Doctoral Thesis

University of Trento

Thesis Title

All-Silicon-Based Photonic Quantum Random Number Generators

PhD Candidate:

Zahra Bisadi

Supervisor:

Prof. Lorenzo Pavesi

Trento, July 2017

Abstract

Random numbers are fundamental elements in different fields of science and technology such as computer simulation like Monte Carlo-method simulation, statistical sampling, cryptography, games and gambling, and other areas where unpredictable results are necessary.

Random number generators (RNG) are generally classified as "pseudo"-random number generators (PRNG) and "truly" random number generators (TRNG). Pseudo random numbers are generated by computer algorithms with a (random) seed and a specific formula. The random numbers produced in this way (with a small degree of unpredictability) are good enough for some applications such as computer simulation. However, for some other applications like cryptography they are not completely reliable.

When the seed is revealed, the entire sequence of numbers can be produced. The periodicity is also an undesirable property of PRNGs that can be disregarded for most practical purposes if the sequence recurs after a very long period. However, the predictability still remains a tremendous disadvantage of this type of generators.

Truly random numbers, on the other hand, can be generated through physical sources of randomness like flipping a coin. However, the approaches exploiting classical motion and classical physics to generate random numbers possess a deterministic nature that is transferred to the generated random numbers. The best solution is to benefit from the assets of indeterminacy and randomness in quantum physics.

Based on the quantum theory, the properties of a particle cannot be determined with arbitrary precision until a measurement is carried out. The result of a measurement, therefore, remains unpredictable and random. Optical phenomena including photons as the quanta of light have various random, non-deterministic properties. These properties include the polarization of the photons, the exact number of photons impinging a detector and the photon arrival times. Such intrinsically random properties can be exploited to generate truly random numbers.

Silicon (Si) is considered as an interesting material in integrated optics. Microelectronic chips made from Si are cheap and easy to mass-fabricate, and can be densely integrated. Si integrated optical chips, that can generate, modulate, process and detect light signals, exploit the benefits of Si while also being fully compatible with electronic. Since many electronic components can be integrated into a single chip, Si is an ideal candidate for the production of small, powerful devices. By complementary metal-oxide-semiconductor (CMOS) technology, the fabrication of compact and mass manufacturable devices with integrated components on the Si platform is achievable.

In this thesis we aim to model, study and fabricate a compact photonic quantum random number generator (QRNG) on the Si platform that is able to generate high quality, “truly” random numbers. The proposed QRNG is based on a Si light source (LED) coupled with a Si single photon avalanche diode (SPAD) or an array of SPADs which is called Si photomultiplier (SiPM). Various implementations of QRNG have been developed reaching an ultimate geometry where both the source and the SPAD are integrated on the same chip and fabricated by the same process.

This activity was performed within the project SiQuro—on Si chip quantum optics for quantum computing and secure communications—which aims to bring the quantum world into integrated photonics. By using the same successful paradigm of microelectronics—the study and design of very small electronic devices typically made from semiconductor materials—, the vision is to have low cost and mass manufacturable integrated quantum photonic circuits for a variety of different applications in quantum computing, measure, sensing, secure communications and services. The Si platform permits, in a natural way, the integration of quantum photonics with electronics. Two methodologies are presented to generate random numbers: one is based on photon counting measurements and another one is based on photon arrival time measurements. The latter is robust, masks all the drawbacks of afterpulsing, dead time and jitter of the Si SPAD and is effectively insensitive to ageing of the LED and to its emission drifts related to temperature variations. The raw data pass all the statistical tests in national institute of standards and technology (NIST) tests suite and TestU01 Alphabit battery without a post processing algorithm. The maximum demonstrated bit rate is 1.68 Mbps with the efficiency of 4-bits per detected photon.

In order to realize a small, portable QRNG, we have produced a compact configuration consisting of a Si nanocrystals (Si-NCs) LED and a SiPM. All the statistical test in the NIST tests suite pass for the raw data with the maximum bit rate of 0.5 Mbps. We also prepared and studied a compact chip consisting of a Si-NCs LED and an array of detectors. An integrated chip, composed of Si p^+/n junction working in avalanche region and a Si SPAD, was produced as well. High quality random numbers are produced through our robust methodology at the highest speed of 100 kcps.

Integration of the source of entropy and the detector on a single chip is an efficient way to produce a compact RNG. A small RNG is an essential element to guarantee the security of our everyday life. It can be readily implemented into electronic devices for data encryption. The idea of “utmost security” would no longer be limited to particular organs owning sensitive information. It would be accessible to every one in everyday life.

Contents

List of Tables

List of Figures

Acronyms

QRNG	Quantum Random Number Generator
QKD	Quantum Key Distribution
QBER	Quantum Bit Error Rate
BER	Bit Error Rate
RNG	Random Number Generator
PRNG	Pseudo Random Number Generator
TRNG	Truly Random Number Generator
CMOS	Complementary Metal-Oxide-Semiconductor
Si-NC	Silicon Nanocrystal
SPAD	Single Photon Avalanche Diode
FPGA	Field-Programmable Gate Array
LED	Light Emittinh Device
LLED	Large area Light Emitting Device
TDC	Time to Digital Converter
ADC	Analog to Digital Converter
AC	Alternate Current
XOR	Exclusive OR
LSB	Least Significant Bit
SL	Semiconductor Superlattice
G-M	Geiger Müler
MOS	Metal-Oxide-Semiconductor
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
BS	Beam Splitter
PMT	Photomultiplier Tube
CW	Continuous-Wave
APD	Avalanche Photodiode
LO	Local Oscillator
MSB	Most Significant Bit
LD	Laser Diode
WSS	Wide-Sense Stationary

JPMF	Joint Probability Mass Function
MI	Mutual Information
NIST	National Institute of Standards and Technology
ANSI	American National Standards Institute
PECVD	Plasma Enhanced Chemical Vapor Deposition
SRO	Silicon Rich Oxide
FN	Fowler-Nordheim
EL	Electroluminescence
CCD	Charge Coupled Device
DCR	Dark Count Rate
FWHM	Full Width at Half Maximum
AP	Afterpulsing Probability
MCS	Multichannel Scaler
nngs	No-Number Generating Subinterval
rngs	Random Number Generating Subinterval
SiPM	Silicon Photomultiplier
PCB	Printed Circuit Board
PET	Positron Emission Tomography
DeCT	Delayed Crosstalk
PDE	Photon Detection Efficiency
TVD	Total Variation Distance
NI	National Instruments
GS	Gain Switching
NZD	Nearly Zero Detuning
pdf	Probability Density Function
cdf	Cumulative Distribution Function
i.i.d.	Independent and Identically Distributed

Chapter 1

Introduction

> ``Anyone who attempts to generate random numbers by deterministic means is, of course, living in a state of sin.''
>
> John von Neumann

A random number is a string of bits taken from a sequence with a distribution which has a constant probability for every bit belonging to that distribution (such a distribution is called a uniform distribution). [1] For such a sequence, it must be impossible to predict the future and past bits based on the knowledge of the present bit. Random numbers are fundamental elements in different fields of science and technology such as computer simulation like Monte Carlo-method simulation, statistical sampling, cryptography, games and gambling, and other areas where unpredictable results are necessary.

1.1 Applications of random numbers

In this section, we explain the applications of random numbers in cryptography, simulations and games. To our particular interest is their application in cryptography which plays a really crucial role in the security of our everyday life and is the main area for which the photonic quantum random number generators (QRNGs) introduced in this thesis are intended for.

1.1.1 Random numbers in cryptography

Cryptography is the art of rendering messages unintelligible or indecipherable to any unauthorized party. Methods and technologies for secure data transmission are developed and studied in this field. Data confidentiality, data integrity, and authentication are different aspects in information security and are very important in modern cryptography.

During on-line transactions, random numbers are used to produce confidential data (encryption), maintain and assure the accuracy and consistency of data during the process (digital signature), and to authenticate the user's identity—e.g. the credit card owner's identity—(challenge-response protocols). [2]

For secure data transmission in cryptography, random numbers are used to encrypt and decrypt secret messages between a sender and a receiver, conventionally called Alice and Bob, respectively (Fig. 1.1). Alice uses a key to encrypt a message (plaintext) and sends the encrypted message (ciphertext) to Bob. In symmetric encryption technique, where both Alice and Bob share a secret key, if an adversary, conventionally called the Eavesdropper, has access to this key and the ciphertext, the security of the transmission channel is lost and the confidential information is known to the irrelevant parties (the Eavesdropper). In asymmetric encryption technique, however, there is a key pair, a public key and a private key. Therefore, the security of this technique is higher than the symmetric encryption. The public key is available to anyone who might want to send a message to Bob, but the private key is kept secret and only Bob knows it. Alice's encrypted message using the public (private) key can only be decrypted using Bob's private (public) key which matches the corresponding public (private) key. Therefore, passing public keys over communication channels (like the Internet) does not create a problem. The weakness of asymmetric encryption is that it is slower than symmetric encryption; a lot more processing power is required to encrypt and decrypt messages than the symmetric encryption technique.

Quantum cryptography is the art and science of exploiting quantum mechanical effects in order to perform cryptographic tasks. It was proposed in the 1980s, first by Stephen Wiesner (1983) and then by Charles H. Bennett and Gilles Brassard (1984, 1985). [3] The rules of quantum physics can be used in "quantum" key distribution (QKD) which is one important application of quantum cryptography. QKD takes advantage of quantum mechanics for secure communications. Let Alice send some information to Bob through a quantum channel (e.g. an optical fiber) using some property of photons (e.g. polarization of photons) to encrypt her message. If the Eavesdropper attempts to have access to the encrypted message by executing some measurement, according to quantum physics, she perturbs the system and hence her presence would be known. However, it is more reasonable if Alice and Bob, before exchanging a message, exchange a key through the quantum channel. When they make sure the channel is safe, they can use the key to encrypt the message and transfer it through the secure quantum channel.

This concept has been dealt with in QKD protocols like the Bennett and Brassard (BB84) protocol. [4] This protocol can be explained using the concept of "qubit" which is the quantum unit of information and its state is defined as the superposition of two states characterizing a single particle (e.g. a spin 1/2 system or the two-state polarization of a photon). For a photon, the qubit can be either the superposition of vertical and horizontal linear polarization or right and left circular polarization—the binary value '0' can be attributed to horizontal and right circular polarization and '1' to vertical and left circular polarization. Alice prepares the states and sends them to Bob one after the other (attributing an index i to each state). Bob chooses randomly one of the two bases (linear or circular), measures the states and attributes binary values of 0 and 1 (outcomes) to them. Then, they check a subset of the key through a classical channel. Bob informs Alice of the basis he chose to measure each qubit without telling her the result. Alice tells him which bases are the same as the ones she used and which ones are different. They discard the outcomes for which they have different bases and keep the remaining ones. This step is called *sifting* and the remaining key is called *sifted* key. Then they need to verify their outcomes: Alice chooses some indices i at random and reveals the outcome to Bob. He compares her outcomes with his own and notifies Alice of the disagreements.

Quantum bit error rate (QBER) is defined as the ratio of discarded bits to the total number of bits received by Bob. QBER was proposed by Beat Perny and Paul Townsend to name the error rate in sifted key to distinguish it from bit error rate (BER) used in standard communications. [3] If QBER$\leq$ 11%, the quantum channel is considered to be safe and Alice can encrypt her message using the transmitted key. Otherwise, the protocol aborts and the key is discarded. It should be noted that in this case, if the Evesdropper intercepts a qubit propagating from Alice to Bob, she just lowers the key rate without gaining any useful information (if Bob does not receive an expected qubit, he informs Alice to disregard it).

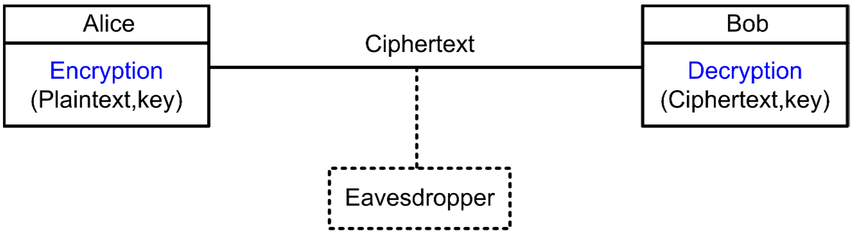


Figure 1.1: Scheme of the communication between a sender (Alice) and a receiver (Bob) in cryptography. Alice encypts a message (plaintext) using a key to produce an encrypted message (ciphertext). Bob decrypts the ciphertext using a secret key (symmetric encryption) or a public and private key (asymmetric encryption). The presence of an Eavesdropper, with the knowledge of the ciphertext and the key, endagers the secure communication between Alice and Bob.

1.1.2 Random numbers in simulations

Random numbers are extensively used in problems with random behavior that require lots of experiments to be executed. Simulation programs run a large number of experiments and record the outcome of events. These programs solve equations in mathematical models and use random numbers to estimate probabilities. The mathematical technique of letting the computer perform lots of experiments based on drawing random numbers is commonly called Monte Carlo simulation. It is a computational algorithm that relies on repeated random sampling to obtain numerical results. It is mainly used in three main areas: optimization, numerical integration and random sampling from certain probability distributions. [5]

Monte Carlo method has been widely used to treat optimization problems. Numerical solutions to optimization problems incur the risk of getting stuck in local minima. [6] Monte Carlo method generates and uses random variables as well as random iterates. The random iterates may enable the method to escape a local optimum and eventually to approach a global optimum. It is a simple and effective way to obtain algorithms with almost certain good performance uniformly across many data sets, for many sorts of problems.

In numerical integration, when the number of dimensions d (equivalently the degrees of freedom in many physical problems) is large, the deterministic numerical integration algorithms no longer work efficiently. Monte Carlo methods are used to solve the problem of the exponential increase in computation times imposed by going from one dimension to a multidimensional region. The number of function values or points increases rapidly with the number of dimensions.

For instance, if in one dimension n points are enough to provide the required accuracy, then n^d points are necessary for d dimensions. On the other hand, the boundary of a multidimensional region may be very complicated making it not very feasible to reduce the problem to an iterated integral. [7] By using Monte Carlo methods, it can be estimated by randomly selecting points in d-dimensional space, and taking some kind of average of the function values at these points (as long as the function in question is reasonably well-behaved). [8]

In sampling, the objective is to gather information about a random object by observing many realizations of it. An example is simulation modeling, where a random process mimics the behavior of some real-life system, such as a production line or telecommunications network. Another example is found in Bayesian statistics, where Markov chain Monte Carlo is often used to sample from a posterior distribution. [5]

1.1.3 Random numbers in games

Unpredictable outcome is highly significant in games of gambling in casinos, online slots games and computer games. It is essential to make the games less “determined” so that the players face a huge challenge to guess the correct answer or to find the best solution to win. If we consider tossing a fair coin or rolling a fair six-sided dice, the outcome has an equal probability of being heads or tails and 1, 2, 3, 4, 5 and 6, respectively. But a coin or a dice cannot be used in a computer, since the computers need to use a sort of algorithm to produce random numbers. If the random numbers are generated with a certain formula, they could be guessed, controlled and eventually the result of the game would not be unpredictable anymore and the player would have a high chance to win. [9] Most numbers used in games seem random enough, but they are not really random. A game program is developed with an implemented random number generator (RNG) which is fed by a seed to generate random numbers. All the subsequent random numbers in a stream of random numbers depend on the seed which if be discovered, results in a potential threat for the game to be controlled by the player. However, in online gaming, the risk of this is fairly low.

Random movements of the objects and creatures or random appearance of some game items, e.g. in video games, affect the course of the game to a great extent. In computer card games different tasks can be done by a RNG: drawing a card at random, drawing a hand of n cards at random from a shuffled deck, shuffling the deck and dealing cards to a set of players. In racing car and motorcycle games, random numbers are used to determine how fast or slow the other cars move. In fact, there would be an implemented RNG in every game today to assure the unpredictability of the outcomes in different stages of the game. [9]

1.2 Pseudo random numbers vs. truly random numbers

As mentioned in the previous section, in the field of cryptography, the key Alice and Bob use to encrypt and decrypt messages is of high significance for secure communications between them. Random numbers are used to produce cryptographic keys. RNGs are generally classified as “pseudo” random number generators (PRNG) and “truly” random number generators (TRNG). [10] Pseudo random numbers are generated by computer algorithms with a (random) seed and a specific formula. The random numbers produced in this way (with a small degree of unpredictability) are good enough for some applications such as computer simulation. However, for some other applications like cryptography they are not completely reliable. When the seed is revealed, the entire sequence of numbers can be produced. The periodicity is also an undesirable property of PRNGs that can be disregarded for most practical purposes if the sequence recurs after a very long period. However, the predictability still remains a tremendous disadvantage of this type of generators.

Truly random numbers, on the other hand, can be generated through physical sources of randomness like flipping a coin. However, the approaches exploiting classical motion and classical physics to generate random numbers possess a deterministic nature that is transferred to the generated random numbers. The best solution is to benefit from the assets of indeterminacy and randomness in quantum physics. Based on the quantum theory, the properties of a particle cannot be determined with arbitrary precision until a measurement is carried out. The result of a measurement, therefore, remains unpredictable and random. [11] Optical phenomena including photons as the quanta of light have various random, non-deterministic properties. These properties include the polarization of the photons, the exact number of photons impinging a detector and the photon arrival times. Such intrinsically random properties can be exploited to generate truly random numbers.

1.3 All-silicon-based approach

Si is considered as an interesting material in integrated optics. Microelectronic chips made from Si are cheap and easy to mass-fabricate, and can be densely integrated. Si integrated optical chips, that can generate, modulate, process and detect light signals, exploit the benefits of Si while also being fully compatible with electronic. [12]

Si photonics has been used to develop various photonic devices based on silicon, such as waveguides, filters, and modulators. It is the leading candidate for optical interconnect (communication by optical fibers) due to its unique combination of low fabrication costs, performance enhancements resulting from electronic–photonic integration, and compatibility with the world's most successful technology for producing electronics, complementary metal-oxide-semiconductor (CMOS). [13] The interconnect market has gone through a transition from electrical to optical technology due to the limitations of copper as an interconnect medium including its high loss, dispersion and low fundamental speed. In addition, germanium photodetectors have been built on a Si photonic platform. These photonic devices have already been monolithically integrated on Si chips. [14] By CMOS technology, the fabrication of compact and mass-manufacturable devices with integrated components on the Si platform is achievable.

1.4 Objective of the thesis

This thesis has been carried out within the SiQuro project [15] (on Si chip quantum optics for quantum computing and secure communications) which aims to bring the quantum world into integrated photonics. By using the same successful paradigm of microelectronics—the study and design of very small electronic devices typically made from semiconductor materials—, the vision is to have low cost and mass manufacturable integrated quantum photonic circuits for a variety of different applications in quantum computing, measure, sensing, secure communications and services. The Si platform permits, in a natural way, the integration of quantum photonics with electronics.

The objective of this thesis, as part of the project SiQuro (Work Package 4), is to model, study and fabricate a compact photonic QRNG on the Si platform able to generate high quality, “truly” random numbers. To achieve this goal, we moved from a macroscopic structure with individual components of the source of entropy and the detector coupled with each other through an optical multimode fiber, to a compact configuration with the source of entropy closely coupled with the detector in free space and, eventually, to a microscopic structure with the source of entropy and detector integrated on a single Si chip.

1.5 Thesis outline

The thesis is divided into seven chapters; the content of each chapter is as follows. The first chapter contains an introduction on the thesis research work and the objective of the thesis. In the second chapter, different types of random number generators are introduced. We classify them as non-quantum physical RNGs and quantum physical RNGs. Quantum physical RNGs are then classified into non-photonic QRNG and photonic QRNG. Different methods are presented and discussed in this chapter.

Chapter three contains the details of our QRNG based on silicon nanocrystals (Si-NCs) LED[1] and a commercial silicon single photon avalanche diode (Si SPAD). The theory and experimental approach of the methodology are explained and results and discussions are provided at the end of the chapter. Chapter four introduces a robust QRNG[2] based on the arrival times of photons. The source of entropy is Si-NCs LED and the randomness extraction is executed in a field-programmable gate array (FPGA). The results of the statistical tests are demonstrated at the end of this chapter.

The fifth and sixth chapters present a compact configuration for random number generation based on arrival times of photons. In chapter five, the compact configuration consisting of Si-NCs large area LED (LLED)[3] and Si photomultiplier (SiPM)[4] is proposed and studied. The QRNG in chapter six uses Si-NCs LLED as the source of entropy. The method consists of 16 SPADs connected to 4 time to digital converters (TDCs)[5] and operates in oversampling regime.

[1] Fabricated by Advanced Photonics and Photovoltaics (APP) group at the Foundation of Bruno Kessler (FBK)

[2] This methodology was proposed by Giorgio Fontana (`http://www.ing.unitn.it/~fontana/`) and improved through our discussions.

[3] Fabricated by Advanced Photonics and Photovoltaics (APP) group at the Foundation of Bruno Kessler (FBK)

[4] Fabricated by Integrated Radiation and Image Sensors (IRIS) group at the Foundation of Bruno Kessler (FBK)

[5] Designed in Integrated Radiation and Image Sensors (IRIS) group at the Foundation of Bruno Kessler (FBK)

In chapter seven, the importance of an integrated, compact QRNG–with the application in everyday life and accessibility to everyone–is expressed. It consists of an emitter containing 16 pixels (Si SPADs) with p^+/n Si junction and a single pixel with the same p^+/n Si junction as the detector[1]. The same approach as the robust methodology in chapter five is used in this chapter as well to generate random numbers. The results and discussions are included at the end of the chapter. The overall conclusions are provided at the end of the thesis. In all the experiments conducted in the above-mentioned chapters, I prepared the setups and performed the statistical analyses.

[1] Fabricated by Integrated Radiation and Image Sensors (IRIS) group at the Foundation of Bruno Kessler (FBK)

Chapter 2

Physical Random Number Generators

Coin flipping, dice, shuffling playing cards and roulette wheel are among the earliest methods to generate random numbers in games, gambling and for scientific purposes. [16] Today, these early methods are mainly used in games and gambling. They are not suitable for statistical and cryptographic applications since they are very slow phenomena.
The invention of computer revolutionized the way random numbers were generated as well as many other amazing achievements by performing lengthy, complicated processes impossible to be done by humans. Algorithmic methods were developed to generate random numbers from an initial value called "seed". As explained in Chapter 1, the mathematical algorithms produce random numbers that *seem* to be random and are called pseudo random numbers. In order to generate truly random numbers, one needs to prepare a hardware random number generator, i.e. an apparatus that generates random numbers from a physical process. Here we classify hardware (true, physical) random number generators (RNGs) as non-quantum, based on the phenomena explained by classical physics, and quantum RNGs (QRNGs), based on the inherent randomness in quantum phenomena explained by quantum physics.

2.1 Non-quantum physical RNGs

As mentioned before, to overcome the problems of determinacy and predictability associated with arithmetic algorithms, physical phenomena have been taken into account to generate true random numbers. The non-quantum physical random number generators have been produced exploiting different types of noise as the source of entropy.

2.1.1 Thermal noise

Thermal noise is the thermal fluctuations of the voltage of a conductor at equilibrium that is originated from the random motion of charge carriers. [17] It is independent of the amount of current flowing and its intensity varies with temperature. Throughout a finite frequency range (up to microwave frequencies), thermal noise has a nearly Gaussian amplitude distribution (see Appendix A) and the power spectral density is nearly constant. It is also called “Johnson-Nyquist” noise after John B. Johnson and Harry Nyquist who discovered and explained it in 1926. [17]

Thermal noise has been amplified to provide a random voltage source to generate random numbers. [18, 19, 20, 21, 22] The schematic of a RNG based on thermal noise of a resistor is presented in Fig. 2.1. [19] V_{noise} is the thermal noise of a resistor and V_{th} is the threshold of the high speed comparator that is equal to the mean voltage of the input noise signal. The input signal is amplified by a low-noise amplifier and then the analog output noise passes through a comparator. The output of the comparator is sampled and latched to a register. One major application of this type of generator is to provide the seeds for PRNGs.

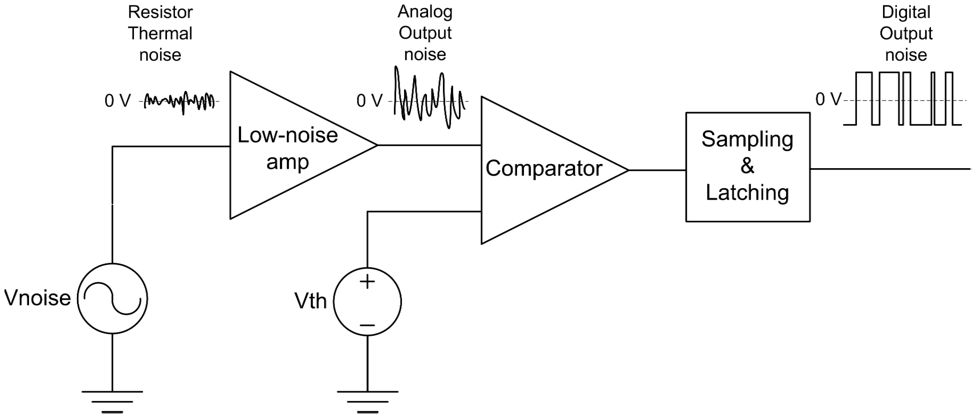


Figure 2.1: Schematic of an RNG based on the noise of a resistor. [19]

Metastability is an unstable equilibrium state in which the logical circuit is not able to settle into a stable ‘0’ or ‘1’ output level for an indeterminate time. The output floats at an intermediate value between ‘0’ and ‘1’ until the system falls into a stable state. [23] In [20, 24], thermal noise and metastability have been used as sources of randomness. In the chip presented in [20], there are 64 switch components of c_0, c_1, c_2,..., c_{63} to form two delay paths in 2^{64} different configurations (since each switch component can take two states of ‘0’ and ‘1’). The two delay paths are excited simultaneously to allow the transitions to race against each other. There is an arbiter block at the end of the delay paths which determines which rising edge arrives first and sets its output to ‘0’ or ‘1’. When the arbiter becomes metastable, it generates random responses.

The proposed generator in [20] is very sensitive to temperature. To minimize the sensitivity to the temperature, eight different RNG circuits, each calibrated at intervals of approximately 10°C, have been used. Therefore, the tolerance of operation was increased to around 80°C. The experiments were conducted in an oven with thermostat control to provide cyclic temperature changes. It is essential to apply post-processing to the original bit streams to eliminate the large difference between the probability of output values of '0' and '1' (this difference is called "bias" that is explained in Section 2.3.6). This results in a reduction of 65-75% of the bit stream and hence a reduction in the efficiency of the RNG. [20]

Even though the electronic chip of the RNG in [20] is easy to be designed and fabricated, there exist some weaknesses including the sensitivity to the temperature, application of post-processing to the raw data and low efficiency. The authors do not estimate or determine the bit rate of the RNG.

In the proposed method in [24], a latch (as the main component of RNG) is tuned into the metastable region in order to minimize the effect of the deterministic noise (external noise, power supply noise, and other non-random events) and to increase the effect of thermal random noise to generate random bits. The metastable operation is controlled by recording the resolution time (defined as the time that it takes the system to resolve from the metastable point to one of the two stable states, either a value of '0' or '1') of each metastable event without observing the value of the generated output. There is a control module which grades the quality of the output bit stream and tunes the system for the maximum probability of randomness.

The RNG in [24] is based on an integrated chip fabricated in a 0.13 μm bulk CMOS technology. The actual bit rate of the RNG is 200 kbps with the theoretical estimation of reaching the maximum value of 50 Mbps. The main drawback of the RNG is the necessity to have a control system to tune the latch into the metastable region to avoid the domination of deterministic noise over the thermal random noise.

In the commercial RNG produced by Intel [25], thermal noise is used to modulate the frequency of a low-frequency oscillator (see Fig. 2.2). The noise-modulated slower oscillator is used to trigger measurements of the fast oscillator and consequently streams of random bits at a rate of 3 Gbps are generated. The RNG takes pairs of 256-bit random samples and applies them to a conditioner which reduces them to a single and more secure 256-bit sample. The secure sample is then used as the seed of a PRNG which makes the numbers ready for use in the Intel RdRand processor instruction. [25]

The problem with the Intel's RNG (and generally with RNGs based on the amplification of thermal noise) is that the noise source is easily influenced by other nearby signals and can be observed and even manipulated by the attacker who would then be able to predict the output or generate his desired output. [26]

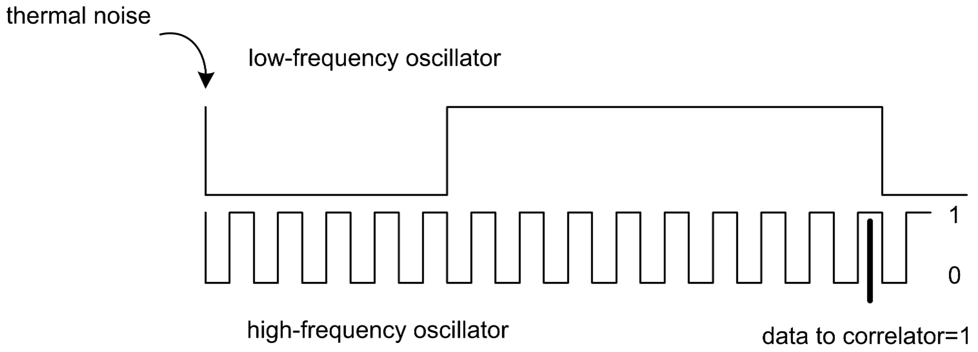


Figure 2.2: Schematic of a dual oscillator in a physical RNG based on frequency jitter in oscillators. [25]

2.1.2 Avalanche noise

Reverse-biased semiconductor p-n junctions, near the breakdown voltage (the voltage beyond which a very small change in voltage results in a sudden, large increase in the current), generate white noise (with Gaussian distribution) at radio frequencies. [27]
The Araneus Alea II [28] uses the avalanche noise of a reverse-biased p-n junction as the source of entropy. The noise is amplified and digitized with an analog-to-digital converter (ADC). The raw data from the ADC are then processed to remove the correlation and bias. Random numbers are generated at the bit rate of 100 kbps. [28]
Chaos Key [29] and TrueRNG [30] use the avalanche noise as the entropy source, as well. In TrueRNG, the noise from the p-n junction (biased at 12 V) is amplified and digitized and then post-processed. The generated random numbers are transmitted through a USB port at the speed of $>$ 350 kbps. [30] Chaos Key sends generated random numbers over a USB port at the speed of approximately 8 Mbps. [29]
As mentioned in the previous section, the problem with RNGs based on the noise amplification is that the noise source can be easily influenced by other nearby signals and can be observed and manipulated by an attacker. [26]

2.1.3 Dark noise

In single photon avalanche diodes (SPADs), which operate at voltages above the breakdown voltage, some pulses are generated even when they are not illuminated. These pulses are called primary and secondary dark counts which are due to thermal generation of carriers in the p-n junction of the SPAD and afterpulses (pulses produced by the recombination of trapped carriers after a dark pulse or a pulse resulted from a photon detection), respectively (see Section 3.2 for more details). [31]
The dark noise of SPAD has been used to generate random numbers in [31]. The random dark pulses are converted into pulses of uniform width transmitted to a PC to register the

digital binary value ‘1’ when the pulse is high and ‘0’ when the pulse is low (0 V). Two excess biases (the excess voltage applied to SPAD above the breakdown) of 5 V and 7 V are considered along with several uniform pulse widths. The percentage of generated ‘0’ is monitored and is found to be close to 50% for the pulse widths of 75 μs and 64 μs at the excess bias of 5 V and 7 V, respectively. [31] Even though the method and electronic parts of the proposed RNG are simple to be implemented, the low efficiency and high dependence on the pulse width (in order to achieve 50% probability of ‘0’ or equivalently ‘1’) are its main drawbacks.

2.1.4 Chaos

Several works have been published based on the chaotic systems as sources of physical randomness. [32, 33, 34, 35, 36, 37]

In [35], the chaos in lasers was used for the first time to achieve efficient and stable generation of random bits at high frequencies. In this device system, there are two semiconductor lasers with chaotic intensity fluctuations. The output intensity of each laser is converted to an alternate current (AC) electrical signal by photodetectors. It is then amplified and converted to a binary signal using an ADC driven by a fast clock. Like the previous methods to generate random numbers based on noise, the converted digital signal is sampled at the rising edge of the clock. Then the binary bit signals obtained from the two lasers are combined by a logical exclusive or (XOR) operation to generate random bit sequences.

At a particular bit rate, some adjustments need to be done to generate random bit sequences with equalized ratio of ‘0’ and ‘1’. Control parameters of the lasers (the injection current, the length of the external cavity and the optical feedback strength) and the threshold levels of the ADC are fixed in order to produce random bit streams that pass the statistical tests. The maximum achievable bit rate of this RNG is 1.7 Gbps corresponding to a clock with a frequency of 1.7 GHz. [35]

The RNG in [37] is based on an integrated chip with two units of chaos generation and post-processing. The core part of the chaos generation unit is a dual-mode amplified feedback laser (a distributed feedback laser with an integrated feedback cavity composed of a phase section and an amplifier section) that can work in the dual-mode state when the two laser modes have comparable threshold gain (by adjusting the DC bias currents of each section of the unit). In the proposed device, there is a fiber-based external optical feedback loop to drive the two lasing modes entering the chaos state. [37]

The chaos signal is then converted into an electronic signal by a broadband photodiode (with 50 GHz bandwidth), is sampled by an oscilloscope at the rate of 160 GS/s and is digitized by an 8-bit ADC. In the post-processing stage, different number of least significant bits (LSBs) are considered and their distributions are monitored. It is seen that by retaining 4 LSBs of each 8-bit sample, a uniform distribution is obtained with a reduced correlation (compared with 8-bit LSBs) generating random numbers at the bit rate of 640 Gbps. [37]

The problem with the proposed RNG by Zhang et al. [37] is that several parameters need to be adjusted (like in the case of the previously-mentioned RNG by Uchida et al. [35]). As mentioned before, the DC bias current to the sections of the dual-mode amplified laser has to be adjusted in order to make the two modes to have comparable threshold gain. The two lasing modes have to be then driven into the chaos state by an external optical feedback loop. The temperature is kept fixed at 25°C by a thermoelectric cooler. [37] All these considerations result in the lack of robustness and security of the RNG.

In [38], spontaneous chaotic oscillations of the current through semiconductor superlattices (SLs) at room temperature are proposed as the source of entropy for RNG. The speed of 6.25 Gbps is achieved using a sampling rate of 1.25 GHz, 4th derivative (discrete time derivative of the digitized current signal) and considering 5 LSBs out of 8 bits. In this implementation the post-processing, derivative and LSB retention were performed in an offline procedure.

A linear combination of the signals of 4 and 6 SL devices results in 40 and 80 Gbps bit rate generation without the use of derivatives. Alternatively, a combination of both methods may be used at 5 GHz sampling rate with 2 SLs, a 3rd order derivative, and retention of 4 LSBs for a generation rate of 20 Gbps. [38] In spite of the high bit rates, the main weak point of this RNG is that a very careful choice of the number of SLs, the order of the derivatives, the number of LSBs retention and the combination of the SL signals (adding or subtracting the signals) is essential in order to minimize the effects of correlation and bias and to pass the statistical tests.

The computer mouse movements have been used as a source of randomness in [39]. The analog signal of the trace of mouse movements is converted to a digital sequence by considering sampled coordinates corresponding to the mouse movements. The value of these coordinates usually ranges from 1 to several thousands. To obtain real values between 0 and 1, every two adjacent points coordinates are considered and an angle is defined as: $arctan\left(|y_{i+1} - y_i| / |x_{i+1} - x_i|\right)$. This angle is then mapped to a real value by dividing by $\pi/2$, so that from n points, n-1 real numbers are achieved.

In order to eliminate the regular patterns in the mouse trace made by the same user, chaotic hash functions[1] are used as the post-processing algorithms. The real numbers (each containing 52 bits in this implementation) are converted into their corresponding bits (x_{bit}) using the formula: $x_{real} = \sum_{bit=1}^{52} x_{bit} \left(1/2\right)^{bit}$. The chaotic signals are then generated by using three chaotic hash functions.

[1] A hash function is a kind of one-way function (a mathematical algorithm) that maps data of arbitrary size to a bit string of a fixed size. Chaotic hash functions are the hash functions integrated with chaotic maps. [40]

The problem with the mouse movements as the source of entropy is that, as mentioned above, there exist similar patterns for the same user. Therefore, careful choice of post-processing algorithms is of high importance to provide very high sensitivity of the output with respect to the input (i.e. even a slight change of the mouse movement should lead to a significant difference in the generated random number.) [39]

2.2 Quantum physical RNGs

Although the methods presented in the previous section are considered good substitutes for the computer-based PRNGs, they are not completely reliable since they are solely based on classical physics. Quantum mechanics, on the other hand, with inherent unpredictability and indeterminacy in quantum phenomena provides the best solution for the generation of non-deterministic, unpredictable, and high quality random numbers.

2.2.1 Non-photonic QRNG

Several non-photonic or non-optical quantum physical properties have been used to generate random numbers such as: the radioactive decay and shot noise.

2.2.1.1 Radioactive decay

In 1956, Isida and Ikeda [41] used the radioactivity to exploit the intrinsic randomness in quantum phenomena and to produce a QRNG. They counted the number of output pulses of a Geiger-Müller (G-M) tube[1] produced by the radioactive decay of Cobalt-60 in a constant time interval. The distribution of these numbers is Poisson (see Appendix. B) with the probability of finding n pulses in a time interval t (seconds) to be $P_n(t) = \frac{(\lambda t)^n}{n!} e^{-\lambda t}$, where λ is the mean number of pulses (in one second).

If the number of pulses is distributed according to a Poisson distribution with a fairly large mean value, the LSBs of the numbers are approximately equally distributed. In the experiment of Isida and Ikeda [41], when the mean value was larger than about 50 the last figures (LSBs) of the sequences of numbers were distributed equally. The autocorrelation of these last figures was calculated and the correlogram was drawn. The LSBs were considered independent without any significant correlation coefficient. Thus, it was considered that the sequences of the LSBs of the numbers would be able to produce random numbers. [41]

Since 1956 different types of QRNGs based on radioactive decay have been proposed. [42,

[1] It is a tube filled with an inert gas such as helium, neon, or argon at low pressure, to which a high voltage is applied. The tube conducts electrical charge when a particle or photon of incident radiation makes the gas conductive by ionization. The ionization is considerably amplified within the tube to produce an easily measured detection pulse, which is fed to the processing and display electronics. [42]

43, 44, 45, 46, 47] The comparison of the time difference between successive pairs of radioactive decays in the G-M detector is used to generate random numbers. [46] If $t_i < t_{i+1}$ a bit with value '1' is generated; otherwise, the generated bit will be '0' (Fig. 2.3). This scheme has been implemented in the so-called HotBits generation hardware which produces random data at the rate of about 800 bps. [46]
In the method proposed in [45], the random number generation is based on the state of a fast clock. When a pulse arrives, if the state of the clock is high, the generator outputs '1' and if it is low, the output is '0'. With a good time resolution, as mentioned before, the LSBs of the digitized bit should be random and post-processing should not be necessary.
In some works in the last decade, G-M detectors have been replaced with semiconductor detectors. [42, 47] The advantage of semiconductor detectors over G-M detectors is that they do not require high voltage. However, their output signal is weaker compared with G-M tubes.
Even though the radioactive decay is a good source of quantum randomness, there are some drawbacks which make the generators based on this quantum phenomenon inconvenient for practical, widespread use. These drawbacks include the need for a radioactive source and the necessity to shield and isolate the QRNG in order to avoid any harm to the users and to prevent the detector from counting the undesirable signals from cosmic rays, radiation from radioactive materials in the Earth's crust, etc. [48]

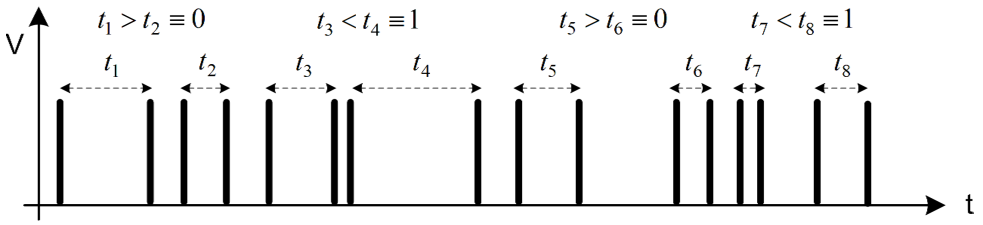


Figure 2.3: The time difference between the two events in the G-M tube is used to generate random bits. If $t_i < t_{i+1}$ a bit value of '1' is generated; otherwise the bit '0' is generated.

2.2.1.2 Shot noise

Shot noise is time-dependent fluctuations in electrical current originating from the discrete nature of electric charge. It is fundamentally non-deterministic, has a Poisson distribution (see Appendix. B) and its amplitude is proportional to the square root of the total current flow and bandwidth. Depending on the conditions it is generated, it may be considered either classical or quantum mechanical or a mixture of both. [49]
Shot noise occurs when charge carriers pass through a potential barrier such as a diode junction in MOS transistors. Shot noise in MOS transistors consists of three major components: sub-threshold leakage, gate leakage and junction leakage. Sub-threshold leakage is the current between the source and drain of a MOS field-effect transistor (MOSFET)

when the transistor is in sub-threshold region or weak-inversion region (where gate-to-source voltages are below the threshold voltage). Gate leakage or gate oxide leakage is due to tunneling across the insulating layer under the gate and the junction leakage is due to high electric fields across reverse-biased p-n junctions. [50]
The ComScire® PQ4000KS [51] is a commercial RNG that uses shot noise due to sub-threshold leakage and gate leakage (quantum phenomena) in MOS transistors along with sources of chaotic entropy (a combination of thermal noise, other types of transistor noise and switching noise) as the source of entropy. Twenty-four independent, high frequency oscillating signal sources continuously operate at different frequencies between 200 and 400 MHz. Each oscillator is sampled separately to produce outputs and the outputs are further combined to produce noisy output signals. Seventy-five of these noisy signals are combined to produce a single sampled binary signal at 128 Mbps. [51]
The PQ4000KS has three independent generators like the one described above. The statistics of each of these three generators is continuously monitored in the generator hardware. The outputs of the three generators are combined to produce one data stream at 128 Mbps, and finally blocks of 32 non-overlapping consecutive bits are XORed together to produce each final output bit at the rate 4 Mbps. [51]
Despite the fact that shot noise is a quantum phenomenon, it is usually not well separated from thermal noise and is affected by environmental noise. The commercial RNG of ComScire® PQ4000KS cannot be considered a purely QRNG since it uses a combination of shot noise and other sources of noise to maximize the entropy (see Section 2.3.5) and enhance the quality of output bits (i.e. to reduce the possible correlation and bias among the random numbers in order to improve the quality of random numbers and pass the statistical tests).

2.2.2 Photonic QRNG

The inherent randomness in quantum mechanical properties of light—as a simple and suitable substitute for other quantum sources of entropy particularly radioactive decay—has been used advantageously to generate random numbers. Optical phenomena, including photons as the quanta of light, have various random, non-deterministic properties. These properties include the polarization of the photons, the exact number of photons illuminating a detector and the photon arrival times. Such intrinsically random properties can be exploited to generate truly random numbers. A classification of these approaches is presented here.

2.2.2.1 Beam splitter and two detectors

A single photon arriving at a beam splitter (BS)—with equal transmissivity and reflectivity T=R=1/2—can take either of two different paths out of the BS (Fig. 2.4). Single photon detection is necessary to detect the single photons transmitted through or reflected from the BS. Depending on which detector clicks, a bit ‘0’ or ‘1’ is produced (we can associate the clicks on detector 1 with bit value ‘0’ and the clicks on detector 2 with bit value ‘1’).

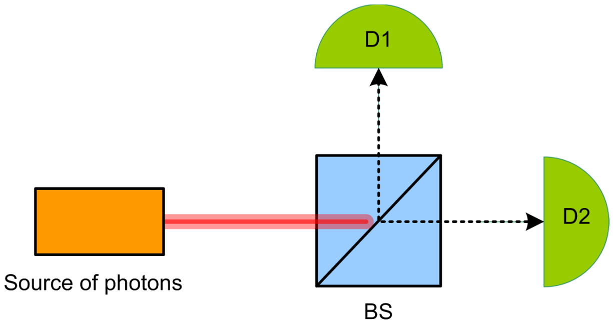


Figure 2.4: Schematic of a beam splitter (BS) and two detectors (D1 and D2). The photon on the balanced BS can take either of the two paths reaching D1 or D2 with equal probability of 0.5. The clicks on D1 and D2 generate the bit values “0’ and “1”, respectively.

A complete implementation of a BS and two photomultiplier tubes (PMTs) was developed as an independent device in 2000. [52] By utilizing either a BS or a polarizing BS, single photon detectors and high speed electronics, binary random numbers were generated at a rate of 1 Mbps. [52] The approach based on a BS and two detectors has been used in several works. [53, 54, 55]
Commercial products, such as the Quantis RNG [53], which are based on a BS and two single photon detectors, generate random numbers at the bit rate of 4 and 16 Mbps after the application of post-processing to the raw data.
Unequal losses, unmatched detection efficiencies and imperfections of the BS and the single photon deterctors affect the random number distribution and would be fatal for the QRNG reliability.

2.2.2.2 Photon counting

Different approaches have been proposed and developed with respect to photon counting. For example, in [56] random bits are generated based on the parity of the total number of counts (of the photons emitted from an LED and detected by a PMT for fast detection) in a fixed period. The intensity of the LED is stabilized and is attenuated to the single photon level. The detected counts during a sampling time interval are interpreted as ‘0’ and ‘1’ for an even and odd number of counts, respectively. After the application of post-processing to reduce the correlation and bias among the bits, the bit rate of 50 Mbps is achieved. [56]

The comparison of the photon numbers in consecutive (attenuated) laser pulses distributed in time [57] is another example of a generator based on photon counting. If there are n_1 photons in a pulse and n_2 photons in the following one, the bit values '1' and '0' are generated if $n_1>n_2$ and $n_1<n_2$, respectively. Random numbers are generated at the rate of 2.4 Mbps.

2.2.2.3 Arrival time of photons

Randomness in photon arrival time has been used in different methods to generate random numbers. [58, 59, 60, 61, 62, 63, 64, 65, 66] One of the first QRNGs that used time detection approach is presented by Stipcevic and Rogina [59] in which the arrival time of photons from an LED at a PMT are compared, in a similar manner to Fig. 2.3. A restartable clock synchronized with the beginning of each random interval is used in this method. Number of pulses of the restartable clock within each random interval is counted and compared to generate random bits. The efficiency of this method is around 0.5 bits per detection. The correlation and bias are eliminated in this approach using the restartable clock and the maximum bit rate of 1 Mbps is achieved. [59]

In [61], a coherent continuous-wave (CW) 1550 nm laser diode, attenuated to the picowatt level, is used as the source of entropy. It is coupled directly into an InGaAs avalanche photodiode (APD) (cooled down to -30°C) via a single mode fiber. The signal from the APD is amplified and sent into time tagging, single photon counting electronics. The time-tagged events acquired are converted into random bits as follows. If a photon is found in an odd clock cycle, the bit value '0' is generated and if it is found in an even cycle, the bit '1' is produced. Sequences of random bits are generated at a rate of 4.01 Mbps. [61]

The implementation in [63] uses a strongly attenuated LED with a PMT for photon detection. For the analysis of the photon arrival times, time tagging electronics is used with a resolution of 1 ps and a throughput of 12.5 Mcps (mega counts per second). In this approach the exponential distribution of the arrival times of photons introduces undesirable bias in the raw data which is removed by post-processing operations. The QRNG generates random numbers at the bit rate of 152 Mbps. [63]

The QRNG scheme of Nie et al. [64] is based on photon arrival time measurement that takes advantage of an external time reference. A SPAD is used to detect photons emitted from a highly attenuated CW laser. The time difference between photon detection and an external time reference is measured as the raw data. Even though the CW laser is highly attenuated, the probability for multi-photon emission is nonzero. It causes bias in the raw data that is reduced by post-processing operation resulting in random bit generation at the maximum rate of 96 Mbps. [64]

In [66] the light emitted by an LED is attenuated to the single photon level, and the intensity of LED is adjusted. To achieve high rates of random number generation, a

PMT with high sensitivity is utilized for detecting the attenuated light. The pulses generated from PMT are amplified and are discriminated by a discriminator module which converts the analog output pulse of the PMT into a digital signal. The discriminator can distinguish two pulses only if two pulses are separated by about the pulse width, otherwise they will overlap and form only one pulse. The output of the discriminator is fed into a time-to-digital-converter (TDC) module which is connected to an FPGA module. In order to increase the min-entropy (see Section 2.3.5) and reduce the correlation and bias, the FPGA selects the period of only one single detection and outputs random numbers corresponding to the arrival time of the selected detection. The bit rate of the random number generation can reach up to 45 Mbps. [66]

The generator developed in [65] comprises a Si-CMOS-LED light source integrated with the Si SPAD on a single chip. An FPGA digitizes the time intervals between random events into bit streams. The maximum bit rate of 1 Mbps is obtained after removing the bias in the raw data (due to the nonuniform distribution of the time intervals) by post-processing in a special configuration of XOR gates. [65]

2.2.2.4 Quantum vacuum fluctuations

Fluctuations of the vacuum state using homodyne detection techniques (in which a weak signal and a strong laser beam, called the local oscillator (LO), interfere on a symmetric BS to form two output beams with balanced powers) has been used to generate random numbers. [67, 68] In [67], a homodyne detector (consisting of the vacuum state as the weak signal and an LO), a bit conversion method and a hash function are used to generate random numbers. The basic source of entropy for random numbers is the vacuum state which is a particle-free state that cannot be influenced by a potential attacker because the vacuum port of the BS is blocked. The LO does not have to be quantum noise limited. Even if it has some excess noise, it will not create any problems since this noise will be rejected in the balanced homodyne measurement scheme. [67]

The quadrature measurement of the vacuum state in the quadrature (or equivalently the position) representation can be written as [67]:

$$|0\rangle = \int_{-\infty}^{\infty} \Psi(x)\,|x\rangle\,dx, \tag{2.1}$$

where $|x\rangle$ are the amplitude quadrature eigenstates ($\langle x|x'\rangle = \delta(x-x')$) and $\Psi(x)$ is the ground-state wavefunction (a Gaussian function centered around $x = 0$). The measurement of the amplitude quadrature collapses the wavefunction into quadrature eigenstates. The outcomes will be unpredictable but biased according to the Gaussian probability function: $|\Psi(x)|^2$. Unbiased numbers can be obtained by binning the measurement outcomes such that the integrated probability associated with each bin is equalized; that is: $\int_{-\infty}^{x_1} |\Psi|^2 dx = \int_{x_1}^{x_2} |\Psi|^2 dx = ... = \int_{x_l}^{\infty} |\Psi|^2 dx$, where $l+1$ is the number of bins. All the measurement outcomes within one bin are assigned a fixed bit combination (see Fig.2.5). The length of this bit combination depends on the number of bins (for $l + 1 = 2^n$ bins, the length of the bit combination is n). In the approach of Gabriel et al. [67], 32 bins (corresponding to 5 bits per sample) are considered. After using the hash functions, the random bits are generated at the rate of ~6.5 Mbps.

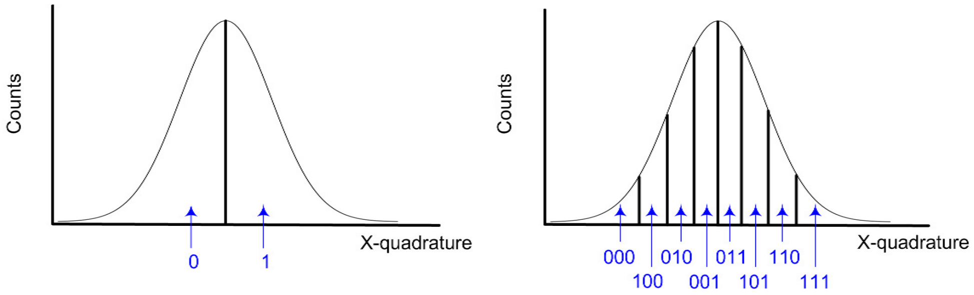


Figure 2.5: The probability distribution of the vacuum state is binned into 2^n equal parts (the same sample size per bin). The random numbers are then produced by assigning a fixed bit combination of length n to each sample point in a certain bin. An example for n=1 (left), n=3 (right) is shown here. [67]

The QRNG based on the fluctuations of the vacuum state using homodyne detection technique in [68] uses an integrated 12 bits, 250 MS/s (mega samples per second) ADC and an FPGA. The results show a uniformly distributed random binary sequence, where 8 bits (keeping only the most significant bits (MSB) as they are the most accurate representation of the vacuum fluctuations) are extracted for each measurement, corresponding to a real-time random bit rate generation of 2 Gbps. [68]

Extraction of random bits from vacuum fluctuations using optical amplification has been demonstrated in [69]. The amplification is phase-insensitive, and the phase of the cavity field remains random. Due to the random phase of the input pulses from the laser diode (LD), the output signals from an unbalanced Mach-Zehnder interferometer (placed after the LD) acquire random amplitudes that are sent to a photodiode. The output of the photodiode is highpass filtered with a cutoff frequency of 40 MHz and it is digitized using the an oscilloscope with input bandwidth of 200 MHz, sampling speed of 2.5 GS/s and a 12-bit ADC. The random bit generation rate of the QRNG reaches 1.11 Gbps. [69]

In spite of the high efficiency and high bit rate of the QRNGs based on quantum vacuum fluctuations, careful conditions should be provided to create vacuum fluctuations and to guarantee the generation of random numbers by vacuum fluctuations and not by deterministic, classical noise.

To have a summary at the end of this section, some of the RNGs proposed in literature or present as commercial products, with their source of entropy and bit rate, are presented in Tables 2.1 and 2.2. From this source, it emerges the need for a photonic QRNG which can be integrated in a single chip.

Table 2.1: A list of some of the RNGs in literature

Author	**Year**	**Source of Randomness**	**Bit Rate**
Tokunaga et al. [24]	2008	Thermal noise	200 kbps
Uchida et al. [35]	2008	Chaos in laser	1.7 Gbps
Li et al. [38]	2013	Chaos in current oscillations	20, 40, 80 Gbps
Zhang et al. [37]	2017	Chaos in laser	640 Gbps
Jennewein et al. [52]	2000	Beam splitter	1 Mbps
Fürst et al. [56]	2010	Photon counting	50 Mbps
Ren et al. [57]	2011	Photon counting	2.4 Mbps
Stipčević and Rogina [59]	2007	Photon arrival time	1 Mbps
Dynes et al. [61]	2008	Photon arrival time	4.01 Mbps
Wahl et al. [63]	2011	Photon arrival time	152 Mbps
Nie et al. [64]	2014	Photon arrival time	96 Mbps
Khanmohammadi et al. [65]	2015	Photon arrival time	1 Mbps
Wang et al. [66]	2015	Photon arrival time	45 Mbps
Gabriel et al. [67]	2010	Quantum vacuum fluctuations	6.5 Mbps
Jofre et al. [69]	2011	Quantum vacuum fluctuations	1.11 Gbps
Symul et al. [68]	2011	Quantum vacuum fluctuations	2 Gbps

Table 2.2: A list of some of the commercial RNGs

Product Name	**Source of Randomness**	**Bit Rate**
IntelRNG [25]	Thermal noise	3 Gbps
Araneus Alea II [28]	Avalanche noise	100 kbps
ChaosKey [29]	Avalanche noise	350 kbps
TrueRNG [30]	Avalnache noise	8 Mbps
HotBits [46]	Radioactive decay	800 bps
ComScire®PQ4000KS [51]	Shot noise	4 Mbps
Quantis [53]	Beam splitter	4, 16 Mbps

2.3 Assessment of RNGs

Assessment of RNGs is very important to evaluate the quality of randomness of the generated bits. Some analyses are described in the following subsections.

2.3.1 Autocorrelation

The first thing to avoid when generating random numbers is the "presence" of correlation. Correlation causes the generation of one bit dependent on another one and, hence, predictable. Autocorrelation is used to check the existence of correlation among the generated bits, codes or symbols with a delayed (lagged) copy of them. The plot of the autocorrelation coefficient ρ_k as a function of the lag k is called the autocorrelation function of $\{\rho_k\}$ of the process. The autocorrelation function is dimensionless, that is, independent of the scale of measurement of the time series. Considering z_t to be the time series, the autocorrelation coefficient between the two values of z_t and z_{t+k}, $k = -K \ \dots \ K$, is calculated as [70]:

$$\rho_k = \frac{\mathbb{E}[(z_t - \mu_t)(z_{t+k} - \mu_{t+k})]}{\sigma_t \sigma_{t+k}} \tag{2.2}$$

where $\mathbb{E}$ is the expected value operator[1], μ_t and μ_{t+k} are the mean values at time t and $t+k$ and σ_t^2 and σ_{t+k}^2 are the variances at time t and $t+k$. This expression is not well-defined for all time series or processes, because the mean may not exist, or the variance may be zero (for a constant process) or infinite (for processes with distribution lacking well-behaved moments, such as certain types of power law). If the function ρ is well-defined, its value must lie in the range [-1, 1], with 1 indicating perfect correlation and -1 indicating perfect anti-correlation. [70]

If z_t is a wide-sense stationary (WSS) process then the mean μ and the variance σ^2 are time-independent, and further the autocorrelation depends only on the lag between t and t+k: the correlation depends only on the time-distance between the pair of values but not on their position in time.

$$\rho_k = \frac{\mathbb{E}[(z_t - \mu)(z_{t+k} - \mu)]}{\sigma^2} \tag{2.3}$$

The autocorrelation function is then an even function with $\rho_{-k} = \rho_k$. The numerator in Eq. 2.3 is called the autocovariance function. If z_t and z_{t+k} are independent, then their autocovariance (and hence autocorrelation) is zero.

[1] The expected value operator over a variable gives the weighted average of all possible values the variable can take on, where each possible value is weighted by its respective probability.

The estimated autocorrelations at time lag k[1] are computed as [71]:

$$\rho_k = \frac{\frac{1}{T-1}\sum_{t=1}^{T-k}(z_t-\mu)(z_{t+k}-\mu)}{\sigma^2} \tag{2.4}$$

where T is the length of the time series. The estimated standard error (se) for the autocorrelation at lag k is:

$$se(\rho_k) = \sqrt{\frac{1}{T}\left(1+2\sum_{j=1}^{q}\rho_j^2\right)} \tag{2.5}$$

where q is the lag beyond which the theoretical autocorrelation function is effectively 0. If the series is completely random, then the standard error reduces to $1/\sqrt{T}$. [71]

2.3.2 Visual analysis

A simple, straightforward way to examine a RNG is to a make a visualisation of the numbers (symbols or codes) it produces. Visualisation, e.g. making a 2-D matrix of the generated codes o symbols, allows spotting any particular, periodic patterns which might exists among the numbers. The 2-D matrix can be created by dividing the vector of random numbers (or a section of it) with k elements into m rows (columns) each containing n elements ($k = m \times n$) and present it in a 2-D graph placing elements of rows (columns) on the x-axis. Since the elements are placed one after the other horizontally and vertically, any pattern existing among them would be observable.

While this type of approach cannot be considered as a formal analysis, it is a quick way to get a rough impression of a given generator's performance. A 512x512, 2-D visualization of the rand() function from PHP on Microsoft Windows, which is a PRNG, can be seen in Fig. 2.6. A clear, periodic pattern is visible in the figure.

2.3.3 Joint probability mass function

In probability theory, considering at least two random variables Z_1 and Z_2 that are defined on a probability space, the joint probability mass function (JPMF) gives the probability that each of Z_1 and Z_2 falls into a discrete set of values specified for that variable. It is computed as [73]:

$$\begin{aligned} &\mathrm{P}(Z_1 = z_1 \text{ and } Z_2 = z_2) \\ &= \mathrm{P}(Z_1 = z_1 | Z_2 = z_2)\mathrm{P}(Z_2 = z_2) \\ &= \mathrm{P}(Z_2 = z_2 | Z_1 = z_1)\mathrm{P}(Z_1 = z_1) \end{aligned} \tag{2.6}$$

[1] Autocorrelations used in Matlab program and used in the following chapters for the autocorrelation of the random sequences

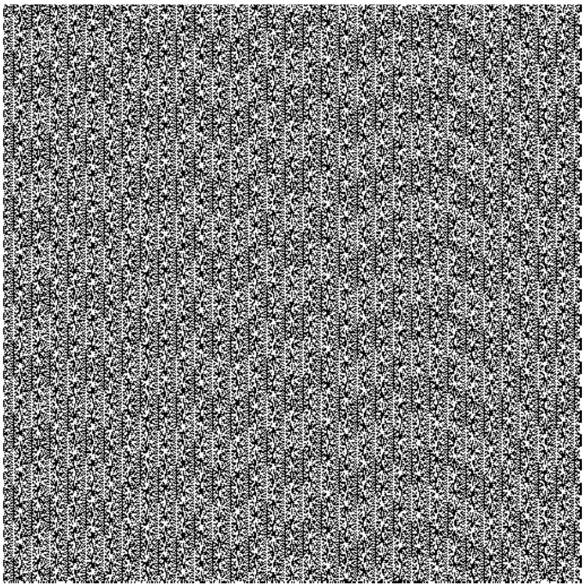


Figure 2.6: A 512x512, 2-D visualization of the rand() function from PHP on Microsoft Windows. The figure is taken from [72].

where $\mathrm{P}(Z_1 = z_1|Z_2 = z_2)$ and $\mathrm{P}(Z_2 = z_2|Z_1 = z_1)$ are conditional probabilities that give the probability of Z_1=z_1 and Z_2=z_2 given that Z_2=z_2 and Z_1=z_1, respectively. In the case of n discrete random variables Z_1 ... Z_n, the joint probability distribution is:

$$\begin{aligned}
&\mathrm{P}(Z_1 = z_1, \ldots, Z_n = z_n) = \prod_{i=1}^{n} \mathrm{P}(Z_i = z_i| \bigcap_{j=1}^{i-1} Z_j = z_j) \\
&= \mathrm{P}(Z_n = z_n|Z_{n-1} = z_{n-1}, \ldots, Z_1 = z_1) \ldots \qquad (2.7) \\
&\cdot \mathrm{P}(Z_3 = z_3|Z_2 = z_2, Z_1 = z_1) \\
&\cdot \mathrm{P}(Z_2 = z_2|Z_1 = z_1) \cdot \mathrm{P}(Z_1 = z_1)
\end{aligned}$$

This is called the chain rule of probability. The sum of all JPMFs over all random variables gives 1. If the discrete random variables are independent, the JPMF will be the product of the marginal probabilities. Therefore, in the case of 2 discrete random variables we can write:

$$\mathrm{P}(Z_1 = z_1 \text{ and } Z_2 = z_2) = \mathrm{P}(Z_1 = z_1) \cdot \mathrm{P}(Z_2 = z_2) \qquad (2.8)$$

If we consider a time series $\mathrm{z_t}$ of 16 hexadecimal symbols $\{1, 2, \ldots, E, F\}$ (as in Chapters 4, 5 and 7), to calculate the JPMF of having any one symbol after the other one for any two variables z_1 and $z_2 \in \{0, 1, \ldots, E, F\}$, one can count the number of times that $\mathrm{z_2}$ occurred after $\mathrm{z_1}$ and dividing by the total number of all possibilities of a pair of symbols, obtains the JPMF for z_1 and z_2.

2.3.4 Mutual information

In probability theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables. More specifically, it quantifies the "amount of information" (in units such as bits) obtained about one random variable, through the other random variable. It is computed by the formula below [74]:

$$\mathrm{I}(Z_1; Z_2) = \sum_{z_1 \in Z_1} \sum_{z_2 \in Z_2} \mathrm{P}(z_1, z_2) \ \log \left(\frac{\mathrm{P}(z_1, z_2)}{\mathrm{P}(z_1)\mathrm{P}(z_2)} \right) \tag{2.9}$$

where $\mathrm{P}(z_1, z_2)$ is the joint probability mass function of random variables Z_1 and Z_2, and $\mathrm{P}(z_1)$ and $\mathrm{P}(z_2)$ are the marginal probability functions of Z_1 and Z_2, respectively.
$\mathrm{I}(Z_1; Z_2) = 0$ if and only if Z_1 and Z_2 are independent random variables. If Z_1 and Z_2 are independent, then $\mathrm{P}(z_1, z_2) = \mathrm{P}(z_1) \cdot \mathrm{P}(z_2)$, and therefore:

$$\log \left(\frac{\mathrm{P}(z_1, z_2)}{\mathrm{P}(z_1)\,\mathrm{P}(z_2)} \right) = \log 1 = 0 \tag{2.10}$$

MI is symmetric, i.e. $\mathrm{I}(Z_1; Z_2) = \mathrm{I}(Z_2; Z_1)$, and is nonnegative, i.e. $\mathrm{I}(Z_1; Z_2) \geq 0$.
In the case of a time series z_t of 16 hexadecimal symbols, the MI among all pairs of symbols is computed as:

$$\mathrm{I} = \sum_{i=0}^{F} \sum_{j=0}^{F} \ \mathrm{P}(i, j) \ \log \left(\frac{\mathrm{P}(i, j)}{\mathrm{P}(i)\mathrm{P}(j)} \right) \tag{2.11}$$

where i and j are hexadecimal symbols $\in \{0, 1, \ldots, E, F\}$.

2.3.5 Entropy

In an ideal random sequence each bit is unpredictable and unbiased. The probability of observing each bit with a particular value is unaffected by the knowledge of the values of all the other bits. Entropy is intuitively defined as a measure of uncertainty. An ideal random sequence of n bits contains n bits of entropy. In information theory, the Rényi entropy of order β is defined as [75]:

$$\mathrm{H}_\beta(Z) = \frac{1}{1 - \beta} \log_b \left(\sum_{i=1}^{\bar{\bar{Z}}} \mathrm{P}^\beta(Z = z_i) \right) \tag{2.12}$$

where $\beta \geq 0$ and $\beta \neq 1$ and Z is a discrete random variable with size $\bar{\bar{Z}} = n$. If b is equal to 2, Euler's number e, and 10, and the unit of entropy is called Shannon, nat, and Hartley, respectively. When $b = 2$, the unit of entropy is referred to as bits.
Depending on the value of β the following entropies are formulated as:

- if $\beta \to 0$, the measure of entropy is called the Hartley entropy or max-entropy [76]:

  $$\mathrm{H} \equiv \mathrm{H}_0 = \log_b(\bar{\bar{Z}}) \tag{2.13}$$

  which is equivalent to the case of a uniform probability distribution where the probabilities are $\mathrm{P}(Z = z_i) = 1/\bar{\bar{Z}}$ for all $z_i = z_1, \ldots, z_n$.

- the limit of $\beta \to 1$, gives the Shannon entropy [77]:

  $$\mathrm{H} \equiv \mathrm{H}_1 = -\sum_{i=1}^{\bar{\bar{Z}}} \mathrm{P}(Z = z_i) \log_b \mathrm{P}(Z = z_i) \tag{2.14}$$

- if $\beta = 2$ the Rényi entropy is known as the collision entropy [78]:

  $$\mathrm{H} \equiv \mathrm{H}_2 = -\log \sum_{i=1}^{\bar{\bar{Z}}} \mathrm{P}^2(Z = z_i) \tag{2.15}$$

  where $\mathrm{P}^2(Z = z_i)$ is called the "collision probability of a random variable Z" since if we let $Z^{'}$ be another random variable with identical associated probability distribution as Z but independent of it, the probability of Z and $Z^{'}$ colliding, i.e. of yielding the same value, is equal to the expression:

  $$\sum_{i=1}^{\bar{\bar{Z}}} \mathrm{P}(Z = z_i, Z^{'} = z_i) = \sum_{i=1}^{\bar{\bar{Z}}} \mathrm{P}(Z = z_i)\ \mathrm{P}(Z^{'} = z_i) = \sum_{i=1}^{\bar{\bar{Z}}} \mathrm{P}^2(Z = z_i) \tag{2.16}$$

  where we used the independence of Z and $Z^{'}$ first and then the fact that the associated probability distributions are the same.

- in the case of $\beta \to \infty$ the Rényi entropy H_β converges to the min-entropy [78, 79]:

  $$\mathrm{H} \equiv \mathrm{H}_\infty = -\log\ \mathrm{P} = -\log \max(p_{z_i}) \tag{2.17}$$

  where p_{z_i} is a given finite probability distribution for a random variable Z.

The various Rényi entropies are all equal for a uniform distribution, but they measure the unpredictability of a nonuniform distribution in different ways. Min-entropy is defined as the lower bound on the entropy of a random variable. It is often used as a worst-case measure of the unpredictability of observations z since it is the negative logarithm of the probability of the most likely outcome. In this sense, it is the strongest way to measure the information content of a discrete random variable. If z_i has the min-entropy m, then the probability of observing any particular value is no greater than 2^{-m}. [80]

2.3.6 Bias

In statistics, bias is a feature of a process or its results where the obtained value of a a parameter differs from its expected value. It is an undesirable effect in random processes causing a departure from the theoretical expected value. For instance, in the case of an "unfair" dice, the probability of the output for each of the six values of {1, 2, 3, 4, 5, 6} would not be exactly 1/6. The difference between the obtained value and the expected, theoretical value defines the bias and the maximum difference is called the "maximum bias". As the bias increases, the process gets farther and farther from a random process. If we have a sequence of 0 and 1 binary bits, the maximum bias is computed as the absolute value of the maximum departure from 1/2. In the case of a sequence of hexadecimal symbols $\{0, 1, \ldots, E, F\}$, the absolute value of maximum deviation from the expected, theoretical probability value of 1/16 gives the maximum bias. Acceptable values of bias are $< 10^{-4}$ in order to pass the statistical tests particularly the frequency test (explained in Section 2.3.7.1.1).

2.3.7 Statistical tests

Various statistical tests can be applied to a sequence to attempt to compare and evaluate the sequence to a truly random sequence. Randomness is a probabilistic property; that is, the properties of a random sequence can be characterized and described in terms of probability. The likely outcome of statistical tests, when applied to a truly random sequence, is known a priori and can be described in probabilistic terms. There is an infinite number of possible statistical tests, each assessing the presence or absence of a "pattern"' which, if detected, would indicate that the sequence is nonrandom. Two typical tests suites of the National Institute of Standards and Technology (NIST) and TestU01 are explained in the following. Since the NIST package contains tests to examine and certify RNGs used in cryptographic applications and the Alphabit battery in TestU01 has been defined to test physical RNGs, we will use these two suites in the thesis.

2.3.7.1 NIST tests suite

The NIST test suite is a statistical package consisting of 15 tests that were developed to test the randomness of (arbitrarily long) binary sequences produced by random number generators. These tests focus on a variety of different types of non-randomness that could exist in a sequence. Some tests are decomposable into a variety of subtests. The 15 tests are [81]:

1. The Frequency (Monobit) Test,
2. Frequency Test within a Block,

3. The Runs Test,
4. Test for the Longest-Run-of-Ones in a Block,
5. The Binary Matrix Rank Test,
6. The Discrete Fourier Transform (Spectral) Test,
7. The Non-overlapping Template Matching Test,
8. The Overlapping Template Matching Test,
9. Maurer's "Universal Statistical" Test,
10. The Linear Complexity Test,
11. The Serial Test,
12. The Approximate Entropy Test,
13. The Cumulative Sums (Cusums) Test,
14. The Random Excursions Test, and
15. The Random Excursions Variant Test

A number of tests in the test suite have the standard normal and the chi-square (χ^2) as reference distributions. If the sequence under test is in fact non-random, the calculated test statistic will fall in extreme regions of the reference distribution. The standard normal distribution (i.e., Gaussian function) is used to compare the value of the test statistic obtained from the RNG with the expected value of the statistic under the assumption of randomness. The test statistic for the standard normal distribution is of the form $z = (x - \mu)/\sigma$, where x is the sample test statistic value, and μ and σ^2 are the expected value and the variance of the test statistic. The χ^2 distribution (i.e., a left skewed curve) is used to compare the goodness-of-fit of the observed frequencies of a sample measure to the corresponding expected frequencies of the hypothesized distribution. The test statistic is of the form $\chi^2 = \sum \left(\frac{(O_i - E_i)^2}{E_i}\right)$, where $\mathrm{O_i}$ and $\mathrm{E_i}$ are the observed and expected frequencies of occurrence of the measure, respectively. [81]

In the following, short descriptions on the 15 statistical tests in NIST tests suite are presented. The order of the application of the tests in the NIST tests suite is arbitrary. However, it is recommended that the Frequency test be run first, since this supplies the most basic evidence for the existence of non-randomness in a sequence, specifically, non-uniformity. If this test fails, the likelihood of other tests failing is high. In the following descriptions, $\bar{\bar{Z}}$ is the length of the sequence of bits ($z = z_1, z_2, \ldots, z_n$) generated by the

RNG. The *P-value* is defined as the probability, under the assumption of a hypothesis (e.g. the hypothesis that a sequence is random), of obtaining a result equal to or more extreme than what is actually observed. For all the tests, if the computed *P-value* is < 0.01, we conclude that the sequence is non-random. Otherwise, we conclude that the sequence is random.

2.3.7.1.1 Frequency test

The Frequency test monitors the proportion of zeroes and ones for a sequence. The purpose of this test is to determine whether the number of ones and zeros in a sequence are approximately the same as would be expected for a truly random sequence. The test assesses the closeness of the fraction of ones to 1/2. All subsequent tests depend on the passing of this test.
The 0 and 1 bits in the sequence are converted to -1 and +1, respectively. The test is executed as follows.

1. X is produced as:

$$X = (2z_1 - 1) + (2z_2 - 1) + \cdots + (2z_n - 1)$$

2. The test statistic (s) is then computed as:

$$s = \frac{|X|}{\sqrt{\bar{\bar{Z}}}}$$

3. The *P-value* is computed as:

$$P - value = \boldsymbol{erfc}\left(\frac{s}{\sqrt{2}}\right)$$

   where ***erfc*** is the complementary error function defined as:

$$\boldsymbol{erfc}(y) = \frac{2}{\sqrt{\pi}} \int_y^{\infty} e^{-u^2} du$$

If the computed*P-value* is < 0.01 (at the 1% level), then conclude that the sequence is non-random. Otherwise, conclude that the sequence is random.
The small *P-value* (< 0.01) would be caused by X or s being large. Large positive values of X are indicative of too many ones, and large negative values of X are indicative of too many zeros

2.3.7.1.2 Frequency test within a block

The purpose of this test is to determine whether the frequency of ones in an M-bit block is approximately M/2, as would be expected under an assumption of randomness. For block size M=1, this test is the same as Frequency (Monobit) test. Considering M to be the block size, the test runs as follows.

1. The sequence is partitioned into $n = \left[\frac{\bar{\bar{N}}}{M}\right]$ non-overlapping blocks with $\left[\frac{\bar{\bar{N}}}{M}\right]$ to be the floor function of $\frac{\bar{\bar{N}}}{M}$ (that gives the largest integer less than or equal to $\frac{\bar{\bar{N}}}{M}$). The unused bits are discarded.

2. The proportion of ones is determined in each block as:

$$\pi_i = \frac{\sum_{k=1}^{M} z_{(i-1)M+k}}{M} \qquad \text{for} \;\; 1 \leq i \leq n$$

3. The test statistic ϱ is computed as:

$$\varrho = 4M \sum_{i=1}^{n} \left(\pi_i - \frac{1}{2}\right)^2$$

4. The *P-value* is computed as:

$$P - value = \boldsymbol{igamc}\left(\frac{n}{2}, \frac{\varrho}{2}\right)$$

   where ***igamc*** is the incomplete gamma function defined as:

$$\boldsymbol{igamc}\ (a, b) = \frac{\Gamma(a, b)}{\Gamma(a)} = \frac{1}{\Gamma(a)} \int_{b}^{\infty} e^{-t} t^{a-1} dt$$

Small P-values (< 0.01) would indicate a large deviation from the equal proportion of ones and zeros in at least one of the blocks.

2.3.7.1.3 Runs test

The purpose of the Runs test is to determine whether the number of "runs" of ones and zeros of various lengths is as expected for a random sequence. A run is an uninterrupted sequence of identical bits. A run of length k consists of exactly k identical bits and is bounded before and after with a bit of the opposite value. The Runs test carries out a Frequency test as a prerequisite (if not initially executed). The test steps are as follows.

1. The pre-test proportion π of ones is computed as:

$$\pi = \frac{\sum_i z_i}{\bar{\bar{Z}}}$$

2. It is determined if the prerequisite Frequency test is passed (if it is not already executed). If $\left|\pi - \frac{1}{2}\right| \geq \frac{2}{\sqrt{\bar{\bar{Z}}}}$, it indicates the failure of the Frequency test and therefore the Runs test is not applicable setting the *P-value* to 0.0000. Otherwise, the Runs test would be applicable and the next step is performed.

3. The test statistic R is computed as:

$$R = \left(\sum_{j=1}^{\bar{\bar{Z}}-1} r(j) \right) + 1$$

   where $r(j) = 0$ if $z_j = z_{j+1}$ and $r(j) = 1$ if $z_j \neq z_{j+1}$.

4. The *P-value* is computed as:

$$P - value = \boldsymbol{erfc} \left(\frac{\left| R - 2\bar{\bar{Z}}\pi(1-\pi) \right|}{2\sqrt{2\bar{\bar{Z}}}\pi(1-\pi)} \right)$$

If the computed *P-value* is < 0.01, then one can conclude that the sequence is non-random. Otherwise, it is concluded that the sequence is random. A large value for R would have indicated an oscillation in the string which is too fast; a small value would have indicated that the oscillation is too slow. An oscillation is considered to be a change from a one to a zero or vice versa. A fast oscillation occurs when there are a lot of changes, e.g., 010101010 oscillates with every bit. A stream with a slow oscillation has fewer runs than would be expected in a random sequence, e.g., a sequence containing 20 ones, followed by 65 zeroes, followed by 115 ones (a total of 200 bits) would have only three runs, whereas 100 runs would be expected.

2.3.7.1.4 Test for the longest run of ones in a block

The purpose of this test is to determine whether the length of the longest run of ones (or equivalently zeros) within the tested sequence is consistent with the length of the longest run of ones that would be expected in a random sequence. The test is executed as follows.

1. The sequence is divided into M-bit blocks (M being preset as M=8, M=128 and M=10^4).

2. The frequencies $\mathrm{f_i}$ of the longest runs of ones in each block are put into categories. For instance, if we consider two blocks of 11100110 and 00101001 in a sequence, the maximum runs of ones are 3 and 1, respectively. Then the number of each maximum runs of ones is counted as $\mathrm{f_0}$, $\mathrm{f_1}$, ...; For M=8, $\mathrm{f_0}$ is the frequency for maximum runs of$\leq$1 and $\mathrm{f_3}$ is the last frequency for maximum runs $\geq$4. For M=128, $\mathrm{f_0}$ is the

frequency for maximum runs of$\leq$4 and f_5 is the last frequency for maximum runs $\geq$9. And for M=10^4, f_0 is the frequency for maximum runs of$\leq$10 and f_6 is the last frequency for maximum runs $\geq$16.

3. The test statistics is computed then as:

$$\chi^2 = \sum_{i=0}^{K} \frac{(f_i - N\pi_i)^2}{N\pi_i}$$

with K and N determined by the value of M as:

- M=8, K=3 and N=16
- M=128, K=5 and N=49
- M=10^4, K=6 and N=75.

π_i is the probability defined for each value of K and M.

4. The *P-value* is computed as:

$$\textit{P-value} = \boldsymbol{igamc}\left(\frac{K}{2}, \frac{\chi^2}{2}\right)$$

If the computed *P-value* is < 0.01, then conclude that the sequence is non-random. Otherwise, conclude that the sequence is random. Large values of χ^2 indicate that the tested sequence has clusters of ones.

2.3.7.1.5 Binary matrix rank test

The purpose of this test is to check for linear dependence among fixed length substrings of the original sequence. The test is run as:

1. The sequence is divided into M.Q-bit (M being the number of rows and Q the number of columns) disjoint blocks. There will exist $N = \left[\frac{\bar{\bar{\bar{Z}}}}{MQ}\right]$ such blocks. Discarded bits will be reported as not being used in the computation within each block. The rows of the matrix are filled with successive Q-bit blocks of the original sequence *z*.

2. The binary rank (R_l l=1,...,N) of each matrix is determined. It is done as follows. The rank of each matrix would be the number of rows it contains. If some rows contain the same binary strings, they are counted only once.

3. The number of matrices with full rank (F_M), full rank minus one ($F_{M\text{-}1}$) and the remaining matrices (N-F_M-$F_{M\text{-}1}$) are counted.

4. The test statistics χ^2 is then computed as:

$$\chi^2 = \frac{(F_M - 0.2888N)^2}{0.2888N} + \frac{(F_{M-1} - 0.5776N)^2}{0.5776N} + \frac{(N - F_M - F_{M-1} - 0.1336N)}{0.1336N}$$

5. *P-value* is computed as:

$$P\text{-}value = e^{-\chi^2/2}$$

If the computed *P-value* is < 0.01, we conclude that the sequence is non-random. Otherwise, we conclude that the sequence is random. Large values of χ^2 (and hence, small *P-values*) would indicate a deviation of the rank distribution from that corresponding to a random sequence.

2.3.7.1.6 Discrete Fourier transform (spectral) test

The purpose of this test is to detect periodic patterns in the tested sequence that would indicate a deviation from the assumption of randomness. The intention is to detect whether the number of peaks exceeding the 95% threshold is significantly different than 5%. The steps taken in this test are:

1. The zeros and ones of the input sequence (z) are converted to values of –1 and +1, respectively, to create the sequence $x = x_1, x_2, ..., x_n$, where $x_i = 2z_i$—1.

2. Discrete Fourier transform is applied on x creating a sequence of complex variables which represents periodic components of the sequence of bits at different frequencies: s=DFT(x).

3. Modulus of s' which is the half substring of s is calculated m=modulus(s'). m is a sequence of peak heights.

4. $T = \sqrt{\left(\log \frac{1}{0.05}\right) n}$=the 95% peak height threshold value. Under an assumption of randomness, 95% of the values obtained from the test should not exceed T.

5. the expected theoretical (95%) number of peaks (under the assumption of randomness) that are less than T is computed as: N_0=0.95n/2.

6. The actual number of peaks in m that are less than T is computed.

7. The normalized difference between the observed and the expected number of frequency components that are beyond the 95% threshold is computed as:

$$d = \frac{(N_1 - N_0)}{\sqrt{n(0.95)(0.05)/4}}$$

8. *P-value* is computed as:
$$P\text{-}value = \boldsymbol{erfc}\left(\frac{|d|}{\sqrt{2}}\right)$$

A d value that is too low would indicate that there are too few peaks (< 95%) below T, and too many peaks (more than 5%) above T.

2.3.7.1.7 Non-overlapping template matching test

The purpose of this test is to detect if too many occurrences of a given non-periodic (aperiodic) pattern are produced by a generator. The test is executed as follows.

1. The sequence is partitioned into N independent blocks of length M.

2. The number of times that B (the template) occurs within the block j is counted as n_j j=1, ..., N. The search for matches proceeds by creating an m-bit (m being the length of the template) window on the sequence, comparing the bits within that window against the template. If there is no match, the window slides over one bit , e.g., if m=4 and the current window contains bits 4 to 7, then the next window will contain bits 5 to 8. If there is a match, the window slides over m bits, e.g., if the current (successful) window contains bits 4 to 7, then the next window will contain bits 8 to 11.

3. Under an assumption of randomness, the theoretical mean μ and variance σ^2 are computed as:
$$M = \frac{(M - m + 1)}{2^m} \qquad \sigma^2 = M\left(\frac{1}{2^m} - \frac{2m-1}{2^{2m}}\right)$$

4. The test statistic which measures how well the observed number of template "hits" matches the expected number of template "hits" under an assumption of randomness, is computed as:
$$\chi^2 = \sum_{j=1}^{N} \frac{(n_j - \mu)^2}{\sigma^2}$$

5. The *P-value* is computed as:
$$P\text{-}value = \boldsymbol{igamc}\left(\frac{N}{2}, \frac{\chi^2}{2}\right)$$

   Multiple *P-values* will be computed, i.e., One*P-value* will be computed for each template. For m=9 for example, up to 148 *P-values* may be computed; for m = 10, up to 284 *P-values* may be computed.

If the computed *P-value* is < 0.01, it shows that the sequence has irregular occurrences of the possible template patterns and we can conclude that the sequence is non-random.

2.3.7.1.8 Overlapping template matching test

This test is meant to check the number of occurrences of pre-specified target strings. Both this test and the non-overlapping template matching test use an m-bit window to search for a specific m-bit pattern. The difference between them is that in this test when the pattern is found, the window slides only one bit before resuming the search. The test steps in order are as follows.

1. The sequence is partitioned into N independent blocks of length M.

2. The number of occurrences of B in each of the N blocks is calculated. The search for matches proceeds by creating an m-bit window on the sequence, comparing the bits within that window against B and incrementing a counter when there is a match. The window slides over one bit after each examination, record the number of occurrences of B in each block by incrementing an array ν_i (where i = 0, ..., 5), such that ν_0 is incremented when there are no occurrences of B in a substring, ν_1 is incremented for one occurrence of B, ... and ν_5 is incremented for 5 or more occurrences of B.

3. The values for λ and η are computed to be used for the theoretical probabilities π_i corresponding to the ν_i:

$$\lambda = \frac{(M - m + 1)}{2^m} \qquad \eta = \frac{\lambda}{2}$$

4. χ^2 is computed as:

$$\chi^2 = \sum_{i=0}^{5} \frac{(\nu_i - N\pi_i)^2}{N\pi_i}$$

   where $\pi_0 = 0.364091$, $\pi_1 = 0.185659$, $\pi_2 = 0.139381$, $\pi_3 = 0.100571$, $\pi_4 = 0.070432$ and $\pi_5 = 0.139865$ for 5 degrees of freedom. [82]

5. The *P-value* is then computed as:

$$P\text{-}value = \boldsymbol{igamc}\left(\frac{5}{2}, \frac{\chi^2}{2}\right)$$

2.3.7.1.9 Maurer's "universal statistical" test

The purpose of this test is to detect whether the sequence can be significantly compressed without loss of information. A significantly compressible sequence is considered to be non-random. The test is run as follows.

1. The sequence (z) is partitioned into two segments: an initialization segment consisting of Q L-bit non-overlapping blocks, and a test segment consisting of K L-bit non-overlapping blocks. Bits remaining at the end of the sequence that do not form a complete L-bit block are discarded.

2. A table is created for each possible L-bit value in the segment partition. The block number of the last occurrence of each L-bit block is put in the table (i.e., For i from 1 to Q, $\mathrm{T_j}$= i, where j is the decimal representation of the contents of the $\mathrm{k^{th}}$ L-bit block).

3. Each of the K blocks in the test segment is examined and the number of blocks since the last occurrence of the same L-bit block is determined. The values in the table are replaced with the location of the current block (i.e., $\mathrm{T_j}$ = i). The calculated distance between re-occurrences of the same L-bit block (in segment partition) is added to an accumulating log2 sum of all the differences detected in the K blocks (i.e., $sum = sum + \log 2(i^{'} – i)$), $i^{'}$ being the number of block in the test segment.

4. The test statistic, which is the sum of the number of digits in the distance between L-bit templates with the reference distribution of half-normal distribution (a one-sided variant of the normal distribution) as in the case for the Frequency test, is computed as:
$$s = \frac{1}{K} \sum_{i^{'}=Q+1}^{Q+K} \log 2(i^{'} - i)$$

5. *P-value* is computed as:
$$\textit{P-value} = \boldsymbol{erfc}\left(\left|\frac{s- < L >}{\sqrt{2}\sigma}\right|\right)$$
where $< L >$ is the expected value of L that is precomputed together with σ. [2]

If s differs significantly from $< L >$, then the sequence is significantly compressible.

2.3.7.1.10 Linear complexity test

The purpose of this test is to determine if the sequence is complex enough to be considered random. Random sequences are characterized by longer linear feedback shift registers (LFSRs). An LFSR that is too short implies non-randomness. The following steps are taken to execute the test.

1. The sequence is partitioned into N independent blocks of M bits, where n = MN.

2. The linear complexity of each of the N blocks ($\mathrm{L_i}$ i= 1, ..., N) is determined by the Berlekamp-Massey algorithm. [2] $\mathrm{L_i}$ is the length of the shortest LFSR sequence that generates all bits in the block i. Some combination of the bits within any $\mathrm{L_i}$-bit sequence, when added together modulo 2, produces the next bit in the sequence (bit $L_i + 1$).

3. Under the assumption of randomness, the theoretical mean μ is calculated:

$$\mu = \frac{M}{2} + \frac{(9 + (-1)^{M+1})}{36} - \frac{\left(\frac{M}{3} + \frac{2}{9}\right)}{2^M}$$

4. For each substring, a value of $\mathrm{T_i}$ is calculated as:

$$T_i = (-1)^M (L_i - M) + \frac{2}{9}$$

5. The $\mathrm{T_i}$ values are recorded in $\nu_0, \ldots, \nu_6$ as follows:

$$\begin{array}{lrl} \text{if:} & T_i \leq -2.5 & \text{increment } \nu_0 \text{ by } 1 \\ & -2.5 < T_i \leq -1.5 & \text{increment } \nu_1 \text{ by } 1 \\ & -1.5 < T_i \leq -0.5 & \text{increment } \nu_2 \text{ by } 1 \\ & -0.5 < T_i \leq 0.5 & \text{increment } \nu_3 \text{ by } 1 \\ & 0.5 < T_i \leq 1.5 & \text{increment } \nu_4 \text{ by } 1 \\ & 1.5 < T_i \leq 2.5 & \text{increment } \nu_5 \text{ by } 1 \\ & T_i > 2.5 & \text{increment } \nu_6 \text{ by } 1 \end{array}$$

6. The test statistic which is a measure of how well the observed number of occurrences of fixed length LFSRs matches the expected number of occurrences under an assumption of randomness is computed as:

$$\chi^2 = \sum_{i=0}^{K} \frac{(\nu_i - N\pi_i)^2}{N\pi_i}$$

where $\pi_0 = 0.010417$, $\pi_1 = 0.03125$, $\pi_2 = 0.125$, $\pi_3 = 0.5$,$\pi_4 = 0.25$, $\pi_5 = 0.0625$ and $\pi_6 = 0.020833$.

7. *P-value* is computed as:

$$\textit{P-value} = \boldsymbol{igamc}\left(\frac{K}{2}, \frac{\chi^2}{2}\right)$$

The computed *P-value* are < 0.01 would indicate that the observed frequency counts of $\mathrm{T_i}$ stored in the ν_i bins varied from the expected values. It is expected that the distribution of the frequency of the $\mathrm{T_i}$ (in the ν_i bins) should be proportional to the computed π_i.

2.3.7.1.11 Serial test

The purpose of this test is to determine whether the number of occurrences of the 2^{m} m-bit overlapping patterns is approximately the same as would be expected for a random sequence. Random sequences have uniformity that means every m-bit pattern has the same chance of appearing as every other m-bit pattern. For m=1, the Serial test is equivalent to the Frequency test. The test is executed as follows.

1. The z' sequence is formed by appending the first m-1 bits to the end of the sequence z.

2. The frequency of all possible overlapping m-bit blocks, all possible overlapping (m-1)-bit blocks and all possible overlapping (m-2)-bit blocks are determined. Let $\nu_{i_1 \ldots i_m}$ denote the frequency of the m-bit pattern $i_1 \ldots i_m$, $\nu_{i_1 \ldots i_{m-1}}$ the frequency of the (m-1)-bit pattern $i_1 \ldots i_m$ and $\nu_{i_1 \ldots i_{m-2}}$ denote the frequency of the (m-2)-bit pattern $i_1 \ldots i_{m-2}$.

3. The following variables are computed:

$$\begin{aligned}
\Psi_m^2 &= \frac{2^m}{n} \sum_{i_1 \ldots i_m} \left(\nu_{i_1 \ldots i_m} - \frac{n}{2^m}\right)^2 = \frac{2^m}{n} \sum_{i_1 \ldots i_m} \nu_{i_1 \ldots i_m}^2 - n \\
\Psi_{m-1}^2 &= \frac{2^{m-1}}{n} \sum_{i_1 \ldots i_{m-1}} \left(\nu_{i_1 \ldots i_{m-1}} - \frac{n}{2^{m-1}}\right)^2 = \frac{2^{m-1}}{n} \sum_{i_1 \ldots i_{m-1}} \nu_{i_1 \ldots i_{m-1}}^2 - n \\
\Psi_{m-2}^2 &= \frac{2^{m-2}}{n} \sum_{i_1 \ldots i_{m-2}} \left(\nu_{i_1 \ldots i_{m-2}} - \frac{n}{2^{m-2}}\right)^2 = \frac{2^{m-2}}{n} \sum_{i_1 \ldots i_{m-2}} \nu_{i_1 \ldots i_{m-2}}^2 - n
\end{aligned}$$

4. The test statistics which are a measure of how well the observed frequencies of m-bit patterns match the expected frequencies of the m-bit patterns is computed as:

$$\begin{aligned}
\nabla \Psi_m^2 &= \Psi_m^2 - \Psi_{m-1}^2 \\
\nabla^2 \Psi_m^2 &= \Psi_m^2 - 2\Psi_{m-1}^2 + \Psi_{m-2}^2
\end{aligned}$$

5. The *P-values* are computed as:

$$\begin{aligned}
\textit{P-value}_1 &= \boldsymbol{igamc}\left(2^{m-2}, \nabla \Psi_m^2\right) \\
\textit{P-value}_2 &= \boldsymbol{igamc}\left(2^{m-3}, \nabla^2 \Psi_m^2\right)
\end{aligned}$$

If $\nabla \Psi_m^2$ or $\nabla^2 \Psi_m^2$ are large then non-uniformity of the m-bit blocks is implied.

2.3.7.1.12 Approximate entropy test

The purpose of the test is to compare the frequency of overlapping blocks of two consecutive lengths (m and m+1) against the expected result for a random sequence. The test is run as follows.

1. n overlapping m-bit sequences are augmented by appending m-1 bits from the beginning of the sequence to the end of the sequence.

2. A frequency count is made of the n overlapping blocks (e.g., if a block containing z_j to $z_{j+m\text{-}1}$ is examined at time j, then the block containing z_{j+1} to z_{j+m} is examined at time j+1). Let the count of the possible m-bit values be represented as C_i^m , where i is the m-bit value.

3. $C_i^m = \frac{\#(i)}{\bar{Z}}$ for each value of i.

4. $\phi^{(m)}$ is calculated as:
$$\phi^{(m)} = \sum_{i=0}^{2^m-1} \pi_i log_{\pi_i}$$
where $\pi_i = C_k^m$ and k is the corresponding m-bit binary to the decimal value of i.

5. Steps 1-4 are repeated for m+1 instead of m.

6. The test statistic which is a measure of how well the observed value of ApEn(m) (approximate entropy (m)) matches the expected value is computed as:
$$\chi^2 = 2n(log2 - ApEn(m)) \qquad ApEn(m) = \phi^{(m)} - \phi^{m+1}$$

7. *P-value* is computed as:
$$P\text{-}value = \boldsymbol{igamc}\left(2^{m-1}, \frac{\chi^2}{2}\right)$$

Small values of ApEn(m) would imply strong regularity and large values would imply substantial fluctuation or irregularity.

2.3.7.1.13 Cumulative sums (cusums) test

The purpose of the test is to determine whether the cumulative sum of the partial sequences occurring in the tested sequence is too large or too small relative to the expected behavior of that cumulative sum for random sequences. This cumulative sum may be considered as a random walk. For a random sequence, the excursions of the random walk should be near zero. For certain types of non-random sequences, the excursions of this random walk from zero will be large. The test steps are as follows.

1. A normalized sequence is formed by replacing the zeros with -1 to have a sequence of -1 and $+1$ using $X_i = 2z_i$–1.

2. Partial sums S_i of successively larger subsequences are computed starting with X_1 (if mode = 0) or X_n (if mode = 1). Mode is a switch for applying the test either forward through the input sequence (mode = 0) or backward through the sequence (mode = 1). That is, $S_k = S_{k-1} + X_k$ for mode 0, and $S_k = S_{k-1} + X_{n-k+1}$ for mode 1.

3. The test statistic r which is the largest excursion from the origin of the cumulative sums in the corresponding (-1, +1) sequence with the reference distribution being the normal distribution is computed as:

   $$r = max_{1 \leq k \leq n} |S_k|$$

   where r is the largest of the absolute values of the partial sums S_k.

4. The *P-value* is computed as:

   $$\begin{aligned} P\text{-}value = 1 - &\sum_{k=\left(-\frac{\bar{\bar{\bar{Z}}}}{r}+1\right)/4}^{k=\left(\frac{\bar{\bar{\bar{Z}}}}{r}-1\right)/4} \left[\phi\left(\frac{(4k+1)r}{\sqrt{\bar{\bar{\bar{Z}}}}}\right) - \phi\left(\frac{(4k-1)r}{\sqrt{\bar{\bar{\bar{Z}}}}}\right)\right] \\ + &\sum_{k=\left(-\frac{\bar{\bar{\bar{Z}}}}{r}-1\right)/4}^{k=\left(\frac{\bar{\bar{\bar{Z}}}}{r}-3\right)/4} \left[\phi\left(\frac{(4k+3)r}{\sqrt{\bar{\bar{\bar{Z}}}}}\right) - \phi\left(\frac{(4k+1)r}{\sqrt{\bar{\bar{\bar{Z}}}}}\right)\right] \end{aligned}$$

When mode=0 (1), large values of this statistic indicate that there are either “too many ones” or “too many zeros” at the beginning stages of the sequence (at the late stages). Small values of the statistic would indicate that ones and zeros are intermixed too evenly.

2.3.7.1.14 Random excursions test

The purpose of this test is to determine if the number of visits to a particular state within a cycle (a sequence of steps of unit length taken at random that begin at and return to the origin) deviates from what one would expect for a random sequence. This test is a series of eight tests, one test and conclusion for each of the states: -4, -3, -2, -1 and +1, +2, +3, +4. The test steps are taken as follows.

1. A normalized sequence X_i is formed by replacing 0 with -1.

2. The partial sums S_i is computed each starting with X_1 such that $S_1 = X_1$, $S_2 = X_1 + X_2$, ..., $S_n = X_1 + \cdots + X_n$.

3. A new sequence of $S^{'}$ is formed by putting one 0 at the beginning and one 0 at the end such that: $S^{'} = 0, S_1, \ldots, S_n$.

4. The total number of zero crossings q (a value of zero in $S^{'}$ that occurs after the starting zero) in $S^{'}$ is counted. q is also the number of cycles in $S^{'}$. A cycle of $S^{'}$ is defined as a subsequence of $S^{'}$ consisting of a zero, followed by no-zero values, and ending with another zero. The ending zero in one cycle may be the beginning zero in another cycle. The number of cycles in $S^{'}$ is the number of zero crossings. If $q < 500$, discontinue the test.

5. The frequency of the state values of $x = -4, -3, -2, -1$ and $x = 1, 2, 3, 4$ within each cycle is computed.

6. For each eight states of x, $\nu_k(x)$ is computed that is the total number of cycles in which state x occurs exactly k times among all cycles. For instance, for $k = 5$, all frequencies ≥ 5 are stored in $\nu_5(x)$ with $\sum_{k=0}^{5} \nu_k(x) = q$.

7. For each of the eight states of x, the test statistic ,which is a measure of how well the observed number of state visits within a cycle match the expected number of state visits within a cycle, under an assumption of randomness, is computed as:

$$\chi^2 = \sum_{k=0}^{5} \frac{(\nu_k(x) - q\pi_k(x))^2}{q\pi_k(x)}$$

   where $\pi_k(x)$ is the probability that the state x occurs k times in a random distribution.

8. For each state of x, a *P-value* is computed as:

$$\textit{P-value} = \boldsymbol{igamc}\left(\frac{5}{2}, \frac{\chi^2}{2}\right)$$

If χ^2 is too large, then the sequence would display a deviation from the theoretical distribution for a given state across all cycles.

2.3.7.1.15 Random excursions variant test

The purpose of this test is to detect deviations from the expected number of visits to various states in the random walk. This test is a series of eighteen tests, one test and conclusion for each of the states: -9, -8, ..., -1 and +1, +2, ..., +9. The test is executed as follows.

1. This step is the same as the first step in the random excursions test (Section 2.3.7.1.16).

2. This step is also the same as the second step in the random excursions test (Section 2.3.7.1.16).

3. This step is the same as the third step in the random excursions test (Section 2.3.7.1.16).

4. For each of the eighteen non-zero states of x, $\xi(x)$ is computed that is the total number of times that state x occurred across all q cycles. The reference distribution for the test statistic is the half normal (for large $\bar{\bar{Z}}$). If ξ is distributed as normal, then $|\xi|$ is distributed as half normal. If the sequence is random, then the test statistic will be about 0. If there are too many ones or too many zeros, then the test statistic will be large.

5. For each $\xi(x)$ the *P-value* is computed as:

$$\textit{P-value} = \boldsymbol{erfc}\left(\frac{|\xi(x) - q|}{2q\,(4|x| - 2)}\right)$$

2.3.7.1.16 Test results interpretation

The final analysis file of the NIST tests suite contains the results of two forms of analysis:

- proportion of sequences passing a test
- uniform distribution of *P-values*

In the event that either of these approaches fails (i.e., the corresponding null hypothesis must be rejected), additional numerical experiments should be conducted on different samples of the generator to determine whether the phenomenon was a statistical anomaly or a clear evidence of non-randomness.

Given the empirical results for a particular statistical test, the proportion of sequences that pass can be computed. First the number of binary sequences with *P-value* > 0.01 is counted and is divided by the total number of sequences to give the proportion. Then, the range of the acceptable proportions is determined using the confidence interval (CI) defined as:

$$CI = \hat{p} \pm 3\ \frac{\sqrt{\hat{p}(1-\hat{p})}}{m}$$

where $\hat{p} = 1 - \alpha$ and m is the sample size (the number of binary sequences). To pass the test, the proportion should lie above the minimum limit of CI. The confidence interval is calculated using a normal distribution as an approximation to the binomial distribution, which is reasonably accurate for large sample sizes $m \geq 1000$.

The distribution of *P-values* is examined to ensure uniformity. The interval between 0 and 1 is divided into 10 sub-intervals, and the *P-values* that lie within each sub-interval are counted. The uniformity is determined via an application of a χ^2 test and the determination of a *P-value* corresponding to the goodness of fit distributional test on the *P-values* obtained for an arbitrary statistical test. This is accomplised by computing:

$$\chi^2 = \sum_{i=1}^{10} \frac{\left(w_i - \frac{m}{10}\right)}{\frac{m}{10}}$$

where w_i is the number of *P-values* in subinterval i and m is the sample size. Then, a *P-value*$_T$ is computed which is the *P-value* of the *P-values*:

$$\textit{P-value}_T = \boldsymbol{igamc}\left(\frac{9}{2}, \frac{\chi^2}{2}\right).$$

If *P-value*$_T \geq 0.0001$, then the sequences can be considered to be uniformly distributed. To provided statistically meaningful results at least 55 sequences must be processed.

2.3.7.2 TestU01 suite

TestU01 is a software library, implemented in the American National Standards Institute (ANSI) C language, and offering a collection of utilities for the empirical statistical testing of uniform random number generators. It provides general implementations of the classical statistical tests for random number generators, as well as several others proposed in the literature, and some original ones. These tests can be applied to the generators predefined in the library and to user-defined generators. The batteries Alphabit and Rabbit can be applied on a binary file considered as a source of random bits. Alphabit has been defined to test physical (hardware) random bits generators. When invoking the battery, one must specify the number of bits (n_B) available for each test. The Alphabit apply 17 different statistical tests of Multinomial Bits Over, Hamming tests and Random Walk. [83] The multinomial Bits Over test is similar to overlapping template matching test in NIST tests suite. Hamming tests and Random Walk, however, provide more stringent testing for random sequences. TestU01 is widely considered as the most comprehensive and stringent battery of tests. [84] It is most suitable for fast RNGs since sequences of at least 1 G-bit length are required to be provided for the tests in the Alphabit and Rabbit batteries.

2.3.7.2.1 Multinomial bits over

The purpose of this test is to determine uniformity between successive bits in overlapping blocks of length L=2, 4, 8 and 16. Let $k = 2^L$ be the number of cells, for L. Each cell number is generated by taking L successive bits from the sequence with overlap. This test is similar to overlapping template matching test described in Section 2.3.7.1.8.

2.3.7.2.2 Hamming tests

Two tests are analyzed here: Hamming correlation and Hamming independence. Hamming correlation applies a correlation test on the Hamming weights of successive blocks of L bits. n blocks of L (fixed as L=32 in the tests suite) bits. Let X_j be the Hamming weight (the numbers of bits equal to 1) of the j$^{\text{th}}$ block, for j = 1,..., n. The test computes the empirical correlation between the successive X_j's as:

$$\hat{\rho} = \frac{4}{(n-1)L} \sum_{j=1}^{n-1} \left(X_j - \frac{L}{2} \right) \left(X_{j+1} - \frac{L}{2} \right)$$

Under H_0 (the null hypothesis under which the random variables are assumed to be uniformly distributed in the set of all possible sequences), as $n \to \infty$, $\hat{\rho}\sqrt{n-1}$ has asymptotically the standard normal distribution. This is what is used in the test. The test is valid only for large n.

Hamming independence applies a test of independence between the Hamming weights of successive blocks of L (fixed as L=16 in the tests suite) bits. 2n blocks of L bits are built. Let X_j be the Hamming weight (the numbers of bits equal to 1) of the j$^{\text{th}}$ block, for j = 1, ..., 2n. Each vector (X_i, X_{i+1}) can take $(L+1) \times (L+1)$ possible values. The test counts the number of occurrences of each possibility among the non-overlapping pairs $\{(X_{2j-1}, X_{2j}\}$, $1 \leq j \leq n$, and compares these observations with the expected numbers under H_0 through a chi-square test.

2.3.7.2.3 Random walk

From a bit sequence of length $\bar{\bar{Z}}$, a random walk is defined as $X_k = \sum_{j=1}^{k}(2z_j - 1)$ for $k > 0$ with $X_0 = 0$, like in the case of random excursions explained in Section 2.3.7.1.16. In the two tests of random walk in Alphabit battery of testU01 with parameters $L_0 = 64$, $L_1 = 64$ and $L_0 = 320$, $L_1 = 320$ (which define blocks of 64 and 320 bits), five *P-values* are computed for each test on the test statistics of H, M, J, R and c. H is the number of steps to the right, M the maximum value reached by the walk, J the fraction of time spent on the right of the origin, R the number of returns to 0 and C the number of sign changes. They are computed as:

$$
\begin{aligned}
&H = \frac{\bar{\bar{z}}}{2} + \frac{X_{\bar{\bar{z}}}}{2} \\
&M = maxX_k \quad 0 \leq k \leq \bar{\bar{z}} \\
&J = 2\sum_{k=1}^{\bar{\bar{z}}/2} \mathbb{I}[X_{2k-1} > 0] \\
&R = \sum_{k=1}^{\bar{\bar{z}}} \mathbb{I}[X_k = 0] \\
&C = \sum_{k=3}^{\bar{\bar{z}}} \mathbb{I}[X_{k-2}X_k < 0]
\end{aligned}
$$

where $\mathbb{I}$ is the indicator function. The empirical distributions of the test statistics are then compared with the corresponding theoretical ones via a chi-square test. [85]

Chapter 3

Silicon Nanocrystals LED for Quantum Random Number Generator

In this chapter, we first describe the formation of silicon nanocrystals (Si-NCs) via plasma enhanced chemical vapor deposition (PECVD). Then, the fabrication and electrical and optical properties of the Si-NCs LED are explained. At the end of the chapter, an approach to generate quantum random numbers exploiting light from the Si-NCs LED is introduced and studied in details.

3.1 Si-NCs LED

Si is an indirect band-gap semiconductor and is inefficient to emit light in the visible range of electromagnetic radiation. [86] It can emit light in the near infrared due to weak band-to-band emission at the energy of the band-gap (1.12 eV). Thanks to quantum confinement effect based on the Heisenberg uncertainty principle ($\Delta x\ \Delta p \geq \hbar/2$), Si-NCs or Si quantum dots (confined in 3 dimensions) can emit light in the visible regime. This phenomenon is due to the increased overlap of the electron and hole wavefunctions in k (reciprocal) space when their size in real space decreases. [86]

Si-NCs are CMOS compatible, they can be easily incorporated in integrated configurations, they emit photons with wavelengths in the spectral range detectable by Si detectors allowing the fabrication of an all-Si-based device and since the spontaneous emission of photons in a Si-NCs LED is a non-deterministic, quantum mechanical and random process, they can be used as a quantum source of randomness to generate random numbers. Si-NCs are formed through several methods such as laser ablation, ion implantation, decomposition of silane, and PECVD. The latter is the method used here[1] to grow the Si-NCs in a silica matrix. In this method, deposited layers of sub-stoichiometric silica (SiO_x) with large excess of Si (called Si rich oxide (SRO) hereafter) are exposed to the precursor gases of silane (SiH_4), nitrous oxide (N_2O) and amonia (NH_3) and are annealed at high temperature to grow Si-NCs in the SiO_2 matrix with controlled excess of both Si and N.

By band-gap engineering as described in [87], graded-size multilayer Si-NCs LED are fabricated with the structure presented in Fig. 3.1 (a) and (b). The substrate (back contact) is a p-type Si (anode) and the top contact is an n-type polycrytalline Si (cathode). The active area of the graded-size multilayer Si-NCs LED is composed of 6 periods of SRO/SiO_2 films with the nominal thickness of (4/2)-(3/2)-(2/2)-(2/2)-(3/2)-(4/2) nm annealed at 1150°C for 30 min and 1000°C for 60 min to grow Si-NCs by PECVD. The main advantage of the graded-size structure is that the larger NCs close to the contacts make the injection of carriers easier to the active area and the smaller NCs in the center make the emission more efficient. [87]

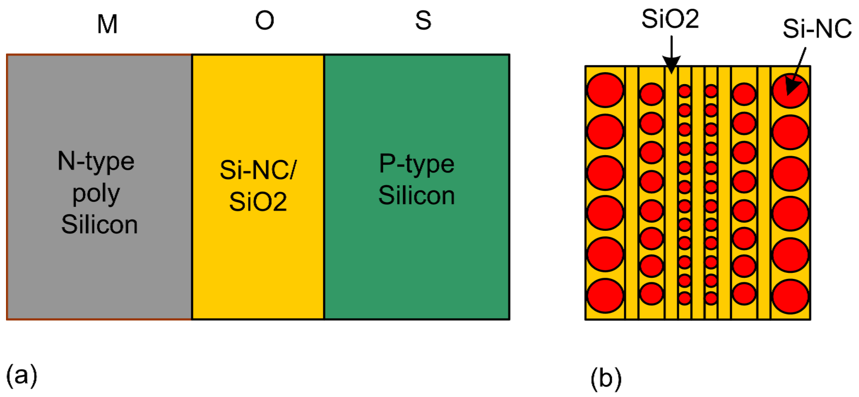


Figure 3.1: (a) The metal-oxide-semiconductor (MOS) structure of the Si-NCs LED. The n-type poly Si and the p-type Si (substrate) layers act as the cathode and anode, respectively. (b) The graded-size structure of the active layer of Si-NCs LED composed of 6 periods of SRO/SiO_2 films with the nominal thickness of (4/2)-(3/2)-(2/2)-(2/2)-(3/2)-(4/2) nm.

Carrier injection into a dielectric like SiO_2, with band-gap energy of 9 eV, occurs only through tunneling. Depending on the voltage applied between cathode and anode (ΔV), there are two different tunneling mechanisms, direct and Fowler-Nordheim (FN) tunneling. As can be seen in Fig. 3.2, direct tunneling is the dominant tunneling mechanism when ΔV is low so that both electrons (e) and holes (h) can tunnel through the oxide layer through a trapezoidal area. At high ΔV, however, FN tunneling dominates and

[1] Fabricated by Advanced Photonics and Photovoltaics (APP) group at the Foundation of Bruno Kessler (FBk)

the tunneling occurs via hot *e* through a triangular area. The probability of *h* tunneling through the oxide layer is low since the potential barrier is higher for *h* than in the case of direct tunneling (when ΔV is lower). The *e* and *h* injected ("bipolar" injection) into Si-NCs create excitons which then recombine radiatively emitting photons through spontaneous emission mechanism. [86]
In FN tunneling, on the other hand, hot *e* tunnel through the oxide (the probability of *h* tunneling is low since the potential barrier for *h* to tunnel is higher here than in the case of direct tunneling), so the final state of the *e* will be the conduction band of the oxide. When this *e* relaxes on the conduction band of the Si-NCs, its excess energy gives rise to an $e - h$ pair (impact ionization) which may recombine radiatively to produce a photon (resulting from "unipolar" injection of carries). The tunneling of hot carriers (with high energy) causes degradation and damage to the oxide layer, and hence low efficiency and short durability of the device. [88]

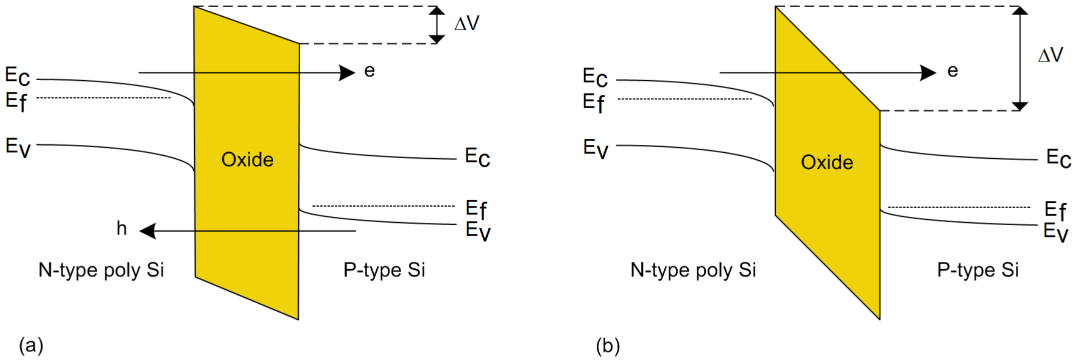


Figure 3.2: (a) Direct tunneling of electrons (*e*) and holes (*h*)("bipolar" injection) through the oxide layer when the voltage difference (ΔV) between the cathode (n-type poly Si) and anode (p-type Si) is low. (b) Fowler-Nordheim (FN) tunneling of hot electrons (e) ("unipolar" injection) through the oxide layer when ΔV is high.

The power efficiency as a function of current density for the graded-size multilayer Si-NCs LED is presented in Fig. 3.3. The actual thickness of the active area is 20.0±0.2 nm and the emitting (gate) area is ~3.2×10^{-3} cm^2. Two distinct regions are observed, low and high current density regions, when power efficiency decreases slowly and a rapidly, respectively. The intersection of linear fits of the two regions in Fig. 3.3 indicates the applied current density of ~0.3 mA/cm^2 or ΔV~2.9 V (corresponding to the applied electric field of 1.45 MV/cm at the forward bias of -2.9 V applied to the cathode while keeping the anode at the ground potential) to the Si-NCs LED. This voltage corresponds to the energy barrier height (band offset) for *e* at the Si/SiO$_2$ interface. The *e* barrier height controls the onset of FN tunneling as it is schematically shown in Fig. 3.2. Above this voltage, the FN tunneling dominates while below it, the direct tunneling is the dominant carrier injection mechanism. Fig. 3.3 shows that the FN regime yields lower efficiency than the direct tunneling regime. The direct tunneling is the dominant charge transport mechanism in LEDs with thin <2.6 nm oxide layers. [89]

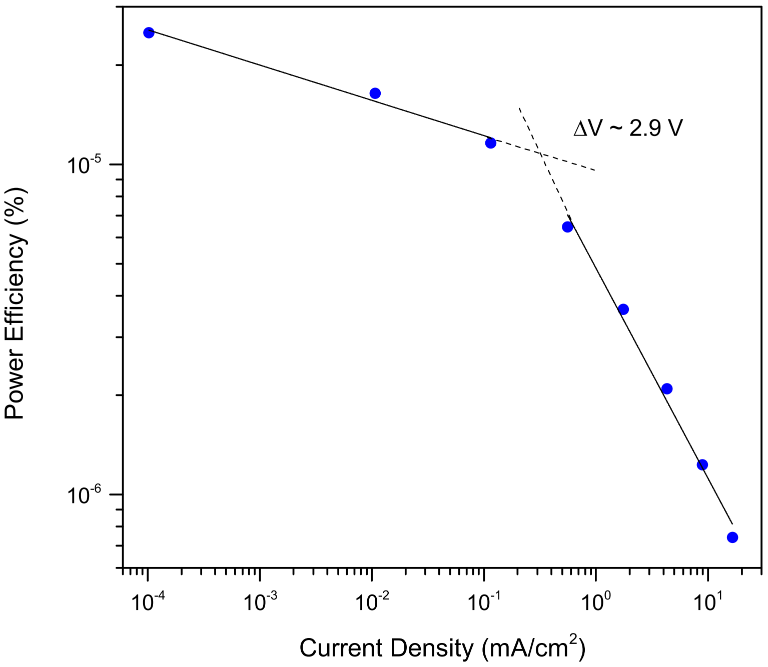


Figure 3.3: Current density versus applied electric field for the graded-size multilayer Si-NCs LED. The intersection of the linear fits of the two regions indicates the applied electric field of ~1.45 MV/cm (equivalently ΔV ~2.9 V) which separates dominant bipolar injection via direct tunneling from dominant unipolar injection through the FN tunneling.

The current density-applied voltage characteristic of the graded-size multilayer Si-NCs LED can be seen in Fig. 3.4. A quite rectifying behavior is observed. At low forward and reverse bias, a hysteresis loop is found in both wafers which extends -1.4 V under forward bias. The hysteresis originates from the charge accumulation within the SRO layer. Injected positive charges from the substrate are accumulated under negative forward bias. This causes a built-in potential that adds to the external bias. This phenomenon causes the weak increase of current in the hysteresis region. Similarly, the hysteresis loop in reverse bias region is due to the accumulation of the negative charges.

The Electrolumnescence (EL) spectra of graded-size multilayer Si-NCs at different applied currents are shown in Fig. 3.5. They were obtained by a Spectra-Pro 2300i monochromator coupled with a nitrogen-cooled charge coupled device (CCD) camera. A broad distribution of wavelengths is observed around the peak at ~ 800 nm attributed to the emission from Si-NCs. The graded-size multilayer Si-NCs LED has some main advantages over the single layer Si-NCs LED such as high density of Si-NCs, more uniformity in Si-NCs sizes, lower turn-on voltages, higher current density at low-applied electric fields, and

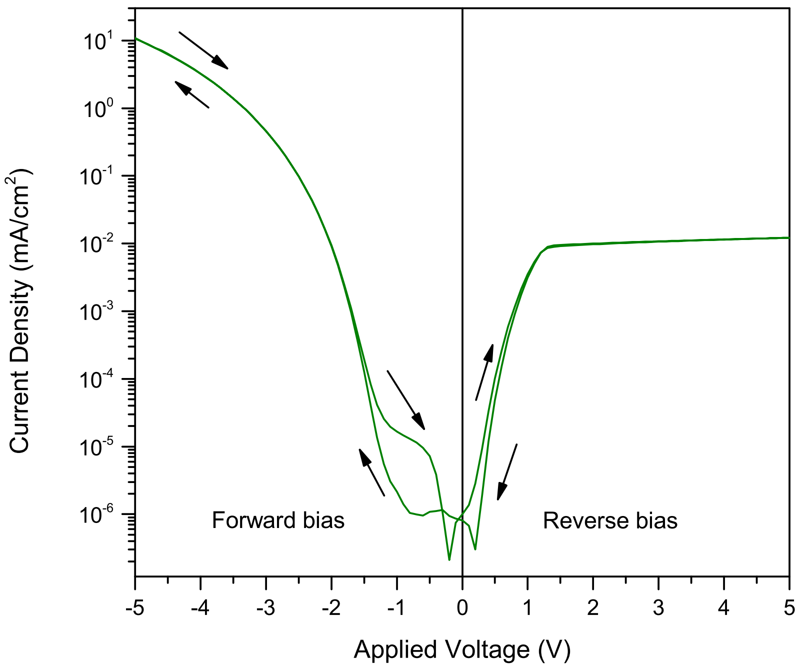


Figure 3.4: The current density-applied voltage characteristic of the graded-size multilayer Si-NCs LED with step-like SRO layers of 4-3-2-2-3-4 nm thickness sandwiched between layers of SiO_2 with the thickness of 2 nm.

higher power efficiency at low applied currents. [90]

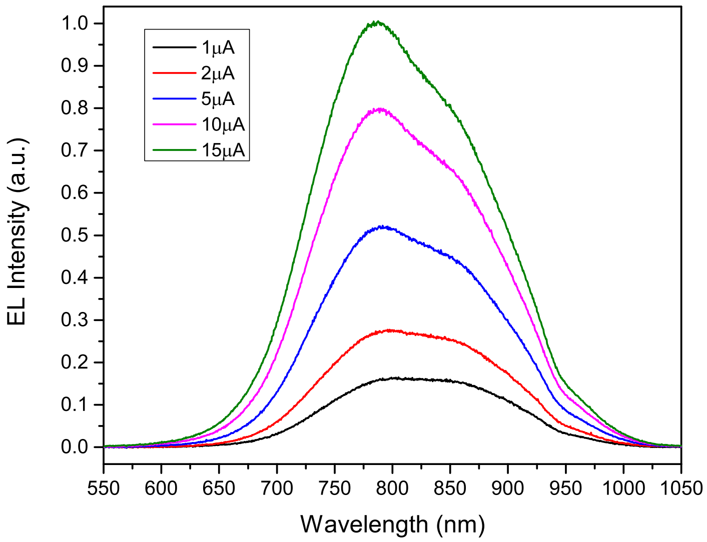


Figure 3.5: Electroluminescence (EL) spectra of the graded-size multilayer Si-NCs LED at different applied currents of 1μA, 2μA, 5μA, 10μA, and 15μA.

3.2 Single Photon Avalanche Diode

A single photon avalanche diode (SPAD) is a reverse-biased p-n junction where a carrier, generated by photon absorption, triggers an avalanche current through impact ionization. Compared with avalanche photodiodes (APDs), SPADs are biased far above the breakdown voltage (the reverse voltage above which a small increase in voltage results in an exponential increase in the leakage current through the diode). When a carrier is generated, ionization occurs and continues until the diode either destroys or a circuit returns the diode to the reverse region. Due to the similar operation of SPAD to a Geiger counter, SPADs are also known as Geiger-mode avalanche photodiode. SPADs have been implemented into CMOS technologies and have helped greatly to reduce the cost of the CMOS-based devices. [91]

As mentioned before, after an avalanche, the SPAD is returned to the condition with the bias voltage above the breakdown voltage in order to be ready to be ignited again. The process through which the SPAD is returned to the initial condition (with the bias voltage above the breakdown voltage) to detect the next photon is called *quenching* and the period of time this process takes is called *dead time* during which the SPAD is blind. The quenching is done in two different ways, *passive* and *active*. Passive quenching is done by a resistor placed in series with the SPAD. The avalanche current is quenched due to a voltage drop on a high ballast resistor of 10 kΩ or more. [92]

Active quenching is used to avoid the slow recovery from avalanche pulses and to exploit fully the inherent performance of the SPAD. [92] In this type of quenching, rise of the avalanche pulse is sensed by a fast discriminator through a 50 Ω resistor and reset transitions in short times providing a digital output pulse (synchronous with the photon arrival time with picosecond time jitter explained in the following). If the SPAD is exposed to a very high photon count rate ($\geq$ 1/dead time), it avalanches immediately the moment the bias voltage is restored (after an avalanche) and hence its count rate depart from a linear relationship with detected light level indicating the *saturation* of the SPAD. [93] Light emission (with a broad spectral distribution) from the avalanche region of the Si SPAD when a coming photon is detected, is a peculiar property of this type of detector. In the measurements with another detector monitoring light, or if the optical system is such that light emitted from the Si SPAD is reflected back on itself, this undesired emitted light can cause some confusion. [91]

The photon detection probability of the SPAD at a particular wavelength λ is the probability that the $e-h$ pairs, generated by absorbed photons (with wavelength λ) in the depletion region, trigger an avalanche. [91] The efficiency of photon detection increases with excess bias voltage (the amount of bias voltage above the breakdown voltage of the SPAD), since a higher electric field enhances the triggering probability. [92]

Even in the absence of light, thermally generated carriers may trigger an avalanche in the depletion region of the SPAD generating some counts known as *dark counts*. The dark count rate (DCR) increases with excess bias voltage. [92] Traps energy levels formed close to the energy bands can hold carriers during the avalanche process with a release lifetime in the order of nanoseconds. [91] The released carriers may then trigger an avalanche resulting in a pulse called *afterpulse*. Afterpulses are undesired pulses that are highly correlated to the real pulses resulting from a photon absorption.

Timing jitter of the SPAD is the variation in delay between the absorption of a photon and the generation of an output electrical pulse. To measure the timing jitter of a detector, a picosecond laser and high-resolution electronics (a high-resolution TDC) are required to ensure that the dominant jitter is that of the detector and to give the instrument response of the SPAD. [94] Full width at half maximum (FWHM) of the instrument response function determines the timing jitter of the detector. It should be noted that many detectors have a non-Gaussian instrument response function. Jitter in the SPAD is an undesirable effect that causes the counts to shift into neighbouring clock cycles. [94]

3.3 QRNG based on Si-NCs LED

As mentioned before in Chapters 1 and 2, the chief assets of intrinsic indeterminacy and randomness in quantum physics can be utilized. Truly random numbers have been generated by exploiting the non-deterministic nature of quantum phenomena. Quantum dots [95, 96], SPAD [97], light emitting devices (LEDs) [54, 63, 98], and laser [61, 99, 100, 101, 57, 102, 103] have been employed as sources of entropy to produce random numbers. In this section, we introduce a QRNG which is able to produce sequences of random bits with a very negligible bias that pass all the NIST tests without the need of a post-processing algorithm for small datasets (∼100 Mbits). It is based on silicon nanocrystals (Si-NCs) LED as the source of entropy coupled with a SPAD as the detector. [104, 105, 106, 107, 108] At low applied currents, the Si-NCs LEDs act like an attenuated source of light with a Poisson distribution (see Appendix. B) in photon counts statistics. [107] Since the spontaneous emission of photons in Si-NCs LEDs is of non-deterministic quantum nature and considering the fact that these LEDs are CMOS compatible, employing them for the production of QRNGs is beneficial. The architecture where a Si LED and a Si SPAD are coupled can yield a compact and cheap QNRG fabricated by standard microelectronic processes. Furthermore, its performance can be greatly increased exploiting CMOS scalability.

3.3.1 Theory

3.3.1.1 Test for the Poisson distribution

As mentioned before, photons are emitted spontaneously in a Si-NCs LED. The spontaneous emission of photons in an LED is a non-deterministic, random process. There are several tests to examine whether a sample of observations comes from a Poisson distribution. [109] To test if the recorded data follow a Poisson distribution, we make use of the chi-squared (χ^2) statistic which compares observed data with the expected data we would obtain according to the null hypothesis that the data comes from a Poisson distribution. Let y_1, ..., y_n be independent, non-negative integer variables from a distribution P_1, the null and the alternative hypotheses then state that the distribution comes from a Poisson distribution P_2 or not, respectively:

$$H_1 : P_1 = P_2 \tag{3.1}$$

$$H_2 : P_1 \neq P_2 \tag{3.2}$$

The p-value which is the chi-square cumulative distribution function is calculated as [110]:

$$P = \int_0^x \frac{t^{(d-2)/2} e^{-t/2}}{2^{d/2}\Gamma(d/2)} dt \tag{3.3}$$

where x is the calculated value of χ^2, d the degree of freedom and Γ the gamma function. Considering a significance level α, if the p-value is larger than this level (p-value$>\alpha$), the null hypothesis is accepted, otherwise the alternative hypothesis is accepted which

indicates that the observed data does not come from a Poisson distribution.
Through the cross-correlation measurement which is the measurement of how similar the two random variables are as a function of the shifts of one relative to the other [111], we can check whether there exists a strong correlation between the two random variables. The measurement is based on the random transmission of the emitted photons from the source (e.g. Si-NCs LED) into two arms of a fiber beam splitter (as presented in Fig. 3.8) each connected to a detector (e.g. a SPAD). The output signal of the two detectors are then sent to a cross-correlator where the cross-correlation function, $g^2(\tau)$, is computed. A peak in the cross-correlogram indicates photon bunching while a dip shows anti-bunching. Photon bunching occurs in the case of chaotic or thermal light which has a super-Poissonian distribution with the mean greater than the variance and photon number fluctuations larger than a coherent light. Photon antibunching, however, refers to a sub-Poissonian distribution with the mean less than the variance and photon number fluctuations smaller than a coherent light. [112] A flat graph in cross-correlogram demonstrates that the photons are emitted randomly with a Poisson distribution. [113]

3.3.1.2 Survival model

If the null hypothesis in section 3.3.1.1 is accepted, meaning that the photon counts show a Poisson distribution, the probability of not observing photons in a given time window t_w is given by the survival function with an exponential distribution (see Appendix. C). Let t denote a non-negative random variable representing the waiting time until the occurrence of an event, then the survival function ($S(t_w)$) gives the probability that the event of interest has not occurred by duration t_w and is defined as in Eq. C.5:

$$S(t_w) = P(t \geq t_w) = e^{-\lambda t_w},$$

where λ is the number of detected photons per unit time. Suppose we fix t_w and that we want the probability of observing (at least) one photon to be equal to the probability of observing no photons, the survival function is then defined as:

$$e^{-\lambda t_w} = \frac{1}{2} \qquad \rightarrow \qquad \lambda t_w = ln(2). \tag{3.4}$$

In this way, by knowing the detected flux of photon, we can fix the integration time window to obtain an equal probability of detecting one photon and no photons.

3.3.2 Experimental procedure

Figure 3.6 shows the setup schematic for generating bit sequences using Si-NCs LED and Si SPAD. The current/voltage source that drives the LED is an Agilent B1500A Semiconductor Device Parameter Analyzer. The photons emitted from the LED are sent to the SPAD through an optical multimode fiber bundle which collects the light from the LED surface. No optics is used between the bundle and the LED. The SPAD is a PerkinElmer SPCM-AQRH-16, with the dead time of ∼35 ns, the DCR of ∼300 Hz and afterpulsing probability (AP) of 0.5%. The pulses from the SPAD are recorded via a multichannel scaler Ortec Easy-MCS with a minimum channel (bin) width of 100 ns and with no dead time between the channels. The scan length is variable from 4 to 65,536 channels. The measurements were performed at room temperature and the optical multimode fiber coupled with Si-NCs were kept inside a dark room.

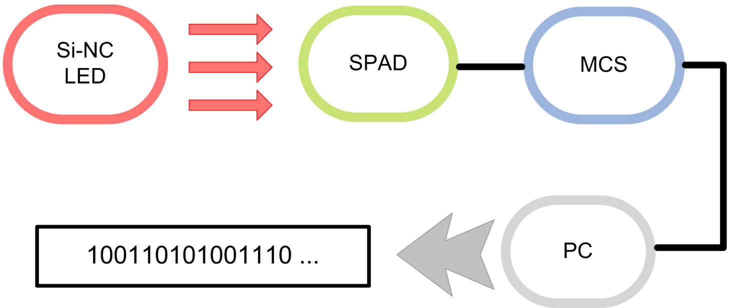


Figure 3.6: Scheme of the setup used to generate the bit sequences. Emitted photons from the Si-NCs LED are detected by a Si SPAD. The electrical signals are then sent to a multichannel scaler (MCS) connected to a PC to generate sequences of random numbers. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

The LED characteristics are fully described in [87] and in Section 3.1. As discussed in Section 3.1, to assure the injection of carriers into Si-NCs under the direct tunneling mechanism and to avoid the FN tunneling which causes degradation to the oxide layer and, hence, inefficiency of the LED and the creation of a thermal LED, the applied forward current to the LED was kept lower than ∼1.7 μA (corresponding to a voltage of ∼2.9 V). These Si-NCs LEDs show remarkable stability over weeks of continuous operation. [114]

3.3.3 Results and discussion

As stated before, since the spontaneous emission of photons in an LED is the origin of randomness, the Si-NCs LED can be used as a quantum source of randomness. First, we demonstrate that it can be described as a Poisson source of entropy. Having fixed the driving current to 1.3μA, we measured the occurrence of counts in a time window t_w of 1 μs. The histogram plot of the counts with the Poisson fit are presented in Fig. 3.7. The statistics was done on 65535 counts recorded each two t_w (the reason of this procedure was to remove correlation as we will detail later). Considering the conventional significance level of α=0.05, we computed a p-value of 0.0663. Since p-value> α, we conclude that the Poisson distribution matches well with our obtained data.

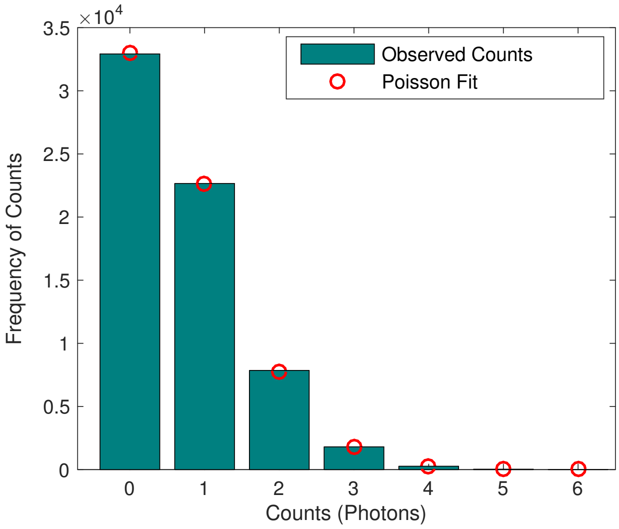


Figure 3.7: The histogram containing 65535 samples of data and the Poisson fit with the mean photon number of 0.69. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

To check further if the detected photons follow a Poisson distribution, we did cross-correlation measurements. Fig. 3.8 shows the schematic setup comprising Si-NCs LED, a 2x2 optical multimode fiber splitter, two Si SPADs and a cross-correlator. The fiber splitter has arms of 0.5 and 15 m, with one of the 0.5 m arms coupled with the Si-NCs LED, the other 0.5 m arm being blocked and and the 15 m arms coupled with the Si SPADs). The fiber arms which are coupled with the Si SPADs, are long enough to prevent the possible correlations between the detected light on one SPAD and the detection of any sorts of reflected light (due to that detection) on the other one. The Si SPADs are Excelitas (SPCM-AQRH-14) with the dead time of 22 ns, the DCR of ∼100 Hz and afterpulsing probability of 0.5%. The cross-correlator is a linear correlator with 103 channels (bins) and the resolution of 1.3 ns.

The output signals of the Si SPADs are sent to the two channels of the cross-correlator to perform the cross-correlation measurement. The cross-correlogram of the output pulses from the Si SPADs, resulted from the detection of the emitted photons from the Si-NCs, is presented in Fig. 3.9. As can be seen in the figure and has been explained in Section 3.3.1.1, the flat cross-correlation graph (with no peak or dip) demonstrates that the detected photons follow a Poisson distribution. This is another proof, in addition to the χ^2 statistic, that the Poisson distribution is a good match for the distribution of the detected photons (emitted from the Si-NCs LED) on the Si SPAD.

The next step was to make use of the survival function to fix the measurement parameters. The emitted photons from the Si-NCs LED were detected by the SPAD. The electrical signals out of the SPAD were then sent to the MCS with a maximum scan length of 65536

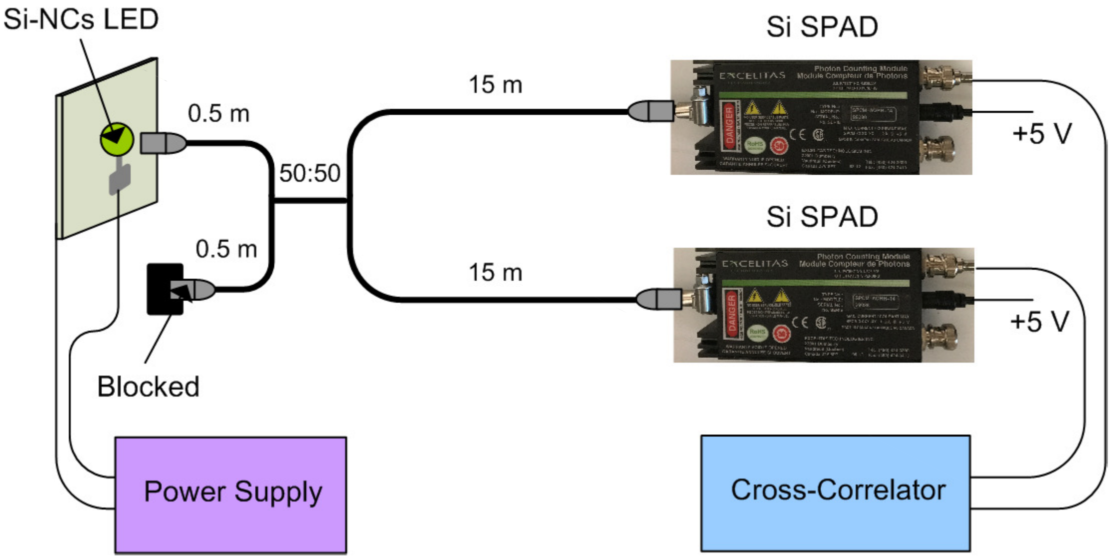


Figure 3.8: The cross-correlation setup consisting of a 50:50 optical multimode fiber splitter coupled with Si-NCs LED from one of the shorter arms (0.5 m) end and with Si SPADs from the longer arms (15 m) ends and a linear cross-correlator to perform cross-correlation measurement of the output signals from the Si SPADs. The other 0.5 m arm is being blocked to light exposure. The cross-correlator is connected to a PC with the Labview software to exhibit and save the cross-correlogram.

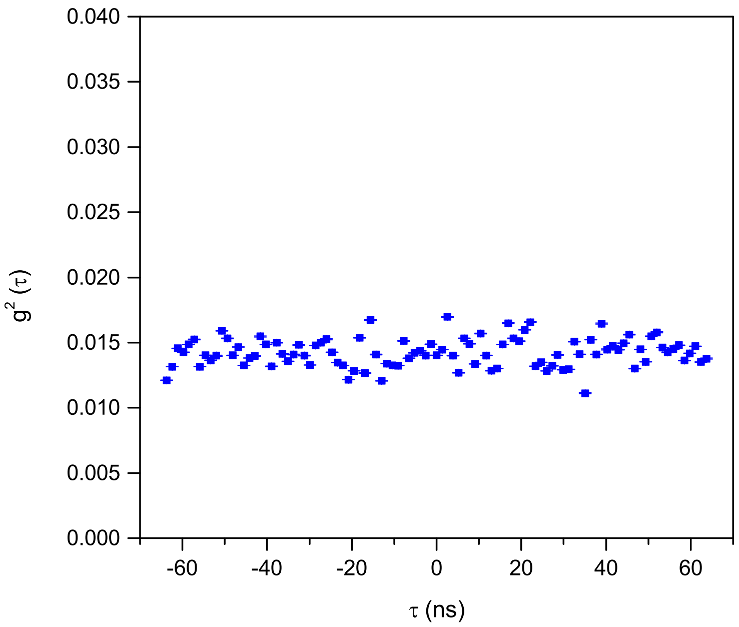


Figure 3.9: The cross-correlogram of the output pulses from the Si SPADs. The flat graph indicates that the detected photons on the Si SPADs (resulted from the detection of the emitted photons from the Si-NCs) follow a Poisson distribution. The plot is not normalized and the error bars can be seen for some data points.

channel (bins) to count the input events (detected photons) in the channels of its digital memory. Fixing the bin width to 1 μs, we find $\lambda \sim 6.9 \times 10^5$ (counts/s) obtained by applying a current of 1.5-1.7 μA to the Si-NCs LED.

Fixing the applied current to the Si-NCs LED and acquiring data using the setup in Fig. 3.6, we measured the probability of ones for 262 sequences each of 10^6 bits (Fig. 3.10). The overall time duration of the sequences is then $262 \times 10^6 \times 1\ \mu s$ (i.e. number of sequences times number of bits times bin width), whereas the actual acquisition time was 28 min due to the time required for the data to be buffered in the MCS and transferred to the PC.

Then, the acquired bit strings were evaluated for randomness by a set of statistical tests. A very popular set is the NIST test suite [81]. The raw sequences passed all the statistical tests in the NIST tests suit except for the Runs test. This failure is usually attributed to the existence of correlation between consecutive bits since the Runs test implies that the probability of a change from a one to a zero to be equal to the probability of a change from a zero to one for a random sequence [81]. This correlation is caused by the detector's afterpulsing. As can be seen in Fig. 3.11 (blue squares), only the first time lag is correlated significantly since the time scale of the bin width is greater than the dead time of the avalanche photodiode [115].

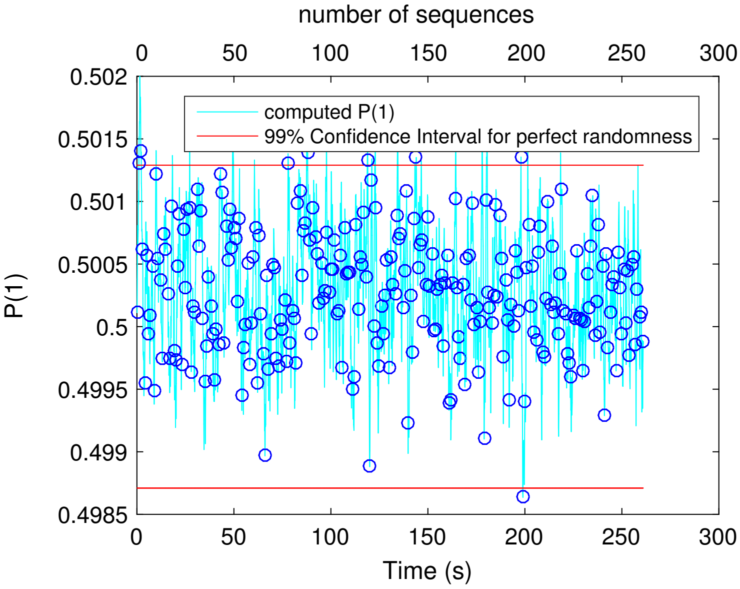


Figure 3.10: The evolution of the probability distribution of one with a 99% confidence interval in a dataset of 262 sequences each of 10^6 bits. Note that the time scale refers to the bit acquisition time and not the actual time of the measurement. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

Considering the 262 sequences as the output of a Markov process[1] of order one, we built the transition matrix $(T)^2$ of the process. As mentioned before, the experimental setup was adjusted to make individual 0 and 1 equally likely to happen, so the transition matrix was symmetric and its stationary distribution is consequently uniform. We could not measure a difference between the uniform distribution and the transition probability after only two steps, T^2, which suggests that any correlation present was low enough that reinforcing a dead time of a single observation window would be appropriate. Hence, one bit in every two in the dataset was removed that is like a simulation of experimentally enforcing a dead time of length equal to the bin width t_w.

Doing so, we halve the bit rate, outputting a single bit every 2 μs instead of 1 μs. The statistical analysis, done on these new sequences, shows no correlation between the bits and a perfect balance between zeros and ones. Figure 3.11 (red circles) represents the removal of the significant correlation after eliminating each alternate bit in the dataset. Therefore, the effect of the afterpulsing of the detector is removed.

The remaining 131 sequences of 10^6 bits pass all the statistical tests in NIST test suite de-

[1] A Markov process is a stochastic process in which the number of outcomes or states are finite, the outcome at any stage depends only on the outcome of the previous stage (i.e. as time goes by, the process loses the memory of the past) and the probabilities are constant over time (time-homogeneous Markov process). [116]

[2] The transition matrix of an n-state Markov process is an nxn matrix T where the i, j entry of T represents the probability of the transition of an object in state j into state i. [116]

scribed in Section 2.3.7.1 with a maximum measured bias of ~ 0.0012 (see Section 2.3.6). The min-entropy is calculated to be ~ 0.9965 bits per bit of data (Eq. 2.17). As previously mentioned in Section 2.3.7.1, to evaluate the success of a test, a significance level (α) of 0.01 is assigned for the test since common values of α in cryptography are about 0.01. If $P\text{-}value_T \geq 0.0001$, then the sequences can be considered to be uniformly distributed. The proportion of passes for each test is presented in Fig. 3.12 (b). All the results have been normalized to 100.

It was shown after more experiments that to remove the correlation between consecutive bits in the dataset, an enforced dead time of 500 ns is sufficient. Doing so, high quality random sequences are generated without the application of any post-processing operations. The results of the NIST tests are illustrated in Fig. 3.13 (a) and (b).

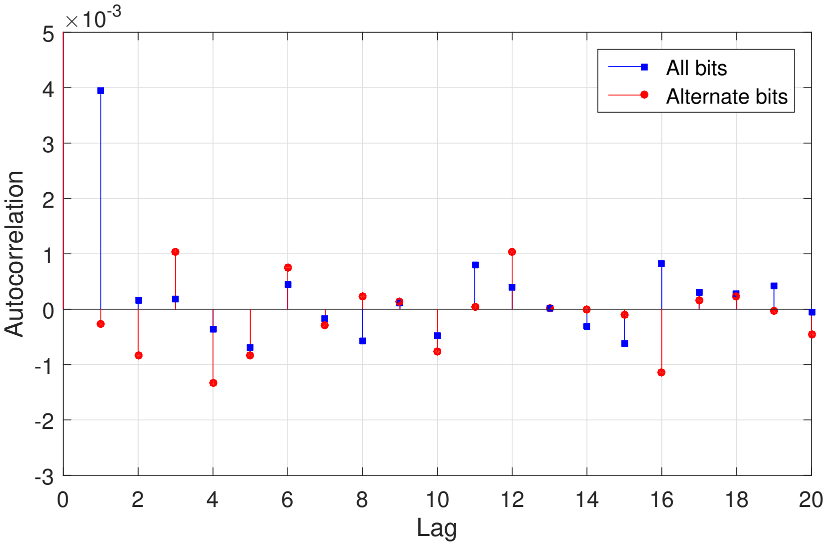


Figure 3.11: Autocorrelation versus time lag for the dataset of 262 sequences each of 10^6 bits (blue squares) showing high correlation at time lag 1 and the dataset of 131 sequences each of 10^6 bits (red circles) showing negligible correlation after eliminating each alternate bit in the dataset of 262×10^6 bits length. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

3.3.4 Discussion on long datasets

The emission intensity of the Si-NCs LED is remarkably stable over continuous operation. Still small variations of the ambient conditions influence their behavior. Note that to acquire the 131 Mbits long sequences that pass the NIST tests, we need to operate the system continuously for 28 min. This implies that the EL should be stable in that time interval. If we acquire 1 Gbits long sequences, the actual measurement time is about 214 minutes. Now EL dependence on ambient interference becomes evident (Fig. 3.15). In

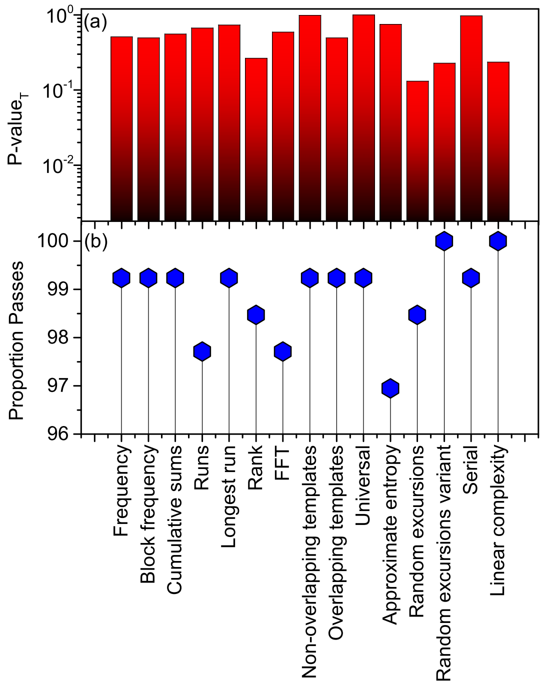


Figure 3.12: (a) Results of the NIST tests for 131 sequences of 10^6 bits for a simulated dead time of 1 μs. All tests are passed at the 0.01 significance level. (b) The proportion of passes for each test. The minimum pass rate for each statistical test is approximately 96 for a sample size of 100 binary sequences. The results of all tests have been normalized to 100. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

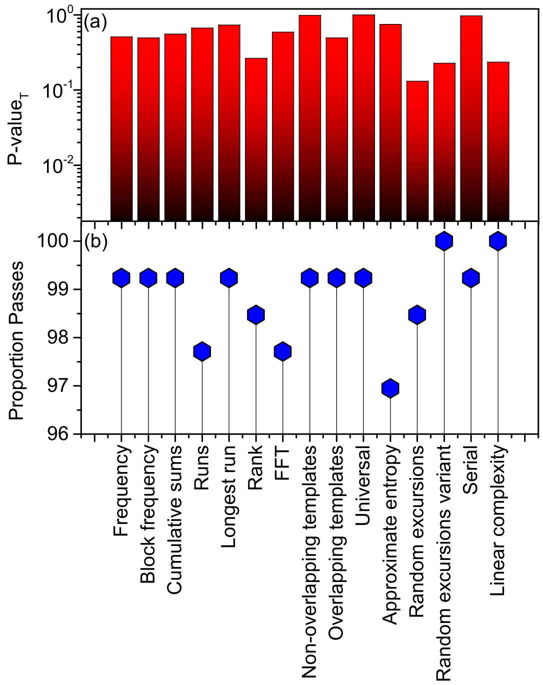


Figure 3.13: (a) Results of the NIST tests for 131 sequences of 10^6 bits for a simulated dead time of 500 ns. All tests are passed at the 0.01 significance level. (b) The proportion of passes for each test. The minimum pass rate for each statistical test is approximately 96 for a sample size of 100 binary sequences. The results of all tests have been normalized to 100. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

fact, considering a long dataset of 1 Gbits failure in the main statistical tests in NIST tests suite is observed which is due to the loss of equal probability of ones and zeros (bias) (Table 3.1). As shown in Fig. 3.14, a drift in the probability distribution of ones is observed. The failure time then precedes t_w and the survival function in Eq. 3.4 no longer applies to our system. The loss of the equal probability is due to a $< 0.4\%$ variation in the EL of the Si-NCs LED. Figure 3.15 (a) reports a day record of the emission intensity of the Si-NCs LED. Fluctuations of the EL intensity in the few per mille range are observed. At the same time, we have recorded also the voltage bias, the ambient temperature, and the ambient humidity (Fig. 3.15 (b), (c), and (d)) to look for correlations. Close observation of the plots shows that the EL is most significantly correlated with the alterations of the ambient temperature. A slight decrease (increase) in ambient temperature results in a decrease (increase) in EL. It can be seen that after about six hours of measurement, a small decrease of 0.5°C in temperature causes a decrease of about 0.2% in the EL of the LED. As the bias factor increases exponentially with λ (Eq. C.5 and 3.4), the randomness is extremely sensitive to the EL intensity and any small variation of the EL intensity will exponentially increase the bias factor. Therefore, the QNRG is possible only when the driving current is precisely set at each condition.

In the literature, to remove the bias on the data either a precise control on the QNRG parameters [54] or post-processing algorithms [117] or simple encoding methods [118] have been proposed. Although the boundary between post-processing algorithms and encoding methods is not clearly defined, the amount of resources needed can be adopted to differentiate. Here we first used the Von Neumann randomness extractor [118] and then more efficiently the information-theoretically provable Toeplitz extractor [119] to extract the randomness in the raw dataset of 1 G bits length.

3.3.4.1 Von Neumann randomness extractor

Some parameter control solutions can also be taken into account for long datasets to overcome the problem of the drift in the probability of ones (zeros) such as stabilizing the Si-NCs LED temperature, resetting the applied current to the Si-NCs LED (or equivalently resetting the bin width in the MCS) and designing a feedback for the system. Stabilizing the temperature, for instance, will cause the EL intensity to remain constant and therefore the equal probability of ones and zeros will be maintained. In the same manner, if the applied current to the LED or the bin width in MCS is reset to keep the EL intensity invariant, the drift in the probability of ones (zeros) will be eliminated and eventually the system would produce long datasets of high quality random numbers.

However, these methods are more resource hungry than the simple and more economic Von Neumann randomness extractor. This extractor takes successive pairs of consecutive bits (non-overlapping) from the input stream. If the two bits match, no output is

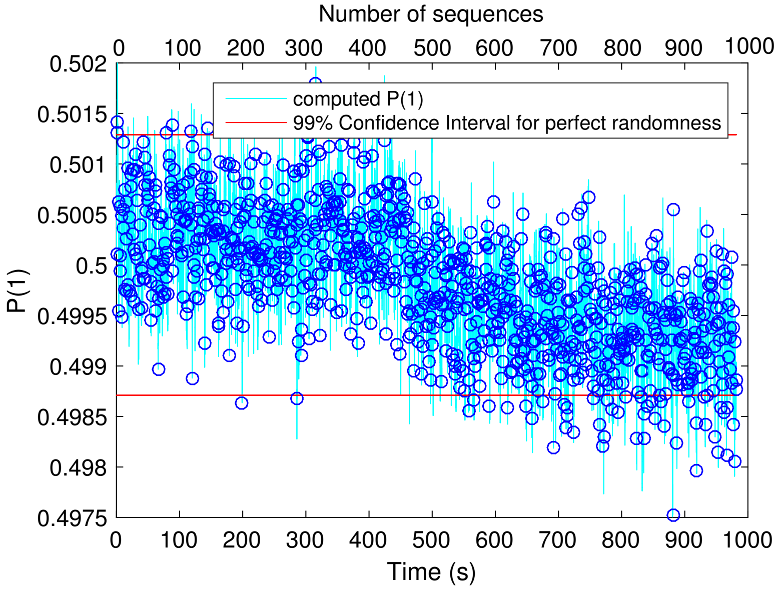


Figure 3.14: The evolution of the probability distribution of one with a 99% confidence interval in a dataset of 1000 sequences each of 10^6 bits. Note that the time scale refers to the bit acquisition time and not the actual time of the measurement. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

generated. If the bits differ, the value of the first bit is output. [120] The Von Neumann extractor makes the binary data unbiased but quarters the bit rate. [118]

We apply the Von Neumann extractor to remove the bias in the raw dataset A with 1 G bits length. The results of the NIST tests for the obtained dataset B are presented in Table 3.1. Since there is a failure in the Runs test, eliminating each alternate bit in dataset B would remove the correlation between consecutive bits resulting in dataset C. Therefore, as can be seen in Fig. 3.16, the correlation is appreciably suppressed and all the NIST tests are passed for dataset C with the reduced bit rate to 125 kbps.

Table 3.1: Results of the four statistical NIST tests for datasets A, B, and C with 1000, 250, and 125 sequences of 10^6 bits, respectively. The *P-value$_T$* has to be larger than 0.0001. The minimum pass rates for each statistical test is approximately 980, 242, and 120 for a sample size of 1000, 250, and 125 binary sequences, respectively.

Statistical test	**Dataset A**			**Dataset B**			**Dataset C**		
	P-value$_T$	Proportion	Result	P-value$_T$	Proportion	Result	P-value$_T$	Proportion	Result
Frequency	0.000000	952/1000	Failed	0.363593	246/250	Passed	0.080108	124/125	Passed
Block frequency	0.000000	968/1000	Failed	0.009136	250/250	Passed	0.826984	125/125	Passed
Cumulative Sums	0.000000	955/1000	Failed	0.779188	246/250	Passed	0.936639	124/125	Passed
Runs	0.000000	431/1000	Failed	0.000000	236/250	Failed	0.158872	125/125	Passed

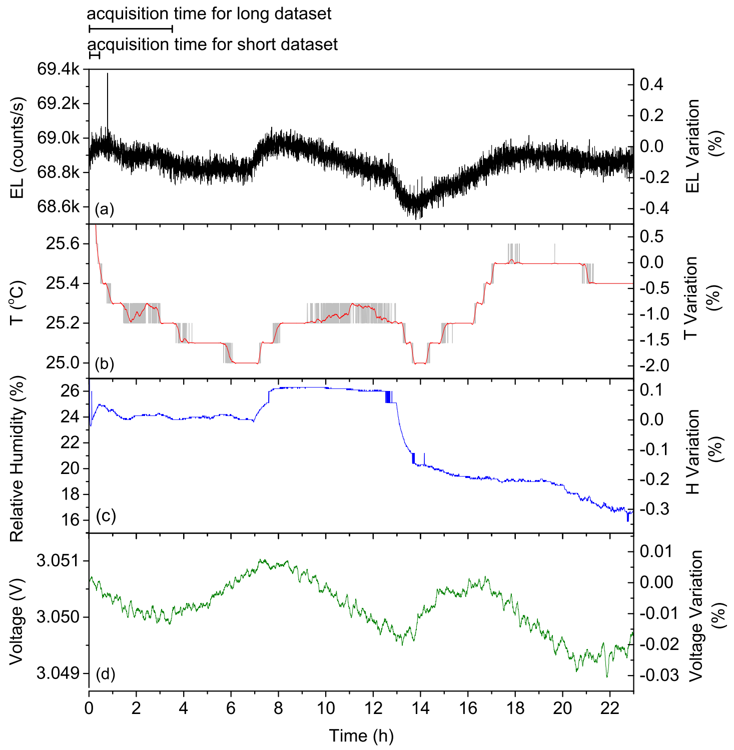


Figure 3.15: Plots showing (a) EL, (b) ambient temperature, (c) ambient humidity, and (d) Si-NCs LED voltage with their variations versus time. The acquisition time scales for small (28 min) and long (214 min) datasets can be seen on top of the plot. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

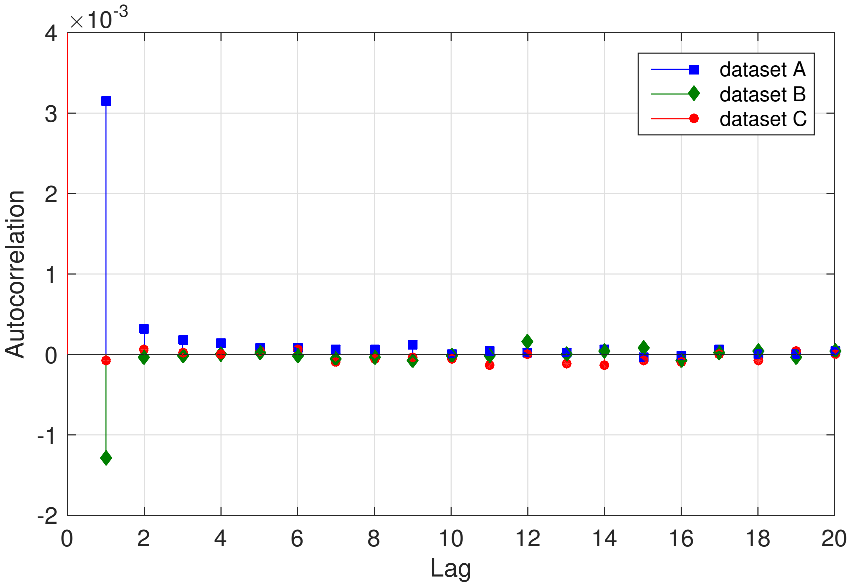


Figure 3.16: Autocorrelation versus time lag for the datasets A (blue squares), B (green diamonds), and C (red circles) with 1000, 250, and 125 sequences of 10^6 bits, respectively. Dataset B is obtained by the application of Von Neumann randomness extractor to the raw dataset A. The elimination of each alternate bit in dataset B results in dataset C.

3.3.4.2 Toeplitz-hashing randomness extractor

It is more efficient to use information-theoretically secure randomness extractors like the Toeplitz-hashing function instead of simple randomness extractors like the Von Neumann extractor. [66] We implemented the Toeplitz-hashing extractor to our QRNG as follows. As mentioned before in Section 3.3.3, the min-entropy of the raw data is calculated to be $\sim$ 0.99 bits per bit. With the input bit-string length of 1000 bits, the output bit-string length is $1000 \times 0.99 \geq 990$. Therefore, we conservatively built the Toeplitz matrix of 1000×940 using some high quality random numbers as the elements of this matrix. Through matrix multiplication of the raw data, put in vectors of 1×1000, by the Toeplitz matrix of 1000×940, we obtained 940 Toeplitz-hashed bits out of each 1000 bits of raw data. As can be seen in Fig. 3.17, the correlation is appreciably suppressed and all the NIST tests are passed for the Toeplitz-hashed dataset (Table 3.2).

3.4 Conclusions

We realized a physical quantum random number generator exploiting Si-NCs LED as the source of randomness. Very negligible bias and simple setup are the chief strengths of our QRNG. With forced dead time of 1 μs and 500 ns, 100 M bits long sequences pass the statistical tests of the NIST suite. The highest bit rate achieved for short datasets, with

Table 3.2: Results of the statistical NIST tests for the datasets of 1000 and 940 strings of 10^6 bits of raw and Toeplitz-extracted data, respectively. The $P\text{-}value_T$ has to be larger than 0.0001. The minimum pass rates for statistical tests is 0.980. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

	Raw data			**Toeplitz-extracted data**		
Statistical test	P-value$_T$	Proportion	Result	P-value$_T$	Proportion	Result
Frequency	0.000000	0.952	Failed	0.114955	0.987	Passed
Block frequency	0.000000	0.968	Failed	0.588505	0.992	Passed
Cumulative Sums	0.000000	0.955	Failed	0.229355	0.989	Passed
Runs	0.000000	0.431	Failed	0.065561	0.988	Passed
Longest run	0.607993	0.992	Passed	0.547061	0.984	Passed
Rank	0.916599	0.989	Passed	0.164541	0.987	Passed
FFT	0.130369	0.985	Passed	0.601722	0.990	Passed
Non-overlapping template	0.009071	0.985	Passed	0.855973	0.984	Passed
Overlapping template	0.000000	0.963	Failed	0.676924	0.989	Passed
Universal	0.975012	0.990	Passed	0.267060	0.990	Passed
Approximate entropy	0.000000	0.986	Failed	0.100084	0.994	Passed
Random excursions	0.131334	0.980	Passed	0.380164	0.984	Passed
Random excursions variant	0.034368	0.986	Passed	0.182977	0.989	Passed
Serial	0.735908	0.984	Passed	0.733513	0.989	Passed
Linear complexity	0.530120	0.993	Passed	0.264222	0.990	Passed

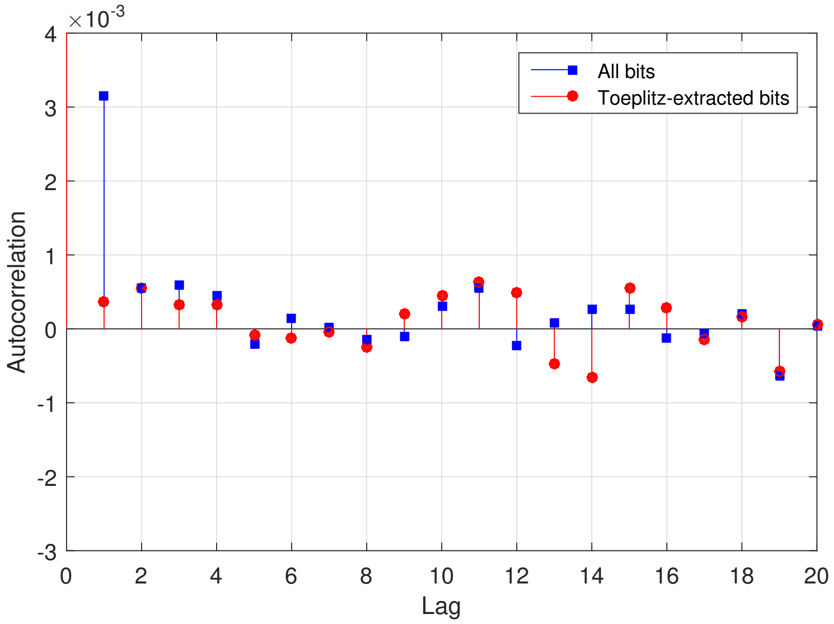


Figure 3.17: Autocorrelation versus time lag for 1 Gbits and 940 Mbits of raw data and Toeplitz-extracted data, respectively. The correlation at the first lag is minimized after the application of the Toeplitz extractor. © Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced from [107] with permission.

a forced dead time of 500 ns, is 0.6 Mbps.

Besides the bit rate, our approach benefits from several advantages: it uses light to stimulate events in the SPAD and avoids a deterministic post-processing of the raw data for small datasets. This fact is extremely remarkable in producing high quality random numbers and compensates for the low bit rate. Furthermore, the approach proposed here uses simple silicon-based LEDs as the light source and its overall bit rate can be easily increased by adopting a parallel architecture and exploiting the CMOS compatibility of all the components.

However, 1 G bits long datasets fail the main statistical tests in the NIST tests suite. This failure is attributed to a per mille drift in the EL of the Si-NCs LED that violates the equal probability of ones and zeros. The randomness is extracted by the application of the Von Neumann and information-theoretically secure Toeplitz extractors. The bias and correlation among bits are removed and all the statistical tests in the NIST tests suite are consequently passed. A number of parameter control solutions such as stabilizing the temperature, resetting the applied current to the Si-NCs LED (or equivalently resetting the bin width in the MCS) and considering a feedback for the system can also be taken into account to overcome the problem of bias and to generate long, high quality random bit streams.

Chapter 4

Robust QRNG based on Si-NCs LED

A class of optical QRNGs based on the timing measurement of the photon arrivals has been proposed in literature. [59, 63, 57, 121, 102] Random bits have been extracted by comparison of the time difference between subsequent random events [59], comparison of the photon numbers in consecutive laser pulses distributed in time [57], random arrival times of photons on a single and an array of photodiodes [63, 102], and encoding the independent and uniformly distributed random phase time. [121] In [102], despite the high bit rate, the mean of the photon flux is larger than one which makes the security of this method arguable for applications in quantum cryptography where the quality of random numbers is of higher importance than the speed. In [63], the exponential distribution of the arrival times of photons introduces bias in the raw data which is removed by post-processing operations. The bit extraction method in [59] makes the efficiency to be around 0.5 bits per detection since using the restartable clock method to eliminate both bias and correlation reduces the efficiency to less than 1 bit per arrival. In a recent method [121], with the maximum generation rate of 128 Mbps, the quality of random numbers is affected by a bias introduced at too high counting rates. In addition, the setup used in this approach is complex.

To improve the state of the art in terms of simplicity, robustness and random numbers quality, we developed a methodology based on the photon arrival time measurements with thorough consideration of the Si SPAD imperfections such as afterpulsing, dead time and jitter (described in Section 3.2). This approach is simple and easy to model, all-silicon based, robust and able to generate high quality random numbers. We focus on a random number generation technique in which the source of entropy is quantum mechanical. It is a Si-NCs LED (see full description in Section 3.1) coupled with a Si SPAD. A dedicated field-programmable gate array (FPGA) performs the random bit extraction. This approach avoids the use of post-processing algorithms used elsewhere. [64, 63] The proposed QRNG is robust against variations of the internal and external parameters such as the aging of the components and changing temperature. The components of the QRNG can be integrated on the silicon platform via CMOS technology allowing the fabrication of a compact device.

4.1 Methodology

4.1.1 Theory

As explained and proved in (Appendix B), the Poisson process has the property that if there is only one single arrival in a time interval $[0\ ,\ t]$, the distribution of the arrival times is uniform throughout the interval. This fact has been considered in several works [64, 121]. However, a robust methodology which takes into account this fact in addition to thorough consideration of the non-idealities of the detector (Si SPAD in our case), did not exist in the literature. Therefore, to develop a methodology[1] based on the studies of the statistics of the detected photons (emitted by the Si-NCs LED) by the Si SPAD (see Section 3.3.3), we consider intervals with a fixed duration having “only” one single arrival; intervals with no arrival or with more than one arrival are discarded. In this way, we cope with the emission variations of the Si-NCs LED at the expense of a reduced random bit generation rate.

4.1.2 Target function

The interval and subinterval structure for an ideal detector is explained in Fig. 4.1 (a). Every interval is composed of 16 subintervals of equal length, each associated with a symbol that generates the random number if a single photon is detected (i.e. a photon arrival occurs) in that specific time interval.

The real Si SPADs exhibit a number of non-idealities. The most important ones are afterpulsing, dead time, jitter, dark counts, light emission during avalanche and efficiency lower than 100%; all dependent on temperature, ageing, bias voltage, etc. Afterpulses are strongly correlated to true pulses and can severely deteriorate the Poisson statistics of the source.

Through autocorrelation analysis of the detector signal, the afterpulsing distribution can be measured. [123] The normalized correlation function of the multitau digital correlator (the correlator used to perform autocorrelation measurement of the output signal of the Si SPAD in the experimental setup shown in Section 4.3) is computed as [124]:

$$g_{ij}^{2}(t_c) = \frac{N_{\text{pairs}}(t_c)}{S_i\ S_j\ (t_{\max} - t_c)\ \Delta t_c}, \tag{4.1}$$

[1] This methodology was proposed by Giorgio Fontana (`http://www.ing.unitn.it/~fontana/`) and improved through our discussions.

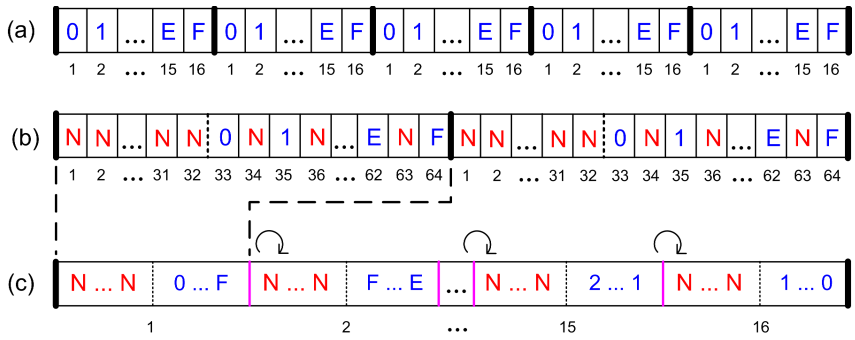

(a)
0 1 ... E F 0 1 ... E F 0 1 ... E F 0 1 ... E F 0 1 ... E F
1 2 ... 15 16 1 2 ... 15 16 1 2 ... 15 16 1 2 ... 15 16 1 2 ... 15 16
(b)
N N ... N N 0 N 1 N ... E N F N N ... N N 0 N 1 N ... E N F
1 2 ... 31 32 33 34 35 36 ... 62 63 64 1 2 ... 31 32 33 34 35 36 ... 62 63 64
(c)
N ... N 0 ... F N ... N F ... E ... N ... N 2 ... 1 N ... N 1 ... 0
1 2 ... 15 16

Figure 4.1: Schematic of (a) a conventional interval with 16 symbols of {0 1 2 ... 9 A B ... F}, (b) the "double length" interval with the first half of no number symbol (N) and the second half with N subintervals between symbols, and (c) the super interval containing 16 "double length" intervals as in (b) with a consecutive one-rotation of symbols. © 2017, IEEE. Reprinted, with permission, from [122].

where $N_{\text{pairs}}(t_c)$ is the number of photon pairs accumulated for the time bin t_c, S_i and S_j are the average signals in channels i and j, respectively, $t_{\max}$ is the total experiment time, and Δt_c is the bin width for t_c. In the case of autocorrelation, only one channel is used and hence in Eq. 4.1, i and j are considered equal.

Dead time can also be measured with autocorrelation analysis of the detector signal. [125] The detector jitter is a random variable that adds to the arrival times of the photons. Therefore, it changes the statistics of the measured arrival times (see Section 3.2).

Compared to the operational photon flux, dark counts are extremely rare events in our detector ($\sim$ 300 counts/s) and they do not alter the overall behavior of the apparatus. Detector efficiency is about 50% and simply adds to the losses of the whole optical chain. The low efficiency of the detector highlights the fact that a large proportion of the arrivals are inherently discarded by the losses. The surviving detections, however, keep their Poisson distribution.

In order to mitigate these detector non-idealities, we modified the simple scheme of Fig. 4.1 (a). If we proceed with the conventional interval scheme in Fig. 4.1 (a), we face some problems with:

1. the afterpulsing which occurs within a single interval that causes multiple detections within the same interval,
2. the afterpulsing which occurs across intervals (i.e. a photon is detected in an interval while the afterpulse is generated in the following time interval),
3. in-interval dead time, and
4. across-interval dead time.

As we discard all intervals with more than one detected photon (Section **??**), the first problem has no effect.The second problem can be defeated by counting the number of photon arrivals in the previous interval. If there is one or more than one detection in the previous interval, the present interval is discarded. Therefore, it may never happen that the afterpulse generated by a legitimate detection in an interval is counted as a legitimate detection in the following one. This mitigation strategy removes the effects of afterpulsing but reduces at the same time the efficiency of the generator.

The discard of an interval provided a photon detection in the previous interval, alleviates also across-interval dead time. In fact, if a legitimate photon is detected at the end of an interval and this arrival generates a random number, the detector dead time makes it impossible to detect a photon in the first subintervals of the following interval, introducing correlation in the random number generation. With the above-mentioned rule, this situation is impossible.

In-interval dead time is masked if the dead time of the detector is shorter than the subinterval duration, so multiple photon detections would generate the same random number. Ideally, our method would discard that number due to more photon detections in an interval, but dead time makes it valid. This is equivalent to the effect of an optical attenuator, which does not alter the uniform distribution of arrival times. The case of in-interval dead time spanning across two subintervals is different; it changes the uniform distribution of arrival times. To overcome this problem, we introduce a no-number generating subinterval (nngs) between random number generating subintervals (rngs), as presented in Fig. 4.1 (b). Doing so, in-interval dead time can mask photons that if detected would generate no number or would cause the number to be discarded, so again mimicking an optical attenuation.

One factor that might affect the results is the inability of the SPAD to resolve multiple photons arriving within the SPAD resolution window. Let us note that the result in Eq. **??** generalizes to any number of arrivals, i.e. if there are multiple arrivals in a time interval, each arrival time is a random variable with uniform distribution. [126] Therefore, if many photons hit the detector in the same resolution window, one of them should be selected at random before detection in order to be a valid photon according to our methodology. If multiple photons arrive at the same subinterval, one should be selected at random as mentioned before. However, multiple photons (not resolved by the detector) will generate the same random symbol as the single photon selected at random. In any case, a low flux of photons, obtained by an attenuator or a low efficiency source, makes this occurrence a low probability event. Thus, the inability to observe the multiple arrivals within the resolution window of the detector or inside the subinterval will not affect the generation method.

As mentioned before, the SPAD measures photon quantity $N > 0$ and not only $N = 1$. The two conditional probabilities of $P(T \leq \tau | N(t) > 0)$ and $P(T \leq \tau | N(t) = 1)$ are merely different for large λ. [126] The observable departure from the uniform distribution in Fig. 4.2 is possibly due to this phenomenon together with the non-uniform duration of the subintervals originated from the electronics as discussed below.

Through initial measurements, we observed a very small departure of $\sim 0.01\%$ in the probability of generated symbols from the ideal value of $P_{\text{ideal}}(\text{symbol}) = 1/16$ (Fig. 4.2) as the result of non-uniformity in the length of subintervals and/or the lack of photon number resolution in SPAD. The non-uniformity is most probably due to periodic fluctuations of rail voltages in the FPGA (the jitter). [127] We do propose in the following a mitigation technique which solves these issues.

Our methodology is developed by defining “double length” periodic time intervals with an associated fully deterministic “target function”. In the case of 16 rngs, the alphabet of the symbols is {N, 0, 1, ... F}, that reads N (no-number), and the hexadecimal numbers 0 to F. Each interval has 32 N subintervals in the first half to mask the afterpulsing distribution of the Si SPAD (Section 4.3) and an alternation of N subintervals and the full set of numeric symbols in the second half, with a total of 64 subintervals (Fig. 4.1 (b). Only if one single detection hits the target function associated with an interval, a random symbol is generated. In addition, since the order of the symbols in the target function is not relevant, a “super-interval” has been defined that is composed of 16 “double length” intervals, in which the random number generating symbols are ordered as {0, 1, ... F} in the first interval, {F, 0, ... E} in the second one and so on.

The consecutive one-rotation of symbols (Fig. 4.1 (c)) does indeed uniformly redistribute the inherent jitter of the physical counter to all symbols (Fig. ??).The super-interval structure is deterministic. Because not all Intervals within a super-interval generate a random number, it is not possible to invert the function from the generated random symbol stream. In fact the information of the number of skipped intervals is not included in the output stream.

4.2 Robustness

Robustness plays a key role in a QRNG designed for cryptographic applications. The method employed to extract random numbers has to be robust against internal defects and external attacks. An external attack might change the dead time of the detector or the afterpulsing distribution, or it might find loopholes in the post-processing algorithms often used by most methods.

By masking the non-idealities of the detector described in the previous section, the internal robustness against the operational system defects is achieved. If the nominal characteristics of the light source and the Si SPAD are maintained within tolerances included in the design of the target function, the generation of high quality random numbers is guaranteed

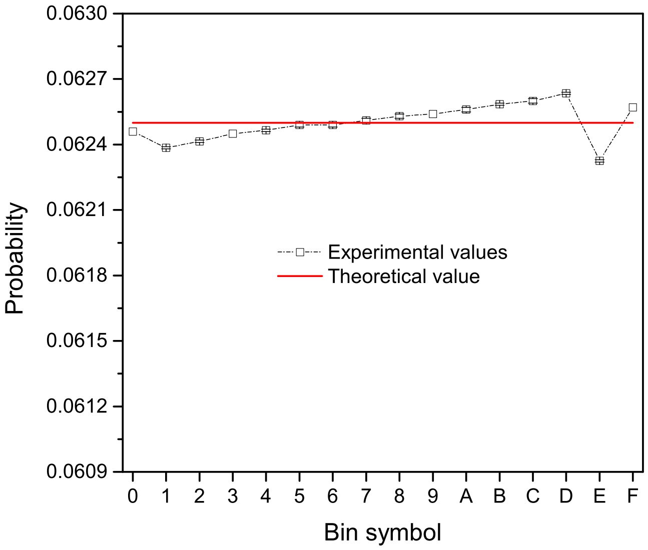


Figure 4.2: Probability distribution of the 16 bin symbols built by analyzing 1 G symbols raw data. The empty squares refer to the experimental values acquired using the target function in Fig. 4.1 (b) and the solid red line to the theoretical value of 1/16=0.0625. The error bars are visible for some data points. The distribution is observed to be nonuniform. © 2017, IEEE. Reprinted, with permission, from [122].

without any consistency check of the generation hardware. The external attacks could change the intensity of light by for instance a change in the temperature, but this effect will not change the quality of the random numbers; it might only change the efficiency of the generator (see Section 4.4).

4.3 Experimental procedure

The experimental setup is presented in Fig. **??**. Photons emitted from a Si-NCs LED are detected by a single photon counting module, PerkinElmer SPCM-AQRH-16, through a multimode fiber bundle. The LED is driven by an Agilent B1500A Semiconductor Device Parameter Analyzer. The EL of the LED is monitored by a Hewlett Packard 53131A Universal Counter. The TTL output of the detector is directly connected to the high speed digital input of the FPGA.

The measurement of the arrival times is performed by a fully synchronous logic. The FPGA continuously samples the detector at the frequency of 100 MHz, which is crystal controlled. A valid arrival is produced by a high analog logic level heralded by one clock cycle (10 ns) of low analog logic level. A Digilent ATLYS FPGA board has been used with the programming language VHDL. The temperature is monitored and controlled by an LCI Light Control 350 Temperature Controller Module. All the measurements are conducted in dark condition inside a probe station.

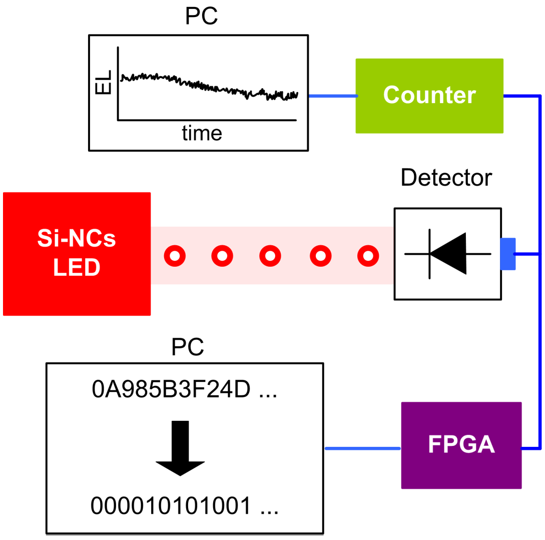


Figure 4.3: Schematic of the setup for random numbers generation comprising a Si-NCs LED, a Si SPAD and an FPGA to generate random numbers. The random hexadecimal symbols produced at the FPGA are then converted into binary “0” and “1” with the efficiency of 4 bits out of each symbol. The electroluminescence (EL) is monitored by a photon counter connected to a PC. © 2017, IEEE. Reprinted, with permission, from [122].

Autocorrelation, $g^2(\tau)$, measurement of the Si SPAD signal was performed via a multitau digital correlator with 4 ns resolution.[124] The afterpulsing distribution exhibits a main peak within 80 ns from the main autocorrelation peak at τ=0 (Fig. 4.4). Additional peaks are possibly related to reflections of the light generated by the SPAD itself when a photon is detected. This light pulse travels forth and back in the fiber and may induce a time shifted correlated detection. The plateau in $g^2(\tau)$ approaches the normalization value of 1 at about 160 ns.

We conservatively determined that after 320 ns afterpulsing and fiber reflections will not contribute to the statistics of the generated random numbers. Therefore, we adopt fixed contiguous “double length” time intervals of 640 ns, embedded in the super-interval structure. A single arrival in the second half of the interval (Fig. **??** (b)) produces 4 bits when arrival occurs in one of the 16 possible active subintervals.

Starting with a Poisson process with intensity λ we need to assess the effect of considering only single arrivals in a time interval with minimum t_0 distance between the last rngs in interval M and the first rngs in interval M+1. The distribution of the times between one arrival and the following one for a Poisson process is given in Eq. B.3 as:

$$f(t) = \lambda\, e^{-\lambda t}.$$

The integral over all times, t, from 0 to ∞ gives 1. It means that the probability of having any possible elapsed time between one arrival and the following one is unity. If we exclude all arrivals that have inter-arrival time lower than t_0 and integrate between

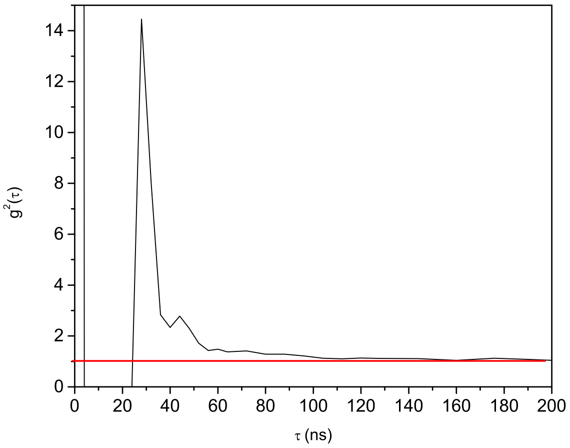


Figure 4.4: Autocorrelation function ($g^2(\tau)$) of the Si SPAD signal (peak at zero is out of scale). Dead time and afterpulsing distribution of the detector can be seen here. © 2017, IEEE. Reprinted, with permission, from [122].

t_0 and infinity, we obtain $e^{-\lambda t_0}$. We can therefore define the average number of arrivals that have inter-arrival time greater than t_0 by multiplying the above probability by the average arrivals per unit time for the Poisson process:

$$\lambda_{t_0} = \lambda e^{-\lambda t_0} \tag{4.2}$$

It maximizes at $\lambda = 1/t_0$ with the value of λ/e. If λ is 1.56 Mcounts/s (corresponding to 1/640 ns), according to the theory (Eq. 4.2) with t_0=320 ns the bit rate would be 3.7 Mbps. However, the experimental bit rate reaches 1.68 Mbps. This discrepancy depends on the losses due to N subintervals and on the detector dead time. Figure 4.5 shows the experimental bit rate at different counting rates. It reaches a maximum of about 1.68 Mbps at the counting rate of 1.88 Mcounts/s. At higher counting rates, the bit rate decreases due to higher discards of multiple arrivals in the time intervals.

4.4 Results and discussion

Using the setup shown in Fig. **??**, long datasets were generated at different counting rates—equivalent to different applied currents to the Si-NCs LED—and different temperatures. The minimum counting rate is defined as the minimum count rate required to have the buffer of the FPGA fully written. The results and discussion are presented here.

4.4.1 Quality of random numbers

It can be seen in Fig. 4.6 that the raw data, generated by the FPGA according to the procedure described, follow a nearly uniform distribution which fits the expectation. Indeed,

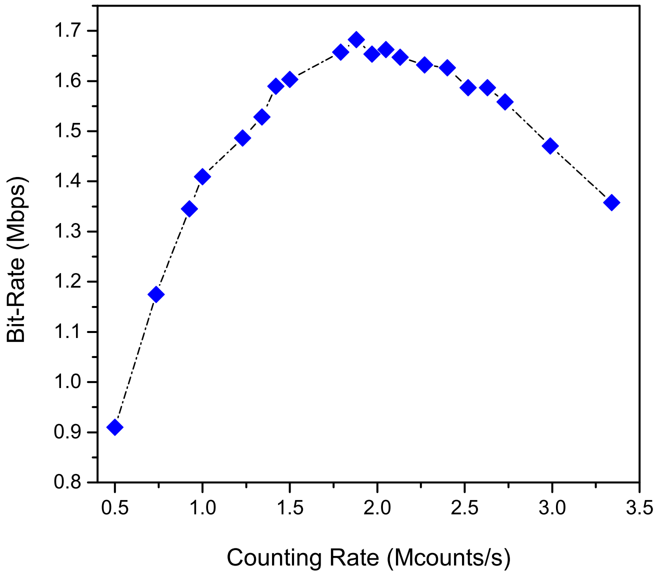


Figure 4.5: The experimental bit rate at different counting rates. It reaches a maximum of about 1.68 Mbps at the counting rate of 1.88 Mcounts/s. At higher counting rates, the bit rate decreases due to higher discards of multiple arrivals in the time intervals. © 2017, IEEE. Reprinted, with permission, from [122].

in the ideal case, the theoretical value for the probability distribution of 16 bin symbols is 1/16 (indicated by a red solid line in Fig. 4.6).

The generated raw data show high quality of randomness. JPMF (Section 2.3.3) is used to look for a potential weakness of this method. The analysis of JPMF shows a very low deviation in the order of $\sim 10^{-6}$ from the expected theoretical value of $(1/16) \times (1/16) = 0.00390625$ (Fig. 4.7). The MI (Section 2.3.4) of the generated random symbols calculated by Eq. 2.11 is $\sim 10^{-7}$ bits considering 1G random symbols.

To test the robustness of our method, we arbitrarily changed the LED driving current or the LED temperature to change the emitted flux of photons. The min-entropy (described in Section 2.3.5) of the raw data taken at different counting rates and temperatures is represented in Fig.4.8 (a) and (b), respectively. We observe that although it is slightly affected by the change in the counting rate of the photon flux and by the temperature variation, the values of the min-entropy are in the range of 3.99907-3.99972 bits per hexadecimal digit (a nibble or 4-bits). This shows the high efficiency of our methodology with respect to entropy. The maximum bias (explained in Section 2.3.6) is calculated to be in the order of $\sim 10^{-5}$.

As mentioned before the legitimate detection of a photon in each subinterval produces 4 bits. Replacing each symbol with its corresponding 4-bit binary values, long sequences of zeros and ones are obtained.

The probability of having ones (zeros) is unaffected by the change of the photon flux or

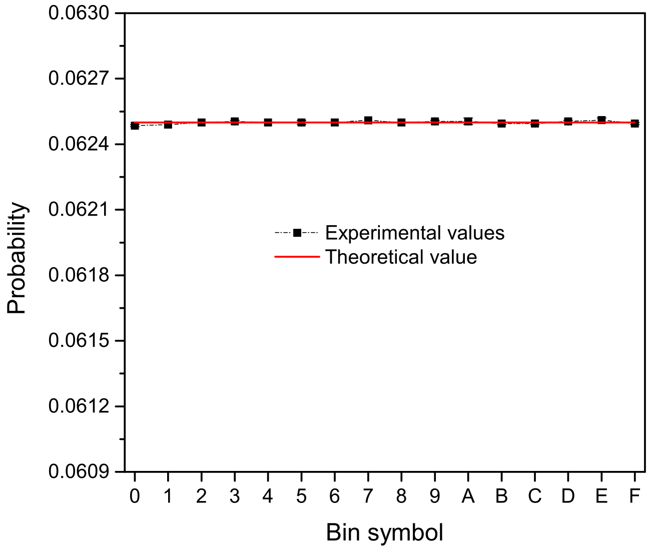


Figure 4.6: Probability distribution of the 16 bin symbols built by analyzing 1 G symbols raw data. The filled squares refer to the experimental values acquired using the target function in Fig. 4.1 (c) and the solid red line to the theoretical value of 1/16=0.0625. The error bars are placed inside the squares. © 2017, IEEE. Reprinted, with permission, from [122].

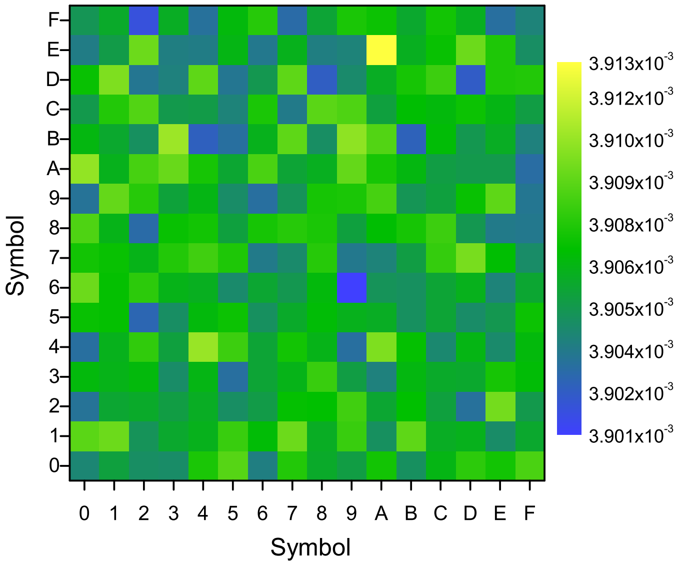


Figure 4.7: Joint probability mass function (JPMF) for 1G generated symbols showing the probability of having each symbol followed by the other one. There is a very low deviation in the order of $\sim 10^{-6}$ from the expected theoretical value of $(1/16) \times (1/16) = 0.00390625$. © 2017, IEEE. Reprinted, with permission, from [122].

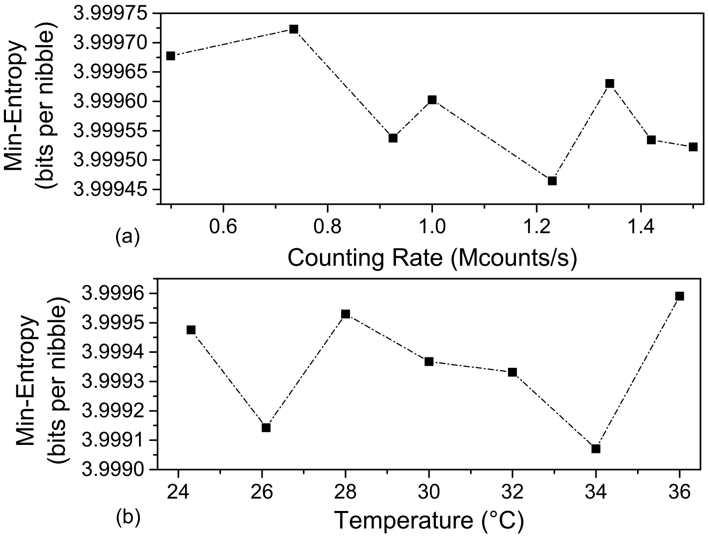


Figure 4.8: Min-entropy of raw data sequences each containing 500 Msymbols taken at (a) different counting rates and (b) different temperatures. © 2017, IEEE. Reprinted, with permission, from [122].

of the temperature (Fig. 4.9).

4.4.2 Statistical tests

4.4.2.1 NIST tests

We apply the 15 statistical tests in NIST tests suite described in Section 2.3.7.1 to the generated raw data. Various datasets with 1 to 10 Gbits length at different applied currents to the LED—from the minimum counting rate to the maximum—and at different temperatures (24°C - 36°C) were obtained. They all passed the NIST tests without the application of a post processing algorithm irrespective of the EL variations of the LED during data acquisition. The results for a dataset of 10 Gbits at the EL intensity of 1.5 Mcounts/s are reported in Table 4.1.

4.4.2.2 Alphabit battery

As previously described in Section 2.3.7.1, the Alphabit battery consists of 17 statistical tests designed primarily for hardware random bits generators. It is much faster than the NIST tests suite; it takes about 1 minute for 2^{30} bits of data. [83]

We considered 200 datasets each of 1 Gbits length. The calculated *P-values* of the tests are presented in Fig. 4.10. In order to pass a test, the *P-value* has to be in the range of [0.001 , 0.9990]. If the deviation of *P-value* ($\Delta_{P-value}$=min{1−P-value, P-value}) is in the range of [10^{-6} , 10^{-2}], the test is considered inconclusive or weak and in the range of

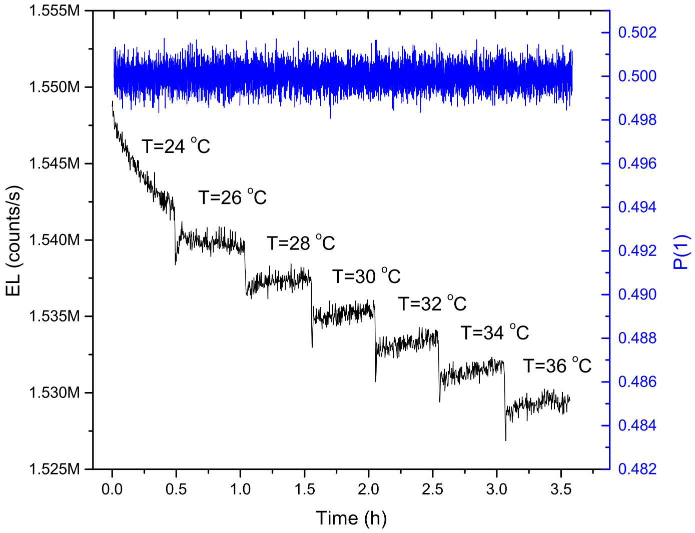


Figure 4.9: The EL variation at different temperatures (black), the probability of ones (blue) can be seen on the right of the plot. © 2017, IEEE. Reprinted, with permission, from [122].

Table 4.1: NIST tests results for 10G random bits (10^{10} bits). The significance level is α=0.01. In order to pass, the p-value$_T$ should be larger than 0.0001 and the minimum proportion should be 0.987. © 2017, IEEE. Reprinted, with permission, from [122].

Statistical test	**P-value$_T$**	**Proportion**	**Result**
Frequency	0.662506	0.9892	Passed
Block frequency	0.072289	0.9916	Passed
Cumulative sum	0.677444	0.9894	Passed
Runs	0.738917	0.9894	Passed
Longest run	0.067300	0.9910	Passed
Rank	0.322594	0.9910	Passed
FFT	0.291282	0.9870	Passed
Non overlapping template	0.581082	0.9909	Passed
Overlapping template	0.268110	0.9891	Passed
Universal	0.334077	0.9878	Passed
Approximate entropy	0.076564	0.9893	Passed
Random excursions	0.155778	0.9926	Passed
Random excursions variant	0.516352	0.9880	Passed
Serial	0.020945	0.9897	Passed
Linear complexity	0.025108	0.9902	Passed

$[10^{-15}, 10^{-6}]$ it fails. [128]

The deviation of *P-value* for 4 tests (each one in a dataset) was $\sim 10^{-4}$ and hence they were considered weak. However, they are passed for all the other datasets. Statistically speaking, if we get the weak result for only one dataset out of 200, we can safely say that the test is passed consistently and the QRNG is considered ideal. All the tests were passed for all the other datasets.

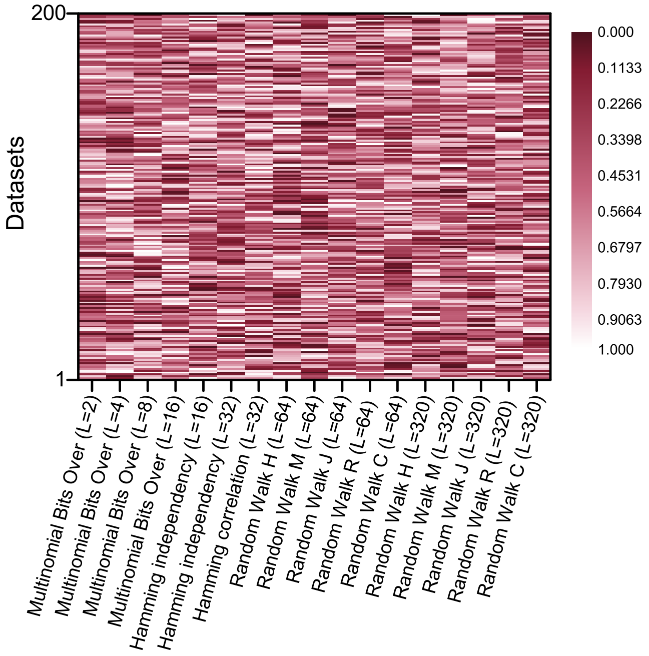


Figure 4.10: The *P-value* of the 17 statistical tests in TestU01 Alphabit battery for 200 datasets each of 1 Gbits length. © 2017, IEEE. Reprinted, with permission, from [122].

4.5 Conclusions

We developed a robust methodology to generate quantum random numbers. The source of entropy is a Si-NCs LED coupled with a Si SPAD connected to an FPGA to extract random numbers. So far in the literature, timing information of the photon arrivals has been utilized to generate random bits through different approaches. However, the lack of a robust methodology with a complete study of the detector imperfections and a simple setup to generate random numbers has been evident. The methodology developed, tested and presented here masks all the defects of afterpulsing, dead time and jitter of the Si SPAD and is effectively insensitive to ageing of the LED and its emission drifts related to temperature variations. A simple, integrable setup is used to produce sequences of random numbers.

Analyses of JPMF, MI and min-entropy show the high quality of generated random numbers and the high efficiency of the methodology. Despite the variations of the LED emission intensity, the system is efficient in producing long bit sequences maintaining the high quality of random numbers.

The raw data pass all the statistical tests in NIST tests suite and TestU01 Alphabit battery without a post processing algorithm. The maximum demonstrated bit rate is 1.68 Mbps with the efficiency of 4-bits per detected photon.

The bit rate can be increased by reducing the subintervals duration, optimizing the number of symbols per interval and decreasing the duration of the Ns between the random number generating subintervals, according to improved photodetector parameters. All these factors enhance the bit rate at the expenses of a more complex and expensive system. Alternatively, parallelization can be employed to improve the generation rate; high bit rate can be achieved by multiple Si-NCs LED/Si SPAD in a single chip. High density integration will benefit from the single detector structure and the use of very simple analog electronics not requiring silicon area for adjustment circuitries.

Chapter 5

A compact configuration with Si-NCs large LED and Si photomultiplier

In almost all photonic/optical QRNG a bulky setup is used to generate random numbers. A small-sized QRNG, easy to be implemented in small electronic devices such as mobile phones and cameras for secure data encryption and decryption as well as other applications, is highly essential for facile accessibility to everyone. In this chapter, we present a QRNG with a novel, compact configuration comprising a Si nanocrystals large area LED (Si-NCs LLED) coupled with a Si photomultiplier (SiPM) in free space to generate high quality random numbers. First, we describe Si-NCs LLED and its electrical and optical properties. We will then explain the SiPM. The experimental setup to generate random numbers will be presented. At the end of the chapter, we will demonstrate the results and discussion with a comparison between the compact configuration with Si-NCs LLED and SiPM in this chapter and the bulky configuration in chapter 4.

5.1 Si-NCs large area LED

In order to get closer to the final goal of Work Package 4 of project SiQuro that aims to make a compact all-Si based QRNG, we made up a setup comprising a Si-NCs large area LED (LLED) and a Si photomultiplier (SiPM) (Fig. 5.1). Si-NCs LLEDs were fabricated in Bruno Kessler Foundation (FBK) in order to illuminate arrays of SPADs (which will be described in the next chapter) and large area detectors (like the 1 mm×1 mm SiPM we use here). The large area of the Si-NCs LLED allows the coupling with and uniform illumination of large area, more efficient CMOS detectors. The Si-NCs LLED has a multilayer structure with 5 periods of SRO/SiO_2 layers of 3.5-4 nm and 2 nm, respectively. The fabrication process of the Si-NCs LLED is the same as that described in section 3.1.

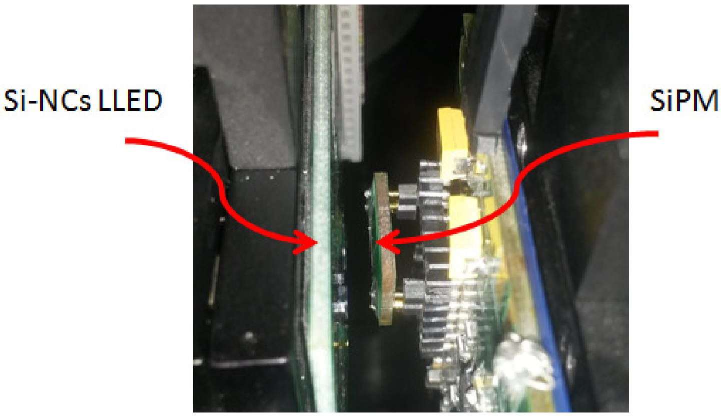


Figure 5.1: The compact configuration of Si-NCs LLED and SiPM.

The Si-NCs LLED we used here have been prepared in three different sizes (big (b), medium (m) and small (s) with gate surface area of 1.3 mmx0.99 mm, 0.99 mmx0.82 mm and 1.02 mmx0.11 mm, respectively) shown in Fig 5.2 (a). Four sets of Si-NCs LLED as in Fig. 5.2 (a) were bonded on a printed circuit board (PCB) (Fig. 5.2 (b)) for the measurements of the compact configuration.

The EL spectra of these LLEDs (b, m and s) can be seen in Fig. 5.3 with a high peak at $\sim$ 850 nm attributed to the emission from Si-NCs. The measurements for random number generation in this chapter were performed on the m Si-NCs LLED. As can be seen in Table 5.1, the m Si-NCs LLED show higher responsivity (computed as EL over the applied current density (EL/J)) than the b and s LLED.

The applied current densities to the LLEDs in Table 5.1 correspond to the voltages of 2.52 V, 2.34 V and 3 V to the b, m and s LLEDs, respectively. The electrical power density is calculated to be 5.87, 0.93 and 0.08 mW/cm^2 for the b, m and s LLEDs, respectively. It should be noted that the applied currents to the b, m and s LLEDs are 30, 3 and 3 μA, respectively. At lower currents than 30μA to the b LLED, no appreciable EL is observed. Therefore, by applying the previously-mentioned currents to the LLEDs, we tried to keep the voltages and hence the electric field through the active area of the LLEDs (with actual thickness of $\sim$22.5 nm) more or less the same. The low current density and high responsivity of the m LLED yield the higher efficiency of this LLED compared with the b and s LLEDs. In addition to the efficiency, the active area of m Si-NCs LLED allows the suitable coupling with the large area SiPM with the dimension of 1 mm$\times$1 mm.

The current-voltage (I/V) characteristics of Si-NCs LLED are presented in Fig. 5.4. They show a quite rectifying behavior with more current density at forward regime–i.e. negative voltage applied to the cathode and zero voltage to the anode–than the reverse regime–i.e. positive voltage applied to the cathode and zero voltage to the anode. It is observed that at a fixed forward voltage, the current density through the b LLED is larger than m and s

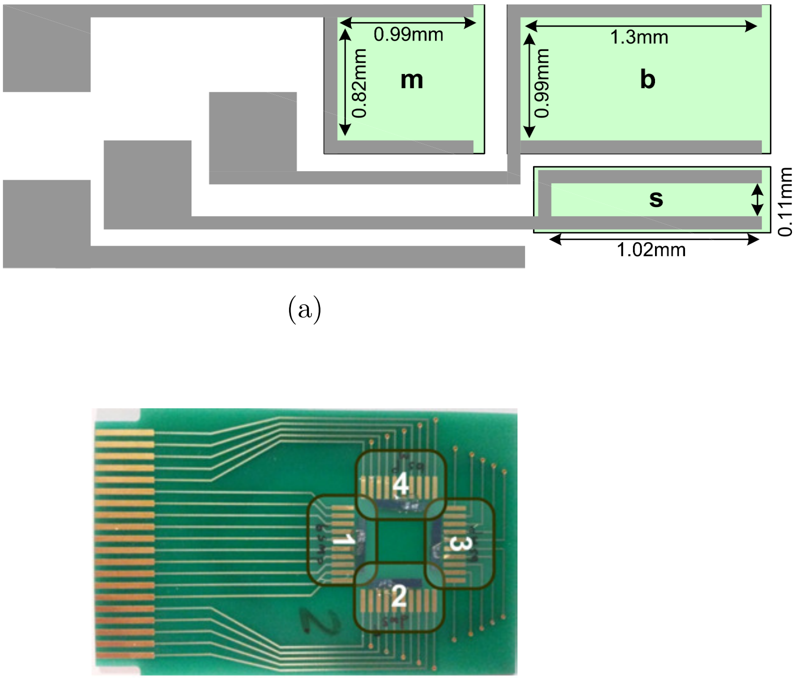


Figure 5.2: (a) The geometry of Si-NCs LLED with different sizes b (big), m (medium) and s (small) and (b) 4 sets of Si-NCs LLED as in (a) bonded on a printed circuit board (PCB).

Table 5.1: The responsivity (EL over current density) of b, m and s Si-NCs LLED.

LLED	**Current Density (mA/cm^2)**	**EL (kc/s)**	**Responsivity (G.cm^2/s.A)**
b	2.33	∼533	∼0.23
m	0.37	∼363	∼0.98
s	2.67	∼364	∼0.14

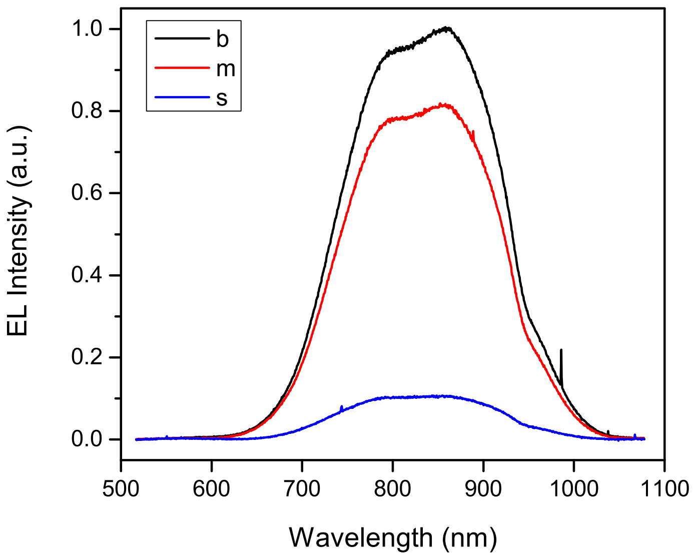


Figure 5.3: The EL spectra of the three different sizes of Si-NCs LLED illustrated in Fig. 5.2 (a) at the applied voltage of -3 V to the cathode of the LLEDs. A high peak at ~850 nm can be seen that is attributed to the emission from Si-NCs.

LLEDs (particularly at 0.5-3 V) that is due to larger free carrier density flowing through the active area in b than m and s LLEDs. In the reverse bias region (1-6 V), however, the b, m and s LLEDs show the same order of magnitude current densities that is related to the inefficient carrier injection to the active area by the accumulation of the charges near the boundaries of cathode and anode with the active area. This effect blocks the carriers from flowing through, recombining and contributing to the net current and consequently makes the current density independent of the gate areas of the LLEDs. [129]

5.2 Si photomultiplier

Si photomultiplier (SiPM)–more precisely analog Si photomultiplier–is an array of many (hundreds) SPADs. They are all connected in parallel to common anode and cathode, but each one with its own quenching resistor. Each cell (i.e. SPAD+resistor) is sensitive to a single photon and provides a defined current at the output. Therefore, the SiPM output is proportional to the number of triggered cells, thus to the number of detected photons. SiPM has obtained a growing attention as an alternative to the traditional photomultiplier tube in the detection of low photon fluxes, thanks to a number of advantages typical of solid state detectors, such as compactness, ruggedness, ease of use, low operational voltage and insensitivity to magnetic fields. [130] SiPM can be used for fast detection of scintillation light, e.g., in nuclear medicine and high energy physics. It allows important advancements in positron emission tomography (PET). [131, 132] Further, the good time

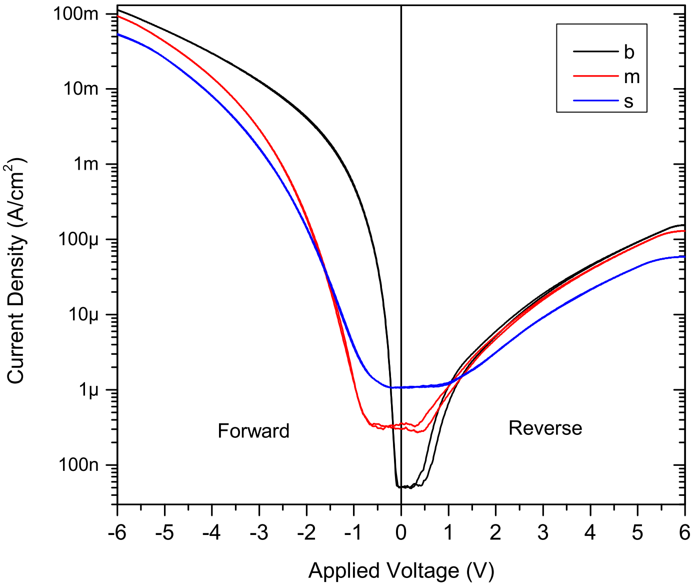
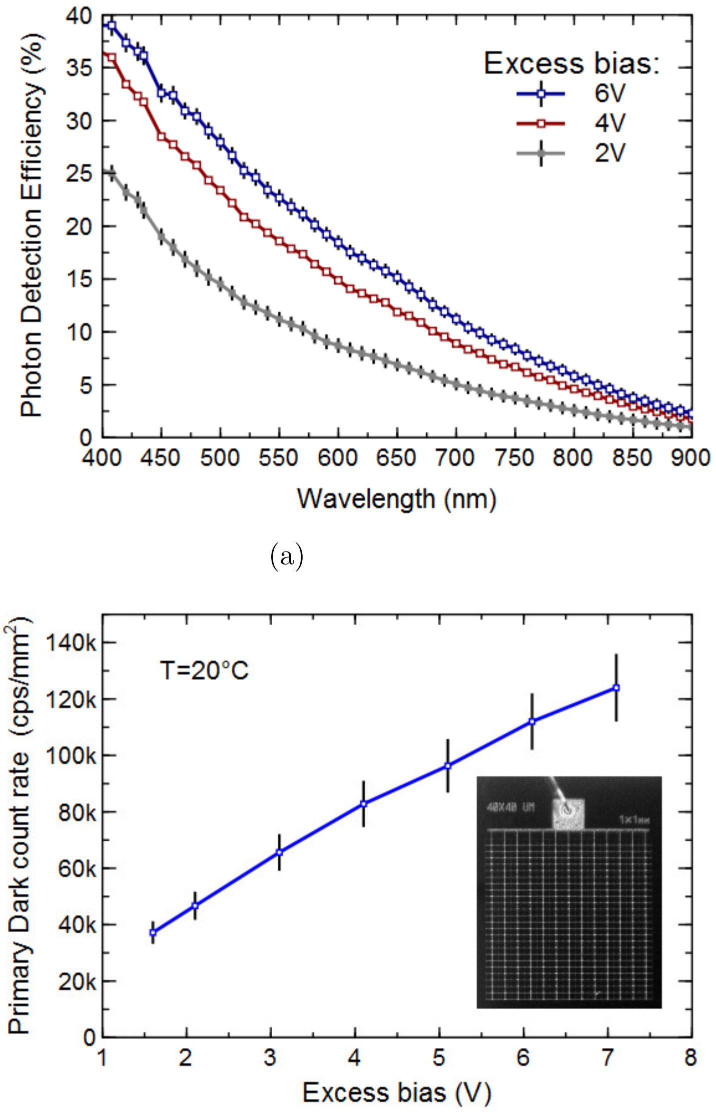


Figure 5.4: The I/V characteristics of the Si-NCs LLEDs. A quite rectifying behavior is observed in the I/V curves with more current density at forward regime–i.e. negative voltage applied to the n-type contact electrode and zero voltage to the p-type contact electrode–than reverse regime–i.e. positive voltage applied to the n-type contact electrode and zero voltage to the p-type contact electrode.

resolution permits time-of-flight PET, which improves image quality. [133] SiPM is also emerging as a very sensitive detector in many single-photon or few-photon applications in physics, biology or other fields [134], and like in the case of the work in this chapter.
Different technologies for SiPM have been developed in Bruno Kessler Foundation (FBK) during the last few years, with peak sensitivity in the green part (RGB-SiPM) or in the blue part (NUV-SiPM) of the visible spectrum, and with different cell sizes. The NUV technology, in particular, benefits from an upgraded silicon material [135], employing an epi/substrate structure with a lower-lifetime substrate. This gives particular benefits in terms of correlated noise reduction, i.e. AP and delayed crosstalk (DeCT) probability.
In this work, we employ a 1 mm×1 mm NUV SiPM (inset in Fig. 5.5 (b)), with cell size of 40 μm, with a fill factor of 60 %, thus a total number of 625 cells (i.e. SPADs). This particular technology has a photon detection efficiency (PDE) not matched to the LLED emission. PDE is about 5% at 800 nm, at 4 V of excess bias (i.e. the difference between bias and breakdown voltage), as shown in Fig. 5.5 (a). However, NUV SiPM has the advantage of a low primary DCR, less than 100 kcps/mm^2 at 5 V of excess bias (see Fig. 5.5 (b)), and a reduced correlated noise probability (overall AP+DeCT probability lower than 5%), which is very important in this kind of application.
To exploit this kind of detector for the application in QRNG, we designed a custom



(a)

(b)

Figure 5.5: (a) Photon detection efficiency (PDE) as a function of wavelength for NUV SiPM at excess bias of 2, 4 and 6 V. (b) Dark count rate (DCR) of NUV SiPM versus excess bias at he temperature of 20°C. The inset shows the 1 mm×1 mm SiPM containing 625 SPADs with cell size of 40 μm.

front-end board to amplify and digitalize the analog output signal from the detector (see Fig. 5.6). This is based on a AD8000 amplifier in a trans-impedance configuration, followed by a comparator with an adjustable voltage threshold and a monostable, creating pulses of 3.3 V and 100 ns width. This gives the maximum count rate of the detection system, which is anyhow limited by the afterpulsing time constant of the detector, giving an overall time to let all traps to empty, thus an overall time to avoid any possible afterpulsing, of few hundreds of nanoseconds, as seen in autocorrelation function. As will be explained in section 5.3, this signal is transferred to an FPGA processing unit for the generation of random symbols.

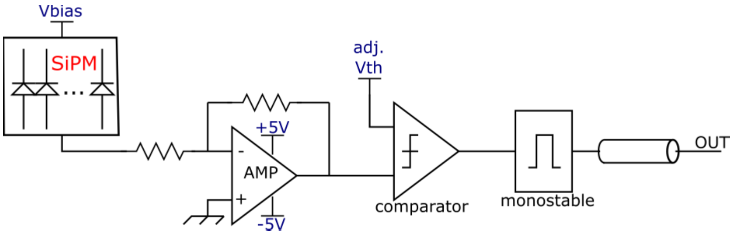


Figure 5.6: The front-end board to amplify and digitalize the analog output signal from the detector. This is based on a AD8000 amplifier in a trans-impedance configuration, followed by a comparator with an adjustable voltage threshold (adj.Vth) and a monostable, creating pulses of 3.3 V and 100 ns width.

5.3 Experimental setup

The experimental setup is schematically shown in Fig. 5.7. The Si-NCs LLED is coupled with the SiPM at a distance of ∼ 1 mm in free space without an optic or diffuser between them. The Si-NCs LLED is driven by an Agilent B1500A Semiconductor Device Parameter Analyzer. The EL of LLED is monitored by a Hewlett Packard 53131A Universal Counter 225 MHz by selecting the Totalize mode, so that true pulse counting is performed. The TTL output of the SiPM is directly connected to the high speed digital input of the FPGA. The SiPM readout board is connected to a ±5 V and a voltage of ∼30-36 V is applied to the SiPM by Agilent E3631A DC Power Supply. The EL spectra were obtained by a Spectra-Pro 2300i monochromator coupled with a nitrogen-cooled CCD camera. The measurements were performed at room temperature in a dark chamber.

The measurement of the arrival times is performed by a fully synchronous logic. The FPGA continuously samples the detector at the frequency of 100 MHz, which is crystal controlled. A valid arrival is produced by a high analog logic level heralded by one clock cycle (10 ns) of a low analog logic level. A Digilent ATLYS FPGA board has been used with the programming language VHDL.

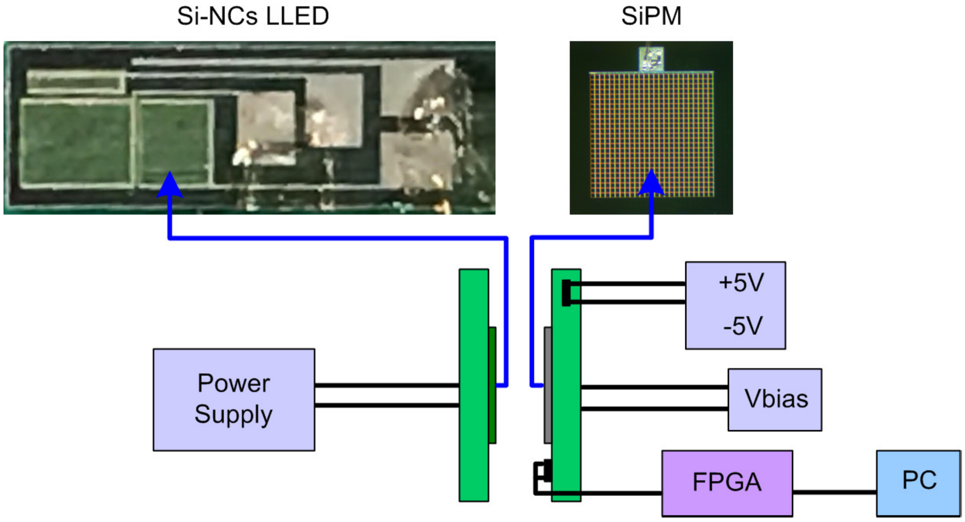


Figure 5.7: Schematic of the experimental setup with Si-NCs LLED coupled with SiPM at a distance of ∼ 1 mm. The SiPM board is connected to ±5 V and a bias voltage (Vbias) is applied to SiPM. Si-NCs LLED is driven by an applied current (voltage) provided by the power supply. The output signal of the SiPM is transmitted to FPGA connected to PC for the generation of random numbers.

The same approach as in section 4.1 is used here to extract random numbers. The autocorrelation, $g^2(\tau)$, measurement of the SiPM signal was performed via a multitau digital correlator with 4 ns resolution.[124] The afterpulsing and crosstalk distribution exhibits a main peak within 140 ns from the main autocorrelation peak at τ=0 (Fig. 5.8). The plateau in $g^2(\tau)$ approaches the normalization value of 1 at about 950 ns.
Considering the $g^2(\tau)$ (Fig. 5.8), we set the length of the "double length" intervals (see Section 4.1.2) to 640 ns, 1280 ns and 1920 ns—with the first half of Ns to be 320 ns, 640 ns and 960 ns, respectively. We see that the correlation coefficient at time lag 1 decreases from 1.29×10^{-4} to 1.45×10^{-5} with the coefficients all inside the 95% confidence interval (Fig. 5.9). It implies that 960 ns is long enough to mask the afterpulsing and crosstalk distribution of SiPM (see Fig. 5.8).

5.4 Results and discussions

The measurements for random number generation were performed on the medium Si-NCs LLED with the active area of ∼ 0.99 mm× 0.82 mm (see Fig. 5.2 (a)) suitable to be coupled with the SiPM with the dimension of 1 mm×1 mm. The applied forward current to LLED was kept below ∼ 45 μA corresponding to the forward voltage of 3 V (see Fig. 5.4), that is the boundary between the direct bipolar and the FN unipolar tunneling, in order to avoid degradation of the oxide layer in the active area of the Si-NCs LLED

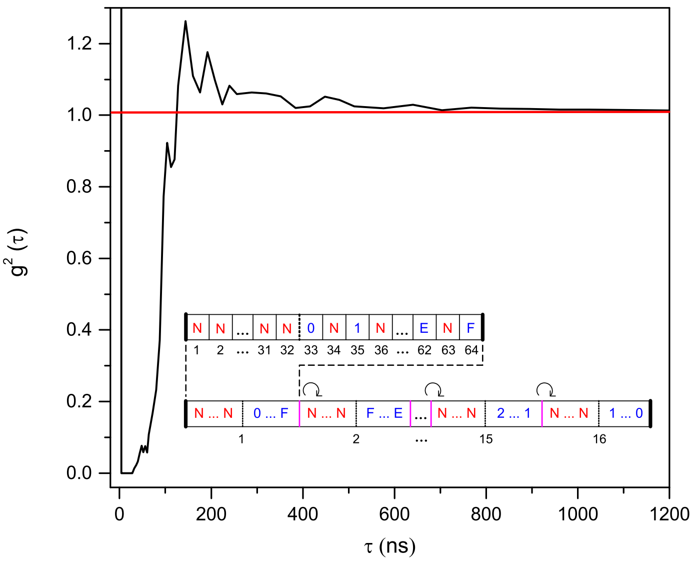


Figure 5.8: Autocorrelation function ($g^2(\tau)$) of SiPM signal (peak at zero is out of scale). Dead time and afterpulsing and crosstalk distribution of the SiPM can be seen here. The dead time of ∼110 ns is not due to limitation of the SiPM, but it is set by the monostable in the electronics (front-end shown in Fig. 5.6).

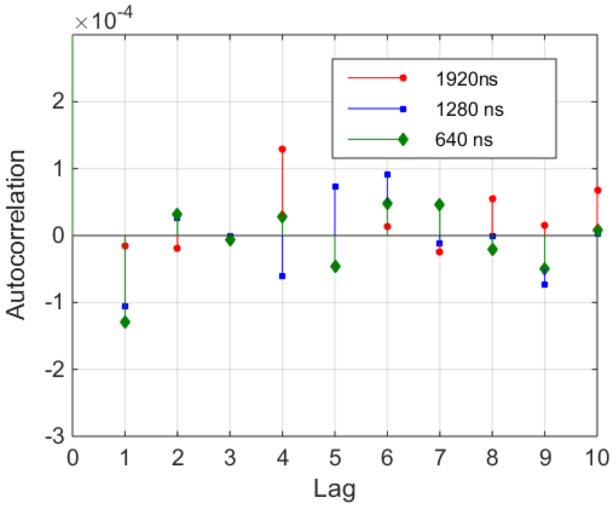


Figure 5.9: Autocorrelation coefficients for the first 10 time lags for three “double length” intervals of 640 ns, 1280 ns and 1920 ns corresponding to the first half interval of Ns with 320 ns, 640 ns and 960 ns, respectively. It is seen that the correlation coefficient at time lag 1 decreases as the length of the interval increases with all the coefficients inside the 95% confidence interval for 1920 ns “double length” interval.

(with the same structure as the multilayer Si-NCs LEDs in [89] composed of 5 periods of SRO/SiO_2 films with the nominal thicknesses of 3 nm/2 nm). The V_{bias} to SiPM was 32 V corresponding to an excess bias of ∼4 V with the DCR of ∼80 kcps/mm^2.
Setting the "double length" interval to 1920 ns and applying the same methodology in Section 4.1 we acquired long sequences of datasets. As mentioned before in Section 2.3.2, a very straightforward way to detect an observable pattern among the random symbols or codes is to create a 2-D matrix of them. A 512 × 512 2-D visualization of the 16 hexadecimal symbols is presented in Fig. 5.10. As can be seen clearly, no particular, periodic pattern is observed among the symbols.
Figure 5.11 shows the probability of the generated hexadecimal symbols in a sequence of 1 G symbols. It is seen to follow a nearly uniform distribution (the theoretical value for the probability distribution of 16 bin symbols (1/16) is indicated by a solid red line in Fig. 5.11).

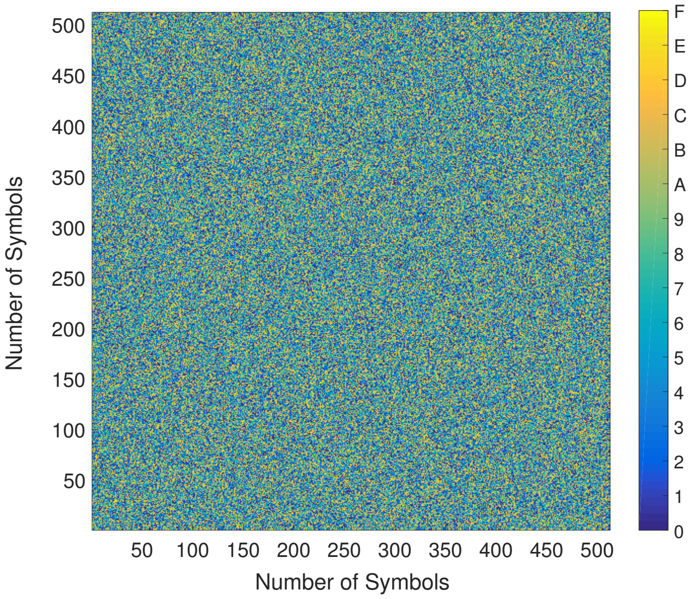


Figure 5.10: A 512×512 map of the hexadecimal symbols. No particular, periodic pattern is observable among the symbols.

The high quality of random symbols is proved through the following analyses. Joint probability mass function (JPMF) (section 4.4.1), shows a very low deviation in the order of $\sim 10^{-6}$ from the expected theoretical value of $(1/16) \times (1/16) = 0.00390625$ (Fig. 5.12). The mutual information (MI) of the generated random symbols (Section 4.4.1) is calculated to be $\sim 10^{-7}$ bits considering 1G random symbols. The maximum bias is in the order of $\sim 10^{-5}$ and the min-entropy (see Section 2.3.5) is $\sim$ 3.9997 bits per nibble or 4-bits. The highest bit rate is calculated to be $\sim$ 0.5 Mbps at the EL intensity of $\sim$ 550 kHz.

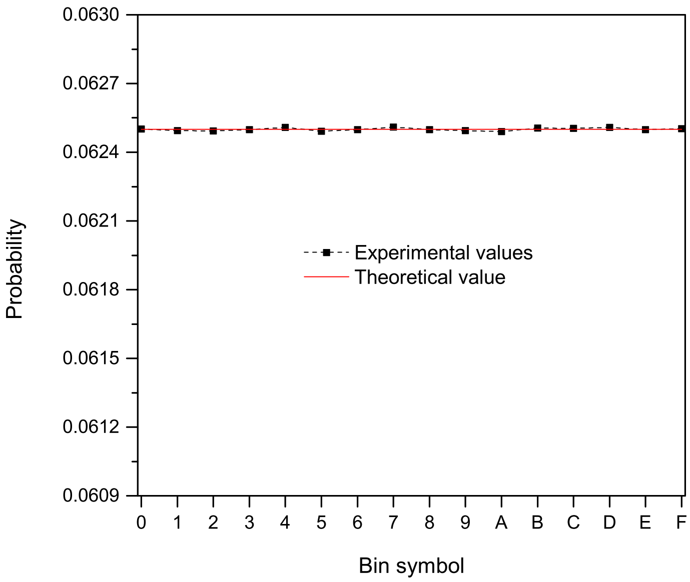


Figure 5.11: Probability distribution of 16 hexadecimal symbols in a sequence of 1 G symbols raw data. The solid red line shows the theoretical value of 1/16.

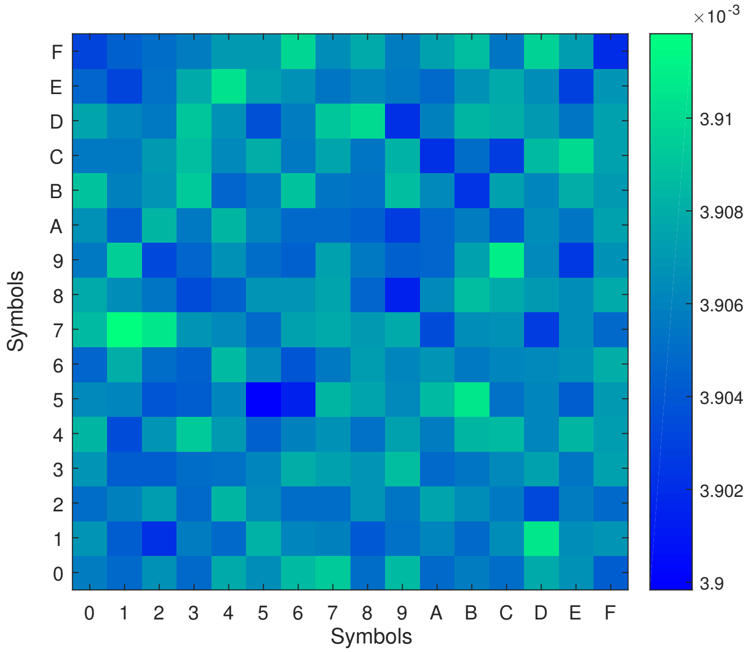


Figure 5.12: Joint probability mass function (JPMF) for 1 G generated symbols showing the probability of having each symbol followed by the other one. There is a very low deviation in the order of $\sim 10^{-6}$ from the expected theoretical value of $(1/16) \times (1/16) = 0.00390625$.

To further analyze the quality of generated random numbers, each symbol is replaced with its corresponding 4-bit binary values. We then apply the 15 statistical tests in NIST tests suite to the generated raw data. Various datasets with 1 and 2 Gbits length at different applied currents to the LLED were obtained. They all passed the NIST tests without the application of a post processing algorithm irrespective of the EL variations of the LLED during data acquisition. The results of the NIST tests for 2 Gbits of raw data are presented in Table 5.2.

Table 5.2: NIST tests results for 2 G random bits. The significance level is α=0.01. In order to pass, the p-value$_\text{T}$ should be larger than 0.0001 and the minimum proportion should be 0.983.

Statistical test	**P-value$_\mathbf{T}$**	**Proportion**	**Result**
Frequency	0.2861	0.9930	Passed
Block frequency	0.2868	0.9935	Passed
Cumulative sum	0.1657	0.9920	Passed
Runs	0.3298	0.9935	Passed
Longest run	0.4817	0.9910	Passed
Rank	0.3611	0.9860	Passed
FFT	0.0401	0.9910	Passed
Non overlapping template	0.5666	0.9905	Passed
Overlapping template	0.4064	0.9900	Passed
Universal	0.1404	0.9850	Passed
Approximate entropy	0.2854	0.9930	Passed
Random excursions	0.5310	0.9938	Passed
Random excursions variant	0.3127	0.9883	Passed
Serial	0.3376	0.9870	Passed
Linear complexity	0.2550	0.9905	Passed

A comparison of the compact configuration designed here and the bulky configuration of Chapter 4 is presented in Table 5.3. The DCR of the SiPM is much higher than the DCR of the commercial Si SPAD (80 kHz vs. 300 Hz) resulting in the average contribution of the dark counts to be ~0.15 per "double length" interval of 1920 ns length (it is ~0.0002 in the case of QRNG in Chapter 4), with the approximate signal to noise ratio of 6.88 at the maximum bit rate. Therefore, the probability of random symbol generation due to the dark counts is much higher in this case than the QRNG of Chapter 4. However, the robustness and efficiency of our methodology make possible the generation of high quality random numbers even when the contribution from the dark counts of the detector (SiPM in this case) is high.

Table 5.3: A comparison of some properties between the bulky configuration with Si-NCs LED and commercial Si SPAD (see Chapter 4) and the compact configuration with Si-NCs LLED and SiPM designed in this chapter.

Property	**Si-NCs LED+Si SPAD**	**Si-NCs LLED+SiPM**
PDE of the detector	~50% at 800 nm	~5% at 800 nm
Detection area	0.024 mm^2	1 mm^2
DCR	300 Hz	80 kHz
LED current density	0.2-0.4 mA/cm^2	0.8-1.2 mA/cm^2
Robustness	Robust	Robust
Compactness	Bulky	Compact
Min-entropy	~ 3.999 bits per 4-bits	~ 3.999 bits per 4-bits
Bias	~ 10^{-5}	~ 10^{-5}
MI	~ 10^{-7} bits	~ 10^{-7} bits
Bit rate	1.68 Mbps	0.5 Mbps

5.5 Conclusions

We developed a compact QRNG consisting of a Si-NCs LLED and a SiPM. The robust methodology introduced in Chapter 4 is found to be robust in this configuration as well. We set the “double length” interval to 1920 ns to mask the afterpulsing and crosstalk distribution of SiPM. The analyses on JPMF, MI and min-entropy show high quality of the generated random numbers. All the statistical test in the NIST tests suite pass for 2 Gbits raw data.

The compact configuration we designed and tested in this chapter, is able to generate high quality random numbers at the maximum bit rate of 0.6 Mbps. The average of dark counts per interval (of 1920 ns length) is ~0.15 which is approximately 750 times more than that of the bulky configuration of Chapter 4. Even though the contribution of the dark counts to generate random hexadecimal symbols is much larger here, the quality of random numbers is still maintained, thanks to our robust and efficient methodology.

The main advantage of the configuration designed in this chapter is the compactness, with Si-NCs LLED coupled closely with large area SiPM in free space, which allows the fabrication of a more compact RNG on Si platform. The bit rate can be increased by parallelization of the Si-NCs LLEDs/SiPM. A more efficient Si large area detector, a cheaper one (like the SiPM here) than the commercial Si SPAD, which, compared to the SiPM, has a lower DCR, a lower afterpulsing distribution and higher PDE matching well the emission from Si-NCs LLEDs, would greatly increase the efficiency of our RNG and guarantee the quantum nature of the generated random numbers.

Chapter 6

A compact chip with Si-NCs large LED coupled with an array of SPADs

In this chapter, a method based on an array of 16 SPADs connected to 4 TDCs is considered to generate random numbers. A compact chip has been prepared with bonded Si-NCs LLED (Section 5.1) coupled with a bonded array of SPADs connected to TDCs. The approach to generate random numbers is based on photon arrival times. At the beginning of this chapter, the theory is explained. Then, the experimental procedure, results and discussions are presented and at the end, some conclusions are drawn.

6.1 The oversampling technique

In the architecture used in this chapter, the time information is calculated with respect to a reference signal (clock) thanks to a TDC. In this scheme, the TDC works as a digital chronometer where the event forces a start and the falling edge of the clock stops the counts. The period of the clock defines the observation window T_w and the measurement rate, meaning that for each period of the clock the detector gives at most one measurement (see Fig. 6.1).

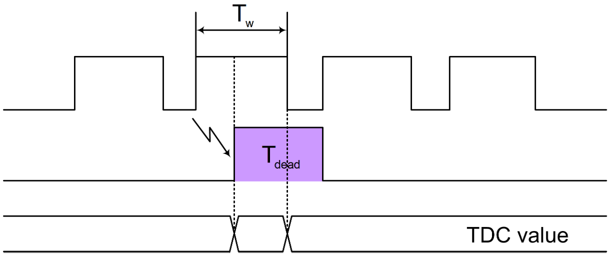


Figure 6.1: Detection scheme used in the sampling model for a detector.

The oversampling technique can be used in a condition of low flux of photons and high sampling rate of the detector, based on a high resolution TDC. From Eq. B.2, the arrival time between events is modeled by an exponential distribution. If we consider the model of the architecture described before–where the arrival time of a photon is computed with respect to a synchronous clock–and let our observation window be T_w and let its mid-point be $t_m = T_w/2$, we might then choose to digitize our photon arrival times, t_a, according to a simple output rule such as:

$$\begin{cases} 0 & \text{if } t_a \in [0 \;\; t_m) \\ 1 & \text{if } t_a \in [t_m \; T_w) \end{cases}$$

The output would have some bias written precisely using the exponential distribution as:

$$\delta = |P(0) - P(1)| = \int_0^{t_m} f(t;\lambda)dt - \int_{t_m}^{T_w} f(t;\lambda)dt \tag{6.1}$$

where $f(t;\lambda)$ is the probability density function (pdf) of an exponential random variable with rate parameter λ at t (see Appendix. B). The first integral is greater than the second because $T_w = 2t_m$, so we can drop the absolute value. This can be easily rewritten as:

$$\begin{aligned} \delta &= 2\int_0^{t_m} f(t;\lambda) - \int_0^{T_w} f(t;\lambda) \\ &= 2\xi(t_m;\lambda) - \xi(T_w;\lambda) \\ &= 2\left(1 - e^{-\frac{\lambda T_w}{2}}\right) - \left(1 - e^{-\lambda T_w}\right) \end{aligned} \tag{6.2}$$

where $\xi(t;\lambda)$ is the cumulative distribution function (cdf) of an exponential random variable with rate parameter λ at t and is equal to $1 - e^{-\lambda t}$ (Eq. B.2). It can be seen clearly from the equation that the bias value δ increases with λ.

If in our approach n equi-distant intervals are considered so that more than one bit at a time is generated, the most appropriate measure of distance from uniformity would be the total variation distance (TVD) rather than the bias. It is computed as [136]:

$$TVD(P,Q) = ||P(x) - Q(x)||_1 = \frac{1}{2}\sum_x |P(x) - Q(x)| \tag{6.3}$$

where P and Q are two probability measures defined on the space Ω of measurable events x. If the TDC resolution is n codes each one with the probability of p(i) i=1 ... n, we can rewrite the TVD in Eq. 6.3 as:

$$TVD = \frac{1}{2}\sum_{i=1}^{n} \left| p(i) - \frac{1}{n} \right| \tag{6.4}$$

As seen before in Eq. 6.2, the bias reduction is almost directly proportional to the reduction in the photon count; in other words, reducing the TVD by an order of magnitude also reduces the count by the same. When $\lambda T_w < 1$ the photon rate is lower than the sampling frequency and we are in weak oversampling region. For values of λT_w lower than 0.1, we are in moderate oversampling region. In this case the exponential distribution can be approximated by a linear decay, meaning that the resulting codes suffer from a bias. When the detector, instead, is working in strong oversampling region, λT_w is lower than 0.01 and the codes produced by the TDC have an approximately uniform distribution.

6.2 Architecture

In this section the structures of the detection and random module as well as source of entropy are presented.

6.2.1 Detection and random module

The architecture under study consists of 18 pixels: the 16 central pixels are connected to the compression tree circuit while dummy pixels in first and last positions are used only for testing purpose. The compression tree is connected to a block of 4 TDCs for time information plus an address finder for recovering the spatial information (Fig. 6.2).

The detector works in this way: internal logic into each pixel allows to enable all SPADs at the same time for a specific observation window $\mathrm{T_w}$. In this period of time the detector can acquire and process photons. The working regime is set in order to detect a small amount of photons per $\mathrm{T_w}$, then, typically, only one of the SPADs triggers. Let us suppose for example that pixel number 8 detects a photon. The front end circuit generates, at the detection, an electrical pulse which propagates along a network (compression tree network of Fig. 6.2), which connects all SPADs of the array to the TDC block. This block determines the arrival time of the photon as soon as the pulse is detected. In the meanwhile another circuit allows determining the position of the pixel has fired (in this case the position is 8). At the end of the measurement the output of the detector will deliver two information: the arrival time of the photon and the position of the pixel that has fired.

In case of multiple detections within the same temporal window $\mathrm{T_w}$, the architecture allows to trigger the first TDC, at the presence of the first photon, then the second TDC when a second photon is detected. The architecture is able to manage up to 4 events. In case of empty measurement (no photon) or more than 4 photons (full condition) a flag is enabled: EMPTY and EXFULL. When High, EMPTY signal informs that the memory is empty. This flag can be useful to fast monitor a presence of photon detection in $\mathrm{T_w}$ or to indicate, during the readout, when the memory has been completely delivered. Flag EXFULL (High) inform about the presence of a missed photon, meaning that the number of TDC was not suitable to address the detection of photons. Figure 6.3 shows

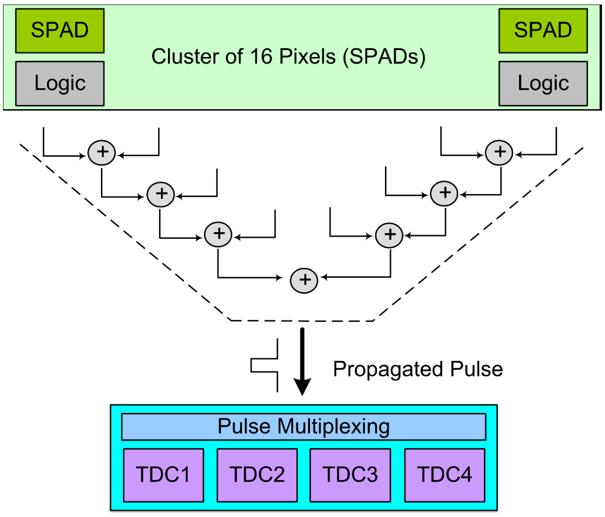


Figure 6.2: Block diagram of the architecture based on the oversampling technique.

the oversampling architecture with maximum 2 TDCs configuration: the third photon enables the EXFULL flag.

Depending on signal configuration (see Table 6.1), the structure can work with 1 up to 4 TDCs working in T_w (NTDC signals). Moreover an anti-collision circuit can be activated in order to avoid the superposition of events. In case of multiple detections, it is clear that information about the arrival times is ordered, meaning that e.g. TDC1 is always lower than TDC2 and so on for all four TDCs. In order to avoid this type of correlation, we can use a second source of entropy which is given by the pixel address: the TDC values are then ordered by the pixel address.

Table 6.1: Signal configuration of the detection and random module architecture of the compact chip.

Signal	**Configuration**	**Description**
MODE	Low	Counting Mode
MODE	High	Oversampling Mode
CONTROL	High	Pixel with Anti-Collision Circuit
NTDC$< 0 >$ NTDC$< 1 >$	Low High	Maximum Number of TDC=1
NTDC$< 0 >$ NTDC$< 1 >$	Low Low	Maximum Number of TDC=2
NTDC$< 0 >$ NTDC$< 1 >$	High High	Maximum Number of TDC=3
NTDC$< 0 >$ NTDC$< 1 >$	High Low	Maximum Number of TDC=4

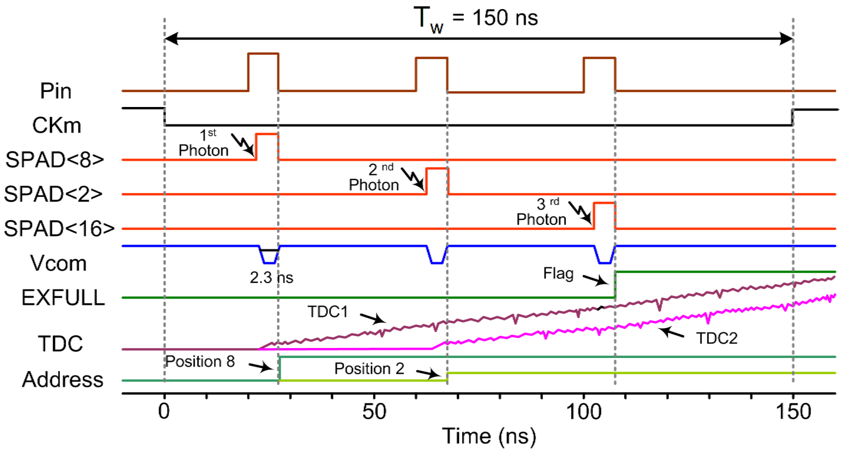


Figure 6.3: simulation of the oversampling architecture imposing a maximum number of TDC equal to two. As a consequence of this setting the third photon will be discarded and the EXFULL flag is activated indicating that the number of TDC was not suitable to address the detection of photons. CKm is the synchronous main clock, Vcom is a common line which is pulled down as soon as a SPAD triggers avoiding other SPADs to be propagated through the network, and P_{in} is the input of the block of 4 TDCs which is the pulse propagated through the compression tree network.

6.2.2 Source of entropy

The source of entropy is Si-NCs LLED with size s (see Section 5.1) coupled with the array of 16 SPADs matching quite well the size of the surface area of the pixels. The compact chip mounted on a PCB board can be seen in Fig. 6.4. The backside of the bonded LLEDs (a diced section containing s, m and b Si-NCs LLED) is visible in the figure. The Si-NCs LLED s on the front side is coupled with the cluster of 16 pixels (SPADs) connected to 4 TDCs. The LLED and SPADs are biased by two power supplies and the main board, on which the compact chip is mounted, is controlled by National Intstruments (NI) Labview programs for counting and random modes.

6.3 Data readout

The main clock (PRE) allows precharging each SPAD at every cycle of measurement. During the counting mode, after a reset period at the beginning (where PRE=High), the signal is forced to be Low, being the dead time mechanism originated asynchronously by the counter block.

Timing diagram of Fig. 6.5 describes how to implement different waveforms in this modality. In this configuration signal MODE=Low and CNTENABLE allows to count the SPAD into each counter. At the same time CNTENABLE defines the integration time meaning

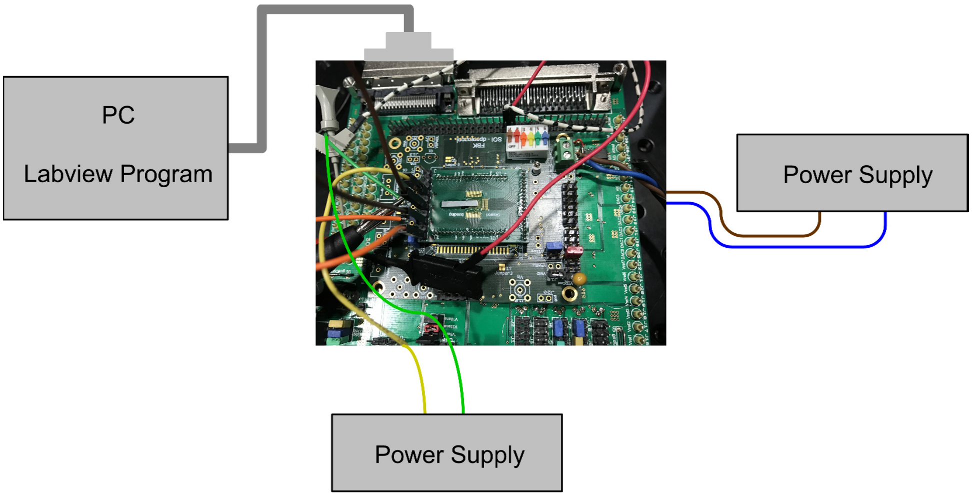


Figure 6.4: The compact chip comprising a Si-NCs LLED as the source of entropy coupled with an array of 16 SPADs connected to 4 TDCs.

that, when High disables the buffers and stops the counting.

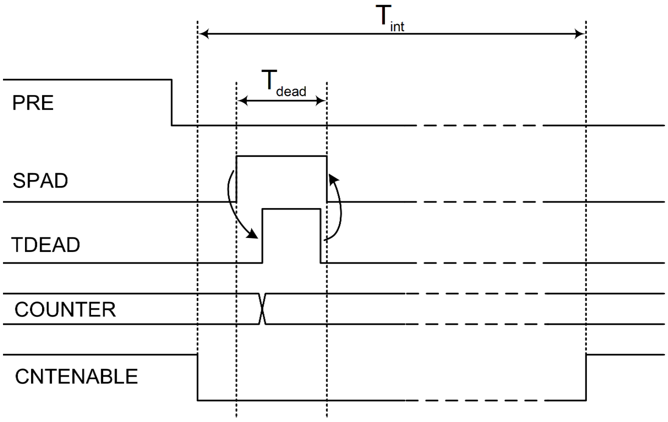


Figure 6.5: Timing diagram for counting mode.

When working in counting mode (see Fig. 6.6), the detector has to deliver the content of the counters. After reset (RN =Low, RN is the digital reset) and the integration time T_{int} defined by signal CNTENABLE, all outputs should be forced to gnd! (digital ground). Signal MODE has to be forced to L in order to properly set the output multiplexers. Signal CKCNT is used in order to scan all counters. We need 18 iterations in order to readout all counters (8 parallel bits). (Fig. 6.6).
In random number generation, the output of the chip is made of 8 (for pixel address and TDC value) plus 2 bits for chip empty and full flag. The saved file is organized in this way:

1. For each frame (a single temporal window T_w) 4 TDC values are acquired (regardless how many photons are acquired);

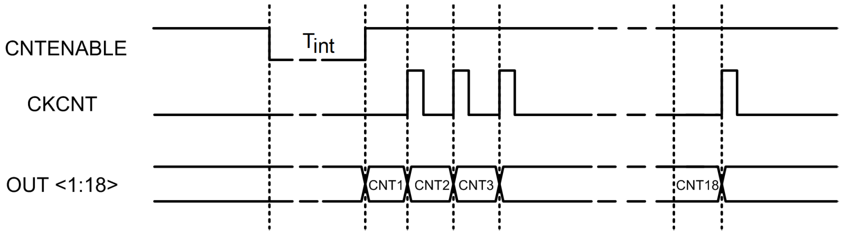


Figure 6.6: Timing diagram for readout counting mode. The integration time T_{int} is defined by signal CNTENABLE. Signal CKCNT is used in order to scan all counters.

2. Every TDC value consists of two words (=32 bits), giving a total of 8 words per frame;

3. The TDC and pixel address information (2 words) is divided in 4 bytes:

   a. First byte contains information about the LSB part of the TDC (4 bis) + pixel address (4 bits)

   b. Second byte contains bit FULL and bit EMPTY, giving information about the status of the measurement;

   c. Third byte has the TDC MSB (8 bits);

   d. Fourth byte contains again bit FULL and bit EMPTY.

First and the second word structures can be seen in Table 6.2. The TDC value is then calculated as:

$$\text{TDC} = \text{MSB} \times 12 + \text{LSB}$$

Flag information:

Table 6.2: First and second word structures containing TDC and pixel address information.

First Word															
LSB of the TDC				Pixel Address				Flags		Not Considered					
LSB_4	LSB_3	LSB_2	LSB_1	Add_4	Add_3	Add_2	Add_1	Full	Empty	–	–	–	–	–	–

Second Word															
MSB of the TDC								Flags		Not Considered					
MSB_8	MSB_7	MSB_6	MSB_5	MSB_4	MSB_3	MSB_2	MSB_1	Full	Empty	–	–	–	–	–	–

When one of the two flag is asserted (=1) a full or empty condition is detected. In particular the Empty flag can help us understand which TDC values have sampled real value and which not.
We can have these different cases:

1. Empty=1 on the first word of the first TDC → all 4 TDCs of the frame has to be discarded;

2. Empty=1 on the second word of one of the TDC value → represents the last valid TDC value of the frame.

Examples:

In the following frame (4 TDCs):
0101 0111 0100 0000   0000 0000 0100 0000   1st TDC empty flag on first word →discarded
0010 0011 0100 0000   0000 0000 0100 0000   2nd TDC → discarded
1000 1011 0100 0000   0000 0000 0100 0000   3rd TDC → discarded
0111 0001 0100 0000   0000 0000 0100 0000   4th TDC → discarded

In the following frame (4 TDCs):
0101 0111 0000 0000   0101 1000 0100 0000   1st TDC empty flag on second word → last valid TDC of the frame
0010 0011 0000 0000   0000 0000 0000 0000   2nd TDC → discarded
1000 1011 0100 0000   0000 0000 0000 0000   3rd TDC → discarded
0111 0001 0100 0000   0000 0000 0000 0000   4th TDC → discarded

In the following frame (4 TDCs):
0101 0111 0000 0000   0101 1000 0000 0000   1st TDC → TDC=88x12+5=1061   Address=7
0010 0011 0000 0000   1100 0100 0000 0000   2nd TDC → TDC=196x12+2=2354   Address=3
1000 1011 0100 0000   0001 0101 0100 0000   3rd TDC → TDC=21x12+8=260   Address=11
0111 0001 0100 0000   0000 0000 0000 0000   4th TDC → discarded

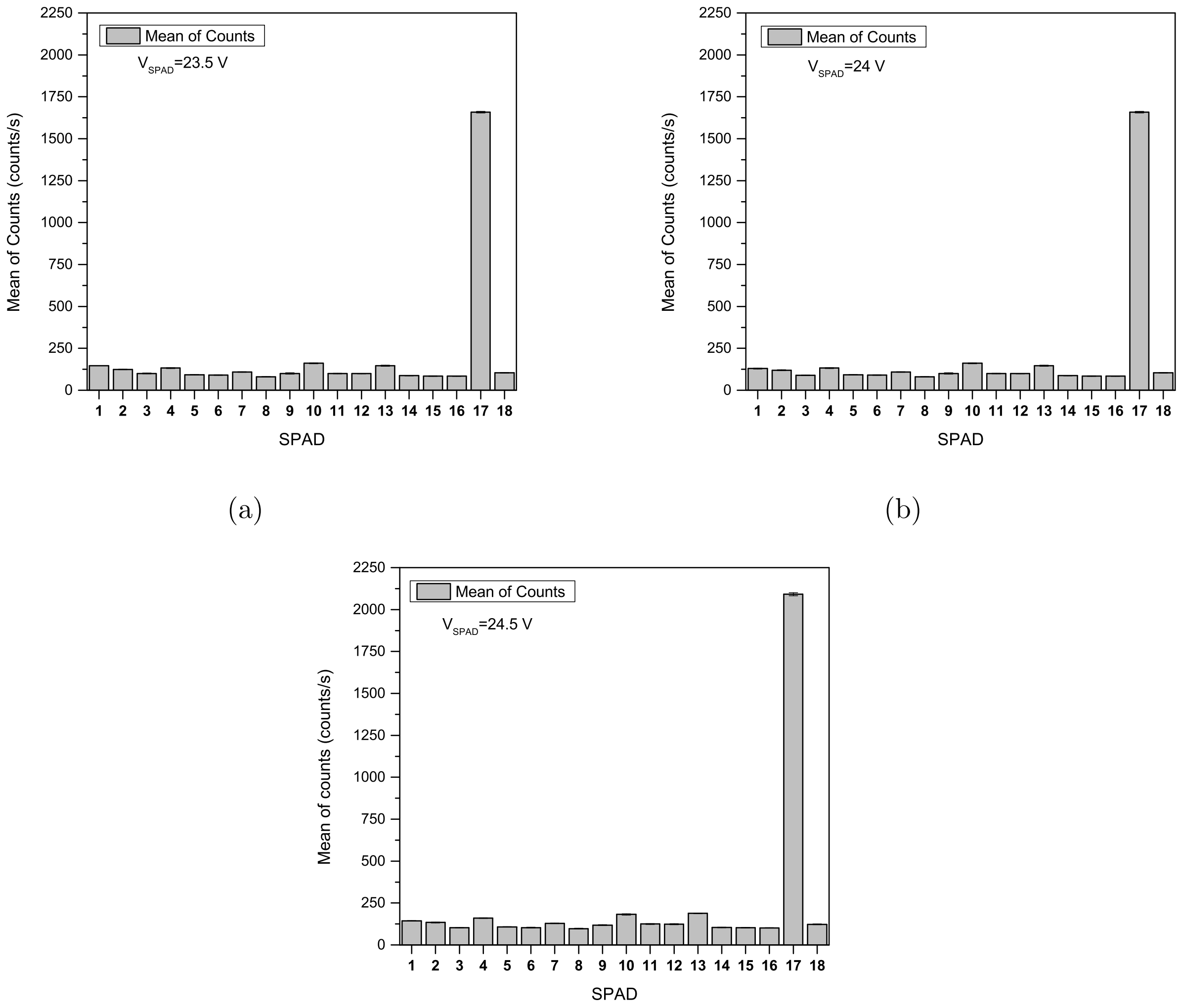
6.4 Experimental procedure

Using the experimental setup schematically shown in Fig. 6.4, we changed some parameters such as the voltage bias to the SPADs and V_c and V_{reg} which regulate the reset pulse of SPAD logic block and voltage for TDCs, respectively, in counting and random mode for data acquisition process. The details are presented in Sections 6.5.1 and 6.5.2 where the SPADs are characterized in dark and light conditions and random codes are acquired, respectively. The limits and drawbacks of the system are explained and future activities to improve the system, with respect to random number generation, are expressed.

6.5 Results and Discussion

6.5.1 SPADs characterization

The DCR of the 18 SPADs (including the first and last dummy ones) at the bias voltages of 23.5, 24 and 24.5 V are shown in Fig. 6.7 (a), (b) and (c), respectively. As can be seen clearly, the SPADs show different DCRs and SPAD17 (the last SPAD in the array of 16 pixels) shows the highest DCR. When working with all SPADs, we can turn off the SPADs which have the highest activities, i.e. highest DCR.



(a)

(b)

(c)

Figure 6.7: The DCR of the array of 18 SPADs (including 2 dummy ones) at the applied bias voltage of (a) 23.5 V, (b) 24 V and (c) 24.5 V. The error bars are shown in the figures and are visible for some data points.

When the LLED is on (at the applied voltage of 2.9 V), we can see in Fig. 6.8 (a), (b) and (c) that the SPADs show different counting rates, as would be expected from DCR analyses with SPAD17 showing the highest counting rate at all applied bias voltages of 23.5, 24 and 24.5 V to the SPADs. Moving from the middle SPADs to the beginning and ending ones, a gradual decrease in the count rate (except for SPAD17 which, as mentioned before, has the highest activity) is observed.

It should be due to two reasons: on the one hand, based on the uniformity measurements of the emission from the s Si-NCs LLEDs (see Fig. 6.9 (b)), the detected photon rate gradually decreases from position 1 to 3 (from right to left) as can be seen in Fig. 6.9 (a). On the other hand, despite the approximately equal length of the s LLED and the array of SPADs (∼1.02 mm and ∼1 mm, respectively), even a tiny misalignment between them would result in a reduced illumination of the SPADs locating further from the ones in the middle. These two reasons most probably cause such behavior in the counting rate of the SPADs.

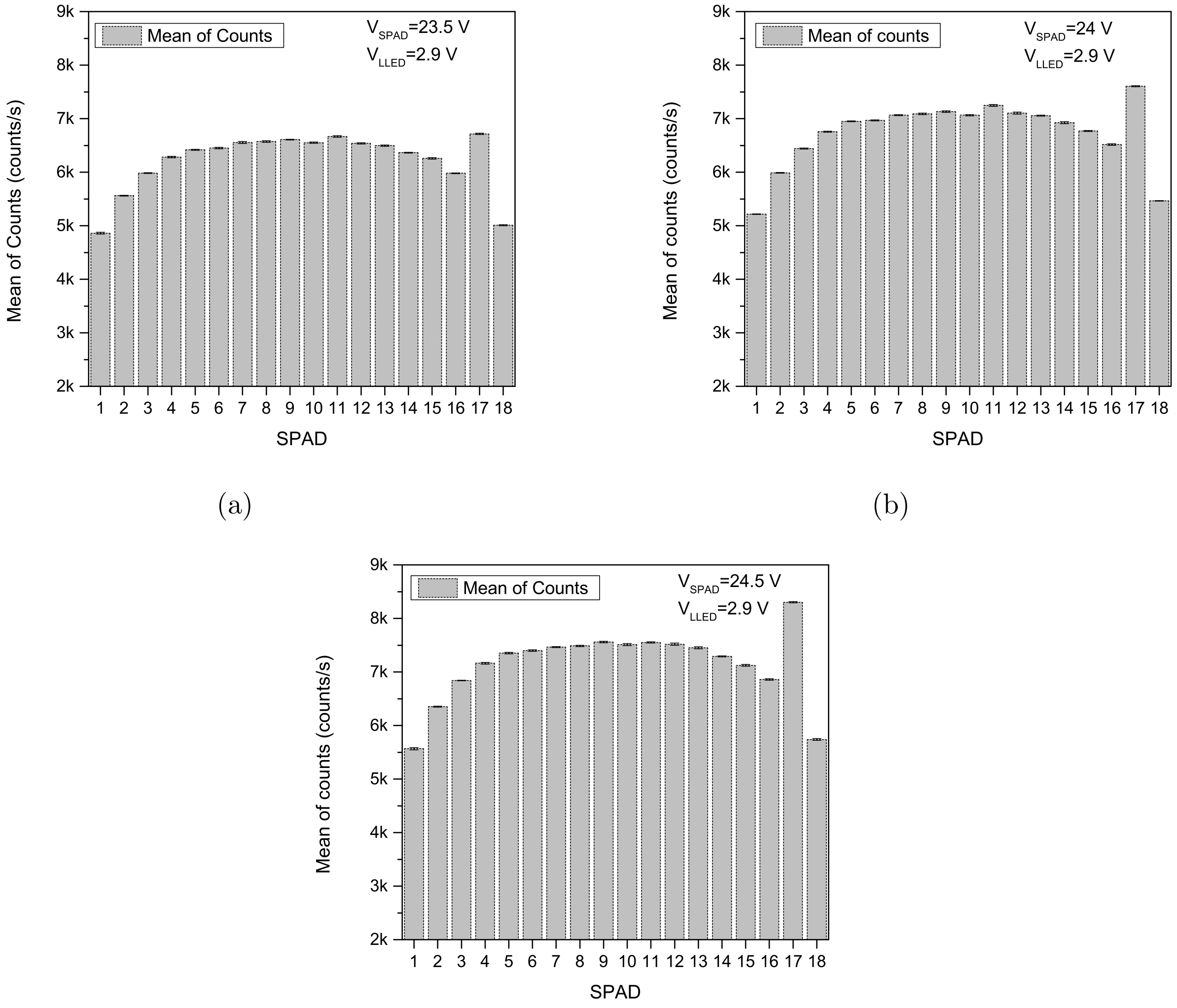
6.5.2 Generation of random numbers

Based on the analyses of the DCR of the array of the detectors, we chose SPAD8 to begin the measurements with since it has the lowest mean of DCR ∼97 cps (at V_{SPAD}=24.5 V) with standard error of 2.9 cps (see Fig.6.7 (c)). Using SPAD8 and TDC1, we tried to fix some parameters. As a single TDC, due to some fabrication restriction, we can choose only TDC1. If we want to use TDC2, it has to work together with TDC1. It is the same for TDC3 and TDC4, as well, that cannot work individually and they have to work with TDC1 and 2 and TDC1, 2 and 3, respectively.

Using the Labview program for the random mode and fixing the applied voltage to the LED to 2.9 V, we acquired sequences of TDC codes. Since we saw some missing TDC codes and some periodic oscillations in the TDC codes distribution due to LSB codes, we just considered the MSB codes of the TDC. At this stage we did not fix the average counting rate in order to satisfy the oversampling regime condition and hence having a nearly uniform distribution. Therefore, the MSB codes follow an exponential distribution with smaller codes having higher probability and bigger codes lower probability. Fixing V_{reg} to 2.75 V, V_c to 1.2 V, V_{LLED} to 2.9 V (with signal to noise ration (SNR) of ∼60–70) and T_w to 140 μs, we applied the bias voltage of 23 and 25 V to SPAD8 and studied the MSB codes probability distribution. We observe that:

- MSB codes of 0 and 1 as well as >241 have very high probabilities making some peaks at the beginning and the end of the distribution and hence are not shown in Fig. 6.10.

- At the beginning and the end of the MSB codes probability distribution when the V_{SPAD} is 23 V, there is a ringing (see Fig. 6.10 (a)). However, as it is seen in

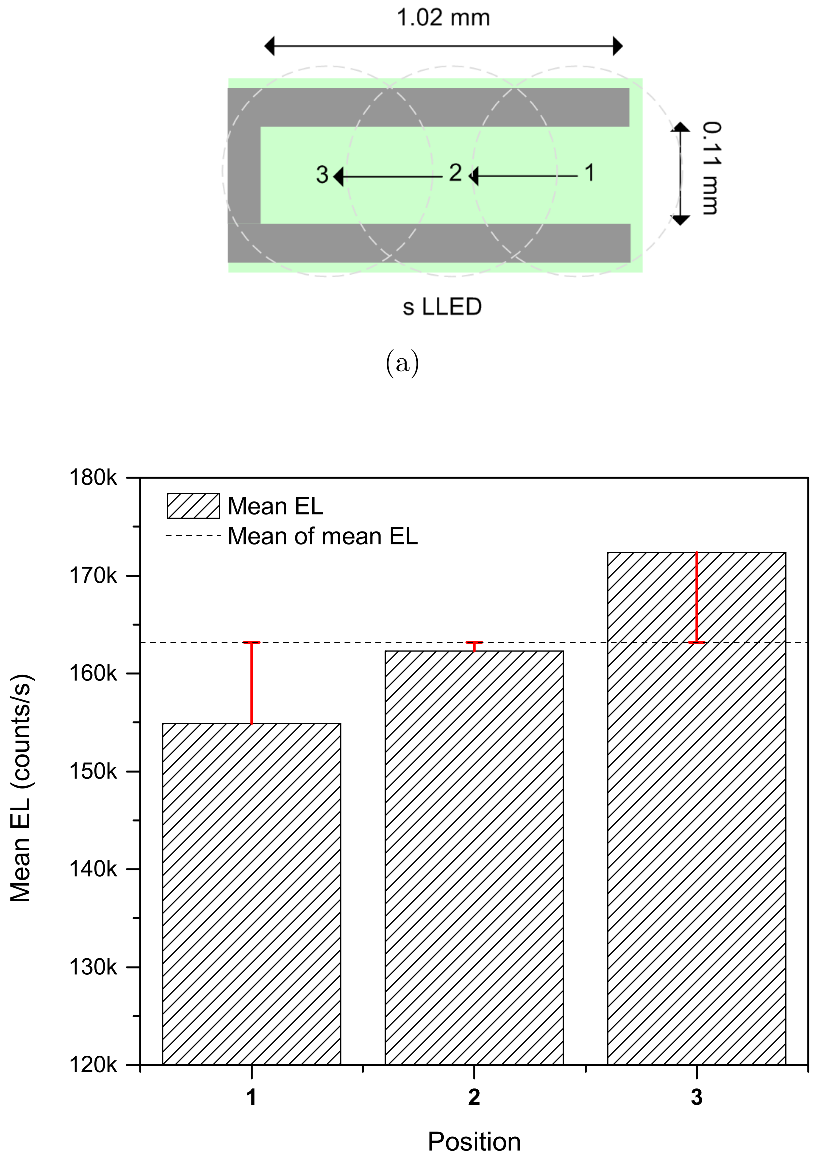


(a)

(b)

(c)

Figure 6.8: Counting rate of the array of 18 SPADs (including 2 dummy ones) at the applied voltage of 2.9 V to Si-NCs LLED and the applied bias voltage of (a) 23.5 V, (b) 24 V and (c) 24.5 V to the SPADs. The error bars are shown in the figures and are visible for some data points.



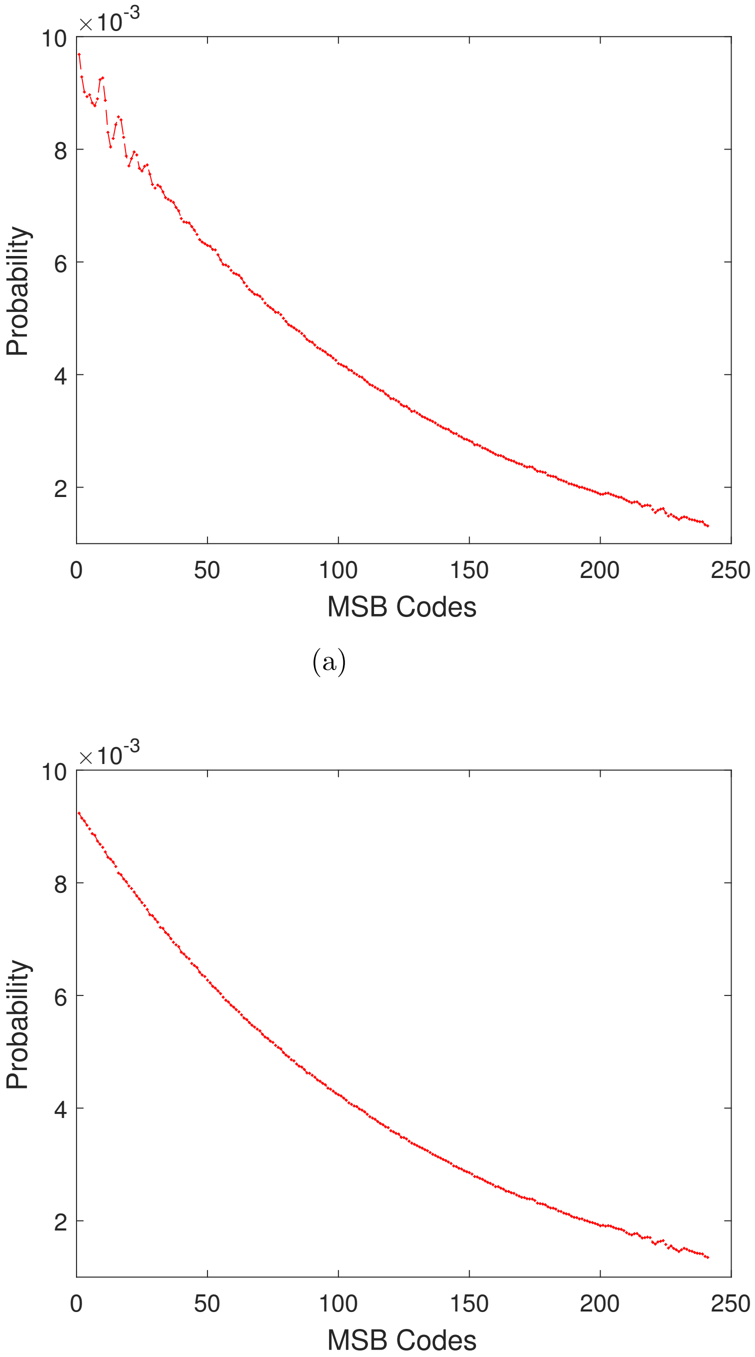
(a)

(b)

Figure 6.9: (a) The three positions on the s Si-NCs LLED to check the uniformity of EL. The spot diameter is 450 μm. (b) EL of the three positions on the s Si-NCs LLED. For the statistics analyses at each position, 20 EL values were obtained, each over 10 seconds of integration time. Then the mean value (the dashed bars) and the mean of this mean value (the dashed line) are computed. The red error bars indicate the deviation of each EL mean value from the mean of mean value.

Fig.6.10 (b) it fades away from the beginning of the distribution when we apply a bias voltage of 25 V to the SPAD. We could not remove the ringing at the end of the distribution by adjusting different parameters since it is most probably due to some parasitic capacitances and inductances caused by the electrical signal of TDC in the circuit (i.e. those that are not part of the design, but just by-products of the materials used to fabricate the circuit) to resonate at their characteristic frequency.

Therefore, we fixed the bias voltage of 25 V to SPAD8 and acquired long sequences of TDC codes considering the MSB codes for further analyses with $V_{\text{LLED}} = 2.9$ V and $T_w = 112.5$ ns (satisfying the oversampling regime of $\lambda T_w \sim 0.0008$). The 512x512 2-D visualization (see Section 2.3.2) of MSB codes (0-255) is illustrated in Fig. 6.11. No particular, periodic pattern is visible among the codes.
The MSB codes probability distribution for codes 2-241 can be seen in Fig. 6.12. The probabilities are distributed around the theoretical value of 1/240. The TVD is computed to be ~ 0.0037 (Eq.6.4).
Replacing each MSB code (2-241) with its corresponding 8-bit binary 0 and 1, we obtained 10^8 bits (data acquisition is very slow and since we consider only MSB codes 2-241, it takes a very long time to acquire sequences of 10^8 bits) and applied the frequency test in NIST tests suite (see Section 2.3.7.1.1). This test fails since although we discarded the MSB codes of 0, 1, and >241, the maximum bias is still high ~0.005.
After the application of the information-theoretically secure randomness extractor of Toeplitz function (see Section 3.3.4.2), all the statistical tests in NIST tests suite pass. The results of the NIST tests before and after the application of Toeplitz are presented in Table 6.3. With all SPADs and 4 TDCs working, however, we do not have the problem of missing codes. Codes from 0 to 3050 appear which make the efficiency of TDC code to $\sim$ 11.5 bits instead of 12 bits. This is another drawback due to the TDC structure. The probability distribution of TDC codes is presented in Fig. 6.13 (a). The ringing is visible at the end of the distribution. The probability of counts detected by each SPAD can be seen in Fig. 6.13 (b). The probabilities vary from one SPAD to the other and SPADs 1 and 16 are found to have the closest probabilities to the expected theoretical probability of 1/16 with deviations of ~0.00082% and ~0.0023%, respectively. The visual presentation of the TDC codes in Fig. 6.14 does not show any particular, periodic pattern. On the other hand, as can be seen in Fig. 6.15 (a), there is correlation among the TDC codes for the first 10 lags (autocorrelation coefficient is outside the 95% confidence interval).
The solution would be to separate the codes based on the address, i.e. the codes produced by each pixel are put together, and are then concatenated to produce a string of TDC codes. This will remove the correlation among the codes as observed in Fig. 6.15 (b). TVD is calculated by Eq. 6.4 to be ~0.087. Since the data acquisition of this type of configuration, with all SPADs and 4 TDCs working together, is very time consuming, we



(a)

(b)

Figure 6.10: MSB codes (100 M codes 2-241) probability distribution of TDC1 fixing V_{reg} to 2.75 V, V_c to 1.2 V, V_{LLED} to 2.9 V and T_w to 140 μs at V_{SPAD} (a) 23 V and (b) 25 V. The ringing at the beginning of the distribution in (a) is removed in (b) when we apply the bias volatge of 25 V to the SPAD.

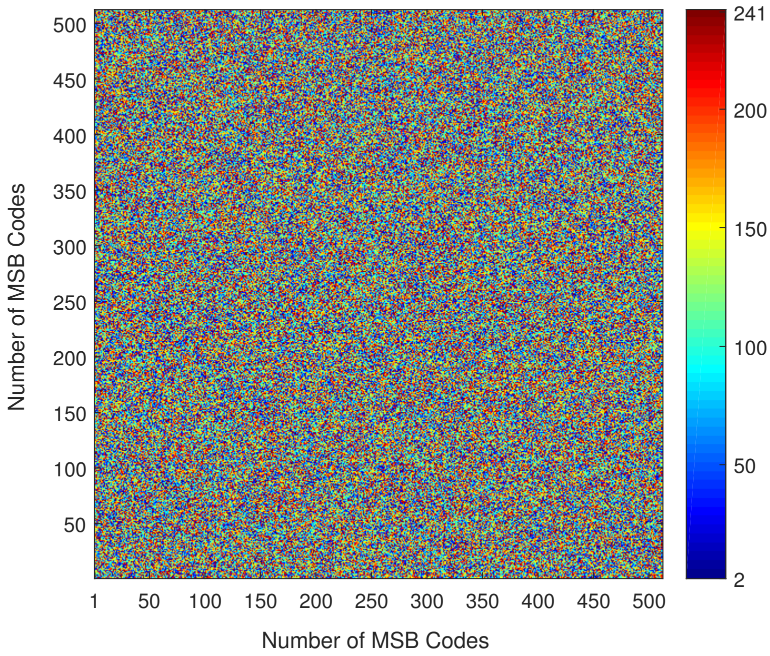


Figure 6.11: 512x512 2-D visualization of MSB codes. No particular pattern is observed among the codes.

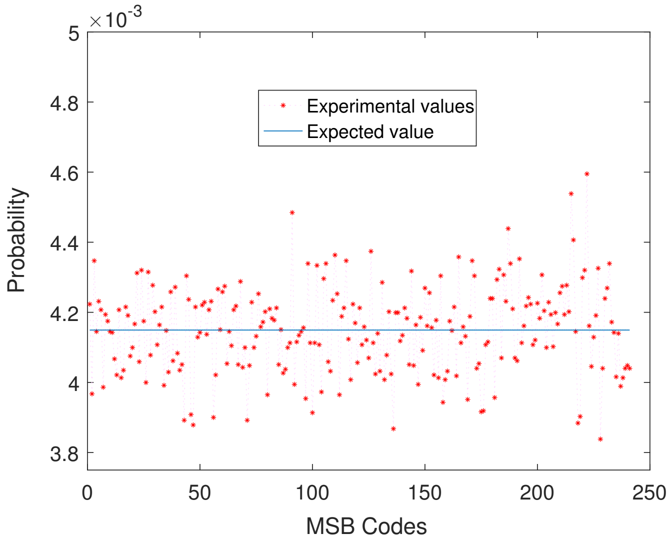


Figure 6.12: MSB codes (100 M codes 2-241) probability distribution of TDC1 with V_{SPAD}=25 V, V_{LLED}=2.9 V, and T_w=112.5 ns with the oversampling condition of ∼0.0008. The theoretical value of 1/240 is indicated by the blue line. The ringing at the end of the distribution is also visible here.

Table 6.3: Results of the statistical NIST tests for the datasets of 100 and 94 strings of 10^6 bits of data before (raw data) and after the application of Toeplitz randomness extractor, respectively. The *P-value$_T$* has to be larger than 0.0001. The minimum pass rates for statistical tests is 0.960. The worst case results are shown here.

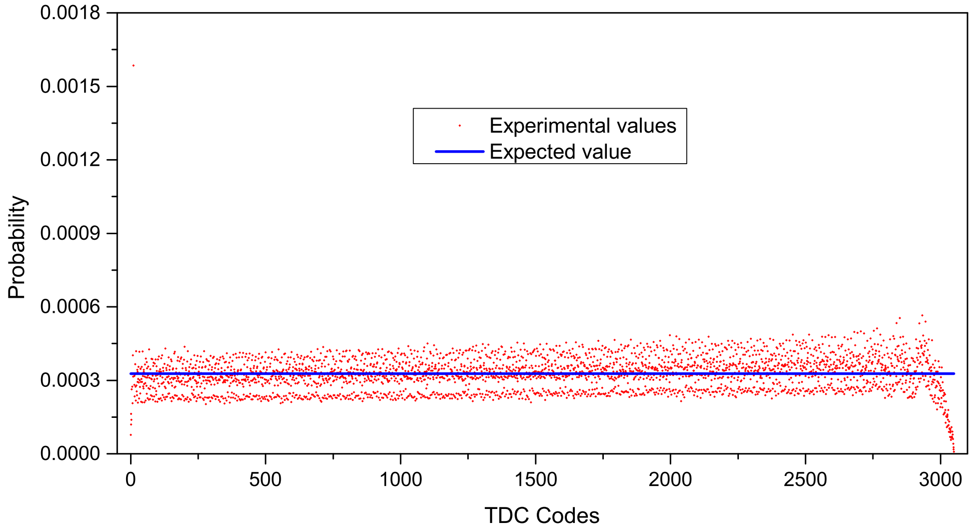
	Raw data			**Toeplitz-extracted data**		
Statistical test	P-value$_{\mathrm{T}}$	Proportion	Result	P-value$_{\mathrm{T}}$	Proportion	Result
Frequency	0.000000	0.000	Failed	0.051942	1.000	Passed
Block frequency	0.000000	1.000	Failed	0.401199	0.990	Passed
Cumulative Sums	0.000000	0.000	Failed	0.249284	1.000	Passed
Runs	0.000000	0.000	Failed	0.816537	0.990	Passed
Longest run	0.000000	0.000	Failed	0.015598	0.990	Passed
Rank	0.637119	1.000	Passed	0.037566	0.990	Passed
FFT	0.145326	0.990	Passed	0.455937	0.960	Passed
Non-overlapping template	0.000000	0.000	Failed	0.678686	1.000	Passed
Overlapping template	0.000000	0.000	Failed	0.334538	1.000	Passed
Universal	0.000000	0.000	Failed	0.964295	0.990	Passed
Approximate entropy	0.000000	0.000	Failed	0.040108	0.990	Passed
Random excursions	n/a	n/a		0.756476	1.000	Passed
Random excursions variant	n/a	n/a		0.619772	0.952	Passed
Serial	0.000000	0.150	Failed	0.010988	0.960	Passed
Linear complexity	0.494392	1.000	Passed	0.181557	0.980	Passed

could not acquire enough data for further statistical analyses. More studies are supposed to be done in the future as expressed in Section 6.6.

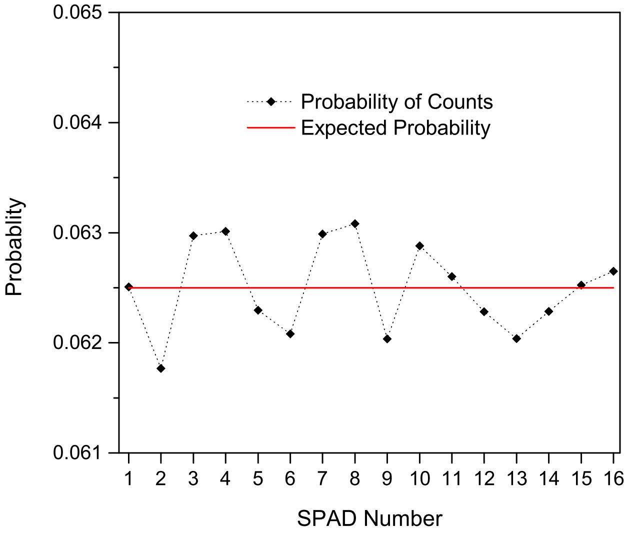
6.6 Conclusions and future activities

We realized a compact chip comprising a bonded Si-NCs LLED coupled with a cluster of 16 pixels (SPADs) connected to 4 TDCs to generate random numbers. Based on the characterization of the SPADs in dark and light conditions, we found out that SPAD8, with the lowest DCR, would be a good candidate to be used as the detector with TDC1 in order to run the random mode. Since we saw some missing TDC codes and some periodic oscillations in the TDC codes distribution due to LSB codes, we just considered the MSB codes of the TDC. We considered the MSB codes >1 and $<$ 242 in order to approach the nearly uniform distribution for the probability of MSB codes. Replacing each MSB code with its corresponding 8-bit binary 0 and 1, we acquired 10^8 bits and executed the statistical NIST tests. They failed due to a high bias of ∼0.005. However, after the application of the information-theoretically secure randomness extractor of Toeplitz function, they all passed.

Going from 1 SPAD/1 TDC to all SPADs/4 TDCs, the problem with missing codes



(a)



(b)

Figure 6.13: (a) TDC codes (100 M codes 0-3050) probability distribution of all pixels and 4 TDCs working together.The ringing is visible at the end of the distribution. (b) Probability of counts for each SPAD when all SPADs are working with 4 TDCs. V_{reg} is fixed to 2.75 V, V_c to 1.2 V, V_{LLED} to 2.9 V, T_w to 112.5 ns and V_{SPAD} to 25 V.

resolves. However, due to some drawback of the TDCs, the effiency of TDC code is not 12 bits per code but ∼11.5 bits per code. The visual representation of the TDC codes doen not show any particular, periodic patterns. The autocorrelation analysis reveals that for the first 10 lags the correlation coefficients stay out of the 95% confidence interval. The correlation is eliminated by ordering the TDC codes based on the pixel address and

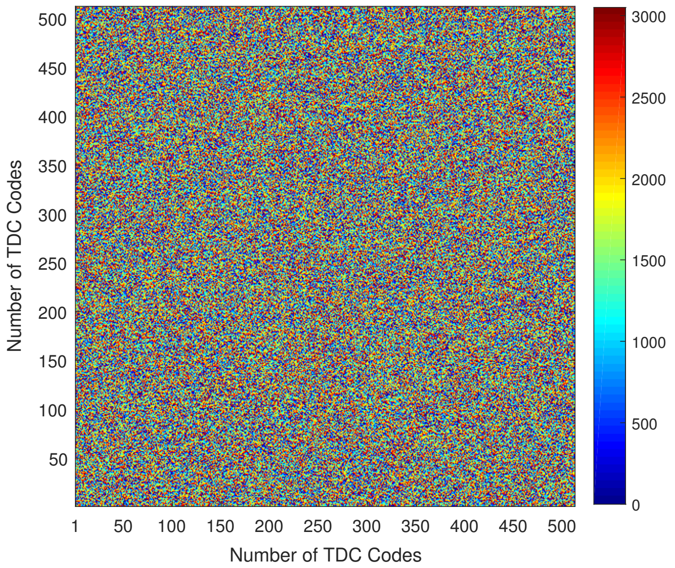
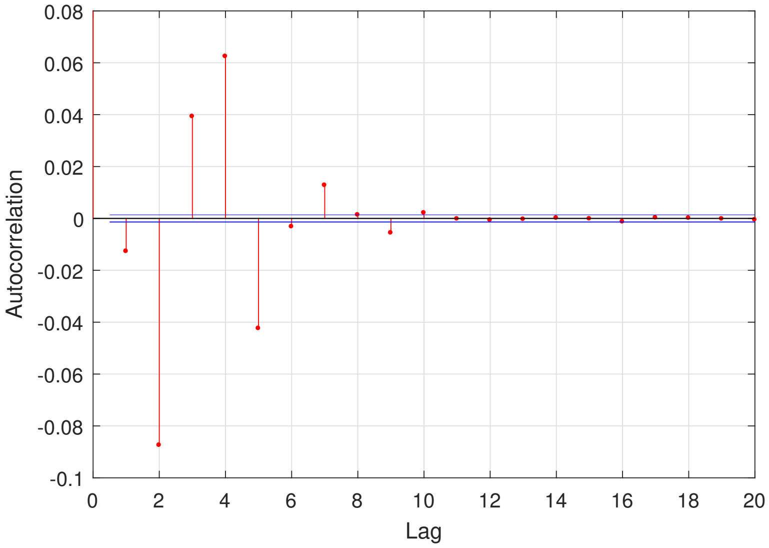


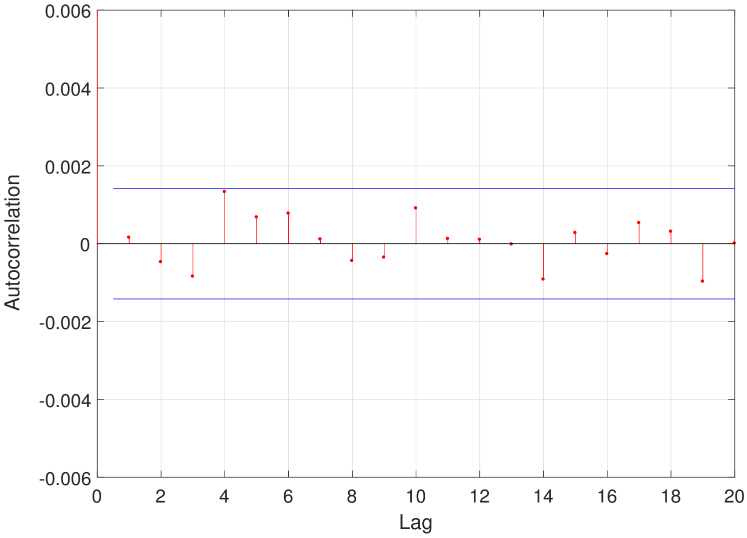
Figure 6.14: 512x512 2-D visualization of TDC codes. No particular, periodic pattern is observed among the codes.

concatenate them to produce a sequence of TDC codes.

The data acquisition is very slow and hence for more than 1 TDC configurations obtaining long sequences of codes is really time consuming. The future outlook is to optimize the data acquisition step in order to speed up the generation of random numbers and eventually the efficiency of our QRNG.



(a)



(b)

Figure 6.15: Autocorrelation function of (a) TDC codes and (b) TDC codes ordered based on the SPAD address. It can be seen that the correlation among the TDC codes at the beginning 10 lags is removed in (b) and the correlation coefficients in are placed inside the 95% confidence interval.

Chapter 7

A Robust QRNG Based on an Integrated Chip

``Small is beautiful.''

Leopold Kohr

Integration of the source of entropy and the detector on a single chip is an efficient way to produce a compact RNG. A small QRNG is an essential element to guarantee the security of our everyday life. It can be readily implemented into electronic devices for data encryption. The idea of "utmost security" would no longer be limited to particular organs owning sensitive information. It would be accessible to every one in everyday life.

7.1 State of the art

Examples of the integrated chips containing the source of entropy and the detection stage can be found in [65, 137]. In [137], a quantum entropy source for random number generation on an indium phosphide (InP) photonic integrated circuit is demonstrated. In the scheme designed by Abellan et al. [137], two distributed feedback lasers are combined on the same chip. One laser is operated in gain switching (GS) mode, while the second one is in CW mode. The interference of the chirped frequency (due to thermal effects) of the GS laser with the stable frequency of the CW laser is used to generate random numbers. When these two frequencies coincide, a nearly zero detuning (NZD) region is observed. Therefore, in the experiments, the NZD region is tuned at the end of the pulse to maximize the detuning frequency between the lasers so as to reduce any residual phase-locking effects.
In the approach in [137], several parameters need to be controlled such as the temperature which is kept at 25°C during the experiment and the relatively low modulation frequency of 100 MHz to capture properly the dynamics of the interference pattern within the GS

pulse. It is claimed that the modulation frequencies up to 2 GHz are reachable allowing for tens of Gbps raw generation rates using current analog-to-digital conversion technologies. In addition, the impossibility of creating a Si laser source does not permit taking advantage of Si photonics which is a promising candidate for building scalable optical applications due to its compatibility with the microelectronics industry.

Reverse-biased Si-LEDs emit light in the visible and near-infrared regions of the light spectrum. [138, 139] In [65], there is an integration of a ring-shaped Si SPAD around the Si-CMOS-LED on a single chip to generate random numbers. Random numbers are produced by considering pairs of non-overlapping random time intervals (t_i, t_{i+1}) of sequential photon arrivals and by defining the digital output as:

$$\begin{cases} 1 & \text{if } t_i > t_{i+1} \\ 0 & \text{if } t_i < t_{i+1} \qquad i = 1, 2, 3, \dots \\ \text{discarded} & \text{if } t_i = t_{i+1} \end{cases}$$

There is a large bias in the raw data due to the non-uniform distribution of the time intervals. The large bias is removed by post-processing in a special configuration of XOR gates to improve the randomness of the generated random bits. [65]

In this chapter, we demonstrate a QRNG based on an integrated chip composed of an emitter containing 16 pixels (Si SPADs) with p$^+$/n Si junction (see the structure in Fig. 7.3) (a) and a single Si SPAD (as the detector) with the same structure as each pixel of the emitter. The same approach as the robust methodology in Chapter 4 is used here as well to generate random numbers. The advantages of our RNG over [137] and [65] are:

- no post-processing operation is used to eliminate the bias and correlation of the raw data, and
- when we set the system to work, we do not need to control the parameters involved in the experiment from then on. The methodology we use to generate random numbers is robust against the variations of temperature, the non-idealities of the detector and some other parameters like the applied voltage bias to the source of entropy.

7.2 Photon emission from a reversed-bias Si LED

As mentioned in the previuos section, light emission from reverse-biased Si-LEDs has been reported in the visible and near-infrared regions of the light spectrum. [138, 139] Photon emission is attributed to some mechanisms (schematically shown in Fig. 7.1) such as direct and indirect interband transitions, intraband bremsstrahlung radiation from hot electrons scattered by charged coulombic centers and intraband transitions of hot holes

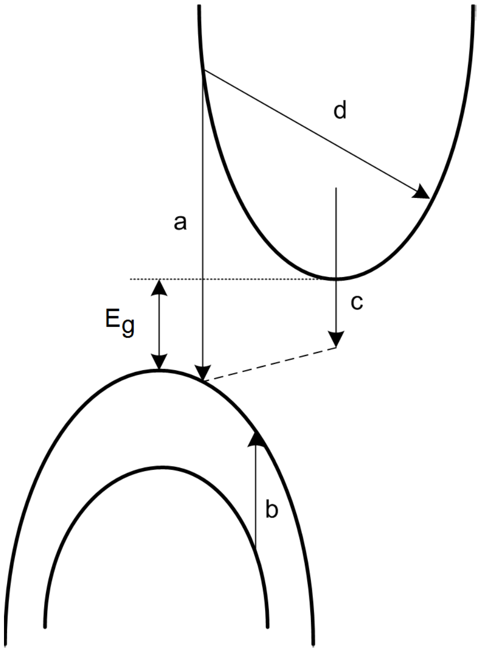


Figure 7.1: Possible mechanisms for the avalanche emission from reverse-biased Si p-n junction: (a) direct interband recombination, (b) intraband h transition, (c) indirect interband recombination and (d) intraband bremsstrahlung. Reproduced, with permission, from [141] © 2011, IEEE.

between the light and heavy-mass valence bands. [140]
The EL spectra of the emitter versus photon energy, at the applied bias ($\mathrm{V_{emitter}}$) of 54.8, 55.1 and 55.3 V to the emitter, are presented in Fig. 7.2.
Based on the spectra of the emitter (a reversed bias $\mathrm{p^+/n}$ Si junction) in Fig. 7.2, three different sections (with fit for indirect interband, intraband bremsstrahlung and direct interband mechanisms for the spectrum plot at $V_{\mathrm{emitter}} = 55.1$ V) can be observed. The fit for direct interband is based on the formula presented in [140]:

$$I(E_{\mathrm{ph}}) = A\sqrt{(E_{\mathrm{ph}} - E_g)}E_{\mathrm{ph}}\left[1 + B\left(\frac{E_{\mathrm{ph}}}{W}\right)\right]\exp\left[-\left(\frac{E_{\mathrm{ph}}}{W}\right)\right],$$

where I is the emission intensity, E_{ph} is the photon energy, E_g the band-gap energy, A is an empirically determined constant and B depends on the exponential integral of E_0/W with E_0 and W being the threshold for $e-h$ pair generation and a number depending on electric field and mean free path, respectively. [140]
The intraband bremsstrahlung region has been fitted by the following formula [140]:

$$I(E_{\mathrm{ph}}) = C\exp\left(\frac{E_{\mathrm{ph}}}{kT_{\mathrm{e}}}\right),$$

where C is an empirically determined constant, k_0 the Boltzmann constant and T_{e} is the electron temperature. The bremsstrahlung intraband response is monotonic and is,

therefore, unable to explain the observed peaks in the spectrum. [140]
The fit for indirect interband (occurring at low photon energies) is based on the formula presented by Gautam et al. [142]:

$$I(E_{\text{ph}}) = G(E_{\text{ph}})(E_{\text{ph}} + k\theta - E_g). \exp - \left(\frac{E_{\text{ph}} + k\theta - \Delta F}{k_0 T}\right). \exp\left(\frac{E_{\text{ph}} + k\theta}{k_0 T_h}\right)$$
$$. \exp\left[-\frac{a}{2}\left(E_{\text{ph}} + k\theta - E_g\right)\right].I_1\left[\frac{a}{2}(E_{\text{ph}} + k\theta - E_g)\right],$$

where G is the gamma function, T the silicon lattice temperature, T_h the hole temperature, k the component of wave vector in the field direction, $a = [1/k_0T_e + 1/k_0T_h]$ and I_1 is the modified Bessel function of order one. [142] The fitting parameters are fixed as $T_e = 2000°$, $T = 280°C$, $W = 0.3\ eV$, $B = -0.013$, $l_{\text{ioe}} = 68$ Åand $l_{\text{ioe}} = 68$ Å(the two latter are the ionization length of e and h, respectively.) [140]
For energies lower than ∼1.8 eV, photon emission is attributed to indirect interband transitions of high field carrier populations. The indirect interband processes depopulate the conduction band and reduce bremsstrahlung intensity at these energies. [140]
However, for energies ∼1.8-2.2 eV, the emission is predominantely through indirect intraband (bremsstrahlung) processes. For energies above ∼2.2 eV, the direct interband transitions dominate and generate the spectra.

7.3 Chip structure

As mentioned before in Section 7.2, at avalanche breakdown Si LED emits photons with visible or near-infrared wavelengths. This effect can be used advantageously to integrate Si SPADs as both the source of entropy and the detector on a single chip. Since the quantum efficiency of Si SPAD is very low, we can use an array of them (emitter) to generate more carriers in order to multiply the efficiency.
The structure of the integrated emitter and detector is shown in Fig. 7.3 (a). The emitter is an array of 16 pixels each with the same specifications of the cells in SiPM presented in Section 5.2. The detector, which is one single Si SPAD, is located on one side of the emitter (on its left in the figure). The emitter is connected to the detector as schematized in Fig. 7.3 (b). Both emitter and detector work in reversed bias region and the overall bias to the emitter is $\text{V}_{\text{emitter}} = \text{V}_{\text{SPAD}} - \text{V}_{\text{e}}$.

7.4 Experimental procedure

The experimental setup is presented in Fig. 7.4. The chip readout board is connected to a ±5 V. Bias voltages of V_{SPAD} ∼30-36 V is applied by Agilent E3631A DC Power Supply and the negative voltages for V_e are provided by an Agilent B1500A Semiconductor

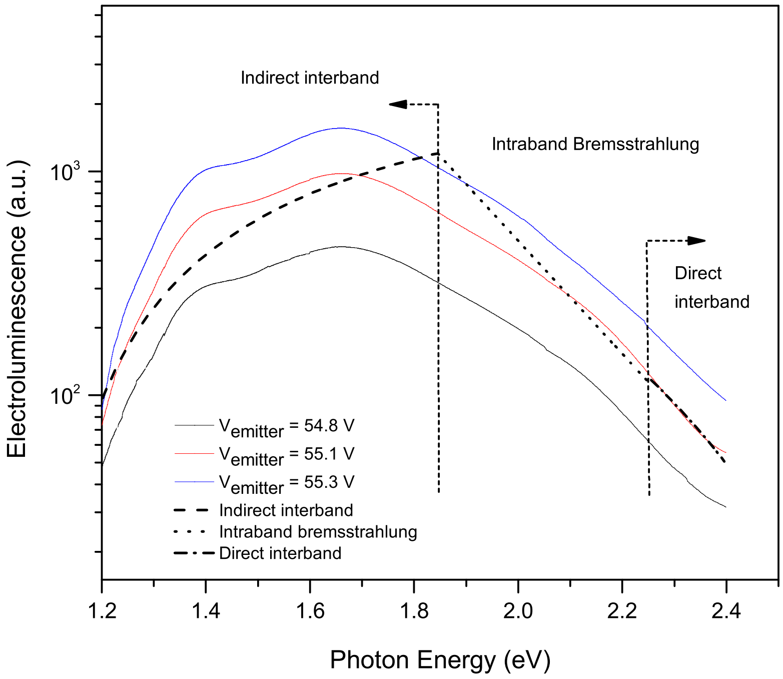
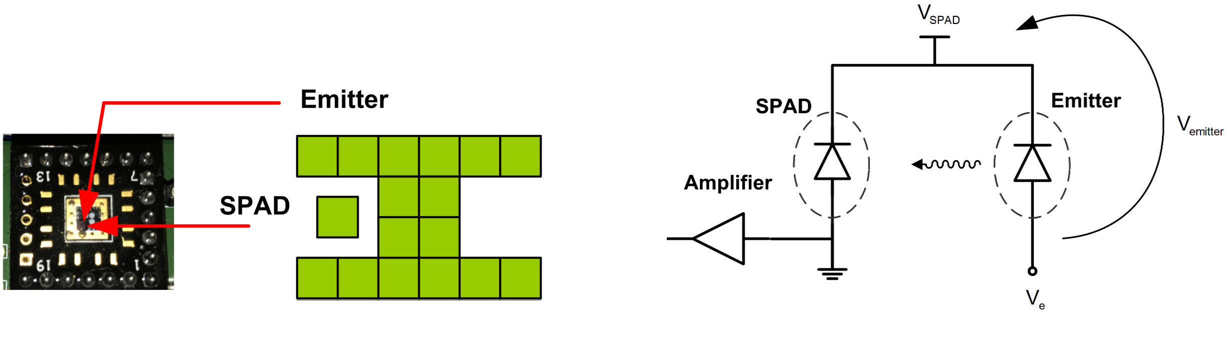


Figure 7.2: EL versus photon energy (eV) for the emitter in the integrated chip at the applied bias V_{emitter} of 54.8, 55.1 and 55.3 V (corresponding to the $V_{\text{SPAD}} = 36.8$ V and V_e of -18, -18.3 and -18.5 V, respectively. The fitted curves for three mechanisms of indirect interband, intraband bremsstrahlung and direct interband can be seen for the red curve with $V_{\text{emitter}} = 55.1\ V$.



(a)

(b)

Figure 7.3: (a) Integrated emitter and SPAD (detector) structure. The structure of the emitter (an array of 16 Si SPADs) and the detector (one single Si SPAD on the left of the emitter) can be seen on the right in a large scale. (b) Schematic of the circuit of the emitter and SPAD connected to each other. The overall bias to the emitter is $V_{\text{emitter}} = V_{\text{SPAD}} - V_{\text{e}}$ and the output signal is amplified before reaching the FPGA for random numbers generation.

Device Parameter Analyzer. The count rate of the SPAD, which is detecting the EL of the emitter, is monitored by a Hewlett Packard 53131A Universal Counter 225 MHz by selecting the Totalize mode, so that true pulse counting is performed. The TTL output of the front-end circuit is directly connected to the high speed digital input of the FPGA. The EL spectra were obtained by a Spectra-Pro 2300i monochromator coupled with a nitrogen-cooled CCD camera. The measurements were performed at room temperature in a dark chamber.

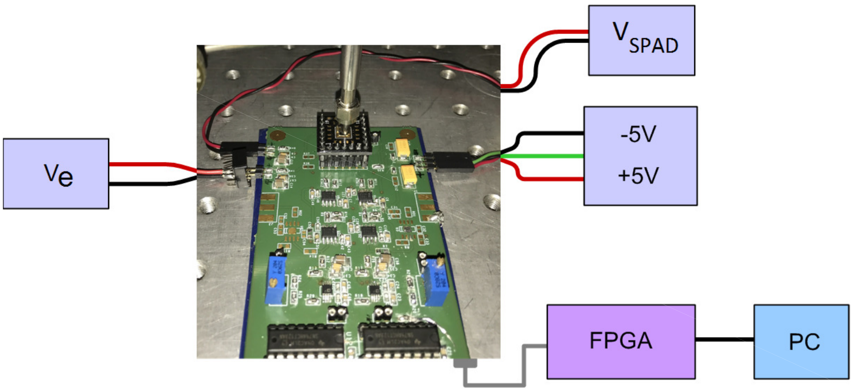


Figure 7.4: Schematic of the experimental setup for random numbers generation. The chip board is connected to ±5 V and a bias voltage (V_{SPAD}) is applied to the SPAD. The bias voltage of V_e is provided by an Agilent B1500A Semiconductor Device Parameter Analyzer. The output signal of the SPAD is transmitted to FPGA connected to PC for the generation of random numbers.

Autocorrelation, $\mathrm{g}^2(\tau)$, measurement of the Si SPAD signal was performed via a multitau digital correlator with 4 ns resolution [124] at the applied voltage $V_{\mathrm{emitter}} = 53$ V (corresponding to $V_{\mathrm{SPAD}} = 36$ V and $V_e = -17$ V). The afterpulsing distribution exhibits a main peak within ∼224 ns from the main autocorrelation peak at τ=0 (Fig. 7.5). The plateau in $\mathrm{g}^2(\tau)$ approaches the normalization value of 1 at about 4.5 μs.
The pulses from the Si SPAD are recorded via a multichannel scaler Ortec Easy-MCS for the analysis of Poisson distribution fully described in Section 3.3.1.1. MCS has a minimum channel (bin) width of 100 ns and has no dead time between the channels. The scan length is variable from 4 to 65,536 channels. The results are presented in the next section.

7.5 Results and discussions

The SPAD signal was sent to the MCS (with bins set to have different mean of photons) to study Poisson distribution. At different mean values from 0.2 to 2, we observed that p-value was always lower than α. Based on the χ^2 statistic, since p-value$< \alpha$ we

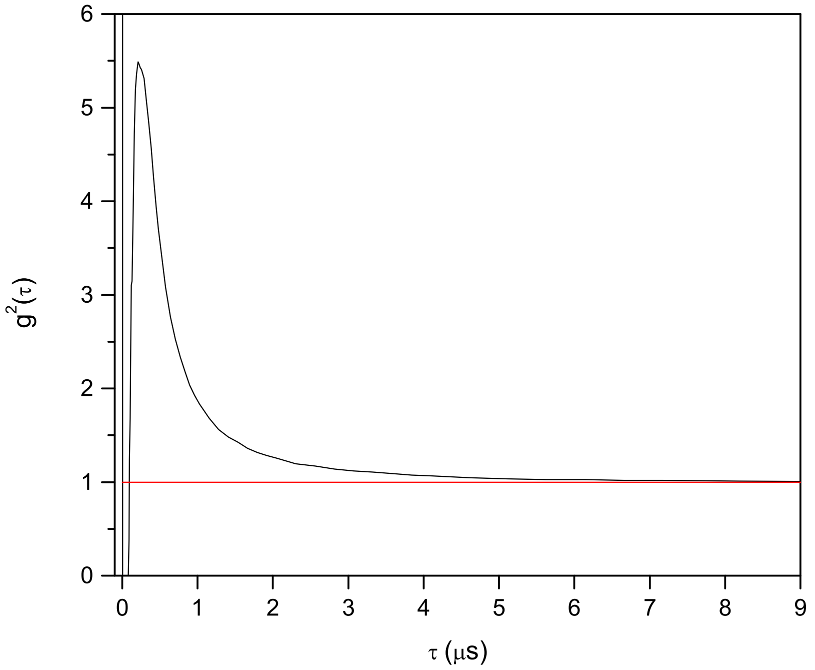


Figure 7.5: Autocorrelation function ($\mathrm{g}^2(\tau)$) of the Si SPAD signal (peak at zero is out of scale) at the applied voltage of $V_{\mathrm{emitter}} = 53$ V (corresponding to $V_{\mathrm{SPAD}} = 36$ V and $V_e = -17$ V). Dead time ($\sim$70 ns) and afterpulsing distribution of the detector can be seen here.

conclude that the Poisson distribution does not match well with the obtained data (see Section 3.3.1.1). And since the variance of data is larger than the mean, we conclude that the distribution could be super-Poissonian. The result of a measurement for mean value of ∼0.96 is illustrated in Fig. 7.6 (b) with the variance of ∼1.13.

In order to understand if this occurrence is attributed to the SPAD or the emitter, we conducted measurements with SPAD dark counts in addition to the SPAD signal when illuminated by Si-NCs LLED and commercial red LED. We observed that, for different mean values, the Poisson fit would be appropriate for distribution of the acquired data (see Fig. 7.6 (a) for SPAD dark counts).

We can conclude that the Poisson distribution does not seem to be a suitable fit for the EL of the emitter due to the bunching of the emitted photons. Furthermore, we conducted the cross-correlation measurement described in Sections 3.3.1.1 and 3.3.3 using the same setup in Fig. 3.8 with Si-NCs LED replaced with the emitter (at the applied bias of $V_{\text{emitter}} = 52\text{V}$). The cross-correlogram is presented in Fig. 7.7. A peak (due to photon bunching) is visible in the plot indicating a super-Poissonian distribution of the detected photons with the mean greater than the variance and photon number fluctuations larger than a coherent light (see Section 3.3.1.1).

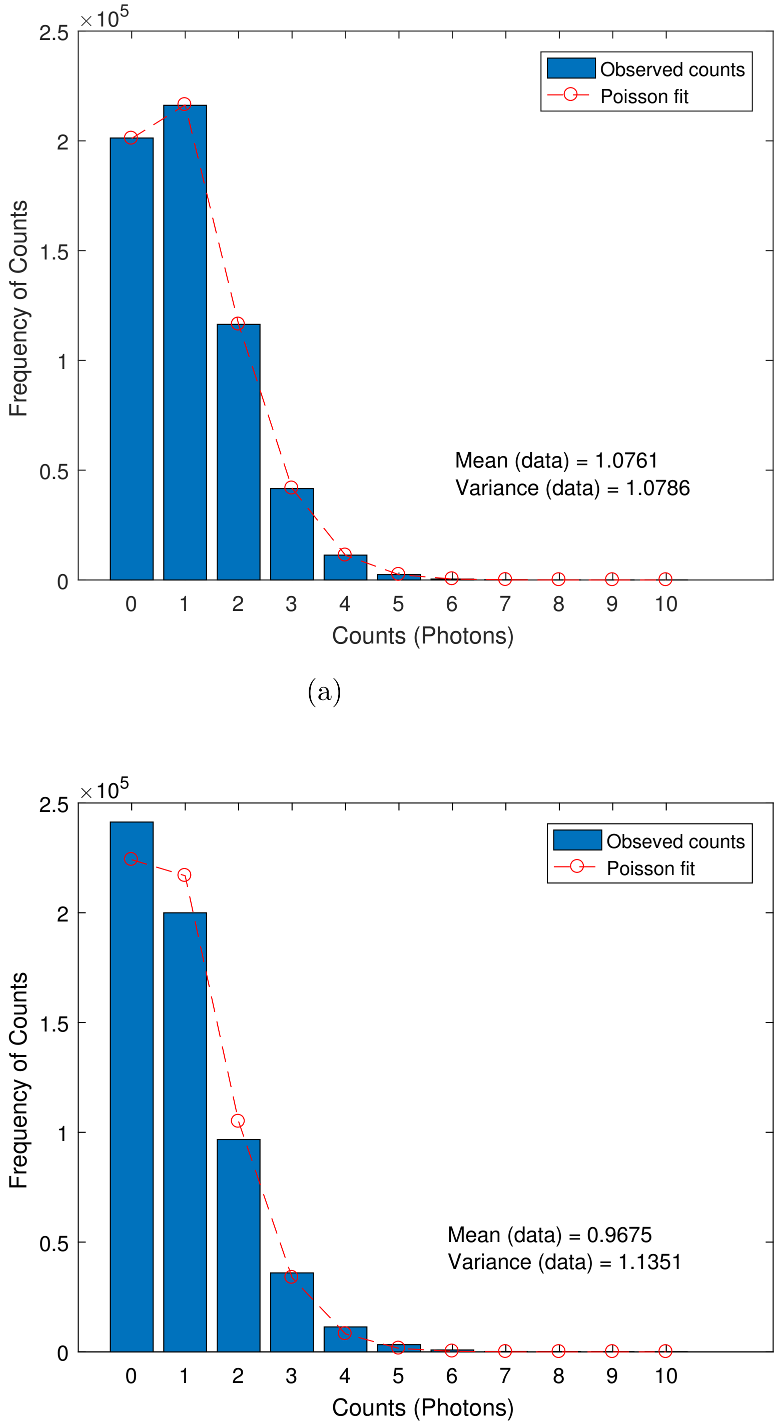
Taking into account the afterpulsing distribution of the SPAD in Fig. 7.5, we set the length of the "double length" interval (see Section 4.1.2) to 8960 ns with subintervals of 140 ns. Having fixed $V_{\text{emitter}} = 54$ V (corresponding to $V_{\text{SPAD}} = 37$ V with DCR ~ 80 cps and $V_e = -17$ V), we acquired sequences of symbols for further analysis. The visual representation can be seen in Fig. 7.8 for a matrix of 512x512 symbols. As it is observed in the figure, no particular, periodic pattern exists among the symbols.

The probability distribution of the symbols is nearly uniform (Fig. 7.9). The analysis of the JPMF (see Section 4.4.1) shows a departure of $\sim 10^{-6}$ from the theoretical value of $1/16 \times 1/16$ (Fig. 7.10).MI of the generated random symbols (section 4.4.1) is calculated to be $\sim 10^{-7}$ bits considering 1G random symbols. The maximum bias is in the order of $\sim 10^{-5}$ and the min-entropy (see Section 2.3.5) is ~ 3.999 bits per 4-bits. The highest bit rate is calculated to be ~ 100 kbps (see Fig. 7.11).

Replacing each hexadecimal symbol with its corresponding 0 and 1 binary, we acquire long sequences of data. The results in Table 7.1 show that all the statistical tests in NIST tests suite pass without the application of a postprocessing algorithm.

7.6 Conclusions

We realized a QRNG based on an integrated chip with a Si SPAD and an emitter which is an array of reversed bias Si p^+/n junctions. Studying the EL spectra, we learn about the different mechanisms through which emission from Si reversed bias avalanche junction



(a)

(b)

Figure 7.6: Poisson fit for (a) the dark counts of the SPAD with both the variance and mean value of ∼1.07 and (b) the output signal of the SPAD illuminated with emitter with mean and variance of ∼0.96 and 1.13, respectively. Based on the chi-squared (χ^2) statistic, Poisson fit is a good match for SPAD dark counts. However, it does not seem to be a suitable fit for the obtained data from SPAD and emitter most probably due to the bunching of the emitted photons from the emitter.

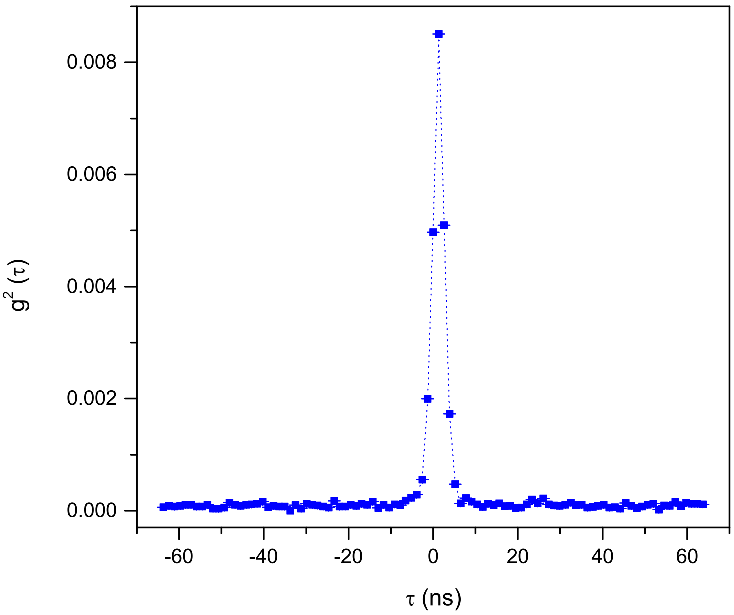


Figure 7.7: The cross-correlogram for the emitter at the applied bias $V_{\text{emitter}} = 52\text{V}$. A peak is visible in the plot indicating photon bunching and a super-Poissonian distribution of the detected photons. The plot is not normalized and the error bars for some data points are visible.

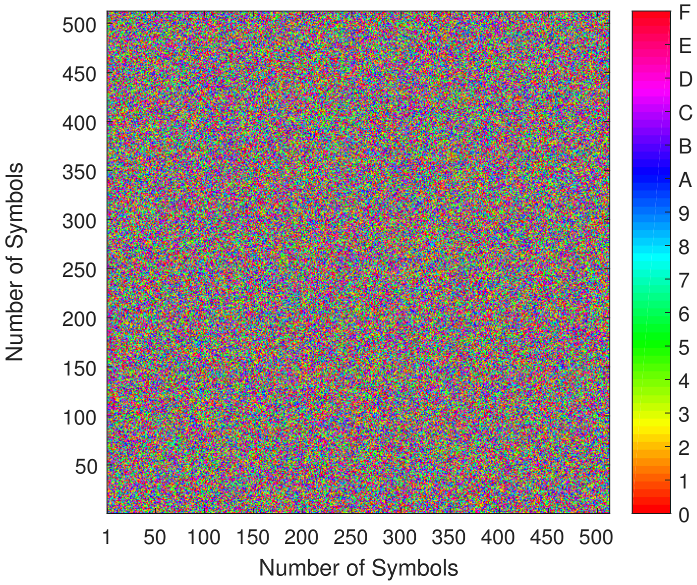


Figure 7.8: A 2-D 512 × 512 map of the hexadecimal symbols. No particular, periodic pattern is observable among the symbols.

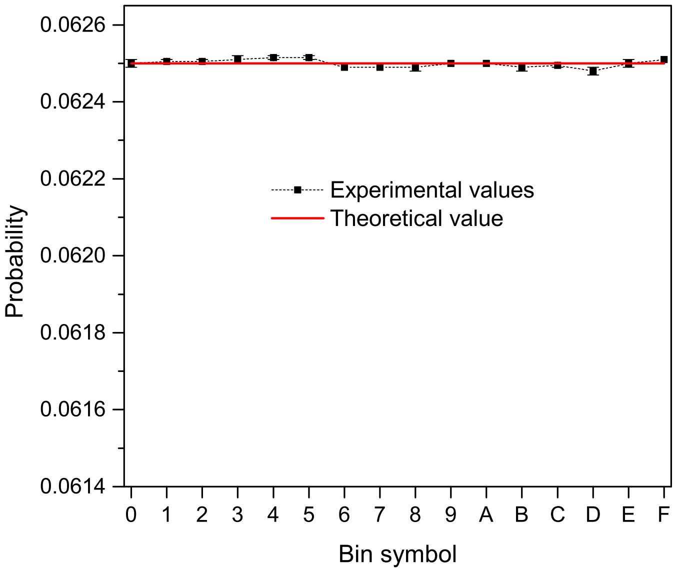


Figure 7.9: The probability distribution of the 16 symbols for 1 G symbols. It follows a nearly uniform distribution. The theoretical value of 1/16 is indicated by a red line in the figure.

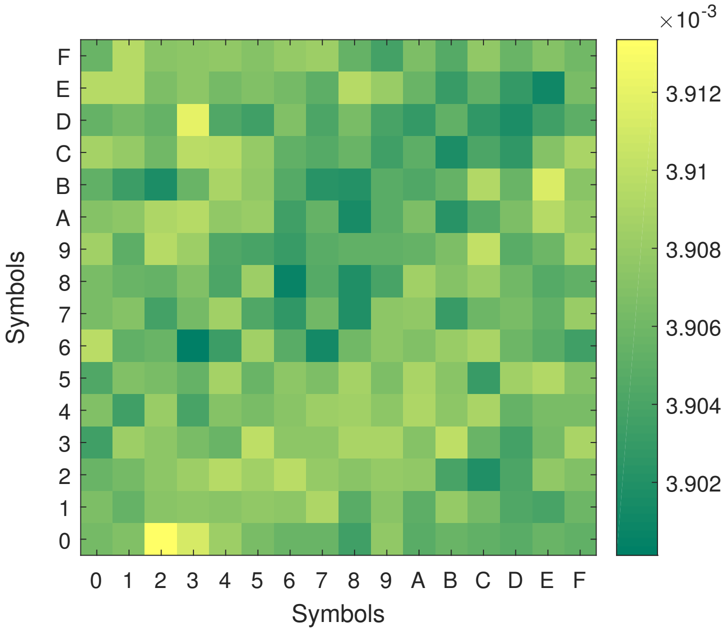


Figure 7.10: JPMF for 1 G generated symbols showing the probability of having each symbol followed by the other one. There is a very low deviation in the order of $\sim 10^{-6}$ from the theoretical value of $(1/16) \times (1/16) = 0.00390625$.

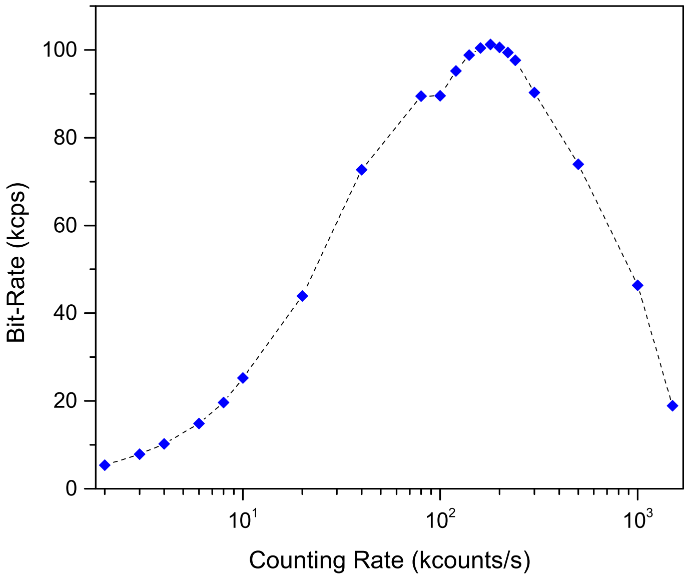


Figure 7.11: The bit rate of the QRNG based on an integrated chip composed of a SPAD and an emitter. It increases as the counting rate (detected photons by the SPAD) increases, reaches a maximum of ∼100 kbps and decreases afterwards due to the growing number of discards of more than one photon arrivals in "super interval".

Table 7.1: NIST tests results for 1 G random bits. The significance level is α=0.01. In order to pass, the p-value$_\mathrm{T}$ should be larger than 0.0001 and the proportion should be more than 0.980.

Statistical test	**P-value$_\mathrm{T}$**	**Proportion**	**Result**
Frequency	0.424453	0.9880	Passed
Block frequency	0.336111	0.9930	Passed
Cumulative sum	0.516113	0.9920	Passed
Runs	0.933472	0.9930	Passed
Longest run	0.686955	0.9910	Passed
Rank	0.075719	0.9940	Passed
FFT	0.715679	0.9880	Passed
Non overlapping template	0.363593	0.9920	Passed
Overlapping template	0.009071	0.9890	Passed
Universal	0.522100	0.9870	Passed
Approximate entropy	0.965083	0.9920	Passed
Random excursions	0.083143	0.9853	Passed
Random excursions variant	0.152493	0.9918	Passed
Serial	0.164425	0.9950	Passed
Linear complexity	0.610070	0.9920	Passed

occurs. For energies lower that ∼1.7 eV, photon emission is attributed to indirect interband transitions of high field carrier populations. However, for energies ∼1.7-2.2 eV, the emission is predominantely through indirect intraband (bremsstrahlung) processes. The direct interband transitions dominate for energies above ∼2.2 eV.

There are several improvements in the QRNG designed and studied in this chapter compared with the QRNG presented in Chapter 5 such as the microscopic size of the integrated chip (with both the source of entropy and the detector on a single chip), the much higher signal to noise ratio of ∼1250 (vs. ∼6.88 of Chapter 5) at the maximum bit rate and the much less dark count contribution of ∼0.0007 (vs. ∼0.5 of Chapter 5) in random number generation.

Even though the detected photons from the emitter do not follow a Poisson distribution, we see that using our robust methodology introduced in Chapter 4, high quality random numbers are generated:

- The 2-D visual representation of the generated hexadecimal symbols does not show any particular, periodic patterns.
- The probability distribution of the symbols is nearly uniform.
- The analysis of the JPMF shows a departure of $\sim 10^{-6}$ from the theoretical value of $1/16 \times 1/16$.
- MI of the generated random symbols is calculated to be $\sim 10^{-7}$ bits considering 1G random symbols.
- The maximum bias is in the order of $\sim 10^{-5}$ and the min-entropy is computed $\sim$ 3.999 bits per 4-bits.
- All the statistical tests in NIST tests suite pass for the generated raw data.

The highest bit rate is calculated to be $\sim$ 100 kbps.

Conclusions

In this thesis, various systems based on Si that differ in the degree of integration have been studied. In the first systems, composed of discrete components, two methodologies based on photon counting and photon arrival time measurements are considered to obtain a compact and reliable QRNG. We realized a physical QRNG based on Si nanocrystals (Si-NCs) LED as the source of randomness. Very negligible bias and simple setup are the chief strengths of our QRNG. With forced dead time of 1 μs and 500 ns, 100 Mbits long sequences pass the statistical tests of the NIST suite. The highest bit rate achieved is 0.6 Mbps.

Despite the low bit rate, this first simple approach benefits from several advantages: it uses light to stimulate events in the SPAD and avoids a deterministic post-processing of the raw data for small datasets. This fact is extremely remarkable in producing high quality random numbers and compensates for the low bit rate. Furthermore, the approach proposed here uses simple silicon-based LEDs as the light source and its overall bit rate can be easily increased by adopting a parallel architecture and exploiting the CMOS compatibility of all the components.

However, 1 G bits long datasets fail the main statistical tests in the NIST tests suite. This failure is attributed to a per mille drift in the electroluminescence (EL) of the Si-NCs LED that violates the equal probability of ones and zeros. By using post-processing, the randomness is recovered by the application of the Von Neumann and information-theoretically secure Toeplitz extractors. The bias and correlation among bits are removed and all the statistical tests in the NIST tests suite are consequently passed. A number of parameter control solutions such as stabilizing the temperature, resetting the applied current to the Si-NCs LED (or equivalently resetting the bin width in the MCS) and considering a feedback for the system can also be taken into account to overcome the problem of bias and to generate long, high quality random bit streams.

By the analysis of physical reasons of the failures in the previous approach, we developed a robust methodology based on photon arrival time measurements to generate quantum random numbers. The source of entropy is again a Si-NCs LED coupled with a Si SPAD connected to a field-programmable gate array (FPGA) to extract random numbers. So far in the literature, timing information of the photon arrivals has been utilized to generate random bits through different approaches. However, the lack of a robust methodology with a complete study of the detector imperfections and a simple setup to generate random numbers has been evident. The methodology developed, tested and presented here masks all the defects of afterpulsing, dead time and jitter of the Si SPAD and is effectively insensitive to aging of the LED and its emission drifts related to temperature variations. A simple, integrable setup is used to produce sequences of random numbers.

Analyses of the joint probability mass function (JPMF), mutual information (MI) and min-entropy show the high quality of generated random numbers and the high efficiency of the methodology. Despite the variations of the LED emission intensity, the system is efficient in producing long bit sequences maintaining the high quality of random numbers. The raw data pass all the statistical tests in NIST tests suite and TestU01 Alphabit battery without a post processing algorithm. The maximum demonstrated bit rate is 1.68 Mbps with the efficiency of 4-bits per detected photon.

The second QRNG system we studied is based on in-house fabricated LED and detectors. Specifically, it comprises a Si-NCs LLED and a Si photomultiplier (SiPM) and the random bits are generated by using the same methodology based on photon arrival times. The QRNG is found to be robust in this configuration as well. We set the double length interval to 1920 ns to mask the afterpulsing and crosstalk distribution of SiPM. The analyses on JPMF, MI and min-entropy show high quality of the generated random numbers. All the statistical test in the NIST tests suite pass for the raw data. Making up this configuration, we approached quite well the objective of the Work Package 4 of project SiQuro.

In the third analyzed system, the degree of integraion was increased. We realized a compact chip comprising a bonded Si-NCs LLED coupled with a cluster of 16 pixels (SPADs) connected to 4 TDCs to generate random numbers. Based on the characterization of the SPADs in dark and light conditions, we found out that SPAD8, with the lowest DCR, would be a good candidate to be used as the detector with TDC1 in order to run the random mode. Since we saw some missing TDC codes and some periodic oscillations in the TDC codes distribution due to LSB codes, we just considered the MSB codes of the TDC. We considered the MSB codes >1 and < 242 in order to approach the nearly uniform distribution for the probability of MSB codes. Replacing each MSB code with its corresponding 8-bit binary 0 and 1, we acquired 10^8 bits and executed the statistical NIST tests. They failed due to a high bias of $\sim$0.005. However, after the application of the information-theoretically secure randomness extractor of Toeplitz function, they all passed.

Going from 1 SPAD/1 TDC to all SPADs/4 TDCs, the problem with missing codes resolves. However, due to some drawback of the TDCs, the effiency of TDC code is not 12 bits per code but $\sim$11.5 bits per code. The visual representation of the TDC codes does not show any particular, periodic patterns. The autocorrelation analysis reveals that for the first 10 lags the correlation coefficients stay out of the 95% confidence interval. The correlation is eliminated by ordering the TDC codes based on the pixel address and concatenate them to produce a sequence of TDC codes.

The data acquisition was very slow and hence for more than 1 TDC configurations obtaining long sequences of codes is really time consuming. The future outlook is to optimize the data acquisition step in order to speed up the generation of random numbers and eventually the efficiency of our QRNG.

The final system, which also concludes my thesis work, is based on an integrated chip with a Si SPAD and an emitter which is an array of reverse-biased Si p^{+}/n junctions. Fitting the EL intensity versus photon energy curve, photon emission from Si reverse-biased junction (the emitter in the integrated chip) is attributed to three mechanisms. For energies lower that ∼1.8 eV, photon emission is due to indirect interband transitions of high field carrier populations. However, for energies ∼1.8-2.2 eV, the emission is predominantely through indirect intraband (bremsstrahlung) processes. The direct interband transitions dominate for energies above ∼2.2 eV.

Even though the detected photons from the emitter do not follow a Poisson distribution, we see that using our robust methodology introduced in Chapter 4 high quality random numbers are generated. The highest bit rate is calculated to be ∼ 100 kbps. Despite the low bit rate, the QRNG designed and studied in this chapter has several advantages over the QRNG presented in Chapter 5 such as the microscopic size of the integrated chip (with both the source of entropy and the detector on a single chip), the much higher signal to noise ratio of ∼1250 (vs. ∼6.88 of Chapter 5) at the maximum bit rate and the much less dark count contribution of ∼0.0007 (vs. ∼0.5 of Chapter 5) in random number generation.

Throughout this thesis, several configurations have been designed and tested for the generation of high quality random numbers. Moving from a macroscopic structure with separated components, coupled and connected to one another, to a microscopic structure, with both the source of entropy and the detector integrated on a single chip, we approached well the final goal of Work Package 4 of the project SiQuro which is the mass production of a small, cheap, high quality, robust and all-Si-based QRNG.

In conclusion, in this work I have reported a step by step development of a Si-based, fully integrated QRNG. I have described the limitations and the advantages of the proposed systems, specifically for what concerns the quality of the produced random numbers. The reduced degree of integration to having a single chip, containing both the source of entropy and the detector, together with robustness and high quality random numbers are prominent achievements of this thesis work.

The outlook is to parallelize the integrated chip in order to produce and commercialize the SiQuro USB QRNG able to generate high quality random numbers at reasonably high bit rate for the applications in cryptography and secure communications. Since the availability of cheap means and tools to provide widespread security for all in everyday communications is absolutely demanding, this QRNG can be also implemented in small electronic devices to guarantee the "utmost security" for *everyone* not only for particular military and political applications which own sensitive, confidential information. The utmost security has to belong to all at a very low expense affordable by every member of the societies.

List of publications

1. “Robust Quantum Random Number Generation with Silicon Nanocrystals Light Source”, Z. Bisadi, G. Fontana, E. Moser, G. Pucker, L. Pavesi, *Journal of Lightwave Technology*, vol. 35, no. 9, pp. 1588–1594, 1 May 2017. doi: 10.1109/JLT.2017.2656866.

2. “A Robust Approach to the Generation of High Quality Random Numbers”, Z. Bisadi, G.Fontana, E. Moser, G. Pucker and L. Pavesi *Proc. of SPIE*, vol. 9996, 2016, doi: 10.1117/12.2242000.

3. “Generation of high quality random numbers via an all–silicon–based approach”, Z. Bisadi, A. Meneghetti, A. Tomasi, A. Tengattini, G. Fontana, G. Pucker, P. Bettotti, M. Sala, and L. Pavesi, *physica status solidi (a)*, vol. 213, no. 12, pp. 3186–3193, 2016. doi: 10.1002/pssa.201600298.

4. “(Invited) Silicon Nanostructures: A Versatile Material for Photonics”, Z. Bisadi, C. Piotto, G. Fontana, M. Scarpa, L.Pavesi, P. Bettotti, *ECS Trans.*, vol. 72, no. 34, pp. 1–6, 2016, doi:10.1149/07234.0001ecst.

5. “Silicon nanocrystals for nonlinear optics and secure communications” , Z. Bisadi, M. Mancinelli, S. Manna, S. Tondini, M. Bernard, A. Samusenko, M. Ghulinyan, G. Fontana, P. Bettotti, F. Ramiro-Manzano, G. Pucker, L. Pavesi, *Physica Status Solidi (a)*, vol. 212, no. 12, pp. 2659–2671, 2015. doi: 10.1002/pssa.201532528.

6. “Quantum random number generator based on silicon nanocrystals LED”, Z. Bisadi, A. Meneghetti, G. Fontana, G. Pucker, P. Bettotti, L. Pavesi, *Proc. SPIE*, vol. 9520, pp. 952004–952004–7, 2015. doi:10.1117/12.2179027.

Appendices

Appendix A

Gaussian distribution

Gaussian distribution or normal distribution is a continuous probability distribution (of a normal random variable Z) with the probability density function (pdf) defined as [143]:

$$f(z;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}}\ exp\ \left[-\frac{(z-\mu)^2}{2\sigma^2}\right],\quad |x|<\infty,\quad |\mu|<\infty,\quad \sigma>0. \tag{A.1}$$

The constants μ, σ and σ^2 are the mean, standard deviation and variance of the random variable Z, respectively.
A special case of a normal distribution with $\mu = 0$ and $\sigma = 1$ is called the *standard* normal distribution with pdf defined as [143]:

$$\phi(z) = \frac{1}{\sqrt{2\pi}}\ exp\ (-\frac{z^2}{2}),\quad |z|<\infty \tag{A.2}$$

The error function, $erf(z)$, is related to the cumulative distribution function (cdf) of the standard normal distribution $(\Phi(z))^2$ as [143]:

$$erf(z) = 2\Phi(z\sqrt{2}) - 1,\qquad z \geq 0, \tag{A.3}$$

corresponding to a normal pdf with variance $\frac{1}{2}$:

$$f(z;\ 0,\frac{1}{2}) = (1/\sqrt{\pi})\ e^{-z^2}.$$

The pdf of a normal distribution is unimodal (with a single mode at which the pdf has

[2] $\Phi(z)$ and $erf(z)$ are computes as:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{z} e^{-t^2/2}\ dt \qquad erf(z) = \frac{2}{\sqrt{\pi}}\int_{0}^{z} e^{-t^2}\ dt$$

its maximum value) with mean, median and mode at $z = \mu$ and $z = 0$ for generic and standard normal distribution, respectively. The normal distribution has a bell-shaped curve with $f(z)$ and $\phi(z)$ to be symmetric about $x = \mu$ and $z = 0$, respectively.

Appendix B

Poisson distribution

Poisson distribution is a discrete probability distribution for the counts (frequency) of an event (occurrences of an independent and identically distributed (i.i.d.) random variable) in a fixed interval (e.g. time interval). The probability of observing n number of occurrences of an i.i.d. random variable in the time interval t_0 is expressed as [144]:

$$P(n) = \frac{e^{-\lambda t_0}(\lambda t_0)^n}{n!}, \tag{B.1}$$

where λ is the average number of occurrences per interval.
The variance and mean of the Poisson distribution are the same and are equal to λ. Poisson distribution, contrary to the Gaussian distribution, is not symmetric and exhibits a positive skew which decreases as λ increases. As λ increases, the Poisson distribution can be approximated by the normal distribution and the pdf (Eq. B.1) becomes more bell-shaped (like the normal distribution).
The cdf of difference between any random arrival time t_0 and the next random arrival time t_1 has an exponential distribution [126]:

$$P(t_1 - t_0 \leq t) = 1 - P(t_1 - t_0 \geq t) = 1 - e^{-\lambda t}, \tag{B.2}$$

where λ is the number of detected events of the detector. By taking the derivative of the cdf, the pdf of the difference between any random arrival time t_0 and the next random arrival time t_1 is written as:

$$f(t) = \lambda\ e^{-\lambda t}. \tag{B.3}$$

The Poisson process has the property that if there is only one single arrival in a time interval [0 , t], the distribution of the arrival times is uniform throughout the interval. This can be proved by writing the conditional probability and substituting the joint probability with independent probabilities of one photon detection in $(0\ ,\ \tau]$ and no photon detection in (τ , t] [126]:

$$\begin{aligned}
P(T \leq \tau \mid N(t) = 1) = \frac{P(T \leq \tau, N(t) = 1)}{P(N(t) = 1)} \\
= \frac{P(1 \text{ event in } (0,\tau], 0 \text{ event in } (\tau, t])}{P(N(t) = 1)} \\
= \frac{P(1 \text{ event in } (0,\tau])\ P(0 \text{ event in } (\tau, t])}{P(N(t) = 1)} \\
= \frac{\lambda \tau e^{-\lambda \tau} e^{-\lambda(t-\tau)}}{\lambda t e^{-\lambda t}} = \frac{\tau}{t},
\end{aligned} \tag{B.4}$$

Given $N(t) = n$, the n arrival times $\tau_1,\ \ldots,\ \tau_n$ have the same distribution as the order statistics corresponding to n independent random variables uniformly distributed on the interval [0 , t]. Suppose $0 < t_1 < t_2 < \cdots < t_n < t_{n+1} \equiv t$ and let δ_i be small enough so that $t_i + \delta_i < t_{i+1}$, $i = 1, 2, \ldots, n$, we can write the conditional probability as [126]:

$$\begin{aligned}
& P(t_i \leq \tau_i \leq t_i + \delta_i, \quad i = 1, 2,\ \ldots,\ n \mid N(t) = n) \\
& = \frac{P(1 \text{ event in } [t_i, t_i + \delta_i], i = 1, 2, \ldots, n, \quad 0 \text{ event elsewhere in } [0, t])}{P(N(t) = n)} \\
& = \frac{\lambda \delta_1 e^{-\lambda \delta_1} \ldots \lambda \delta_n e^{-\lambda \delta_n} e^{-\lambda(t - \delta_1 - \delta_2 - \cdots - \delta_n)}}{e^{-\lambda t}(\lambda t)^n / n!} = \frac{n!}{t^n}\ \delta_1.\delta_2.\ \ldots\ .\delta_n
\end{aligned} \tag{B.5}$$

Therefore, dividing the conditional probability by $\delta_1.\delta_2.\ \ldots\ .\delta_n$ and by letting $\delta_i \rightarrow 0$, the conditional density of $\tau_1,\ \tau_2,\ \ldots, \tau_n$ at $t_1,\ t_2,\ \ldots,\ t_n$ would be [126]:

$$f_{\tau_1,\ \tau_2,\ \ldots,\ \tau_n}(t_1,\ t_2,\ \ldots,\ t_n \mid N(t) = n) = \frac{n!}{t^n}, \tag{B.6}$$

which is the pdf of the order statistics from a sample size n with the uniform distribution on [0, t].

Appendix C

Survival function

Let T be a non-negative random variable representing the waiting time until the occurrence of an event. We assume that T has the pdf, $f(t)$ and the cdf, $F(t) = P(T < t)$ giving the probability that the event has occurred by duration t. The *survival* function is the complementary cdf of T written as [145]:

$$S(t) = P(T \geq t) = 1 - F(t) = \int_t^\infty f(z)\ dz. \tag{C.1}$$

Which gives the probability of being *alive* just before duration t (the probability that the event of interest has not occurred by duration t). Let us define the instantaneous rate of occurrence of the event (called *hazard* function) as the conditional probability of the occurrence of the event in the interval $[t,\ t + dt)$ given that it has not occurred before, divided by the width of the interval dt [145]:

$$\begin{aligned}
\lambda(t) &= \lim_{dt \to 0} \frac{P(t \leq T \leq t + dt \mid T \geq t)}{dt} \\
&= \lim_{dt \to 0} \frac{P(t \leq T \leq t + dt \ \text{ and } \ T \geq t)}{P(T \geq t)} . \frac{1}{dt} \\
&= \lim_{dt \to 0} \frac{f(t)\ dt}{S(t)} . \frac{1}{dt} \\
&= \frac{f(t)}{S(t)},
\end{aligned} \tag{C.2}$$

where the conditional probability in the numerator is replaced by the ratio of the joint probability that T is in the interval $[t,\ t + dt)$ and $T \geq t$ (which is the same as the probability that t is in the interval), to the probability of the condition $T \geq t$. The former

may be written as $f(t)\ dt$ for small dt and the latter is $S(t)$ by definition (Eq. C.1). [145]
From Eq. C.1, it is seen that $f(t)$ is the derivative of $S(t)$. Therefore, the hazard function can be written as:

$$\lambda(t) = -\frac{d}{dt}\ \ln S(t), \tag{C.3}$$

which can be solved as:

$$\begin{aligned} &-\int_0^t \lambda(z)\ dz = \int_0^t d\ln S(t) = \ln S(t) - \ln S(0) \\ &\rightarrow \quad S(t) = \exp\{-\int_0^t \lambda(t)\ dt\}, \end{aligned} \tag{C.4}$$

where $S(0) = 1$ since $F(t) = 0$ in Eq. C.1. This formula gives the probability of surviving to duration t as a function of the hazard at all durations up to t. [145]
Assuming a constant rate of occurrence of the event, $\lambda(t) = \lambda$, the survival function is:

$$S(t) = exp(-\lambda t), \tag{C.5}$$

which is an exponential distribution with parameter λ. The pdf is obtained multiplying the survival function by λ: $f(t) = \lambda\ \exp(-\lambda t)$ and the mean turns out to be $1/\lambda$. [145]

Acknowledgements

Firstly, I would like to express my heartfelt gratitude to my supervisor Prof. Lorenzo Pavesi for the continuous support of my PhD study and related research. His guidance helped me in all the time of research and writing of this thesis. I appreciate all his contributions of time, ideas, and funding to make my PhD experience productive and stimulating.

My sincere thanks go to my dear colleague Dr. Giorgio Fontana for all his support and encouragement. I really enjoyed working with him in the laboratory and I appreciate all our fruitful discussions particularly during the development of our robust methodology to generate high quality random numbers. The experience of working with Giorgio has been undoubtedly one of the best experiences in my scientific life.

I would like to thank Nicola Massari, Dr. Fabio Acerbi and Dr. Georg Pucker from the Bruno Kessler Foundation (FBK) for the design and fabrication of the compact and the integrated chips of chapters 6 and 7 of this thesis. Their ideas and support really helped us to approach well the final goal of the Work Package 4 of the project SiQuro: the fabrication of a compact all-silicon-based quantum random number generator (QRNG).

The members of the NanoLab group have also contributed to my personal and professional time at UniTN. I am especially grateful to Dr. Fernando Ramiro Manzano, Stefano Biasi and Enrico Moser for their help with software programming, hardware installation and the laboratory facilities usage and maintenance. I would also like to thank Dr. Alessandro Tomasi and Dr. Alessio Meneghetti from the department of mathematics for their helpful discussions. I gratefully acknowledge the project SiQuro for the funding of my PhD work, as well.

My time in Trento and at UniTN was made enjoyable in large part due to the many friends and groups that became a part of my life. I would never forget the amazing trips we took to visit the beautiful cities in Italy and breathtaking places in Trentino. I particularly appreciate the scientific discussions on Matlab programming with my dear friend Dr. Payam Nazemzadeh.

Lastly, I would like to express my thanks to my family for all their love and encouragement. For my parents especially my dad who always supported me in all my decisions and pursuits. I highly appreciate all their profound love and tremendous support in my life.

Bibliography

[1] http://mathworld.wolfram.com/RandomNumber.html.

[2] Alfred J Menezes, Paul C Van Oorschot, and Scott A Vanstone. *Handbook of applied cryptography.* CRC press, 1996.

[3] Nicolas Gisin, Grégoire Ribordy, Wolfgang Tittel, and Hugo Zbinden. Quantum cryptography. *Reviews of modern physics*, 74(1):145, 2002.

[4] Charles H. Bennett and Gilles Brassard. Quantum cryptography: Public key distribution and coin tossing. In *Proceedings of the IEEE International Conference on Computers, Systems and Signal Processing*, pages 175–179. IEEE, 1984.

[5] Dirk P Kroese, Tim Brereton, Thomas Taimre, and Zdravko I Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6(6):386–392, 2014.

[6] SP Li. A guided monte carlo method for optimization problems. *International Journal of Modern Physics C*, 13(10):1365–1374, 2002.

[7] James E Gentle. *Random number generation and Monte Carlo methods.* Springer Science & Business Media, 2006.

[8] John F. Monahan. *Numerical Integration and Monte Carlo Methods*, page 257–302. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2 edition, 2011.

[9] http://www.gamasutra.com/blogs/CharlotteWalker/20141021/228292/Developing_Games_with_Random_Number_Generators.php, 2017.

[10] Atsushi Uchida. *Optical communication with chaotic lasers: applications of nonlinear dynamics and synchronization.* John Wiley & Sons, 2012.

[11] Mario Stipčević and Rupert Ursin. An on-demand optical quantum random number generator with in-future action and ultra-fast response. *Scientific Reports*, 5, 2015.

[12] Simply silicon. *Nature Photonics*, 4(8):491, 2010.

[13] Graham T Reed, G Mashanovich, FY Gardes, and DJ Thomson. Silicon optical modulators. *Nature photonics*, 4(8):518–526, 2010.

[14] Jurgen Michel, Jifeng Liu, and Lionel C Kimerling. High-performance ge-on-si photodetectors. *Nature Photonics*, 4(8):527–534, 2010.

[15] Project SIQURO. `http://events.unitn.it/en/siquro`.

[16] Francis Galton. Dice for statistical experiments. *Nature*, 42:13–14, 1890.

[17] Hermann A Haus. *Electromagnetic noise and quantum optical measurements*. Springer Science & Business Media, 2012.

[18] W Timothy Holman, J Alvin Connelly, and Ahmad B Dowlatabadi. An integrated analog/digital random noise source. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 44(6):521–528, 1997.

[19] Pong P Chu and Robert E Jones. Design techniques of fpga based random number generator. In *Military and Aerospace Applications of Programmable Devices and Technologies Conference*, volume 1, pages 28–30. Citeseer, 1999.

[20] DC Ranasinghe, D Lim, S Devadas, D Abbott, and PH Cole. Random numbers from metastability and thermal noise. *Electronics Letters*, 41(16):1, 2005.

[21] Robert H Miller Jr. Random bit stream generation by amplification of thermal noise in a cmos process, February 28 2006. US Patent 7,007,060.

[22] Cheol-min Kim. Low-power random bit generator using thermal noise and method thereof, April 20 2010. US Patent 7,702,701.

[23] Thomas J Chaney and Charles E Molnar. Anomalous behavior of synchronizer and arbiter circuits. *IEEE Transactions on computers*, 100(4):421–422, 1973.

[24] Carlos Tokunaga, David Blaauw, and Trevor Mudge. True random number generator with a metastability-based quality control. *IEEE Journal of Solid-State Circuits*, 43(1):78–85, 2008.

[25] The Intel random number generator. cryptography research, Inc. white paper prepared for Intel Corporation. `https://42xtjqm0qj0382ac91ye9exr-wpengine.netdna-ssl.com/wp-content/uploads/2015/08/IntelRNG.pdf`.

[26] `https://github.com/waywardgeek/infnoise/`.

[27] Lev I Berger. *Semiconductor materials.* CRC Press, 1997.

[28] `https://www.araneus.fi/products/alea2/en/`.

[29] `http://altusmetrum.org/ChaosKey/`.

[30] `http://ubld.it/products/truerng-hardware-random-number-generator/`.

[31] Shelan Khasro Tawfeeq. A random number generator based on single-photon avalanche photodiode dark counts. *Journal of Lightwave Technology*, 27(24):5665–5667, 2009.

[32] Don Davis, Ross Ihaka, and Philip Fenstermacher. Cryptographic randomness from air turbulence in disk drives. In *Annual International Cryptology Conference*, pages 114–120. Springer, 1994.

[33] Béla Vizvári and Géza Kolumban. Quality evaluation of random numbers generated by chaotic sampling phase-locked loops. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 45(3):216–224, 1998.

[34] Toni Stojanovski and Ljupco Kocarev. Chaos-based random number generators-part i: analysis [cryptography]. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 48(3):281–288, 2001.

[35] Atsushi Uchida, Kazuya Amano, Masaki Inoue, Kunihito Hirano, Sunao Naito, Hiroyuki Someya, Isao Oowada, Takayuki Kurashige, Masaru Shiki, Shigeru Yoshimori, et al. Fast physical random bit generation with chaotic semiconductor lasers. *Nature Photonics*, 2(12):728–732, 2008.

[36] Luis L Bonilla, Mariano Alvaro, and Manuel Carretero. Chaos-based true random number generators. *Journal of Mathematics in Industry*, 7(1):1, 2016.

[37] Limeng Zhang, Biwei Pan, Guangcan Chen, Lu Guo, Dan Lu, Lingjuan Zhao, and Wei Wang. 640-Gbit/s fast physical random number generation using a broadband chaotic semiconductor laser. *Scientific Reports*, 8, 2017.

[38] Wen Li, Igor Reidler, Yaara Aviad, Yuyang Huang, Helong Song, Yaohui Zhang, Michael Rosenbluh, and Ido Kanter. Fast physical random-number generation based on room-temperature chaotic oscillations in weakly coupled superlattices. *Physical review letters*, 111(4):044102, 2013.

[39] Qing Zhou, Xiaofeng Liao, Kwok-wo Wong, Yue Hu, and Di Xiao. True random number generator based on mouse movement and chaotic hash function. *information sciences*, 179(19):3442–3450, 2009.

[40] Di Xiao, Xiaofeng Liao, and Shaojiang Deng. Chaos based hash function. In *Chaos-Based Cryptography*, pages 137–203. Springer, 2011.

[41] Masatugu Isida and Hiroji Ikeda. Random number generator. *Annals of the Institute of Statistical Mathematics*, 8(1):119–126, 1956.

[42] Glenn F Knoll. *Radiation detection and measurement*. John Wiley & Sons, 2010.

[43] CH Vincent. The generation of truly random binary numbers. *Journal of Physics E: Scientific Instruments*, 3(8):594, 1970.

[44] RS Maddocks, S Matthews, EW Walker, and CH Vincent. A compact and accurate generator for truly random binary digits. *Journal of Physics E: Scientific Instruments*, 5(6):542, 1972.

[45] Ammar Alkassar, Thomas Nicolay, and Markus Rohe. Obtaining true-random binary numbers from a weak radioactive source. In *International Conference on Computational Science and Its Applications*, pages 634–646. Springer, 2005.

[46] J. Walker. HotBits: Genuine random numbers generated by radioactive decay [online]. `https://www.fourmilab.ch/hotbits/`.

[47] Gerhard Lutz. *Semiconductor radiation detectors*. Springer-Verlag Berlin Heidelberg, 2007.

[48] Miguel Herrero-Collantes and Juan Carlos Garcia-Escartin. Quantum random number generators. *arXiv preprint arXiv:1604.03304*, 2016.

[49] M Reznikov, R De Picciotto, M Heiblum, DC Glattli, A Kumar, and L Saminadayar. Quantum shot noise. *Superlattices and microstructures*, 23(3):901–915, 1998.

[50] Scott A Wilber. Entropy analysis and system design for quantum random number generators in cmos integrated circuits. 2013.

[51] `https://comscire.com/product/pq4000ks/`.

[52] Thomas Jennewein, Ulrich Achleitner, Gregor Weihs, Harald Weinfurter, and Anton Zeilinger. A fast and compact quantum random number generator. *Review of Scientific Instruments*, 71(4):1675–1680, 2000.

[53] `http://www.idquantique.com/random-number-generation/quantis-random-number-generator/`.

[54] André Stefanov, Nicolas Gisin, Olivier Guinnard, Laurent Guinnard, and Hugo Zbinden. Optical quantum random number generator. *Journal of Modern Optics*, 47(4):595–598, 2000.

[55] Ma Hai-Qiang, Wang Su-Mei, Zhang Da, Chang Jun-Tao, Ji Ling-Ling, Hou Yan-Xue, and Wu Ling-An. A random number generator based on quantum entangled photon pairs. *Chinese Physics Letters*, 21(10):1961, 2004.

[56] Harald Fürst, Henning Weier, Sebastian Nauerth, Davide G Marangon, Christian Kurtsiefer, and Harald Weinfurter. High speed optical quantum random number generation. *Optics Express*, 18(12):13029–13037, 2010.

[57] Min Ren, E Wu, Yan Liang, Yi Jian, Guang Wu, and Heping Zeng. Quantum random-number generator based on a photon-number-resolving detector. *Physical Review A*, 83(2):023820, 2011.

[58] Hai-Qiang Ma, Yuejian Xie, and Ling-An Wu. Random number generation based on the time of arrival of single photons. *Applied Optics*, 44(36):7760–7763, 2005.

[59] Mario Stipčević and B Medved Rogina. Quantum random number generator based on photonic emission in semiconductors. *Review of Scientific Instruments*, 78(4):045104, 2007.

[60] NM Thamrin, G Witjaksono, A Nuruddin, and MS Abdullah. A photonic-based random number generator for cryptographic application. In *Software Engineering, Artificial Intelligence, Networking, and Parallel/Distributed Computing, 2008. SNPD'08. Ninth ACIS International Conference on*, pages 356–361. IEEE, 2008.

[61] James F Dynes, Zhiliang L Yuan, Andrew W Sharpe, and Andrew J Shields. A high speed, postprocessing free, quantum random number generator. *Applied Physics Letters*, 93(3):031109, 2008.

[62] Michael A Wayne and Paul G Kwiat. Low-bias high-speed quantum random number generator via shaped optical pulses. *Optics Express*, 18(9):9351–9357, 2010.

[63] Michael Wahl, Matthias Leifgen, Michael Berlin, Tino Röhlicke, Hans-Jürgen Rahn, and Oliver Benson. An ultrafast quantum random number generator with provably bounded output bias based on photon arrival time measurements. *Applied Physics Letters*, 98(17):171105, 2011.

[64] You-Qi Nie, Hong-Fei Zhang, Zhen Zhang, Jian Wang, Xiongfeng Ma, Jun Zhang, and Jian-Wei Pan. Practical and fast quantum random number generation based on photon arrival time relative to external reference. *Applied Physics Letters*, 104(5):051110, 2014.

[65] Abbas Khanmohammadi, Reinhard Enne, Michael Hofbauer, and Horst Zimmermanna. A monolithic silicon quantum random number generator based on measurement of photon detection time. *IEEE Photonics Journal*, 7(5):1–13, 2015.

[66] Jian-min Wang, Tian-yu Xie, Hong-fei Zhang, Dong-xu Yang, Chao Xie, and Jian Wang. A bias-free quantum random number generation using photon arrival time selectively. *Photonics Journal, IEEE*, 7(2):1–8, 2015.

[67] Christian Gabriel, Christoffer Wittmann, Denis Sych, Ruifang Dong, Wolfgang Mauerer, Ulrik L Andersen, Christoph Marquardt, and Gerd Leuchs. A generator for unique quantum random numbers based on vacuum states. *Nature Photonics*, 4(10):711–715, 2010.

[68] Thomas Symul, SM Assad, and Ping K Lam. Real time demonstration of high bitrate quantum random number generation with coherent laser light. *Applied Physics Letters*, 98(23):231103, 2011.

[69] M Jofre, M Curty, F Steinlechner, G Anzolin, JP Torres, MW Mitchell, and V Pruneri. True random numbers from amplified quantum vacuum. *Optics Express*, 19(21):20665–20672, 2011.

[70] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. *Time series analysis: forecasting and control.* John Wiley & Sons, 2015.

[71] Autocorrelation function in matlab. `https://it.mathworks.com/help/econ/autocorr.html#btzjb3t`.

[72] `https://www.random.org/analysis/`.

[73] Geoffrey Grimmett and David Stirzaker. *Probability and random processes.* Oxford university press, 2001.

[74] Robert M. Gray. *Entropy and information theory.* Springer, 2011.

[75] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 547–561, 1961.

[76] Masahiko Higashi and George J Klir. Measures of uncertainty and information based on possibility distributions. *International Journal of General Systems*, 9(1):43–58, 1982.

[77] E Claude Shannon and Weaver Weaver. The mathematical theory of communication. *Urbana: University of Illinois Press*, 29, 1949.

[78] GM Bosyk, M Portesi, and A Plastino. Collision entropy and optimal uncertainty. *Physical Review A*, 85(1):012108, 2012.

[79] `http://csrc.nist.gov/publications/nistpubs/800-90A/SP800-90A.pdf`.

[80] http://csrc.nist.gov/publications/nistpubs/800\protect\kern+.1667em\relax90a/sp80090a.pdf/.

[81] http://csrc.nist.gov/groups/st/toolkit/rng/documents/sp80022rev1a.pdf/.

[82] Kenji Hamano and Toshinobu Kaneko. Correction of overlapping template matching test included in nist randomness test suite. *IEICE transactions on fundamentals of electronics, communications and computer sciences*, 90(9):1788–1792, 2007.

[83] http://www.iro.umontreal.ca/~lecuyer/myftp/papers/testu01.pdf.

[84] Christophe Guyeux, Qianxue Wang, and Jacques M Bahi. A pseudo random numbers generator based on chaotic iterations: application to watermarking. In *International Conference on Web Information Systems and Mining*, pages 202–211. Springer, 2010.

[85] http://www.iro.umontreal.ca/~simardr/testu01/guideshorttestu01.pdf.

[86] Lorenzo Pavesi and David J Lockwood. *Silicon photonics*, volume 1. Springer Science & Business Media, 2004.

[87] A Anopchenko, A Marconi, M Wang, G Pucker, P Bellutti, and L Pavesi. Graded-size si quantum dot ensembles for efficient light-emitting diodes. *Applied Physics Letters*, 99(18):181108, 2011.

[88] Lorenzo Pavesi and Rasit Turan. *Silicon nanocrystals: fundamentals, synthesis and applications*. John Wiley & Sons, 2010.

[89] A Marconi, A Anopchenko, Minghua Wang, G Pucker, P Bellutti, and L Pavesi. High power efficiency in si-nc/sio2 multilayer light emitting devices by bipolar direct tunneling. *Applied Physics Letters*, 94(22):221110, 2009, Reprinted with the permission of AIP Publishing.

[90] M Wang, A Anopchenko, A Marconi, E Moser, S Prezioso, L Pavesi, G Pucker, P Bellutti, and L Vanzetti. Light emitting devices based on nanocrystalline-silicon multilayer structure. *Physica E: Low-dimensional Systems and Nanostructures*, 41(6):912–915, 2009.

[91] Matthew W Fishburn. *Fundamentals of cmos single-photon avalanche diodes*. fishburn, 2012.

[92] Sergio Cova, M Ghioni, A Lacaita, C Samori, and F Zappa. Avalanche photodiodes and quenching circuits for single-photon detection. *Applied Optics*, 35(12):1956–1976, 1996.

[93] Andreas Eisele, Robert Henderson, Bernd Schmidtke, Tobias Funk, Lindsay Grant, Justin Richardson, and Wolfgang Freude. 185 mhz count rate 139 db dynamic range single-photon avalanche diode with active quenching circuit in 130 nm cmos technology. In *Proc. Int. Image Sensor Workshop*, pages 278–280, 2011.

[94] Robert H Hadfield. Single-photon detectors for optical quantum information applications. *Nature photonics*, 3(12):696–705, 2009.

[95] RM Stevenson, RM Thompson, AJ Shields, I Farrer, BE Kardynal, DA Ritchie, and M Pepper. Quantum dots as a photon source for passive quantum key encoding. *Physical Review B*, 66(8):081302, 2002.

[96] Makoto Naruse, Song-Ju Kim, Masashi Aono, Hirokazu Hori, and Motoichi Ohtsu. Chaotic oscillation and random-number generation based on nanoscale optical-energy transfer. *Scientific Reports*, 4, 2014.

[97] Simone Tisa and Franco Zappa. One-chip quantum random number generator. In *SPIE OPTO: Integrated Optoelectronic Devices*, pages 72360J–72360J. International Society for Optics and Photonics, 2009.

[98] Bruno Sanguinetti, Anthony Martin, Hugo Zbinden, and Nicolas Gisin. Quantum random number generation on a mobile phone. *Physical Review X*, 4(3):031056, 2014.

[99] Bing Qi, Yue-Meng Chi, Hoi-Kwong Lo, and Li Qian. High-speed quantum random number generation by measuring phase noise of a single-mode laser. *Optics Letters*, 35(3):312–314, 2010.

[100] Ido Kanter, Yaara Aviad, Igor Reidler, Elad Cohen, and Michael Rosenbluh. An optical ultrafast random bit generator. *Nature Photonics*, 4(1):58–61, 2010.

[101] Yi Jian, Min Ren, E Wu, Guang Wu, and Heping Zeng. Two-bit quantum random number generator based on photon-number-resolving detection. *Review of Scientific Instruments*, 82(7):073109, 2011.

[102] Simone Tisa, Federica Villa, Andrea Giudice, Georg Simmerle, and Franco Zappa. High-speed quantum random number generation using cmos photon counting detectors. *IEEE Journal of Selected Topics in Quantum Electronics*, 21(3):23–29, 2015.

[103] MJ Applegate, O Thomas, JF Dynes, ZL Yuan, DA Ritchie, and AJ Shields. Efficient and robust quantum random number generation by photon number detection. *Applied Physics Letters*, 107(7):071106, 2015.

[104] Zahra Bisadi, Alessio Meneghetti, Giorgio Fontana, Georg Pucker, Paolo Bettotti, and Lorenzo Pavesi. Quantum random number generator based on silicon nanocrystals led. In *SPIE Microtechnologies*, pages 952004–952004. International Society for Optics and Photonics, 2015.

[105] Zahra Bisadi, Alessio Meneghetti, Alessandro Tomasi, Giorgio Fontana, Andrea Tengattini, Paolo Bettotti, Georg Pucker, Massimiliano Sala, and Lorenzo Pavesi. A post-processing free si nanocrystals based quantum random number generator. In *European Quantum Electronics Conference*, page EA_P_32. Optical Society of America, 2015.

[106] Zahra Bisadi, Mattia Mancinelli, Santanu Manna, Stefano Tondini, Martino Bernard, Alina Samusenko, Mher Ghulinyan, Giorgio Fontana, Paolo Bettotti, Fernando Ramiro-Manzano, Georg Pucker, and Lorenzo Pavesi. Silicon nanocrystals for nonlinear optics and secure communications. *physica status solidi (a)*, 212(12):2659–2671, 2015, Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.

[107] Zahra Bisadi, Alessio Meneghetti, Alessandro Tomasi, Andrea Tengattini, Giorgio Fontana, Georg Pucker, Paolo Bettotti, Massimiliano Sala, and Lorenzo Pavesi. Generation of high quality random numbers via an all-silicon-based approach. *physica status solidi (a)*, 213(12):3186–3193, 2016, Copyright Wiley-VCH Verlag GmbH & Co. KGaA. Reproduced with permission.

[108] Zahra Bisadi, Chiara Piotto, Giorgio Fontana, Marina Scarpa, Lorenzo Pavesi, and Paolo Bettotti. (invited) silicon nanostructures: A versatile material for photonics. In *Meeting Abstracts*, number 42, pages 2092–2092. The Electrochemical Society, 2016.

[109] Lawrence D Brown and Linda H Zhao. A test for the poisson distribution. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 611–625, 2002.

[110] Christian Walck. Handbook on statistical distributions for experimentalists, 2007.

[111] Ronald Newbold Bracewell and Ronald N Bracewell. *The Fourier transform and its applications*, volume 31999. McGraw-Hill New York, 1986.

[112] H Paul. Photon antibunching. *Reviews of Modern Physics*, 54(4):1061, 1982.

[113] DN Qu and JC Dainty. A multichannel detector for photon correlation. *Advances in Electronics and Electron Physics*, 74:107–118, 1988.

[114] Aleksei Anopchenko, Alessandro Marconi, Fabrizio Sgrignuoli, Laura Cattoni, Andrea Tengattini, Georg Pucker, Yoann Jestin, and Lorenzo Pavesi. Electroluminescent devices based on nanosilicon multilayer structures. *physica status solidi (a)*, 210(8):1525–1531, 2013.

[115] Robert GW Brown, Kevin D Ridley, and John G Rarity. Characterization of silicon avalanche photodiodes for photon correlation measurements. 1: Passive quenching. *Applied Optics*, 25(22):4122–4126, 1986.

[116] `https://people.math.osu.edu/husen.1/teaching/571/markov_1.pdf`.

[117] Feihu Xu, Bing Qi, Xiongfeng Ma, He Xu, Haoxuan Zheng, and Hoi-Kwong Lo. Ultrafast quantum random number generation based on quantum phase fluctuations. *Optics Express*, 20(11):12366–12377, 2012.

[118] Fang-Xiang Wang, Chao Wang, Wei Chen, Shuang Wang, Fu-Sheng Lv, De-Yong He, Zhen-Qiang Yin, Hong-Wei Li, Guang-Can Guo, and Zheng-Fu Han. Robust quantum random number generator based on avalanche photodiodes. *Journal of Lightwave Technology*, 33(15):3319–3326, 2015.

[119] Xiongfeng Ma, Feihu Xu, He Xu, Xiaoqing Tan, Bing Qi, and Hoi-Kwong Lo. Postprocessing for quantum random-number generators: Entropy evaluation and randomness extraction. *Physical Review A*, 87(6):062327, 2013.

[120] John Von Neumann. Various techniques used in connection with random digits. *Applied Math Series*, 12:36–38, 1951.

[121] Qiurong Yan, Baosheng Zhao, Zhang Hua, Qinghong Liao, and Hao Yang. High-speed quantum-random number generation by continuous measurement of arrival time of photons. *Review of Scientific Instruments*, 86(7):073113, 2015.

[122] Zahra Bisadi, Giorgio Fontana, Enrico Moser, Georg Pucker, and Lorenzo Pavesi. Robust quantum random number generation with silicon nanocrystals light source. *Journal of Lightwave Technology*, 2017.

[123] *Advanced Photon Counting: Applications, Methods, Instrumentation.* Springer International Publishing, 2015.

[124] Stanislav Kalinin, Ralf Kühnemuth, Hayk Vardanyan, and Claus AM Seidel. Note: A 4 ns hardware photon correlator based on a general-purpose field-programmable gate array development board implemented in a compact setup for fluorescence correlation spectroscopy. *Review of Scientific Instruments*, 83(9):096105, 2012.

[125] Sergey V Polyakov, Michael Ware, and Alan Migdall. High-accuracy calibration of photon-counting detectors. In *Optics East 2006*, pages 63720J–63720J. SPIE-International Society for Optics and Photonics, 2006.

[126] Sheldon M Ross. *Applied probability models with optimization applications*. Courier Corporation, 2013.

[127] `https://www.sitime.com/support2/documents/AN10007-Jitter-and-measurement.pdf`.

[128] Carlos Abellán, Waldimar Amaya, Daniel Mitrani, Valerio Pruneri, and Morgan W Mitchell. Generation of fresh and pure random numbers for loophole-free bell tests. *Physical review letters*, 115(25):250403, 2015.

[129] `http://www.ieee.li/pdf/essay/pin_diode_handbook.pdf`.

[130] Dieter Renker. Geiger-mode avalanche photodiodes, history, properties and problems. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 567(1):48–56, 2006.

[131] DW Townsend. Multimodality imaging of structure and function. *Physics in medicine and biology*, 53(4):R1–R39, 2008.

[132] Carel WE Van Eijk. Radiation detector developments in medical applications: inorganic scintillators in positron emission tomography. *Radiation protection dosimetry*, 129(1–3):13–21, 2008.

[133] William W Moses. Recent advances and future advances in time-of-flight pet. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 580(2):919–924, 2007.

[134] Alberto Dalla Mora, Davide Contini, Simon Arridge, Fabrizio Martelli, Alberto Tosi, Gianluca Boso, Andrea Farina, Turgut Durduran, Edoardo Martinenghi, Alessandro Torricelli, et al. Towards next-generation time-domain diffuse optics for extreme depth penetration and sensitivity. *Biomedical Optics Express*, 6(5):1749–1760, 2015.

[135] Fabio Acerbi, Alessandro Ferri, Gaetano Zappala, Giovanni Paternoster, Antonino Picciotto, Alberto Gola, Nicola Zorzi, and Claudio Piemonte. Nuv silicon photomultipliers with high detection efficiency and reduced delayed correlated-noise. *IEEE transactions on Nuclear Science*, 62(3):1318–1325, 2015.

[136] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[137] Carlos Abellan, Waldimar Amaya, David Domenech, Pascual Muñoz, Jose Capmany, Stefano Longhi, Morgan W Mitchell, and Valerio Pruneri. Quantum entropy source on an inp photonic integrated circuit for random number generation. *Optica*, 3(9):989–994, 2016.

[138] Roger Newman. Visible light from a silicon p- n junction. *Physical Review*, 100(2):700–703, 1955.

[139] AG Chynoweth and KG McKay. Photon emission from avalanche breakdown in silicon. *Physical Review*, 102(2):369–376, 1956.

[140] Nader Akil, Sherra E Kerns, David V Kerns, Alain Hoffmann, and J-P Charles. A multimechanism model for photon generation by silicon junctions in avalanche breakdown. *IEEE Transactions on Electron Devices*, 46(5):1022–1028, 1999.

[141] VE Houtsma, J Holleman, N Akil, LM Phoung, V Zieren, A van den Berg, H Walling, and PH Woerlee. Visible light emission from reverse-biased silicon nanometer-scale diode-antifuses. In *Semiconductor Conference, 1999. CAS'99 Proceedings. 1999 International*, volume 2, pages 461–465. IEEE, 1999.

[142] DK Gautam, WS Khokle, and KB Garg. Photon emission from reverse-biased silicon pn junctions. *Solid-state electronics*, 31(2):219–222, 1988.

[143] Jagdish K Patel and Campbell B Read. *Handbook of the normal distribution*, volume 150. CRC Press, 1996.

[144] Frank Avery Haight. Handbook of the poisson distribution. 1967.

[145] `http://data.princeton.edu/wws509/notes/c7.pdf`.