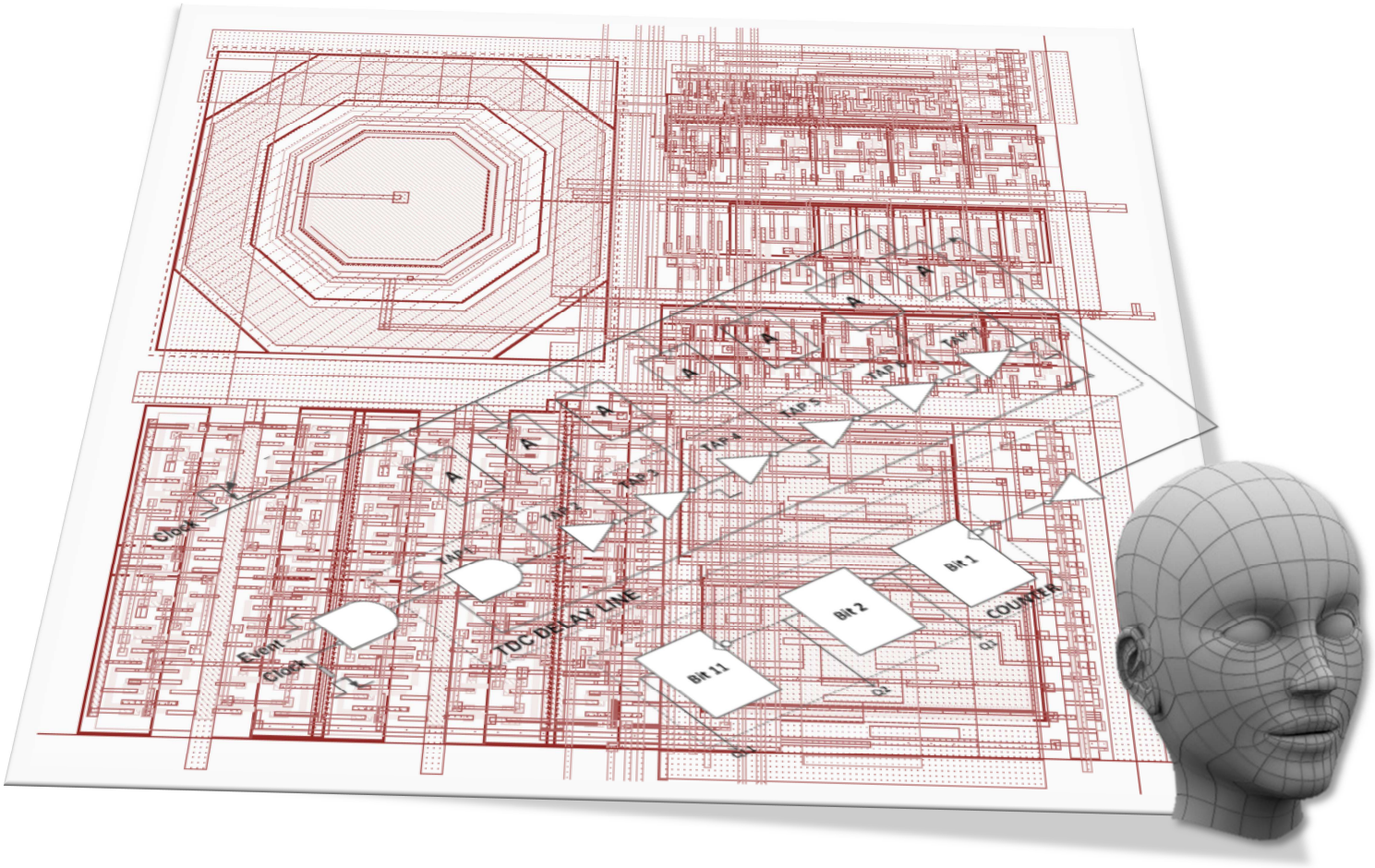
Time-of-Flight 3D Imaging based on a SPAD-TDC Pixel Array in Standard 65 nm CMOS Technology

Priyanka Kumar



TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science

Delft University of Technology

Priyanka Kumar: *Time-of-Flight 3D Imaging based on a SPAD-TDC Pixel Array in Standard 65 nm CMOS Technology,* September 2011

Time-of-Flight 3D Imaging based on a SPAD-TDC Pixel Array in Standard 65 nm CMOS Technology

Thesis

submitted in partial fulfilment of the

requirements for the degree of

Master of Science

in

Microelectronics

by

Priyanka Kumar

born in Jhansi, India

Electronics Research Laboratory
Department of Microelectronics & Computer Engineering
Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology

Several concepts used in this thesis are protected by patents.

Delft University of Technology

Department of

Microelectronics & Computer Engineering

The undersigned hereby certify that they have read and recommend to the Faculty of Electrical Engineering, Mathematics and Computer Science for acceptance a thesis entitled **"Time-of-flight 3D Imaging based on a SPAD-TDC Pixel Array in Standard 65 nm CMOS Technology"** by **Priyanka Kumar** in partial fulfillment of the requirements for the degree of Master of Science.

Dated: September 9, 2011

Chairman and Co-Advisor:

prof. dr. ir. Edoardo Charbon

Co-Advisor:

dr. R. Bogdan Staszewski

Committee Members:

dr. ir. Nick van der Meijs

Andre Borowski

Abstract

The interest in high performance three-dimensional (3D) imaging has grown in recent years due to immense demand in engineering, science, medicine and entertainment domains. The driving goals of state-of-the-art 3D imagers are high sensitivity to light and fine depth resolution at long range. Furthermore, high level of integration is desirable to achieve low system cost. To meet these demands, the industry has started to transition from traditional analog techniques to standard CMOS based solutions.

In this thesis, a novel time-of-flight 3D CMOS imager is proposed. The focus of the design is low power consumption while maintaining human-eye safety requirements. The imager targets security applications, primarily facial recognition, but is also suitable for automotive vision and robotics. In this work, a prototype of a 32x32 pixel array is designed, where each pixel consists of a single-photon avalanche diode as photodetector and a time-to-digital converter (TDC) for fast image acquisition. The imager is expected to achieve millimeter-level depth resolution for range as long as 30 m and has a maximum frame-rate of 1000 fps. Each pixel is constructed within 25x25 μm^2 area and has a fill factor of 5.76%. The layout implementation has been carried out in a 65 nm CMOS technology and would be the first of its kind at this process node.

To my loving parents,

Contents

List of Figures

List of Tables

Acknowledgments

Working on my thesis has been a fascinating and an extremely rewarding experience. As I bring a closure to my research over the past one year, I agree with Sir Isaac Newton in saying that I was able to see so far only because I was standing on the '*shoulders of giants*'. While the whole year was full of challenges and long working hours, it has been enveloped with the support and guidance of many a people. The new learnings from each of them has enriched me both personally and professionally.

First and foremost, I would like to express my utmost gratitude towards my supervisors - Prof. Edoardo Charbon and Dr. R. Bogdan Staszewski for giving me a wonderful opportunity to work on an interesting topic. Prof. Charbon's experience and knowledge in 3D imaging added a great deal to my learning in the past one year. His strong vision always inspired me to achieve better and his continuous motivation boosted my confidence greatly. Under his guidance, I have learnt to be meticulous and have practical thinking in my work. Dr. Bogdan gave me an initial motivation to work on this project and it proved extremely valuable to my graduate experience. The brain-storming sessions with him have enhanced my creative side in work. His industry exposure added a perspective of competitiveness in the project and his advice on corporate life would surely help me in my career ahead. I am fortunate to have received such supervision and I can say that I have learnt to dream bigger.

I would like to thank Andre Borowski for his idea to build an ambitious 3D camera and his financial support over the last one year. It really helped me to focus better to reach the desired goals.

I am thankful to Dr. Nick van der Meijs for serving on my defence committee and giving me an opportunity to work as his teaching assistant. It was an enriching experience which showed me the other side of student world. I am also grateful to Prof. Kofi A. Makinwa for his career guidance at the time I needed the most; it instilled confidence in me to make good decisions.

I am grateful to Antoon Frehe and Atef Akhnoukh for their technical support. It eased my working with 65nm technology in the initial phase of thesis project. I am also thankful to Marion de Vleiger for her warm smile every morning and her help in the administrative issues.

I would like to acknowledge the contributions of Yuki Maruyama, HyungJune Yoon, Shingo Mandai and Chockalingam Veerappan in SPAD related work. The technical discussions with them helped a great deal in my circuit design. Thanks to Matthew Fishburn for patiently solving all my doubts and for his valuable comments on my thesis report.

I am also grateful to my colleagues in ELCA group - Wanghua, Senad, Vincent, Wannaya, Akshay and Yi for their encouragement over last twelve months and their help in smallest of issues. The humorous coffee time discussions with them always left me with more energy to work. A special thanks to all my friends especially Harsh, Shishir, Vishwas and Parag for supporting me during my M.Sc. study and for good memories that I will always cherish.

One person whom I cannot thank enough is my dearest friend - Gaurav. It would have been difficult to understand electronic design without those endless technical discussions. A source of strength and encouragement, he believed in me always. I really hope we can together achieve the best in life.

I am truly amazed by the patience and encouragement from my family during the past few years. To venture out far from one's own land is not easy; and I was able to do so only because of my wonderful parents. My heart-felt gratitude for all the sacrifices they made, their words of inspiration and their belief that I could excel. My study and stay in Netherlands would not have been possible without their continuous support. I extend the same thankfulness to my little brother, my elder sister and my brother-in-law for always being there for me and motivating me during my study at Delft. Also to my nephew, Aditya, who taught me how simple life can be. A special thanks to Skype and Google technologies - for they allowed remaining in touch with the dear ones and reduced distances to a great extent.

Finally, I would dedicate my thesis to my both sets of grand-parents who touched my life in countless ways. I wish I could spend more time with them...

Acronyms

2D	Two Dimensional
3D	Three Dimensional
ADC	Analog-to-Digital Converter
APD	Avalanche Photo-diode
CCD	Charge Coupled Device
CMOS	Complementary Metal Oxide Semiconductor
CW	Continuous Wave
DCR	Dark Count Rate
DNL	Differential Non-linearity
DL	Delay Line
DSM	Deep Sub-micron
DFF	D Flip-flop
DLL	Delay Locked Loop
FWHM	Full Width at Half Maximum
INL	Integral Nonlinearity
LPI	Local Passive Interpolation
PDP	Photon Detection Probability
PMT	Photon Multiplier Tube
PET	Positron Emission Tomography
PVT	Process, Voltage and Temperature
PLL	Phase Locked Loop
ToA	Time-of-Arrival
TCSPC	Time Correlated Single Photon Counting
TDC	Time-to-Digital Converter
TFF	Toggle Flip-flop
ToF	Time-of-Flight
TG	Transmission Gate
SPSD	Single Photon Synchronous Detection
SPAD	Single Photon Avalanche Diode
VDL	Vernier Delay Line

1. Introduction

The ability to reproduce and extend the natural senses has always fascinated mankind and it has arguably been the most important factor for development of sensors and related technologies. One of the growing areas of interest is that of 3D imaging which is gaining popularity for its ability to capture the objects as we see them in the real world. Although 3D imaging techniques have existed for several decades, it has been mostly restricted to research domain and low volume applications till recently. The recent advent of low-cost solid-state 2D imagers along with technological progress has propelled the idea to build low-cost and compact three-dimensional (3D) imaging systems. Such systems have increasing demand in varied domains ranging from face recognition to machine vision, from 3D gaming to nuclear security and from molecular imaging to land and sea surveying. The diverse scope of applications makes the field of 3D imaging even more challenging as every application has stringent requirements in terms of performance and cost.

1.1 Applications of 3D Imaging

The applications for depth map acquisition and evaluation are widespread. A depth map is an image consisting of information pertaining to the distance of the objects in a scene from a certain viewpoint. A 3D imaging imager creates depth maps through range estimation of the contours on the object. The challenge is to build a very fine depth measurement system in order to detect the surface reliefs on the object. In this section, some of the main applications of 3D imaging are presented.

1.1.1 Machine Vision

One of the important functions of an imager is related to enhanced vision. While 2D cameras existing in market boast of high resolution, they cannot capture depth information of an object. The depth information is desirable for machine vision in several domains, the most notable being the automotive industry where 3D vision can be employed for pedestrian detection, rear vision for parking assistance, lane departure warning and adaptive cruise control. However, such applications are very sensitive to robustness and cost of the solution and therefore, there has not yet been widespread adoption of 3D vision systems in automobile industry.

3D vision is also useful in robotics to assist in better maneuvering of autonomous robots. Conventionally, ground robots rely only on 2D sensors to estimate their position with respect to environment and avoid collisions. But, the use of 2D sensors restricts the field of view to a single plane parallel to the floor. To enhance robot vision systems, multiple sensors need to be integrated to produce denser data. With use of 3D imaging, the process of depth map creation, self-localization and collision avoidance can be made more accurate [1]. Moreover, solid-state 3D imagers provide wide field of view along with compactness, speed and low power consumption, making it perfectly suitable for autonomous robots.

1.1.2 Security

One of the main advantages of 3D imagers in security is their ability to perform in dim-lit environment. Enhanced vision is desirable in security or surveillance applications for physical access control in homes, banks or production plants, border control surveillance, baggage checks at airports and e-passport programs. At present, CCTV cameras are generally employed for surveillance activities. A 3D camera adds value to the existing system by giving depth information to estimate the exact position of objects in focus. Similarly, 3D face recognition systems can boost the security in access control applications.

1.1.3 Human-Computer Interaction

Advancements in 3D vision can lead to better experience in the human-computer interaction domain. The existing solid-state 3D sensors are capable of creating real-time depth maps and multiple functionalities can be built upon them to boost the user experience in consumer applications. Recognizing user gestures in gaming, virtual keyboards, mouse, and interactive home entertainment devices are some of the applications that can benefit from low-cost 3D imagers.

1.1.4 Biomedical Imaging

A number of medical applications based on time-resolved optical imaging have benefited from 3D imaging. In particular, sub-nanosecond timing resolution along with high sensitivity is useful for specific applications such as positron emission tomography (PET), fluorescence lifetime imaging microscopy (FLIM) and fluorescence correlation spectroscopy (FCS).

While the above discussed applications show the present scope of 3D imaging, the possible applications for 3D imaging will only grow with availability of low-cost solutions.

1.2 3D Imaging Techniques

1.2.1 General Classification of Optical 3D Image Sensors

The range-finding and acquisition of 3D images is enabled by different optical techniques. Various authors have presented several classifications based on the different physical principles involved. Besl [2] identifies six basic optical techniques which are shortly described below:

Triangulation

It is the oldest technique for measuring distance to a point in the scene. It determines the location of a point by measuring angles to it from two known points on either end of a known baseline. The point on the object is fixed as the third point of a triangle with one known side and two known angles (see Figure 1.1a). Nearly all animals obtain 3D information based on triangulation technique and it has been widely used in applications related to mobile robotics, navigation, surveying and 3D photography. The disadvantages of

this technique are that its measurement range is limited to millimeter range only and it has high dependency on the source of illumination.

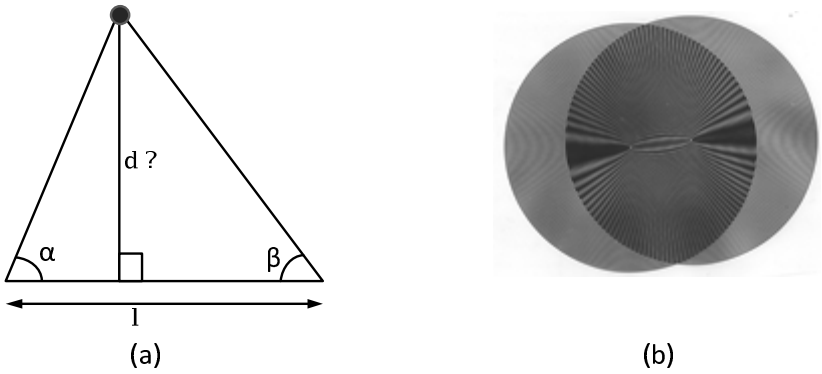


Figure 1.1 - (a) Triangulation method used to calculate coordinates and distance of a point from baseline; (b) Moiré pattern.

Moiré

This technique extracts shape of an object from the interference patterns. A Moiré pattern is a superimposed image of highly regular interference patterns which differ in relative angle, size or spacing. An example of such a pattern is shown in Figure 1.1b. This method is well known for applications which need high resolution for smooth surfaces [1] but qualifies for limited dynamic range applications only.

Holographic Interferometry

This technique has the ability to measure displacements in the order of a wavelength of light or better. Thus, it provides highest absolute accuracy and resolution. The setup, as shown in Figure 1.2, includes coherent light from laser sources which produces interference patterns due to phase-frequency differences in different optical paths [3], [4]. This produces two slightly different scenes with minute differences between them. It is a powerful scheme and is widely used in identification of vibrational modes, surface displacements, and motion geometries. Nevertheless, this technique is only suitable for short range applications (in the order of micrometres to millimeter range).

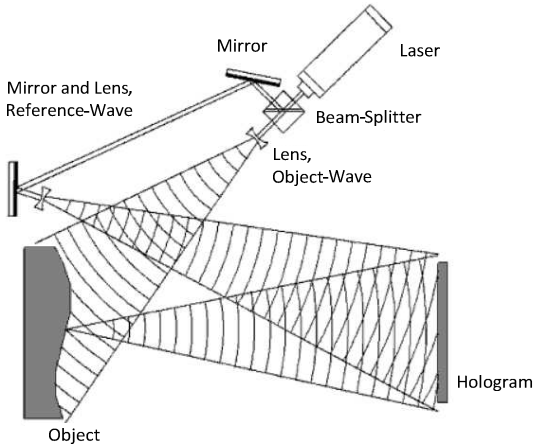


Figure 1.2 - Basic holographic setup [3].

Focusing

Various techniques exist in the literature for this approach to 3D imaging [5]. All of them share the basic concept of optical physics of imaging: the distance from the optical centre of the lens to a path on 3D scene is related to the degree of sharpness or blurring of the corresponding components of the image [2].

Fresnel Diffraction

It refers to the modification in the direction and intensity of a group of waves after passing by a slit or through an aperture whose size is of the order of the wavelength of the waves [2], [6].

Radar or Time-of-Flight

Acronym for 'Radio detection and ranging', this technique dates back to early 1900s when Christian Hülsmeyer first experimented with radio waves to detect the presence of ships in dense fog conditions. The underlying principle of radar based imaging is to measure the time elapsed between an emitted pulse of light and its reflected pulse from the object. It is suitable for objects at large distances, in the range of kilometres depending upon the wavelength used. The basic range equation for radars is:

$$v \cdot \tau = 2 \cdot d \tag{1.1}$$

where 'v' s the velocity of signal propagation, 'τ' is the transit time taken for the pulse to hit the object and return and 'd' is the distance to the object. This technique is also known as Time-of-Flight (ToF).

Apart from [2], several other classifications of optical 3D image sensors exist in the literature. In [5], Jarvis classifies the range sensing methodology in three sets:

(a) active vs. passive – based on illumination technique e.g. ambient lighting or structured laser source;
(b) image based vs. direct – based on if image analysis is done for range extraction;
(c) monocular vs. multiple view – based on if single or multiple viewpoints are considered for range determination.

Originally, 3D imaging systems were based on non-solid-state devices. However in recent times, design of 3D imagers has significantly evolved due to the advances made in IC technology. Solid-state implementation of imagers allows low system cost, high lateral and depth resolution, short acquisition time, insensitivity to mechanical stress along with fast image processing. A discussion is made in the next section about the state-of-the-art solid-state ToF 3D imagers.

1.2.2 Solid-State Time-of-Flight Imagers

Much work has been carried out in the field of solid-state ToF 3D imagers [7], [8]. To analyze these contributions to the art, it is often useful to classify the existing work in ToF imagers as a function of (i) the type of laser modulation, (ii) the ToF principle, and (iii) the nature of

photo-detection. Although a mix of techniques is possible, a simple classification is made and is shown in Figure 1.3.

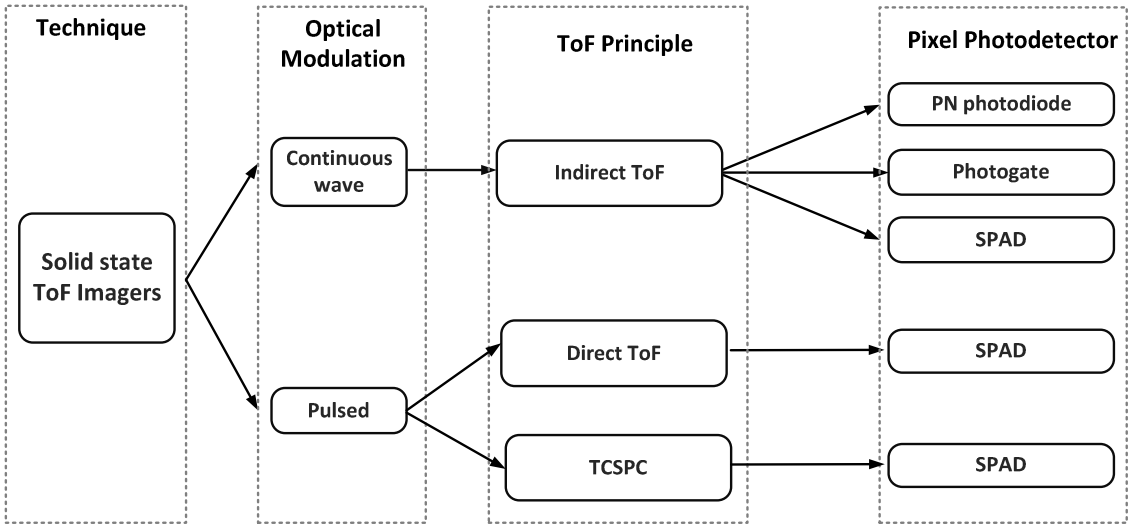


Figure 1.3 - Classification of state-of-the-art solid-state optical ToF image sensors [1][9].

1.2.2.1 Laser Modulation Techniques

The two common modulation techniques used in ToF image sensors are continuous wave (CW) modulation methods and pulsed modulation methods. Both modulation schemes require dedicated detector architecture, measurement logic and an evaluation scheme.

In CW modulation technique (also known as 'lock-in'), the modulation signal is a periodic optical signal, typically a sine wave [9], [10] whereas in the pulsed modulation technique, the modulation signal is a sequence of pulses. A third modulation technique called 'Pseudo noise modulation' is described in [9] and is very robust in nature. Here, the modulation signal is a coded sequence of pulses and can be used to avoid interference between pulses from different 3D cameras.

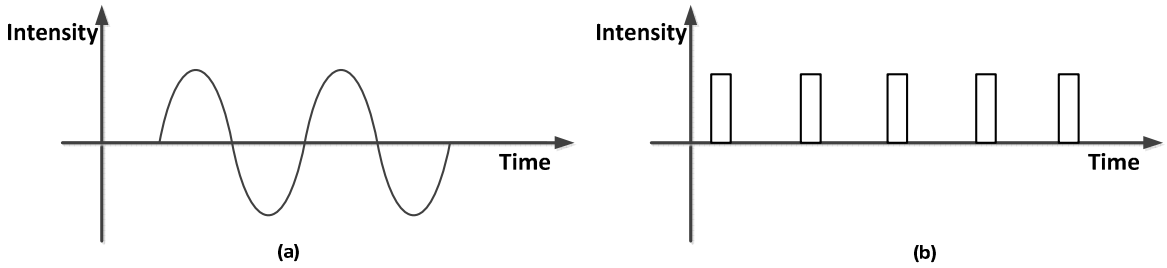


Figure 1.4 - Laser modulation techniques: (a) continuous wave modulation; (b) pulsed modulation.

1.2.2.2 ToF Measurement Principles:

The measurement principle defines the basic scheme and setup for ToF estimation. The main techniques used for ToF measurement are:

Indirect ToF:

It refers to the comparison of *phase* of the emitted modulating signal and the detected modulated signal. The range is measured by calculating the phase shift of the first harmonic in the returned wave and is estimated by:

$$z = \frac{c}{4\pi f}\varphi \tag{1.2}$$

where φ is the phase shift due to time of flight and *f* is the modulation frequency.

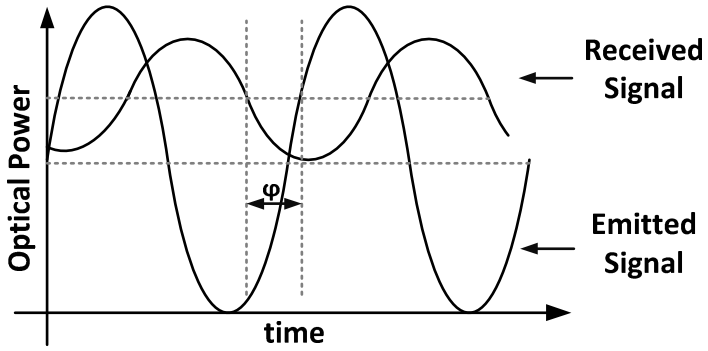


Figure 1.5 - Range estimation through phase measurement [9].

The first implementation of lock-in photo-gate structure based on phase shift measurement was reported in [11], [12]. The traditional indirect ToF methods are analog in nature. They are also vulnerable to uncorrelated background illumination. A digital approach to phase detection, called 'single photon synchronous detection' (SPSD) was proposed in [13]. It achieves enhanced phase accuracy compared to analog approaches but suffers from harmonic distortion in the illumination system.

Direct ToF:

As the name suggests, it refers to the direct ToF measurement of optical *pulses* sent by a light source. The range is determined by time measurement between emitted and received pulse and is given by:

$$z = c \cdot \frac{t_{TOF}}{2} \tag{1.3}$$

where c is the speed of light and t_{ToF} is the time interval between emitted pulse and received photon.

The illumination sources required by direct ToF need to have extremely fast rise and fall times. Additionally, very fast photodetectors and front end circuits are required to detect precise optical pulses. For instance, if a single-shot distance resolution of 1 mm (~3 ps) is needed to be measured, a clock frequency of ~150 GHz is required for the counter to measure ToF. This is a very tough requirement to meet and is close to impossible in the existing technology nodes. This principle also suffers from high sensitivity to background light when single shot measurements are required [1]. A solid-state 3D image sensor that works on direct ToF principle is reported in [14].

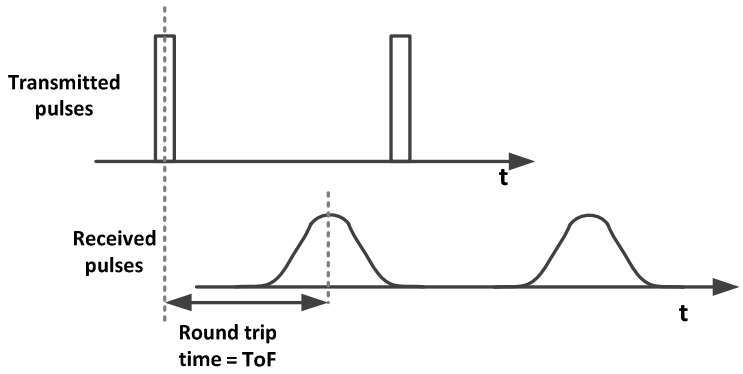


Figure 1.6 - Range estimation through time interval measurement.

Time Correlated Single Photon Counting (TCSPC)

One of the concepts in 3D imaging is based on correlation of time-of-arrival (ToA) of incident photons [1]. It is a pulsed based scheme specific to the detection of single photons and correlation of their ToA to obtain the distance information.

In this scheme, generally, multiple single-photon measurements are obtained and high resolution can be achieved with statistical processing. This helps to realize millimeter level accuracy even at low photon counts or in the presence of large background illumination [15].

TCSPC is useful for a variety of applications such as measuring the lifetime of a fluorescent marker while studying a biological specimen or for understanding the metabolic activity in positron emission tomography (PET). Apart from this, this approach finds various other applications such as optical range-finding [15], fluorescence correlation spectroscopy (FCS) [16], Förster resonance energy transfer (FRET) [17], and fluorescence lifetime imaging microscopy (FLIM) [18].

The work presented in this thesis extends the art in TCSPC applications. The proposed 3D imager is aimed at security and surveillance applications and the system objectives can be met using TCSPC approach to design. The system is required to be sensitive to wavelengths in non-visible range and must work for low photon counts or dim lit areas with high resolution at low power. The next chapter discusses the building blocks of TCSPC based 3D imagers.

1.3 Imager Performance Parameters

To evaluate the performance of state-of-the-art 3D imagers, it is important to understand the different performance parameters. Following is a brief description of the primary performance parameters.

1.3.1 Range (Z_{MAX})

It is the maximum range that the imager can capture an image from. This may vary from few millimetres to hundreds of meters depending upon the application. For instance, in surveillance activities at airport or shopping malls, large range is desirable to capture movement of people at distances as large as 20-30 m.

1.3.2 Depth Resolution (ΔZ)

Also known as "step-size", it is the smallest change in depth that the sensor can measure. In face recognition systems, a fine depth resolution is required for identification of facial contours. In modern imaging systems, the resolution may be of the order of micrometres to millimetres.

1.3.3 Lateral Resolution ($N_x \times N_y$)

This parameter is a measure of density of information points in a unit area. In a colour 2D camera, the information in each point is the colour and brightness while in a 3D camera, it is the depth resolution of the device.

In consumer applications, this is also known as dots per inch (DPI). A 100x100 pixel image printed on a 1-inch square sheet is said to have 100 DPI. Good quality photographs usually require 300 dots per inch or higher when printed.

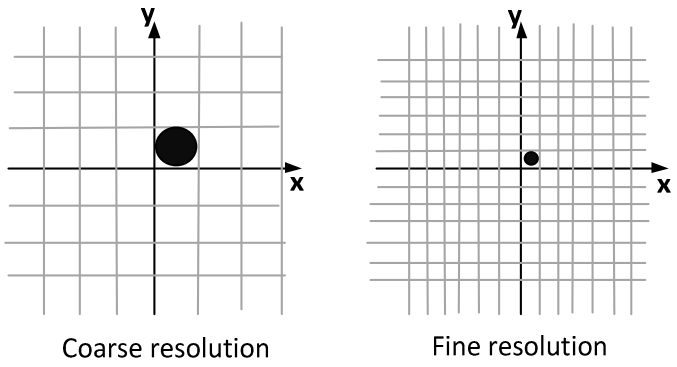


Figure 1.7 - Graphical representation of resolution [19].

1.3.4 Field of View

It refers to the linear or angular extent of the scene of interest that can be seen by the imager at a particular position and orientation in space [Source: Wikipedia].

1.3.5 Fill Factor

Fill factor in the context of imaging arrays is defined as the ratio of the photo-sensitive or active area to the total array area. Ideally, a 100% fill factor is desirable to avoid any information loss. However, high image acquisition rate mandates additional circuitry in the array which lowers the fill factor of an imaging array. Techniques such as back-side illumination allow 100% fill factor with photo-sensor and additional circuitry on different sides of wafer. But it involves a high manufacturing cost.

1.3.6 Repeatability

It refers to statistical variations when a sensor makes repeated measurements of the exact same distance under identical conditions. It indicates the reliability of the sensor and Figure 1.8 shows the basic concept of repeatability.

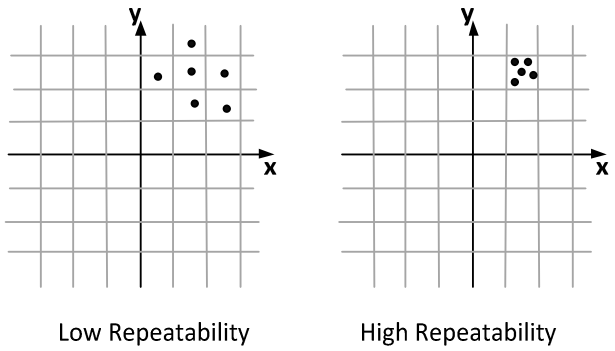


Figure 1.8 - Graphical representation of accuracy and repeatability [19].

1.3.7 Frame Rate (N fps)

The rate at which the moving target is captured and an image is formed is referred to as the frame rate. In other words, it is the number of images or frames that can be produced per second.

Human eye can typically resolve 24 frames per second and any higher frame rate makes the motion appear more fluid to the human eye. The typical video frame rate is 30 fps and higher frame rates are desired for real time applications such as steering a guided missile, surveillance/security systems and biomedical imaging. Gaming industry is also increasingly demanding higher frame rates for high end action games such as Star Wars: Battlefront, FIFA and QUAKE. Frame rate in an imaging sensor is determined principally by the number of pixels and the pixel readout speed through an IO pad. If the number of pixels is greater, the image produced is finer but it leads to lower frame rate. On the other hand, if the pixel readout rate decreases, frame rate is decreased as well.

1.3.8 Constraints:

The above described concepts describe the performance of the imager. However, it is also important to consider the constraints presented by an application. The design of 3D imaging systems can drastically vary if the constraints are altered. The common constraints for imaging systems are:

Cost

Low-cost is always desirable in consumer applications for higher profitability. With CMOS technology, higher level of integration is possible and mass production leads to low-cost systems.

Size

This parameter directly implies the area occupied by the system. Although size is less significant in some applications such as cars or space-vehicles, it is a bottleneck in handheld devices and digital cameras. Thus, a highly compact imager with varied features is the goal of any low-cost imager.

Robustness

The ability of the system to work at desired performance in spite of undesirable magnetic fields, radiation, electro-magnetic interference, external vibrations or mechanical stress defines the robustness of the system.

Eye - safety

Some 3D imaging techniques require structured lighting in the form of lasers. The human eye is extremely sensitive to laser radiation and it is of extreme importance to cater to safety requirements.

1.4 Motivation

The above discussion on the different performance parameters of 3D imagers helps to evaluate the performance of 3D imagers. In this section, the motivation to work on a ToF based 3D real-time imager is presented. The challenges faced during the execution of the project are also stated.

The CCD image sensors are a mature technology producing high quality images. But their high cost and low level of integration encourages the imaging industry to venture into CMOS based solutions. CMOS technology is cost-effective and allows the photodetector such as APDs or SPADs to be fabricated on a single chip along with imaging logic circuitry such as TDCs. Moreover, lithography and process control in CMOS fabrication have reached levels that allows CMOS sensor image quality to be comparable to that of CCDs.

The latest SPAD-TDC array reported in the literature uses 130 nm CMOS [8]. Although 90 nm SPADs had been developed, a SPAD-TDC imaging array in this technology does not exist in the literature till date. The state-of-the-art image sensors are summarized in Table 1.1.

In this thesis, an investigation of a 65 nm implementation of SPAD-TDC pixel array is carried out which is first of its kind. The small CMOS feature size allows better TDC resolution and therefore, a higher depth resolution. This also leads to a higher speed of image acquisition. Since CMOS dimensions are small, higher fill factor is possible. Furthermore, it leads to lower power consumption due to smaller supply voltage. However, design in deep submicron technology presents challenges in terms of parasitic dominance, performance degradation due to process variations and other related issues.

In this thesis, the challenge is to build a 3D imager specific to security applications. The goal of this work is to lay the ground-work for building a 1 Mpixel 3D real-time camera in 65 nm CMOS technology with less than 3 mm depth resolution. However, a 32x32 prototype is first implemented as a proof of concept and to verify the correctness of the proposed imager architecture.

Table 1.1 - State-of-the-art in 3D TCSPC (SPAD) imaging. The shaded cell indicates the best result of the summary for a given parameter.

	[20]	[7]	[21]	[8]	[22]
Technology	0.35 µ	0.13 µ	0.18 µ	0.13 µ	0.13 µ
Array Size	60x48	32x32	340x96	160x128	128x96
Range	5m	15m (100 ns)	128m	8m (55 ns)	45m
Resolution	38 mm (1σ distance resolution)	< 20 mm	< 32 mm	<8.5 mm	160 mm (1σ distance resolution)
Time Jitter at FWHM	-	185 ps	-	140 ps	
Frame Rate (fps)	-	-	10	25k (raw)	20
Median Dark Count Rate	245 Hz	-		Below 50 Hz	Below 100 Hz
DNL / INL (LSB)	INL: 110 mm upto 2.4m	±0.4/±0.2	-0.52/0.73	±0.3/±2	INL: 5 mm upto 2.4 mm
Chip power dissipation	35 mW	94 mW (78 mA)	-	550 mW	40 mW
Chip area	6.5 x 5.5 mm^2	-	3.3 x 3.3 mm^2	11 x 2.3 mm^2	7.2 x 7.5 mm^2
Pixel pitch	85 µ	85 µ	-	50 µ	44.65 µ
Type of TDC used	Phase domain approach	Coarse-fine	Flash TDC	Ring Oscillator	Phase domain ΔΣ approach
Fill factor	0.5%	-	70% (macrocells)	1%	3.17%
Year	2009	2009	2011	2011	2011

1.5 Design Methodology

A top-down design methodology was undertaken to realize the prototype. At first, an extensive literature survey of the existing work on TCSPC 3D imagers was carried out along with system analysis to understand the working and requirements of the system. The understanding gained from the study was used to draw out the overall system architecture, its different blocks, their functionality and the desired specifications. During this process, multiple new concepts were also incorporated to increase the efficiency of the system.

The system analysis was followed by circuit level implementation of the system. A bottom-up approach was employed in the circuit design in 65 nm technology. Each of the blocks identified in system level analysis was independently designed to meet the required specification. To achieve the desired performance after layout was one of the biggest challenges of the work. It required multiple iterations between circuit design and layout before the desired specifications could be realized.

The circuit and layout of the pixel was followed by integration at the system level. A modular approach to constructing a 2D array was chosen. The system level blocks were then integrated with the pixel array to realize the full implementation. At each step of circuit and layout design, the performance was verified against the desired system specifications. This

approach finally led to a successful design of the prototype which meets the desired system requirements.

1.6 Organization

The organization of the thesis report is along similar lines as the design methodology.

Chapter 2 explains the working of a SPAD photodetector and the conventional methods to build its front-end circuitry. It also describes the state-of-the-art TDC architectures and explains the choice of TDC architecture for a million pixel array.

Chapter 3 draws out the system specifications for a ToF-based 3D real-time camera. It also discusses the architecture of the imaging array and determines the required circuit performance.

Chapter 4 focuses on the design process of a single pixel with respect to desired system specifications. It also discusses the implementation issues at pixel as well as system level. Finally, the layout of the whole system is presented.

Chapter 5 demonstrates the design and post-layout simulation results obtained.

Chapter 6 concludes this dissertation providing the summary of the work done along with the contributions to the state of the art and its future scope.

2. Background

In this chapter, the background of the basic building blocks of a TCSPC SPAD based 3D imager is presented. A detailed discussion of imager architecture and pixel design is made in Chapter 3 and Chapter 4.

This chapter is organized as follows: Section 2.1 gives an introduction to the basic architecture of imager architecture. Section 2.2 describes the working of SPAD and how its front-end circuit is typically designed in the literature. Section 2.3 surveys various TDC architectures and discusses the choice of suitable TDC for a 3D imager.

2.1 Generic TCSPC Imager Architecture

A TCSPC approach to 3D imaging involves statistical processing on the collection of a large number of measurements. Such processing helps to achieve picosecond level time resolution. Although the traditional photomultiplier tubes and multi-channel plates are capable of high time resolution, they are relatively bulky and costly. Single photon avalanche diodes in a multi-pixel array are increasingly being adopted for their small cost and comparable performance. Along with a SPAD, a TDC with picosecond resolution is employed for time interval measurement and thus, the structure allows for high-speed image acquisition.

The block diagram in Figure 2.1 shows the basic architecture of a SPAD based TCSPC 3D imager. Each pixel consists of a photo-detector along with its front-end circuitry. In order to measure the time-of-flight (ToF) of photons hitting the imager, a time-to-digital converter is employed in every pixel. The pixel receives inputs from the global controller w.r.t. better control of interface circuitry and its readout. The data readout from the pixel undergoes signal treatment for final image reconstruction.

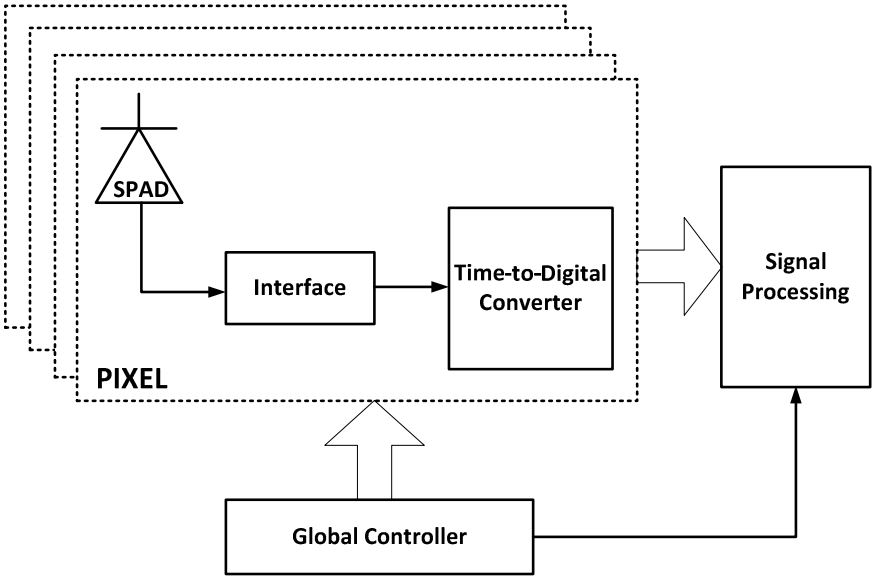


Figure 2.1 - Block diagram of SPAD based TCSPC 3D imager.

2.2 Single Photon Avalanche Diode

One of the integral blocks of any imager is a photodetector. In this work, a single-photon avalanche diode (SPAD) has been employed. A SPAD is a p-n junction diode biased above breakdown voltage V_{BD}, commonly known as operating in *Geiger mode*. When biased in this mode, a high electric field exists within the junction. If a photon hits the semiconductor surface and is absorbed in the diode's depletion region, it may trigger an avalanche producing electron-hole pairs by impact ionization which discharge the depletion region capacitance rapidly [10].

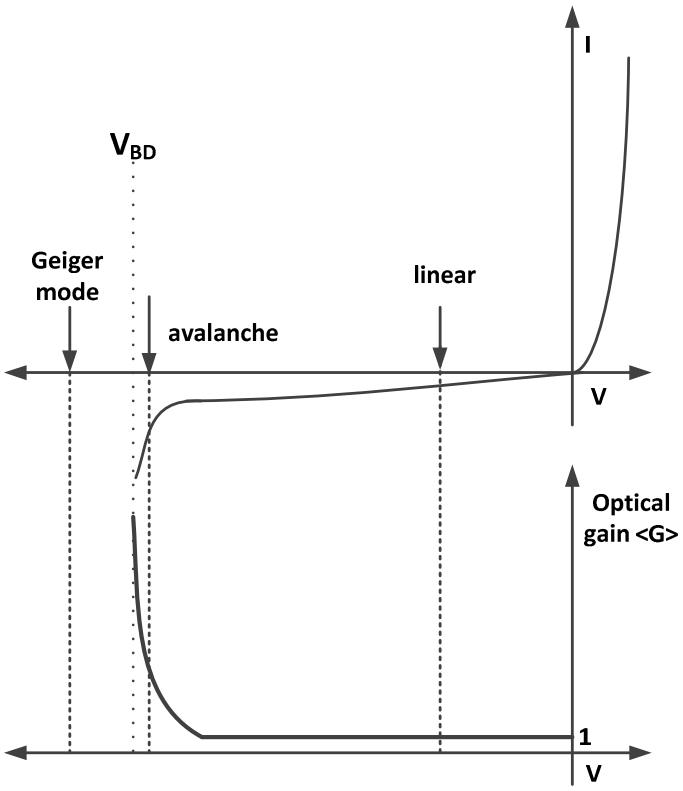


Figure 2.2 - I-V characteristic and gain of SPAD in Geiger mode of operation [23].

SPADs offer a near-infinite gain as compared to the linear gain avalanche photo-diodes (APDs), as shown in Figure 2.2. Furthermore, SPADs have shown to achieve picosecond level resolution [24]. Thus, SPAD acts like a trigger device for a single photon producing a sharp current pulse on being hit by a photon. This current pulse can be measured to determine the exact arrival time of photon in TCSPC applications.

2.2.1 Device Structure

The device structure used for building 65 nm SPADs is shown in Figure 2.3 [25]. The avalanche occurs in the depletion region on the edge of n+ layer and native p-substrate junction. It consists of a guard ring which helps to reduce electric field at the edges. The guard ring is built with low concentration n-well and helps to prevent edge breakdown. The

native p-substrate has low concentration allowing larger depletion width; thereby allowing higher sensitivity to longer wavelengths.

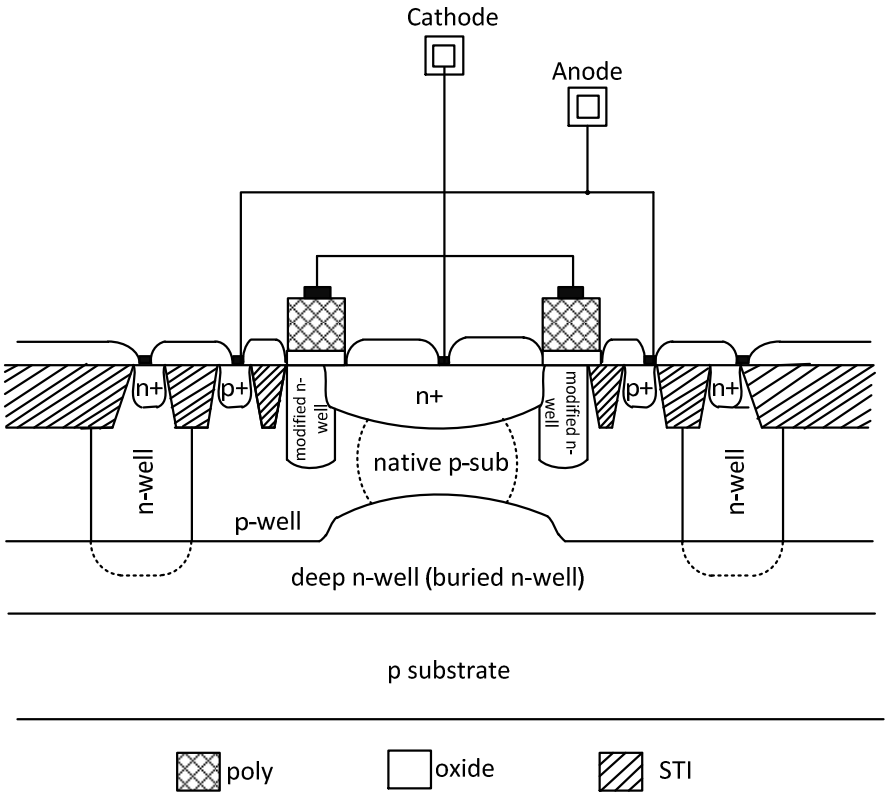


Figure 2.3 - Device structure of SPAD [25].

2.2.2 SPAD Performance Parameters

To analyse the operation of SPAD and its interface, it is important to understand its performance determining factors. Few of the performance factors are influenced by excess bias voltage of SPAD (V_E) which is the difference between the operating bias of SPAD V_{OP} and the breakdown voltage V_{BD}. Following is a brief discussion of the major influencing factors:

Dark Count Rate (DCR)

Dark counts refer to the triggering of an avalanche by secondary carriers in the absence of illumination. These carriers produced through tunnelling or thermal generation effects can be captured in traps and their random release leads to avalanche current pulses unrelated to a photon event. Such dark counts have a Poissonian distribution and can be seen as a noise source of the device [26].

In Geiger mode, these spurious pulses cannot be distinguished from regular photon triggered pulses and are produced at a mean frequency known as DCR. The SPAD DCR increases with higher V_E because of two effects: (a) an increase in avalanche triggering probability and (b) higher emission rate from generation centers due to larger field. DCR is strongly dependent on temperature and defines the minimum detectable signal, thereby limiting the dynamic range of the imager.

Photon Detection Probability (PDP)

PDP refers to the probability of a photon being absorbed in the detector active area to generate an electron-hole pair capable of triggering an avalanche. The probability of photon detection depends on diode's quantum efficiency as well as the excess bias voltage V_E. PDP increases with higher V_E since a higher electric field increases the sensitivity and thus its triggering probability. Usual values for PDP in CMOS SPADs are in the range of 1% to 30% [10][27].

SPAD Timing Jitter

It refers to the statistical fluctuation of the time interval between photon arrival at the sensor and pulse generation by the interface circuitry. Also known as timing resolution of SPAD, it improves with increasing V_E. At higher V_E, the depletion width is greater and thus, only photons of longer wavelengths can be absorbed, giving a sharper timing jitter pulse. At smaller V_E, saturation occurs giving a large timing jitter.

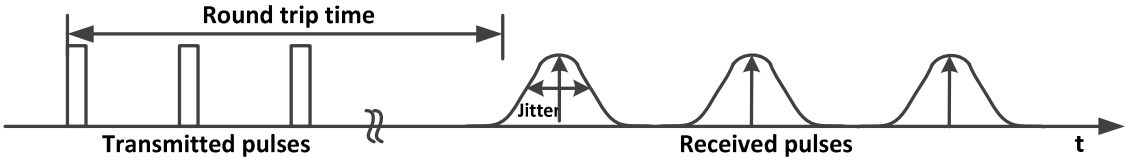


Figure 2.4 - Transmitted and received pulses [28].

SPAD jitter is typically expressed at full width half maximum (FWHM) as shown in Figure 2.5. FWHM is a common expression in signal processing to define bandwidth and describes the extent of a function. It is given by the difference between two extreme values of the independent variable at which the dependent variable is equal to half of its maximum value.

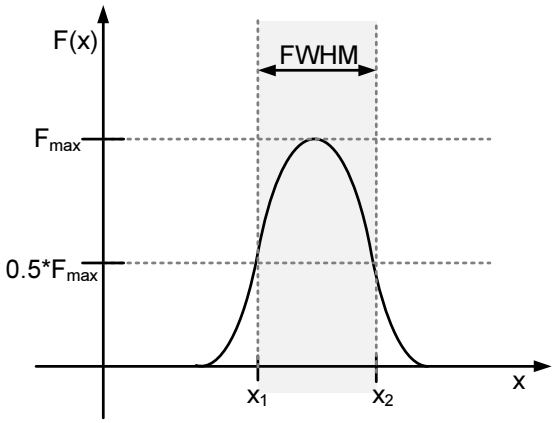


Figure 2.5 - Full width at half maximum.

Afterpulsing

During a photon triggered avalanche, some carriers can be captured by deep levels in the junction depletion layer and be released at a later arbitrary time. These released carriers can retrigger the avalanche, generating false current pulses correlated with the previous

avalanche pulse. Such an after-effect is termed as afterpulsing and denotes the spurious pulses related to previous photon events.

The number of carriers captured by deep levels during an avalanche pulse depends on the trap concentration and the total number of carriers generated during a Geiger pulse i.e. the total charge of the pulse. Longer is the quenching time of avalanche current, higher is the total charge in the pulse and therefore, higher is the afterpulsing effects. To sum, afterpulsing increases with reducing quenching time or with higher avalanche current intensity.

Dead time

After the onset of avalanche in SPAD device, the avalanche current needs to be controlled and the SPAD needs to be restored to its Geiger mode operating bias in order to detect subsequent photon arrival. The time required to quench the avalanche and recharge the diode upto 90% of its excess bias is defined as the dead time of SPAD [10]. Dead time should be as small as possible as it limits the rate of detection of photons, thus restricting the acquisition rate or imager. However, a small dead-time implies larger afterpulsing and therefore, a trade-off exists.

Cross-talk in SPAD farms

Cross-talk in SPAD farms can be both electrical and optical in nature. Electrical cross-talk occurs when photon absorbed in the substrate generates carriers which diffuse laterally in adjacent pixels and trigger avalanche. The mean penetration depth of photons depends strongly on the photon wavelength and therefore, photons in red and infrared ranges induce more electrical cross-talk than short wavelength photons [1]. However, electrical cross-talk can be eliminated by proper SPAD device design techniques.

Optical cross-talk refers to trigger of spurious avalanche current in the neighbouring SPADs. This is due to electroluminescence effect [10] where secondary photons can be emitted and absorbed by the nearby pixels in an array of SPADs. Optical isolation may be employed among the SPAD devices to reduce such an effect.

Optical cross-talk can be estimated by measuring the signal correlation between two adjacent pixels while electrical cross-talk needs sophisticated optical testing.

2.2.3 SPAD Readout

Once the photon hit creates an avalanche effect in the SPAD device, its front-end circuitry has the following tasks:

i. **Sensing:** The photon absorption indicates an event occurrence which needs to be propagated to time interval measurement circuitry.
ii. **Quenching:** When avalanche occurs, it needs to be halted to avoid large flow of current.
iii. **Recharge:** The SPAD needs to be recharged so that it is biased in Geiger mode and can detect a photon again.

2.2.3.1 Sensing

The signal generated at the anode of SPAD needs to be sensed and propagated to the processing circuitry. Typically, an inverter is employed to capture the SPAD signal. However, different circuit topologies may be applied depending upon the signal amplitude of SPAD. A detailed discussion is made in Chapter 4.

2.2.3.2 Quenching

Quenching is the process of halting SPAD's avalanche breakdown effect. Different techniques have been proposed in the literature and fast quenching is preferred to avoid self-heating, trapping and optical cross-talk [29]. There are two approaches to quenching - passive and active, as described below.

Passive Quenching (PQ)

When an avalanche occurs in SPAD, a potential drop can be developed if a serial load is connected. This potential drop moves the SPAD outside Geiger mode of operation. Thus, the SPAD avalanche current can be quenched on its own by developing a potential drop across a high impedance load R_B [26] as shown in Figure 2.6a.

The advantage of passive quenching is that the technique is self-quenching in nature and it involves passive components or active component resembling passive behaviour (a ballast resistance). However, it suffers from a relatively long and poorly defined dead time. In addition, discharging the capacitance through SPAD junction increases the number of carriers, thereby increasing the afterpulsing effect.

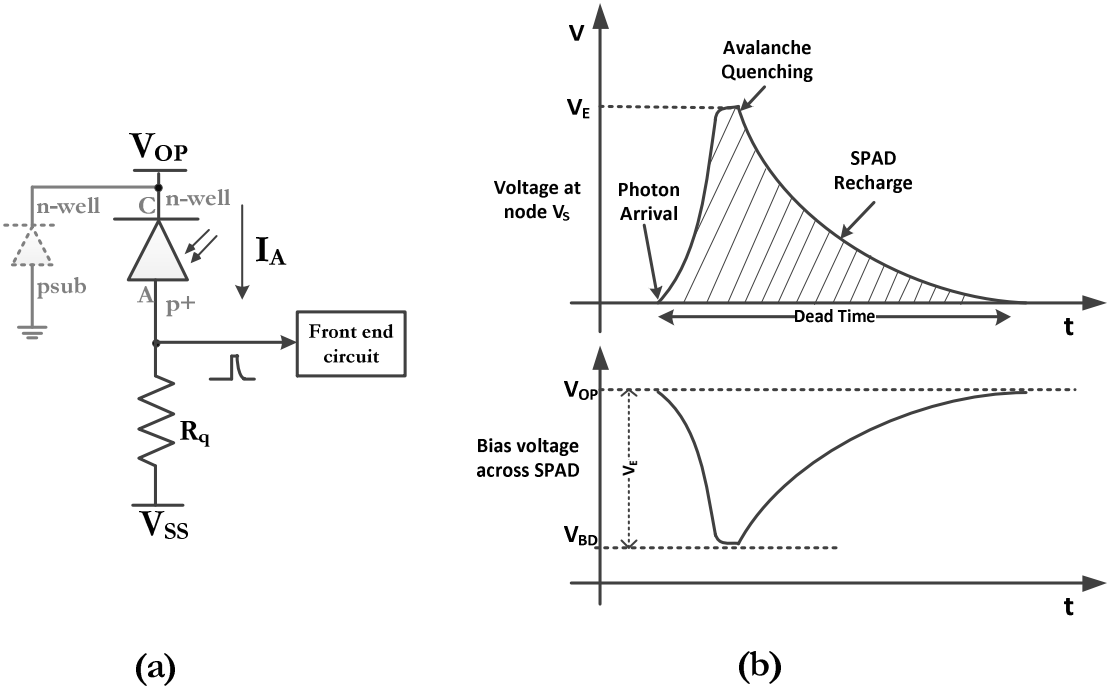


Figure 2.6 - SPAD operation with (a) passive quenching; (b) build-up of voltage and the concept of dead time [23].

Active Quenching (AQ)

An active quenching circuit tracks the SPAD voltage or current by comparing the signal with a threshold to identify an occurrence of avalanche. After an avalanche is detected, it quickly discharges the total capacitance seen by the SPAD through a secondary path, thereby quenching the avalanche. To achieve this, high speed sensing and feedback circuits are employed. The underlying principle of active quenching is to minimize the number of carriers generated in the SPAD and force the carriers to flow through the secondary path. It is a faster way of quenching through positive feedback mechanism and the dead time is small, well-defined and reproducible.

2.2.3.3 Recharge

After the avalanche current in SPAD is quenched, recharge of SPAD is needed to prepare it for next photon detection. This is done by biasing the SPAD back in Geiger mode. In literature, multiple recharge schemes exist and are typically classified into active and passive nature.

Passive Recharge (PR)

A passive recharge is possible by the same circuit as passive quenching i.e. a resistor. However, large parasitic capacitance introduces unfavourable effects in terms of inaccurate timing resolution and maximum possible count rate [1]. In TCSPC applications, high count rates are extremely desirable. It is possible to reduce unwanted parasitics to an extent by integration of front-end circuits and SPADs on a single chip.

With passive recharge, the dead time τ is given by,

$$\tau_{dead} = RC \tag{2.1}$$

A small R gives a shorter dead time but larger afterpulsing, as the trapped carriers are released after SPAD recovers from avalanche.

Active Recharge (AR)

The underlying principle of active recharge is as follows. As soon as the avalanche is quenched, a secondary low resistance path quickly recharges the SPAD back to Geiger mode. AR can be performed to advantage with a small dead time implementation. In [30], a passive quenching/active recharge circuit is described and it achieves a dead time of 6 ns with a dynamic range of 116 dB at an acquisition time of 20 ms.

2.3 Time to Digital Converter

This section discusses another critical block of a 3D imager - time-to-digital converter (TDC). A basic TDC quantizes a time interval into a digital code and its ideal transfer function is shown in Figure 2.7. The design challenges for the TDC in this application are a high dynamic range, small die area, a high time resolution, high linearity, small offset and gain error and low power consumption.

TDCs finds its applications in varied domains such as PLLs, logic analyzers, space science instruments, jitter measurement and ToF based particle detection. Diverse methods of digitizing time intervals have been proposed in the literature depending upon the application requirements. The techniques proposed in past include utilization of fast counters [31], analog techniques based on generating a voltage ramp, and various CMOS tapped delay line configurations. The tapped delay line technique is of particular interest as it is based on standard digital CMOS process resulting in multiple advantages of low-cost, high integration level and low power dissipation. In this section, the important delay line based approaches have been discussed.

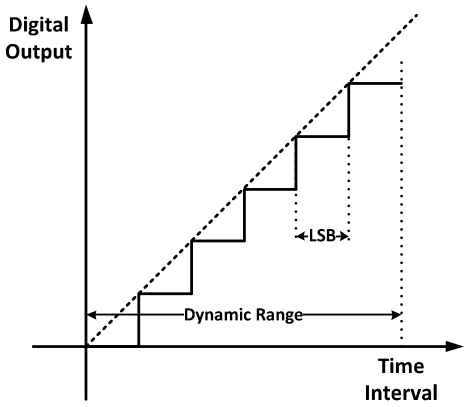


Figure 2.7 - Ideal transfer characteristic of TDC.

2.3.1 Delay Line (DL) Based Architecture

Figure 2.8 shows the simplest way to measure a time interval between START and STOP signals. The START signal propagates through a string of buffers, also known as 'delay taps', to produce a delayed output after each delay tap. When the STOP signal arrives, it samples the state of delay line. The output thermometer code determines the time interval in terms of buffer delay. The sampling of delay line is generally done using arbiters such as flip –flops or latches.

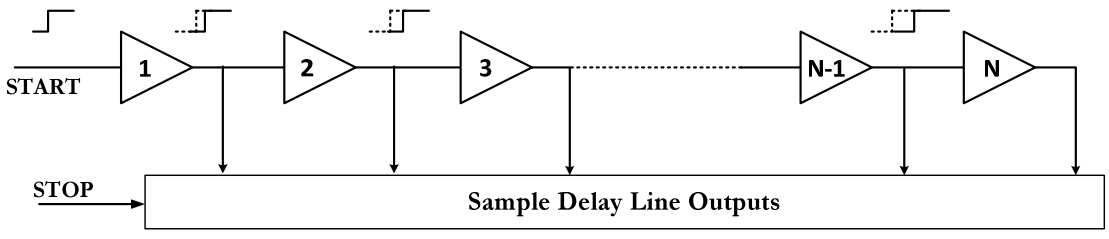


Figure 2.8 - Conceptual diagram of delay line.

The TDC resolution is defined by the minimum time interval that is possible to be measured. The resolution of TDC shown in Figure 2.8 is one buffer delay. The dynamic range is given by the time interval between the first and the last delay tap.

Consider N delay taps in a delay line each with a propagation delay τ, the dynamic range '*T*' is given by:

$$\mathrm{T} = \mathrm{N} \cdot \tau_{delay-tap} \tag{2.2}$$

A higher TDC resolution is often desirable and an inverter based delay line is employed as it provides the smallest gate delay. However, inverter based structure complicates TDC design due to uneven rising and falling edges of the delayed signal along with unbalanced meta-stability resolution of flip-flops. The resulting output code deviates from the conventional string of 1s based thermometer code and is pseudo-thermometer in nature. In [32], a delay of 20 ps in 90 nm technology using a differential delay line structure is reported and it produces a pseudo-thermometer code.

It is possible to achieve sub-gate resolution in TDCs and several architectures have been proposed in the literature. One of the most common architecture is Vernier delay line TDC which utilizes a technique known as the 'Nonius principle' in the context of slide gauges [Source: Wikipedia][33]. A Vernier TDC consists of two delay lines: one for START and other for STOP signal. Each delay line is built with buffers; the delay elements in the first delay line have a delay τ_{d1} which is slightly larger than the delay τ_{d2} of the second delay line. The time interval measurement begins with the propagation of START signal in the first delay line. The STOP signal arrives later but it experiences smaller delays in the delay line and chases the START signal. At each stage, the skew between the two signals is reduced by:

$$t_{LSB} = \tau_{d1} - \tau_{d2} \tag{2.3}$$

To measure a maximum time interval ΔT, the number of stages required is

$$N = \frac{\Delta T}{t_{LSB}} = \frac{\Delta T}{\tau_{d1} - \tau_{d2}} \tag{2.4}$$

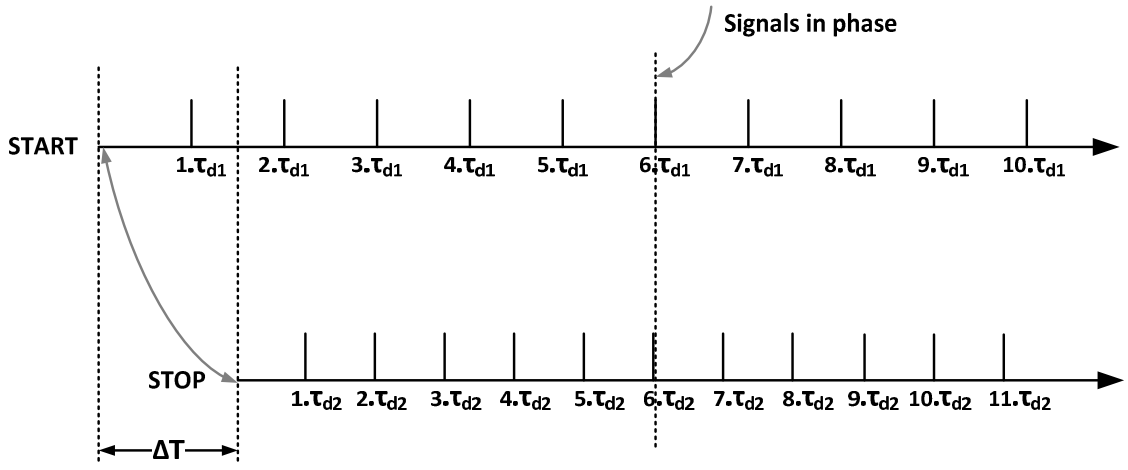


Figure 2.9 - Operating principle of Vernier delay line based TDC.

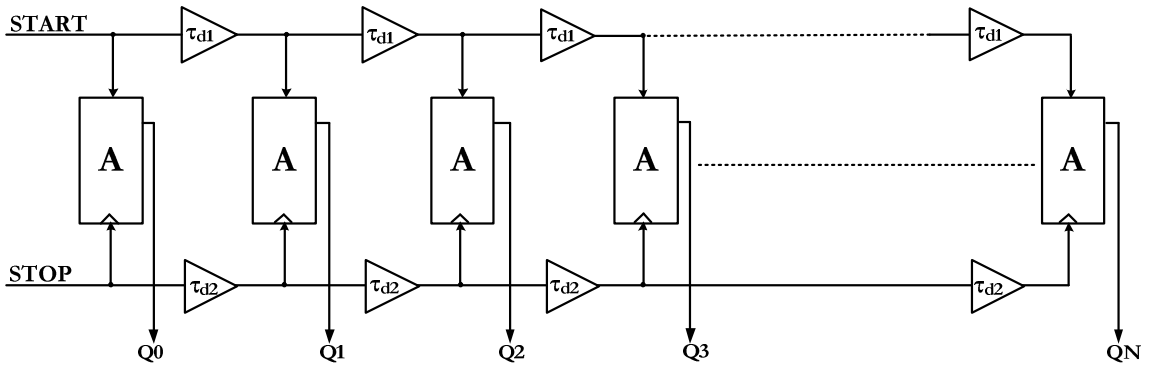


Figure 2.10 - A basic Vernier delay line based TDC architecture.

With a Vernier delay line (VDL) based TDC, the TDC resolution is the difference between two gate delays which in principle, can be made very small. In [34], a 7 bit two-dimension VDL based TDC for a digital PLL is reported. It achieves 5 ps resolution for a conversion rate of 50 Msps and a power consumption of 1.65 mW in 65 nm technology.

Although the Vernier technique overcomes the resolution limitations presented by the technology, it occupies a large area with two delay taps and an arbiter for each stage. Moreover, for a large dynamic range, long delay lines are required leading to a high power consumption. Furthermore, if the resolution is very high, the mismatch in gate delays play a critical role and may need additional stabilizing circuitry such as DLLs [35].

To conclude, VDL structures are sought for their simple modular design and sub-gate level resolution. However, its two long delay lines and impact of process variation on the resolution makes it undesirable for low-cost robust applications. There are other TDC architectures as well such as pulse-shrinking [36] and local passive interpolation [33] which achieve sub-gate resolution. However, each of the architectures is costly in terms of area and a large dynamic range is not possible to be achieved.

From the above discussion, we can conclude that linear delay line structures are suitable only for short dynamic range only. Also, sub-gate TDC architectures exist in the literature but are high cost in terms of area and design. If a larger dynamic range of 13-15 bits is sought in small area with high resolution, alternate TDC configurations need to be explored. Two such possible methods are (a) coarse-fine TDC and (b) a looped TDC. They are discussed at length in the following sections.

2.3.2 Coarse-Fine Architecture

A coarse-fine TDC is one of the common architectures employed by applications demanding large dynamic range along with high resolution. The underlying principle is - the coarse TDC covers the time interval in bigger time steps to cover the range whereas the fine TDC quantizes the time residue between the input and its closest coarse level in smaller time steps, resulting in high resolution.

In [32], another topology of coarse-fine TDC is introduced where the fine TDC measures the time interval to the next clock edge with 't' resolution. The coarse TDC is a synchronous digital counter whereas the fine TDC is built with pseudo-differential delay line structure

giving a TDC resolution of 20 ps in 90 nm technology. If the coarse and fine TDC count 'N' and 'm' cycles respectively, the time interval is equal to:

$$\text{Total Time} = [N \cdot T + (T - (m \cdot t))] \tag{2.5}$$

where T is the clock period or coarse resolution and 't' is the fine resolution.

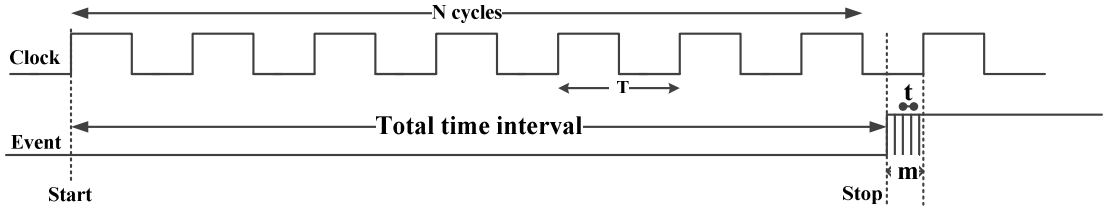


Figure 2.11 - Concept of coarse-fine TDC.

In [37], a simple delay line achieves the coarse count whereas a folded Vernier delay line has been employed for realizing a high resolution of 10 ps in compact area. However, the linearity is strongly dependent upon the matching of delays between coarse and fine networks and the performance degrades with PVT variations.

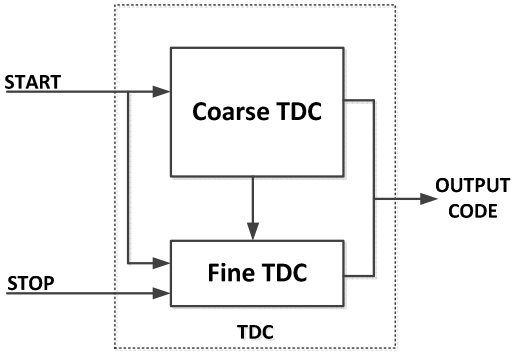


Figure 2.12 - Block diagram of coarse-fine TDC.

The coarse-fine TDC can be implemented in various configurations - in [38], the concept of time amplification has been used where the amplified residue is quantized with the coarse resolution. Although it achieves 1.25 ps resolution in 90 nm CMOS without any analog circuitry and has a 9 bit dynamic range, it utilizes large area in its string of Time Amplifiers employed to realize a higher resolution.

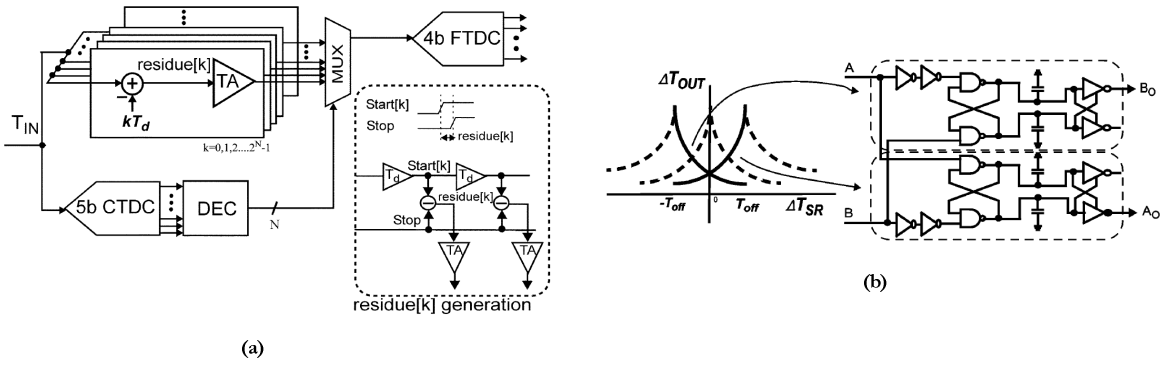


Figure 2.13 - Coarse-fine TDC: (a) conceptual diagram of time amplification; (b) time amplifier [35].

The above discussed architectures achieve a large dynamic range with high resolution but each of them consumes a large area as well. To realize such a TDC in every pixel impacts the fill factor of imager tremendously and thus, techniques to fit TDCs in compact area are sought. One of the techniques that can be used in a pixel array to achieve low-cost is 'sharing of resources'. With respect to TDCs, the technique can be applied by sharing the coarse TDC among a group of pixels.

2.3.3 Looped TDC Architectures

A looped configuration is desirable in imaging application due to its compact size. Any of the delay line structures (including sub-gate resolution architectures as in Section 2.3.1) can be put in a looped configuration and the dynamic range can be extended using a counter, as shown in Figure 2.14 [39].

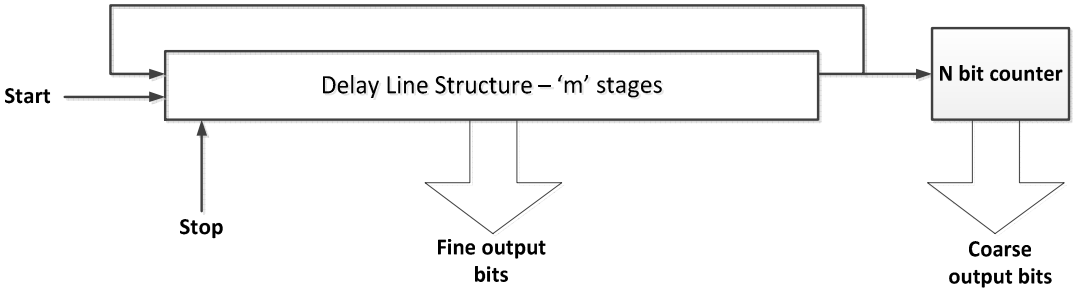


Figure 2.14 - Concept of a looped TDC.

However, there are only a few practical implementations in the literature. This is because controlling the loop dynamics increases design effort and adds non-linearity among other design issues. For example, folded Vernier structures are complex and non-linear due to mismatch in delays and need a stabilizing circuit such as DLLs. Inverter delay line based looped architectures are most common due to their simple structure. Several looped TDC architectures are present in the literature.

In [8], a single delay line ring oscillator based on cross-coupled inverters is used and is shown in Figure 2.15. However, such a TDC has very low resolution, of the order of 55 ps and does not meet high resolution requirements.

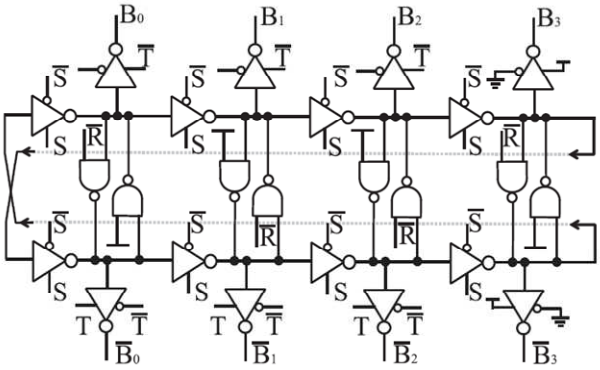


Figure 2.15 - Ring oscillator based on cross coupled inverters [8].

In [40], a ring oscillator based TDC in 0.35µm technology is implemented. Here, the ring toggles continuously and the start/stop events capture the state of ring for time interval measurement. The power consumption of such a TDC is very high due to continuous oscillation of the ring. If such a TDC is incorporated in every pixel of the array, the power consumption would be massive due to continuous operation of all TDCs. Therefore, this architecture is highly unsuitable for a million pixel array. Further, the delay tap is cross coupled inverter which results in very low resolution of 70 ps. In [41], an improvement in the resolution is achieved by implementing a Vernier ring oscillator but it is at the cost of larger area.

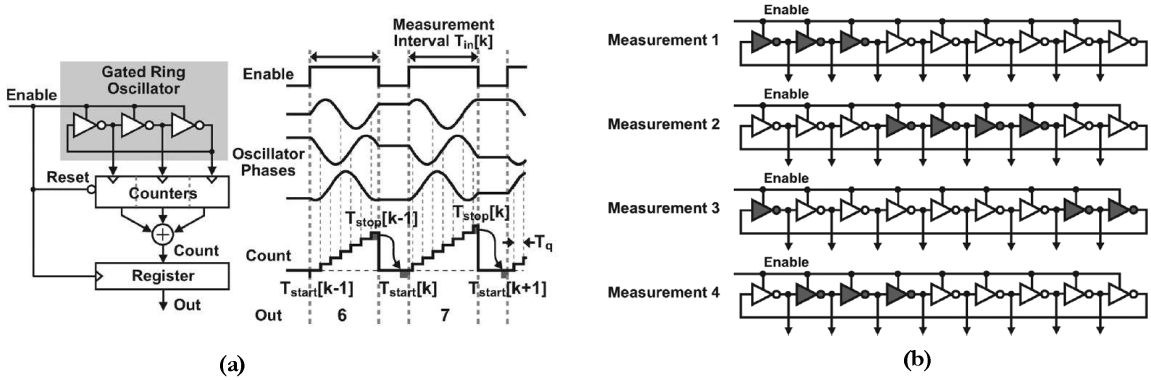


Figure 2.16 - (a) Gated ring oscillator TDC; (b) the concept of barrel shifting of GRO delay elements to achieve first order mismatch shaping [42].

In [42], a 6 ps, 1st order noise shaping, 11-bit TDC is presented in which delay between two signals is given by the number of delay element transitions within the measurement interval. Although high resolution is possible, the design is not suitable for an imaging application for three reasons. Firstly, the design relies on the memory of the previous state to measure the current input. This demands memory storage capacity in terms of registers for each TDC and translates to larger area of TDC. Secondly, the delay element used in this TDC is a switched inverter which gives lower resolution than a normal inverter. The first order noise shaping is useful in this scheme only when the input signal is oversampled by a factor of 50-100. Therefore, the architecture is not appropriate from area and readout perspective.

2.4 Summary

In this chapter, the basic building blocks of 3D imager, a SPAD photodetector and a TDC were discussed. Different techniques to build the front-end circuitry of SPAD were described. Furthermore, various TDC architectures which may be suitable for a large 2D pixel array were considered. Two TDC configurations: a coarse-fine TDC and a looped TDC were shown to be appropriate for a 3D imager.

The next chapter presents the system architecture of the 3D real-time imager built in this project. A detailed discussion is made to derive the system specifications and explain the working of the system.

3. System Design

This chapter discusses the system level aspects of building a time-of-flight (ToF) based 3D imager. We have already covered the immense scope of 3D imagers from Section 1.1. One of the important areas of application is that of security where a high resolution ToF based 3D camera installed on a docking station could be useful in many scenarios such as physical access control in homes, banks or production plants, border control surveillance, baggage checks at airports and e-passport/visa programs. Obtaining the information in real-time adds value to the system.

This chapter is organized as follows: Section 3.1 derives the imager system specifications to make the imager suitable for multiple security applications. Section 3.2 discusses the imager architecture and presents the new concepts introduced at the system level for high depth resolution and low power consumption.

3.1 Imager Specifications

Face recognition is a critical requirement for many security applications. In order to detect the distinctive features of the face, a high depth resolution, in the order of a few millimeters, is desirable. In this work, a depth resolution of less than 5 mm is desired to allow the facial contours to be easily distinguished. A 3-5 mm depth resolution implies a timing resolution of 20-30 ps for an optical ToF 3D-imager. The depth resolution for an imager is limited by the noise floor of the system. The main noise sources in the system are the timing jitter of SPAD and the time resolution of TDC. Therefore, the total noise is given by:

$$\Delta z = \sqrt{t_{noise}^2} = \sqrt{t_{spad-jitter}^2 + t_{tdc}^2} \tag{3.1}$$

From past measurement results, it is known that SPAD jitter is of the order of 100-150 ps [43]. In 65 nm technology, the SPAD jitter is expected to be greater than 100 ps whereas a single inverter delay is approximately 20 – 30 ps. As SPAD jitter is the dominant source of noise in the system in this technology, the TDC resolution of a single inverter delay is good enough for the system. Although it is possible to achieve sub-gate TDC resolution, as discussed in Section 2.2, it requires complex architectures with high cost in terms of area. Moreover, it does not result in a significant improvement in the overall resolution.

Along with a high depth resolution, a high lateral resolution is necessary to capture sufficient number of points for face recognition or for a larger field of view that may be required in surveillance activities. The lateral resolution is determined by the number of pixels in the imager. Considering the complexity of a 3D imager pixel compared to a 2D imager, achieving high megapixel resolution may not be practical. At this stage, achieving a lateral resolution comparable to that of a 'Video Graphics Array' (VGA) camera is a significant contribution to 3D imaging field. Therefore, an array of 1000x1000 pixels is targeted to surpass the 640x480 lateral resolution of a VGA camera.

For surveillance activities in banks, shopping malls, or airports, a long linear field of view is desirable. Therefore, the imager must have a large maximum range to be able to capture images at a large distance. However, the number of photons lost in environment (link attenuation) can be prohibitive in achieving a very large range. To meet the large dynamic range requirement, the maximum range of the imager is fixed as 30 m. This corresponds to a photon time-of-flight of 200 ns. Therefore, the TDC must achieve a dynamic range of 14 bits, assuming a 20 ps resolution.

The fill factor is the defining parameter for sensitivity of any imaging system. In linear avalanche photo-diodes (APD) based imagers, fill factor as high as 20% - 60% is possible. SPAD based pixels generally have a smaller fill factor of 1-10% to accommodate more imaging logic. However, this parameter does not degrade the performance of the imager. This is because SPADs are sensitive to single photon [44] as opposed to linear APDs which depend on intensity of light for image formation. In case higher photon flux is required for more measurements, higher power laser can be employed for increasing the intensity. Another way to increase the probability of photons hitting the SPAD detector is by using micro-lenses. Such optical compensation increases the probability of photon being incident on SPAD to about 80% [28], but it raises the system cost.

For real-time operation of the 3D imager required for surveillance applications, it is required to have a high frame rate. Typically, the human eye cannot resolve beyond 24 frames per second. Standard video frame rates are 30 fps which is the typical specification this system is aimed at. However, to advance the state of the art, a maximum frame rate of 1000 fps is targeted to allow a high end real time feature in the 3D camera.

The chip-area is the major cost driving factor of the system built in CMOS technology. To have small area for low-cost imager, it is of great interest to have small sized pixels. The pixel size is limited by the in-pixel readout circuit and therefore, a smaller pixel size will limit the possible fill factor. This results in a trade-off between the fill-factor and the imager cost. A pixel pitch of 25µm was chosen which allows a fill-factor greater than 5% and is considered reasonable for so called 'smart pixels' capable of high performance. This choice leads to the core area of 1000x1000 pixels in the order of ~625 sq. mm.

With such a complex array of pixels, the challenge is to limit the power consumption to 1 Watt. Beyond this limit, a heat sinking facility may be required with every IC which increases the system cost tremendously and is highly undesirable for low-cost 3D imagers.

The final set of specifications is summarized in Table 3.1. Although the specifications have been derived for security applications; the design with this set of specifications can be easily extended to different applications such as automotive rear vision or space-craft landing station.

Table 3.1 - Imager Specifications.

Parameter	Value
Technology	65 nm CMOS
Number of pixels	10^6 (1000x1000 array)
Core Area	~ 25 mm x 25 mm
TDC Resolution	Inverter Delay (Less than SPAD Jitter)
TDC Dynamic Range	14 bits
Maximum Range (z_{max})	30 m
Imager Resolution	< 5 mm
Average Power	~ 1 W
Frame Rate	30 fps (typical)
	~ 1000 fps (maximum)
Fill factor	> 5 %

3.2 Imager Architecture

The system aims to build a pulsed ToF based 3D real time camera for security dock stations. It is necessary to perform system analysis for better understanding of system working along with its requirements. In this section, the system operation is described and its various trade-offs are discussed.

3.2.1 At The Light Source

A pulsed laser source emits pulses towards the object of interest in a cone beam of light to illuminate the scene of interest as shown in Figure 3.1. The cone beam of light can be perceived as a collection of discs incident on the object of interest. Assuming partial collimation, for a given laser power emitted from the source, the intensity of light varies with the radius, *r,* of the disk as:

$$\text{Intensity} = \frac{\text{Power}}{\text{Area}} = \frac{\text{Power}}{\pi r^2} \tag{3.2}$$

Thus, the number of photons incident on the object vary with the distance of the object from the laser source. As the photons emitted from laser source hit the object, they are reflected back on the imager. Farther the object from laser source, lower would be the intensity of light and, consequently, lower the reflected photon-flux.

The impinging photon energy is

$$E_0 = \frac{hc}{\lambda} \tag{3.3}$$

where $h = 6.62 \cdot 10^{-34}$ J · s and λ = 0.78-0.9 µm. This wavelength falls in non-visible near-infrared range to which SPADs are sensitive to and it also meets eye-safety requirements. To achieve sensitivity to a higher wavelength band such as 1.5 µm, a Germanium layer, in principle, can be added on top of the design as in [45].

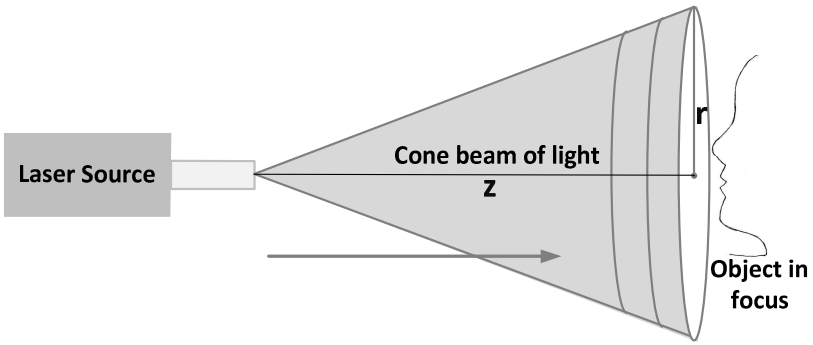


Figure 3.1 - Illumination of the scene using cone beam of light.

3.2.2 At The Detector End

After the photons from the source hit the object, they are reflected back on the imager. Most of the reflected photons are lost in the environment, and only few photons hit the imager. This is also known as 'link attenuation'. Its magnitude can be estimated by calculating the percentage of light energy received by the imager of the total light energy reflected. Assuming that the light reflection is in form of hemisphere, having a radius equal to its distance from the imager '*z*', the attenuation in the energy reflected from the object is given by:

$$\frac{\text{Energy}_{\text{imager}}}{\text{Energy}_{\text{reflection}}} = \frac{\text{Area}_{\text{imager}}}{\text{Area}_{\text{hemisphere}}} = \frac{(25\ \mu \cdot 25\ \mu) \cdot 10^{6}}{4\pi \cdot 30^{2}} = 5.5 \cdot 10^{-8} \tag{3.4}$$

This corresponds to 60 - 80 dB of link attenuation. The time-of-flight information from the photons incident on the imager is used for calculating the distance of the object for image reconstruction.

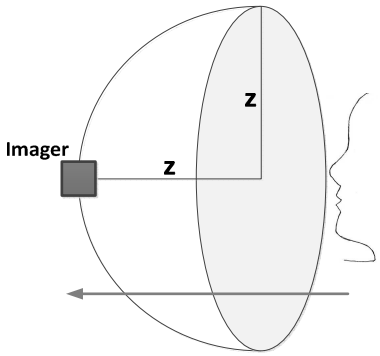


Figure 3.2 - A side-view of the reflected light from the 3D object. The reflection forms a hemisphere with radius equal to object range.

The intensity I of laser light incident on the imager can be translated into number of photons hitting the square matrix of area A_{IMAGER} by:

$$\frac{I \cdot A_{\text{imager}}}{E_{0}}\ \textit{photons per second} \tag{3.5}$$

where E_o is photon energy, A_{IMAGER} is the area of detector and I is the intensity of incident laser light.

As the system works with pulsed laser source, the light is not emitted continuously. Therefore, it is more appropriate to consider the received laser pulse energy [J] rather than received laser intensity [J/s]. For an imager size of 1000x1000 pixels (625 sq. mm), the relationship between the energy of incident laser light per pulse and effective number of photons hitting each pixel is:

$$\frac{\text{Energy received by imager}}{E_0} \textit{ photons per SPAD per pulse} \tag{3.6}$$

Figure 3.3 shows this relation with and without the use of 20x microlens.

From the system level analysis, we can say that:

(a) A maximum of only one event per SPAD per pulse can be read out,
(b) The intensity of received light should be such that the number of events on every SPAD should be as close to one as possible in order to maximize the number of data points,
(c) To achieve this, we see from Figure 3.3 that the optical energy received on the imager should be ~1 pJ,
(d) Therefore, the laser intensity emitted from the source and received at the imager should fit the following specifications:

i. Received laser energy per pulse: 1 pJ
ii. Link (round-trip) attenuation: 60 - 80 dB,
iii. Emitted laser energy per pulse: 100 µJ (assuming 80 dB link attenuation).

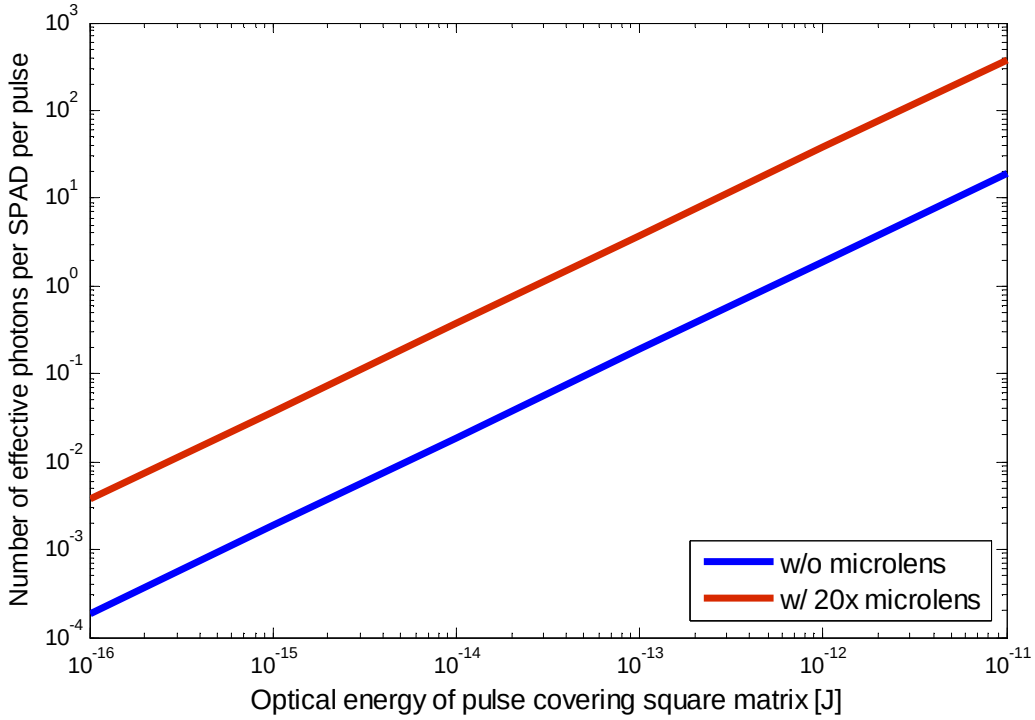


Figure 3.3 - Relationship between the number of effective photons impinging on a single detector and laser pulse energy covering the entire imaging matrix of 1000x1000 pixels [28]. The fill factor is assumed to be 5% and PDP is approximated to 50%.

In order to achieve at least 1 photon per SPAD per pulse, we have three possible options:

(a) either the intensity of received pulse is made higher or in other words, the laser emission power is increased,
(b) or a micro-lens is employed to focus the photons on imager array which translates to a higher system cost,
(c) or a burst of pulses is employed to make the probability of a SPAD event per burst as close to 1 as possible instead of SPAD event per pulse.

A high power laser pulse as suggested in option 'a' may be hazardous to human eye and is not desirable in imaging system. Option 'b', as stated, is a high cost solution. Therefore, both the choices are not favourable from system level. Option 'c' proposes to be an effective way of ensuring high number of photons reflected back at imager. This concept, also known as 'train of pulses' implies that the pulsed emission of laser beam is in the form of 'train of pulses' instead of low frequency high intensity pulses. Figure 3.4 presents the conceptual idea of 'train of pulses' and is favorable for many reasons:

(a) Constructing closely spaced (less than round trip time) and low intensity pulses is easier than a single powerful pulse in terms of laser power [28],
(b) A high intensity pulse may be hazardous to human body and not cater to human-eye safety requirements [28],
(c) A high frequency train of pulses compensates for the link attenuation and low PDP. For example, emitting 1000 pulses, each with en energy of 100 nJ, would imply the total emitted laser energy per burst is 100 µJ. However, since each pulse has only 100nJ energy, it is much safer for human eye than a high single intensity pulse of 100 pJ energy.

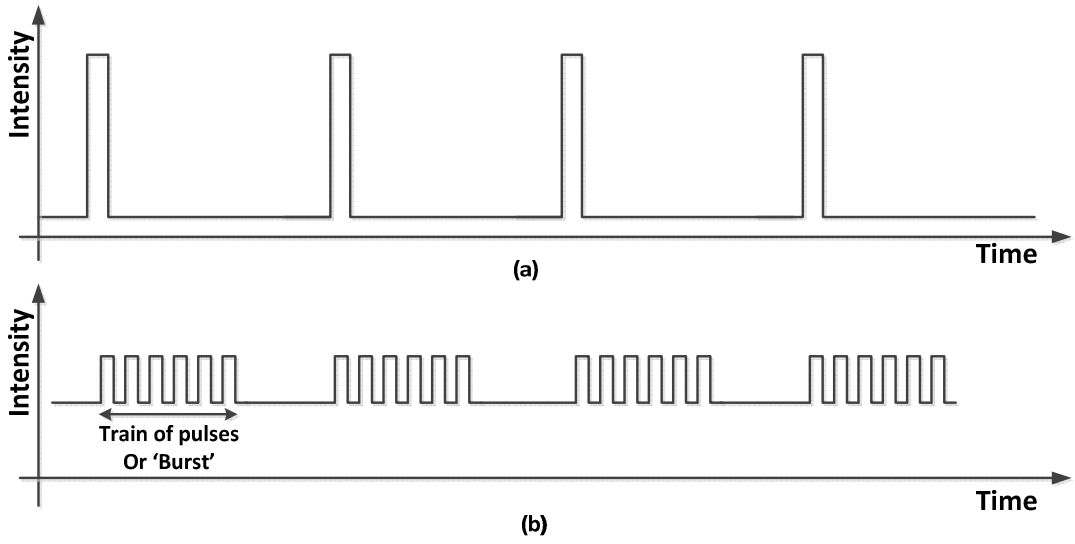


Figure 3.4 - Pulsed mode of operation: (a) high intensity, low frequency pulses; (b) low intensity, high frequency pulses (a preferred mode of operation).

Thus, the concept of high frequency low power train of pulses is advantageous from system level and is preferred for pulsed mode of operation for ToF estimation. From the above calculation, we can summarize the following:

(a) Emitted laser energy per pulse: 100-400 nJ [28],
(b) Link (round-trip) attenuation: 60-80 dB,
(c) Received laser energy per pulse: 1-4 fJ (assuming 80 dB link attenuation).
(d) Number of pulses in a single burst: 1000

The next section discusses how the time interval measurement can be performed with low power consumption.

3.2.3 Time of Flight Estimation

The time interval measurement for a 3D imaging system is performed using a time-to-digital converter. Different configurations are possible for ToF estimation and are discussed in this section.

3.2.3.1 Coarse-Fine Architecture based Time Interval Measurement

In Section 2.3.2, a conventional method suitable for coarse-fine TDC architecture was discussed to measure time interval, refer Figure 2.11. In this configuration, the positive edge of the clock was the START signal whereas the STOP signal was a SPAD photon detection event. In coarse-fine architecture, the coarse TDC works by counting the clock edges, while the fine TDC performs interpolation to achieve high time-resolution.

It is important to judge if coarse-fine TDC architecture is appropriate for an imaging pixel array. The most critical parameter is frequency of clock operation as it presents a trade-off exists between area, speed, power and jitter parameters. A higher clock frequency implies a faster coarse counter resulting in more power but has less interpolation at fine TDC level, implying lower power. Table 3.2 summarizes the performance parameters in two sample cases with frequency of operation equal to 10 GHz and 2.5 GHz. Depending upon the desired performance, the system level frequency can be chosen.

From the table, we see that the coarse-fine architecture based TDC is surely one of the possible ways for time interval measurement in a ToF based imager and can be designed to match the range-resolution-area requirements of imager. However, it is not suitable for large pixel arrays as it is power hungry by design. In its range estimation, the coarse counter begins to count the integer clock cycles to estimate the ToF and this continuous operation leads to very high power consumption. Power consumption can be reduced to some extent if the coarse counter is shared by a group of pixels. Each pixel with its own fine TDC taps the Nth value of counter when it receives a photon. But such sharing is also not very power-efficient. For example, if the coarse counter is shared among 100 pixels; for a 1000x1000 array, 10^4 counters running at 2.5 GHz or 10 GHz would still be required, which translates into massive power consumption.

Table 3.2 - Performance parameters of TDC with variation in frequency.

Parameter	f = 10 GHz	f = 2.5 GHz	Comments
Area	x (= 5 ; with 20 ps inverter resolution)	4x (= 20)	Longer the clock period, more inverters required to interpolate one clock period
TDC power (Coarse and Fine TDC)	Coarse: higher power; Fine: lower power	Coarse: lower power; Fine: higher power	For a fixed time interval, total power consumption depends on final counts in coarse and fine TDCs
Speed of operation	Higher	Lower	Depending upon the clock frequency
TDC Resolution	Equal		Equal to an inverter delay ~ 20 ps
Jitter	$\sqrt{5} \cdot \sigma_{inv}$	$\sqrt{20} \cdot \sigma_{inv}$	$\text{Jitter} = \sqrt{\#\text{inv}} \cdot \sigma_{inv}$
Clock Distribution	Complex; more power consumption in clock drivers	Simpler; lower power consumption in clock drivers	Clock tree consumes more power and has increasing complexity at higher frequencies due to parasitics

3.2.3.2 Alternative Scheme for Time Interval Measurement

In order to conserve power, another approach to time interval measurement is proposed here. In this configuration, the photon arrival triggers the START signal while the following clock edge is used as the STOP signal. In this scheme, the TDC is event driven and only becomes active when a SPAD event occurs, thereby reducing the overhead of an always-running counter as in the coarse-fine TDC approach. The clock period, in this technique, can be equivalent to the maximum range of imager (say 10 m or 66 ns). This implies a system clock frequency of ~15 MHz which results in much lower power for a clock distribution circuit than a 10 GHz. However, it also means that each TDC is active for a longer period (upto 66 ns). The time-of-flight measurement is given by:

$$\text{Total Time} = [\, T_{clock} - (N_{output} \cdot t_{inv}) \,] \tag{3.7}$$

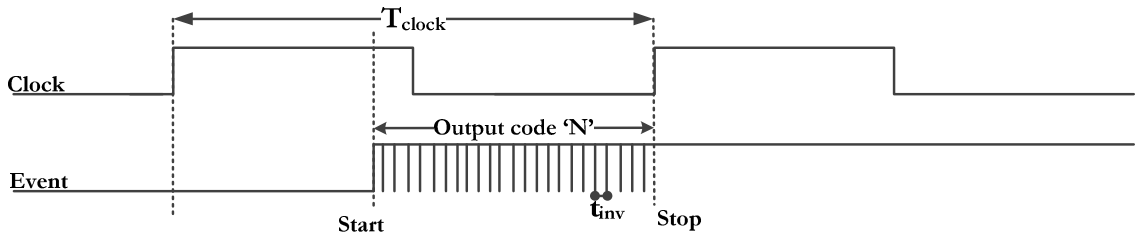


Figure 3.5 - ToF measurement technique for lower power consumption.

From the above discussion, we can conclude that the 'sharing of resources' does not seem profitable in TDC implementation. With the above described time interval measurement scheme, an event driven TDC is desired for every SPAD detector, i.e. an independent TDC must exist in every pixel. Since the power consumption is a concern with this architecture, a new scheme at system level is introduced to lower the power consumption and we call it '*Reconnaissance*'. The next section explains this technique in detail.

3.2.4 Reconnaissance [46]

This concept is introduced to reduce the power consumption of the system consisting of million pixel array. In this technique, the range of object is first estimated and then the system adapts to this known range for its further operation.

It is known that the lowest clock frequency in this system is 5 MHz which corresponds to 200 ns (or 30 m) of maximum object range. However, if the object range is less than 30 m, then the clock can be increased to match the range. To give an example, if the range of object in focus is 15 m, the clock frequency can be increased to 10 MHz from the initial frequency of 5 MHz. With the modification in clock frequency, the time duration for which TDC is in operation can be reduced by half, thereby reducing the power consumption drastically. Figure 3.6 depicts the concept of reconnaissance.

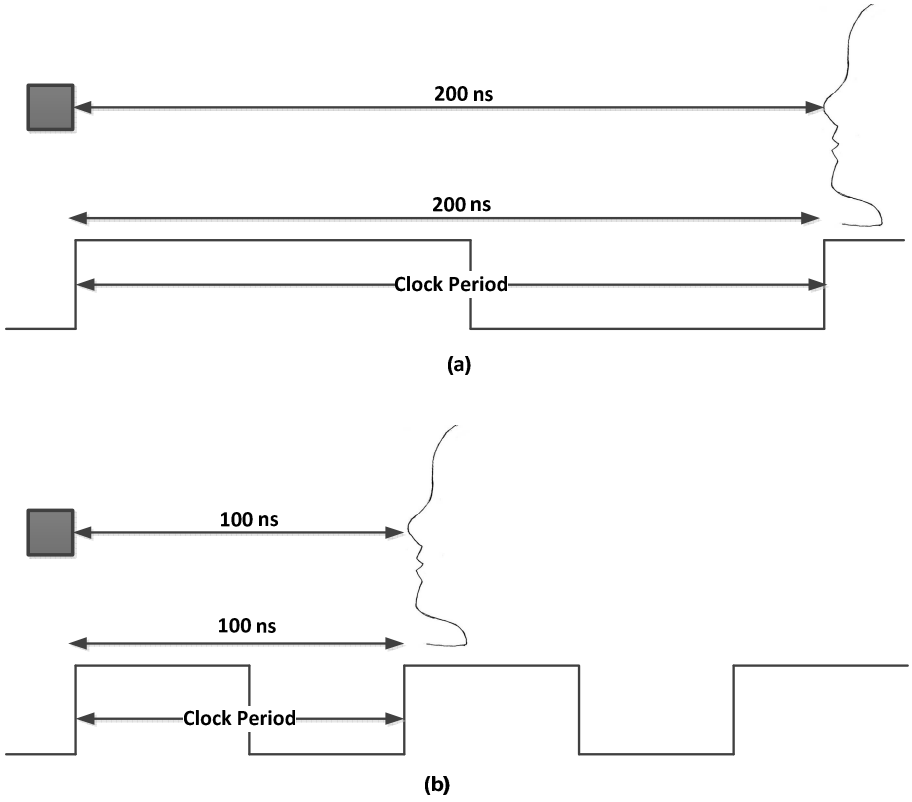


Figure 3.6 - The concept of reconnaissance: (a) when the object is 200 ns apart, (b) when the object is 100 ns apart.

To summarize, the range of object is estimated in the first few measurements. The clock frequency is adjusted to this distance (less than z_{MAX} = 30 m) and therefore, the TDC is in operation for smaller period of time. For minimal power consumption, the time interval for

which the TDC in each pixel works can be fixed to an average of 5 ns (corresponds to ~1m distance). The implication is that a time interval of 5 ns is good enough to capture an image with a depth range of ~1m. This is a reasonable assumption for a 3D camera to operate in low power mode. Further, issues related to aliasing may lead to higher power consumption, but the intensity of such photons incident on the imager from further distance would be much less. Further investigation and characterization of this scheme is needed and it is one of the important future works of the project.

3.2.5 Improving Resolution

The system works on 'train of pulses' concept where each burst of pulses result in at least one event at every pixel on the imager. With such collection of data, averaging can be done to snap-back to the correct time-of-flight. However, as discussed in Section 3.1, the resolution of the system is limited by the SPAD jitter.

In this work, new techniques have been proposed to overcome this limitation and improve the performance of the system. From statistics and probability theory, it is known that by averaging N samples of a random number with a standard deviation of σ, the mean standard deviation σ_{MEAN} can be reduced by $\sqrt{N}$ times (refer Appendix A),

$$\sigma_{MEAN} = \frac{\sigma}{\sqrt{N}} \tag{3.8}$$

In this imager, a SPAD photodetector is employed and it has a timing uncertainty associated with it, called as SPAD jitter. As discussed in Section 3.1, this is the main cause of noise in the system. The jitter creates an random error in the time-of-flight measurement and has a Gaussian distribution [47]. This error can, therefore, be reduced if multiple samples of same measurement are averaged.

Another major source of noise in the system is TDC quantization noise. This noise can be suppressed by oversampling and averaging, a technique commonly used in $\Sigma\Delta$ ADCs and other oversampled data converters. If a measurement is repeated 'N' times with each having uncorrelated noise, then averaging these 'N' samples reduces the noise power by a factor of 1/N. Thus, the total noise or effective depth resolution after averaging 'N' samples would be:

$$\Delta z = \frac{1}{\sqrt{N}} t_{NOISE} = \sqrt{\frac{t_{SPAD\text{-}JITTER}^2 + t_{TDC}^2}{N}} \tag{3.9}$$

With averaging, the SNR improves by $10\log N$. With this basic understanding of averaging and how it can be used to reduce the noise level in the imager and therefore, improve the depth resolution, we propose two types of averaging in the imager as described below:

Spatial Averaging

It implies averaging in space i.e. averaging of measurements from different SPADs. To achieve such an averaging, multiple SPADs are grouped in a single pixel to indicate the ToF from single point on object. However, this implies that the total number of pixels in an array reduces. Therefore, if more spatial averaging is done, it leads to better depth resolution but at the cost of poorer lateral resolution.

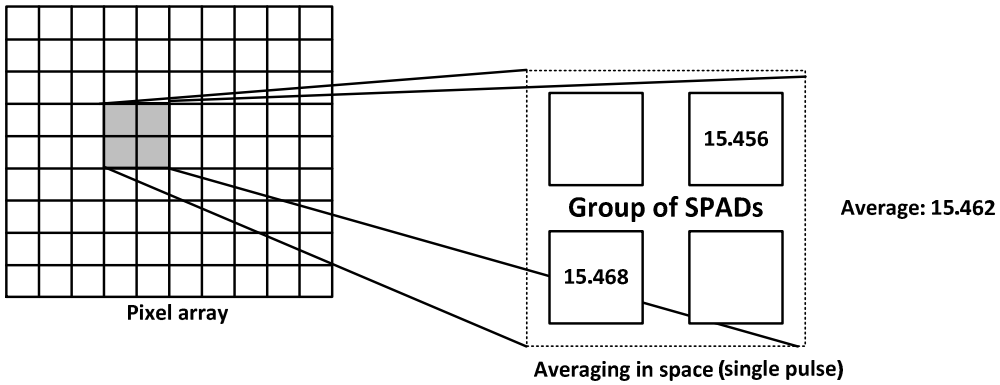


Figure 3.7 - Averaging in space (single pulse).

Temporal Averaging

It implies averaging in time i.e. averaging of measurements of a single SPAD occurring at different instants of time. As discussed earlier, this work employs the concept of '*train of pulses'* emitted in bursts. The measurement of ToF in all the pixels during a burst constitutes a frame. Assuming that the object of interest does not move over the duration of multiple bursts, which in the order of µs, the ToF measurements of consecutive frames can be averaged to reduce the error and improve resolution. However, multiple frames being merged together by averaging, lead to lower overall frame rate. Therefore, more the temporal averaging, higher is the depth resolution but smaller is the frame rate.

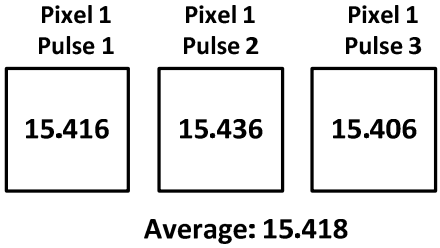


Figure 3.8 - Averaging in time (multiple pulses hitting the same pixel).

A mix of both the temporal and spatial techniques can be employed by the controller and the conceptual idea is shown in Figure 3.9.

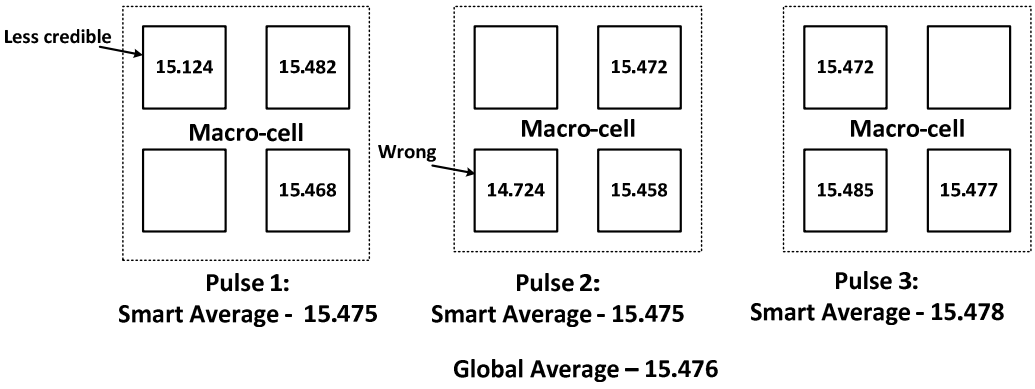


Figure 3.9 - Averaging in time and space across multiple pulses.

The controller has the intelligence to assess the less credible measurement and reject the wrong measurement after few measurements. This is possible by SNR computation.

Apart from temporal and spatial averaging, the imager is capable of many other functions. Most of such functions are handled by an on-chip controller and are discussed later in the chapter. We, first, look into readout methodology implemented in this imager.

3.2.6 Readout in 32x32 Pixel Array

Generally, an imager can include processing logic in every pixel to simplify the readout methodology. In this work, each pixel is a complex structure with its own SPAD and a TDC. Thus, we shift the complexity to the readout for an effective implementation of imager. Raw data is read out and high end processing is performed at readout level. This is a feasible approach from speed perspective given the high end processing possible in recent times.

3.2.7 Readout Mechanisms

The typical readout mechanisms for an array of pixels are:

Event driven readout

In this readout architecture, a pixel is read only when an event has occurred. This translates to a need for signalling mechanism through which the pixel communicates the occurrence of the event to the controller. It is a low power architecture as any pixel is read only when it has a new data from a photon hit but the complexity of the signalling mechanism overhead is counterproductive. Such a requirement demands explicit hardware and needs better timing control from the controller.

Sequential readout

This readout refers to serial readout of each pixel in an array. Figure 3.10 shows the two architectures possible for sequential readout: pixel readout and column readout. While the pixel based readout is very slow, its advantage lies in its low-cost as only one output pin is needed. A column readout requires more number of output pins but the parallelism leads to significantly faster readout.

The power consumption in both architectures is higher than event based readout as data from all pixels is read out irrespective of the occurrence of new event.

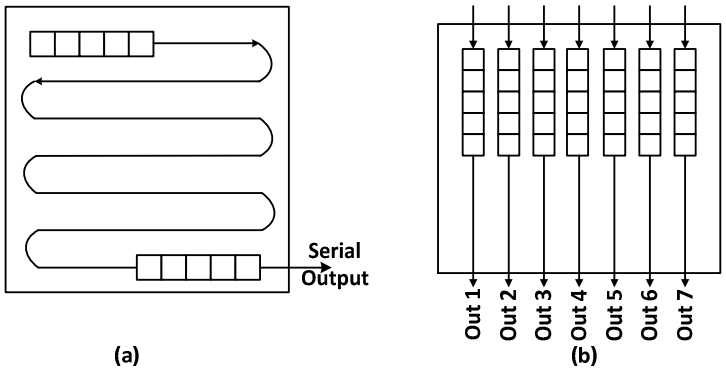


Figure 3.10 - Sequential readout: (a) pixel based readout (no parallelism); (b) column readout (limited parallelism) [23].

Since the pixel of this imager is complex and a low-cost 3D imager is desirable, a column readout is preferred. This architecture allows faster readout and is simple to design as well.

3.2.8 Readout Blocks

To enable column based readout, a row decoder and a serializer are required to select the pixel and then, convert its parallel data into serial bits. Thus, in a 32x32 imager array as shown in , the readout has the following functions:

a. A row selection is made to select a pixel in the column by a row decoder,
b. The multi-bit pixel data is loaded in the column bus through access transistors,
c. The data is latched in the serializer FFs,
d. The serializer shifts the parallel data serially out of the chip.

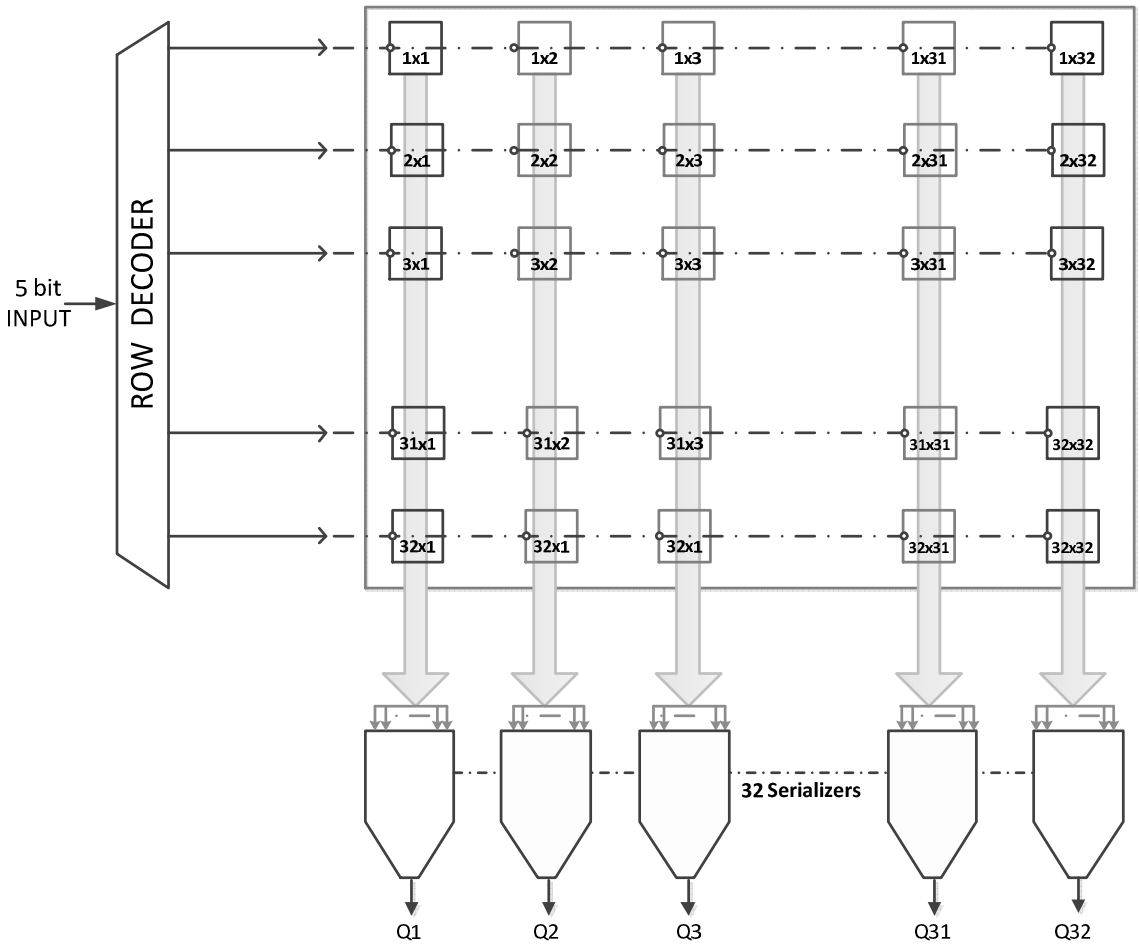


Figure 3.11 - Block diagram of system readout.

3.2.9 Estimation of Readout Speed

In this prototype, a column based readout has been employed. It is important to consider the serializer speed as it places a limit on the frame rate. For a 100x improvement in imager resolution using temporal averaging (to achieve 1 mm accuracy with SPAD jitter ≈ 300 ps), 10^4 raw frames are required. A 100x averaging will increase the resolution to ~3 ps which is equivalent to 1 mm. Furthermore, if the required overall frame rate is 30 fps, it implies that

$30 \cdot 10^4$ raw frames per second must be readout. In other words, the entire SPAD array must be read out $3 \cdot 10^5$ times per second. It also implies that $30 \cdot 10^4 \cdot 10^3$ SPAD events occur per second approximately for an array of 32x32 (~1000 pixels).

A readout speed of 300 kHz implies that the readout of full array must be within 3.33 µs. As the column readout mechanism has been employed, the time required for readout of full array is relaxed to readout of single column and thus, the readout of 32 pixels in each column must be within 3.33 µs. This implies that the pixel readout must be complete within 100 ns. Therefore, a serializer must receive parallel data stream, convert to serial bits and shift it within 100 ns.

Furthermore, if each pixel has 'm' data bits for readout, each pixel bit must be read out in $\frac{100}{m}$ ns. Suppose 'm' is 20 bits, then the readout of each pixel-bit must be completed in less than 5 ns. This is possible only with an IO pad speed of ~200 MHz. Therefore, a serializer working at 200 MHz ensures the desired millimetre accuracy and frame rate of 30 fps within the constraints placed by IO speed. If higher overall frame rate is desired, with the same given IO pad speed, the depth resolution would be compromised as the improvement in resolution by averaging would be:

$$\sqrt{\frac{300\,\mathrm{k}}{1000\,\mathrm{frames}}} = \sqrt{300} \sim 17x \tag{3.10}$$

Thus, the trade-off between accuracy and frame rate lowers the accuracy improvement to ~15x for a frame rate of 1000 fps.

3.2.10 Power Consumption

The upper limit of power consumption by the full imager is 1W. Extrapolating the target average power per pixel for a 1000x1000 array implies that each pixel can consume 1 µW maximum power.

Looking back, from the readout calculations, it was shown that $30 \cdot 10^4$ events/pixel/second are required for a typical frame rate of 30 fps and 100x improvement in resolution. The following calculations analyse the energy consumed per event,

$$\text{Power Budget} = 1\,\mu\text{W/pixel}$$

$$\text{Number of events per pixel per second} = 30 \cdot 10^4$$

$$\text{Average energy per pixel per event} = \frac{1\,\mu\text{W}}{30 \cdot 10^4} = 3.4\,\text{pJ}$$

The above calculated energy per pixel per event indicates the average energy consumption that must be targeted for a pixel.

To meet such energy requirements, it is of foremost importance that each pixel is event-driven i.e. the pixel operates only when a photon hits its photodetector. Only in this way, the average energy consumption can be related to the number of photons hitting the SPAD surface. Moreover, clock gating by event is also important to minimize dynamic power consumption in pixels all over the array.

The average energy per pixel per event is 3.3 pJ. This is a very stringent requirement but is tackled at system level through the technique of *Reconnaissance*. With this technique, the clock period can be reduced to much lower than 200 ns and thus, the TDC in each pixel operates for a much shorter duration burning much smaller energy.

3.2.11 Clock Distribution

The task of a clock network in an imager is to provide the clock signal to each pixel with tolerable rise time and minimum skew. The skew is unavoidable in any clock distribution network; however, it can be calibrated for if it is deterministic in nature. The effectiveness of clock distribution network is strongly dependent on the clock frequency. Higher the frequency, stronger is the dominance of parasitics and larger is the area consumption along with design effort required in designing the distribution network.

With the ToF scheme devised in Section 3.2.3, the clock frequency is equal to the range of object in focus. As the range equal to 200 ns, the clock frequency is 5 MHz. For lower power consumption, the clock frequency is increased to 200 MHz. This is relatively easy to distribute in 65 nm technology.

While building the clock tree network, it is important to consider the clock requirements of a pixel in the array. In Chapter 4, we look into the detailed pixel architecture and the clock requirement of each pixel building block. To summarize, each pixel needs a differential clock signal with minimum skew.

The above set of pixel requirements place strong timing constraints on the clock tree. Firstly, two clock trees - one for clock and other for $\overline{\text{clock}}$ are required. Even if large effort is spent on optimal design of such clock trees conforming to the desired timing requirements, on-chip PVT variations and random mismatch can easily alter the performance and in turn, affect system performance.

In order to avoid timing issues associated with clock tree and simplify the clock tree design, an additional clock-conditioning block in each pixel is added to relax the requirements on clock tree network. By this approach, the global constraints are transferred to be handled locally in each pixel [46]. For this purpose, an additional block is introduced at pixel level called as *'Clock Retiming and Edge Aligner'* block which is responsible for following:

(a) It retimes the clock such that the clock is presented to the entire pixel circuitry only when an event (a photon arrival) has taken place. This ensures that the TDC state remains unchanged until the next photon hits the SPAD. It also reduces redundant switching in the circuitry lowering the power consumption.
(b) It aligns the clock edges with smaller skew to ensure correct TDC performance.

With such clock retiming and edge alignment in each pixel, the task of clock distribution reduces to single clock distribution at the global level. The architecture for clock tree network is shown in Figure 3.12 where two columns of clock buffers feed the 32x32 array. These columns are placed on either side of imager such that the pixel array is the core of the system. Each clock buffer drives 16 pixels in a column.

The two column structure reduces the skew and the 16th pixel in each row now has the worst skew instead of 32nd pixel. The clock distribution has a symmetric configuration and the skew is deterministic in nature and can be calibrated for off-chip. For the final array of 1000x1000 array, similar approach can be extended or an advanced clock tree design can be sought.

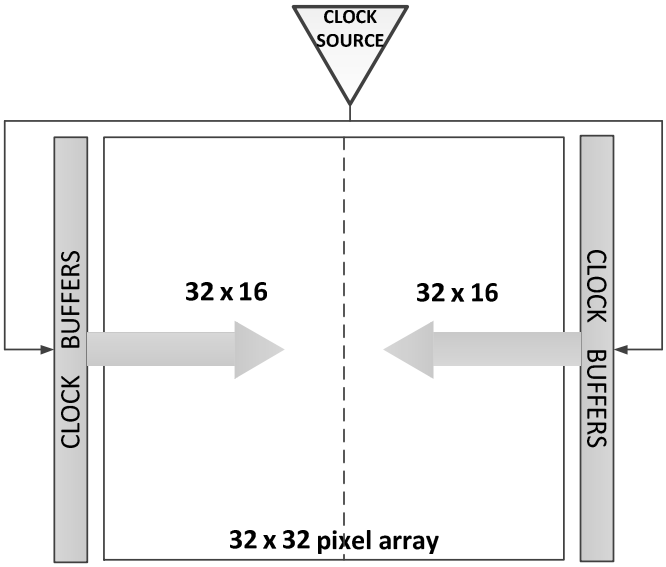


Figure 3.12 - Clock distribution network.

3.2.12 Controller

One of the critical blocks of the imager is the controller which regulates the imager operation and its performance. Multiple functions can be incorporated in the controller for high speed performance and few of the important ones are discussed below [28].

At Input Side:

a) **Mode of operation**
   According to [28], two modes of operation can exist – *crude acquisition* and *full detail acquisition mode*. The crude mode has a lower resolution but can capture objects quickly at any distance in the range of the imager (~ 30 m or above) whereas the full detail mode willingly limits its capture range to specific distances, has better resolution and smaller frame rate.

b) **Reconnaissance**
   As explained in Section 3.2.4, the concept can be applied to reduce the power consumption once the range of object is detected. The clock frequency can be altered depending upon the distance between the object and imager. The controller is responsible for the modification in clock frequency and with such scheme, the time duration for which TDC is in operation can be reduced to a great extent, thereby reducing the power consumption drastically.

c) **Blanking**
   The concept of *blanking* is employed to achieve higher SNR. When the imager is evaluating the range of the object, the photodetectors i.e. SPADs can be blanked to eliminate the dark counts in SPAD device. It translates to an effective reduction in

the noise of the imager, thereby increasing the SNR. The controller would also pause blanking at random instants of time to detect any change in field of view.

d) Just-in-time clocking and powering up
With this feature, the controller activates the cells only when it expects photon events. The global controller may also limit the number of acquisitions if the chip is becoming too hot. Thus, with this feature, it inherits intelligence to limit the power consumption of the imager ensuring an optimal performance while keeping the system secure.

At Output Side:

a) Readout
The controller drives the control inputs of the readout mechanism and also handles the output data from the pixel array for image construction.

b) Controlling Resolution and Frame Rate
The controller regulates the pixel configuration within the array to achieve a higher resolution or frame rate. The averaging mechanism can be either or both spatial and temporal depending upon the external user input. With this, the controller can adjust the spatial resolution, the depth resolution as well as the frame rate.
Under the assumption that the number of photons hitting the imager surface is constant (possible if emitting source has the same conditions), either more temporal averaging can be performed or a higher speed frame readout can be achieved. Higher averaging of samples implies higher depth resolution while faster frame readout indicates higher frame rate.

c) Signal Treatment
It controls the signal treatment and its further processing. In specific, more accurate timestamp from TDC can be obtained after processing or averaging of data.

Apart from handling the pixel array, the global controller can also regulate the laser source for controlling the power and frequency of emitted laser beam. This is important to meet the human-eye safety requirements.

3.2.13 Imager Architecture

Various concepts employed in the imager along with the functions of the controller, DSP, readout and clock-tree blocks have been discussed in this chapter. Figure 3.13 shows the interfacing between the pixels, controller, readout, clock-tree and DSP blocks.

Each pixel consists of a SPAD photodetector, a TDC and a readout mechanism which is controlled by the global controller. The controller also regulates the signal processing to obtain a compressed timestamp for image reconstruction. The controller also handles the clock distribution to the pixel array.

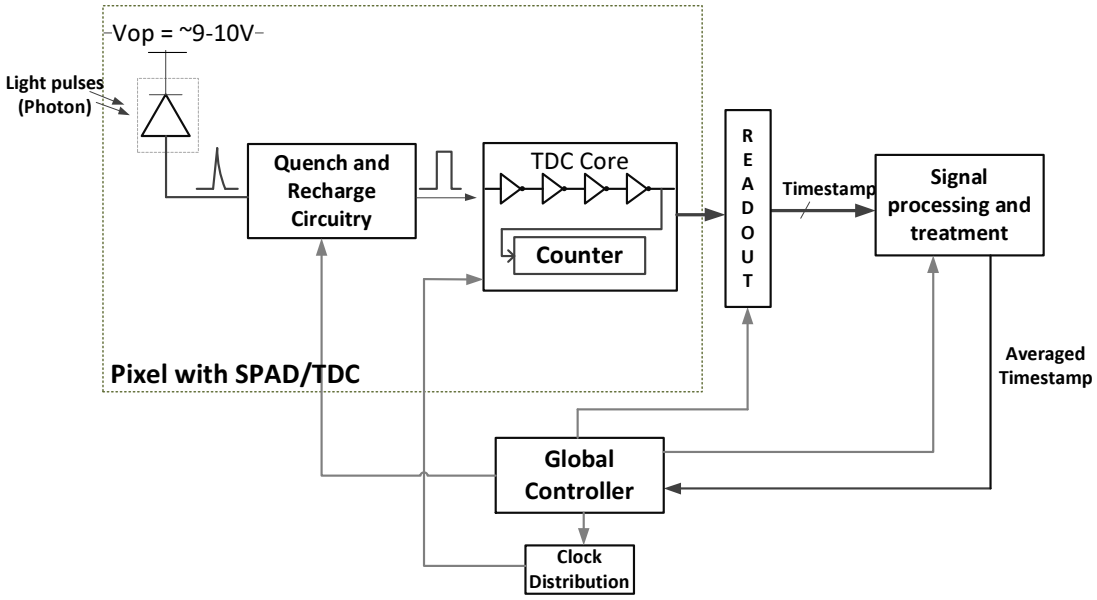


Figure 3.13 - Imager architecture [28].

3.3 Summary

In this chapter, the system design was carried out to derive system specifications. The imager architecture is discussed and the performance for different circuit blocks is drawn out. To enhance the performance of the system, multiple concepts are introduced such as temporal/spatial averaging, reconnaissance and blanking. Furthermore, calculations at system level show that the chosen readout scheme is a feasible implementation to achieve the desired power and speed.

4. Circuit Design & Implementation

This chapter describes the design and implementation of different blocks of a 3D imager. First, the appropriate design choice is made for each pixel building block. This is followed by its design to meet the desired specifications and finally, all the building blocks are integrated to fit within a 25 μm x 25 μm pixel area. A modular approach is employed to construct the 32x32 array and system level blocks are incorporated for correct operation of the 3D imager.

This chapter is organized as follows: Section 4.1 discusses the architectural aspects of pixel building blocks, whereas Section 4.2 explains the detailed design process. Section 4.3 presents the implementation aspects for each designed block. Section 4.4 discusses the design of system level blocks and concludes with the final integrated array in 65 nm technology.

4.1 Pixel Level Building Blocks

In this section, the system level constraints are considered for the choice of architecture for each building block of a single pixel. A 65 nm SPAD photodetector has been already developed at the beginning of the design process and can be employed directly into the pixel.

4.1.1 Design of SPAD – TDC Interface Circuitry

When a SPAD biased above breakdown, in Geiger mode, absorbs a single photon in the visible and near-infrared wavelength range, an avalanche is created which indicates occurrence of an event. To control the SPAD operation, a front end interface circuitry is required for the following tasks:

iv. **Sensing:** The photon absorption indicates an event occurrence which needs to be propagated to TDC.
v. **Quenching:** After avalanche occurs, it needs to be halted to avoid large flow of current in the circuit.
vi. **Recharge:** The SPAD needs to be re-biased in Geiger mode to be able to detect a photon again.
vii. **External control**: Unique features such as blanking of SPAD may be required by many applications.

Each of the above described tasks is crucial for the front end circuitry of SPAD. Various integrated quenching and recharge circuits have been proposed in the literature. Active quenching/recharge schemes occupy a large area and therefore, in this design, passive quenching and recharge schemes have been implemented to realize a pixel in smallest possible area with sub-optimal performance in terms of dead time.

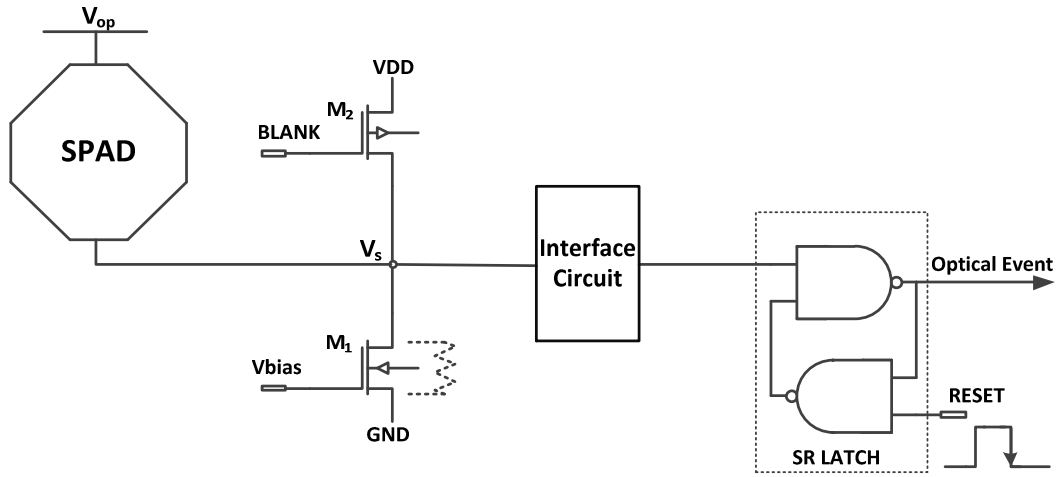


Figure 4.1 – The SPAD-TDC interface circuit.

In the SPAD-TDC interface circuitry shown in Figure 4.1, the transistor M_1 acts as ballast resistance. Its gate voltage is tunable from off-chip and thus, the resistance can be varied externally. M_2 is used as a blanking control for SPAD. If BLANK signal is logic LOW, M_2 conducts and biases the SPAD below breakdown. In this way, it turns off the SPAD. The interface circuit connected to node V_S needs design focus as it senses the SPAD signal and indicates the occurrence of an event. It acts as a START signal for TDC. Finally, the SR latch keeps the START signal to logic HIGH until clock arrives to reset it.

4.1.2 Interface Circuitry

The 65 nm SPAD devices are sensitive and noisy and limit the applied excess bias voltage to 0.2V-0.3V [48]. As the maximum SPAD signal amplitude is equal to excess bias voltage, it is important to detect such small amplitude. Several interface circuits were considered to suit this requirement and few of the important ones are discussed as follows.

a) Inverter

This is one of the traditional ways to interface a SPAD with the processing circuitry. However, in this implementation, the inverter required for sensing avalanche needs a low threshold so that it can detect a SPAD signal of 0.2V-0.4V amplitude (node V_S in Figure 4.2).

The typical inverter threshold V_M for a minimum sized inverter is 0.55V. To reduce V_M, it is necessary to shift the transfer characteristic to the left, i.e. its pull-down needs to be stronger. To achieve this, NMOS width has to be as large as possible to reduce the on-resistance. At the same time, longer PMOS length increases pull-up resistance adding to the desired effect of small inverter threshold.

However, the influence of sizing the transistor on inverter threshold is restricted to few tens of millivolts only and therefore, an inverter, is not appropriate as an interface circuit. An additional technique to reduce inverter threshold is to decrease the supply voltage. However, this translates to an extra power line and is costly in terms of implementation.

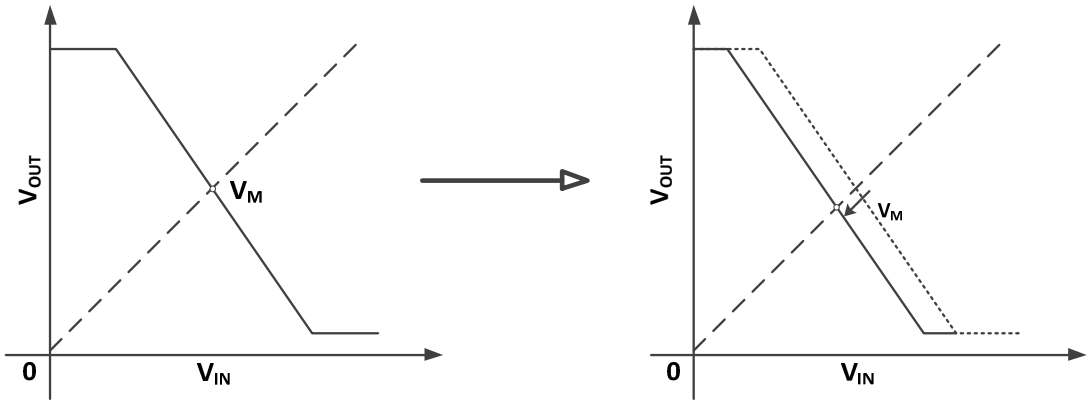


Figure 4.2 - Transfer characteristic of an inverter. The inverter threshold is modified by shifting the curve left or right.

b) PMOS Source Follower

In a PMOS source follower, V_{BIAS} signal controls the current through the signal NMOS transistor. The circuit allows small signal amplitudes to be measured and the output is a shifted version of input signal. The source follower is attractive for its small size but its small output swing (limited by current and threshold voltage) disqualifies it.

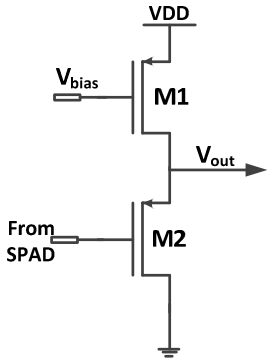


Figure 4.3 - PMOS source follower.

c) Comparator

Inspired from differential amplifier, the circuit shown in Figure 4.4 compares the SPAD signal against a small V_{REF} voltage. The current is controlled with V_{BIAS} and V_{REF} is kept approximately 0.1V-0.2V so that SPAD signal can be detected.

When SPAD signal is 0V, most current flows through transistor M1 resulting in low output voltage. When the SPAD signal exceeds V_{REF} voltage, a larger current flows through M2-M4 branch which produces a high output voltage. Due to the current mirror load and small amplitude input signals, the output swing is limited and an inverter is required at the output for regeneration.

Figure 4.4b shows a 4 transistor comparator circuit without current biasing. Thus, it translates to smaller area along with a reduction in an external bias control input. It is therefore, preferred over the previous configuration. One of the disadvantages of such a

circuit is the static current consumption, in the order of few tens of µA. This architecture, therefore, results in sub-optimal performance in terms of power consumption. However, this static consumption is suitable for a 32x32 pixel array prototype and tries to verify the correctness of architecture in smallest possible area.

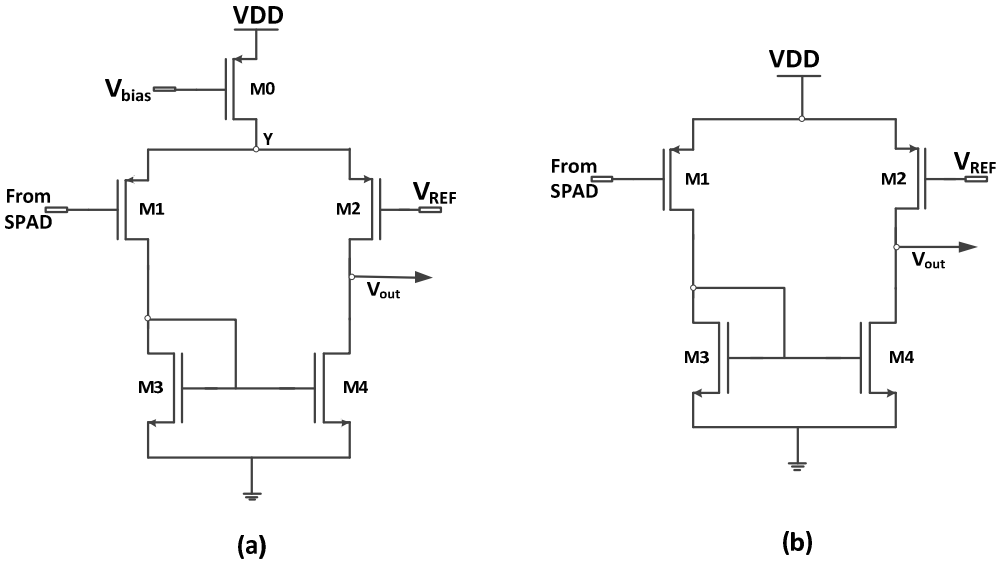


Figure 4.4 - Comparator circuits: (a) with current source; (b) without current source;

From the above discussion, it can be concluded that a low threshold inverter is an ideal way to interface SPAD with TDC. However, it is only possible with LVT NMOS transistors in this implementation. For this prototype, a four transistor comparator circuit proves to be the most robust way to interface 65 nm SPAD with TDC in small area. Although it has static power consumption, it is a worth-while trade-off for this prototype.

SR Latch

A cross-coupled pair of NAND gates is used to build SR latch where input 'S' is the photon event and input 'R' is the clock signal. In idle state, S=R=1 and the output is held as previous state (Q=0). When photon arrives, S goes from 1 $\rightarrow$ 0 with R=1, changing Q to 1. Due to quenching and recharge action by ballast resistance, S may change back 0 $\rightarrow$ 1 while R = 1, and therefore Q is held at 1. Only when clock arrives or R switches from 1 $\rightarrow$ 0, Q switches to 0. The clock can toggle back to its initial state and Q is held at 0.

With such a latch, it is possible to trigger the TDC ring operation. The output of SR latch is the 'START' signal of TDC and maintains logic HIGH until clock arrives and resets it. The arrival of clock edge indicates that time interval measurement by TDC is finished.

Table 4.1 – Truth-table of SR latch.

Sb	Rb	Q	Qb
0	0	1	1
0	1	1	0
1	0	0	1
1	1	HOLD	

4.1.3 Design of TDC Architecture [46]

In Section 2.3.3, it was concluded that the existing ring oscillators [8], [40-42] lack at least one of the desired criterion for a TDC appropriate for an million pixel imager in 65 nm technology. Thus, a novel architecture is proposed in the next few sections which meets the desired application requirements.

This section proposes a ring oscillator based TDC suitable for a million pixel 3D imager. To begin with, we summarize the specifications derived in Section 3.1 with respect to TDC design requirements.

Table 4.2 - Specifications for TDC.

Parameter	Value
Pixel Area	25 µm x 25 µm
Resolution	Inverter Delay (Less than SPAD jitter of 100 ps - 300 ps)
Dynamic Range	14 bits
Average Energy	3.4 pJ/pixel/event

In common terms, a TDC behaves like a digital stopwatch which counts a time interval between 'Start' and 'Stop' inputs. A photon hitting the SPAD surface implies occurrence of an event. This acts as a 'Start' signal for the TDC to begin counting. The 'Stop' signal for TDC is given by the system clock which halts the counting. A looped architecture of TDC is shown in Figure 4.5. As the ring starts to oscillate, the binary counter increments every alternate cycle. With such a counter, it is easy to extend the dynamic range of the imager to higher bits. The delay line produces an N-bit pseudo-thermometer code which can be converted to $\log_2(N+1)$ bit binary code to be merged with M-bit counter code.

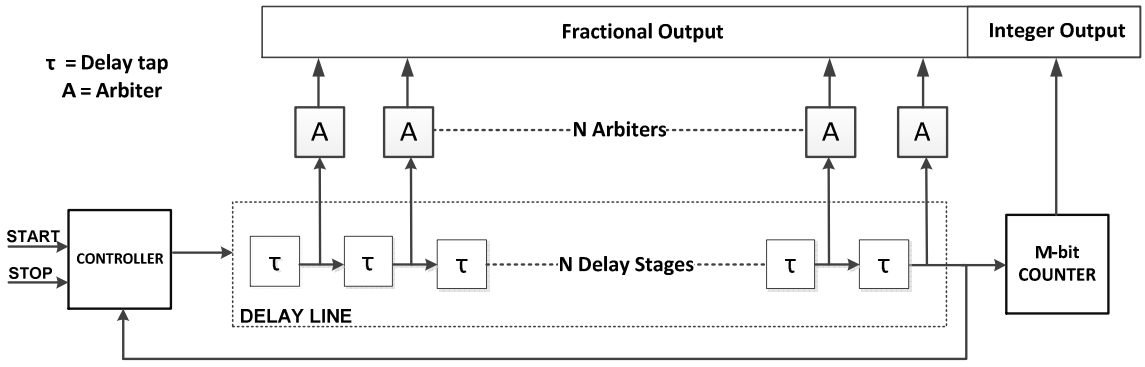


Figure 4.5 - Block diagram of ring oscillator based TDC.

As this 3D imager is the first ever implementation in 65 nm, we decided to implement the simplest possible topology to reduce the risk of unforeseen non-idealities, at a risk of sub-optimal and over-sized design. For this reason, static CMOS logic has been preferred in the TDC design. Although transmission gate logic is small in area, it is dynamic in nature with

back-driving and non-restoring (floating nodes) properties. Therefore, other topologies such as dynamic logic or pseudo-NMOS logic have not been considered for this design.

4.1.3.1 Choice of Delay Tap

Constructing a delay line in the looped configuration is critical to achieve the desired performance. In literature, several kinds of delay elements have been used depending upon TDC architecture and its requirements. Following is a brief summary of the performance for each delay element and the architecture where it is most suitable.

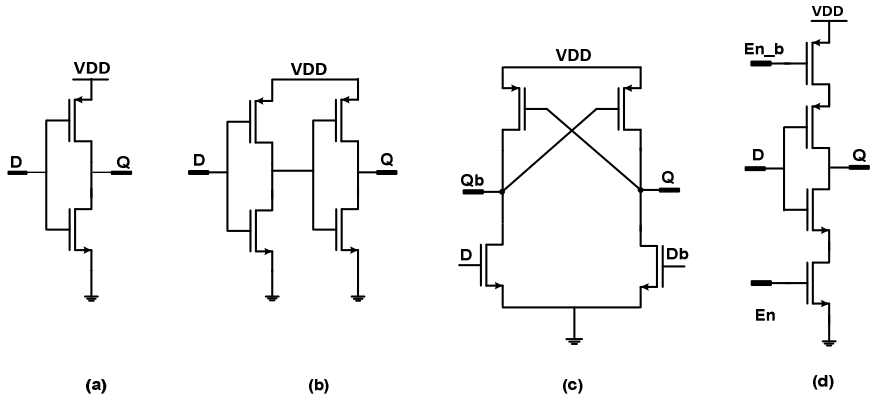


Figure 4.6 - Different types of delay taps: (a) inverter; (b) buffer; (c) differential inverter; (d) switched inverter.

Table 4.3- Comparison of different delay taps.

	Inverter	**Buffer**	**Switched Inverter**	**Differential Inverter**
Architecture	Simplest delay element	Two inverters in a row	Inverter with an enable	Cross coupled inverters
Resolution	Highest gate level resolution, of the order of 15 ps	Twice the inverter resolution, of the order of 30 ps	Lower than inverter resolution, of the order of 25 ps	Lower than buffer resolution, of the order of 40 ps
Area	1 PMOS + 1 NMOS	2 PMOS + 2 NMOS	2 PMOS + 2 NMOS	2 PMOS + 2 NMOS
Existing TDCs	Single delay line [32]	Vernier delay line [34]	Ring oscillators [42]	Ring oscillators [8]

From the summary shown in Table 4.3, a simple inverter proposes to be the best choice for delay line based on achievable area and resolution. The final choice is made after considering the loop dynamics.

4.1.3.2 Choice of Number of Stages in Ring

The number of stages in delay line depends on the following factors:

(a) Odd number of inverter stages is required for oscillation.
(b) The counter should be able to toggle at the frequency determined by the number of stages.

Based on the first parameter, a number of stages S = {1, 3, 5, 7, 9} is easily achievable. Beyond 9 stages, the area required by the delay line would increase beyond the given system specifications in Table 3.1. The second parameter implies that the maximum frequency by which a counter is limited is given by the propagation time of its flip-flop. To have a higher clock frequency, it is desirable to have a smaller clock-to-Q delay of flip-flop. This implies stronger switching speed of gates which is possible with larger area.

In this design, a trade-off appears between number of delay taps and the counter size. A standard flip-flop designed in 65 nm gives a propagation delay of ~300 ps i.e. the counter can operate close to 3.33 GHz. If the counter speed is to be doubled, it implies fewer delay taps but at the cost of larger flip-flops. As the counter resolves more bits than delay line, large sized flip-flops are less preferred. An optimum configuration for smallest area was reached with a delay line of 7 inverters along with an 11-bit counter to obtain a 14-bit dynamic range.

4.1.3.3 Controlling the Ring Oscillator

The control of ring oscillation is required for correct TDC operation and also to reduce power consumption. There are few possible methods to achieve this control and are as follows:

(a) Switch as a controller:

A switch enabled by output of AND gate controls the ring oscillation. In this approach, START is an active high signal whereas STOP is an active low signal (see Figure 4.7)

Advantages:

(i) All delay taps are built from identical inverters.

Disadvantages:

(i) Initial state is not known and reset of ring not possible.
(ii) Systematic error due to extra TG in propagation path - The last delay tap is equivalent to two inverter delays.
(iii) The STOP (or clock signal) has 2 AND gate input load.

(b) AND gate as a controller:

Advantages:

(i) Initial state is known as AND gate output is logic '0' when START is '0'.
(ii) No extra logic in the ring and therefore, systematic error is less than two inverter delays.
(iii) The STOP (or clock signal) has a single AND gate input load.

Disadvantages:

(i) Delay mismatch between NAND gate and an inverter brings unavoidable systematic error. However, sizing can be done such that the error is minimal.

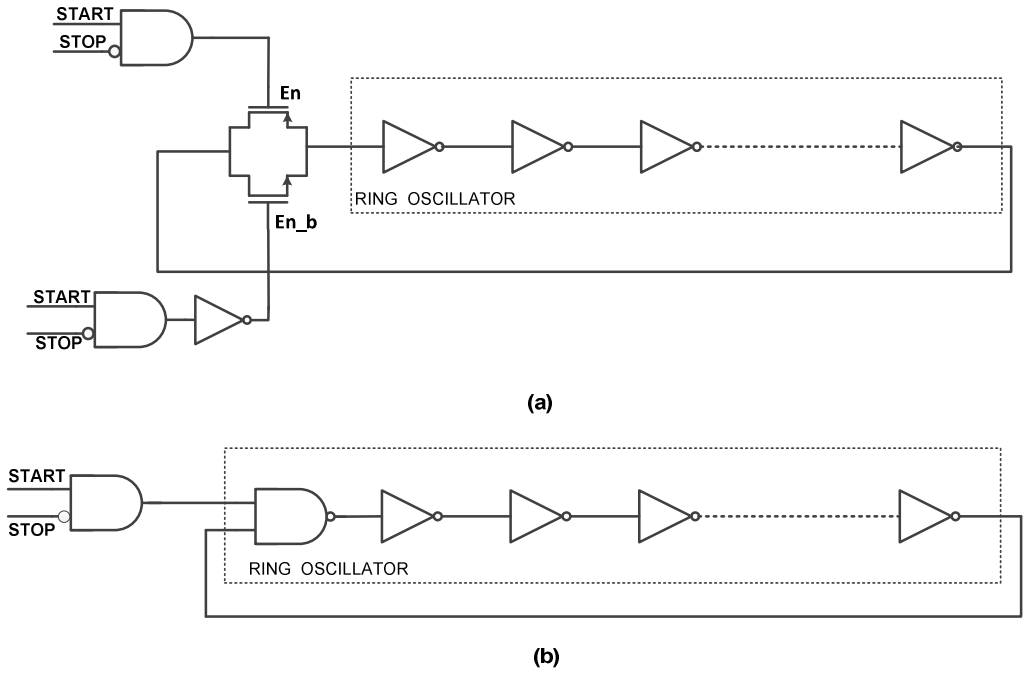


Figure 4.7- Methods to control the ring oscillation: (a) using TG; (b) using AND gate.

The second method is more advantageous and offers easy control of ring with appropriate transistor sizing. Therefore, it was adopted as the controlling mechanism for ring oscillator.

4.1.3.4 Choice of Delay Line Arbiter - A Discussion

The choice of delay line arbiter is critical in TDC design from both area and power perspective. An arbiter structure is required to capture the state of the delay line when the clock arrives. In the TDC architecture being worked upon, an inverter is the delay tap of the ring. Since there is a single delay line, as shown Figure 4.5, only single phase data input is available for arbiter input. Furthermore, the delay line changes its state at a high frequency due to ring oscillation behaviour. Different flip-flop and latch configurations are discussed in this section to arrive at the most suitable arbiter design. They are as follows.

Sense Amplifier based flip-flop [49]

This flip-flop is suitable for differential inputs and single phase clock, as shown in Figure 4.8. The pulse generating stage senses true and complementary differential inputs following the leading clock edge and assures accurate timing. The slave latch captures the transition and holds till next clock edge arrives.

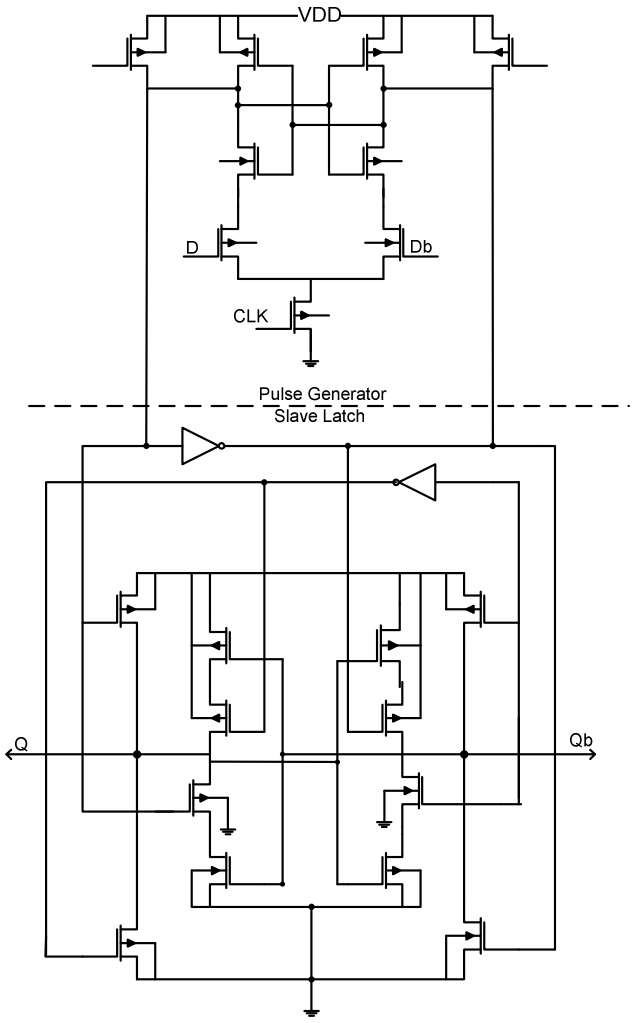


Figure 4.8 - Circuit diagram of high resolution symmetric flip-flop [49].

Since the first stage is a sense amplifier, it assures accurate timing which is critical at high frequencies. Moreover, it has a small clock load of a single transistor and allows reduced clock swing operation [49]. The slave latch is configured such that the flip-flop is symmetrical along the vertical axis providing equal resolution of the rising and falling edge metastability with respect to input data. This high speed architecture is useful in TDCs for digital PLLs [32].

However, this flip-flop is costly in terms of area with 24 transistors required for its working. It also does not fit the requirement of a single phase data input as desired by the TDC in Figure 4.5.

Master-slave Flip-flop

It is the most common architecture for a flip-flop and is constructed with two latches working in master-slave topology. The architecture suits the input requirements presented by delay line of TDC and can be optimized for timing requirements (setup/hold) of an application. But it occupies a large area of 14 transistors which restricts its use in this TDC.

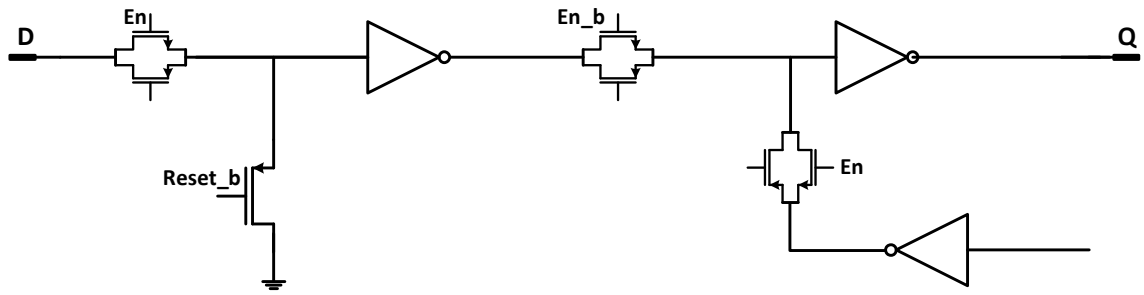


Figure 4.9 – Dynamic master-slave flip-flop.

From the above two discussed flip-flops, it can be seen that large area flip-flops are not suitable for compact TDCs aimed at imaging applications. Imagers are highly dense structures and a high fill factor demands each sub-block to be area effective. At this point, latch based designs were considered to see if they are suitable for TDC design. Some of the latch designs considered to capture the state of delay line are discussed as follows:

SRAM Cell (6T)

It is a very compact 6 transistor structure as shown in Figure 4.10. However, the architecture has a limitation due to the presence of race condition. As seen in Table 4.4, contention is possible in design and in order to overcome this contention, the size of access transistors M5 and M6 needs to be very large such that the M5/M3 and M6/M4 is very high. This leads to large area consumption.

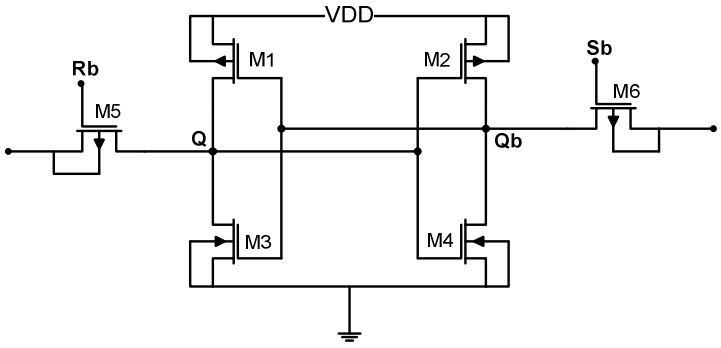


Figure 4.10 - SRAM based latch.

Table 4.4 - Truth -table of 6T latch.

Sb	Rb	Q	Qb
0	0	HOLD	
0	1	1	0
1	0	0	1
1	1	RACE	

NAND based SRAM Cell (8T)

This is a more robust version of SR latch and can be built with eight minimum sized transistors. A NOR based SR latch is also a possible configuration but the area required is larger.

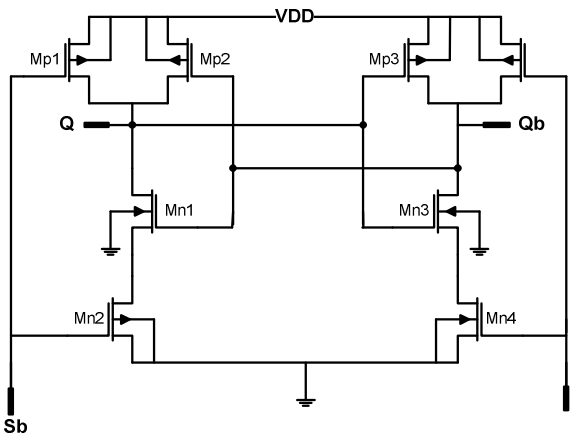


Figure 4.11 - NAND based SR latch.

Table 4.5 - Truth table of 8T latch.

S_B	R_B	Q	Q_B
0	0	1	1
0	1	1	0
1	0	0	1
1	1	HOLD	

The analysis of SR latch based arbiter structures shows that they suffer from several issues. To begin with, a latch is level-sensitive and requires that the state of the delay line remain constant after clock edge arrives in order to hold the valid data. For instance, consider the case when a NAND latch is used as an arbiter and its R_B input is used as the clock. When the clock is HIGH, the NAND latch can hold the value during a '0→1' transition but may change state in a '1→0' transition. In an inverter based ring structure, the ring may continue to toggle from '0→1' or vice-versa even after the clock edge arrives (until the transition reaches the first delay tap). Thus, to incorporate a latch to capture the state of delay line in an inverter based looped TDC, one of the following conditions must be met:

(i) either the readout of the latches must be done before the delay line next toggles, -

It implies that the readout must happen at clock edge and within one inverter delay. This is not a viable option at the system level as it is almost impossible to readout every pixel of array at the instant when clock arrives.

(ii) or the delay tap should be a buffer,

This translates to reduction in resolution to twice the original value. More importantly, it is not possible to use buffers in a ring oscillator based TDC.

(iii) or a combination of NOR and NAND based SR latch must be used.

NAND latch is sensitive to 0 and NOR latch is sensitive to 1. A delay tap experiences both 0→1 and 1→0 signals during ring oscillation. Thus, at each delay tap, a combination of NAND/NOR latches with additional control circuitry is required. This leads to a very complex implementation, and very costly in terms of area.

A flip-flop can be used to avoid these issues. However, it has high cost in terms of area and power as discussed previously. Alternate structures for the TDC delay line were considered that can capture the delay line state with as small area as possible.

Switch in Delay Line

In this structure, the ring propagation stops when its 'Enable' (or system clock) switches off.

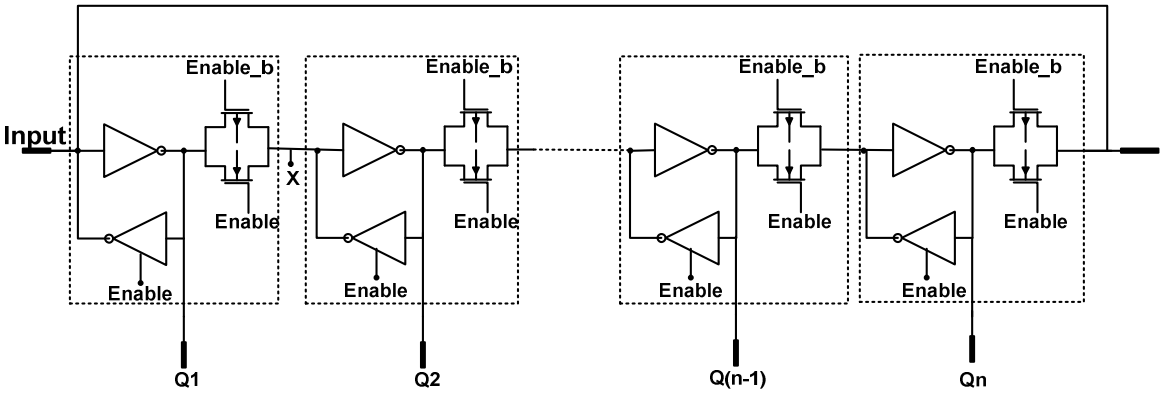


Figure 4.12 - Switches in delay line.

Although it is a highly compact static structure, it has the following drawbacks:

(i) The delay tap includes an inverter with a TG. In this technology, the delay presented by TG is similar to a single inverter and thus, the TDC resolution decreases to 40-60 ps.

(ii) The TG gate is dynamic in nature with back-driving ability and therefore, contention may exist between one stage and the feedback inverter of the next stage; as shown by node X in Figure 4.12 where contention may occur.

Switched Inverter based Delay Line

This implementation is similar to Figure 4.12. However, the approach is static in nature and is devoid of TGs. Switched inverter structure as shown in Figure 4.6(d) has an inherent control with an 'Enable' signal. The propagation stops when clock edge arrives and thus, an inverter or back-to-back inverters can be used to hold the state till readout occurs.

However, the resolution of a switched inverter is higher than that of a single inverter, of the order of 25 ps. A minimum sized normal inverter has a resolution of 15 ps. Thus, such a delay line reduces the resolution. Also, metastability is dominant at the stage where clock edge is closest to the change in state.

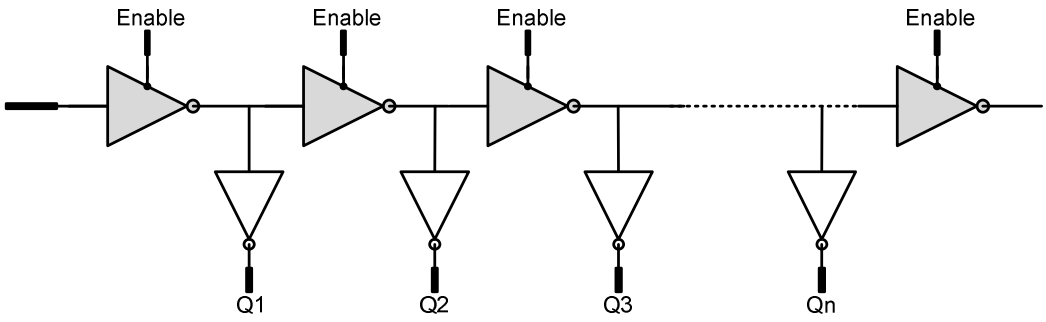


Figure 4.13 - Switched inverter based delay line.

The following observations can be made about the architecture of delay line and in order to identify the appropriate flip-flop.

(i) a single inverter in the delay line seems to be the best way to implement a delay line TDC to achieve highest gate-level resolution.
(ii) TG based implementation should be avoided due to its dynamic nature.

Proposed Latch – Based on Switched Inverter

The latch shown in Figure 4.14 propagates D to Q when its 'Enable' is HIGH. When the 'Enable' signal is LOW, the feedback inverter holds the output. It differs from SR latch structures as there is no change in output once clock arrives and it holds the data. Furthermore, it has symmetric behavior w.r.t. rising and falling input signals. To summarize, it has the following properties:

(i) it requires single phase data along with true and complementary clock signal,
(ii) the architecture is compact and needs only 10 transistors as opposed to master-slave flip-flop or Bora's flip-flop [49],
(iii) the TDC resolution remains best possible as it loads the delay line by a single inverter,
(iv) it is static in nature and holds data as long as clock gating exists. Thus, the readout can be done at a later point of time.

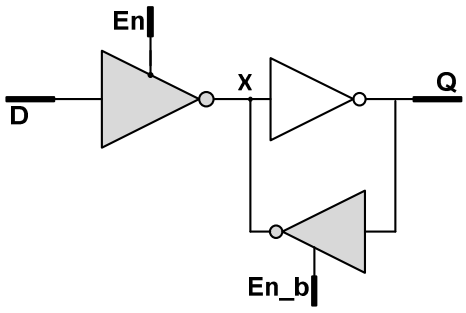


Figure 4.14 - Switched inverter based latch.

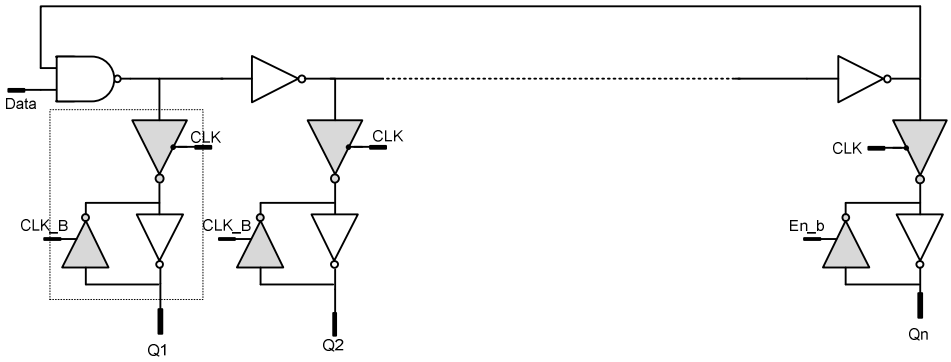


Figure 4.15 - Switched inverter based latch in delay line.

It is known that clock gating is required in the ring-oscillator based TDC proposed in this section to avoid over-writing of data held in the latches or flip-flops. If any flip-flop is employed, it translates to larger area than a simple latch structure as discussed in the

previous sections. In this architecture, a latch configuration symmetric to rising and falling input edge is sufficient for correct TDC operation. Thus, a switched inverter based latch is employed as it satisfies all the requirements of a compact size, inverter delay line based TDC.

4.1.3.5 Choice of Counter

In the TDC architecture, the counter helps to extend the dynamic range of the TDC to 14 bits. It interfaces with the delay line and as soon as ring completes two cycles, it increments by one bit. The trigger from the delay line can act as a synchronous clock to the DFLIP-FLOPs of counter or be an asynchronous clock input to the LSB flip-flop of the counter.

The two common types of counters are synchronous and asynchronous (ripple) counters. A synchronous counter changes its state under the control of clock. A ripple counter, on the other hand, is not controlled by a single clock but the state bits of a flip-flop clock the subsequent flip-flop.

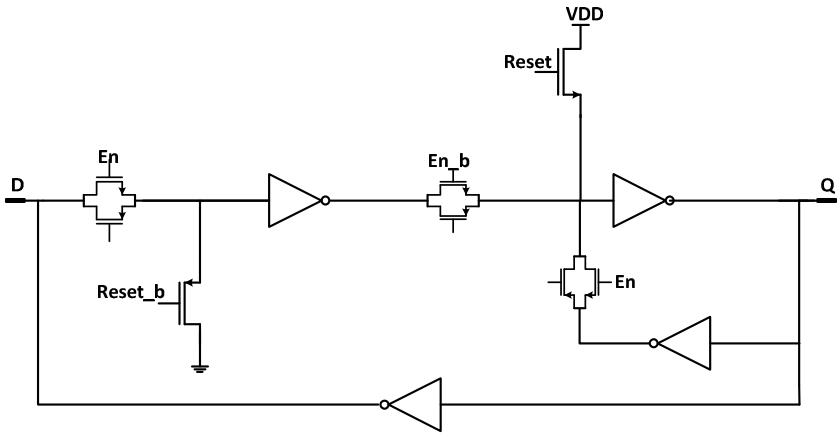


Figure 4.16 - A toggle flip-flop circuit.

While choosing the appropriate configuration for TDC, it was observed that the last delay stage interfaces with the counter and acts as the counter clock. The load on the last stage would be 10x times if a synchronous counter is used as opposed to asynchronous counter. Moreover, synchronous counter demands extra logic in terms of AND gates apart from flip-flops [50]. Thus, a ripple counter was chosen for its simplicity and low-cost to extend the dynamic range of TDC.

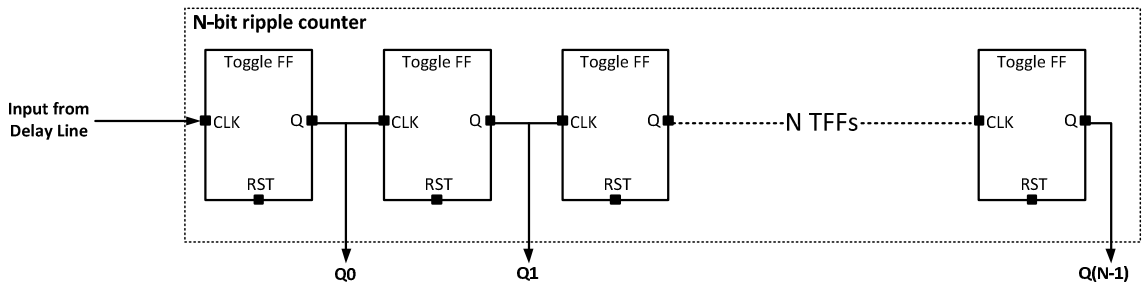


Figure 4.17 - Circuit of binary ripple counter.

A ripple counter is built from toggle flip-flops (T flip-flop) which toggle their output state at each clock pulse. Figure 4.16 shows a typical circuit of T flip-flops which is static in nature such that it can hold data until reset occurs. A series of N 'Tflip-flops' form a N-bit binary ripple counter as shown in Figure 4.17.

4.1.3.6 Interfacing Ring with Counter

The output code of TDC is binary in nature as the counter increments by one bit when the ring completes two cycles. As shown in Figure 4.18, the easiest placing for counter is at the output of any delay stage. However, placing the counter at the last delay tap, as shown by node Y in Figure 4.18, seems a logical way to ease the output code generation. The placing of the counter at the output of any delay tap creates two issues:

(i) The loading faced by delay tap is different w.r.t other delay taps. In case of the last delay tap, it feeds a bigger NAND gate along with a counter flip-flop. Thus, it experiences a higher loading than other delay taps.

(ii) The offset in counter output code is different from that of delay line latches' output code. This can be understood with an example - every arbiter (latch or a flip-flop) has a setup time, say 100 ps. If a transition arrives at instant 'T' ps at the last delay tap, the delay line outputs would correspond to (T-100) ps whereas the counter outputs would correspond to T ps. This brings in a mismatch in offset among two output codes.

In order to solve the above mentioned issues, a different placement of the counter was proposed as indicated by node X in Figure 4.18. If the counter is placed at the output of last stage's arbiter, the mismatch in offset is zero. Furthermore, the loading of last stage of delay line is reduced.

4.1.3.7 Control Signals

The control signals for the TDC are:

(a) START: It indicates the occurence of an event and is generated when a photon hits the SPAD. A leading edge of 'START signal triggers the ring oscillation.

(b) STOP: It indicates the arrival of system clock and stops the TDC ring oscillation. The clock also captures the state of delay line. Both the delay line latches and TDC operation is controlled by the falling edge of the clock.

(c) RESET: The leading edge resets the counter flip-flops after every frame is readout. Its complementary signal '$\overline{\text{Reset}}$ ' is also required.

4.1.3.8 Proposed Circuit - Complete Picture

Figure 4.18 illustrates the architecture of single delay line based ring oscillator. It consists of 7 delay taps each being a single inverter resulting in highest possible gate level resolution in a technology. The switched inverter based latches sample the state of delay line at the negative edge of clock and store the value for further processing. An 11-bit counter extends the dynamic range of the TDC to 14 bits. Event gating is also employed in this architecture to

minimize power consumption [51]. Thus, the ring oscillates only between the START and STOP signals of TDC for time interval measurement.

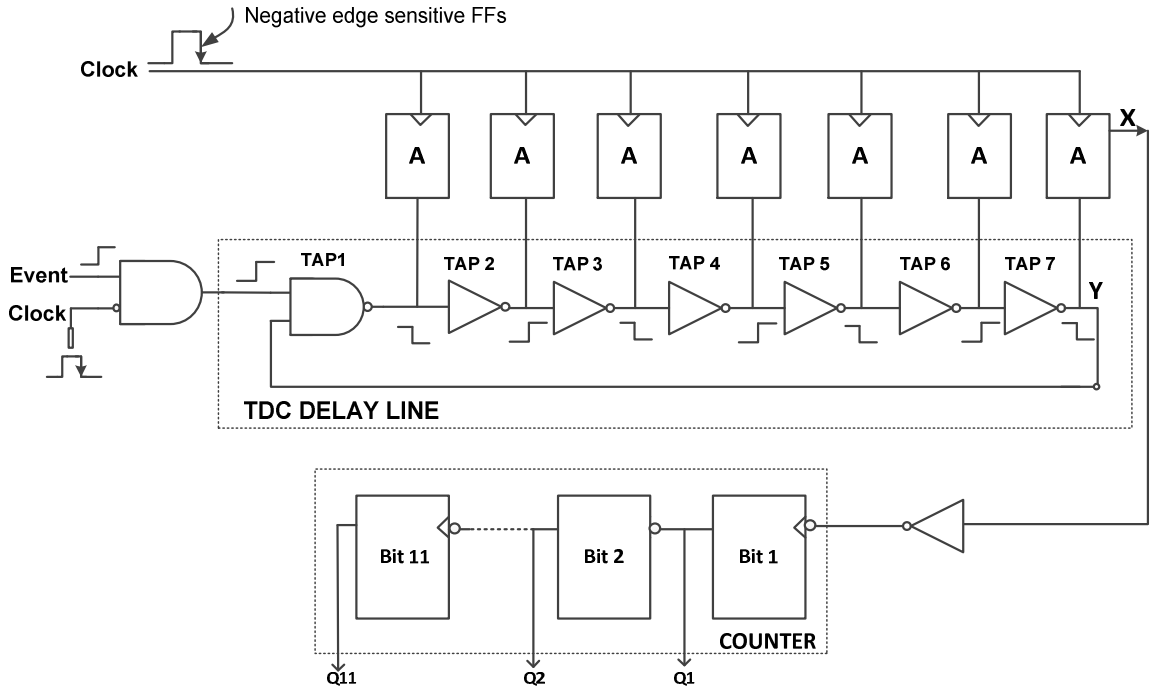


Figure 4.18 - Ring oscillator based TDC.

The delay line presents a pseudo-thermometer code whereas the counter generates a 10-bit binary code. The 7-bit pseudo thermometer code is equivalent to 3 binary bits. It should also be noted that the last delay tap output is the inverse of the LSB bit of binary code. Thus, the counter has been designed for 10 bits instead of 11 bits. The conversion of pseudo-thermometer code to binary code along with accounting for 11th bit can be done during post-processing outside the chip.

4.1.4 Reducing Power Consumption

The switched inverter based latches used in the TDC are transparent when clock is logic HIGH. Thus, the activity in these latches is high as each transition in the delay line is propagated to the latch output, as shown in Figure 4.19. However, such propagation of the ring oscillation transitions is redundant and expensive in terms of energy.

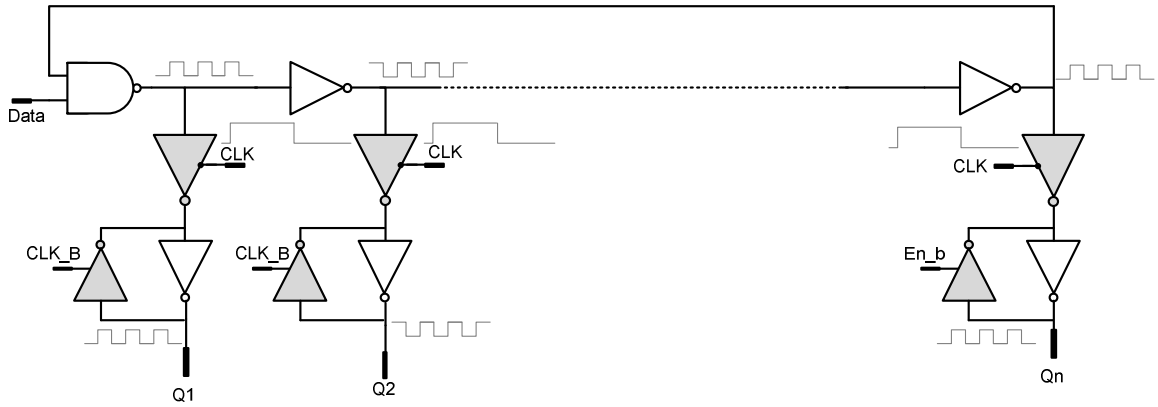


Figure 4.19 - High energy consumption with clock having 50% duty cycle.

The high energy consumption is due to a clock duty cycle of 50%. A new technique was investigated to reduce the energy consumption. If the duty cycle is reduced to much less than 50%, it is possible to lower the required energy. This can be made possible by a pulse generation circuit as shown in Figure 4.20. With such a clock pulse, the latch is transparent for a small time in order to capture the state of delay line at the negative edge of clock.

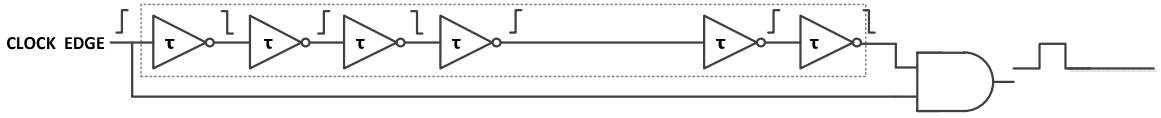


Figure 4.20 - Low duty cycle clock generation.

In the delay line, such a clock configuration is possible for the first six delay stages only. The last delay stage feeds a counter and needs to be transparent as long as ring oscillation goes on. Thus, the last latch needs 50% duty cycle clock. This requirement adds penalty to clock distribution networks and poses the following issues:

(i) If the clock pulse generation is done globally, the clock distribution network for the last delay stage of all pixels needs to be different from the network of first six delay stages of all pixels. This brings skew in clock input of flip-flops which is unavoidable even if the design is done carefully. Moreover, two distribution networks would be required, one for clock and other for $\overline{\mathrm{clock}}$. Furthermore, unequal loading of clock and $\overline{\mathrm{clock}}$ signals (due to difference in parasitics and routing) adds complexity to the clock distribution networks.

(ii) If the clock pulse generation is done locally in each pixel, the global clock distribution is simplified but at the cost of complexity increase at local level. Also, an offset would still exist between the clocks generated from two different networks built locally. Moreover, the large area required to incorporate a clock generation circuit may dominate over TDC and SPAD circuits, thereby lowering the fill factor of pixel tremendously.

4.1.5 Simplifying Clock Distribution Network [46]

It can be seen that in order to lower the power consumption, the area cost w.r.t. clock distribution increases tremendously. However, if a clock with 50% duty cycle for all the DL latches is suggested, the power consumption of the TDC still meets system requirements.

Therefore, a simplified clock distribution scheme is possible from system level which eliminates clock related issues such as skew and unequal loading of $\mathsf{clock}\,/\,\overline{\mathsf{clock}}$ signals.

One of the other observations made during system design was that the TDC delay line counter outputs must change only when an event has occurred. As explained earlier, the ring continues to oscillate even after the arrival of the STOP signal until the ring reaches its initial state. Now, if a new clock pulse arrives before the readout has happened, the state of the delay line latches would change giving a wrong output during readout. Thus, it is important to gate the clock signal in every pixel.

From the above discussion, the clock distribution has the following requirements at the pixel level:

(a) the clock signal must be differential in nature,
(b) the differential clock signals must not be separated by more than 20 ps, for correct TDC operation,
(c) clock gating is required for lower power consumption as well as to ensure TDC state is not overwritten before its readout,
(d) as the TDC latches are sensitive to positive level, the negative edge of clock must arrive earlier than its positive edge,
(e) it acts as a 'STOP' signal to TDC and resets the SPAD-TDC interface memory cell. In both cases, negative edge of clock is required.

4.1.6 Clock Retiming And Edge Aligner [46]

Keeping in mind the above stated observations, a pixel level clock retiming circuit along with an edge aligner is proposed. With such a circuit in each pixel, all the requirements from clock distribution perspective are met and the entire pixel as well as system design simplifies to a great extent.

The retiming circuit is built by considering the desired requirements of the local clock in each pixel. Figure 4.21 shows that only when an event has occurred in a pixel, the clock edge is propagated. The desired output depends on past state of the inputs and therefore, a sequential element is needed. A typical D flip-flop with global clock feeding its clock and pixel event feeding its data input suffices the requirement. The output $\overline{\mathsf{Q}}$ gives the desired output.

The latches in TDC need two-phase clock signal. While the D-flip-flop produces both phases, the loading of clock and $\overline{\mathsf{clock}}$ is unequal. The individual load for both signals is as follows.

clock load: 7 latches + 1 AND gate + 1 SPAD-TDC interface memory cell
$\overline{\mathsf{clock}}$ load: 7 latches

The unequal load signals brings offset in the complementary versions of the same signal and affects the output of delay line. In order to minimize the effect of this offset, an edge aligner was built which assists the TDC in giving the correct output code.

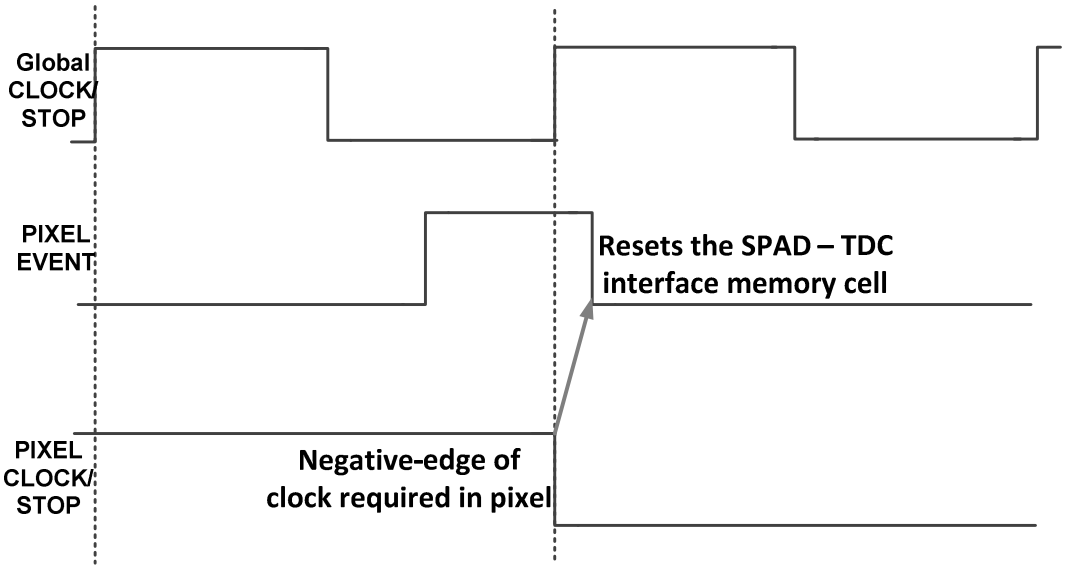


Figure 4.21 - Clock retiming waveform.

The edge aligner is built with cross-coupled inverters as shown in Figure 4.22. The D flip-flop along with edge aligner circuit are small in size and can be easily incorporated in every pixel without affecting the fill-factor. It fixes the skew issues between clock and $\overline{\text{clock}}$ as well. It also reduces the effort on global clock distribution as a single clock signal is only required by each pixel. Furthermore, the global clock skew is not an issue any more as it is deterministic and each pixel operates independent of it.

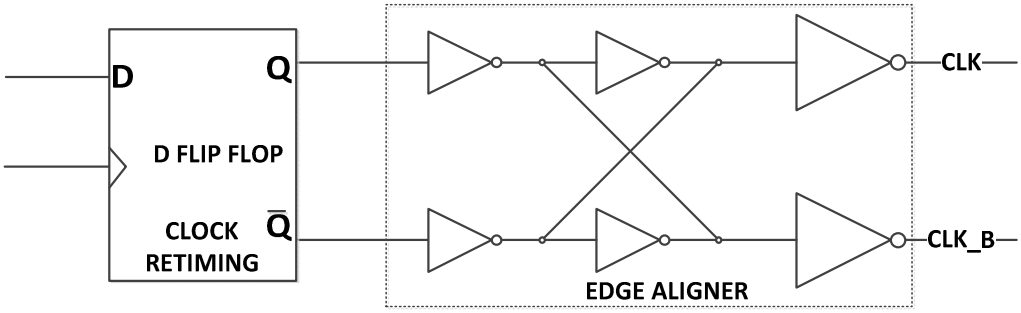


Figure 4.22 - Clock retiming and edge aligner circuit.

4.1.7 Pixel Design

After constructing each block individually, all the circuits are integrated. The complete pixel level circuit is shown in Figure 4.23. A MUX gate is added at the input of TDC to allow selection between an optical or electrical input. The electrical input from off-chip can be used for evaluating TDC performance and for its calibration.

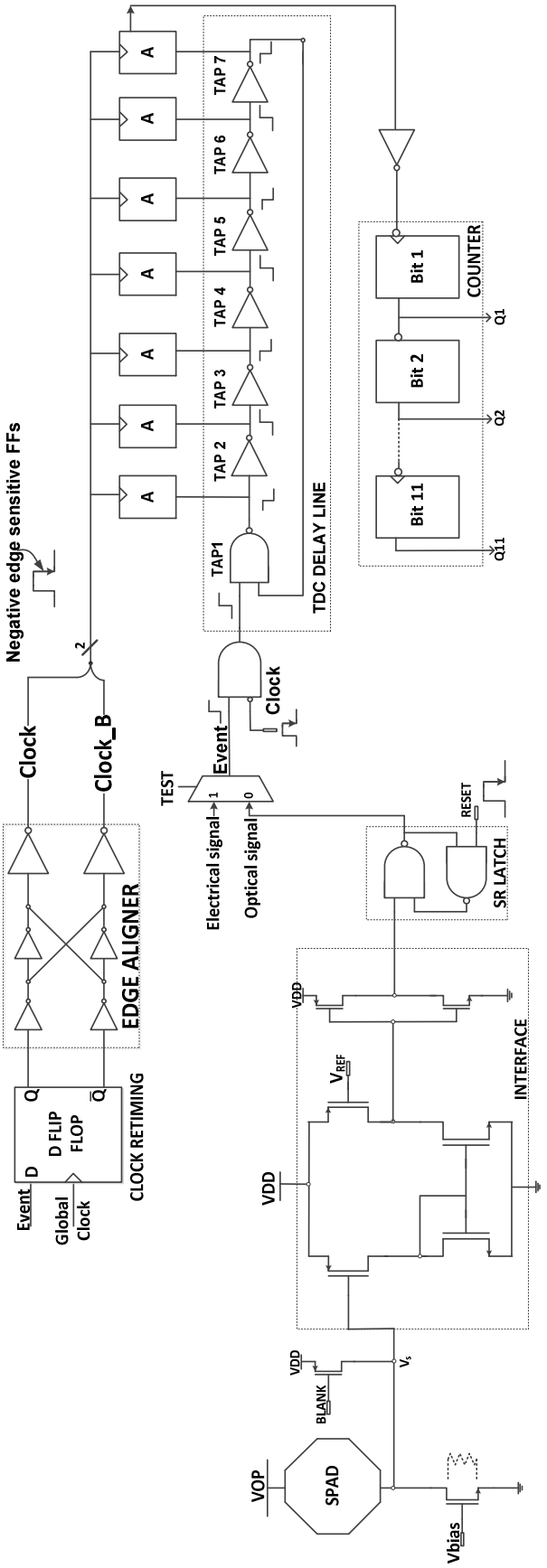


Figure 4.23 - Full Pixel Circuit.

4.2 Detailed Design of Sub-Blocks

This section presents the detailed design process for each block in the pixel. The pixel level design process was iterative in nature mainly due to dominance of parasitics in 65 nm technology at high frequency TDC ring oscillation.

An initial circuit was built to achieve the system specifications. However, after the first layout, the circuit performance was far away from the desired specifications. In order to achieve the specifications post layout, modeling of parasitics was required to allow appropriate sizing of critical blocks.

4.2.1 Parasitic Modeling

Modeling of parasitics is critical in a high frequency design; more so in DSM technologies where the parasitic capacitances become comparable to gate capacitances. In 65 nm process, depending upon the size of the transistor, gate capacitances vary from 0.5-2 fF, whereas parasitic capacitances range from 0.1-1 fF. This proves the dominance of parasitic capacitances as they are of the same order as gate capacitances. In order to have a correct circuit design, it is critical to account for parasitics as well. To consider parasitic behaviour during design process, parasitic modeling was carried out.

The high frequency block in the system is the ring oscillator based TDC with the frequency of oscillation of the order of 3-4 GHz. Therefore, it is most affected by the parasitic capacitances. To start with the parasitic modeling, a simple test-bench as shown in Figure 4.24 was built. It consists of an inverter having a load of two inverters and resembles a single stage of ring oscillator where a delay line inverter has two inverters as its load - one to denote the next delay line inverter and the other to denote the inverter of the latch. The variation in gate capacitance for increasing gate width is shown in Figure 4.25.

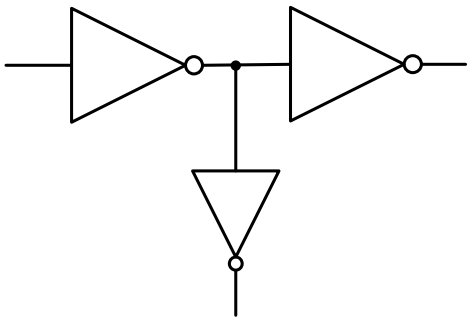


Figure 4.24 - Test-bench for parasitic modelling (single stage of TDC).

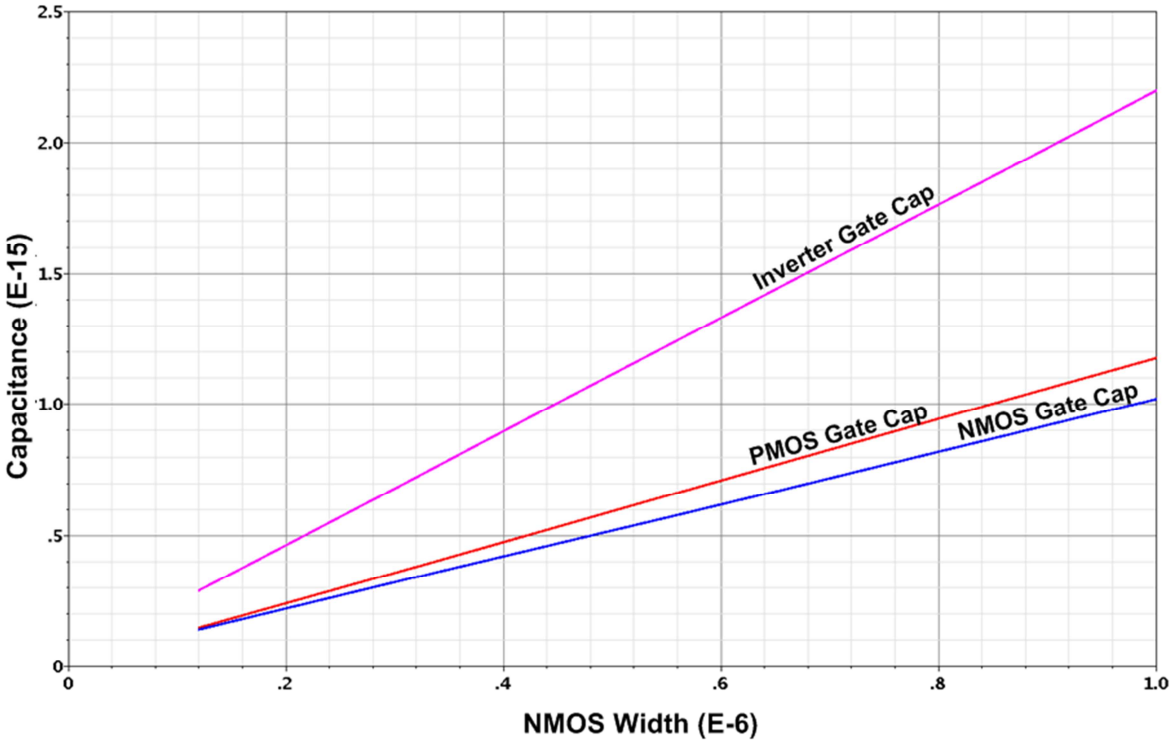


Figure 4.25 - Variation in inverter gate capacitance with increasing gate width (PMOS-width = 2*NMOS-width).

In Cadence environment, only lumped parasitic values are visible at each node during back-annotation. These back-annotated lumped parasitics are distributed into three components:

(a) Parasitic resistance R_{PAR}
(b) Parasitic capacitance C_{PAR}
(c) Parasitic coupling capacitance $C_{C\text{-}PAR}$

Such lumped modeling makes the task of parasitic modeling more difficult as it presents very little information for the designer. In order to proceed with modeling, few cases of inverter design were done manually for the test-bench shown in Figure 4.24. As parasitic capacitance influence the delay of inverter, such data provided a starting point for extrapolating the behavior of parasitic capacitance. The results of manual experiments are shown in Table 4.6. It can be seen that C_{PAR} at the output node remains almost same with increasing width while the corresponding $C_{C\text{-}PAR}$ increases with increasing size. Thus, the overall parasitic capacitance increases with increasing transistor widths.

Table 4.6 - Extracted parasitic lumped components for different gate widths at input-output nodes.

W_{pmos}	W_{nmos}	t_{PLH}	t_{PHL}	Input Node			Output Node		
(nm)	(nm)	(ps)	(ps)	$R_{PAR}(\Omega)$	C_{PAR}(fF)	$C_{C\text{-}PAR}$(fF)	$R_{PAR}(\Omega)$	C_{PAR}(fF)	$C_{C\text{-}PAR}$(fF)
320	150	27.3	29	227	0.08	0.17	69	0.075	0.12
450	200	28.4	28.1	307	0.1	0.16	66	0.09	0.13
900	400	25.9	26.8	330	0.08	0.32	27	0.15	0.25
1100	500	25	26	472	0.11	0.16	70.9	0.15	0.16
2000	1000	25	23.3	542	0.08	0.3	64	0.09	0.33

Figure 4.26 illustrates the lumped parasitics at the input and output node of a single inverter as given by the Cadence PEX tool. In each $C_{C\text{-}PAR}$ component at each node, there is coupling between the node and other connected nodes. To model coupling between the input and output node, we assume that half the coupling exists between input and output nodes whereas other half coupling is between the power rails and the input/output node. Thus, if $C_c = C_{c\text{-}in} + C_{c\text{-}out}$, a $\frac{C_c}{4}$ capacitance exists between the input and output node as shown in Figure 4.26(b). With such a feedback capacitance, the Miller effect occurs resulting in twice the swing at the input w.r.t output node and vice versa. If the input node switches from $0 \rightarrow 1$, the output node would switch from $1 \rightarrow 0$, thus, the change in output is twice that of input leading to a capacitance of $\frac{C_c}{2}$ at each node.

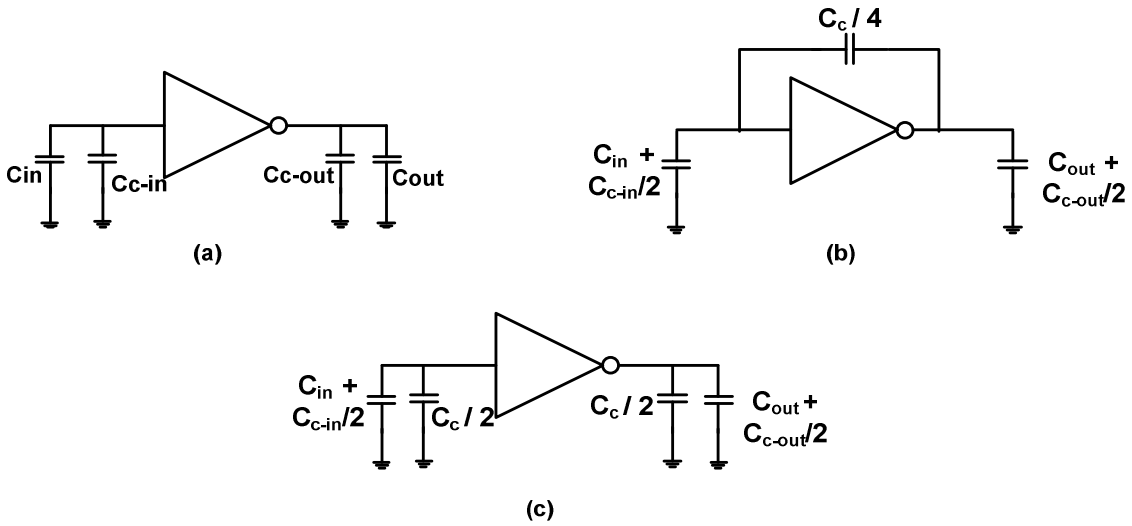


Figure 4.26 - Parasitic modelling: (a) lumped components provided by Cadence PEX and its back-annotation; (b) feedback capacitance to model coupling between input and output nodes; (c) Miller effect to derive the complete parasitic model.

Merging the parasitic model shown above in the testbench, the first inverter in the testbench can be shown to have the capacitances as shown in Figure 4.27 where

$$C_{par} = 2C_{in} + C_{out} + \frac{3}{2}C_c + C_{c-in} + \frac{1}{2}C_{c-out} \approx 2C_{in} + C_{out} + \frac{5}{2}C_c \tag{4.1}$$

exists at node X. Merging the model with the manual observed values, we can extrapolate the value of

$$\begin{aligned} C_{par} &= 0.3\text{f} + [3 \cdot (0.56\text{f}) - (c \cdot (400\text{n} - W_{nmos}) \cdot 10^{-9} \cdot 10^{-15})]; \\ W_{pmos} &= 2.2 W_{nmos} \end{aligned} \tag{4.2}$$

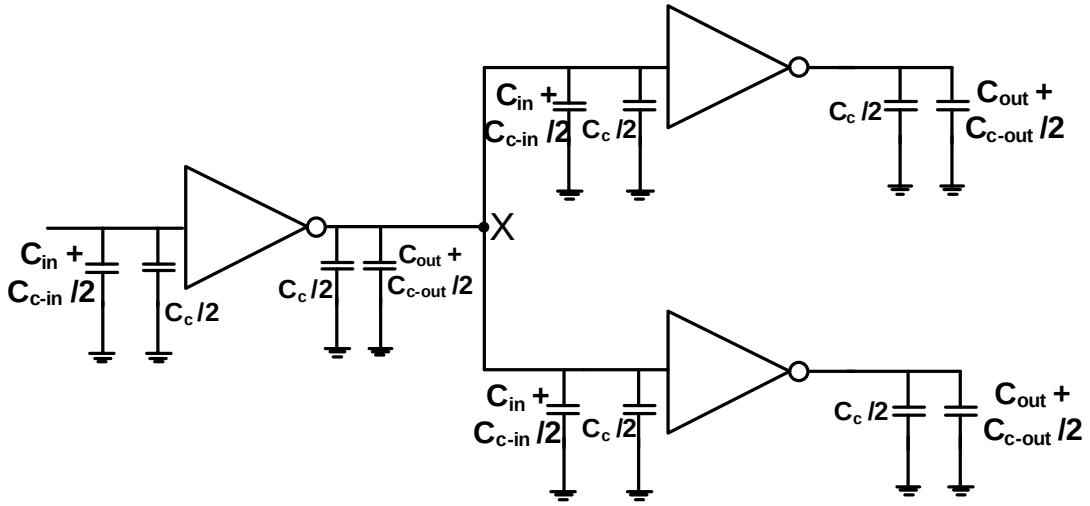


Figure 4.27 - Lumped parasitic components in a single stage of TDC.

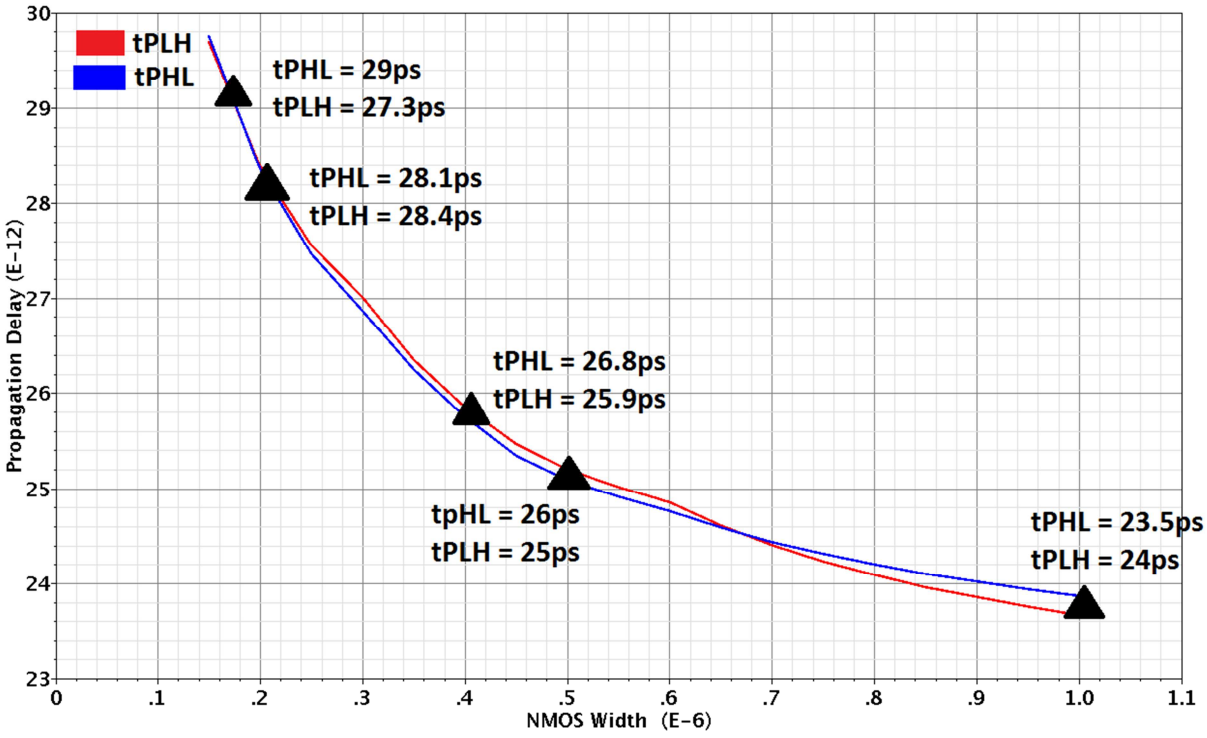


Figure 4.28 - The parasitic modelling of TDC single stage where the plot shows the variation in delay with increasing size.

The test-bench was simulated with an additional C_{PAR} to account for parasitics. The variation in propagation delay of inverter with increasing inverter size is shown in Figure 4.28. The hand-calculated values for different widths are denoted as well and it can be seen that the parasitic modeling done is quite close to the real case.

Post parasitic-modeling for high frequency ring oscillator, a detailed discussion is now made on the design of each of the pixel building blocks.

4.2.2 SPAD-TDC Interface Circuit

The interface circuitry has four tasks:

(a) sense the photon arrival by generating a step signal
(b) passively quench the avalanche current and recharge the SPAD back to Geiger mode through ballast resistance
(c) hold the signal at logic HIGH until the clock arrives, through SR latch
(d) blank the SPAD when required, using a PMOS transistor

Ballast resistance

The ballast resistance R_B is designed using an NMOS transistor. Since, the tunable nature of resistance allows control of dead time, an external control of the gate voltage of the NMOS transistor is provided. The resistance of MOS in linear region is given by:

$$R = \frac{1}{\frac{dI_{DS}}{dV_{DS}}} = \frac{1}{\frac{d[k \cdot \frac{W}{L} \cdot ((V_{GS}-V_T)V_{DS} - \frac{V_{DS}^2}{2}]}{dV_{DS}}} = [k \cdot \frac{W}{L} \cdot ((V_{GS}-V_T) - V_{DS})]^{-1} \quad (4.3)$$

Thus, longer channel in transistor or smaller V_{GS} leads to larger resistance. The simulation plot in Figure 4.29 shows the range of resistance possible, from 16 kΩ - 650 kΩ over different V_{DS} values. As a large resistance is desirable for faster quenching, an NMOS with

$$\frac{W}{L} = \frac{120n}{250n} = \frac{1}{2} \quad (4.4)$$

was chosen for quenching.

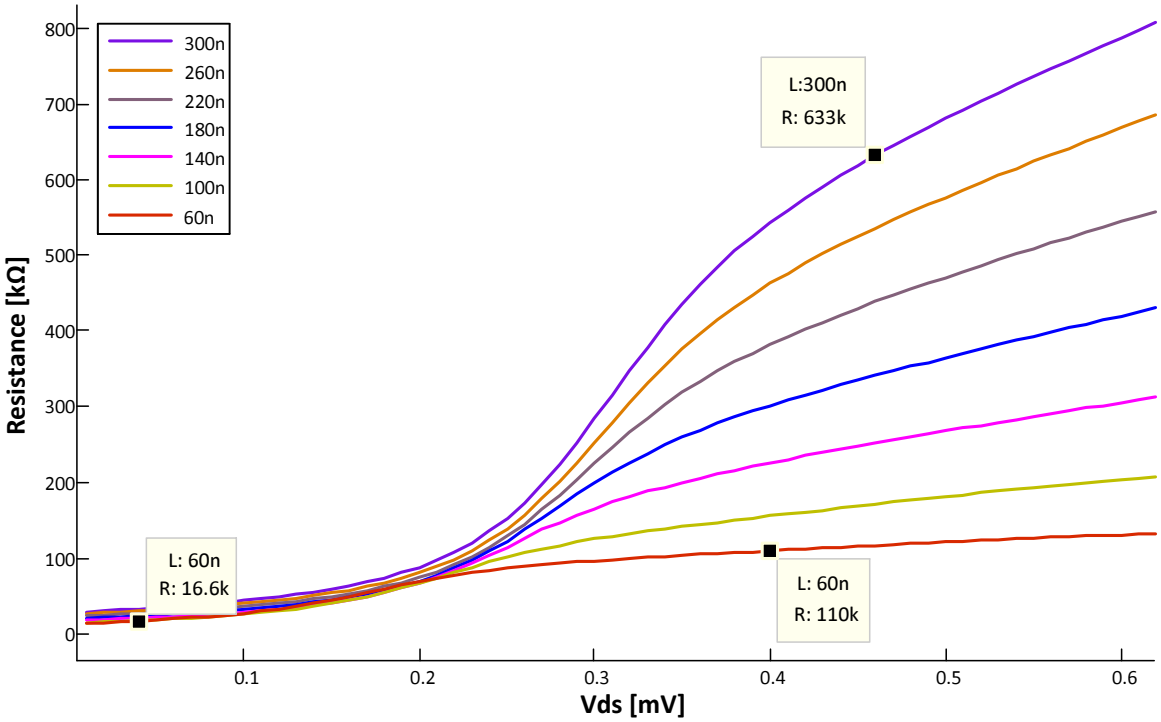


Figure 4.29 - Change in resistance with increasing length with V_{GS} = 0.7 V.

Such sizing of NMOS transistor allows larger tuning range. Longer channel reduces the threshold voltage of NMOS and thus, enhances the tuning range of V_{GS} is possible. The change in threshold variation V_T with increasing length is shown in Figure 4.30. At NMOS length = 250 nm, the threshold voltage is 0.45 V. The change in the resistance with increasing V_{GS} is shown in Figure 4.31.

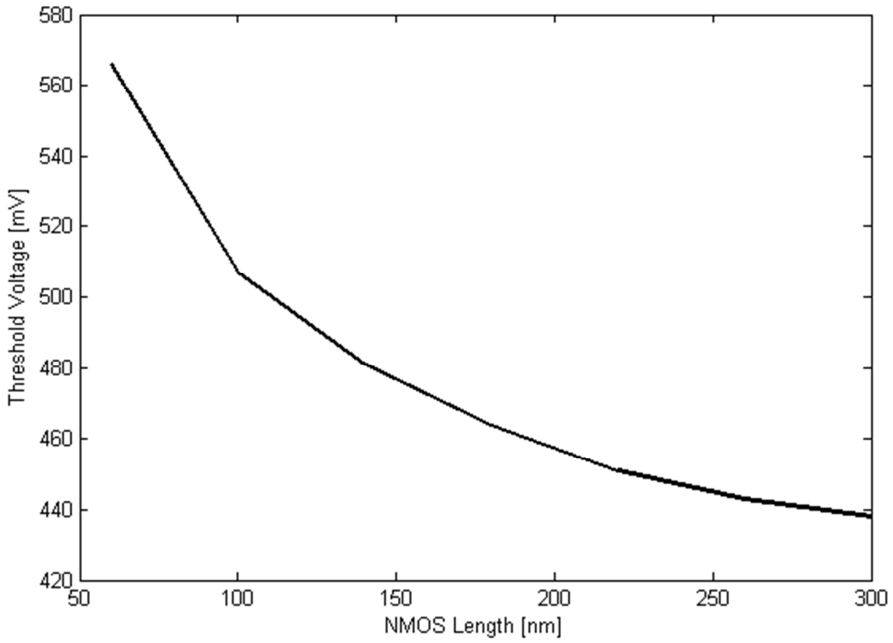


Figure 4.30 - Change in NMOS threshold voltage with increasing length.

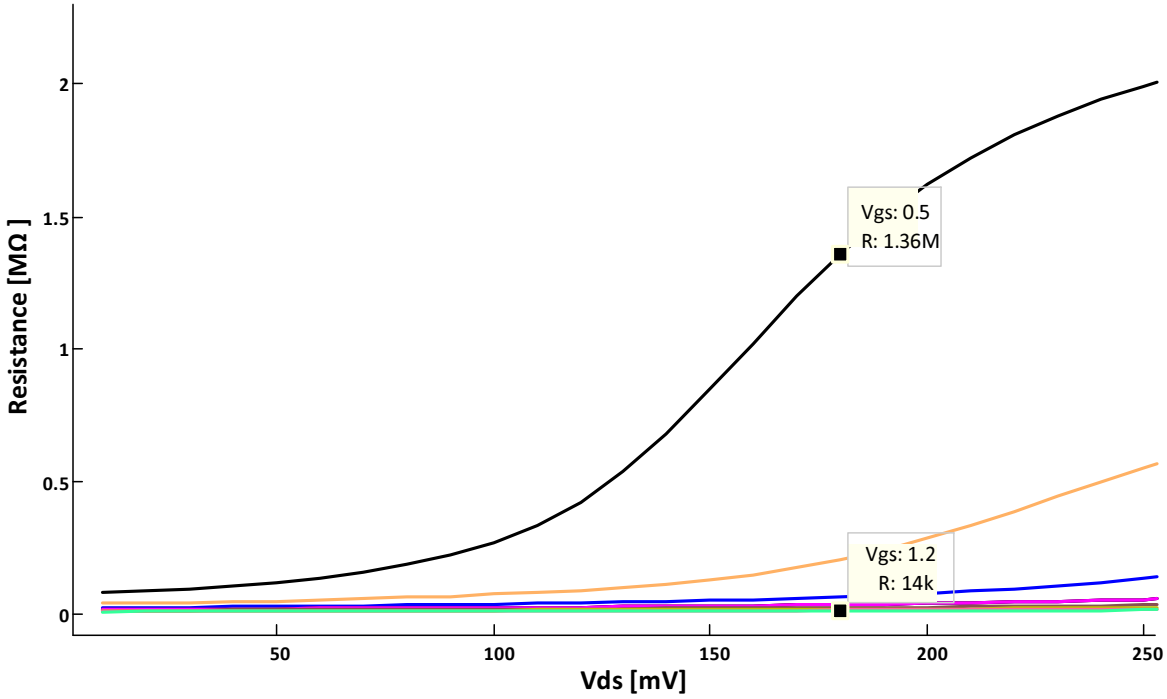


Figure 4.31 - Range of tunable resistance by varying V_{GS} with L = 250 nm.

4.2.3 Photon Sensing Interface

a) Comparator

The comparator, as discussed in Section 4.1.2, is built with minimum size transistors to reduce area and is chosen for its high speed along with its ability to sense very small signal amplitudes. Although the static current consumption is of the order of 50 µA - 70 µA, it is still within the tolerance limits for a 32x32 pixel array. As indicated in the previous section, the priority is laid on the correct functionality of the imager in this prototype. Although the comparator has sub-optimal performance in terms of power consumption, it offers to be one of the definite ways to integrate 65 nm SPAD with a TDC for an imager. For a 1000x1000 array, better SPAD device will allow a single inverter to be employed to sense the SPAD signal.

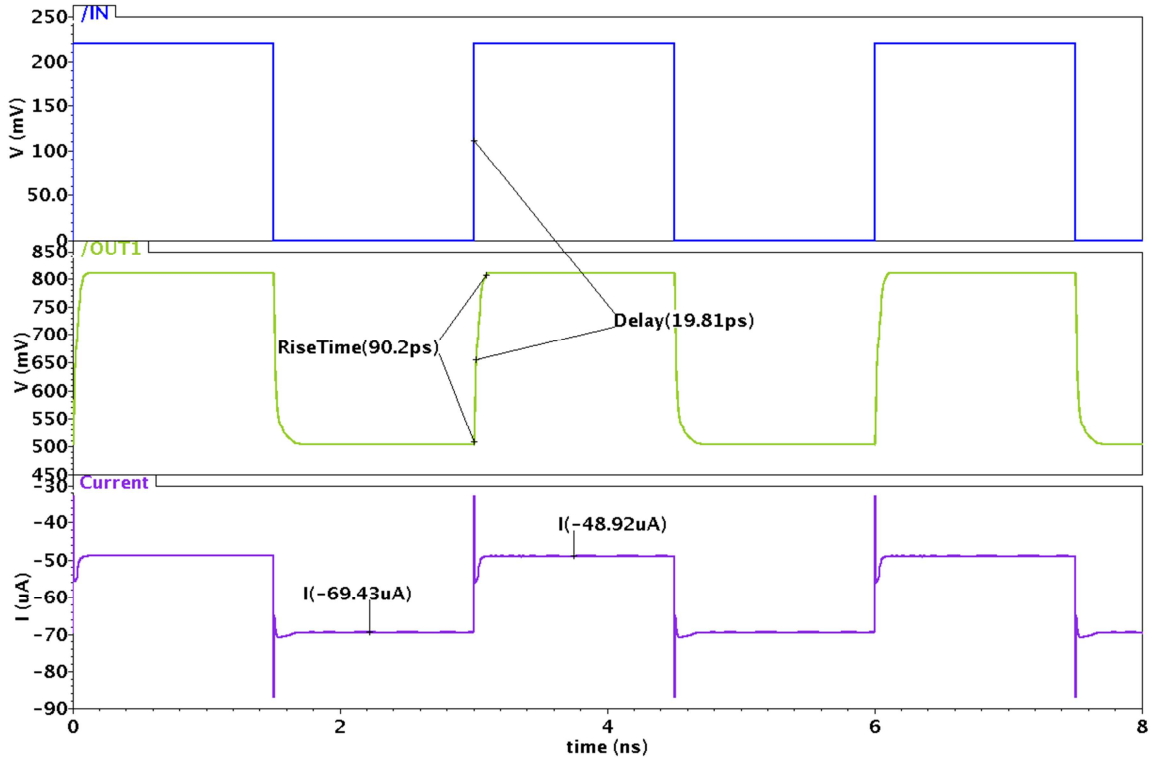


Figure 4.32 - Transient response of comparator (pre-layout).

4.2.4 Time to Digital Converter [46]

NAND gate sizing

The first delay tap of TDC is a NAND gate and is different from other inverter based delay taps. As the typical delay of NAND gate is greater than an inverter, the difference in delay brings a non-linearity in TDC resolution. In order to minimize this non-linearity error, it is important to size the NAND gate such that it has similar delay as that of an inverter.

To meet this requirement, the NAND gate was modified: the feedback input of the ring was placed closer to the output. The second input transistors were sized large so that the ring input experiences the same delay as that of an inverter, as shown in Figure 4.33.

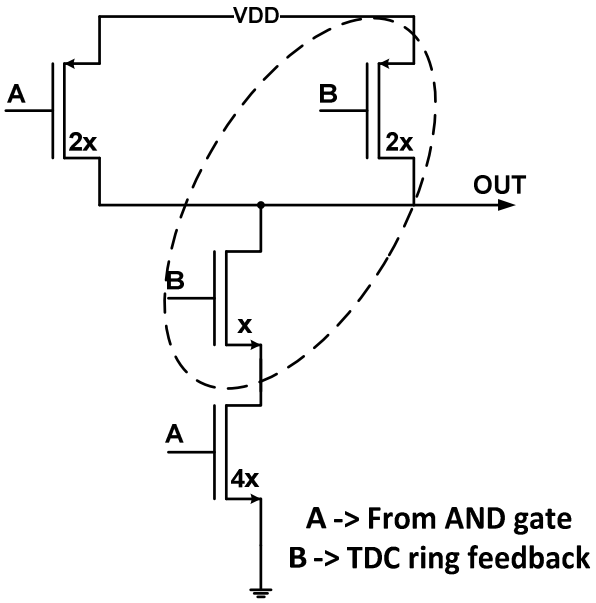


Figure 4.33 - Sizing of NAND for the first delay tap.

With such a design, the NAND gate delay is ~24 ps in post-layout simulations and is close to that of an inverter (~23 ps).

Inverter sizing

In ring oscillation, it is important that the inverter is symmetric in nature i.e. it has equal rise time and fall time along with equal propagation delay for both rising and falling input. Such a requirement is met with a PMOS:NMOS ratio of ~2.1 in 65 nm technology.

As every inverter has the same input and output load, each inverter was sized equally in the ring. The high frequency ring oscillation increases parasitic dominance and therefore, larger sized inverters are required to control the inverter delay. An optimal inverter design was constructed based on the knee point in Figure 4.28 to achieve a TDC resolution of 23 ps.

Latch sizing

The switched inverter at the input needs to be such that it is strong enough to drive the output $\overline{Q}$ without loading the delay line inverter. Thus, it is designed bigger than a minimum sized inverter. The inverter driving output Q is again sized bigger than a minimum sized inverter to be able to strongly drive the output Q. However, the feedback switched inverter is sized minimum to reduce contention issues.

One of the key design parameters in flip-flop design is its metastability. If both clock and D input arrive simultaneously, it can ideally take infinite amount of time to resolve. However, noise in the system eventually resolves the state. Metastability is defined as the resolution time taken to resolve signals within a specified 'metastability window'. For this latch, the metastability window is less than 0.1 ps for a resolution time of 200 ps. The metastability curves for this latch are shown in Figure 5.6. The metastability window is much smaller than inverter resolution and ensures no bubbles in the TDC code. It also makes the inverter delay the only limiting factor for TDC resolution.

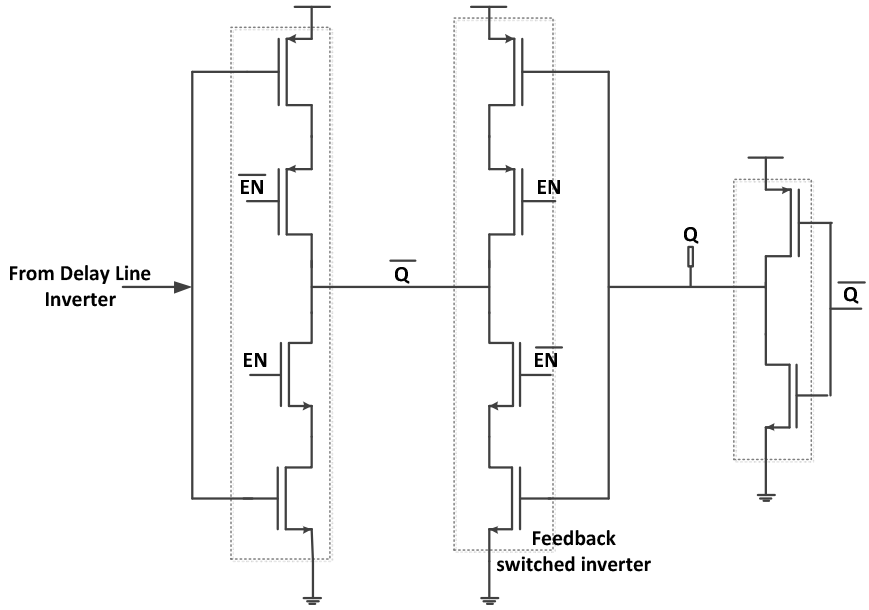


Figure 4.34 - Circuit diagram of switched inverter based latch.

4.2.5 Clock Retiming and Edge Aligner [46]

The clock is essential in the pixel for three purposes:

(a) To stop TDC ring oscillation
(b) To capture the delay line outputs in latches
(c) To reset the SPAD latch cell

In each of the three cases, the negative edge of clock is required. A *'retiming circuit'* is built to enforce clock gating in the pixel so that delay line latches are not overwritten when a new clock pulse arrives. Furthermore, an edge aligner is also needed to align the differential clock signal required by latches. If the skew between clock and $\overline{\text{clock}}$ is larger than 25 ps, then the delay line latches do not generate the correct output. It is also important in the latch that clock arrives earlier than $\overline{\text{clock}}$ for valid output. If two global clock distribution trees are designed, it is not possible to meet such requirements at every pixel considering the on-chip PVT variations.

In order to design this block, the load for clock and $\overline{\text{clock}}$ is calculated:

- Clock load:
  - in SPAD latch cell - 1 eq. inverter
  - in AND gate - 1 eq. inverter
  - in latches - 7 eq. inverters
- $\overline{\text{Clock}}$ load:
  - in latches - 7 eq. inverters

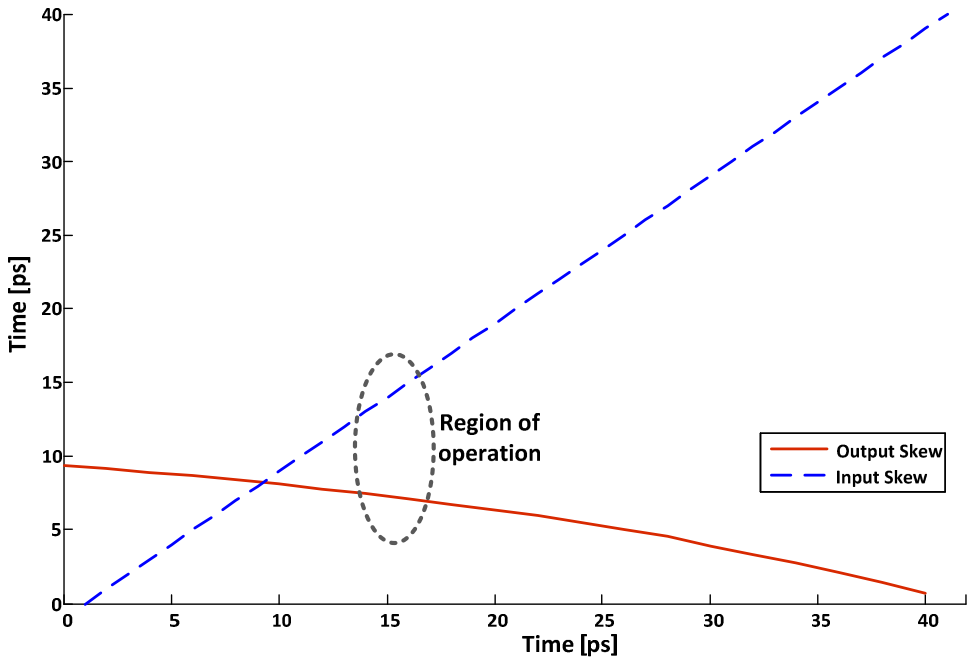


Figure 4.35 - Output skew with increasing input skew.

Thus, clock load is higher than $\overline{\text{clock}}$ but the parasitics seen by $\overline{\text{clock}}$ are greater due to routing in the layout. The clock conditioning circuit was built using standard library cells and is shown in Figure 4.22. In the employed D flip-flop, Q arrives earlier than $\overline{\text{Q}}$ and they are aligned by minimum sized cross-coupled inverters. Stronger inverters are added to drive the high output load. The performance of edge aligner is evaluated by measuring the variation in output skew with respect to input skew. The post-layout performance plot the edge aligner is shown in Figure 4.35. The plot shows that for an increasing input skew between the two differential signals, the output skew reduces with the use of edge aligner which is desirable for the correct operation of TDC.

4.3 Implementation of Pixel Level Blocks

The implementation of each block requires considerations with respect to speed, power consumption, delay and area and is discussed in detail in this section.

4.3.1 Layout of SPAD -TDC Interface Circuitry

The interface circuitry consists of -

a. ballast resistance formed by NMOS transistor
b. blanking control using PMOS transistor
c. comparator to detect a small amplitude SPAD signal
d. S-R latch

The layout of the interface circuitry was done adjacent to SPAD to avoid long routing, as shown in Figure 4.36.

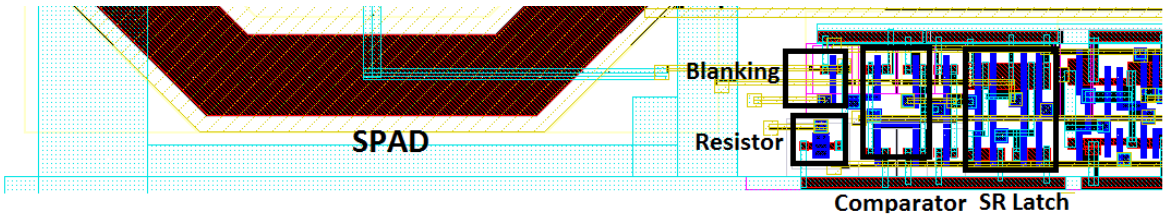


Figure 4.36 - SPAD-TDC interface circuitry.

4.3.2 Layout of TDC [46]

Since the delay line inverters define the TDC resolution, the layout of ring oscillator based TDC delay line is critical to achieve the final TDC performance. Moreover, in an imager where an array of pixels forms the structure, it is crucial that all pixels have same orientation to attain similar behaviour of each one. In order to realize this goal, the driving factor for TDC layout is the identical direction and placement of delay taps and that the routing between its elements is minimized.

A simple layout strategy for a delay line is shown in Figure 4.37. The layout is distributed in two levels - the top level is delay line inverters whereas the second level is their corresponding latches. Although it is the simplest placement of blocks, the routing is very long for the last inverter, thus, making it less matched to other delay line inverters. Therefore, it is not the best delay line placement and an alternative was required.

Looking into analog design techniques for nanometer CMOS [52], the matching of inverter delay was possible if they had same orientation, followed a common centroid configuration and had same overlaying metal layers to get similar inter-metal coupling effects.

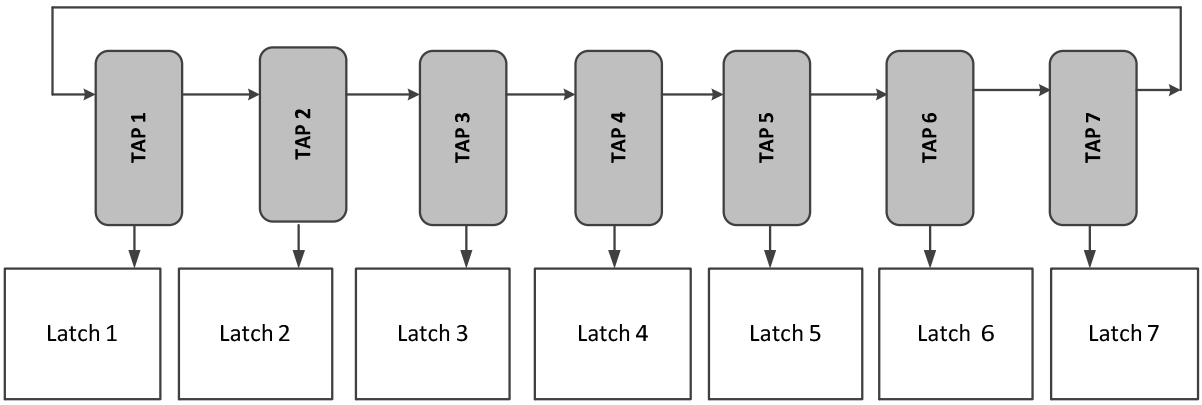


Figure 4.37 - Placement of delay line taps - Configuration I.

An inter-digited approach to placement of inverters, as shown in Figure 4.38, provided equal routing between delay taps. Such a placement preserves the orientation along with equalizing routing delay between inverters. Its division into three layers allows equal routing from inverter to latches as well as compact size which can be fitted easily in the pixel (adjacent to SPAD).

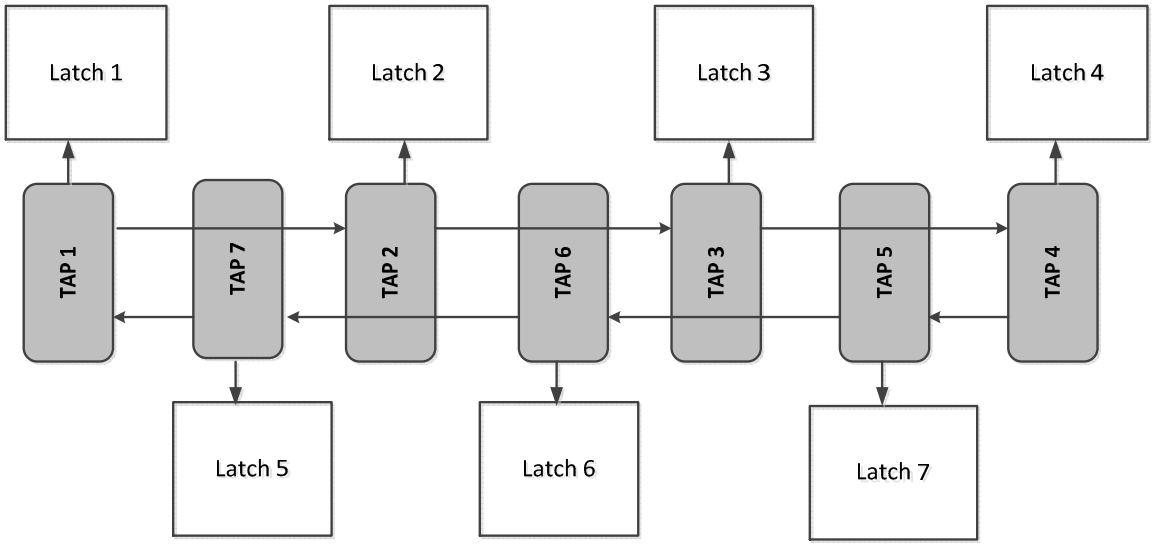


Figure 4.38 - Placement of delay line taps - Configuration II.

As the placement of TDC delay line is known, it is appropriate to visit the floor-plan of the pixel, as shown in Figure 4.39.

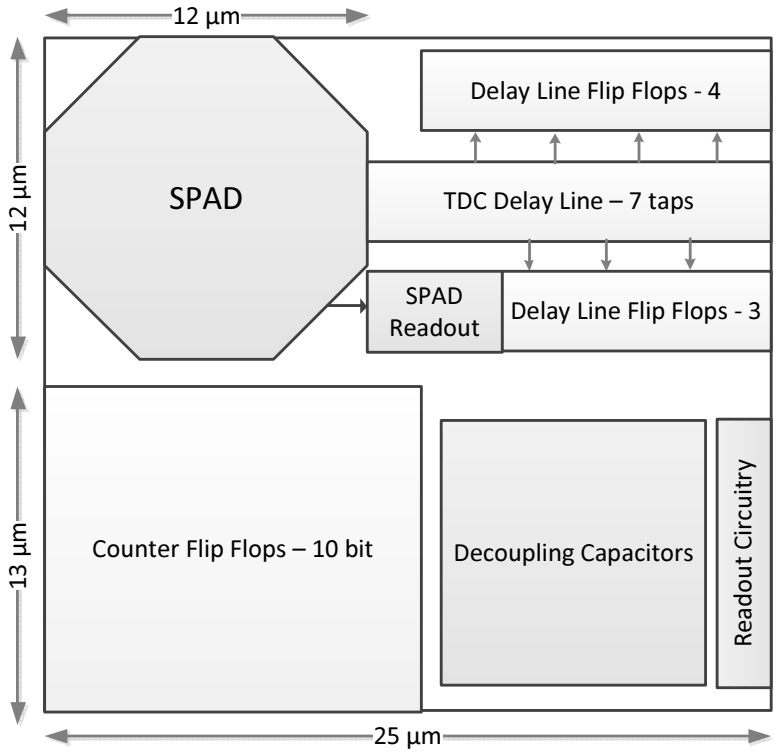


Figure 4.39 - Pixel Floor-plan.

To build the complete TDC, the basic blocks were first designed and a modular approach, similar to circuit design, was chosen. The three basic blocks are: inverter, switched inverter based latch and a counter flip-flip.

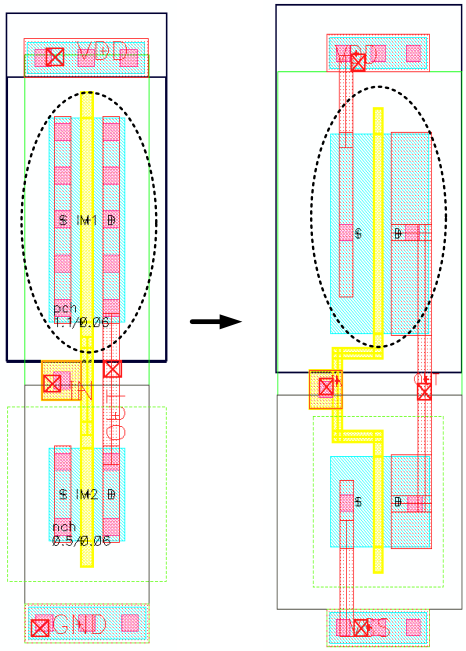


Figure 4.40 - Optimized inverter layout.

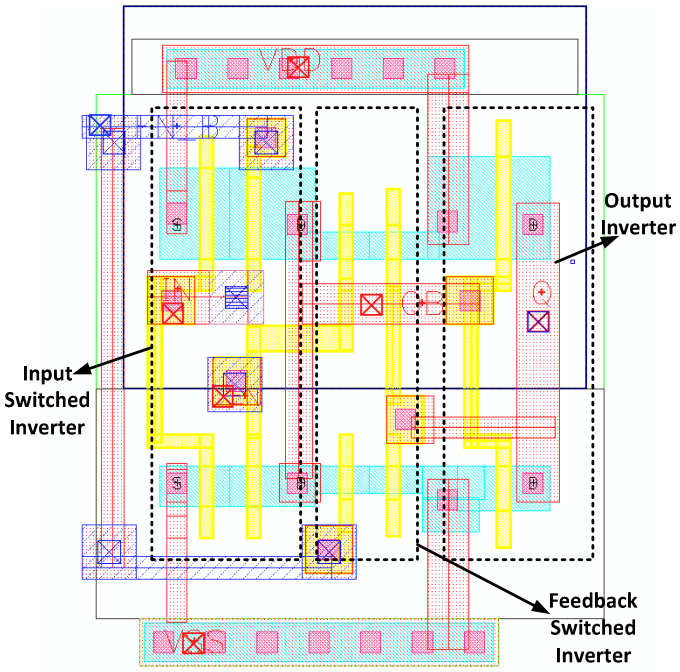


Figure 4.41 - Layout of switched inverter based latch.

The inverter layout was optimized for minimum delay by reducing coupling in two ways:

(i) increasing the spacing between input and output to decrease poly-M1 coupling,
(ii) reducing the number of drain-source contacts to reduce the contact-poly coupling.

The optimized inverter layout is Figure 4.40. The switched inverter based latch is also built in order to have minimum coupling effects and is compact in size (refer Figure 4.41). The final layout of TDC delay line is shown in Figure 4.44.

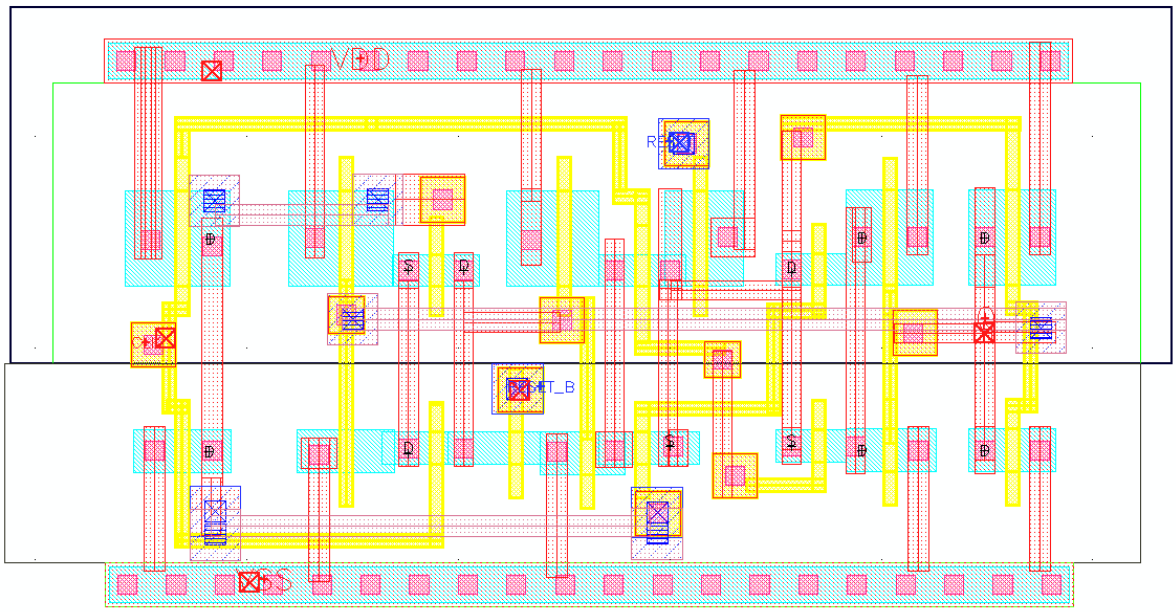


Figure 4.42 - Layout of a toggle flip-flop.

The 10-bit counter required to extend the dynamic range of TDC to 14 bits is built using T flip-flops. The placement of counter flip-flops is done across 5 levels in two rows and its ripple nature is evident in Figure 4.43. The counter is placed below the SPAD sensor.

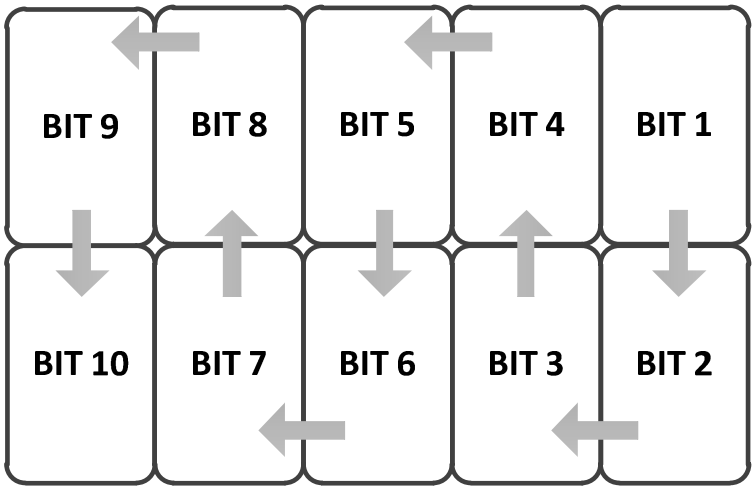


Figure 4.43 - Placement of 10 bit ripple counter.

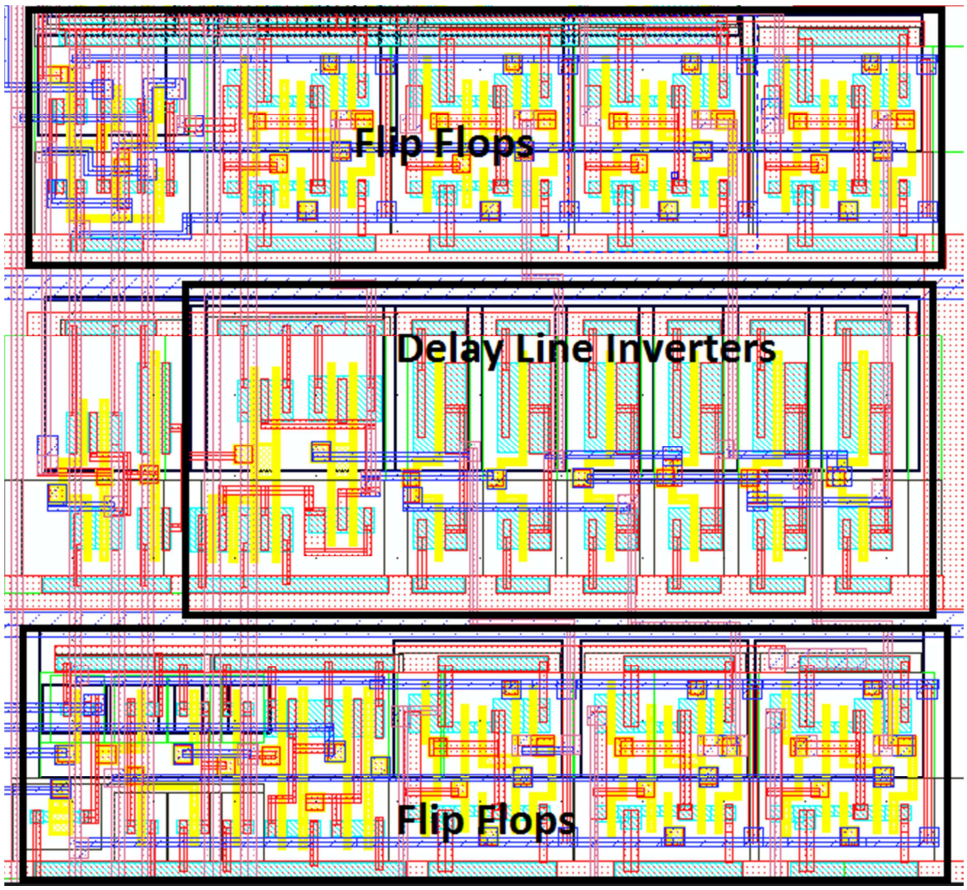


Figure 4.44 - TDC delay line layout.

4.3.3 Layout of Clock Retiming Circuit[46]

The clock edge retiming and edge aligner circuit was constructed with standard library gates, as precise timing could be achieved in minimum area. The block has two outputs - clock and $\overline{\text{clock}}$. These signals are internally routed to delay line latches, SPAD-TDC interface memory cell and delay line controller AND gate. The layout of this block is shown in Figure 4.45.

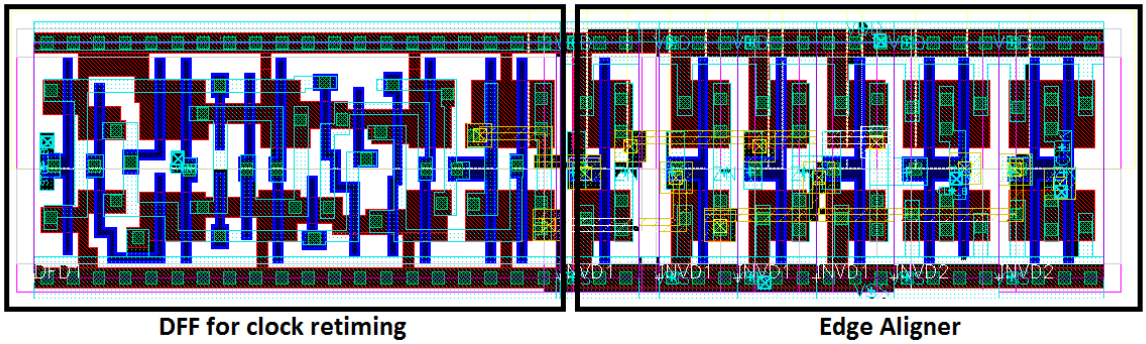


Figure 4.45 - Placement and layout of edge aligner.

4.3.4 Decoupling Capacitors

In order to reduce noise caused by power rails in an array of pixels, a decoupling capacitor plays an important role as pixel's local energy storage. This energy responds quickly to changing current demands and avoids undesirable circuit performance.

The simplest capacitors available in 65 nm technology are the MOS capacitors. Typical density of MOS capacitors is 14 $fF/\mu m^2$ in 65 nm technology.

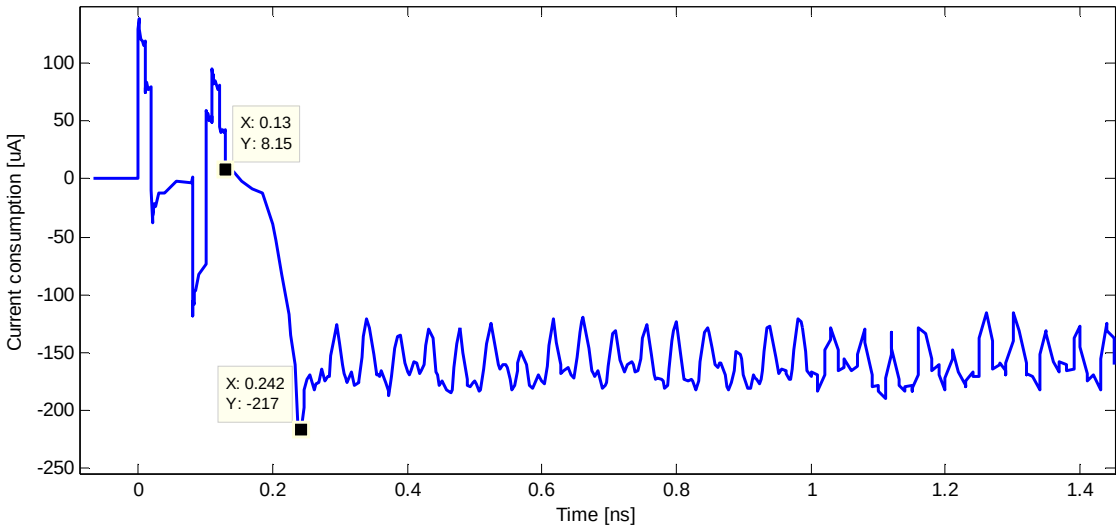


Figure 4.46 - Current consumption at pixel level for delay line when it begins ring oscillation

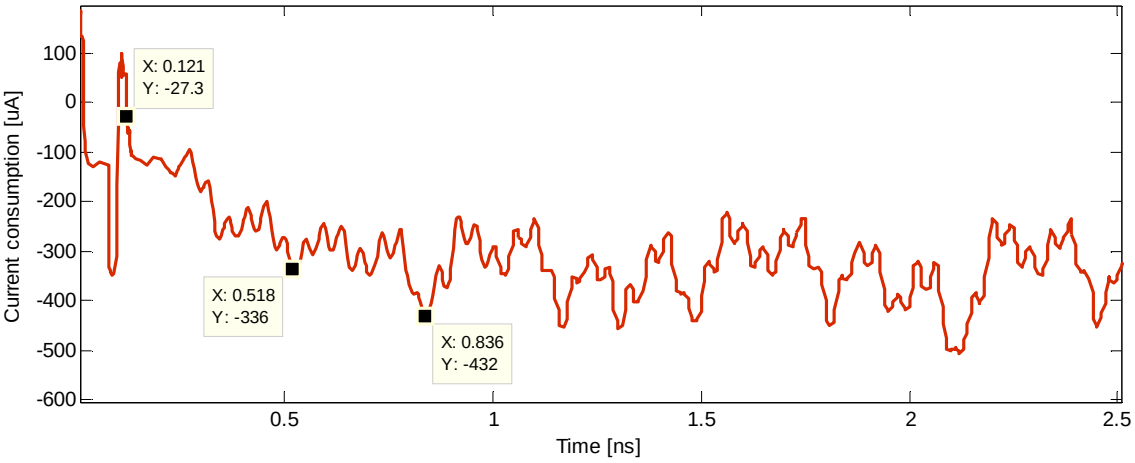


Figure 4.47 - Current consumption at pixel level for whole circuit (apart from delay line).

The following calculations lead to the sizing of decoupling capacitors for the two power lines.

For delay line VDD,

$$\begin{aligned} &\text{Height of current spike} = 230\ \mu A \\ &\text{Width of current spike} = 100\ ps \\ &\text{Charge } Q = I \cdot t = 230 \cdot 10^{-6} \cdot 10^{2} \cdot 10^{-12} \sim 25\ fC \end{aligned} \tag{4.5}$$

$$\text{Required MOS cap } C = \frac{Q}{V_{dd}} = \frac{25f}{1.2} \sim 21 \text{ fF}$$

$$\text{Required MOS cap area} \sim \frac{21fF}{14 \text{ fF}/\mu m^2} \sim 2 \ \mu m^2$$

For VDD common to other circuit in pixel,

$$\begin{aligned} &\text{Height of current spike} = 400 \ \mu A \\ &\text{Width of current spike} = 500 \text{ ps} \\ &\text{Charge } Q = I \cdot t = 400 \cdot 10^{-6} \cdot 500 \cdot 10^{-12} \sim 200 \text{ fC} \\ &\text{Required MOS Cap } C = \frac{Q}{V_{dd}} = \frac{200f}{1.2} \sim 167 \text{ fF} \\ &\text{Required MOS Cap area} = \frac{167 \text{ fF}}{14 \text{ fF}/\mu m^2} \sim 12 \ \mu m^2 \end{aligned} \tag{4.6}$$

Although the area required is 2 μm^2 and 12 μm^2 for VDD_{DL} and VDD_{common}, careful routing and placement of blocks allowed bigger decoupling capacitances - 22 μm^2 and 27 μm^2 respectively. The decoupling capacitors were placed on the diagonal right of the SPAD as shown in Figure 4.51.

4.3.5 Full Pixel Layout Assembly

Each pixel consists of a SPAD sensor, its front-end circuitry, a TDC, decoupling capacitors for each of the power rails and access transistors for readout.

In layout, it is important to consider the placing of metal lines as a sensitive SPAD photodetector is present in the pixel. A 45° incident angle of light on SPAD is desirable. To achieve this, the metal lines are placed such that their height from substrate is same as the distance from SPAD. Therefore, metal 1 is allowed to be closest to SPAD than other metals. The metal stack of 65 nm technology (upto metal 5) is shown in Figure 4.48 and throughout the pixel array layout, such spacing has been maintained.

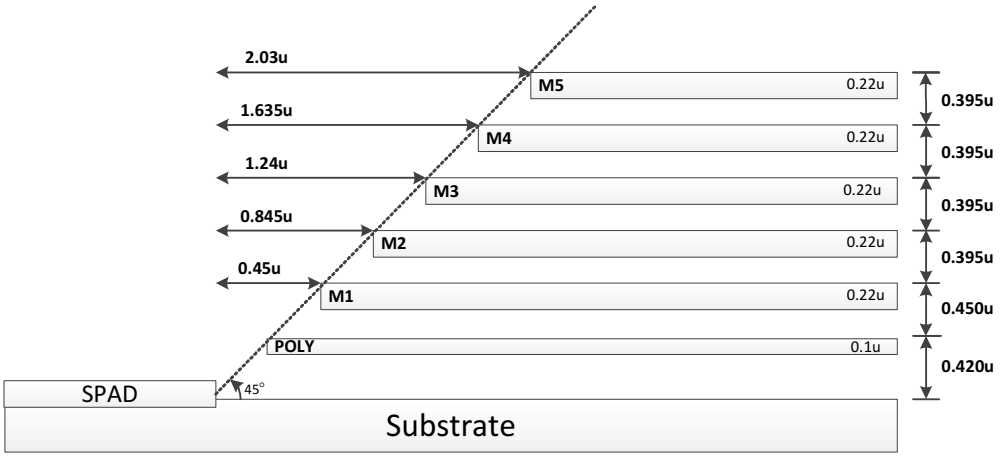


Figure 4.48 - Metal Stack (upto metal 5) for 65 nm technology.

Apart from the layout of each block, routing of control inputs and outputs is critical as well. In the pixel layout, the control signals are laid out in vertical direction in metal 3 whereas the output lines are laid out in horizontal direction in metal 4. This optimizes the area as fewer contacts are required.

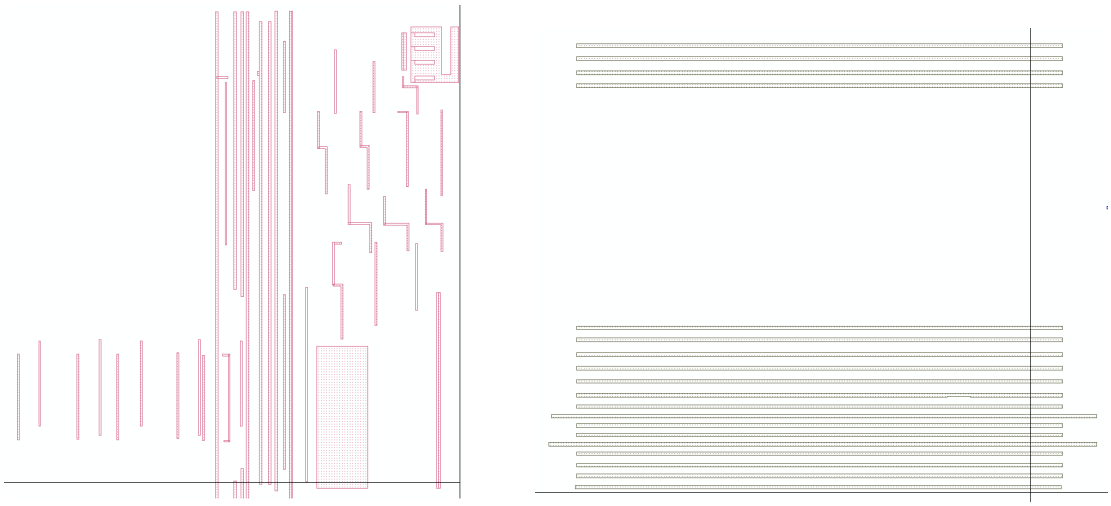


Figure 4.49 - Routing across the pixel: metal 3 for control inputs (left); metal 4 for outputs (right).

There are two power rails - one for TDC delay line and other for the remaining circuit and both VDD rail runs in metal 5. Such a division reduces IR drop in TDC delay line and maintains the best possible resolution in layout. Another power rail required is the operating bias of SPADs, known as V_{OP}. This has been routed in metal 2 and runs across the pixel.

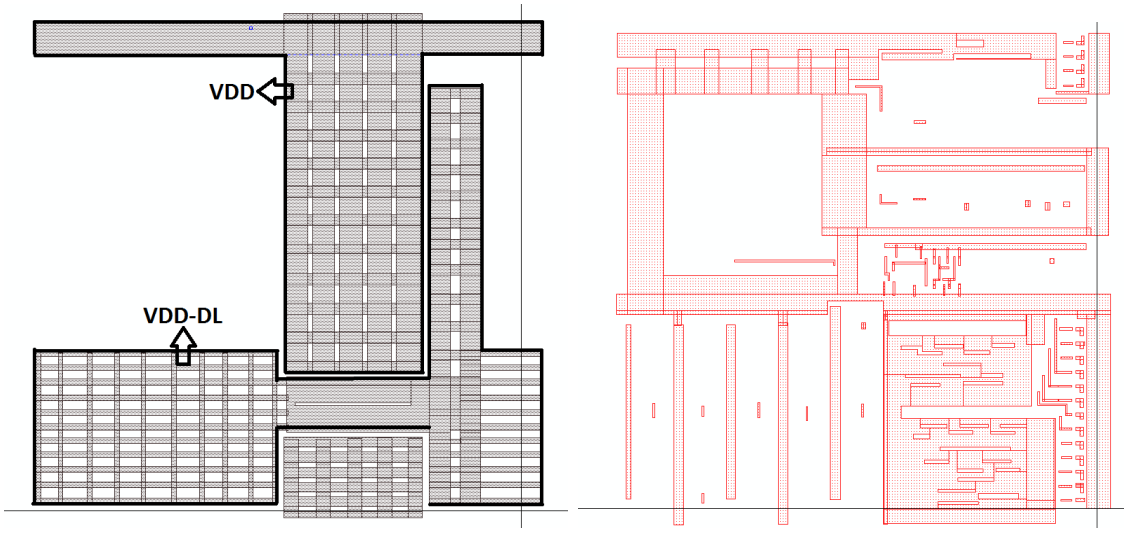


Figure 4.50 - Power lines across pixel: metal 5 for power lines (left); and metal 1 for GND (right).

Furthermore, no metal is overlaid upon TDC delay line to avoid coupling effects. The GND plane is laid in metal 1 as it can maintain the closest distance to SPAD. The final pixel layout with all blocks and routing is shown in Figure 4.51.

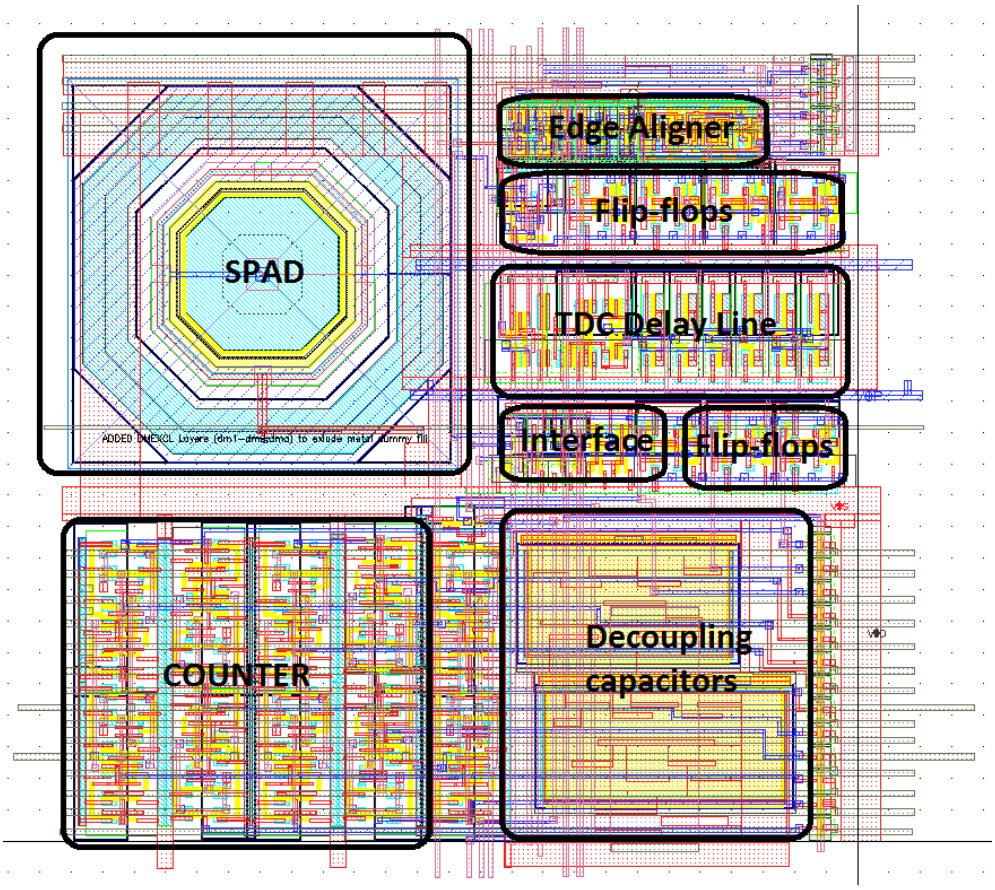


Figure 4.51 - Pixel layout.

4.4 System Integration

4.4.1 Clock Distribution

With the clock retiming and edge alignment in each pixel, the task of clock distribution reduced to single clock signal distribution. The simplest way to design clock distribution network was through RC modeling.

The architecture for clock distribution network is very simple. Two columns of clock buffers feed the 32x32 array, refer Figure 3.12. These columns are placed on either side of imager such that each clock buffer of a column drives 16 pixels. With such a symmetric configuration, the 16th pixel in every row has the worst clock skew. Such skew is deterministic in nature and can be calibrated.

4.4.1.1 Delay Modeling

With respect to the clock tree design described previously, a careful clock buffer is crucial. The clock buffer must consider the interconnect resistance, its parasitic capacitance along with pixel's clock load and parasitics. In order to estimate the delay, RC delay modeling is performed. The common methods to calculate the delay are:

Lumped RC model - Total wire resistance is lumped into a single R and total capacitance into a single C. The delay using lumped RC model can be calculated as:

$$\tau_d = 0.69 \cdot R \cdot C_{load} \tag{4.7}$$

(4.7) holds for short wires and becomes inaccurate as the interconnect length increases.

Elmore delay model - It is a simple approximation to the delay through an RC network and is more accurate as delay through the wire is modeled. It is often useful in interconnect optimization as it uses the delay of individual wire segments to estimate the final delay between two nodes. The Elmore delay between node 1 and node N is calculated as:

$$\tau_d = \sum_{k=1}^{N} kR \cdot C \tag{4.8}$$

where k refers to a wire segment.

4.4.1.2 Circuit Resistance and Capacitance

The wire resistance R in each pixel can be known by calculating the length and width of clock signal wire. From the pixel layout, the clock signal wire has the following dimensions:

$$L = 25\ \mu m; W = 150\ nm;$$

$$\#squares = \frac{L}{W} = \frac{25\ \mu m}{150\ nm} \sim 170$$

$$R_{pixel} = R_S \cdot W = 0.14 \cdot 170 = 25\Omega$$

The capacitance C in each pixel can be calculated as:

$$C_{pixel} = C_{gate} + C_{wire} + C_{parasitic} = C_{clk\text{-}DFF} + C_{extracted} = 0.75\ fF + 7\ fF$$

$$C_{pixel} = 8\ fF$$

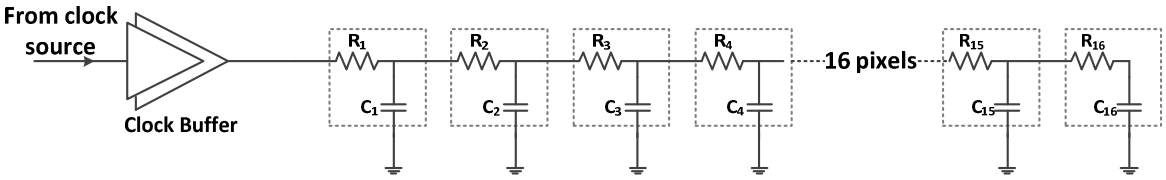


Figure 4.52 - Elmore delay based RC modelling.

Clock Skew Estimation

Applying lumped RC modeling, the clock skew between the first and 16th pixel is,

$$\tau_d = 0.69 \cdot (16 \cdot R_{pixel})(16 \cdot C_{pixel}) = (0.69) \cdot 25 \cdot 16 \cdot 25 \cdot 8f$$

$$\text{Clock skew } \tau_d = 35\ ps$$

Applying Elmore delay model to calculate the clock skew between pixel 2 and pixel 16 (N=16),

$$\tau_d = \sum_{k=1}^{N} k \cdot R_{pixel} \cdot C_{pixel} = R_{pixel} \cdot C_{pixel} \cdot (N+1) \sum_{k=1}^{N} k$$

$$\tau_d = R_{pixel} \cdot C_{pixel} \cdot [\frac{N \cdot (N+1)}{2}]$$

$$\tau_d = 25 \cdot (8f) \cdot (\frac{16 \cdot 17}{2})$$

$$\text{Clock skew } \tau_d = 27.2 \text{ ps}$$

The skew presented by each of the models is independent of the clock source and depends only on the number of wire segments i.e. the number of intermediate RC nodes between the two nodes.

The simulated clock skew in the RC modeled circuit of Figure 4.52 is ~30 ps, as shown in Figure 4.53 and is closer to Elmore delay model than the lumped RC model calculations.

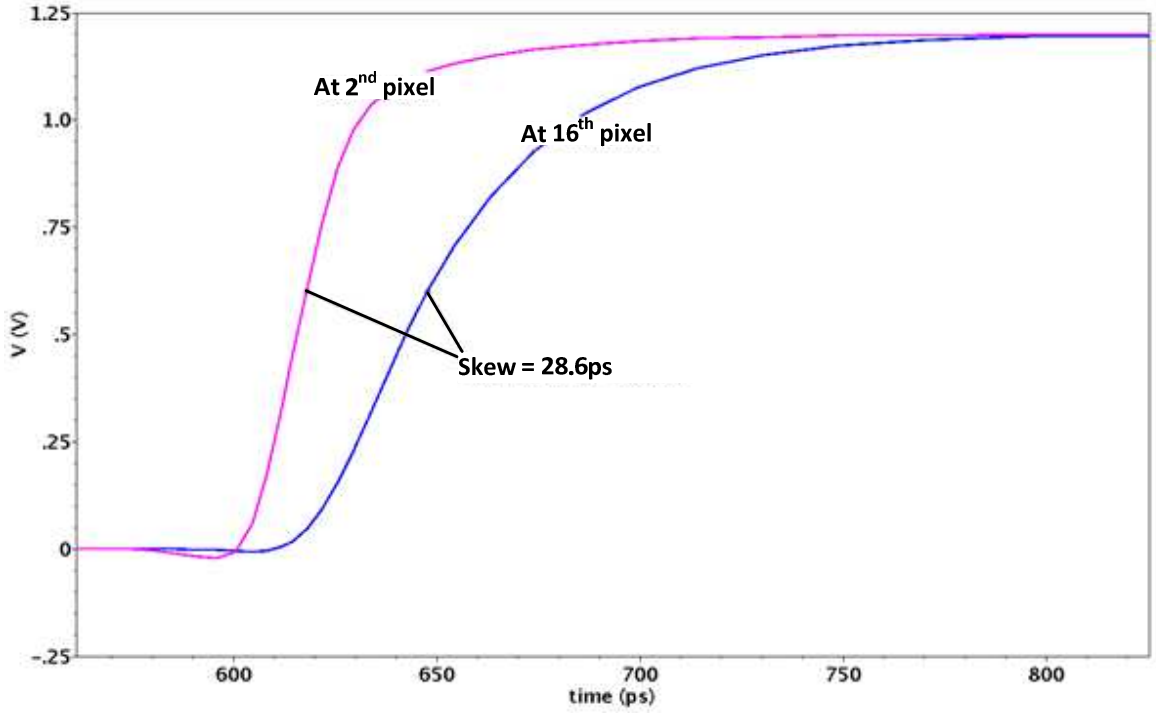


Figure 4.53 - Clock skew between the 2nd and the 16th pixel of a column.

4.4.1.3 Clock Buffer Design

To design a clock buffer capable to drive 16 pixels, it is necessary to know the load that each pixel presents. From previous calculations, C_{PIXEL} was calculated as 8 fF and therefore, the lumped load for a column of 16 pixels would be:

$$C_{load} = 16 \cdot 8 \text{ fF} = 128 \text{ fF}$$

The minimum sized inverter has an input gate capacitance

$$C_{in} = 0.5 \text{ fF}$$

The electrical effort for minimum sized inverter is given by:

$$\text{Electrical effort H} = \frac{128\text{f}}{0.5\text{f}} = 256$$

For minimum delay, a *'fan-out of four'* (Fo4) is the typical buffer sizing standard and it implies a 4-stage buffer proves to be good to achieve desirable performance with the following input capacitances:

$$C_{in} \rightarrow 4C_{in} \rightarrow 16C_{in} \rightarrow 64C_{in} \rightarrow 256C_{in} (= C_{load}) \tag{4.9}$$

As seen in Figure 4.55, the delay between the clock source and the clock arriving at the first pixel is ~102 ps. This delay is deterministic in nature and therefore, can be calibrated.

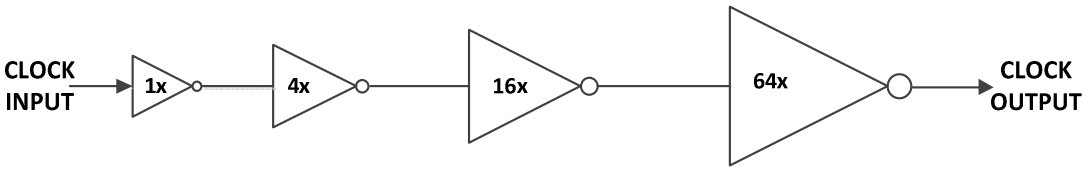


Figure 4.54 - Logical effort based clock buffer sizing.

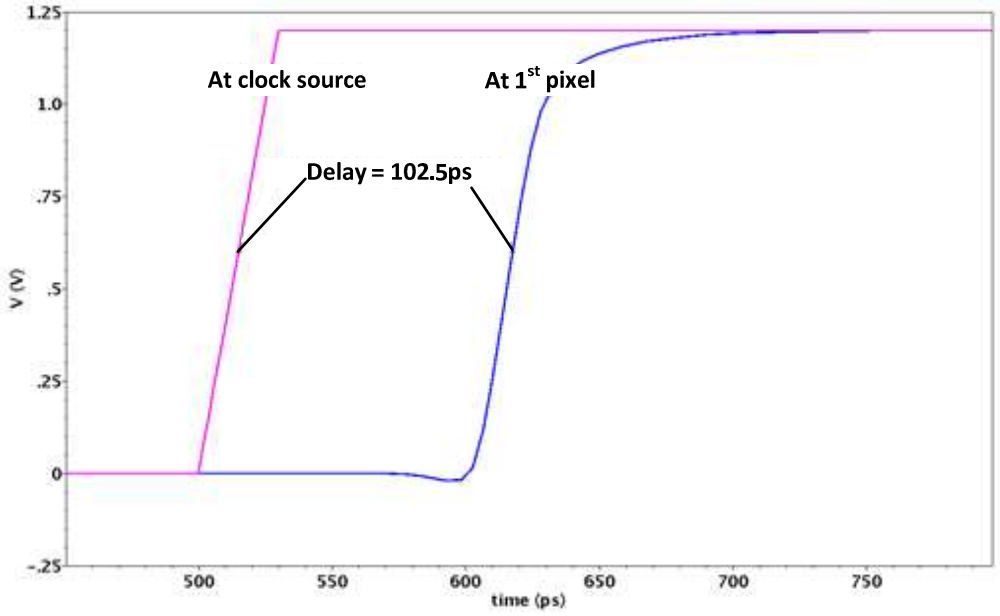


Figure 4.55 - Minimum clock delay between clock and the first pixel - with clock buffer.

4.4.2 System Readout Design

The readout has the following blocks:

4.4.2.1 Access Transistors

In order to load the TDC output data in every pixel on the column bus, access transistors are required. Minimum sized NMOS transistors were used as access transistors as they

provide the smallest possible area. Even though a V_T drop exists across an NMOS transistor, the signal is good enough to be identified as logic HIGH by the serializer flip-flop.

4.4.2.2 Row Decoder

Different decoder designs exist in the literature depending upon the logic family. A static logic row decoder was chosen for its fast design time and ease of implementation. Although dynamic logic based decoders are smaller as well as faster, they pose issues such as leakage and their dynamic nature adds uncertainty to the performance.

Since a 5 bit row decoder was needed for a 32x32 array of pixels, a pre-decoding based static implementation of row decoder was carried out to reduce area - by minimizing the number of common inputs. NOR gates were employed for pre-decoding and 3 input NAND gates performed the final decoding.

With such architecture, one word-line (WL) is always logic HIGH and therefore, a specific row is always selected. This means more power consumption as well as more loading for the TDC. In order to control the operation of row decoder, a column of latches was added to the row decoder. This addition allowed row selection only during readout (when EN is logic HIGH) and at other times, all WLs would be deactivated.

Implementation of 5 bit row decoder

Standard library gates were used for the implementation. Each of the NOR gates feed 4 NAND gates, each having a pin capacitance of 0.7 fF. The NOR gate was sized twice the minimum size to achieve a propagation delay of ~100 ps. The 3 input NAND gates were minimum sized as they have small load capacitance of 0.8 fF (D input of a latch). The latch output feeds 32 pixels, each having 17 access transistors. The pixel capacitance presented to word-line is ~10 fF and thus, each latch has a load of ~350 fF. It takes ~2 ns for the word-line to switch on all the 32 pixels in a row.

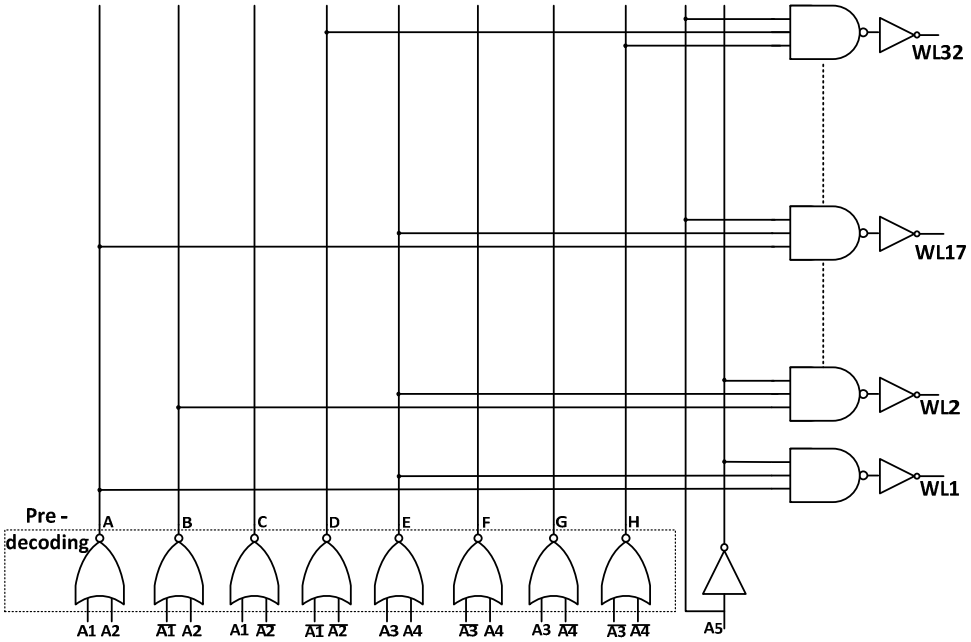


Figure 4.56 - A 5-bit static row decoder.

4.4.2.3 Serializer

A serializer takes parallel set of data and converts it into a serial bitstream. This is commonly done by Scan flip-flops (see Figure 4.57) which has an 'Enable' signal to control the parallel feed of inputs and serially shift the data. Since 17 bits form the output of each pixel, 17 Scan flip-flops were required to build the serializer.

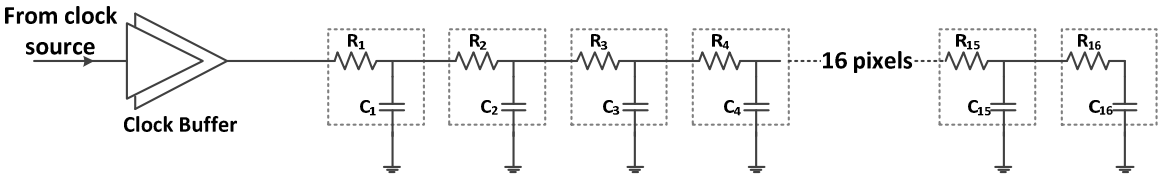


Figure 4.57 - Serializer using scan flip-flops.

4.4.3 Building The Imager Array

With clock distribution, readout and pixel in place, the system was built by integrating all blocks. A modular design approach was sought to build the full system. Emphasis was laid to make a symmetric layout. The floor-plan for the full system is shown in Figure 4.58.

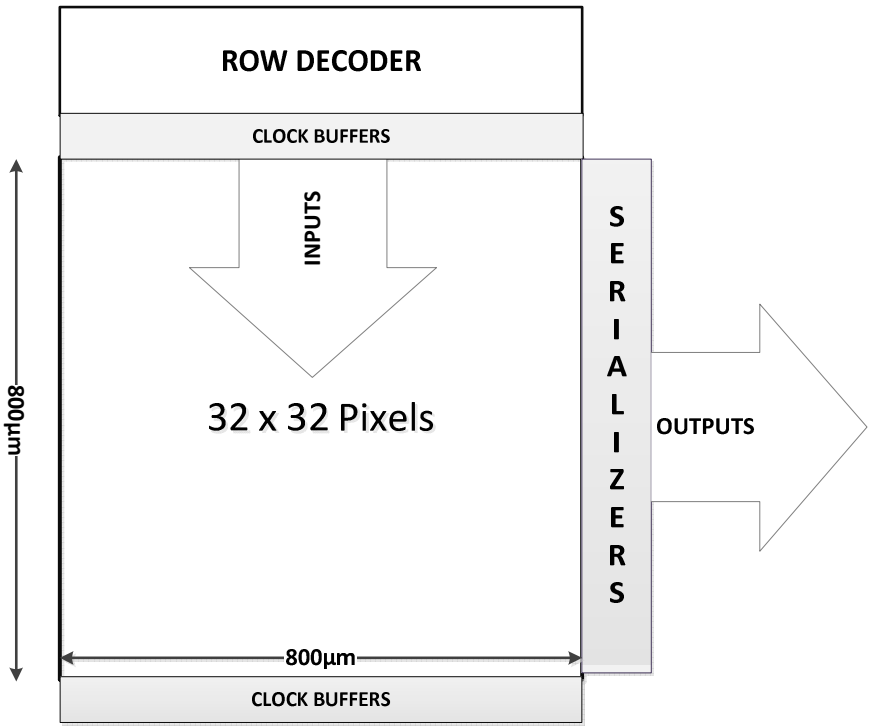


Figure 4.58 - System floor-plan.

At first, a single column was constructed with 32 pixels and then, 32 such column instantiations built the 32x32 pixel array as shown in Figure 4.59. Five instantiations of row decoder were used to avoid long metal routing. All the row decoders are placed on top and are horizontally aligned, thereby occupying no extra vertical area. It also ensures that the available area is not wasted and best put to use. 32 clock buffers are divided into two rows of 16 and is placed on top - bottom of array; each buffer feeds 16 pixels. The array of serializers is placed on the right side of the array.

Power rails run across the array. VDD bus runs vertically in metal 5 while GND is in metal 1 all across the imager (refer Figure 5.20) to limit the IR drop.

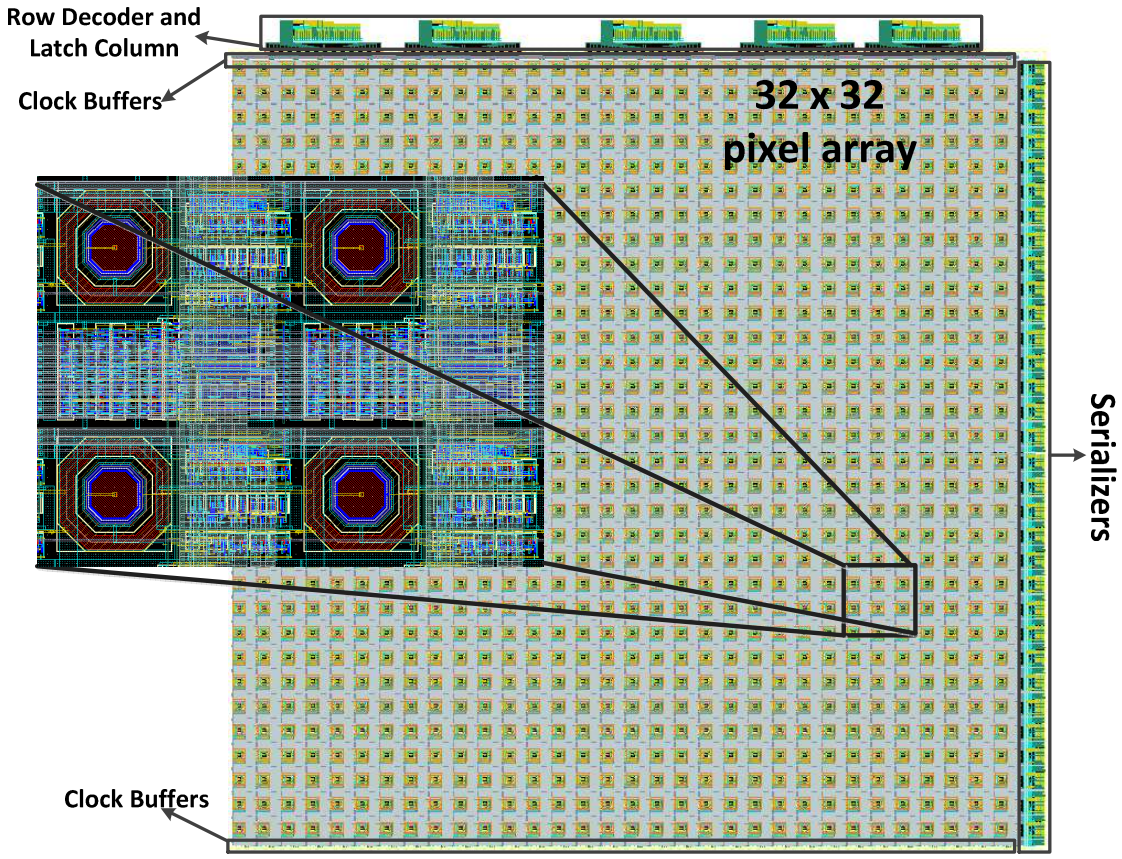


Figure 4.59 - Top level layout (the inset shows the arrangement of pixels).

4.5 Summary

In this chapter, a detailed discussion was made on the circuit design of pixel array and its implementation in 65 nm CMOS technology. We first discussed the pixel level building blocks and then, the overall system design. The next chapter presents the post-layout simulation results from the implemented layout.

5. Post-Layout Simulation Results

In the previous chapters, a detailed discussion of the design and implementation of a 3D real-time imager was made. In this chapter, the obtained post-layout results are presented and evaluated with respect to system requirements.

This chapter is organized as follows: Section 5.1 presents the performance of pixel level building blocks. Section 5.2 shows the system level performance. Section 5.3 summarizes the post-layout simulation result and Section 5.4 states a few recommendations for the future work.

5.1 Performance at Pixel Level

The pixel consists of SPAD photodetector to detect a photon, a TDC for time interval measurement and a SPAD-TDC interface circuitry. Their performance is analyzed in the following sections.

5.1.1 65 nm SPAD Performance

A 65 nm SPAD signal from the measurement setup [48] is shown in Figure 5.1. The SPAD signal has increasing amplitude with increasing excess bias voltage. The signal amplitude is very small due to high dark count rate (greater than 100 kHz for V_E of 0.3V) in the measurements. The small amplitude of SPAD signal makes the interface circuitry critical to the operation of imager.

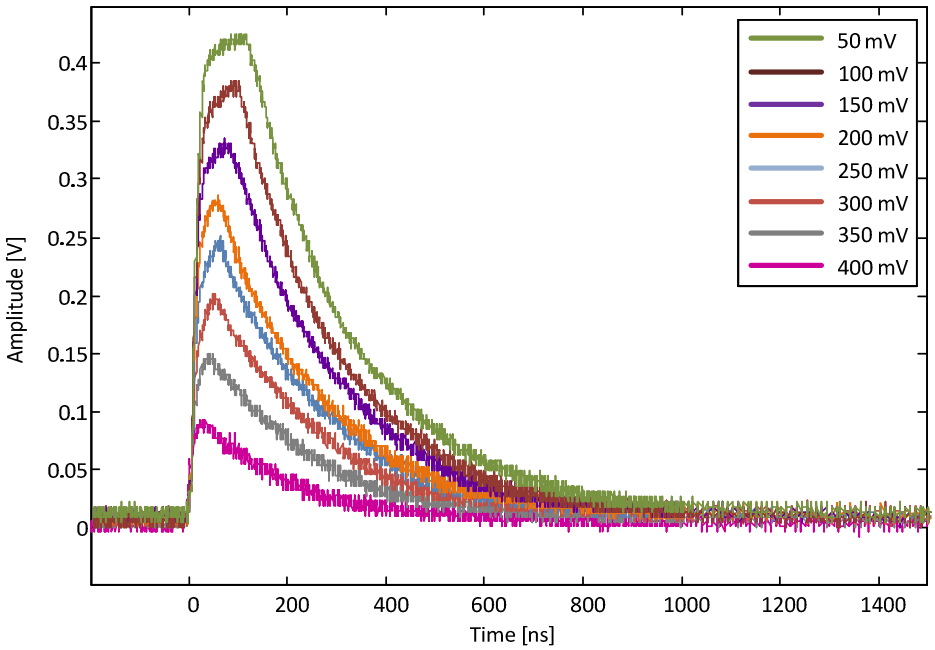


Figure 5.1 - Change in SPAD signal amplitude with increasing excess bias voltage. The inset shows the excess bias voltage (V_{OP} - V_{BD}).

Figure 5.2 depicts the I-V characteristic plot obtained from the measurement of 65 nm SPAD detector [48]. The curve shows that the breakdown voltage is close to 9.2V and the avalanche current is of the order of 10 µA.

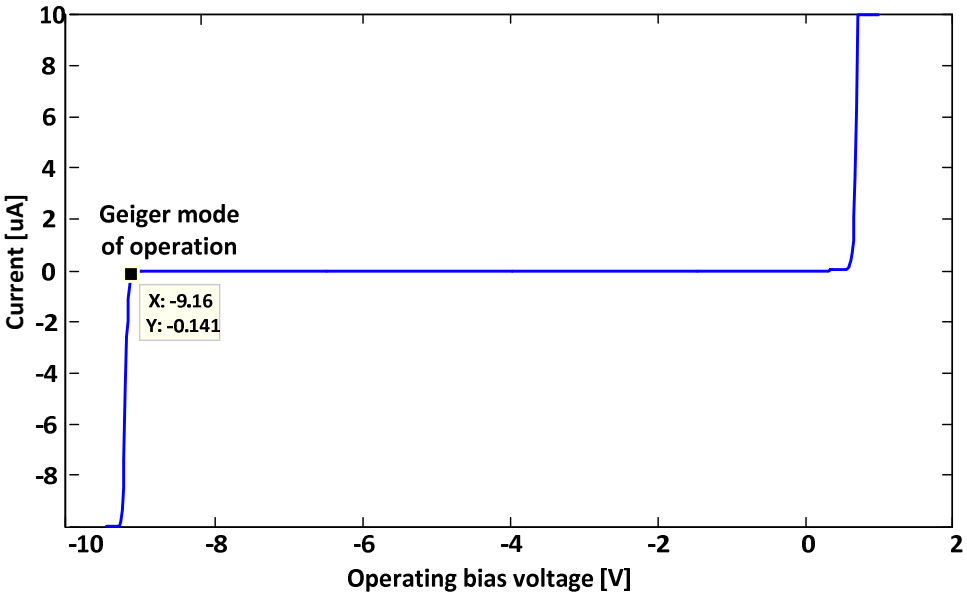


Figure 5.2 - I-V characteristic of 65 nm SPAD device.

5.1.2 SPAD-TDC Interface

Figure 5.1 shows the working of SPAD-TDC interface circuitry. The sequence is indicated by numbers 1-6.

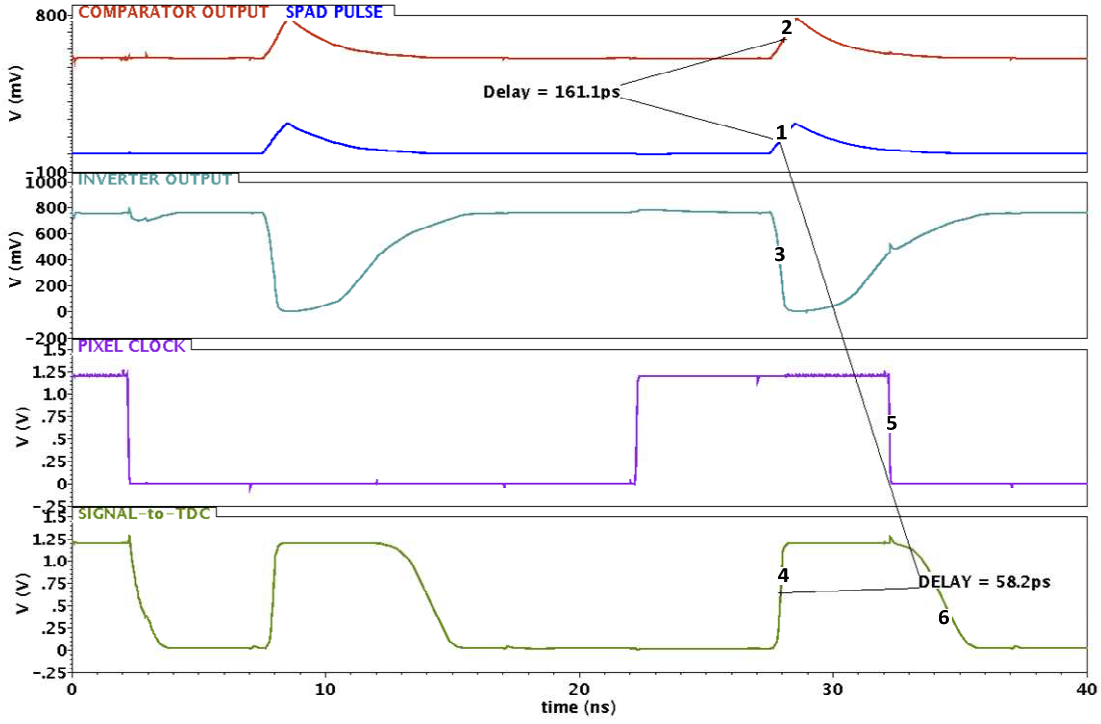


Figure 5.3 - Transient response of SPAD-TDC interface circuit.

For designing a suitable SPAD-TDC interface, it is important to model the behaviour of the SPAD. To represent the SPAD signal amplitude, as shown in Figure 5.1, a model [26] consisting of a pulsed current source of 10 µA is used. The diode resistance and junction capacitance are modeled with 250 kΩ resistor in parallel with a capacitance of 60 fF.

This model generates small SPAD amplitude (edge 1 in Figure 5.3) similar to measured signal. The reference voltage can be controlled externally and is set to 0.2V to sense this small SPAD signal amplitude. The SPAD pulse is detected by the comparator and it generates a level shifted signal (edge 2) very quickly, in about 160 ps. The inverter at the output of the comparator regenerates the voltage levels (edge 3) and signals the photon hit to the TDC (edge 4). The propagation of pulse from SPAD to TDC takes ~60 ps. The interface circuitry is reset by the negative edge of pixel clock (edge 6).

5.1.3 Time-to-Digital Converter

In this section, the working of ring oscillator based TDC is shown and then, analysis on its performance is conducted.

5.1.3.1 Working

The plot in Figure 5.4 shows the input-output behaviour of the proposed ring oscillator based TDC. The characteristic plot is stair-case shaped with the LSB size of 23 ps. The dynamic range of TDC is 14 bits. A snapshot of the full dynamic range is shown.

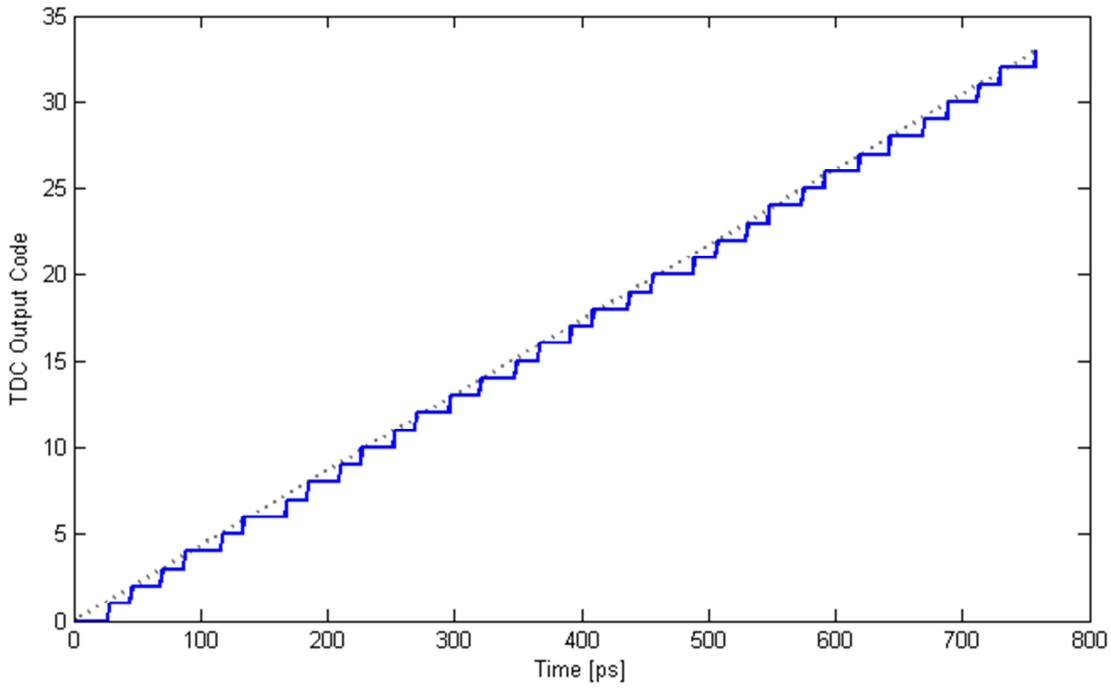


Figure 5.4 - Output characteristic plot of proposed TDC.

The TDC consists of a delay line of 7 inverters, each with an average delay of 23 ps. Therefore, one cycle propagation takes approximately 160 ps. Two such cycles increment the

counter by 1 bit and therefore, one TDC ring oscillation period is close to 320 ps. The frequency of ring oscillation is equal to 3.125 GHz.

5.1.3.2 Metastability of Latch

In the proposed TDC, a switched inverter based latch has been employed to capture the state of the delay line. It is often important to analyze the metastability of the latch or a flip-flop. The metastability window should be much smaller than TDC resolution to minimize the probability of bubbles in the output code. In [53], the metastability characterization of latches is discussed in detail. Following the similar approach, the metastability for switched inverter based latch is analyzed and the obtained plots for rising and falling inputs are shown in Figure 5.5.

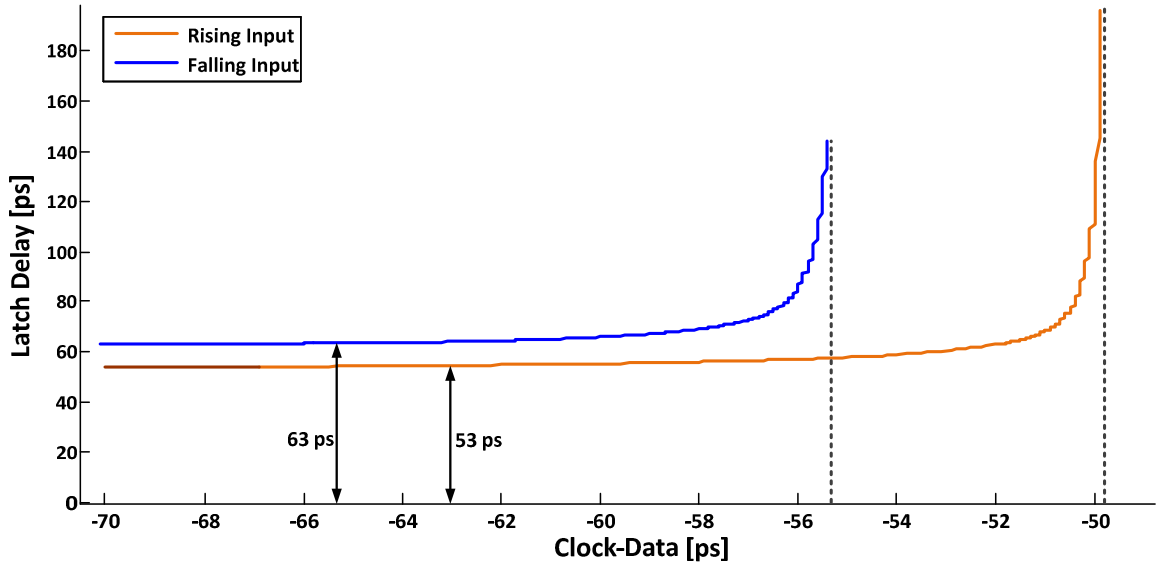


Figure 5.5 - Metastability curve of latch for 0→1 input and 1→0 input signal.

In a latch, as clock-to-data delay is decreased, the latch delay or data-to-Q (on y-axis) increases exponentially. The latch delay approaches infinity at the point of metastability point. In the plot shown in Figure 5.5, the metastability points are ~49.9 ps and ~55.4 ps respectively. To the right of the metastability point, the latch holds the previous state and therefore, the latch delay is zero. In the worst case, the metastability window for this latch is less than 0.1 ps for a resolution time of 200 ps which is much smaller than an average inverter delay of 23 ps.

5.1.3.3 TDC Performance across Corners

The variation in delay of TDC delay taps in ring configuration at different corners is shown in Table 5.1. The change in delay across a single TDC oscillation consisting of two cycles is shown in Figure 5.6. A single cycle across a delay line consists of signal propagation across 7 delay taps and the maximum delay at any corner is at half ring oscillation or after 1 cycle.

It can be observed that the typical behaviour lies at the mean of slow-fast and fast-slow corners as expected.

Table 5.1 - TDC resolution at typical and 4 corners.

Corner	Average Resolution [ps]	Worst Resolution [ps]
typical	22.9	24.6
ss	28.3	30.9
ff	18.6	19.6
sf	23.7	27.8
fs	22.6	25.7

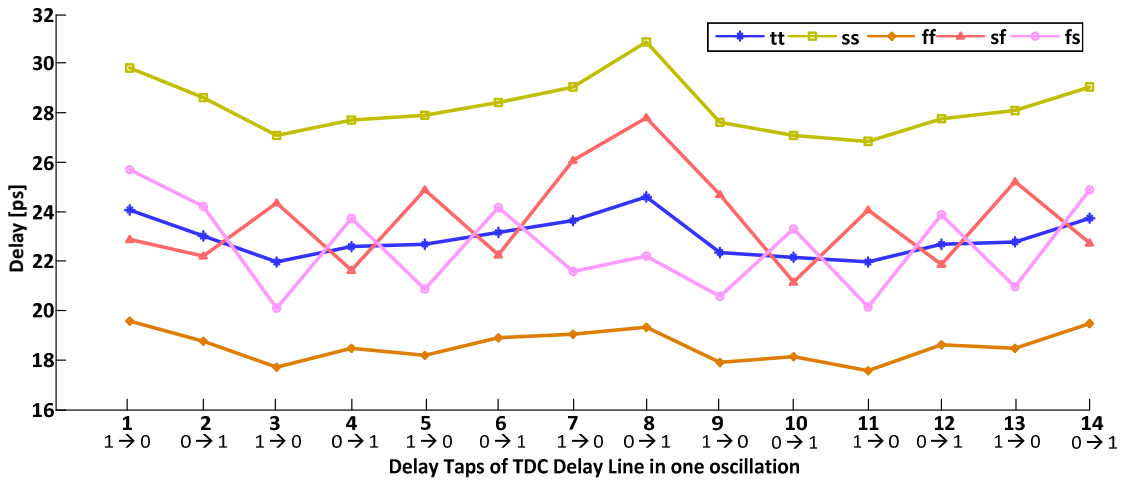


Figure 5.6 - Propagation delay across each delay tap in the ring oscillator at different corners.

5.1.3.4 TDC Linearity

The proposed ring oscillator based TDC achieves a good linearity with DNL_{MAX} = 0.48 LSB and INL_{MAX} = 0.3 LSB. The linearity analysis is conducted over 2 TDC oscillations where each oscillation has 14 delay taps. Since the results obtained in both TDC oscillations are coherent with each other, it can be assumed that every cycle of TDC has the similar behavior across all its oscillations.

In the DNL plot shown in Figure 5.7, we observe that the DNL behavior of both oscillations is similar with a maximum DNL of 0.48 LSB. The maximum DNL occurs at half oscillation or when the propagation returns back to the first delay tap (NAND gate). The INL plot shown in Figure 5.8 also shows the same effect.

The linearity results are as expected from the TDC architecture and since INL < 1 LSB, the accuracy of the system is not limited by TDC linearity.

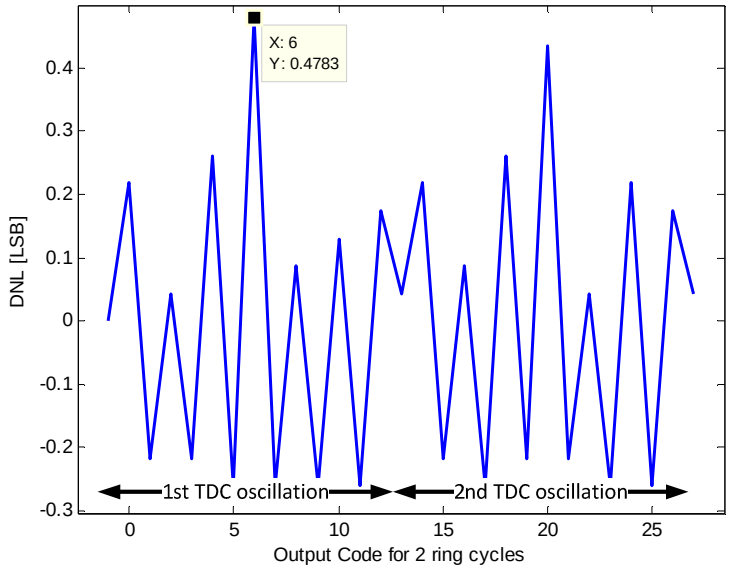


Figure 5.7 - Differential non-linearity for TDC ring oscillation.

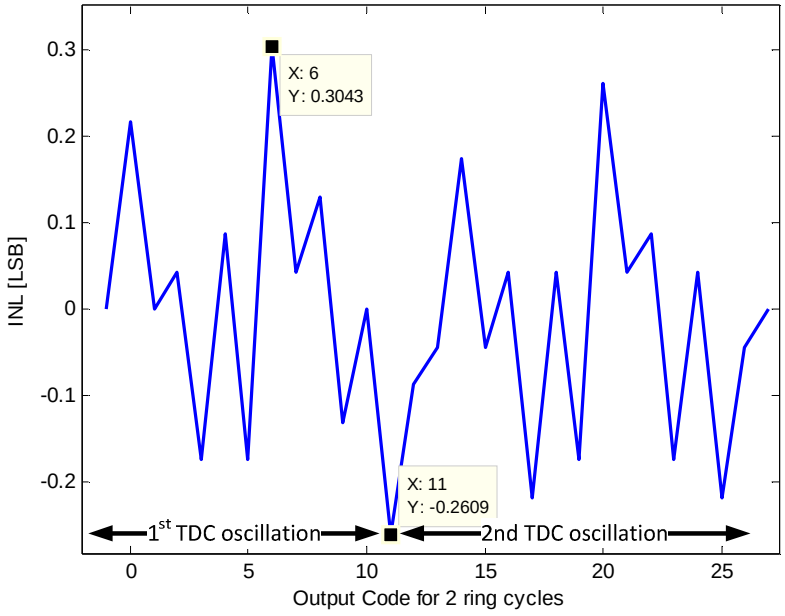


Figure 5.8 - Integral non-linearity for TDC ring oscillation.

5.1.3.5 TDC Mismatch

The mismatch in delay for the center-most delay tap is shown in Figure 5.9. The plot was obtained by performing Monte Carlo simulation on the propagation delay for temperature and process variations. The 1σ deviation of approximately 1 ps and if we approximate 3σ to 4 ps, then the maximum percentage mismatch is 10-17%. This does not affect the system performance as SPAD jitter is the limiting factor for depth resolution. The accumulation of mismatch across one oscillation is shown in Figure 5.10 where the 3σ percentage mismatch is 4% of the total delay only. A summary of mismatch across all delay taps is shown in Table 5.2. The worst mean delay is 24.76 ps and worst σ is 1.32 ps. The plots for each case are shown in Appendix B.

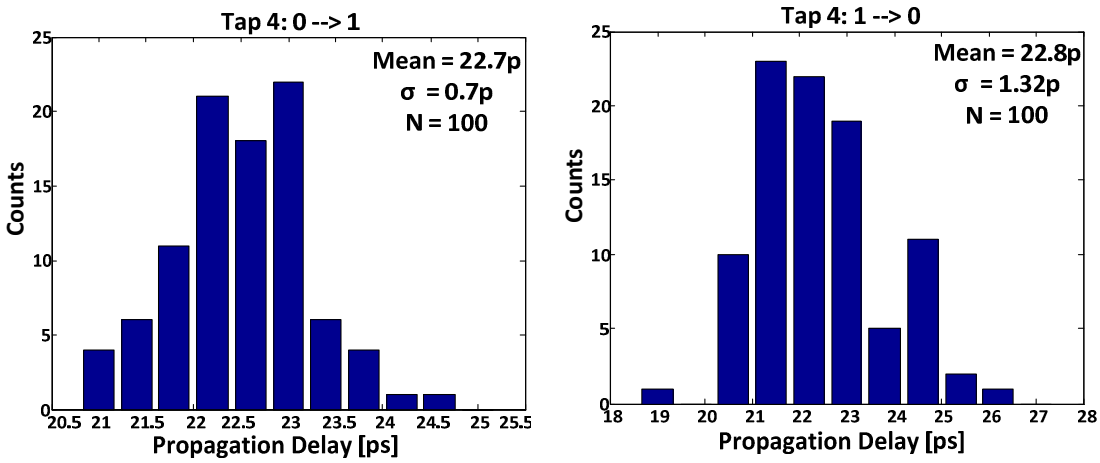


Figure 5.9 - Mismatch in delay for the centre-most delay tap for 0→1 (left) and 1→0 (right) transition.

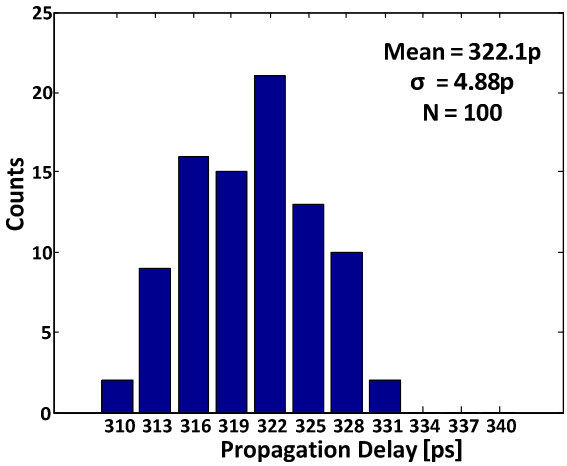


Figure 5.10 - Mismatch in the delay of a single TDC cycle of 14 delay taps.

Table 5.2 - Summary of mismatch in delay across all seven inverters for each type of propagation.

		Tap 1	**Tap 2**	**Tap 3**	**Tap 4**	**Tap 5**	**Tap 6**	**Tap 7**
$\mathbf{t_{PLH}}$	Mean	23.08 ps	22.13 ps	22.6 ps	22.7 ps	23.26 ps	23.6 ps	24.04 ps
	σ	0.8 ps	0.7 ps	0.77 ps	0.7 ps	0.73 ps	0.7 ps	0.73 ps
$\mathbf{t_{PHL}}$	Mean	22.38 ps	22.1 ps	22.01 ps	22.8 ps	22.77 ps	23.85 ps	24.76 ps
	σ	1.29 ps	1.15 ps	1.07 ps	1.32 ps	1.26 ps	1.2 ps	0.99 ps

5.1.3.6 Delay Variation with Temperature

The TDC is robust and works across a large temperature range of -40° C to 125° C, as shown in Figure 5.11. The behaviour across worst case temperatures is similar to that at room temperature (27° C). The maximum variation in average delay from normal room temperature of 27° C is less than 4 ps (at temperature of 125° C). This is tolerable from system level as SPAD jitter is the limiting factor for resolution in this system.

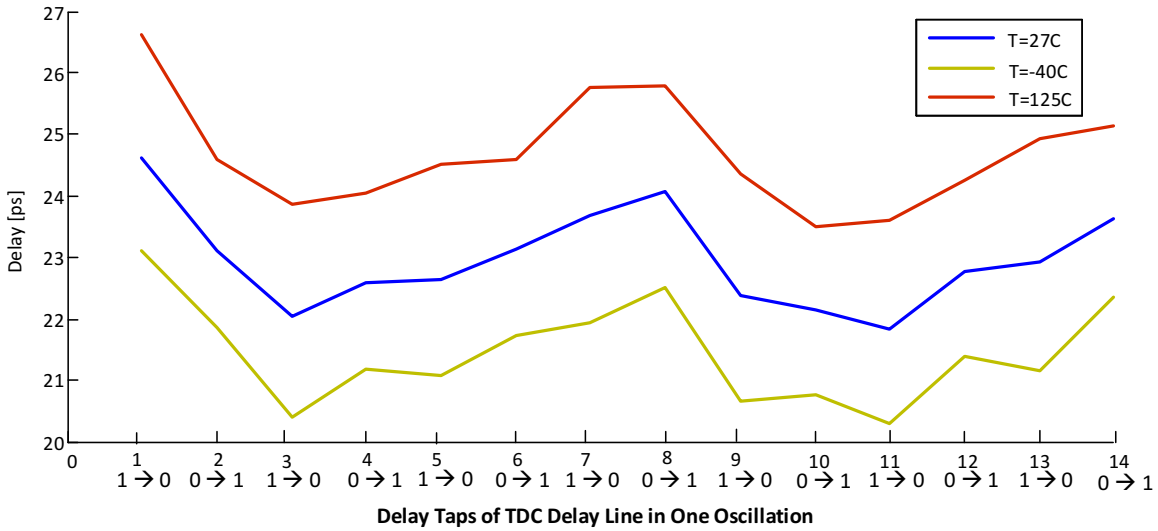


Figure 5.11 - TDC delay line behaviour across different temperatures.

In Figure 5.12, the linear variation in average delay is shown across the temperature range of -40° C to 125° C. The maximum variation is less than 2 ps which is appropriate from system level.

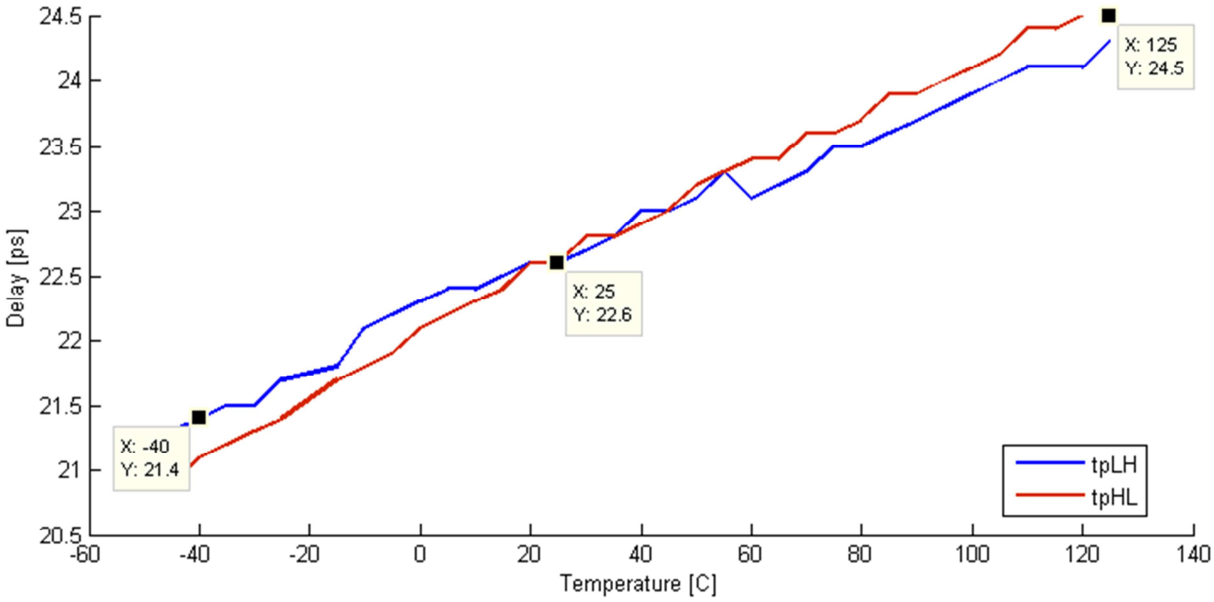


Figure 5.12 - Change in delay across temperature range of -40° C to 125° C.

5.1.3.7 Delay Variation with Supply

The reduction in power supply due to IR drop or noise affects the inverter delay as shown in Figure 5.13. The degree of change in delay is similar for both high-to-low and low-to-high propagation. While t_{pLH} is affected slightly more at lower supply voltages due to weaker PMOS conduction, t_{pHL} is affected more at higher voltages due to weaker NMOS conduction. The change in full TDC oscillation delay (2 cycles or 14 delay taps) is shown in Figure 5.14.

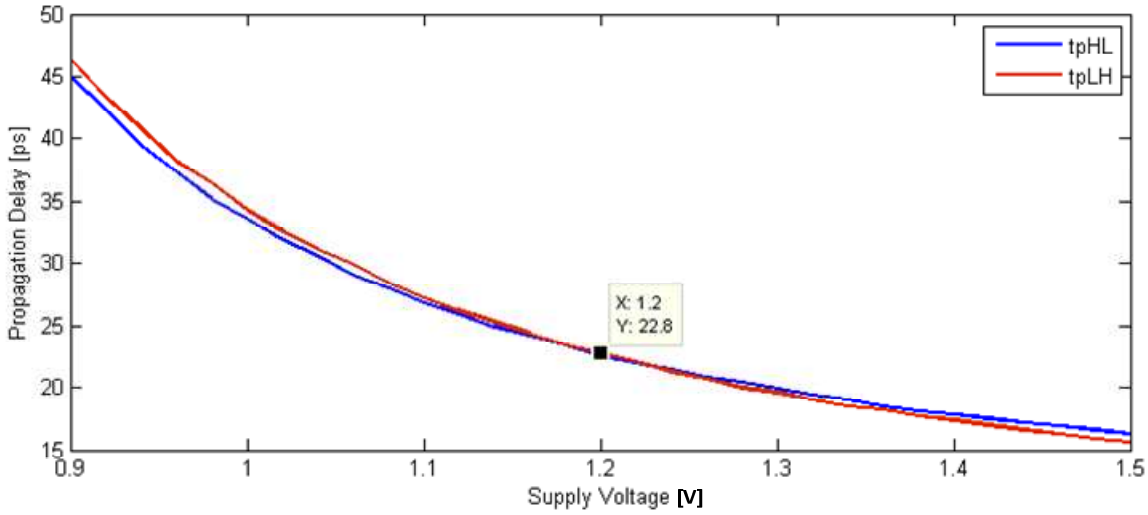


Figure 5.13 - Variation in inverter propagation delay of centre-most delay tap with reducing supply voltage.

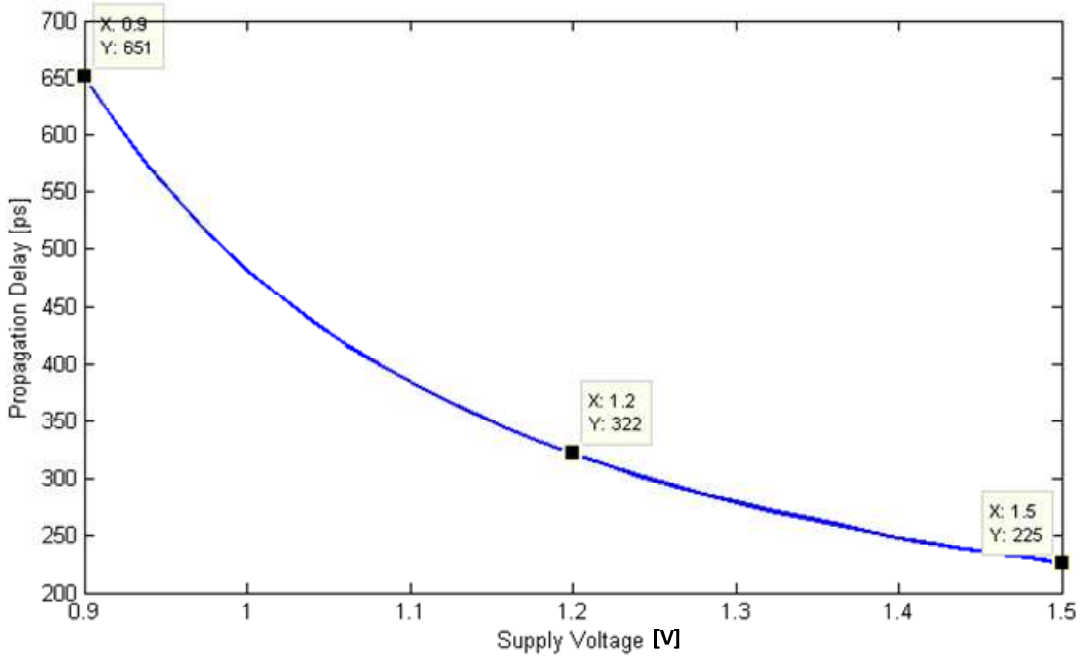


Figure 5.14 - Variation of one ring oscillation delay (sum of all delay taps) with supply voltage.

5.1.4 Energy Consumption

In Section 3.2.10, the energy consumption requirements per pixel per event were drawn out. It was shown that for a 1000x1000 array, the average energy consumption must be ~3.5 pJ for a time interval measurement of 5 ns. The desired specification is met with the obtained post layout simulation result shown in Figure 5.15. The energy plots are linear in nature and if required, extrapolation can be done to estimate the energy consumption at a larger time interval. The slight bump in the energy curve at 5 ns is due to change in state of counter bits from 0111→1000.

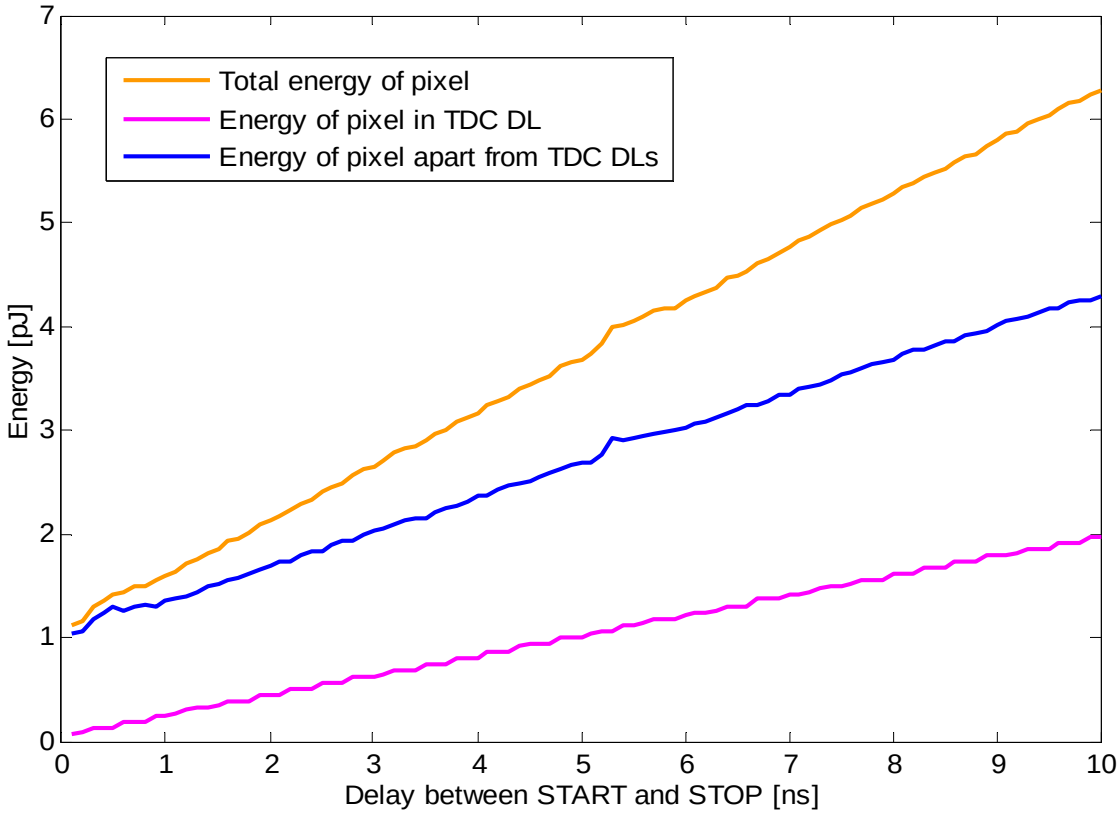


Figure 5.15 - Energy consumption in pixel over time

The total energy consumption per pixel per event is about 3.4-3.6 pJ. If a frame-rate of 30 fps and a 100x averaging is required to achieve 1 mm depth resolution, the number of events per pixel per second required are 30.10^4. The calculations in Table 5.3 give an indication of the power consumed per pixel.

Table 5.3 - Power consumption analysis at system level.

Average energy per pixel per event	3.5 pJ
Number of events per pixel per second	$30 \cdot 10^4$
Average power per pixel	1.05 µW
Power consumed by 10^6 pixels	1.05 W

The leakage current consumption of each TDC in every pixel is ~40 nA. For a 1000x1000 pixel array, it translates to ~50 µW of extra power. Therefore, the total power consumed for a 1000x1000 array is still be close to 1 W and is within system specifications.

5.2 System Level Performance

This section discusses the working of system and verifies through post-layout results.

The following plot shows the sequence of steps when a photon hits the SPAD device. Edge 1 represents the arrival of photon on SPAD and a pulse is generated, as on node V_s of Figure 4.1. The SPAD-TDC interface output (edge 2) indicates the event arrival and triggers event signal (edge 3). This starts the TDC ring oscillation as shown by edge 4. When the following clock edge arrives, it generates a negative edge on the pixel level clock (edge 6). The negative edge of pixel clock resets the SPAD-TDC interface signal and also stops the TDC ring oscillation (edge 7). This action makes the TDC ready for next photon detection.

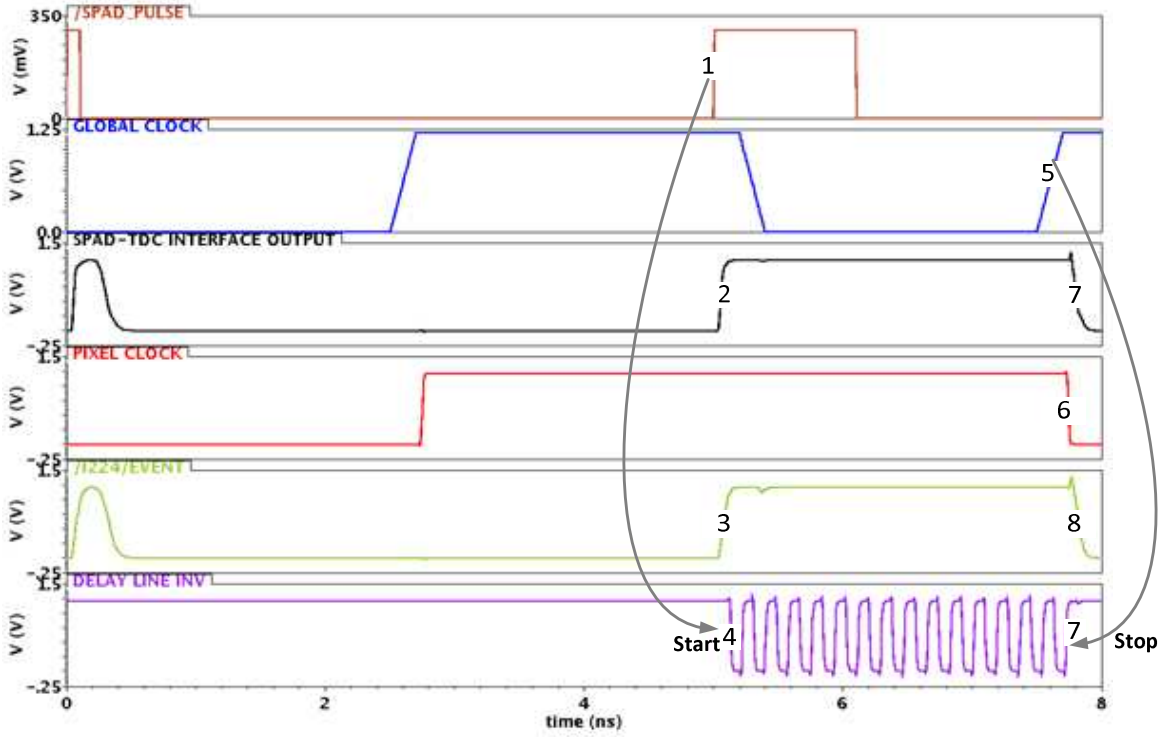


Figure 5.16 - Sequence of steps indicating the operation of a pixel after the photon hit on the SPAD.

5.2.1 IR Drop

Each pixel in the array of 32x32 pixels consists of two power lines and one ground line:

(a) VDD for delay line inverters - VDD_{DL}
(b) VDD for other circuitry in pixel - VDD_{COMMON}
(c) Ground line common to whole pixel circuitry - GND

A separate VDD for delay line has been made so that it is not influenced by IR drop of VDD supplying power to other circuitry in the pixel. To achieve small IR drop in all power and ground lines, it is important that these lines are distributed well across the pixel array. The layout was carried out considering the requirement and the final pixel layout for power and ground line is shown in Figure 5.20. For IR drop analysis, every pixel is divided into its serial R_S and parallel R_P components. The grid resistive network is shown in Figure 5.17. The worst IR drop is experienced by the center-most pixel.

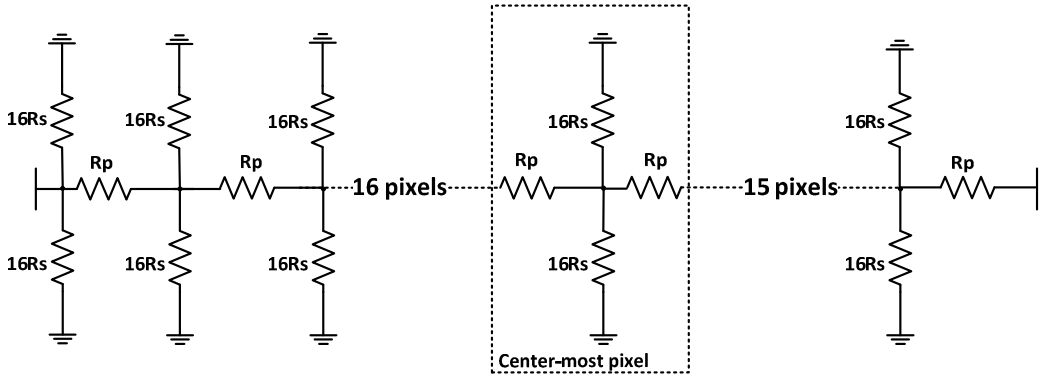


Figure 5.17 - Resistive network of the centre-most pixel of the array.

Assuming $R_P < R_S$, the equivalent resistive network is shown in Figure 5.18.

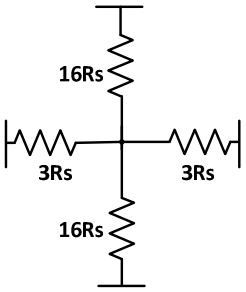


Figure 5.18 - Equivalent resistive network as seen at the centre-most pixel.

5.2.1.1 IR Drop for VDD_{DL}

Metal layer 5 is used to supply VDD to delay line. Metal 5 has a sheet resistance of 0.14 Ω/□. In each pixel, the series and parallel components are:

$$R_s = \rho_s \cdot \square = 0.14 \cdot 15.2 = 2.12\,\Omega$$
$$R_p = \rho_s \cdot \square = 0.14 \cdot 3.5 = 0.5\,\Omega$$

The equivalent resistive network is shown in Figure 5.18. The contribution of R_P is small and the final resistance is approximately equal to $2R_S$ or 4.5Ω.

The current consumption in each pixel during normal TDC operation is ~150 µA and therefore, the IR drop experienced by the centremost pixel when one pixel is working is given by:

$$I \cdot R = 150\ \mu A \cdot 4.25 \simeq 0.65\ mV$$

If 10 pixels work in the imager at one point of time, the worst case IR drop is 6.5 mV, which is acceptable at system level.

The plot in Figure 5.16 shows the change in TDC delay with reduced supply voltage. It can be observed that the propagation delay of low input to high output (t_{pLH}) is affected more than t_{pHL}. This is because PMOS conducts less strongly with reducing supply voltage. The change in delay is ~5 ps and it can be tolerated at system level because the SPAD jitter is the limiting factor of the system.

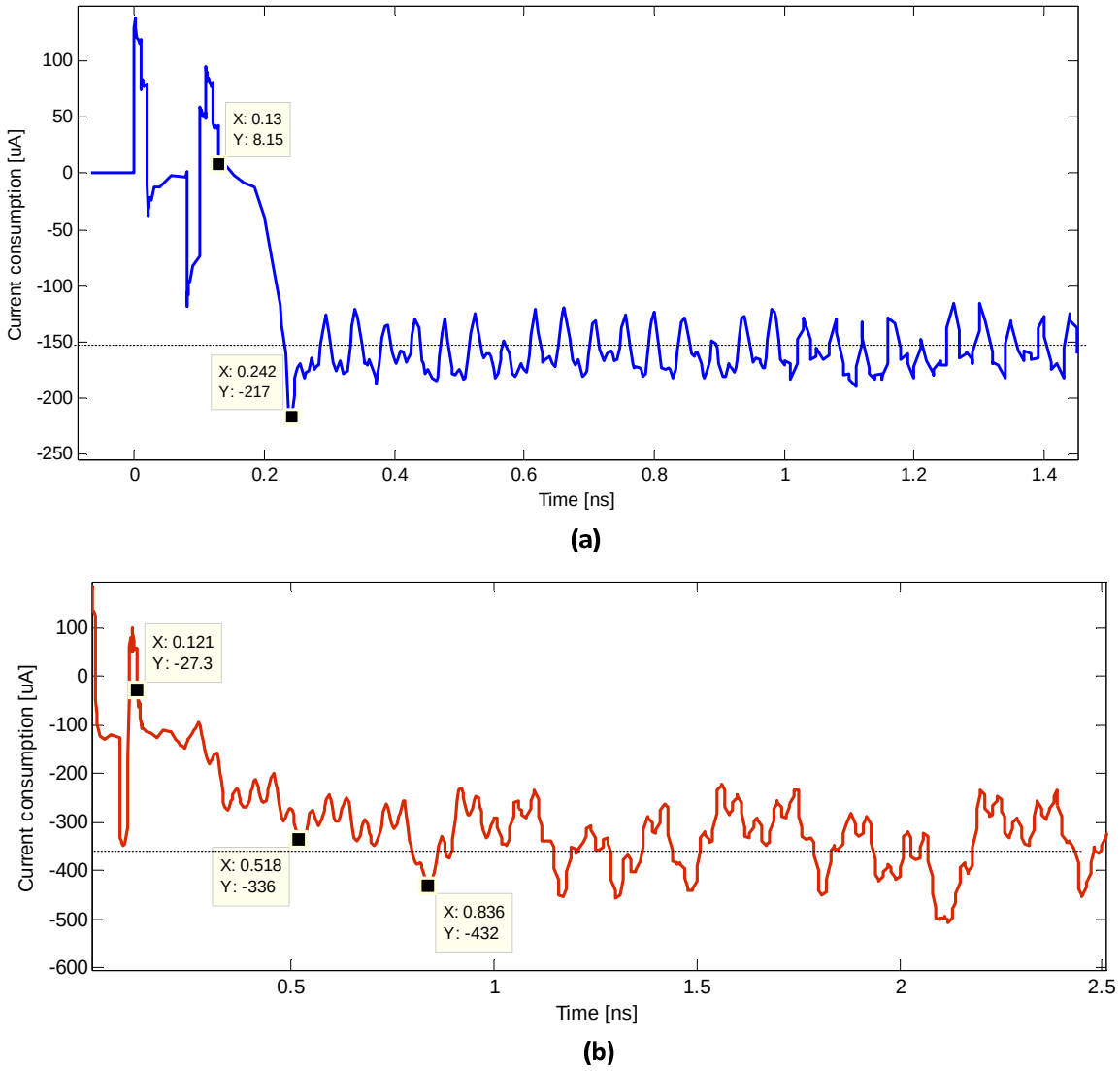


Figure 5.19 - Current absorption in each pixel for: (a) VDD for delay line; (b) VDD for other circuitry.

5.2.1.2 IR Drop for VDD_{COMMON}

For supplying VDD to other circuitry in the pixel, metal 5 is used. In each pixel, the series and parallel components are:

$$R_s = \rho_s \cdot \square = 0.14 \cdot 16 = 2.24\,\Omega$$
$$R_p = \rho_s \cdot \square = 0.14 \cdot 10 = 1.4\,\Omega$$

With similar calculation for VDD_{DL}, the equivalent resistance is 4.5Ω. The current absorption in each pixel during normal TDC operation is ~400 µA and therefore, the IR drop experienced by the centremost pixel when one pixel is working is given by:

$$I \cdot R = 400\ \mu A \cdot 4.5 \simeq 1.8\ mV$$

If 10 pixels work in the imager at one point of time, the worst case IR drop is close to 18 mV which is acceptable at system level.

5.2.1.3 IR Drop for GND

Metal layer 1 is used for ground connections. Metal 1 has a sheet resistance of 0.16 Ω/□. In each pixel, the series and parallel components are:

$$R_s = \rho_s \cdot \square = 0.16 \cdot 27 = 4.32\,\Omega$$
$$R_p = \rho_s \cdot \square = 0.16 \cdot 100 = 16\,\Omega$$

With similar calculation for VDD_{DL}, the equivalent resistance is $2R_S$ or equal to 8.65Ω. The current flowing through ground line in each pixel during normal TDC operation is ~550 µA and therefore, the IR drop experienced in the worst case by a single pixel is,

$$I \cdot R = 550\ \mu A \cdot 8.65 \simeq 4.7\ mV$$

If 10 pixels operate in the imager at one point of time, the IR drop is close to 47 mV which is acceptable at system level.

Figure 5.21 shows the VDD and GND power rails in the imager. VDD bus runs vertically in metal 5 while GND is in metal 1 all across the imager to limit the IR drop.

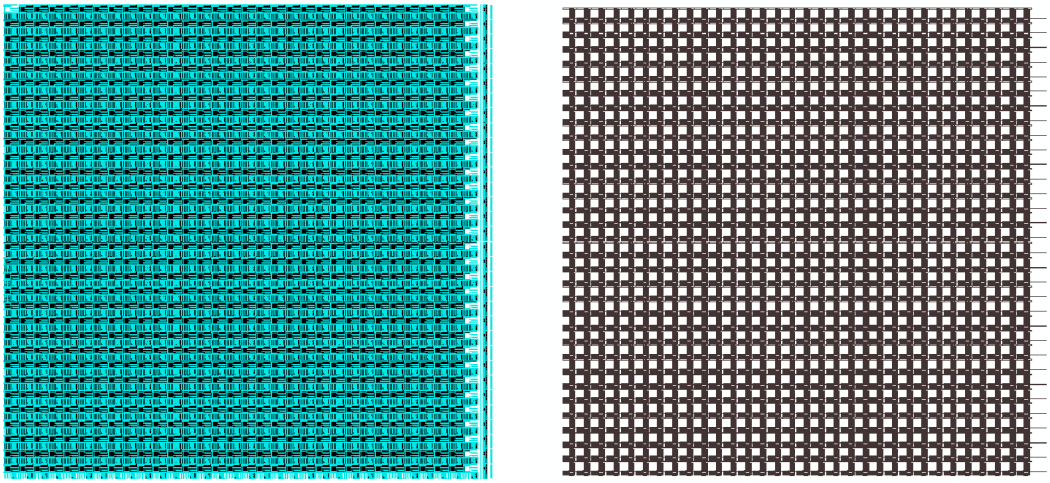


Figure 5.20 - Layout of GND (left) and VDD (right) grids.

5.2.2 Row Decoder

Figure 5.21 shows the performance of 5 bit row decoder. Static logic based implementation results in high speed and is verified from the post-layout simulation result. The delay between the input and its output is less than 0.5 ns which is fast enough to meet the frame rate requirements; one pixel readout must finish within 100 ns to achieve <3 mm depth resolution and a frame rate of 30 fps.

The 'enable' signal ensures no readout operation of the pixel array. The word-lines are activated only when the 'enable' signal is externally asserted.

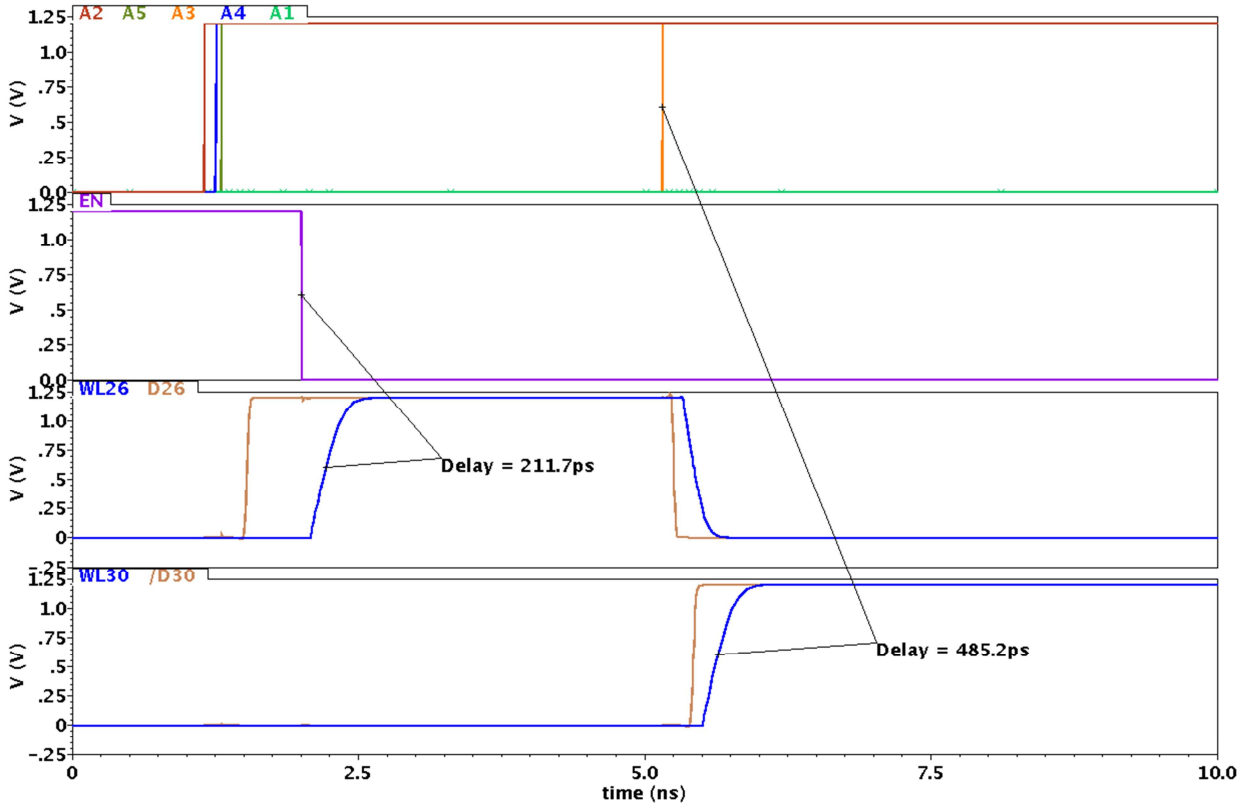


Figure 5.21 - Operation of row decoder

5.2.3 Serializer

The plot in Figure 5.22 shows the working of serializer in post-layout simulations. When the wordline WL31 goes high, all the pixels in 31st row are selected and the flip-flops of the serializer capture the pixel data. When signal SCAN_EN is asserted, the serializer switches into serial mode shifting out the bits sequentially at every positive edge of SCAN_CLK. It can be observed that the data from a single pixel is captured and shifted out within 100 ns and therefore, fits the system requirements of high frame-rate.

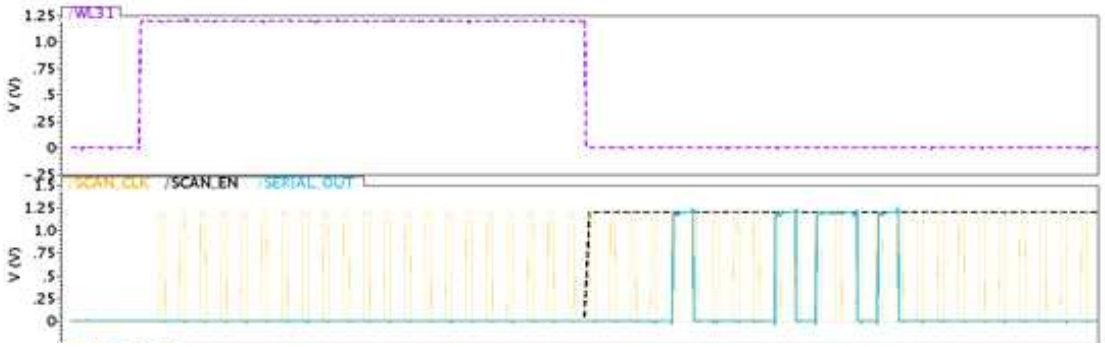


Figure 5.22 - A serializer shifting pixel bits. One pixel readout is completed in less than 100 ns.

5.3 Summary

This chapter presented post-layout simulation results of the 32x32 pixel array which verify the correct working of the implemented 3D imager in 65 nm technology. The results also substantiate the correct behavior of the SPAD-TDC interface circuitry and the proposed TDC architecture. Furthermore, all the obtained results are in line with the system requirements and thus, establish the accurate operation of the circuit from system level perspective. The final post-layout results are summarized in Table 5.4.

Table 5.4 - Summary of post-layout results.

Parameter	**Value**
Technology	65 nm CMOS
Number of pixels	32x32
Pixel Area	25 µm x 25 µm
SPAD Jitter at FWHM	235 ps
Median DCR	< 100 kHz
TDC Resolution	22.9 ps (mean)
TDC Dynamic Range	14 bits
TDC Jitter (1σ)	1.3 ps
Maximum Range (z_{max})	30 m
TDC energy	3.5 pJ for 5 ns
TDC current consumption	550 µA
DNL	0.48 LSB
INL	0.3 LSB
Chip Power	70 mW (for 32x32 array)
Frame rate	30 fps (typical)
	~ 1000 fps (maximum)
Fill factor	5.76 %

5.4 Future Work

The work done in this thesis project achieves a set of goals. However, a lot more work can be directed for its growth into a commercial product. In this section, a few recommendations are made:

a) The chip needs to be fabricated and tested. The obtained measurement results would need to be evaluated against the desired system requirements.
b) In this work, a prototype chip of 32x32 array of pixels was built. However, the final goal is to realize a VGA camera or, even better, a 1000x1000 pixel array. This scaling up would need considerable effort in terms of an efficient floor-planning along with a suitable clock distribution network.
c) Data processing can be integrated on-chip to treat the read out data from pixels and reduce the I/O pin count.

d) In the current work, only temporal averaging is possible. To incorporate spatial averaging, a group of pixels need to work together and such mechanism can be introduced in the layout of 1000x1000 pixel array.

e) The global controller is critical for the desired commercial product to regulate different blocks of the system. Design and integration of such a controller having multiple features of averaging, blanking, reconnaissance and event driven readout would bring the system closer to the final product.

f) The current SPADs are experimental and they still show several drawbacks mostly in terms of noise. In future implementations, it will be advisable to replace the current devices with new more reliable ones, possibly operating in the IR range (the current SPADs are optimized for operation in the visible spectrum.)

g) An appropriate clock distribution scheme for a 1000x1000 array needs to be analyzed for minimal skew, area and power consumption.

6. Conclusions

In this dissertation, a novel time-of-flight 3D real-time imager was proposed for security camera applications, specifically for face recognition and for surveillance activities. A 32x32 array of pixels was laid out in 65 nm CMOS technology and it is first of its kind. The prototype laid the ground-work for building a 1000x1000 pixel array for realizing a 3D camera. Each pixel consists of a SPAD detector, along with its dedicated TDC for high speed image acquisition. Multiple ideas were proposed at system level to lower power consumption. The technique of reconnaissance modifies the clock frequency to reduce the time of operation of event-driven TDC. The pixel level clock-conditioning simplifies the global clock distribution network and adds clock gating in each pixel.

Furthermore, a novel TDC architecture was proposed to suit the requirements imposed by a million pixel array. The proposed architecture is based on a ring oscillator and is highly compact in size. It achieves 23 ps resolution with 14 bits of dynamic range in 25 µm x 25 µm pixel area. The implementation was done such that the inter-metal coupling is minimal and routing delays between delay taps of TDC is equal.

In this 3D imager implementation, the requirements of achieving 1 mm depth resolution at 30 m range were met by the proposed TDC architecture, column readout scheme and with required averaging. The fill factor achieved is 5.76% and is almost twice as better than the state-of-the-art SPAD based 3D imagers.

The post-layout simulated results are within the system objectives and are better than the state-of-the-art SPAD based imagers in most respects. The summary is shown in Table 6.1.

.

Table 6.1 - Summary of the performance with respect to state-of-the-art 3D SPAD imagers.

	[20]	[7]	[21]	[8]	[22]	**This Work (post-layout)**
Technology	0.35 µ	0.13 µ	0.18 µ	0.13 µ	0.13 µ	65 nm
Array Size	60x48	32x32	340x96	160x128	128x96	32x32
Range	5m	15m (100 ns)	128m	8m (55 ns)	45m	30 m
Resolution	38 mm (1σ distance resolution)	< 20 mm	< 32 mm	<8.5 mm	160 mm (1σ distance resolution)	< 1 mm
Time Jitter at FWHM	-	185 ps	-	140 ps		235 ps
Frame Rate (fps)	-	-	10	25 k (raw)	20	30
Median Dark Count Rate	245 Hz	-		Below 50 Hz	Below 100 Hz	
DNL / INL (LSB)	INL: 110 mm upto 2.4m	±0.4/±0.2	-0.52/0.73	±0.3/±2	INL: 5 mm upto 2.4 mm	+0.48/+0.3
Chip power dissipation	35 mW	94 mW (78 mA)	-	550 mW	40 mW	70 mW*
Chip area	6.5 x 5.5 mm^2	-	3.3 x 3.3 mm^2	11 x 2.3 mm^2	7.2 x 7.5 mm^2	1.8 x 1.8 mm^2
Pixel pitch	85 µ	85 µ	-	50 µ	44.65 µ	25 µ
Type of TDC used	Phase domain approach	Coarse-fine	Flash TDC	Ring Oscillator	Phase domain ΔΣ approach	Ring Oscillator
Fill factor	0.5%	-	70% (macrocells)	1%	3.17%	5.76%
Year	2009	2009	2011	2011	2011	2011

*The TDC power consumption is close to 1 mW. Majority of the power consumption is due to static current consumption in the comparator structure placed at SPAD-TDC interface. With better SPAD devices, a simpler interface consisting of a single inverter is possible which would drastically improve the power consumption.

A.

It is often desirable to know the precision of the mean of a distribution. This can be done by determining the standard deviation of the sampled mean. Assuming statistical independence of the underlying random variables, the standard deviation of the mean is related to the standard deviation of the distribution by:

$$\sigma_{MEAN} = \frac{1}{\sqrt{N}}\sigma$$

where N is the number of observations used to estimate the mean. The above relation can be easily proven with:

$$\begin{aligned}
&\mathrm{var}(X) \equiv \sigma_x^2 \\
&\mathrm{var}(X_1 + X_2) \equiv \mathrm{var}(X_1) + \mathrm{var}(X_2) \\
&\mathrm{var}(cX_1) \equiv \mathrm{c}^2\ \mathrm{var}(X_1)
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathrm{var}(mean) &= \mathrm{var}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right) = \frac{1}{N^2}\mathrm{var}\left(\sum_{i=1}^{N} X_i\right) \\
&= \frac{1}{N^2}\mathrm{var}\left(\sum_{i=1}^{N} X_i\right) = \frac{N}{N^2}\mathrm{var}(X) = \frac{1}{N}\mathrm{var}(X)
\end{aligned}$$

This results in:

$$\sigma_{mean} = \frac{1}{\sqrt{N}}\sigma$$

B.

The mismatch for each delay tap of the delay line in TDC ring oscillator is shown here. The plots are obtained by Monte Carlo simulations performed on the propagation delay of each inverter over process and temperature variations.

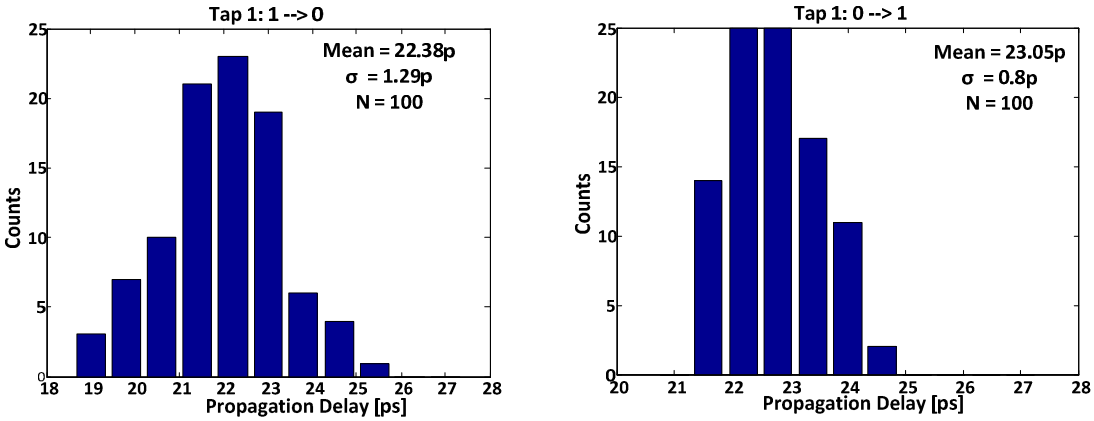


Figure B.1 -Mismatch in delay for delay tap 1 for low-to-high (left) and high-to-low (right) transition.

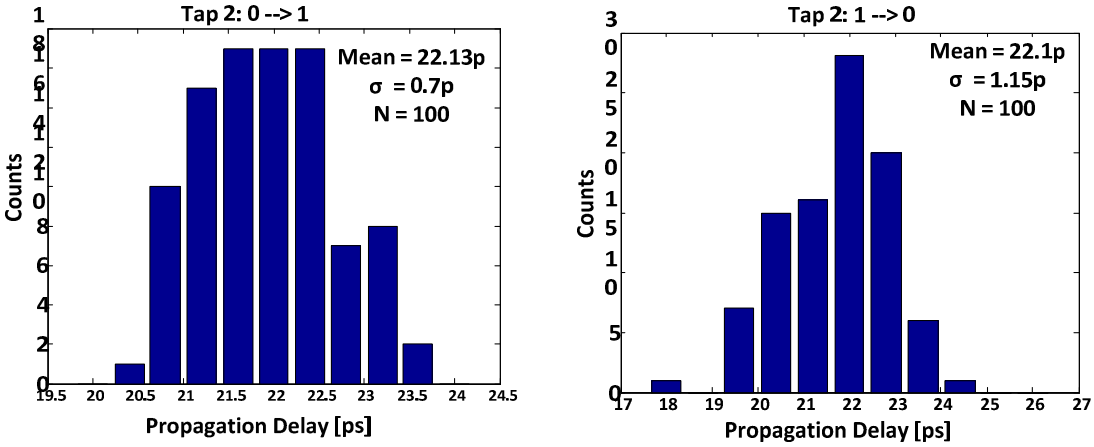


Figure B.2 - Mismatch in delay for delay tap 2 for low-to-high (left) and high-to-low (right) transition.

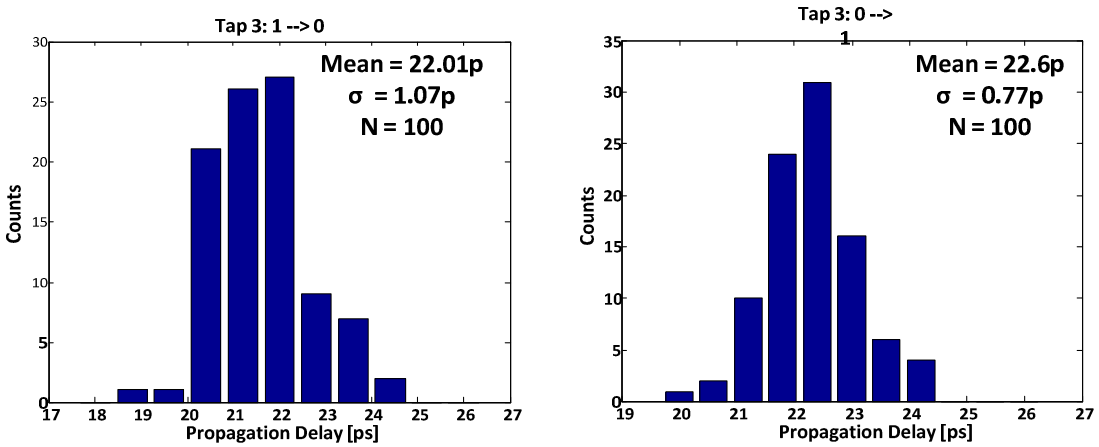


Figure 6.1 - Mismatch in delay for delay tap 3 for low-to-high (left) and high-to-low (right) transition.

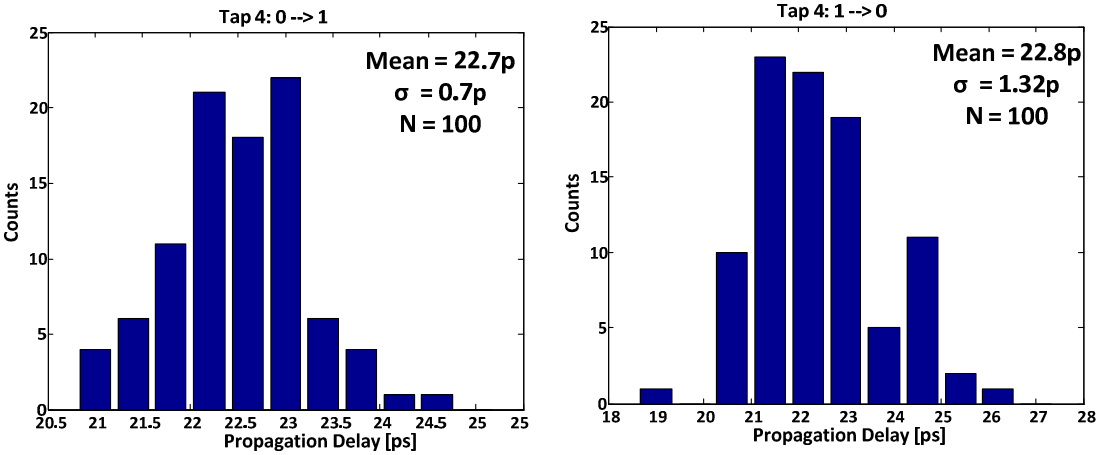


Figure 6.2 - Mismatch in delay for delay tap 4 for low-to-high (left) and high-to-low (right) transition.

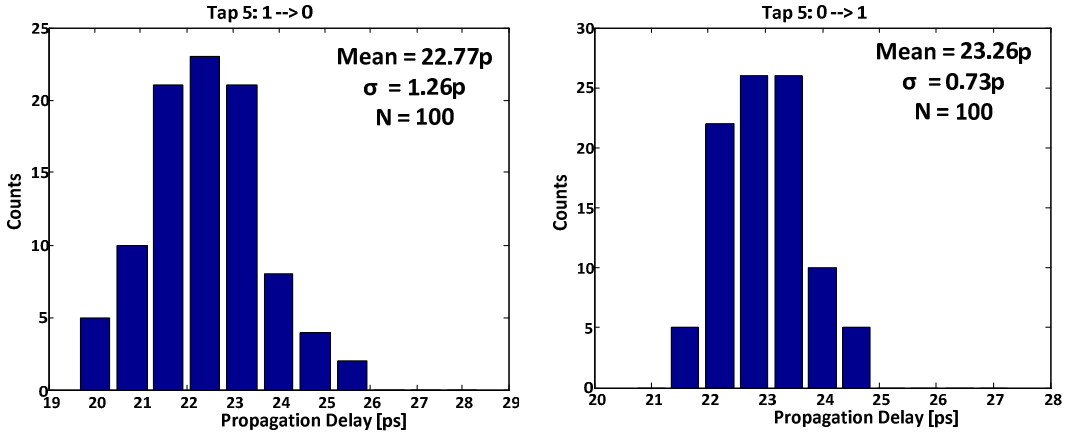


Figure 6.3 - Mismatch in delay for delay tap 5 for low-to-high (left) and high-to-low (right) transition.

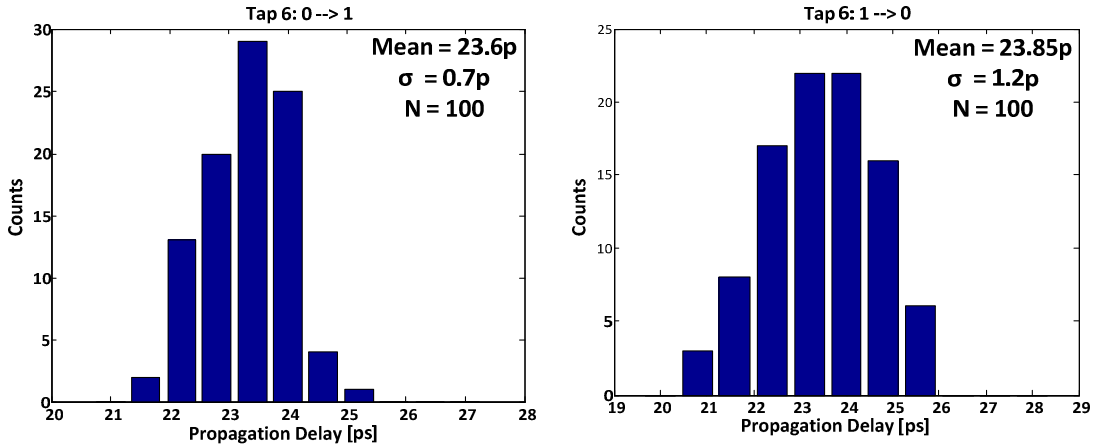


Figure 6.4 - Mismatch in delay for delay tap 6 for low-to-high (left) and high-to-low (right) transition.

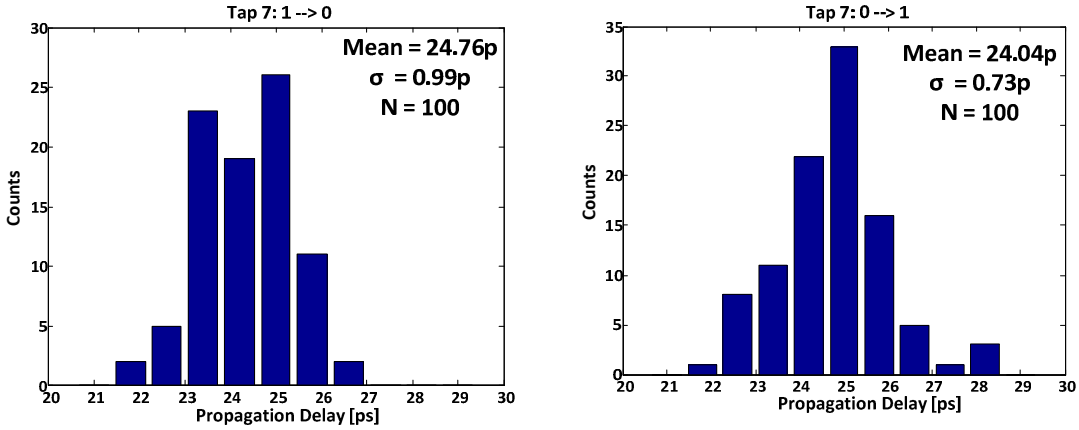


Figure 6.5 - Mismatch in delay for delay tap 7 for low-to-high (left) and high-to-low (right) transition.

Bibliography

[1] C. L. Niclass, "Single-Photon Image Sensors in CMOS : Picosecond Resolution for Three-Dimensional Imaging," EPFL, 2008.

[2] P. J. Besl, "Active, optical range imaging sensors," *Machine Vision and Applications*, vol. 1, no. 2, pp. 127-152, Jun. 1988.

[3] T. Kreis, *Handbook of holographic interferometry: optical and digital methods*. Wiley-VCH, 2005, p. 542.

[4] C. Vest and D. Sweeney, "Holographic interferometry of transparent objects with illumination derived from phase gratings," *Applied Optics*, vol. 9, no. 10, pp. 2321–2325, 1970.

[5] A. K. Jain, *Three-Dimensional Object Recognition Systems, Volume 1 (Advances in Image Communication)*. Elsevier Science.

[6] J. W. Goodman and J. W. Goodman, *Introduction to Fourier optics*. New York: McGraw-Hill, 1996.

[7] M. Gersbach et al., "A parallel 32x32 time-to-digital converter array fabricated in a 130 nm imaging CMOS technology," in *ESSCIRC 2009 - 34th European Solid-State Circuits Conference*, pp. 196–199.

[8] C. Veerappan et al., "A 160x128 Single-Photon Image Sensor with On-Pixel 55ps 10b Time-to-Digital Converter," in *Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2011 IEEE International*, 2011, pp. 312–314.

[9] B. Büttgen, T. Oggier, M. Lehmann, R. Kaufmann, and F. Lustenberger, *CCD/CMOS Lock-in Pixel for Range Imaging: Challenges, Limitations and State-of-the-Art*. 2005.

[10] 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, Sep. 2005.

[11] R. Schwarte, "New Electrooptical Mixing and Correlating Sensor: Facilities and Applications of the Photonic Mixer Device (PMD)," *Proceedings of SPIE*, vol. 3100, pp. 245-253, 1997.

[12] R. Schwarte, "New Optical Four-Quadrant Phase Detector Integrated into a Photogate Array for Small and Precise 3D Cameras," *Proceedings of SPIE*, vol. 3023, pp. 119-128, 1997.

[13] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, "Single-Photon Synchronous Detection," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 7, pp. 1977-1989, 2009.

[14] M. Vaidyanathan et al., "High performance ladar focal plane arrays for 3D range imaging," *2004 IEEE Aerospace Conference Proceedings (IEEE Cat. No.04TH8720)*, pp. 1776-1781, 2004.

[15] C. Niclass, M. Sergio, and E. Charbon, "A CMOS 64x8 Single Photon Avalanche Diode Array with Event-Driven Readout," *2006 Proceedings of the 32nd European Solid-State Circuits Conference*, pp. 556-559, Sep. 2006.

[16] P. Schwille, U. Haupts, S. Maiti, and W. W. Webb, "Molecular Dynamics in Living Cells Observed by Fluorescence Correlation Spectroscopy with One- and Two-Photon Excitation," *Biophysical journal*, vol. 77, no. 4, pp. 2251-65, Oct. 1999.

[17] W. Becker, "FRET Measurements by TCSPC Laser Scanning Microscopy," *Proceedings of SPIE*, vol. 4431, no. June, pp. 94-98, 2001.

[18] A. V. Agronskaia, L. Tertoolen, and H. C. Gerritsen, "Fast Fluorescence Lifetime Imaging of Calcium in Living Cells.," *Journal of biomedical optics*, vol. 9, no. 6, pp. 1230-7, 2004.

[19] Aerotech Report, "Resolution , Accuracy , and Repeatability." [Online]. Available: http://www.aerotech.com/products/pdf/EngineeringRef.pdf.

[20] C. Niclass, C. Favi, T. Kluter, F. Monnier, and E. Charbon, "Single-photon synchronous detection," *Solid-State Circuits, IEEE Journal of*, vol. 44, no. 7, pp. 1977–1989, 2009.

[21] C. Niclass, M. Soga, H. Matsubara, and S. Kato, "A 100m-Range 10-frame / s 340x96-pixel Time-of-Flight Depth Sensor in 0.18u CMOS," in *ESSCIRC 2011 - 36th European Solid-State Circuits Conference*.

[22] R. J. Walker, J. A. Richardson, and R. K. Henderson, "A 128×96 pixel event-driven phase-domain ΔΣ-based fully digital 3D camera in 0.13μm CMOS imaging technology," in *2011 IEEE International Solid-State Circuits Conference*, 2011, vol. 21, no. 14, pp. 410-412.

[23] E. Charbon, "CMOS Quantum Imagers for Picosecond Sensing Applications," *Slide Presentation*. [Online]. Available: http://si.epfl.ch/files/content/sites/si/files/shared/CMOSQuantumImagers_Charbon.pdf.

[24] E. Charbon, "CMOS Single Photon Detectors," *Slide Presentation*. [Online]. Available: http://si.epfl.ch/files/content/sites/si/files/shared/Nanoscale Sys and Tech/QuantumDevices_Charbon.pdf.

[25] H. Yoon, *65nm SPADs Design - Internal Report - TU Delft*. 2011.

[26] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche Photodiodes and Quenching Circuits for Single-Photon Detection.," *Applied optics*, vol. 35, no. 12, pp. 1956-76, Apr. 1996.

[27] A. Rochas, "Single Photon Avalanche Diodes in CMOS Technology," EPFL, 2003.

[28] A. Borowski, "Celeritas3D Real-Time Imagers, Sensors and Detectors - Patent Description Document," 2010.

[29] F. Zappa, M. Ghioni, S. Cova, C. Samori, and A. C. Giudice, "An Integrated Active-Quenching Circuit for Single-Photon Avalanche Diodes," *IEEE Transactions on Instrumentation and Measurement*, vol. 49, no. 6, pp. 1167-1175, 2000.

[30] C. Niclass and M. Soga, "A Miniature Actively Recharged Single-Photon Detector Free of Afterpulsing Effects with 6ns Dead Time in a 0.18 um CMOS Technology," in *IEEE International Electron Devices Meeting*, 2010, pp. 340-343.

[31] O. Sasaki et al., "1.2 GHz GaAs Shift Register IC for Dead-time-less TDC Application," in *IEEE Transactions on Nuclear Science*, 1989, vol. 36, no. 1, pp. 512–516.

[32] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, "1.3 V 20 ps Time-to-Digital Converter for Frequency Synthesis in 90 nm CMOS," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2006, vol. 53, no. 3, pp. 220-224.

[33] S. Henzler et al., "A Local Passive Time Interpolation Concept for Variation-Tolerant High-Resolution Time-to-Digital Conversions," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 7, pp. 1666-1676, 2008.

[34] L. Vercesi, A. Liscidini, and R. Castello, "Two-Dimensions Vernier Time-to-Digital Converter," *IEEE Journal of Solid-State Circuits*, vol. 45, no. 8, pp. 1504-1512, 2010.

[35] P. Dudek, S. Szczepanski, and J. V. Hatfield, "A high-resolution CMOS time-to-digital converter utilizing a Vernier delay line," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 2, pp. 240-247, 2000.

[36] S. Henzler, *Time-to-Digital Converters*, vol. 29. Dordrecht: Springer Netherlands, 2010, pp. 103-113.

[37] V. Ramakrishnan and P. T. Balsara, "A Wide-Range, High-Resolution, Compact CMOS, Time to Digital Converter," in *19th International Conference on VLSI Design held jointly with 5th International Conference on Embedded Systems Design (VLSID'06)*, 2006, pp. 197-202.

[38] M. Lee, S. Member, and A. A. Abidi, "A 9b, 1.25ps Resolution Coarse-Fine Time-to-Digital Converter in 90 nm CMOS that Amplifies a Time Residue," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 769-777, 2008.

[39] S. Henzler, *Time-to-Digital Converters*, vol. 29. Dordrecht: Springer Netherlands, 2010, pp. 1-3.

[40] I. Nissinen, A. Mantyniemi, and J. Kostamovaara, "A CMOS Time-to-Digital Converter Based on a Ring Oscillator for a Laser Radar," *ESSCIRC 2004 - 29th European Solid-State Circuits Conference (IEEE Cat. No.03EX705)*, pp. 469-472.

[41] R. Rashidzadeh, M. Ahmadi, and W. C. Miller, "An All-Digital Self-Calibration Method for a Vernier-Based Time-to-Digital Converter," *IEEE Transactions on Instrumentation and Measurement*, vol. 59, no. 2, pp. 463-469, Feb. 2010.

[42] M. Z. Straayer and M. H. Perrott, "A Multi-Path Gated Ring Oscillator TDC With First-Order Noise Shaping," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 4, pp. 1089-1098, 2009.

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

[44] C. Niclass and E. Charbon, "A Single Photon Detector Array with 64 x 64 Resolution and Millimetric Depth Accuracy for 3D Imaging," in *Solid-State Circuits Conference, 2005. Digest of Technical Papers. ISSCC. 2005 IEEE International*, 2005, pp. 364–604.

[45] A. Sammak, M. Aminian, L. Qi, W. D. de Boer, E. Charbon, and L. K. Nanver, "A CMOS Compatible Ge-on-Si APD Operating in Proportional and Geiger Modes at Infrared Wavelengths," in *IEEE International Electron Devices Meeting*, 2011.

[46] "Patent Pending," 2011.

[47] M. W. Fishburn and E. Charbon, "System Tradeoffs in Gamma-Ray Detection Utilizing SPAD Arrays and Scintillators," *Nuclear Science, IEEE Transactions on*, vol. 57, no. 5, pp. 2549–2557, 2010.

[48] Y. Maruyama, *65 nm SPAD Measurements - Internal Report - TU Delft*. 2011.

[49] B. Nikolic, V. G. Oklobdzija, V. Stojanovic, and M. Ming-Tak Leung, "Improved sense-amplifier-based flip-flop: design and measurements," *IEEE Journal of Solid-State Circuits*, vol. 35, no. 6, pp. 876-884, Jun. 2000.

[50] Arun Kumar Singh and A. K. Singh, *Digital Principles Foundation of Circuit Design and Application*, First Edit. New Age International (P) Limited Publishers, 2006.

[51] M. F. Takashi Tokairin, Mitsuji Okada, Masaki Kitsunezuka, Tadashi Maeda, "A 2.1-to-2.8GHz All-Digital Frequency Synthesizer with a Time-Windowed TDC," in *Vlsi Design*, 2010, vol. 53, no. 3, pp. 188-189.

[52] J. B. C. Hurwitz, "Layout: The other half of Nanometer CMOS Analog Design," in *ISSCC Tutorial*, 2011.

[53] M. Bhushan, M. B. Ketchen, and K. K. Das, "CMOS Latch Metastability Characterization at the 65 nm Technology Node," *2008 IEEE International Conference on Microelectronic Test Structures*, pp. 147-151, Mar. 2008.