EPFL

Thèse n° 10 193

Modelling and design of CMOS SPAD sensors for quantum random number generation

Présentée le 1er juin 2023

Faculté des sciences et techniques de l'ingénieur
Laboratoire d'architecture quantique
Programme doctoral en microsystèmes et microélectronique

pour l'obtention du grade de Docteur ès Sciences

par

Pouyan KESHAVARZIAN

Acceptée sur proposition du jury

Prof. D. Atienza Alonso, président du jury
Prof. E. Charbon, Dr M. Stipcevic, directeurs de thèse
Prof. F. Regazzoni, rapporteur
Dr L. Gasparini, rapporteur
Prof. K. Choo, rapporteur

École polytechnique fédérale de Lausanne

2023

Attention is a state of openness that assumes there is something new to be seen, it is also true that this state must resist our tendency to declare our observations finished– to be done with it.
— Jenny Odell

Anyone who considers arithmetical methods of producing random digits is, of course, in a state of sin.
— John von Neumann

زن زندگی آزادی

To Mali.

Acknowledgements

Over the course of my graduate studies, I have had the great privilege of meeting, working, exploring, and celebrating with a number of extraordinary professionals. This story that has passed with many highs and lows, it is now time to express my gratitude for all who have been a part of the journey. First, to my thesis director, Prof. Edoardo Charbon and my thesis co-director Dr. Mario Stipčević. Edoardo, you pushed me to seize opportunities but also gave me the freedom to explore, which I am grateful for. Mario, I was quite 'green' in terms of QRNGs at the start, so without your breadth of knowledge, this work would not have been possible. Also, to my M.Sc. supervisors, Prof. Michal Okoniewski and Prof. John Nielsen back home in Calgary, thank you for inspiring me to pursue research in the first place.

I arrived at AQUA lab in February 2019, on the heels of what was, by all accounts, the greatest Christmas party of all time. I slowly met the organizers of that party (my colleagues), many of whom have become close friends and confidants. In the early days, the old alumni of Delft set the standard for quality and culture. A special thank you to Ivan Michel Antolovic and Augusto Carimatto for helping me find my bearings in the beginning and for your kind friendship as time went on. Another special thank you to Francesco Gramuglia and Ekin Kizilkan for the rock-solid collaboration we had together on the 55 nm project. On Fridays *au Galop*, a newer generation of colleagues, Ming-Lo Wu, Paul Mos, Bariş Can Efe, Halil Kerim Yildirim, Yang Lin, Vladimir Pesic, Tommaso Milanese, Utku Karaca, Emanuele Ripiccini, Jad Benserhir, Feng Liu, Won-yong Ha and others, came together for beers, burgers and chess, which was always good fun. Thank you Arin Ulku, Ermanno Bernasconi, Preethi Padmanabhan, Carlo Alberto Fenoglio, Andrei Ardelean, Andrada Muntean, Michael Wayne, Scott Lindner, Jiuxuan Zhao, Simone Frasca, Kodai Kaneyasu, Chufan Zhou, Kazuhiro Morimoto, and Yasemin Uzun, Yating Zou. It has been a pleasure getting to know you all. To my friends Bedirhan Ilik and Andrea Ruffino, I've cherished our many discussions, laughs and adventures around Switzerland. Thank you also to Claudio Bruschini, Samuel Burri and the team at Qrypt for the collaborations that led to the development of the QRNG sensors. In this same vein, I would like to thank GlobalFoundries and the team in Singapore for the excellent opportunity to develop SPADs in the 55 nm BCD process, along with Myung-Jae Lee and the group at KIST for collaborating on this project. I owe a huge debt of gratitude to the two lab administrators at AQUA, Brigitte Khan, and her successor, Begonia Tora, for all the help, kindness and professionalism, over the years.

It's certainly not an exaggeration to say that, in early 2020, the world changed in a way that our generation has not yet experienced. Covid-19 turned the *train-train-quotidien* into *restez-chez-vous*. The process of professional and personal adaptation was frustrating, anxiety-ridden,

long and uncertain. What's more, it also just happened to coincide with a period of great personal difficulty. Not to mention that, as PhD students in the field of microelectronics, we had to adapt to the resulting 'chip shortage' and 'semiconductor crisis' (still waiting on those level shifters ordered in 2021...). Getting through this stretch would not have been possible without the support of my friends and family (those weekend calls/Zooms were fundamental). Maman, Baba, Naz, Bryce, Leila, Mina, thank you for your love, encouragement and support. To my dear friends back home and across the globe, Brett, Rebecca, Stephen, Ally, Cassie, Megan, Jules, Mike, Michael, Marie, and Amar, you know I can't express how grateful I am for you all. And of course to Auntie Sanja, Uncle Anuj and Nimaya, I'm so blessed to have your consistent, loving presence.

Switzerland is a country that is unique and wonderful in many ways. One of those is the existence of four national languages. Although not initially concerned with the challenge (hubris) of not speaking a word of any of those languages, I quickly learned the daunting struggle (humbled) of such an endeavor. I'm grateful for my french teacher, Abigail, who put up with my constant, « J'ai pas eu le temps de faire mes devoirs cette semaine...» and patiently helped me rectify (at least to some extent) my language deficiency.

Finally, and most importantly, to my partner-in-crime, Malithi. You have been part of this journey since Day 1 (in fact, since Day 1 of undergrad). Whether it was weekend trips from GVA to LBA, or the TGV between Lausanne and Gare de Lyon, we hustled relentlessly to prioritize each other along the way. Your love and care during the lowest of the lows, your ability to drive many of our joint projects (including the wedding), your patience with the years-long journey, your reminders to stay persistent in the face of adversity, and of course, your passion for the adventure, has made it all possible.

Neuchâtel, 09 March 2023 P. K.

Abstract

Modern digital connectivity has necessitated the creation of robust methods for securely storing and transferring data. At the heart of all security infrastructure is the random number generator (RNG). While random numbers find use in a variety of applications, from scientific simulation to gambling, it is in the hardware security domain, where their performance is most critical. As a fundamental security primitive, the requirements for an RNG are exhaustive. Any practical RNG realization, which extracts entropy from a physical source of randomness, requires intensive modelling, characterization and robustness against environmental changes and adversarial attacks. Existing implementations often rely on heavy levels of post-processing of the digital bits produced to meet performance requirements. Furthermore, with the advent of Quantum Computing, which promises to shatter the security of many existing encryption protocols, the demands placed on RNG designs is fortifying considerably. Quantum random number generators (RNGs) hope to meet those demands. A variety of techniques, that exploit various quantum mechanical phenomena, can be used to generate random numbers. However, many require complex, bulky setups that are not amenable to small form factor implementations or scalability. Single-photon avalanche diodes (SPADs) are a CMOS-compatible modern detector technology enabling versatile sensing of optical photons. SPAD-based sensors, which have the ability to co-integrate a myriad of digital timing and processing functions, are a promising potential solution for QRNG implementations.

This thesis focuses on modelling and design of quantum random bit generators in a 55 nm Bipolor-CMOS-DMOS (BCD) process. First, several new detectors are developed in a 55 nm BCD process. Effective optimization of the designs, without changes to the standard process, are proposed and tested, resulting in excellent noise and sensitivity performance. These detectors are then combined with pixel circuits that exploit photon-timing statistics for generation of random bits. Intensive modelling on bias and system modelling are performed. In particular, an analytical method for determining serial correlation of bits, when considering detector dead time and afterpulsing, is proposed and validated by way of simulation and experiments. Based on the analysis of bias, a dynamic comparator based sampling flip-flop approach, for reduction of bit bias, is introduced as a component in a quantum random flip-flop (QRFF), configuration. This method is compared and contrasted to a first photon-arrival comparison bit generation method. The designs are scaled to SPAD sensor arrays that, when combined with external illumination, function as gigabit-per-second QRNGs. The FortunaSPAD QRNG is a dual-interface design, capable of a combined output data rate of

3.3 Gbps. The FortunaSPAD2 QRNG sensor implements a macro-pixel design to improve robustness against pixel failure.

Key words: Random number generation, quantum random number generation, single-photon avalanche diodes, hardware security, BCD technology, pixel-electronics, dead time modelling, afterpulsing effects, quantum random flip-flop.

Résumé

La connectivité numérique moderne a nécessité la création de méthodes robustes pour stocker et transférer des données en toute sécurité. Au cœur de toute infrastructure de sécurité se trouve le générateur de nombre aléatoires (GNA). Alors que les chiffres aléatoires trouvent une utilisation dans une variété d'applications, comme de la simulation scientifique au jeu de hasard, c'est dans le domaine de la cryptographie que leurs performances sont les plus critiques. En tant que primitive de sécurité fondamentale, les exigences pour un GNA sont exhaustives. Toute réalisation GNA pratique, qui extrait l'entropie d'une source physique aléatoire, nécessite une modélisation, une caractérisation et une robustesse intensives contre les changements environnementaux et les attaques adverses. Les implémentations existantes reposent souvent sur des niveaux élevés de post-traitement des chiffres numériques produits pour répondre aux exigences de performances. De plus, avec l'avènement de l'informatique quantique, qui promet de briser la sécurité de nombreux protocoles de cryptage existants, les exigences imposées aux conceptions GNA se renforcent considérablement. Les générateurs de nombres aléatoires quantiques (GNAQ) espèrent répondre à ces demandes. Une variété de techniques, qui exploitent divers phénomènes de mécanique quantique, peuvent être utilisées pour générer des nombres aléatoires. Cependant, beaucoup nécessitent des configurations complexes et volumineuses qui ne se prêtent pas à des implémentations ou à une évolutivité à petit facteur de forme. Les diodes diode à avalanche monophotonique sont une technologie de détection moderne compatible avec CMOS, qui permet une détection polyvalente des photons optiques. Les capteurs basés sur SPAD, qui ont la capacité de co-intégrer une myriade de fonctions de synchronisation et de traitement numériques, constituent une solution potentielle prometteuse pour les implémentations GNAQ.

Cette thèse porte sur la modélisation et la conception de générateurs de nombres aléatoires quantiques dans un processus Bipolaire-CMOS-DMOS (BCD) 55 nm. Tout d'abord, plusieurs nouveaux détecteurs sont développés dans un procédé BCD à 55 nm. Une optimisation efficace des conceptions, sans modification du processus standard, est proposée et testée, ce qui se produits de performances de bruit et de efficacité. Ces détecteurs sont ensuite combinés avec des circuits de pixels qui exploitent les statistiques de synchronisation des photons pour la génération de chiffres aléatoires. Une modélisation intensive sur les biais et la corrélation est effectuée. En particulier, une méthode analytique pour déterminer la corrélation en série des chiffres, compte tenu du temps mort du détecteur et de la post-impulsion, est proposée et validée par simulation et expérience. Sur la base de l'analyse du biais, une approche de flip-flop

d'échantillonnage basée sur un comparateur dynamique, pour la réduction du biais de bit, est introduite en tant que composant dans une configuration de bascule aléatoire quantique. Cette méthode est comparée et mise en contraste avec une première méthode de génération de bits de comparaison d'arrivée de photons. Les conceptions sont mises à l'échelle des réseaux de capteurs SPAD, qui, lorsqu'ils sont combinés avec un éclairage externe, fonctionnent comme des conceptions QRNG gigabit par seconde. Le FortunaSPAD GNAQ est une conception à double interface, capable d'un débit de données de sortie combiné de 3,3 Gbps. Le détecteur FortunaSPAD2 GNAQ implémente une conception macro-pixel pour améliorer la robustesse contre la défaillance des pixels.

Mots clefs : Génération de nombres aléatoires, génération de nombres aléatoires quantiques, diode avalanche monophotonique, sécurité matérielle, technologie BCD, électronique des pixels, modélisation des temps morts, effets post-impulsion, flip-flop aléatoire quantique

Contents

List of Figures

List of Tables

List of Acronyms

ADPLL	All-Digital Phase-Locked Loop
AES	Advanced Encryption System
AMS	Analog Mixed-Signal
APP	Afterpulsing Probability
ASIC	Application-Specific Integrated Circuit
BCD	Bipolar CMOS DMOS
BIW	Barak/Impagliazzo/Wigderson
BNW	Buried N-Well
BSI	Backside Illumination
DAC	Digital-to-Analog Converter
DCR	Dark Count Rate
DI	Device-Independent
DNW	Deep N-Well
DPW	Deep P-Well
EaaS	Entropy-as-a-Service
FA	First Arrival
FF	Fill Factor
FoM	Figure of Merit
FPGA	Field-Programmable Gate Array
FSI	Frontside Illumination
FWHM	Full Width at Half Maximum
IC	Integrated Circuit
iid	Independent and Identically Distributed
IoT	Internet of Things
LED	Light-Emitting Diode
LET	Light Emission Testing
LVDS	Low-Voltage Differential Signalling
NIR	Near-Infrared
NIST	National Institute of Standards and Technology
NUV	Near Ultra-violet
PDE	Photon Detection Efficiency
PDF	Probability Density Function
PDP	Photon Detection Probability
PEB	Premature Edge Breakdown

PET	Positron Emissions Tomography
PLL	Phase-Locked Loop
PQAR	Passive-Quench Active Recharge
PRNG	Pseudo-Random Number Generator
PSD	Pyramid Surface for Diffraction
PVT	Process Voltage Temperature
QE	Quantum Efficiency
QKD	Quantum Key Distribution
QRFF	Quantum Random Flip-Flop
QRNG	Quantum Random Number Generator
RBG	Random Bit Generator
RFF	Random Flip-Flop
RNG	Random Number Generator
RSA	Rivest–Shamir–Adleman
RTS	Random Telegraph Signal
SC	Slow Clock
SoC	System-on-Chip
SOI	Silicon-on-Insulator
SP	Special Publication
SPAD	Single-Photon Avalanche Diode
SPTR	Single-Photon Timing Resolution
SPW	Shallow P-Well
STS	Statistical Test Suite
TCAD	Technology Computer Aided Design
TCSPC	Time-correlated single-photon counting
TDC	Time-to-Digital Converter
TRNG	True Random Number Generator
TSV	Through-Silicon Vias

List of Symbols

α	Afterpulsing probability
A_a	SPAD active area
a_j	Serial correlation coefficient with j-lag
b	Bias of bit string
C_{as}	Anode parasitic capacitance
C_j	Diode junction capacitance
C_{ks}	Cathode parasitic capacitance
C_L	Load capacitance
DR_s	SPAD dynamic range
$\lvert\overline{\mathrm{E}_d}\rvert$	Depletion region average electric field magnitude
E_g	Bandgap energy
E_s	SPAD energy consumed per detection
f	Frequency
f_{BG}	Bit generation frequency
F_c	Dynamic range correction factor
$\mathrm{G}_{\mathrm{btb}}$	Band-to-band generation
G_{therm}	Thermal carrier generation rate
h	Planck constant
H_1	Shannon entropy
H_∞	Min-entropy
I_s	Diode static current
k	Number of detections per period
k_{B}	Boltzmann constant
k_{max}	Maximum number of detections in an integration period
λ	Wavelength of light
λ_{A}	Photon arrival rate
λ_{D}	Photon detection rate
L_e	Electron diffusion length
L_h	Hole diffusion length
m	Mass
$\mu_d(\lambda)$	Photon penetration depth in silicon of light
n_i	Intrinsic doping level
$n_{t,i}$	i^{th}-trap
N_t	Recombination center density

η	Normalized threshold voltage level
η_c	Breakdown probability empirical correction factor
π	Pi
P_{av}	Junction avalanche probability
P_{ab}	Photon absorption probability
P_b	Junction breakdown probability
P_{diff}	Probability carrier reaches depleted region
P_e	Probability of electron-initiated avalanche
P_h	Probability of hole-initiated avalanche
P_r	Optical power
P_s	SPAD power consumed per detection
$P_{t,i}$	i^{th}-trap exponential pre-factor probability
q	Normalized charge
Q_t	Total charge
$\mathbb{R}$	Real numbers
R_d	Diode resistance
R_L	Load resistance
R_{XX}	Autocorrelation function of random variable, X
σ	Standard deviation
σ_0	Carrier capture cross-section
t	Time
T	Temperature
T_{BG}	Bit generation period
t_f	Fall time
t_r	Rise time
τ_b	Avalanche build-up time constant
τ_{dead}	Detector dead time
τ_{det}	Overall detection time constant
τ_p	Discriminator pulse time constant
τ_{par}	Paralysable dead time
τ_{sat}	Saturation dead time
τ_t	Carrier transit time constant
$\tau_{t,i}$	i^{th}-trap exponential decay time constant
τ_q	Quenching time constant
τ_r	Recharge time constant
θ_0	Angle of incidence
$T(\lambda, \theta_0)$	Transmittance function
V_{BR}	SPAD breakdown voltage
V_{EX}	SPAD excess bias voltage
V_{OP}	SPAD operating voltage
v_{th}	Thermal velocity
W_d	Depletion region width
X	General random variable
$\mathbb{Z}_0^+$	Positive integers

1 Random number generation: fundamentals, methods and systems

1.1 Preliminaries

The concept of randomness extends back several millennia, when dice-throwing and games of chance were used as methods for determining fate. Indeed, throughout human history, the many aspects of day-to-day life, which remain unpredictable, have greatly influenced our spirituality, medical practices, relationships, political systems and more. With the dedicated analysis of randomness starting in the early twentieth century, which led to the eventual formulation of mathematical axioms, randomness has now become a fundamental aspect of our scientific pursuits. In the computer sciences, randomness first stimulated the creation of developments in how data is quantified, stored, and transmitted, which is now classified as information theory.

More unexpected developments followed with the discovery of algorithmic randomness. Here, researchers showed that the introduction of a random sequence could improve an algorithm's ability to perform certain computations. For physicists, probability and stochastic theory are a fundamental part of the lexicon used to develop quantum mechanics. The advent of quantum science has not only shaped our perception of reality, but has also shifted our views on what we deem to be a scientifically conceivable system versus those that remain, for the time being, in the realm of science fiction.

It is this very ability to exploit quantum phenomena in order to engineer potentially revolutionary technologies that have given birth to the relatively new field of Quantum Engineering. A quantized particle of light, known as a photon, is of particular interest in this space, owing to the development, in the last half-century, of integrated circuit (**IC**) technologies that can readily detect and manipulate light. Finally, as the world has grown unfathomably connected through information technologies, the role of randomness in the *secure* storage and transmission of data has taken on a critical role in the systems we deploy.

It is at this intersection between randomness, optical quantum technologies, and integrated circuit design that this thesis is situated. It presents the development, modeling and design of the monolithic single-photon avalanche diode (**SPAD**)-based quantum random number

generators (**QRNG**s). Any of these disciplines merits a study extending far past the scope of this thesis. Nevertheless, an attempt is made to adequately cover the topics required to explain and justify the research conducted. Therefore, the remainder of this chapter attempts to present the necessary principles of randomness and information theory relevant to the development, modeling and design of the monolithic single-photon avalanche diode (SPAD)-based quantum random number generators (QRNGs) presented in the following chapters.

By virtue of the definition of non-determinism, the difficulty of verifying that any given bit string is perfectly random, remains a virtually impossible task. It is only possible to compare the statistics of such a sting to those which would be expected from one produced by a perfect random number generator. For this, and other reasons, a fierce debate continues on the classification, utility of and requirements for, a random number generator (**RNG**). Thus, any analysis on *how-random* typically starts with a discussion on entropy, which quantifies the uncertainty contained in any random variable.

1.1.1 Entropy

The formulation of information entropy (H_1) was introduced by Claude Shannon in 1948 [1] and is shown in Equation (1.1).

$$\mathrm{H}_1(X) = -\sum_{i=1}^{n} p_i \log p_i \tag{1.1}$$

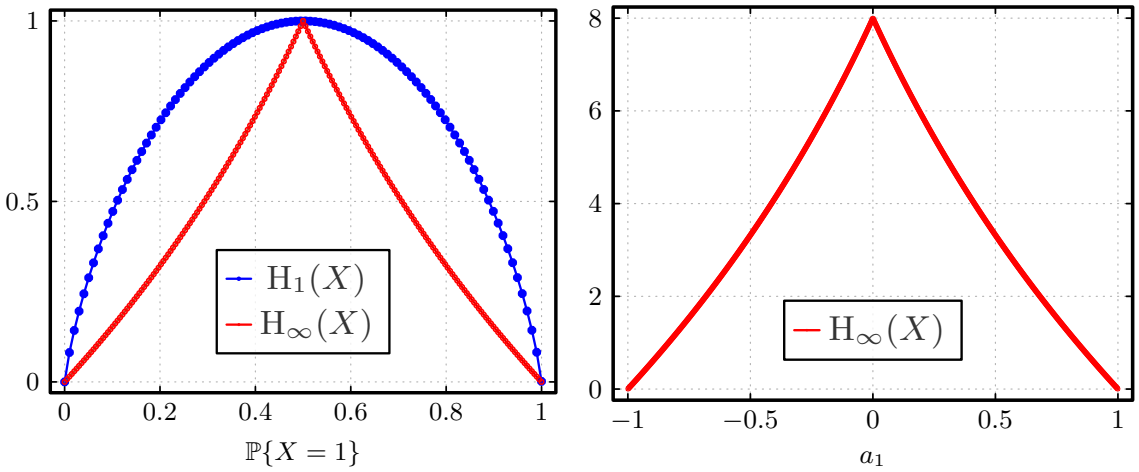
Here, X denotes a random variable, the basis of which also determines the log basis, and p_i is the probability of occurrence of the event. Therefore, for a typical binary computing system, the basis chosen is 2 and X becomes a discrete random variable representing the probability distribution of a generated bit string. The summation denotes the accumulation over all the random variable's possible outcomes. Although this equation appear highly abstract, it quantifies the predictability of outcomes and therefore presents a mathematical limit on how well data, with a given entropy, can be encoded onto a noiseless channel. Several decades later, the Hungarian mathematician, Alfréd Rényi generalized several methods for calculating the information contained within a sequence, into a family of entropies known as Rényi entropy [2].

$$\mathrm{H}_\alpha(X) = \frac{1}{1-\alpha} \cdot \log\left(\sum_{i=1}^{n} p_i^\alpha\right) \tag{1.2}$$

An instance of Equation (1.2), which corresponds to the most conservative measure of the unpredictability of outcomes, is when $\alpha \to 0$. This form is known as *min*-entropy, H_∞.

$$\mathrm{H}_\infty(X) = \log \max(p_i) \tag{1.3}$$

From a practical perspective, these functions remain fairly abstract, particularly in their utility for analyzing the performance of a random number generator. To elucidate these concepts



(a) Shannon and min entropies of a biased generator

(b) Min entropy of serially correlated bits in an 8-bit symbol

Figure 1.1: Shannon and min entropy plots for a binary system in the biased and serially correlated cases. Clearly, to ensure that a certain entropy requirement is met, min entropy places more stringent bounds.The y-axis shows denotes the entropy.

within the context of this research, the equations need to be evaluated using the statistical properties of a random number produced by a RNG. First, the bias in a binary system is defined by Equation (1.4) and describes the non-uniformity in the distribution of ones and zeroes.

$$b = \frac{p_1 - p_0}{2} \tag{1.4}$$

A bit string can be defined as a set, $\mathcal{X}$, with binary bits $x \in \mathcal{X}$ and bias, b, where $p_1 = \mathbb{P}\{x = 1\}$. The Shannon entropy of $\mathcal{X}$-valued random variable, X, is calculated simply using Equation (1.5).

$$\mathrm{H}_1(X) = -\sum_{i=1}^{n} p_i \log_2 p_i = -(p_1 \cdot \log_2(p_1) + (1 - p_1) \cdot \log_2(1 - p_1)) \tag{1.5}$$

The evaluation of min entropy of a physical RNG can be performed in a similar manner for a biased binary system using Equation (1.6).

$$\mathrm{H}_\infty(X) = \begin{cases} -\log_2(1 - p_1), & p_1 < 0.5 \\ -\log_2(p_1), & p_1 > 0.5 \end{cases} \tag{1.6}$$

The Shannon and min entropies are plotted in Figure 1.1a. It can be observed that achieving any given entropy requirement, requires a more stringent bound on bit bias when using the min entropy calculation. The relevance of this distinction is discussed briefly in the security parameter section.

While bias is one source of degradation to entropy, correlation is another. The autocorrelation

function of data generated from a time-sampled random variable, X_t, with j-lag is defined by Equation (1.7).

$$a_j = \frac{Cov(X_t, X_{t-j})}{\sqrt{\sigma^2(X_t)\sigma^2(X_{t-j})}} \tag{1.7}$$

There exists, in principle, numerous lag j values corresponding to the length of the bit string that is being evaluated. This presents the challenge of determining the relevant values for consideration. Typically, only lower order lags and in particular $j = 1$ are sufficient for evaluating random numbers produced using systems which exploit photon-timing statistics. The justification for this will be discussed in subsequent chapters, but for now we move forward with examples using $j = 1$. Therefore, Equation (1.7) can be, for a length of N bits, explicitly evaluated with Equation (1.8).

$$a_1 = \frac{N\left(\Sigma_{i=0}^{N-1} x_i x_{i+1}\right) - \left(\Sigma_{i=0}^{N-1} x_i^2\right)}{\left(N\Sigma_{i=0}^{N-1} x_i^2\right) - \left(\Sigma_{i=0}^{N-1} x_i\right)^2} \tag{1.8}$$

The process of calculating the correlation coefficient requires an additional padded bit for the shifted string. A zero padding method (zeros augmented onto the end) can be implemented, however, it is typical to wrap the first value in the set, as is done by the popular program ENT battery of tests [3].

Commonly, an RNG can be used for generation of a *key*, which can be used for a variety of security functions, such as encryption. We can further investigate the effect correlation can have on entropy by evaluating Equation (1.3) for keys generated by an RNG. For example, if a generator produces an n-bit key, Z, then the min entropy is simply the logarithm of the probability of occurrence for the *symbol*, z_i, $\quad i \in N, \quad N = 2^n$, which occurs with the highest probability in the family of symbols $\mathcal{Z} = \{z_0, z_1, ..., z_{n-1}\}$. Given an unbiased bit stream with a serial correlation value of a_1, we can define the probability of a subsequent bit x_{i+1} being the same as the previous x_i with Equation (1.9).

$$\mathbb{P}\{x_{i+1} = x_i\} = \frac{a_1}{2} + 0.5 \tag{1.9}$$

Furthermore, for an n-bit symbol, the most probable symbol values can be calculated, for both positive and negative a_j values, with Equation (1.10).

$$P_{z,\max} = \mathbf{max}\left(\left(\frac{a_1}{2} + 0.5\right), 1 - \left(\frac{a_1}{2} + 0.5\right)\right)^n \tag{1.10}$$

Min entropy calculations for serially correlated data are plotted using Equation (1.10) and shown in Figure 1.1b.

1.1.2 Keys in a cryptographic system

The significance of entropic properties can be illustrated with a simple practical example. First, imagine a hypothetical system containing a cryptographic algorithm that has no exploitable vulnerability. In such a situation, an adversary would need to resort to brute force measures to guess the key. Generally speaking, the computational effort entailed in such an endeavor has an exponential relationship with the effective key length, defined as keylength · entropy [4].

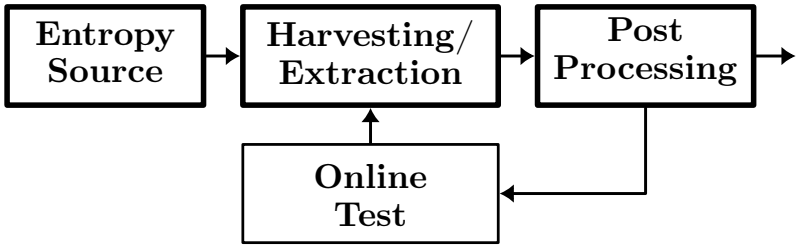
$$\text{key space size} = 2^{\text{keylength}\cdot\text{per-bit-entropy}} \tag{1.11}$$

Private key cryptography such as the Advanced Encryption Standard (**AES**) standard is a representative example for evaluating the requirements for the effective key length. Typically, the resources required to perform a brute force attack on a system with an effective key length of 128 bits is in the order of low millions of dollars [4]. As a rule of thumb, if the effective key length is entropy degraded by less than 1 bit, a 256-bit key is deemed to be sufficiently secure for most, if not all, *contemporary* applications, given the extremely high resource requirement for guessing the key. More detail on cryptographic keys and encryption systems is discussed in one of the following sections. Equation (1.5) can be used to calculate the acceptable bias in a system requiring a 1-bit entropy degradation on a 256-bit key. If the most conservative estimate for entropy is required, then Equations (1.6) and (1.10) are used to determine acceptable bias and serial correlation of a generated string. In the example of a 256-bit key with an effective key length keylength · entropy = 255 bits, the corresponding entropy is ≈ 0.996. Therefore, in the case of a biased generator, this corresponds to an acceptable bias, $b \approx 0.035$ when using Shannon entropy, and $b \approx 0.001$, in the min entropy case. The acceptable serial correlation in this example is calculated as $|a_1| \approx 0.0025$.

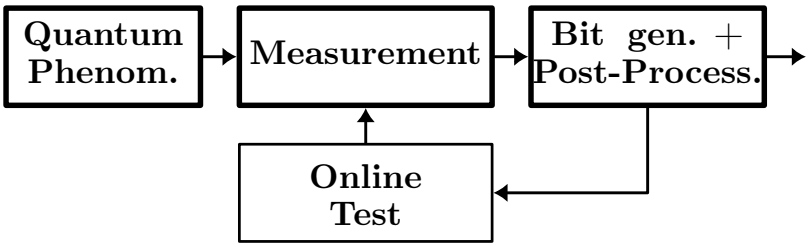
This simple example demonstrates a method for performing quick analysis on bit strings to test a random number generator's performance. However, it is far too rudimentary to verify, let alone certify, a physical RNG design. From a statistical analysis point of view, a more rigorous test suite is required to compare the bits produced by a RNG against one which would possess the statistical properties of a perfect RNG. The certification of RNGs is an even more expansive topic. In this thesis, rigorous statistical testing of the QRNG designs is conducted. However, as a way of providing quick feedback when performing modelling of the detectors, circuits and systems used for random bit generation, bias and serial correlation are calculated. Furthermore, entropy of the generated bits are benchmarked against those outlined by the upcoming revision to the AIS-31 standard. In practice, any physical RNG needs a source of entropy, and a method(s) for *extracting* randomness from the source to produce a stream of bits.

1.2 Taxonomy of generator designs

Figure 1.2a shows the basic block diagrams highlighting the components in a typical random number generator along with a QRNG. The entropy source is the physical process from which the random behavior derives. Every physical random number generator requires this



(a) Block diagram of a typical RNG. Some physical source of entropy is harvested and conditioned algorithmically to generate random bits.



(b) QRNG general structure. A measurement is performed on the quantum entropy source. A variety of techniques are used to the generate random bit strings.

Figure 1.2: Anatomy of a random number generator and the slight differences between the general structures of typical RNGs and QRNGs.

component. Bit generation, extraction and post-processing stages are solution specific, and can consist of a simple bit generation circuit or a complex, multi-stage block. Requirements for these stages remain fiercely debated, and are influenced heavily by the entropy source chosen. Sources can be broadly classified as either classical or quantum.

A classification chart specifying the many classes of RNG types is shown in Figure 1.3 Pseudo-random number generators (**PRNG**s), aptly named in that they only mimic the statistical properties of randomness, use algorithms for generation of bits from an entropy source. True random number generators (**TRNG**s) exploit a classical non-deterministic entropy source, such as electrical noise, that can often comprise multiple physical phenomena. Evidently, a QRNG uses randomness extracted from a discrete quantized source that is fundamentally random in nature.

1.2.1 Pseudo-random number generators

Pseudorandom numbers highlight an omnipresent problem in hardware security systems: distinguishing entropy from pseudorandomness can be very difficult, if not impossible. That is to say, any given bit stream may possess the statistical properties expected from an ideal source, even if the generator itself is highly flawed. In general terms, PRNGs can be classified into those which are/are not cryptographically secure. Simple generators are often implemented in software tools for operations requiring an approximate uniform random variable. An example of this is the linear congruential generator.

More complex algorithms, such as those outlined in the SP800-90A standard, implement

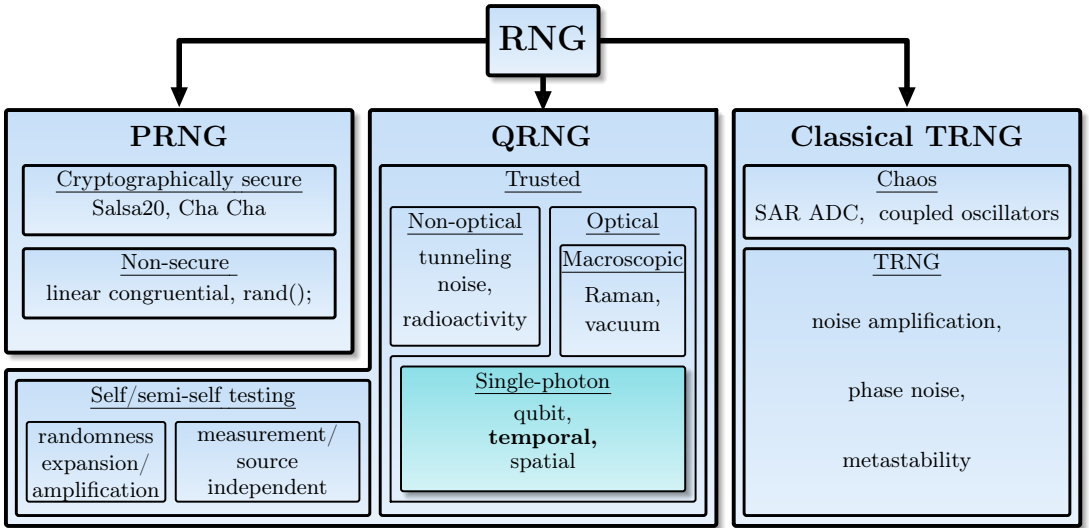


Figure 1.3: Classification chart of random number generators. This thesis focuses on techniques based on photon-timing statistics.

hash functions. This standard was first published by National Institute of Standards and Technology (NIST) in 2006 [5]. Many revisions have been made, stemming from flaws with the generator designs highlighted by industry and academia. Furthermore, a back door within the algorithm was revealed [6], [7]. Cryptanalysis has been performed on many PRNGs. In [8], a cryptanalysis of Salsa20 showed it could break 8 out of 20 rounds to recover the 256-bit secret key. A summary of cryptanalyses of PRNGs is provided in [9].

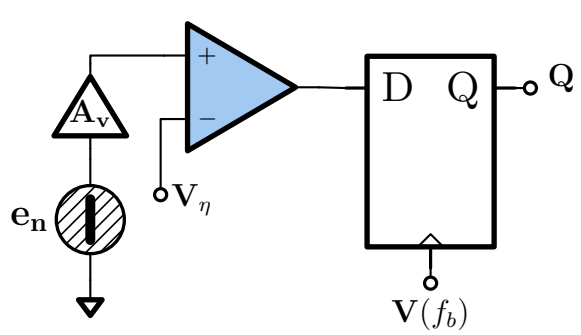
For an observer who is in possession of the key used within the PRNG design, the output stream can be predictable.

1.2.2 Classical TRNGs and chaotic maps

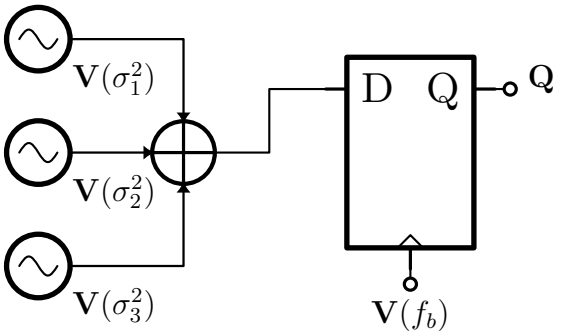
The proliferation of microelectronic devices has necessitated the co-integration of hardware random number generators. Therefore, CMOS based microelectronic devices typically exploit easily producible random behaviors of discrete silicon devices and circuits to generate random numbers.

The simplest classical true random number generator consists of an amplified noisy source with a clocked comparator, as shown in Figure 1.4a. The operation of this design is simple. Noise from a semiconductor device ($\mathbf{e}_n$), such as a resistor or diode, is first amplified and then a comparator is used to generate a digital signal with random edge transitions. Sampling this signal produces a random bit. The simplicity of the device comes with significant performance flaws. In general, noisy sources are highly unstable. Moreover, to reduce bias, the noise must be heavily amplified to overcome the mismatch of the comparator, which increases power consumption [10]. They are also sensitive to surrounding electromagnetic (EM) noise/interference, which can cause correlations between neighboring generators [11].

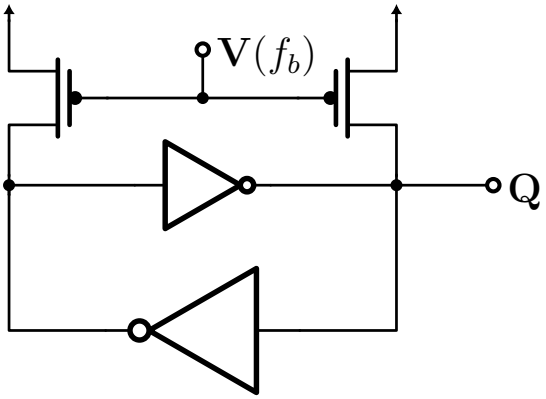
Metastability-based random number generators offer a more reliable, less power hungry solution.



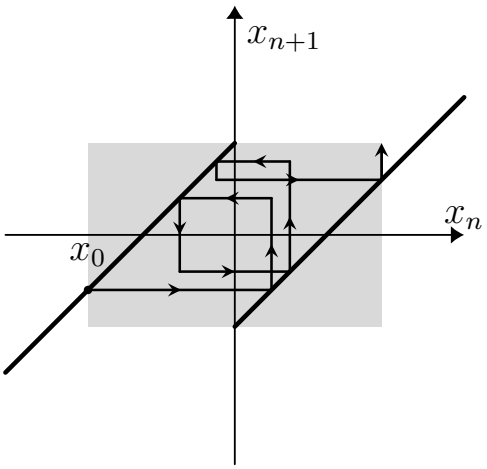
(a) A noise source, $\mathbf{e_n}$, is amplified then digitized by a comparator with threshold $\mathbf{V}_\eta$. A random bit is generated by the clock signal, $\mathbf{V}(f_b)$.



(b) Jitter based TRNG example with multiple ring oscillators XOR'd together. A random bit is generated by the clock signal, $\mathbf{V}(f_b)$.



(c) A simple latch-based circuit realization of a metastability based RNG. Noise forces the output to an unpredictable state. Mismatch between inverters (PVT), creates bias.



(d) Illustration of a chaotic map. The initial state is denoted as x_0. The next state x_{n+1} is some random evolution of the previous x_n.

Figure 1.4: Examples showing each of the four fundamental classes of TRNG designs.

Metastability is a well-defined phenomenon which describes the sensitivity of circuits close to their latching threshold. A clocked bistable latch, such as the one shown in Figure 1.4c will fall to an unpredictable state. A slight voltage offset caused by the inherent electronic noise present inside any semiconductor device causes the latch to settle to either position. TRNG implementations based on this principle remain popular, as they have shown the ability to operate in the low gigabit-per-second (Gbps) range while achieving low or even sub pJ/bit consumption [12]–[18]. However, their popularity is declining as higher mismatch in reduced technology nodes increases the bias of these generators, which increases the level of post-processing required to meet acceptable entropy bounds [19].

Another popular technique is the exploitation of phase noise or jitter in CMOS oscillator designs. Fluctuations in the period of an oscillator are caused by a combination of thermal, flicker and shot noise from devices [20]. A drawback of this method is robustness, as frequency can easily change due to small variations in process, voltage and temperature (**PVT**). Methods

have been suggested for alleviating degradation, such as XORing multiple unit cells of jitter based random bit generators to create a single noise source [21]. A schematic illustration is shown in Figure 1.4b. Furthermore, these generator designs typically have lower throughput compared to their metastability counterparts.

The fourth fundamental class of TRNGs is the chaotic map. The theory of chaos in non-linear dynamics claims that low-dimensional dynamic systems contain unpredictable behavior [22]. A chaotic system that is deterministic in microscopic space, becomes random in macroscopic space when an initial condition is applied (such as electronic noise). This can allow for representative modelling of physical random number generators that implement this technique. Early analog implementations used switched current circuits to demonstrate a piece-wise linear map with unpredictable output current value [23]–[25]. In Figure 1.4d, the random trajectory and output of a variable, x_n, which can represent a voltage or current, is shown. More modern implementations of chaos-based TRNGs are complex systems that are hybrid in method. Examples include the non-linear behavior of capacitively coupled ring oscillators [26] and data-converters, which use the residue functions to create a linear map [25].

Monolithic TRNGs have undergone two plus decades of research and development and have recently shown promising results in terms of throughput. However, they are still fundamentally limited by the determinism induced by the entropy sources. As technology has scaled, it has allowed for lower power and faster electronics. This advantage is offset by the fact that technology scaling also increases mismatch issues, manifesting itself in bias [4]. To combat this effect, virtually all credible TRNG implementations require a significant amount of post-processing in order to de-bias and remove correlations. These complex systems then become reliant on extraction and de-biasing methods. Furthermore, these de-biasing techniques rely on one-way functions that are used in PRNGs, which have been shown to be victims of security flaws. For these reasons, QRNGs have garnered interest. There is agreement in the scientific community that quantum entropy sources possess characteristics that are fundamentally random. Therefore, it is proposed that QRNGs can alleviate many of the challenges in performance, speed and robustness posed by classical random number generators.

1.2.3 Quantum random number generators

The promise of quantum entropy sources, along with the ever-increasing demand for more robust hardware security primitives, has led to an explosion of research into QRNG methods and systems. As outlined by Figure 1.3, there are numerous demonstrated QRNG types. A general review of QRNG techniques and publications is provided in [27], with more recent developments highlighted by [28]. The review provided in [29] summarizes well the state-of-the-art for so-called trusted optical QRNGs designed specifically in silicon-photonic technologies. A dissection of the relevant QRNG classes outlined by Figure 1.3 is provided here.

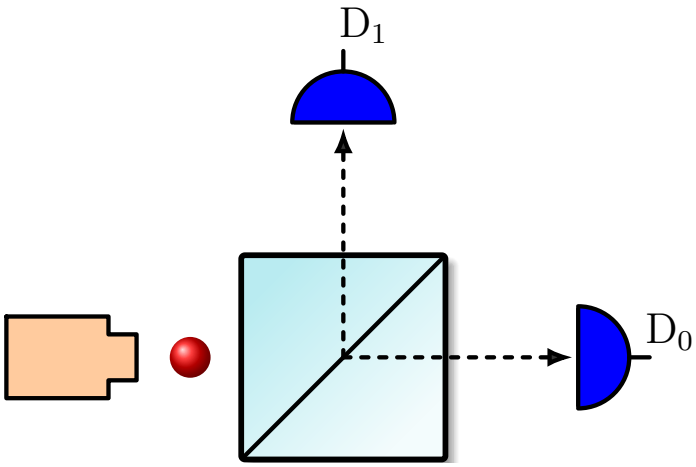


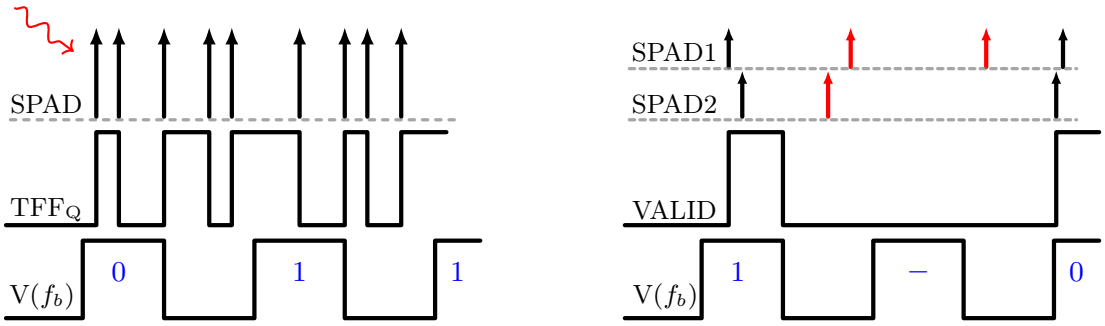
Figure 1.5: Archetype QRNG implementation consisting of a photon source, a beam splitter and two detectors. When a detection arrives at D_0, the output bit is 0, and a 1 when detected by D1. The uncertainty of the photon path produces the random behavior. In [38], it was shown that only afterpulsing, twilight events and dead time of detectors cause correlations in the beam splitter-based QRNG.

Self-testing QRNGs

Testing, in some form, is required for all practical QRNGs. Performance can be determined using simple online health monitoring, where a subset of generated bits is continuously tested against the entropy benchmarks. However, results that more closely correspond to the physical nature of the source can also be implemented. For example, in a Geiger-mode based system, where photons are counted, the time of arrival distribution can be validated against a Poisson distribution [30]. One of the remarkable aspects of quantum theory is that certain measurement outcomes are unpredictable through observation. It is with this principle that a new class of self-testing QRNG designs, based on an *untrusted device*, has emerged. These realizations are referred to as device-independent (**DI**)QRNGs [31]. The aim of these generators is to guarantee randomness by performing Bell tests [32]. An interesting concept in the research domain, they remain completely impractical for applications requiring small form-factor as they require large, bulky setups and only generate bits on the order of kbps or low Mbps [33]. To provide an alternative to the paranoid DI approach, where the aspects of the device provided by a vendor are nefarious or non-functional, an alternative known as source/measurement-independent QRNGs have been proposed [34], [35]. In these cases either a trusted source or measurement device is used in the generator design with achievable rates in the Gbps range [36], [37]. Nevertheless, these QRNG architectures are not considered in this work as they require complex optical setups not conducive for a monolithic approach.

Trusted QRNGs

Assuming that the devices, such as photodetectors, provided by a vendor are functional and non-malicious in nature, many practical QRNG designs can be proposed. The archetypal example used to demonstrate a photon detection-based QRNG is shown in Figure 1.5. Within



(a) The slow clock method. Random events are detected which in turn cause a toggle ($\mathrm{TFF_Q}$). Some sampling clock $\mathrm{V}(f_b)$ samples the state of the TFF and generates a corresponding bit. Conversely the sampling time can be controlled by the random event to realize the fast clock method.

(b) First arrival bit generation principle. Two detectors are used in this example for comparison. Although a single detector comparing the time between detections can also be implemented. The VALID bit represents the probability of arrival method, another way to generate a random bit.

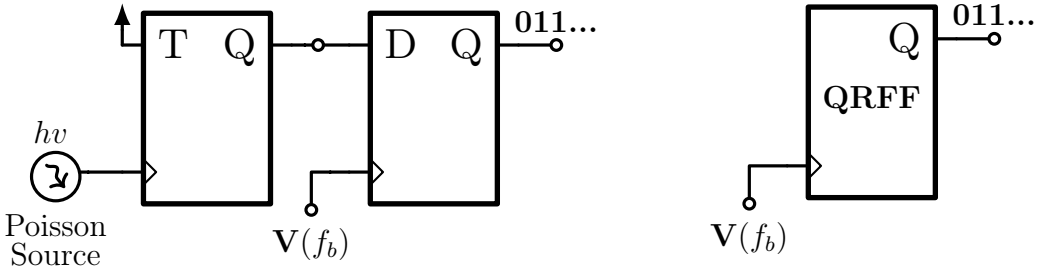
Figure 1.6: Timing diagrams demonstrating random bit generation techniques based on photon-timing statistics. SPADs are assumed to be ideal detectors in this simple illustration.

this configuration, a light source emits photons that pass through a beam splitter, which are then incident on single-photon sensitive detectors. Several variations of this setup exit. In the simplest form, classical light is sent through a balanced beam splitter with equal transmission and reflection, so that light is split into streams of equal optical power. Detections arriving at $\mathrm{D_0}$ produce a logic 0, while those at $\mathrm{D_1}$ produce a 1 [39]. Similar results can be obtained using polarized photons and beam splitters.

Perhaps, owing to their versatility, the most interesting QRNG designs exploit photon-timing statistics. Integrated circuits in modern deep sub-micron CMOS processes allow for picosecond order logic transitions. Therefore, high-performance single-photon sensitive detectors, such as SPADs, can be co-integrated with many different time sensing and conversion functions. A highly representative example is a chronometer circuit, such as the time-to-digital converter (**TDC**). Modern TDCs are capable of resolving events with a timing resolution in the order of 10 picoseconds. Scaling of SPADs and circuits has allowed for pixel-wise integration of time-based circuits. These advancements have given researchers and designers tremendous flexibility to exploit the random arrival times of photons.

Random flip-flop concept

Several conceptual bit generation techniques based on random event arrivals are illustrated in Figure 1.6. The first example shown in Figure 1.6a, known as the slow clock method, is an analogue to the digitized noise based TRNG. In this configuration, a random event arrival produces a toggle, creating a waveform that has edges occurring at random times. This stochastic process is referred to as a random telegraph signal (**RTS**) process. A strobe clock, $\mathbf{V}(f_b)$, that samples the RTS then produces a random bit, at a rate of f_{BG}. Figure 1.7a is a simple circuit configuration that achieves this functionality. This circuit is a realization of the



(a) A realization of a random flip-flop.

(b) Abstract circuit symbol of a QRFF.

Figure 1.7: The quantum random flip-flop (QRFF) concept along with a practical realization, consisting of an ideal source with Poisson-distributed arrival statistics, and two conventional flip-flops.

random flip-flop (**RFF**) concept, introduced in [40]. The RFF abstraction describes a random bit generation circuit capable of produces a single random bit upon the arrival of a clock signal, as shown in Figure 1.7b. The fast clock method applies the same concept in reverse, such that the random event arrival is the strobe that samples a periodic oscillation. Examining the timing characteristics of two separate events can also be employed as a method for generating a random bit. For example, a comparison of the time-difference between consecutive events is used to determine the generated bit. If two detection devices are available, this concept can be tweaked for a similar effect; the bit is then determined by the first of two independent detectors to receive a photon, as displayed in Figure 1.6b. Finally, for a given photon flux, a precise detection window can be defined so that a detection occurs with $p = 0.5$. Various circuit realizations of these concepts have been demonstrated. Each method comes with its own advantages and challenges in terms of performance, along with practical limitations. Analysis, design and characterization of circuits that implement these concepts, with particular emphasis on QRFFs based on the slow clock method, is explored in Chapter 4.

1.3 Extraction methods, and post-processing algorithms

The terms entropy extraction and post-processing are often used interchangeably to describe the conditioning of biased and/or correlated bitstrings to achieve *near-full entropy output data.* In RNG designs, some level of post-processing, is often necessary. This is due to circuit or detector imperfections that can augment classical noise. The general principle of entropy extraction is simple. Given a weak entropy source, i.e. one that does not contain perfect entropy, it is possible to implement circuits or algorithms that produces data with improved entropy-per-bit. In 1990 McInnes and Pikas proved that you cannot use a single non-ideal entropy sources to create a improved entropy output [41]. Therefore, multiple sources are needed for robust random number generation. The cost for this is a reduced output data rate compared to the input. Common methods used in RNG designs are described below.

XOR

The simplest extraction circuit is the XOR gate. The utility of this circuit is demonstrated with a simple example. Two independent entropy sources A and B each with bias, b, are inputs into an XOR gate $C = A \oplus B$. The output bias of C will be b^2 [42]. This technique is also useful for reducing serial correlations, but only if there is no cross-correlation present between the two entropy sources. While certainly not a robust enough solution to be standalone, XOR can be a useful tool in RNG design where slightly biased independent bits are plentiful. This will become relevant when, the FortunaSPAD2, an improved SPAD array architecture for random bit generation, is presented in Chapter 6.

Von Neumann whitening

This algorithm is used to take biased independent bits and turn them into a smaller, unbiased sequence. It works as follows. A m-length sequence of bits Z_m is first divided into $n = m/2$ pairs of bits X_n. All pairs which are 00 and 11 are discarded. Then, the second bit of each pair is taken as the final output. An obvious drawback of this technique is that the output data rate is not well-defined, seeing as the initial number of 11 and 00 pairs is unknown. Moreover, the properties of the algorithm require input data to have zero serial correlation. This is not always a practical requirement from entropy sources.

Yuval Peres algorithm

The Yuval Peres algorithm proposed a method for feeding back the unused bits in the Von Neumann method [43]. The recursive function works as follows. First, the function $\Psi_1(x_0, ..., x_{n-1})$ is defined as the Von Neumann de-biased bits. Then another calculation is performed as, $\Psi_u(u_0, ..., u_{n-1}/2)$, where u_i is the XOR-ing of pairs of bits with x_n. Finally, $\Psi_v(v_0, ..., v_{n-1}/2)$ is defined as the selection of a single-bit selection from the discarded pairs. Recursively iterating this algorithm until no discarded bits remain reaches the Shannon limit i.e. full Shannon entropy.

Multiple input extractors

Many involved algorithms have emerged to overcome the limitations of the previously described methods. Some examples include the 2-EXT [44], Barak/Impagliazzo/Wigderson (**BIW**) [45], and the Trevisan extractors [46]. These architectures utilize a variety of mathematical principles such as blenders, hash functions and Galois Fields. They have proved useful for turning weakly random digital noise sources in computers into efficient high-entropy RNG designs. However, they are not appropriate for QRNG applications. Implementing complex, on-chip, mathematical functions, to improve the output data entropy, arguably defeats the purpose of using quantum entropy sources.

This work proposes, and later demonstrates, that a properly implemented QRNG design **can**

require only minimal post-processing. Practically, post-processing is largely used to mitigate against present, but statistically outlying physical phenomena, the effectiveness of an adversary to attack the device, or changes in environment that can cause entropy degradation during normal operation.

1.4 RNGs in contemporary and future security systems

There are many applications that involve high-quality random numbers, ranging from scientific methods requiring Monte Carlo simulations to lottery machines. However, the most prevalent application, and one of great contemporary concern, is the secure storage and transmission of digital data. Hardware security is a large, multidisciplinary topic, covering aspects of computer science, engineering and physics. It has become a crucial element of system-on-chip (**SoC**) design given the proliferation of connected devices. At a general level, systems can be divided into private (symmetric) or public (asymmetric) key cryptography. In a public key cryptographic systems, a shared publicly-available key is used to encrypt data while a private key is used to decrypt, whereas, in the private key system, there exists only a single key for both functions. Therefore, in a private key based system, a method of secure *key distribution* is required. This is known as the *key distribution problem*. Public key cryptography's simple solution to this challenge has buoyed it to be the method of choice for many modern encryption systems. However, with the advent of quantum computing, the requirements for robust RNG designs have been heightened, and the ability of public key algorithms to address security requirements has been placed in peril [47], [48].

In 1994 Peter Shor proved mathematically that a quantum computer could, in principle, break public key based security protocols [49]. An example of such an encryption protocol, in wide use today, is the **RSA** cryptosystem. As a result, it is commonly believed that so-called *bad actors* are currently harvesting encrypted data to be decrypted, at a future date, upon the availability of quantum computers. Indeed, this has increased urgency and demand for novel cryptographic protocols. Soon after, Grover proposed an algorithm that can be used to speed up the process of finding the private key in a symmetric key encryption system, such as AES [50]. AES is a widely used algorithm for encrypting blocks of data at a time (128, 192, or 256 bits) and is designed to be resistant to attacks from classical computers. However, unless the key size is increased to 256 bits, it is vulnerable to attacks from quantum computers [47], [51]. Needless to say, given the widespread risk posed to many existing encryption algorithms, an international effort to mitigate threats by creation of so-called *post-quantum cryptographic systems*, which are resilient against attacks from quantum computers, has been called for. In 2016, NIST solicited submissions of post-quantum cryptographic algorithms [48], [52]. Four quantum-resistant algorithms from these submissions were announced in July 2022.

Alternatively, a different approach, is the use of one-time pads. A one-time pad is the only theoretically proven unbreakable cipher [53]. This simple approach involves a modulo addition (XOR) of data with the random key. If the key is perfectly random, then the encrypted data is unbreakable. The challenge is that the key must be at least equal in length to the data. This principle is shown in Figure 1.8. A new class of enterprise, known as **Entropy-as-a-service**

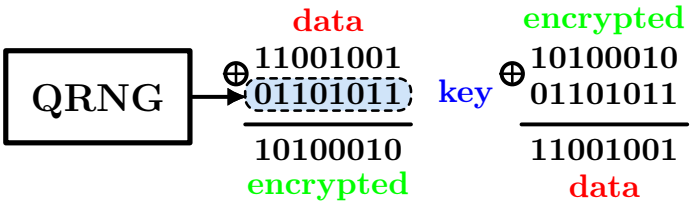


Figure 1.8: Demonstration of the one-time pad encryption method. When a modulo addition (XOR) is performed with a perfectly random key, the result is a provably secure cipher-text, assuming an adversary does not possess the key.

(**EaaS**), has risen to meet the demand for random bits [54]. EaaS aims to provide a framework where keys can be securely generated by a client, from a secure server, without the server being able to gain information about the key. Quantum key distribution (**QKD**), is another breakthrough technology which makes use of QRNGs, promising a method for secure delivery of generated keys [55]. Internet-of-things (**IoT**) devices manufacturers have also started to consider these implications. In 2021 the Samsung became the first smartphone manufacturer to place a QRNG within one of their phones. Finally, it is worth noting that the research on post-quantum cryptography is constantly evolving, and new developments are being published regularly. Regardless of the system solutions implemented to overcome contemporary and future security threats, high-speed, scalable QRNG designs are an attractive solution for meeting the increased key-length requirement.

Standardization

In 2012, **NIST** drafted the Special Publication (**SP**) 800-90B, *Recommendation for the Entropy Sources Used for Random Bit Generation*, with the aim of moving towards a framework for random number generator standardization [56]. This document came with a test suite for determining if samples of a random number generator were independent and identically distributed (**iid**) and various methods for estimating min entropy. After a period of review and feedback, a final version was published in 2018. Previously, the NIST statistical test suite (**STS**) (SP-800 22 *A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications*) was used for characterizing entropy and overall benchmarking of random number generators. Since its original release in 2010, there have been many works highlighting its flaws, limitations and bugs [57]–[61]. In response to the scrutiny, NIST announced in April 2022 that they plan to update the 800-22 statistical test suite. Moreover, the development of SP 800-90C [62], *Recommendation for Random Bit Generator (**RBG**) Constructions* is in DRAFT status, allowing for a period for public comments. Furthermore, a new version of the German Federal Office for Information Security's AIS 31 standard, *A Proposal for Functionality Classes for Random Number Generators,* is under development [63]. This standard requires RNGs to fulfill the entropy requirements $H_1 = 0.9998$ and $H_\infty = 0.98$ per bit, using a sample size of N = 10 000 000 bits. The draft AIS standard can be accessed with this link.

1.5 Thesis goals and contributions

This thesis asserts that many aspects of random number generation, such as hardware security, cryptography, privacy amplification, etc., are outside the scope of work. A QRNG's utility is indeed part of a much wider interdisciplinary study on the application. The focus of research is rather on the modelling, design, development, and characterization through measurement of monolithic SPAD-based QRNG designs capable of Gbps operation. Therefore, the goals of this thesis are as follows:

1. Study the advantages and limitations of silicon SPAD sensors, both from the detector and circuit perspectives, as a method from which to extract entropy.
2. Develop novel detectors and circuits for quantum random bit generation, which combat the degradation in entropy caused by classical characteristics of the system.
3. Develop models that accurately predict the performance of proposed solutions based on detector and system imperfections.
4. Scale bit generation techniques to full system-on-chip arrays capable of gigabit/s operation

As a result from this research, the following contributions are summarized.

The **first contribution** is the design and characterization of four separate SPAD detectors in a 55 nm bipolar-CMOS-DMOS (**BCD**) process. The author shared equally in the development of this work with Francesco Gramuglia and Ekin Kizilkan. The author was responsible for TCAD simulations, the implementation of the designs (GDS), and partook in measurements. The pixel circuits used to characterize the detectors were also designed by the author. The results show excellent performance commensurate with the state-of-the-art for deep sub-micron detectors. Chapter 3 elaborates this work.

Rigorous analytical modelling and simulation, including a new way to predict correlation of randomly generated bits for detectors when taking into account dead time and afterpulsing (**second contribution**), was conducted. This modelling also led to the **third contribution**, which was the design and validation, in silicon, of a novel clocked comparator-based random bit generator **QRFF**, capable of eliminating bias in a single-pixel. The author was responsible for the modelling, simulation, implementation of the designs in silicon along with measurement. This work is detailed in Chapter 4.

The **fourth and final contribution** are two **ASIC**s. To the best of the author's knowledge, the first ASIC, eclipses the fastest data rates previously published for SPAD-based QRNG designs which integrate the entropy extraction/bit generation on-chip. They have been Christened as the FortunaSPAD, and it's successor, the FortunaSPAD2. The FortunaSPAD is a dual-interface 40×70 pixel QRNG design demonstrating 3.3 Gbps data rate, when illuminated with an external LED, and the highest per-pixel data rate published to date. Further developments showed that a macro-pixel design containing four QRFF instances XOR'd together could produce a highly robust random bit generator with negligible correlation and bias. This new macro pixel, which is resilient against pixel failure, was then scaled to a full 64×64 SPAD array (the FortunaSPAD2). Furthermore, the design of a version of the

FortunaSPAD2 QRNG, which includes integrated micro-LEDs, is also presented. All QRNG systems were designed and characterized, in their entirety, by the author. These two chips are discussed in Chapters 5 and 6, respectively.

1.6 Thesis organization

The following chapter details silicon SPAD devices in CMOS. The history of development along with recent advances are described. An emphasis is placed on detector and array characteristics that are relevant for random number generation. In Chapter 3, three distinct 55 nm SPAD detectors that were designed, fabricated and measured, are presented. Each detector demonstrates adequate noise performance for QRNG applications. A method for dramatically improving the sensitivity of a deep-junction SPAD was implemented. Chapter 4 goes through a detailed modelling approach that was developed for QRFF designs based on photon-timing statistics. The concept of a random flip-flop is further analytically developed. A particular emphasis is placed on one circuit architecture, which implements the slow clock method. A random bit generator that implements the first photon arrival principle is also designed and characterized, for comparison to the slow clock method. Results for QRFF circuits in 55 nm, including a novel method for removing bias, are presented. Finally, Chapters 4 and 5 scale the previously designed QRFF into large arrays. The FortunaSPAD design achieves an output data rate of 3.3 Gbps. It's successor, uses a macro-pixel based QRFF to improve robustness.

2 Silicon SPADs: background and QRNG considerations

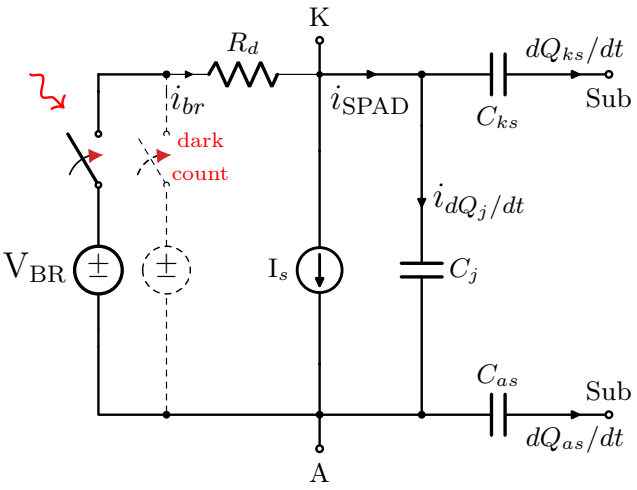
In this chapter, an effort is made not to simply regurgitate previous detector literature, but to contextualize figures of merit (**FoM**) within the application of random number generation, when designing SPADs in a standard process. A thorough description of the device operation, construction, and limitations provides a foundation for understanding the QRNGs presented in this thesis. Finally, recent innovations in detector design are also highlighted to demonstrate the general direction of research.

2.1 SPAD Device Physics, Operation, and Characterization

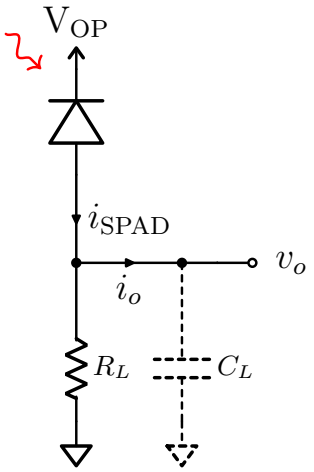
2.1.1 Basic operation and circuit abstraction

A single-photon avalanche diode (SPAD) is a device used to detect optical photons. When reverse biased with an excess bias voltage above its electrical breakdown potential, a SPAD is capable of detecting a single particle of light by generating a macroscopic current that is measurable with a pixel circuit. The detection process commences when an absorbed photon generates an electron-hole pair through the photoelectric effect. By operating above breakdown, in the so-called *Geiger mode*, a single carrier, may experience virtually infinite gain. This physical process is known as avalanche breakdown, and it contains distinct phases, known as buildup, spread, quench, and recharge.

SPAD sensors can, in certain instances, be modelled as asynchronous digital systems with pulses representing photon arrivals. However, for applications, in particular, those which are timing critical or reliant on Poisson statistics, the electrical dynamics of a detection should be properly understood. This understanding then facilitates the design of a pixel circuit suitable for an application. Figure 2.1 displays circuit abstractions for a SPAD and the simplest detector configuration. A reverse bias, V_{OP}, with an excess voltage V_{EX} above the PN junction's breakdown voltage, V_{BR}, applied to the diode creates a high-field region where the avalanche occurs. A circuit capable of modelling the static and dynamic electrical behavior of the SPAD avalanche is presented in Figure 2.1a. The model comprises a static diode current, I_s, a junction capacitance, C_j, parasitic substrate capacitances C_{as}, C_{ks}, and



(a) A circuit representation of a SPAD highlighting the major components used to model its static and dynamic behavior. A photon detection or dark count event closes the switch and initiates an avalanche.



(b) The simplest functional SPAD detector. A load resistance, R_L, is utilized to passively quench the avalanche.

Figure 2.1: Circuit representations of a standalone SPAD along with the simplest SPAD detector, which consists of a SPAD and a passive quench resistor. The internal capacitance is dominated by the junction and is in the order of $\simeq$ 50 fF for modern silicon SPADs. The static diode current, I_s, is in the range of picoamperes, although photo generated current can increase to nanoamperes. The breakdown voltage can vary greatly between devices ($\simeq 12 - 40$ V). Diode resistance, R_d, can vary in the range of a few hundred ohms to low kΩ depending on the thickness of the SPAD [64]. An avalanche causes the output voltage to rise to $\mathrm{V_{EX}} = \mathrm{V_{OP}} - \mathrm{V_{BR}}$.

resistance, R_d. The closing of the switch denotes the initiation of an avalanche caused either by a photon arrival or by noise (dark count), which will be covered in a subsequent section. At the onset of an avalanche, carriers are accelerated by the electric field, they gain kinetic energy and produce more carriers through impact ionization. This process denotes buildup, which spreads to the rest of the physical structure as carriers multiply. As the avalanche spreads, the voltage at the anode begins to rise, lowering the potential difference across the diode while charging the load capacitance. To prevent destruction of a device, a quenching circuit, that forces the current and voltage to exponentially approach their asymptotic values, is required. The simplest functional SPAD detection circuit is highlighted by Figure 2.1b, where a load resistance, R_L is used for properly quenching the avalanche. The value of, R_L, sometimes referred to as a ballast resistor, must be on the order of ~ 100 kΩ to ensure the device reaches its latching current, $\mathrm{I_{latch}}$. Below this value, which is typically on the order of 100 μA for silicon SPADs, the avalanche is self-quenching [65]. A scheme where a resistor is used to quench is denoted as *passive quenching*. The final voltage, $v_{o,f}$, and current, $i_{o,f}$, values at the output of the detector circuit can be approximated with Equations (2.1) and (2.2), respectively.

$$v_{\mathrm{o,f}} = \mathrm{R}_L \cdot i_{o,f} \simeq \mathrm{V_{EX}} \tag{2.1}$$

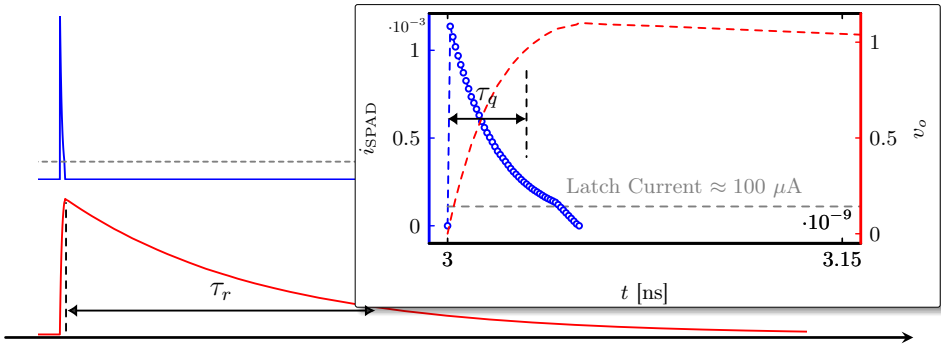


Figure 2.2: Results from a SPICE compatible Verilog-A simulation of a passively quenched SPAD. The graph highlights current that flows through the diode, i_{spad}, and the corresponding output voltage, v_o, across the quench resistor, as a function of time. Since the diode resistance is low, typically $\simeq 100$ kΩ, the peak current flowing through the SPAD can be in the mA range, although this is for a short period of time ($10 - 100$ps). Quench, τ_q, and recharge, τ_r, times are denoted on the plot. The simulation is performed with a quench resistor of 100 kΩ and a junction capacitance value of 50 fF.

$$i_{o,f} = \frac{\mathrm{V_{OP}} - \mathrm{V_{BR}}}{R_d + R_L} = \frac{\mathrm{V_{EX}}}{R_d + R_L} \simeq \frac{\mathrm{V_{EX}}}{R_L} \tag{2.2}$$

Once the total current flowing through the diode, i_{SPAD}, is below the latching current value, the probability of a carrier reaching the high field region is continually reduced until no more impact ionization events occur. Moreover, charge flow inside the SPAD directly impacts another FoM, known as afterpusling probability, which introduces correlated noise and is a critical consideration for QRNG design. This will be elaborated on in a coming section. The quenching time constant can be estimated using a simple capacitor charging model, as shown in Equation (2.3).

$$\tau_q = (C_d + C_L) \cdot \frac{R_d R_L}{R_d + R_L} \simeq (C_d + C_L) \cdot R_d \tag{2.3}$$

When the avalanche has been quenched, the final phase, referred to as recharging, commences. The principle is simple and describes the resetting of the SPAD to its idle state by recharging the junction so that the potential difference across the diode is restored to $\mathrm{V_{EX}}$. The time constant for this process is calculated with (2.4).

$$\tau_r = (C_d + C_L) \cdot R_L \tag{2.4}$$

As the diode is recharged, the probability of another avalanche initiation increases. The duration of time, from the start of an avalanche, until the detector is capable of generating another pulse i.e. detecting another photon, is referred to as the dead time. Dead time is dependent on several factors determined by the SPAD circuit interface, i.e. the pixel circuit design. It is clear that the load capacitance, C_L, has a significant effect on the quench, τ_q, and recharge, τ_r, times, therefore the dead time. However, the threshold of the discriminator circuit, which denotes a countable detection, ultimately decides the point where, during quench, a pulse is registered, and during recharge, a subsequent detection can occur. Therefore, the

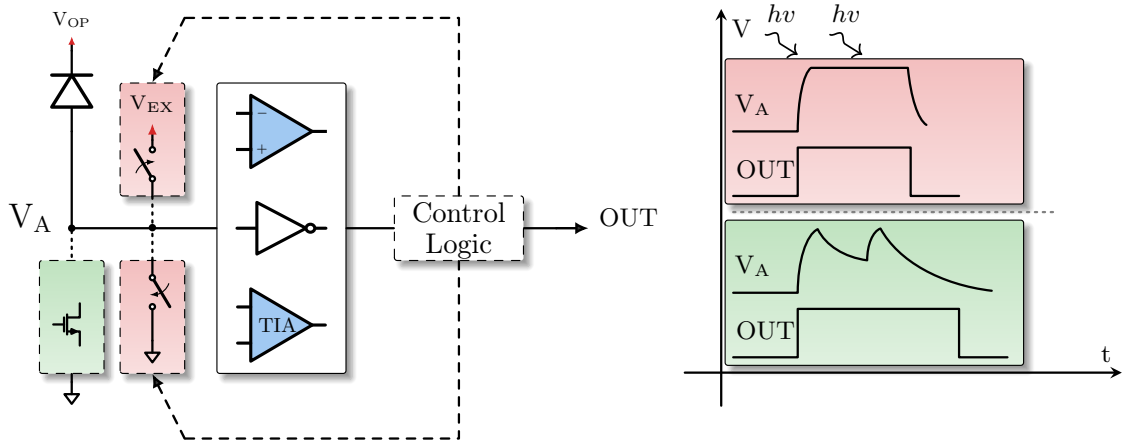


Figure 2.3: Diagram of a SPAD circuit interface outlining the various pixel circuit elements. Passive quenching/recharge can be performed with a resistive element i.e. biased transistor (green). However, feedback electronics can also be implemented in order to actively quench and recharge (red). Different discriminator circuits such as transimpedance amplifiers (TIAs) or comparators can be implemented. The timing diagram demonstrates a consequence in terms of dead time when choosing passive techniques. If a photon arrival happens during the recharge stage, this can extend the dead time (paralysable). More on this in Chapter 4. A thorough summary of pixel topologies is presented in [75].

SPAD circuit interface design is the dominant factor for ensuring a low dead time. Clearly, this determines the achievable count rates and as such, extensive research into the design of pixel circuitry has been conducted [66]–[74]. Certain pixels prioritize compactness, while others are designed for dead time speed or versatility. Regardless of the application specific considerations, each pixel implements quenching and recharge methods that are either passive, such as the resistive example in Figure 2.2, or active.

Figure 2.3 illustrates various blocks implemented in pixel circuits to achieve these functions. Active components, in place of a resistor, can be used for both quenching and recharge. The benefit provided by active quenching is an increase in quenching speed, which can be advantageous for application that require excellent timing performance, such as Positron Emission Tomography (**PET**). However, it comes at the cost of increased current flow during the quench phase. Furthermore, given the low diode resistance and capacitance of integrated circuits, the quench time of modern SPAD pixels is $\tau_q \leq 1$ ns. Therefore, few applications justify the use of an active quench circuit, as the practical gain is limited. To decrease the recharge time phase caused by the quench resistance, it is common to implement an active recharge circuit. The activation of the recharge can be controlled externally, or using a feedback mechanism in the circuit for automatic operation. Moreover, a tunable delay element can also be implemented within the recharge loop for controllable dead time. Sub 10 ns dead times are readily achievable when applying active recharge. More recently, pixels capable of sub-ns performance has been demonstrated [74].

Various discriminator circuits have been demonstrated, although the two most common are a simple inverter and a comparator. An inverter provides a compact solution that is suitable for many applications. However, for timing critical performance, the threshold pin of a comparator

can provide a way to shorten the period of time between an avalanche initiation and a countable detection.

Perhaps more relevant to QRNGs, is the effect the pixel circuit has on the statistics of photon counting. While an ideal detector can be modelled as a Poisson process, dead time complicates the counting statistics [76]. If passive techniques are used, the dead time is difficult to determine, as the SPAD is capable of initiating another avalanche during the recharge phase, but before the discriminator is capable of producing a pulse, thereby extending the dead time. This phenomenon is referred to as *paralysable dead time* and is illustrated by the timing diagram in Figure 2.3. When active circuits are used, the dead time can be controlled more precisely, and an essentially non-paralyzable regime can be achieved. The dead time effects on counting statistics, and consequently bit generation statistics, are analyzed extensively in Chapter 4. While the avalanche dynamics and pixel circuits are important for understanding the counting behavior of a SPAD, the device is further characterized by FoMs that describe it's photo-sensitivity, noise and timing performance [77]–[80].

2.1.2 Detection efficiency and fill factor

Photon detection probability (**PDP**) denotes the ratio of detected photons versus those incident. To understand the photon detection process, a diagram denoting the physical regions within the SPAD, can be analyzed. Figure 2.4 displays a basic two-dimensional cross-section of a SPAD along with a sample electric field profile when reversed biased above breakdown. As discussed in [81] a distinction is made between the multiplication and drift regions. In the drift region, the electric field only needs to be above a certain value (10^4V/cm) to guarantee saturation of the carrier's velocities. This minimizes transit time to the multiplication region, where the carriers are multiplied (gain) and initiate an avalanche current. The design of the drift and multiplication regions directly influence not only the PDP but noise and timing performance of any detector. In general terms, it is desirable to fully deplete the depth range where photons can be absorbed. This allows for the drift of generated charge carriers to the multiplication region, increasing the likelihood that an avalanche will be initiated, therefore maximizing PDP. However, this is not always possible, particularly in standard technologies where custom implants are not available. A method to help overcome this challenge is demonstrated in Chapter 3. When a photon reaches the SPAD, if it is not reflected by the silicon dioxide interface, it may be absorbed within the silicon, thereby generating an electron-hole pair (charge carrier). The average depth, $\mu_d(\lambda)$, at which the absorption happens is defined exclusively for any given material, as a wavelength (λ) dependent function. Depending on its absorption depth, the carrier can then either drift or diffuse towards the multiplication region, where it may ignite an avalanche. Indeed, all of these steps are probabilistic in nature and are modeled as such.

The PDP is often defined either using the product of a detector's quantum efficiency (**QE** and the probability of a breakdown event (P_b) or the product of absorption (P_{ab}) and avalanche

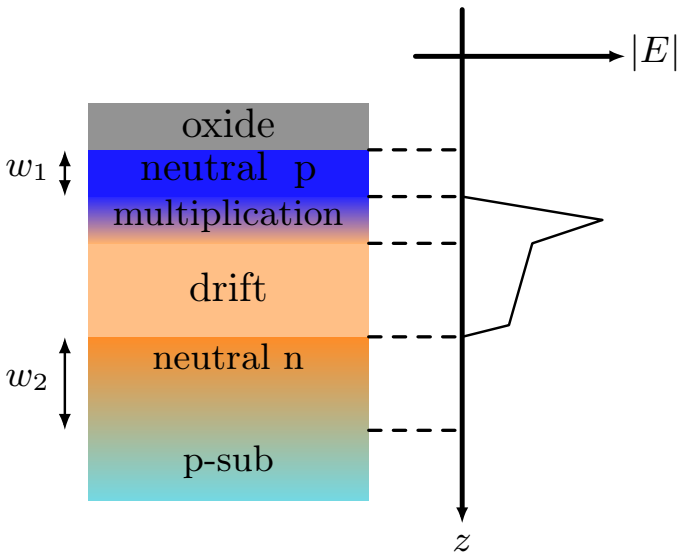


Figure 2.4: A PN junction SPAD, under reverse bias, outlining the relevant regions when considering charge collection (P_{diff}). Equation (2.7) can be used to analyze the charge collection probability, i.e. the probability that a photo generated charge carrier will enter the depleted region before recombining.

statistics (P_{av}).

$$\text{PDP}(\lambda, \theta_0, z) = \text{QE}(\lambda, \theta_0, z) \cdot P_{\text{b}}(\lambda, z, \text{V}_{\text{EX}}) = P_{\text{ab}}(\lambda, \theta_0, z) \cdot P_{\text{av}}(z, \text{V}_{\text{EX}}) \tag{2.5}$$

The formulation of PDP that contains (QE) is more abstract, whereas the definition containing absorption and avalanche probabilities is quite intuitive. Moreover, understanding the relationships between the probabilities contained in Equation (2.5) is useful during design and simulation of SPADs. Quantum efficiency (QE) denotes the fraction of photon flux that contributes to the photocurrent generated in a pixel. For any arbitrary region that extends to a depth of z_e, it can be defined by Equation (2.6).

$$\text{QE} = \int_0^{z_e} \underbrace{T(\lambda, \theta_0)\mu_d(\lambda)e^{-z\mu_d(\lambda)}}_{P_{\text{ab}}(\lambda, z)} \cdot P_{\text{diff}}(z) \cdot dz \tag{2.6}$$

The angle of incidence of a photon is defined as θ_0. $T(\lambda, \theta_0)$ is the transmittance function determined by isolation (oxide) layers and also passivization layers on top of the SPAD. If the SPAD surface interface is not optimized, $T(\lambda, \theta_0)$ will induce a standing wave pattern on the PDP. The probability that a generated carrier reaches the depletion region before recombination is defined by P_{diff}. Furthermore, P_{diff} is dependent on the region within the SPAD where the photon is absorbed, i.e. where the charge carrier is generated [82], [83]. For the practical physical construction shown in Figure 2.4, P_{diff} can be evaluated for individual regions using Equation (2.7).

$$P_{\text{diff}}(z) = \begin{cases} \int_0^T e^{-(w_1 - z/L_e)} \cdot dz, & \text{neutral p-type (above)} \\ \int_0^T e^{-(z - w_2/L_h)} \cdot dz, & \text{neutral n-type (below)} \\ 1, & \text{inside depletion region} \end{cases} \tag{2.7}$$

The lengths L_h and L_e are distances the holes and electrons must diffuse to reach the depleted region, respectively. The probability of an avalanche occurring is the product of the diffusion and breakdown probabilities $P_{\text{av}}(z, \text{V}_{\text{EX}}) = P_b(z, \text{V}_{\text{EX}}) \cdot P_{\text{diff}}(z)$. Therefore, in order to obtain the avalanche probability, the breakdown probability must be modeled. Originally, this was explored in [84] and [85]. More recently, in [86], it was shown that the non-local (dead space) model proposed by McIntyre in [87] did a much better job at predicting P_b for narrow junctions, which are not well modeled by local ionization rates. For this reason, technology CAD (**TCAD**) simulations of P_{b} in Chapter 4 of this thesis are performed using the dead-space model. Moreover, P_{b} is dependent on the excess bias applied to the SPAD, and can be roughly modelled using an exponential ratio with an empirical correction factor, η_c: $P_{\text{b}} \simeq 1 - exp(-\text{V}_{\text{EX}}/(\eta_c \cdot \text{V}_{\text{BR}})$ [88]. However, this probability saturates at a point when, an excess bias that causes full depletion of the junction, is applied. Note that in literature, the breakdown probability is sometimes referred to as the avalanche triggering probability.

In a practical detector design, there will be some regions of the silicon, such as the guard ring and periphery regions, that are insensitive to photons. Therefore, the fill factor (**FF**) of a SPAD is defined simply as the ratio of active area over total area. Therefore, the concept of detector efficiency is extended to include the metric known as photon detection efficiency (**PDE**), which is simply the PDP multiplied by the FF: $\text{PDE}(\lambda) = \text{PDP}(\lambda) \cdot \text{FF}$. Furthermore, in a sensor design containing an 2D array of pixels, the region of the pixel containing electronics indeed increases the dead region. Evidently, more recent research has focused on increasing FF for both the detector and the array, using a variety of techniques, which are outlined further on.

2.1.3 Noise Performance

A single-photon sensitive detector can have spurious pulses in the absence of present photons, known as dark counts. In SPADs, the rate of these noise contributions, known as the dark count rate (**DCR**), comprises thermal generation, band-to-band tunneling and trap-assisted mechanisms. Dark counts, which are statistically indistinguishable from photon detections, in the sense that they arrive with exponentially-distributed inter-arrival times, are known as primary dark counts. However, carriers, which become trapped in energy levels caused by process defects, can create pulses that are temporally correlated. These are defined as secondary dark counts, or afterpulses.

Primary Dark Counts

Shockley-Read-Hall theory is used to model the rate of thermally generated carriers inside SPAD detectors [89], [90]. As the name suggests, this carrier generation rate is highly dependent on temperature with a typical increase by a factor of 2-4 per 10°C with temperature, which is aided by traps [91]. Therefore, implantation and annealing processes play a vital role in the detector noise performance [92]. For a device containing a density of generation and recombination centers, N_t, assuming that the electron/hole concentration is much lower than the intrinsic level ($n \ll n_i$) in the depletion region, and the cross-section, σ_0, for capturing carriers are equal, then the thermal generation rate, G_{therm}, can be estimated with Equation (2.8) [93].

$$\mathrm{G_{therm}} \simeq \frac{n_i v_{th} \sigma_0 N_t}{2} \cdot A_a \cdot W_d \tag{2.8}$$

The SPAD active area is A_a and W_d is the width of the depletion region. The thermal velocity of electrons can be calculated using $v_{th} = \sqrt{3k_{\mathrm{B}}T/m}$ where m is the electron mass.

The other contributor to primary dark counts is band-to-band tunneling ($\mathrm{G_{btb}}$). Typically, band-to-band tunneling is not a significant contributor to primary dark counts, particularly in deep junction SPAD designs which have lightly doped junctions. However, this can change at low temperatures where thermal generation is reduced. Band-to-band tunneling carrier generation can be estimated with Equation (2.9).

$$\mathrm{G_{bbt}} = \frac{\sqrt{2m^*}q^2\overline{\mathrm{E}}_d \mathrm{V_{OP}}}{h^2\sqrt{E_g}} \exp\left(-\frac{8\pi\sqrt{2m^*}E_g^{3/2}}{3q|\overline{\mathrm{E}}_d|}\right) A_a \tag{2.9}$$

The bandgap energy is, E_g, normalized charge, q, and Planck's constant is denoted as h. A higher average electric field magnitude $|\overline{\mathrm{E}}_d|$, in the depletion region, increases band to band tunneling. Primary DCR contribution can then be estimated by multiplying the triggering probability with the total carrier generation rate Equation (2.10).

$$\mathrm{DCR_p} = (\mathrm{G_{bbt}} + \mathrm{G_{therm}}) \cdot P_b \tag{2.10}$$

Technology node scaling has had an adverse effect on DCR performance in SPADs. This is due to the increased doping levels and curtailed annealing steps used in deep sub-micron processes. For this reason, until recently, low noise SPADs in technology nodes below 65 nm were not prevalent. Chapter 3 details how these challenges were overcome to design low noise SPADs using a variety of junction types.

Secondary Dark Counts (Afterpulsing)

Afterpulsing, its causes, characterization methods, and modeling has been researched intensely in silicon avalanche photodetectors [94]–[98]. As defect concentration in the crystals, which

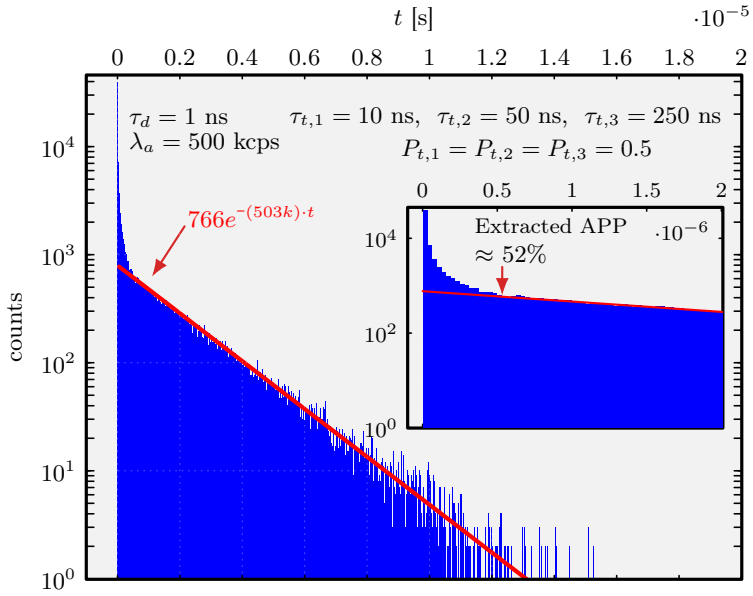


Figure 2.5: A Verilog-A simulation used to generate photon arrivals at $\lambda = 500$ kcps. The inter-arrival histogram of detections is plotted. Here, 3 traps with varying lifetimes are used to mimic afterpulsing behavior. The pre-exponential factors $P_{t,1}, P_{t,2}, P_{t,3}$ are exaggerated to clearly show the hyper exponential behavior for temporally bunched arrivals, resulting in a high APP (52%). A fitted exponential (red) is used to determine the afterpulsing probability, α. The simulation was performed with a negligible dead time, $\tau_d = 1$ ns. The resulting plot mimics closely a measurement taken with a time tagging device.

allow carriers to be trapped at energy levels close to the Fermi-level, are heavily dependent on material properties and process parameters, afterpulsing behavior can vary greatly between detectors. For QRNG design, where any form of correlated noise can manifest as correlated bits, the effects of afterpulsing, must be well understood. Simple principles are outlined here, while a more detailed review of literature and considerations for QRNG design are left for Chapter 4. While equations that model the temporal dependence of afterpulsing effects can differ [98], [99], it is generally agreed upon that each trap, $n_{t,i}$, has a lifetime, $\tau_{t,i}$, along with some initial probability of generating an afterpulse, $P_{t,i}$, which decays with time until the trap is released. The lifetimes can be determined through fitting of a histogram of inter-arrival times using a time-tagging device, and the activation energies can be extracted using an Arrhenius plot of the dark count temperature dependence [100]. The initial probability of an avalanche ignition from a trap, $P_{t,i}$, is an abstraction of a combination of physical parameters such as defect concentration and charge flow. As current flows through the detector, defect centers become populated with trapped carriers. When these carriers are released, they may cause a spurious avalanche. It is common to model this probabilistic behavior using a summation of probabilities for all traps, as shown in Equation (2.11).

$$\mathbb{P}(A)\{t\} := \sum_{i}^{N_{t,t}} P_{t,i} \cdot e^{(-t/\tau_{t,i})} \quad \text{for } t \in \mathbb{R} \tag{2.11}$$

Here, $\mathbb{P}\{t\}$ represents the probability, at any given time during the decay process i.e. after an avalanche, that a released trapped carrier, generates an afterpulse. From a FoM perspective, a SPAD is typically characterized by its overall apfterpulsing probability, (α or **APP**) which denotes the percentage of pulses that can be statistically attributed to being an afterpulse. There are several measurement techniques used to determine APP Equation (2.12).

$$\alpha = \mathrm{APP} = \frac{\mathrm{N_A}}{\mathrm{N_T}} \tag{2.12}$$

The simplest method is to calculate the percentage of pulses, in an inter-arrival histogram of detections, that are augmented on top of an ideal exponential fit, as shown in Figure 2.5. This figure shows a histogram generated from a simulation of a SPAD that contains three traps with various lifetimes. The pre-exponential probabilities, $P_{t,i}$, were exaggerated in order to visualize well the hyper-exponential behavior of the inter-arrival times in the presence of afterpulses.

2.1.4 Timing Performance

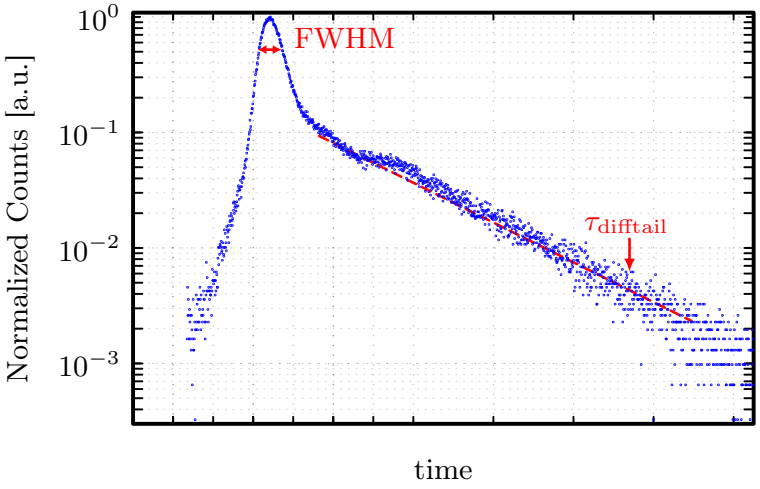


Figure 2.6: An example timing jitter histogram of a SPAD detector. The FWHM of the Gaussian component is highlighted. A sample fit showing the time constant of the diffusion tail, τ_{difftail}, is also shown.

If we define an arrival event as the time when an absorbed photon generates a charge carrier, then there exists a temporal delay between this event and the generation of a pulse at the discriminator, which is what constitutes a detection. As previously denoted, the avalanche must build up and spread from its seed point until the discriminator circuit registers a pulse. The statistical characterization of this delay is known as the timing performance or timing jitter. During measurement, a histogram, using a time-correlated single-photon counting (**TCSPC**) technique, is built to denote the variation in detection time and is benchmarked using the full width at half maximum (**FWHM**) of the timed detection. The physical processes that contribute to the timing performance are broken down into three components. The transit time, τ_t, describes the amount of time the charge carrier takes to drift or diffuse into the

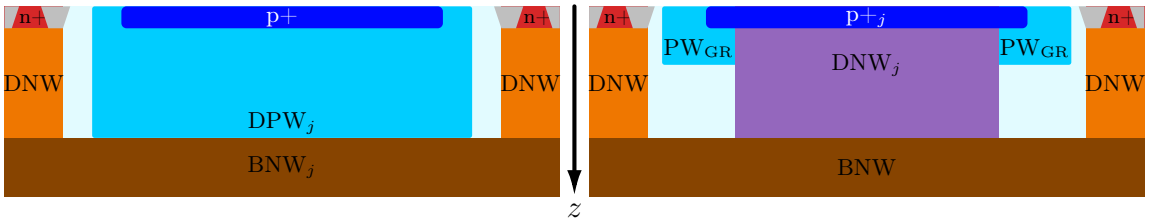


Figure 2.7: Sample cross-sections of popular SPAD designs in CMOS. The SPAD on the left can be broadly classified as a deep junction, with the multiplication region formed by the DPW$_j$/BNW$_j$ interface. In this design, a virtual guard ring is used, signifying that no dedicated implant is used on the periphery of the device. On the right, a shallow junction is formed between the p+$_j$/DNW$_j$ interface. An example of an explicit guard ring is drawn as PW$_{GR}$. Many variations of these structures have been demonstrated. z denotes the relative depth within the silicon wafer. These structures are explored in more detail in Chapter 3.

multiplication region. For photo-generated carriers outside the multiplication region, τ_t is longer and possesses an exponentially-distributed delay in the timing response. The build up of the avalanche is described by a time, τ_b, where charge carriers are accelerated and multiply. Subsequently, the avalanche spreads laterally until a time, τ_p, where the discriminator is able to detect the current pulse. Therefore, the overall description of detection time, τ_{det}, can be defined as

$$\tau_{det} = \tau_t + \tau_b + \tau_p \tag{2.13}$$

When measurement of this metric is performed, it is typical to attenuate the laser pulse in order to achieve a single-photon operating region and therefore measure the single-photon timing resolution (**SPTR**). The resulting figure, demonstrated by Figure 2.6 can be broken up into two separate regions. The Gaussian component characterizes the statistical variation of the avalanche statistics when a photon is absorbed in the drift region. The *diffusion tail* outlines the timing performance of the detector for photons that diffuse towards the multiplication region in order to initiate and avalanche. Timing performance is dependent on both wavelength and the diode physical construction and will be discussed further in Chapter 3.

2.1.5 SPADs in CMOS

Typical construction and practical considerations

Single-photon avalanche diodes exit in a variety of physical configurations. However, the general implant requirements for a SPAD remains similar, regardless of the specific construction. in Figure 2.7, two common structures are shown. During design, the first consideration is the SPAD junction, which forms the multiplication region. When the implants used form a shallow junction, the multiplication region is formed with a depth of $0.1 - 1$ μm within the silicon. The ability to achieve a higher doping level at shallow depths (due to ion implantation energies) has certain advantages. A SPAD's breakdown voltage can be lowered with higher doping levels. However, this can increase tunneling noise. A more critical concern is the proximity of the junction to the silicon dioxide interface. In CMOS processes, it is possible to

have a higher defect concentration in this region, therefore afterpulsing is relatively higher for shallow junction SPADs [101]. Moreover, due to the absorption function, $\mu(\lambda)$, described earlier on, shallow junction designs are limited in their ability to achieve high sensitivity. Conversely, for deep junction SPADs, which are formed using high-energy Boron implantation, can achieve a junction depth of $\sim 2-3$ μm [102]. While the low doping levels can result in an inconveniently high V_{BR}, the resulting sensitivity, particularly in the red/near-infrared (**NIR**), is improved. Note, that only p/n SPADs are diagrammed here. The reverse configuration, where the junction is formed with an n/p junction, can also be considered. However, in particular for 2D ICs, where SPADs and circuits are on the same substrate, p/n configurations are preferred. This is due to their inherent isolation from circuits, which simplify the pixel circuit design requirements.

In order to avoid what is referred to as *premature edge-breakdown* around the periphery of the multiplication region, a guard-ring design needs to be employed. A guard-ring reduces the electric field magnitude around the multiplication periphery. This can be a dedicated implant that forms a doping gradient around the edge of the junction. Or, a *virtual guard ring* can be implemented by with enlarged periphery space and a doping profile of implants which naturally sums to a smooth transition. Shallow trench isolation can be used as an guard ring in order to improve the device FF, but it has shown to have significant drawbacks in terms of noise performance [103]. Some more recent trends in guard ring designs will be introduced further on.

2.2 SPAD sensors

SPAD sensors are system-on-chip designs which can scale from arrays of a few dozen pixels [104], to megapixel order [105]. The demand and interest for these sensors seemingly increase monotonically due to their utility for use in a variety of applications. Many additional considerations, both from the detector and system perspective, must be considered in order to design practical, functional sensors.

2.2.1 Additional FoMs for arrays

Scaling

Scaling the detector design to an array brings additional considerations and challenges. As previously mentioned, the requirement for pixel electronics reduces FF. Therefore, arrays have, when possible, employed *resource-sharing* in order to increase the total surface area that is sensitive to photons. However, the ultimate method for improving FF is to move to a 3D stacked technology, where only detectors are placed on a single die and the electronics, which are designed using a separate technology node, are stacked either below or above the detectors. Details on this technique will be discussed further on. Advanced theory on scaling laws for pixel miniaturization is available from [66].

Power consumption

Power consumption becomes a more prevalent issue as systems scale. Moreover, for QRNGs presented in this thesis, which require high photon flux, the power consumption must be considered carefully. The energy consumed by a SPAD during an avalanche can be estimated with (2.14). Furthermore, given a detection rate of λ_{D}, the per detection power consumption is given by (2.15) [65].

$$E_s = Q_t \cdot \left(\mathrm{V_{BR}} + \frac{\mathrm{V_{EX}}}{2} \right) \simeq Q_t \mathrm{V_{EX}} \tag{2.14}$$

$$\mathrm{P}_s = Q_t \cdot \left(\mathrm{V_{BR}} + \frac{\mathrm{V_{EX}}}{2} \right) \cdot \lambda_{\mathrm{D}} \tag{2.15}$$

The total charge Q_t can be calculated as $Q_t \simeq \mathrm{V_{EX}} \cdot (C_j + C_L)$. The consequences of the bit generation method used in terms of power consumption will be elaborated on in Chapter 4.

Dynamic range

Dynamic range is a critical quantity for many imaging applications, as it directly impacts image quality and the ability to image in low light conditions. The dynamic range of a SPAD using an integration time, t, can be calculated, when taking into account a correction factor, F_c, which is based on the recharge mechanism, with (2.16) [104].

$$\mathrm{DR}_s = \frac{F_c \cdot t}{\tau_{\mathrm{dead}} \sqrt{t \cdot \mathrm{DCR}}} \tag{2.16}$$

As DCR is not necessarily detrimental, as long as afterpulsing is very low/negligible, and high flux conditions are needed in order to generate at high-speeds, DR_s is not a large concern for QRNG designs. Nevertheless, by virtue of the fact that defect concentration must be kept low for reduced afterpulsing, and dead time must be controlled and low for high flux, QRNGs can have DR_s values acceptable for imaging applications as well.

Crosstalk

Recombination of electrons/holes during avalanche multiplication causes either photons or phonons to be emitted so that the energy conservation law is preserved. A detection of one of these photons by an adjacent detector is called a crosstalk event and is another form of correlated noise. As previously mentioned, correlated noise must be avoided for QRNG development as it can manifest as entropy degradation. Due to silicon being an indirect bandgap material, the number of generated photons is relatively low (tens of photons for every million avalanche carriers). However, crosstalk can still be significant in arrays with values that range from $(\sim 0.1 - 4)\%$, depending on the pixel pitch chosen for the array [106]. In RNG applications a smaller active area can be desirable to reduce crosstalk, and to reduce

capacitance, which can increase the count rates.

2.3 Recent advances and trends

3D Stacking and Backside Illumination

Although high-performance SPADs and circuits are capable of monolithic integration, the optimized conditions for performance involve 3D stacking of chips. In this configuration, detectors are placed on a separate die, using a process that is amenable to SPAD requirements. Moreover, this allows all circuitry to be placed below/above the detector, automatically providing a FF improvement and allowing for additional space for integration of complex circuit/processing functions. Two main configuration exist. When ICs are placed face-to-back, the detectors are illuminated from the frontside with circuits on top, therefore referred to as frontside illumination (**FSI**). Conversely, when the bonding between dies is done in a face-to-face manner, the detectors are illuminated from the backside (**BSI**) with circuits below. Although either configuration is possible, it is usually preferred to have 3D stacking in the BSI configuration. BSI SPADs can bury the junctions at much deeper depths, facilitating improved NIR performance. Moreover, FSI 3D ICs require through-hole vias (**TSV**s) for each individual pixel, which complicates the process flow. Consequently, in the last half-decade, there have been many demonstrated BSI detectors [107]–[112] and systems [113]–[118] in various processes. A review of process considerations for 3D ICs is given by [119]. While 3D stacking is an attractive option for NIR performance demanding applications, it is not suitable for every sensor given the complexity, risk and cost associated with the additional process steps. For this reason, monolithic design in deep sub-micron CMOS remains an active area of research and development.

PDE enhancement

An illustration outlining the various sections of a contemporary SPAD's cross-section, used in image sensing applications, is displayed by Figure 2.8. The technologies and techniques are described here. Improving quantum efficiency can be broken down into three categories, **photon absorption, carrier collection and breakdown probability**. Research has focused on all of these areas in order to continually improve SPAD performance. In particular, driven by the demand for automotive LiDAR, a considerable effort has been put into enhancement in the NIR area of the spectrum.

Microlenses, used to guide light that would normally enter the SPAD periphery or pixel circuits, have become a cornerstone of recent image sensors [117], [120]–[124]. These lenses therefore *recover fill factor,* thereby improving PDE. Novel concepts, and those borrowed from other sciences such as photonic technologies and applied electromagnetics, are becoming prevalent in SPAD sensor design to improve absorption. For example, a light trapping SPAD proposed in [125] was used for enhanced absorption by patterning nanostructures to create a horizontal

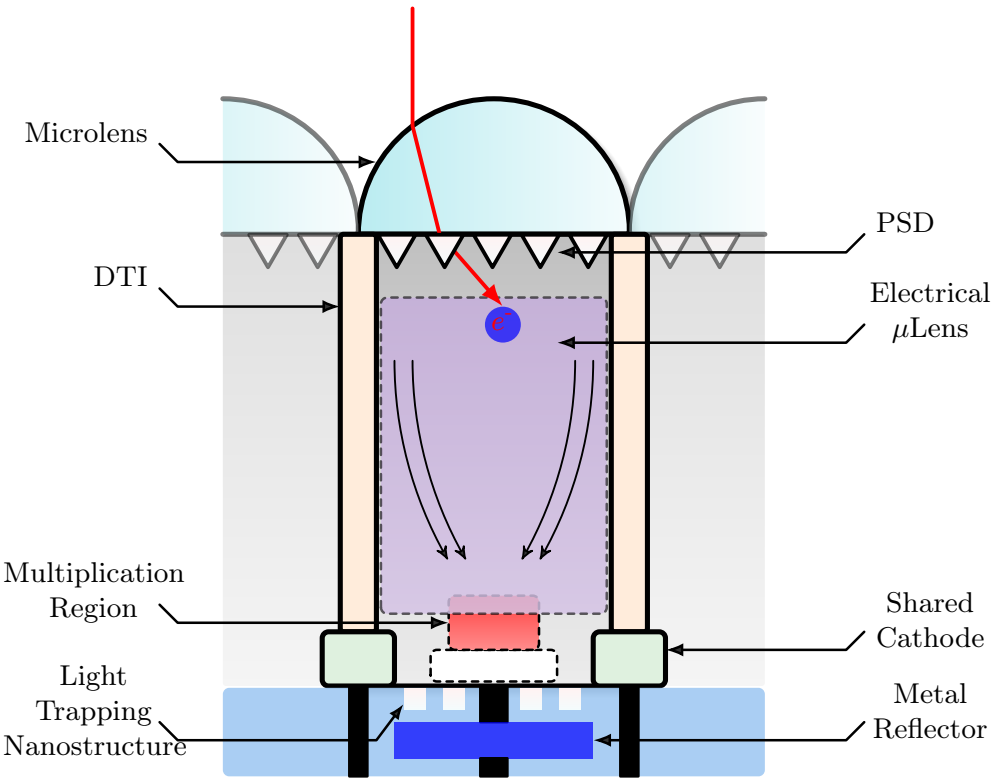


Figure 2.8: Anatomy of a modern SPAD in 3D BSI. An optical microlens is shown to help recover fill factor by focusing light into the photocollector region. Moreover, PSDs can be implemented for diffracting NIR photons, which increases sensitivity but reduces timing performance. Electrical microlensing extends the photo collector region. Deep trench isolation enables small pixel pitch while reducing cross talk. Light trapping structures, such as a simple metal layer or more advanced nanostructures, reflects back photons that have passed through the silicon.

waveguide mode from the incident photon. The increased absorption length improved on an inherent trade-off between junction depth and timing performance while achieving NIR enhancement. However, the compatibility of this method is limited for standard processes. Nanostructurization was also used in [126] by altering the STI mask to generate sub-wavelength patterns in the silicon, although with a detrimental effect on afterpulsing. Resonant cavities, such as those in [127], where a Bragg reflector using a silicon-on-insulator (**SOI**) substrate was used to prolong the optical path, can also increase PDE. However, the improvement achieved from these resonant structures are narrowband. Finally, a reflective structure that has gained commercial interest due to its amenability for CMOS process manufacturing is the pyramid surface for diffraction (**PSD**) [128]. A review of these nanophotonic structures is provided in [129]. A review of photon manipulating nanostructures including those not covered here, such as dielectric mirrors, waveguide based avalanche photodiodes, and vertical cavities with lateral collection, is provided in the two-part manuscript [130], [131].

Perhaps the largest breakthrough in PDE-enhacement technology for silicon SPAD structures in standard processes has been the electrical microlens [132], otherwise referred to as a charge-focusing SPAD [133]. This technique is an advancement in charge collection. In this

configuration, the SPAD multiplication area is reduced to a relatively small cross-section. However, engineering of the electrostatic potential with tailored implant doping profiles allows for charge carriers to be swept into the multiplication region with high-probability, regardless of where the photon was absorbed. Essentially, this allows the avalanche to propagate laterally so that carriers are collected from so-called dead regions. Results have been promising in recent years. A 130 nm design demonstrating 13% at 905 nm at 3.5 V_{EX} was presented in [134] and then improved with a BSI version in [135] achieving 27% at 904 nm at 3.5 V_{EX}. These designs were then scaled to large arrays, where a 3 Megapixel design was shown [136] and subsequently, a 0.37 W 143dB-Dynamic-Range 1 Megapixel BSI image sensor with pixel-wise exposure control and adaptive clocked recharging [118].

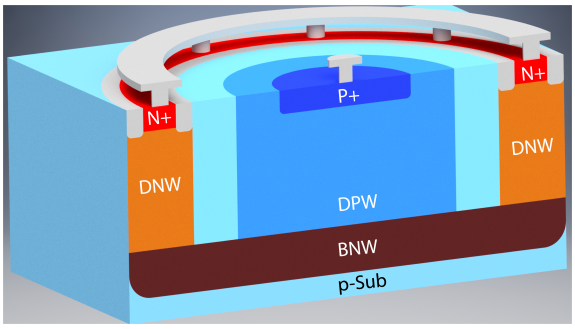
Although, from a system perspective, the multi-die approach is potentially advantageous for QRNG applications, due to the ability to optimize components and perhaps integrate illumination within a package, the remainder of this thesis focuses on monolithic approaches. The SPAD parameters presented in this chapter, in particular dead time and afterpulsing, are carefully considered in order to reduce bias and correlation in the bit generation technique used for the large arrays (FortunaSPAD, FortunaSPAD2).

3 Silicon SPADs in 55 nm BCD

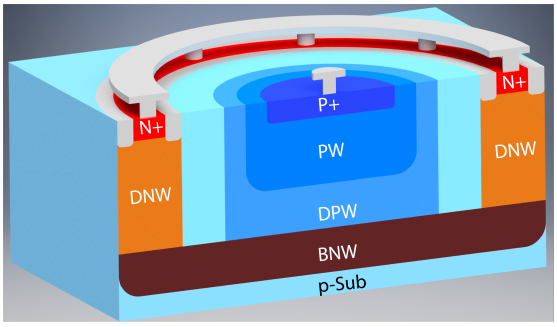
The work presented in this chapter has been published in [137] and [101].

Design and characterization of four SPAD structures fabricated in the GF 55 nm BCDlite process is detailed in this chapter. Bipolar-CMOS-DMOS (BCD) are emerging process technologies, sometimes referred to as 'smart-power' technologies because of their flexible offering of implants. This provides designers with the ability to integrate high voltage devices, while maintaining the integration capability/density of CMOS. Therefore, BCD processes are promising for the development of deep submicron SPAD devices and sensors. Moreover, the design of a pixel circuit, used for proper characterization of timing critical metrics such as timing jitter and afterpulsing, is presented.

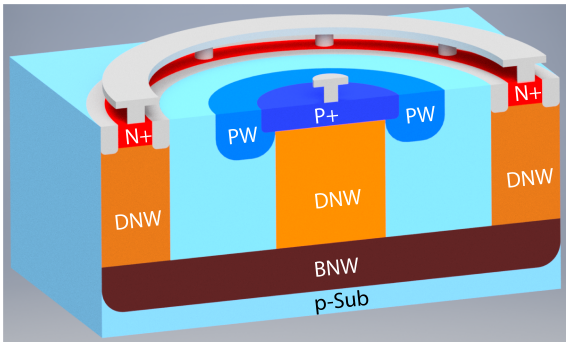
Cross-sections of the four devices are shown in Figure 3.1. As considered in the previous chapter, CMOS SPADs can be broadly classified as either deep or shallow junctions, depending on the relative depth of the junction that forms the multiplication region. In this work, Figures 3.1a and 3.1b can be considered as deep junction devices. The cross-sections displayed in Figures 3.1c and 3.1d are, in contrast, shallow junctions. Two advancements are presented and described. First, the ability to engineer the breakdown probability, P_{b}, of a SPAD, in a standard process is presented. This single implant is highlighted by the additional PW that differentiates deep junctions, shown in Figures 3.1a and 3.1b. Another advancement is demonstrated by extending the depleted region of a shallow junction SPAD. This is performed with the addition of a shallow p-well (**SPW**), which differentiates the devices displayed in Figures 3.1c and 3.1d. Finally, a performance summary and comparison to the state of the art is provided. Certain quantities, such as junction depth/width, and doping levels are referred to in relative rather than absolute terms, to avoid revealing confidential foundry process information.



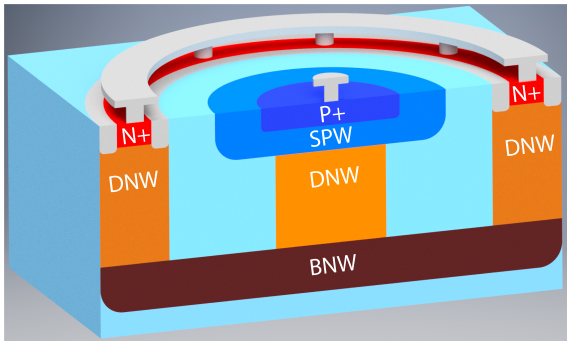
(a) Deep SPAD with junction formed by DPW/BNW interface. Referred to as non-opt.



(b) Deep junction SPAD with engineered breakdown probability. Refferred to as opt.



(c) A shallow device with junction formed by the p+/DNW interface.



(d) Shallow SPAD with enhanced depletion region using SPW implant.

Figure 3.1: Cross-sections of SPADs fabricated and characterized in a 55 nm process. Two devices (a,b) with junctions formed deep within the silicon using a buried n-well (BNW), deep p-well (DPW) interface are shown. Two shallow junctions (c,d) are also presented.

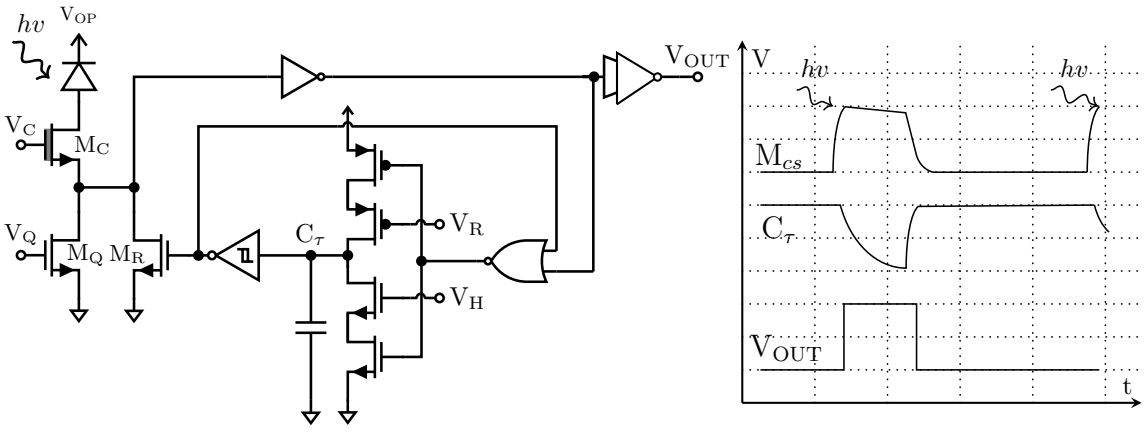


Figure 3.2: PQAR circuit integrated with SPADs for accurate characterization of afterpulsing and timing jitter. A cascode transistor is used to enable testing at $V_{EX} \leq 5$ V and a tunable delay element is added in the feedback loop for controllable dead time. The waveform displays the general operation upon detection of a photon.

3.1 Passive-quench active-recharge pixel circuit

A pixel was designed for characterizing select SPAD designs. Moreover, by doing so, the capability of the 55 nm process for monolithic integration of CMOS circuits and high-performance SPADs is demonstrated. The passive-quench, active-recharge (**PQAR**) pixel is displayed in Figure 3.2 with a simple timing diagram outlining relevant voltages. For proper characterization of afterpulsing, very low dead time is desirable. Moreover, as previously described, this increases dynamic range. Therefore, an active recharge circuit is implemented. When an avalanche is initiated, the anode voltage starts to rise. Once the inverter threshold is reached and a pulse is registered, the feedback capacitor starts to discharge, as shown in the timing diagram. This part of the pixel represents a tunable delay element, allowing for precise control of the dead time through controls V_H and V_R. By decreasing V_H, the rate of discharge of C_τ is decreased, i.e. the pulse-width hold time is extended, which increases the dead time. Moreover, changing V_R adjusts the duration where M_R is on, thereby controlling the recharge time. The cascode transistor, M_C acts as a voltage clamp that enables characterization of the SPAD at higher excess biases, without damaging the electronics that operate at VDD = 1.2 V. The passive quench transistor, M_Q, which has a high impedance for fast quenching, is typically turned off as the entire pixel is functional without it. However, it was included in the design for additional flexibility in dead time control. The achievable dead time range measured with a single device is between $\tau_{\text{dead}} \simeq 2 - 100$ ns.

3.2 Deep junction SPADs

Chapter 2 described some of the recent advancements in SPAD design of the photocollector regions, using, for example, electrical microlensing. Moreover, light trapping/absorption can be performed/improved with the addition of nanostructures. However, these methods are often not available in standard processes. Therefore, for FSI SPAD sensors, methods that are amenable to existing fabrication steps must be used to improve sensitivity (PDP). Here, a SPAD (Figure 3.1a) and it's optimized counterpart (Figure 3.1b) are presented. It is shown that the simple addition of the PW implant can dramatically improve PDP. For simplicity, the devices are referred to as optimized (**opt**) and non-optimized (**non-opt**), going forward.

3.2.1 Device description

The SPAD junction is formed using the **DPW**/**BNW** interface for both SPADs. In general, the availability of a BNW layer is greatly advantageous for SPAD design. First, it provides a method to isolate the multiplication region from the p-doped substrate, thereby reducing optical crosstalk and electrical interference with the pixel circuit [138]. High energy Boron implantation have enabled the creation of retrograde-doped BNW implants, which help build a thick multiplication region while reducing noise [91]. Furthermore, the design of the BNW layer is critical for timing performance as the thickness of the n-well region contributes to the diffusion tail, while the overall resistivity will impact the full-width at half maximum

[139], [140]. The region in between the p+ anode and the junction can be broadly classified as the photocollector or quasi-neutral region. TCAD simulations, the change in breakdown probability, P_b, can be investigated when the PW implant is added in the photocollector region.

3.2.2 TCAD simulation

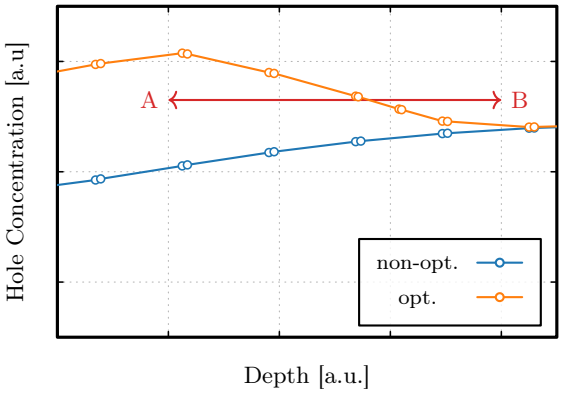
Ideally, a designer would carefully control the distribution of impurities (also known as "doping") in a semiconductor material to fully deplete the region where photons are absorbed. This would allow the charge carriers generated by the absorbed photons to drift to the region where multiplication occurs, leading to a higher probability of breakdown and a higher overall PDP. However, in practice, there are usually only a limited number of masks available to create the desired doping pattern, which can result in variations in the actual doping distribution. One way to study the impact of these variations on the detection efficiency is to use simulations (such as TCAD) to model the resulting band diagram and estimate the corresponding breakdown probability. The SPAD designs presented in this work use only unaltered standard implants inherent to the BCD process.

To simplify the analysis, the SPAD junction can be modelled in a single dimension to provide quick evaluation of a SPADs detection efficiency [86], [141]. Differential equations that describe the voltage dependent triggering probabilities are shown in Equations (3.1) and (3.2) [84]. The ionization coefficients of electrons and holes are denoted by α and β, respectively. Moreover, the probabilities that either an electron or hole initiate an avalanche are P_e and P_h, respectively.

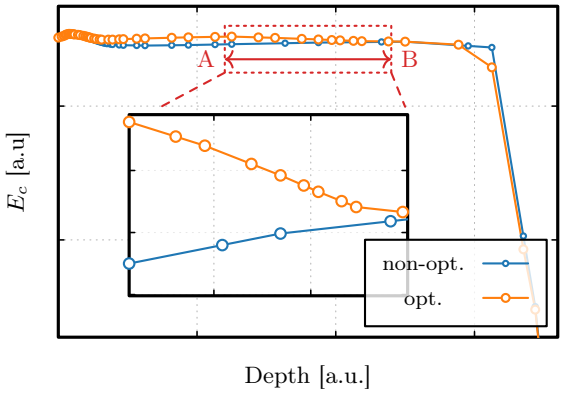
$$\frac{dP_e}{dP_x} = -(1 - P_e)\alpha(P_e + P_h - P_e \cdot P_h) \tag{3.1}$$

$$\frac{dP_h}{dP_x} = (1 - P_h)\beta(P_e + P_h - P_e \cdot P_h) \tag{3.2}$$

Over the course of multiple decades, McIntyre studied and modelled impact ionization, eventually proposing a history-dependent ionization coefficient [85], [87]. In TCAD, the McIntyre model is used to simulate these quantities along with the band diagram and doping profile. The concentration of majority carriers (holes) in both the **opt** and **non-opt** designs is compared in Figure 3.3a. Figure 3.3b shows the corresponding conduction band diagram. This depth within the device outlines the transition from the P+ to PW implants. It can be seen that the hole concentration in **opt** decreases monotonically towards the junction after a certain depth (point A), while there is a region of increased hole concentration in **non-opt**. This increase in holes is due to the net concentration caused by the profile of all the implants used. It is important to note that the depth range of interest (A-B) is outside the high-field multiplication region, which can be seen in Figure 3.4a. The energy band diagram in Figure 3.3b shows that there is an evident barrier that inhibits electron diffusion towards the multiplication region, due to photons absorbed close to the surface. In contrast, **opt**, has

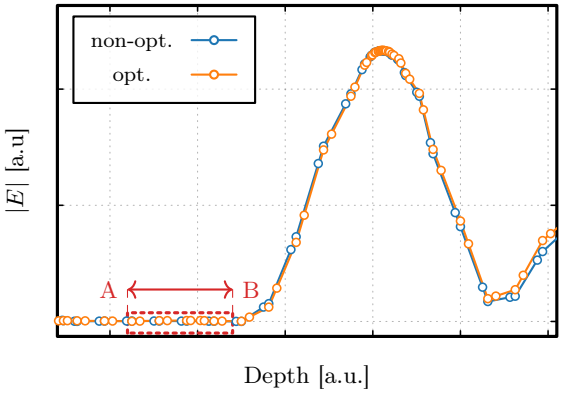


(a) Hole concentration comparison with the p+ to PW region highlighted by AB.

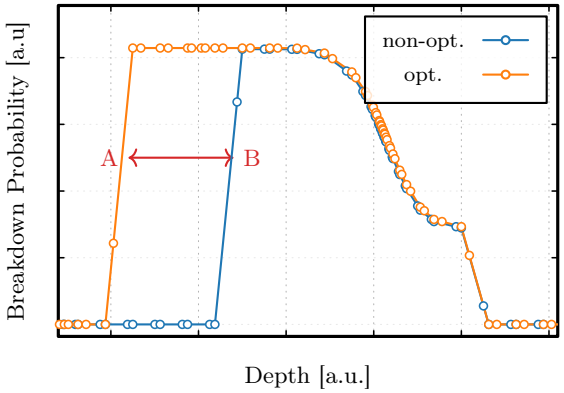


(b) Energy band diagram of the conduction band demonstrating electron barrier in non-opt.

Figure 3.3: Diagrams showing the TCAD simulation results for both the optimized (orange) and non-optimized (blue) deep junction SPADs at $V_{EX} = 5$ V. The simulation is performed after all implants are combined, i.e. with the net relative doping. The monotonic decrease in hole concentration with the addition of the PW enables electron diffusion towards the multiplication region.



(a) Electric field magnitude of SPADs as a dunction of depth.



(b) P_b simulation showing highlighting enhanced performance in opt.

Figure 3.4: Electric field of two deep junction SPADs simulated at $V_{EX} = 5$ V and the corresponding breakdown probability as a function of depth. The region AB is clearly shown to be outside the high field multiplication region. Nevertheless, the probability of avalanche from carriers generated in this region remains high.

a much wider photocollector region, allowing photo-generated electrons to easily move towards the multiplication region. As a result, the breakdown probability is greatly improved in **opt**. Simulation confirms this principle, with the combined breakdown probability of holes and electrons showing a higher likelihood of igniting an avalanche breakdown within a wider range of depths in **opt**, corresponding to a wider spectral response. Finally, the space charge region of the SPAD, when biased at $V_{EX} = 5$ V, is simulated and plotted in Figure 3.5. Clearly, from simulation, it is shown that the space charge regions and electric field magnitude of the two devices are very similar in the multiplication region. Therefore, it is concluded that the differences in measured PDP are largely due to the doping variation in the photo-collector

region.

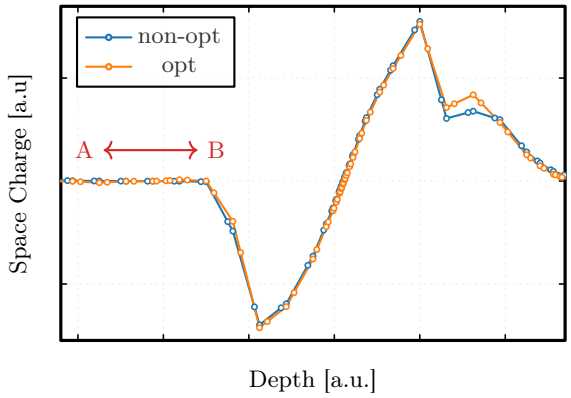


Figure 3.5: Space charge plot of both deep junction SPADs, at $V_{EX} = 5$ V. This confirms that the depleted regions for both detectors under excess bias are similar.

This relatively simple, but powerful analysis demonstrates how the breakdown probability can be engineered, by mixing different implants to engineer the photocollector region. Therefore, it clearly outlines a method amenable to design in CMOS nodes for enhanced PDP, without the addition of custom implants.

3.2.3 Measurements

Both devices are fabricated and characterized, with an emphasis on demonstrating the enhanced sensitivity of the proposed **opt** SPAD.

IV and LET

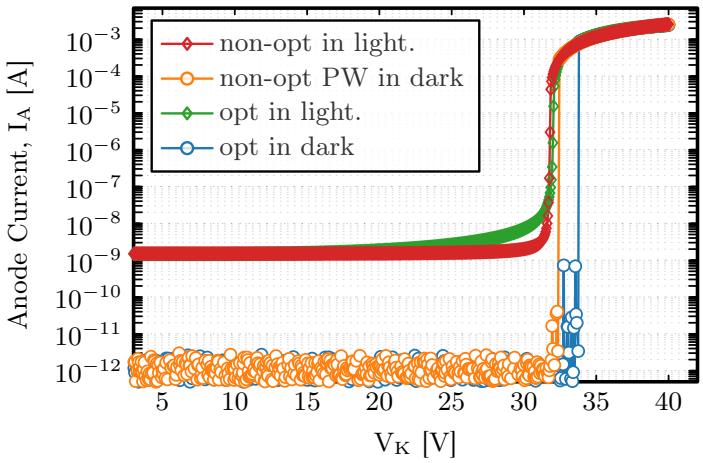
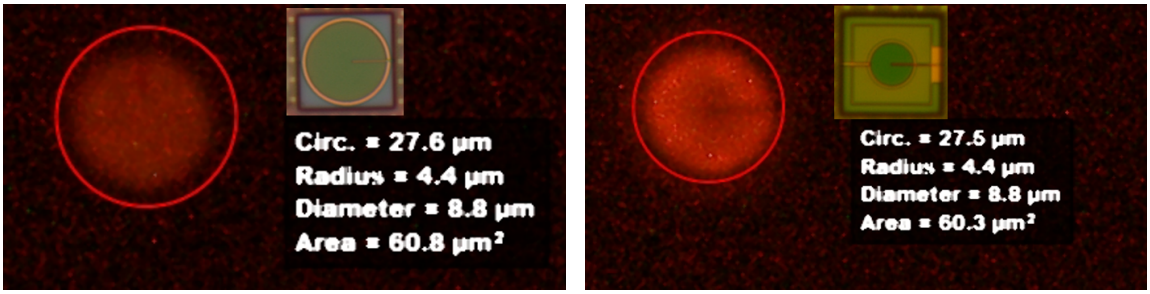


Figure 3.6: IV curve of opt and non-opt deep Junction SPADs under dark and illuminated conditions. The opt design demonstrates higher photo-current near the breakdown voltage.

To understand the diode characteristics of the SPADs, the IV-curves are first measured using the Keysight B1500A semiconductor analyzer. The results are shown in Figure 3.6 with

measurements performed under dark and illuminated conditions. A few observations are made. First, the breakdown voltages of both SPADs are very similar at $V_{BR} \simeq 32$ V. This validates that the additional PW has little impact on the doping level at the depth where the junction is formed. However, it can be seen that the photo-current close to the breakdown region of the **opt** SPAD is increased, owing to the enhanced breakdown probability. Moreover, the value of the breakdown voltage matches the expectation from simulation. To confirm



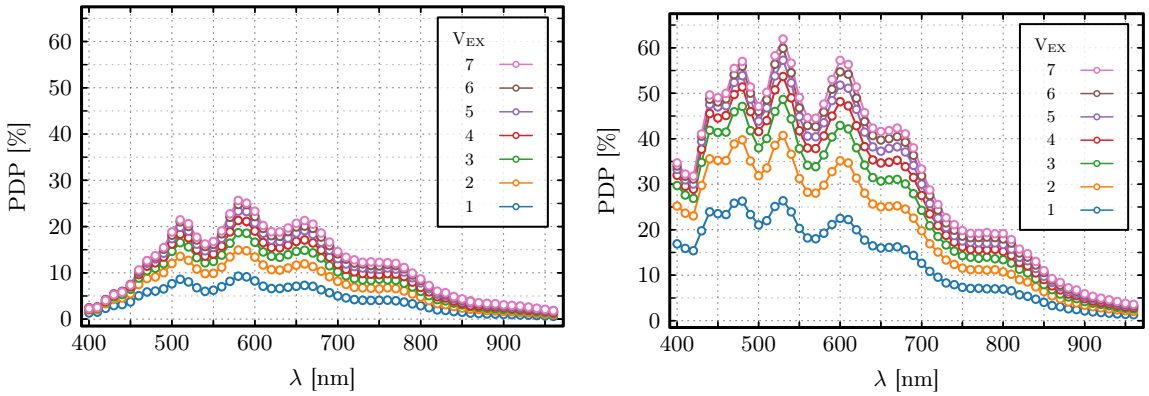
(a) non-opt. **(b)** opt.

Figure 3.7: LET results for deep junction SPADs. Good light emission uniformity around an active radius of 4 μm is observed, with no evidence of premature edge breakdown. Testing is performed at $V_{EX} = 3$ V. Micrographs of the devices are included.

the proper operation of the detector, light emission testing (**LET**) is performed. If the guard ring/periphery region is not designed properly, the SPAD can exhibit premature edge breakdown (**PEB**). This describes a situation where the edge of the multiplication region has a breakdown voltage lower than the center, resulting in considerably reduced sensitivity. LET is a rudimentary method for verifying that PEB is not an issue. As current flows through the SPAD, photons are emitted. Observing the light emission profile under a microscope can outline where breakdown is occurring. Results for both deep junction designs are shown in Figures 3.7a and 3.7b. Micrographs of the two detectors are also displayed.

PDP

PDP is measured using the continuous light method. First, a monochromator and integrating sphere select a temporally coherent and spatially uniform light source. The device under test is placed at a calibrated distance from the source. Pulses from the SPAD are counted using a universal counter (Keysight 53230A) while current through a reference photo-diode (Hamamatsu 2281) is monitored to perform the PDP calculation. A more detailed explanation and diagram of this technique is provided in [142]. The results are shown in Figure 3.8. The scales of the results for each SPAD are kept the same and placed side-by-side to highlight the clear performance difference. Due to the fact that this was the first time SPADs had ever been developed in this 55 nm process, the dielectric layers placed above the SPAD were not optimized for optical transmission. Therefore, a standing wave pattern (hills and valleys) is present in the response throughout the spectrum. The results show a significant improvement across the spectrum. The **non-opt** SPAD achieves a peak PDP of 26 % at 580 nm. In contrast, with the additional implant, the PDP is enhanced to 62 % (530 nm) at 7 V excess



(a) PDP results for non-opt SPAD

(b) PDP results for opt SPAD

Figure 3.8: PDP measurements of deep opt and non-opt SPADs measured at room temperature with $V_{EX} = 1 - 7$ V and a 220 kΩ passive quench resistor. Measurements are taken at 10 nm intervals. The resulting standing wave pattern is due to a non-optimized optical stack (dielectrics) placed above the SPADs.

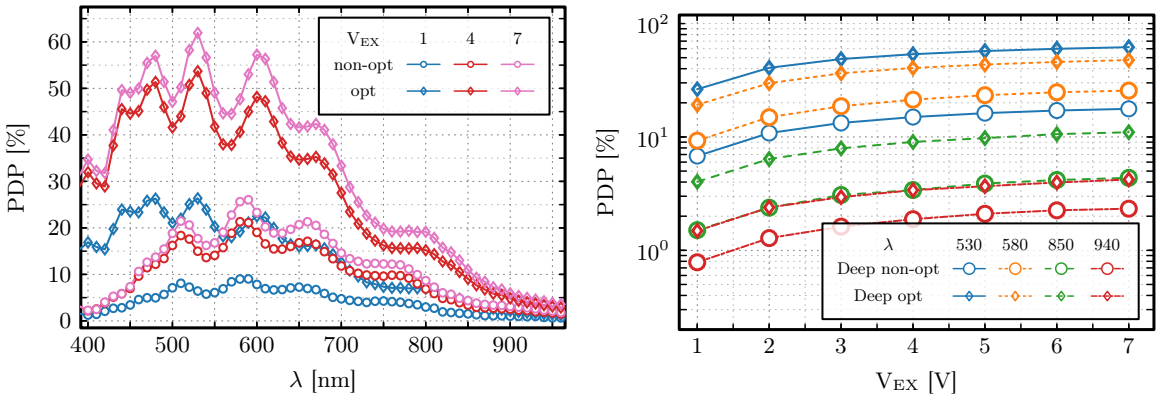
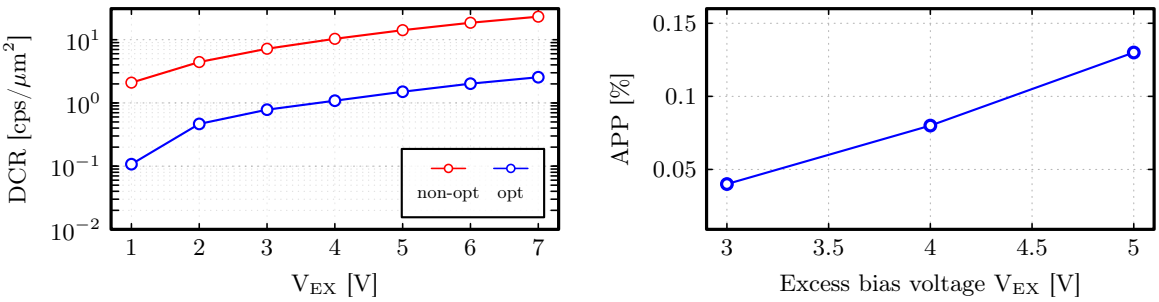


Figure 3.9: PDP comparison of deep junction SPADs. Improved sensitivity is achieved at every wavelngth. The non-opt SPAD notably has very low sensitivity at NUV and blue wavelengths, owing to the barrier for carrier transit outlined in simulation.

bias. Moreover, a value > 19 % is maintained up to 800 nm. The excellent NIR performance is particularly interesting for LiDAR applications, as the PDP at 940 nm is 4.2 %. A more direct comparison can be made by viewing Figure 3.9, where the results for each SPAD at chosen excess biases and wavelengths are plotted together. These plots highlight how the difference in PDP of the two SPADs becomes less pronounced at longer wavelengths. This is caused by lower energy photons, which are absorbed deeper inside the silicon, igniting avalanches with a similar probability. These measurements were performed without the pixel circuit so that excess bias values, up until the point where compression is observed (7 V), could be tested.



(a) DCR across excess bias of bot opt and non-opt SPADs

(b) Afterpulsing measuremend results of opt SPAD

Figure 3.10: Noise performance of deep junction SPADs in 55 nm. Ten separate devices for each SPAD were measured for DCR, with the median value plotted. Due to its superior performance, APP was measured only for the opt design.

Noise performance

The DCR for both detectors, measured at room temperature and across an excess bias range $V_{EX} = 1 - 7$ V, are shown in Figure 3.10a. An oscilloscope (Teledyne LeCroy WaveMaster 813 Zi-B) was used to count pulses for measuring DCR. Ten of each die were measured, with the median result plotted. Both SPADs demonstrate excellent noise performance. Previously, few detectors demonstrated in literature using process nodes below 65 nm were able to achieve low noise. This is largely due to the high doping levels and curtailed annealing steps in advanced CMOS nodes. However, the optimized design achieves 1 cps/μm^2 at $V_{EX} = 4$ V and 2.6 cps/μm^2 at $V_{EX} = 7$ V. Due to its superior sensitivity and noise performance, the remainder of the characterization focused on the optimized design. The inter arrival histogramming technique, as described in the previous chapter, was used for measurement of the SPAD afterpulsing. Initially, the device appeared to have no distinguishable hyper-exponential behavior. Only when the dead time was tuned to the minimum value ($\tau_{dead} \simeq 2$ ns) did afterpulsing become evident. The results are plotted in Figure 3.10b. Typically, in silicon SPADs, the APP can range anywhere from 0.1 to 10 %. Extremely low afterpulsing is present in the **opt** SPAD, staying below a value of $\alpha = 0.15\%$ up until $V_{EX} = 5$ V.

Temperature testing of DCR was performed on the **opt** design, with results displayed in two plots by Figure 3.11. A single die is chosen randomly for this measurement. The first image (Figure 3.11a) outlines the strong exponential dependence of DCR across temperature. Moreover, in this design, the low relative doping levels at the junction results in a low electric field at the breakdown voltage ($1 < MV/cm^3$). Therefore, due to the dominant thermal generation component of DCR, an Arrhenius plot can be used for extraction of the activation energies [143]. From Figure 3.11b, the extracted activation energy is in the range of $\simeq 38 - 46$. Traps, with activation energies with this value, have been shown to be caused by phosphorus vacancies in the silicon [144], [145].

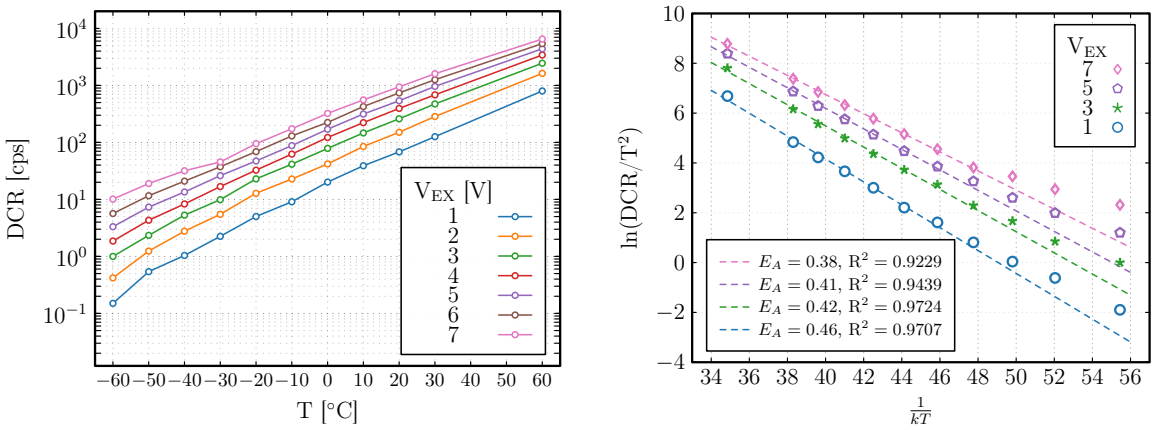


Figure 3.11: Temperature dependence of deep opt SPAD in 55 nm BCD process. Measurements are performed across an excess bias range of $V_{EX} = 1 - 7$ and a temperature range of $-60°$C to $60°$C. The Arrhenius plot is shown, highlighting that trap assisted thermal generation is the dominant contributor to DCR until low temperatures, where tunneling becomes more significant.

Timing Jitter

Timing performance was characterized using a femtosecond laser and a fast reference photodiode at a wavelength of 780 nm. The components of the setup and measurement procedure are described in [142]. The measurement results of the FWHM, for an excess bias range $V_{EX} = 1 - 5$ V, is shown in Figure 3.12. The device achieves 30 ps FWHM at $V_{EX} = 5$ V, an improvement from 96 ps FWHM at $V_{EX} = 1$ V. The mean time to breakdown, along with the current density of the avalanche, increase with increased excess bias, explaining the improved timing performance [146], [147].

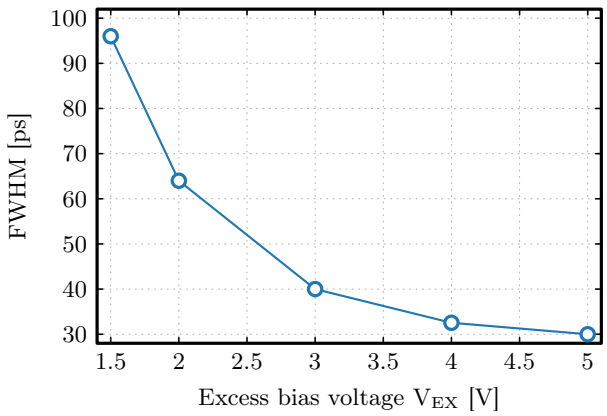
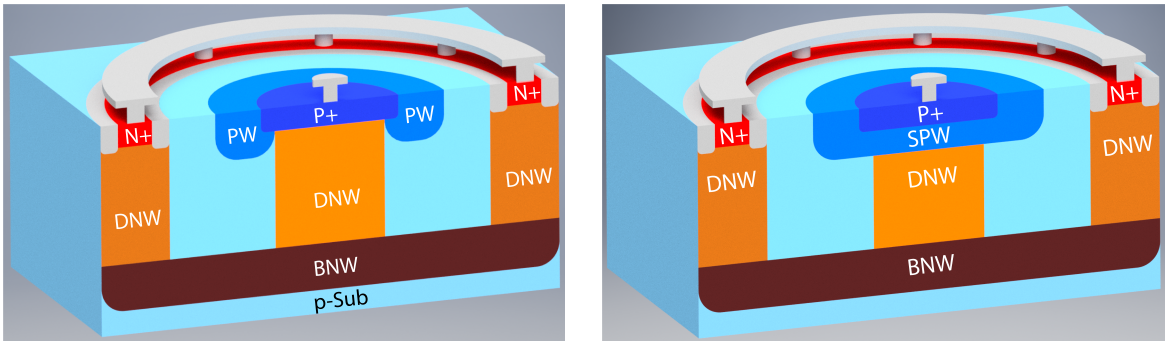


Figure 3.12: Single-photon timing jitter (FWHM) of deep opt SPAD, measured at $\lambda = 780$ nm across an excess bias range of $V_{EX} = 1 - 5$ V.



(a) A shallow device with junction formed by the p+/DNW interface. Device refferred to as SJ1.

(b) Shallow SPAD with enhanced depletion region using SPW implant. Device refferred to as SJ2.

Figure 3.13: 55 nm BCD shallow junction cross-sections. SJ1 has an abrupt junction formed using the p+/DNW interface. A PW implant is used as an explicit guard ring. SJ2 improves the detection efficiency of this junction with the addition of a shallow p-well (SPW) implant.

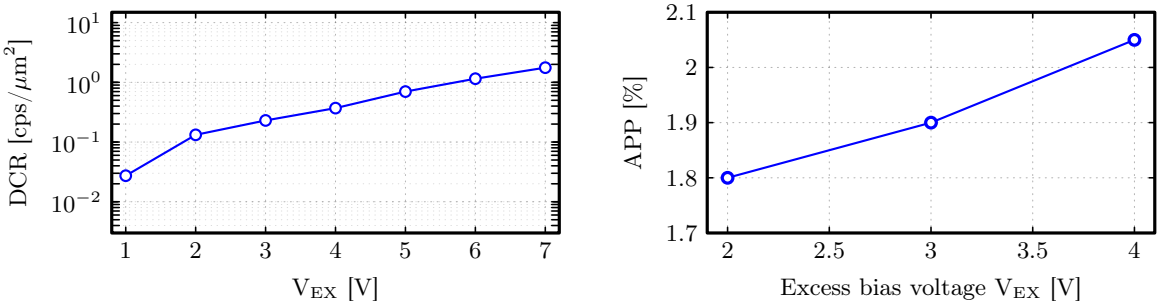
3.3 Shallow Junction SPADs

The shallow junction SPAD characterized in this work are shown again in Figure 3.13. The first device, **SJ1**, was fully characterized, including afterpulsing and jitter measurements using a pixel circuit. Moreover, the effects of pixel scaling on two geometric variations of SJ1 was also studied. A second device, **SJ2**, was then fabricated with an additional implant (SPW), to extend the depletion region width, thereby improving the PDP. The breakdown voltages of SJ1 and SJ2 are $\simeq$ 18 V and $\simeq$ 20 V, respectively. The small increase in breakdown voltage is due to the lightly doped SPW layer added in SJ2. Moreover, these V_{BR} values are considerably lower than the 32 V values for the deep junction, which is advantageous for applications requiring simplified systems (high voltage generation). Overall, the devices are compared at the end of this chapter.

3.3.1 SJ1

Noise Performance

The measured dark count rate and afterpulsing probability of a single SJ1 sample is shown in Figure 3.14. The results show very low DCR, remaining below 1 cps/μm^2 until $V_{EX} = 6$ V. However, the extracted afterpulsing probability is considerably higher when compared to the deep junction devices. APP is 1.8 % at $V_{EX} = 2$ V and 2.05 % at $V_{EX} = 4$ V. Previously, it has been shown that multiplication regions that are close to oxide layers can have increased afterpulsing, due to defects introduced at the oxide interface [103]. Therefore, this could be a potential cause of the increase in APP, despite the low noise performance. Defects caused from the **DNW** layer, which is not present at the junction of the deep SPADs could also contribute to the increase in APP, although more investigation would be required to understand the principal cause(s). Nevertheless, $\simeq$ 2 % APP is acceptable for many applications.



(a) DCR of a single die measured at room temperature

(b) Afterpulsing probability extracted from an inter-arrival histogram.

Figure 3.14: Noise performance of an SJ1 SPAD with an active radius of 4.5 μm, measured across excess bias.

PDP

Using the same measurement setup/methodology for the deep junctions, the PDP of an SJ1 SPAD with an active radius of 4.5 μm was measured, with the results shown in Figure 3.15. Near ultra-violet (**NUV**) and blue photons, which have a shallow penetration depth, are detected with a higher efficiency, as expected. The peak achieved PDP, at an excess bias value $V_{EX} = 7$ V, is 25 % at 430 nm. When compared to the optimized deep junction device, the PDP, at the same wavelength and excess bias value, is lower by $\simeq$ 5 %. Both devices show have extremely low noise performance, although, the afterpulsing of SJ1 is considerably higher. Therefore, the principal advantage of SJ1 compared to the opt deep SPAD is its lower breakdown voltage.

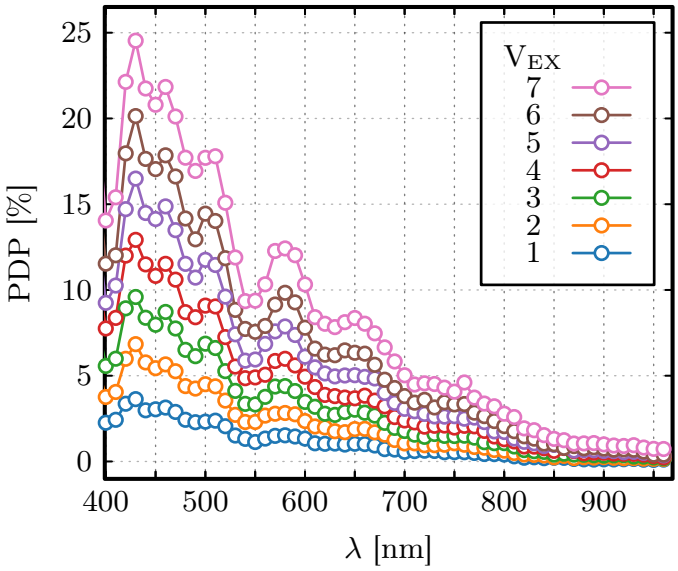
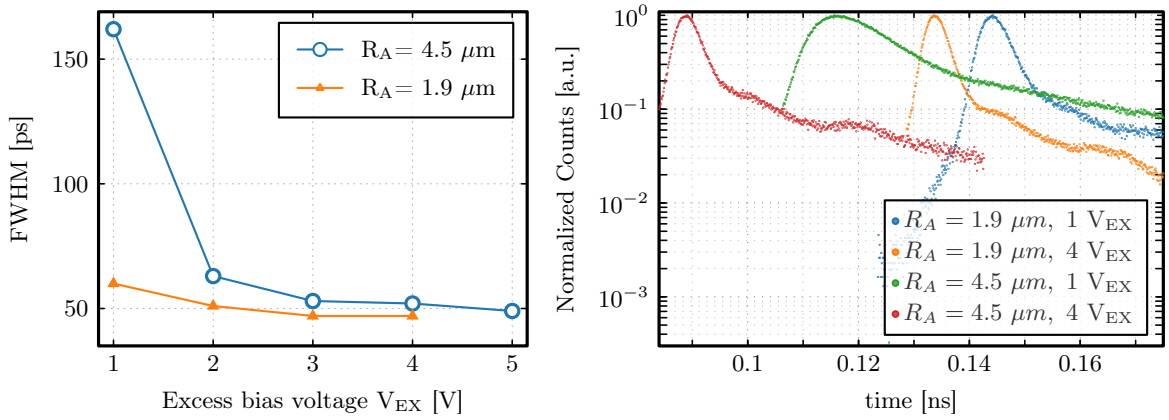


Figure 3.15: PDP measurements of the SJ1 SPAD with an active radius of 4.5 μm, at room temperature, with an excess bias range of $V_{EX} = 1 - 7$ V at 10 nm wavelength intervals. Quenching performed with a 220 kΩ resistor.

Timing Performance

As described in Chapter 2, the timing performance of SPADs is dependent on many factors such as doping, SPAD geometry, and the SPAD circuit interface. The timing performance for two separate SJ1 SPADs, with active radius $R_A = 1.9$ μm and $R_A = 4.5$ μm, were measured with the results shown in Figure 3.16. At higher excess bias values, the FWHM of the timing jitter is very similar for both devices. However, at $V_{EX} = 1$ V, the larger device has achieves 160 ps whereas the smaller device results in $\simeq$ 60 ps. This demonstrates well the effects of excess bias and device geometry on the avalanche build-up and spread. Evidently, given an initial photon injection position, the later propagation time is smaller for a SPAD with smaller active radius. The carrier multiplication rate (and current) increases with increased excess bias, reducing the dependence of the lateral propagation time on the injection position. [148]. Similar results to those shown in Figure 3.16a, where two shallow junction devices with different active radius had their timing performance characterized, were observed in [149].



(a) FWHM results summary for two SJ1 devices.

(b) Timing jitter histograms at 1 and 4 V_{EX}.

Figure 3.16: Timing performance of two separate SJ1 SPADs with active radius $R_A = 1.9$ μm and $R_A = 4.5$ μm. At low excess bias, the timing performance of the larger SPAD is significantly larger.

3.3.2 SJ2

Earlier, the effects of adding a single implant to increase the breakdown probability of deep SPADs was demonstrated. Here, a similar methodology is used, but this time for the purposes of extending the drift region of collected photons that penetrate deeper into the silicon. This is done with the addition of the SPW implant, which fully depletes an extended region of depth within the DNW layer. Although the electric field is narrow, as is the case for SJ1, it remains high enough so that the drift region is extended until the BNW layer. Therefore, as the excess bias increases, the PDP should continue to increase until the saturation of ionization coefficients in silicon [150].

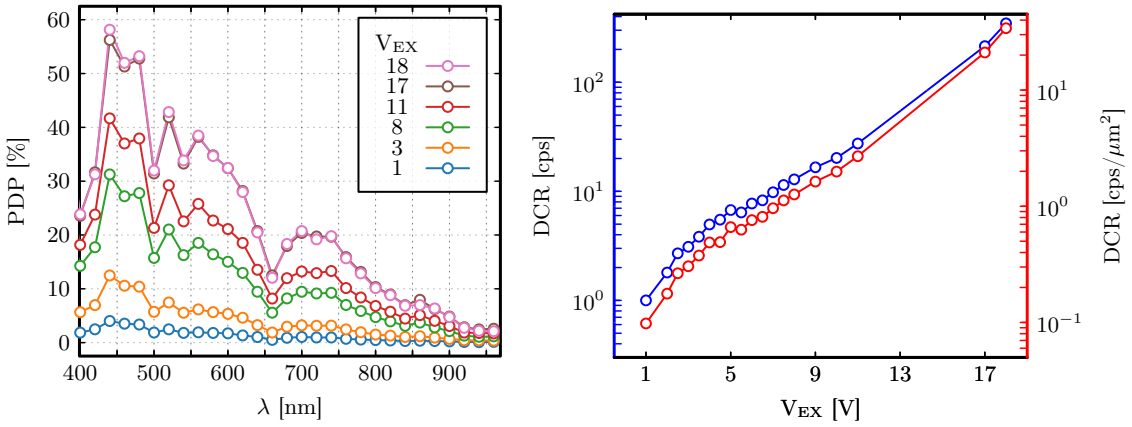


Figure 3.17: DCR and PDP measurements for the fully depleted SJ2 shallow junction SPAD in 55 nm BCD. The results show an increase in PDP until an excess bias voltage $V_{EX} = 18$ V, achieving $\simeq$ 59 % at 440 nm.

Measurements

PDP and DCR measurements for a SJ2 device with an active area radius of 1.8μm are displayed in Figure 3.17. It can be seen that saturation of the PDP does not occur until $V_{EX} = 18$ V, where the maximum measured value is $\simeq$ 59 % at 440 nm. This is an excellent result, when considering the dark count rate, shown in Figure 3.17b does not increase past $\simeq$ 34 cps/μm^2 at the same excess bias.

3.4 Comparison

Table 3.1 outlines the figures of merit of the SPADs presented in this work, compared to other silicon FSI SPADs that have been demonstrated in literature. This table is ordered chronologically to provide a simple way of visualizing advancements over time. Moreover, a comparison of detector DCR vs PDP, at the maximum measured excess bias, is shown in Figure 3.18. The improved SPADs, both for the deep (opt) and shallow (SJ2) junctions, demonstrate some of the best performing FSI detectors published in the literature in terms of noise and DCR. Moreover, these designs were implemented using only existing implant profiles in the process. For SPADs presented in a technology process node below 65 nm, the achieved PDP and DCR are best in class, to the author's knowledge. Moreover, from the application standpoint the deep junction designs are excellent candidates for QRNG detectors, considering their ulta-low afterpulsing ($< 0.2\%$). The pixel circuit presented in this chapter had a minimum measurable dead time of $\tau_{\text{dead}} \simeq 1.5$ ns. This fortifies the claim of low afterpulsing, and enables fast single pixel random bit generation rates, presented in the next chapter.

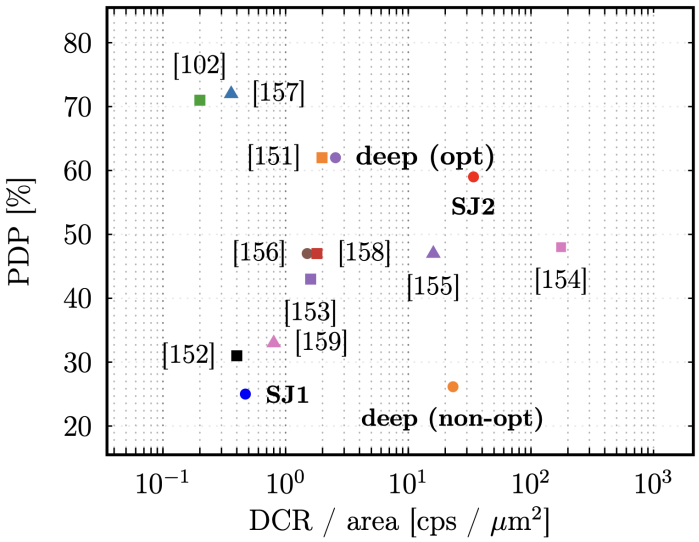


Figure 3.18: Comparison of recently published silicon SPAD performance in terms of DCR and PDP.

Table 3.1: Comparison table of Silicon SPADs in literature.

Reference	Year	Tech. [nm]	Junc.	Guard Ring	Active Diam. [μm]	V_{EX} / V_{BD} [V]	Peak PDP [%] @λ [nm]	PDP [%] @940 [nm]	DCR / area [cps / μm^2]^{af}	DCR @V_{EX}^{f} [V]	AP [%]	Jitter [ps]	Jitter λ[nm]	Jitter V_{EX} [V]
Gersbach et al. [151]	2008	130	p+/nw	STI^{m}	8.6	1-2 /9	18-30 @480	2	1.5k - 11.5k	670k @2	<1^{g}	125	637	1
Richardson et al. [152]	2009	130	pw/ DNW	N.A.^{c}	8	0.6-1.4 /14	18-28 @500	2	0.24 - 0.6	30 @1.4	0.02^{h}	200	815	N.A
Richardson et al.^{y} [153]	2011	130	pw/ DNW^{v} p+/nw	N.A.^{c}	8^{x}	0.2-1.2 /12-18	18-33 @450-560	2-3	0.5-0.97	40-47 @0.8	0.02^{i}	183 - 237	470	1.2
Webster et al. [154]	2012	90	DNW/ p-epi^{k}	N.A.^{c}	6.4	2.46 /14.9	44 @690	12	~ 4.6	~150 @1	0.38^{l}	51	470	2.46
Webster et al. [155]	2012	130	p-epi^{k}	N.A.^{c}	8	2-12 /20	72 @560	12	0.36	18.0 @2	<4	60	654	12
Leitner et al. [156]	2013	180	p+/nw	N.A.^{c}	10	1-3.3 /21	35-47 @450	N.A.	0.3-1.8	~180 @4	N.A.	N.A.	N.A.	N.A.
Charbon et al. [157]	2013	65	n+/pw	nw	8	0.05-0.4 /9	2-5.5 @420	0.2	340-15.6k	105 @.05	1^{n}	235	637	0.4
Villa et al. [158]	2014	350	p+/nw	pw	10-500	2-6 /25	37-55 @420	2	0.05	1.0 @4	1^{o}	56-4470	780	6
Veerappan et al. [159]	2014	180	p+/nw	pw	12	2-10 /23.5	24-48 @480	N.A.	0.16-12.8	20 @2	0.03-0.3^{p}	86	637	10
Lee et al. [160]	2015	140^{d}	p+/nw	pw	12	0.5-3 /11	10-25 @500	1-3	0.9-244	30k @3	1.7^{p}	65	405	3
Veerappan et al. [161]	2015	180	p+/nw	pw	12	1-4 /14	23-47 @480	N.A.	16	2k @4	0.2^{q}	95	405	4
Veerappan et al. [162]	2016	180	p-epi/ DNW	N.A.^{c}	12	1-12 /25	18-47 @520	N.A.	1.5	200 @11	7.2^{q}	97	637	11
Takai et al. [163]	2016	180	n/p^{r}	N.A.^{c}	16	1.5-5 /20.5	62 @210	8	0.5	100 @5	0.35	161	635	5
Pellegrini et al. [164]	2017	130	pw/ DNW	N.A.^{c}	8	0.5 /14.2	43 @480	1.4	1.6	80 @0.5	0.08	100	N.A.	2.4
Xu et al. [165]	2017	150	p+/nw	N.A.^{c}	10	5 / 18.0	31 @450	2	0.4	200 @5	0.85	42	831	4
Pellegrini et al. [122]	2017	40	pw/ DNW	N.A.^{c}	N.A.	1.0 /15.5	45^{s} @500	3	N.A.	50 @1.0	0.1	140	850	1

Continued on next page

Table 3.1 – continued from previous page

Reference	Year	Tech. [nm]	Junc.	Guard Ring	Active Diam. [μm]	V_{EX} / V_{BD} [V]	Peak PDP [%] @λ [nm]	PDP [%] @940 [nm]	DCR / area [cps / μm^2]$^{a\,f}$	DCR @V_{EX}^{f} [V]	AP [%]	Jitter [ps]	Jitter λ[nm]	Jitter V_{EX} [V]
Sanzaro et al.^{y} [102]	2018	160^{e}	p+/n p/ DNW	N.A.^{c}	10-80	3-9/ 25-36	58-71 @450-490	3	0.1-0.2	100 @5	0.02-1.59	28-41	820	5
This Work (opt)	2021	55^{e}	DPW/ BNW	N.A.^{c}	8.8	7/ 31.5	62 @530	4.2	2.6^{a,f}	156 @7	$\sim$0.13^{t}	30	780	5
This Work (SJ1)	2021	55^{e}	p+/ DNW	PW	9-3.8	7/18	25 @430	0.8	0.47^{a,f}	30 @8	$\sim$2$^{\beta}$	52	780	5
This Work (SJ2)	2021	55^{e}	SPW/ DNW	N.A.^{c}	3.6	18/ 20	60 @440	2.6	34^{a,f}	350 @18	NA	230$^{\gamma}$	780	10

^{a}Taken at max excess bias if not range of excess bias values not specified. ^{c}Virtual guard ring structure. ^{d}Silicon-on-insulator. ^{e}BCD. ^{f}At 20°C. ^{g}180 ns dead time. ^{h}200 ns dead time. ^{i}50 ns dead time. ^{k}Not substrate isolated. ^{l}15 ns dead time. ^{m}Shallow trench isolation with passivation implants to create p-type glove structure. ^{n}At 1 μs dead time. ^{o}30 μm diameter. ^{p}200 ns dead time. ^{q}300 ns dead time. ^{r}Surface-isolated n-spad/p-spad junction. ^{s}With microlens. ^{t}4.5 ns dead time with 50 % level @ 3 V_{EX}. ^{v}Two different deep structures presented one with an epi layer and one with a pw implant. ^{x}Multiple diameters demonstrated. ^{y}3 SPAD structures proposed. $^{\beta}$at 1.5 ns dead time and 4 V excess bias.$^{\gamma}$at 10 V excess bias, measured with external load.

4 Modelling and design of random bit generators in 55 nm

The parts of this chapter which pertain to correlation modelling and the SC QRFF have been placed on arXiv:2209.04868 [166] and has been accepted to IEEE Journal of Solid State Circuits for publication (DOI: 10.1109/JSSC.2023.3274692). The FA circuit presented is in preparation for the IEEE Journal of Solid-State Circuits Letters (SSC-L).

In this chapter, several methods for generating random bits by exploiting the random arrival times of photons, are discussed. A heavy emphasis is placed on the slow clock method, due to its architectural simplicity, which allows for modelling of the degradation in entropy, caused by device and circuit imperfections. Verilog-A models of the SPAD and bit generation methods are developed so that long bias and correlation simulations, that generate enough statistics, can be run in the Cadence analog mixed-signal (**AMS**) simulator. Moreover, the first photon arrival method, implemented as a first SPAD to detect a photon in a pair of independent pixels, is also discussed. This method has been popular in literature, and it contrasts well in terms of performance as well as general advantages and disadvantages for array implementation. These two methods will be referred to as the slow clock (**SC**) and first arrival (**FA**) circuits. QRFF realizations, containing a SPAD, pixel, and random bit generation circuitry are implemented and characterized for each method in 55 nm BCD. The SPAD used is the previously presented deep junction opt SPAD (Chapter 3). Two key innovations are presented. First, a new method to analytically calculate the correlation of the slow clock method when considering dead time and afterpulsing is developed. This is validated in simulation, and to a certain extent, in measurements. Second, the simulation of the bias model led to the development of a novel clocked comparator based implementation of the slow clock method, which can eliminate bias in a single pixel.

4.1 Counting statistics of SPAD detectors

Before diving into the specifics of bit generation methods, an overview of SPAD counting statistics, including recent analytical advances, are outlined. This will then facilitate the modelling of the proposed methods.

Nomenclature of the counting equations are slightly modified so that they can be more easily linked to bit generation parameters. Therefore, the bit generation period, i.e. some integration period of arbitrary time during which photons are counted, is referred to as T_{BG}. Moreover, the arrival rate, which constitutes SPAD photon detections in the absence of dead time, is denoted as λ_{A}. Therefore, the probability of receiving k detections in an integration interval $(0, T_{\mathrm{BG}})$ is defined by the well known Poisson distribution (4.1).

$$p(k, \lambda_{\mathrm{A}}) = \frac{(\lambda_{\mathrm{A}} \cdot T_{\mathrm{BG}})^k}{k!} e^{-\lambda_{\mathrm{A}} \cdot T_{\mathrm{BG}}} \tag{4.1}$$

The arrival rate can be estimated with the PDP of the detector and the optical power, P_r, of the impinging signal.

$$\lambda_{\mathrm{A}} = \frac{\mathrm{PDP} \cdot \mathrm{P}_r}{hf} \tag{4.2}$$

However, in practice the SPAD dead time obscures these statistics, and as explained in Chapter 2, the type of pixel circuit also influences dead time. Several models for calculating the detection rate (λ_{D}), i.e. the rate of avalanches based on the dead time, have been proposed. From nuclear particle physics instruments, it was shown that the count rate of a paralysable detector can be calculated with [167].

$$\lambda_{\mathrm{D}} = \lambda_{\mathrm{A}} \cdot e^{-\lambda_{\mathrm{A}} \cdot \tau_{\mathrm{dead}}} \tag{4.3}$$

A more accurate calculation, which differentiates between the avalanche generation time, where the detector is non-paralysable, and the recharge time, was introduced as the hybrid model [168], [169]. In this case the total (saturation) dead time is denoted as τ_{sat} and the paralysable dead time as τ_{par}

$$\lambda_{\mathrm{D}} = \frac{\lambda_{\mathrm{A}} \cdot e^{-\lambda_{\mathrm{A}} \cdot \tau_{\mathrm{par}}}}{1 + \lambda_{\mathrm{A}} \cdot \tau_{\mathrm{sat}}} \tag{4.4}$$

In the actively quenched scenario, where τ_{par} becomes negligible, the detected count rate can be calculated as:

$$\lambda_{\mathrm{D}} = \frac{\lambda_{\mathrm{A}}}{1 + \lambda_{\mathrm{A}} \cdot \tau_{\mathrm{sat}}} \tag{4.5}$$

Recently, new counting equations for paralysable and non-paralysable SPADs, developed using renewal theory, were proposed in [76]. The notation is annotated in this thesis for consistency with the bit generation circuits presented further on in the chapter.

$$p_{\mathrm{paralyse}}(k, \lambda_{\mathrm{A}}, \tau_{\mathrm{dead}}) = \sum_{i=k}^{k_{\max}-1} (-1)^{i-k} \binom{i}{k} \cdot \frac{\lambda_{\mathrm{A}}^i (T_{\mathrm{BG}} - i \cdot \tau_{\mathrm{dead}})^i}{i!} \cdot e^{-i\lambda_{\mathrm{A}} \tau_{\mathrm{dead}}} \tag{4.6}$$

$$p_{\text{nonparalyse}}(k, \lambda_k, \tau_{\text{dead}}) = \sum_{i=0}^{k} \frac{\lambda_{k+1}^{i}(T_{\text{BG}} - (k+1) \cdot \tau_{\text{dead}})^{i}}{i!} \cdot e^{-\lambda_{k+1}} - \sum_{i=0}^{k-1} \frac{\lambda_k^i}{i!} \cdot e^{-i\lambda_k} \tag{4.7}$$

where the dead time is τ_{dead} and $k_{\max}$ is defined as the maximum number of rising edges that can fall within an integration window given a deadtime i.e. $k_{\max} = T_{\text{BG}}/\tau_{\text{dead}} + 1$. For the non-paralysable case, λ_k is defined as $\lambda_k = \lambda_{\text{A}}(T_{\text{BG}} - k \cdot \tau_{\text{dead}})$.

More recently, counting equations for active quenched SPADs were also proposed in [170]. This rigorous analysis considered several case scenarios, such as the probabilities of whether the detector is sensitive or dead at the beginning of the counting window. Furthermore, in their analyses, afterpulses and twilight pulses were augmented to the counting equations. Those counting equations, which extend to a length of a full page, are not reprinted here, for brevity. The availability of these equations, which accurately predict the probability that a number, k, photons have arrived in the period chosen where a bit must be generated, T_{BG} allows for more accurate modelling of correlation, particularly for the slow clock method. This analysis is shown further on in this chapter.

4.2 Overview of methods

Many techniques, which exploit photon timing statistics, have been implemented as QRNGs. In this thesis, methods which are amenable to array implementations are considered. Some interesting works have shown the possibility of generating multiple bits per photon by timestamping the arrivals [174], [175]. However, this produces correlated bits that require post-processing. This method can be improved, by making sure that only one photon is present in the time interval, which then makes the distribution of arrival times uniform [176]. In [177], a robust algorithm that discarded time windows that contained 0 or more than 1 photon detections (also used in [178]), was introduced. They achieved four generated bits per one detected photon. However, bit generation circuits that require discarding of bits are not suitable for array implementation, as they result in significantly increased system complexity for combining pixels to achieve a consistent total data rate. Moreover, the fast clock method, which uses the random arrivals as the clocking signal, is asynchronous. Therefore, the practicality of synchronizing an array of individual asynchronous bit generators, is very low due to the uncertainty of the moment at which a valid bit has been produced. It is the author's view that the most scalable single-pixel random bit generators based on photon-timing statistics are those summarized in Table 4.1. The **slow-clock** method [40], [179], is a flexible solution that has its bit generation rate adjustable based on the photon flux. Moreover, this design's sensitivity to changes in flux is excellent so long as a minimum λ_{D} is maintained (will be detailed further on). This versatility is conducive to array implementation.

To the author's knowledge, prior to the work performed over the course of this research, a monolithic array with SPADs and circuits on-chip using the slow clock method had not been explored. The principal disadvantage of this method is that it requires continuous operation, i.e. on average, multiple photons are required per generated bit. In contrast, if a time of arrival arbiter is used to generate a bit based on the FA method, as demonstrated in [180], [181], then free running operation can be avoided. This is advantageous from a power consumption standpoint. Moreover, in principle, k, comparisons can be performed between n SPADs, increasing the bit per pixel efficiency to $\binom{n}{k}$. However, the principal drawback of this technique is bias. Considering negligible arbiter (comparator) offset, the bias from the FA method is determined by the difference in count rates between two detectors [178], [181], [182]. In principle, this method should not generate serially correlated bits. However, to-date, there has not been extensive analytical correlation model of this method demonstrating this to be true. Moreover, if a photon is not present in the correlation window then there will be serial correlations. For context, to achieve the bias benchmark of 10^-3, the count rate separation between two SPADs must be $|\lambda_2 - \lambda_1|/\lambda_1 = 0.4\%$. Finally, the time difference, between consecutive photons, can be measured, to see whether a 1 or 0 should be generated. This presents a challenge for synchronizing, as three detections must be made in order to generate a bit. However, this can be overcome by placing an evaluation window (bit generation clock) that is long enough such that the probability of having detected three photons is very high. Moreover, in [173], a ranking algorithm was presented to continually compare timestamps of previous arrivals so that the bit per photon rate could be increased (in principle to 1 photon per bit). The main drawback of the time difference method is that it requires per pixel time tagging devices. Analog processing can be used to overcome this [172], although the pixel

Method	b (theoretical)	a_1 (theoretical)	bits/ photon
Slow clock	$0^\star$ $b = \frac{t_f - \eta(t_r + t_f)}{2} \cdot {\lambda_\mathrm{D}}^\gamma$	$e^{-2\cdot\lambda_\mathrm{D}/f_\mathrm{BG}}$ * $\mathrm{R}_{XX}(T_\mathrm{BG}) = \mathbb{P}(K_\mathrm{even}) - \mathbb{P}(K_\mathrm{odd})^\chi$	$< 1^\dagger$
First arrival photon	$\lambda_1/(\lambda_1 + \lambda_2)^\beta$	0	1^ζ
Time-difference between photons	0	0	$< 1 \ddagger \alpha \bowtie$

* Ideal model.
$^\gamma$ Equation first introduced in [171], explored further in this work.
$^\chi$ Developed during the course of this research.
$^\beta$ Ignoring circuit imperfections such as comparator offset.
$^\zeta$ multiple comparisons can increase bits per photon.
† Free-running operation required i.e. multiple photons needed per bit, relatively higher power consumption compared to other methods.
$^\alpha$ Requires per pixel TDC.
$^\bowtie$ [172] used analog processing to avoid TDC for optimized area.
‡ Ranking algorithm can improve the bit/per photon ratio as done in [173].

Table 4.1: A comparison table of SPAD-based random bit generation techniques that are considered for array implemenation.

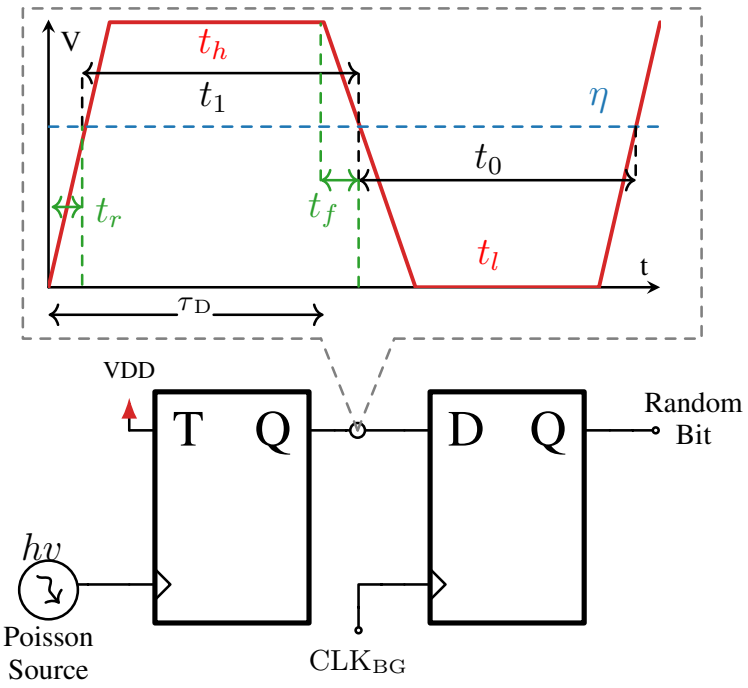


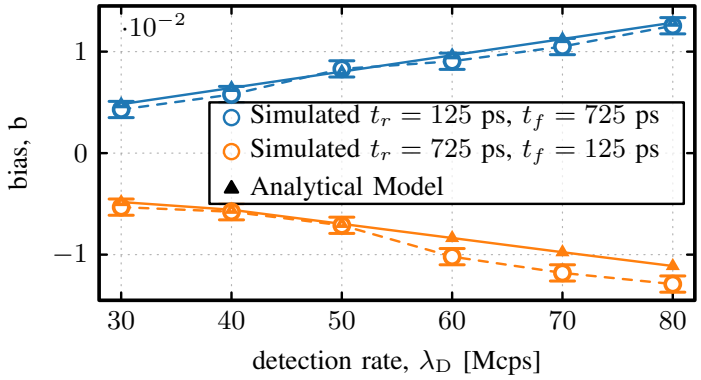
Figure 4.1: A realization of the slow-clock QRFF circuit with a source that has ideal Poisson arrival times. A realistic waveform of the output of the TFF is shown, given electronics with finite rise/fall times. These times are denoted as t_r and t_f, respectively. Edge transitions happen, on average, at intervals decided by the detection rate, i.e. $\lambda_D = 1/\tau_D$. The normalized sampling threshold, i.e. the point at which the sampling DFF determines the signal to be a zero or one, is highlighted by η. A bit is generated upon the arrival of the clock signal (CLK_{BG})

complexity is still relatively high. The formulation and meaning of bias and correlation equations shown, in Table 4.1, for the SC method, are discussed in the following sections.

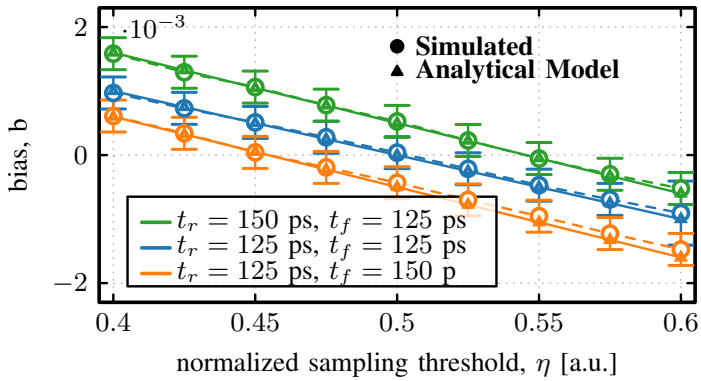
4.3 Comprehensive modelling of SC QRFF

Bias model

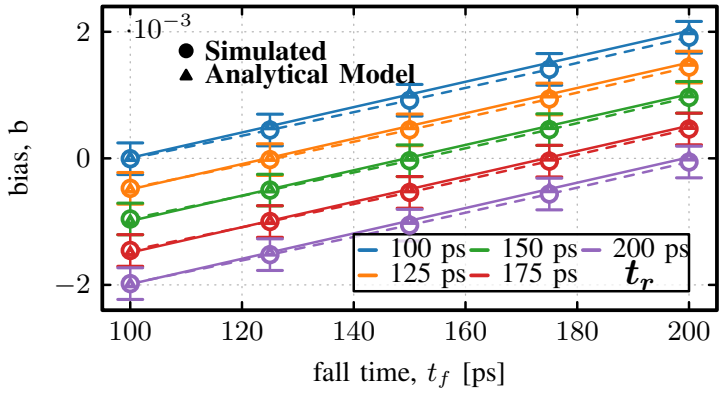
The simple circuit that is a QRFF realization of the slow-clock method, as discussed in Chapter 1, is shown here again in Figure 4.1, although this time with a realistic waveform of the TFF output, which will be referred to as TFF_Q. Assuming a Poisson source with exponentially distributed inter-arrival times, the output of the TFF_Q will, over a sufficient integration period, result in a uniform distribution i.e. detections split TFF_Q into two equal half-periods. Therefore, TFF_Q is a digital random process reminiscent of the random-telegraph signal. Given this ideal case, bits generated by the sampling flip-flop would exhibit zero bias. However, given the potential for asymmetric rise and fall times (t_r, t_f), along with a sampling threshold (η) that deviates from the normalized center, TFF_Q can spend unequal times in the high and low states, which manifests as bias. In [171] a bias equation, based on the waveform shown in Figure 4.1, was presented and is shown in Equation (4.8).



(a) Simulation of bias vs detection rate. The difference between rise and fall times are deliberately exaggerated to emphasis bias, which scales linearly with increased detection rate.



(b) Simulation showing that adjusting of the normalized sampling threshold, η, can result in elimanted bias for unmatched rise and fall times.



(c) Simulation showing bias from mismatched rise and fall times of at a constant η.

Figure 4.2: Verilog-AMS simulation of bias model of the slow-clock QRFF circuit using the bias model presented in [171]. An ideal source, which generates exponentially distributed inter-arrival times, are used for the simulation. The results demonstrate that η can be used to eliminate bias caused by mismatched rise and fall times.

$$b = \frac{t_f - \eta(t_r + t_f)}{2} \cdot \lambda_{\mathrm{D}} \tag{4.8}$$

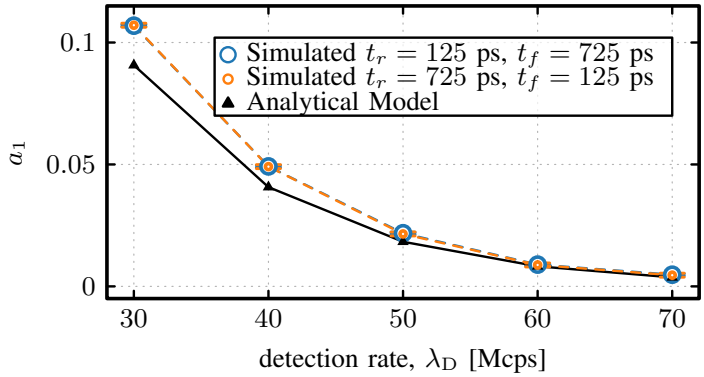
To validate this bias equation, and to understand the bounds of TFF circuit performance, which results in acceptable random bit generation, a Verilog-A models of the SPAD, TFF and DFF, with controllable variables for the parameters in Equation (4.8), was developed. Three separate simulation modes are run, with the results shown in Figure 4.2. Binary bit generation is a Binomial process with N trials. Therefore, variance of bias from simulation is calculated with $\sigma^2 = 1/(4N)$. The error bars shown in the results denote $\pm\sigma$. Key propositions from the bias model are validated in these simulations. First, it is clear that a mismatch in rise and fall times results in bias. Acceptable values for t_r and t_f are in the range of 150 ps or less, depending on the magnitude of the mismatch. Perhaps the most relevant insight from simulation is the ability to compensate for bias by adjusting the sampling threshold. This can be viewed as a form of post-processing, although, in contrast to other entropy correction methods, **this technique does not require combining multiple independent sources**. Furthermore, this bolsters a critical advantage, which is that bias can be reduced even as the photon flux increases, which enables faster bit generation. The limitations of this are discussed in the following section.

Ideal correlation model

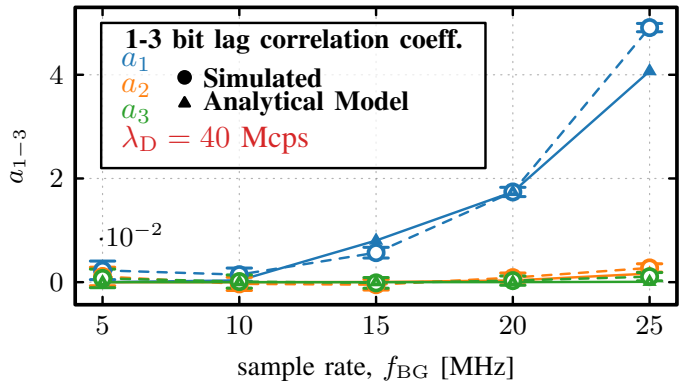
Autocorrelation of a binary RTS is well-defined in statistics [176], with the equation shown in (4.9).

$$a_i = e^{(-2\cdot\lambda_{\mathrm{D}}/(i\cdot f_{\mathrm{BG}}))}. \tag{4.9}$$

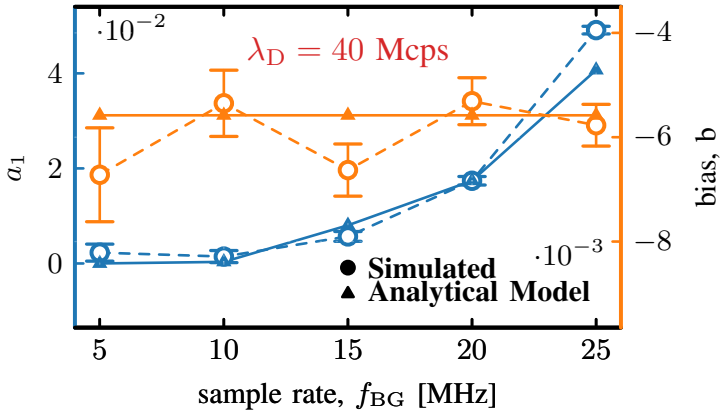
Clearly, for reduction of correlation, an appropriate detection rate/sampling rate ratio ($\lambda_{\mathrm{A}}/f_{\mathrm{BG}}$), must be chosen. Moreover, it can be seen that the one bit lag correlation coefficient, a_1, should be most prominent as higher order correlation coefficients decrease exponentially. An interesting consequence presented by the analysis so far is that there exists an inherent tradeoff between correlation, which requires increased count rate for faster generators and reducing bias, which scales with count rate. However, this is an ideal model, which does not take into account detector properties such as dead time and afterpulsing. Data from the bias simulations was used to calculate correlation coefficients and compared with the ideal analytical model. The results are shown in Figure 4.3. The exponential decay of correlation, with increased flux, is demonstrated by the simulation results shown in Figure 4.3a. Finally, Figure 4.3c validates that bias should not be affected by sampling rate, when the flux is held constant.



(a) Exponential decay of correlation coefficient at a constant sampling rate, with increased flux.



(b) Comparison between a_1 and higher order coefficients.



(c) Bias and correlation at constant flux with increased bit generation rate.

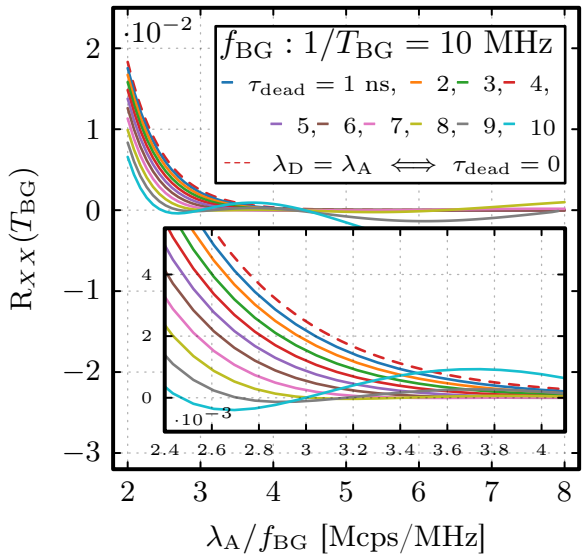
Figure 4.3: Comparison of ideal correlation model with results from Verilog-A circuit simulations. The exponential relationship of correlation with the count rate bit generation rate λ_A/f_{BG} ratio is clearly highlighted. Moreover, it can be see that even as correlation increases due to higher sampling rate at a constant flux, bias remains constant, as expected.

Correlation model with dead time

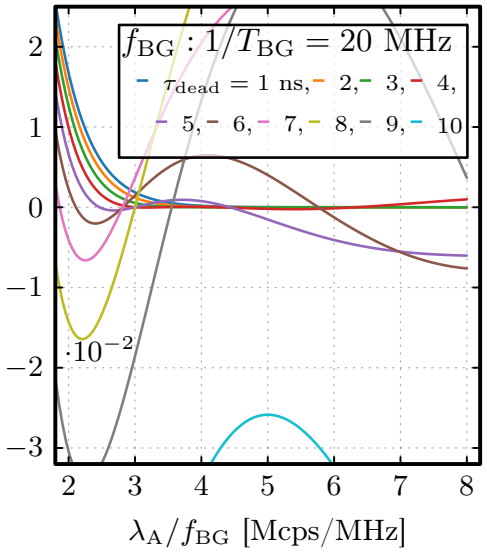
A fundamental goal of this thesis was to take the ideal RTS-like correlation modelling presented in the previous section, and augment it with realistic limitations of the SPAD detector, namely dead time and afterpulsing. Therefore, the statistics of the RTS are revisited. A more general approach is required for analyzing the behavior of the circuit employing the QRFF shown in Figure 4.1. Considering the RTS-like digital signal, $\mathrm{TFF_Q}$, as shown in Figure 4.1, it can be inferred that although the edge transitions can no longer be modeled as purely Poisson, they still occur at random times. As previously performed, correlation is considered within a generation period, $\tau : T_{\mathrm{BG}} = 1/f_{\mathrm{BG}}$. It can be seen that a correlated bit i.e. a bit that is the same binary value as the previous bit, occurs when $\mathrm{TFF_Q}$ is in the same logical state at the end of the period as it was at the start i.e. exclusively when an even number of photon detections happened during the period T_{BG}. $\mathrm{TFF_Q}$ is generalized as a stochastic process, $\{X_{K(t)}\}_{t\in T} : \mathbb{R} \to [0,1]$ with the random variable K, $k \in \mathbb{Z}_0^+$, which denotes the number of edge transitions (detections) that have occurred, in the interval T_{BG}. In this case, given that a generated bit was $x_i = 1$, then $x_{i+1} = 1$ happens only when the number of detections in the following sampling period is even i.e. if k, in the period T_{BG} is $k = \{0 \cup 2 \cup 4 \cup ... \cup k_{\max}\}$. As outlined previously, the maximum number of possible detections in the period is limited by the dead time, τ_{dead}, such that $k_{\max} = T_{\mathrm{BG}}/\tau_{\mathrm{dead}}$. Consequently, the probability of each of $k = \{0, 1, ..., k_{\max}\}$ detections can be calculated. The corresponding autocorrelation function, R_{XX} can then be evaluated as the probability of an even number of detections minus the probability of odd (which are summations), in the interval T_{BG}.

$$\mathrm{R}_{XX}(T_{\mathrm{BG}}) = \mathbb{P}(K_{\mathrm{even}}) - \mathbb{P}(K_{\mathrm{odd}}) \tag{4.10}$$

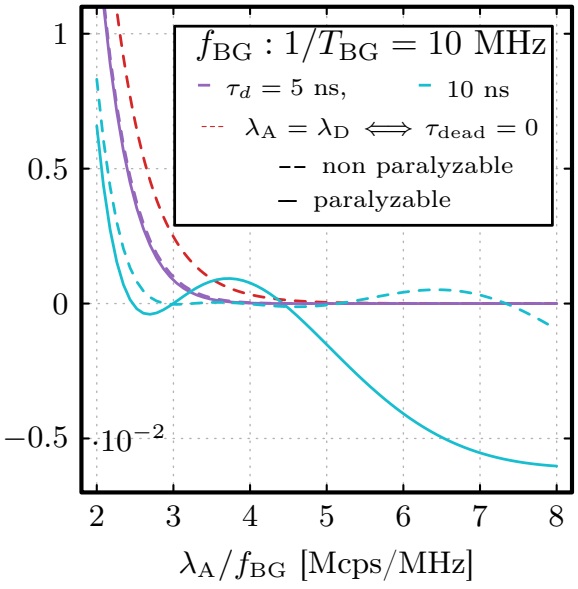
Therefore, the counting equations developed in [76], [170], mentioned earlier in this chapter, can be used to solve Equation (4.10). First, a paralysable detector is considered. Using swept values for τ_{dead}, the autocorrelation function is plotted against the $\lambda_{\mathrm{A}}/f_{\mathrm{BG}}$ ratio, with results shown in Figure 4.4. The first interesting observation, seen from Figure 4.4a is that the $\lambda_{\mathrm{A}}/f_{\mathrm{BG}}$ ratio required to achieve acceptable correlation is decreased as dead time is increased i.e. the serial correlation function decays quicker as dead time is increased. This provides a clear advantage by allowing for reduced detection rate requirements, which translates to lower power consumption, by increasing dead time. However, from Figure 4.4b, a clear limitation to this is shown because of pile-up effects for the paralysable case, manifesting as high fluctuations in the correlation function. Moreover, it is shown that this effect is exacerbated as sampling rates increases, where heavy oscillations are visible at low arrival rate/sampling rate ratios, with longer dead times. In this scenario, subsequent arrivals with shorter mean inter-arrival times will have a higher probability of detection during dead time. In the paralysable case, this can extend the dead time of the detector, further prolonging the time between toggles. Therefore, the sampling flip-flop over samples the 'stuck' toggle state, which results in higher correlation. Figure 4.4c shows a comparison between the paralysable and non-paralysable case at $\tau_{\mathrm{dead}} = 5$ and $\tau_{\mathrm{dead}} = 10$ ns. The correlation at lower dead times are similar, however the oscillation previously viewed, is improved for the non-paralysable



(a) $f_{\mathrm{BG}} = 10$ MHz paralysable dead times.



(b) $f_{\mathrm{BG}} = 20$ MHz paralysable dead times.



(c) $f_{\mathrm{BG}} = 10$ MHz comparison of dead time models.

Figure 4.4: Autocorrelation function of the QRFF circuit, at varying dead times, calculated using Equation (4.10). The counting models presented in [76], which model passive and active quenching scenarios, are used to calculate the probability of k detections.

case, as dead time extension from frequent arrivals is not a contributing factor to oversampling of a toggle state. Finally, a comparison is made using the counting equations presented in [170]. These comparisons are shown in Figures 4.5 for only the non-paralysable detector and Figure 4.6 for the non-paralysable detector. Clearly, implementation of a non-paralysable dead time is desirable for design of a random bit generator that uses this method, as it enables consistently low correlation at lower $\lambda_{\mathrm{A}}/f_{\mathrm{BG}}$, even when dead times are increased.

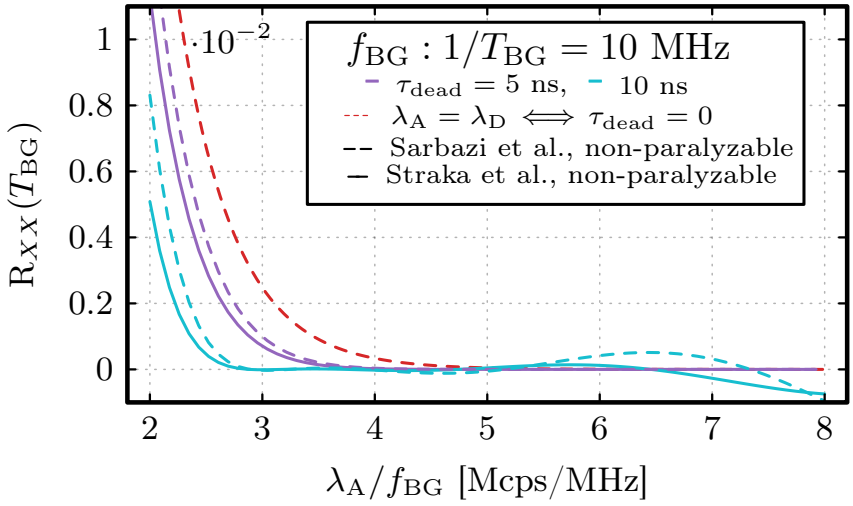


Figure 4.5: Comparison of autocorrelation function with non-paralysable detector using the counting models presented in [76] (dashed line) and [170] (solid line) for calculation of Equation (4.10).

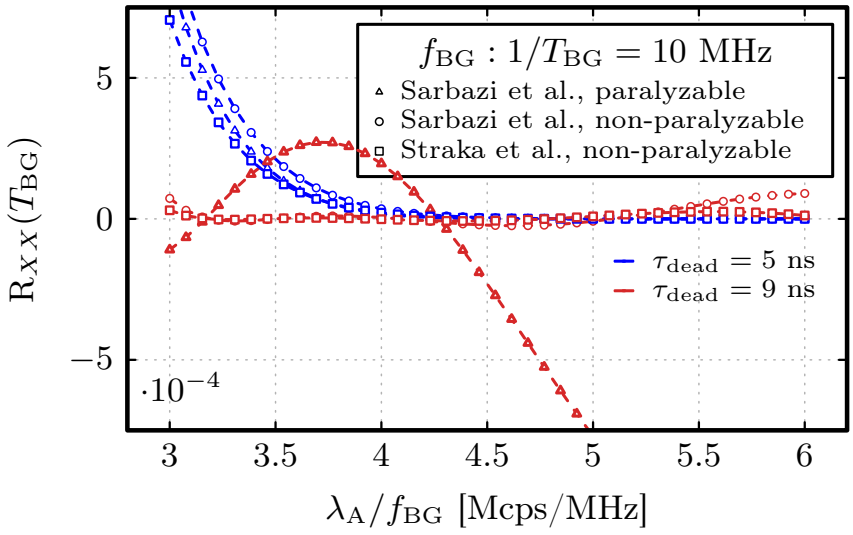


Figure 4.6: Comparison of autocorrelation function for a paralysable and non-paralysable detector using both counting models presented in [76] and [170] used to calculate Equation (4.10).

The analyses of correlation in this thesis, moving forward, use only counting equations derived in [170], given their mathematical rigor as well as the ability to augment afterpulsing. The simulation result of the non-paralysable case, using an updated Verilog-A model that includes detector dead time, is shown in Figure 4.7. Realistic dead time values of $\tau_{dead} = 5$ ns $\tau_{dead} = 10$ ns are chosen for this analysis. The key phenomenon highlighting faster correlation decay with larger dead time is confirmed in simulation.

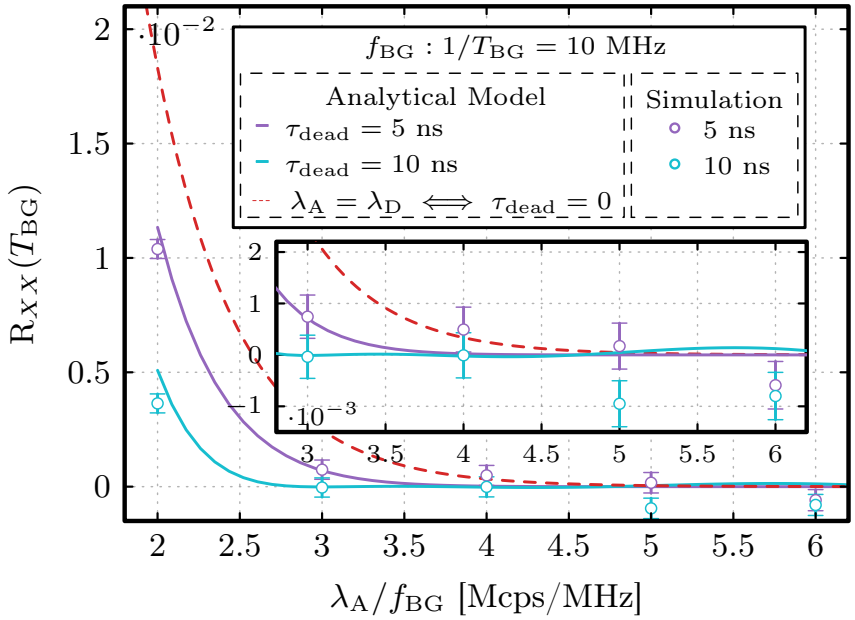


Figure 4.7: Verilog-A simulation results of autocorrelation values compared to analytical calculation ([170] counting). The analysis is performed using the non-paralyzable counting model at τ_{dead} values of 5, and 10 ns. The ideal autocorrelation function, described by Equation (4.9), based on a pure Poisson counting process, is shown by the dashed trace.

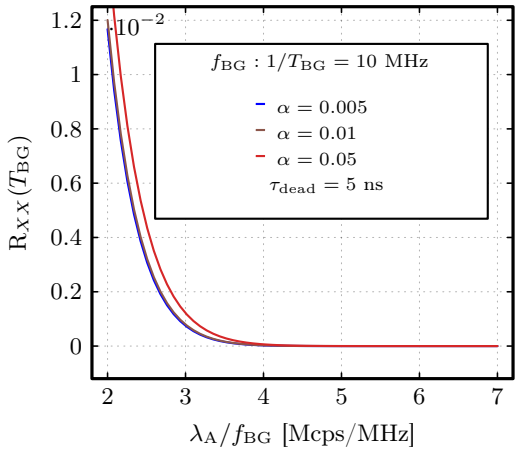
Correlation model with dead time and afterpulsing

As introduced in Chapter 2, afterpulsing describes the effect of avalanches that are caused by the release of trapped carriers. Due to the fact that afterpulsing is highly dependent on both the process and material properties, it is generally difficult to form a universal model. Furthermore, multiple methods to determine afterpulsing in experiment have been proposed [98], [183]. Typically, either the power law [184] or exponential models [185] are used to fit measured data in order to extract lifetime and prefactor values. Regardless of the methods employed, the release of a certain percentage of trapped carriers will initiate a subsequent avalanche. The ratio of the number of these afterpulses, N_A, to the total counts, N_T, is known as the afterpulsing probability:

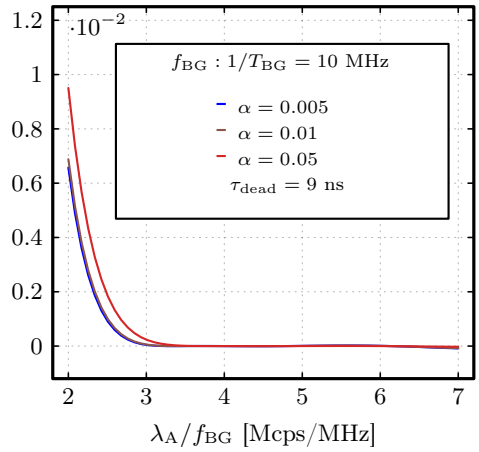
$$\alpha = \frac{N_A}{N_T} \tag{4.11}$$

In [170], to simplify the derivation of a counting equation, the probability density function (**PDF**) of afterpulsing was not considered, i.e. the analysis considered all afterpulses to arrive at time zero with probability α. Therefore, from the analytical perspective of autocorrelation, the trap lifetime is not considered. However, the impacts of this assumption is explored in the subsequent simulation section.

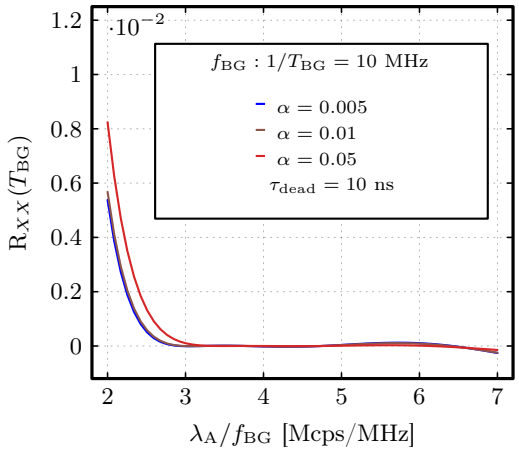
As afterpulsing probability increases, the probability of an even number of detections in a given T_{BG} also increases, due to the fact that an avalanche has a higher probability of triggering a secondary pulse. Based on this fundamental intuition, it would be expected that the



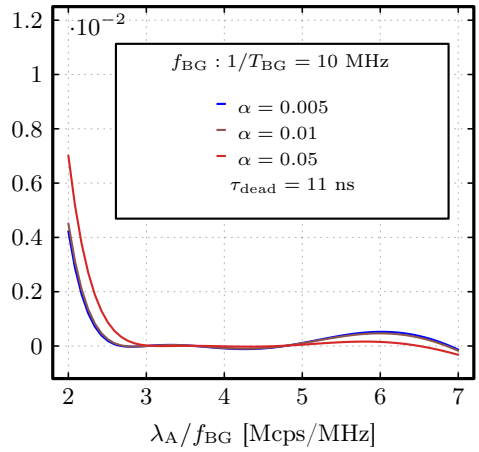
(a) $\tau_{dead} = 5$ ns.



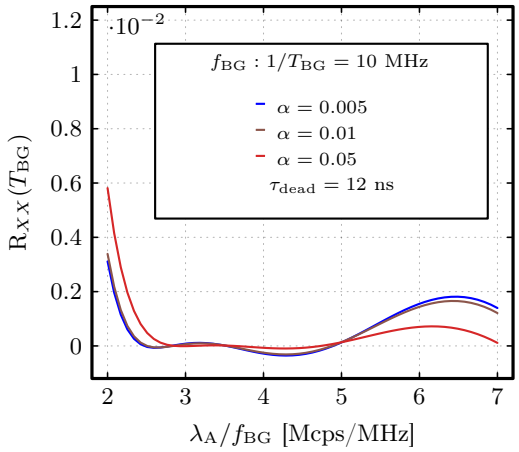
(b) $\tau_{dead} = 9$ ns.



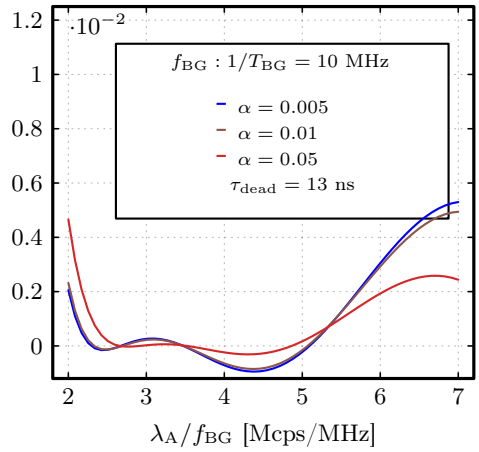
(c) $\tau_{dead} = 10$ ns.



(d) $\tau_{dead} = 11$ ns.



(e) $\tau_{dead} = 12$ ns.



(f) $\tau_{dead} = 13$ ns.

Figure 4.8: Autocorrelation combining dead time and afterpulsing probabilities (α) using $f_{BG} = 10$ MHz. Counting equations from [170] i.e. no consideration of afterpulsing PDF.

afterpulsing would increase autocorrelation. In Figure 4.8, the results from calculating (4.10) using the counting equations derived in [170] are displayed. Each individual plot illustrates α values of 0.5, 1, and 5 % at a given dead time.

A few observations are made from this analysis. First, it appears that afterpulsing is largely additive in terms of autocorrelation, although the effect is mostly pronounced at low flux values $\lambda_A/f_{BG} < 3$. In a scenario where afterpulsing probability is relatively high, while photon flux is relatively low, the probability of an even number of total detections in a period where an initial avalanche is present, is higher. However, at higher relative illumination, the count rate statistics are dominated by photon arrivals. Moreover, the analytical calculations show the same high sensitivity behavior at longer dead times, as was seen previously. Although increased afterpulsing seems to have a damping effect on this characteristic. It is possible that this behavior is nonphysical, as the equations assume arrival of afterpulses at time, $t = 0$ i.e. ignoring the PDF of afterpulses. While in practice, there will be some lifetime which greatly impacts the count statistics. In general this analysis suggests that, whether the afterpulsing probability is high or low, the result of correlations in a certain region where the arrival rate/sampling ratio is within $\lambda_A/f_{BG} = 3.5$ to 6, and the dead time bounded by 5 ns $\leq \tau_{dead} \leq$ 10 ns shows performance within the desired entropy benchmark.

Verilog-A simulation of afterpulsing

A few choice assumptions are made to model the physical behavior of afterpulses. For simplicity, only a single trap with a corresponding lifetime, that is always populated after an avalanche, is considered. An initial probability P_1 of igniting an avalanche is hard coded, which then decays exponentially with a given lifetime, τ_1, depending on the time after being trapped, t, it took to be released. In a fabricated SPAD, P_1 is determined by physical parameters such as defect concentration and current flow during an avalanche. It, along with the lifetime of the trap, τ_1, are typically extracted via experiment [94]. During runtime in simulation, the *$rdist_ exponential* function with a random seed is used to determine the trap release time, t [186]–[188]. Therefore, t is an occurrence of the exponential random variable $T \sim \text{Expo}(1/\tau_1)$ and the probability of afterpulsing becomes a function of a random variable, $Y_A(T) = P_1 \cdot e^{-T/\tau_1}$, with mean $\mathbf{E}[Y_A(T)] = P_1 \cdot \tau_1/(2 \cdot \tau_1) = P_1/2$. The probability measure of afterpulsing, A, in this model is defined by Equation (4.12). In the case where a trapped carrier is not released at the time of another photon arrival, the choice is made to release the carrier (restarts the decay process) without the ignition of another subsequent avalanche. Practically, during a realization of this scenario, the dynamic and physical behavior of the SPAD is complex. The band diagram shifts during the quench and recharge phases, and the trapped carriers may be released as they cross the quasi-fermi level. At this point, an avalanche can be ignited depending on a variety of variables such as the time of release and the excess bias value before the SPAD, which is low before completion of the recharge process. Therefore, to simplify these dynamics during modelling, a trap release can only cause an avalanche if it occurs after the dead time but before the subsequent photon arrival. The last assumption that is made is to ignore quench time, τ_q as it is relatively small ($\leq$ 1 ns) when using integrated circuitry. This signifies that the decay timing in simulation starts as soon as an avalanche is

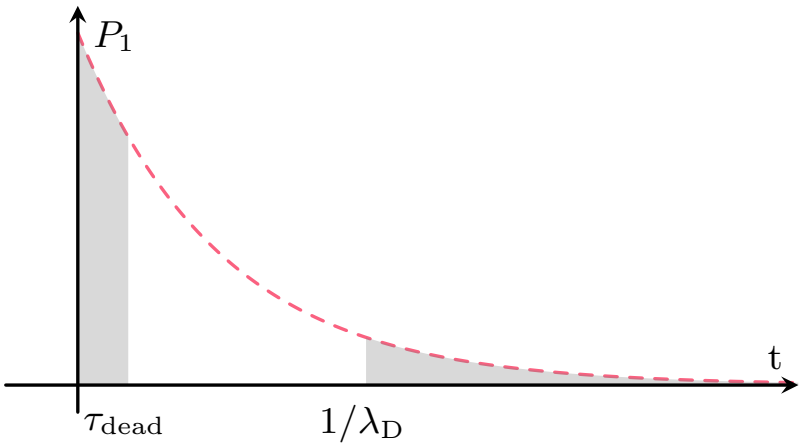


Figure 4.9: Diagram illustrating the probability measure of afterpulsing implemented in the model. A carrier which has a decaying probability P_1 and lifetime τ_1 can only ignite an avalanche if it is released at time, t_r, $t_r > \tau_q$ but before the next photon arrival. The non-shaded section corresponds to probability values that can ignite an avalanche.

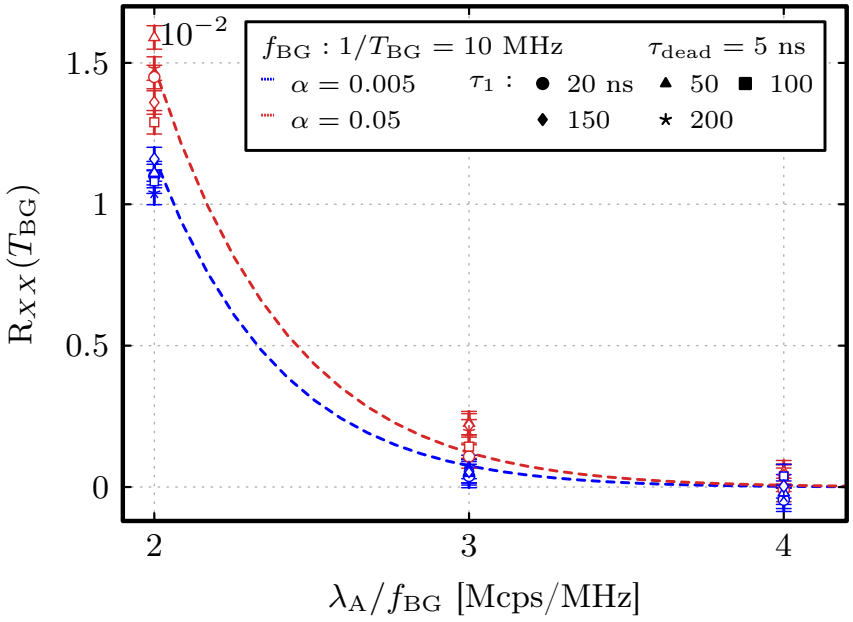
initiated. Therefore, the afterpulsing probability, α, in simulation, is determined by the range of probability values capable of igniting an avalanche, which is illustrated by Figure 4.9. This modified distribution, highlighted by the non-shaded region is denoted as $Y_A^{'}(T)$

$$\mathbb{P}(A)\{t\} := P_1 \cdot e^{(-t/\tau_1)} \quad \text{for } t \in \mathbb{R} \tag{4.12}$$

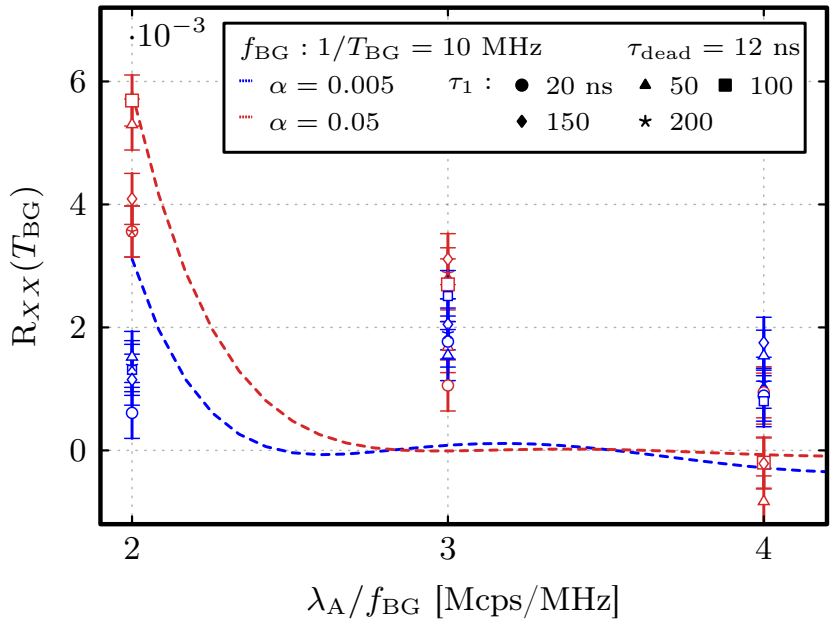
$$\begin{aligned} \alpha &\approx \mathbf{E}[Y_A^{'}(T)] \\ &= \int_{\tau_{\text{dead}}}^{1/\lambda_{\text{D}}} \frac{P_1}{2 \cdot \tau_1} \cdot te^{-t/\tau_1} \cdot dt \end{aligned} \tag{4.13}$$

In order to achieve a desired afterpulsing probability, for comparison to the analytical models in the previous section, P_1 must be chosen (hard coded) based on (4.13) so that α is representative of the analytical equations with the specified λ_{A}. During runtime, counter variables, that keep track of the total number of avalanches and successfully initiated afterpulses, are incremented within the code, in order to ensure that the value hard coded for P_1 corresponds to a simulated α value within $\pm$ 5% of the intended value.

Using lifetime values of $\tau_1 = 20, 50, 100, 150$ and 200 ns, the QRFF circuits are simulated with P_1 values in order to achieve afterpulsing probability of 5% and 0.5%. This was performed using a dead time of 5 and 12 ns, at $\lambda_{\text{A}} = 20, 30$ and 40 Mcps with a sampling (bit generation) rate of 10 MHz. Results are plotted in Figure 4.10 with comparison to the corresponding analytical calculations. From Figure 4.10a, it can be seen that the simulation and analytical calculations for both α values are very similar, with relatively no difference in results caused by afterpulsing lifetime. Moreover, the effect of increased afterpulsing on autocorrelation is largely additive, as predicted by the analysis. However, as dead time is increased, the ability of the analytical calculation to accurately predict autocorrelation decreases. This is highlighted



(a) $\tau_{\mathrm{dead}} = 5$ ns.



(b) $\tau_{\mathrm{dead}} = 12$ ns.

Figure 4.10: Comparison between analytical calculation of autocorrelation and simulation. Analysis is performed with $\tau_{\mathrm{dead}} = 5$ ns and 12 ns and with 0.5 % and 5 % afterpulsing probabilities. Lifetime values of τ_1 of 20, 50, 150, and 200 ns are used for each arrival rate. The dashed line indicates the analytical calculation while the markers are simulated values at λ_a values of 20, 30 and 40 Mcps.

by Figure 4.10b. For longer dead times, a higher count rate does not necessarily translate to reduced R_{XX}, for a given afterpulsing probability, and in general, the correlation becomes highly dependent on lifetime.

The critical conclusions from these analyses on bias and autocorrelation are summarized as follows:

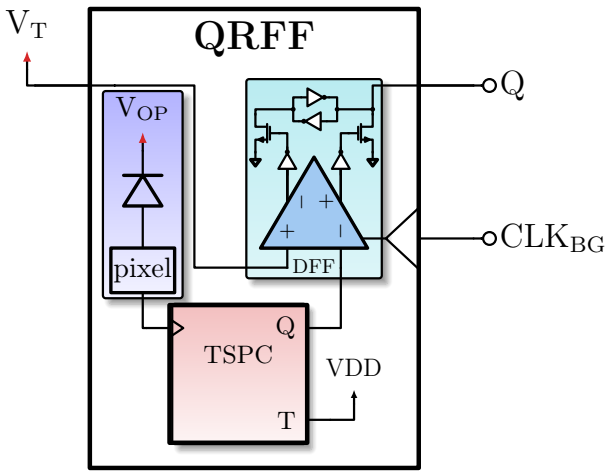
1. A fast TFF should be designed to reduce bias.
2. An adjustable threshold can compensate for asymmetric TFF rise/fall times.
3. A precise, non-paralysable dead time between $\tau_{\text{dead}} = 5 - 10$ ns is desirable, to enable a variety of $\lambda_{\text{A}}/f_{\text{BG}}$ ratios, decreasing both power consumption and performance sensitivity to flux variation.
4. Given the analysis on afterpulsing in this chapter, and the afterpulsing measurements of the deep junction SPADs in Chapter 2, afterpulsing induced correlation is not expected to be a concern in 55 nm designs.

4.4 Quantum random bit generators in 55 nm BCD

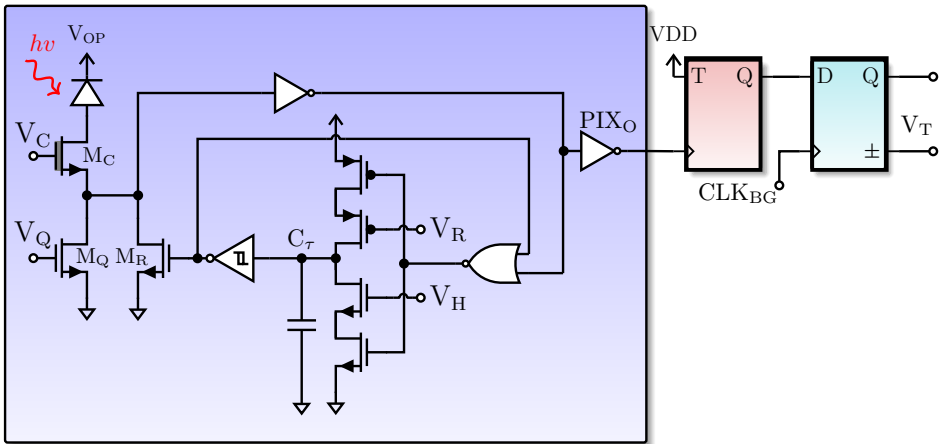
4.4.1 SC QRFF

Design and principle of operation

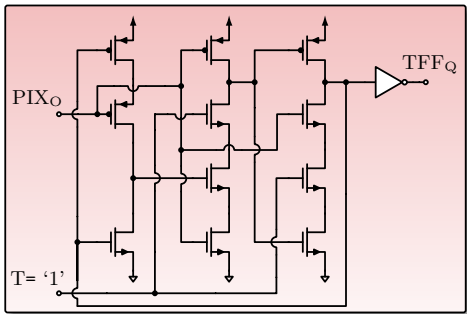
The architecture of the SC QRFF (method described in Chapter 1 consisting of 1 toggle flip-flop and 1 sampling flip-flop) design is shown in Figure 4.11d. The components within this random bit generator are highlighted by the block diagram in Figure 4.11a. It comprises a pixel, which has the same general structure as that presented in Chapter 3, along with a TFF, that realized the RTS-like digital signal and the sampling DFF. Based on the modelling performed, the continuous time characteristics of this digital circuit are intimately linked with the quality of generated bits. Therefore, a full-custom design approach is undertaken. The TFF is created using a compact 13 transistor true single-phase clock (TSPC) logic DFF with feedback from the $\overline{\text{Q}}$ output. This enables fast edge transitions and clock-to-q time. Moreover, a strongARM latch based clocked comparator is used for the DFF. The threshold pin, V_{T}, is implemented to experimentally validate the bias propositions presented earlier in this chapter. Although, other dynamic comparator architectures would also likely be suitable, a strongARM latch was chosen for it's fast latching time. The pixel design uses a passive-quench, active recharge structure. A fully non-paralysable detector would require active quenching to quickly starve the SPAD of charge and hold the anode at breakdown. However, an essentially non-paralysable regime can be implemented using a PQAR circuit. This is demonstrated by Figure 4.12, which shows a simulated pulse of the anode after an avalanche. Due to leakage, even when the quench resistor is off, there is some recharge that happens during the hold period. However, the excess bias across the SPAD is very low during this period, which makes the probability of an avalanche initiation very low ($<<$ 1%, when considering PDP). Moreover, the recharge time until the inverter threshold is crossed is in the range of 100ps. Therefore, it is assumed that modelling the detector as non-paralysable is valid. Finally, a timing diagram, which demonstrates the overall operation of the SC QRFF design, is shown in Figure 4.13.



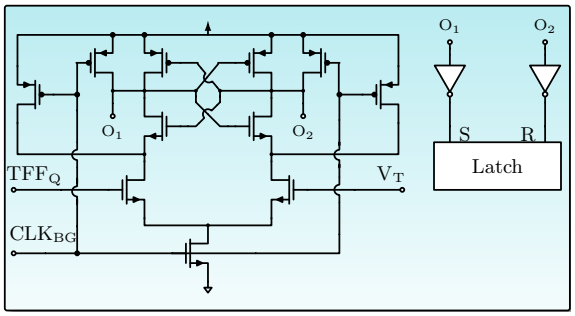
(a) Block diagram highlighting of SC QRFF design in 55 nm.



(b) Pixel architecture with 2 flip-flop configuration for big generation.



(c) TSPC TFF design using DFF with feedback division.



(d) StrongARM latched based clocked comparator DFF.

Figure 4.11: Slow clock QRFF design presented in this work to generate random bits. A block diagram highlights the major components: a SPAD+pixel, a TSPC TFF and a clocked comparator based DFF. $\mathrm{V_T}$ is the threshold control voltage ($\eta = \mathrm{V_T}/1.2$). A bit is generated at Q upon the arrival of an edge from the bit generation clock, $\mathrm{CLK_{BG}}$.

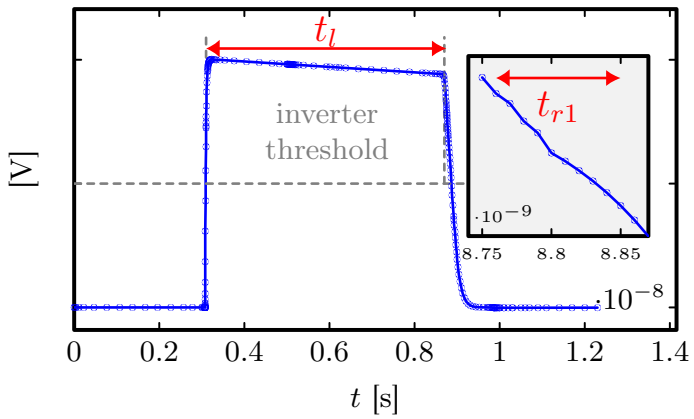


Figure 4.12: Plot showing simulated pulse width of the SPAD anode when coupled with the PQAR pixel. While technically paralysable in the time interval $t_l + t_{r1}$. However, in the hold-off interval, t_l, the excess bias across the SPAD is very low (< 100 mV). Therefore, absorbed photons have a low probability of initiating an avalanche. The recharge time is denoted by t_{r1}, until the inverter threshold is crossed (order of 100 ps). Therefore, the assumption is made that modelling this detector as non-paralysable is acceptable so long as flux is tuned to reduce pile-up effects.

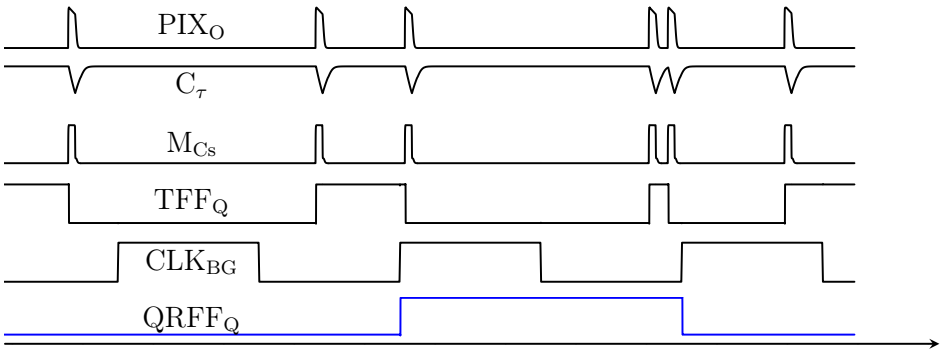


Figure 4.13: Timing diagram highlighting the general function of the SC QRFF design.

Measurements

Initial validation of bias and correlation, as a function of photon flux and normalized threshold (η), are preformed on a single pixel, before the QRFF is scaled to a large array (Chapter 5). The PDP of a single SPAD inside the QRFF design is shown in Figure 4.14. Clearly, there is a performance dip compared to the results shown in Chapter 3. This is caused by an increase in the optical stack thickness, owning to the use of additional metals. A LED (Cree C503B-BAN-CZ0A0452) is chosen for testing as the wavelength (470 nm) falls in a range where the PDP is high and spectrally separated from sensitivity troughs. The SPAD counting performance is characterized using an FPGA, with results shown in Figure 4.15. Two different pixel control settings ($\mathrm{V_H} = 0.65$ V and $\mathrm{V_H} = 0.7$) V values are chosen for characterization, corresponding to dead times of $\tau_{\mathrm{dead}} = 10$ and $\tau_{\mathrm{dead}} = 8$, respectively. When $\mathrm{V_H}$ is increased past 0.7 V, the pulse-width becomes too narrow to accurately determine dead time. Based on this count rate and dead time values, the generation rates of $f_{\mathrm{BG}} = 5 - 10$ MHz should be achievable, with 2 to 3 mA setting of the illumination current ($\mathrm{I_{LED}}$), while maintaining the entropy benchmarks.

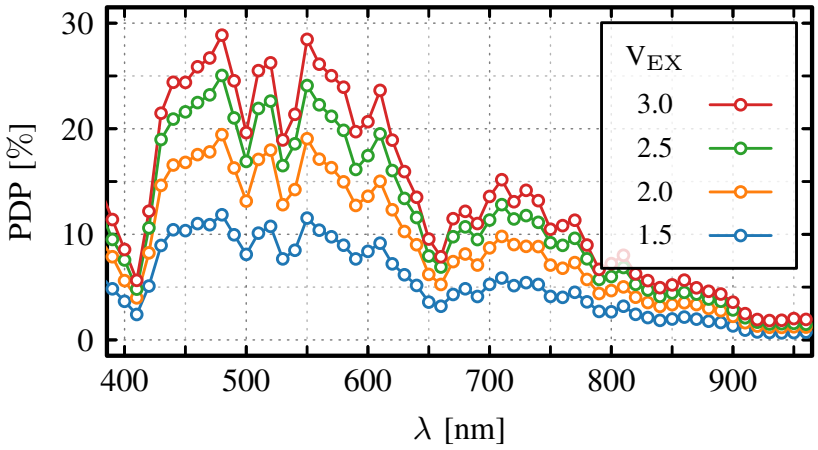


Figure 4.14: Measured PDP of a single QRFF with excess bias in the range $V_{EX} = 1 - 3$ V.

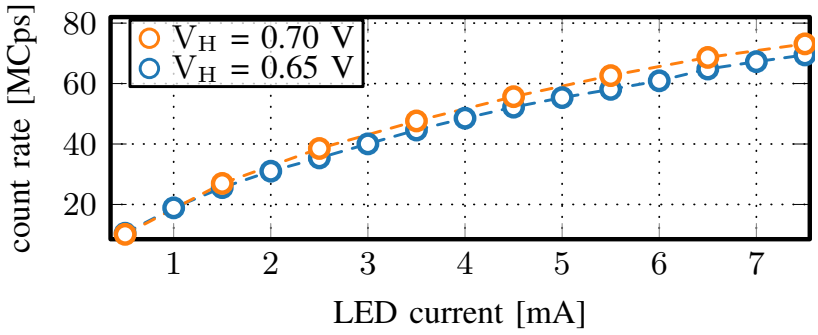


Figure 4.15: Counting performance of a single QRFF with hold voltage settings ($V_H = 0.65$ V and $V_H = 0.7$ V) vs LED current (I_{LED}).

Bias is calculated using data generated at a sampling rate of $f_{BG} = 5$ MHz. The results are shown in Figure 4.16. First, it is observed, as predicted, that by adjusting the sampling threshold, bias can be reduced. At a normalized threshold of $\eta = 0.75$, the measured bias is on the order of 10^{-5}. However, the normalized threshold required is considerably higher than expected. This would suggest that phenomena that has not been included in the model, such as metastability, comparator offset, or uneven high-to-low/low-to-high clock-to-q, could be contributing to bias. Nevertheless, it is shown that the magnitude of bit bias also scales (increases) with increased count rates. The adjustable threshold, while a novel and effective solution for reducing bias, **without the need to combine multiple entropy sources**, has some limitations. For example, it provides one more method, for an adversary, to try and attack/gain control of the device. However, this could be mitigated by generating and controlling the threshold internally, using, for example, a bandgap reference.

Validation of the proposed analytical model of autocorrelation, which includes dead time, was performed at ($V_H = 0.65$ V and $V_H = 0.7$ V). The arrival rate λ_A, was calculated from the detection (count) rate measured in Figure 4.15 and Equation (4.5), so that correlation could be characterized at various λ_A/f_{BG} ratios. Measurements are compared to analytical calculations and plotted in Figure 4.17. The results match well with the expectation, thereby

semi validating the proposed model. Increasing the dead time to 10 ns allows for diminished serial autocorrelation of bits, with a λ_A/f_{BG} ratio of 2.5.

Entropy estimation is performed on the data generated by the single QRFF for various two separate η and ILED values with a generation rate of $f_{BG} = 10$ MHz. Note, a more thorough description of randomness testing and suites is provided in the following chapter, where complete evaluation of a full QRNG chip is provided. Results are summarized in Table 4.2. Min entropy is estimated using the NIST SP 800-90B test suite [56]. In each of the test cases shown, the results pass i.i.d., chi squared and longest run tests. At $\eta = 0.5$, $I_{LED} = 3$ mA, the achieved min entropy is $H_\infty = 0.978$, which is slightly below the requirement (0.98) for min entropy of the upcoming AIS standard [63]. Therefore, it can be deduced that a range between $\eta = 0.5 - 0.75$ can produce acceptable results. The scalability of this solution, along with a wider illumination intensity study, is conducted in Chapter 5.

Table 4.2: Single-QRFF Entropy Characterization at $f_{BG} = 10$ MHz and $V_{OP} = 33.3$ V

η	I_{LED} [mA]	H_1	H_∞
0.75	2	0.999999997	0.995
0.50	2	0.999988458	0.985
0.75	3	0.999997114	0.994
0.50	3	0.999953833	0.978

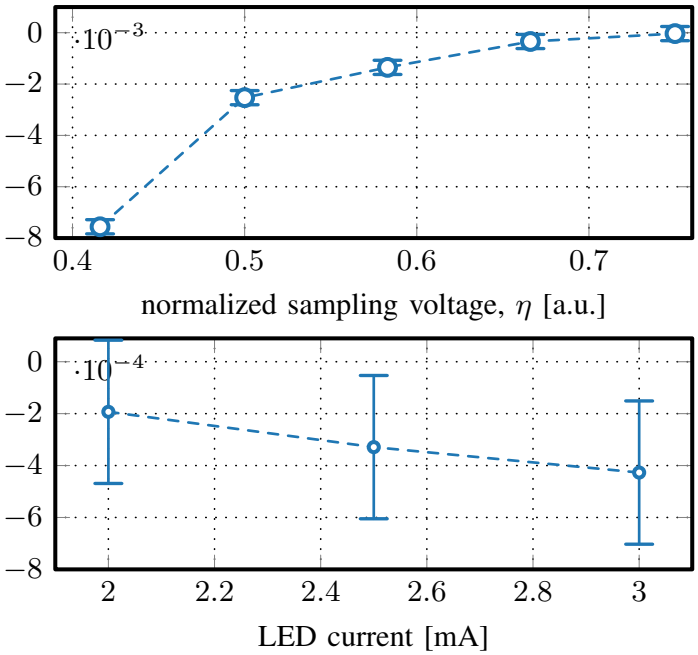


Figure 4.16: Bias measurements from $P(X = 1) = 0.5$ at $f_{BG} = 5$ MHz vs normalized sampling threshold and LED current. Upper plot performed with $I_{LED} = 2$ mA and lower plot with $V_T = 0.9$ V ($\eta = 0.75$).

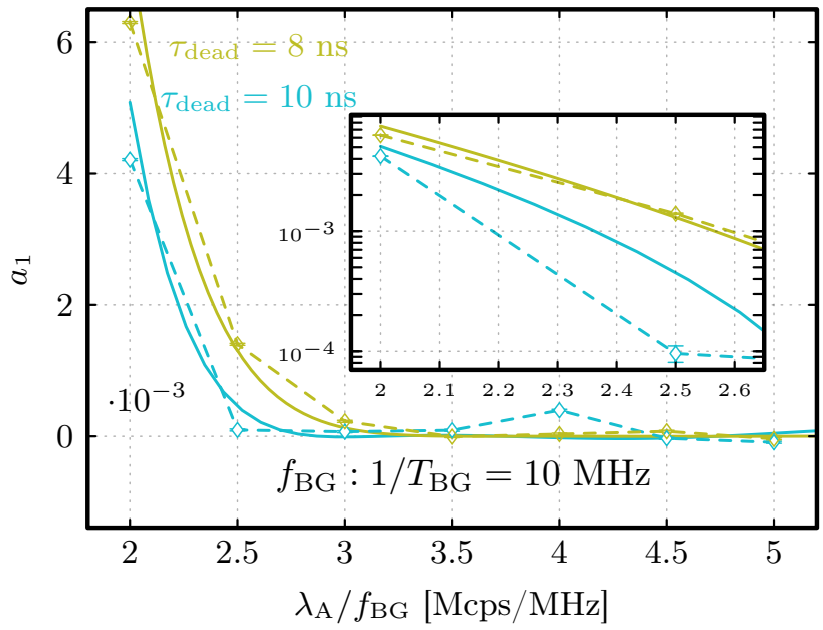


Figure 4.17: Serial correlation measurement results for a single QRFF compared to expected results based on proposed correlation model with dead time.

4.4.2 FA QRFF

Design and principle of operation

First-arrival (FA) QRFF circuit, designed again in 55 nm BCD, is proposed and analyzed. The general block diagram of the QRFF is shown in 4.18. The random bit generation operation is performed by the decision cell, which determines which one of two identical pixels was the first to detect a photon, within an evaluation window, $\overline{\text{RST}}$. Dividers, implemented as simple clock dividers, are used for included so that counting can be performed by an external FPGA, even when the SPAD dead time is only a few nanoseconds. A simplified timing diagram that demonstrates the random bit operation of the QRFF is provided in Figure 4.19. The circuit schematics for each of the blocks are shown in Figure 4.20. The pixel includes a recharge loop, similar in design to others in this thesis, although without the hold transistor M_H, as the functionality is not necessary. When, $\overline{\text{RST}}$ goes high, the SPAD becomes active as M_{SO} is turned off and the recharge loop very quickly recharges the SPAD. A short pulse is sent to the pixel output (Q_1/Q_2) i.e. to the decision cell. Therefore, the $\overline{\text{RST}}$ signal functions as a gate. Careful thought has gone into the design of this cell. This method for generating random bits has been demonstrated in literature [180], [181]. However, to the best of the author's knowledge, this fast decision cell architecture has not been implemented in a random bit generator design. Blue transistors denote low threshold-voltage devices, while red are high-threshold voltage. The cell can be divided into three general sections. First, the input stage consists of pre-charged logic transistors M_1, M_2,M_3, and M_6. The remainder of the low threshold-voltage transistors create two regenerative negative impedance amplifications steps, where the difference is then latched by the final stage. When two inputs rise close to each

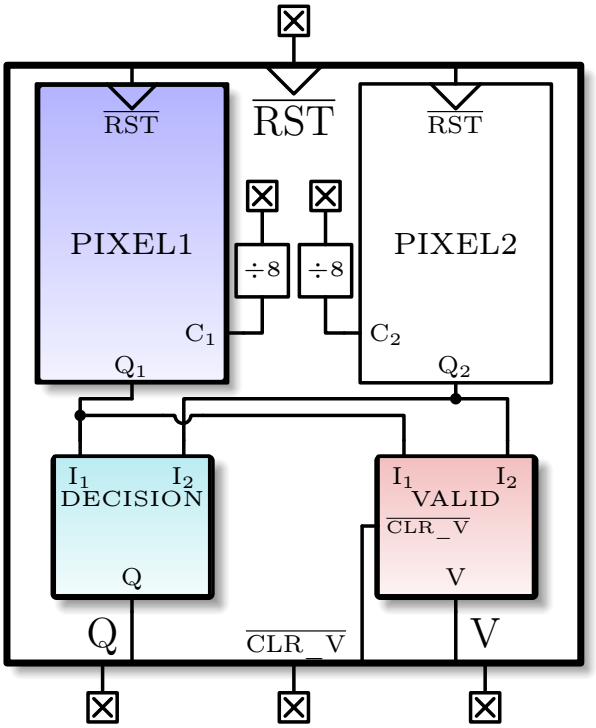


Figure 4.18: First arrival (FA) based QRFF block diagram. It consists of two identical SPAD pixels, a decision cell to decide which was the first to fire, along with a valid cell that checks that a photon arrival occurred. Count dividers are added to the pixel outputs so that small pulses τ_{dead} can be consistently counted by an FPGA. Circuit schematics for each block are illustrated by Figure 4.20.

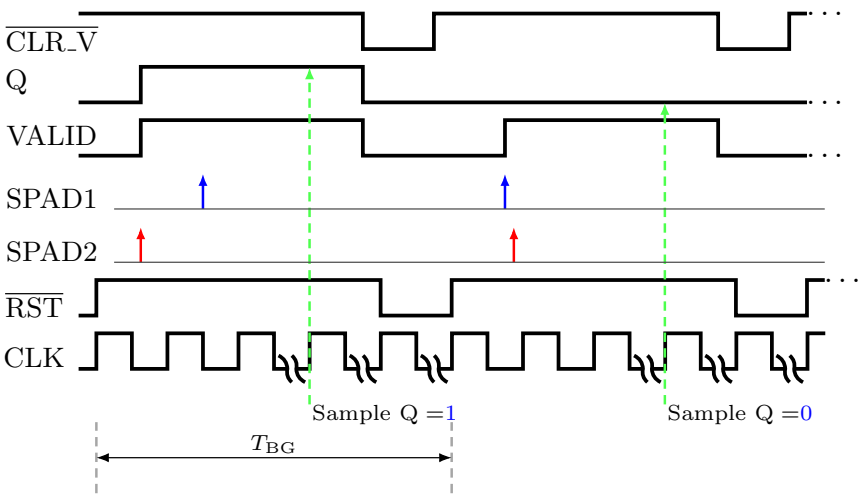
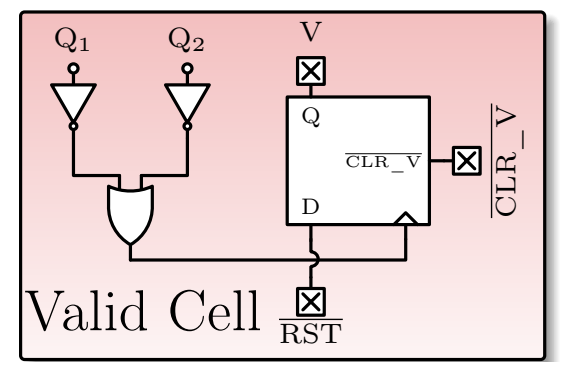
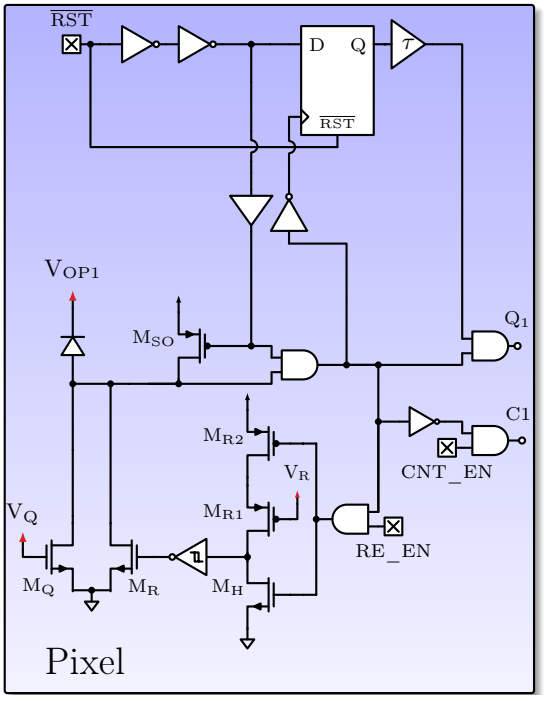


Figure 4.19: Timing diagram of the FA QRFF design. The comparison is made within the evaluation window set by $\overline{\text{RST}}$. A valid signal is generated when a photon is detected.

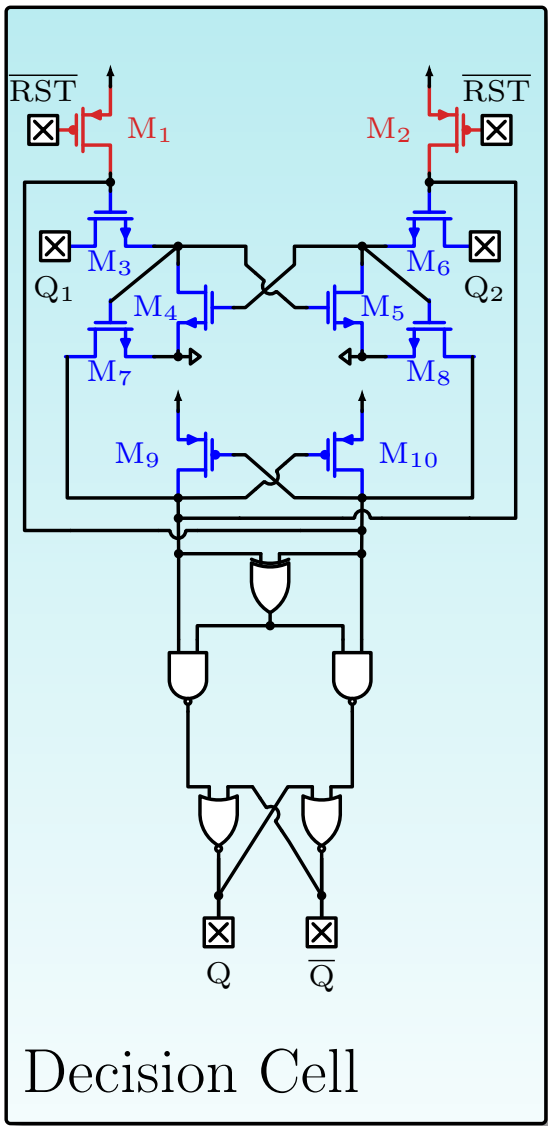
other, the drain voltages of M_4 and M_5 start to rise. Then, the drains of cross-coupled pmos cells M_9 and M_{10}, which both originally start at logic high, perform a decision based on the



(a) Valid cell.



(b) Pixel circuit.



(c) Decision cell.

Figure 4.20: The FA QRFF circuit schematic. The evaluation period is set by $\overline{\text{RST}}$. An automatic recharge loop is used to quickly activate the SPADs after M_{SO} is turned off. The valid cell check for a photon detection from either of the SPADs. A two stage regenerative amplification process in the decision cell is used to discriminate between closely spaced events.

first arrival. The result is a cell which can discriminate between events which are < 5 ps separation. To reduce effects of metastability, buffering on the $\overline{\text{RST}}$ clock is included within the cell. The layout of the FA QRFF is displayed by Figure 4.21. Compactness is compromised so that a symmetric layout could be implemented. This helps prevent bias caused by path length mismatch or process variation. The entire cell (excluding SPADs) covers an area of $\simeq 132.6\ \mu\text{m}^2$. Moreover, for testing purposes, the SPAD bias V_{OP} for each pixel is connected separately as a method for equalizing count rates. This provides a qualitative method for understanding entropy degradation caused by the electronics or other phenomena that has not been modelled.

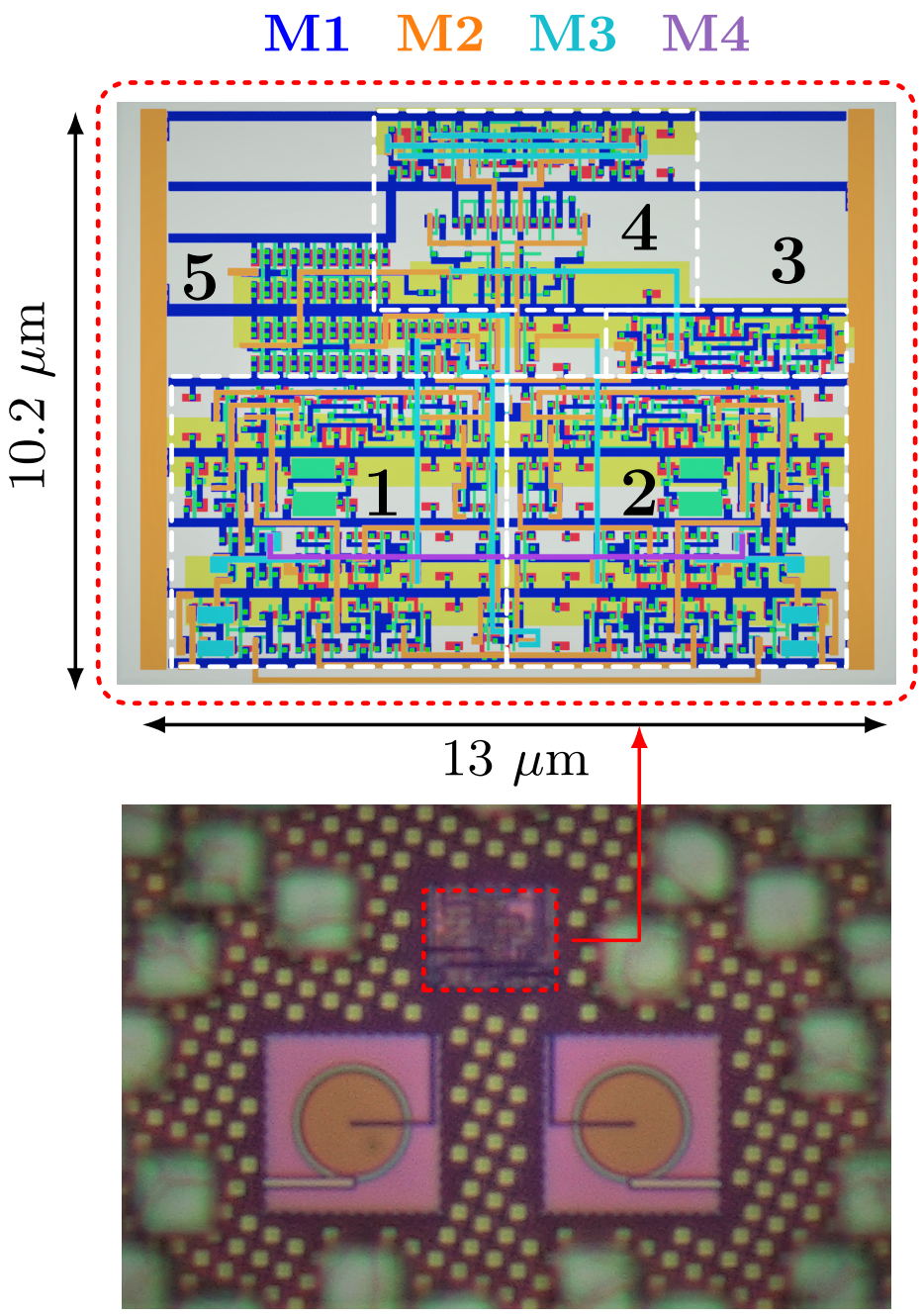


Figure 4.21: Layout of FA QRFF cell. 1: Pixel 1, 2: Pixel 2, 3: Valid cell, 4: Decision cell, 5: $\overline{\text{RST}}$ clock buffer.

Measurements

To produce randomness statistics that arise primarily from a photon arrival, the FA QRFF should be operated in a regime where there is a high probability that a photon detection has occurred in a sampling period. For any given count rate, this sets a limitation on the operable bit generation rate. This can be estimated with the cumulative distribution function (CDF) of

the exponential distribution, as shown in Figure 4.22. As an example, at $\lambda_D = 10$ Mcps, the wait time required to ensure a photon arrival is $\simeq 400$ ns. This translates to a maximum bit generation rate of 2.5 MHz. Conversely, at $\lambda_D = 80$ MHz, bit generation can be performed up to $\simeq f_{BG} = 20$ MHz. The counting performance, using the same blue LED that was used for

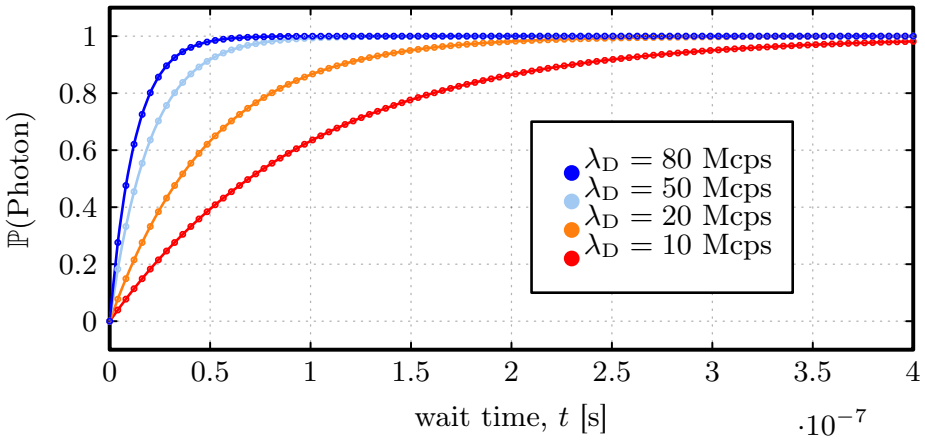


Figure 4.22: CDF of the exponential distribution with detection rates $\lambda_D = 10$, $\lambda_D = 20$, $\lambda_D = 50$, and $\lambda_D = 80$ Mcps.

the SC QRFF, is used in measurement of the FA QRFF. Count rate difference between the two SPADs hovers around 1%, suggesting a theoretical bias in the range of $\simeq 3 \cdot 10^{-3}$. This is considerably higher than the SC QRFF (orders of magnitude higher when threshold correction is used). However, in principle, the design should still be capable of passing NIST SP 800-90B. Moreover, the theoretical absence of correlations is advantageous. With the ability to count at rates above 100 Mcps, the QRFF should be capable of generating high-entropy bits at a rate above $\simeq f_{BG} = 20$ MHz. Random data is generated at a constant bit generation rate $f_{BG} = 15$ MHz as a function of LED current. Before each test, the count rate is measured so that the resulting bit bias can be compared to the theoretical value. Results are shown in Figure 4.24. As expected, bias remains high, until the requisite flux for a high probability of detection, is reached. Between 4-9 mA of LED current, it can be seen that the measured bias remains stable and close to the theoretical value (described by Equation (4.8)) based on the measured count rates. Serial correlation and bias are measured for data generated in a range of $f_{BG} = 5 - 22.5$ MHz at a constant $I_{LED} = 10$ mA, shown in 4.25. As expected from theory, correlation remains low regardless of sample rate. This is true for bias as well, until a point at which the illumination intensity is no longer adequate when compared to the sampling rate (22.5 MHz in Figure 4.25). In these measurements, all bits are kept regardless of the presence of photons, in order to maintain a constant bit rate. These results show reasonable bit generation statistics and speed. Clearly, there is a degradation in terms of bias, when compared to the SC QRFF circuit demonstrated earlier. However, an advantage is that only one photon is required, compared to $\simeq 2.5$ shown previously for the SC method. Finally, to better understand the limits of the FA circuit, H_∞ is estimated across a variety of operating parameters. The results are detailed in Table 4.3. At a constant sampling rate of $f_{BG} = 15$ MHz, the SPAD bias of the first pixel is swept while V_{OP2} is constant at 33.3

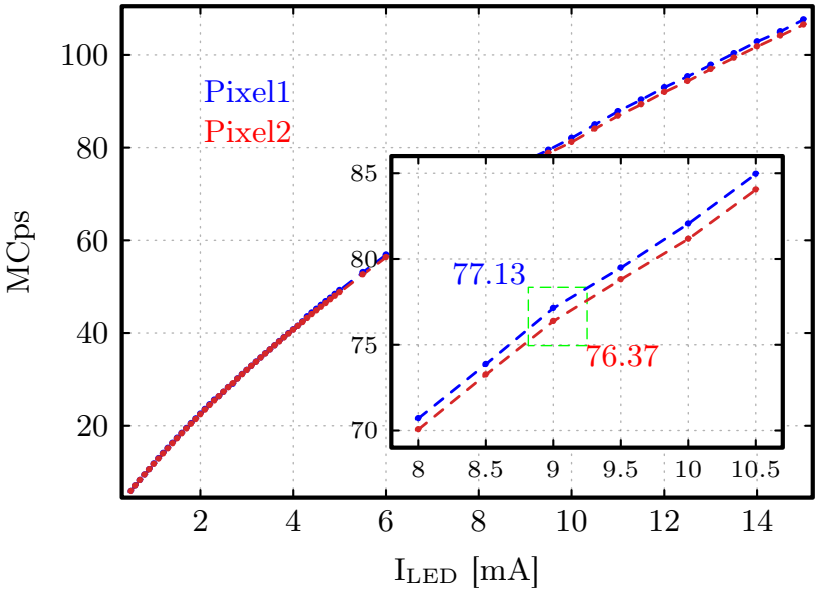


Figure 4.23: Counting performance of both detectors in the FA QRFF design, as a function of LED current. Both SPADs are connected to $V_{OP} = 33.3$ V.

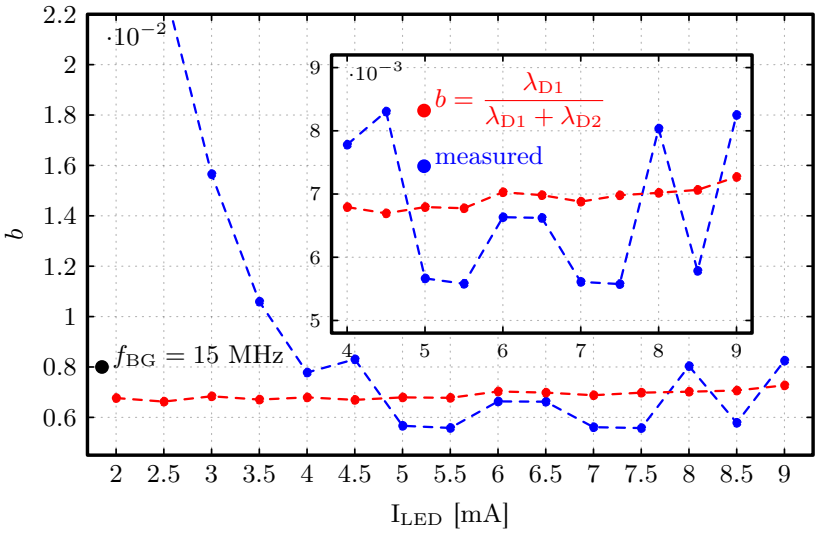


Figure 4.24: FA QRFF bias measurement at a constant bit generation rate of $f_{BG} = 15$ MHz, compared to the theoretical value based on count rates.

V. As intuited, the bias can be reduced significantly when the SPAD voltage bias are kept separate (corresponding to $V_{OP1} = 32.28$ V). Interestingly, this is not the point at which the count rates of the SPADs are equalized, suggesting that comparator offset or metastability are contributing to bias. Furthermore, when bias is reduced, the min-entropy of generated bits is also maximized, $H_\infty \simeq 0.995$.. Finally, even when entropy is degraded, due to bias, by increasing f_{BG} to 22.5 MHz, the SP 800-90B tests still are passed. This is likely due to the low correlation between bits.

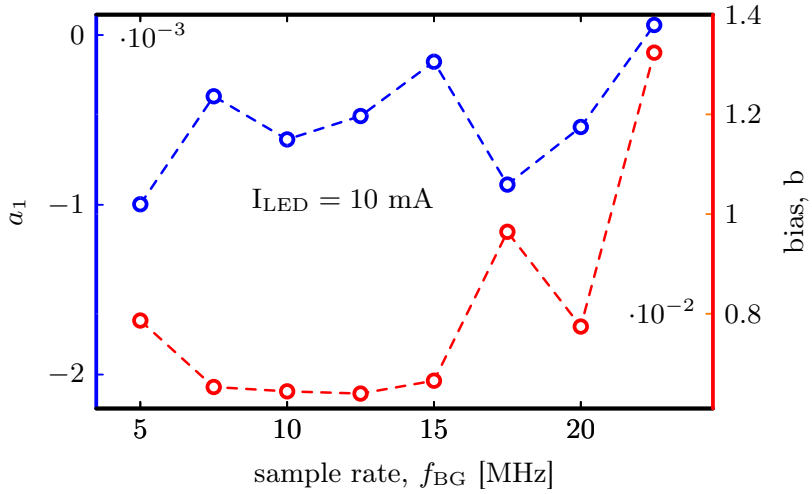


Figure 4.25: FA QRFF bit bias and serial correlation with swept frequency $f_{\mathrm{BG}} = 5 - 22.5$ MHz at a constant illumination of $\mathrm{I_{LED}} = 10$ mA.

$\mathrm{V_{OP1}}$	C1-C2 [MCps]	b (theoretical)	b (measured)	a_1 (measured)	$\mathrm{H_1}$	$\mathrm{H_\infty}$
$f_{\mathrm{BG}} = 15$ MHz						
33.270	-4.3419	-0.01576	-0.00439	-0.00035	0.99994	0.985†
33.275	-3.554	-0.01319	-0.00231	-0.00025	0.999984	0.988†
33.280	-2.406	-0.0089	-0.00007	-0.00025	0.999999985	0.995†
33.285	-1.597	-0.00597	0.00092	-0.0003	0.9999975	0.994†
33.290	- 0.515	-0.0019	0.00296	-0.0002	0.9999747	0.989†
33.295	0.1019	0.00037	0.00358	0.00055	0.999963	0.987†
33.300	1.106	0.00394	0.00573	0.00063	0.999905	0.981†
$f_{\mathrm{BG}} = 17.5$ MHz						
33.300‡	1.272	0.00449	0.00573	0.00009	0.999905	0.979†
$f_{\mathrm{BG}} = 20$ MHz						
33.300	1.4170	0.00508	0.00736	0.00025	0.99984	0.976†
$f_{\mathrm{BG}} = 22.5$ MHz						
33.300	1.236	0.00453	0.01576	0.00014	0.99928	0.953†

† Passed all SP 800 90B tests.

Table 4.3: Detailed results of FA QRFF circuit in terms of bias, correlation and entropy. The SPAD bias ($\mathrm{V_{OP}}$) of the first pixel is adjusted to compensate for count rate difference.

5 FortunaSPAD: A dual-interface QRNG with 3.3 Gbps output rate

The FortunaSPAD QRNG, which is a single-die sensor containing two independent arrays is presented in this Chapter. The design and full characterization has been accepted in IEEE JSSC (DOI: 10.1109/JSSC.2023.3274692). The arxiv preprint arXiv:2209.04868 is also available [166].

5.1 System architecture

The SC QRFF presented in Chapter 4 was scaled to a QRNG sensor containing 2800 individual QRFF pixels and fabricated in the GF 55 nm process. Figure 5.1 is a block diagram describing the architecture of the chip. Two independent arrays, containing separate bit generation clocks (CLK_{BG}) and readout methods, were implemented. They are denoted as A1 and A2. Array A1 contains 32×70 QRFF pixels. In this sub-array, a multiplexer, controlled by the COUNT signal, which bypasses the TFF/DFF bit generation circuit in each pixel is included, so that counting can be monitored. Moreover, this enables a more quantitative way for comparison of measured results vs expectation, based on the models presented earlier. A QRFF serialization readout scheme is implemented for A2. This array contains 8×70 QRFF pixels. A low speed serialization clock is multiplied by an on-chip all-digital **PLL** (**ADPLL**, an IP block provided by GlobalFoundries) that is used to then serialize the bits generated by A2 onto 8 low-voltage differential signaling (**LVDS**) channels i.e. 70 individual QRFF pixels become serialized onto a single channel. Therefore, the frequency of bit generation f_{BG} is performed at a 70× slower rate compared to the output data rate of each channel. Moreover, a KNOWN_PATTERN control flag is included in the serialization circuit. When enabled, a pre-determined 8-bit pattern is continuously output by every channel of the transmitter. This allows the FPGA to first tune the input/output (IO) delays of each individual receiver channel until the known pattern is detected, at which point the flag can be disabled and the bit generation can commence. Another channel, that contains as strobe pulse every time the serializer returns to the first QRFF (bit 0), is also included. This is useful for being able to analyze bits generated by individual QRFFs. All pixels controls (V_C, V_Q, V_R, V_H, V_T) are controlled globally. Ideally, the threshold control for each QRFF could be set individually with

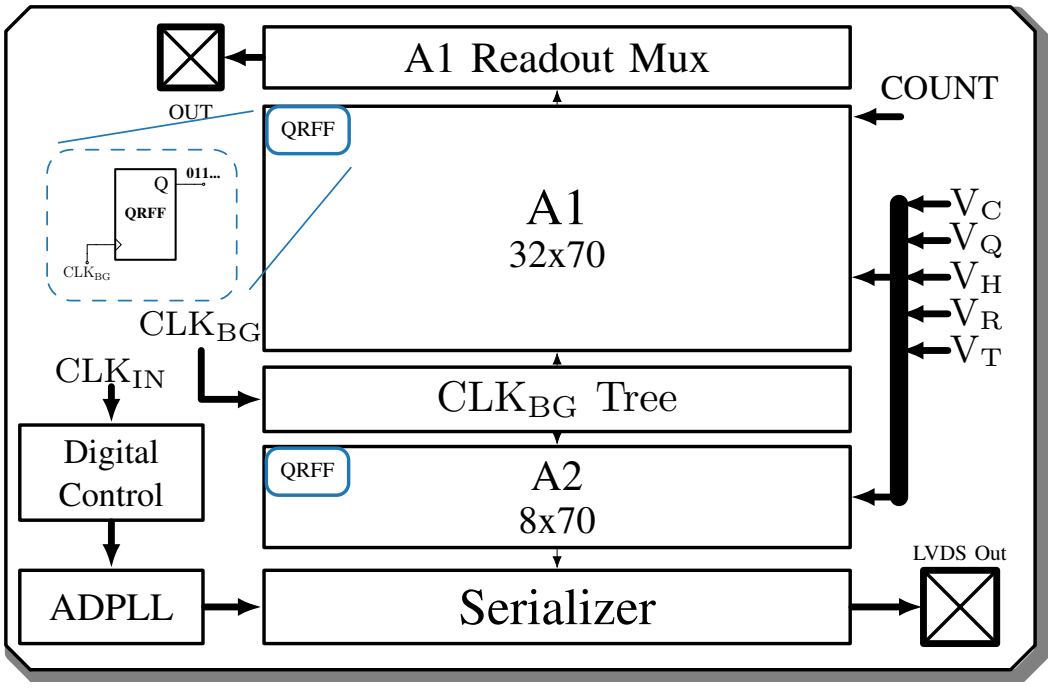
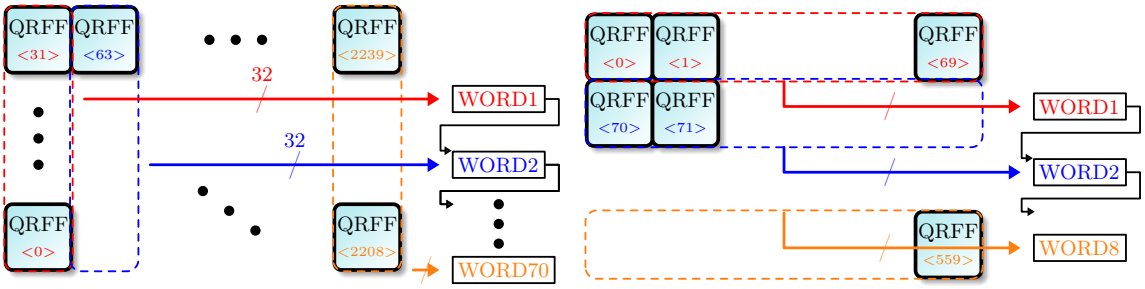


Figure 5.1: FortunaSPAD chip architecture. Two independent arrays (A1) and (A2) capable of generating bits concurrently are included. A1 implements a simple readout structure where all pixels are connected to an output readout multiplexer. The readout scheme for A2 uses a high-speed serialization clock generated by the ADPLL to serialize 70 QRFFs onto a single channel.



(a) A1 readout scheme. Columns QRFFs generate a word with each subsequent column concatenated.

(b) A2 readout scheme. Row QRFFs are serialized with each subsequent row concatenated.

Figure 5.2: A diagram that illustrates how random bits generated from individual QRFF pixels are read-out i.e. how serial data is turned into spatial data.

a digital-to-analog converter (**DAC**), although the complexity of this would be prohibitive. It will be shown that a global threshold setting of the FortunaSPAD QRFF threshold is sufficient for achieving the entropy benchmarks.

Extensive statistical analysis of bias and correlation for serial data generated by each individual QRFF is performed. However, to ensure there are no correlations between neighboring pixels caused by crosstalk or other phenomena, spatial analysis must also be performed. Figure 5.2 illustrates how serial data from each pixel is spatially combined by the readout scheme to

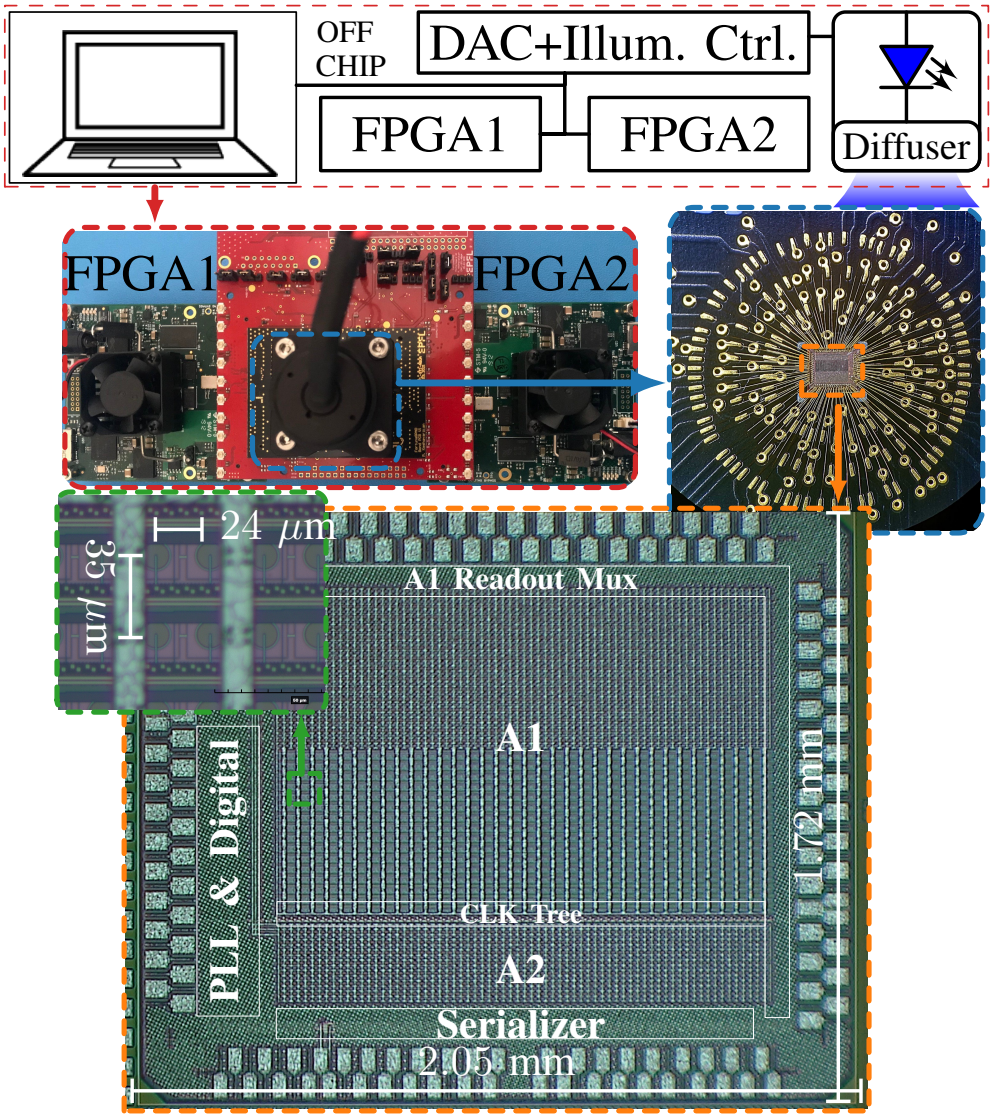


Figure 5.3: Test characterization setup and micrograph of the FortunaSPAD QRNG. A FPGA is used for read-out of each array. The LED is housed inside an optical tube. A diffuser is used for more uniform illumination. Red motherboard contains all voltage generation and illumination control required for testing. The total die area is 2.05 mm $\times$ 1.72 mm.

create bit strings (words). For A1, each column is read-out to form a 32-bit word. When the readout multiplexer setting is changed, the next column is selected. A2 uses the reverse spatial mode, where every QRFF in a single row is serialized to create a 70-bit word.

The entire system, including the optics and electronics used for measurement of the sensor, is displayed in Figure 5.3. The motherboard generates all the required voltages and the **LED** illumination control needed for measurement of the chip. A micrograph outlines the total die area (2.05 mm$\times$1.72 mm) along with the individual blocks within the design. Two (Opal Kelly XEM7350) field-programmable gate arrays (**FPGA**s) are used to acquire data from the individual arrays, although there is nothing in the design that precludes use of a single FPGA.

The pixel pitch is 24 μm horizontal and 35 μm vertical. These pitches are artificially extended within the design for additional isolation, resulting in a relatively low fill-factor of $\simeq 7$ %. All generated data is saved to a PC over a USB interface from the FPGA. The speed analysis in this chapter is defined with reference to the data rate transfer from the chip to the FPGA rather than the FPGA to PC.

5.2 Measurements and characterization

Noise analysis

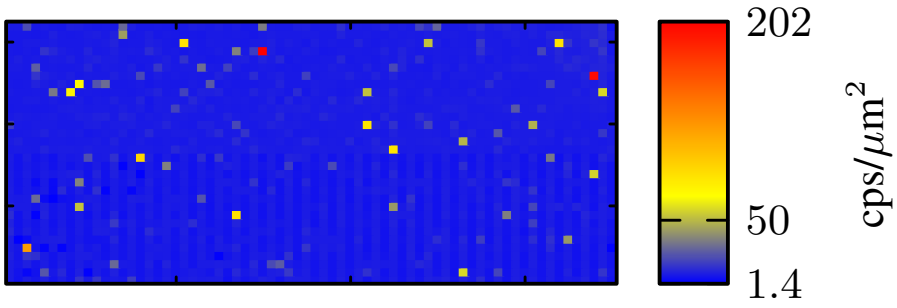


Figure 5.4: DCR spatial distribution of A1 array when tested with $V_{OP} = 33.3$ V and at room temperature.

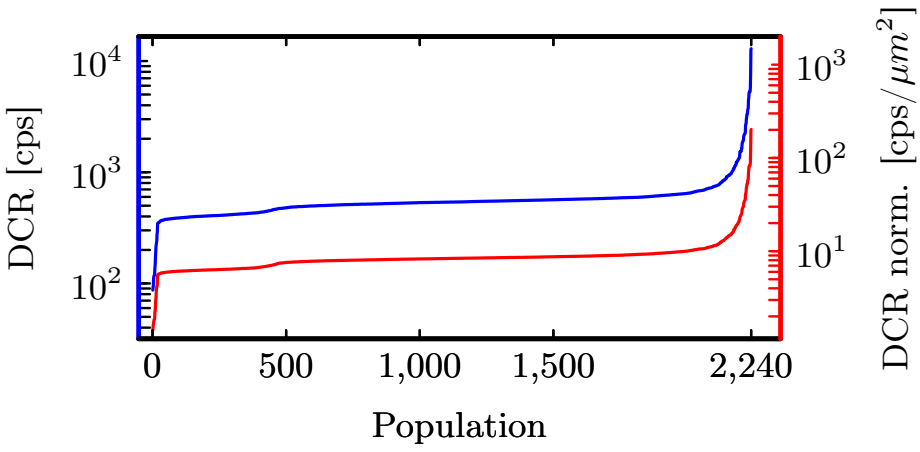


Figure 5.5: DCR showing A1 results for both normalized and non-normalized case at room temperature. This plot provides a simple visualization of hot pixel population.

The FortunaSPAD was the first SPAD sensor design in the GF 55 nm process, containing a large array ($>$ 1000 pixels). Therefore, this was the first time a large amount of statistics on detector noise and breakdown voltage spread could be obtained. These measurements were first performed for understanding the reliability of the 55 nm process before bit generation performance is assessed. Measurement of the DCR for array A1 was first performed by

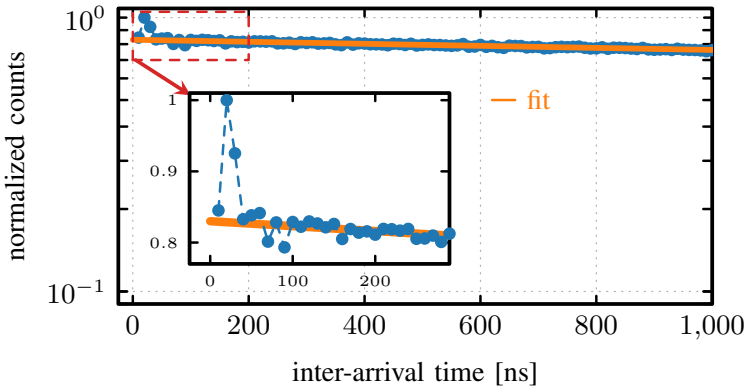


Figure 5.6: Inter-arrival time histogram of Fortuna test pixel measured at $\tau_{\text{dead}} = 8$ ns and room temperature.

bypassing the random bit generation circuitry. The results for room temperature analysis is shown in Figure 5.4 and Figure 5.5. DCR remains relatively low with $\simeq$ 95 % of pixels staying below < 700 cps and only three so-called 'hot' pixels having > 10000 cps. Moreover, no so-called 'screamers', where DCR is $>>$ 100000 cps the mean value, are present. This further highlights the suitability of the 55 nm BCD process for SPAD sensor development.

As was performed for individual SPAD pixels in Chapter 3, measurement of afterpulsing is performed on a FortunaSPAD test pixel using the inter-arrival histogramming method. An active probe is used with a fast oscilloscope (40 GS/s oscilloscope Teledyne LeCroy WaveMaster 813 Zi-B) with a bin width of 10 ns. The pixel dead time was $\simeq \tau_{\text{dead}} = 8$ ns. A low amount of light was added to achieve a count rate of $\simeq$ 1 kcps to speed up the measurement time. As expected, the afterpulsing is very low with an extracted value of 0.005 %. Moreover, it can be seen that the traps decay after approximately 100 ns. Based on the afterpulsing analysis performed in Chapter 4, correlation should not be adversely affected by afterpulsing.

Breakdown voltage non-uniformity

Non-uniformity of the breakdown voltage is performed by observing serial bias of individual QRFFs at a constant illumination. This way, a suitable V_{OP} value can be set for adequate performance of each pixel. The spatial analysis at three separate V_{OP} values is shown in Figure 5.7 by plotting the bit bias of each pixel at a constant illumination of $I_{LED} = 2$ mA.

(a) $V_{OP} = 32.8$ V. **(b)** $V_{OP} = 33.1$ V. **(c)** $V_{OP} = 33.3$ V.

Figure 5.7: Spatial bias analysis of bias at $V_{OP} = 32.8$, $V_{OP} = 33.1$, and $V_{OP} = 33.3$ V for analysis of the breakdown voltage spread. Parameter settings: $I_{LED} = 2$, $\eta = 0.71$, $f_{BG} = 5$ MHz.

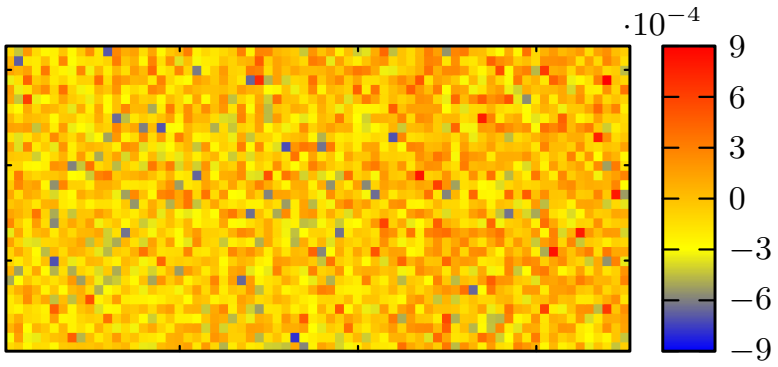


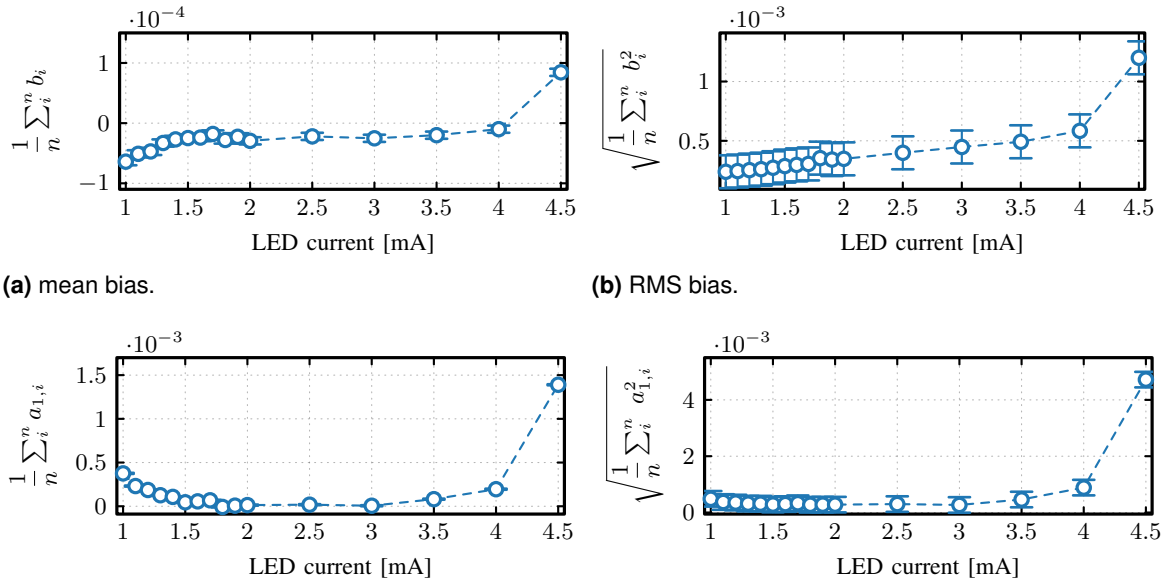
Figure 5.8: A1 array spatial bias map from $P(X = 1) = 0.5$ (Figure 5.7c re-plot with appropriate scale) at $I_{LED} = 2$ and a normalized voltage setting $\eta = 0.71$ and a bit generation rate of $f_{BG} = 5$ MHz.

Clearly, there exists a non-insignificant spread of breakdown voltages across the array. At 32.9 V, there are many pixels that are not operating at an excess bias which is high enough to consistently trigger the inverter threshold. This results in large bit bias. However, when the SPAD bias is increased to 33.3 V, the uniformity is improved with all pixels functioning as expected. Process improvements, which result in more consistent doping levels of the buried and deep layers that form the junction, could improve the spread of breakdown voltages. The results from Figure 5.7c are re-plotted with an appropriate scale in Figure 5.8. This measurement is performed with an illumination current of $I_{LED} = 2$ mA and $\eta = 0.71$. Under these conditions, it is shown that every single pixel achieves a bit bias within the desired benchmark of 10^{-3}, with no obvious spatial concentration centers. A more complete analysis of bias and correlation with swept illumination and threshold settings is detailed in the following section.

A1 bit generation performance

Ideally, all characterization would be performed by knowing the exact count rate of each pixel under any given illumination setting. The FortunaSPAD showed large variation ($> 10\%$) in dead time across the array caused by non-ideal feedback loop design in the pixel circuits. As mentioned previously, the seperation between SPAD and pixel circuits was artificially increased for isolation purposes. In order to maintain a reasonable fill factor and overall die size, minimum sized transistors were used in the tunable delay element of the recharge loop path. Moreover, as shown in Chapter 4, the ideal dead time is between 5-10 ns. Therefore, the hold voltage setting (V_H) was set to a value where all pixels were operating with a low dead time ($V_H = 0.65$ V). From the variation in the dead time of pixels and the long path lengths of the counting bus within the chips, it was not possible to get accurate count values for all the pixels at these settings, due to the short pulse width. Therefore, all subsequent measurements are with reference to the LED current as opposed to absolute count rate. For analyses performed in this section, a column word is repeatedly generated at the sampling rate (f_{BG}) until a statistically significant number of bits is generated.

The root-mean-square (RMS) and mean values of each QRFF bias, b, and serial correlation, a_1



(a) mean bias.

(b) RMS bias.

(c) mean 1-bit lag serial correaltion.

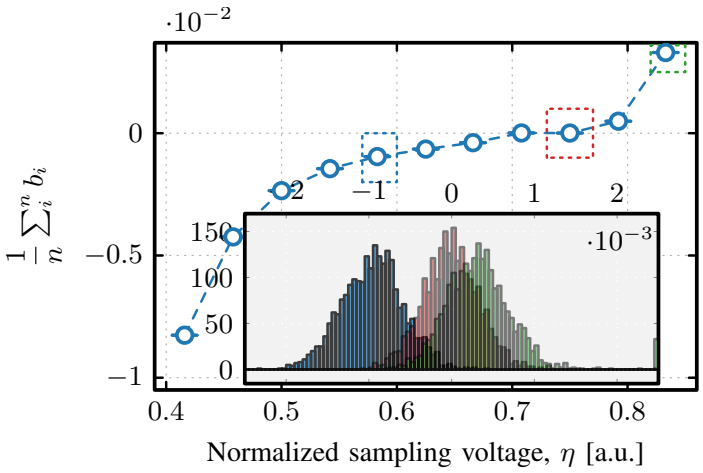
(d) mean 1-bit lag serial correaltion.

Figure 5.9: Bias and correlation analysis using serial data sorted pixel wise of all QRFFs in A1 as a function of LED current. Data is generated at $f_{\mathrm{BG}} = 5$ MHz. Threshold setting is held constant at $\eta = 0.71$ i.e. $\mathrm{V_T} = 0.85$ V.

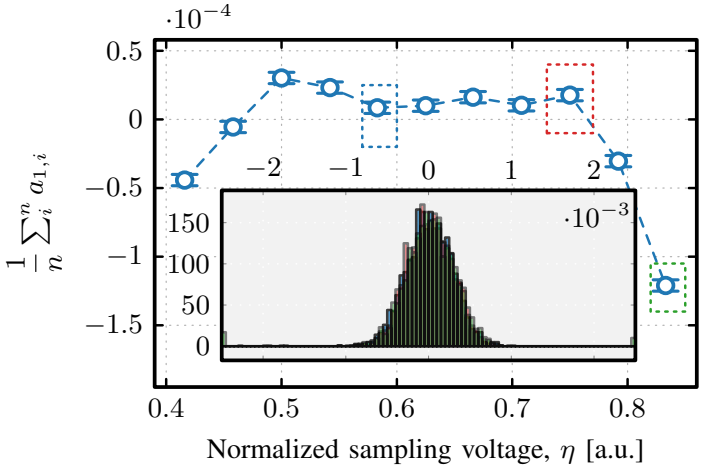
are detailed as a function of the illumination current ($\mathrm{I_{LED}}$) in Figure 5.9. The results verify certain aspects of the model, while also revealing a deviation from the expected performance under certain circumstances. First, from the perspective of bias, the RMS plot shows that the bias increases linearly with count rate between 2-4 mA, as expected based on the analysis in Chapter 4. However, this is not completely the case at low illumination settings, where the magnitude of bias appears to increase non-linearly. Nevertheless, the bias remains very low up until 4 mA. At this point, pile-up effects appear to degrade entropy significantly, as verified by the plot of correlation in Figure 5.9, where both the mean and RMS values increase. The larger relative increase in the RMS value shows that a selected number of pixels become effectively paralyzed. Overall, these measurements highlight well an advantage of generating random bits using the SC QRFF method, which is that it can perform well under a relatively large flux range (1-4 mA) in an array of thousands of SPADs. Moreover, improved pixel designs with precise dead time selections with less variance on PVT would improve on this range.

Threshold voltage analysis

A similar analysis is performed for the threshold voltage setting of the QRFF. Previously, in Chapter 4, it was shown that adjustment of the threshold voltage could effectively reduce bias to a negligible value. Now, the bias variation is observed over an entire array. The results for mean bias of all QRFFs are shown in Figure 5.10. Three separate points along the curve also contain a histogram of bias for all individual QRFFs. This shows that increasing the



(a) b. Centering of bias happens at roughly $\eta \simeq 0.7$.



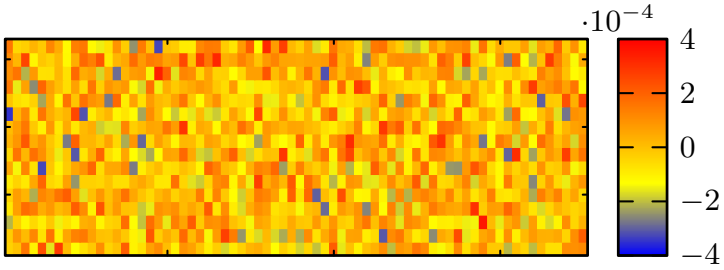
(b) a_1, Correlation remains unaffected by change in threshold until η is increased to the point where the comparator offset causes QRFFs to become stuck in a praticular state.

Figure 5.10: Bias and serial correlation of QRFFs in the A1 array as a function of threshold voltage setting $\mathrm{V_T}$. Measurements performed with $\mathrm{I_{LED}} = 2.5$ mA and f_{BG}=5 MHz.

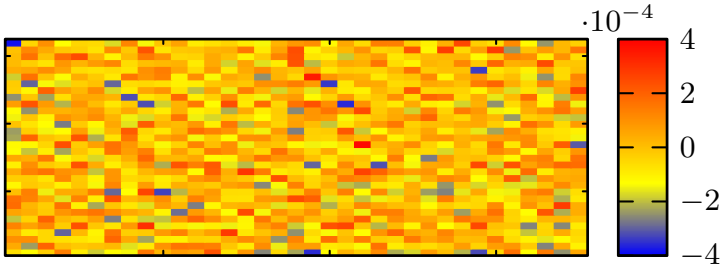
threshold voltage centers the histogram of bias. At a certain point ($\eta \simeq 0.85$), the comparator offset of certain pixels becomes significant enough that the state of the dynamic comparator in the sampling flip-flop is no longer switched. Therefore, a select number of pixels have very high bias and correlation (shown also by Figure 5.10b). Up until this point, serial correlation remains unaffected by the threshold voltage setting, as predicted. While the strength of the dynamic comparator approach is validated by these analyses, there clearly also remains some room for improvement. The model predicts a linear relationship between the threshold setting and bias. Moreover, this behavior was verified in simulation. However, from Figure 5.10a, this is clearly not the case. However, metastability, comparator offset, and asymmetric clock-to-q times were not modeled. This provides an opportunity for further research that can help facilitate the design of an improved QRFF circuit architecture.

Spatial correlation

Until this point, serial data of each QRFF was analyzed, without attention to potential spatial correlations that can also degrade performance. For this analysis, cross-correlation of adjacent columns and rows are analyzed when generated bit strings are saved in a single $\mathrm{CLK_{BG}}$ cycle. This analysis is conducted by generating words in two columns at a time, and then calculating the cross correlation once the data is sorted into bits produced by individual pixels. The results



(a) Vertically adjacent cross correlation analysys.



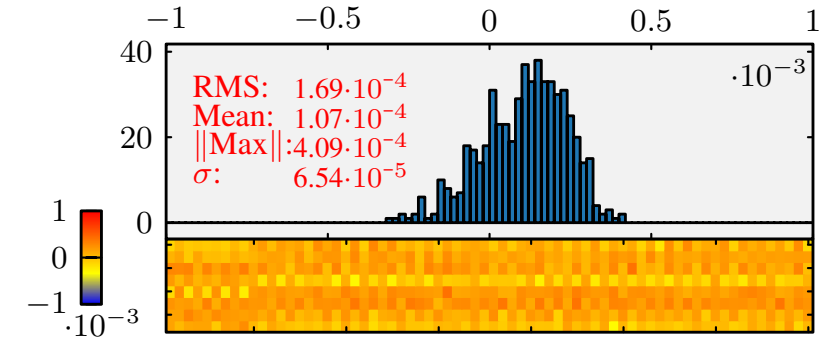
(b) Horizontally adjacent cross-correlation analysys.

Figure 5.11: Spatial bias maps showing the calculated cross-correlation of bits generated by adjacent pixels. Two full columns of data are generated in a single cycle at $f_{\mathrm{BG}} = 5$ MHz.

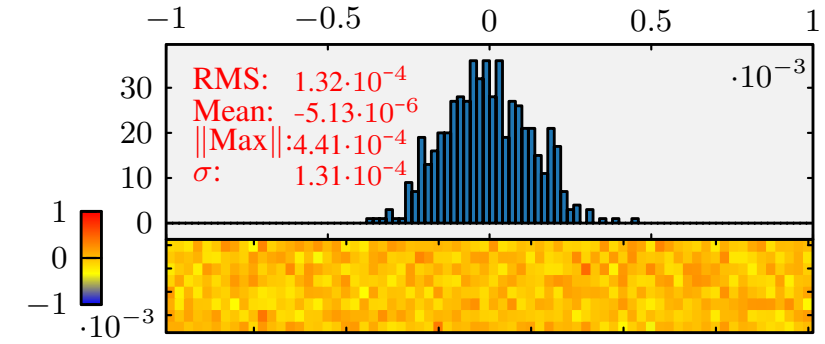
suggest that there are no evident spatial correlations present. Therefore, in the final test, where bits are continuously generated for a full frame in a single $\mathrm{CLK_{BG}}$ cycle, it is expected that cross talk should not adversely degrade the overall performance of the FortunaSPAD. This is verified with statistical testing and entropy estimation. Moreover, this spatial analysis is considered representative of the A2 array as well, since the pixel pitch/ QRFF structure is identical.

A2 bit generation performance

Originally, the serialized array was designed to work at a speed of 400 Mbps/per channel. However, a locking issue with the on-chip PLL at higher frequencies prevented the intended data rate from being achieved. The maximum speed per channel tested was 160 Mbps, although the most consistent results came at 140 Mbps. This translates to a bit generation rate of $f_{\mathrm{BG}} = 2$ MHz. Using the strobe signal described earlier for sorting, pixel-wise bias and correlation are calculated, as was performed for the A1 array. Results are show in Figure 5.12. All pixels in A2 achieve a bias and serial correlation coefficient well within benchmark of



(a) b



(b) a_1

Figure 5.12: Spatial bias and correlation analysis of the A2 array. Each channel generates 140 Mbps of data i.e. $f_{BG} = 2$ MHz, $I_{LED} = 2$ mA and $\eta \simeq 0.71$. Sorting of data is performed using the STROBE signal output from the ASIC that aligns with the first serialized bit in the word (QRFF[0]).

10^{-4}. The max calculated bias and correlation for any QRFF are, $4.09{\cdot}10^{-4}$ and $4.41{\cdot}10^{-4}$, respectively, with RMS values across the array of $1.69{\cdot}10^{-4}$ and $1.32{\cdot}10^{-4}$, respectively. A non-zero mean in the bias suggests that the ideal threshold setting for this array is slightly lower than is the case for A1.

Overall, both arrays demonstrate very low bias and correlation when operated within an illumination range of $I_{LED} = 2 - 4$ mA and a threshold setting of $\eta \simeq 0.5 - 0.8$ with the best performance coming at $I_{LED} = 2$ and $\eta \simeq 0.71$. Under these conditions, all 2800 QRFFs achieve the correlation and bias benchmarks of 10^{-3} which translates to an estimated $H_1 \geq 0.999997$ and $H_\infty \geq 0.9986$ based on the analysis introduced in Chapter 1 i.e Equations (1.6) and (1.10).

It is assumed, since no evident spatial correlation is available, that worst case serial data can be used as a rudimentary entropy evaluation. Per QRFF characterization, for both arrays, within a normalized threshold range 0.65 - 0.8 (V_T = 0.76 - 0.92 V), and an illumination range of 1.5 - 3 mA, shows that the poorest performing pixel results to b = $-2.33{\cdot}10^{-3}$. This is an estimated $H_1 \simeq 0.99998$. Correspondingly, the worst QRFF serial autocorrelation is $a_1 = 4.26{\cdot}10^{-3}$. This results in an estimated $H_\infty \simeq 0.994$ using Equation (1.10). Formal estimation of entropy using NIST SP 800-90B is shown in the following section.

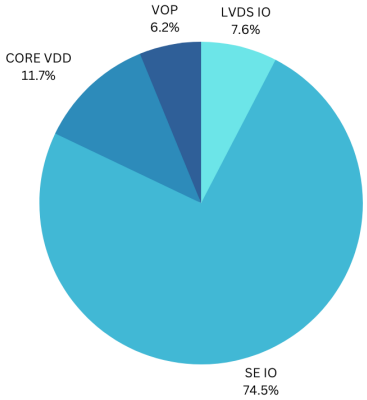


Figure 5.13: FortunaSPAD power consumption measurement. Measurement performed at $I_{LED} = 2$ mA, $\eta = 0.71$, and $V_{OP} = 33.3$ V. Overall the power consumption is 243 mW.

Power consumption

Power consumption of the FortunaSPAD (without FPGA or LED contributions) was characterized by monitoring the current of each individual supply: single-ended (SE) IOs, LVDS IOs, SPAD operating voltage (V_{OP}) and the core supply VDD. The results are displayed in 5.13. The overall power consumption is 243 mW with the majority being taken up by the single ended IOs (32 in total). The measurement was performed with the FortunaSPAD operating at it's maximum output data rate with 2 mA illumination, at room temperature.

5.3 NIST SP 800-22 and SP 800-90B randomness testing

5.3.1 Theory

Bias and correlation of generated bit strings from the FortunaSPAD have been extensively analyzed within this chapter. However, randomness qualification requires a more comprehensive analysis which makes use of randomness testing suites. As discussed in Chapter 1, formal standardization of RNGs is ongoing both in NIST and AIS, and test suites such as the NIST SP 800-22 are being updated to address limitations and flaws pointed out by industry/academia. Nevertheless, offline randomness testing is still an accepted *part* of randomness qualification. Moreover, it is now recommended by NIST to use the STS (800-22), in conjunction with the SP 800-22 test suite.

Both suites implement a variety of tests that calculate a *test statistic* (such as the number of 1s or 0s in a sequence i.e. bias) and then comparing the distribution of the test statistic to the distribution that would be expected if the sequence were truly random. Therefore, the null hypothesis (H0) is defined as the sequence under test being random. Conversely, the alternative hypothesis (Ha) is the sequence not being random. Indeed, there exists four

different conclusive scenarios, described by random/not random data accepting H0/Ha. As a result, there can be two possible error types. Type I errors are considered as acceptance of Ha (rejection of H0) when the data was in fact indeed random. A Type II is the inverse, where H0 was accepted when the data is not random. The probability of a Type I error is referred to as the significance level, α, whereas β outlines the probability of a Type II error. The calculation of the Type II error β is evidently difficult compared to, α, as there exists many different types/signs of non-randomness. SP 800-90B only checks for Type I errors. In 800-22 tests, β is calculated internally by the suite depending on the chosen α, the sequence length, n, and the type of test, in an effort to minimize the probability of Type II errors. In every test, a p-value is calculated, based on the test statistic, to determine the likelihood against the null hypothesis. Therefore, given a significance level, α, if the determined p-value $\geq \alpha$, then the sequence can be considered random with a confidence level of $(1 - \alpha \cdot 100)$ %. The p-value is determined by comparing the calculated test statistic to the critical values of the test's probability distribution using either the incomplete gamma function or complementary error function. For example, in the Frequency (Monobit) Test, the test statistic is the number of ones in a sequence of bits. The p-value is calculated by comparing the calculated test statistic to the probability distribution of the number of ones in a truly random sequence of bits.

The NIST SP 800-22 STS includes a set of 15 statistical tests that can be divided into four categories. A clear pass/fail criteria for each test. However, it's important to note that, no single test or analysis can prove that a random number generator is truly random. The categories of tests included in the suite are:

1. Frequency: these tests are used to check whether the distribution of bits in the string is uniform. They include the Frequency (Monobit), Block Frequency, Run, Longest run of ones.
2. Repetitive pattern/spectral: These tests check for periodic ties or repetitive patterns. They include the Discrete Fourier Transform, and Binary Matrix Rank tests.
3. Pattern matching: These test for too many occurrences of an aperiodic pattern and entropy degradation caused by correlations. They include Overlapping/Non-Overlapping Template matching tests, Maurer's, Linear Complexity, Serial, and Approximate Entropy.
4. Random walk: These tests check for correlations and patterns and returns/visits to a particular value. They include the Cumulative Sums, Random Excursion and Random Excursion Variant tests.

The major components of the 800-90B suite are entropy estimation and testing of the iid assumption. Testing of this assumption is performed using a variety of permutation tests and additional Chi-square statistical tests. A permutation test is a type of non-parametric analysis used to determine whether two groups of data are significantly different from each other. It is a distribution-free test, meaning that it does not rely on any assumptions about the underlying distribution of the data. A test statistic is calculated for the original sequence, and then the test is repeated by randomly permuting the bits in the sequence and recalculating the test statistic, 10000 times. Then the distribution of the test statistic is compared to the distribution that would be expected if the sequence were truly random. The result is then considered to be consistent with randomness if the test statistic for the original sequence falls

Test	Min. pass rate	p-value	Pass rate
Frequency	996	0.8831	998/1000
Block frequency	996	0.0278	1000/1000
Cumulative sums	996	0.1855	997/1000
Runs	996	0.4521	999/1000
Longest run	996	0.4885	998/1000
Rank	996	0.9723	998/1000
FFT	996	0.1364	999/1000
Non overlapping template	996	0.8429	997/1000
Overlapping template	996	0.6454	998/1000
Universal	996	0.7830	1000/1000
Approximate entropy	996	0.5769	1000/1000
Random excursions	616	0.3258	618/619
Random excursions variant	616	0.5457	616/619
Serial	996	0.9737	1000/1000
Linear complexity	996	0.5523	999/1000

Table 5.1: Sample summary of FortunaSPAD NIST SP 800-22 results. Data generated at 3.3 Gbps overall rate with parameters: $\eta \simeq 0.71$, $\mathrm{I_{LED}} = 2$ mA.

within the range of the permuted test statistics. Testing can be done for a wide range of properties of randomness such as the distribution of bits, the distribution of runs, and the distribution of autocorrelations. The SP 800-90B permutation tests use a variety of these subtests to *find evidence that* the samples are not iid. The chi-squared test compares the observed frequencies of events in a bit string to the expected frequencies. A test statistic is calculated as the sum of the squares of the differences between the observed and expected frequencies, divided by the expected frequencies. Two types of general subtests exist, one set that aims to find dependencies in the data and another the length of longest repeated sub-strings. If iid. testing fails, then it is recommended by the publication to use a separate set of estimation methods (non-iid track) for the min-entropy tests, although some of those estimators have proven to greatly underestimate the min entropy [189]–[191]. Otherwise, if iid tests are successful, the most-common value (MCV) method is used to estimate min-entropy. Additionally, NIST 800-90B recommends the use of block-based estimations, which involve breaking the data into fixed-size blocks and calculating the entropy of each block separately. This method helps detect non-random patterns that may not be apparent in the entire dataset, and it's especially useful when the data has temporal dependencies.

5.3.2 Results

NIST STS testing is performed with a strict significance level, $\alpha = 0.001$ using 1 Gb of data generated with a full frame for both arrays, as explained above, split into 1000 bit strings. The results for the NIST test are outlined in Table 5.1 with all tests passing. The third, p-value

column shows the value from the Chi-squared test, which checks the uniformity of the p-values.

The same binary file was also used for testing using the 800-90B test suite. All tests including the chi squared, longest repeated, and permutation, are passed, thereby confirming the random data is independent and identically distributed (iid). The estimated min entropy calculated by the suite, using MCV method, is $H_\infty \simeq 0.9954$.

5.4 Discussion and state-of-the-art comparison

New QRNG implementations are appearing regularly from both industry and academia. The discussion and comparison of related works and the state-of-the-art is based off SPAD array sensors, which integrate the bit generation methods on the same die as the SPADs. This allows for the most relevant comparison. A summary of works within these criteria is outlined in Table 5.2. In what perhaps was the first work to demonstrate the use of SPAD image sensors as QRNGs, Burri et al., used 512×128×2 pixels in a gated mode operation to see the probability that a photon was detected by each individual pixel. However, Von Neumann filtering was required to achieve acceptable bias, with a total output data rate of 5 Gbps (only 40 kbps per pixel) [192]. Tisa et al., improved on the per pixel throughput by using an LFSR based counter of SPAD pulses [193]. A whitening algorithm was used and a relatively arbitrary choice of post-processing was implemented by XOR of the SPAD generated bits with a PRNG. Very little modeling of the bit generation method was provided. The FA bit generation method presented in Chapter 4 was used in [180], [181]. The dynamic range of light intensity in [181] of 46 was improved upon from [180] (40). However, in that work an acceptable bias is set to 0.01., to calculate this value, which is not acceptable for cryptographic applications. No detailed correlation modelling or analysis is provided in either of those works. Regazzoni et al., performed detailed modelling of entropy in their system by evaluating the min-entropy conditioned on the available side information such as the photon distribution. Thus, a proof of the quantum nature of their generator is provided, with an output data rate of 400 Mbps achieved. Research is trending towards full monolithic integration of these sensors, i.e. with illumination included on die. Although still nascent, some promising works have been demonstrated. In [178], an emitter structure was used to produce photons for a SPAD to detect, achieving a single pixel output rate of 100 kbps. This concept was scaled to a full system including on chip control and post-processing of a $16 \times 8 \times 2$ sensor was presented in [194]. In that work, TDCs were included on chip to measure the time difference between photon arrival events, and a correlater circuit with a programmable window was used to check that detected photons arrived from the emitter. However, only 400 kbps total data output was achieved.

In this work, the FortunaSPAD was designed with an Entropy-as-a-service application in mind. Therefore, the larger relative power consumption of a system that uses free running SPAD detections, could be tolerated. Moreover, as opposed to relying on the quantum nature of the device, bias and correlation were systematically modelled, with simulation and experiment showing the efficacy of the model. All pixels on the FortunaSPAD were shown to operate well when generating 5 Mbps, with single pixel analysis showing 10 Mbps as viable as well.

These are some of the highest per pixel bit generation rates demonstrated to date. The key to accomplishing these rates, while maintaining entropy commensurate with those required for cryptographic applications, was the QRFF threshold control, and the dead time control of the SPAD. The combined data rate from both arrays on the FortunaSPAD QRNG is 3.3 Gbps.

Ref. & Year	Physical Principle	Circuit/Extraction Implementation	Array Size	Bitrate (per pixel)	Further Post Processing	Evaluation Method
[192] Burri, 2013.	SPAD triggering prob.	Frame readout	512×128 $\times 2^{\alpha}$	5 Gbps ($\simeq$ 0.04 Mbps)	Von Neumann filter	NIST STS DIEHARD
[193] Tisa, 2015.	SPAD triggering prob.	LFSR counter	32x32	200 Mbps ($\simeq$ 0.2 Mbps)	Whitening algorithm	DIEHARDER TestU01
[180] Massari, 2016.	First detected photon	Inter-arrival arbiter	16×16	128 Mbps (0.5 Mbps)	none	NIST STS
[178] Acerbi, 2018.	SPAD triggering prob.	Frame readout	1	$\simeq$ 0.1 Mbps	none	NIST STS
[181] Xu, 2018.	First detected photon	Inter-arrival arbiter	16×16	18 Mbps ($\simeq$ 0.07 Mbps)	none	NIST STS
[195] Regazzoni, 2021	Photon detection	Vector matrix multiplication†	128×128†	400 Mbps ($\simeq$ 0.024 Mbps)	none	NIST STS Diehard
[194] Massari, 2022	Photon timing statistics	Time-difference of photon arrival$^{\beta}$	16×8×2	400 kbps ($\simeq$ 1.57 kbps)	Linear corrector Elias' Integer addition	NIST SP 800-90B AIS 31
This work$^{\zeta}$	Photon timing statistics	QRFF	40×70‡	3.3 Gbps ($\simeq$ 1.2 Mbps)$^{\bowtie}$ (10 Mbps)*	none	NIST STS NIST SP 800-90B

Table 5.2: Published Integrated SPAD-Based QRNGs with bit-generation/extraction on chip.

† a fixed matrix and reconfigurable matrix are included on chip
‡ two independent arrays (70 x 32 & 70 x 8) with different readout architectures
$\bowtie$ calculated per pixel throughput with both arrays readout, limited by readout and IO speeds
* single pixel capable of $H_1 \geq 0.99997$ and $H_\infty \geq 0.98$.
$^{\alpha}$ two of the same die were used in parallel.
$^{\beta}$ a correlator circuit is used to validate that photons arrived from the emitter.
$^{\zeta}$ **Bias and correlation modelled with considerations for detector and circuit. Validated by simulation and measurement.**

6 FortunaSPAD2: A SPAD-based QRNG with robust macro-QRFF pixels

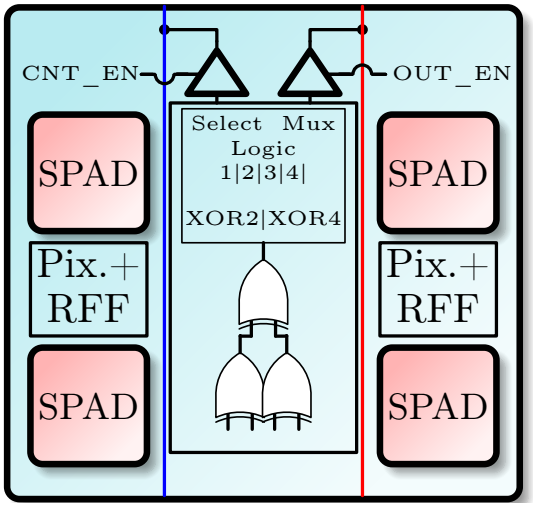
This chapter presents a follow-up to the FortunaSPAD sensor. Two main research goals were set for this part of the thesis:

1. Improve robustness, i.e. fault tolerance due to environmental changes or pixel failure.
2. Improve total throughput with a simplified architecture/readout scheme.

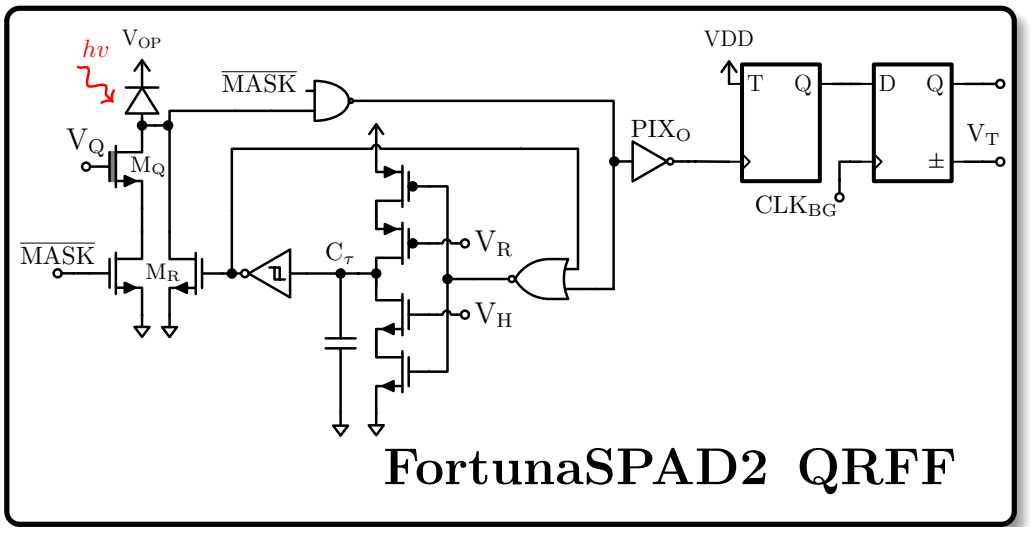
The FortunaSPAD, presented in the previous chapter, demonstrated excellent random bit generation statistics, in terms of bias and correlations. Statistical testing and entropy estimation validated the overall performance of the chip. It was shown that the FortunaSPAD was capable of functioning well across a range of photon flux, threshold voltages, and generation rates. However, no form of post-processing or capability to mitigate against pixel failure or changes in environment, were included in the design. For this reason, the FortunaSPAD2 is proposed. Moreover, a method to monitor the count rate of any individual pixel, without disruption of the bit generation process, is included. The goal of this part of the thesis is to demonstrate that, with the future integration of illumination, and online health tests, this architecture could be a candidate QRNG technology for cryptographic applications.

6.1 Design of an improved macro-pixel QRFF

As mentioned in Chapter 1, the simplest randomness extractor is the XOR gate, which provides a very area efficient way of combining independent entropy sources [196]. Two independent sources, each with bias, b, produce a string with a result bias, b^2, when they are passed through an XOR gate. However, due to its incapability of removing cross-correlations between inputs, the XOR gate is not suitable as a post-processing method for many entropy sources [197]. Given that the individual QRFFs in the FortunaSPAD exhibited no distinct cross-correlation between neighboring pixels, the design can be modelled as an array of independent entropy sources. Therefore, this allows for an area-efficient implementation of QRFF-wise XOR as a post-processing scheme. A slightly modified QRFF pixel was designed in the FortunaSPAD for enabling this functionality, as shown in Figure 6.1. In this design, a macro-pixel based random bit generator utilizing four QRFFs is presented. It consists of two output connections to a



(a) Block diagram of macro-pixel. Blue denotes the counting bus while the red signal is the generated bit.

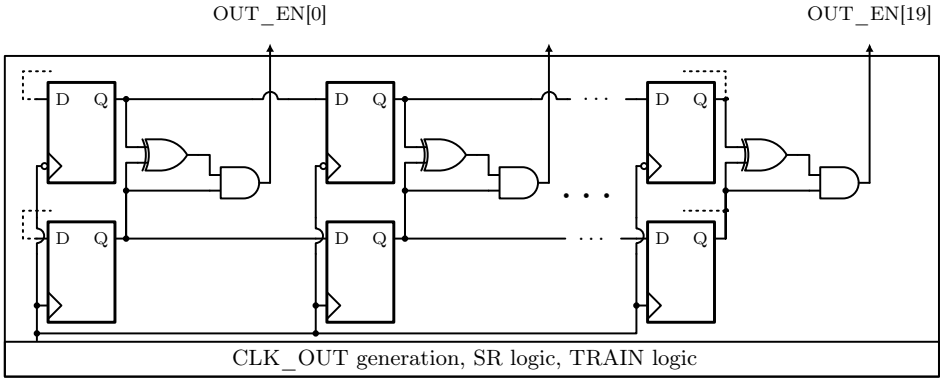


(b) Single QRFF with modified pixel electronics.

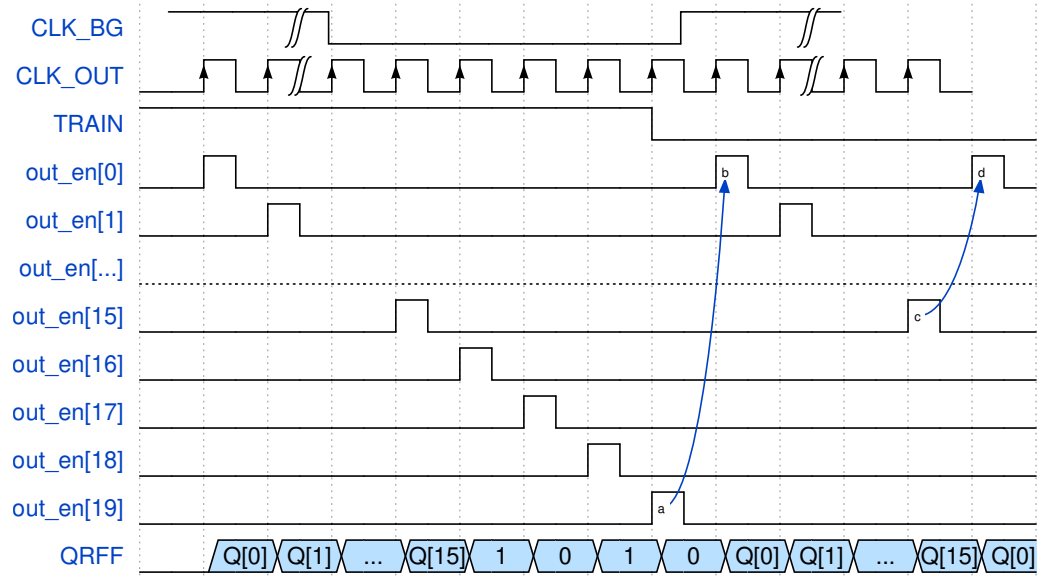
Figure 6.1: Macro-pixel based QRFF design for the FortunaSPAD2. Four QRFF designs are combined. Count rates and random data are output on the column blue/red buses, respectively. Count rates and generated random bits from any individual QRFF can be selected. The two QRFFs in a column can also have their random bits XOR'd, or all 4 can be selected for XOR ($Q = 1 \oplus 2, 3 \oplus 4, (1 \oplus 2) \oplus (3 \oplus 4)$). Pixel-wise masking is also implemented using a 1-bit memory cell in each pixel.

shared column bus. The first, denoted by the blue bus connection, is that counts coming from the pixels. Control logic, i.e. a simple mux, can be programmed to select any of the four pixels to output onto the column bus, when the tri-state buffer, controlled by CNT_EN, is activated. The generated bit is output on the shared column bus shown by the red signal. The pixel electronics from individual QRFFs include a 1-bit memory for masking, i.e. deactivating a SPAD from firing. Pixel masking is performed to avoid entropy degradation caused by outlier noisy pixels. Moreover, whenever a single pixel within a macro-QRFF, that is using the XOR feature, is masked, the generated bit statistics are governed by the remaining active QRFFs.

This is another advantage of the XOR macro-pixel structure. Monitoring of the count rate can be performed to determine which pixels are operating in an acceptable counting regime, based on the desired generation frequency, f_{BG}. Finally, the cascode transistor that was included in the FortunaSPAD, which allowed for operation at higher excess biases, is removed. Given the typical operation at $\sim V_{EX} = 1$ V, it was deemed unnecessary. Moreover, the loop dynamics of the pixel circuit are simplified, potentially improved stability depending on the recharge control. The remaining aspects of the QRFF are left identical.



(a) Output enable shift register general architecture.



(b) Timing diagram illustrating QRFF column readout.

Figure 6.2: Column readout circuit and timing diagram for the FortunaSPAD2. Readout is performed by enabling the tri-state buffer of each individual macro-QRFF output onto the column bus. A custom shift-register is designed to perform this with the general circuit architecture shown here. Tri-state buffers are only enabled for a clock-cycle for fast operation while avoiding bus contention. A train option is additionally available to allow for 4-bits of known output.

6.2 FortunaSPAD2 architecture and design

6.2.1 Readout circuitry

The readout of data in the FortunaSPAD2 is performed by sampling 16 macro-QRFFs onto a single column. The scheme works as follows. As implemented in the original FortunaSPAD, a global bit generation signal, CLK_{BG} is used to clock every QRFF. Subsequently, each macro-QRFF has its tri-state buffer enabled by a shift register that is controlled by the signal CLK_OUT. A timing diagram of this operation, along with the general structure of a custom shift-register, is shown in Figure 6.2. The shift-register only enables the output of each buffer for a half-cycle to facilitate fast operation while avoiding bus contention during the transition from one reading to another. Moreover, a training feature is included in the readout scheme. When the signal TRAIN is active, the 16 bit QRFF word from each column is augmented with four bits of known data output. This allows the FPGA to periodically calibrate the IO delay to ensure that valid data is being received. Each column is designed to be capable of handling at least 6 MHz QRFF clocking rate within one cycle, i.e. every column outputs data at 96 MHz.

6.2.2 System design

A block diagram of the complete sensor design is shown in Figure 6.3. The array is symmetric across the x-axis, with a total of 4096 QRFF pixels split into a $2 \times 32 \times 16$ configuration of macro-QRFFs. The columns in each half of the array go through a buffer tree and then directly to outputs. To reduce the total number of IOs, only 4 total counting line outputs are provided (2 top, 2 bottom), where a $32:2$ multiplexer is used to select the specific column. Moreover, the count column buses go through $\div 8$ blocks, as the pixels are capable of achieving dead times as low as 1.5 ns. This allows the FPGA to accurately count, even when the pulse of the SPAD is too narrow compared to the bandwidth limitations of the IOs (and additional PCB trace capacitance). Masking, and selection of the row, which has its count buffer enabled, are controlled by simple decoder circuits. The die area is 2.57 mm $\times$2.27 mm with vertical and horizontal pixel pitches of 31.5 μm and 28.8 μm, respectively. The die micrograph, along with the bonding PCB, are displayed in Figure 6.4. Moreover, the layout (only up to METAL2 layer) of the full-custom macro-QRFF is also shown to demonstrate the complexity and density of the pixel. A total of 64 random data single-ended outputs enables a designed total throughput of $\sim$ 6.4 Gbps with continuous operation. FortunaSPAD2 features were designed with flexibility in mind, i.e. to provide reconfigurable operation, depending on the entropy requirements, the acceptable power consumption, and the desired throughput. As a conservative measure, configuration pins on the macro-QRFF can be set for XOR of 4 individual QRFFs. However, pixels can also be masked, reducing total activity, while still allowing for XOR of 2 pixels. Moreover, as demonstrated in Chapters 4 and 5, the advantage of the slow-clock QRFF architecture is that it allows for adjustment of the throughput with varied photon detection rate. This provides another variable for addressing the power consumption/throughput trade-off.

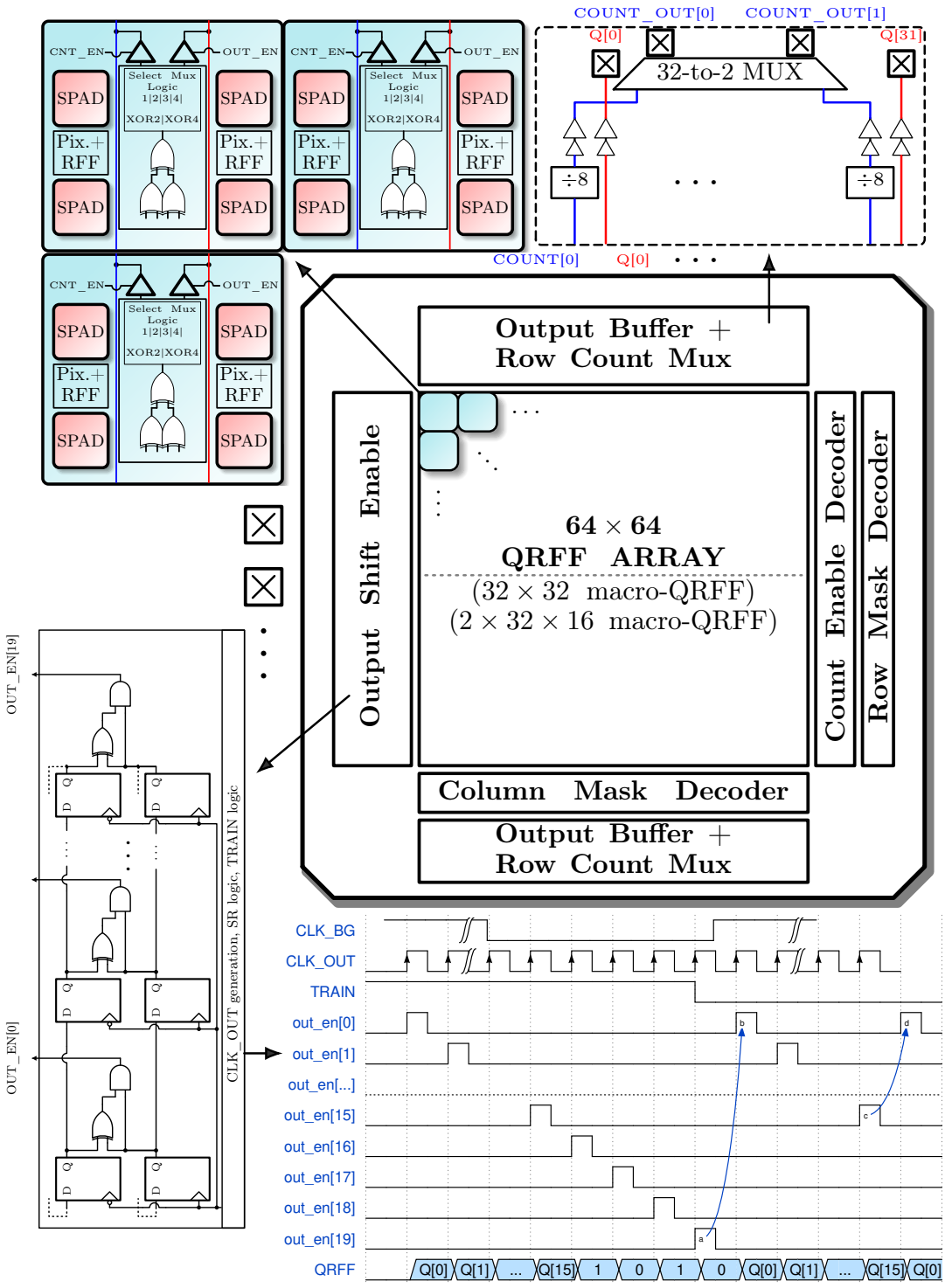


Figure 6.3: Block diagram of the FortunaSPAD2. The 64×64 pixel (SPADs) sensor is symmetrical across the x-axis and consists of a total of $2 \times 32 \times 16$ macro-QRFFs.

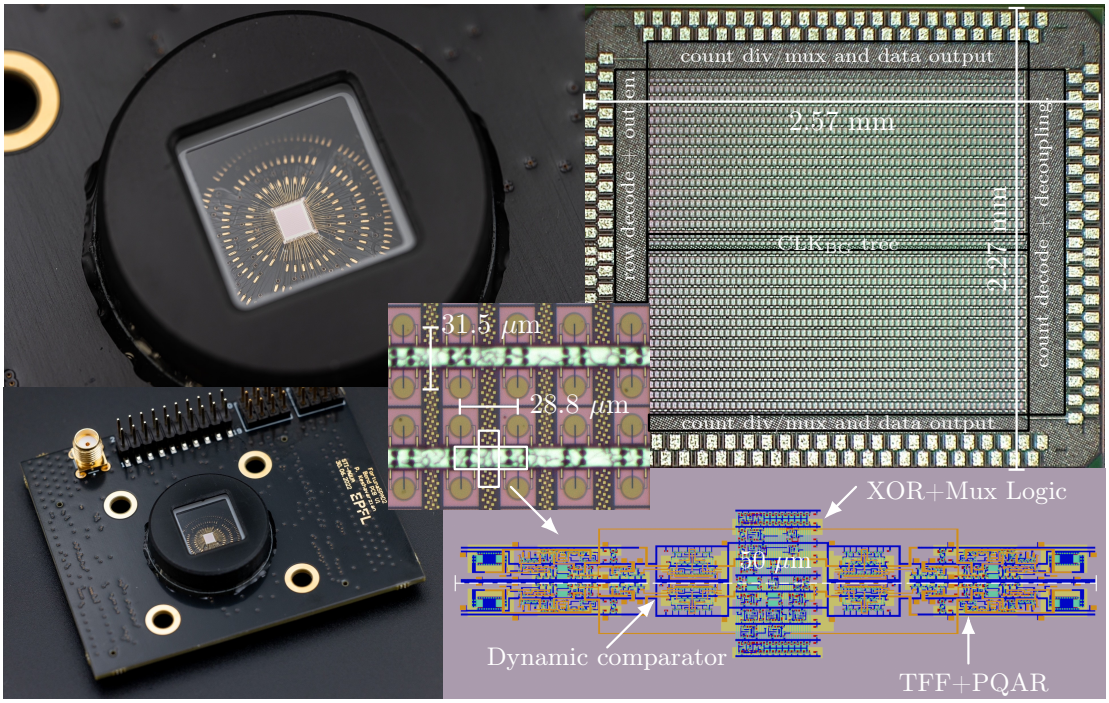


Figure 6.4: Micrograph and bonding PCB of the FortunaSPAD2. Layout of the macro-QRFF and its corresponding pixel circuit blocks are also shown.

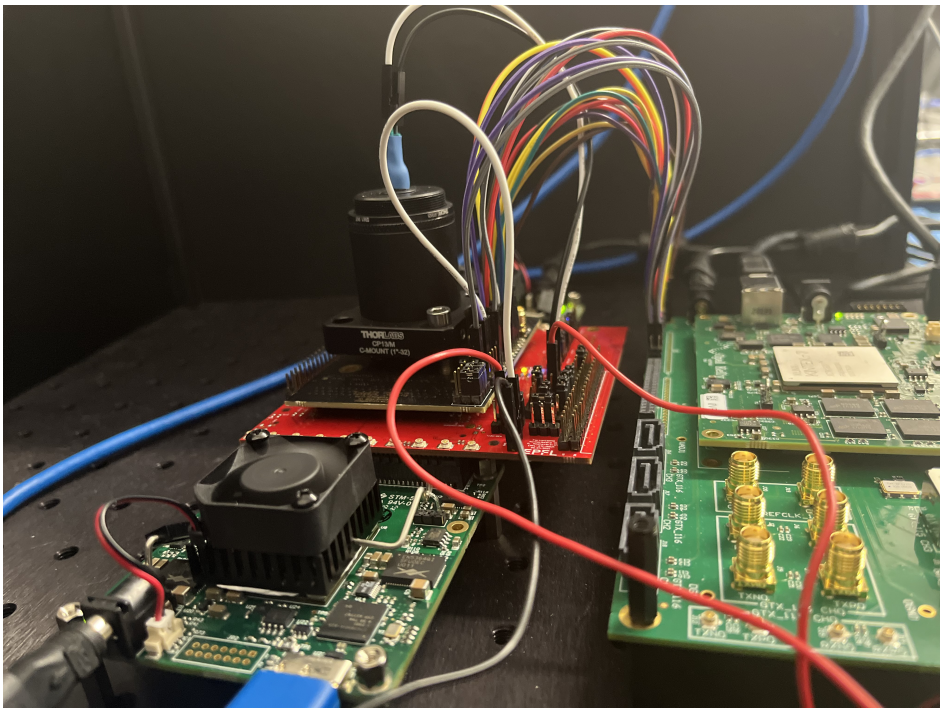


Figure 6.5: Complete characterization setup for FortunaSPAD2. The die is placed below the black optical tube which houses the LED.

6.3 Measurements

Characterization of the FortunaSPAD2 is in its early stages. In this thesis, the measurement results of a macro-QRFF is presented, with a particular emphasis on demonstrating how it can operate under a larger range of λ_A/f_{BG} ratios when using its pixel XOR capability.

6.3.1 Single macro-pixel

The counting capability of all four pixels inside of a macro-QRFF is shown in Figure 6.6. High non-linearity of the current-starved inverter is visible at $V_H = 0.45$ V, which is close to the threshold of the NMOS transistor. At $V_H = 1$ V count rate, compression is seen around 4 mA, although maintaining a dead time low enough for counting above 70 Mcps. The recharge control voltage was kept constant at $V_R = 0.35$ V for these measurements. These results are used to determine illumination settings for correlation measurements at specific λ_D/f_{BG} ratios.

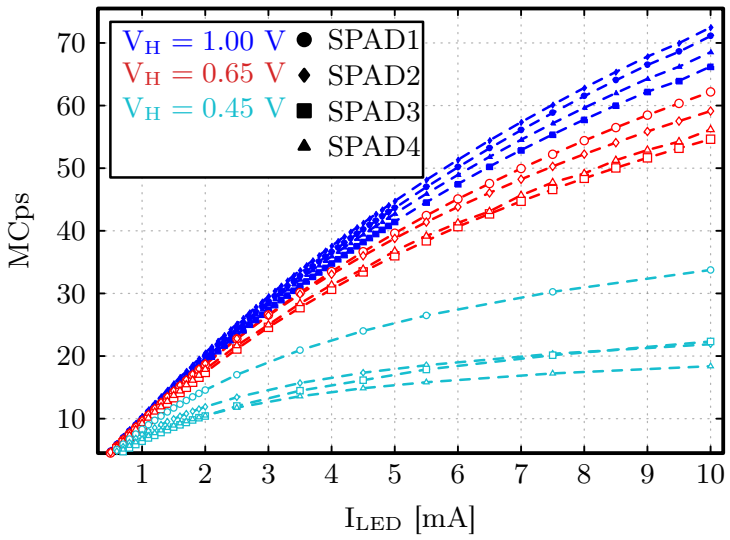


Figure 6.6: Measured counts of all four pixels in a single macro-QRFF as a function of illumination current with three separate settings for the hold voltage control V_H.

An analysis on the threshold voltage on bias is also performed, to verify that the QRFF is operating as expected based on results from the FortunaSPAD. Measurement results are shown in Figure 6.7. In general, the behavior of the four QRFFs represent what was observed in the FortunaSPAD, which is that a relatively high $\eta \simeq 0.75$, is required to minimize the bit bias. Nevertheless, these three QRFFs again demonstrate low bias ($< 10^{-3}$) within a range of $0.55 < \eta < 0.75$. Moreover, as mentioned previously, XOR of two independent random bit generators, with bias b, results in a reduced bias of b^2. In the following section, the statistical independence of QRFFs is analyzed by viewing the serial correlation of XOR'd bits in a macro-QRFF design.

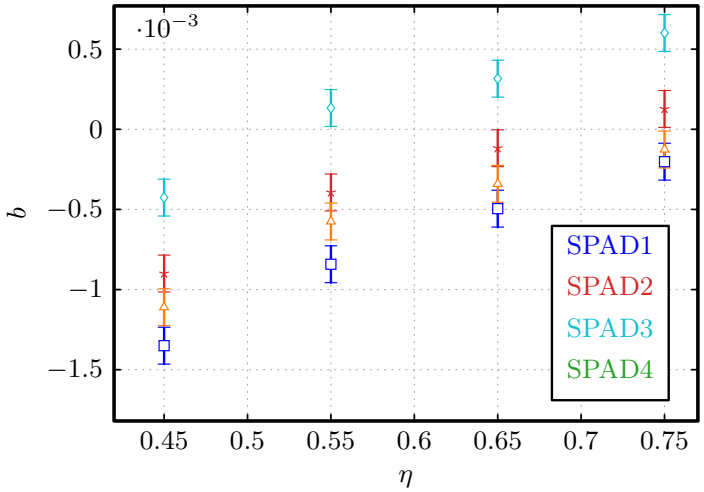
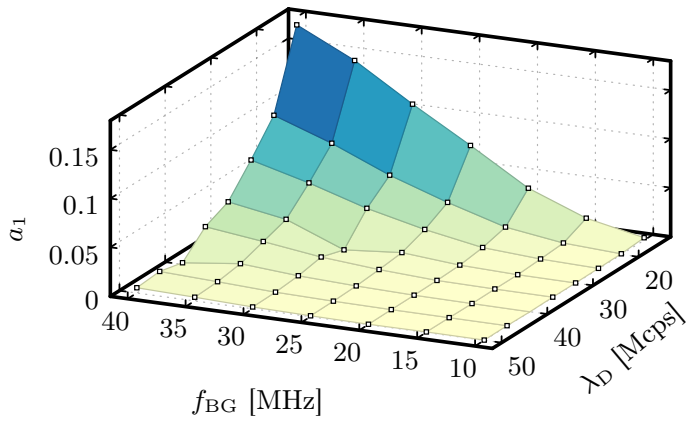


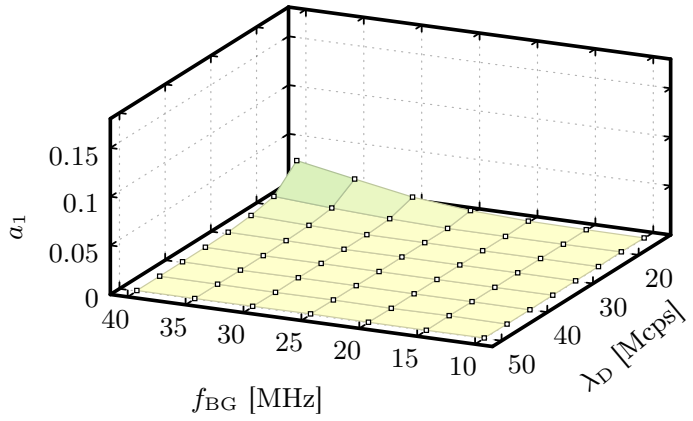
Figure 6.7: Measured bias as a function of threshold voltage, at a constant illumination of $I_{LED} = 2$ mA, $V_H = 0.65$ V, $V_R = 0.35$ V, and $f_{BG} = 5$ MHz.

Statistical independence of QRFFs in a macro-pixel

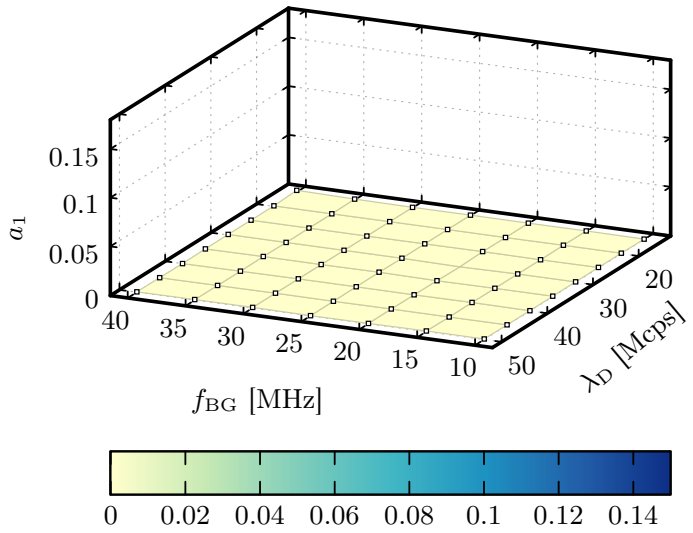
Three sub-plots are shown in Figure 6.8. These measurements were conducted with the same objective as in Chapter 4, which is to determine the arrival rate/sampling ratio required to achieve acceptable serial correlation. However, the measurements are now presented with a detection rate of $50 \leq \lambda_D \leq 15$ Mcps **and** a swept sampling frequency $30 \leq f_{BG} \leq 10$ MHz. The purpose of this is to observe whether, as postulated, the QRFFs are statistically independent. In Figure 6.8a, the exponential relationship previously seen between the count/sampling ratio and the corresponding serial correlation coefficient is present. For any given sampling rate, the required detection rate should be 2–3 times to result in $a_1 \leq 10^{-3}$. Therefore, it is shown that at $f_{BG} = 40$ MHz, and $\lambda_D \simeq 15$ Mcps, the 1-bit lag serial correlation coefficient is measured to be $a_1 \simeq 0.17$. At the same sampling frequency, when the detection rate is increased to $\lambda_D \simeq 50$ Mcps, the correlation is only reduced to a value of $a_1 \simeq 0.0051$. Figure 6.8b highlights the same measurement, with the macro-QRFF logic control set so that two standalone QRFFs are XOR'd together. A dramatic improvement is shown in the serial correlation. Under the same stress test conditions, where $\lambda_D \simeq 15$ and $f_{BG} = 40$ MHz, the resulting correlation coefficient is $a_1 \simeq 0.031$. Note, in this case, the detection rate plotted is that of QRFF1. The correlation bound is now met with a detection rate of $\lambda_D \simeq 35$ Mcps, where the result is $a_1 \simeq 5.97 \cdot 10^{-4}$. Finally, the measurement is performed again with all four sub-QRFF bits XOR'd on-chip. The result demonstrates superior performance to other random bit generator construction proposed in this thesis. With $f_{BG} = 40$ MHz, and $\lambda_D \simeq 15$ Mcps, the measured serial correlation now becomes $a_1 \simeq 1.31 \cdot 10^{-3}$, with a value of 10^{-4} reached as soon as the detection rate is increased to 20 Mcps. These results bolster the claim that this slow-clock QRFF is a versatile and high-performing bit generation primitive, which can enable quantum random number generating sensors that can have variable bit generation rates and are robust against pixel failure.



(a) Single QRFF



(b) Two QRFF with XOR.



(c) 4 QRFF with XOR.

Figure 6.8: Autocorrelation result for a single macro-QRFF in the FortunaSPAD2. A 3D visualization of the λ_D/f_{BG} relationship is shown with a variety of plotted ratios. The results demonstrate that two round of XOR reduces serial correlation to acceptable levels even at $\lambda_D/f_{BG} \simeq 0.5$.

6.4 FortunaSPAD2 with integrated micro-LEDs

A research platform, for determining the feasibility of integrating illumination on-chip, was designed and taped-out in the 55 nm BCD platform. Due to the indirect band-gap structure of silicon, high-brightness light sources in the visible range have yet to be demonstrated in a CMOS process. Recently, it was shown that high-brightness micro-LEDs in the NIR could be designed in the 55 nm BCD process [198]–[200]. Therefore, this presented an opportunity to investigate generation of quantum random bits with a complete monolithic approach. In [199], the maximum achieved emission intensity, P_E was published as 40 mW/cm^2 at a wavelength of $\lambda = 1$ μm. A rudimentary estimation of the available photon budget can be performed. Several assumptions are made to simplify the analysis. In [199], the active radius of a circular structure published, which is also implemented in this work, is 10 μm^2. The energy emitted per second, E_E, is defined as the emission intensity, multiplied by the active area of the LED. Therefore, the photons radiated per second can then be estimated by dividing E_E with the energy per photon, hc/λ as shown in Equation (6.1).

$$\begin{aligned} \frac{\text{Photons}}{s} &= \frac{P_E \cdot \lambda \cdot \text{A}_{\text{LED}}}{hc} \\ &= \frac{40 \text{ mW/cm}^2 \cdot 1 \ \mu m \cdot (10 \ \mu\text{m})^2 \pi}{hc} \simeq 6.31 \cdot 10^{11} \text{ photons/s} \end{aligned} \tag{6.1}$$

The speed of light is denoted by c, and h, is Planck's constant. Indeed, this analysis ignores required optical modelling. The directivity of the LED is undefined, i.e. assumed to be isotropic. However, a half-angle, θ, where the radiative power drops 3 dB should also be included in the analysis. Conversion of the emission intensity would then transform $\text{W/A}_{\text{LED}}{}^2$ to W/sr by multiplying the surface area of the active region of the micro-LED is A_{LED} and the solid angle, sr, in steradian, which is $2\pi \cos(\theta/2)$. Moreover, since the micro-LEDs and SPADs are placed on the same planar structure, most of the photons are radiated away from the SPAD. Finally, the performance of the SPAD in the NIR at 1 μm is highly insensitive; measurements of SPADs presented in this work, at an excess bias of $\text{V}_{\text{EX}} = 1$ V and $\lambda = 1$ μm is estimated at $\simeq 0.5$ %. Nevertheless, even if the value calculated here is reduced by 5 orders of magnitude, in principle, Mcps operation is still attainable. The purpose of this analysis is to demonstrate the viability of using these NIR micro-LEDs to achieve fully monolithic bit generation in the MHz-per-pixel range.

An electrically identical version of the FortunaSPAD2 is designed, with arrays of monolithically integrated CMOS micro-LEDs. A micrograph is shown in Figure 6.9. As little optical modelling was performed before the tape-out date, a variety of spatial patterns with respect to the LED and SPAD placement is included on chip. Adjacent to each column of macro-QRFFs, there is a column of micro-LEDs, with unit cells placed in parallel. The anodes (A) of the unit cells for each column are connected in parallel and attached directly to an output pad. To reduce the total number of pads the cathodes (K) of four adjacent columns are connected together. The

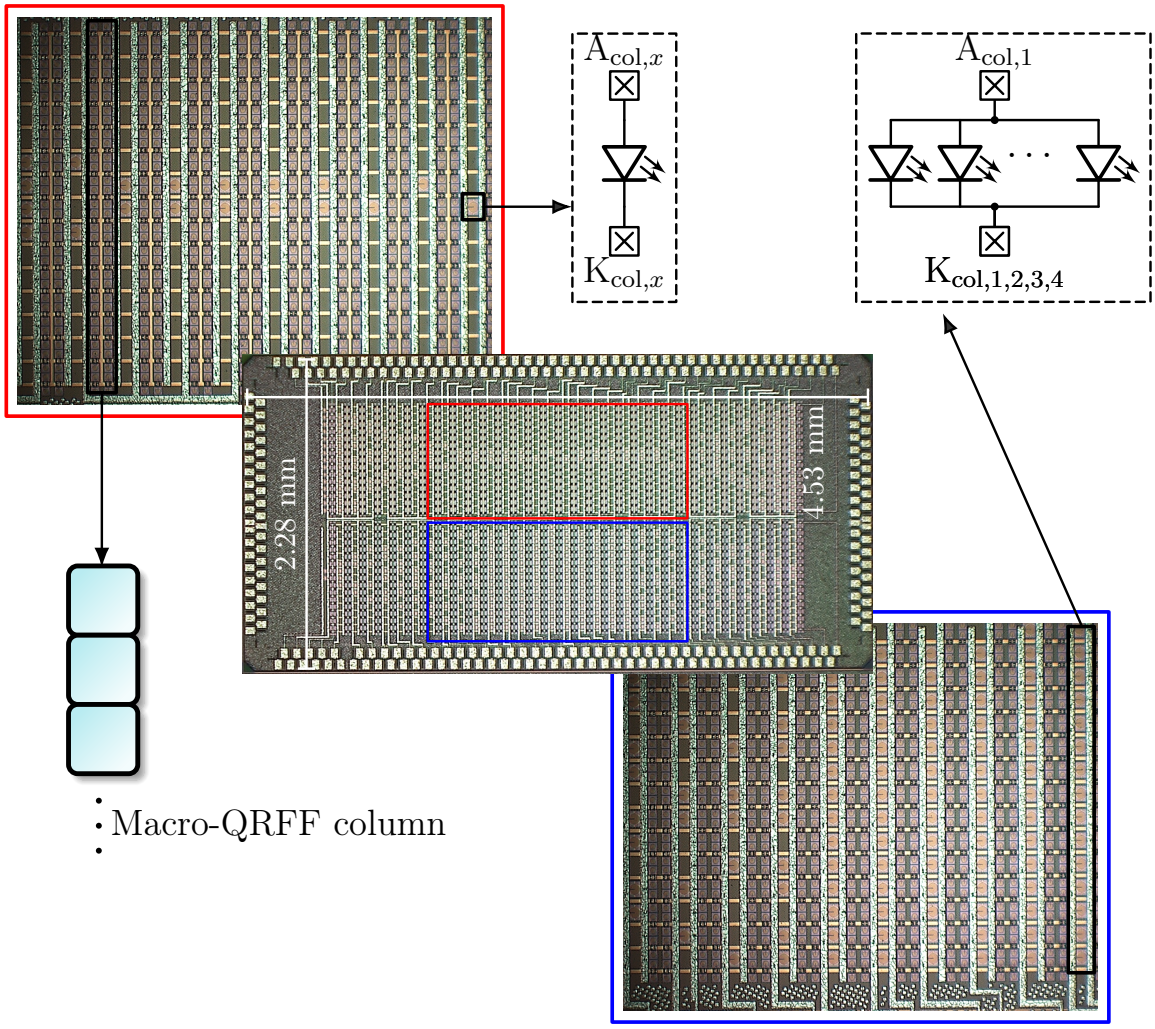


Figure 6.9: Micrograph of the FortunaSPAD2 version, which includes silicon μ-LEDs.

top half of the array has a single micro-LED placed at the center, with the number per column increasing moving towards the array. The bottom half take the opposite approach, where the maximum number (15) unit cells are placed in the center, with a decreasing per-column density moving towards the edge. All functions, such as bit generation, masking, and readout, are identical to the FortunaSPAD2 design. In principle, it would be advantageous to pattern metal overtop to LEDs and SPADs, in an attempt to reflect and radiation back towards the silicon surface. However, this is avoided for two reasons.

1. For proper characterization of the device, it would be beneficial to measure the emission intensity of all columns containing micro-LED unit cells, as a function of its IV characteristics.
2. After the desired/optimal mode of electrical operation is determined, metal patterning could be performed in the clean room. Moreover, enhanced light trapping structures such as meta-surfaces could be researched.

Although limited by the wavelength of radiative emission, the relatively large power consumption of the unit cells ($\simeq$ 40 mA at 2.5 V for 40 mW/cm^2), and the planar placement, this chip can function as a promising research platform for:

1. Experimenting with optical nanostructures that reflect radiation emitted by sources placed on the same planar surface as a SPAD.
2. Understanding the power/consumption per-bit generated trade-off when using NIR micro-LEDs as a entropy source for silicon SPADs.
3. Determining whether per pixel bit generation in the mega-bits-per-second order is possible using the slow clock QRFF and entropy sources that are monolithically integrated.

7 Conclusions

7.1 Summary of thesis outcomes

This thesis explored devices, circuits, and methods towards the full monolithic integration of scalable, high-speed quantum random number generators. The general block diagram of a QRNG from Chapter 1 is shown again in Figure 7.1, annotated with the various corresponding components presented within this work. F our specific SPAD detectors, designed in a 55 nm BCD process, were proposed. Simple methods, i.e. the addition of already existing implants were used to improve sensitivity of a deep and shallow junction SPAD. Moreover, all presented designs demonstrated excellent noise performance. The optimized deep junction designed was used for QRNG development, due to its superior sensitivity and afterpulsing performance.

The majority of the emphasis was the modelling and design of photon-timing statistic-based

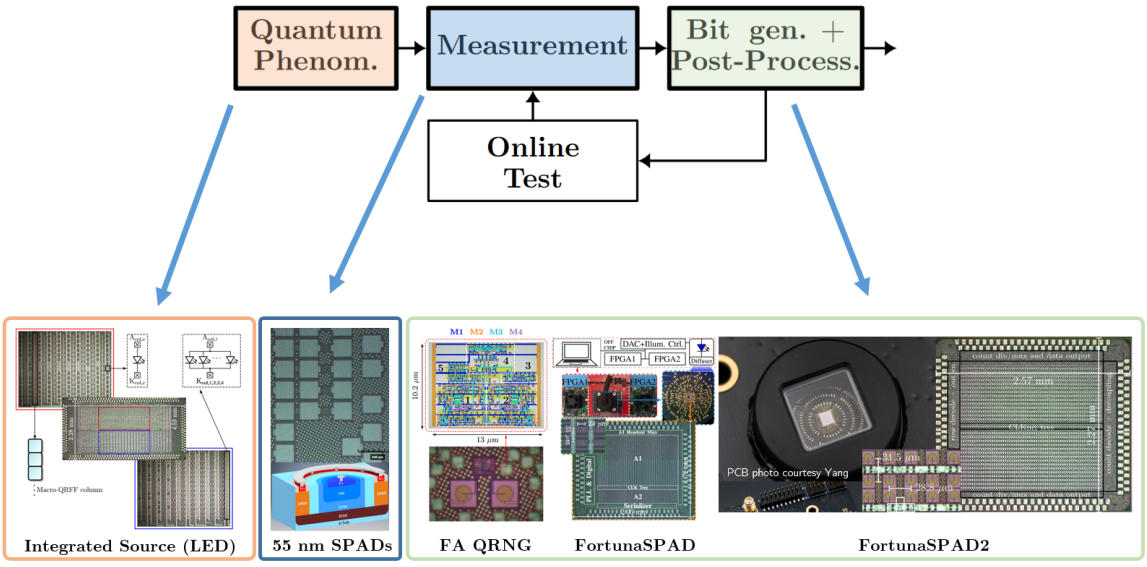


Figure 7.1: QRNG block diagram with integrated solutions proposed in this thesis. 55 nm SPADs, various QRFF circuits, FortunaSPAD and FortunaSPAD2 for random number generation, along with a research platform for studying integrated illumination.

random bit generating solutions. Two different realizations of the quantum random flip-flop (QRFF), which generates one random bit every clock-cycle, were studied, designed and measured. A significant effort was placed on further developing the slow-clock based random bit generating concept. Two major contributions were proposed in this respect. The first, was a dynamic comparator based sampling flip-flop, used to either essentially eliminate bit bias in a single pixel, or center the mean bias of pixels in an array to zero. The second, was a method for modelling the serial correlation of bits using a detector when considering dead time, and to a certain extent, afterpulsing. Both these contributions led to improvements in performance. The first-arrival photon bit generation method was studied, although to a lesser extent, in contrast to the slow clock method. The SC method, when combined with the comparator sampling, achieves improved bias compared to the FA method. However, the FA method has the advantage that it only requires one photon detection per bit. The modelling of the SC proposed in this thesis showed that the average number of photons per bit could be reduced by increasing dead time, although with a limit of $\sim \lambda_A/f_{BG} \simeq 2.5$.

Two QRNG sensor arrays based on the SC QRFF were designed and taped-out. FortunaSPAD contained two independent arrays on the same die, which employed different readout methodologies. The PLL did not lock at the desired high frequency $f \geq 400$ MHz, therefore the total throughput of the chip was limited. Nevertheless, Gbps operation was achieved, and the performance was validated using NIST SP 800-22 and SP-800-90B test suites. The FortunaSPAD2 implemented a macro-QRFF, which improved robustness by XOR-ing pixels together. Testing of the FortunaSPAD2 is ongoing with the majority of detailed characterization, including the achievable total throughput, to be determined. Moreover, a version of the FortunaSPAD2 that included silicon micro-LEDs, first introduced in [198], was designed as a platform for determining the viability of full monolithic operation. Characterization of this chip is commencing soon.

7.2 Future work

QRFF model improvements and overall comparisons

The SC QRFF demonstrated in this thesis has demonstrated good performance in terms of bias and correlation, and is modelled well by the proposed methods from this research, when combined with previous work [171]. However, measurements have shown that there is certainly room for improvement in terms of matching theory with practice. Most notably, some improvements are desirable for the bias calculation. Although, as expected, adjustment of the threshold voltage can reduce/eliminate bias, this happens at a normalized value quite far (up to $\sim 50\%$) from what would be predicted by the model, i.e. $\eta \simeq 0.75$. Therefore, there are likely other phenomena, such as metastability, mismatched clock-to-Q times, or threshold offset, which are affecting the results significantly and therefore should be understood. Moreover, the significant drawback of this method is its free-running operation (multiple photons per bit), which reduces its scalability. Perhaps there are sampling techniques, or gated mode operation, which could provide similar results in terms of randomness performance, while reducing the

number of avalanches.

QRNGs, particularly SPAD-based solutions, are, in a sense, still in their infancy. Therefore, as the technology advances, broad analyses, in terms of the outlook, and comparison between state-of-the-art classical TRNGs, are required. Most notably, Table 4.1, which provided a general comparison of practical methods, should be expanded to include missing models and figures-of-merit. For an objective analysis, common FoMs, such as bit/energy, should be determined. This leads to the next major research item, which is the development of full monolithic integration.

Full monolithic integration

QRNGs can potentially be a technology to replace the use of established classical TRNGs in certain or perhaps all applications. Quantum has the advantage of true verifiable randomness, and from a practical perspective, as demonstrated in this thesis, evaluation of entropy can be modelled quite well using equations that describe photon-timing statistics. Although, to surpass established technologies, it is likely that full monolithic integration, is required. As mentioned, some previous works [194], [201], have demonstrated this, although at low throughput levels. Using the platform designed in this thesis, a study on the ability to use silicon micro-LEDs could be performed. Moreover, if silicon proves to be too-great a hurdle, then perhaps 3D-stacking with other semiconductor flavors, as done for many image sensing technologies, provides a path for a fully packaged solution.

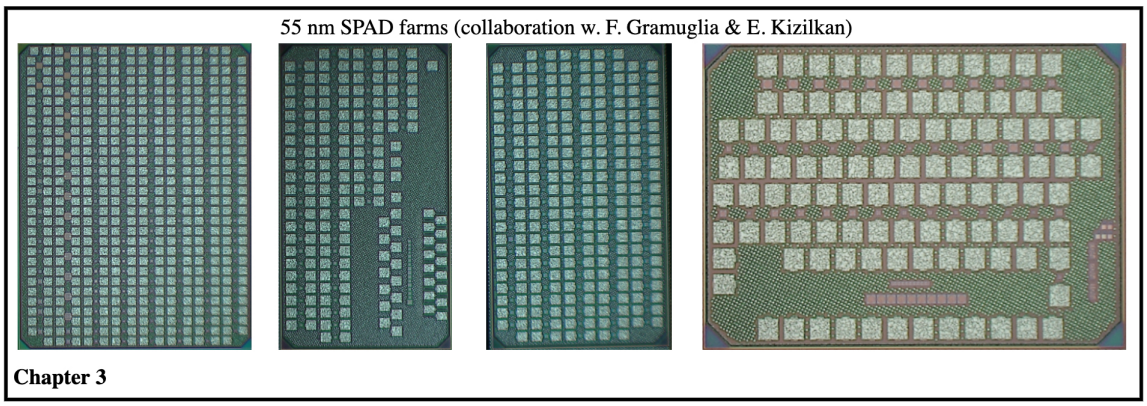
A more detailed look at robustness

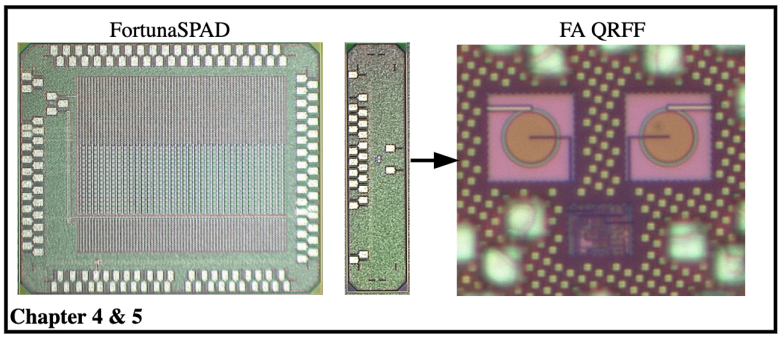
Side-channel attacks and detailed characterization of robustness in terms of PVT variations were out of scope for this work and yet, perhaps essential, for a final solution. A detailed analysis of the FortunaSPAD's vulnerabilities, and ways to protect/mitigate them, should be researched. The modelling of entropy degradation performed here could be adapted to comply as a stochastic model to determine a bound on min-entropy of raw bits, as required by AIS certification. Moreover, conditional entropy could be included in the analysis as well, to consider if system information gained by a potential adversary, such as photon flux, could degrade the performance of the generator when used in a key generation application [36], [202].

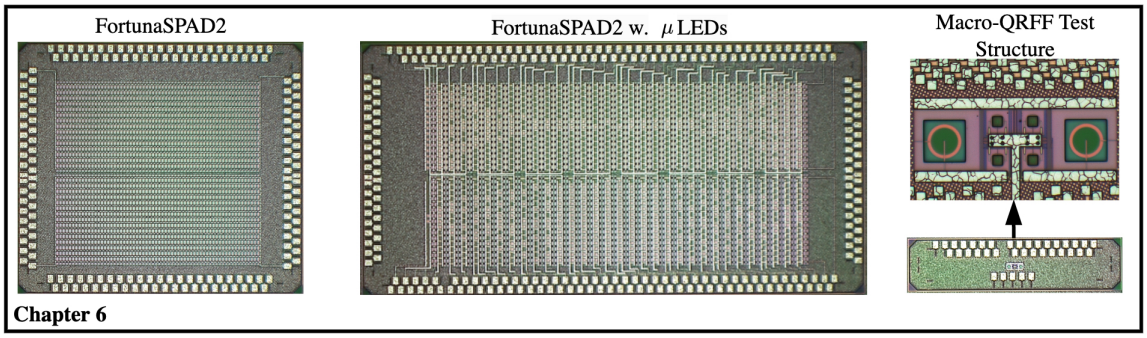
Real-time entropy estimation

One of the advantages of the simple SC method presented in this work is that bias, and to a certain extent, entropy, can be estimated by simply monitoring the count rate of pixels. However, a standalone QRNG implementation would likely require more comprehensive online testing/monitoring. Efficient estimation of entropy is an ongoing and active field of research, with the use of neural networks as a potential method being explored recently [189], [203]. Exploration into the hardware implementation of such algorithms, for QRNGs exploiting photon-timing statistics, could potentially be a fruitful field of research.

Chip Gallery







Publications

Peer-Reviewed Journal

1. F. Gramuglia[†], **P. Keshavarzian**[†], E. Kizilkan[†], C. Bruschini, S. S. Tan, M. Tng, E. Quek, M.-J. Lee, and E. Charbon, "Engineering breakdown probability profile for PDP and DCR optimization in a SPAD fabricated in a standard 55 nm BCD process", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 2, pp. 1–10, 2021.

2. **P. Keshavarzian**, K. Ramu, D. Tang, C. Weill, F. Gramuglia, S. S. Tan, M. Tng, L. Lim, E. Quek, D. Mandich, M. Stipčević, and E. Charbon, "A 3.3 Gbps SPAD-based quantum random number generator", *arXiv preprint arXiv:2209.04868, accepted for publication in IEEE JSSC DOI: 10.1109/JSSC.2023.3274692*,

3. **P. Keshavarzian** *et al.*, "A first-photon-arrival comparison SPAD-based quantum random bit generator in 55 nm", **In preparation**.

4. **P. Keshavarzian** *et al.*, "FortunaSPAD2: A SPAD-based QRNG with robust macro-QRFF pixels.", **In preparation**.

5. J. Zhao, T. Milanese, F. Gramuglia, **P. Keshavarzian**, S. S. Tan, M. Tng, L. Lim, V. Dhulla, E. Quek, M.-J. Lee, and E. Charbon, "On analog silicon photomultipliers in standard 55-nm BCD technology for LiDAR applications", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 5, pp. 1–10, 2022.

6. M.-L. Wu, E. Ripiccini, E. Kizilkan, F. Gramuglia, **P. Keshavarzian**, C. A. Fenoglio, K. Morimoto, and E. Charbon, "Radiation hardness study of single-photon avalanche diode for space and high energy physics applications", *Sensors*, vol. 22, no. 8, p. 2919, 2022.

7. J. Zhao, F. Gramuglia, **P. Keshavarzian**, S. S. Tan, M. Tng, L. Lim, V. Dhulla, E. Quek, M.-J. Lee, and E. Charbon, "A gradient-gated SPAD array for non-line-of-sight imaging", *Submitted-IEEE Journal of Selected Topics in Quantum Electronics*,

[†] Equal contributions

Conference Proceedings

1. **P. Keshavarzian**, F. Gramuglia, E. Kizilkan, C. Bruschini, S. S. Tan, M. Tng, D. Chong, E. Quek, M.-J. Lee, and E. Charbon, "Low-noise high-dynamic-range single-photon avalanche diodes with integrated PQAR circuit in a standard 55nm BCD process", in *Advanced Photon Counting Techniques XVI*, vol. 12089, 2022, pp. 73–82.

2. A. Morelle, F. Gramuglia, **P. Keshavarzian**, C. Bruschini, D. Chong, J. Tan, M. Tng, E. Quek, and E. Charbon, "Deep cryogenic operation of 55 nm CMOS SPADs for quantum information and metrology applications", in *Quantum Information and Measurement*, Optica Publishing Group, 2021, M2B–7.

Conference Talks and Poster Presentations

1. **P. Keshavarzian** and E. Charbon, "SPAD-based high-speed quantum random number generation in DSM CMOS", International SPAD Sensor Workshop (ISSW), 2020.

2. F. Gramuglia, **P. Keshavarzian**, E. Kizilkan, M.-L. Wu, C. Bruschini, S. Tan, M. Tng, E. Quek, M.-J. Lee, and E. Charbon, "A 7.5 %. 60% PDP low-noise SPAD fabircated in CMOS technology", International SPAD Sensor Workshop (ISSW), 2022.

Bibliography

[1] C. E. Shannon, "A mathematical theory of communication", *The Bell system technical journal*, vol. 27, no. 3, pp. 379–423, 1948.

[2] A. Rényi *et al.*, "On measures of entropy and information", in *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, Berkeley, California, USA, vol. 1, 1961.

[3] J. Walker, "ENT: a pseudorandom number sequence test program", *Software and documentation available at/www. fourmilab. ch/random/S*, 2008.

[4] M. Alioto, "Trends in hardware security: from basics to ASICs", *IEEE Solid-State Circuits Mag.*, vol. 11, no. 3, pp. 56–74, 2019.

[5] E. Barker, A. Roginsky, *et al.*, "Transitions: recommendation for transitioning the use of cryptographic algorithms and key lengths", *NIST Special Publication*, vol. 800, 131A, 2011.

[6] J. P. Degabriele, K. G. Paterson, J. C. Schuldt, and J. Woodage, "Backdoors in pseudorandom number generators: possibility and impossibility results", in *Advances in Cryptology–CRYPTO 2016: 36th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2016, Proceedings, Part I 36*, 2016, pp. 403–432.

[7] T. C. Hales, "The NSA back door to NIST", *Notices of the AMS*, vol. 61, no. 2, pp. 190–192, 2013.

[8] Y. Tsunoo, T. Saito, H. Kubo, T. Suzaki, and H. Nakashima, "Differential cryptanalysis of Salsa20/8", in *Workshop Record of SASC*, vol. 28, 2007.

[9] A. B. Orúe, L. Hernández Encinas, V. Fernández, and F. Montoya, "A review of cryptographically secure prngs in constrained devices for the iot", in *International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, September 6–8, 2017, Proceeding 12*, Springer, 2018, pp. 672–682.

[10] L. Gong, J. Zhang, H. Liu, L. Sang, and Y. Wang, "True random number generators using electrical noise", *IEEE Access*, vol. 7, pp. 125 796–125 805, 2019.

[11] M. Dichtl and J. D. Golic, "High-speed true random number generation with logic gates only", in *CHES*, Springer, vol. 7, 2007, pp. 45–62.

[12] S. K. Satpathy, S. K. Mathew, R. Kumar, V. Suresh, M. A. Anders, H. Kaul, A. Agarwal, S. Hsu, R. K. Krishnamurthy, and V. De, "An all-digital unified physically unclonable function and true random number generator featuring self-calibrating hierarchical Von Neumann extraction in 14-nm tri-gate CMOS", *IEEE J. Solid-State Circuits*, vol. 54, no. 4, pp. 1074–1085, Apr. 2019.

[13] S. K. Mathew, S. Srinivasan, M. A. Anders, H. Kaul, S. K. Hsu, F. Sheikh, A. Agarwal, S. Satpathy, and R. K. Krishnamurthy, "2.4 Gbps, 7 mW all-digital PVT-variation tolerant true random number generator for 45 nm CMOS high-performance microprocessors", *IEEE Journal of Solid-State Circuits*, vol. 47, no. 11, pp. 2807–2821, 2012.

[14] V. R. Pamula, X. Sun, S. Kim, F. ur Rahman, B. Zhang, and V. S. Sathe, "An all-digital true-random-number generator with integrated de-correlation and bias correction at 3.2-to-86 Mb/s, 2.58 pJ/bit in 65-nm CMOS", in *2018 IEEE Symposium on VLSI Circuits*, IEEE, 2018, pp. 1–2.

[15] S. Taneja and M. Alioto, "Fully synthesizable unified true random number generator and cryptographic core", *IEEE Journal of Solid-State Circuits*, vol. 56, no. 10, pp. 3049–3061, 2021.

[16] A. Olgun, M. Patel, A. G. Yağlıkçı, H. Luo, J. S. Kim, F. N. Bostancı, N. Vijaykumar, O. Ergin, and O. Mutlu, "QUAC-TRNG: high-throughput true random number generation using quadruple row activation in commodity DRAM chips", in *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*, IEEE, 2021, pp. 944–957.

[17] R. Zhang, X. Wang, K. Liu, and H. Shinohara, "A 0.186-pJ per bit latch-based true random number generator featuring mismatch compensation and random noise enhancement", *IEEE Journal of Solid-State Circuits*, 2022.

[18] X. Wang, R. Zhang, Y. Wang, K. Liu, X. Wang, and H. Shinohara, "A 0.116 pJ/bit latch-based true random number generator with static inverter selection and noise enhancement", in *2022 International Symposium on VLSI Design, Automation and Test (VLSI-DAT)*, IEEE, 2022, pp. 1–4.

[19] M. Alioto, "Aggressive design reuse for ubiquitous zero-trust edge security-from physical design to machine learning-based hardware patching", *IEEE Open Journal of the Solid-State Circuits Society*, 2022.

[20] N. Da Dalt and A. Sheikholeslami, *Understanding Jitter and Phase Noise: A Circuits and Systems Perspective*. Cambridge University Press, 2018.

[21] B. Sunar, W. J. Martin, and D. R. Stinson, "A provably secure true random number generator with built-in tolerance to active attacks", *IEEE Transactions on computers*, vol. 56, no. 1, pp. 109–119, 2006.

[22] T. Stojanovski and L. Kocarev, "Chaos-based random number generators-Part I: analysis [cryptography]", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 3, pp. 281–288, 2001.

[23] T. Stojanovski, J. Pihl, and L. Kocarev, "Chaos-based random number generators. Part II: practical realization", *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 48, no. 3, pp. 382–385, 2001.

[24] T. Addabbo, M. Alioto, A. Fort, S. Rocchi, and V. Vignoli, "A feedback strategy to improve the entropy of a chaos-based random bit generator", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 53, no. 2, pp. 326–337, 2006.

[25] M. Kim, U. Ha, K. J. Lee, Y. Lee, and H.-J. Yoo, "A 82-nW chaotic map true random number generator based on a sub-ranging SAR ADC", *IEEE Journal of Solid-State Circuits*, vol. 52, no. 7, pp. 1953–1965, 2017.

[26] S. N. Dhanuskodi, A. Vijayakumar, and S. Kundu, "A chaotic ring oscillator based random number generator", in *2014 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, IEEE, 2014, pp. 160–165.

[27] M. Herrero-Collantes and J. C. Garcia-Escartin, "Quantum random number generators", *Reviews of Modern Physics*, vol. 89, no. 1, p. 015 004, 2017.

[28] A. Saini, A. Tsokanos, and R. Kirner, "Quantum randomness in cryptography—a survey of cryptosystems, RNG-based ciphers, and QRNGs", *Information*, vol. 13, no. 8, p. 358, 2022.

[29] J. Aldama, S. Sarmiento, I. H. L. Grande, S. Signorini, L. T. Vidarte, and V. Pruneri, "Integrated QKD and QRNG photonic technologies", *Journal of Lightwave Technology*, 2022.

[30] B. Sanguinetti, A. Martin, H. Zbinden, and N. Gisin, "Quantum random number generation on a mobile phone", *Physical Review X*, vol. 4, no. 3, p. 031 056, 2014.

[31] Y. Liu, Q. Zhao, M.-H. Li, J.-Y. Guan, Y. Zhang, B. Bai, W. Zhang, W.-Z. Liu, C. Wu, X. Yuan, *et al.*, "Device-independent quantum random-number generation", *Nature*, vol. 562, no. 7728, pp. 548–551, 2018.

[32] R. Colbeck and A. Kent, "Private randomness expansion with untrusted devices", *Journal of Physics A: Mathematical and Theoretical*, vol. 44, no. 9, p. 095 305, 2011.

[33] V. Mannalath, S. Mishra, and A. Pathak, "A comprehensive review of quantum random number generators: concepts, classification and the origin of randomness", *arXiv preprint arXiv:2203.00261*, 2022.

[34] Z. Cao, H. Zhou, X. Yuan, and X. Ma, "Source-independent quantum random number generation", *Physical Review X*, vol. 6, no. 1, p. 011 020, 2016.

[35] Z. Cao, H. Zhou, and X. Ma, "Loss-tolerant measurement-device-independent quantum random number generation", *New Journal of Physics*, vol. 17, no. 12, p. 125 011, 2015.

[36] D. Drahi, N. Walk, M. J. Hoban, A. K. Fedorov, R. Shakhovoy, A. Feimov, Y. Kurochkin, W. S. Kolthammer, J. Nunn, J. Barrett, *et al.*, "Certified quantum random numbers from untrusted light", *Physical Review X*, vol. 10, no. 4, p. 041 048, 2020.

[37] P. R. Smith, D. G. Marangon, M. Lucamarini, Z. Yuan, and A. Shields, "Simple source device-independent continuous-variable quantum random number generator", *Physical Review A*, vol. 99, no. 6, p. 062 326, 2019.

[38] M. Stipčević and D. J. Gauthier, "Precise monte carlo simulation of single-photon detectors", in *Advanced Photon Counting Techniques VII*, SPIE, vol. 8727, 2013, pp. 68–75.

[39] J. G. Rarity, P. Owens, and P. Tapster, "Quantum random-number generation and key sharing", *Journal of Modern Optics*, vol. 41, no. 12, pp. 2435–2444, 1994.

[40] M. Stipčević, "Quantum random flip-flop and its applications in random frequency synthesis and true random number generation", *Review of Scientific Instruments*, vol. 87, no. 3, p. 035 113, 2016.

[41] J. L. McInnes and B. Pinkas, "On the impossibility of private key cryptography with weakly random keys", in *Conference on the Theory and Application of Cryptography*, Springer, 1990, pp. 421–435.

[42] S. K. Mathew, D. Johnston, S. Satpathy, V. Suresh, P. Newman, M. A. Anders, H. Kaul, A. Agarwal, S. K. Hsu, G. Chen, *et al.*, "μRNG: a 300–950 mV, 323 Gbps/W all-digital full-entropy true random number generator in 14 nm FinFET CMOS", *IEEE Journal of Solid-State Circuits*, vol. 51, no. 7, pp. 1695–1704, 2016.

[43] Y. Peres, "Iterating Von Neumann's procedure for extracting random bits", *The Annals of Statistics*, pp. 590–597, 1992.

[44] Y. Dodis, A. Elbaz, R. Oliveira, and R. Raz, "Improved randomness extraction from two independent sources", in *Approximation, randomization, and combinatorial optimization. Algorithms and techniques*, Springer, 2004, pp. 334–344.

[45] B. Barak, R. Impagliazzo, and A. Wigderson, "Extracting randomness using few independent sources", *SIAM Journal on Computing*, vol. 36, no. 4, pp. 1095–1118, 2006.

[46] L. Trevisan, "Construction of extractors using pseudo-random generators", in *Proceedings of the thirty-first annual ACM symposium on Theory of computing*, 1999, pp. 141–148.

[47] V. Mavroeidis, K. Vishi, M. D. Zych, and A. Jøsang, "The impact of quantum computing on present cryptography", *arXiv preprint arXiv:1804.00200*, 2018.

[48] L. Chen, L. Chen, S. Jordan, Y.-K. Liu, D. Moody, R. Peralta, R. Perlner, and D. Smith-Tone, *Report on post-quantum cryptography*. US Department of Commerce, National Institute of Standards and Technology . . ., 2016, vol. 12.

[49] P. W. Shor, "Algorithms for quantum computation: discrete logarithms and factoring", in *Proceedings 35th annual symposium on foundations of computer science*, Ieee, 1994, pp. 124–134.

[50] L. K. Grover, "A fast quantum mechanical algorithm for database search", in *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, 1996, pp. 212–219.

[51] X. Bonnetain, M. Naya-Plasencia, and A. Schrottenloher, "Quantum security analysis of AES", *IACR Transactions on Symmetric Cryptology*, vol. 2019, no. 2, pp. 55–93, 2019.

[52] G. Alagic, G. Alagic, J. Alperin-Sheriff, D. Apon, D. Cooper, Q. Dang, Y.-K. Liu, C. Miller, D. Moody, R. Peralta, *et al.*, *Status report on the first round of the NIST post-quantum cryptography standardization process*. US Department of Commerce, National Institute of Standards and Technology . . ., 2019.

[53] C. E. Shannon, "Communication theory of secrecy systems", *The Bell system technical journal*, vol. 28, no. 4, pp. 656–715, 1949.

[54] A. Vassilev and R. Staples, "Entropy as a service: unlocking cryptography's full potential", *Computer*, vol. 49, no. 9, pp. 98–102, 2016.

[55] C. Cheng, R. Lu, A. Petzoldt, and T. Takagi, "Securing the internet of things in a quantum world", *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 116–120, 2017.

[56] M. S. Turan, E. Barker, J. Kelsey, K. A. McKay, M. L. Baish, M. Boyle, *et al.*, "Recommendation for the entropy sources used for random bit generation", *NIST Special Publication*, no. 800 90B 3rd Draft, 2018.

[57] D. Hurley-Smith and J. Hernandez-Castro, "Quantum leap and crash: searching and finding bias in quantum random number generators", *ACM Trans. Privacy. & Security*, vol. 23, no. 3, 2020.

[58] K. Marton and A. Suciu, "On the interpretation of results from the NIST statistical test suite", *Science and Technology*, vol. 18, no. 1, pp. 18–32, 2015.

[59] M.-J. O. Saarinen, "NIST SP 800-22 and GM/T 0005-2012 tests: clearly obsolete, possibly harmful.", *IACR Cryptol. ePrint Arch.*, vol. 2022, p. 169, 2022.

[60] K. A. Kowalska, D. Fogliano, and J. G. Coello, "On the revision of NIST 800-22 test suites", *Cryptology ePrint Archive*, 2022.

[61] A. L. Rukhin, "Statistical testing of randomness: new and old procedures", in *Randomness through Computation*, H. Zenil, Ed., 1st ed., Singapore: World Scientific, 2011, pp. 160–174.

[62] E. Barker, J. Kelsey, K. McKay, A. Roginsky, and M. S. Turan, "Recommendation for random bit generator (RBG) constructions", *NIST Special Publication*, no. 800 90C 3rd Draft, 2022.

[63] W. Killmann and W. Schindler, "A proposal for: functionality classes for random number generators", *Tech. Rep., Bundesamt für Sicherheit in der Informationstechnik (BSI)*, vol. 2.35 DRAFT, 2022.

[64] F. Zappa, A. Tosi, A. Dalla Mora, and S. Tisa, "SPICE modeling of single photon avalanche diodes", *Sensors and Actuators A: Physical*, vol. 153, no. 2, pp. 197–204, 2009.

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

[66] K. Morimoto and E. Charbon, "A scaling law for SPAD pixel miniaturization", *Sensors*, vol. 21, no. 10, p. 3447, 2021.

[67] M. Perenzoni, L. Pancheri, and D. Stoppa, "Compact SPAD-based pixel architectures for time-resolved image sensors", *Sensors*, vol. 16, no. 5, p. 745, 2016.

[68] L. Pancheri, N. Massari, and D. Stoppa, "SPAD image sensor with analog counting pixel for time-resolved fluorescence detection", *IEEE Transactions on Electron Devices*, vol. 60, no. 10, pp. 3442–3449, 2013.

[69] S. Lindner, S. Pellegrini, Y. Henrion, B. Rae, M. Wolf, and E. Charbon, "A high-PDE, backside-illuminated SPAD in 65/40-nm 3D IC CMOS pixel with cascoded passive quenching and active recharge", *IEEE Electron Device Letters*, vol. 38, no. 11, pp. 1547–1550, 2017.

[70] A. Gallivanoni, I. Rech, and M. Ghioni, "Progress in quenching circuits for single photon avalanche diodes", *IEEE Transactions on nuclear science*, vol. 57, no. 6, pp. 3815–3826, 2010.

[71] F. Villa, B. Markovic, S. Bellisai, D. Bronzi, A. Tosi, F. Zappa, S. Tisa, D. Durini, S. Weyers, U. Paschen, *et al.*, "SPAD smart pixel for time-of-flight and time-correlated single-photon counting measurements", *IEEE Photonics Journal*, vol. 4, no. 3, pp. 795–804, 2012.

[72] K. Buckbee, N. A. Dutton, and R. K. Henderson, "An indirect time-of-flight SPAD pixel with dynamic comparator re-use for a single-slope ADC", *IEEE Solid-State Circuits Letters*, 2022.

[73] L. Parmesan, N. A. Dutton, N. J. Calder, A. J. Holmes, L. A. Grant, and R. K. Henderson, "A 9.8 μm sample and hold time to amplitude converter CMOS SPAD pixel", in *2014 44th European Solid State Device Research Conference (ESSDERC)*, IEEE, 2014, pp. 290–293.

[74] F. Severini, I. Cusini, D. Berretta, K. Pasquinelli, A. Incoronato, and F. Villa, "SPAD pixel with sub-NS dead-time for high-count rate applications", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 28, no. 2, pp. 1–8, 2021.

[75] F. Gramuglia, "High-performance CMOS SPAD-based sensors for time-of-flight PET applications", EPFL, Tech. Rep., 2022.

[76] E. Sarbazi, M. Safari, and H. Haas, "Statistical modeling of single-photon avalanche diode receivers for optical wireless communications", *IEEE Transactions on Communications*, vol. 66, no. 9, pp. 4043–4058, 2018.

[77] D. Bronzi, F. Villa, S. Tisa, A. Tosi, and F. Zappa, "SPAD figures of merit for photon-counting, photon-timing, and imaging applications: a review", *IEEE Sensors J.*, vol. 16, pp. 3–12, Jan. 2016.

[78] I. Cusini, D. Berretta, E. Conca, A. Incoronato, F. Madonini, A. A. Maurina, C. Nonne, S. Riccardo, and F. Villa, "Historical perspectives, state of art and research trends of single photon avalanche diodes and their applications (part 1: single pixels)", *Frontiers in Physics*, vol. 10, p. 906 675, 2022.

[79] F. Ceccarelli, G. Acconcia, A. Gulinatti, M. Ghioni, I. Rech, and R. Osellame, "Recent advances and future perspectives of single-photon avalanche diodes for quantum photonics applications", *Advanced Quantum Technologies*, vol. 4, no. 2, p. 2 000 102, 2021.

[80] D. P. Palubiak and M. J. Deen, "CMOS SPADs: design issues and research challenges for detectors, circuits, and arrays", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 409–426, 2014.

[81] A. Gulinatti, I. Rech, M. Assanelli, M. Ghioni, and S. Cova, "A physically based model for evaluating the photon detection efficiency and the temporal response of SPAD detectors", *Journal of Modern Optics*, vol. 58, no. 3-4, pp. 210–224, 2011.

[82] M. Mazzillo, A. Piazza, G. Condorelli, D. Sanfilippo, G. Fallica, S. Billotta, M. Belluso, G. Bonanno, L. Cosentino, A. Pappalardo, *et al.*, "Quantum detection efficiency in Geiger mode avalanche photodiodes", *IEEE Transactions on Nuclear Science*, vol. 55, no. 6, pp. 3620–3625, 2008.

[83] H. Mahmoudi, S. S. K. Poushi, B. Steindl, M. Hofbauer, and H. Zimmermann, "Optical and electrical characterization and modeling of photon detection probability in CMOS single-photon avalanche diodes", *IEEE Sensors Journal*, vol. 21, no. 6, pp. 7572–7580, 2021.

[84] W. G. Oldham, R. R. Samuelson, and P. Antognetti, "Triggering phenomena in avalanche diodes", *IEEE Trans. Electron Devices*, vol. 19, no. 9, pp. 1056–1060, Sep. 1972.

[85] R. J. McIntyre, "On the avalanche initiation probability of avalanche diodes above the breakdown voltage", *IEEE Transactions on Electron Devices*, vol. 20, no. 7, pp. 637–641, 1973.

[86] L. Pancheri, D. Stoppa, and G.-F. Dalla Betta, "Characterization and modeling of breakdown probability in sub-micrometer CMOS SPADs", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 20, no. 6, pp. 328–335, 2014.

[87] R. J. McIntyre, "A new look at impact ionization-Part I: a theory of gain, noise, breakdown probability, and frequency response", *IEEE Trans. Electron Devices*, vol. 46, no. 8, pp. 1623–1631, Aug. 1999.

[88] H. Dautet, P. Deschamps, B. Dion, A. D. MacGregor, D. MacSween, R. J. McIntyre, C. Trottier, and P. P. Webb, "Photon counting techniques with silicon avalanche photodiodes", *Applied optics*, vol. 32, no. 21, pp. 3894–3900, 1993.

[89] W. Shockley and W. Read Jr, "Statistics of the recombinations of holes and electrons", *Physical review*, vol. 87, no. 5, p. 835, 1952.

[90] G. Giustolisi, R. Mita, and G. Palumbo, "Behavioral modeling of statistical phenomena of single-photon avalanche diodes", *International Journal of circuit theory and applications*, vol. 40, no. 7, pp. 661–679, 2012.

[91] A. Schenk, "An improved approach to the Shockley–Read–Hall recombination in inhomogeneous fields of space-charge regions", *Journ. Appl. Phys.*, vol. 71, no. 7, pp. 3339–3349, 1992.

[92] M. W. Fishburn, *Fundamentals of CMOS single-photon avalanche diodes*. 2012.

[93] J.-P. Colinge and C. A. Colinge, *Physics of semiconductor devices*. Springer Science & Business Media, 2005.

[94] A. Giudice, M. Ghioni, S. Cova, and F. Zappa, "A process and deep level evaluation tool: afterpulsing in avalanche junctions", in *ESSDERC'03. 33rd Conference on European Solid-State Device Research, 2003.*, IEEE, 2003, pp. 347–350.

[95] D. P. Palubiak, Z. Li, and M. J. Deen, "Afterpulsing characteristics of free-running and time-gated single-photon avalanche diodes in 130-nm CMOS", *IEEE Transactions on Electron Devices*, vol. 62, no. 11, pp. 3727–3733, 2015.

[96] D. Horoshko, V. Chizhevsky, and S. Y. Kilin, "Afterpulsing model based on the quasi-continuous distribution of deep levels in single-photon avalanche diodes", *Journal of Modern Optics*, vol. 64, no. 2, pp. 191–195, 2017.

[97] M. Anti, A. Tosi, F. Acerbi, and F. Zappa, "Modeling of afterpulsing in single-photon avalanche diodes", in *Physics and Simulation of Optoelectronic Devices XIX*, SPIE, vol. 7933, 2011, pp. 371–378.

[98] A. W. Ziarkash, S. K. Joshi, M. Stipčević, and R. Ursin, "Comparative study of afterpulsing behavior and models in single photon counting avalanche photo diode detectors", *Scientific Reports*, vol. 8, no. 1, p. 5076, Dec. 2018.

[99] G. Humer, M. Peev, C. Schaeff, S. Ramelow, M. Stipčević, and R. Ursin, "A simple and robust method for estimating afterpulsing in single photon detectors", *Journal of Lightwave Technology*, vol. 33, no. 14, pp. 3098–3107, 2015.

[100] S. Cova, A. Lacaita, and G. Ripamonti, "Trapping phenomena in avalanche photodiodes on nanosecond scale", *IEEE Electron device letters*, vol. 12, no. 12, pp. 685–687, 1991.

[101] P. Keshavarzian, F. Gramuglia, E. Kizilkan, C. Bruschini, S. S. Tan, M. Tng, D. Chong, E. Quek, M.-J. Lee, and E. Charbon, "Low-noise high-dynamic-range single-photon avalanche diodes with integrated PQAR circuit in a standard 55nm BCD process", in *Adv. Photon Counting Techn. XVI*, vol. 12089, SPIE, May 2022, pp. 73–82.

[102] M. Sanzaro, P. Gattari, F. Villa, A. Tosi, G. Croce, and F. Zappa, "Single-photon avalanche diodes in a 0.16 μm BCD technology with sharp timing response and red-enhanced sensitivity", *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 2, pp. 1–9, Mar. 2018.

[103] H. Finkelstein, M. J. Hsu, and S. C. Esener, "STI-bounded single-photon avalanche diode in a deep-submicrometer CMOS technology", *IEEE Electron Device Letters*, vol. 27, no. 11, pp. 887–889, 2006.

[104] I. M. Antolovic, C. Bruschini, and E. Charbon, "Dynamic range extension for photon counting arrays", *Optics Express*, vol. 26, no. 17, pp. 22 234–22 248, 2018.

[105] K. Morimoto, A. Ardelean, M.-L. Wu, A. C. Ulku, I. M. Antolovic, C. Bruschini, and E. Charbon, "Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications", *Optica*, vol. 7, no. 4, pp. 346–354, 2020.

[106] K. Morimoto, "Megapixel SPAD cameras for time-resolved applications", EPFL, Tech. Rep., 2021.

[107] M.-J. Lee, A. R. Ximenes, P. Padmanabhan, T.-J. Wang, K.-C. Huang, Y. Yamashita, D.-N. Yaung, and E. Charbon, "High-performance back-illuminated three-dimensional stacked single-photon avalanche diode implemented in 45-nm CMOS technology", *IEEE Journal of selected topics in quantum electronics*, vol. 24, no. 6, pp. 1–9, 2018.

[108] M.-J. Lee, P. Sun, G. Pandraud, C. Bruschini, and E. Charbon, "First near-ultraviolet-and blue-enhanced backside-illuminated single-photon avalanche diode

based on standard SOI CMOS technology", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 25, no. 5, pp. 1–6, 2019.

[109] T. Al Abbas, N. Dutton, O. Almer, S. Pellegrini, Y. Henrion, and R. Henderson, "Backside illuminated SPAD image sensor with 7.83 μm pitch in 3D-stacked CMOS technology", in *2016 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2016, pp. 8–1.

[110] E. Van Sieleghem, G. Karve, K. De Munck, A. Vinci, C. Cavaco, A. Süss, C. Van Hoof, and J. Lee, "A backside-illuminated charge-focusing silicon SPAD with enhanced near-infrared sensitivity", *IEEE Transactions on Electron Devices*, vol. 69, no. 3, pp. 1129–1136, 2022.

[111] Y. Liu, M. Liu, R. Ma, J. Hu, D. Li, X. Wang, and Z. Zhu, "A wide spectral response single photon avalanche diode for backside-illumination in 55-nm CMOS process", *IEEE Transactions on Electron Devices*, vol. 69, no. 9, pp. 5041–5047, 2022.

[112] W. Sun, Y. Wang, M. Liu, and Y. Yang, "A back-illuminated 4μm p+ n-well single photon avalanche diode pixel array with 0.36 hz/μm^2 dark count rate at 2.5 V excess bias voltage", *IEEE Electron Device Letters*, vol. 43, no. 9, pp. 1519–1522, 2022.

[113] A. R. Ximenes, P. Padmanabhan, M.-J. Lee, Y. Yamashita, D.-N. Yaung, and E. Charbon, "A 256× 256 45/65nm 3D-stacked SPAD-based direct TOF image sensor for LiDAR applications with optical polar modulation for up to 18.6 dB interference suppression", in *2018 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2018, pp. 96–98.

[114] T. Al Abbas, O. Almer, S. W. Hutchings, A. T. Erdogan, I. Gyongy, N. A. Dutton, and R. K. Henderson, "A 128× 120 5-wire 1.96 mm 2 40nm/90nm 3D stacked SPAD time resolved image sensor SoC for microendoscopy", in *2019 Symposium on VLSI Circuits*, IEEE, 2019, pp. C260–C261.

[115] P. Padmanabhan, C. Zhang, M. Cazzaniga, B. Efe, A. R. Ximenes, M.-J. Lee, and E. Charbon, "7.4 a 256× 128 3D-stacked (45nm) SPAD FLASH LiDAR with 7-level coincidence detection and progressive gating for 100m range and 10klux background light", in *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, vol. 64, 2021, pp. 111–113.

[116] R. K. Henderson, N. Johnston, S. W. Hutchings, I. Gyongy, T. Al Abbas, N. Dutton, M. Tyler, S. Chan, and J. Leach, "5.7 a 256× 256 40nm/90nm CMOS 3D-stacked 120dB dynamic-range reconfigurable time-resolved SPAD imager", in *2019 IEEE International Solid-State Circuits Conference-(ISSCC)*, IEEE, 2019, pp. 106–108.

[117] I. Gyongy, N. Calder, A. Davies, N. A. Dutton, R. R. Duncan, C. Rickman, P. Dalgarno, and R. K. Henderson, "A 256 × 256, 100-kfps, 61% fill-factor SPAD image sensor for time-resolved microscopy applications", *IEEE Transactions on Electron Devices*, vol. 65, no. 2, pp. 547–554, 2017.

[118] Y. Ota, K. Morimoto, T. Sasago, M. Shinohara, Y. Kuroda, W. Endo, Y. Maehashi, S. Maekawa, H. Tsuchiya, A. Abdelahafar, *et al.*, "A 0.37 W 143dB-dynamic-range 1Mpixel backside-illuminated charge-focusing SPAD image sensor with pixel-wise exposure control and adaptive clocked recharging", in *2022 IEEE International Solid-State Circuits Conference (ISSCC)*, IEEE, vol. 65, 2022, pp. 94–96.

[119] S.-G. Wuu, H.-L. Chen, H.-C. Chien, P. Enquist, R. M. Guidash, and J. McCarten, "A review of 3-dimensional wafer level stacked backside illuminated CMOS image sensor process technologies", *IEEE Transactions on Electron Devices*, 2022.

[120] I. M. Antolovic, A. C. Ulku, E. Kizilkan, S. Lindner, F. Zanella, R. Ferrini, M. Schnieper, E. Charbon, and C. Bruschini, "Optical-stack optimization for improved SPAD photon detection efficiency", in *Quantum Sensing and Nano Electronics and Photonics XVI*, SPIE, vol. 10926, 2019, pp. 359–365.

[121] M. W. Fishburn and E. Charbon, "System tradeoffs in gamma-ray detection utilizing SPAD arrays and scintillators", *IEEE Transactions on Nuclear Science*, vol. 57, no. 5, pp. 2549–2557, 2010.

[122] S. Pellegrini, B. Rae, A. Pingault, D. Golanski, S. Jouan, C. Lapeyre, and B. Mamdy, "Industrialised SPAD in 40 nm technology", in *Proc. 2017 IEEE Int. Electron Devices Meeting*, 2017, pp. 16.5.1–16.5.4.

[123] I. Gyongy, A. Davies, B. Gallinet, N. A. Dutton, R. R. Duncan, C. Rickman, R. K. Henderson, and P. A. Dalgarno, "Cylindrical microlensing for enhanced collection efficiency of small pixel SPAD arrays in single-molecule localisation microscopy", *Optics express*, vol. 26, no. 3, pp. 2280–2291, 2018.

[124] R. K. Henderson, N. Johnston, F. M. Della Rocca, H. Chen, D. D.-U. Li, G. Hungerford, R. Hirsch, D. Mcloskey, P. Yip, and D. J. Birch, "A 192 × 128 time correlated SPAD image sensor in 40-nm CMOS technology", *IEEE Journal of Solid-State Circuits*, vol. 54, no. 7, pp. 1907–1916, 2019.

[125] K. Zang, X. Jiang, Y. Huo, X. Ding, M. Morea, X. Chen, C.-Y. Lu, J. Ma, M. Zhou, Z. Xia, *et al.*, "Silicon single-photon avalanche diodes with nano-structured light trapping", *Nature communications*, vol. 8, no. 1, pp. 1–6, 2017.

[126] L. Frey, M. Marty, S. André, and N. Moussy, "Enhancing near-infrared photodetection efficiency in SPAD with silicon surface nanostructuration", *IEEE Journal of the Electron Devices Society*, vol. 6, pp. 392–395, 2018.

[127] M. Ghioni, G. Armellini, P. Maccagnani, I. Rech, M. K. Emsley, and M. S. Unlu, "Resonant-cavity-enhanced single-photon avalanche diodes on reflecting silicon substrates", *IEEE Photonics Technology Letters*, vol. 20, no. 6, pp. 413–415, 2008.

[128] I. Oshiyama, S. Yokogawa, H. Ikeda, Y. Ebiko, T. Hirano, S. Saito, T. Oinoue, Y. Hagimoto, and H. Iwamoto, "Near-infrared sensitivity enhancement of a back-illuminated

complementary metal oxide semiconductor image sensor with a pyramid surface for diffraction structure", in *2017 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2017, pp. 16–4.

[129] S. Yokogawa, "Nanophotonics contributions to state-of-the-art CMOS image sensors", in *2019 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2019, pp. 16–1.

[130] H. Cansizoglu, E. P. Devine, Y. Gao, S. Ghandiparsi, T. Yamada, A. F. Elrefaie, S.-Y. Wang, and M. S. Islam, "A new paradigm in high-speed and high-efficiency silicon photodiodes for communication—Part I: enhancing photon–material interactions via low-dimensional structures", *IEEE Transactions on Electron Devices*, vol. 65, no. 2, pp. 372–381, 2017.

[131] H. Cansizoglu, E. P. Devine, Y. Gao, S. Ghandiparsi, T. Yamada, A. F. Elrefaie, S.-Y. Wang, and M. S. Islam, "A new paradigm in high-speed and high-efficiency silicon photodiodes for communication—Part I: enhancing photon–material interactions via low-dimensional structures", *IEEE Transactions on Electron Devices*, vol. 65, no. 2, pp. 372–381, 2017.

[132] C. Veerappan, Y. Maruyama, and E. Charbon, "Silicon integrated electrical micro-lens for CMOS SPADs based on avalanche propagation phenomenon", in *International Image Sensor Workshop*, 2013.

[133] K. Morimoto *et al.*, "Charge-focusing SPAD image sensors for low light imaging applications", in *Int. SPAD Sensor workshop*, 2020.

[134] E. Van Sieleghem, A. Süss, P. Boulenc, J. Lee, G. Karve, K. De Munck, C. Cavaco, and C. Van Hoof, "A near-infrared enhanced silicon single-photon avalanche diode with a spherically uniform electric field peak", *IEEE Electron Device Letters*, vol. 42, no. 6, pp. 879–882, 2021.

[135] E. Van Sieleghem, G. Karve, K. De Munck, A. Vinci, C. Cavaco, A. Süss, C. Van Hoof, and J. Lee, "A backside-illuminated charge-focusing silicon SPAD with enhanced near-infrared sensitivity", *IEEE Transactions on Electron Devices*, vol. 69, no. 3, pp. 1129–1136, 2022.

[136] K. Morimoto, J. Iwata, M. Shinohara, H. Sekine, A. Abdelghafar, H. Tsuchiya, Y. Kuroda, K. Tojima, W. Endo, Y. Maehashi, Y. Ota, T. Sasago, S. Maekawa, S. Hikosaka, T. Kanou, A. Kato, T. Tezuka, S. Yoshizaki, T. Ogawa, K. Uehira, A. Ehara, F. Inui, Y. Matsuno, K. Sakurai, and T. Ichikawa, "3.2 megapixel 3D-stacked charge focusing SPAD for low-light imaging and depth sensing", in *2021 IEEE International Electron Devices Meeting (IEDM)*, 2021.

[137] F. Gramuglia, P. Keshavarzian, E. Kizilkan, C. Bruschini, S. S. Tan, M. Tng, E. Quek, M.-J. Lee, and E. Charbon, "Engineering breakdown probability profile for PDP and DCR optimization in a SPAD fabricated in a standard 55 nm BCD process", *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 2, pp. 1–10, Mar. 2022.

[138] M.-J. Lee and E. Charbon, "Progress in single-photon avalanche diode image sensors in standard CMOS: from two-dimensional monolithic to three-dimensional-stacked technology", *Japanese Journal of Applied Physics*, vol. 57, no. 10, 1002A3, 2018.

[139] M. Assanelli, A. Ingargiola, I. Rech, A. Gulinatti, and M. Ghioni, "Photon-timing jitter dependence on injection position in single-photon avalanche diodes", *IEEE J. Quantum Electron.*, vol. 47, no. 2, pp. 151–159, Jan. 2011.

[140] C. H. Tan, J. S. Ng, G. J. Rees, and J. P. R. David, "Statistics of avalanche current buildup time in single-photon avalanche diodes particle detection", *IEEE J. Sel. Topics Quantum Electron.*, vol. 13, no. 4, pp. 903–910, Sep. 2007.

[141] A. Spinelli and A. L. Lacaita, "Physics and numerical simulation of single photon avalanche diodes", *IEEE J. Electron Devices Soc.*, vol. 44, no. 11, pp. 1931–1943, Nov. 1997.

[142] F. Gramuglia, M.-L. Wu, C. Bruschini, M.-J. Lee, and E. Charbon, "A low-noise CMOS SPAD pixel with 12.1 ps SPTR and 3 ns dead time", *IEEE J. Sel. Topics Quantum Electron.*, vol. 28, no. 2, pp. 1–10, Mar. 2022.

[143] M. Liu, C. Hu, X. Bai, X. Guo, J. C. Campbell, Z. Pan, and M. M. Tashima, "High-performance InGaAs/InP single-photon avalanche photodiode", *IEEE Journal of selected topics in quantum electronics*, vol. 13, no. 4, pp. 887–894, 2007.

[144] E. A. Webster, R. L. Nicol, L. Grant, and D. Renshaw, "Per-pixel dark current spectroscopy measurement and analysis in CMOS image sensors", *IEEE Transactions on electron devices*, vol. 57, no. 9, pp. 2176–2182, 2010.

[145] E. A. Webster and R. K. Henderson, "A TCAD and spectroscopy study of dark count mechanisms in single-photon avalanche diodes", *IEEE transactions on electron devices*, vol. 60, no. 12, pp. 4014–4019, 2013.

[146] J. D. Petticrew, S. J. Dimler, X. Zhou, A. P. Morrison, C. H. Tan, and J. S. Ng, "Avalanche breakdown timing statistics for silicon single photon avalanche diodes", *IEEE Journal of Selected Topics in Quantum Electronics*, vol. 24, no. 2, pp. 1–6, 2017.

[147] F. Sun, Y. Xu, Z. Wu, and J. Zhang, "A simple analytic modeling method for SPAD timing jitter prediction", *IEEE Journal of the Electron Devices Society*, vol. 7, pp. 261–267, 2019.

[148] A. Ingargiola, M. Assanelli, A. Gallivanoni, I. Rech, M. Ghioni, and S. Cova, "Avalanche buildup and propagation effects on photon-timing jitter in Si-SPAD with non-uniform electric field", in *Advanced Photon Counting Techniques III*, SPIE, vol. 7320, 2009, pp. 103–114.

[149] A. Lacaita, A. Spinelli, and S. Longhi, "Avalanche transients in shallow p-n junctions biased above breakdown", *Applied physics letters*, vol. 67, no. 18, pp. 2627–2629, 1995.

[150] Y. Okuto and C. Crowell, "Ionization coefficients in semiconductors: a nonlocalized property", *Physical review B*, vol. 10, no. 10, p. 4284, 1974.

[151] M. Gersbach, C. Niclass, E. Charbon, J. Richardson, R. Henderson, and L. Grant, "A single photon detector implemented in a 130nm CMOS imaging process", in *Proc. 38th Eur. Solid-State Device Res. Conf.*, Nov. 2008, pp. 270–273.

[152] J. A. Richardson, L. A. Grant, and R. K. Henderson, "Low dark count single-photon avalanche diode structure compatible with standard nanometer scale CMOS technology", *IEEE Photon. Technol. Lett.*, vol. 21, no. 14, pp. 1020–1022, Jul. 2009.

[153] J. A. Richardson, E. A. G. Webster, L. A. Grant, and R. K. Henderson, "Scaleable single-photon avalanche diode structures in nanometer CMOS technology", *IEEE Trans. Electron Devices*, vol. 58, no. 7, pp. 2028–2035, Jul. 2011.

[154] E. A. G. Webster, J. A. Richardson, L. A. Grant, D. Renshaw, and R. K. Henderson, "A single-photon avalanche diode in 90-nm CMOS imaging technology with 44% photon detection efficiency at 690 nm", *IEEE Electron Device Lett.*, vol. 33, no. 5, pp. 694–696, May 2012.

[155] E. A. G. Webster, L. A. Grant, and R. K. Henderson, "A high-performance single-photon avalanche diode in 130-nm CMOS imaging technology", *IEEE Electron Device Lett.*, vol. 33, no. 11, pp. 1589–1591, Nov. 2012.

[156] T. Leitner, A. Feiningstein, R. Turchetta, R. Coath, S. Chick, G. Visokolov, V. Savuskan, M. Javitt, L. Gal, I. Brouk, S. Bar-Lev, and Y. Nemirovsky, "Measurements and simulations of low dark count rate single photon avalanche diode device in a low voltage 180-nm CMOS image sensor technology", *IEEE Trans. Electron Devices*, vol. 60, no. 6, pp. 1982–1988, Jun. 2013.

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

[158] F. Villa, D. Bronzi, Y. Zou, C. Scarcella, G. Boso, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, U. Paschen, and W. Brockherde, "CMOS SPADs with up to 500 μm diameter and 55% detection efficiency at 420 nm", *J. Mod. Opt.*, vol. 61, no. 2, pp. 102–115, Jun. 2014.

[159] C. Veerappan and E. Charbon, "A substrate isolated CMOS SPAD enabling wide spectral response and low electrical crosstalk", *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 6, pp. 299–305, Nov. 2014.

[160] M.-J. Lee, P. Sun, and E. Charbon, "A first single-photon avalanche diode fabricated in standard SOI CMOS technology with a full characterization of the device", *Opt. Express*, vol. 23, no. 10, pp. 13 200–13 209, May 2015.

[161] C. Veerappan and E. Charbon, "CMOS SPAD based on photo-carrier diffusion achieving PDP >40% from 440 to 580 nm at 4 v excess bias", *IEEE Photon. Technol. Lett.*, vol. 27, no. 23, pp. 2445–2448, Dec. 2015.

[162] C. Veerappan and E. Charbon, "A low dark count p-i-n diode based SPAD in CMOS technology", *IEEE Trans. Electron Devices*, vol. 63, no. 1, pp. 65–71, Jan. 2016.

[163] I. Takai, H. Matsubara, M. Soga, M. Ohta, M. Ogawa, and T. Yamashita, "Single-photon avalanche diode with enhanced NIR-sensitivity for automotive LIDAR systems", *Sensors*, vol. 16, no. 4, Mar. 2016.

[164] S. Pellegrini and B. Rae, "Fully industrialised single photon avalanche diodes", in *Proc. Adv. Photon Count. Techniq. XI*, vol. 10212, 2017, pp. 16.5.1–16.5.4.

[165] H. Xu, L. Pancheri, G.-F. D. Betta, and D. Stoppa, "Design and characterization of a p+/n-well SPAD array in 150nm CMOS process", *Opt. Express*, vol. 25, no. 11, pp. 12 765–12 778, May 2017.

[166] P. Keshavarzian, K. Ramu, D. Tang, C. Weill, F. Gramuglia, S. S. Tan, M. Tng, L. Lim, E. Quek, D. Mandich, M. Stipčević, and E. Charbon, "A 3.3 Gbps SPAD-based quantum random number generator", *arXiv preprint arXiv:2209.04868, accepted for publication in IEEE JSSC (DOI: 10.1109/JSSC.2023.3274692)*, Jul. 2022.

[167] W. R. Leo, *Techniques for nuclear and particle physics experiments: a how-to approach*. Springer Science & Business Media, 2012.

[168] S. H. Lee and R. P. Gardner, "A new G–M counter dead time model", *Applied Radiation and Isotopes*, vol. 53, no. 4-5, pp. 731–737, 2000.

[169] L. Neri, S. Tudisco, F. Musumeci, A. Scordino, G. Fallica, M. Mazzillo, and M. Zimbone, "Note: dead time causes and correction method for single photon avalanche diode devices", *Review of Scientific Instruments*, vol. 81, no. 8, p. 086 102, 2010.

[170] I. Straka, J. Grygar, J. Hloušek, and M. Ježek, "Counting statistics of actively quenched SPADs under continuous illumination", *Journal of Lightwave Technology*, vol. 38, no. 17, pp. 4765–4771, 2020.

[171] M. Stipčević, I. M. Antolović, C. Bruschini, and E. Charbon, "Scalable quantum random number generator for cryptography based on the random flip-flop approach", *arXiv 2102.12204*, 2021.

[172] H. Xu, N. Massari, L. Gasparini, A. Meneghetti, and A. Tomasi, "A SPAD-based random number generator pixel based on the arrival time of photons", *Integration*, vol. 64, pp. 22–28, 2019.

[173] A. Tontini, L. Gasparini, N. Massari, and R. Passerone, "SPAD-based quantum random number generator with a Nth-order rank algorithm on FPGA", *IEEE Trans. Circuits Syst. II*, vol. 66, no. 12, pp. 2067–2071, Dec. 2019.

[174] Q. Yan, B. Zhao, Z. Hua, Q. Liao, and H. Yang, "High-speed quantum-random number generation by continuous measurement of arrival time of photons", *Review of Scientific Instruments*, vol. 86, no. 7, p. 073 113, 2015.

[175] A. Stanco, D. G. Marangon, G. Vallone, S. Burri, E. Charbon, and P. Villoresi, "Efficient random number generation techniques for CMOS single-photon avalanche diode array exploiting fast time tagging units", *Physical Review Research*, vol. 2, no. 2, p. 023 287, 2020.

[176] S. M. Ross, *Applied probability models with optimization applications*. Courier Corporation, 1970.

[177] Z. Bisadi, G. Fontana, E. Moser, G. Pucker, and L. Pavesi, "Robust quantum random number generation with silicon nanocrystals light source", *Journal of Lightwave Technology*, vol. 35, no. 9, pp. 1588–1594, 2017.

[178] F. Acerbi, Z. Bisadi, G. Fontana, N. Zorzi, C. Piemonte, and L. Pavesi, "A robust quantum random number generator based on an integrated emitter-photodetector structure", *IEEE J. Sel. Topics Quantum Electron.*, vol. 24, no. 6, pp. 1–7, Nov. 2018.

[179] V.-L. Dosan, M. Mihăilescu, N. Tarbă, M.-A. Ungureanu, and R. Ionicioiu, "Quantum random number generation with down converted photon pairs", in *Advanced Topics in Optoelectronics, Microelectronics and Nanotechnologies X*, SPIE, vol. 11718, 2020, pp. 185–192.

[180] N. Massari, L. Gasparini, A. Tomasi, A. Meneghetti, H. Xu, and D. Perenzoni, "16.3 16 × 16pixel SPAD-based 128-Mb/s quantum random number generator with -74dB light rejection ratio and -6.7ppm/°C bias sensitivity on temperature", in *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*, San Francisco, CA, USA, Jan. 2016, pp. 292–293.

[181] H. Xu, D. Perenzoni, A. Tomasi, and N. Massari, "A 16 × 16pixel post-processing free quantum random number generator based on SPADs", *IEEE Trans. Circuits Syst. II*, vol. 65, no. 5, pp. 627–631, May 2018.

[182] A. Tomasi, A. Meneghetti, N. Massari, L. Gasparini, D. Rucatti, and H. Xu, "Model, validation, and characterization of a robust quantum random number generator based on photon arrival time comparison", *J. Lightw. Technol.*, vol. 36, no. 18, pp. 3843–3854, Jun. 2018.

[183] A. V. Losev, V. V. Zavodilenko, A. A. Koziy, A. A. Filyaev, K. I. Khomyakova, Y. V. Kurochkin, and A. A. Gorbatsevich, "Dead Time Duration and Active Reset Influence on the Afterpulse Probability of InGaAs/InP Single-Photon Avalanche Diodes", *IEEE Journal of Quantum Electronics*, vol. 58, no. 3, pp. 1–11, Jun. 2022.

[184] M. A. Itzler, X. Jiang, and M. Entwistle, "Power law temporal dependence of InGaAs/InP SPAD afterpulsing", *Journal of Modern Optics*, vol. 59, no. 17, pp. 1472–1480, Oct. 2012.

[185] T. Ferreira da Silva, G. B. Xavier, and J. P. von der Weid, "Real-Time Characterization of Gated-Mode Single-Photon Detectors", *IEEE Journal of Quantum Electronics*, vol. 47, no. 9, pp. 1251–1256, Sep. 2011.

[186] Z. Cheng, X. Zheng, D. Palubiak, M. J. Deen, and H. Peng, "A Comprehensive and Accurate Analytical SPAD Model for Circuit Simulation", *IEEE Transactions on Electron Devices*, vol. 63, no. 5, pp. 1940–1948, May 2016.

[187] G. Giustolisi, R. Mita, and G. Palumbo, "Verilog-A modeling of SPAD statistical phenomena", in *2011 IEEE International Symposium of Circuits and Systems (ISCAS)*, IEEE, May 2011, pp. 773–776.

[188] J. M. Lopez-Martinez, I. Vornicu, R. Carmona-Galan, and A. Rodriguez-Vazquez, "An Experimentally-Validated Verilog-A SPAD Model Extracted from TCAD Simulation", in *2018 25th IEEE International Conference on Electronics, Circuits and Systems (ICECS)*, Dec. 2018, pp. 137–140.

[189] S. Zhu, Y. Ma, X. Li, J. Yang, J. Lin, and J. Jing, "On the analysis and improvement of min-entropy estimation on time-varying data", *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1696–1708, 2019.

[190] S. Zhu, Y. Ma, T. Chen, J. Lin, and J. Jing, "Analysis and improvement of entropy estimators in NIST SP 800-90B for non-IID entropy sources", *IACR Transactions on Symmetric Cryptology*, pp. 151–168, 2017.

[191] Y. Kim, C. Guyot, and Y.-S. Kim, "On the efficient estimation of min-entropy", *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3013–3025, 2021.

[192] S. Burri, D. Stucki, Y. Maruyama, C. Bruschini, E. Charbon, and F. Regazzoni, "Jailbreak imagers: transforming a single-photon image sensor into a true random number generator", in *Proc. Int. Image Sensors Works. (IISW)*, Snowbird, UT, USA, Jun. 2013, pp. 1–4.

[193] S. Tisa, F. Villa, A. Giudice, G. Simmerle, and F. Zappa, "High-speed quantum random number generation using CMOS photon counting detectors", *IEEE J. Sel. Topics Quantum Electron.*, vol. 21, no. 3, pp. 23–29, May 2015.

[194] N. Massari, A. Tontini, L. Parmesan, M. Perenzoni, M. Gruijé, I. Verbauwhede, T. Strohm, D. Oshinubi, I. Herrmann, and A. Brenneis, "A monolithic SPAD-based random number generator for cryptographic application", in *ESSCIRC 2022-IEEE 48th European Solid State Circuits Conference (ESSCIRC)*, IEEE, 2022, pp. 73–76.

[195] F. Regazzoni, E. Amri, S. Burri, D. Rusca, H. Zbinden, and E. Charbon, "A high speed integrated quantum random number generator with on-chip real-time randomness extraction", *arXiv 2102.06238*, 2021.

[196] M. Grujić and I. Verbauwhede, "TROT: a three-edge ring oscillator based true random number generator with time-to-digital conversion", *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 6, pp. 2435–2448, 2022.

[197] V. Rožić and I. Verbauwhede, "Hardware-efficient post-processing architectures for true random number generators", *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 66, no. 7, pp. 1242–1246, 2018.

[198] J. Xue, J. Kim, A. Mestre, K. Tan, D. Chong, S. Roy, H. Nong, K. Lim, D. Gray, D. Kramnik, *et al.*, "Low voltage, high brightness CMOS LEDs", in *2020 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2020, pp. 33–5.

[199] J. Xue, J. Kim, A. Mestre, K. M. Tan, D. Chong, S. Roy, H. Nong, K. Y. Lim, D. Gray, D. Kramnik, A. Atabaki, E. Quek, and R. J. Ram, "Low-voltage, high-brightness silicon micro-LEDs for CMOS photonics", *IEEE Trans. Electron Devices*, vol. 68, no. 8, pp. 3870–3875, Jun. 2021.

[200] R. J. Ram and J. Xue, "CMOS nanoLEDs", in *Photonic and Phononic Properties of Engineered Nanostructures XII*, SPIE, 2022, PC1201002.

[201] A. Khanmohammadi, R. Enne, M. Hofbauer, and H. Zimmermanna, "A monolithic silicon quantum random number generator based on measurement of photon detection time", *IEEE Photon. J.*, vol. 7, no. 5, pp. 1–13, Oct. 2015.

[202] D. Frauchiger, R. Renner, and M. Troyer, "True randomness from realistic quantum devices", *arXiv preprint arXiv:1311.4547*, 2013.

[203] H. Li, J. Zhang, Z. Li, J. Liu, and Y. Wang, "Improvement of min-entropy evaluation based on pruning and quantized deep neural network", *IEEE Transactions on Information Forensics and Security*, 2023.

Pouyan Keshavarzian

Last Updated on 28th February 2023

Mixed-Signal Engineer

Address
Neuchâtel
Switzerland

Contact Info
+41 76 727 4220

pouyan.keshavarzian
@epfl.ch

LinkedIn
\in\keshavarzian

GitHub
\pkeshava

CAD TOOLS
Cadence Virtuoso, Ansys HFSS,
Keysight ADS/Momentum, Altium, TCAD, LTSpice, AutoCAD, CADstar

Skills
♥ IC Design, Discrete RF & Microwave Circuits, Antenna Design, Algorithm Design, Leadership, Programming, GIT, Latex, Advanced Testing/Debugging using Oscilloscopes, VNA, VSA, etc.

Programming Languages
♥ Python, C/C++, Matlab, SKILL, Verilog, Verilog-AMS

Professional Organizations
IEEE Microwave Theory and Techniques Society

IEEE Solid State Circuits Society

Graduate Awards and Scholarships
Teaching Excellence Award (2018)
Alberta Graduate Student Award (2018)
University of Calgary 3 Min. Thesis 2nd Place (2018)
IEEE MTT-S Pre-Graduate Award (2017)
Queen Elizabeth II Schol. (2016 & 2017)

Education

2019–Present	PhD in Microelectronics CMOS SPAD Circuits & Systems	École Polytechnique Fédérale de Lausanne (EPFL)
2016–2018	M.Sc. in Electrical Engineering Thesis: Active-loaded Phase-Conjugating Rotman Lens for Intelligent Transportation System Backscattering Applications	University of Calgary
2010–2015	B.Sc. in Electrical Engineering	University of Calgary

Experience

Feb 2019-Present — **École Polytechnique Fédérale de Lausanne (EPFL)** — Neuchâtel, Switzerland
Doctoral Assistant

Sept 2016-Dec 2018 — **University of Calgary** — Calgary, Canada
Graduate Research Assistant

May 2013-Aug 2016 — **Garmin Canada** — Cochrane, Canada
Hardware Engineer 2 (June-Aug 2016)
Hardware Engineer 1 (June 2015-May 2016)
Hardware Engineering Contractor (Sept 2014-May 2015)
Hardware Engineering Intern (May 2013-Aug 2014)