# A No-History, Low Latency Photonic Quantum Random Bit Generator for use in a Loophole Free Bell Tests and general applications

Mario Stipčević[1,*], Rupert Ursin[2]
[1]Photonics and Quantum Optics unit of the Center of Excellence for Advanced Materials and Sensing Devices, Ruđer Bošković Institute, Bijenička 54, 10000 Zagreb, Croatia
[2]Institute for Quantum Optics and Quantum Information,
Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria
*E-mail: stipcevi@irb.hr

**Published as**: Stipčević M., Ursin R. (2020) A No-History, Low Latency Photonic Quantum Random Bit Generator for Use in a Loophole Free Bell Tests and General Applications. In: Kollmitzer C., Schauer S., Rass S., Rainer B. (eds) Quantum Random Number Generation. Quantum Science and Technology. Springer, Cham. DOI 10.1007/978-3-319-72596-3_5

**Abstract**. Random numbers are essential for our modern information-based society. Unlike frequently used pseudo-random generators, physical random number generators do not depend on deterministic algorithms but rather on a physical process to provide true randomness. In this work we present a conceptually simple optical quantum random number generator that features special characteristics necessary for application in a loophole-free Bell inequality test, namely: (1) very short latency between the request for a random bit and time when the bit is generated; (2) all physical processes relevant to the bit production happen after the bit request signal; and (3) high efficiency of producing a bit upon a request (100% by design). This generator is characterized by further desirable characteristics: ability of high bit generation rate, possibility to use a low detection-efficiency photon detector, a high ratio of number of bits per detected photon (≈2) and simplicity of the bit generating process. Generated sequences of random bits pass NIST STS test without further postprocessing.

## Introduction

Digital data processing in computers, mobile devices, ATM machines etc., does have a huge impact on our information-based society. Digital processing is strictly deterministic. But sometimes randomness is required. Ability to generate random numbers is required for cryptographic protocols which are necessary to ensure digital security, privacy and integrity of communicated data, as well as for many other digital applications including but not limited to: internet trade, crypto currency, cloud computing, e-banking, secure e-mail access, online gambling, Monte Carlo modeling of natural phenomena, randomized algorithms and scientific research. While computers can generate long sequences of numbers that have good statistical properties via so-called pseudo-random algorithms, such number sequences remain deterministic and thus predictable.

In contrast to computational methods used by pseudo-random number generators, physical random numbers generators derive random numbers from a physical source of a reasonably random process e.g. flipping a coin. However, systems relying on classical motion actually do have a component of deterministic prediction that will be transferred to the random numbers obtained thereof. On the other extreme is the quantum theory (also known as Quantum Mechanics or QM for short), a branch of physics that strives to understand and predict the properties and behavior of tiny objects, such as elementary particles. One intriguing aspect of QM is that properties of a particle are not determined with arbitrary precision until one measures them, consequently the individual result of a measurement contains an inevitable intrinsic random component. This characteristic of the quantum theory provides fundamental randomness that can be used for generating true random numbers. While QM allows for completely random number generators, in practice "the devil is in the detail" of practical realization: whether a certain part is doing what it is supposed to do in theory and with what precision. It is therefore crucial to investigate and build such parts and bit generating methods that come as close as possible to their theoretical ideal.

Quite generally, quantum random number generators (QRNG) can be divided into two broad categories depending on their type of operation: firstly *continuous* which produce random numbers at their own pace and secondly *triggered* which produce a single random number (eg. one bit or one set of bits) upon a request, as illustrated in Fig. 1. Both, continuous and triggered RNGs feature the Strobe output which generates a short logic pulse when the new random bit is available at the Random Bit output. Additionally, the triggered type features a Trigger input. When a pulse is sent to that input it triggers a series of physical events and measurements - resulting in generation of a new random bit. Examples of continuous generators include those that extract random numbers from time-wise random events such as radioactive decay [1], photon arrival [2], or beamsplitter based [3-4] RNG's. Examples of a triggered RNG include sampled time-wise random toggling flip-flop [5-6] (bit generating probability equal to 1) and beamsplitter with a pulsed or on-demand single-photon light source [4] (bit gen. probability < 1). An important consideration is the latency between a moment of trigger and the moment when the random bit is available for readout (technically the delay between the Trigger and the Strobe pulses).

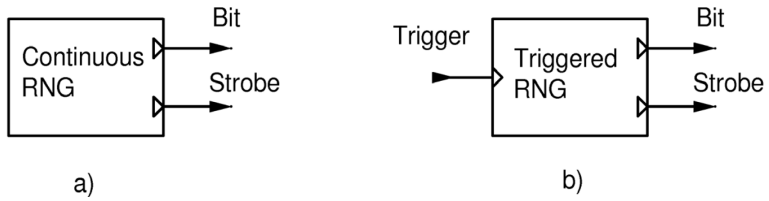


Fig. 1. Continuous (a) and triggered (b) random number generators.

An interesting further requirement does come from experimental loophole-free Bell inequality tests. Bell test allows distinguishing quantum mechanics from local hidden variable theories. These experiments are also quite important for future implementation of quantum key distribution devices [7]. Experimental tests performed so far do suffer from so called "loopholes" [8]. In order to close the "locality" [9] as well as the "freedom-of-choice" loophole [10] one needs to decide on random setting of detection basis by means of a RNG that satisfies three properties: (1) all physical processes required for production of a bit must happen completely *in the future* of the trigger, that is anything that happened before the trigger

must not have any influence on the generated bit value; (2) a random bit is produced upon a request with certainty within a *bounded* time ($\tau_L$); (3) in order to enable realistic experimental implementation of a loophole-free Bell test, including detection loophole [11], [12], the delay $\tau_L$ must be shorter than qubit flight times from the production to the detection sites which is typically a few tens of nanoseconds. None of the generators or generating principles known so far satisfies all those requirements simultaneously to that extent.

This work is based on our previous research of a QRNG whose randomness can be brought close to theoretical perfection by suitable tuning of the device's controllable parameters in order to minimize effect of the hardware imperfections [32] and thus mitigate the "devil in the detail" problem. This particular QRNG is unique in simultaneously satisfying three characteristics mentioned above, having a 100% efficiency of producing a bit upon a request by design and a latency $\tau_L = (9.8 \pm 0.2)$ ns. Generated bits pass the NIST Statistical Test Suite (STS) [9]. All this make this QRNG suitable for even the most demanding applications, including the loophole-free Bell test.

**Concept of a low-latency QRNG**

Our generator, shown in Fig. 2, comprises: a bit request input (Trigger Input), a laser diode (LD), a single photon detector (PD), and a coincidence circuit consisting of a single AND gate. It functions in the following way. The external trigger signal causes LD to emit a short (sub-nanosecond) light pulse. We define that one random bit is generated upon *every* trigger signal. The value of the random bit is defined as the state of the detector's output at the moment of positive-going edge of the synchronous Strobe signal which is derived from the Trigger signal by a suitable delay (latency). Note, if emission and detection of light were classical processes then detection would either happen every time (if pulse energy is higher than some given threshold) or never (if below the threshold). However, due to the quantum nature of light, detection of a photon arising from the laser pulse is a binomial process with success probability $p_1$ that can take on any value in the range [0, 1]. The energy of the light pulse falling upon the detector is carefully set such that the probability $p_1$ of detecting a photon (and thus generating a value of “1”) is as close as possible to the ideal value of $p_1 = 0.5$. We assumed that the laser is stable in power and the detector's efficiency is constant during the measurement time. Note, the detection efficiency of the chosen PD is irrelevant since it is always possible to set pulse power such that the above condition is met. This is in contrast with e.g. pulsed beam-splitter method [4] where efficiency of detector affects the bit generation rate and thus it can never reach unity. In this device however, for each and every trigger signal we get a bit from the QRNG, hence we call the device 100% efficient.

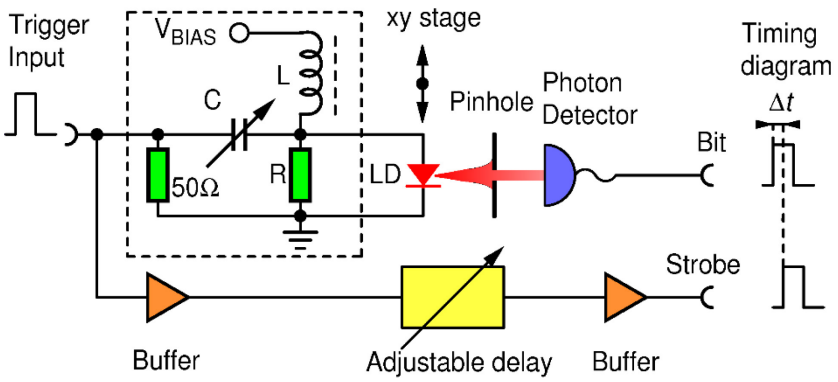


Fig. 2. Conceptual schematic diagram of the low-latency quantum random number generator.

Under the assumption that both, the light source and the detector, are completely reset to their initial conditions between subsequent triggers, it is impossible for generated bit values to "communicate", i.e. influence each other. Consequently there would be no correlation among successive bits. We will elaborate later how this assumption can be guaranteed in practice. Having these two characteristics (probability of ones equal to 0.5 and absence of correlation among successive bits) a pool of generated bits has no other possibility than to be random. Namely, according to min-entropy theory, laid out in Ref. [31], a sufficient condition for a RNG to generate truly random bits is that it generates any $n$-bit string with an *a priori* probability of $1/2^n$. Now, for $n = 1$ this is simply a condition that probability of ones is equal to $1/2$, which is probably the most intuitive characteristic of a random bit string. Interesting thing happens for $n \geq 2$ where the RNG must have (at least) $n$ bits of memory to store the substring and be able to recognize it as one that needs to be generated with probability different than some other substring of the same length. Note that this type of behavior is absolutely impossible without a memory. But what if we make sure there is no memory in our RNG and yet engineer it such that it generates ones and zeros with equal probability? To the extent to which we can make these

We conclude that for the generator depicted n Fig. 2., in principle, a bit generated upon a trigger has no history prior to that trigger because all relevant physical processes, namely: (1) powering of the laser diode and subsequent light pulse emission, (2) photon detection and (3) detector-strobe coincidence, are all happening *after* the trigger. In practice, we will make sure that it has no memory either. The bit-generating efficiency of the method is high: it yields two random bits per photon detection as compared to $\leq 1$ bit for beamsplitter [4] and $\leq 0.5$ for arrival-time [2] methods. Even though this high efficiency does not allow for higher bit generation rate, because the ultimate rate is bounded by inverse of the dead time, it does put a less strain to the detector reducing its power consumption and possibly extending its lifetime.

**Experimental setup**

**Sub-nanosecond pulsed laser**

In the experimental realization of the QRNG shown in Fig. 2, light pulses are obtained from a single mode laser diode LD (Sony DL3148-025, 650 nm). The laser diode is driven by a sub-nanosecond current pulse formed by a simple RLC circuit upon each positive-going edge of the trigger pulse. Passive driver design

ensures smallest delay between the driving electrical pulse and the light pulse. Coarse adjustment of the energy of light pulses is made by the variable capacitor C. The RLC network circuit and the laser diode are mounted on an XY translation stage and can move relative to a 50 μm pinhole placed in front of the photon detector thus allowing for a fine tuning of the pulse detection probability $p_1$. The attenuation of light is performed by means of geometric misalignment between the laser mode and the aperture of the pinhole. The goal of this adjustment is to have $p_1$ as close as possible to the ideal value of 0.5. The energy of the light pulse also depends on the bias voltage $V_{BIAS}$ which, in principle, allows for automatic bias zeroing via a negative feedback loop. The simplicity of the electrical and mechanical designs is intended to minimize the time lapse between a trigger pulse and the arrival of the optical pulse to the single-photon detector.

The optical pulse from the laser diode circuit, shown in Fig. 2, features a jitter of 190 ps FWHM with respect to the trigger raising-edge. In order to avoid degradation of pulse power and shape, shortest period between two consecutive triggers should be $\geq$ 40 ns. The combined delay between the trigger input and photon detector output corresponding to detected photon(s) from the light pulse is (6.5 +- 0.2) ns and has a jitter of 370 ps FWHM for detectors with SLiK SPAD, as shown in Fig. 3, and about 750 ns FWHM for detectors with SUR500. The SUR500 diode has significantly smaller diffusion tail than SLiK diode, thus in both cases virtually all detection pulses are contained within 8.5 ns delay from the laser trigger pulse.

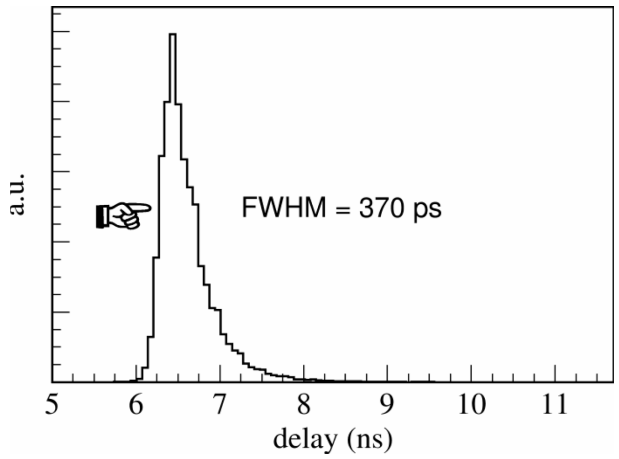


Fig. 3. Time profile of the optical pulse emitted by the laser diode convoluted with the jitter of a detector with a SLiK SPAD. The main part has full width at half maximum of 370 ps. Virtual all pulses are emitted within ≈2 ns.

The laser can be triggered at will with a shortest period between two consecutive pulses of about 40 ns (maximum 25 MHz repetition rate). At shorter delays pulse power and shape degrades.

**Single-photon detectors**

For this study we use two types of home-made single-photon detectors which differ in the silicon single-photon avalanche photodiode (SPAD) that is used as a sensor. Each detector consists of a silicon single-photon avalanche diode (SPAD) operated in Geiger mode and actively quenched. We will show how different characteristics of the detectors affect quality of the generated random numbers.

The distinctive and important characteristic of the avalanche quenching circuits (AQC) used in this study is that the delay between the avalanche and the output pulse (the detection delay) is quite small, equal to about 6.5 ns and that they contain an integrated pulse shaping circuit, shown in Fig. 4, which allows setting the output pulse to any value between 8 ns and 50 ns by means of potentiometer P1, without changing the detection delay.

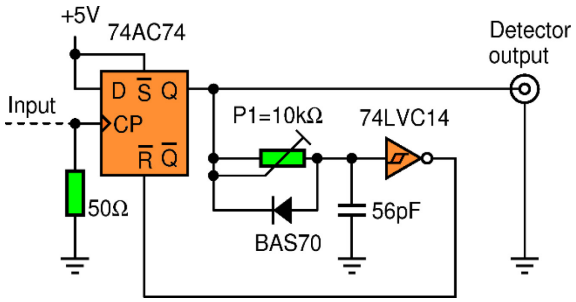


Fig. 4. Pulse stretching/blanking part of the avalanche quenching circuit.

The first detector type makes use of SPADs recovered from PerkinElmer SPCM-AQR modules, also known as "SLiK". We have built two detectors of this type. Active quenching circuit is a modification of the AQC described in Ref. [9] optimized for this particular SPAD type. A photodiode is operated at -10 °C and excess voltage of 19 V. Characteristics of this detector type are: dead time $\tau_{dead} = 22$ ns, output pulse width $\tau_{pd}$ = 8-50 ns (adjustable), detection efficiency of 71% at 650 nm, jitter of 320 ps FWHM and dark counts of 200 cps and 750 cps for each detector respectively.

The second detector type makes use of a silicon avalanche photodiode SUR500 manufactured by Laser Components. Even though the manufacturer states that this SPAD cannot be used for photon counting in photon-counting Geiger mode, we succeed to obtain reproducible avalanches that correspond to detection of photons. The avalanche current triggered by a single photon is quite small, a few times smaller than that of the SLiK diode, so we use a modification of the AQC described in Ref. [33] optimized for this type of SPAD. A photodiode is operated at -10 °C and an excess voltage of 18 V. Characteristics of this detector type are: dead time $\tau_{dead} = 25.5$ ns, output pulse width $\tau_{pd}$ = 8-50 ns (adjustable), detection efficiency of 38% at 650 nm, jitter of 730 ps FWHM, and dark counts of 84.6 kHz.

Since, as it will become clear later, afterpulsing has a crucial impact on performance of the QRNG, we also measure afterpulsing probability and afterpulsing lifetime for the detectors, using single lifetime afterpulsing model [19] and method described in Ref. [34]. Measured distributions of time intervals between successive detections, for the two detector types, are shown in Fig. 5. From this, we obtain total afterpulsing probability $P$ = 0.047 and lifetime $\tau_a$ = 33 ns for SliK detectors, while $P = 0.016$, $\tau_a = 8.0$ ns for the SUR500 detector. Note that $P$ is just a parameter of exponential distribution and by itself does not give a realistic estimate of afterpulsing. Namely, dead time absorbs a fraction of afterpulses that depends on $\tau_a$. For SLiK detector visible afterpulses amount 2.3% of all pulses, while for SUR500 detector visible afterpulses constitute only 0.055% of all pulses.

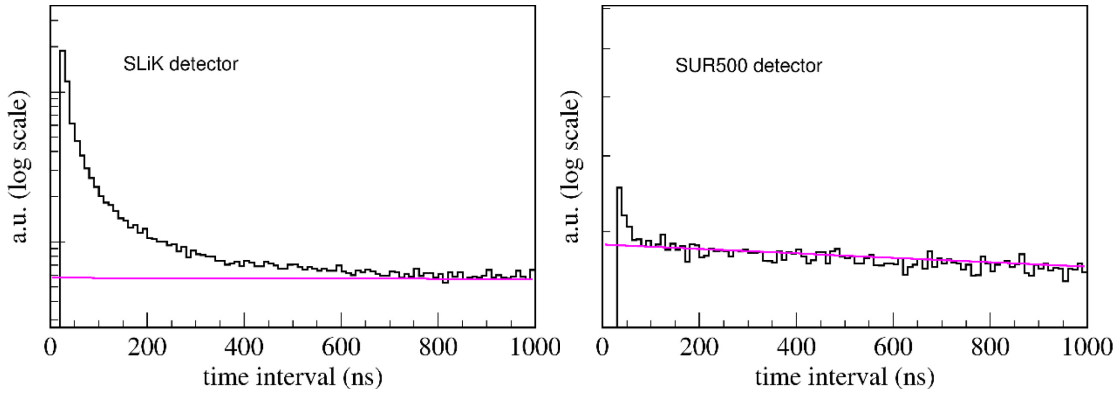


Fig. 5. Distributions of time intervals between successive detections, for the two detector types: SLiK (dashed line) and SUR500 (full line). Afterpulses are apparent as peaks above the flat background.

**Strobe signal**

Due to the laser jitter and intrinsic time resolution of single-photon detectors, photon detections jitter with respect to the trigger signal. Therefore, the Strobe signal should appear a bit later than the detector's output (as shown in Fig. 6) in order to read a well-defined bit value. For both types of detectors, taking the delay of 2 ns, the total latency budget between the Trigger signal and the Strobe has to be set to about 8.5 ns. The variable delay line, depicted in Fig. 2, is realized as a coaxial cable of suitable length.

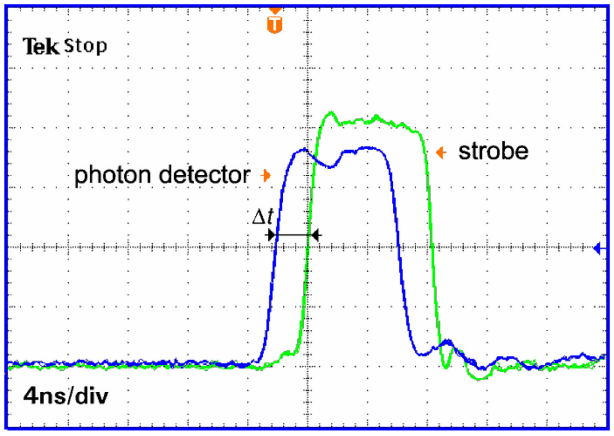


Fig. 6. Strobe – detector timing detail (4 ns/div time scale). Width of the photon counter pulse is $\tau_{pd}$ = 8 ns, and duration of the strobe pulse is $\tau_{trig}$ = 8 ns. By means of the variable delay line (shown in Fig. 2) the relative delay $\Delta t$ between the two signals is set to 2 ns, which is enough to overcame mutual jitter of the pulsed laser and the photon detector, thus enabling readout of a well-defined random state of the detector's output.

**Random number generation modeling and practical realization**

Even though, in theory, as explained above, there should be no correlation among generated bits, due to inevitable memory effects in realistic devices some autocorrelation appears also in experimental realization of the QRNG. Successive pulses of a pulsed laser diode are phase randomized exhibiting a

Poisson statistics of number of emitted photons per pulse ($n$) [23-24]. The detection of such a state is ether supposed to be ballistic ($n$ independent detection trials) or superlinear [14]. Crucial insight into the present QRNG is that any details of photon emission or detection are irrelevant as long as all physical processes pertaining to one emission and subsequent detection event are completed (i.e. die off) before the next trigger (a random bit request moment). This would ensure no correlations among generated bits. However, while the turn-on and turn-off processes in a laser diode have typical lifetimes on the order of <100 ps [15], a photon detection imperfections (dark counts, dead time, afterpulsing) involve effects on a time scale of tens to hundreds of nanoseconds that ultimately limit the achievable trigger rate and randomness. Dark counts are randomly distributed in time and therefore do not carry *per se* any correlating information and are furthermore greatly suppressed by tight coincidence between strobe and detector pulses. However, dead time and afterpulsing may cause correlations among bits. Since afterpulsing probability of the used APD dies-off nearly exponentially in time [16], in the limit of long enough trigger period, only neighboring bits may be non-negligibly correlated. Under that condition, correlations among bits is characterized by the serial autocorrelation between neighboring bits, namely coefficient $a_1$, defined as [17]:

$$a_k = \frac{\sum_{i=1}^{N-k}(x_i - \bar{x})(x_{i+k} - \bar{x})}{\sum_{i=1}^{N-k}(x_i - \bar{x})^2} \tag{1}$$

where $x_i$ are generated bits and lag $k = 1$. Throughout the paper we use statistics of $N = 10^9$ bits for each measurement point, leading to statistical error of $1/\sqrt{N-k} \approx 3.2 \cdot 10^{-5}$. Random bits have been generated upon a periodic trigger with frequency spanning from 1 to 25 MHz. Statistical bias, defined as $b = p_1 - 0.5$, was manually adjusted to zero within ±0.0005 before each measurement point. The generated bits were transferred to a PC computer via a USB2 controller. Correlation coefficient $a_1$ has been evaluated using ENT software [18]. Results are shown as hollow dots in Fig. 7.

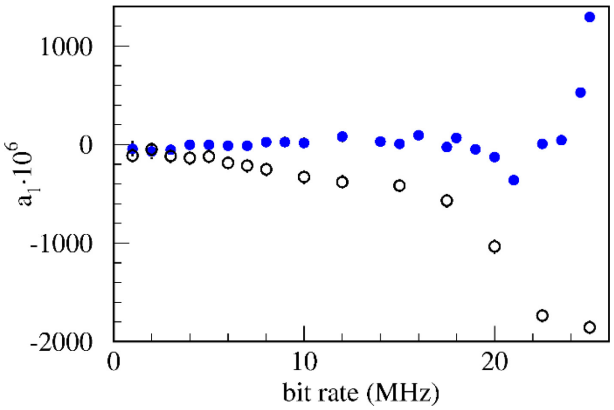


Fig. 7. A series of autocorrelation coefficients $a_1$ as a function the triggered bit rate, measured for two distinct detector pulse widths ($\tau_{pd}$): 8 ns (hollow dots) and 21 ns (filled dots), for a detectors based upon SLiK SPAD. Statistics per coefficient is $10^9$ bits. One sigma error bars are barely visible being roughly equal to the dot size.

We see that $a_1$ is generally small, negative and that its magnitude rises with the rate. To explain this behavior, we start by considering a successful detection of a photon (bit value "1") as shown in Fig. 8,

where $\Delta t$ is the delay between expected photon detection and the strobe pulse (explained in Fig. 5), $\tau_{pd}$ and $\tau_{dead}$ are the detector pulse width and dead time respectively, while $T$ is the bit generation period.

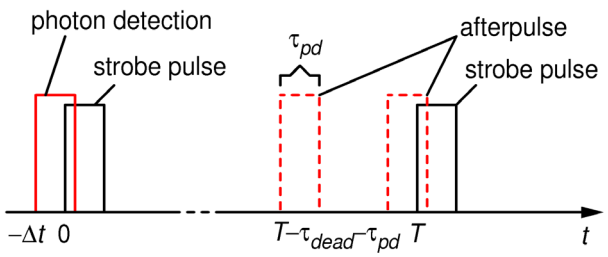


Fig. 8. Time sequence of detection and afterpulse events that cause negative autocorrelation between subsequent random bits.

The next bit value is requested/generated a period *T* later. Afterpulsing in conjunction with dead time causes two competing effects. First, at time *T* there will be an enhanced probability $P_+$ to generate "1" due to an afterpulse appearing in coincidence with the strobe. Second, with probability $P_-$, an afterpulse appearing less that the dead time $\tau_{dead}$ and *before* the strobe will cause the detector to miss the next photon whose probability would otherwise be ½. The total correlation is then given as:

$$a_1 = \frac{1}{2}[P_+ - P_-] = \frac{1}{2}\left[\int_{T+\Delta t-\tau_{pd}}^{T+\Delta t} P_a(t)dt \; - \frac{1}{2}\int_{T+\Delta t-\tau_{pd}-\tau_{dead}}^{T+\Delta t-\tau_{pd}} P_a(t)dt\right] \tag{2}$$

where $P_a(t)$ is probability density function for appearance of an afterpulse at time *t* after a detection event. The overall factor ½ stems from the fact that two photons are generated on average per photon detection. We note that higher lag coefficients ($k > 1$) are obtained by shifting the boundaries of both integrals in Eq. 2 by $T$, that is: $a_k = a_1 \exp\left(\frac{(k-1)T}{\tau_a}\right)$, that is bit generation is a Markov process and correlation among bits can indeed be characterized well by only the serial correlation coefficient with lag 1, namely $a_1$. In case of the SLiK detector, where $\tau_{pd}$ = 8 ns and $\tau_{dead}$ = 22 ns, the net autocorrelation $a_1$ is negative because the integration interval of the second term (of length $\tau_{dead}$) is longer than that of the first term (length $\tau_{pd}$) and because $P_a(t)$ is larger in the second integral. However, since the two integrals are the contiguous parts of an integral over a fixed interval (of length $\tau_{pd} + \tau_{dead}$) it could be possible to choose $\tau_{pd}$ such that the correlation vanishes. If a simple exponential model of afterpulsing is assumed, i.e. $P_a(t) = \frac{P}{\tau_a} e^{-t/\tau_a}$ [19] where $P$ is the total afterpulsing probability, by requiring $a_1 = 0$ one gets:

$$e^{\frac{\tau_{pd}}{\tau_a}}\left[3 - e^{\frac{\tau_{dead}}{\tau_a}}\right] = 2 \tag{3}$$

from which $\tau_{pd}$ can be expressed as:

$$\tau_{pd} = \tau_a \ln\left[\frac{2}{3 - e^{\frac{\tau_{dead}}{\tau_a}}}\right]. \tag{4}$$

We note that if $P = 0$ in Eq. (2) then $a_1 = 0$ regardless of all other parameters. Interestingly, there is yet another possibility that leads to the same effect: for a hypothetical detector with an overwhelming afterpulsing lifetime (i.e. $\tau_a \to \infty$) Eq. (3) would be satisfied even if $P > 0$ and any value of $\tau_{pd}$ would be optimal. This is because afterpulses would then be virtually randomly distributed over time, like dark counts, not correlated to any particular detection and thus not able to cause correlations. However, in our realistic, SLiK SPAD based detector diode we have $\tau_a = 33$ ns, $P = 0.047$ and $\tau_{dead} = 22$ ns. Inserting $\tau_a$ and $\tau_{dead}$ in Eq. (4) yields $\tau_{pd} \approx 21$ ns. Apparently, the value of $\tau_{pd}$ so obtained, is optimal for cancelation of $a_1$ is independent of *T*.

To verify that experimentally, we vary the width of the detector's output pulse at the AQC and a evaluate autocorrelation as a function of $\tau_{pd}$ for several bit rates (10 MHz, 15 MHz, 17.5 MHz and 20 MHz). Experimental results shown in Fig. 9 indicate that an overall minimum of the autocorrelation is indeed obtained for $\tau_{pd} \approx 21$ ns and that is rather insensitive on the bit rate.

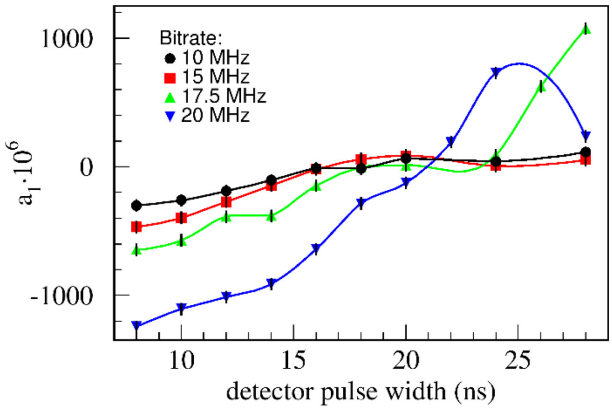


Fig. 9. Serial autocorrelation coefficient $a_1$ as a function of detector's pulse width ($\tau_{pd}$), measured for a set of bit rates. An overall minimum is obtained for $\tau_{pd} \approx 21$ ns.

We further note that following a detection of a photon at $-\Delta t$, the detector goes into the dead time and therefore afterpulses would contribute to the second integral in Eq. (1) only if its starting range ($T + \Delta t - \tau_{pd} - \tau_{dead}$) is greater than $\tau_{dead}$, that is:

$$T > 2\tau_{dead} + \tau_{pd} - \Delta t \tag{5}$$

which corresponds to bit rate of about $1/T < 16$ MHz. For higher trigger rates the second integral in Eq. (1) would become smaller and the autocorrelation would rise sharply, as indeed observed for bitrates of 17.5 MHz and 20 MHz.

After setting $\tau_{pd}$ to the optimal value of 21 ns, correlation coefficient $a_1$ has been evaluated again as a function of bit generation rates in the range 1-25 MHz. Results displayed in Fig. 7 (filled dots) show a significant improvement with respect to the result obtained with the original pulse width of 8 ns (hollow dots). The absolute value of $a_1$ is less than $1.25 \cdot 10^{-4}$ for bit rates all the way up to 20 MHz. At higher rates correlation quickly diverges because our simple model fails due to the effects explained above and possibly other smaller imperfections not taken into account.

In practice Eq. (4) cannot be exactly satisfied for physical devices. It is therefore interesting to investigate the sensitivity of autocorrelation to variation of parameters such as detector pulse width ($\tau_{pd}$), dead time ($\tau_{dead}$) and bit generation period ($T$). By substituting the exponential afterpulsing model in Eq. (2) and taking partial derivative of $a_1$ with respect to $\tau_{pd}$ we get:

$$\frac{\partial a_1}{\partial \tau_{pd}} = \frac{P}{4\tau_{pd}}\left[3 - e^{\tau_{dead}/\tau_a}\right]e^{-(T+\Delta t-\tau_{pd})/\tau_a}. \quad (6)$$

Evaluated at $\tau_{pd} = 21$ ns, for $T = 100$ ns, $\tau_{dead} = 22$ ns, $\tau_a = 33$ ns, $\Delta t = 2$ ns and $P = 0.047$, Eq. (6) predicts sensitivity of $a_1$ with respect to $\tau_{pd}$ of $32 \cdot 10^{-6}\ \mathrm{ns}^{-1}$ which is indeed in a good agreement with the slope of the 10 MHz curve in Fig. 9. Similar analysis for dead time yields a sensitivity of $-59 \cdot 10^{-6}\ \mathrm{ns}^{-1}$, whereas for generation period the variation sensitivity is $0.2 \cdot 10^{-6}\ \mathrm{ns}^{-1}$only. Since the three parameters ($\tau_{pd}$, $\tau_{dead}$, $T$) can be engineered with high precision and stability on the order of 1 ns, randomness quality of the present generator is predominantly affected by stability of bias which is about $500 \cdot 10^{-6}$. We find that serial correlation coefficients $a_k$ with lag $1< k \leq 64$ are consistent with zero within statistical error for $T = 100$ ns and $N = 10^9$. This is to be expected since with every lag the afterpulsing probability (and consequently the serial correlation) drops roughly by a factor of $\exp(T/\tau_a) \approx 21$, and thus the second and all further serial coefficients are much smaller than our statistical error.

In order to further improve on both the statistical bias and the autocorrelation, one could use the Von Neumann extractor [23]. However, while on average it takes a block of 4 bits to generate one corrected bit, the time to gather enough bits to generate one corrected bit is not bounded and can span anywhere from 4T to infinity. In our case that would result in lowering of the bit production efficiency to only 25% and enlargement of the delay between the request and availability of the random bit. Therefore we chose an alternative, well known approach, which enabled us to keep the 100% efficiency and bounded latency: we built two independent generators of the type shown in Fig. 2, distributed the same trigger signal to their inputs and logically XORed their outputs. The XOR gate added another 1.3 ns of propagation delay, therefore the delay between the trigger and strobe was enlarged by the same amount, i.e. to 9.8 ns. According to [20] XORing two independent random strings each with bias $b$ and autocorrelation $a_1$ results in a new string with an improved bias $b'$ and autocorrelation $a_1'$ :

$$b' = -2b^2 \quad (7)$$

$$a_1' = a_1^2 + 8a_1b^2 \quad (8)$$

At 10 Mbit/s (i.e. *T* = 100 ns) for a single QRNG we measured: $b \leq 5 \cdot 10^{-4}$; $a_1 \leq 5 \cdot 10^{-5}$. Higher lag correlations were consistent with zero, within statistical errors, as expected in our model. By applying Eqs. (7-8) we estimate the upper bounds for the residual bias and autocorrelation of the XORed QRNGs to be: $|b'| \leq 5 \cdot 10^{-7}$ and $|a_1'| \leq 3 \cdot 10^{-9}$, respectively.

In our theoretical model of thus QRNG, illustrated in Fig. 8, there are no deviations from randomness other than bias and serial autocorrelation and we saw that coefficients with lag $k > 2$ contribute to non-randomness negligibly, both theoretically and as confirmed by measurements. To detect statistically the

above imperfections as a 3 sigma effect, one would need to generate at least $10^{13}$ bits for bias, and $10^{18}$ for correlation, showing that bias is the leading imperfection. However, afterpulsing is generally more complex [19] and there could be other imperfections in the setup that were not accounted for in our model, all of which could limit the achievable randomness.

**Results with SLiK-based detectors**

As explained above, in order to arrive to a long sequence of random bits that is statistically indistinguishable from a perfectly random one, we may resort to XORing of two independent QRNGs. To that end, outputs of two independent and identical QRNGs are built and their outputs XORed. Since the XOR gate adds 1.3 ns propagation delay, the overall delay budget (latency) of the XORed SLiK-based QRBGs arrives at $\tau_L = (9.8 \pm 0.2)$ ns. This is the delay that has to be set between the Trigger and Strobe. As deduced in Ref [35], XORing two independent Markov processes each with bias $b$ and correlation $a_1$ results in an improved bias: $b' = -2b^2$ and correlation: $a_1' = a_1^2 + 8a_1b^2$. At 10 Mbit/s (or $T$ = 100 ns) for a single QRNG we measured: $b \leq 5 \cdot 10^{-4}$; $|a_1| \leq 1 \cdot 10^{-4}$, from which we conclude that the tandem performs: $|b'| \leq 5 \cdot 10^{-7}$ and $|a_1'| \leq 1.1 \cdot 10^{-8}$. At that level, at least $\sim 10^{13}$ bits are required to statistically detect deviation from randomness which is orders of magnitude more than would be required by any conceivable Bell test. A sequence of $10^9$ bits (1000 samples of 1 Mbits) generated by the tandem generator at 10 Mbit/s passes NIST's randomness test suite STS-2.1.2 with high scores. Typical results are shown in Table 1.

| Statistical test | p-value | Proportion/Threshold | Result |
|---|---|---|---|
| Frequency | 0.784927 | 994/980 | Pass |
| Block frequency | 0.096578 | 992/980 | Pass |
| Cumulative sums | 0.767582 | 997/980 | Pass |
| Runs | 0.775337 | 995/980 | Pass |
| LongestRun | 0.103138 | 991/980 | Pass |
| Rank | 0.657933 | 994/980 | Pass |
| FFT | 0.251837 | 993/980 | Pass |
| NonOverlappingTemplate | 0.574903 | 994/980 | Pass |
| OverlappingTemplate | 0.867692 | 987/980 | Pass |
| Universal | 0.697257 | 994/980 | Pass |
| ApproximateEntropy | 0.348869 | 993/980 | Pass |
| RandomExcursions | 0.588541 | 626/615 | Pass |
| RandomExcursionsVariant | 0.235040 | 625/615 | Pass |
| Serial | 0.637119 | 990/980 | Pass |
| LinearComplexity | 0.880145 | 986/980 | Pass |

Table 1. Typical results of NIST statistical test suite STS-2.1 for 1000 samples of 1 Mbits generated by XORing outputs of two independent QRNGs based on SLiK-based single-photon detector. For each statistical test an overall p-value as well as proportion of samples that passed the test versus theoretical threshold are given.

The test results confirm that indeed this tandem QRBG performs un-distinguishably from perfect randomness as long as the shortest time lapse between two bit requests is >= 100 ns and for string length of $10^9$ bits.

Finally, as an alternative approach to improve randomness, non-overlapping pairs of bits from a single QRNG operated at 10 Mbit/s have been XORed. In that case, the resulting bias and correlation are given by [20]:

$$b' \approx -2b^2 - a_1/2 \tag{9}$$

$$a_1' \approx 4a_1 b^2 \tag{10}$$

which gives $b' \approx -2.6 \cdot 10^{-6}$ and $a_1' \approx 5 \cdot 10^{-11}$. Again, 1000 samples of 1 Mbits have passed NIST test suite. The drawback of this approach is halving of the effective bit rate (to 5 Mbits/s) and doubling the latency, while the good side is requirement for only one photon detector.

**Results with SUR500-based detector**

We now realize the QRBG shown in Fig. 2 with the second type of detector, namely the one based on SUR500 SPAD. In the discussion above we realized that afterpulsing is the dominant effect that generates correlations among bits. The main advantage of the detector based on SUR500 is its low and short-lived afterpulsing. Because of that, for a long enough bit generating period $T$, probability to encounter an afterpulse at the next strobe signal becomes negligible. This intuitive argument, in fact, points out to even a third possible solution of Eq. (2) which yields $a_1 = 0$. Namely, if $P_\alpha(t)$ tends to 0 when $t$ satisfies:

$$t > T + \Delta t - \tau_{pd} - \tau_{dead} \tag{11}$$

then both integrals in Eq. (2) also tend to zero, and the pulse width $\tau_{pd}$ does not matter anymore. In practice, the shorter $\tau_{pd}$ the better, since then condition in Eq. (11) is satisfied to a greater extent. For practical reasons of clean readout we chose $\tau_{pd} = 10$ ns. With this setting, random bits have been generated upon a periodic trigger with frequency spanning from 1 to 25 MHz in the same manner as for the SLiK detector. Obtained autocorrelation coefficient $a_1$ shown in Fig. 10 features lower absolute value further towards high bit rate end, when compared to the performance of the SLiK-based QRNG shown in Fig. 7. On top of that, now we do not need to adjust $\tau_{pd}$ because it does not affect the autocorrelation unless it is so large that Eq. (5) is violated, which we confirmed by measurements. According to Eq. (5), for $\tau_{pd} = 10$ ns, we expect that the highest bit generation rate is about 17 MHz. Indeed, we see that after that point correlation rises towards positive values, as expected, while above 22 MHz dead time proximity ($1/T \approx \tau_{dead}$) starts to cause large anti-correlation and our generator becomes useless.

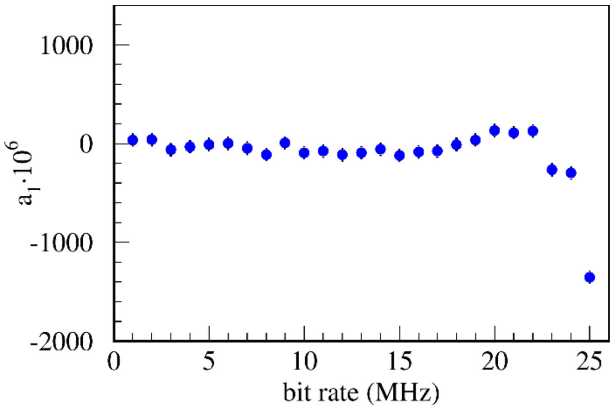


Fig. 10. A series of autocorrelation coefficients $a_1$ as a function the triggered bit rate, measured for the pulse width $\tau_{pd}$ = 10 ns, for a detector based upon SUR500 SPAD. Statistics per coefficient is $10^9$ bits. One sigma error bars are barely visible being roughly equal to the dot size.

In order to improve on both the bias and the correlation, we generate two sets of $10^9$ random bits at a rate of 17 MHz and XOR them bit-by-bit in order to obtain a single string of $10^9$ bit. For a typical string we measure: $|b|$ and $|a_k|$ for lags $1 \leq k \leq 64$ to be consistent with zero within statistical errors of $1.6 \cdot 10^{-5}$ and $3.2 \cdot 10^{-5}$ respectively. Table 2 summarizes test results obtained by the NIST test suite of a typical string of $10^9$ bits obtained in this manner.

| **Statistical test** | **p-value** | **Proportion/Threshold** | **Result** |
|---|---|---|---|
| Frequency | 0.745908 | 991/980 | Pass |
| Block frequency | 0.897763 | 994/980 | Pass |
| Cumulative sums | 0.619590 | 990/980 | Pass |
| Runs | 0.996996 | 989/980 | Pass |
| LongestRun | 0.603841 | 985/980 | Pass |
| Rank | 0.735908 | 992/980 | Pass |
| FFT | 0.556460 | 983/980 | Pass |
| NonOverlappingTemplate | 0.474837 | 990/980 | Pass |
| OverlappingTemplate | 0.643366 | 986/980 | Pass |
| Universal | 0.834308 | 990/980 | Pass |
| ApproximateEntropy | 0.932333 | 987/980 | Pass |
| RandomExcursions | 0.573467 | 622/614 | Pass |
| RandomExcursionsVariant | 0.499546 | 620/614 | Pass |
| Serial | 0.765922 | 996/980 | Pass |
| LinearComplexity | 0.572847 | 995/980 | Pass |

Table 2. Typical results of NIST statistical test suite STS-2.1 for 1000 samples of 1 Mbits generated by XORing outputs of two independent QRNGs based on SUR500-based single-photon detector.

For each statistical test an overall p-value as well as proportion of samples that passed the test versus theoretical threshold are given.

To test possibility to generate high-quality bits with the SUR500-based photon detector, we generate a string of $2 \cdot 10^9$ bits at 17 MHz and XOR neighboring bits to arrive to a new string of $10^9$ bits. This new string also has bias and serial correlations within statistical errors and passes NIST statistical test, confirming that a statistically good random bits can be generated at a pace of 8.5 MHz or every 118 ns (or more).

We see that the QRNG realized with SUR500-based detector performs significantly better and faster than the one utilizing SLiK diode, allowing bit generation of up to 17 MHz for the (simulated) tandem configuration or 8.5 MHz for a configuration with successive bit XORing. We conclude that this improvement in performance is solely due to lower afterpulsing of SUR500 SPAD, even though it is inferior as single-photon sensor, having only half the quantum efficiency and over two orders of magnitude higher dark counts rate than SLiK SPAD.

**Discussion**

A conceptually simple, on-demand optical quantum random number generator is presented that simultaneously features: (1) ultra-fast response upon a bit request (9.8 ns), (2) 100% bit generation efficiency upon the trigger and (3) in-future-of-request random action. While its characteristics are of particular relevance to some applications (such as Bell tests or random logic [25]), it can be used for a much wider range of applications. It can deliver random bits at a maximum rate of currently 10 MHz featuring very low randomness errors without post-processing. Sources of randomness errors and their sensitivity to variations in hardware components have been studied, modeled and shown to be small. In comparison, other post-processing free-running QRNGs have achieved 100% efficiency and nanosecond scale response by quick sampling of a randomly toggling flip-flop [6], [26], but with all relevant physical processes happening hundreds of nanoseconds in the past of the request due to long delays in optical and electrical paths or long range correlations among bits. A post-processing-free QRNG based on self-differencing technique [27] operated at a clock 1.03 GHz delivers bits randomly at an average rate of 4.01 Mbit/s thus having efficiency of only about $4‰$. In a setup having a similar topology to ours [28] a gain-switched laser diode feeds an asymmetric Mach-Zender interferometer whose output intensity is measured by a photodiode and digitized by 8-bit ADC, whereas in [29] an in-future-of-request continuous-variable QRNG is based on phase diffusion in a laser diode. Both QRNGs feature unavoidable requirement for ADC conversion followed by complex post-processing which results in long response times. Furthermore, none of the above discussed constructs has been tested random for strings longer than $\sim 10^9$ bits, which can be too short for applications like Monte Carlo calculations and simulations. For the XORed QRNG, assuming the validity of our model, we estimated that randomness imperfections can not be statistically detected for a sequence of generated bits shorter than $\sim 10^{13}$ bits. A notable success in randomness estimation is achieved in [30] by calculating propagation of min-entropy through privacy amplification claiming randomness for strings of up to $\sim 10^{96}$ bits, but at the expense of time-consuming post-processing and long history of physical events prior to the bit request. Finally, achieved delay between a request and availability of random bit in our QRNG is arguably the shortest possible with a

given state of technology since only a logically minimal sequence of processes is required to generate one bit, namely a light pulse emission followed by a photon detection.

**Further research - next steps**

We have demonstrated that using our concept, one can generate a random bit every 59 ns (or more) with a latency below 10 ns. The cost of this generator is two pulsed lasers and two photon detectors per bit, which seems high when considered in terms of usual bulk components that we used in this study. On top of that, low bias of this QRNG cannot be guaranteed "off the shelf", rather it must be obtained by a careful adjustments of each laser or, alternatively, by some kind of electrical auto adjustment via a negative feedback loop, that was not demonstrated here.

While expensive when made with bulk components, thanks to its simplicity, this QRNG seems ideally suited for realization on a chip. Namely, recent advances in silicon CMOS-process-based camera chips, allow for thousands of independent single-photon counting pixels on a single silicon chip along with all required logic circuits and analog amplifiers needed for automatic bias adjustment, thus reducing the price for detectors. The need for a large number of lasers can be eliminated by use of a single pulsed laser that uniformly illuminates all pixels at the same time.

The presented bit generating method, in principle, allows for miniaturization of the QRNG to a chip level with the existing technology. This would open possibility for wider range of applications.

**References**


1. Figotin A. *et al.*, inventors; The Regents of the University of California, asignee; *A random number generator based on spontaneous alpha-decay*. PCT patent application WO0038037A1.
2. Stipčević M., Medved Rogina B., Quantum random number generator based on photonic emission in semiconductors, *Rev. Sci. Instrum.* **78**, 045104:1-7 (2007).
3. Rarity J. G., Owens P. C. M., Tapster P. R., Quantum random-number generator and key sharing, *J. Mod. Opt.* **41**, 2435-2444 (1994).
4. Stefanov A., Gisin N., Guinnard O., Guinnard L., Zbinden H., Optical quantum random number generator, *J. Mod. Opt.* **47**, 595-598 (2000).
5. Fürst H. *et al.*, High speed optical quantum random number generation, *Opt. Express* **18**, 13029-37 (2010).
6. Stipčević M., Fast nondeterministic random bit generator based on weakly correlated physical events, *Rev. Sci. Instrum.* **75**, 4442-4449 (2004).
7. Scarani V. *et al.*, The security of practical quantum key distribution, *Rev. Mod. Phys.* **81**, 1301–1350 (2009).
8. Merali Z., Quantum mechanics braces for the ultimate test, *Science* **331**, 1380–1382 (2011).
9. Weihs G., Jennewein T., Simon C., Weinfurter H., Zeilinger A., Violation of Bell's Inequality under Strict Einstein Locality Conditions, *Phys. Rev. Lett.* **81**, 5039–5043 (1998).
10. Scheidl, T. *et al.*, "Violation of local realism with freedom of choice," *Proc. National Academy of Sciences* **107**, 19708–19713 (2010).
11. Giustina M. et al., Bell violation using entangled photons without the fair-sampling assumption, *Nature* **497**, 227–239 (2013).
12. Christensen B. G. et al., Detection-Loophole-Free Test of Quantum Nonlocality, and Applications, *Phys. Rev. Lett.* **111**, 130406 (2013).
13. Stipčević M., Active quenching circuit for single-photon detection with Geiger mode avalanche photodiodes, *Appl. Opt.* **48**, 1705-1714 (2009).
14. Lydersen L. *et al.*, Superlinear threshold detectors in quantum cryptography, *Phys. Rev. A* **84**, 032320 (2011).

15. Pesquera L., Revuelta J., Valle A., and Rodriguez M. A., *Theoretical calculation of turn-on delay time statistics of lasers under PRWM*, Proc. SPIE 2994: Physics and Simulation of Optoelectronic Devices V [Osinski M., Chow W. W., (eds.)] (SPIE, San Jose, 1997).
16. Stipčević M., Gauthier D. J., *Precise Monte Carlo Simulation of Single-Photon Detectors*, Proc. SPIE Vol. 8727: Advanced Photon Counting Techniques VII [Itzler M. A., Campbell J. C. (eds.)] (SPIE, Baltimore, 2013).
17. Knuth D., *The art of computer programming Volume 2: Seminumerical Algorithms*, Third Edition [70-71] (Addison-Wesley, Reading, 1997).
18. Walker, J., *ENT - A Pseudorandom Number Sequence Test Program*, (2003) URL: http://www.fourmilab.ch/random/, Date of access: 05/02/2014.
19. Giudice A. C., Ghioni M., and Cova S., *A process and deep level evaluation tool: afterpulsing in avalanche junctions*, Proc. European Solid-State Device Research 2003 (ESSDERC 03). 16–18 Sept. 2003 p. 347–350.
20. Davies R., *Exclusive OR (XOR) and hardware random number generators*, February 28, 2002, URL: http://www.robertnz.net/pdf/xor2.pdf, Date of access: 05/02/2014.
21. Rukhin A. *et al.*, NIST Special Publication 800-22rev1a (April 2010), URL: http://csrc.nist.gov/rng , Date of access: 01/02/2012.
22. Von Neumann J., *Various techniques for use in connection with random digits*, [Von Neumann Collected Works, Vol 5] [768-770] (Macmillan, New York, 1963).
23. Henry C., "Theory of the line width of semiconductor lasers," *IEEE J. Quantum Electron.* **18**, 259–264 (1982).
24. Henry C., "Phase noise in semiconductor lasers," *J. Lightwave Technol.* **4**,298–311(1986).
25. Stipčević M., "Quantum random flip-flop based on random photon emitter and its applications", arXiv:1308.5719 [quant-ph]
26. Jennewein T., Achleitner U., Weihs G., Weinfurter H., Zeilinger A., "A Fast and Compact Quantum Random Number Generator", *Rev. Sci. Instrum.* **71**, 1675-1680 (2000).
27. Dynes J. F., Yuan Z. L., Sharpe A. W., and Shields A. J., "A high speed, postprocessing free, quantum random number generator", *Appl. Phys. Lett.* **93**, 0311109 (2008).
28. Yuan, Z. L. et al., "Robust random number generation using steady-state emission of gain-switched laser diodes", *Appl. Phys. Lett.* **104**, 261112 (2014).
29. Abellán C. et al., "Ultra-fast quantum randomness generation by accelerated phase diffusion in a pulsed laser diode", *Opt. Express* **22**, 1645-1654 (2014).
30. Sanguinetti B., Martin A., Zbinden H., and Gisin N., "Quantum Random Number Generation on a Mobile Phone", *Phys. Rev. X* **4**, 031056 (2014).
31. D. Frauchiger, R. Renner, and M. Troyer, "True randomness from realistic quantum devices," SPIE Security+Defense Conference Proceedings, Volume 8899 (2013), arXiv:1311.4547 [quant-ph].
32. M. Stipčević, R. Ursin, "An On-Demand Optical Quantum Random Number Generator with In-Future Action and Ultra-Fast Response", Scientific Reports **5**, 10214:1-8 (2015).
33. M. Stipčević, B. G. Christensen, P. G. Kwiat, D. J. Gauthier, "An advanced active quenching circuit for ultra-fast quantum cryptography", Opt. Express **25**, 21861-21876 (2017).
34. G. Humer, M. Peev, C. Schaeff, S., M. Stipčević, R. Ursin, "A simple and robust method for estimating afterpulsing in single photon detectors", J. Lightwave Technol. **33**, 3098-3107 (2015).
35. M. Stipčević, J. Bowers, "Spatio-temporal optical random number generator", Opt. Express **23**, 11619-11631 (2015).