Single photon detection for quantum technologies

Amri, Emna

Université de Genève
Groupe de Physique Appliquée

Faculté des Sciences
Professeur Hugo Zbinden

Single photon detection for quantum technologies

Thèse

présentée à la Faculté des sciences de l'Université de Genève
pour obtenir le grade de Docteur en sciences, mention Physique

par

Emna Amri

de SFAX (Tunisie)

Thèse N 5510

Genève
Centre d'impression de l'Université de Genève
2021

Abstract

During this thesis work, two approaches to single-photon detection at telecom wavelengths have been investigated. The first one focuses on improving the performance of commercially available semiconductor single photon avalanche diodes (SPADs) to meet quantum communication requirements, and the second one describes the development and characterization of high performance superconducting devices. On the application side, quantum random number generation (QRNG) was implemented using two schemes based on two different types of single photon detectors. In the first part of the thesis I investigated the operation of free-running InGaAs/InP negative feedback avalanche diodes (NFADs) with a particular focus on timing jitter and afterpulsing. Through an extensive study of the low temperature behavior of these detectors, insights into the fundamental origin of timing jitter were given. The key finding is that NFADs with a breakdown voltage higher than $\sim$ 67 V can combine a very low timing jitter (52 ps) with an extremely low dark count rate (1 c.p.s) when operated at low temperatures. A new method to decrease afterpulsing in free-running NFADs was also implemented and the preliminary results showed a reduction of afterpulsing probability at low delays. The second part of the thesis concentrated on the development, fabrication and characterization of large active-area superconducting nanowire single-photon detectors (SNSPDs). This type of detector is very beneficial for applications requiring free-space and multimode fibers coupling. A high kinetic inductance is however unavoidable for long nanowires which leads to a long recovery time. The approach proposed in this work is to make large sensitive-area Parallel-SNSPDs, where several nanowires are connected in parallel and cover a large area. This design would mitigate the speed problem while guaranteeing the excellent attributes provided by standard SNSPDs. Using flood-illuminated 50 μm^2 active-area devices, we demonstrated a saturated detection efficiency with a count rate as high as 40 MHz.
The final part of the thesis looks at the application of single photon detectors to quantum random number generation. Two implementations are proposed, both using photo-sensitive transducers activated by a LED. The first QRNG is based on a CMOS SPADs matrix and can be regarded as a stand-alone system offering high

level of integration with a throughput of 400 Mbit/s of random data. The second one is based on quanta image sensors (QIS) and it showed promising advantages over previous QRNG technologies combining the advantages of both SPADs and conventional CMOS image sensors. This QRNG showed high randomness quality (near-unity entropy/bit for raw data) and promising data rate (in an array of 2.5 mm^2 area we can fit millions of jots and reach up to 12 Gb/s throughput).

Résumé

Au cours de cette thèse, deux approches pour la détection de photon unique ont été étudiées. La première approche se focalise sur l'amélioration des performances des diodes à avalanches déclenchées par photon unique, disponibles sur le marché. La deuxième approche est basée sur des détecteurs supraconducteurs à hautes performances développés en interne. Le but général est de répondre aux exigences des applications de communication quantique. Du point de vue application de détecteurs de photons, on a présenté deux architectures de générateurs de nombres aléatoires quantiques basées sur deux types de détecteurs de photons.
Dans la première partie de cette thèse, j'ai étudié le mode d'opération asynchrone des photodiodes à avalanche composées des matériaux InGaAs/InP et intégrant les éléments passifs permettant d'arrêter l'avalanche (désignées dans la suite par leur acronyme anglais "NFADs"). Je me suis particulièrement intéressée à leur résolution temporelle et au phénomène de "Afterpulsing". D'abord, une étude approfondie du fonctionnement de ces détecteurs à basse température nous a permis de comprendre l'origine de leur gigue temporelle et on a pu montrer que les NFADs ayant une tension de claquage supérieure à 67 V peuvent combiner une gigue temporelle aussi basse que 52 ps avec un bruit thermique extrêmement faible (1 c.p.s) quand elles sont utilisées à basses températures. On a pu aussi mettre en place une nouvelle méthode pour diminuer l'Afterpulsing des NFADs asynchrones et les résultats préliminaires sont prometteurs. La deuxième partie de cette thèse se focalise sur le développement, la fabrication et la caractérisation de nano-fils supraconducteurs capables de détecter des photons uniques (désignés dans la suite par leur acronyme anglais "SNSPDs") à larges surfaces photosensibles. Ce type de détecteur peut être bénéfique pour les applications qui nécessitent un couplage de lumière direct ou à travers des fibres optiques multimodes. Malheureusement, la grande taille de ces détecteurs implique une grande inductance cinétique qui va dégrader la vitesse de détection. Pour surmonter ce problème et garder les bonnes performances des SNSPDs standards, on propose ici une solution originale qui consiste à utiliser plusieurs nano-fils connectés en parallèle pour couvrir la même grande surface. Cette approche nous a permis de démontrer une courbe de

détection saturée et un taux de comptage de 40 MHz avec un détecteur parallèle couvrant une surface de 50 μm^2.
La dernière partie de cette thèse traite de la génération de nombres aléatoires quantiques en utilisant des détecteurs de photons. La première implémentation proposée ici est basée sur une matrice de photodiodes à avalanche intégrée sur le même substrat que l'électronique qui effectue l'extraction de l'entropie en temps réel. Ce QRNG offre un haut taux de génération de données aléatoires s'élevant à 400 Mbit/s. Le deuxième QRNG est basé sur un nouveau type de capteurs d'images sensible au photon unique qui lui a garanti un taux de génération très élevé (5-12 Gb/s) avec une consommation d'énergie minimale.

Contents

Chapter 1

Introduction

Light measurement and manipulation at the single photon level has supported and enabled an expanding range of applications covering several topics of science and engineering. The most straightforward field one can think of is low-level light sensing such us medical imaging, astronomy and low ambient light surveillance. Another more demanding category includes applications where the quantum nature of light is the key component. This is mainly applicable to optical quantum information [1] and quantum metrology. For instance quantum information tech-

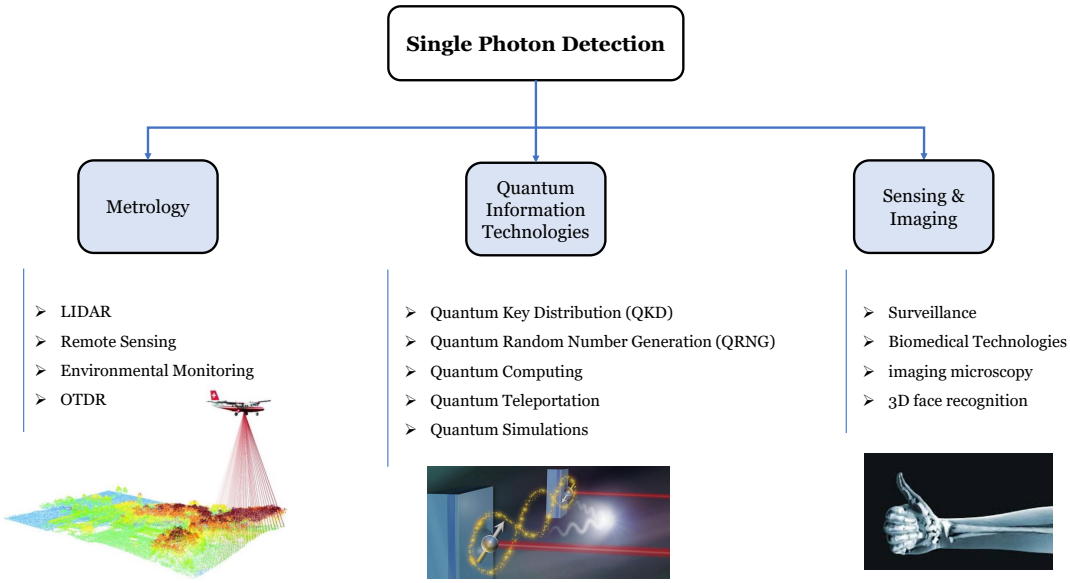


Figure 1.1: Main applications of Single Photon Detection

nologies (QIT) use photons to encode, process and generate data according to the laws of quantum physics [2]. Within this innovative field lies Quantum key distribution (QKD) [3], the most secure communication framework reported to date, and quantum random number generators (QRNGs) [4, 5], that use the intrinsic random nature of light to produce true randomness. These two commercially mature technologies are among the major drivers for the development and improvement of single-photon detectors.

Single Photon Detectors

A single photon detector (SPD) is an extremely sensitive device able to register energies as low as 10^{-19}J. Besides its high sensitivity, an SPD should exhibit the highest performance in terms of spectral range, count rate and time resolution in order to keep up with the extreme demands of rapidly-expanding applications (Figure 1.1). To assess the compatibility of a single photon detector with a specific application we should characterize it according to the following properties:

- *Photon detection efficiency (PDE):* The probability that an incident photon will be detected and will generate an output signal. An ideal SPD would have a unity detection efficiency which is not possible in practical implementations because of hardware imperfections and the dependence of this property on the fluctuating operation conditions.
- *Dark count rate (DCR):* The probability that the SPD registers a detection event in the absence of incoming light. These false counts are usually dependent on the temperature, the biasing solution and the material properties.
- *Afterpulsing*: The probability of false counts correlated to previous photon detections. This characteristic is relevant for III/V semiconductor SPDs (See section 2.3 for more details).
- *Timing jitter:* The temporal variation between the absorption of a photon and the generation of an output electrical pulse. It can be also defined as the time uncertainty on the registering time of the detection. For a typical Gaussian distribution, the jitter can be quantified as the full width at half maximum (FWHM) or additionally as the 1/100 maximum of the distribution.
- *Dead-time:* The time period that follows a detection event, during which the detector is not active. For some SPDs, the dead-time designates also the recovery-time but we like to describe the recovery-time as the shortest possible dead-time which is, in the optimal case, limited by the detector physics. In the case of semiconductor-based SPD, the dead-time is made

long enough to suppress afterpulsing at the expense of the maximum count rate (MCR).

- *Spectral range:* The range of absorbed energies determined by the band-gap of the constituent materials. For semiconductor SPDs, Silicon is the best choice for visible applications while InGaAs has given the best results for Infra-Red (IR) sensing.
- *Maximum exposure level:* The incoming light intensity above which the detector may undergo temporary or permanent damage.
- *Photon number resolution:* The ability to distinguish and count the number of photons in each incoming optical pulse. This criteria is important for advanced quantum information protocols [6].

Some of these characteristics delineate intrinsic properties while others are wavelength and temperature dependent and can also vary across the spatial dimensions of the detector. Note that some technologies may only demonstrate a subset of the features listed above. For instance, superconducting SPDs do not suffer from afterpulsing phenomena.

Many technologies have been developed for single photon detection. Some of them are already well-established and others are still emerging, driven by the new cutting-edge applications.

Vacuum photo-multiplier tubes (PMTs) were the first-demonstrated and commercialized single photon detectors [7]. They are commonly used for ultraviolet and visible radiations measurement and are not efficient with low-energy radiations (infrared and microwave). The detection process starts when the photon flux strikes the photo-active cathode and dislodges electrons. The photocurrent signal is then amplified through a series of dynodes before being sensed at the anode. The main advantage of a PMT is its ability to detect very weak signals thanks to the amplification effect. However, this same effect can be problematic since any spurious signal is also amplified leading to a lower signal to noise ratio (SNR). Despite their good performance at visible wavelengths [8], PMTs still have large dimensions and are quite expensive, which is why they have been replaced by single photon avalanche diodes in most of the applications.

Single photon avalanche diodes (SPADs) are the most common and commercially available technology for single photon detection. These highly sensitive semiconductor devices generate a measurable photo-current when an irradiation arrives at its active area [9]. SPADs are reverse-biased above their breakdown voltage which initiates an impact ionization process when an electron-hole pair is photo-generated. The self-sustainable avalanche must be stopped and the de-

vice reset using an appropriate quenching circuit (more details are provided in Section 2.1). Silicon SPADs have replaced PMTs in the visible range thanks to their integration potential, high performance and lower power consumption. These detectors can also operate at telecom wavelengths using lower-band-gap semiconductor materials, such as Ge and InGaAs/InP (see Chapter 2). However, their performance cannot compete with the recently-developed superconducting SPDs.

Superconducting nanowire single photon detectors (SNSPDs) demonstrate single-photon sensitivity from X-ray to mid-infrared wavelengths, together with a high efficiency ($> 90\%$) [10, 11], low dark count rate, low timing jitter and fast recovery (see Chapter 3). These nano-devices do not suffer from afterpulsing effect and, depending on their design, can exhibit photon-number resolving capability. A major limitation is that they must be operated at cryogenic temperatures, which can be difficult to implement. The superconducting nanowire is then biased just below its critical current, and a localized resistive hot-spot is created when a photon is absorbed, triggering an output voltage-pulse.

Table 1.1 reports the best performance of these three technologies. Note that the detectors do not necessarily combine simultaneously all of the performance listed in the table.

Technology	Spectral range	Op. Temp.	SDE	DCR	Jitter	MCR
PMT (VIS-NIR)	VIS-NIR	300 K	40% at 550 nm	100 kHz	300 ps	10 MHz
PMT (IR)	IR	200 K	2% at 1550 nm	200 KHz	300 ps	10 MHz
Si SPAD	VIS	300 K	65% at 650 nm	25 Hz	35 ps	10 MHz
InGaAs SPAD	NIR	240 K	55% at 1550 nm	1 Hz	50 ps	100 MHz
SNSPD	X-ray to NIR	1-3 K	95% at 1550 nm	$10^{-3}\ Hz$	3 ps	200 MHz

Table 1.1: Summary of the performances of photomultiplier tube (PMT) [12, 13], single photon avalanche diode (SPAD) [14, 15, 16, 17] and superconducting nanowire single photon detector (SNSPD) [18, 19, 20].

In addition to these well-established SPDs, a host of new single-photon detector technologies such as visible-light photon counters (VLPCs) [21], superconducting transition-edge sensors (STESs) [22] and quantum dot field-effect transistors (QDFETs) [23] have been demonstrated and start to be deployed. An exhaustive review of these technologies is provided in [24, 25, 26].

Outline of the thesis

In this thesis, I focused on Indium Gallium Arsenide single photon avalanche diodes (InGaAs SPADs) and superconducting nanowires single photon detectors (SNSPDs), being the most commonly used for quantum communication applications and at the core business of Id Quantique Quantum Sensing division.
The industrial framework of this thesis gave me the opportunity to work with other SPDs (CMOS SPADs and Quanta Image Sensors (QIS)) used for new implementations of quantum random number generators. In chapter 2, I focused on improving the performance of free-running InGaAs/InP SPADs operating at low temperature (IDQ - ID230). After an extensive study of the timing jitter mechanism, I showed that these detectors can combine a very low jitter with the lowest dark count when operating in well-defined conditions. Then, I demonstrated a reduction of afterpulsing probability using a quantum cascade laser to release the charges trapped inside the material defects.

Chapter 3 is dedicated to SNSPDs, it covers the main aspects of this technology in terms of operation principle, design, fabrication and characterization. A special focus was brought in large active-area SNSPDs for multimode-fibers coupling that were developed within a collaborative project between Id Quantique and the University of Geneva.
The importance of a technology is illustrated by its applications and that is why I dedicated chapter 4 to quantum random number generation using single photon detectors. Two different implementations are proposed with a full description of the randomness generation process and a characterization of the systems in terms of bit rate, SNR, scalability and power consumption.

The last chapter summarizes the results of this work and discusses future research directions and open problems. Finally, a list of peer-reviewed articles, pre-prints and patents issued during the course of this thesis is provided.

Chapter 2

Single-photon detectors based on InGaAs/InP avalanche diodes

Indium-Gallium-Arsenide/Indium-Phosphide (InGaAs/InP) single-photon avalanche diodes (SPADs) are a popular detector choice for the near-infrared range (1000 nm - 1700 nm) thanks to their commercial maturity, compact size, good performances and ease of operation (cryogenic temperatures are not required). These III-V heterostructure devices have separate photon absorption, charge and multiplication regions as shown in Figure 2.1 When operating in Geiger mode, the reverse-bias voltage of the SPAD is larger than its breakdown voltage V_{br} and the electric field in the multiplication layer (InP) is large enough to trigger impact ionization phenomena. The incident photons will cross the InP layer and will be absorbed in the narrower-gap InGaAs ($In_{0.53}Ga_{0.47}As$) layer producing electron–hole pairs. Photogenerated carriers created in the absorption layer will drift toward the multiplication region under the effect of the electric field. In the case of InP, holes have higher impact ionization coefficient and will trigger a self-sustaining avalanche [27, 28]. However, a direct holes transport from the InGaAs to the InP layer would be very difficult because of the valence band discontinuity between the two materials ($E_{g,InGaAs}$=0.7eV and $E_{g,InP}$=1.35eV at room temperature). That is why an additional layer (InGaAsP) with an intermediate bandgap, called grading layer, is added between the two materials to grade the valence-band discontinuity and, hence, increase the transition rate of the holes.

InGaAs/InP SPADs can be operated in synchronous (gated) and asynchronous (free-running) modes depending on the application requirements. In the gated scheme, the SPAD is working in Geiger mode only during a specific time gates (typically less than a nanosecond [29, 30]) while in the free-running regime, the detector is constantly working, that is why either active or passive quenching is

required to stop the avalanche and reset the detector [31].

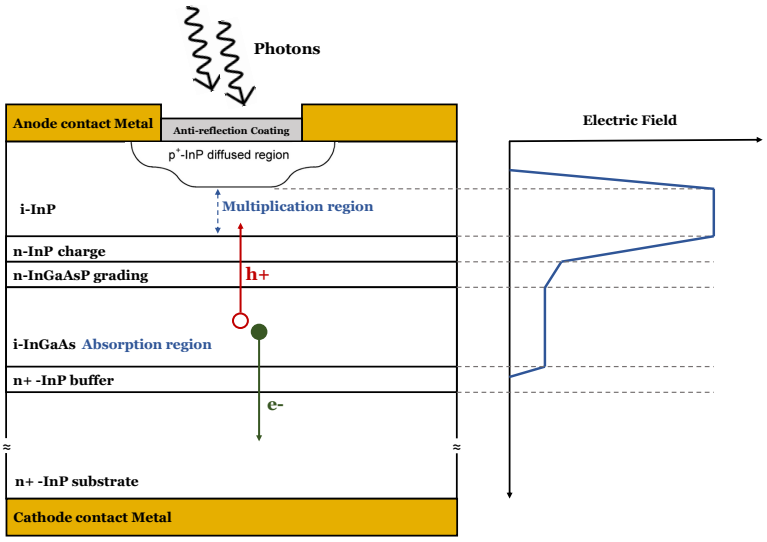


Figure 2.1: Cross-section schematic of a front-illuminated planar InGaAs/InP SPAD. The electric field along the different layers is shown on the right.

In this work I mainly used SPADs operating in the free-running regime implemented with passive quenching.

2.1 Free-running operation and passive quenching

For many applications, where the time of arrival of the photon is unknown, operation in the free-running regime is required. For instance, in time resolved photoluminescence (TRPL) [32] lifetimes may range from subnanosecond to tens of nanoseconds so the detector should be constantly armed ready to detect the emission of the material after excitation. Same thing for Lidar [33] where reflected photons are asynchronous, biomedical imaging [34, 35], quantum dot emission and many other applications.

When a photon is detected and an impact ionization process is triggered, the current continues to flow across the diode and has to be quenched to allow the detector to be reset and armed for the next detection. This is usually done electrically using active or passive electrical components.
In this work I only used passive quenching (PQ) where a high voltage resistor ($\sim$1 MΩ) is serially connected to the SPAD with the DC bias source used to set the reverse-bias V_{bias} ($V_{bias} \geq V_{br}$) as depicted if Figure 2.2. When avalanche events

occur the same current flows across the SPAD and the quenching resistor. The voltage across the resistor increases and the voltage difference between the anode and the cathode of the diode decreases simultaneously until falling below V_{br}, quenching the avalanches.
This passive quenching technique can be implemented using discrete components or more efficiently, by integrating the resistor on the same substrate as the SPAD, to obtain Negative Feedback Avalanche Diodes (NFADs) known as the most effective self-quenching implementation [36, 37].

NFADs have many advantages over traditional SPADs. Not only they provide photon counting operation with just a simple DC bias voltage but they also allow higher density integration of several adjacent SPADs and, more importantly, they have a shorter recovery time after avalanche since the use of integrated resistors dramatically reduces the parasitic capacitance with respect to discrete components.

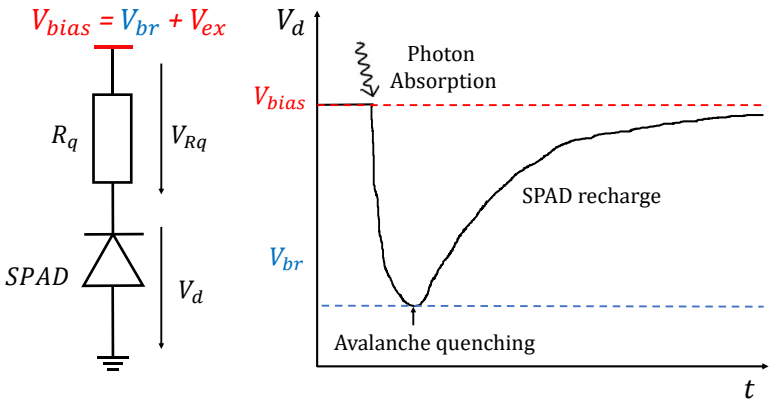


Figure 2.2: Electonic modelling of Passive Quenching Circuit where the quenching resistor is placed serially with the SPAD. At the right, the voltage diagram across the SPAD during the avalanche event is shown.

In this chapter I mainly used negative feed-back avalanche diodes to study first, timing jitter and then Afterpulsing effect. The results are presented in the sections 2.2 and 2.3.

2.2 Temporal jitter in free-running InGaAs/InP SPADs

Negative-feedback avalanche diodes (NFADs) provide a practical solution for different single-photon counting applications requiring free-running mode operation with low afterpulsing probability. Unfortunately, their timing jitter has never been as good as for gated InGaAs/InP single-photon avalanche diodes [38]. Moreover, the mechanism of jitter at low temperatures has not been thoroughly studied in NFADs (and in InGaAs/InP SPADs as a whole) , which is crucial if operation with a low DCR is an additional requirement.

The timing jitter is the time resolution of the photon arrival on the SPAD. A detailed definition is given in the first chapter.

In this section, I present a characterization of the time response of different InGaAs/InP based NFADs with particular focus on the temperature dependence and the effect of carrier transport between the absorption and multiplication regions. Based on this study, we were able to define the connection between the lowest achievable DCR and the temporal jitter, when operating at low temperatures. More details about this work are found in [17].

We characterized four different NFADs manufactured by Princeton Lightwave which have different feedback resistances and active areas. Table 2.1 summarizes the characteristics of these devices.

Device	Code	Diameter (μm)	R_s (kΩ)	V_b at -130°C
#1	E2G2	25	500	64.6 V
#2	E3G7	30	1700	65.3 V
#3	E2I1	25	860	70.2 V
#4	E2I9	25	1150	71.6 V

Table 2.1: Characteristics of tested devices

2.2.1 Efficiency and DCR characterization

Before doing the jitter measurements, I first characterized the Photon Detection Efficiency (PDE) and the Dark Count Rate (DCR) of the NFADs under test (table 2.1) at 1550 nm for temperatures between -110°C and -50°C as a function of the excess bias voltage V_{ex} (the difference between the bias and the breakdown voltages). These information are very important to take into account to make sure we do not improve jitter at the expense of the other performances.

The characterization measurements gave efficiencies between 10% and 30% for excess bias between 1 V and 3.5 V. The lowest value of the DCR (1 cps) was obtained at the lowest efficiency for the NFAD from batch $E2G2$, a full characterization of this detector is available in Ref [19]. This lowest value can be explained by the fact that the diode $E2G2$ has the lowest breakdown voltage (see table 2.1) meaning that the absolute electric field in the amplification region is lower with respect to other diodes for the same excess bias voltage. The main contributions to the DCR are the thermal-carrier generation in the absorption region and the field-dependent trap-assisted-tunnelling (TAT). Below -70°C the TAT becomes the dominant effect and, although it is not directly temperature dependent, it is reduced at lower temperatures due to the reduction of the breakdown voltage. So it is clearly beneficial to operate NFADs at low temperatures, if the application calls for low DCR and does not require high count rates, since the required hold-off time increases in order to avoid Afterpulsing effect (see section 2.3 for more details). But how does time jitter evolve at low temperature? And how is it connected to the operating bias voltage?

2.2.2 Timing jitter mechanisms

There are two dominant contributions to the timing jitter in InGaAs/InP SPADs. The first is attributed to the time distribution of the transit time of the photo-generated carriers (holes) from the absorption region (InGaAs) to the multiplication region (InP) [39]. Due to the band gap difference between the two materials there exists a valance-band energy step which has to be overcome by the holes travelling to the multiplication region (see section 2.1 for more details). Such a barrier leads to charge pile-up which can be liberated through thermionic emission, giving rise to a temporal distribution with an exponential tail. This energy barrier is reduced with increasing electric field increasing the emission rate. In addition, the grading layer at the heterojunction of the NFAD helps easier cross of the energy step, meaning that lower field-strengths are needed to cancel this energy barrier. Nevertheless, there exists a given field-strength whereupon the barrier is non-zero and in the following I shall probe the onset of this effect.
Subsequently, upon the arrival of the hole in the multiplication region, a fundamental build-up time is needed for the avalanche amplitude to reach a predetermined threshold level, signalling the detection event. The temporal distribution of this process is Gaussian. **Therefore, the system temporal jitter is expected to be a convolution of an exponential (thermionic emission during the holes transport) and Gaussian (impact ionization) distributions.**

2.2.3 Experiment

The NFADs were placed inside a Stirling cooler (Twinbird SC-UE15R) that enables cooling of the detectors down to -130°C. The light source used to excite the detector is a pulsed laser emitting at 1538 nm with a repetition rate of 76 MHz. The laser signal was attenuated to the single photon level prior arriving at the detectors. The readout circuit is implemented at room temperature and is described in [40], its digital output is fed to a time-correlated single-photon counting (TCSPC) module (SPC-130 from Becker Hickl), which generates a histogram of the delay between the NFAD detection and the synchronization signal generated by a fast photodiode illuminated by the pulsed laser. The instrument response function (IRF) of the measurement setup has a full-width-half-maximum (FWHM) of 7 ps (given by the contribution of 3 ps FWHM from the optical signal and 6.5 ps FWHM from the TCSPC card) which is negligible in comparison to the detector jitter.

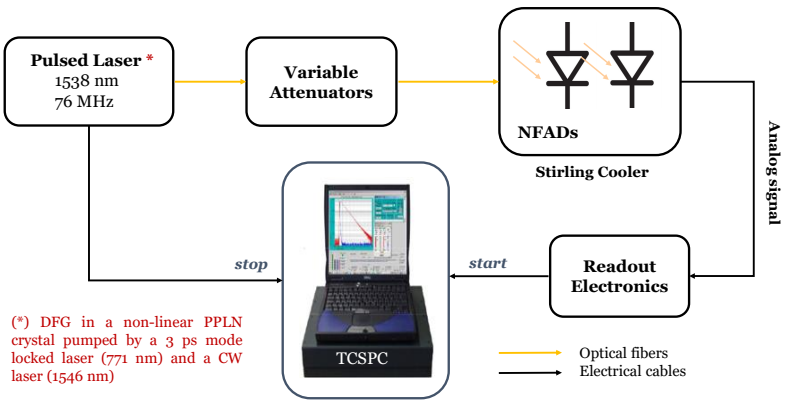


Figure 2.3: Schematic of the jitter measurement setup: the pulsed laser signal is attenuated at the single photon level befor arriving at the NFADs cooled down inside a Stirling Cooler. The discriminated output of the detectors is fed to a TCSPC module synchronized to the laser in order to generate histogram of the delays.

2.2.4 Results and Discussion

First we investigated the IRF response of the NFADs for different temperatures and different excess bias voltages in order to check its agreement with the theoretical model. Figure 2.4(a) shows the temporal jitter histogram for a temperature of -130°C, for diode #1 (E2G2). It is indeed clear that there exists two different contributions, where at small time delays the distribution is Gaussian, whilst at longer delays a clear exponential tail is visible.

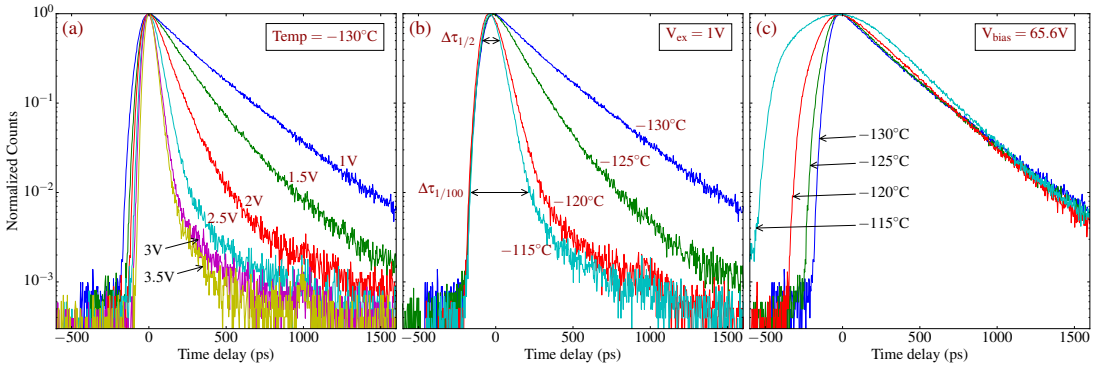
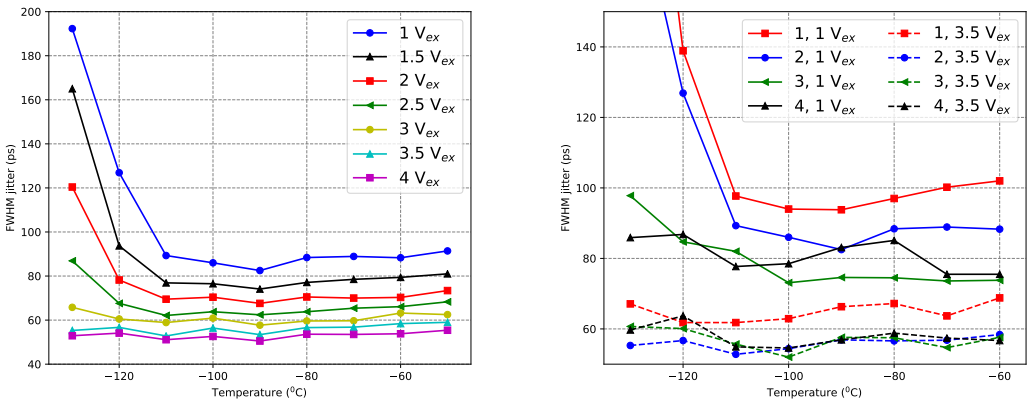


Figure 2.4: Jitter histograms for NFAD #1 in different conditions. (a) Varying excess bias voltage at a constant temperature of -130°C. (b) A constant excess bias voltage of 1 V for different temperatures. (c) A constant bias voltage of 65.5 V for different temperatures.

In order to isolate the two effects, we can exploit their temperature and field dependency and we can clearly see that impact ionization is dependent on the excess bias voltage while thermionic emission is dependent on the absolute bias voltage of the NFAD. To confirm this statement, you can look at Figure 2.4(b), where the excess bias is kept constant at different temperatures, and see that the histograms overlap at short delays, meaning the impact ionization process is unaffected. On the other hand, at longer delays, the exponential time-constant is changing significantly at different temperatures, which is due to varying absolute bias voltage caused by a temperature dependent breakdown voltage in the multiplication region (temperature coefficient is 0.134 V/K). Indeed, if the bias voltage is kept constant for the same temperature range, the contrary is true: the exponential tail is almost unchanged, whilst the Gaussian component changes, as can be seen in Figure 2.4(c).
The temperature dependence of the decay rate (from Fig. 2.4(b)) gives a measure of the energy barrier experienced by the holes [41], which is approximately 0.03 eV at 65.6 V and drops to near zero at around 67 V, leaving the impact ionization as the dominant effect.

To illustrate the overall effects of temperature and excess bias variation, we can plot the FWHM of the jitter histograms. The results obtained with diode #2 are depicted in Figure 2.5a. For a constant temperature, the jitter decreases with increasing excess bias voltage as expected, due to the speed up of the impact ionization process [42]. For decreasing temperature and fixed excess bias, it reduces slightly (about 10%) in the range of -50°C to -100°C. This can be explained by the increase of ionization coefficients with lower temperature in the multiplication region [43], which make the avalanche build-up process yet again faster. At temperatures below about -110°C, for low excess bias voltages, one can see the

increase of the jitter due to the significant hole trapping between the absorption and multiplication regions. This effect is clearly negated through the increase of the excess bias, which reduces the energy barrier, as discussed earlier.
Note that for applications, such as QKD, requiring a very high extinction ratio, it is important to consider the jitter width at a lower level than the half-maximum, especially when the jitter histogram shows non-gaussian behaviour. In our case we measured the full-width at 1/100 of the maximum ($\Delta\tau_{1/100}$) for our devices [17] and it showed similar temperature behaviour as the FWHM. At the highest excess bias, a $\Delta\tau_{1/100}$=200 ps was achieved.



(a) FWHF time jitter ($\Delta\tau_{1/2}$) for device #2 at different V_{ex}

(b) Comparison of the FWHM jitter for four devices at 1 V and 3.5 V_{ex}

Figure 2.5: NFADs time jitter versus temperature for different excess bias voltages.

One strategy to reduce the time jitter contribution coming from the hole-trapping phenomena is to use NFADs with a higher breakdown voltage. Figure 2.5b shows the FWHM jitter for all four NFADs tested for the same range of temperatures and two excess bias voltages, 1 V and 3.5 V. At high excess bias, all the detectors have the same behaviour with the minimum jitter being between 52 ps and 67 ps. However, for low excess bias voltage we see that two of the NFADs do not exhibit the sharp increase in the jitter at low temperatures (devices #3 and #4). This is due to the fact that these diodes have a breakdown voltage of around 5-6 V higher (see Table 2.1) than the diodes which do exhibit the low temperature jitter increase, hence the bias voltage remains sufficiently high in order to keep the energy barrier at the heterojunction below zero, preventing hole pile-up.
Note that the difference in breakdown voltage is mainly due to run-to-run fabrication variations since the design structure is the same for all the devices.

These results suggest that for optimum operation of the NFAD, the bias voltage

should be sufficiently large in order to avoid the hole pile-up jitter effects, however, it should not be too high, in order to minimize TAT contribution to the DCR. In order to keep the thermally generated DCR well below the 1 cps level, the NFAD should be operated at -130°C. Hence, the optimum breakdown voltage would be around 67 V for this temperature, which would allow operation at any excess bias, without any hole pile-up effects.

Lower jitter for Longer-distance QKD

In QKD the signal-to-noise ratio (SNR) is a crucial characteristic as it defines the maximum transmission distance of the system. In order to maximize the signal, it is preferable to operate at the maximum possible clock rate, which is limited by the jitter of the detectors. This means the signal of a QKD system is proportional to $\eta/\Delta\tau_{1/100}$, where η is the detection efficiency and we consider the FW1/100M jitter in order to ensure low error rates. The noise is given by the DCR (r_{dc}) within the detection time window, hence the SNR = $\eta/(r_{dc}{\Delta\tau_{1/100}}^2)$. This shows that the timing jitter is the most important characteristic for a long distance QKD system. Given the jitter demonstrated in this work, QKD operation at 5 GHz and an increase of the maximum distance would be possible.

2.2.5 Conclusion

In this section, we showed that free-running InGaAs/InP NFADs can achieve a temporal jitter as low as 52 ps, which was comparable to the best gated-mode devices [44] and only a factor of 2 larger than the record-holding superconducting devices [45] at the time when this work was published (2016). We have also analysed the low-temperature performance of the NFAD jitter which has enabled the understanding of the jitter contribution due to charge-carrier pile-up between the absorption and multiplication regions, a phenomena which has been rarely studied. Finally, we have shown that in order to avoid degradation of the temporal resolution due to this effect, the operating voltage of these devices should be greater than 67 V at the lowest operation temperature. Given this, excess bias voltage and temperature can be chosen freely according to the applications requirements.

2.3 Afterpulsing

Afterpulsing is another important feature of InGaAs/InP detectors that we should take into account when operating these detectors in the free-running regime.

> Afterpulsing (AP) defines spontaneous dark counts coming from the release of carriers that get trapped inside the defects of the multiplication region during a photo-triggered avalanche event.

This problem is well known for IR detectors because III/V materials technology is less mature than Silicon (for visible detectors) which gives higher defect concentration, resulting in higher afterpulsing. To reduce this effect it is common to operate the detectors in gate-mode with sub-nanosecond gates [29] which quickly reduces the charge flow inside the detector resulting in a subsequent reduction of the amount of carriers getting trapped in the first place. However, it is often more convenient to operate the detector in the free-running regime where a very efficient quenching, like the one obtained with NFADs (see section 2.1), is required to reduce afterpulsing probability. Another commonly-used technique for mitigating afterpulsing of free-running SPADs is the implementation of a hold-off time, also called "dead time", following the detector quenching during which the bias voltage is held below V_{br} to prevent new avalanches from occuring while trapped carriers are getting released. Long hold-off times allow efficient decrease of afterpulsing probability but severely reduce the maximum count rate.

In the previous sections we showed that free-running NFADs can acheive an extremely low DCR (1 cps) [19] with a very low time jitter (52 ps) [17] when operated at low temeratures. So what about the temporal dependence of afterpulsing? The answer to this question was brought by B. Korzh and al. [46] in 2015 as they found out that at low temperatures, the AP dependency is exponential at short delays before following a power-law, whereas at higher temperatures the exponential behaviour is observed at long delays. While the power-law dependence confirms the idea of a dense spectrum of trap levels, they explained the exponential behaviour at short and long delays by the detrapping rate of defects at the edges of the energy spectrum. Finally, they measure the activation energies of these outermost trap levels: $E_{shallow} = 0.05$ eV and $E_{deep} = 0.22$ eV.

In this work, we decided to use the results obtained by B. Korzh and al. to try to experimentally reduce afterpulsing in InGaAs/InP NFADs. The idea of the experiment is based on shining on the NFAD a quantum cascade laser (QCL) whose energy is enough to induce carrier de-trapping of multiple defect levels but not high enough to trigger photo-generated avalanches.

2.3.1 Experiment

The detectors used in this study are InGaAs/InP based NFADs provided by Politecnico Milano, Italy. A full description of these devices is available in [37]. These NFADs have low-value integrated quenching resistors that can reduce most of the avalanche current but are not high enough for a complete avalanche quenching. That is why an external active circuit is used to completely stop the avalanche. The proposed mixed-quenching approach with selectable well-defined hold-off times, is described in ref [37].

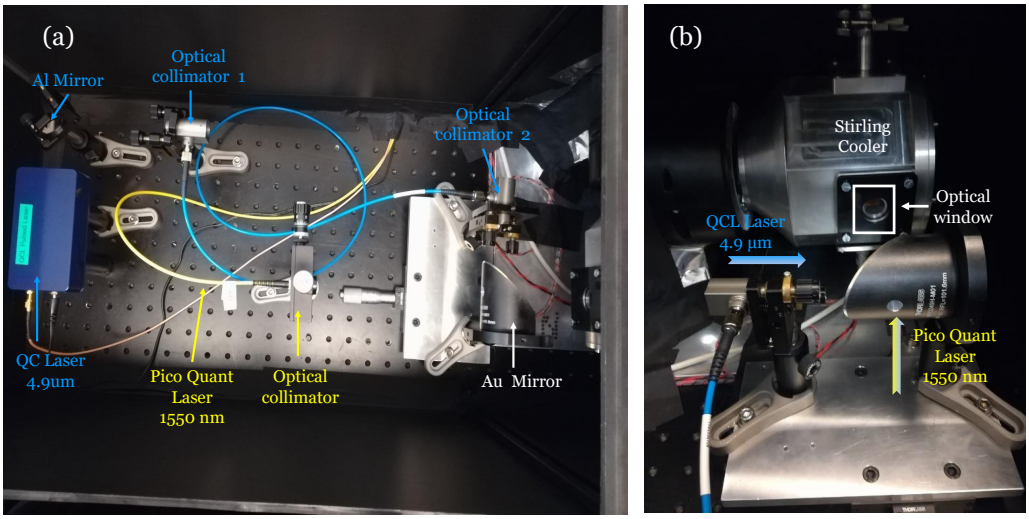


Figure 2.6: Picture of the afterplsing measurement setup: the two optical signals coming from the picoQuant laser and the QCL laser are free-space coupled to the NFAD through the optical window. (a) A top view of the setup inside a black box for optinum light isolation. (b) Side view of the stirling cooler showing how the two laser beams arrive at the detector.

The NFAD was placed inside a Stirling cooler (Twinbird SC-UD08) that allows cooling the detectors down to -130°C. Two optical signals were free-space coupled to the detector: the first one was a pulsed laser (PicoQuant PDL-800) emitting at 1550 nm wavelength and used for optical excitation; this fiber-coupled signal was fed to an optical collimator mounted on an XYZ stage and sent through the central hole of an off-axis parabolic gold mirror at the active area of the detector. The second optical signal was coming from a pulsed 4.9 μm quantum cascade laser (ALPES LASERS) used to empty the carrier traps during the avalanche events. This optical signal is collimated into a 100 μm diameter InF_3 optical fiber and then sent at the detecor after being reflected on the parabolic mirror used for the two beams superposition. Figure 2.6 shows a picture of the experimental setup.

We used an FPGA to send and collect most of the inputs/outputs of the setup including the temperature setting, the NFAD bias voltage, the hold-off time, the detection discrimination threshold, the QCL supply voltage and frequency. The digital output of the NFAD and the PicoQuant synchronization signal are fed to a time to digital convertor (TDC - ID800) and the recorded data is used to plot the histogram of detections over time. The entire setup was placed inside a black box completely opaque to light.

2.3.2 Results and discussion

After characterizing the breakdown voltage dependency on temperature we started by measuring the dark count rate of the detector under test for different excess bias voltages (V_{ex}) and hold-off times. These measurements were done for -50°C, -75°C and -90°C. As mentioned previously, afterpulsing are false counts that add to

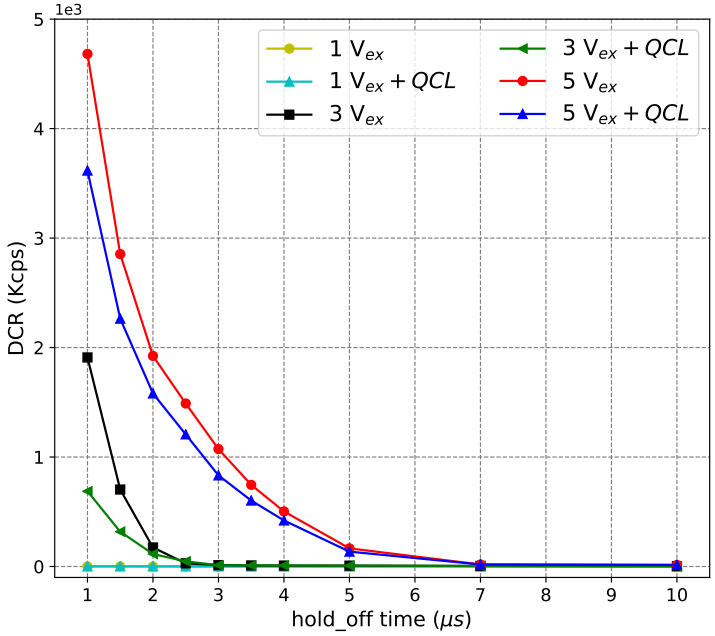
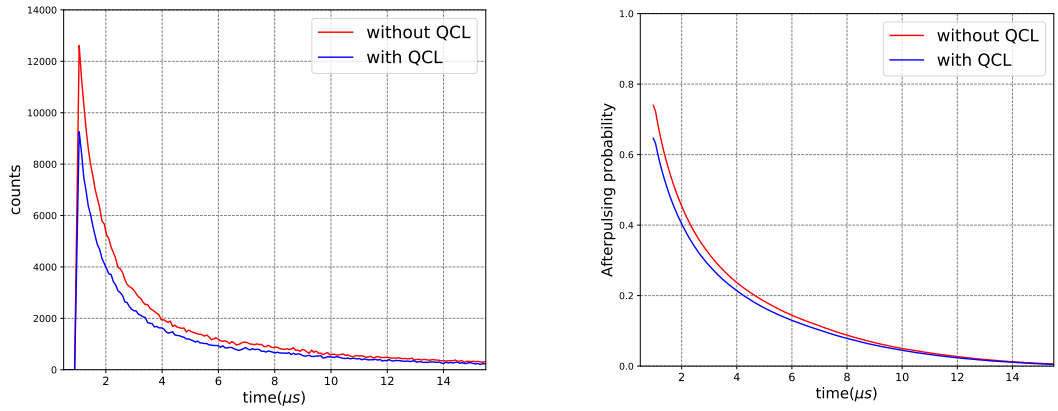


Figure 2.7: Dark count rate variation with hold-off time measured at -75° different excess bias. The measurements are done with and without a pulsed quantum cascade laser (QCL, 4.9 μm, 200 ns pulse width) shining on the detector active area.

the thermally induced ones and increase the overall dark counts. The afterpulsing contribution is indicated by the notable increase of the dark count rate for shorter delay times. If we do the same measurement while shining the QCL laser on the

NFAD we should be able to see a drop of the overall dark counts. The results of these measurements are shown in Figure 2.7. With the QCL lasing at 10 kHz frequency and 200 ns pulse width we could see a slight decrease of the dark counts depending on the operation conditions. For 1 V excess bias, corresponding to $\sim$ 5% efficiency, we saw almost no effect of the QCL on afterpulsing which can be explained by the fact that afterpulsing probability is negligible at low efficiencies. At 3 V and 5 V excess bias (12% and 20% respective efficiencies), afterpulsing contribution to dark counts decreased for short delays (up to 5 μs) which indicates that the QCL is actually helping with the detrapping process but its effect is not enough to allow the use of InGaAs/InP NFADs at lower delays. To make sure



(a) (b)

Figure 2.8: Comparaison of afterpulsing measurements for an InGaAs/InP NFAD at -75°C for different operation conditions, with and without a pulsed quantum cascade laser (4.9 μm, 200 ns pulse width) shining on the detector active area. (a) Afterpulsing counts histogram measured at 1 μs hold-off time and 20% efficiency. (b)Afterpulsing probability measured at 1 μs hold-off time and 20% efficiency.

that the QCL is not locally heating the detector and consequently, increasing the thermally-induced dark counts, we characterized the breakdown voltage variation with and without the QCL shining on the active area. We measured, for 200 ns pulse width, a variation of 0.051 V corresponding to only 0.34 °C temperature rise We repeated the measurements for different QCL pulse widths and we saw that the effect increased with the pulse width but it was never powerful enough to allow operation at short delays. To better quantify the QCL effect on afterpulsing probability we compared the detection histograms after the arrival of the detected PicoQuant photons and we could clearly see the afterpulsing time distribution. Figure 2.8 shows the AP histogram and the AP probability with and without the QCL at -75°C for $V_{ex} = 5\ V$ and 1 μs hold-off time. From the counts histogram

we computed a maximum afterpulsing probability of 74% without QCL vs 64.6% with QCL.

Unfortunately I was not able to pursue this work because of technical problems and I had to switch to another project.

2.3.3 Conclusion and Outlook

Using a quantum cascade laser emitting at 4.9 μm and sent at the NFAD after avalanche events, we were able to show a decrease of afterpulsing at short delays by triggering the de-trapping of trapped carriers in the multiple defect levels of the InP layer. Unfortunately, the measured effect was not powerful enough to allow operating the InGaAs/InP based NFADs at lower hold-off times.

One of the possibilities to obtain better results could be the use of other lasers emitting at 3 μm and 2 μm with the optimization of the timing delay between the PicoQuant Laser and the QCL laser.
This project stays on the to-do list of the Lab and the work will be resumed as soon as possible.

Chapter 3

Superconducting Nanowire Single-Photon Detectors

Superconducting nanowire single-photon detectors (SNSPD) have emerged in the past decade and have experienced tremendous performance improvements since their first implementation in 2001 by Gregory Gol'tsman and colleagues [47]. These nano-devices have stood out as highly-promising single photon detectors for a wide range of wavelengths (from X-ray to mid infra-red) thanks to their high detection efficiency [10, 11], low dark count rate [48] excellent time resolution [18] and fast recovery [49]. SNSPDs have became the first choice of many Time-Correlated Single-Photon Counting (TCSPC) applications showing greater performance than their first competitor, Semiconductor Single Photon Avalanche Diodes (SPADs). Besides their wide-range sensitivity, SNSPDs have the best signal-to-noise ratio, are the fastest and do not suffer from afterpulsing phenomena. The only drawback of these detectors is their operation at cryogenic temperatures (around 4.2 K, the boiling point of liquid Helium) which can be sometimes difficult to implement. Fortunately, closed-cycle cooling technologies are concurrently growing and improving enabling a better ease-of-use [50, 51].
Superconducting nanowire single photon detectors have driven the development and the upgrade of many applications such as quantum communication [52], light detection and ranging [53, 54], single photon spectroscopy [55] and integrated circuits testing [56].

In this chapter, we will describe the operating principle of SNSPD and give an overview of its performance metrics. We will then focus on large-area nanowires for multimode-fiber coupling. We will present the main fabrication steps before reporting the results obtained with single meander and parallel designs, including a saturated detection efficiency and a max count rate as high as 40 MHz for par-

allel nanowires.
This work was carried out within a collaborative project between Id Quantique and the University of Geneva that includes the technological transfer of standard SNSPD technology and the development of large active-area nanowires as a potential new industrial product.

3.1 Device physics

3.1.1 Operation principle

Superconducting nanowires single photon detectors are built by patterning a thin film of a superconducting material into a meander of narrow nanowires (80 - 160 nm). Highly disordered superconductors with short electrothermal time con-

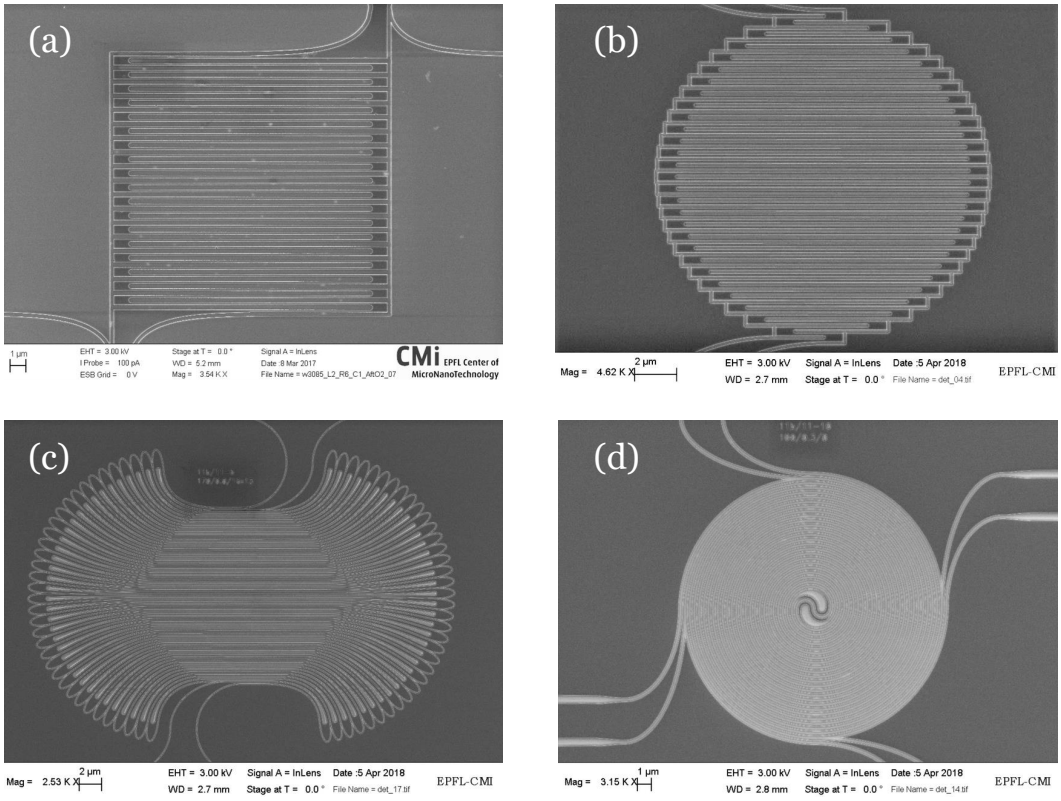


Figure 3.1: SEM images of different meander structures for single-mode fiber coupling designed and manufactured by the University of Geneva. (a) Square meander of 16 μm x 16 μm active area.(b) Circular meander of 15 μm diameter for optimal coupling with the circular core of the fiber. (c) "Loop" meander designed with higher fill factor in the center for better absorption efficiency. (d) "Spirale" meander designed for polarization independence.

stants are usually chosen such as WSi, NbN and MoSi. The active area of the meander is designed in such a way to maximize the coupling efficiency of incoming light (through a single mode or a multimode fiber). Figure 3.1 shows different meander shapes designed in the University of Geneva for different purposes. The

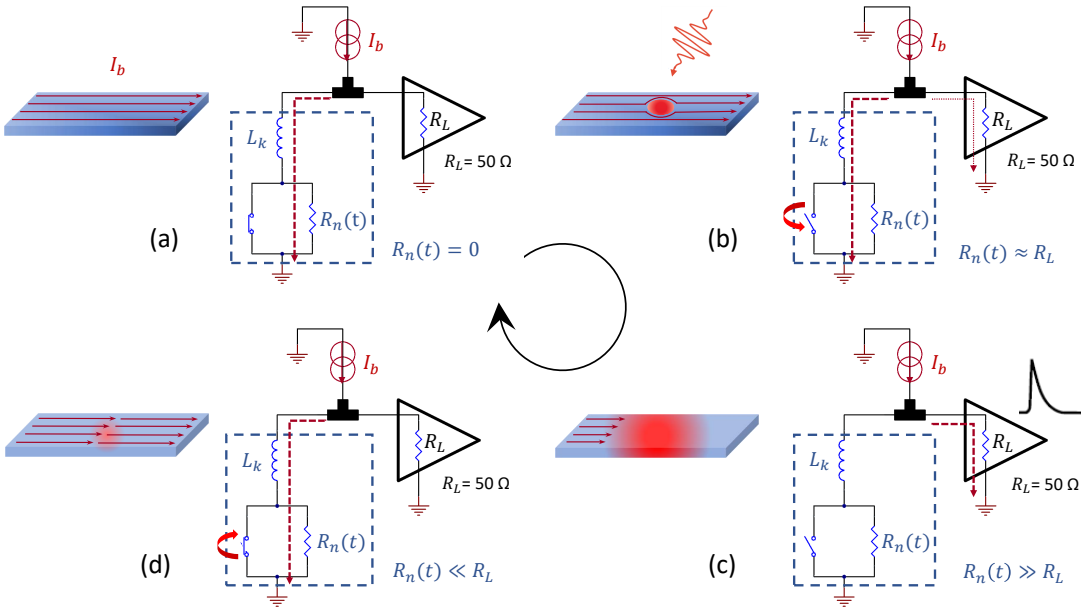


Figure 3.2: Operation principle of the superconducting nanowire single-photon detector (SNSPD). (a) The SNSPD is in the superconductive state. (b) When a photon is absorbed, a resistive region called hotspot is created. (c) The hotspot propagates across the nanowire width and creates a resistive barrier that diverts the current into the load impedance. (d) The bias current shunt outside the nanowire allows the resistive region to subside and the superconductive state is restored.

nanowire is cooled down below the critical temperature[1] of the superconducting material and biased just below its critical current[2]. In the superconductive state, the SNSPD can be electrically seen as an inductor L_k representing the kinetic inductance of the superconducting nanowire and defined as $L_k = l_k \int \frac{ds}{A(s)}$, where l_k is the kinetic inductivity of the superconducting material and $A(s)$ is the cross-sectional area of the nanowire (Figure 3.2(a)). When a photon is absorbed in the device, its energy breaks hundreds of Cooper pairs resulting in a local resistive region called a "hotspot". This hotspot will force the current to flow around it increasing the peripheral current density beyond the critical current density as de-

[1]The critical temperature of superconducting materials is the temperature at which its electrical resistivity drops to zero.

[2]The critical current is defined as the maximum current that can be passed in the nanowire without destroying its superconductivity

scribed in the "normal-domain" growth model [57] (Figure 3.2(b)). Together with Joule heating, this effect will create a resistive barrier across the entire width of the nanowire and the resistance value can reach several kilo Ohms within picoseconds (Figure 3.2(c)). As illustrated in Figure 3.2, the SNSPD is modeled as an inductor L_k serially connected to a time dependant resistance R_n(t) in parallel with a switch . The switch opens when the photon is absorbed and the superconductive state is disturbed and this will divert the current into the Load impedance R_L (which is mostly a 50 Ω readout amplifier) generating a measurable output voltage pulse (Figure 3.2(c)). Once the current has been shunted, the nanowire cools down and returns to thermal equilibrium, and the SNSPD is ready for the next detection (Figure 3.2(d)).

3.1.2 Performance metrics

The detection process described in the previous section can be divided into three steps: photon absorption, resistive region creation and output pulse generation. Each one of these steps gives insight into some of the SNSPD performance metrics.

System Detection Efficiency (SDE) is a crucial characteristic for most of the applications and is highly impacted by the photon absorption probability. SDE is defined as:

$$\eta_{sde} = \eta_{coupling} * \eta_{absorption} * \eta_{registration} \tag{3.1}$$

The coupling efficiency $\eta_{coupling}$ can be improved with a perfect optical alignment between the fiber core and the detector active area, known as the self-alignment technique [58]. The absorption efficiency $\eta_{absorption}$ can be maximized by stacking the nanowire inside an optical cavity optimized for the intended wavelength [59]. As for the registration efficiency $\eta_{registration}$, defined as the probability of registering an electrical pulse correlated to the photon absorption, it is mainly dependant on the nanowire characteristics (material, width and thickness) and the fabrication process quality. Near-unity system detection ($\geq 95\%$) efficiency has been recently reported by the NIST [11] and the same result was obtained in IDQ with Molybdenum Silicide (MoSi) single meander at 1550 nm wavelength.
A full description of the setup we have been using for SDE measurement is available in [60]. The same setup has been used to measure the **Dark Count Rate (DCR)**, known as the the average rate of "false counts" generated by SNSPDs when light is blocked. The main origin of dark counts comes from blackbody radiation at room temperature that propagates through the optical fiber to the detectors inside the cryostat. This dominant contribution can be reduced and almost suppressed with cryogenic pass-band filters [48] or simply using the cold fiber filter technique which achieved good results with our MoSi detectors, enough to meet the requirements of long distance QKD experiment [52].

Timing jitter is another key metric of SNSPD and is defined as the time uncertainty between the photon absorption and the generation of the output pulse. This feature is very important for time-resolved applications such us LIDAR [53, 54] and quantum key distribution (QKD) [52]. The time resolution of SNSPDs includes the measurement setup contribution given mainly by the jitter of the readout electronics and the laser, the geometric jitter related to the propagation path of the signal inside the nanowire and the superconducting material intrinsic jitter which represents the time uncertainty of the duration taken by the hotspot to propagate across the nanowire width. An extensive study of jitter in Molybdenum Silicide devices fabricated in the University of Geneva is provided by Caloz *et al* in [61]. A record value of 2.7 ps timing jitter at 400 nm wavelength has been recently reported by Jet Propulsion Laboratory (JPL) using a short NbN nanowire [18].

The **Recovery time** τ of an SNSPD sets the limit on the maximum achievable count rate and needs to be minimized for high-speed applications. The recovery time defines the minimum time required by the SNSPD to recover its superconducting state after a detection event and is limited by the time charge duration of the RL circuit composed of L_k and R_L (see the electrical model of SNSPD shown in Figure 3.2). This reset time is equal to L_k $/R_L$. To make faster detectors we can either reduce the kinetic inductance of the device by making shorter nanowires or increasing the load resistance, e.g. using high impedance readout. However, if τ is excessively reduced, the current will return too rapidly into the nanowire and the hotspot will not have enough time to cool down which causes the nanowire to "latch" into a permanent resistive state preventing subsequent detection. The recovery time can also be improved using parallel wires architecture which dramatically reduces the overall kinetic inductance, this approach will be discussed in Section 3.2.
To characterize the recovery time of the efficiency of SNSPDs we implemented a hybrid auto-correlation method that allowed us to have a direct insight into the time dynamics of the current inside the detector after one or multiple detections. Note that this technique can be applied to any type of single-photon detector, and could be considered as a universal benchmarking method to measure and compare the recovery time of single-photon detectors. A detailed description of this method and its advantages is given in the paper "Direct measurement of the recovery time of superconducting nanowire single-photon detectors"(See section: Preprint articles).

3.1.3 High-speed detectors: Parallel-SNSPDs

The standard SNSPD design is a single nanowire patterned into a meander shape. We achieved outstanding performance using this design (Section 3.1), notably a near-unity SDE (95%) and sub-20 ps timing jitter at 1550 nm wavelength [62]. However, the maximum count rate of a single meander is usually limited to several tens of MHz, and may not meet the requirements of some high-speed applications. As explained above, the maximum count rate of a detector is limited by its recovery time that is mainly set by its intrinsic kinetic inductance L_k. That is why we need to explore different designs with lower L_k if we want to boost the count rate without reducing the SDE.

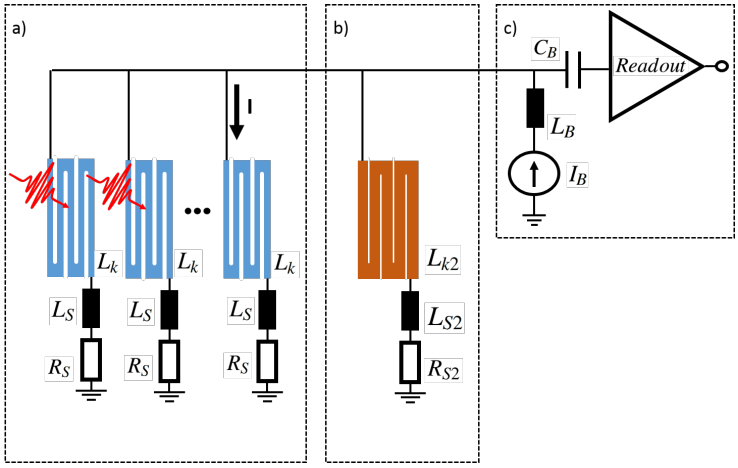


Figure 3.3: Schematic of the electrical model of Parallel-SNSPD. a) Schematic of a basic parallel SNSPD design[63], which consists of a limited number of photosensitive nanowires with kinetic inductance L_k. An additional serial inductor L_s can be added to choose the overall inductance of each section which has an impact on the output signal amplitude. Serial resistors R_s ensure that the biasing current is evenly split among the nanowires. b) Additional nanowires with low inductance L_{k2} are added in order to decrease the electronic crosstalk between the nanowires during detection events. The values of L_{s2} and R_{s2} can be chosen to optimize the trade-off between the crosstalk and the output signal amplitude c) A bias tee is used to bias the detector and amplify the output signal with the same coaxial line.

Parallel-SNSPDs design, depicted in Figure 3.3.a, brings an elegant solution to the recovery time limitation. Several nanowires are connected in parallel and together they cover the same active area of a single meander, which makes the individual inductance of one nanowire much smaller. Moreover, only part of the detector undergoes dead-time after a detection event which leaves the remaining nanowires available to detect another photon at their full detection efficiency, and this make the overall recovery time even smaller. This implementation has already been

explored [64, 65] but the maximum count rate has never been properly investigated due to the cascade-switch effect [66] that appears at high illumination and can lead to the detector latching, i.e. all nanowires end up in a steady resistive state where the whole detector is effectively disabled.

Perrenoud *et al*, from Geneva University demonstrated recently a new parallel design for SNSPDs that overcomes this limitation [20]. As shown in Figure 3.3, cascade-switch effect between the detector sections is prevented by limiting the number of photosensitive nanowires and adding wider nanowires positioned outside the optical fiber spot which makes them unexposed to light. After a detection, part of the cross-current is redirected into these insensitive nanowires, which effectively reduces the total cross-current seen by the active nanowires. Additional serial inductances L_s are added in order to tune the output signal amplitude while R_s guarantees an even split of the bias current between the nanowires. L_{s2} and R_{s2} are serially connected to the non-photosensitive nanowires and are used to adjust the current inside the active nanowires. Note that thermal crosstalk is avoided with proper spacing of the different parallel nanowires.
In this design the nanowires are biased well below their I_c. When one or multiple photons are detected by one or multiple nanowires, their resistances will increase as explained previously and their current will be distributed among the other parallel sections. The redirected current will add up to the initial bias current and the total current inside each nanowire should not exceed the value of I_c in order to prevent the cascade-switch phenomena and keep the detector constantly active. The reduction of the electrical crosstalk between the nanowires is crucial to increase the maximum count rate. Using this design for detectors covering 15 μm x 15 μm active area, we demonstrated detection rates over 200 MHz without any latching, a fibre-coupled system detection efficiency (SDE) as high as 77%, and more than 50% average SDE per photon at 50 MHz detection. More details about this work are provided in the preprint "High detection rate and high efficiency with parallel-SNSPDs" available in The section "Preprint articles". The design was developed by Perrenoud, M. and I helped with the fabrication and characterization of the devices. A patent application about this parallel-design has been co-filed by Id Quantique and the University of Geneva.

Parallel-SNSPDs design gave promising results when coupled to single mode fibers and seem to mitigate the main issues of large area SNSPDs. In this thesis, I decided to use this parallel design to make large active-area SNSPDs. This idea is further explored in Section 3.2.

3.2 Large active-area SNSPDs

Superconducting nanowire single-photon detectors with small active areas have achieved outstanding performance when coupled to single mode fibres (SMF), and have been broadly used in numerous fields. However, many applications require the detection of irregularly emitted photons which makes coupling to SM fibers very challenging. For such applications, multimode fibers (MMF) offer better coupling efficiency thanks to their larger core and wider numerical aperture (NA). For instance, coupling photons emitted by III-V quantum dots to SM fibers has been a big challenge due to the nature of the emission process of QDs and also to the high refractive indices of III-V semiconductor materials. MM fibers on the other hand offer necessary features to considerably relax the task of light coupling. Similarly, in SNSPD-based Lidar systems, MM fibers are widely used as they provide easier coupling to telescopes thanks to their large core compared to SM fibers. Yet, the SNSPDs used in most of these applications are smaller than the core size of the MM fiber and have an active area of around 15 μm diameter. This size mismatch leads to a dramatic loss of coupling efficiency when the self-alignmenent technique is used, and requires the implementation of complex optical alignment otherwise. **Thus, large-area SNSPDs coupled to multimode fibers will enable a wide range of applications while guaranteeing the same high performance obtained with single mode detectors.**
Large-area SNSPDs are also required for applications using free-space light coupling, such as satellite laser ranging [67] and ground-satellite Quantum Key Distribution [68].

3.2.1 State-of-the-Art

Many attempts have been made to increase the sensitive-area of SNSPDs using one of the two approaches: increasing the length of the nanowire to cover a bigger area or making arrays of multi-pixel SNSPDs.
The main two problems with the first approach are the defect density in nanofabrication process, which makes it difficult to produce a defect-free nanowire structure over a large area, and the large kinetic inductance which is proportional to the nanowire length. Nevertheless, quite good results obtained with large area single meander have been reported, mainly for visible and near-infrared wavelengths. For instance, the Shanghai Institute of Microsystems and Information Technology (SIMIT) reported NbN SNSPDs with a sensitive area diameter of 50 μm fabricated on a photonic crystal for 850 nm detection [69]. The MMF coupled SNSPDs exhibited a SDE of 82% which is the highest DE reported for a large-area SNSPD at 850 nm. The same group demonstrated two years later

(2017) a 100 μm active-area SNSPD which achieved 65% SDE at 532 nm when coupled to 105 μm multi-mode optical fiber [70]. On the other hand, Delft University of Technology together with Single Quantum demonstrated a 20 μm NbTiN SNSPD with 80% SDE at visible wavelengths and 50% at the telecom range [71]. Their devices, coupled to 25 μm core MM graded-index fiber, exhibited very low time jitter using cryogenic amplifiers. The demonstrated results are unarguably very good but a 20 μm active-area will probably increase the coupling efficiency but will not resolve the problem when coupled to standard MM fibers with 50 μm, 62.5 μm and 105 μm core sizes.

The second approach uses arrays of standard-size SNSPDs to extend the overall sensitive area. Yet, this implementation requires a complex-readout circuit as well as many coaxial cables which drastically increases the cooling power needed to operate the system. The biggest SNSPD array reported to date has 64 pixels and a diameter of 320 μm. It was fabricated by the Jet Propulsion Laboratory (JPL) for deep-space optical communication and exhibited a free-space SDE of 40% at 1550 nm [72]. More recently, SIMIT developed a NbN SNSPD array with a circular active area of 300 μm divided into nine pixels. When coupled to a 200 μm multi-mode fiber, the superconducting array achieved 42% SDE at 1064 nm with a maximum count rate exceeding 40 MHz [73]. Shortly after, they increased the number of the pixels from 9 to 16 and they attained an SDE of 72% at a wavelength of 1550 nm and a low-photon-flux limit, with a DCR of 100 Hz and a single-pixel jitter of 59 ps. The new 16-Pixel Interleaved SNSPDs array achieved a MCR exceeding 1.5 GHz with an SDE of 12% [74].

3.2.2 New approach for high-speed large active-area SNSPD

The idea I propose and discuss in this work is to make large sensitive-area parallel-SNSPDs. With this approach, we would decrease the kinetic inductance of the device as explained in Section 3.1. For instance, in the case of 6 parallel nanowires covering a total area of 50 μm^2, the single nanowire's kinetic inductance is 6 times lower than the kinetic inductance of a single meander covering the same area. In fact, the total inductance of the 6 parallel nanowires is even lower. Furthermore, if one or more of the sections get affected by the fabrication defects, the detector does not loose its entire efficiency because the unaffected sections would continue to detect photons and generate output signals. With this original solution we would mitigate the problems of speed and fabrication defects to obtain high efficiency large sensitive-area SNSPDs with a much shorter recovery times than what single meanders can provide
Note that the parallel-SNSPDs detector uses the same readout circuit as for single meanders and does not require multi cryogenic amplifiers and coaxial cables with

a huge cooling power as is the case for the multi-pixel scheme.

3.2.3 Nano-Fabrication steps

At the opposite of InGaAs SPADs (See Chapter 2) which were acquired from external manufacturers to be used in our experiments, SNSPDs are designed in the University of Geneva and IDQ and manufactured in the clean room of CMi-EPFL (ISO 7-6). That is why nano-fabrication has been a substantial component during a big part of this thesis.

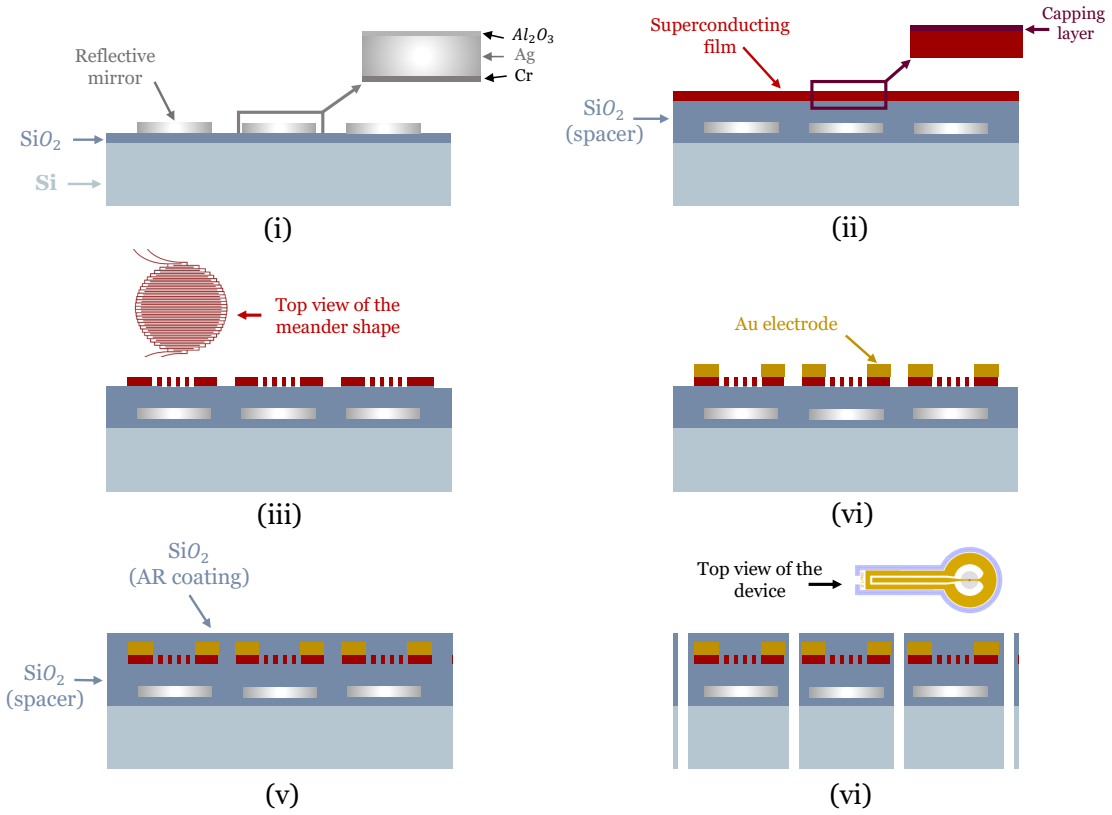


Figure 3.4: Schematic of the nano-fabrication steps. (i) A Cr-Ag-Al_2O_3 mirror is evaporated. (ii) The SiO_2 Spacer layer is deposited by RF sputtering followed by the superconducting film. (iii) The superconducting film is patterned into meanders using E-beam photolithography and Ion Beam Etching. (iv) Au electrodes are evaporated. (v) An AR coating SiO_2 layer is deposited to encapsulate the optical cavity. (vi) The lollipop detector shape is etched through the entire wafer thickness.

The SNSPDs manufactured by our team are designed for telecom wavelengths and we have been mainly using Molybdenum Silicide (MoSi) as superconducting

material [60, 20]. Nevertheless, we have decided recently to investigate Niobium Titanium Nitride (NbTiN), driven by its lower inductivity, in order to improve the maximum count rate. An extensive comparison of the two materials is given by Caloz, M. in his PhD thesis [62]. The main fabrication steps are the same for both materials with some minor adjustments. The SNSPDs are fabricated out of a 6-7 nm-thick film of superconducting material (MoSi or NbTiN) deposited by co-sputtering with a DC and RF bias in the case of MoSi and reactive sputtering in the case of NbTiN. The crystalline and electrical properties of both films are well studied as explained in [62]. The superconducting film is embedded inside an optical cavity build on the top of a reflective mirror, designed to maximize the absorption efficiency.

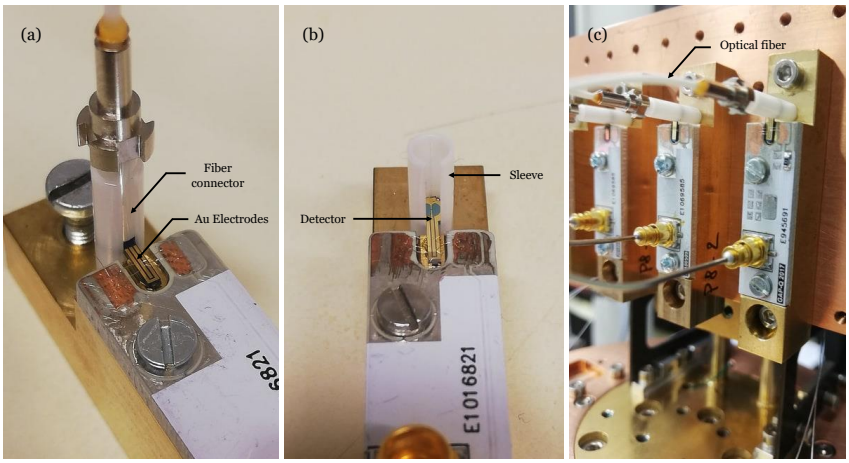


Figure 3.5: Pictures of packaged detectors. (a) The optical fiber is perfectly aligned to the detector using the zircon sleeve (self-alignment) and the gold electrodes are wire-bonded to the PCB. (b) Same picture without the optical fiber to show the "lollipop" detector. (c) Packaged detectors mounted on the 0.8 k plate inside the cryostat.

The nano-fabrication steps are shown in Figure 3.4. First, the metallic mirror is evaporated on a thin film of Silicon dioxide (SiO_2) (i). This mirror should have high reflectivity at telecom wavelengths and good adhesion to SiO_2, that is why we use Silver (Ag) stacked between two adhesive layers (Cr and Al_2O_3). Then, a SiO_2 layer with a $\sim \lambda/4$ thickness is deposited by RF sputtering, followed by the superconducting film and its capping layer (ii). The film is patterned into a meander structure by a combination of e-beam lithography and reactive ion etching (iii). Figure 3.9 shows a SEM image of a 50 μm-diameter single meander SNSPD and Figure 3.11 shows a parallel SNSPDs with 6 photosensitive nanowires covering an area of 50 μm x 50 μm. Afterwards, gold electrical pads are evaporated (iv)

and the optical cavity is encapsulated with $\sim$ 20 nm anti-reflection (AR) SiO_2 layer (v). Note that the thicknesses of the SiO_2 layers should be optimized to ensure constructive interference inside the cavity, which maximises the absorption in the superconducting film [59]. Finally, the detector "lollipop" structure is etched through the entire wafer thickness (vi) and coupled to the fiber using the self-alignment technique [58]. Figure 3.5 shows the SNSPDs packaged and ready to be used.

Nano-fabrication challenges

As explained in the previous section, fabricating large active-area SNSPDs has been a quite difficult task because of the high defect density. Even though we paid a lot of attention during the fabrication of our devices, we still lost some detectors (single meanders and parallel-SNSPDs). Figures 3.6.a-b-c show some of the encountered defects.

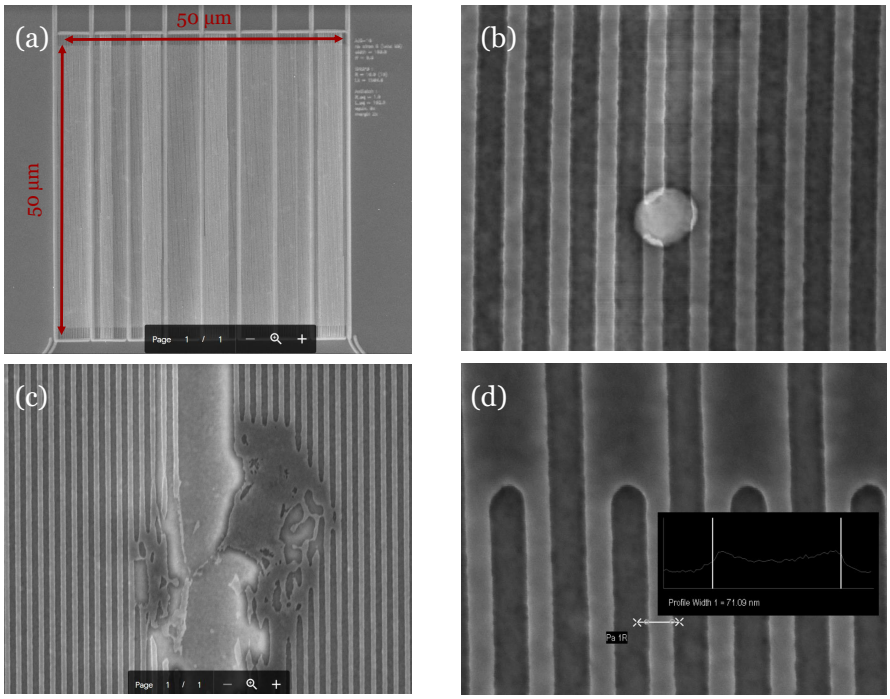


Figure 3.6: SEM images of some fabrication problems registered for large active-area SNSPDs. (a)The difference in the color shade indicates that some parts of the nanowire are not connected. (b) Zoom on image (a) to show the nanowire discontinuity because of a sub-100 nm- diameter defect. (c) Etching anomalies in some regions of the detector. (d) Over-etching problem because of proximity effect in E-beam lithography: we measured 70 nm nanowire width for a design of 100 nm.

The Scanning Electron Microscopy (SEM) performed on our devices also showed a global over-etching effect: we measured a 70 nm nanowire width for a design of 100 nm like depicted in Figure 3.6.d. This problem came from the proximity effect in E-beam lithography, where patterns receive more than the intended dose due to the design high density. Fortunately, we performed the Proximity Effect Correction (PEC) when preparing the job for the E-beam and we were able to retrieve widths very close to the design.

3.2.4 Characterization results

After fabrication and packaging, the SNSPDs are placed inside a sorption-equipped closed-cycle cryocooler and cooled down to 0.8 K. The detectors are biased close to their critical current through a Bias-Tee and their output signal is amplified using a cryogenic amplifier at 40 K and a Mini-circuits amplifier placed at room-temperature.

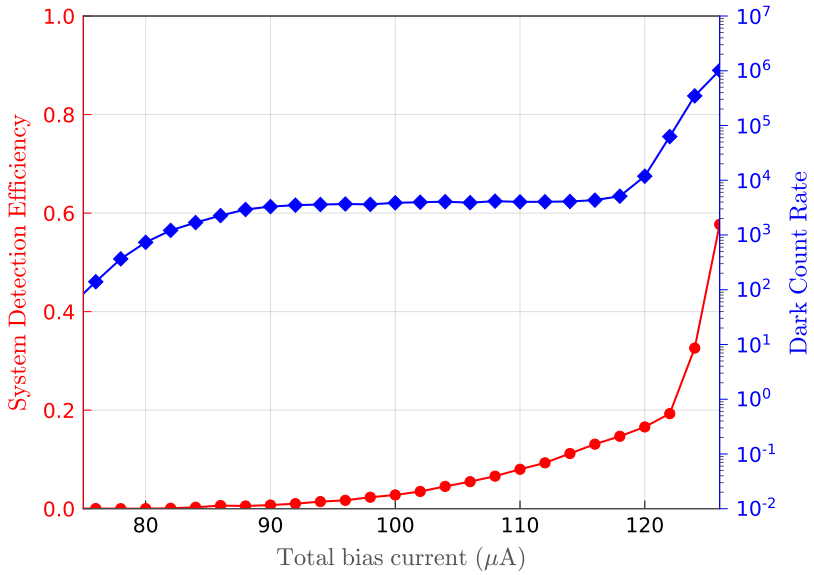


Figure 3.7: System Detection Efficiency (red curve) and Dark Count rate (blue curve) as a function of the bias current I_b measured at 0.8 K and 1550 nm wavelength for a MoSi parallel-SNSPD having 6 photosensitive nanowires (100 nm width and 0.6 FF) covering $50\mu m^2$ active-area surrounded by 10 non active large nanowires.

Our first attempt to make large sensitive-area SNSPDs did not give the best results. We fabricated a full wafer of MoSi parallel-SNSPDs with different numbers of parallel nanowires and serial resistances R_s (see section 3.1.3). All the nanowires

had a width of 100 nm and a fill factor of 0.5. After SEM and AFM[3] characterization, we noticed a general over-etch of 30/35 nm but most of the nanowires still looked in a good shape.
The SDE of the tested devices did not exhibit any plateau as expected with MoSi devices [60] and the maximum achieved SDE was equal to 17% (for 8 tested parallel-SNSPDs). We noticed however a clear plateau showed by the DCR curve indicating a SDE saturation at shorter wavelengths (blackbody radiation propagating through the fiber end placed at room temperature). Figure 3.7 shows the SDE and DCR curves of a parallel-SNSPDs having 6 photosensitive nanowires (100 nm width and 0.6 FF) covering $50\mu m^2$ active-area surrounded by 10 non active large nanowires.

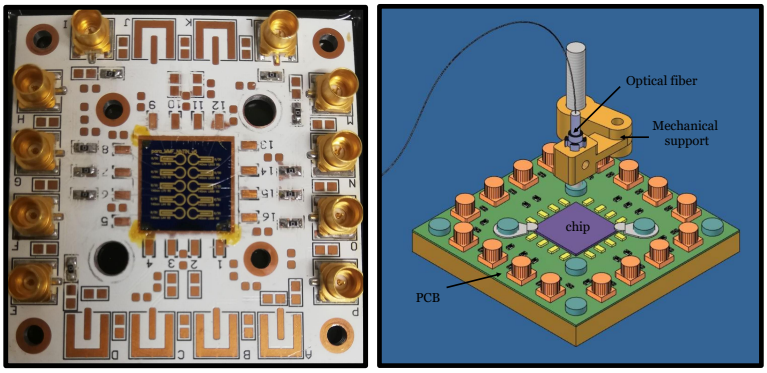


Figure 3.8: (a) A 12 x 12 mm chip comprising 10 SNSPDs is wirebonded on an Insulated Metal Substrate (IMS) PCB. (b) 3-D modeling of the flood-illuminated chip as it is implemented inside the cryostat at 0.8 K. A mechanical support is used to hold the MM fiber at a predefined distance from the chip.

We were not able to determine accurately the origin of this unsatisfactory performance due to the fact that we were using a new design, with new parameters and we did not have any prior experience with large-area SNSPDs. The problem could have come from the design parameters, notably the number of sensitive nanowires with respect to the total parallel nanowires which could have induced premature switch-cascade effect or other unknown effects. On the other hand, the over-etching problem that made the nanowires very thin, could have affected their behaviour. In order to investigate the issue in a fast iterative way, we decided to switch to chip fabrication instead of full wafer fabrication. A 12 mm x 12 mm chip can be fabricated in less than three days against a month for a full wafer. However, it can only accommodate 10 devices. We also decided to use NbTiN as the

[3]Atomic Force Microscopy.

superconducting film because we have been able to deposit this material ourselves in EPFL while the MoSi deposition is done in Basel University and requires longer time. Moreover, it has been shown that NbTiN typically features a low inductivity and a fast recovery time that lead to a fast detector with low jitter [75]. The chip does not allow perfect alignment between each detector and its fiber, we have used instead flood-illumination where one fiber is placed at a the focal distance from the chip and used to illuminate all the detectors (See Figure 3.8). Note that we decided to remove the reflective mirror and the optical cavity when fabricating the chips because they do not play a big role if we are not aiming to optimize the system detection efficiency but their fabrication does consume time.

Many iterations were done using chips with 7-nm thick NbTiN layer deposited by reactive sputtering on a NbTi target in a N_2 atmosphere. These iterations allowed us to make adjustments on the design in term of nanowire width, fill-factor, total number of sensitive nanowires placed in parallel, serial resistance and inductance in the case of parallel-design... We also investigated both large area single meander and parallel-SNSPDs designs.
Note that the first step of the characterization process consisted in verifying the single photon response of the device through the linear relation between the count rate and the relative light intensity.

Large-area single meander SNSPD

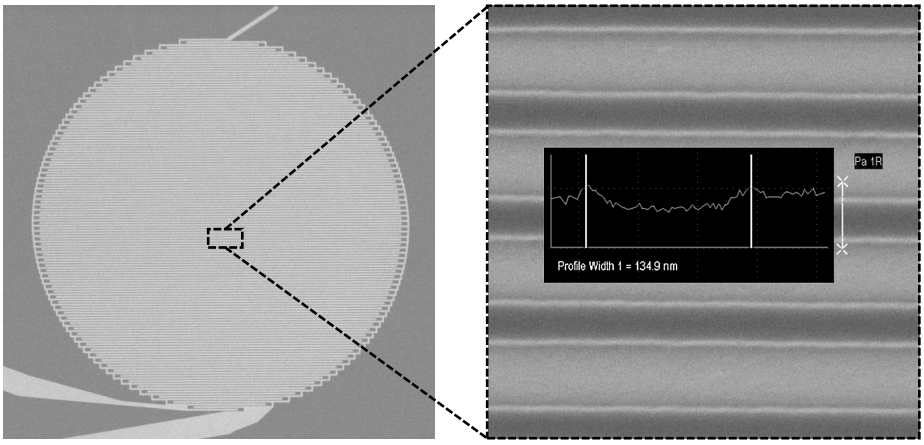


Figure 3.9: (a)SEM image of a NbTiN nanowire patterned into a meander of 50 μm diameter. b) Zoom on the nanowire which has a 140 nm width and 0.4 fill factor. Note that due to the etching rate fluctuation we measured a width of 140 nm ± 10 nm.

Using the chip approach, we were able to demonstrate a 50 μm-diameter single

meander with saturated photon counting curve. The designed nanowire has 140 nm width and 0.4 fill factor as shown in Figure 3.9. We noticed low values of critical current (11 - 14 μA) for these large-area meanders with respect to the standard ones (16 μm-diameter meanders have 15 - 20 μA critical current). This critical current reduction comes from the fact that to cover a larger area patterned into a meander, the nanowire undergoes higher number of "bendings" where the current density increases and the cumulative effect induces a decrease in the critical current [76].

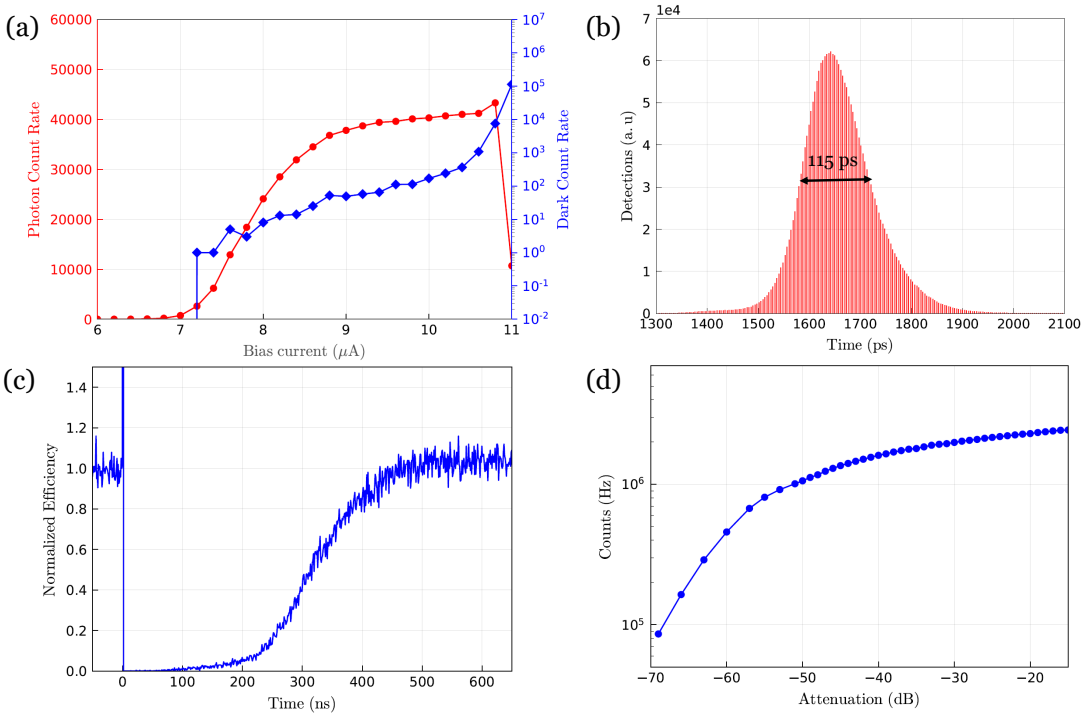


Figure 3.10: Characterization results of 50μm-diameter SNSPD at 0.8 K and 1550 nm wavelength.(a) Photon count eate (red curve) and dark count rate (blue curve) as a function of the bias current I_b. (b) Timing jitter measured at 10 μA. (c) Recovery time measured at 10 μA with the hybrid-autocorrelation method described in [77]. (d) Count rate measurement with CW laser for several attenuations.

Besides the photon count rate (PCR) characterization, we also measured the time jitter and the recovery time of our large sensitive-area meanders. The obtained results are depicted in Figure 3.10. The most important criteria to have promising detectors is a saturated PCR curve. This saturation presented by a "plateau" over a range of currents just before I_c indicates the registering detection efficiency saturation of the nanowire which hints a high SDE once the coupling and absorption efficiencies are optimized. Moreover, previous works showed that biasing the detector on the edge of the plateau slightly below I_c gives a better timing jitter

[61] and shorter recovery time [77]. In our case, we were able to secure this criteria as shown in Figure 3.10.a over a range of 3 μA before critical current. The time jitter measured at 10 μA gave a full-width half-maximum (FWHM) value of the Gaussian distribution equal to 115 ps (Figure 3.10.b). This high value was expected for large sensitive-area meander because of its low critical current (low SNR), high kinetic inductance inducing a higher rising time of the output pulse (SR) which increases the noise jitter, and because it is simply longer which gives higher geometric jitter. The high kinetic inductance of large-area meanders also led to a quite high recovery time (90% of the maximum efficiency is recoverd after 420 ns) limiting the max count rate to 2.3 MHz as shown in Figures 3.10.c-d.

Large-area Parallel-SNSPDs

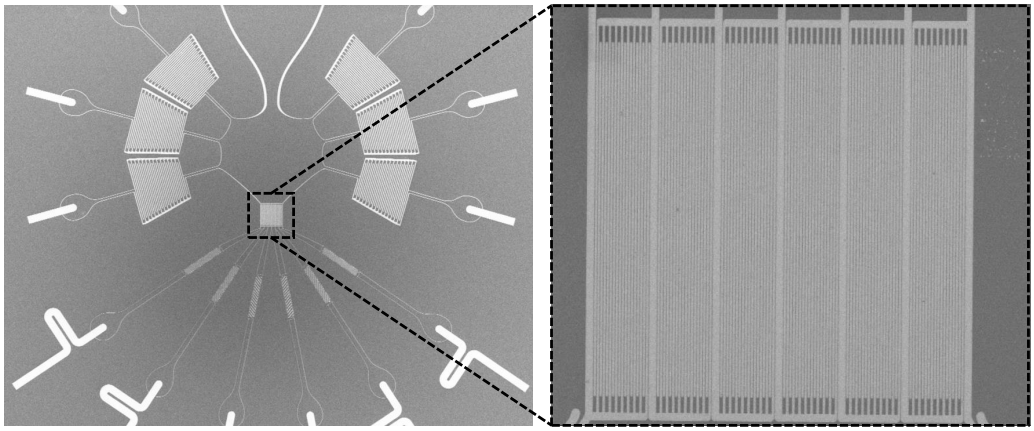


Figure 3.11: (a)SEM image of a NbTiN parallel-SNSPDs with 6 photosensitive nanowires and 6 non photosensitive additional large wires. The 6 active nanowires can be seen in the center of the image (dashed square). The 6 additional nanowires are positioned on a circle around the center. Each non photosensitive nanowire is equivalent to 4 sensitive nanowires. b) Zoom on the photosensitive area of the detector where we can see the 6 active nanowires connected in parallel and covering a total area of 50 μm x 50 μm. The active nanowires width is equal to 140 nm with 0.4 fill-factor.

The same characterization was performed on a parallel-SNSPD having 6 active nanowires and 6 non photosensitive large nanowires as depicted in Figure 3.11. The active nanowires have a width of 140 nm and 0.4 fill-factor. The photon count rate measurement gave a saturated curve over more than 50 μA as shown in Figure 3.10.a. This "plateau" indicates a saturation of the registering detection efficiency. The total critical current was equal to 407 μA which is equivalent to

13 μA for one section. The time jitter curve, depicted in Figure 3.12.b, demonstrated a FWHM of 104 ps, slightly lower than the single meander due to the lower kinetic inductance that reduced the noise jitter contribution. A non-Gaussian tail is however observed and becomes more obvious when we reduce the bias current. Even though similar behaviors have been reported in many studies [78, 61, 18], its origin remains unclear to date.

The global recovery time of the detector which designates the recovery time of

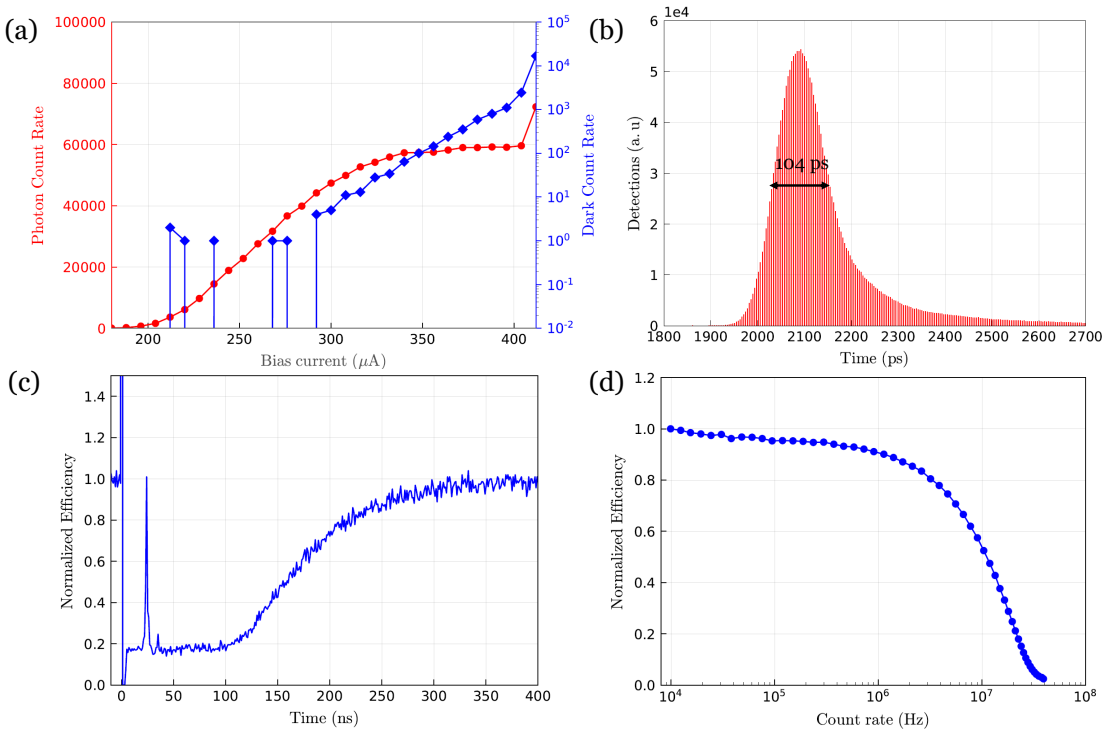


Figure 3.12: Characterization results of $50\mu m^2$-active area parallel-SNSPD with 6 sensitive sections. The measurements were done at 0.8 K for 1550 nm wavelength.(a) Photon count rate (red curve) and dark count rate (blue curve) as a function of the bias current I_b. (b) Timing jitter histogram measured at 390 μA. (c) Recovery time measured at 380 μA with the hybrid-autocorrelation method described in [77]. The peak apparent at around 25 ns is due to an optical reflection in the measurement setup (d) Normalized efficiency measured at 380 μA as a function of the detection rate.

the 6 parallel photosensitive nanowires after they all click, is measured using the hybrid autocorrelation method described in [77] and Figure 3.12.c shows the corresponding histogram. 90% of the total relative detection efficiency is recovered after 260 ns, much faster than a large active-area single meander. This reduction was expected since a single active section is shorter than the nanowire patterned into a single meander. The efficiency drop to zero is not shown here because we

were not able to provide high laser power (enough to force all the active sections to click) using flood-illumination because of heating effects.

At high detection rates, several consecutive photons can be absorbed in the same active nanowire before it could fully recover its efficiency. This effect can be seen simultaneously in several sections. As a result, the average efficiency will decrease with the increasing rate of incident photons. To characterize this effect, the rate of incident photons is progressively increased for a given bias current and the average detection efficiency per photon is calculated from the detection rate value. Figure 3.12.d. shows the normalized efficiency vs the detection rate of our device and we can see an efficiency drop to 50% of the nominal value at 10 MHz and to 2.5% at 40 MHz. This effect can be avoided by increasing the number of photosensitive nanowires which reduces the probability of having multiple detections in the same section, however such a detector can undergo latching effect if the nonsensitive wide nanowires are not optimized. More details about this efficiency drop effect is provided in [20].

3.2.5 Conclusion and outlook

This chapter was dedicated to Superconducting Nanowire Single Photon Detectors (SNSPDs) for telecom wavelengths. The first part introduced the operational principle of the device and the essential metrics used to characterize its performance and evaluate its compatibility with specific applications. The second part focused on large active-area SNSPDs for multimode fiber coupling and free-space detection. We presented the key fabrication steps of these devices, carried out at CMi-EPFL, and we also showed some of the challenges we came across during the nano-fabrication process.
After several design and fabrication iterations using flood-illuminated chips, we were able to demonstrate large active-area meanders and parallel-SNSPDs with saturated photon counting rate indicating a saturated registering detection efficiency. With the $50\mu m$-diameter single meanders, we registered 115 ps jitter and a maximum count rate of 2.3 MHz. On the other hand, the high-speed parallel design, featuring a low kinetic inductance, gave 104 ps time resolution and a maximum count rate of 40 MHz. These preliminary results can highly benefit many applications when the detectors are coupled to 50 μm core multimode fibers. However, we believe that the performance of our devices can be greatly enhanced by some design adjustments. For instance a higher fill factor than the 40% used with the tested SNSPDs is expected to improve the SDE.
The next step of this on-going project consists in bringing the fabrication to the wafer level where the nanowires are stacked inside the optical cavity and the overall design is compatible with the self-alignment technique to individual multimode

fibers. This approach should allow an accurate characterization of the system detection efficiency and the dark count rate of large active-area detectors which would improve the time jitter and the recovery time. We trust that our experience with SNSPDs [62, 20] will help us fabricate large sensitive-area detectors with state-of-the-art performance thanks to our original approach using parallel-nanowires design.

Chapter 4

Quantum Random Number Generators based on Photon Detectors

Random number generators (RNGs) are the key components in many applications such as quantum cryptography whose the security entirely relies on the use of high-quality random numbers. Genuine randomness cannot be generated by a deterministic algorithm [79] no matter how complex it may be, for the reason that, given the seed, the output sequence becomes predictable. Hardware-based RNGs can generate true randomness only under the assumption that the core physical process cannot be fully described by deterministic classical physics, which is not the case for instance of RNG based on thermal noise [80] or on clock drift [81].
In contrast to this, Quantum Random Number Generators (QRNGs) offer the ultimate solution for true random number generation thanks to the intrinsic probabilistic nature of quantum processes. Many practical implementations of QRNGs have been demonstrated in the past decades and have been based on specialized devices such as single photon sources and detectors [82, 83], CMOS image sensors [5], optical parametric oscillator [84] and homodyne detection [85].

In this chapter, we will present two different implementations of quantum random number generator, both using single photon detectors activated by a LED. The first QRNG is based on an array of CMOS Single Photon Avalanche Diodes (SPADs), and the second one uses CMOS Quanta Image Sensor (QIS) [86]. In both configurations the generation process is discussed, the quantum entropy[1] is evaluated and the true random data is submitted to statistical tests.

[1]The quantum entropy designates the entropy created by a quantum process.

This work was developed as part of a collaborative Eurostar project between Id Quantique SA and TU Delft. The project's main goal was to reduce the size and increase the throughput of RNGs, in order to make them more compatible with mobile devices and meet the mass market requirements.

4.1 Theoretical concept

4.1.1 Quantum entropy

If we want to define randomness, we would find many different approaches in the literature. For philosophers, randomness is a property of any event that happens by chance [87] and one can't always generate it or measure it. In information theory, a sequence of bits is called random if its Kolmogorov complexity is maximal [88], this definition does not guarantee the unpredictability of the bit sequence, an essential criteria for many applications. In this chapter, we define the randomness as **a property of a physical process whose the outcome is uniformly distributed and independent of all information available in advance** [89] and we show that the QRNGs described in the following sections output random data compliant to this definition.

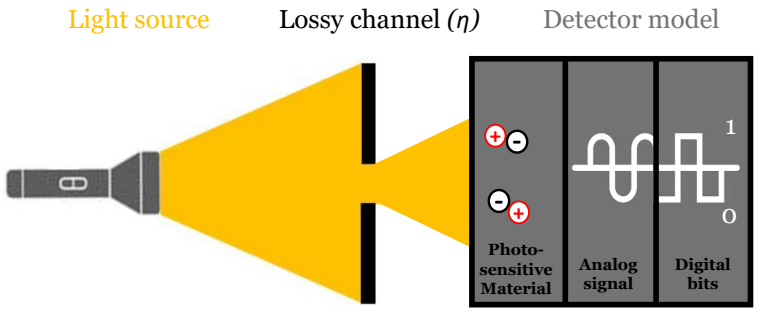


Figure 4.1: Modelling of a quantum random number generator based on photon detectors: a light source illuminates a photon detector through a lossy channel with a transmission probability η including all the optical coupling losses and the photon detection efficiency of the detector. The photon detector plays the role of the transducer and outputs digital bits correlated to the amount of detected photons.

A QRNG based on photon detectors can be modelled as a light source emitting photons according to a certain statistical time distribution and illuminating a photon detector through a lossy channel with a transmission probability η including

all the optical coupling losses and the photon detection efficiency of the detector. The photon detector plays the role of the transducer and outputs digital bits correlated to the amount of detected photons. Figure 4.1 shows a demonstration of this model.
To quantify the randomness of a sequence of bits, we refer to the concept of entropy in information theory, first introduced by Shannon [90]. Entropy measures the uncertainty associated with a random variable and is expressed in bits. Depending on the type of light source and on the detection mechanism of the transducer the origin of quantum entropy varies and so does its amount. that is why a perfect modelling of the physical process is required and should take into account the system imperfections. In fact, in practical implementation of QRNG, the desired quantum process cannot be created or measured perfectly, there are always hardware imperfections and different noise sources that cannot be controlled and that can affect the raw output of the TRNG by introducing a bias, a pattern or some classical noise. All these side information should be taken into account for accurate quantum entropy quantification.

4.1.2 Entropy extraction

The hardware imperfections and noise sources mentioned above can dramatically reduce the amount of generated quantum entropy. Fortunately, it is still possible to obtain true randomness by applying an appropriate randomness extraction to the raw random bits generated by the device. Block-wise hashing functions are widely used and for our QRNGs we used a hash function based on vector-matrix multiplication (Toeplitz matrix) [4]. This extractor is applied to vectors of n raw bits and output shorter vectors of k true random bits. The extraction matrix elements (n x k) are usually constant and generated by an independent RNG. Once n, the length of the raw string, is chosen, the parameter k will be bound by the probability that the output bit string deviates from perfectly random output bits ϵ. If the extraction function is taken from a two-universal family of hash functions, it is possible to quantify this failure probability by the Leftover Hash Lemma with side information:

$$\epsilon = 2^{-(h \cdot n - k)/2} \tag{4.1}$$

Where h is the min-entropy/bit of the input vector of n raw bits. Since a value of ϵ=0 is generally unachievable, we try to keep ϵ below 2^{-100} implying that even using millions of photon detectors during a time longer than the age of the universe, we won't be able to see any deviation from perfect random sequence of bits.

4.1.3 Statistical tests

After efficient extraction, the final sequences of bits should be uniformly distributed and independent of all information available in advance. However, this statement is not sufficient for applications requiring high level of security such as cryptography. The generated random numbers still need to be tested for particular weaknesses. On the other side, if the theoretical model of the QRNG gives near-unity entropy/bit, the extraction may not be necessary and in that case another proof is needed to "certify" the generated data for use and increase the level of trust in the generator.
Many standard batteries of statistical tests have been conceived to bring the solution for this concern. These tests characterize the properties of a random sequence in terms of probabilities. Some of the tested criteria are the proportion of zeroes and ones in the entire sequence or within a block of bits, the frequency of sequences with identical bits, the presence of periodic and aperiodic patterns and the compressibility. The NIST tests suite (SP 800-90B) [91] is one of the most used set of tests to characterize physical RNGs, but we can also cite the "Diehard" battery of tests [92] and the "dieharder" tests [93] that combines both; NIST and Diehard with some extra tests.

Note that there is no "complete" set of tests for perfect randomness evaluation and that some of the sequences are expected to fail the statistical tests with a limited probability. Moreover, the results of these tests should be interpreted cautiously to avoid incorrect outcomes.

4.2 QRNG based on SPADs matrix

The first QRNG presented in this chapter was designed and manufactured in TU Delft and characterized in Id Quantique in collaboration with the University of Geneva.
Our system is based on a parallel array of independent CMOS SPADs (the transducer), homogeneously illuminated by a DC-biased LED (the light source). The digital postprocessing including data readout and entropy extraction is integrated on the same Si chip and outputs true random data with high throughput.

We first discuss the randomness generation process and estimate the amount of generated quantum entropy, then we evaluate the influence that the co-integrated digital postprocessing may have on the raw randomness quality, and finally we do statistical analysis on final data after randomness extraction. More details about this work are provided in the preprint "A Fully Integrated Quantum Random Number Generator with on-Chip Real-Time Randomness Extraction" (see Preprint articles)
Our SPADs-based QRNG can output up to 400 Mbit/s with significant area reduction and low power consumption.

4.2.1 Randomness generation process

The quantum process exploited in this implementation is the distribution of photon over the surface of the detectors. The position of the photon is not deterministically known before its detection. The probability distribution of its position is given by the intensity profile of the electromagnetic field impinging on the SPADs.
The distance between the detectors and the light source is chosen so that the light intensity profile on the detectors is uniform. The quantum state emitted by the LED is a mixed state in the Fock space, where the probability of having n photons is given by the Poisson distribution P_N:

$$P_N(n) = e^{-\lambda}\frac{\lambda^n}{n!} \tag{4.2}$$

The matrix of SPADs is modelled as an array of independent detectors. Their efficiency η is considered as the probability that one photon is detected. If n photons arrive on one detector, the probability that at least one of them is detected (probability that the detector clicks) is given by $P^n_{det} = 1-(1-\eta)^n$. Similarly, we can model the dark counts probability with a random variable $S_i = 0, 1$ for each detector. In the case of $S_i = 1$ the i^{th} detector will click independently of the light

coming into it. This event has probability p_{dark} equal to the probability of having a dark count on one pixel.

Following the work of Frauchiger *et al.* in order to evaluate the amount of quantum randomness present in our system we have to evaluate the min-Entropy of our distribution conditioned on all possible side information that could be predicted by a third party. In this work we are going to consider the side information as classical, determined by the random variable E. In this case the conditional min-Entropy takes the form:

$$H_{min}(X|E) = -\sum_{e} P_E(e) \log_2[\max_x P_{X|E=e}(x|e)] \tag{4.3}$$

where X represents the random variable of the output sequence, and $P_{X|E}$ is the conditional probability distribution of X knowing the variable E. Moreover the quantity $2^{-H_{min}(X|E)}$ represents the maximum guessing probability of X given E.

In our analysis we consider the photon distribution of the source and the random variables corresponding to the dark counts and the crosstalk as the principal source of classical side information denoted by the variable E, and we obtain a min-Entropy $H_{min}(X|E)/m \approx 77\%$ for m independant SPADs ($m \geq 30$), $\eta = 0.12 \pm 0.03$ (considering a possible deviation of 25% from the average efficiency), $p_{dark} = 8.45 \cdot 10^{-5}$ per detection window, $P_{cross} = 0.001$ (this corresponds to the worst case scenario where each click provoked by the click of another detector is known by an adversary) and a mean photon number arriving on the detectors λ set experimentally to give a probability of 0.5 for each detector to click. More details about the entropy calculation are provided in the supplementary material of the prerpint "A Fully Integrated Quantum Random Number Generator with on-Chip Real-Time Randomness Extraction". This value gives the lower bound on the possible extractable randomness from the raw generated data.

Entropy extraction parameters

With the obtained value of min-Entropy and for a vector of 1024 raw bits, the parameter k should be lower than 588 for efficient entropy extraction.

4.2.2 Experiment

The QRNG chip comprises an array of SPADs organized in a matrix of 128 x 128 pixels illuminated by a DC-biased LED placed vertically at a distance of 5 mm. The entire matrix is read out every 1.3 μs and the raw data is stored in a 512-bit

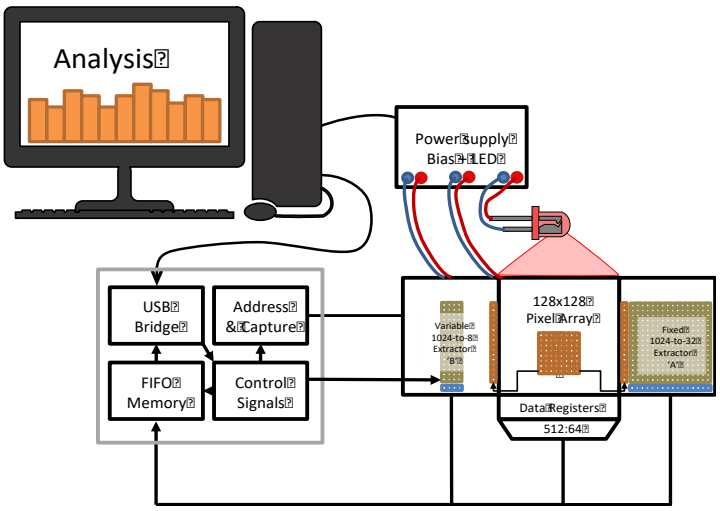


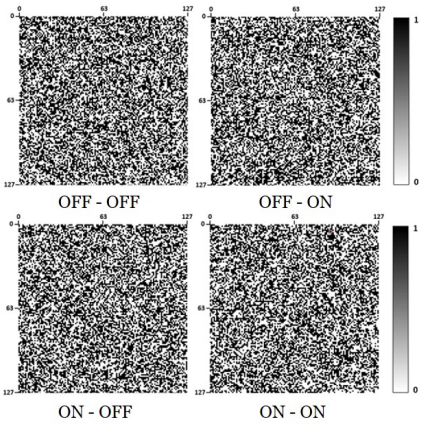
Figure 4.2: Data collection setup: The QRNG IC chip comprises an array of 128 x 128 SPADs with two extractors denominated 'A' and 'B'. The detectors are illuminated by a DC-biased LED and the final random data are read in packets of 8 bits and 32 bits at a rate of 12.5 MHz.

register whose content is diverted to two extractors based on vector-matrix multiplication as described previously. The first is a fixed extractor, denominated 'A'; it is realized from standard logic gates using a hard-coded 1024 x 32 matrix pre-generated from an independent QRNG; this extractor generates 400 Mbit/s of random data and is intended for applications requiring higher throughput and minimal area. The second is a variable extractor, denominated 'B' and is a matrix of 1024 x 8 memory elements, realized by means of scan chain registers with 100 Mbit/s throughput and is intended for applications requiring one to extract an ad hoc matrix in the field. The block diagram of the measurement setup is depicted in Figure 4.2. Note that if we did not have area limitation, we would have used bigger extractors up to 1024 x 588 which would have given higher generation rate.
The QRNG integrated circuit (IC) was fabricated in standard, 0.35μm 1P4M HV CMOS technology, where the SPADs exhibit low noise, negligible afterpulsing, and low crosstalk.

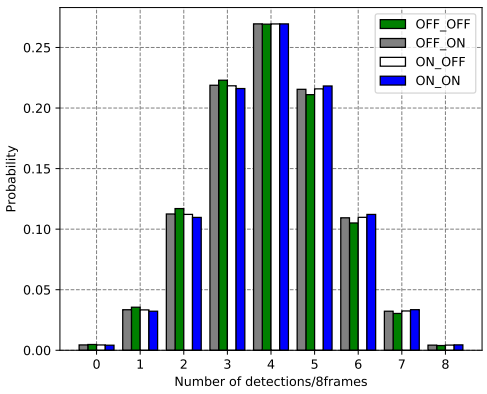
4.2.3 Results and tests

First we needed to evaluate the influence of on-chip digital post-processing on the quality of generated randomness in order to demonstrate the feasibility of a fully integrated QRNG in a commercially relevant CMOS technology. To do that, we compared the statistical distribution of raw data (before post-processing) when the two extractors are not powered on, when only one of them is working and when

both are performing on-the-fly extraction. Figure 4.3a shows the constellation



(a) Raw binary constellations.



(b) Detection histogram (8-bits Hamming Weight distributions) of raw data for different operation conditions

Figure 4.3: Comparison of raw data distribution when both extractors are OFF (OFF - OFF), when the fixed extractor is ON and the variable is OFF (ON - OFF), when the fixed extractor is OFF and the variable is IN (OFF - ON) and when both extractors are OFF (OFF - OFF)

generated by the SPAD array for these same conditions. From the Figure we can see the equi-probability of '1's and ò's (computed probability of 50.36% of '1' vs 49.64% of '0') with a homogeneous distribution among the pixels, irrespectively of the state of the extractors, at better than 2σ deviation from the 50% distribution mark. This observation is confirmed with the detection histograms (8-bits hamming weight distributions) in all four cases shown in Figure 4.3b. Therefore, we can conclusively confirm that the analog-digital co-integration does not have any serious effect on the quantum process generation and measurement.

The effect of temperature on the QRNG was also evaluated and we saw a slight decrease of entropy/bit when increasing temperature due to a higher thermally induced noise that can lead to a bias toward '1'.

Before extraction, the entropy/bit was considerably low as predicted by the theory, that is why entropy extraction was compulsory to increase and maintain the mean quantum entropy. Afterwards, about 1 Gbit of extracted data (using both extractors) were tested using the NIST statistical test suite [91] and the Diehard test battery [92] and it passed all of them.

4.2.4 Conclusion and Outlook

The integration of the entropy source with the post-processing logic allowed us to demonstrate the feasibility of CMOS integrated quantum random number generator. The QRNG chip achieved 400 Mbit/s throughput with the lowest energy per bit ever reported on standard CMOS technology for the nominal 3.3 V supply voltage at room temperature. This characteristics make our chip suitable for portable low-power applications.

The next step will target the integration of the LED on the same Si substrate and an efficient packaging of the whole chip to fulfill the compactness requirement and target the System-in-Package (SiP) approach. The generation rate could also be improved by decreasing the number of SPADs and using the available space to implement a bigger Toeplitz matrix.

4.3 QRNG based on Quanta Image Sensor

Another quantum random number generation method is proposed in this section. It follows the same scheme described in Section 4.1 using an array of LEDs as the light source and a Quanta Image Sensor (QIS) as a the transducer.

First introduced in 2005 by E. R. Fossum, the inventor of CMOS cameras, Quanta Image Sensor is an array of sub-micron pixels called "jots" [94]. Each jot can count incident photons and outputs single-bit or multi-bit digital signal reflecting the number of photoelectrons [95]. A QIS can include over one billion pixels read out at high speed, e.g. 1000 fps, with extremely low power consumption (2.5 pJ/bit) [96]. To guarantee the photon counting capability, deep sub-electron read noise (DSERN) is a prerequisite and it has been achieved with the pump-gate (PG) jot device designed by the Dartmouth group [97, 98] and used in this work.

The idea of using CMOS image sensors for random numbers generation was first brought by Bruno *et al.* [5] from the University of Geneva; the quantum entropy extracted from their QRNG was limited by the technical noise (dark current and 1/f noise) which required high compression. That is why I thought that the DSEN of the QIS would help solving this problem. Moreover, the huge number of jots should boost the final throughput of the device.

4.3.1 Randomness generation process

First of all, our devise is a hardware-based RNG that makes use of photon emission to generate randomness. Photon emission is a quantum process which is intrinsically probabilistic. In our case, the light source is modeled as a mixed density operator over Fock states and emits photons according to Poisson statistics. The probability P[k] of k photoelectron generated in a QIS jot is given by:

$$P[k] = e^{-\lambda}\frac{\lambda^k}{k!} \tag{4.4}$$

where λ is the average number of photoelectrons collected in each jot per frame. So under the illumination of a stable light source, randomness exists in the number of photoelectrons arriving in each frame.
The output signal U emitted by the jots is corrupted by the readout noise and the readout signal probability distribution function (PDF) becomes a convolution of the Poisson distribution with an average number of photoelectrons λ and a normal distribution with read noise u_n (measured in e- r.m.s. after normalization by the conversion gain (V/e-)). The result is a sum of constituent PDF components, one for each possible value of k and weighted by the Poisson probability for that k

[99]:

$$P[U] = \sum_{k=0}^{\infty} \frac{1}{\sqrt{2\pi u_n^2}} \left[-\frac{(U-k)^2}{2u_n^2} \right] \cdot e^{-\lambda} \frac{\lambda^k}{k!} \tag{4.5}$$

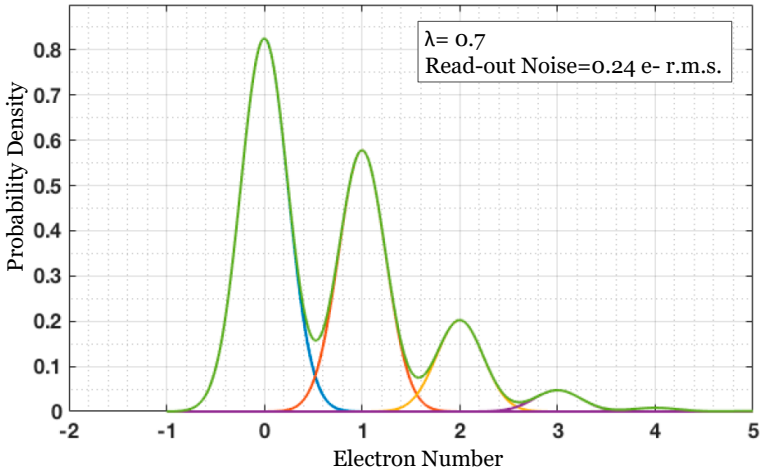


Figure 4.4: Readout signal probability distribution function (PDF) from Poisson distribution corrupted with readout noise for $\lambda = 0.7$ and read noise $u_n = 0.24 er.m.s.$

An example of a Poisson distribution for λ= 0.7 corrupted with readout noise u_n= 0.24 e- r.m.s. is shown in Figure 4.4.
While jots are sensitive to multiple photons and can output signal of multiple photoelectrons, subsequent electronics can be used to discriminate the output to two binary states: "0" meaning no photoelectrons and "1" meaning at least one photoelectron, by setting a threshold U_t between 0 and 1, typically 0.5 and comparing U to this threshold. In this case the probabilies of the two states are given by:

$$\text{"0"}state : P[U < U_t] = \sum_{k=0}^{\infty} \frac{1}{2} \left[1 + erf\left(\frac{U_t - k}{u_n \sqrt{2}} \right) \right] \cdot e^{-\lambda} \frac{\lambda^k}{k!} \tag{4.6}$$

$$\text{"1"}state : P[U \geq U_t] = 1 - P[U < U_t] \tag{4.7}$$

The minimum quantum entropy of this distribution is given by:

$$H_{min} = -log[max(P[U \geq U_t], P[U < U_t]] \tag{4.8}$$

4.3.2 Experiment

A chip composed of 32 x 32 PG jots was used for data collection. The output signal from the 32 columns was selected by a multiplexer and then amplified by

a switch-capacitor programmable gain amplifier (PGA) before being sent off-chip for digitization using a 14-bit ADC. A complete description of readout electronics can be found in [98]. A 3 x 3 array of green LEDs was used as light source and was placed at 2 cm above the test chip. The intensity of the light source was controlled by a precision voltage source.

First a single jot was selected and read out at a speed of 10 ksamples/s and

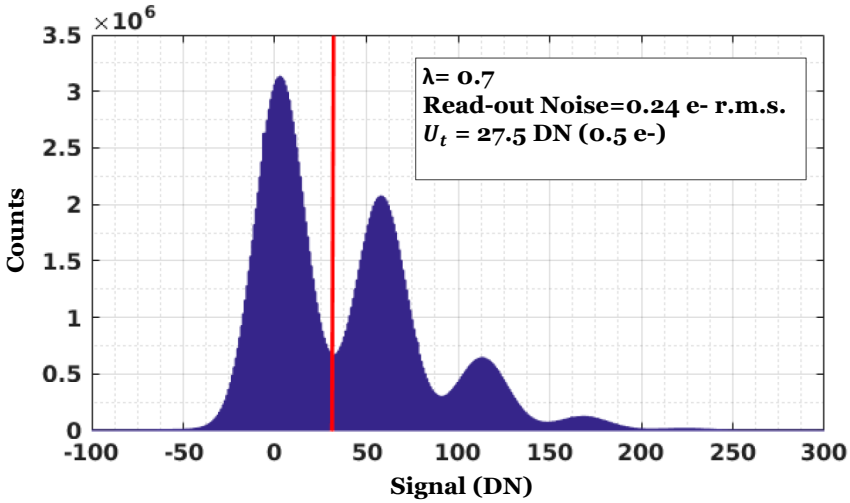


Figure 4.5: Photon counting histogram (PCH) of the first 200,000,000 samples

a 14 bits raw digital output was collected for each sample. The average number of photoelectrons collected in the jot was obtained from The Photon Counting Histogram of the first 20000 test values and the threshold U_t was determined as the median of the testing samples and then used in the following measurements. We chose a light intensity that gave an average number of photoelectrons collected in each jot per frame equal to 0.7 in order to obtain $P[U \geq U_t] \approx P[U < U_t] \approx 0.5$. Note that other combinations of λ and U_t are possible but the ones chosen gave better stability, more details are found in [86].
The experimental photon counts histogram (PCH) created by 200,000,000 samples is shown in Figure 4.5 and it fits with the theoretical model.
Once the measurements parameters fixed, the data was collected from the whole chip and sent off-chip for further analysis.

4.3.3 Results

Using equation 4.8 for 500 Mbyte of random raw data, we computed a minimum quantum entropy per output bit equal to 0.9845 with a mean photoelectron number $\lambda = 0.7$ and $u_n = 0.24$ e- r.m.s. Then we used the obtained value in the formula 4.1

with n= 1024 and it gave a compression factor equal to 1.23 (k= 832) which corresponds to losing only 18% of the input raw data.

After extraction, the NIST tests (see section 4.1 for more details) were performed and the obtained random bits passed all these tests.

4.3.4 Conclusion and Outlook

A new quantum random number generation method based on the QIS is proposed. Taking advantage of the randomness in photon emission process and the photon counting capability of the quanta image sensor, it showed high randomness quality (near-unity entropy/bit for raw data) and promising data rate (In an array of 2.5 mm^2 area size we can fit millions of jots and reach up to 12 Gb/s throughput).

More efforts have been done since the publication of this work (2016) to improve the performance of QIS in terms of speed, noise and scalability [100, 101, 102] which would equally enhance the performance of the QRNG based on this technology. Moreover, Quanta Image Sensors are now commercially available by Gigajot [103], a startup founded in 2018 by the co-writers of our paper " Quantum random number generation using Quanta Image Sensors" [86] and this would potentially drive the industrialization of our QRNG chip. An application patent has been already co-filed by Id Quantique and Thayer School of Engineering at Dartmooth about the use of QIS for random number generation [104].

4.4 Our QRNGs and the state of the art

In this chapter we presented two different implementations for quantum random number generation . The first QRNG is based on SPADs matrix (Section 4.2) and can be regarded as a stand-alone system. It has the advantage of integrating the entropy source and the digital post-processing on the same small chip (10 mm x 4.5 mm) with an overall power at full speed less than 500 mW for 3.3V supply voltage, which make the chip suitable for low-power portable applications. Although it has been demonstrated that this device produce data of a satisfactory randomness quality, more work needs to be done to enhance the generation process, especially on the improvement of output data rate and device scalability because implementing bigger extractors would increase the generator throughput but also require bigger area and longer processing time. The SPADs array could also be fabricated using more advanced nodes which would enable significant area reductions and could compensate for the extractor size.

The second QRNG is based on Quanta Image Sensors (Section 4.3) and it showed promising advantages over previous QRNG technologies because QIS seems to cover the advantages of both SPADs and conventional CMOS image sensors ((best trade-off between data rate and scalability, single photon detection and CMOS manufacturing line) while providing solutions for most of their problems (speed, dark count rate, detection efficiency). Table 4.1 summarizes the comparison of SPADs (cite paper charbon), QIS [86] and Cmos Image Sensors (CIS) [5, 105] under the assumption of being used as RNGs. Note that the generation processes are different which limits the comparison points.

A more exhaustive comparison of most of the technologies that have been used for quantum random number generation is provided in the supplementary material of the prerpint “A Fully Integrated Quantum Random Number Generator with on-Chip Real-Time Randomness Extraction”.

Criteria	SPADs	QIS	CIS
Bit rate	400 Mb/s	5-12 Gb/s	4.9 Mb/s
Read Noise(r.m.s.)	< 0.15 e-	< 0.25 e-	> 1 e-
On-chip extraction	yes	no	yes
Dimensions(mm^2)	45	6.25	22.5
Power (nJ/bit)	1.25	0.030	17
Certifications	NIST, Diehard	NIST	NIST, AEC-Q100

Table 4.1: SPADs array vs QIS vs CIS for quantum random number generation

Chapter 5

General Conclusion and Outlook

During this thesis two approaches to single-photon detection have been investigated, using semiconductor and superconducting technologies. Despite the huge amount of applications relying on single-photon detectors, the technology studied here was quantum random number generation (QRNG), since this is a relatively young field at the core activity of Id Quantique and the University of Geneva.

5.1 Summary of the results

In the first part of this work, I discussed the low-temperature behavior of free-running InGaAs/InP negative feedback avalanche diodes (NFADs) with a special focus on their time resolution. This analysis has enabled the understanding of the jitter contribution due to charge-carrier pile-up between the absorption and multiplication regions, a phenomena which has been rarely studied. We have shown that in order to avoid degradation of the temporal resolution due to this effect, the operating voltage of these devices should be greater than 67 V at the lowest operation temperature. Given this, excess bias voltage and temperature can be chosen freely according to the applications requirements and a jitter as low as 52 ps, which is comparable to the best gated-mode devices, can be achieved.
One of the major drawbacks of the InP based SPADs is the significant amount of afterpulsing making operation at the free-running regime challenging. Using a quantum cascade laser emitting at 4.9 μm and sent at the NFAD after avalanche events, I showed a decrease of afterpulsing at short delays by releasing a part of trapped carriers in the defects of the InP layer. These preliminary results are quite promising but more work is required to optimize the experiment and improve this AP reduction effect.

The second part of the thesis focused on the development, fabrication and characterization of superconducting nanowire single-photon detectors. These nanodevices have showed a tremendous potential in the near-infrared range and beyond, owing to their unique combination of high detection efficiency, low dark count rate, excellent time resolution, high speed and no afterpulsing. Standard MoSi SNSPDs with active areas of 15 - 20 μm-diameter have demonstrated important breakthroughs in the last few years [10, 48, 61] and in this thesis I tried to build the road leading to similar performance using large sensitive-area SNSPDs that can be coupled to MM fibers. The secret recipe being the use of high-speed parallel-SNSPDs design [20] covering larger area (50 μm x 50 μm). For this proof-of-principle study, I used flood-illuminated NbTiN chips and I was able to demonstrate large active-area meanders and parallel-SNSPDs with saturated photon counting rate, indicating a saturated registering detection efficiency. With the 50 μm-diameter single meanders, we registered 115 ps jitter and a maximum count rate of 2.3 MHz. On the other hand, the high-speed parallel design, featuring a low kinetic inductance, gave 104 ps time resolution and a maximum count rate of 40 MHz. These first results open up a host of applications requiring free-space or MM fibers coupling. An accurate characterization of the system detection efficiency and the dark count rate is still necessary before turning this research into a possible industrial product.

Both detector technologies have their place in various applications, since it is not always necessary to combine all state-of-the-art attributes simultaneously. The user has to choose between the ease-of-use of SPADs and the unequalled performance of SNSPDs.

The final part of the thesis looks at two implementations of quantum random number generators taking advantage of the randomness in photon emission process. The first QRNG is based on an array of independent SPADs, homogeneously illuminated by a DC-biased LED. This module has the advantage of integrating the entropy source and the digital post-processing on the same small chip (10 mm x 4.5 mm) with an extremely low energy per bit (1.25 nJ/bit), and can be regarded as a stand-alone system suitable for low-power portable applications. The second QRNG is based on Quanta Image Sensors (QIS) [98] and it showed promising assets over previous QRNG technologies since QIS seems to cover the advantages of both SPADs and conventional CMOS image sensors, while providing solutions for most of their problems. This QRNG showed high randomness quality (near-unity entropy/bit for raw data) and promising data rate (in an array of 2.5 mm^2 area we can fit millions of jots and reach up to 12 Gb/s throughput). A patent application was co-filed by Id Quantique and Thayer School of Engineering at Dartmooth as a first step toward the possible industrialization of this QRNG.

The time-frame of my thesis arrives to an end but there are still many unfinished developments, open questions and new directions which are left to be explored. This shall now be briefly discussed.

5.2 Outlook into the future of the studied technologies

InGaAs/InP single photon avalanche diodes

Following the work of Korzh, B. [106] and my own work [17] about the performance of free-running InGaAs/InP SPADs at low temperatures, we conclusively showed that low temperature operation (between -90°C and -110°C) is very favorable to optimize the performances of these detectors. Keeping this in mind, the structure of the device should be adapted for this temperature range. For instance, the absorption region thickness can be increased in order to increase the photon absorption efficiency without side effects. This has become possible because at low temperatures thermally generated dark counts are negligible [19] so we do not expect a dramatic increase of the DCR. On the other hand, the ionization coefficients in the multiplication region are higher [43] which would lead to shorter avalanche build-up time and would reduce the timing jitter.
From a system point of view, the electronic jitter contribution could be reduced by using low-temperature readout electronics cooled-down at the same temperature as the SPAD. This idea can be pushed further by developing integrated quenching and readout circuits on the same substrate as the detector. This integration would have many advantages including cost-effectiveness, detector miniaturization, parasitic capacitance minimization and power reduction. It will also enable the implementation of III–V SPAD arrays which would be beneficial for applications requiring multi-pixel near-infrared single-photon detection.

Superconducting nanowires single photon detectors

SNSPDs are already available as off-the-shelf products in the single photon detectors market thanks to their outstanding performance that has surpassed all competition. Yet, there are still several challenges that are awaiting to be addressed.
The nano-fabrication process is the first to be optimized. For instance, making extremely clean mirrors with the highest refelctivity, together with depositing homogeneous dielectric layers with the precise intended thickness will definitely improve the absorption efficiency and help reaching near-unity SDE. Similarly, the absorption efficiency of other wavelengths can be improved. As for the dark

count rate, it is actually limited by the black body radiation at room temperature that propagates through the optical fiber to the detectors inside the cryostat. One effective solution to suppress this limitation would be the use of low-loss filters directly deposited on the fiber tip. In fact, we have started investigating this option for multimode fibers. Finally, the parallel nanowires design introduced in this thesis [20] seems to offer the ultimate solution to maximize all the attributes in one device and I strongly believe that this design, with some adjustments, will replace the standard meander structure in most of the applications.
From a system point of view, integrated cryogenic electronics [107, 108] would definitely improve the SNR and the timing jitter and a high-impedance readout approach would push the limitation of the intrinsic recovery time and increase the maximum count rate. More efforts should be put in the miniaturization of cryostats and the compactness of the total cryogenic system which would promote the use of SNSPDs in technically-challenging applications.

Quantum Random Number Generation

The importance of quantum random number generators has been proven in several fields, and we see more and more applications opening up to this type of TRNG. Mobile devices and Internet-of-Things (IoT) offer an enormous potential to expand the use of QRNGs and bring the concept to the next level. To make this diffusion possible, QRNG system should be optimized in terms of integration, scalability power consumption and cost-effectiveness, while keeping a high throughput rate. A QRNG system should translate into a QRNG IC chip that outputs true random data ready to be used. The chip should be completely independent from the environment where it shall be integrated (System In Package approach) while being compatible with universal communication interfaces. The chip should also dispose of enough power for on-chip extraction without exceeding the standard of IC components used in mobile devices. In my opinion, QRNGs will become more of a challenge for the electronic IC design field than for physicists and quantum theorists.

Bibliography

[1] A. Migdall, "Introduction to journal of modern optics special issue on single-photon: Detectors, applications, and measurement methods," *J. Mod. Opt*, vol. 51, p. 1265–1266, 2004.

[2] M. Nielsen and I. Chuang, *Quantum Computation and Quantum Information Ch. 1.* Cambridge Univ. Press, 2000.

[3] R. G. T. W. Gisin, N and H. Zbinden, "Quantum cryptography," *Rev. Mod. Phys.*, vol. 74, p. 145–195, 2002.

[4] M. Troyer and R. Renner, "Id Quantique technical report," 2012.

[5] B. Sanguinetti, A. Martin, H. Zbinden, and N. Gisin, "Quantum random number generation on a mobile phone," *Phys. Rev. X*, vol. 4, p. 031056, 2014.

[6] P. Kok, W. J. Munro, K. Nemoto, T. C. Ralph, J. P. Dowling, and G. J. Milburn, "Quantum cryptography," *Rev. Mod. Phys.*, vol. 79, p. 135, 2007.

[7] G. A. Morton, "Photomultipliers for scintillation counting," *RCA Rev.*, vol. 10, p. 525–553, 1949.

[8] A. Fukasawa, J. Haba, A. Kageyama, H. Nakazawa, and M. Suyama, "High speed hpd for photon counting," *IEEE Trans. Nucl. Sci.*, vol. 55, p. 758–762, 2008.

[9] S. Cova, A. Longoni, and A. Andreoni, "Towards picoseconds resolution with single-photon avalanche diodes," *Rev. Sci. Inst.*, vol. 52, p. 408–412, 1981.

[10] F. Marsili, V. B. Verma, J. A. Stern, S. Harrington, A. E. Lita, T. Gerrits, I. Vayshenker, B. Baek, M. D. Shaw, R. P. Mirin, and S. W. Nam, "Detecting single infrared photons with 93 % system efficiency," *Nat. Photon.*, vol. 7, p. 210–214, 2013.

[11] D. V. Reddy, R. R. Nerem, A. E. Lita, S. W. Nam, R. P. Mirin, and V. B. Verma, "Exceeding 95% system efficiency within the telecom C-band in superconducting nanowire single photon detectors," in *OSA Conference on Lasers and Electro-Optics*, p. paper FF1A.3, 2019.

[12] Hamamatsu, *H7422-40 - PMT*. https://www.hamamatsu.com/.

[13] Hamamatsu, *H10330C-25 - PMT*. https://www.hamamatsu.com/.

[14] J. Blazej, "Photon number resolving in geiger mode avalanche photodiode photon counters.," *J.Mod. Opt.*, vol. 51, p. 1491–1498, 2004.

[15] G. Ribordy, J. D. Gautier, H. Zbinden, and N. Gisin, "Performance of ingaas/inp avalanche photodiodes as gated-mode photon counters," *Appl. Opt.*, vol. 37, p. 2272–2277, 1998.

[16] N. N. Akiko Tada and S. Inoue, "Sinusoidally gated InGaAs/InP avalanche photodiode with 53% photon detection efficiency at 1550 nm," in *Conference on Lasers and Electro-Optics*, Optical Society of America, 2016.

[17] E. Amri, G. Boso, B. Korzh, and H. Zbinden, "Temporal jitter in free-running InGaAs/InP single-photon avalanche detectors," *Opt. Lett*, vol. 41(24), pp. 5728–5731, 2016.

[18] B. Korzh, Q. Zhao, S. Frasca, J. Allmaras, T. Autry, E. Bersin, M. Colangelo, G. Crouch, A. Dane, T. Gerrits, F. Marsili, G. Moody, E. Ramirez, J. Rezac, M. Stevens, E. Wollman, D. Zhu, P. Hale, K. Silverman, R. Mirin, S. Nam, M. Shaw, and K. Berggren, "Demonstrating sub-3 ps temporal resolution in a superconducting nanowire single-photon detector," *ArXiv e-prints*, vol. 1804.06839, 2018.

[19] B. Korzh, N. Walenta, T. Lunghi, N. Gisin, and H. Zbinden, "Free-running InGaAs single photon detector with 1 dark count per second at 10% efficiency," *Appl. Phys. Lett.*, vol. 104, p. 081108, 2014.

[20] M. Perrenoud, M. Caloz, E. Amri, H. Zbinden, and F. Bussieres, "High detection rate and high efficiency with parallel-SNSPDs," *Preprint to be submitted to Appl. Phys. Lett.*, vol. List of "Preprint articles", 2020.

[21] S. Takeuchi, J. Kim, Y. Yamamoto, and H. H. H, "Development of a high-quantum-efficiency single-photon counting system," *Appl. Phys. Lett*, vol. 74, p. 1063–1065, 1999.

[22] B. Cabrera, R. M. Clarke, P. Colling, A. J. Miller, S. Nam, and R. W. Romani, "Detection of single infrared, optical and ultraviolet photons using su-

perconducting transition edge sensors," *Appl. Phys. Lett*, vol. 73, p. 735–737, 1998.

[23] E. J. Gansen, M. W. Rowe, M. B. Greene, D. Rosenberg, T. E. Harvey, M. Y. Su, R. H. Hadfield, S. W. Nam, and R. P. Mirin, "Photon-number-discriminating detection using a quantum-dot, optically gated, field-effect transistor," *Nat. Photonics*, vol. 1, p. 585–588, 2007.

[24] R. Hadfield, "Single-photon detectors for optical quantum information applications," *Nat. Photonics*, vol. 3, p. 696–705, 2009.

[25] C. J. Chunnilall, I. P. Degiovanni, S. Kück, I. Müller, and A. G. Sinclair, "Metrology of single-photon sources and detectors: a review," *Opt. Eng.*, vol. 53, p. 081910, 2014.

[26] M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov, "Invited review article: Single-photon sources and detectors," *Rev. Sci. Instrum.*, vol. 82, p. 071101, 2011.

[27] R. J. McIntyre, "Multiplication noise in uniform avalanche diodes," *IEEE Trans. Elec. Dev.*, vol. 13, pp. 164–168, 1966.

[28] I. Umebu, A. N. M. M. Choudhury, and P. N. Robson, "Ionisation coefficients measured in abrupt InP junctions," *Appl. Phys. Lett.*, vol. 36, pp. 302–303, 1980.

[29] N. Walenta, T. Lunghi, O. Guinnard, R. Houlmann, H. Zbinden, and N. Gisin, "Sine gating detector with simple filtering for low-noise infra-red single photon detection at room temperature," *J. Appl. Phys.*, vol. 112, p. 063106, 2012.

[30] A. Restelli, J. C. Bienfang, and A. L. Migdall, "Single-photon detection efficiency up to 50% at 1310 nm with an InGaAs/InP avalanche diode gated at 1.25 ghz," *Appl. Phys. Lett*, vol. 102, p. 141104, 2013.

[31] S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, "Avalanche photodiodes and quenching circuits for single-photon detection," *Appl.Opt.*, vol. 35, pp. 1956–1976, 1996.

[32] S. S. Buller, S. J. Fancey, J. S. Massa, A. C. Walker, S. Cova, and A. Lacaita, "Time-resolved photoluminescence measurements of InGaAs/InP multiple-quantum-well structures at 1.3-m wavelengths by use of germanium single-photon avalanche photodiodes," *Appl.Opt.*, vol. 35, pp. 916–921, 1996.

[33] C. Yu, M. Shangguan, H. Xia, J. Zhang, X. Dou, and J. W. Pan, "Fully integrated free-running InGaAs/InP single-photon detector for accurate lidar applications," *Opt. Exp.*, vol. 25, pp. 14611–14620, 2017.

[34] D. Bronzi, F. Villa, S. Tisa, A. Tosi, F. Zappa, D. Durini, S. Weyers, and W. Brockherde, "100 000 frames/s 64×32 Single-Photon Detector Array for 2-D imaging and 3-D ranging," *IEEE J. Sel. Top. Quant. Elec.*, vol. 20, p. 3804310, 2014.

[35] G. Boso, D. Ke, B. Korzh, J. Bouilloux, N. Lange, and H. Zbinden, "Time-resolved singlet-oxygen luminescence detection with an efficient and practical semiconductor single-photon detector," *Bio. Opt. Exp.*, vol. 7, pp. 211–224, 2016.

[36] M. A. Itzler, X. Jiang, B. Nymanand, and K. Slomkowski, "InP-based negative feedback avalanche diodes," *Proc. SPIE*, vol. 7222, p. 72221K, 2009.

[37] M. Sanzaro, N. Calandri, A. Ruggeri, and A. Tosi, "InGaAs/InP SPAD with Monolithically Integrated Zinc-Diffused Resistor," *IEEE J. Quantum Elec.*, vol. 52, p. 4500207, 2016.

[38] A. D. Mora, A. Tosi, F. Zappa, S. Cova, D. Contini, A. Pifferi, L. Spinelli, A. Torricelli, and R. Cubeddu, "Fast-gated single-photon avalanche diode for wide dynamic range near infrared spectroscopy," *IEEE J. QUANTUM ELECT.*, vol. 16, pp. 1023–1030, 2010.

[39] M. A. Itzler, X. Jiang, M. Entwistle, K. Slomkowski, A. Tosi, F. Acerbi, F. Zappa, and S. Cova, "Advances in InGaAsP-based avalanche diode single photon detectors," *J. Mod. Opt.*, vol. 58, pp. 174–200, 2011.

[40] T. Lunghi, C. Barreiro, O. Guinnard, R. Houlmann, X. Jiang, M. A. Itzler, and H. Zbinden, "Free-running single-photon detection based on a negative feedback InGaAs APD," *J. Mod. Opt.*, vol. 59, p. 1481, 2012.

[41] S. R. Forrest, O. K. Kim, and R. G. Smith, "Optical response time of $In_{0.53}$ $Ga_{0.47}$As/InP avalanche photodiodes," *Appl. Phys. Lett.*, vol. 41, p. 95, 1982.

[42] C. H. Tan, J. S. Ng, G. J. Rees, and J. P. R. David, "Statistics of avalanche current buildup time in Single-Photon Avalanche Diodes," *J. Sel. Topics Quantum Electron*, vol. 13, p. 906, 2007.

[43] F. Zappa, P. Lovati, and A. Lacaita, "Temperature dependence of electron and hole ionization coefficients in InP," in *Eighth International Conference on Indium Phosphide and Related Materials*, IEEE, 1996.

[44] A. Tosi, F. Acerbi, M. Anti, and F. Zappa, "InGaAs/InP Single-Photon Avalanche Diode with reduced afterpulsing and sharp timing response with 30 ps tail," *IEEE J. Quantum Elec.*, vol. 48, pp. 1227–1232, 2012.

[45] L. You, X. Yang, Y. He, W. Zhang, D. Liu, W. Zhang, L. Zhang, L. Zhang, X. Liu, S. Chen, Z. Wang, and X. Xie, "Jitter analysis of a superconducting nanowire single photon detector," *J. AIP Adv.*, vol. 3, p. 072135, 2013.

[46] B. Korzh, T. Lunghi, K. Kuzmenko, G. Boso, and H. Zbinden, "Afterpulsing studies of low-noise InGaAs/InP single-photon negative-feedback avalanche diodes," *J. Mod. Opt.*, vol. 62, pp. 1151–1157, 2015.

[47] G. N. Gol'tsman, O. Okunev, G. Chulkova, A. Lipatov, A. Semenov, K. Smirnov, B. Voronov, and A. Dzardanov, "Picosecond superconducting single-photon optical detector," *Appl. Phys. Lett.*, vol. 79, p. 705, 2001.

[48] H. Shibata, K. Shimizu, H. Takesue, , and Y. Tokura, "Ultimate low system dark-count rate for superconducting nanowire single-photon detector," *Opt. Lett.*, vol. 40, pp. 3428–3431, 2015.

[49] A. Vetter, S. Ferrari, P. Rath, R. Alaee, O. Kahl, V. Kovalyuk, S. Diewald, G. N. Goltsman, A. Korneev, C. Rockstuhl, and W. H. P. Pernice, "Cavity-enhanced and ultrafast superconducting single-photon detectors," *Nano Lett.*, vol. 16, pp. 7085–7092, 2016.

[50] M. Hills, T. Bradshaw, S. Dobrovolskiy, S. Dorenbos, N. Gemmell, B. Green, R. Heath, T. Rawlings, K. Tsimvrakidis, V. Zwiller, M. Crook, and R. Hadfield, "A compact 4 k cooling system for superconducting nanowire single photon detectors," *IOP Conf. Series: Materials Science and Engineering*, vol. 502, p. 012193, 2019.

[51] V. Kotsubo, J. Ullom, and S. W. Nam, "Compact low-power Cryo-Cooling Systems for superconducting elements," US20190226724A1, Jul. 2019.

[52] A. Boaron, G. Boso, D. Rusca, C. Vulliez, C. Autebert, M. Caloz, M. Perrenoud, G. Gras, F. Bussieres, M. J. Li, D. Nolan, A. Martin, and H. Zbinden, "Secure Quantum Key Distribution over 421 km of Optical Fiber," *Phys. Rev. Lett.*, vol. 121, p. 190502, 2018.

[53] J. Zhu, Y. Chen, L. Zhang, X. Jia, Z. Feng, G. Wu, X. Yan, J. Zhai, Y. Wu, Q. Chen, X. Zhou, Z. Wang, C. Zhang, L. Kang, J. Chen, and P. Wu, "Demonstration of measuring sea fog with an SNSPD-based Lidar system," *Nat. Scient. Rep.*, vol. 7, p. 15113, 2017.

[54] G. G. Taylor, D. Morozov, N. R. Gemmell, K. Erotokritou, S. Miki, H. Terai, and R. H. Hadfield, "Photon counting LIDAR at 2.3 μm wavelength with superconducting nanowires," *Opt. Exp.*, vol. 27, pp. 38147–38158, 2019.

[55] R. Cheng, C. L. Zou, X. Guo, S. Wang, X. Han, and H. X. Tang, "Broadband on-chip single-photon spectrometer," *Nat. Commun.*, vol. 10, p. 4104, 2019.

[56] J. Zhang, N. Boiadjieva, G. Chulkova, H. Deslandes, G. N. Gol'tsman, A. Korneev, P. Kouminov, M. Leibowitz, W. Lo, R. Malinsky, O. Okunev, A. Pearlman, W. Slysz, K. Smirnov, C. Tsao, A. Verevkin, B. Voronov, B. Wilsher, and R. Sobolewski, "Noninvasive cmos circuit testing with nbn superconducting single-photon detectors," *Electron. Lett.*, vol. 39, p. 1086, 2003.

[57] A. J. Kerman, J. K. W. Yang, R. J. Molnar, E. A. Dauler, and K. K. Berggren, "Electrothermal feedback in superconducting nanowire single photon detectors," *Phys. Rev. B*, vol. 79, p. 100509, 2009.

[58] A. J. Miller, A. E. Lita, B. Calkins, I. Vayshenker, S. M. Gruber, and S. W. Nam, "Compact cryogenic self-aligning fiber-to-detector coupling with losses below one percent," *Opt. Exp.*, vol. 19, p. 9102–9110, 2011.

[59] K. Rosfjord, J. K. Yang, E. Dauler, A. Kerman, V. Anant, B. Voronov, G. N. Gol'tsman, and K. K. Berggren, "Nanowire single-photon detector with an integrated optical cavity and anti-reflection coating," *Opt. Exp.*, vol. 14, pp. 527–534, 2016.

[60] M. Caloz, M. Perrenoud, C. Autebert, B. Korzh, M. Weiss, C. Schönenberger, R. J. Warburton, and H. Zbinden, "High-detection efficiency and low-timing jitter with amorphous superconducting nanowire single-photon detectors," *Appl. Phys. Lett.*, vol. 112, p. 061103, 2018.

[61] M. Caloz, B. Korzh, E. Ramirez, C. Schönenberger, R. J. Warburton, H. Zbinden, M. D. Shaw, and F. Bussieres, "Intrinsically-limited timing jitter in molybdenum silicide superconducting nanowire single-photon detectors," *Appl. Phys. Lett.*, vol. 126, p. 164501, 2019.

[62] M. Caloz, "Superconducting nanowire single-photon detectors for quantum communication applications," *Geneva University - GAP Quantum Technologies*, 2019.

[63] A. Korneev, "Ultrafast and high quantum efficiency large-area superconducting single-photon detectors," *Proc. of SPIE*, vol. 6583, pp. 65830I–1, 2007.

[64] A. Divochiy, F. Marsili, D. Bitauld, A. Gaggero, R. Leoni, F. Mattioli, A. Korneev, V. Seleznev, N. Kaurova, O. Minaeva, G. Gol'tsman, K. G. Lagoudakis, M. Benkhaoul, F. Lévy, and A. Fiore, "Superconducting nanowire photon-number-resolving detector at telecommunication wavelengths," *Nat. Photonics*, vol. 2, pp. 302–306, 2008.

[65] M. Tarkhov, J. Claudon, J. P. Poizat, A. Korneev, A. Divochiy, O. Minaeva, V. Seleznev, N. Kaurova, B. Voronov, A. V. Semenov, and G. Gol'tsman, "Ultrafast reset time of superconducting single photon detectors," *Appl. Phys. Lett.*, vol. 92, p. 241112, 2008.

[66] M. Ejrnaes and R. Cristiano, "A cascade switching superconducting single photon detector," *Appl. Phys. Lett.*, vol. 91, p. 262509, 2007.

[67] S. Pellegrini, G. S. Buller, J. M. Smith, A. M. Wallace, and S. Cova, "Laser-based distance measurement using picosecond resolution time-correlated single-photon counting," *Opt. Exp.*, vol. 13, pp. 17301–17308, 2015.

[68] Nauerth, F. Moll, M. Rau, C. Fuchs, J. Horwath, S. Frick, and H. Weinfurte, "Air-to-ground quantum communication," *Nat. Photonics*, vol. 7, pp. 382–386, 2013.

[69] H. Li, L. Zhang, L. You, X. Yang, W. Zhang, X. Liu, S. Chen, Z. Wang, and X. Xie, "Large-sensitive-area superconducting nanowire single-photon detector at 850 nm with high detection efficiency," *Opt. Exp.*, vol. 23, pp. 17301–17308, 2015.

[70] C. L. Lv, H. Zhou, L. X. You, X. Y. Liu, Y. Wang, W. J. Zhang, S. J. Chen, Z. Wang, and X. M. Xie, "Large active area superconducting single-nanowire photon detector with a 100 μm diameter," *Supercond. Sci. Technol.*, vol. 30, p. 115018, 2017.

[71] J. Chang, I. E. Zadeh, J. W. N. Los, J. Zichi, A. Fognini, M. Gevers, S. Dorenbos, S. F. Pereira, P. Urbach, and V. Zwiller, "Multimode-fiber-coupled superconducting nanowire single-photon detectors with high detection efficiency and time resolution," *App. Opt.*, vol. 58, pp. 9803–9807, 2019.

[72] J. P. Allmaras, A. Beyer, F. Marsili, and M. D. Shaw, "Large-area 64-pixel array of WSi superconducting nanowire single photon detectors," in *Conference on Lasers and Electro-Optics (CLEO), San Jose, CA*, IEEE, 2017.

[73] C. Zhang, W. Zhang, J. Huang, L. You, H. Li, C. Lv, T. Sugihara, M. Watanabe, H. Zhou, Z. Wang, and X. Xie, "Nbn superconducting nanowire single-photon detector with an active area of 300 μm-in-diameter," *AIP Advances*, vol. 9, p. 075214, 2019.

[74] W. e. a. Zhang, "A 16-pixel interleaved superconducting nanowire single-photon detector array with a maximum count rate exceeding 1.5 ghz," *IEEE Transactions on Applied Superconductivity*, vol. 29, pp. 1–4, 2019.

[75] X. Yang, L. You, L. Zhang, C. Lv, H. Li, X. Liu, H. Zhou, and Z. Wang, "Comparison of superconducting nanowire single-photon detectors made of NbTiN and NbN thin films," *IEEE Transactions on Applied Superconductivity*, vol. 28, p. 1–6, 2018.

[76] D. Henrich, P. Reichensperger, M. Hofherr, J. M. Meckbach, K. Il'in, M. Siegel, A. Semenov, A. Zotova, and D. Yu. Vodolazov, "Geometry-induced reduction of the critical current in superconducting nanowires," *Phys. Rev. B*, vol. 86, p. 144504, 2012.

[77] C. Autebert, G. Gras, E. Amri, M. Perrenoud, M. Caloz, H. Zbinden, and F. Bussieres, "Direct measurement of the recovery time of superconducting nanowire single-photon detectors," *Preprint to be submitted to J. App. Phys.*, vol. List of "Preprint articles", 2020.

[78] M. Sidorova, A. Semenov, H. W. Hübers, I. Charaev, A. Kuzmin, S. Doerner, and M. Siegel, "Physical mechanisms of timing jitter in photon detection by current-carrying superconducting nanowires," *Phys. Rev. B*, vol. 96, p. 184504, 2017.

[79] D. Knuth, *Art of Computer Programming, Volume 2: Seminumerical Algorithms*. Addison-Wesley Professional, 2014.

[80] H. Zhun and C. Hong, "A truly random number generator based on thermal nois," in *Proceedings of the 4th International Conference on ASIC*, p. 862–86, 2001.

[81] B. Jun and P. Kocher, "The Intel® random number generator," *White Paper for Intel. C.*, 1992.

[82] A. Stefanov, N. Gisin, O. Guinnard, L. Guinnard, and H. Zbinden, "Optical Quantum Random Number Generator," *J. Mod. Opt.*, vol. 47, pp. 595–598, 2000.

[83] W. Wei and H. Guo, "Bias-free true random-number generator," *Opt. Lett.*, vol. 34, p. 1876, 2009.

[84] M. Marandi, N. C. Leindecker, K. L. Vodopyanov, and R. L. Byer, "All-Optical Quantum Random Bit Generation from Intrinsically Binary Phase of Parametric Oscillators," *Opt. Exp.*, vol. 20, pp. 19322–19330, 2012.

[85] C. Gabriel, C. Wittmann, D. Sych, R. Dong, M. Mauerer, U. L. Andersen, M. Marquardt, and G. Leuchs, "A generator for unique quantum random numbers based on vacuum states," *Nat. Ph.*, vol. 4, p. 711–715, 2010.

[86] E. Amri, Y. Felk, D. Stucki, J. Ma, and E. R. Fossum, "Quantum Random Number Generation using a Quanta Image Sensor," *Sensors*, vol. 16, p. 1002, 2016.

[87] J. Earman, *A Primer on Determinism*. Springer Science Business Media, 1986.

[88] M. Li and P. Vitányi, *An introduction to Kolmogorov complexity and its applications*. Springer, 2008.

[89] D. Frauchiger, R. Renner, and M. Troyer, "True randomness from realistic quantum devices," *arXiv*, vol. 1311.4547 [quant-ph], 2013.

[90] C. Shannon, *A Mathematical Theory of Communication*. Bell Syst. Tech. J., 1948.

[91] L. E. Bassham and al, "A Statistical Test suite for Random and Pseudorandom number Generators for Cryptographic Applications," *NIST SP*, vol. 800-22 Rev 1a, 2010.

[92] G. Marsaglia, *The Marsaglia random number CDROM: including the diehard battery of tests of randomness*. Florida State University, 1995.

[93] R. G. Brown, *Dieharder: A Random Number Test Suite*. Duke University, 2017.

[94] E. Fossum, "The quanta image sensor (qis): Concepts and challenges," in *OSA Topical Meeting on Computational Optical Sensing and Imaging, Toronto*, 2011.

[95] E. R. Fossum, "Modeling the performance of single-bit and multi-bit quanta image sensors.," *IEEE J. Elec. Dev. Soc.*, vol. 1, p. 166–174, 2013.

[96] S. Masoodian, A. Rao, J. Ma, K. Odame, and E. R. Fossum, "A 2.5pj/b Binary Image Sensor as a Pathfinder for Quanta Image Sensors," *IEEE Trans. Elec. Dev.*, vol. 63, p. 100–105, 2015.

[97] J. Ma and E. R. Fossum, "Quanta Image Sensor Jot with Sub 0.3e- r.m.s.Read Noise and Photon Counting Capability," *IEEE J. Elec. Dev. Lett.*, vol. 36, p. 926–928, 2015.

[98] J. Ma, D. Starkey, A. Rao, K. Odame, and E. Fossum, "Characterization of Quanta Image Sensor Pump-Gate Jots with Deep Sub-Electron Read Noise," *IEEE J. Elec. Dev. Soc.*, vol. 3, p. 472–480, 2015.

[99] E. Fossum, "Photon counting error rates in Single-bit and Multi-bit Quanta Image Sensors," *IEEE J. Elec. Dev. Soc.*, vol. 4, pp. 136–143, 2016.

[100] D. Wei and E. Fossum, "1/f noise modelling and characterization for Cmos Quanta Image Sensors," *Sensors*, vol. 19, p. 5459, 2019.

[101] J. Ma, S. Masoodian, D. A. Starkey, and E. Fossum, "Photon-number-resolving megapixel image sensor at room temperature without avalanche gain," *Optica*, vol. 4, p. 1474, 2017.

[102] J. Ma and E. Fossum, "Analytical modeling and TCAD simulation of a quanta image sensor jot device with a JFET source-follower for deep sub-electron read noise," *IEEE J. Elec. Dev. Soc.*, vol. 5, p. 69–78, 2017.

[103] S. Masoodian, J. Ma, and E. R. Fossum, "Gigajot technology," 2018.

[104] E. Amri, Y. Felk, D. Stucki, J. MA, and E. R. Fossum, "Quanta image sensor quantum random number generation," WO2017193106A1 Nov. 2017.

[105] IdQuantique, "Quantis QRNG Chip Brochure," 2019.

[106] B. Korzh, "High-performance single-photon detectors and applications in quantum communication," *Geneva University - GAP Quantum Technologies*, 2016.

[107] A. N. McCaughan and K. K. Berggren, "A superconducting-nanowire three-terminal electrothermal device," *Nano Lett.*, vol. 14, p. 5748–5753, 2014.

[108] D. Zhu, M. Colangelo, B. Korzh, Q. Y. Zhao, S. Frasca, A. Dane, A. Velasco, A. Beyer, J. Allmaras, E. Ramirez, W. Strickland, D. Santavicca, M. Shaw, and K. Berggren, "Superconducting nanowire single-photon detector with integrated impedance-matching taper," *Appl. Phys. Lett.*, vol. 114, p. 042601, 2019.

A. List of Papers and patent

Papers:

1. E. Amri, G. Boso, B. Korzh and H. Zbinden,"Temporal jitter in free-running InGaAs/InP single-photon avalanche detectors," *Opt. Lett.*, vol. 41, pp. 5728-5731, 2016.
2. E. Amri, Y. Felk, D. Stucki, J. Ma and E. R. Fossum, "Quantum Random Number Generation Using a Quanta Image Sensor," *Sensors*, vol. 16, p. 1002, 2016.
3. F. Regazzoni, E. Amri, S. Burri, D. Rusca, H. Zbinden and F. Regazzoni, "A Fully Integrated Quantum Random Number Generator with on-Chip Real-Time Randomness Extraction," *Pre-print to be submitted to Nature Electronics*, 2020.
4. M. Perrenoud, M. Caloz, E. Amri, C. Autebert, H. Zbinden and F. Bussieres, "High detection rate and high efficiency with parallel-SNSPDs," *Preprint to be submitted to App. Phys. Lett*, 2020.
5. C. Autebert, G. Gras, E. Amri, M. Perrenoud, M. Caloz, H. Zbinden and F. Bussieres, "Direct measurement of the recovery time of superconducting nanowire single-photon detectors," *Preprint to be submitted to J. App. Phys.*, 2020.

Application patent:

E. Amri, Y. Felk, D. Stucki, J. MA, and E. R. Fossum, "Quanta image sensor quantum random number generation," WO2017193106A1, Nov. 2017.

Peer-reviewed articles

P1. Temporal jitter in free-running InGaAs/InP single-photon avalanche detectors

Temporal jitter in free-running InGaAs/InP single-photon avalanche detectors

EMNA AMRI[1,2], GIANLUCA BOSO[1], BORIS KORZH[1], AND HUGO ZBINDEN[1,*]

[1] Group of Applied Physics (GAP), University of Geneva, Switzerland
[2] ID Quantique SA (IDQ), Switzerland
[*] Corresponding author: hugo.zbinden@unige.ch

Compiled October 7, 2016

Negative-feedback avalanche diodes (NFADs) provide a practical solution for different single-photon counting applications requiring free-running mode operation with low afterpulsing probability. Unfortunately, the time jitter has never been as good as for gated InGaAs/InP single-photon avalanche diodes (SPADs). Here we report on the time jitter characterization of InGaAs/InP based NFADs with a particular focus on the temperature dependence and the effect of carrier transport between the absorption and the multiplication regions. Values as low as 52 ps were obtained at an excess bias voltage of 3.5 V.

OCIS codes: *(040.5160) Photodetectors; (040.1345) Avalanche photodiodes; (030.5260) Photon counting.*

http://dx.doi.org/10.1364/ao.XX.XXXXXX

Single-photon detectors (SPDs) at telecom wavelengths play an important role in many applications. Among these are quantum key distribution (QKD) [1, 2], singlet-oxygen dosimetry for photodynamic therapy [3], photon counting optical communication [4], optical time domain reflectometry [5], eye-safe laser ranging [6, 7], testing of integrated circuits [8], biomedical imaging [9] and general quantum optics experiments. A popular detector choice for the near-infrared range (NIR, 1000 nm – 1700 nm) are InGaAs/InP single-photon avalanche diodes (SPADs), thanks to their ease-of-use (cryogenic temperature not required), compact size and competitive performance. It is often beneficial to operate these detectors in the free-running regime, especially when the tasks are asynchronous. The simplest solution for this is to implement a passive circuit [10], such as a series resistance which is able to quench the avalanche current following a detection event. So far, the most successful self-quenching NIR SPAD is known as the negative-feedback avalanche diode (NFAD) [11], which has an integrated monolithic thin-film resistor. Recently, it was demonstrated that NFADs can achieve extremely low dark-count rates (DCR) when cooled to temperatures below -100°C [12], which has pushed the limits of several applications [3, 13].

For a vast majority of SPD applications a critical characteristic is the temporal resolution, also know as timing jitter. Although this has been extensively studied in gated-mode SPADs [14], which has shown that these detectors can be a competitive alternative to superconductor-based devices [15], it has not been clear if free-running NFADs can achieve a jitter of $<$100 ps. Moreover, the nature of jitter at low temperatures has not been thoroughly studied in NFADs, which is crucial if operation with a low DCR is an additional requirement.

In this Letter, we present a characterization of the timing response for different NFADs with a particular focus on the temperature dependence. Moreover, we analyse the effect of carrier transport between the absorption and multiplication regions in these devices, which limits the minimum operating bias voltage for a given timing jitter. Subsequently this gives rise to the connection between the lowest achievable DCR and the temporal jitter, when operating at low temperatures.

Table 1. Characteristics of tested devices

Device	Code	Diameter (μm)	R_s (kΩ)	V_b at -130°C
#1	E2G2	25	500	64.6 V
#2	E3G7	30	1700	65.3 V
#3	E2I1	25	860	70.2 V
#4	E2I9	25	1150	71.6 V

We have characterized four different NFADs manufactured by Princeton Lightwave which have different feedback resistances and active areas (see Table 1). The NFADs were placed inside a dry chamber of a free-piston Stirling cooler (FPSC) (Twinbird SC-UE15R or SC-UD08) that enables cooling of the detectors down to -130°C. The readout circuit is described in Ref. [16]. To measure the timing jitter of the complete SPD system, we used an optical probe signal, generated by difference-frequency generation (DFG) in a nonlinear PPLN crystal that is pumped by a 3 ps mode-locked laser operating at a wavelength of 771 nm and a continuous wave laser operating at 1546 nm. The resulting optical pulse at 1538 nm has the same pulse duration and repetition rate (76 MHz) as the pulsed laser. This probe signal

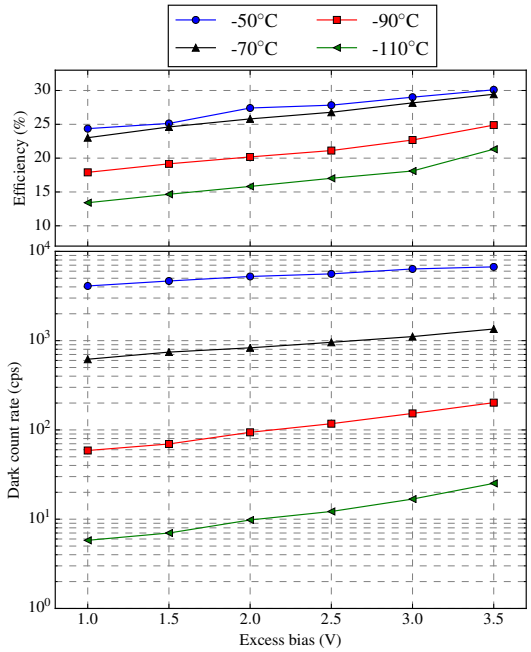


Fig. 1. Efficiency at 1550 nm and dark count rate versus excess bias voltage for different temperatures for diode #3.

was attenuated to the single photon level prior to arriving at the NFAD. A synchronization signal is generated by a fast photodiode illuminated by the pulsed laser. For the acquisition, we used a time-correlated single-photon counting (TCSPC) module (SPC-130 from Becker & Hickl), which generates a histogram of the delay between the NFAD detection and the synchronization signal. The instrument response function (IRF) of the measurement setup has a full-width-half-max (FWHM) of 7 ps (given by the contribution of 3 ps FWHM from the optical signal and 6.5 ps FWHM from the TCSPC card) which is negligible in comparison to the detector jitter.

Prior to discussing the temporal jitter behaviour, we present the efficiency (characterized at 1550 nm) and DCR results for a typical NFAD (#3 in this case), which motivates the reason for using these devices at low temperatures. Figure 1 shows the two characteristics as a function of excess bias (difference of the bias voltage and the breakdown voltage) for different temperatures between -50°C and -110°C. The measured efficiencies were between 12% and 30% and a DCR of 6 cps is achieved at the lowest temperature and excess bias. Note that we have presented the efficiency/DCR data for other NFAD devices in previous studies, for E2G2 (#1 here) in Ref. [12] and E3G7 (#2 here) in Ref. [3]. Of the three, the E2G2 device exhibited the lowest DCR at a temperature of -110°C, which was 1 cps at around 12% efficiency. It is believed that this difference is due to the fact that the breakdown voltage of device #1 is significantly lower than that of device #3 (see Table 1), meaning that the absolute electric field in the amplification region is lower. The main contributions to the DCR are the thermal-carrier generation in the absorption region and the field-dependent trap-assisted-tunnelling (TAT) in the amplification region, since these devices utilize a separate absorption and multiplication (SAM) structure [17]. Below -70°C the TAT becomes the dominant effect and, although it is not directly temperature dependent, it is reduced at lower temperatures due to the reduction of the breakdown voltage. As seen in Fig. 1, below -70°C, the detection efficiency starts to reduce for a given excess bias. This is due to a blue-shift in the spectral response [18] at lower temperatures, since 1550 nm lies at the edge of the detection spectrum. Therefore, if shorter wavelengths were of interest, there would be no reduction of efficiency. These results demonstrate that it is beneficial to operate NFADs at low temperatures, if the application calls for low DCR and does not require high count rates, since the required hold-off time increases [19]. We shall now analyse the temporal jitter dependence on the operating bias voltage, which will allow us to make a connection between the minimum possible DCR at a given jitter.

There are two dominant contributions to the timing jitter in InGaAs/InP SPADs. The first is attributed to the time distribution of the transit time of the photo-generated carriers (holes) from the absorption region to the multiplication region [17]. Due to the band gap difference between the two regions (InGaAs and InP, respectively), there exists a valance-band energy step which has to be overcome by the holes travelling to the multiplication region [20]. Such a barrier leads to charge pile-up which can be liberated through thermionic emission, giving rise to a temporal distribution with an exponential tail. With increasing electric field, the effective barrier is reduced, increasing the emission rate. In addition, NFAD structures are implemented with grading layers at the heterojunction, in order to decrease the slope of the energy step [11], meaning that the effective barrier reduces to zero at lower field-strengths. Nevertheless, there exists a given field-strength whereupon the barrier is non-zero and in the following we shall probe the onset of this effect.

Subsequently, upon the arrival of the hole in the multiplication region, a fundamental build-up time is needed for the avalanche amplitude to reach a predetermined threshold level, signalling the detection event. The temporal distribution of this process is Gaussian. Therefore, the system temporal jitter is expected to be a convolution of an exponential (thermionic emission during the hole transport) and Gaussian (impact ionization) distributions.

Figure 2(a) shows the temporal jitter histogram for a temperature of -130°C, for diode #1. It is indeed clear that there exist two distinct contributions, where at small delays the distribution is Gaussian, whilst at longer delays a clear exponential tail is visible. In order to isolate the two effects, we can exploit their temperature dependence. Both effects are field-dependent, however, the thermionic emission is dependent on the absolute bias voltage of the NFAD, whilst the impact ionization is dependent on the excess bias. As can be seen in Fig. 2(b), when the excess bias is kept constant at different temperatures, the histograms overlap at short delays, meaning the impact ionization process is unaffected. On the other hand, at longer delays, the exponential time-constant is changing significantly at different temperatures, which is due to varying absolute bias voltage caused by a temperature dependent breakdown voltage in the multiplication region (temperature coefficient is 0.134 V/K). Indeed, if the bias voltage is kept constant for the same temperature range, the contrary is true: the exponential tail is almost unchanged, whilst the Gaussian component changes, as can be seen in Fig. 2(c). The temperature dependence of the decay rate (from Fig. 2(c)) gives a measure of the energy barrier experienced by the holes [20], which is approximately 0.03 eV at 65.6 V and drops to near zero at around 67 V, leaving the impact ionization as the dominant effect.

To illustrate the overall effects of temperature and excess bias

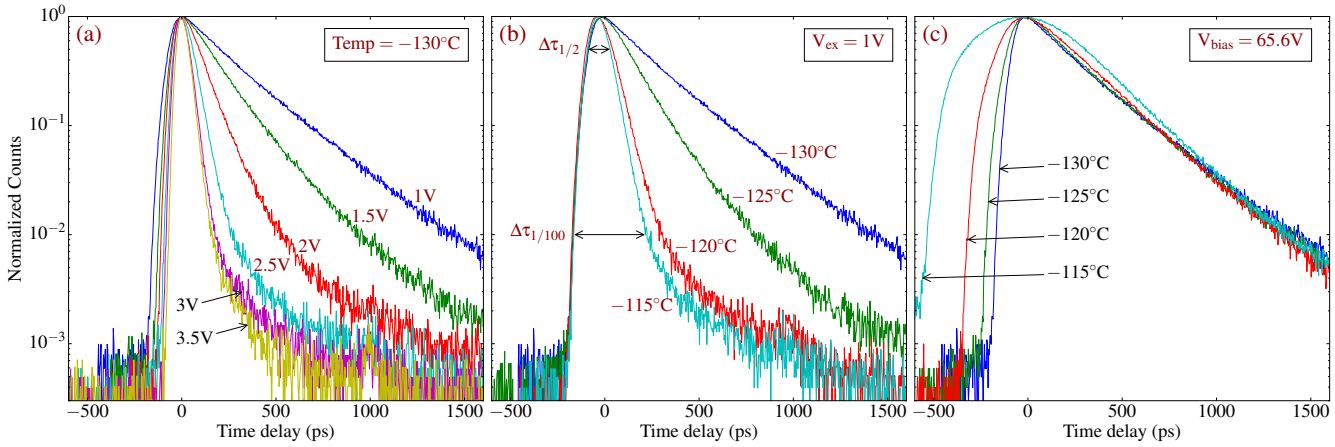


Fig. 2. Jitter histograms for NFAD #1 in different conditions. (a) Varying excess bias voltage at a constant temperature of -130°C. (b) A constant excess bias voltage of 1 V for different temperatures. (c) A constant bias voltage of 65.5 V for different temperatures.

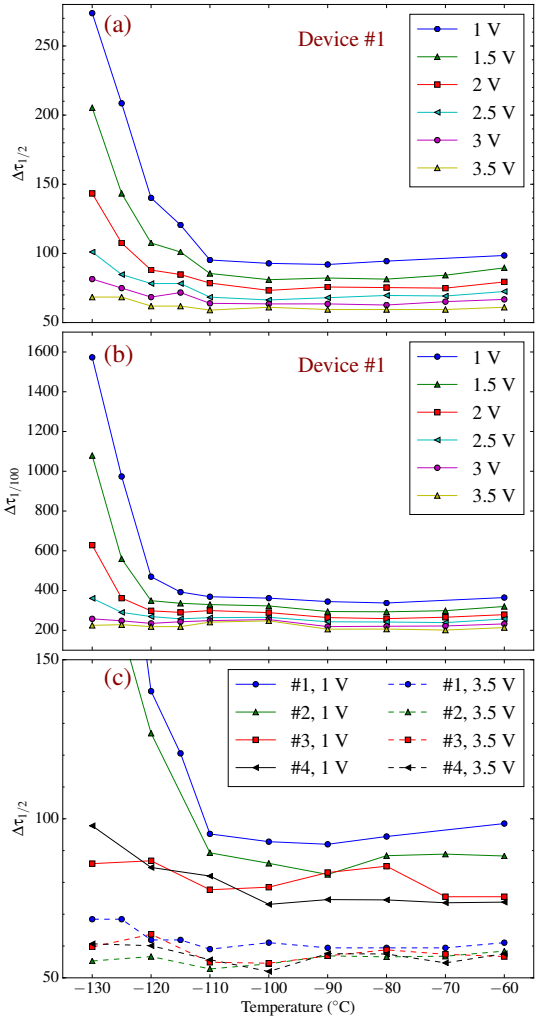


Fig. 3. NFADs time jitter versus temperature for different excess bias voltages. (a) FWHM ($\Delta\tau_{1/2}$) for device #1. (b) FW1/100M ($\Delta\tau_{1/100}$) for device #1. (c) Comparison of the FWHM jitter for four devices at 1 V and 3.5 V excess bias.

variation, we can plot the FWHM of the jitter histograms. Figure 3(a) shows the results obtained with diode #1 for different temperatures between -60°C and -130°C. For a constant temperature, the jitter decreases with increasing excess bias voltage as expected, due to the speed up of the impact ionization process [21]. For decreasing temperature and fixed excess bias, it reduces slightly (about 10%) in the range of -60°C to -100°C. This can be explained by the increase of ionization coefficients with lower temperature in the multiplication region [22], meaning the avalanche build-up process is yet again faster. At temperatures below about -110°C, for low excess bias voltages, one can see the increase of the jitter due to the significant hole trapping between the absorption and multiplication regions. This effect is clearly negated through the increase of the excess bias, which reduces the energy barrier, as discussed earlier. For many applications, such as QKD, the detection scheme requires a very high extinction ratio, therefore it is important to quantify the jitter width at a lower level than the half-max, especially when the jitter histogram shows non-gaussian behaviour. To analyze this, Fig. 3(b) shows the the full-width at 1/100 of the maximum (FW1/100M, $\Delta\tau_{1/100}$) for device #1, which shows similar temperature behaviour as the FWHM. At the highest excess bias, a $\Delta\tau_{1/100} = 200$ ps is achieved.

Since the hole-trapping phenomena is mainly dependent on the operating bias voltage, one strategy to reduce this effect is to use a diode with a higher breakdown voltage. Figure 3(c) shows the FWHM jitter for all four NFADs tested for the same range of temperatures and two excess bias voltages, 1 V and 3.5 V. At high excess bias, all the detectors have the same behaviour with the minimum jitter being between 52 ps and 67 ps.We also performed measurements at 4V of excess bias obtaining slightly lower timing jitter at the expense of a disproportionate increase of DCR and afterpulsing. However, for low excess bias voltage we see that two of the NFADs do not exhibit the sharp increase in the jitter at low temperatures (devices #3 and #4). This is due to the fact that these diodes have a breakdown voltage of around 5-6 V higher (see Table 1) than the diodes which do exhibit the low temperature jitter increase, hence the bias voltage remains sufficiently high in order to keep the energy barrier at the heterojunction below zero, preventing hole pile-up. Note that the design structure of all the devices was the same and the differences in breakdown voltage are most probably due to

run-to-run fabrication variations.

These results suggest that for optimum operation of the NFAD, the bias voltage should be sufficiently large in order to avoid the hole pile-up jitter effects, however, it should not be too high, in order to minimize TAT contribution to the DCR. In order to keep the thermally generated DCR well below the 1 cps level, the NFAD should be operated at -130°C. Hence, the optimum breakdown voltage would be around 67 V for this temperature, which would allow operation at any excess bias, without any hole pile-up effects.

In most applications the signal-to-noise ratio (SNR) is a crucial characteristic. In QKD the SNR defines the maximum transmission distance of the system. In order to maximize the signal, it is preferable to operate at the maximum possible clock rate, which is limited by the jitter of the detectors. This means the signal of a QKD system is proportional to $\eta/\Delta\tau_{1/100}$, where η is the detection efficiency and we consider the FW1/100M jitter in order to ensure low error rates. The noise is given by the DCR (r_{dc}) within the detection time window, hence the SNR = $\eta/r_{dc}\Delta\tau_{1/100}{}^2$. This shows that the timing jitter is the most important characteristic for a long distance QKD system. Given the jitter demonstrated in this work, QKD operation at 5 GHz and an increase of the maximum distance would be possible.

To conclude, we have demonstrated that free-running InGaAs/InP NFADs, which achieve efficient passive quenching, can operate with a temporal jitter as low as 52 ps, which puts them on par with the best gated-mode devices [23–26] and is only a factor of 2 larger than the record-holding superconducting detector [27]. We have also analysed the low-temperature performance of the NFAD jitter and this has enabled the understanding of the jitter contribution due to charge-carrier pile-up between the absorption and multiplication regions, a phenomena which has been rarely studied. Afterwards, we have shown that in order to avoid degradation of the temporal resolution due to this effect, the operating voltage of these devices should be greater than 67 V at the lowest operation temperature. Given this, excess bias voltage and temperature can be chosen freely according to the applications requirements.

1. FUNDING INFORMATION

This work has been sponsored by the the Swiss NCCR QSIT project, the European MSCA ITN PROMIS as well as the EMPIR 14IND05 MIQC2 project co-funded by the European Union's Horizon 2020 research and innovation program and the EMPIR Participating States. aswfgwh

2. ACKNOWLEDGMENT

The authors would like to thank M. Itzler and X. Jiang for their helpful discussions.

REFERENCES

1. N. Gisin, G. Ribordy, W. Tittel, and H. Zbinden, Rev. Mod. Phys. **74**, 145 (2002).
2. V. Scarani, H. Bechmann-Pasquinucci, N. J. Cerf, M. Dušek, N. Lütkenhaus, and M. Peev, Rev. Mod. Phys. **81**, 1301 (2009).
3. G. Boso, D. Ke, B. Korzh, J. Bouilloux, N. Lange, and H. Zbinden, Biomed. Opt. Express **7**, 211 (2016).
4. B. S. Robinson, A. J. Kerman, E. A. Dauler, R. J. Barron, D. O. Caplan, M. L. Stevens, J. J. Carney, S. A. Hamilton, J. K. Yang, and K. K. Berggren, Opt. Lett. **31**, 444 (2006).
5. P. Eraerds, M. Legré, J. Zhang, H. Zbinden, and N. Gisin, J. Lightwave Technol. **28**, 952 (2010).
6. R. E. Warburton, A. McCarthy, A. M. Wallace, S. Hernandez-Marin, R. H. Hadfield, S. W. Nam, and G. S. Buller, Opt. Lett. **32**, 2266 (2007).
7. A. McCarthy, R. J. Collins, N. J. Krichel, V. Fernández, A. M. Wallace, and G. S. Buller, Appl. Opt. **48**, 6241 (2009).
8. F. Stellari, A. Tosi, F. Zappa, and S. Cova, IEEE Trans. Instrum. Meas. **53**, 163 (2004).
9. I. Bargigia, A. Tosi, A. B. Shehata, A. Della Frera, A. Farina, A. Bassi, P. Taroni, A. Dalla Mora, F. Zappa, R. Cubeddu, and A. Pifferi, J. App. Spect. **66**, 944 (2012).
10. S. Cova, M. Ghioni, A. Lacaita, C. Samori, and F. Zappa, Appl. Opt. **35**, 1956 (1996).
11. M. A. Itzler, X. Jiang, B. Nyman, and K. Slomkowski, Proc. SPIE **7222**, 72221K (2009).
12. B. Korzh, N. Walenta, T. Lunghi, N. Gisin, and H. Zbinden, Appl. Phys. Lett. **104**, 081108 (2014).
13. B. Korzh, C. C. W. Lim, R. Houlmann, N. Gisin, M. J. Li, D. Nolan, B. Sanguinetti, R. Thew, and H. Zbinden (2014).
14. A. Tosi, A. D. Mora, F. Zappa, and S.Cova, J. Mod. Opt. **56**, 299 (2009).
15. M. D. Eisaman, J. Fan, A. Migdall, and S. V. Polyakov, Rev. Sci. Instrum. **82**, 071101 (2011).
16. T. Lunghi, C. Barreiro, O. Guinnard, R. Houlmann, X. Jiang, M. A. Itzler, and H. Zbinden, J. Mod. Opt. **59**, 1481 (2012).
17. M. A. Itzler, X. Jiang, M. Entwistle, K. Slomkowski, A. Tosi, F. Acerbi, F. Zappa, and S. Cova, J. Mod. Opt. **58**, 174 (2011).
18. F. Acerbi, M. Anti, A. Tosi, and F. Zappa, Photon. J. **5**, 6800209 (2013).
19. B. Korzh, T. Lunghi, K. Kuzmenko, G. Boso, and H. Zbinden, J. Mod. Opt. **62**, 1151 (2015).
20. S. R. Forrest, O. K. Kim, and R. G. Smith, Appl. Phys. Lett. **41**, 95 (1982).
21. C. Tan, J. Ng, G. Rees, and J. David, J. Sel. Topics Quantum Electron. **13**, 906 (2007).
22. F. Zappa, P. Lovati, and A. Lacaita, in "Indium Phosphide and Related Materials, 1996. IPRM '96., Eighth International Conference on," (1996), pp. 628–631.
23. A. Tosi, F. Acerbi, M. Anti, and F. Zappa, IEEE J. Quantum Elec. **48**, 1227 (2012).
24. N. Namekata, S. Adachi, and S. Inoue, Opt. Express **17**, 6275 (2009).
25. Y. Liang, E. Wu, X. Chen, M. Ren, Y. Jian, G. Wu, and H. Zeng, IEEE Photon. Technol. Lett. **23**, 887 (2011).
26. M. A. Itzler, r. Ben-Michael, C. F. Hsu, K. Slomkowski, A. Tosi, S. Cova, F. Zappa, and R. Ispasoiu, J. Mod. Opt. **54**, 283 (2007).
27. L. You, X. Yang, Y. He, W. Zhang, D. Liu, W. Zhang, L. Zhang, L. Zhang, X. Liu, S. Chen, Z. Wang, and X. Xie, J. AIP Adv. **3**, 072135 (2013).

P2. Quantum Random Number Generation Using a Quanta Image Sensor

sensors

MDPI

Article

Quantum Random Number Generation Using a Quanta Image Sensor

Emna Amri [1,*], Yacine Felk [1], Damien Stucki [1], Jiaju Ma [2] and Eric R. Fossum [2]

[1] ID Quantique SA, Ch. de la Marbrerie 3, 1227 Carouge, Switzerland; yacine.felk@idquantique.com (Y.F.); damien.stucki@idquantique.com (D.S.)

[2] Thayer Engineering School at Dartmouth College, Hanover, NH, USA; Jiaju.Ma.TH@dartmouth.edu (J.M.); eric.r.fossum@dartmouth.edu (E.R.F.)

* Correspondence: emna.amri@idquantique.com; Tel. +41-22-301-83-71; Fax: +41-22-301-83-79

Academic Editor: Albert Theuwissen

Received: 6 April 2016; Accepted: 23 June 2016; Published: 29 June 2016

Abstract: A new quantum random number generation method is proposed. The method is based on the randomness of the photon emission process and the single photon counting capability of the Quanta Image Sensor (QIS). It has the potential to generate high-quality random numbers with remarkable data output rate. In this paper, the principle of photon statistics and theory of entropy are discussed. Sample data were collected with QIS jot device, and its randomness quality was analyzed. The randomness assessment method and results are discussed.

Keywords: QRNG; random number generator; QIS; quanta image sensor; photon counting; jot; entropy; randomness

1. Introduction

The generation of high-quality random numbers is becoming more and more important for several applications such as cryptography, scientific calculations (Monte-Carlo numerical simulations) and gambling. With the expansion of computers' fields of use and the rapid development of electronic communication networks, the number of such applications has been growing quickly. Cryptography, for example, is one of the most demanding applications. It consists of algorithms and protocols that can be used to ensure the confidentiality, the authenticity and the integrity of communications and it requires true random numbers to generate the keys to be used for encoding. However, high-quality random numbers cannot be obtained with deterministic algorithms (pseudo random number generator); instead, we can rely on an actual physical process to generate numbers. The most reliable processes are quantum physical processes which are fundamentally random. In fact, the intrinsic randomness of subatomic particles' behavior at the quantum level is one of the few completely random processes in nature. By tying the outcome of a random number generator (RNG) to the random behavior of a quantum particle, it is possible to guarantee a truly unbiased and unpredictable system that we call a Quantum Random Number Generator (QRNG).

Several hardware solutions have been used for true random number generation, and some of them are exploiting randomness in photon emission process. This class of QRNG includes beam splitters and single-photon avalanche diodes (SPADs) [1–3], homodyne detection mechanisms [4,5] and conventional CMOS image sensors (CIS) [6]. Although it has been demonstrated that these devices produce data of a satisfactory randomness quality, more work needs to be done to enhance the generation process, especially on the improvement of output data rate and device scalability. Practically, in an RNG utilizing image sensors, the photon emission is not the only source of randomness, and some noise sources in the detector, such as dark current and $1/f$ noise, will act as extra randomness sources and reduce the randomness quality since they have a strong thermal dependency. Therefore,

an ideal detector should have high photon-counting accuracy with low read noise and low dark current to completely realize quantum-based randomness.

The Quanta Image Sensor (QIS) can be regarded as a possible solution to meet these goals because of its high-accuracy photon-counting capability, high output-data rate, small pixel-device size, and strong compatibility with the CIS fabrication process.

Proposed in 2005 as a "digital film sensor" [7], QIS can consist of over one billion pixels. Each pixel in QIS is called a "jot". A jot may have sub-micron pitch, and is specialized for photon-counting capability. A QIS with hundreds of millions of jots will work at high speed, e.g., 1000 fps, with extremely low power consumption, e.g., 2.5 pJ/bit [8]. In each frame, each jot counts incident photons and outputs single-bit or multi-bit digital signal reflecting the number of photoelectrons [9]. The realization of QIS concept relies on the photon-counting capability of a jot device. As photons are quantized particles in nature, the signal generated by photons is also naturally quantized. However, with the presence of noise in the read out electronics, the quantization effect is weakened or eliminated. To realize photon-counting capability, deep sub-electron read noise (DSERN) is a prerequisite, which refers to read noise less than 0.5 e− r.m.s. But, high-accuracy photon-counting requires read noise of 0.15 e− r.m.s. or lower [10,11].

The pump-gate (PG) jot device designed by the Dartmouth group achieved 0.22 e− r.m.s. read noise with single correlated double sampling (CDS) read out at room temperature [12,13]. The low read noise of PG jot devices was fulfilled with improvements in conversion gain (CG) [14], and the photoelectron counting capability was demonstrated with quantization effects in the photon counting histogram (PCH) [15].

2. Randomness Generation Concept

To quantify the randomness in a sequence of bits, we refer to the concept of entropy, first introduced by Shannon [16]. Entropy measures the uncertainty associated with a random variable and is expressed in bits. For instance, a fair coin toss has an entropy of 1 bit, as the exact outcome—head or tail—cannot be predicted. If the coin is unfair, the uncertainty is lower and so is the entropy. And when tossing a two-headed coin, there is no uncertainty which leads to 0 bit of entropy.

To compute the value of the entropy, we need to have full information about the random number generation process. In a photon source, the photon emission process obeys the principle of Poisson statistics [10], and the probability $P[k]$ of k photoelectron arrivals in a QIS jot is given by:

$$P[k] = \frac{e^{-H}H^k}{k!} \tag{1}$$

where the quanta exposure H is defined as the average number of photoelectrons collected in each jot per frame. So under the illumination of a stable light source, randomness exists in the number of photoelectrons arriving in each frame.

During readout, the photoelectron signal from the jot is both converted to a voltage signal through the conversion gain (V/e−) and corrupted by noise. Let the readout signal U be normalized by the conversion gain and thus measured in electrons. The readout signal probability distribution function (PDF) becomes a convolution of the Poisson distribution for quanta exposure H and a normal distribution with read noise u_n (e− r.m.s.). The result is a sum of constituent PDF components, one for each possible value of k and weighted by the Poisson probability for that k [11]:

$$P[U] = \sum_{k=0}^{\infty} \frac{1}{\sqrt{2\pi u_n^2}} \exp\left[-\frac{(U-k)^2}{2u_n^2}\right] \cdot \frac{e^{-H}H^k}{k!} \tag{2}$$

An example of a Poisson distribution corrupted with read noise is shown in Figure 1. While in practice the photodetector may be sensitive to multiple photoelectrons, subsequent circuitry can be used to discriminate the output to two binary states (either a "0" meaning no photoelectron, or a

"1" meaning at least one photoelectron) by setting a threshold U_t between 0 and 1, typically 0.5 and comparing U to this threshold. From a stability perspective, it is better to choose the threshold U_t at a valley of the readout signal PDF, such as at a 0.50 e− when H = 0.7, so that small fluctuations in light intensity have minimal impact on the value of entropy. The probability of the "0" state is given by:

$$P\left[U < U_t\right] = \sum_{k=0}^{\infty} \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{U_t - k}{u_n\sqrt{2}}\right)\right] \cdot \frac{e^{-H}H^k}{k!} \tag{3}$$

and the probability of the "1" state is just:

$$P\left[U \geqslant U_t\right] = 1 - P[U < U_t] \tag{4}$$

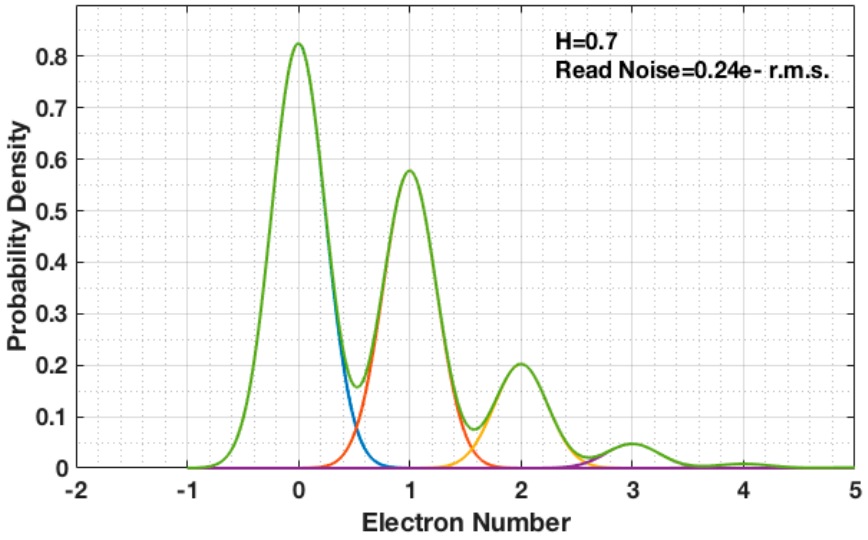


Figure 1. Readout signal probability distribution function (PDF) from Poisson distribution corrupted with read noise. Quanta exposure H = 0.7 and read noise u_n = 0.24 e− r.m.s.

The minimum quantum entropy of this distribution is given by [6]:

$$S_{min} = -\log_2[\max(P\left[U \geqslant U_t\right], P[U < U_t])] \tag{5}$$

If the measured value U will be encoded over b bits, the quantum entropy per bit of output will be, on average, equal to:

$$\overline{S} = \frac{S_{min}}{b} < 1 \tag{6}$$

where b = 1 for the single-bit QIS. It is, therefore, optimal to choose a quanta exposure H such that $P\left[U < U_t\right] = P\left[U \geqslant U_t\right] = 0.5$. These two conditions of stability and entropy lead to a preferred quanta exposure $H \cong 0.7$. An example of the cumulative probability function for the readout signal for $H = 0.7$ is shown in Figure 2. It should be noted that other combinations of H and U_t such as $H = 2.67$ and U_t = 2.5 e− are also viable options. For read noise u_n above 1 e− r.m.s., where the photon-counting peaks of Figure 1 are fully "blurred" by noise (e.g., conventional CMOS image sensors), the optimum settings of U_t and H converge so that the resultant Gaussian readout signal PDF is split in half at the peak, as one might deduce intuitively.

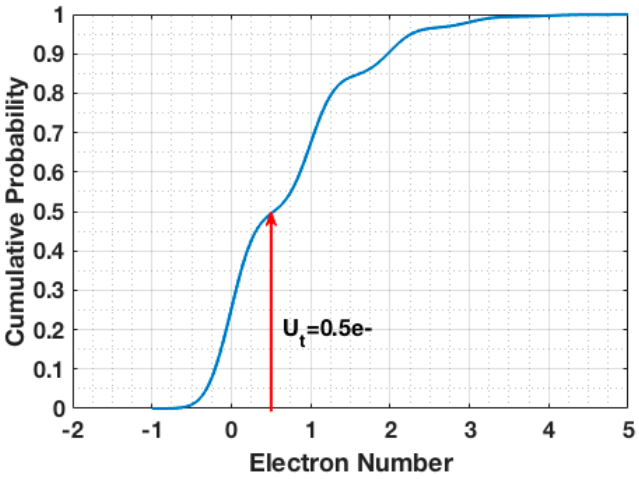


Figure 2. Cumulative probability of readout signal with read noise u_n = 0.24 e− r.m.s. and quanta exposure H = 0.7.

Stability is illustrated by comparing two cases with different quanta exposures and respective thresholds: H = 0.7 and H = 1.2. As shown in Figure 3, the thresholds for each case were selected to maximize the binary data entropy: U_t = 0.5 is located at a valley of PCH for H = 0.7, and U_t = 1 is located at a peak of PCH for H = 1.2. With 2% variation of quanta exposure in both cases, the output data of H = 0.7 showed better stability in entropy.

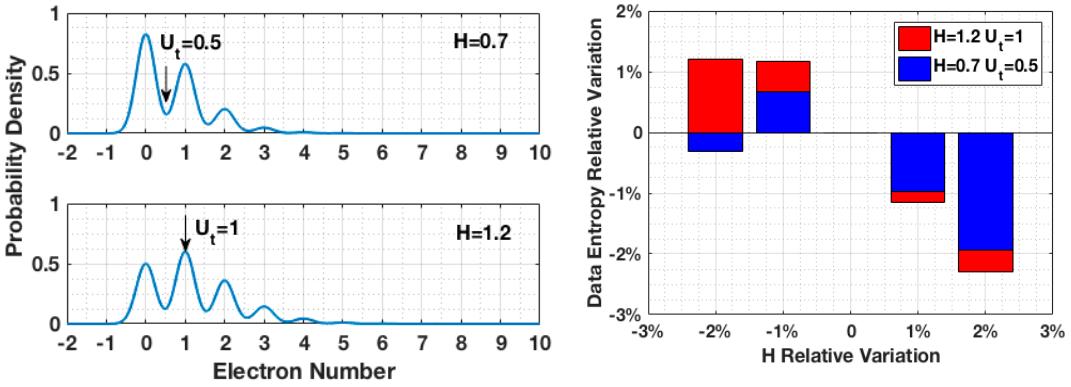


Figure 3. Binary data entropy variation caused by quanta exposure fluctuation during data collection.

It should be noted that only perfectly random bits will have unity quantum entropy, otherwise an extractor is required. A randomness extractor is a mathematical tool used to post-process an imperfect sequence of random bits (with an entropy less than 1) into a compressed but more random sequence. The quality of a randomness extractor is defined by the probability that the output deviates from a perfectly uniform bit string. This probability can be made arbitrarily very small by increasing the compression factor. The value of this factor depends on the entropy of the raw sequence and the targeted deviation probability and must be adjusted accordingly.

In this paper, we used a non-deterministic randomness extractor based on Universal-2 hash functions [17]. This extractor computes a number q of high-entropy output bits from a number $n > q$ of lower-entropy (raw) input bits. This is done by performing a vector-matrix multiplication between the vector formed by the raw bit values and a random n x q matrix M generated using multiple entropy sources. The compression ratio is thus equal to the number of lines divided by the number of columns of M. After extraction, statistical tests are run in order to make sure that randomness specifications are fulfilled.

3. Data Collection

The feasibility of applying the QIS to the QRNG application was tested with PG jot devices. In the PG jot test chip, an analog readout approach is adopted. The output signal from 32 columns is selected by a multiplexer and then amplified by a switch-capacitor programmable gain amplifier (PGA) with a gain of 24. The output signal from the PGA is sent off-chip and digitized through a digital CDS implemented with an off-chip 14-bit ADC. A complete description of readout electronics can be found in [13]. A 3 × 3 array of green LEDs was used as light source, located in front of the test chip. The distance from the light source to the sensor was 2 cm, and the intensity of the light source was controlled by a precision voltage source. During the data collection, a single jot with 0.24 e− r.m.s. read noise was selected and read out repeatedly, and a 14-bit raw digital output was collected. Under the limitation of the readout electronics on this test chip, the single jot was readout at a speed of 10 ksample/s. The testing environment was calibrated with 20,000 testing samples, and the quanta exposure H was obtained using the PCH method. In order to improve the randomness entropy of the data, the threshold U_t was determined as the median of the testing samples and then used with later samples to generate binary random numbers. The experimental PCH created by 200,000,000 samples is shown in Figure 4, which shows quanta exposure H of 0.7, and a read noise of 0.24 e− r.m.s. The threshold was set to 27.5DN, or 0.5 e−. The binary random numbers generated by first 10,000 samples are shown in Figure 5.

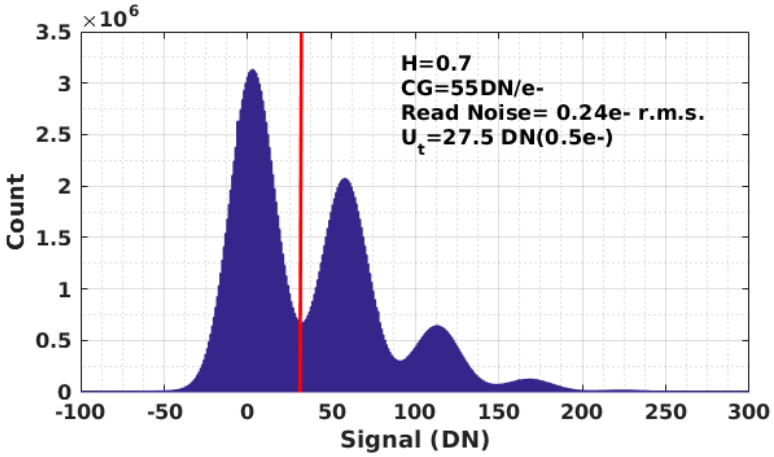


Figure 4. Photon counting histogram (PCH) of the first 200,000,000 samples.

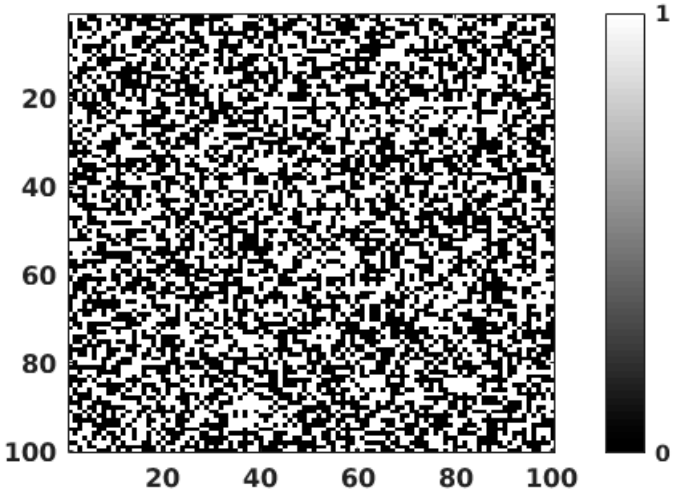


Figure 5. The binary output of the first 10,000 samples.

Although the light source was controlled by a stable voltage source, there was still a small fluctuation inferred in the light intensity. As shown in Figure 6, the quanta exposure H of 200 datasets is depicted, in which each dataset contains 1,000,000 samples and H is determined for each data set using its PCH. During the data collection, about 2.1% variation in quanta exposure was observed. To minimize the impact of light source fluctuation, the testing environment was calibrated to have an average quanta exposure H close to 0.7, for which the threshold U_t is located at a valley between two quantized peaks in the PCH.

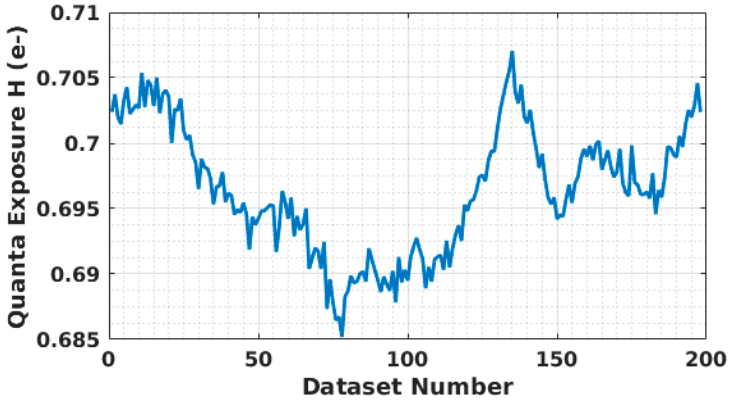


Figure 6. Quanta exposure fluctuation during data collection. Each dataset contains 1,000,000 samples.

4. Results

For a first test, we collected 500 Mbyte of raw random numbers by reading the jot at 5 ksamples/s (200 h of data collection). Using Equation (5), we were able to compute a minimum quantum entropy per output bit equal to 0.9845 for $H = 0.7$ and $u_n = 0.24$ e− r.m.s. Then we used the obtained value in the formula of the probability that the extractor output will deviate from a perfectly uniform q-bit string:

$$\varepsilon_{hash} = 2^{-(\overline{S}n-m)/2} \tag{7}$$

where n is the number of raw bits and m the number of extracted random bits.

Since a value of $\varepsilon_{hash} = 0$ is generally unachievable, we try to keep ε_{hash} below 2^{-100} implying that even using millions of jots one will not see any deviation from perfect uniform randomness in a time longer than the age of the universe. This gave a compression factor for $n = 1024$ equal to 1.23 which corresponds to losing only 18% of the input raw bits.

After extraction, we perform NIST tests [18] on the obtained random bits. This set of statistical tests evaluate inter alia, the proportion of 0 s and 1 s in the entire sequence, the presence of periodic or non-periodic patterns and the possibility of compression without loss of information. The QIS-based QRNG passed all these tests.

5. Comparison with Other Technologies

The idea of using an optical detector for random number generation is not new and has been driven by the intrinsic quantum nature of light. Single Photon Avalanche Diode (SPAD) arrays illuminated by a photon source and operating in Geiger mode have been widely used for this purpose [19,20]. Besides the single photon detection capability and technology maturity, SPAD matrices offer high-quality random data and can be fabricated in standard CMOS manufacturing line. However, these SPAD sensors require high supply voltage (22–27 V) for biasing above breakdown, suffer from after-pulsing phenomena, and have lower throughput per unit area than other optical detectors because of larger pixel size (600 Mbits/s for a matrix size of 2.5 mm^2 [19] and 200 Mbits/s for a matrix size of 3.2 mm^2 [20]).

Another technology exploiting optical quantum process has been recently introduced by the University of Geneva [6] and it consists of extracting random numbers of a quantum origin from an illuminated CIS. This low-power technology is more compatible with consumer and portable electronics since cameras are currently integrated in many common devices. Unfortunately, conventional image sensors are not capable of single-photon detection and provide lower randomness quality [6], which requires higher compression factor and hence lower output data rate. The choice of using QIS for random number generation was driven by the results obtained with SPADs and CIS since we noticed that QIS covers the advantages of both technologies (best tradeoff between data rate and scalability, single photon detection and CMOS manufacturing line) while providing solutions for most of their problems (speed, dark count rate, detection efficiency). Table 1 summarizes the comparison of the three techniques performances under the assumption of being used as RNGs. Note that the generation processes are different which limits the comparison points.

Table 1. The three technologies main comparison points.

Criteria	QIS	CIS	SPADs Matrix
Data Rate [1]	5–12 Gb/s	0.3–1 Gb/s	0.1–0.6 Gb/s
Read Noise	<0.25 e− r.m.s.	>1 e− r.m.s.	<0.15 e− r.m.s.
Dark Current/Count Rate [2]	0.1 e−/(jot·s)	10–500 e−/(pix·s)	200 counts/(pix·s)
Power Supply	2.5/3.3 V	2.5/3.3/5 V	22–27 V
Single Photon Counting	YES	NO	YES

[1] For a device with 2.5 mm^2 area size; [2] We define Dark Current for QIS/CIS and Dark Count Rate for SPADs, these values are measured at room temperature.

6. Summary

A new quantum random number generation method based on the QIS is proposed. Taking advantage of the randomness in photon emission and the photon counting capability of the Quanta Image Sensor, it shows promising advantages over previous QRNG technologies. Testing data was collected with QIS pump-gate jot device, and the randomness quality was assessed. Both randomness assessment method and data collection process are discussed, and the results show good randomness quality.

Acknowledgments: ID Quantique work has been sponsored by the Swiss State Secretariat for Education, Research, and Innovation (SERI) grants received for IDQ participation to European Marie Skłodowska-Curie Actions (MSCA), Innovative Training Network (ITN), Postgraduate Research on Dilute Metamorphic Nanostructures and Metamaterials in Semiconductor Photonics (PROMIS) and Eurostars project Quantum Random Number Generator (QRANGER). The QIS project at Dartmouth is sponsored by Rambus Inc. (Sunnyvale, CA, USA).

Author Contributions: Emna Amri and Damien Stucki co-conceived the random number data assessment experiments; Yacine Felk provided data for comparing technologies; Emna Amri performed the randomness experiments on the QIS data and co-wrote the paper; Jiaju Ma and Eric R. Fossum co-conceived, co-designed and performed the data collection experiments and co-wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Stefanov, A.; Gisin, N.; Guinnard, O.; Guinnard, L.; Zbinden, H. Optical quantum random number generator. *J. Modern Opt.* **2000**, *47*, 595–598. [CrossRef]
2. Dultz, W.; Hidlebrandt, E. Optical Random-Number Generator Based on Single-Photon Statistics at the Optical Beam Splitter. U.S. Patent No. 6,393,448, 21 May 2002.
3. Wei, W.; Guo, H. Bias-Free true random-number generator. *Opt. Lett.* **2009**, *34*, 1876–1878. [CrossRef] [PubMed]
4. Gabriel, C.; Wittmann, C.; Sych, D.; Dong, R.; Mauerer, W.; Andersen, U.L.; Marquardt, C.; Leuchs, G. A generator for unique quantum random numbers based on vacuum states. *Nat. Photonics* **2010**, *4*, 711–715. [CrossRef]

5. Shen, Y.; Tian, L.A.; Zou, H.X. Practical quantum random number generator based on measuring the shot noise of vacuum states. *Phys. Rev.* **2010**, *61*. [CrossRef]
6. Sanguinetti, B.; Martin, A.; Zbinden, H.; Gisin, N. Quantum random number generation on a mobile phone. *Phys. Rev.* **2014**, *4*. [CrossRef]
7. Fossum, E.R. The quanta image sensor (QIS): Concepts and challenges. In Proceedings of the 2011 Optical Society of America Topical Meeting on Computational Optical Sensing and Imaging, Toronto, ON, Canada, 10–14 July 2011.
8. Masoodian, S.; Rao, A.; Ma, J.; Odame, K.; Fossum, E.R. A 2.5 pJ/b binary image sensor as a pathfinder for quanta image sensors. *IEEE Trans. Electron. Devices* **2015**, *63*, 100–105. [CrossRef]
9. Fossum, E.R. Modeling the performance of single-bit and multi-bit quanta image sensors. *IEEE J. Electron. Devices Soc.* **2013**, *1*, 166–174. [CrossRef]
10. Fossum, E.R. Application of photon statistics to the quanta image sensor. In Proceedings of the International Image Sensor Workshop (IISW), Snowbird Resort, UT, USA, 12–16 June 2013.
11. Fossum, E.R. Photon counting error rates in single-bit and multi-bit quanta image sensors. *IEEE J. Electron. Devices Soc.* **2016**. [CrossRef]
12. Ma, J.; Fossum, E.R. Quanta image sensor jot with sub 0.3 e− r.m.s. read noise. *IEEE Electron. Device Lett.* **2015**, *36*, 926–928. [CrossRef]
13. Ma, J.; Starkey, D.; Rao, A.; Odame, K.; Fossum, E.R. Characterization of quanta image sensor pump-gate jots with deep sub-electron read noise. *IEEE J. Electron. Devices Soc.* **2015**, *3*, 472–480. [CrossRef]
14. Ma, J.; Fossum, E.R. A pump-gate jot device with high conversion gain for a Quanta Image Sensor. *IEEE J. Electron. Devices Soc.* **2015**, *3*, 73–77. [CrossRef]
15. Starkey, D.; Fossum, E.R. Determining conversion gain and read noise using a photon-counting histogram method for deep sub-electron read noise image sensors. *IEEE J. Electron. Devices Soc.* **2016**. [CrossRef]
16. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *3*, 379–423. [CrossRef]
17. Troyer, M.; Renner, R. A Randomness Extractor for the Quantis Device, ID Quantique. Available online: http://www.idquantique.com/wordpress/wp-content/uploads/quantis-rndextract-techpaper.pdf (accessed on 27 June 2016).
18. Rukhin, A.; Soto, J.; Nechvatal, J.; Smid, M.; Barker, E. A Statistical Rest Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. National Institute of Standards and Technology (NIST), Special Pub. 800-22, 15 May 2001. Available online: http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA393366 (accessed on 27 June 2016).
19. Stucki, D.; Burri, S.; Charbon, E.; Chunnilall, C.; Meneghetti, A.; Regazzoni, F. Towards a high-speed quantum random number generator. *Proc. SPIE* **2013**, *8899*. [CrossRef]
20. Tisa, S.; Villa, F.; Giudice, A.; Simmerle, G.; Zappa, F. High-Speed quantum random number generation using CMOS photon counting detectors. *IEEE J. Sel. Top. Quant. Electron.* **2015**, *21*. [CrossRef]

Preprint articles

P3. A Fully Integrated Quantum Random Number Generator with on-Chip Real-Time Randomness Extraction

A Fully Integrated Quantum Random Number Generator with on-Chip Real-Time Randomness Extraction

Francesco Regazzoni[1], Emna Amri[2,3], Samuel Burri[4], Davide Rusca[2], Hugo Zbinden[2] and Edoardo Charbon[4]

[1]*ALaRI, Università della Svizzera italiana, Lugano (Switzerland)*

[2]*Group of Applied Physics (GAP), University of Geneva, Geneva (Switzerland)*

[3]*Id Quantique SA (IDQ), Geneva (Switzerland)*

[4]*École Polytechnique Fédérale de Lausanne, Lausanne (Switzerland)*

With the explosive growth of mobile devices, security has become a serious concern. And it is security that will dominate the next technological milestone, the pervasive diffusion of electronic devices connected to form the Internet of Things (IoT). In fact, such pervasive diffusion will only be possible with robust security protocols at device level. Random Number Generators (RNGs) are the fundamental primitive in most secure protocols, but often the weakest one. Establishing security in billions of devices will require RNGs with sufficiently high throughput but also an unprecedented level of integration to remove, in real time, potential biases in the entropy source. However, current RNGs are often very sensitive, and their output can be significantly altered by neighboring circuitry. We present the first fully integrated Quantum RNG (QRNG) in a standard CMOS technology node. The QRNG is based on a parallel array of independent Single-Photon Avalanche Diodes (SPADs), homogeneously illuminated by a DC-biased LED, and co-integrated logic for random number extraction. We

describe the randomness generation process for different operating conditions and we analyze the parasitic effects of such logic on the quality of the random number generation. Our CMOS QRNG can reach up to 400 Mbit/s with high integration level and low power consumption. Thanks to the use of standard CMOS technology and a modular architecture, our technology is shown to be highly scalable.

With the ubiquity and density of mobile devices, security has become a very serious concern and, at the same time, a key enabler for next generation IoT and cyber-physical systems. True random number generators (TRNGs) are the core building block in almost all security schemes today[1,2]. A typical TRNG comprises at least a source of entropy and a digital interface to generate usable bit sequences. Several implementations have been proposed, while the specific requirements are depending on target applications. Numerous works focus on extremely high-speed architectures [3], some on using optical entropy sources [4], others on extracting entropy from existing devices such as FPGAs [5] or general-purpose processors [6]. We argue that the large majority of applications do not require an extremely high-speed random number generator but a physically-secure, compact, integrated architecture, capable of guaranteeing the required throughput in real time [2].

In this paper, we present a fully integrated Quantum Random Number Generator (QRNG) with on-chip real-time randomness extraction. The QRNG is based on a parallel array of independent CMOS SPADs, homogeneously illuminated by a DC-biased LED. Digital post-processing is implemented on chip; it generates a true random bitstream through two parallel digital interfaces. We discuss the randomness generation process and the quantum entropy extraction, we propose a

model to ensure that the produced randomness is indeed originated from a quantum phenomenon and we evaluate the influence that the co-integrated digital post-processing may have on the raw randomness quality. We perform statistical analysis on final data after randomness extraction. Our CMOS QRNG can output up to 400 Mbit/s with significant area reduction and low energy per generated bit.

QRNG with Integrated Entropy Extraction Architecture

Our system is depicted in Figure 1; it features a QRNG IC implementing the SPADs matrix, the two extractors, and a high-speed digital interface. The entropy source is obtained from a LED illuminating an array of 16,634 single-photon avalanche diodes (SPADs), which convert photons chaotically distributed in space and Poissionian in time, onto a randomized 2D constellation of binary digits '1' if at least a photon is detected and '0' otherwise. The chip features two extractors based on vector-matrix multiplication to generate high entropy bit sequences starting from the original constellation. The two extractors target different application requirements. The first is a fixed extractor, denominated 'A'; it is realized from standard logic gates using a hard-coded n x k matrix pre-generated from an independent QRNG; this extractor is intended for applications requiring higher throughput and minimal area. The second is a variable extractor, denominated 'B'; it is a matrix of memory elements, realized by means of scan chain registers to minimize the amount of pads needed for programming it and to allow the use of design tools; this extractor is intended for applications requiring one to extract an ad hoc matrix in the field. The QRNG IC was fabricated using a technology where the SPADs exhibit low noise, negligible afterpulsing, and

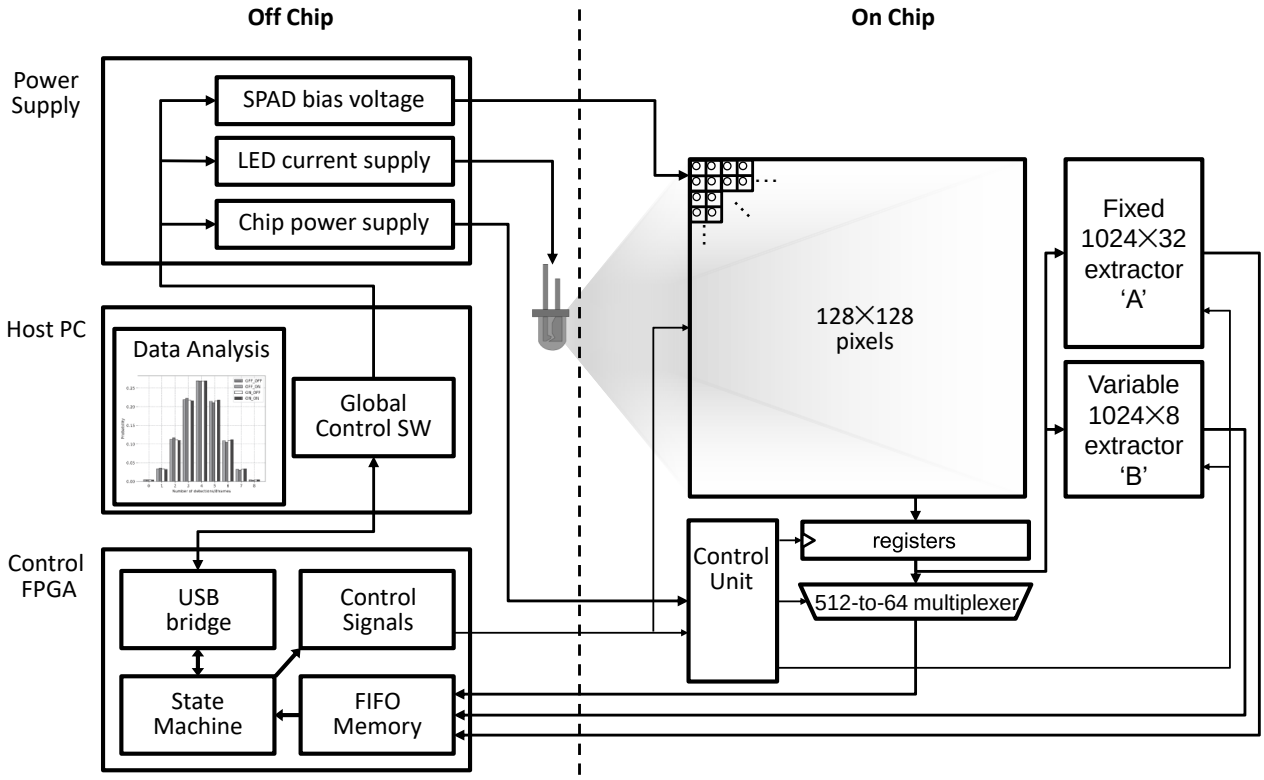


Figure 1: Detailed overview of the QRNG system use for our experiments. The QRNG system used for the experiments presented in the paper comprises a LED source, the QRNG chip realized using photo sensitive material, and a custom made FPGA used to control the QRNG chip and to format and transmit the bit stream to the host PC where the sequence of data are analyzed.

low crosstalk. These characteristics are essential to enable the maximum speed with the desired randomness quality.

The block diagram of the proposed QRNG is shown in Figure 2. The chip comprises an array of SPADs organized in a matrix of 128x128 pixels, whose schematic is shown in the inset of the figure. In each pixel a SPAD, implemented as a P+/N-well junction with lightly doped N wells as guard rings, is passively quenched and recharged by means of T1. T2 is used to trigger the static embedded all-NMOS memory (T3, T4, T5, T6), while T7 is used as reset and T8 and T9 are used to read out the content of the memory in a random access fashion. The matrix is read out four-rows-in-parallel every 40ns and the raw data is stored in a 512-bit register whose content is diverted to the two extractors; the entire matrix is read out in 1.3μs.

Extractor 'A' comprises 1024x32 XOR reduction cells fed by the register through an intermediate 1024-bit register. Extractor 'B' is an array of 8x1024 SRAM cells whose content is provided externally. Both extractors are operating within the chip in real time: extractor 'A' can reach a throughput of 400Mbit/s, extractor 'B' of 100Mbit/s, both producing random words at 12.5MHz. The resulting bit streams are read out from the chip in 8-bit and 32-bit packets, using two parallel interfaces. In test mode, we can independently access the raw entropy source output at 1.6Gbit/s to verify the quality of raw data at the source in accordance with high security standards. The SPADs are characterized in Figure 3. The figure shows a plot of the breakdown voltage distribution and of the dark count rate (DCR) across the entire population of pixels. The afterpulsing probability (APP) was characterized by means of the autocorrelation of several pixels in time based on

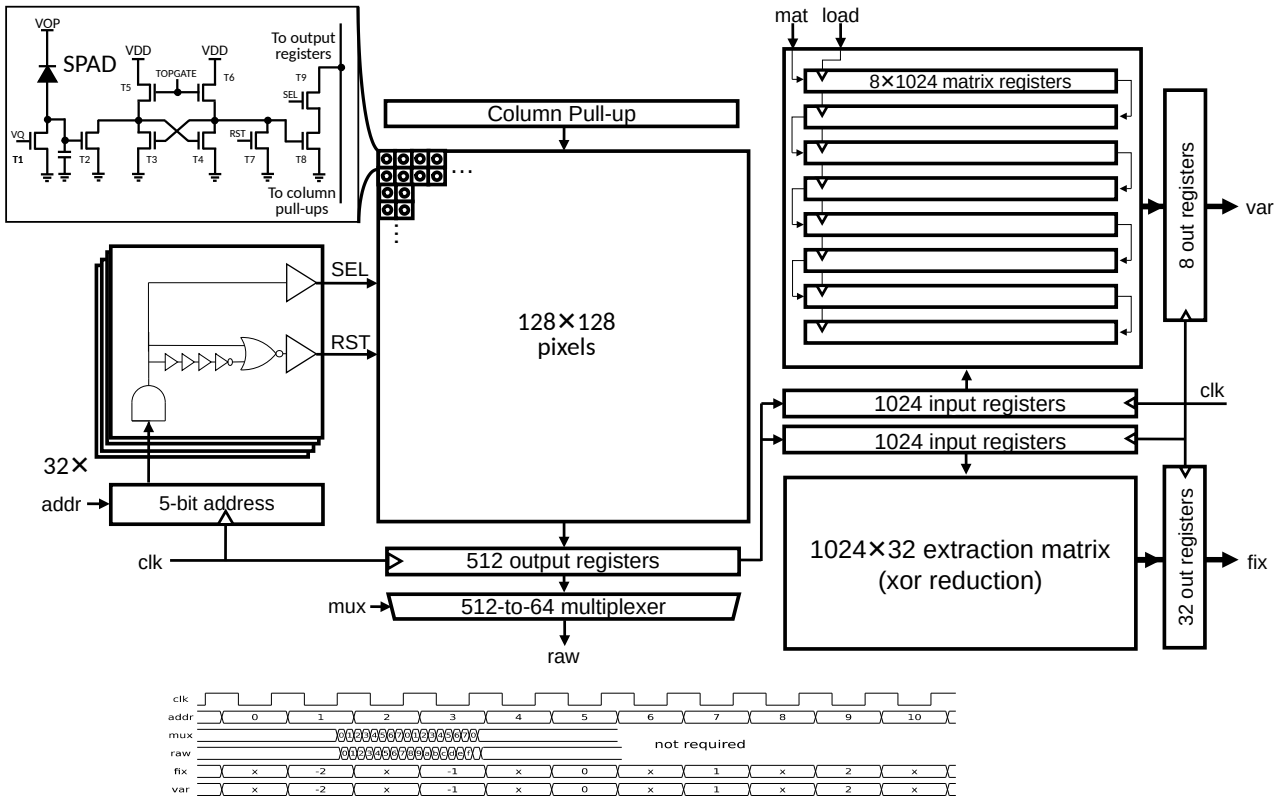


Figure 2: Block and timing diagram of the proposed QRNG chip. The chip comprises an array of 16,384 independent SPADs illuminated through a diffuser made of a standard 4μm polyimide layer. The SPADs act as entropy sources that generate a fast bit stream, which is then transformed onto a fully randomized 400Mbit/s binary stream via an extractor integrated on chip. The timing diagram of the readout process is shown in the inset. .

their digital output. Crosstalk was measured as the cross-correlation of a central pixel with all the surrounding of pixels. The plot in the figure shows central pixel (7,7) within a surrounding 11x11 pixel matrix. The excess bias voltage and dead time used in this chip were 2.1V and 1.3μs, respectively, so as to keep APP and crosstalk below 0.1%, as indicated in the table of Figure 3.

Effects of Integration on Entropy

The main goal of the chip was to conclusively show the negligible influence of on-chip digital post-processing on the quality of generated randomness, thus rejection the common thought of co-integration adding noise and heating effect and demonstrating the feasibility of a fully integrated QRNG in a commercially relevant CMOS technology. To this end, we carried out extensive statistical analysis of the random sequence before and after extraction under various conditions of operation and at a wide range of temperatures.

We first started by comparing the statistical distribution of raw data (before post-processing) when the two extractors are not powered on, when only one of them is working and when both are performing on-the-fly extraction. Figure 4 shows the constellations generated by the SPAD array for these same conditions. From the figure the equi-probability of ‘1’s and ‘0’s is apparent (and actually we computed a probability of 50.36% of ’1’ vs 49.64% of ’0’) with a homogeneous distribution among the pixels, irrespective of the state of the extractors, at better than 2σ deviation from the 50% distribution mark. This observation was confirmed when we compared the detection histograms (8-bits hamming weight distributions) in all four cases (shown in Figure‘5.

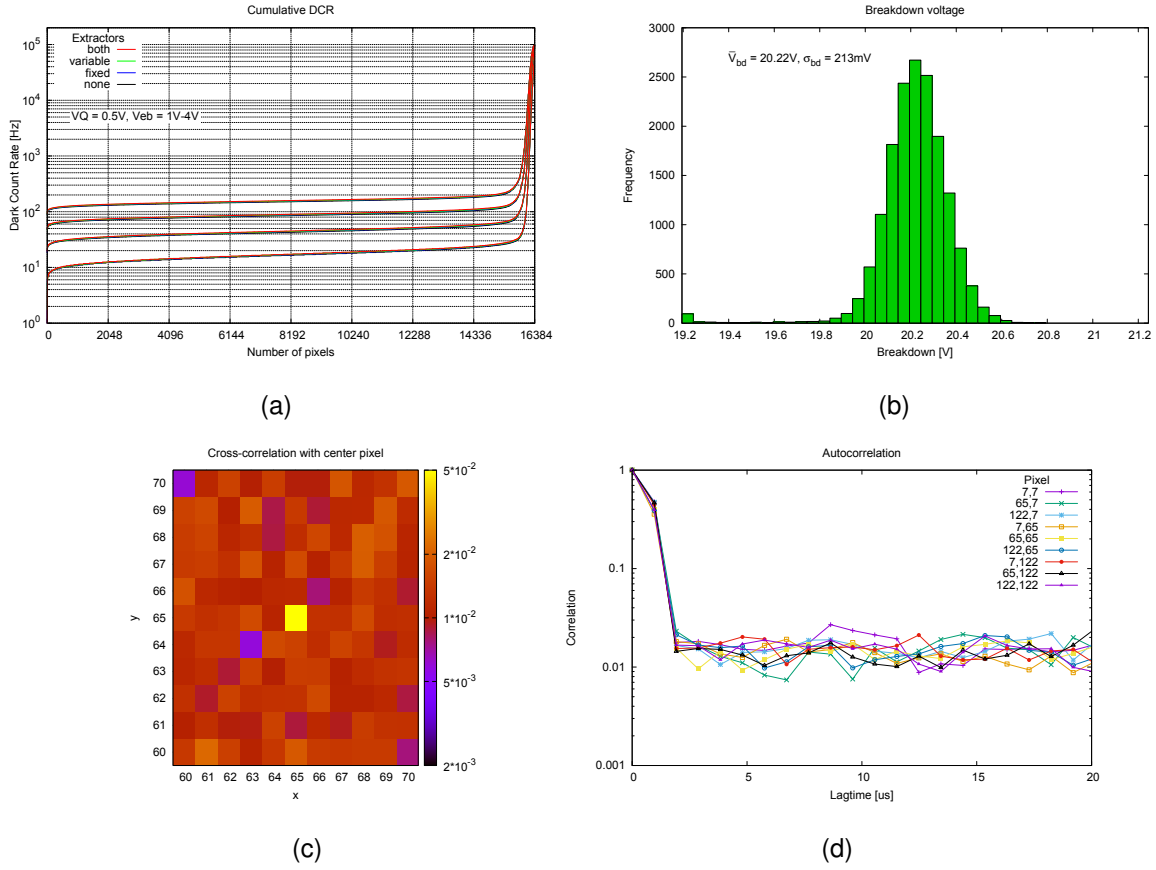


Figure 3: Performance of the SPAD used in this work. Clockwise from top-left: (a) dark count rate frequency vs. excess bias under four operation conditions; (b) breakdown voltage distribution; (d) autocorrelation function for several pixels vs. lag time; (c) cross-correlation function with respect to surrounding pixels. All measurements are at room temperature.

Therefore, we can conclusively confirm that the analog-digital co-integration doesn't have any quality-deterioration effect on the quantum process generation and measurement.

The effect of temperature on the QRNG was also evaluated and we saw a slight decrease of entropy/bit when increasing temperature. This decrease is due to a higher thermally induced noise that can lead to a bias toward '1'. Before extraction, the entropy/bit was relatively low due to errors coming from correlations between pixels signals and constant values given by dead pixels, that's why entropy extraction was compulsory to increase and maintain the mean quantum entropy. Afterwards, about 1 Gbit of extracted data were tested using the NIST statistical test suite [7] and the DIEHARD test battery [8] and it passed all of them.

Our chip achieved an extremely low energy per bit compared to the ones reported on standard CMOS technology, while the overall power at full speed was less than 499mW for the nominal 3.3V supply voltage at room temperature, making the chip suitable for low-power applications. The full integration of the entropy source and the extractors enabled us to achieve the highest level of integration to date with negligible impact on the quality of the random sequence. The micrograph of the QRNG chip is shown in Figure 6; the chip measures $10\times4.5\text{mm}^2$ in this CMOS technology, while more advanced nodes would enable significant area reductions, provided similar noise, AP, and crosstalk performance in SPADs.

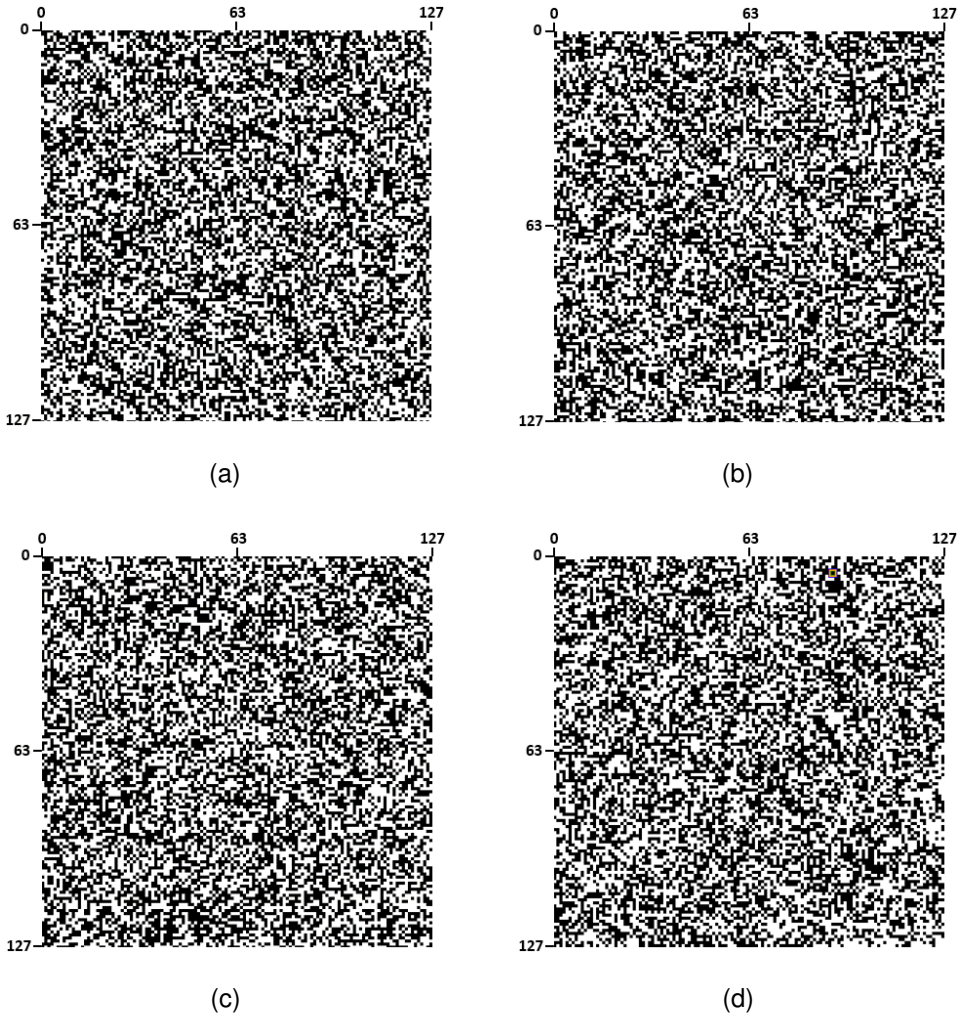


Figure 4: Performance of the QRNG when different extractors are active. Raw binary constellations. (a) Both extractors are OFF; (b) the fixed extractor is ON and the variable extractor is OFF; (c) the fixed extractor is OFF and the variable extractor is ON; (d) both extractors are ON.

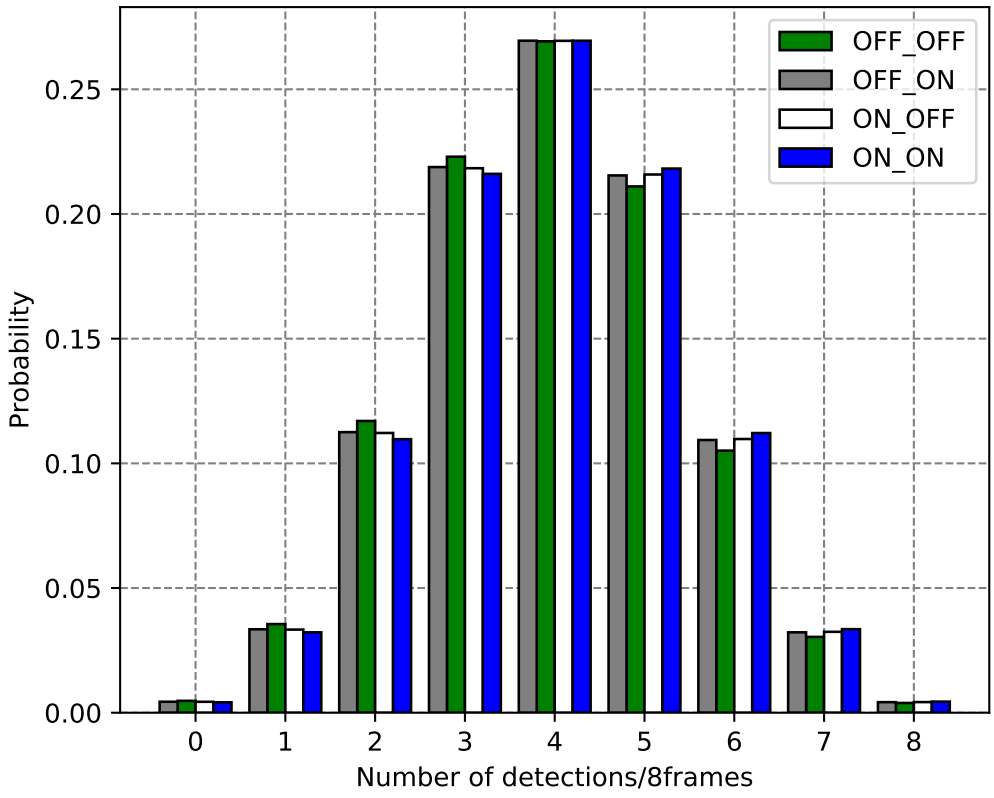


Figure 5: Detection histogram. Detection histogram (8-bits Hamming Weight distributions) of raw data for different operating conditions: (green) Both extractors are OFF; (grey) the fixed extractor is OFF and the variable extractor is ON; (white) the fixed extractor is ON and the variable extractor is OFF; (blue) both extractors are ON.

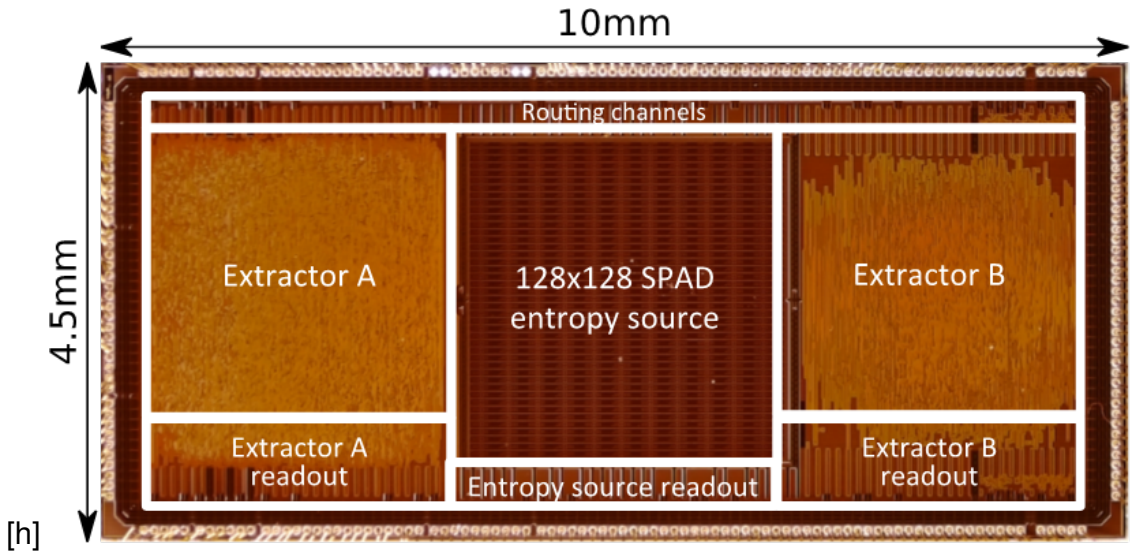


Figure 6: Chip micrograph. Complete chip micrograph of the proposed QRNG chip implemented in standard 0.35μm CMOS technology. From left, it is possible to see the Extractor A and its read out at the bottom, the entropy source composed by a matrix of 128$\times$128 SPADs and its readout at the bottom, the Extractor B and its readout at the bottom

Modeling the Quantum Randomness

Randomness has been previously defined in several ways. For Philosophers, randomness is a property of any event that happens by chance [9]. As such, it is not always possible to generate it or measure it. In information theory, a sequence of bits is called random if its Kolmogorov complexity is maximal [10]. However, this asymptotic definition does not include the unpredictability of the bit sequence, that is instead a fundamental property for many applications, including cryptography. When it comes to entropy source used for the generation of random numbers, randomness should be defined as the a property of the physical process whose the outcome is uniformly distributed and independent of all information available in advance [11]. Quantum Random Number Generators offer a sound solution for high-quality randomness generation, since they are compliant with this definition. The main characteristic of a QRNG is the use of quantum phenomena as entropy source, and, as direct consequence, the use of quantum mechanics to prove the randomness of the generated bits. Chaotic systems, even when not robust with respect to their initial conditions, can be completely described by a deterministic model. Quantum mechanic, instead, is intrinsically probabilistic, and the measurement results of quantum processes can not be predicted, even by a malicious third party. This characteristic makes QRNGs extremely suitable for cryptography applications. However, we need to ensure, by modeling the entropy source, that the produced randomness is due to a quantum process (and not coming , for instance, from classical side noise). To do so, we propose a model to demonstrate that, under given assumptions of the possible knowledge of an adversary, the randomness produced by our QRNG is indeed quantum.

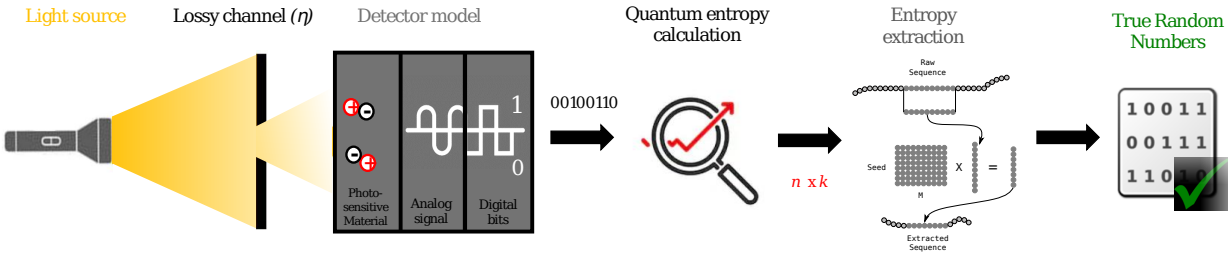


Figure 7: High level view of the proposed QRNG and its operational principle. Schematic illustration of the QRNG proposed in this paper, which consists of a light source that illuminates a matrix of photo-sensitive pixels. Once at least a photon collapses on a detector, an analog signal is generated and then converted into a digital value and a stream of data is produced. Raw data, that are directly accessible for testing purpose, are immediately forwarded into an entropy extractor, also implemented on chip, that produces the sequence of true random numbers.

In the QRNG that we propose here, the entropy comes from an LED that illuminates a matrix of SPADs. The quantum process that is exploited is the distribution of photons over the surface of the detectors. Figure 7 The position of each photon is not deterministically known before its detection. The probability distribution of its position is given by the intensity profile of the electromagnetic field impinging on the detectors. The distance between the detectors and the source is chosen such that the light intensity profile on the detectors is uniform. Considering the time interval defined by the acquisition time and the space mode spanned by the area of the detectors matrix, the quantum state emitted by the LED is a mixed state in the Fock space, where the probability of having n photons is given by the Poisson distribution P_N:

$$P_N(n) = e^{-\lambda}\frac{\lambda^n}{n!} \tag{1}$$

We model the matrix of SPADs as a collection of independent detectors. Their efficiency η is considered in our model, as the probability that one photon is detected. If n photons arrive on one detector, the probability that at least one of them is detected (probability that the detector clicks) is given by $P_{det}^n = 1 - (1-\eta)^n$. Similarly, we can model the dark counts probability with a random variable $S_i = 0, 1$ for each detector. In the case of $S_i = 1$ the i^{th} detector will click independently of the light coming into it. This event has probability p_{dark} equal to the probability of having a dark count on one pixel.

Using the methods proposed by Frauchiger *et al.* [11] to evaluate the amount of quantum ran-

domness of our device we need to evaluate the min-Entropy of the output distribution conditioned on the distribution of all the predictable side information. In our case, the side information is modeled as classical and determined by the random variable E. The conditional min-Entropy that we need to compute is thus the following:

$$H_{min}(X|E) = -\sum_{e} P_E(e) \log_2[\max_x P_{X|E=e}(x|e)] \tag{2}$$

where X represents the random variable of the output sequence, $P_{X|E}$ is the conditional probability distribution of X knowing the variable E, and the quantity $2^{-H_{min}(X|E)}$ represents the maximum guessing probability of X given E. In order to find the probability distribution of the output sequence we have to characterize the quantum state of our device ρ, and the measurement represented by the operators $\Pi^{\overline{x}}$, where $\overline{x}$ is a possible output bit string. Moreover, all side information is represented by the value e which is model after the result of the measurement E^e. Given the physical characterization of the device the probability distribution is given by the Born rule:

$$P_{\overline{X},E}(\overline{x}, e) = Tr[\Pi^{\overline{x}} E^e \rho (E^e)^\dagger] \tag{3}$$

without any need for a stochastic model (see Supplementary material for more information). The most critical part of a QRNG is however defining all the possible side information. In the device here analyzed the photon distribution of the source and the random variables corresponding to the dark counts are the principal source of classical side information and are considered as completely foreseeable. The event of a photon to be detected (described by the efficiency η) is considered to

be independent of any possible observer. However, in the model, we consider a possible shot to shot uncertainty with respect to the average observed value of the efficiency.

In the device here studied all the possible sources of side information have been characterized. The following parameters have been measured accurately in order to certify that the extracted randomness is of quantum nature: $\eta = 0.12 \pm 0.03$ (considering a possible deviation of 25% from the average efficiency), $p_{dark} = 8.45 \cdot 10^{-5}$ per detection window, to this probability we added also the probability of cross-talk of around $P_cross = 0.001$ (this corresponds to the worst case scenario where each click provoked by the click of another detector is known by an adversary), the mean photon number λ of photon arriving on the detectors chosen in such a way that the probability of each of them to click is equal to 0.5 (as set experimentally). With this characterization of the device the min-Entropy per bit has been estimated to be $H_{min}(X|E)/m \approx 77\%$ which is a lower bound on the possible extractable randomness from the raw generated sequence of bits.

In practical implementation of QRNG, the quantum process used to generate entropy cannot be perfectly created or perfectly measured. In fact, there are always hardware imperfections and different noise sources that cannot be controlled. These two elements can affect the raw output of the RNG introducing a bias or a pattern. These effects can be removed using randomness extraction function, such as the hash function based on vector-matrix multiplication [12] used in this work. This extractor is applied to vectors of n raw bits and output shorter vectors of k random bits. The (n x k) elements composing the extraction matrix are constant and they are generated by an independent RNG. Once the length of the raw string n is chosen, the parameter k will be bound

by the probability that the output bit string deviates from perfectly random output bits ϵ. If the extraction function is taken from a two-universal family of hash functions, it is possible to quantify this failure probability by the Leftover Hash Lemma with side information:

$$\epsilon = 2^{-(H_{min}(X|E)\cdot n - k)/2} \tag{4}$$

For instance with the obtained value of min-Entropy and for a vector of 1024 raw bits, the parameter k should be lower than 588 for efficient entropy extraction. In our case the parameter k is equal to 32 for the fixed extractor and equal to 8 for the variable extractor which confirms the high efficiency of our on-chip extraction.

Conclusions

We have presented the first fully integrated Quantum random number generator (QRNG) in a standard CMOS technology node. RNGs are a fundamental primitive in most secure protocols and the ones based on quantum technology will enable the high levels of security required by the emerging applications. To demonstrate the possibility to integrate advanced functionalities with the entropy sources, we developed a modular architecture comprising an array of independently operating SPADs, illuminated by a DC-biased homogeneous LED, and logic for random number extraction. The fabricated QRNG reaches a throughput of 400 Mbit/s with an extremely low energy per bit. We characterized the randomness generation process for different operating conditions and we analyzed the potential effects extraction logic on the quality of the random number generation.

Methods

Fabrication of the Integrated QRNG The QRNG integrated circuit was fabricated in a standard 0.35μm 1P4M high-voltage CMOS technology through Europractice multi-project wafer foundry service. The SPAD device junction is formed between a p+ anode and deep n-well cathode embedded in the standard p-type substrate. A p-well guard ring is used to realize a uniform electric field and prevent premature edge-breakdown of the device. The SPAD is of circular shape with an active diameter of 6μm.

Measurements For measurements the IC was bonded in a ceramic PGA-256 package and inserted in a socket on a adapter PCB for connection with the FPGA main board. The FPGA main board contains a Xilinx Spartan 6 FPGA connecting directly to the QRNG I/O and a Cypress FX3 USB transceiver for communication with a control computer. Voltage supplies for the chip are provided through the main board with separate jumpers for the extractor supply.

A Rohde&Schwarz HMP2030 dc power supply was used to provide controllable operating bias and quenching voltage for the SPAD and pixel circuit. A custom-built matrix of relay switches (Kemet EC2-12TNU) and a USB-SMU (Agilent/Keysight U2723A) were used to respectively switch power of the extractor matrices and control the LED illumination current. For measurements in a temperature-controlled environment the QRNG together with the FPGA board was put in an ESPEC SU-262 temperature control chamber. The temperature chamber is outfitted with a nitrogen purge to avoid condensation.

To evaluate the breakdown voltage of the individual detectors the voltage was sweeped from

a voltage where no events are detected to where all detectors show activity in steps of 50mV. Through readout of individual detectors, the number of events during one second at each voltage step has been recorded and fitted with a two-piece linear function. The function evaluates to zero below the breakdown voltage and increases thereafter, thus providing the breakdown voltage for each detector.

Data from the extraction matrices is acquired through the control FPGA in the same way used to acquire the raw data needed to evaluate the breakdown voltage. The operating conditions, namely excess bias and quenching voltage and LED illumination current are set. Then synchronous control signals are applied to read out the data from the output register of the matrices. The control FPGA relays the data to a computer where they are stored on disk for analysis.

Randomness Evaluation The random data sequences were stored off-chip and tested using the NIST Statistical Test Suite (SP 800-22) downloaded from the NIST website (`https://csrc.nist.gov/Projects/Random-Bit-Generation/Documentation-and-Software`). The parameters used to perform these tests are the followings: blockFrequencyBlockLength = 12800, nonOverlappingTemplateBlockLength = 10, overlappingTemplateBlockLength = 9, approximateEntropyBlockLength = 10, serialBlockLength = 16 and linearComplexitySequenceLength = 1000. The results are stored in an dedicated file, automatically generated when the tests are launched. The same data sequences were tested using the Diehard battery of tests available `https://webhome.phy.duke.edu/~rgb/General/dieharder.php`

Data Availability The data that support the findings of this study described in the this paper are available from the authors upon reasonable request.

1. Sunar, B. True random number generators for cryptography. In *Cryptographic Engineering*, 55–73 (Springer, 2009).

2. Verbauwhede, I., Balasch, J., Roy, S. S. & Van Herrewege, A. 24.1 circuit challenges from cryptography. In *2015 IEEE International Solid-State Circuits Conference-(ISSCC) Digest of Technical Papers*, 1–2 (IEEE, 2015).

3. Wei, W., Xie, G., Dang, A. & Guo, H. High-speed and bias-free optical random number generator. *IEEE Photonics Technology Letters* **24**, 437–439 (2011).

4. Massari, N. *et al.* 16.3 a 16× 16 pixels spad-based 128-mb/s quantum random number generator with- 74db light rejection ratio and- 6.7 ppm/ °Cbias sensitivity on temperature. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, 292–293 (IEEE, 2016).

5. Rožić, V., Yang, B., Dehaene, W. & Verbauwhede, I. Iterating von neumann's post-processing under hardware constraints. In *2016 IEEE International Symposium on Hardware Oriented Security and Trust (HOST)*, 37–42 (IEEE, 2016).

6. Mechalas, J. Intel digital random number generator (drng): Software implementation guide revision 2.0 (2014).

7. Bassham III, L. E. *et al. Sp 800-22 rev. 1a. a statistical test suite for random and pseudo-random number generators for cryptographic applications* (National Institute of Standards & Technology, 2010).

8. Marsaglia, G. The marsaglia random number cdrom including the diehard battery of tests of randomness, 1995. *URL https://web.archive.org/web/20160125103112/http://www.stat.fsu.edu/pub/diehard* (2008).

9. Earman, J. *A Primer on Determinism* (Springer Science & Business Media, 1986).

10. Li, M. & Vitányi, P. *An introduction to Kolmogorov complexity and its applications* (Springer, 2008).

11. Frauchiger, D., Renner, R. & Troyer, M. True randomness from realistic quantum devices. *arXiv* **1311.4547 [quant-ph]**, 164–168 (2013).

12. Troyer, M. & Renner, R. Id quantique technical report. *Phys. Rev. X* **4**, 031056 (2012).

Authors Contributions Francesco Regazzoni devised and coordinated the project, conceived the idea of integration, planned the chip architecture, designed the digital part of the chip, contributed to the final chip integration and discussed the results of the experiments. Emna Amri evaluated the generated entropy under different operation condition (temperature, bias voltage, distance between the light source and the detector), determined the optimum operation condition for maximum entropy, did statistical analysis on raw data to evaluate the influence of integration, and conducted the statistical tests on data after extraction. Samuel Burri designed the analog part of the chip, carried out the final integration of the chip used in the experiments, designed the main FPGA board and the software for collecting the data, and carried out the measurements.

Davide Rusca provided the theoretical model for quantum entropy quantification. Edoardo Charbon and Hugo Zbinden supervised the work. All the authors contributed to the writing of the manuscript.

Acknowledgements The work presented in this paper has been partially supported by the Eurostar framework (Progect Q-RANGER, Quantum - RAndom Number GEneRator)

Competing Interests The authors declare that they have no competing financial interests.

Correspondence Correspondence and requests for materials should be addressed to Francesco Regazzoni (email: regazzoni@alari.ch).

Supplementary Material 1 : Results of Statistical tests

1 NIST SP 800-22 tests

The tests were performed using 1000 sequences of 1 Mbit each.

The minimum pass rate for each statistical test with the exception of th erandom excursion (variant) test is approximately = 980 for a sample size = 1000 binary sequences.

The minimum pass rate for the random excursion (variant) test is approximately = 599 for a sample size = 613 binary sequences.

The tests results for data generated by the fixed and the variable extractor are shown in Table 1 and Table 2 respectively.

2 Diehard battery of tests

Diehard tests were performed on two data sets of 500 Mbits generated by the fixed extractor (P-Value 'A') and the variable extractor (P-value 'B'). The results are presented in Table 3.

Table 1: NIST tests results: The uniformity of P-values and the proportion of passing sequences for 1000 samples of 1 Mbit each generated by the fixed extractor 'A'

Statistical test	P-Value	Proportion
Frequency	0.99577	988/1000
Block-Frequency	0.239266	990/1000
CumulativeSums1	0.941144	989/1000
CumulativeSums2	0.062427	986/1000
Runs	0.928857	989/1000
LongestRun	0.473064	993/1000
Rank	0.301194	990/1000
FFT	0.029205	983/1000
NonOverlappingTemplate	0.662091	993/1000
OverlappingTemplate	0.735908	988/1000
Universal	0.829047	991/1000
ApproximateEntropy	0.767582	988/1000
RandomExcursions	0.349160	606/613
RandomExcursionsVariant	0.359855	607/613
Serial	0.800005	994/1000
LinearComplexity	0.747898	994/1000

Table 2: NIST tests results: The uniformity of P-values and the proportion of passing sequences for 1000 samples of 1 Mbit each generated by the variable extractor 'B'

Statistical test	P-Value	Proportion
Frequency	0.614226	991/1000
Block-Frequency	0.616305	995/1000
CumulativeSums1	0.626709	990/1000
CumulativeSums2	0.980341	994/1000
Runs	0.190654	989/1000
LongestRun	0.612147	986/1000
Rank	0.452173	988/1000
FFT	0.870856	989/1000
NonOverlappingTemplate	0.757790	993/1000
OverlappingTemplate	0.039073	988/1000
Universal	0.014550	988/1000
ApproximateEntropy	0.221317	987/1000
RandomExcursions	0.822122	607/613
RandomExcursionsVariant	0.367493	610/613
Serial	0.777265	994/1000
LinearComplexity	0.809249	992/1000

Table 3: Diehard tests results for 500 Mbit of true random data

Statistical test	P-Value 'A'	P-Value 'B'	Result
Birthday Spacing	0.865773	0.842883	SUCCESS
Overlapping 5-permutation	0.513342	0.471146	SUCCESS
Binary Rank for 31 x 31 matrices	0.634721	0.755819	SUCCESS
Binary Rank for 32 x 32 matrices	0.415482	0.352107	SUCCESS
Binary Rank for 8 x 8 matrices	0.576012	0.361938	SUCCESS
Bitstream	0.970227	0.051626	SUCCESS
Overlapping-Paris-Spares-Occupancy	0.219893	0.015640	SUCCESS
Overlapping-Quardruples-Spares-Occupancy	0.173054	0.015065	SUCCESS
DNA	0.150832	0.016380	SUCCESS
Count-the-1's	0.944274	0.058056	SUCCESS
Count-the-1's for specific bytes	0.703417	0.365542	SUCCESS
Parking lot	0.229559	0.239266	SUCCESS
Minimum distance	0.794391	0.320832	SUCCESS
3D spheres	0.448424	0.064255	SUCCESS
Squeeze	0.317565	0.455628	SUCCESS
Overlapping sums	0.250307	0.988284	SUCCESS
Runs	0.814724	0.194590	SUCCESS
Craps	0.840551	0.823860	SUCCESS

Supplementary Material 2 : Entropy calculation

The model used in our paper follows the method introduced in Frauchiger et al. (CITE). We specify first the density operator ρ corresponding to our QRNG. We consider our state as represented by different subsystems. A subsystem I that, following a Poissonian distribution, encodes the number of photons. A subsystem D that encodes the state of the detectors given by the random variable S. If $S = 1$ the detectors will experience a dark count and it will click independently of the absorption of a photon, otherwise the detector will behave normally. The string $\overline{s}$ corresponds to the state of each detector. If the detectors are labelled from 0 to $m - 1$. Finally the subsystem P consider the position of the photons with respect to the matrix of SPADs.

$$\rho = \sum_{n,\overline{s}} P_N(n) P_S(\overline{s}) |n\rangle\langle n|_I \otimes |\overline{s}\rangle\langle\overline{s}|_D \otimes |\phi\rangle\langle\phi|_P^{\otimes n} \tag{1}$$

The state of the n photons is in the form $|\phi\rangle^{\otimes n}$. This corresponds to consider the photons as distinguishable. This is due to the fact that in beam-splitting experiment the behaviour of photons can be in-principle described in this way as it is shown in the work of (CITE Leonhardt). Since the illumination of the SPAD matrix have been calibrated to be uniform, we can consider each photon to have equal probability to be in front of any detectors. This means that we model the state $|\phi\rangle$ to be the $|W\rangle$ state of dimension m. If the illumination is imbalanced this state should be adjust accordingly.

The detection POVM can be described in terms of different operators for each detector D_i. The following expressions describe the POVM's elements relative to the detection or no detection of a photon:

$$P_{D_i}^{n,1} := (\eta(\mathbb{1} - |0\rangle\langle 0|))^{\otimes n} \quad P_{D_i}^{n,0} := ((1-\eta)\mathbb{1} + \eta|0\rangle\langle 0|)^{\otimes n} \tag{2}$$

For each photon arriving on the detector we consider a probability η that the photon is absorbed by the detector. This event is considered to be unforeseeable by a third party.

The operator related to generating a sequence $\overline{x}$ is now given by:

$$\Pi^{\overline{x}} = \sum_{n,\overline{s}} |n,\overline{s}\rangle\langle n,\overline{s}|_D \bigotimes_i P_i^{n,\overline{x}_i,\overline{s}_i} \tag{3}$$

where the sum over $\overline{s}$ consider all the sequences consistent with the output sequence $\overline{x}$ (if $\overline{s}_i = 1$ this means $\overline{x}_i$ is necessarily 1) and.

$$P_i^{n,\overline{x}_i,\overline{s}_i} = \begin{cases} P^{n,\overline{x}_i}, & \text{if } \overline{s}_i = 0 \\ id^{\otimes n}, & \text{if } \overline{s}_i = 1 \end{cases} \tag{4}$$

meaning that if the random variable $\overline{s}_i = 1$ the detector will click independently of the photon state $|\phi\rangle^{\otimes n}$.

Once that the measurement operator and the preparation state of the device are define we have

to consider the classical noise. In order to model this we consider the following measurement:

$$E^{n,\overline{s}} = |n,\overline{s}\rangle\langle n,\overline{s}| \bigotimes_i \mathbb{1}^{\otimes n} \tag{5}$$

the classical noise is define as its outcome and corresponds to the random variable $N \in \{0, ..., \infty\}$ corresponding to the number of photon and $\overline{S}$ which is the sequence of random variables correspond the state of each detector with respect to a possible dark-count.

The probability to have a certain output string of bits at each sampling event is given now by the Born rule:

$$P(\overline{x}) = Tr(\Pi^{\overline{x}}\rho) \tag{6}$$

which corresponds in the observed experimental distribution of the obtained bits strings.

However the probability distribution of importance in the presented model is given by the joint probability of the output variable and the classical noise, given again by the Born rule:

$$P(\overline{x}, n, \overline{s}) = Tr(\Pi^{\overline{x}} E^{n,\overline{s}} \rho (E^{n,\overline{s}})^\dagger) \tag{7}$$

By simple combinatorial calculation it is possible to obtain the expression given in the main text:

$$P(\overline{x}, n, \overline{s}) = P_N(n) P_S(\overline{s}) \sum_{i=0}^{H(\overline{x})-H(\overline{s})} (-1)^i \binom{H(\overline{x}) - H(\overline{s})}{i} \left(1 - \eta - \frac{H(\overline{x}) - i}{m}\eta\right)^n \quad (8)$$

where the vector $\overline{s} = s_1, s_2, ..., s_m$ collects all the variables for dark counts for each detector, $H(\cdot)$ is the Hamming weight function and m is the number of detector considered. This model corresponds to a generalization of the two model previously proposed for two detectors [?].

Supplementary Material 3: Comparison with the state of the art

Table 1: Comparison Table. Comparison of the presented Quantum Random Number Generator with the state of the art

Performance	**This Work**	**Stefanov[1]**	**Amri[2]**	**Sanguinetti[3,4]**	**qStream[5]**	**Wei[6]**	**Nie[7]**	**Dynes[8]**	**Matsumoto[9]**
Bit Rate	400Mb/s	4-16Mb/s	5-12Gb/s	4.9Mb/s	1Gb/s	280Gb/s	96Mb/s	4Mb/s	20Mb/s
Thermal Noise	$< 0.1\%$	$< 1\%$	$< 1\%$	$< 3\%$	Not specified	$< 0.1\%$	$< 0.1\%$	$< 2\%$	N/A
Full Integration	yes	no	no	no	no	no	no	no	no
On-chip Extractor	yes	no	no	yes	yes	no	no	no	no
Dimension (mm^2)	45	7728	25	22.5	10,892.4	N/A	N/A	N/A	0.012
Standard Tests/ Certifications	NIST, DIEHARD	NIST, DIEHARD, METAL, CTL, AIS31	NIST	NIST, ISTO/ IEC-18031	NIST	DIEHARD STS	NIST	NIST, DIEHARD	NIST
Power (mW)	499	1,500	25	83.44	12,780	N/A	N/A	N/A	1.9
FOM (nJ/bit)	1.25	93.7	0.0025	17	12.8	N/A	N/A	N/A	0.095

1. Stefanov, A., Gisin, N., Guinnard, O., Guinnard, L. & Zbinden, H. Optical quantum random number generator. *Journal of Modern Optics* **47**, 595–598 (2000).

2. Amri, E., Felk, Y., Stucki, D., Ma, J. & Fossum, E. R. Quantum random number generation using a quanta image sensor. *Sensors* **16**, 1002 (2016).

3. Sanguinetti, B., Martin, A., Zbinden, H. & Gisin, N. Quantum random number generation on a mobile phone. *Physical Review X* **4**, 031056 (2014).

4. IdQuantique. Quantis QRNG Chip Brochure. *https://www.idquantique.com/random-number-generation/products/quantis-qrng-chip/* (2019).

5. QuintessenceLab. qStream. https://www.quintessencelabs.com/. Accessed: 2020-02-27.

6. Wei, W., Xie, G., Dang, A. & Guo, H. High-speed and bias-free optical random number generator. *IEEE Photonics Technology Letters* **24**, 437–439 (2011).

7. Nie, Y.-Q. *et al.* Practical and fast quantum random number generation based on photon arrival time relative to external reference. *Applied Physics Letters* **104**, 051110 (2014).

8. Dynes, J. F., Yuan, Z. L., Sharpe, A. W. & Shields, A. J. A high speed, postprocessing free, quantum random number generator. *applied physics letters* **93**, 031109 (2008).

9. Matsumoto, M. *et al.* 1200μm 2 physical random-number generators based on sin mosfet for secure smart-card application. In *2008 IEEE International Solid-State Circuits Conference-Digest of Technical Papers*, 414–624 (IEEE, 2008).

P4. High detection rate and high detection efficiency with parallel-SNSPDs

High detection rate and high detection efficiency with parallel-SNSPDs

Matthieu Perrenoud,[1, a)] Misael Caloz,[1] Emna Amri,[1,2] Claire Autebert,[1,2] Hugo Zbinden,[1] and Félix Bussières[1,2]
[1)]*Group of Applied Physics, University of Geneva, CH-1211 Geneva, Switzerland*
[2)]*ID Quantique SA, CH-1227 Geneva, Switzerland*

(Dated: 2 March 2020)

Recent progress in the development of superconducting nanowire single-photon detectors (SNSPD) has delivered excellent performance, and their increased adoption has had a great impact on a range of applications. One of the key characteristic of SNSPDs is their detection rate, which is typically higher than other type of free-running single-photon detectors. The maximum achievable rate is limited by the detector recovery time after a detection, which itself is linked to the superconducting material properties and to the geometry of the meandered SNSPD. One potential approach to increase the detection rate further is based on parallelizing smaller meander sections. In this way, a single detection temporarily disables only one subsection of the whole active area, thereby leaving the overall detection efficiency mostly unaffected. In practice however, cross-talk between parallel nanowires typically leads to latching, which prevents high detection rates. Here we show how this problem can be avoided through a careful design of the whole SNSPD structure. We demonstrate highly efficient molybdenum silicide-based superconducting nanowire single-photon detectors capable of detecting at more than 200 MHz using a single coaxial line. This significantly outperforms detection rates achievable with single meander SNSPDs while maintaining high efficiency and low jitter.

Superconducting nanowire single-photon detectors[1] (SNSPDs) are known for yielding overall excellent performance thanks to their high efficiency[2], low dark count rate[3], short dead time[4] as well as timing jitter in the order of picoseconds[5–7]. This makes them a key technology for application requiring highly efficient single-photon detection such as optical quantum information processing[8], deep-space optical communication[9] and optical quantum computing[10]. In particular, their typical dead time of a few tens of nanoseconds is of great interest in various applications requiring high detection rates with free-running detectors such as quantum key distribution[11,12].

Although SNSPDs offer high detection rate, their dead-time still limits their use to rates of several tens of MHz. Indeed, the detection efficiency is directly dependent on the biasing current into the nanowire, which drops to zero after a detection and recovers over time. Consequently the detection rate of SNSPDs is ultimately limited by their recovery time[13]: when a photon is absorbed the nanowire becomes resistive and the biasing current rapidly leaves the nanowire (with a typical time of about 1 ns or less), this effectively leads to a zero efficiency right after the detection. After the current left the nanowire, it rapidly cools back to its superconducting state, and the current flows back inside the nanowire. At this point, the kinetic inductance of the superconducting meander forces the bias current to recover with a time constant $\tau = L_k/R$, where R the overall series impedance of the readout circuit and L_k is the kinetic inductance of the nanowire. The timing constant τ typically is in the range of a few tens of nanoseconds. A longer nanowire exhibits a larger kinetic inductance, which slows its current dynamic down and hence increases the time needed before the detector recovers its nominal efficiency. Consequently, at high detection rates where successive detection can happen in time intervals of the order of τ, the system detection efficiency drastically drops. Sub-nanosecond recovery times can be obtained with extremely short nanowires coupled to integrated waveguides[4], but such detectors have so far not shown high system detection efficiency when coupled to an optical fiber, due to the fibre to waveguide coupling loss. Arrays of detectors can be used to keep the efficiency to high detection rates[14,15], but this comes to the cost of using multiple coaxial readout lines drastically increasing the cooling power needed to operate the system.

A potential solution to solve this limitation is to use a parallel SNSPDs design, which consists of splitting the SNSPD into several nanowires connected in parallel[16]. This idea is represented on Fig. 1a. Series resistors ensure that the biasing current is split evenly between the different nanowires. Splitting the whole detector in several nanowires can effectively reduce their individual lengths, hence lowering their respective kinetic inductance, which leads to a shorter recovery time for each nanowire. Moreover only part of the detector is disabled after a detection event which leaves the remaining nanowires available to detect another photon at their full detection efficiency. Similar designs have also been used to demonstrate photon-number-resolving detection[17] and fast recovery time[16,18]. However, to the best of our knowledge, their potential for achieving very large detection rates has not been demonstrated before. As we observed and report on below, the problem seems to be that at high detection rate, electronic crosstalk between the different nanowires cumulates, and this leads to a cascading effect between the different nanowires. Ultimately, this leads to latching, i.e. all nanowires end up in a steady resistive state where the whole detector is effectively disabled.

In this work, we demonstrate a fiber-coupled parallel SNSPD design that overcomes this limitation. Cascading effect between the nanowires is mitigated by adding carefully designed superconducting nanowires to the structure.We demonstrate detection rates over 200 MHz without any latching, and a fibre-coupled system detection efficiency (SDE) as high as 77%, and more than 50% average SDE per photon at 50 MHz detection rate under continuous wave illumination.

[a)]matthieu.perrenoud@unige.ch

To operate parallel SNSPD design at high detection rates, we must consider the electronic crosstalk between the different nanowires : after a photon absorption in one nanowire, the electronic current in that nanowire will be partly redirected in every other nanowires as well as in the readout circuit. In each nanowire, this signal will add-up with the already present bias current. At high detection rates, when multiple absorption occur in different nanowires at time intervals shorter than the recovery time of each nanowire, this effect will stack-up and can eventually bring the current in the remaining nanowires above their critical current value I_c. This cascading effect will lead to a latched state. As the total current in any nanowire has to remain under I_c, only a finite number of photons can be detected in a short time interval for a given biasing current per nanowire $I_b < I_c$. With a larger difference $I_b - I_c$, a larger number of photons can be detected in a given time interval without cascades. The use of lower I_b values thus allows for operation at higher detection rates. Nonetheless low I_b values are not sufficient to obtain saturated efficiency and lower I_b values thus lead to lower detection efficiencies.

The solution we propose here ensures that the total current will never exceed the critical current I_c in each nanowire while using a bias current I_b high enough to maintain the saturated efficiency of the detector. To achieve this, additional nanowires are added in parallel to the structure as shown in Fig. 1b. They are positioned outside the optical fiber spot which makes them unexposed to light. Part of the redirected current after a detection is therefore split into these unexposed nanowires, which effectively reduces the additional current seen by the exposed nanowires. This effect can be enhanced by reducing the kinetic inductance of the unexposed nanowires. By designing them with larger widths than the exposed nanowires, it can be ensured that even if every exposed nanowire detects a photon, this additional structure will carry the excess current without reaching its critical current value. Thanks to the reduced crosstalk seen by the exposed nanowires, our design effectively allows for high biasing current I_b with respect to I_c. However this comes at a cost: lowering the crosstalk also lowers the current flowing into the readout, which inevitably decreases the output signal amplitude[16]. Hence the number of parallel nanowires and their minimum kinetic inductance is limited by the performances of the amplification electronics.

We present results obtained with two detectors which characteristics are shown in table I. Device A is made of few exposed wires and very large unexposed nanowires with low inductance. This device is designed to be resistant to cascade effects and latching. Device B is made with a large number of exposed nanowires, which minimizes the probability of consecutive absorption in the same nanowire. Thus device B is expected to maintain its nominal efficiency at higher rates. However the width of every nanowire (exposed and unexposed) is the same. This device is sensitive to cascade and requires a lower biasing current I_b to be operated at high detection rates.

In the design, we also need to consider that the heat generated by the detection mechanism can trigger other nanowires located in the vicinity of the absorption. This thermal crosstalk, can be avoided with proper spacing of the different parallel nanowires. We observed no thermal crosstalk between nanowires when spaced by a 800 nm gap and implemented this distance in our designs. As the fraction of the detection area required by the gaps increases with the number of exposed nanowires, designs with a larger number of exposed nanowires have a lower fill factor which in turn decreases the detection efficiency.

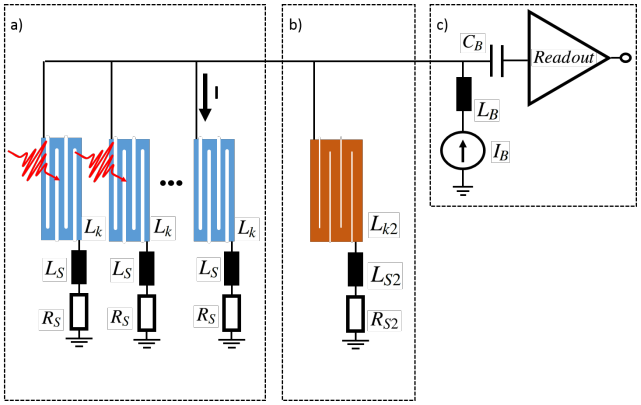


FIG. 1. Parallel SNSPD design. a) Schematic of a basic parallel SNSPD design[16], which consists of a limited number of photosensitive nanowires with a specific inductance L_k connected in parallel. An additional series inductor L_s can be added to choose the overall inductance of each section which has an impact on the output signal amplitude. Series resistors R_s ensure that the biasing current is evenly split among every nanowire. b) Additional nanowires with low inductance L_{k2} are added in order to decrease the electronic crosstalk between the nanowires during detection events. The values of L_{S2} and R_2 can be sized to optimize the crosstalk reduction while keeping output signal of sufficient amplitudes. c) A bias tee is used to bias the detector and amplify the output signal with the same coaxial line.

TABLE I. Characteristics of the two devices presented in this work.

	Device A	Device B
Exposed nanowires	6	16
Unexposed nanowires	8	24
Width of exposed nanowires	100 nm	150 nm
Width of unexposed nanowires	1600 nm	150 nm
Protected from cascade effect	Yes	No

The parallel SNSPD designs are patterned using electron beam lithography on 6 nm MoSi film. Series resistors and electrodes are created by lift-off of a 10 nm Ti and 90 nm Au evaporated double layer, with photo-lithography techniques used to pattern resistive lines of 5 µm width and various lengths. Fig. 2 shows a scanning electron microscope (SEM) picture of different parts of the structure. Gold resistors of different values can be seen on the outside, while the photosensitive area (16 µm× 16 µm) is found at the center of the image. An optical cavity is built, which consists of a silver mirror and a SiO2 spacer integrated under the device, and a single SiO2 capping layer on top[19]. The devices are then separated from the wafer using deep reactive ion etching (DRIE), and pack-

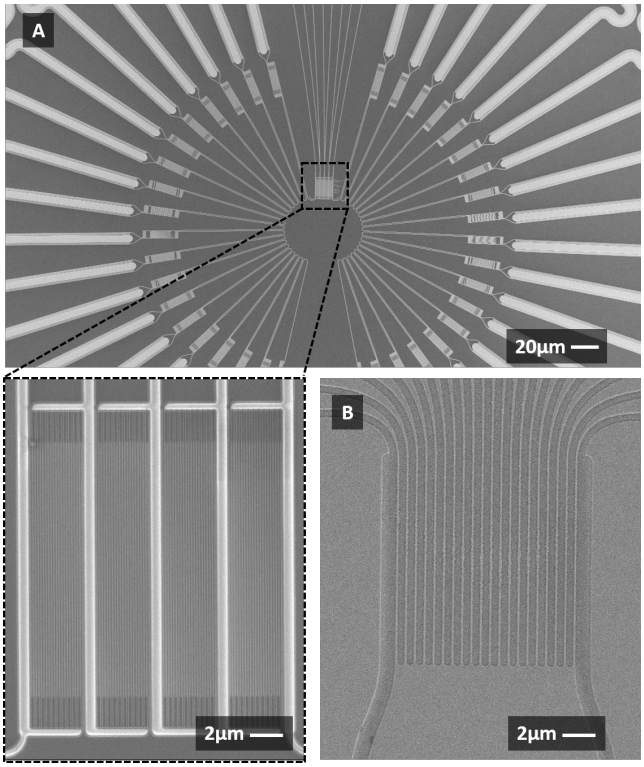


FIG. 2. SEM pictures of the devices. a) Overview of a detector similar to device A presented in the text, with 4 photosensitive nanowires and 36 non photosensitive additional large wires. The 4 photosensitive nanowires can be seen in the center of the image (dotted square). The 36 additional nanowires are positioned on a circle around the center, large gold resistive lines can be seen on the edge of the image. The dotted lines show a zoom on the photosensitive area of this design, the four nanowires are shaped in a meander and connected in parallel. b) Zoom on the photosensitive area of a detector, similar to device B presented in the text, made of 20 parallel straight nanowires.

aged using a self-alignment package technique[20]. The packaged SNSPD is then cooled down to 0.8 K using a sorption-equipped closed-cycle cryocooler. Amplification of the output pulses is done through a first amplification stage cooled at 40 K, and a second amplification circuit at room temperature.

The efficiency at low detection rate of different parallel design devices has been characterized using a calibrated power-meter and three variable attenuators[7]. The incoming light polarization was optimized to maximize the detection efficiency. Fig. 3 shows the SDE vs bias current at low detection rate of device A. A saturation plateau is reached with a maximum efficiency of 77%. High efficiency is obtained with this design thanks to a relatively high fill factor. Indeed, this design consists of nanowires in a meander shape with a fill factor of 60% and five large spacing of 800 nm between each nanowire leading to a global filling factor of 42%. In comparison, device B is made of 16 exposed wires, all of which are spaced by a 800 nm spacing. The fill factor of device B is therefore only 16% which drastically reduces the maximum efficiency obtained. At low detection rate, device B exhibits a saturating

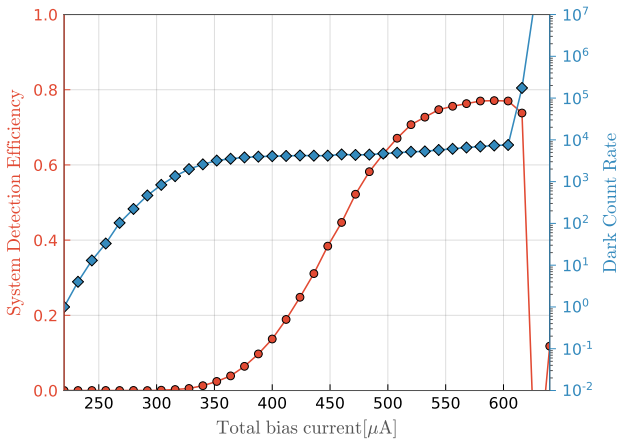


FIG. 3. System detection efficiency of device A. The average fill factor of the device is 42%. The efficiency is improved by an integrated optical cavity. DCR could be reduced using simple fiber loops to filter unwanted wavelengths.

plateau at around 30% SDE.

At high detection rate, several consecutive photons can be absorbed in the same exposed nanowire of the detector before they could fully recover their efficiency. As a result, the average efficiency will decrease with the rate of incident photons. To characterize this effect, the rate of incident photons illuminating the detector is progressively increased for a given bias current. The detection rate is monitored for each different illumination rate and from this the average SDE per photon is calculated. Fig. 4 shows the SDE vs detection rate of device A and device B. An efficiency drop of 10% of the nominal SDE is observed at 13.4 MHz for device A. An efficiency of 33% is obtained at a detection rate of 100 MHz. As desired, the additional large parallel wires protect the device from latching at high detection rate. As a result, the detector can still be operated with rates above 200 MHz.

The operation of device B at high rates was not possible with a bias current close to I_c because the device is sensitive to cascading effect, as explained above. To operate this detector, the bias current I_b was decreased well below I_c, which reduces its efficiency to a value of 12% at low detection rates. In this regime the detector was able to operate above 100 MHz before latching due to the cascading effect caused by the accumulation of current in the nanowires following multiple detections in short time intervals. The large number of exposed nanowires of device B however reduces the probability of fast consecutive absorptions in the same nanowire, which maintains the efficiency of the detector at higher rates than what was observed for device A. This design exhibits a 10% efficiency drop at detection rates as high as 88 MHz.

This results confirms the importance of designing wider parallel wires to protect them from cascading and latching, as well as reducing the number of exposed nanowires to limit the possible accumulation of current at high detection rates. An optimized detector would consist in maximizing the number of exposed nanowires while still allowing operation with

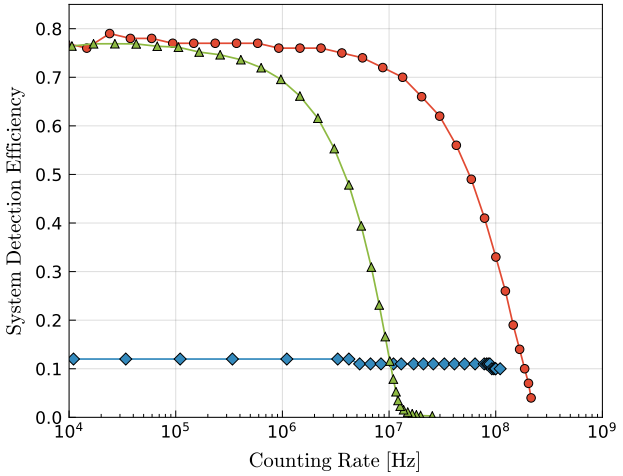


FIG. 4. System detection efficiency of different devices versus detection rate. Device A (red dots) exhibit an SDE at low detection rate almost as high as single meander designs, with an efficiency drop of 10%. The device never latches even above 200 MHz. Device B (blue squares) maintains its efficiency at higher counting rates than device A, but latches when operated at high rates.

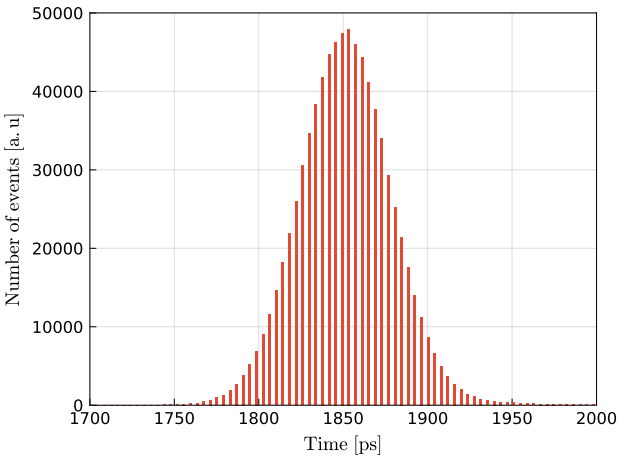


FIG. 5. Timing distribution of the photon detection event. Full-width at half maximum is 62 ps

a bias current I_b high enough to reach the saturated efficiency. This can be done by minimizing the inductance of unexposed nanowires to the limits allowed by the output signal amplification capabilities of the setup, and use detectors geometries with high I_c with respect to the saturated efficiency current.

Timing fluctuation (jitter) in the detection event has an important impact for many application. We characterized the jitter of the system at low detection rates using a TCSPC card with 10 ps time resolution. We measured a jitter of 62 ps full width at half maximum (FWHM) with a detector which design is similar to device A (Fig.5). This result is higher than what is usually obtained with MoSi single meanders SNSPDs[7]. The jitter due to the propagation of the signal along the nanowire is expected to be smaller in our devices than for single meanders due to the shorter nanowires required[6,21]. Thus the relatively high jitter measured could be a consequence of the lower output signal of our devices compared to single meander SNSPDs, indeed the signal over noise ratio is an important contribution of the total jitter of the system[22,23].

This measurement was performed at an approximate detection rate of 2 MHz. We expect that the jitter deteriorates at higher counting rates in the order of $1/\tau$. Indeed in this regime, many detection occur in several nanowires during their recovery time, and the actual biasing current redirected in the readout can take any value below the critical current.

A detector with a detection efficiency as high as 77% have been fabricated and operated with detection rate above 200 MHz. This device was prevented from cascading effects at very high detection rates thanks to the addition of larger nanowires unexposed to light. The necessity of protecting the additional nanowires from cascading by increasing their width has been verified. We also showed that the efficiency can be maintained at higher detection rates with more parallel nanowires exposed to light. Better performances are expected to be achievable with designs adjustments. Limitations due to the lower filling factor of our designs can be overcome by improving the optical cavity, which is a key factor for any detector aiming to near perfect detection efficiency. Improvement of the amplification scheme can lead to better jitter at high detection rate.

[1] G. N. Gol'tsman, "Picosecond superconducting single-photon optical detector," Appl. Phys. Lett. **79**, 705 (2001).

[2] F. Marsili, "Detecting single infrared photons with 93% system efficiency," Nat. Photonics **7**, 210–214 (2013).

[3] H. Shibata, "Ultimate low system dark-count rate for superconducting nanowire single-photon detector," Opt. Letters **40**, 3428–3431 (2015).

[4] A. Vetter, "Cavity-enhanced and ultrafast superconducting single-photon detectors," Nano Lett. **16, 11**, 7085–7092 (2016).

[5] I. E. Zadeh, "Single-photon detectors combining high efficiency, high detection rates, and ultra-high timing resolution," APL. Photonics **111301** (2017).

[6] B. A. Korzh, "Demonstrating sub-3 ps temporal resolution in a superconducting nanowire single-photon detector," arXiv:1804.06839 (2018).

[7] M. Caloz, "High-detection efficiency and low-timing jitter with amorphous superconducting nanowire single-photon detectors," Appl. Phys. Lett. **112**, 061103 (2018).

[8] R. H. Hadfield, "Single-photon detectors for optical quantum information applications," Nat. Photonics **3**, 696–705 (2009).

[9] M. E. Grein, "An optical receiver for the lunar laser communication demonstration based on photon-counting superconducting nanowires," SPIE proceedings **9492**, 949208 (2015).

[10] X. Qiang, "Large-scale silicon quantum photonics implementing arbitrary two-qubit processing," Nat. Photonics **12**, 534–539 (2018).

[11] H. Takesue, "Quantum key distribution over a 40-db channel loss using superconducting single-photon detectors," Nat. Photonics **1**, 343 (2007).

[12] A. Boaron, "Secure quantum key distribution over 421 km of optical fiber," Phys. Rev. Lett. **121**, 190502 (2018).

[13] A. J. Kerman, "Kinetic-inductance-limited reset time of superconducting nanowire photon counters," Appl. Phys. Lett. **88**, 111116 (2006).

[14] A. J. K. R. J. M. D. Rosenberg and E. A. Dauler, "High-speed and high-efficiency superconducting nanowire single photon detector array," Optics Express **21, Issue 2**, 1440–1447 (2013).

[15] W. Z. et al., "A 16-pixel interleaved superconducting nanowire single-photon detector array with a maximum count rate exceeding 1.5 ghz," IEEE Transactions on Applied Superconductivity **29, no.5**, 1–4 (2019).

[16]A. Korneev, "Ultrafast and high quantum efficiency large-area superconducting single-photon detectors," Proc. of SPIE **6583**, 65830I–1 (2007).
[17]A. Divochiy, "Superconducting nanowire photon-number-resolving detector at telecommunication wavelengths," Nat. Photonics **2**, 302–306 (2008).
[18]M. Tarkhov, "Ultrafast reset time of superconducting single photon detectors," Appl. Phys. Lett. **92**, 241112 (2008).
[19]E. A. D. J. K. W. Y. K. M. R. Vikas Anant, Andrew J. Kerman and K. K. Berggren, "Optical properties of superconducting nanowire single-photon detectors," Optics Express **16, Issue 14**, 10750–10761 (2008).
[20]B. C. I. V. S. M. G. Aaron J. Miller, Adriana E. Lita and S. W. Nam, "Compact cryogenic self-aligning fiber-to-detector coupling with losses below one percent," Optics Express **19, issue 19**, 9102–9110 (2011).
[21]N. Calandri, "Superconducting nanowire detector jitter limited by detector geometry," Appl. Phys. Lett. **109**, 152601 (2016).
[22]L. You, "Jitter analysis of a superconducting nanowire single photon detector," AIP Advances **3**, 072135 (2013).
[23]M. Caloz, "Intrinsically-limited timing jitter in molybdenum silicide superconducting nanowire single-photon detectors," arxiv:1906.02073 (2019).

P5. Direct measurement of the recovery time of superconducting nanowire single-photon detectors

Direct measurement of the recovery time of superconducting nanowire single-photon detectors

Claire Autebert,[1] Gaëtan Gras,[1,2] Emna Amri,[1,2] Matthieu Perrenoud,[1] Misael Caloz,[1] Hugo Zbinden,[1] and Félix Bussières[1,2]

[1)]*Group of Applied Physics, University of Geneva, CH-1211 Geneva, Switzerland*

[2)]*ID Quantique SA, CH-1227 Carouge, Switzerland*

(Dated: 2 March 2020)

One of the key properties of single-photon detectors is their recovery time, i.e. the time required for the detector to recover its nominal efficiency. In the case of superconducting nanowire single-photon detectors (SNSPDs), which can feature extremely short recovery times in free-running mode, a precise characterisation of this recovery time and its time dynamics is sometimes essential to use them in certain quantum optics or quantum communication experiments. We introduce a fast and simple method to characterise precisely the recovery time of SNSPDs. It provides full information about the recovery of the efficiency in time for a single or several consecutive detections. We also show how the method can be used to gain insight into the behaviour of the bias current inside the nanowire after a detection, which allows predicting the behaviour of the detector and its efficiency in any practical experiment using these detectors.

I. INTRODUCTION

Single-photon detectors are a key component for optical quantum information processing. Among the different technologies developed for single-photon detection, superconducting nanowire single-photon detectors (SNSPDs) have become the first choice of many applications showing performances orders of magnitude better than their competitors. These nano-devices have stood out as highly-promising detectors thanks to their high detection efficiency[1], low dark count rate[2], excellent time resolution[3,4] and fast recovery[5]. Superconducting nanowire single photon detectors have already had an important impact on demanding quantum optics applications such as long-distance quantum key distribution[6], quantum networking[7], optical quantum computing[8], device-independent quantum information processing[9,10] and deep space optical communication[11].

Depending on the application, some metrics become more important than others and can require extensive characterisation. One example is quantum key distribution (QKD), where the recovery time of SNSPDs limits the maximum rate at which it can be performed. In such a case, studying the time evolution of the SNSPD efficiency after a detection becomes important and would give us insight into the detector's future behaviour, allowing the prediction of experimental performances. Obtaining accurate information is however a non-trivial task because the recovery time is intrinsically linked to the time dynamics of the bias current flowing inside the detector, from which extracting a faithful information is not possible because of the electrical readout influence that affects the shape of the current pulse as initially generated by the SNSPD.

There are several methods used to characterise the recovery time of the efficiency of a SNSPD. The first one uses the output pulse delivered by the readout circuit to gain knowledge about the recovery time dynamics. As mentioned above, we cannot fully trust this method since the time decay of the output voltage pulse is inevitably affected by the amplifier's bandwidth and by all other filtering and parasitic passive components. In the best case we can only have an indirect estimation of the efficiency temporal evolution. A second method consists of extracting the recovery time behaviour from the measurement of the detection rate as a function of the incident photons rate. This method can be performed with either a continuous-wave or a pulsed laser source. The main problem with the pulsed source configuration is that we can only probe the efficiency at time stamps multiple of the pulse period which does not give full information about the continuous time dynamics. Both methods have the drawback of only providing an average efficiency per arriving photon. They can moreover be very sensitive to external parameters such as the discriminator's threshold level. Hence, using one of these measurements does not allow one to make unambiguous predictions about the outcome of any other experiment using the detectors. Another method is based on measuring the auto-correlation in time between two subsequent detections when the detector is illuminated with a continuous-wave laser[12] or a pulsed laser[13]. This method has the clear advantage over all other methods of allowing a direct observation of the recovery of the efficiency in time, and it can therefore reveal additional details (for example the presence of afterpulsing). While the implementation of this auto-correlation method is relatively simple, the acquisition time can however be very long.

In this article we introduce and demonstrate a novel method, simple in both its implementation and analysis, to fully characterise the recovery time dynamics of a single-photon detector. This method is based on an "upgrade" of the autocorrelation method mentioned above, and has the advantage of a much shorter acquisition time with no need of data post-processing. We apply it to characterise the recovery time of SNSPDs under different operating conditions and for different wavelengths. We can also use it to estimate the variation of the current inside the detector after a detection, and consequently, gain insights into what happens to the bias current when two detections occur within the time period needed by the efficiency to fully recover. This method also allows us to reveal details that are otherwise difficult to observe, such as afterpulsing or oscillations in the bias current's recovery as well as predict the outcome of the count rate measurement.

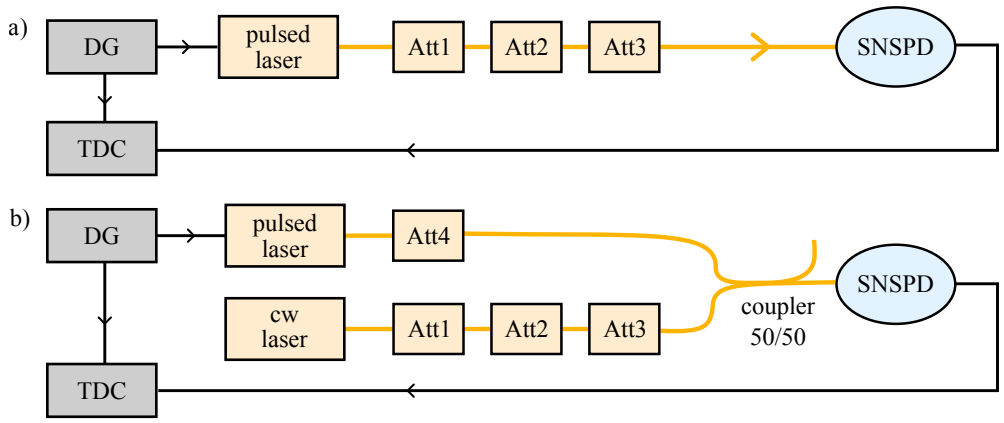


FIG. 1. Schematics of the experimental setups for the a) pulsed-autocorrelation method and for the b) hybrid-autocorrelation method. DG: delay generator, TDC: time-to-digital convertor, Att: attenuators.

II. HYBRID AUTOCORRELATION METHOD

To investigate the time-dependence of the detection efficiency after a first detection event, a useful tool is the normalised time autocorrelation of one detector, which is proportional to the expected probability value of having two detection events separated in time by Δt on the same detector. For an ideal detector with a zero recovery time, the detection events occuring at times t and $t+\Delta t$ are independent when illuminated with coherent light. In this case the autocorrelation will be equal to one for any value of Δt. For a detector with a non-zero recovery time, the autocorrelation function will be equal to zero at $\Delta t = 0$, and then it will recover towards one with a shape that is directly indicative of the value of the efficiency after a detection occurring at time zero.

This method can be implemented with a continuous wave (CW)[12] or pulsed laser[13] and it has the advantage of allowing a direct observation of the recovery of the efficiency in time. Its implementation requires a statistical analysis of the inter-arrival time between subsequent detections. A schematic of an implementation of this method with a pulsed laser is shown in Fig. 1a, and we use it for comparison with the novel method we introduce hereafter. A delay generator (DG) is used to generate two laser pulses with a controllable time delay between them. The triggerable laser is generating short pulses that are then attenuated down to ≈ 0.1 photon per pulse by calibrated variable attenuators. The output signal of the detector is fed to a time-to-digital converter (TDC) that records the arrival times of the detections.

To reconstruct the recovery of the efficiency in time after a first detection, we analyse the time stamps to estimate the probability of the second detection as a function of its delay with respect to the first one. This method can be significantly time consuming because only one given delay can be tested at once. Moreover, one needs a detection to occur in the first pulse to count the occurrences. It also requires to have the same power in both pulses and that this power needs to be very stable during the whole duration of the experiment, which can be difficult to guarantee with some triggered laser such as gain-switched laser diodes.

Here we introduce a new method, named *hybrid-autocorrelation*, that combines the pulsed and CW autocorrelation methods. The advantages of this hybrid measurement are its rapidity, flexibility in terms of wavelengths, ability to faithfully reveal the shape of the recovery of the efficiency as well as tiny features such as optical reflections in the system or even oscillations of the bias current after the detection and most importantly, it doesn't require any post-processing to extract information. In the hybrid autocorrelation method (Fig. 1b), a light pulse is used to make the detector click with certainty at a predetermined time, which greatly reduces the total collection time needed to build the statistics. This pulse is combined on a beamsplitter with a weak but steady stream of photons (typically about 10^6 photons/second or less) coming from an attenuated CW laser. These photons are used to induce a second detection after the one triggered by the pulsed laser, and the detection probability is proportional to the efficiency at this given time. There are no big constraints on the pulsed laser; its pulse width needs only to be much smaller than the recovery time, it does not have to be at the same wavelength as the one required for the recovery time measurement (which is determined by the CW laser) and its power and polarisation do not need to be highly stable (because its only role is to create a detection at a given time with certainty). To record the detection times we use a TDC building start-stop histograms configuration, where the start is given by the DG triggering the pulsed laser.

III. RESULTS

We implemented the pulsed and hybrid autocorrelation methods using a 1550 nm gain-switched pulsed laser diode with 300 ps pulse width and a CW laser at 1550 nm (for the hybrid method). We also used meandered and fibre-coupled molybdenum silicide (MoSi) SNSPDs fabricated by the U. of Geneva group[4]. The arrival times of the detections were recorded with a TDC (ID900 from IDQ) with 100 ps-wide time bins. Figure 2 shows the temporal evolution of the normalised efficiency after a first detection obtained with the

pulsed and hybrid autocorrelation methods. The detector was biased very closely to the switching current I_{SW}, defined as the current at which the dark counts start to rise quickly. Both methods yielded similar results in the trend of the curves, but the pulsed autocorrelation method gave a much larger scatter in the data. This scatter is caused by the instability of the laser power over the duration of the measurement (about 6 hours). The hybrid-autocorrelation method measurement required only about one minute of acquisition time and gave the exact shape of the recovery of the efficiency. We also noticed that the detector does not show any afterpulsing effects, otherwise the normalized efficiency curve could momentarily reach values larger than one.

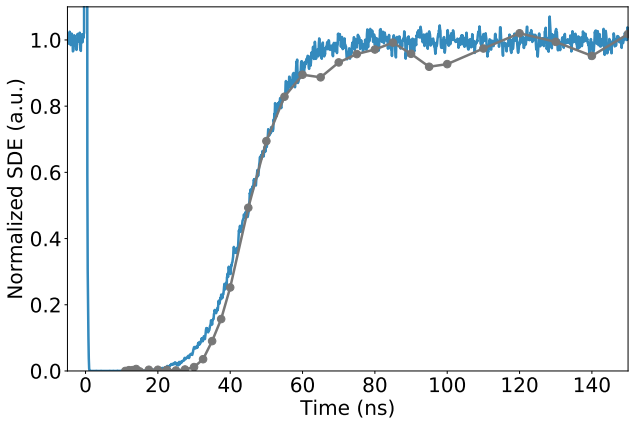


FIG. 2. Normalized SDE as a function of the time delay between two events for the pulsed autocorrelation method (grey points) and the hybrid-autocorrelation method (dark blue curve).

A. Current and wavelength dependency

Using the hybrid autocorrelation method we could also investigate the dependency of the recovery time on different operating conditions. First we looked at the behaviour with different bias currents. Fig. 3a shows the time recovery histograms for different bias currents from 8.5 µA to 13.0 µA, which correspond to the switching current I_{SW} of our detector. Fig. 3b shows the time needed by the detector to recover 50% (red curve) and 90% (blue curve) of its maximum efficiency as a function of the bias current. The results show that the SNSPD recovery time is shorter for increasing bias current, which is expected from the shape of the efficiency curve with respect to the bias current (Fig. 5b). Indeed this curve exhibits a plateau, allowing the current that is re-flowing into the nanowire after a first detection, to reach the full efficiency faster.

Second we vary the wavelength of the CW laser used for background detection. The results are shown in Fig. 4. As expected, we can see that the lower the wavelength, the faster the recovery time. This is due to the reduction of the critical current with decreasing wavelength, while the switching current stay unchanged. This leads to an increase of the plateau

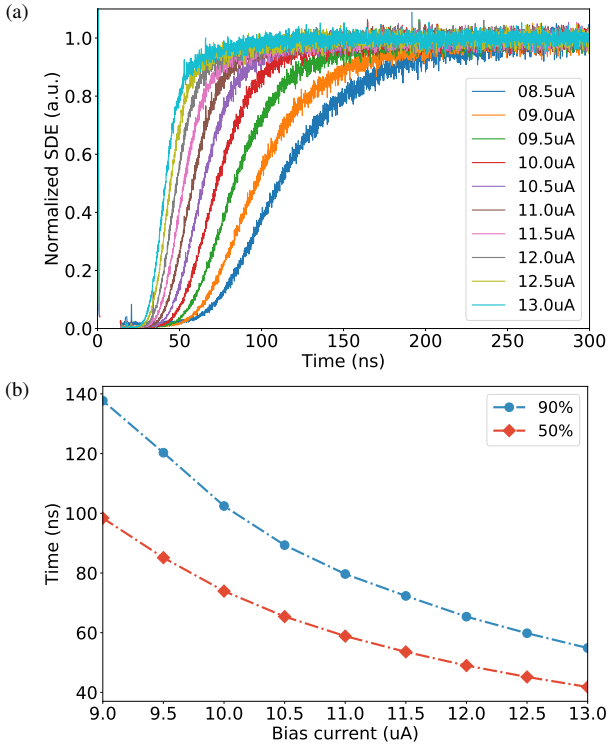


FIG. 3. a) Recovery of the normalized SDE at different bias currents; b) shows the the time to recover 50% (red diamonds) and 90% (blue dots) of the maximum efficiency as a function of the bias current.

length, allowing a faster recovery of the full efficiency. Interestingly, the curve at 850 nm seems to reveal some small oscillations of the efficiency around 30 ns after the trigger detection. While the origin of this small oscillation is not entirely clear (and we did not investigate this further), it nevertheless illustrates the capacity of the method to reveal some specific transient details of the efficiency recovery dynamics.

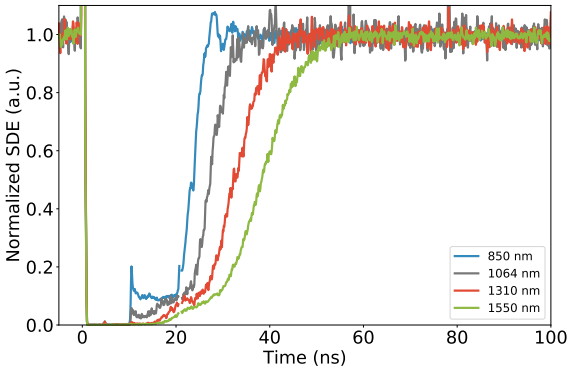


FIG. 4. Recovery of the normalized SDE at different wavelengths.

B. Current inside the SNSPD after detection

The SNSPD can be at first order modelled by an inductance L_k presenting the kinetic inductance of the nanowire, serially connected to a variable resistor whose value depends on the state of the nanowire (0 if it is superconductive, $R_{\text{hs}} \sim 1$ kΩ otherwise)[13]. The bias current I_b is provided by a current generator through a bias tee (see Fig. 5a). When a photon is absorbed and breaks the superconductivity, it creates a local resistive region called "hotspot". The current is then deviated to the readout circuit with a time constant $\sim L_k/R_{\text{hs}} \sim 1$ ns. Once the current has been shunted, the nanowire cools down and returns to thermal equilibrium allowing the current to return to the nanowire with a time constant of $\tau = L_k/R_L$, where $R_L = 50$ Ω is the typical load resistance. Note that, in practice, there may be other series resistance of a few Ohms due to the coaxial cables connecting the SNSPD to the amplifier, which might increase slightly the effective value of R_L, and therefore slightly decrease the value of τ. Also, the amplifiers are typically capacitively coupled, which is not shown here on the drawing. The drop and the recovery of the efficiency of the SNSPD after a detection are therefore directly linked to the variation of the current and to the relation between the detection efficiency and the bias current.

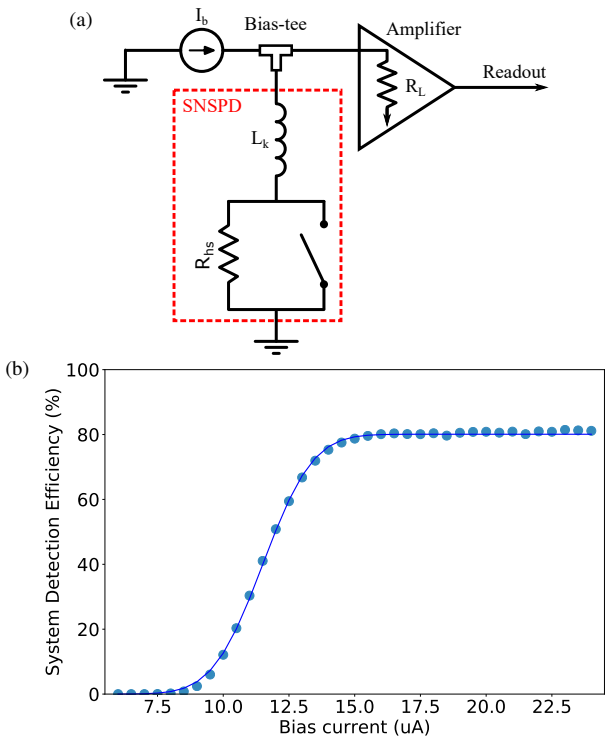


FIG. 5. (a) Simple equivalent electrical circuit of the detector and readout. (b) Relation between the SDE and bias current of a typical MoSi-based SNSPD.

On Fig. 5b, we plot the system detection efficiency as a function of the bias current of a given MoSi SNSPD, and we observe that it follows a sigmoid shape[14]. We can therefore fit that curve using the equation

$$\eta = \frac{\eta_{max}}{2}\left(1+\text{erf}\left(\frac{I-I_0}{\Delta I}\right)\right), \tag{1}$$

where I_0 and ΔI are parameters for the sigmoid and η_{max} is the maximum efficiency of the detector. After a detection, the equivalent circuit of Fig. 5a indicates that the current variation after a detection should be described by

$$I = (I_b - I_{drop})\left(1-\exp\left(-\frac{t}{\tau}\right)\right) + I_{drop}. \tag{2}$$

where I_b is the nominal bias current of operation of the detector just before a detection, I_{drop} is the current left in the nanowire immediately after a detection and τ is the time constant for the return of the current. Here, we neglect the time formation of the hotspot (and therefore the time for I to go from I_b to I_{drop}) as, according to the electro-thermal model of Ref.[15], its lifetime is expected to be short (typically a few hundreds of ps) compared to the recovery of the current τ. By fitting the curve of the efficiency versus the current with Eq. (1) (Fig. 5b) we can infer I_0 and ΔI; by inserting Eq. (1) in Eq. (2) and fitting the recovery time measurement (Fig. 6a) we can estimate I_{drop} and τ. Here, we used $I_b = 23.5$ µA and the best fit is obtained with $I_{drop} = 3.1$ µA and $\tau = 56$ ns. Then using both results, we can infer the value of the current in the nanowire versus time as shown on Fig. 6b. It is worth noting that this method predicts that $I_{drop} > 0$. Physically, this would mean that the current did not have time to completely leave the SNSPD before it became superconductive again. This is the kind of detail that is very difficult to measure directly. Admittedly, this prediction made with our method is not direct and therefore difficult to fully confirm. We note however that we obtained values for I_{drop} greater than zero for all the tested detectors. Moreover, with the values obtained for I_{drop} and τ, thanks to Eq. (1) and Eq. (2) and the efficiency versus bias current and time recovery measurements, it is possible to accurately predict the behavior of a detector at high detection rates, as shown in Section III C. This gives us an increased confidence in the method proposed here.

When a photon strikes the nanowire and a detection occurs, the current inside the detector drops to a percentage of its original value and not to zero. An interesting measurement possible with our hydrid-autocorrelation method consists in sending a train of pulses (here two) with varying delay between them to compare the recovery of efficiency when the wire detects a third photon while it has not yet recovered its full current. With several consecutive detections, we might expect some cumulative effect with the current dropping to lower and lower values. This would lead to a longer recovery time of the detector. The results of this measurement are shown in Fig. 7. The red curves correspond to the cases where two strong pulses where sent, with different time delays between them, and the blue curves correspond to the cases where only one strong pulse was sent. We can see that the shape of the autocorrelation curve for the third detection (in the case of 2 pulses) matches perfectly the one for the second detection (in the case of 1 pulse). This gives us good confidence that the current drops always to the same value. This observation,

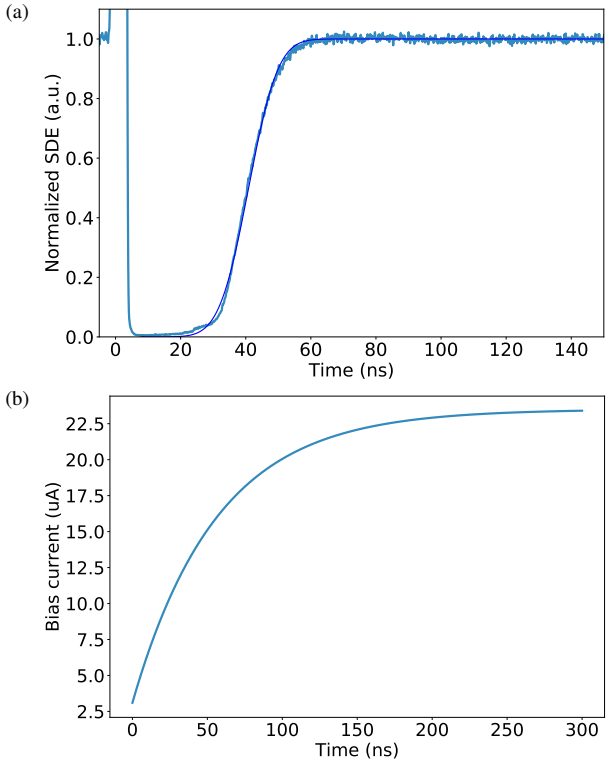


FIG. 6. (a) Normalized efficiency as a function of time after a first detection. (b) Reconstructed bias current of the detector as a function of time after a first detection.

made possible using the hybrid-autocorrelation method, has not been observed as clearly before despite being important for performance characterisation at high count rates. Indeed for experiment where the photons arrive with very short delays between them, it is important to know that the recovery time after any detection is the same and is not affected by the time delay between detections.

C. Predicting the counting rate with a continuous wave source

To illustrate the predictive power of the hybrid-autocorrelation method proposed here, it is interesting to look at how it can be used to predict the behaviour of SNSPDs at high counting rate, when the average time between two detections becomes comparable to the recovery time of the SNSPD. For example, one can consider an experiment where the light of a continuous-wave laser is sent to the detector and the detection rate is measured as a function of the incident photons rate. Only a precise knowledge of the recovery of the efficiency in time after a detection, combined with the observation that the value I_{drop} is always the same, can be used to make an accurate prediction. To estimate the count rate versus incident photon rate from the hybrid-autocorrelation method, we run a Monte-Carlo simulation. We randomly select the time t of arrival of the photon since

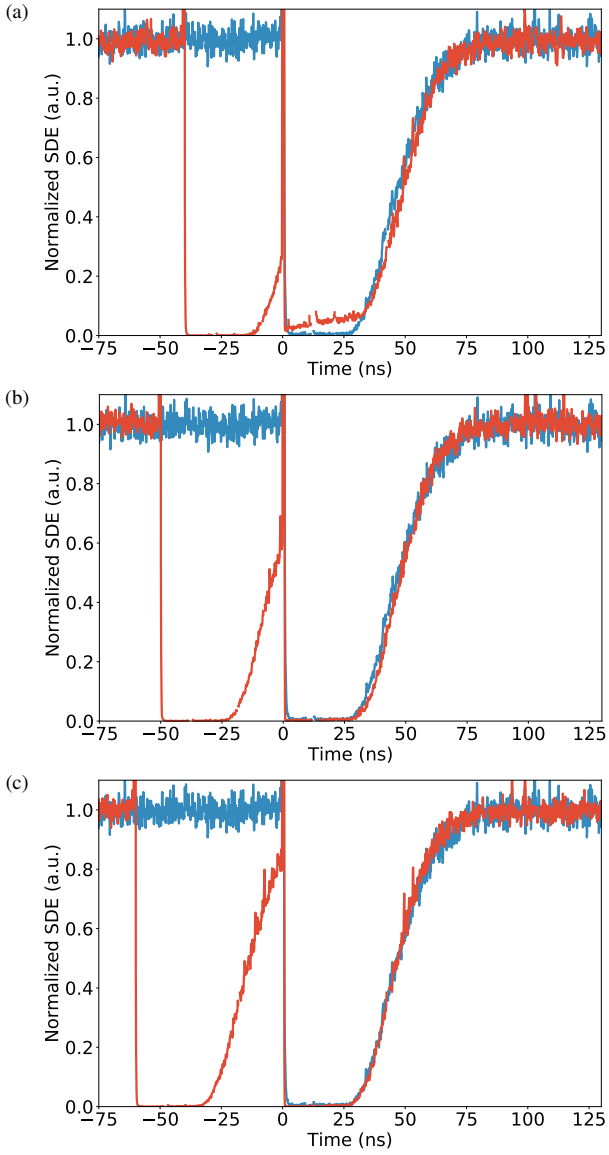


FIG. 7. Recovery of the normalized SDE for one pulse (blue curve) and for two trigger pulses (red curve) with different delays between the pulses: (a) 40 ns, (b) 50 ns and (c) 60 ns.

the last detection using the exponential distribution (which gives the probabilitity distribution of time intervals between events in a Poissonian process). Thanks to the autocorrelation measurement, we know the probability of a successful event (i.e. a detection) at time t. In case of unsuccessful event, we look at the time $t+t'$ of arrival of the next photon. Once we have a detection, we start over. We run this until we have $N = 10\,000$ detections to estimate the count rate of the detector.

Figure 8 shows, for one of the SNSPD we tested, the comparison between the experimental detection rate versus incident photon rate of the SNSPD and its prediction from the hybrid-autocorrelation measurement. We can see that the count rate data and the extrapolation from the autocorrelation

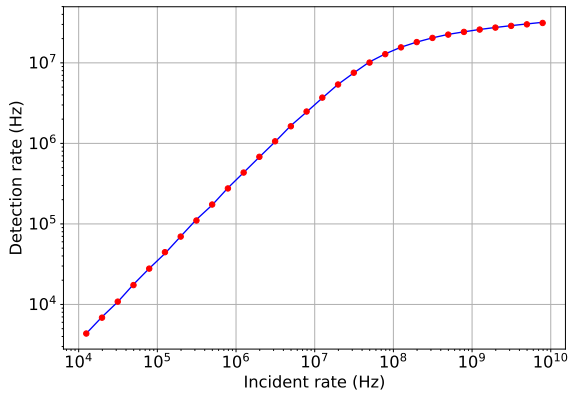


FIG. 8. Count rate with continuous wave laser: the red dots correspond to the count rate measurement versus incident photon rate, and the blue curve correspond to the prediction from the hybrid-autocorrelation measurement.

measurement with $I_{drop} = 2.9$ µA and $\tau = 58$ ns fits very well together, giving a high trust in the model and in the predictive power of the method.

If we can reproduce the counting rate curve, then it appears possible to reproduce all the results of any experiment with pulsed or continuous wave light. Therefore we can see the importance of obtaining a complete characterization of the efficiency, as is allowed by our hybrid-autocorrelation method.

IV. CONCLUSION

The method we proposed here provides a fast, simple and most importantly direct characterisation of the recovery of the efficiency of a SNSPD detector. The measurements showed that the recovery of a SNSPD is faster with larger bias current and shorter wavelengths. We demonstrated that the current through a given detector always drop to the same non-zero value after detection even when subjected to several consecutive pulses all arriving within a fraction of the total recovery time of the SNSPD. We also showed that our method can be used to correctly predict how the detection rate of an SNSPD behaves when it becomes impeded by its recovery time. Therefore, we trust our method to allow predicting the behavior of the SNSPD in other experiments where the variation of the efficiency in time is of importance. Finally, it is also worth noting that this method can be applied to any type of single-photon detector, and could be considered as a universal benchmarking method to measure and compare the recovery time of single-photon detectors.

ACKNOWLEDGMENTS

This project was funded from the European Union's Horizon 2020 programme (Marie Skłodowska-Curie grant 675662).

[1] F. Marsili, V. B. Verma, J. A. Stern, S. Harrington, A. E. Lita, T. Gerrits, I. Vayshenker, B. Baek, M. D. Shaw, R. P. Mirin, and S. W. Nam, "Detecting single infrared photons with 93% system efficiency," Nature Photonics **7**, 210–214 (2013).

[2] H. Shibata, K. Shimizu, H. Takesue, and Y. Tokura, "Ultimate low system dark-count rate for superconducting nanowire single-photon detector," Optics letters **40**, 3428–3431 (2015).

[3] B. A. Korzh, Q. Y. Zhao, S. Frasca, J. P. Allmaras, T. M. Autry, E. A. Bersin, M. Colangelo, G. M. Crouch, A. E. Dane, T. Gerrits, F. Marsili, G. Moody, E. Ramirez, J. D. Rezac, M. J. Stevens, E. E. Wollman, D. Zhu, P. D. Hale, K. L. Silverman, R. P. Mirin, S. W. Nam, M. D. Shaw, and K. K. Berggren, "Demonstrating sub-3 ps temporal resolution in a superconducting nanowire single-photon detector," arXiv preprint arXiv:1804.06839 (2018).

[4] M. Caloz, M. Perrenoud, C. Autebert, B. Korzh, M. Weiss, C. Schönenberger, R. J. Warburton, H. Zbinden, and F. Bussières, "High-detection efficiency and low-timing jitter with amorphous superconducting nanowire single-photon detectors," Applied Physics Letters **112**, 061103 (2018).

[5] A. Vetter, S. Ferrari, P. Rath, R. Alaee, O. Kahl, V. Kovalyuk, S. Diewald, G. N. Goltsman, A. Korneev, C. Rockstuhl, and W. H. P. Pernice, "Cavity-enhanced and ultrafast superconducting single-photon detectors," Nano letters **16**, 7085–7092 (2016).

[6] A. Boaron, G. Boso, D. Rusca, C. Vulliez, C. Autebert, M. Caloz, M. Perrenoud, G. Gras, F. Bussières, M.-J. Li, D. Nolan, A. Martin, and H. Zbinden, "Secure quantum key distribution over 421 km of optical fiber," Physical review letters **121**, 190502 (2018).

[7] F. Bussières, C. Clausen, A. Tiranov, B. Korzh, V. B. Verma, S. W. Nam, F. Marsili, A. Ferrier, P. Goldner, H. Herrmann, C. Silberhorn, W. Sohler, M. Afzelius, and N. Gisin, "Quantum teleportation from a telecom-wavelength photon to a solid-state quantum memory," Nature Photonics **8**, 775 (2014).

[8] X. Qiang, X. Zhou, J. Wang, C. M. Wilkes, T. Loke, S. O'Gara, L. Kling, G. D. Marshall, R. Santagati, T. C. Ralph, J. B. Wang, J. L. O'Brien, M. G. Thompson, and J. C. F. Matthews, "Large-scale silicon quantum photonics implementing arbitrary two-qubit processing," Nature photonics **12**, 534 (2018).

[9] L. K. Shalm, E. Meyer-Scott, B. G. Christensen, P. Bierhorst, M. A. Wayne, M. J. Stevens, T. Gerrits, S. Glancy, D. R. Hamel, M. S. Allman, K. J. Coakley, S. D. Dyer, C. Hodge, A. E. Lita, V. B. Verma, C. Lambrocco, E. Tortorici, A. L. Migdall, Y. Zhang, D. R. Kumor, W. H. Farr, F. Marsili, M. D. Shaw, J. A. Stern, C. Abellán, W. Amaya, V. Pruneri, T. Jennewein, M. W. Mitchell, P. G. Kwiat, J. C. Bienfang, R. P. Mirin, E. Knill, and S. W. Nam, "Strong loophole-free test of local realism," Physical Review Letters **115**, 250402 (2015).

[10] H.-L. Yin, T.-Y. Chen, Z.-W. Yu, H. Liu, L.-X. You, Y.-H. Zhou, S.-J. Chen, Y. Mao, M.-Q. Huang, W.-J. Zhang, H. Chen, M. J. Li, D. Nolan, F. Zhou, X. Jiang, Z. Wang, Q. Zhang, X.-B. Wang, and J.-W. Pan, "Measurement-device-independent quantum key distribution over a 404 km optical fiber," Physical Review Letters **117**, 190501 (2016).

[11] M. E. Grein, A. J. Kerman, E. A. Dauler, M. M. Willis, B. Romkey, R. J. Molnar, B. S. Robinson, D. V. Murphy, and D. M. Boroson, "An optical receiver for the lunar laser communication demonstration based on photon-counting superconducting nanowires," in *Advanced Photon Counting Techniques IX*, Vol. 9492 (International Society for Optics and Photonics, 2015) p. 949208.

[12] S. Miki, M. Yabuno, T. Yamashita, and H. Terai, "Stable, high-performance operation of a fiber-coupled superconducting nanowire avalanche photon detector," Optics express **25**, 6796–6804 (2017).

[13] A. J. Kerman, E. A. Dauler, W. E. Keicher, J. K. W. Yang, K. K. Berggren, G. Gol'Tsman, and B. Voronov, "Kinetic-inductance-limited reset time of superconducting nanowire photon counters," Applied physics letters **88**, 111116 (2006).

[14] M. Caloz, B. Korzh, N. Timoney, M. Weiss, S. Gariglio, R. J. Warburton, C. Schönenberger, J. Renema, H. Zbinden, and F. Bussières, "Optically probing the detection mechanism in a molybdenum silicide superconducting nanowire single-photon detector," Applied Physics Letters **110**, 083106 (2017).

[15] J. Yang, A. Kerman, E. Dauler, V. Anant, K. Rosfjord, and K. Berggren, "Modeling the electrical and thermal response of superconducting nanowire single-photon detectors;modeling the electrical and thermal response of su-

perconducting nanowire single-photon detectors," Applied Superconductivity, IEEE Transactions on;Applied Superconductivity, IEEE Transactions on **17**, 581;581 – 585;585 (2007).

Application patent

WO2017193106A1 - Quanta image sensor quantum random number generation

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
09 November 2017 (09.11.2017)

W!P O | PCT

(10) International Publication Number
WO 2017/193106 Al

(51) International Patent Classification:
G06F 7/58 (2006.01)

(21) International Application Number:
PCT/US2017/031456

(22) International Filing Date:
05 May 2017 (05.05.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/332,077 05 May 2016 (05.05.2016) US

(72) Inventors; and

(71) Applicants: **AMRI, Emna** [TN/CH]; Route de Frontenex 43, Chez Sonia Rahban, 1207 Geneve (CH). **FELK, Yacine** [MA/CH]; Rue des Artisans 2C, 1299 Crans-Pres-Celigny (CH). **STUCKI, Damien** [CH/CH]; Rue Vigier 4, 1205 Geneve (CH). **MA, Jiaju** [CH/US]; Gradute Student, 14 Engineering Dr., 26 Ralston Ln., West Lebanon, NH (US). **FOSSUM, Eric, R.** [US/US]; 198 Forest Road, Wolfeboro, NH 03894 (US).

(74) Agent: **ROSSI, David, V.;** Haug Partners, 745 Fifth Avenue, New York, NY 10151 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available):* AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available):* ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM,

(54) Title: QUANTA IMAGE SENSOR QUANTUM RANDOM NUMBER GENERATION

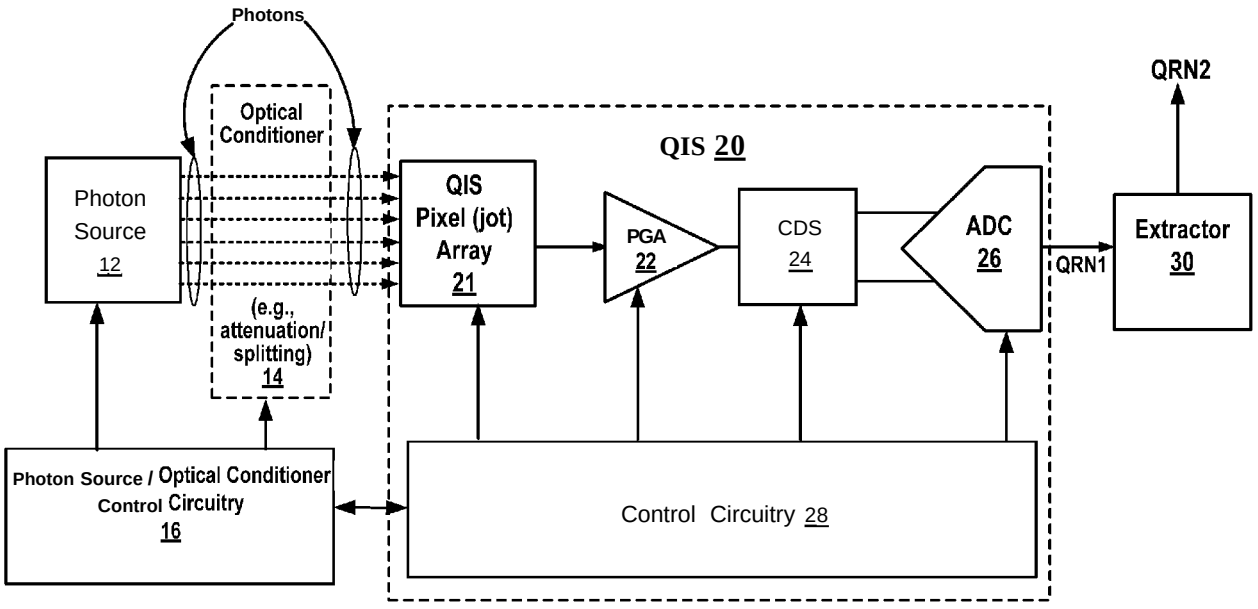


FIG. 1

(57) Abstract: Some embodiments provide methods and apparatus for quantum random number generation based on a single bit or multi bit Quanta Image Sensor (QIS) providing single-photon counting over a time interval for each of an array of pixels of the QIS, wherein random number data is generated based on the number of photons counted over the time interval for each of the pixels.

WO 2017/193106 A1

[Continued on next page]

TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

QUANTA IMAGE SENSOR QUANTUMRANDOM NUMBER GENERATION

RELATED APPLICATIONS

[0001] This application claims the benefit of US Provisional Application No. 62/332,077, filed May 5, 2017, which is hereby incorporated herein by reference in its entirety for purposes of each PCT member state and region in which such incorporation by reference is permitted or otherwise not prohibited.

BACKGROUND

[0002] The present disclosure generally relates to random number generation (RNGn), random number generation using photo-detectors, and more particularly to highly-random, non-deterministic, photon-emission-based random number generation.

[0003] Generating high quality random numbers is becoming more and more important for several applications such as Cryptography, scientific calculations (Monte-Carlo numerical simulations), and gambling. With the expansion of computers' fields of use and the rapid development of electronic communication networks, the number of such applications has been growing quickly. Cryptography, for example, is one of the most demanding applications. It involves algorithms and protocols for ensuring the confidentiality, the authenticity, and the integrity of communications, which requires true random numbers for generating encryption. High-quality random numbers, however, cannot be obtained with deterministic algorithms (e.g., a pseudo-random number generator (PRNG)); instead, an actual physical process may be relied on to generate high-quality random numbers. The most reliable processes are quantum physical processes which are fundamentally random. In fact, the intrinsic randomness of subatomic particles' behavior at the quantum level is one of the few completely random processes in nature. By tying the outcome of a random number generator (RNG) to the random behavior of a quantum particle, it is possible to guarantee a truly unbiased and unpredictable system, which may be referred to as a Quantum Random Number Generator (QRNG).

[0004] The emission of photons is a Poisson process and has been used as the source of randomness in RNGn. Photon detectors used in previous photon-emission-based RNGn technologies include single-photon avalanche diodes (SPADs) and conventional CMOS image sensors (CISs). SPADs can provide single-photon detection capability and realize QRNGn based

on photon quantum effects, but the relatively large size (e.g., 7-20μιη pixel pitch in a SPAD array) limits the data output rate per unit area size. Also, the high dark count rate (e.g., -1000 counts/sec) in SPADs degrades the randomness quality. A conventional CIS is limited by a relatively high noise floor (e.g., > 1 e- r.m.s.) in the readout electronics and does not have single-photon detection capability. In this case, the photon signal is significantly corrupted by read noise, and as read noise is also randomly distributed, the RNGn process using a conventional CIS is not fully quantum-effects based, thus limiting the randomness quality and stability of the output.

[0005] As such, there is a need for further developments and improvements in QRNGs to, for example, provide QRNGs that more fully exploit and/or realize quantum-based randomness. And such developments and improvements may provide for increasing photon-counting accuracy, reducing noise, reducing dark current, increasing the output data rate, and/or increasing scalability.

SUMMARY OF SOME EMBODIMENTS

[0006] To, for example, address at least one or more of the above-described and/or other limitations of QRNGs, some embodiments of the present disclosure provide methods and apparatus for quantum random number generation based on a single-bit or multi-bit Quanta Image Sensor (QIS) providing single-photon counting over a time interval for each of an array of pixels of the QIS, wherein random number data is generated based on the number of photons counted over the time interval for each of the pixels.

[0007] In some embodiments, a QRNG comprises (i) a QIS that includes an array of pixels, wherein each pixel is configured to convert a single photon incident on the pixel into a single photocharge-carrier (an electron or a hole) that is stored in the pixel, and wherein the QIS is configured to readout from each pixel, with single-photocharge-carrier sensitivity (thereby providing for single-photon sensitivity), the photocharge-carriers, if any, stored in the pixel within a time interval, so as to generate a pixel signal (e.g., an analog voltage signal or a digital number/signal) corresponding to the number of stored photocharge-carriers; and (ii) comparison circuitry configured to compare (e.g., in the analog or digital domain), for each pixel, the pixel signal with a threshold level to generate for each pixel a bit having a binary value that depends on whether or not the pixel signal is less than the threshold level or not less than the threshold

level, wherein the binary values are substantially equiprobable based on the threshold level, thereby providing for binary output data having high quality randomness (e.g., with bit entropy ~1).

[0008] All or part of the comparison circuitry may be monolithically integrated with the pixel array of the QIS; in some embodiments, the QIS readout circuitry may embrace or comprise the comparison circuitry.

[0009] In some embodiments, the QRNG may also comprise one or more of (i) a photon source configured to generate the photons incident on the QIS pixel array, (ii) an optical conditioner disposed such that photons emitted by the photon source impinge on the optical conditioner prior to impinging on the pixel array, and (iii) a randomness extractor configured to process data (e.g., the random output data, or digital pixel signals) generated from readout of the QIS.

[0010] In various embodiments, the QRNG may include control circuitry configured to, for example, adjust or control one or more of, the threshold level, the time interval over which the pixel accumulates photocharge-carriers, the photon source emission intensity, and/or the optical conditioner, to maximize the randomness (e.g., according to the bit entropy metric) of the random number data generated by the QIS. In some embodiments, such adjustment control may be based on, for example, monitoring the quanta exposure and/or measuring/monitoring the randomness of the generated random number data.

[001 1] In accordance with some embodiments, the single-bit or multi-bit QIS comprises an array of pixels (e.g., jots), each pixel being configured for photoconversion of an incident photon into a corresponding photocharge (e.g., electron (e-) or hole (h+)), and having sufficient in-pixel conversion gain, without in-pixel avalanche gain, to provide for readout of the photocharge with single-electron sensitivity and resolution, thereby providing for single-photon counting over the time interval. In-pixel conversion gain, according to various embodiments, may be at least 420 μv/charge-carrier (e- or h+), and may be more than 500 μY/ charge-carrier (e- or h+), and may further be more than IOOOμv/ charge-carrier (e- or h+). And, in accordance with various embodiments, the read noise associated with each QIS pixel is about 0.5 charge carriers (e- or h+) rms or less, and may be about 0.3 e- or h+ rms or less, and may further be about 0.15 e- or h+ rms or less. Each QIS pixel may include a charge storage (accumulation)

region configured to store (accumulate) the photocharge that is generated in the pixel over the time interval and that is readout from the pixel following the time interval. The full well charge storage capacity of the pixel storage region may be vary depending on the implementation (e.g., single-bit or multi-bit QIS, conversion gain, voltage limits on readout chain, target threshold level, etc.).

[0012] Throughout the description and claims, the following terms take at least the meanings explicitly associated herein, unless the context dictates otherwise. The meanings identified below do not necessarily limit the terms, but merely provide illustrative examples for the terms. The phrase "an embodiment" as used herein does not necessarily refer to the same embodiment, though it may. In addition, the meaning of "a," "an," and "the" include plural references; thus, for example, "an embodiment" is not limited to a single embodiment but refers to one or more embodiments. Similarly, the phrase "one embodiment" does not necessarily refer the same embodiment and is not limited to a single embodiment. As used herein, the term "or" is an inclusive "or" operator, and is equivalent to the term "and/or," unless the context clearly dictates otherwise. The term "based on" is not exclusive and allows for being based on additional factors not described, unless the context clearly dictates otherwise.

[0013] In addition, as used herein, unless the context clearly dictates otherwise, the term "coupled" refers to directly connected or to indirectly connected through one or more intermediate components and, in some contexts, may also denote or include electrically coupled, such as conductively coupled, capacitively coupled, and/or inductively coupled. Further, "conductively coupled" refers to being coupled via one or more intermediate components that permit energy transfer via conduction current, which is capable of including direct current as well as alternating current, while "capacitively coupled" refers to being electrostatically coupled through one or more dielectric media, and possibly also via one or more intervening conductors (e.g., via a series of capacitive components), that permit energy transfer via displacement current and not via direct current. Those skilled in the art will further understand that elements may be capacitively coupled intentionally or unintentionally (e.g., parasitically) and that in some contexts, elements said to be capacitively coupled may refer to intentional capacitive coupling. In addition, those skilled in the art will also understand that in some contexts the term "coupled" may refer to operative coupling, through direct and/or indirect connection. For instance, a conductor (e.g., control line) said to be coupled to the gate of a transistor may refer to the

conductor being operable to control the gate potential so as to control the operation of the transistor (e.g., switching the transistor between "on" and "off states), regardless of whether the conductor is connected to the gate indirectly (e.g., via another transistor, etc.) and/or directly.

[0014] In this regard, for ease of reference, as used herein, two layers, regions, or other structures/elements may be referred to as being "adjacent" if they do not include one or more intervening layers, regions (e.g., doped regions), or other structures/elements. In other words, two layers, regions, or other structures/elements referred to spatially (e.g., "on," "above," "overlying," "below," "underlying," etc.) with respect to each other may have one or more intervening layers, regions, or other structures/elements; however, use of the term "adjacent" (or, similarly, "directly," such as "directly on," "directly overlying," and the like) denotes that no intervening layers, regions, or other structures/elements are present.

[0015] It will be appreciated by those skilled in the art that the foregoing brief description and the following description with respect to the drawings are illustrative and explanatory of some embodiments of the present invention, and are neither representative nor inclusive of all subject matter and embodiments within the scope of the present invention, nor intended to be restrictive or characterizing of the present invention or limiting of the advantages which can be achieved by embodiments of the present invention, nor intended to require that the present invention necessarily provide one or more of the advantages described herein with respect to some embodiments. Thus, the accompanying drawings, referred to herein and constituting a part hereof, illustrate some embodiments of the invention, and, together with the detailed description, serve to explain principles of some embodiments of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] Aspects, features, and advantages of some embodiments of the invention, both as to structure and operation, will be understood and will become more readily apparent in view of the following description of non-limiting and non-exclusive embodiments in conjunction with the accompanying drawings, in which like reference numerals designate the same or similar parts throughout the various figures, and wherein:

[0017] FIG. 1 schematically depicts a high-level block diagram of an illustrative Quanta Image Sensor (QIS) Quantum Random Number Generator (QRNG), in accordance with some embodiments according to the present disclosure; and

[0018] FIG. 2 depicts an illustrative example of an output signal distribution and binary data conversion in connection with implementing a QIS QRNG, in accordance with some embodiments according to the present disclosure.

DETAILED DESCRIPTION OF SOME EMBODIMENTS

[0019] In accordance with some embodiments according to the present disclosure, random numbers are generated based on a Quanta Image Sensor (QIS) providing single-photon detection of photons emitted from a photon source having Poisson photon-emission statistics, wherein for each QIS pixel (e.g., jot) the number of photons detected by the QIS pixel within a time interval is the quantum random variable used for quantum random number generation (QRNGn). As will be understood by those skilled in the art in view of the present disclosure, such QIS QRNGn embodiments overcome (among other things) limitations of known SPAD-based and conventional CIS-based random number generators (RNGs) (e.g., such as limitations discussed above), and provide for quality quantum random number generation.

[0020] And, more particularly, as will be understood in view of the ensuing description, in some embodiments of a QIS quantum random number generator (QRNG) according to the present disclosure, each pixel (e.g., jot) of the QIS QRNG provides for generating a random number based on a comparison between a threshold (e.g., threshold Ut in hereinbelow illustrative embodiments) and a signal that is generated by reading out the pixel and that corresponds to a number of individual photons detected by the pixel within a given time interval.

[0021] For a given average number of photoelectrons collected in each jot per given time interval (e.g., frame), referred to herein as the quanta exposure (H), the threshold may be selected such that the probability of the signal being less than the threshold is equal (or substantially equal) to the probability of the signal being greater than or equal to the threshold, thus providing for high quality binary quantum random number generation. As will also be understood in view of the present disclosure, in some embodiments, the threshold may be controlled (e.g., dynamically, based on feedback) to ensure quality randomness in view of, for example, temporal fluctuations in the quanta exposure (e.g., due to varying average emission intensity of the photon source). Alternatively or additionally, the quanta exposure may be controlled to ensure quality randomness (e.g., by controlling the photon source emission rate and/or the pixel integration time).

[0022] In addition, in some embodiments the random data generated from the QIS output may be input to and processed by a randomness extractor to provide random number data having further improved randomness qualities. The randomness quality of the random data output by the QIS may require greatly reduced post-processing (e.g., randomness extraction processing) compared to, for example, prior SPAD and CIS based RNGs. And, some embodiments may provide high quality quantum random number generation without requiring a randomness extractor.

[0023] Referring now to FIG. 1, depicted is a schematic, high-level block diagram of an illustrative QIS QRNG, in accordance with some embodiments according to the present disclosure. As shown, the illustrative QIS QRNG embodiment comprises a photon source 12 and an optical conditioner 14 under control of control circuitry 16, a Quanta Image Sensor (QIS) 20, and a randomness extractor 30. Photon source 12, conditioner 14, and QIS 20 are configured (e.g., according to position/alignment, and/or use of reflection, optical waveguiding (e.g., using an optical fiber or other waveguide structure), and/or other optical components) such that photons emitted by source 12 impinge via optical conditioner 14 on a QIS pixel array 21 of QIS 20. In various embodiments, QIS QRNG may be implemented monolithically (e.g., formed on a common semiconductor substrate), or as two or more separate chips (e.g., die) or other components. For instance, in some embodiments, photon source 12 may be formed on a first die, QIS 20 may be formed as a backside-illuminated imager on a second die, and extractor 30 (and possibly additional processing and/or buffering circuitry) may be formed on a third die, with the first, second, and third dies vertically stacked (in sandwich-like fashion) and integrated.

[0024] Photon source 12 may be implemented as any of various optical sources that emits photons according to Poisson statistics, such as one or more light emitting diodes (LEDs; e.g., an silicon (Si) LED device or Si LED array), or one or more laser diodes (e.g., driven with above-threshold drive current). In such embodiments, for example, the intensity of the photon signal emitted by photon source 12 may be controlled by circuitry 16 according to the LED or laser drive current. Alternatively or additionally, the photocarrier (e.g., photoelectron) rate generated in the QIS may be adjusted based on the relative location of the photon source to the pixel array, and in some embodiment this relative location is configured to be controllable/adjustable (e.g., manually and/or automatically (e.g., without user input)). In some embodiments, however, the photon source 12 (and any photon source control circuitry) may be

independent of the QIS QRNG apparatus; for example, in some such embodiments photon source 12 may be an independent photon source, such as a source of ambient light that may be detected by QIS 20. In other words, in such embodiments, the photon source may not be considered as being part of the QIS QRNG apparatus, although QIS QRNG embodiments according to the present disclosure may be configured to include photon source 12 as well as its drive/power and/or control circuitry.

[0025] Optical conditioner 14, which is an optional component, may be included to provide additional control over the photon signal that impinges on the QIS array 21. For example, optical conditioner may be a controllable attenuator, splitter, or the like.

[0026] QIS 20 is schematically depicted as comprising a pixel (e.g., jot) array 21, a readout chain comprising a programmable gain amplifier (PGA) 22, correlated-double-sampling (CDS) circuitry 24, and analog-to-digital converter (ADC) 26, all under control of control circuitry 28, which may also be coupled with control circuitry 16 for purposes of coordinated control of photon source 12, optical conditioner 14, and QIS 20 in providing a quality quantum random number data signal QRNl output from QIS 20 (e.g., by controlling H and/or Ut, as will be further understood below). It will be understood that the simplified block diagram of QIS 20 is set forth for clarity of exposition in describing the operation of the QIS QRNG with respect to readout of an individual pixel (jot) within the QIS pixel array 21, which may comprise around a billion or more (e.g., several billion) sub-diffraction limit pixels that may be read out row-wise in column-parallel manner, with each column of pixels being associated with a readout chain comprising PGA, CDS, and ADC circuitry (though, e.g., in some embodiments all columns may not be read out in parallel simultaneously as groups of two or more columns may share readout chain circuitry (e.g., such as sharing at least an ADC)). In addition, for example, QIS 20 may comprise additional circuitry, such as an output buffer and/or image/data processing circuitry coupled between the output of ADC 26 and input to extractor 30 (e.g., by way of non-limiting example, data QRNl may be input to a buffer that is accessible to image/data processing circuitry that is configured to process QRNl data and write the processed data back to the buffer for output to extractor 30).

[0027] Depending on the implementation, QIS 20 may be a single-bit QIS or multi-bit QIS. Each pixel/jot of QIS 20 has single-electron sensitivity (e.g., ~0.15e- r.m.s.) which may be

obtained from high, in-pixel conversion gain, e.g., more than 500 μv/e-, and more than IOOOμv/e- in some embodiments. As described, QIS 20 may comprise at least one billion jots (at least 1 G-jot, such as several G-jots), though some embodiments may employ less than 1 G-jot (e.g., -0.1 G-jots or more). And the readout speed may be more than lOOOfps, which yields an output data rate of (e.g., for a single-bit QIS) about lOOGb/s to more than 1 Tb/s (e.g., several Tb/s). Depending on the application, the output data rate may be varied according to the number of jots in the QIS array and/or the readout scan rate may be varied.

[0028] The QIS jots may be implemented as pump-gate jots; however, any suitable jot device for implementing a single-bit or multi-bit QIS (e.g., having sufficient conversion gain for single photocarrier detection) may be employed. Additional aspects and details concerning implementations of a QIS in a QIS QRNG in accordance with embodiments of the present disclosure may be understood by those skilled in the art in view of, for example, each of the following publications, each of which is hereby incorporated by reference herein in its entirety: (i) PCT international application publication no. WO/2015/153806 (corresponding to PCT international application no. PCT/US20 15/023945), "CMOS Image Sensor with Pump Gate and Extremely High Conversion Gain," published Oct. 8, 2015, (ii) J. Ma and E.R. Fossum, A Pump-Gate Jot Device with High Conversion Gain for Quanta Image Sensors, IEEE J. Electron Devices Society, vol. 3(2), pp. 73-77, March 2015, (iii) J. Ma and E.R. Fossum, Quanta image sensor jot with sub 0.3e- r.m.s. read noise and photon counting capability, IEEE Electron Device Letters, vol. 36(9), pp. 926-928, September 2015, (iv) J. Ma, D. Starkey, A. Rao, K. Odame, and E.R. Fossum, Characterization of quanta image sensor pump-gate jots with deep sub-electron read noise, IEEE J. Electron Devices Society, vol. 3(6), pp. 472-480, November 2015, and (v) S. Masoodian, A. Rao, J. Ma, K. Odame and E.R. Fossum, A 2.5pJ/b binary image sensor as a pathfinder for quanta image sensors, IEEE Trans. Electron Devices, vol. 63(1), pp. 100-105, January 2016.

[0029] As will be understood, readout of the jots in the QIS array 21 is analogous to readout of accumulated charge from pixels in conventional CISs. During readout, the jot output signal (e.g., output from an in-jot source-follower amplifier, not shown) corresponding to the charge accumulated in the jot may be coupled to a column bus (e.g., corresponding to the input to PGA 22), resulting in a corresponding analog signal being coupled to the input of ADC 26 via

PGA 22 and CDS 24 circuitry. ADC 26 converts the input analog signal into an n-bit digital signal.

[0030] In a single-bit QIS, the bit width (n) is one (1), and the binary output of the ADC corresponds to whether or not the analog signal input to the ADC 26 from CDS 24 (the "ADC signal input") is less than Ut or not less than Ut. As described above, and as may be further understood in view of the ensuing disclosure (as well as the Appendix of priority US Provisional Application No. 62/332,077, filed May 5, 2017, which is hereby incorporated herein by reference in its entirety), Ut may be selected such that these cases are equiprobable, thus providing for the binary output (e.g., QRN1) having high quality randomness (e.g., with bit entropy ~1) based on the quantum optical randomness of the photon source.

[003 1] In some multi-bit QIS embodiments, the bit width (n) may be an integer value between, for example, two and about 6 (e.g., $\mathbf{1} \leq n \leq 6$). In some such embodiments, the LSB may correspond to one photoelectron. It will be understood, however, that in various alternative embodiments, it is also possible to configure the ADC such that the LSB is less than the equivalent of one photoelectron (e.g., 0.2 electrons). Some multi-bit QIS embodiments may employ more than 6 bits, and the DN output by the ADC may be linearly scaled over the range of the analog signal range input to the ADC based on the number of photoelectrons that can be detected/counted by the jot, the readout noise, and the gain (e.g., jot conversion gain, PGA gain).

[0032] In some embodiments, control circuitry 28 may provide threshold Ut as an analog signal to ADC 26 (e.g., control circuitry 28 may convert a digital Ut signal to the analog Ut signal; or Ut may originate as an analog signal in control circuitry 28), which may compare the threshold Ut in the analog domain to the analog signal input to the ADC 26 from CDS 24 (the "ADC signal input"), and output a binary value (e.g., "0" or "1") according to whether the ADC signal input is less than Ut or not less than Ut. Similarly, such analog domain comparison may be implemented wherein control circuitry 28 provides a digital Ut value to ADC 24, which may comprise digital-to-analog converter (DAC) circuitry to convert Ut to an analog signal.

[0033] In some embodiments, the comparison with Ut may be executed in the digital domain. For example, ADC 26 may convert the ADC signal input into an multi-bit digital number (DN). That multi-bit digital number may be compared with a digital representation of Ut (e.g., which may be generated from an analog Ut signal, or may originate as a digital value), and

a binary value may be generated based on whether the DN is less than or not less than Ut. It will be understood that such digital comparison may be implemented within the QIS (e.g., in circuitry following the ADC output (not shown); or, in some implementations, such circuitry may also be embodied in (or considered as being logically part of) ADC 26). Alternatively, for example, such digital comparison circuitry may be implemented external to (e.g., off-chip) from the QIS, and may be embodied within the extractor 30 in some embodiments.

[0034] Accordingly, in some multi-bit QIS embodiments, the output of the ADC may be a multi-bit digital number (DN) (e.g., representing the ADC input signal) that may be provided to additional circuitry to generate a single-bit random number bit stream based on comparison with a threshold value. And, in some multi-bit QIS embodiments, the output of the ADC may be a single-bit random number bit stream (e.g., where the ADC incorporates digital-domain comparison circuitry).

[0035] As noted, in some embodiments, such post-ADC digital-domain comparison circuitry may be embodied in randomness extractor 30, which may further process the random number bit stream to provide a random number bitstream QRN2 having improved randomness using techniques known to those skilled in the art (e.g., compression algorithms based on hashing and/or matrix multiplication). In some embodiments, however, randomness extractor 22 may not be required (and thus may not be included as part of QIS QRNG). For example, QIS QRNG may be configured to periodically or aperiodically adjust/control one or more of, for example, the threshold value (Ut), the jot detection time interval, and the photon source emission intensity to maximize the randomness (e.g., according to the bit entropy metric) of the generated random number data (e.g., QRN1). In some embodiments, such adjustment control may be based on, for example, monitoring the quanta exposure and/or measuring/monitoring the randomness of the generated random number data.

[0036] Design and operational principles for implementing a QIS QRNG according to some embodiments of the present disclosure may be further understood in view of the following, as well as in view of the Appendix of priority US Provisional Application No. 62/332,077, filed May 5, 2017, which is hereby incorporated herein by reference in its entirety.

[0037] In a QIS (as in a conventional CIS), the photon signal is converted to a voltage signal in one pixel/ jot and corrupted with noise in the readout chain. The distribution of the

output signal is a convolution between the Poisson distribution of the arrival of photoelectrons and a normal distribution of noise. An example of signal distribution is shown in FIG. 2. The average rate of photoelectrons is defined as quanta exposure $\boldsymbol{H}$. In a single-bit QIS, an artificial threshold $\boldsymbol{U_t}$ (e.g., 0.5e-) is set in the readout chain to convert the output signal to binary data: output signal higher than $\boldsymbol{U_t}$ will be converted into "1" and to "0" when it is below $\boldsymbol{U_t}$. The probability of "1" state is:

$$P[U < U_t] = \sum_{k=0}^{\infty} \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{U_t - k}{u_n\sqrt{2}}\right)\right] \cdot \frac{e^{-H}H^k}{k!}$$

where u_n is the read noise of the sensor, and the probability of "0" state is:

$$\boldsymbol{P[U \geq U_t] = 1 - P[U < U_t]}$$

[0038] Given a proper quanta exposure $\boldsymbol{H}$ (e.g. $\boldsymbol{H}$<1), a 1-bit random number can be generated from one readout of one jot. The QIS based random number generator can include a QIS device and a stable light source (e.g., such as described hereinabove in connection with embodiments according to FIG. 1 and some variations thereof). An ideal random number generator is expected to generate "0s" and "Is" with equal probability; otherwise, an extractor may need to be applied to select the useful data. The minimum entropy indicates the percentage of useful data, which is given by:

$$\boldsymbol{Smin} = -\log_2[\max(P[U \geq U_t], P[U < U_t])]$$

[0039] An entropy close to 1 is ultimately desired. To achieve that, a quanta exposure $\boldsymbol{H} = \log_e(2)$ may be set up by the illumination condition (e.g., light source, packaging, QIS integration time).

[0040] As noted above, further description of a QIS QRNG, including its principles of operation, according to some embodiments of the present disclosure is presented in the Appendix of priority US Provisional Application No. 62/332,077, filed May 5, 2017, which is hereby

incorporated herein by reference in its entirety, which Appendix is set forth as an article entitled " Quantum Random Number Generation Using Quanta Image Sensor," which illustrates some embodiments of the present invention as well as various features and advantages that may be associated with some embodiments, and is not intended to limit the present invention.

[0041] In view of the present disclosure, it will be understood that a QIS QRNG provides many features and advantages that, among other things, overcome limitations of known SPAD and conventional CIS based RNGs. For example, as discussed, a QIS may include, for example, lOOMjots to one or more Giga-jots having photon-counting capability, with the jot array having submicron pitch (e.g., 200nm-500nm), and the QIS can be readout at a high frame rate (lOOOfps). Accordingly, these features (e.g., small jot size and high speed) provide the QIS QRNG with extremely high data output rate. And the photon-counting capability of jot device can ensure the QRNG is fully photon quantum effects based. Further, the low dark current (e.g., 0.1 e-/sec at room temperature) provides for improved randomness quality and stability. In short, some embodiments of a QIS- based QRNG device provides for, among other things, high data rate (e.g., 5-12 Gb/s), low dark current error, and high stability.

[0042] Accordingly, although the above description of illustrative embodiments of the present invention, as well as various illustrative modifications and features thereof, provides many specificities, these enabling details should not be construed as limiting the scope of the invention, and it will be readily understood by those persons skilled in the art that the present invention is susceptible to many modifications, adaptations, variations, omissions, additions, and equivalent implementations without departing from this scope and without diminishing its attendant advantages. For instance, except to the extent necessary or inherent in the processes themselves, no particular order to steps or stages of methods or processes described in this disclosure, including the figures, is implied. In many cases the order of process steps may be varied, and various illustrative steps may be combined, altered, or omitted, without changing the purpose, effect or import of the methods described. Similarly, the structure and/or function of a component may be combined into a single component or divided among two or more components. It is further noted that the terms and expressions have been used as terms of description and not terms of limitation. There is no intention to use the terms or expressions to exclude any equivalents of features shown and described or portions thereof. Additionally, the present invention may be practiced without necessarily providing one or more of the advantages

described herein or otherwise understood in view of the disclosure and/or that may be realized in some embodiments thereof. It is therefore intended that the present invention is not limited to the disclosed embodiments but should be defined in accordance with claims that are based on the present disclosure, as such claims may be presented herein and/or in any patent applications claiming priority to, based on, and/or corresponding to the present disclosure.

What is claimed is:

1. A quantum random number generator(QRNG), comprising:

a Quanta Image Sensor (QIS) comprising a pixel array, wherein each pixel of the QIS is configured to convert photons emitted from a photon source into charged photocarriers, wherein the QIS is configured to readout each pixel to provide a signal representing a count of the number of photocarriers with single-photocarrier sensitivity; and

wherein the QRNG is configured to output random number data having randomness based on the number of collected photocarriers within a time interval.

2. The QRNG according to claim 1, wherein for each pixel the number of photocarriers collected within the time interval is converted to a voltage and then into binary signal using a threshold level.

3. The QRNG according to claim 2, wherein the voltage is compared to the threshold level in the analog domain.

4. The QRNG according to claim 2, wherein the conversion is performed in the digital domain, wherein the photocarrier signal is converted to a digital signal or digital number (DN) by an ADC which has a bit depth higher than 1-bit, and the digital signal or digital number is converted to a 1-bit random number.

4. The QRNG according to any of the preceding claims, wherein the photon source intensity is tunable to realize an ideal randomness entropy of the random number data and/or to realize greater than a minimum value of the randomness entropy of the random number data.

5. The QRNG according to any of the preceding claims, wherein the photocarrier collection rate and the threshold level are tunable to realize an ideal randomness entropy of the random number data and/or to realize greater than a minimum value of the randomness entropy of the random number data.

6. The QRNG according to any of the preceding claims, wherein the photocarrier rate is capable of being adjusted based on the relative location of the photon source to the pixel array.

7. The QRNG according to any of the preceding claims, wherein the threshold level can be adjusted by a reference voltage supplied by an on-chip or off-chip DAC.

8. The QRNG according to any of the preceding claims, wherein the generated random number data's adjustable levels are periodically or aperiodically reset to maximize randomness entropy.

9. The QRNG according to any of the preceding claims, wherein one or more of the following are periodically or aperiodially adjusted to maximize randomness entropy of the random number data: the time interval over which photocarriers are collected, the photon source intensity, and the threshold level used to determine the value of the binary output.

10. The QRNG according to any of the preceding claims, wherein the QRNG includes the photon source.

11. The QRNG according to any of the preceding claims, wherein the QRNG includes a randomness extractor.

12. The QRNG according to any of the preceding claims, wherein the QRNG includes an optical conditioner disposed such that photons emitted by the photon source impinge on the optical conditioner prior to impinging on the pixel array.

13. A QRNG, comprising:

a Quanta Image Sensor (QIS) comprising an array of jots that are each configured to provide single-photon detection of photons emitted from a photon source having Poisson photon-emission statistics; and

wherein the QRNG is configured to output random number data, wherein for each jot the number of photons detected by the jot within a time interval is the quantum random variable used for generation of the random number data.

14. A QRNG comprising:

a QIS that includes an array of pixels, wherein each pixel is configured to convert a single photon incident on the pixel into a single photocharge-carrier that is stored in the pixel, and wherein the QIS is configured to readout from each pixel, with single-photocharge-carrier sensitivity, the photocharge-carriers, if any, stored in the pixel within a time interval, so as to generate a pixel signal corresponding to the number of stored photocharge-carriers; and

comparison circuitry configured to compare for each pixel, the pixel signal with a threshold level to generate for each pixel a bit having a binary value that depends on whether or not the pixel signal is less than the threshold level or not less than the threshold level, wherein the binary values are substantially equiprobable based on the threshold level, thereby providing for binary output data having high quality randomness.

15. The QRNG according to claim 14, further comprising one or more of (i) a photon source configured to generate the photons incident on the QIS pixel array, (ii) an optical conditioner disposed such that photons emitted by the photon source impinge on the optical conditioner prior to impinging on the pixel array, and (iii) a randomness extractor configured to process data generated from readout of the QIS.

16. The QRNG according to claim 14 or 15, wherein the QRNG includes control circuitry configured to adjust or control at least one of (i) the threshold level, (ii) the time interval, (iii) the photon source emission intensity, and (iv) the optical conditioner, to maximize the randomness of the random number data generated by the QIS.

17. The QRNG according to any one of claims 14 to 16, wherein each pixel has sufficient in-pixel conversion gain, without in-pixel avalanche gain, to provide for readout of the photocharge with single-electron sensitivity and resolution.

18. The QRNG according to any one of claims 14 to 17, wherein the read noise associated with each QIS pixel is at least one of about 0.5 charge carriers rms or less, about 0.3 charge carriers rms or less, and about 0.15charge carriers rms or less.

19. A method for quantum random number generation, the method comprising:

generating for each of a plurality of pixels of a single-bit or multi-bit Quanta Image Sensor (QIS) a signal representing the number of individual photons incident on the pixel over a time interval; and

generating random number data based on the signals representing the number of photons detected over the time interval for each of the pixels.

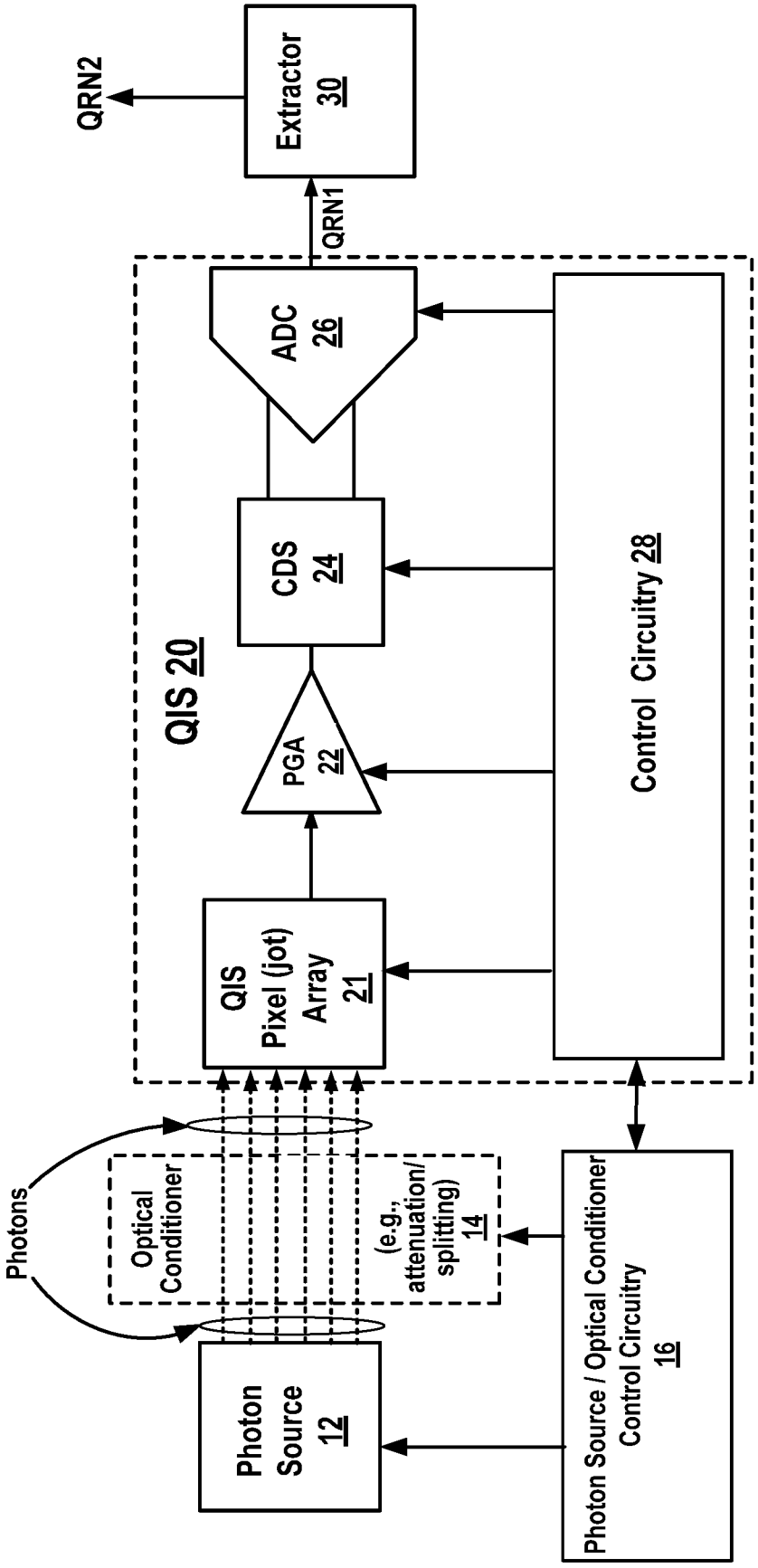

QRN2
Extractor 30
QRN1
ADC 26
CDS 24
PGA 22
QIS Pixel (jot) Array 21
QIS 20
Control Circuitry 28
Photons
Optical Conditioner
(e.g., attenuation/ splitting) 14
Photon Source 12
Photon Source / Optical Conditioner Control Circuitry 16

FIG. 1

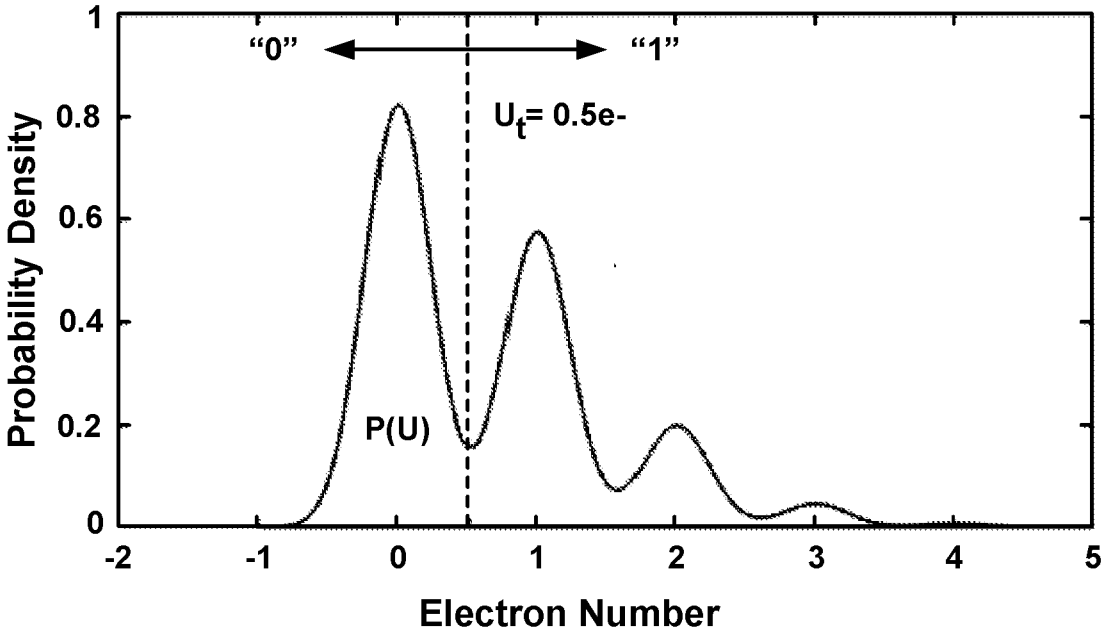

"0"
"1"
U_t= 0.5e-
P(U)
Probability Density
Electron Number
1
0.8
0.6
0.4
0.2
0
-2
-1
0
1
2
3
4
5

FIG. 2

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US17/31456

A. CLASSIFICATION OF SUBJECT MATTER
IPC - G06F 7/58 (2017.01)
CPC - G06F 7/58, 7/588; G06N 99/002

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
See Search History document

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
See Search History document

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
See Search History document

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2015/0261502 A1 (SONY CORPORATION) 17 September 2015; paragraphs [0021], [0023], [0028], [0044], [0063], [0066], [0083], [0084], [0088]	1-3, 4A, 4B/1-4B/3, 4B/4A, 13-15, 16/14, 16/15, 19
A	US 2012/0075134 A1 (ROGERS, D et al.) 29 March 2012; entire document	1-3, 4A, 4B/1-4B/3, 4B/4A, 13-15, 16/14, 16/15, 19
A	US 2014/0287816 A1 (NOVOMATIC AG) 25 September 2014; entire document	1-3, 4A, 4B/1-4B/3, 4B/4A, 13-15, 16/14, 16/15, 19

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
29 June 2017 (29.06.2017)	**03 AUG 2017**
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (January 2015)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 17/3 1456

Box No. [I Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful **intRrnarinnal** search can he carried out. specifically:

3. ☒ Claims Nos.: 5-12, 17-18
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. Ill Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.