Random Sampling-and-Averaging Techniques for Single-Photon Arrival-Time Detections in Quantum Applications: Theoretical Analysis and Realization Methodology

Tony Wu, *Graduate Student Member, IEEE*, Ruoman Yang, *Graduate Student Member, IEEE*, and Tzu-Chien Hsueh , *Senior Member, IEEE*

Abstract— A random sampling-and-averaging (RSA) technique based on stochastic Monte Carlo methods is described in this paper for enhancing the accuracy of single-photon arrival-time measurements down to sub-picosecond ranges in emerging quantum applications. The theoretical variances of both synchronous and asynchronous RSA techniques are presented in the mathematical formats and experimentally verified by the Monte Carlo simulations. Meanwhile, the methodology of converting the mathematical models into an almost all-digital low-power integrated-circuit is elaborated by a circuit-level example with the instruction of setting circuit parameters. Along with the superior measurement resolution, scalable dynamic ranges, high linearity, high noise immunity, and low power/area consumption, the primary limitation of the RSA techniques has also been addressed for the forthcoming conversion-rate enhancement techniques.

Index Terms— Correlated random variable, independent and identically distributed, joint probability density function, Monte Carlo method, quantum probability amplitude, stochastic random sampling, time-correlated single-photon counting, time-domain modulo operation, time-to-digital converter.

I. INTRODUCTION

BECAUSE of the high demands for time-correlated single-photon counting (TCSPC) [1]–[4] functionality in a variety of emerging quantum technology research and commercial market segments, including quantum 2D/3D imaging/ranging/sensing [5]–[7], quantum-bit-state probability amplitude measurements [8]–[11], quantum cryptography [12]–[14], positron emission tomography (PET) [15], [16], time-resolved spectroscopy [17], fluorescence-lifetime imaging (FLIM) [18], [19], diffusive optical tomography, molecular imaging, live-cell/tissue microscopy [2], free-space time-of-flight (TOF) measurements [20], and light detection-and-ranging (LiDAR) [5], [21], the performance specifications of the time-to-digital converter (TDC), which is one of the most essential integrated-circuit building blocks,

are increasingly targeting at pico-seconds fine resolutions, micro-seconds wide dynamic ranges, high linearity, and fast conversion-rates under aggressive silicon-area, voltage, and power constraints.

Generally, the state-of-the-art TDC designs can achieve high standard qualities for a certain aspect but come with trade-offs among the other performance metrics. For instance, the pulse-shrinking delay-line [22]–[24], tapped delay-line [25]–[27], and Vernier delay-line [28]–[30] TDC approaches offer low sensitivities to process, voltage, and temperature (PVT) variations but require a phase-locked loop (PLL) [27] or delay-locked loop (DLL) [6], [22], [23], [25], [29]–[34] to lock their voltage-controlled delay-cells to a low-jitter reference clock. The approach of phase-interpolation between delay-cells [6], [18], [24]–[26], [32]–[37] first-orderly enhances the TDC resolution but increases area cost and power consumption; the linearity of this type of TDC can be further improved by employing the sliding-scale technique [3], [5], [32], [33], with the downsides of doubling the number of phase-interpolators and converting the non-linearity into stochastic jitter. The time-amplifier based approach [38]–[40] improves the TDC resolution for a price of limited linear time-amplification regions and high PVT sensitivities. The simple pseudo-differential free-running ring-oscillator based approach with the period and intermediate phase-edge counters [16]–[18], [20] allows one TDC to serve an individual or multiple single-photon-avalanche-diode (SPAD) pixels for high fill-factor imaging, but the TDC resolution and linearity are dominated by the delay-cell bandwidth and PVT variation, respectively. The two-step TDC approach exploits time-to-amplitude (TAC) and analog-to-digital (ADC) conversions [1], [2], [41] to achieve a pico-second resolution with the assistance of well-developed ADC non-ideality calibration techniques [42], [43], which are suitable for high-end TCSPC instruments [2], [44] but not for highly integrated silicon-photonics quantum applications due to the mixed-signal circuit area and complexity; also, both dynamic range and accuracy of each time-interval measurement are mainly dominated by the ADC specification.

To resolve the problems of performance trade-offs and high-cost implementations, this paper proposes utilizing the random sampling-and-averaging (RSA) technique to meet the growing requirements of time-interval measurements in

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

2

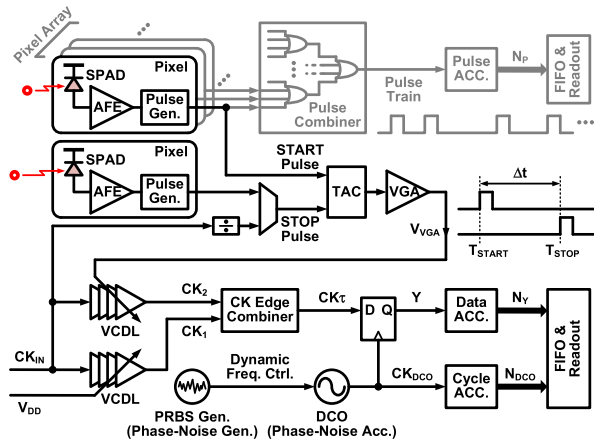IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS

Fig. 1. The block diagram of the TCSPC system including the features of both coarse single-photon counting (gray) and high-accuracy time-interval measurements using the RSA-based TDC technique (black).

emerging quantum applications. The RSA technique has been used in a variety of stationary signal detections mainly for ultra-high accuracy measurements and ultra-low energy sensing [45]–[49] associated with extremely strict power and area constraints. To comprehensively understand the feasibility of using the RSA technique in emerging quantum applications, this paper describes the fundamental concepts of both synchronous and asynchronous RSA techniques by deriving the mathematical expressions of their theoretical measurement variances, which are all experimentally verified by the simulations. Meanwhile, the methodology of converting the mathematical models into an almost all-digital and low-power integrated-circuit is introduced by a circuit-level example with a guideline of setting circuit parameters. Finally, the slow conversion-rate of the RSA technique has been properly addressed, which indicates additional enhancement techniques are required to fully elevate the RSA technique being applicable to not only high-accuracy quantum communication and cryptography applications [8]–[14] but also high frame-rate/fill-factor quantum imaging and ranging systems [5]–[7], [15]–[21].

The remainder of the paper is organized as follows. The circuit-and-system level overview of an TCSPC using the asynchronous RSA technique is introduced in Section II. The fundamental probability principles and theoretical variances of the RSA techniques are derived and presented in Section III. The behavioral-model simulations and RSA performance comparisons are discussed in Section IV. The conclusions and potential future work are summarized in Section V.

II. SYSTEM-LEVEL OVERVIEW

The conceptual block diagram of the proposed TCSPC system is shown in Fig. 1, which offers both coarse single-photon counting (gray) and high-accuracy time-interval (black) measurements. The whole system consists of single-photon detection pixels, time-to-amplitude conversion (TAC) circuits, and a time-to-digital conversion mechanism with the RSA technique. Each detection pixel includes a single-photon avalanche diode (SPAD) with the quenching/clamping circuits [3], [5]–[7] for the optical-to-electrical power domain

transition followed by an analog front-end (AFE) and pulse generator to convert the received single photons to electrical voltage pulses; the AFE circuit contains inverter-based transimpedance amplifiers [4] and high-bandwidth CMOS pulse generators. In the path of coarse single-photon counting, the pulse combiner, i.e., the OR tree in Fig. 1, merges multiple pulse outputs occurring at different timings into a pulse train contributed by the entire array of the single-photon detection pixels, and then the following pulse accumulator counts the number of pulses within a specific time to accomplish the feature of direct photon counting [20] and readout circuit reduction [16]. In the path of high-accuracy time-interval measurement, the timing of T_{START} is set by the START pulse from a specific single-photon detection pixel [2]; meanwhile, the multiplexer for the STOP pulse can select the timing of T_{STOP} from either the other specific single-photon detection pixel [2] or the system input clock, CK_{IN}, [3]–[7]. In any configuration, the time-interval, Δt, between T_{START} and T_{STOP} is the primary quantity under the measurement as shown in Fig. 1. The TAC circuit, containing a current source and capacitor banks with low-resolution static controls for coarse dynamic-range/conversion-gain tunability, converts the time-interval information into a constant DC voltage buffered by the variable-gain amplifier (VGA) offering noise-rejection and driving capabilities with additional tunability if required.

The TDC mechanism, which is the main focus of this paper, is illustrated at the lower-half of Fig. 1. The two identical voltage-controlled delay lines (VCDL) are both driven by the input clock, CK_{IN}, so the clock periods of CK_1 and CK_2 are identical, but the time-domain delays of CK_1 and CK_2, which carry the Δt information, are functions of the DC voltages, V_{DD} and V_{VGA}, respectively. After the rising clock-edge combiner, the Δt information can be periodically represented by the positive duty-cycle of $CK\tau$. Meanwhile, a free-running ring-based digital-controlled oscillator (DCO) generates the asynchronous clock, CK_{DCO}, to sample the waveform of $CK\tau$ through a single (i.e., 1-bit) D flip-flop. The data and cycle accumulators count the numbers of ones at the outputs of the D flip-flop (i.e., Y) and DCO (i.e., CK_{DCO}), respectively. Then, the averaging result is basically the ratio of the counter outputs. At this point, the primary process of the random sampling-and-averaging TDC has been completed. At first glance, the process is simple, which indeed indicates the superior power/area efficiency, high linearity and noise immunity of the RSA technique. However, the theory and implementation knowhow in behind are quite complicated and unintuitive.

III. RANDOM SAMPLING-AND-AVERAGING TECHNIQUES

The RSA measurement technique is originated from the principle of the Monte Carlo methods which are broadly used in the fields of applied mathematics and financial engineering [50]. Briefly speaking, the Monte Carlo methods are based on the analogy between probability and volume. In a stochastic or random process, the mathematics of a measure, e.g., probability density functions (PDF), formalizes the intuitive concept of a probability defining the volume of the possible outcomes. On the other hand, the Monte Carlo methods use this attribute in reverse by obtaining the volume from experiments and

then interpreting the volume as a probability. The relationship between the theoretical probability and experimental Monte Carlo method are summarized in (1), (2) and (3) by examining the expectation and mean values of a random variable, Y.

$$E[Y] = \int_{-\infty}^{\infty} y \cdot f(y) \cdot dy \tag{1}$$

$$\bar{Y} = \frac{1}{N} \cdot \sum_{n=1}^{N} Y_n \tag{2}$$

$$E[Y] = E[\bar{Y}] = \lim_{N \to \infty} \frac{1}{N} \cdot \sum_{n=1}^{N} Y_n \tag{3}$$

where $E[Y]$ is the expectation of the random variable obtained from the PDF, $f(y)$; Y_n is the n-th experimental sample of Y; N is the total number of the samples; $\bar{Y}$ is the mean of Y_n and can be treated as the Monte Carlo estimate of the "ideal" expectation, $E[Y]$. Also, when N increases toward infinity, $\bar{Y}$ converges to $E[Y]$, which is also the expectation of $\bar{Y}$, $E[\bar{Y}]$, as shown in (3). If Y_1, Y_2, ..., Y_N is a sequence of independent and identically distributed (I.I.D.) random variables, theoretically the variances, $Var[Y]$ and $Var[\bar{Y}]$, and standard deviations, σ_Y and $\sigma_{\bar{Y}}$, of Y and $\bar{Y}$ can be respectively expressed by (4) and (5) [51].

$$Var[Y] = \sigma_Y^2 = E\left[(Y - E[Y])^2\right]$$
$$= \int_{-\infty}^{\infty} (y - E[Y])^2 \cdot f(y) \cdot dy \tag{4}$$

$$Var[\bar{Y}] = \sigma_{\bar{Y}}^2 = E\left[(\bar{Y} - E[\bar{Y}])^2\right] = E\left[(\bar{Y} - E[Y])^2\right]$$
$$= Var\left[\frac{1}{N} \cdot \sum_{n=1}^{N} Y_n\right] = \frac{1}{N^2} \cdot \sum_{n=1}^{N} Var[Y_n]$$
$$= \frac{1}{N^2} \cdot N \cdot Var[Y] = \frac{1}{N} \cdot \sigma_Y^2 \tag{5}$$

Equations (4) and (5) provide two key aspects. First, since the delta between the Monte Carlo estimate and ideal expectation, $(\bar{Y} - E[Y])$, represents the "error" of the experimental approximation, the variance and standard deviation of $\bar{Y}$ shown in (5) provide the quality merits of the Monte Carlo method about the degree of accuracy in the estimate after a finite number of the samples, N, based on the central limit theorem [51]. Second, the variance of $\bar{Y}$, $Var[\bar{Y}]$, or the power of the error reduces with increasing N in a reciprocal manner, which is a well-known probability theory: the weak law of large numbers [51]. This law ensures that the Monte Carlo estimate in (2) converges to the correct value as the number of samples increases as shown in (3). This is the fundamental reason why the RSA technique can achieve high resolution time-to-digital conversion with a very small amount of hardware cost by simply increasing the number of samples, but obviously the downside is its slow conversion-rate.

To realize the Monte Carlo method in a single-photon time-interval measurement, the RSA technique shown in Fig. 1 utilizes a TAC, VCDLs and an edge combiner to convert the one-time captured Δt information, which is the quantity under measurement, into a periodic signal, $CK\tau$, carrying a scaled version of Δt within each clock cycle for the RSA system to enable a process for an unlimited number of samples. The simplified TAC schematic shown in Fig. 2(a) first generates a single pulse, INT, whose pulse width equals the time difference, Δt, between the rising edges of the START and
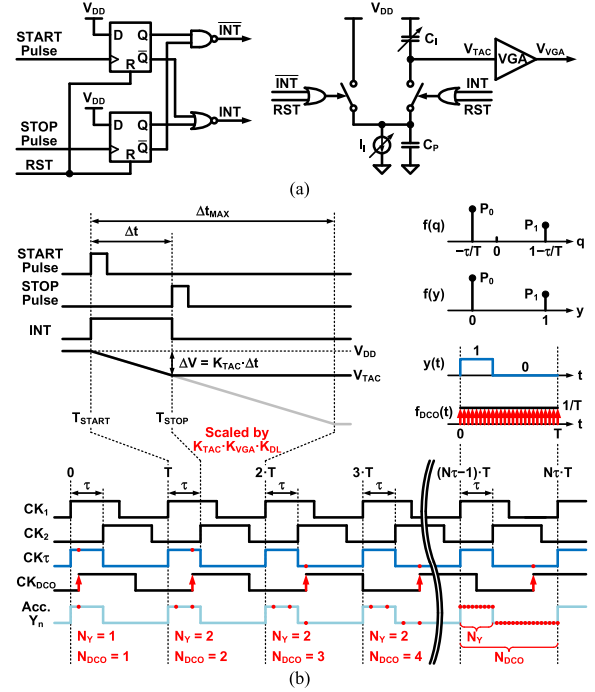

Fig. 2. (a) The simplified circuit schematics of the START/STOP pulse-edge detection and single-pulse-width to DC voltage convertor in TAC. (b) The signal waveforms and PDFs of the RSA time-interval measurement technique.

STOP pulses. Then, the INT pulse width enables an analog integrator implemented by a tunable constant current source, I_I, charging the integration and parasitic capacitors, C_I and C_P, to form the DC voltage, V_{TAC}. As the waveforms shown in Fig. 2(b), the time-interval, Δt, is converted and retained in the voltage domain as a differential DC voltage, $\Delta V = V_{DD} - V_{TAC} = K_{TAC} \cdot \Delta t$, where K_{TAC} is the conversion-gain of the TAC set by the magnitudes of I_I and C_I. A VGA buffers the constant voltage information with its gain, K_{VGA}, to one of the following VCDLs as shown in Fig. 1. Because of the control voltage difference between V_{DD} and V_{VGA}, these two identical VCDLs generate two clock signals, CK_1 and CK_2, with a common frequency of CK_{IN} (i.e., 1/T) and a constant delay, $\tau = K_{TAC} \cdot K_{VGA} \cdot K_{DL} \cdot \Delta t$, where K_{DL} is the conversion-gain of the VCDLs. After a rising-edge combiner, the $CK\tau$ signal merged from CK_1 and CK_2 is a periodic pulse carrying the scaled time-interval information, τ, as its duty-cycle in every T as shown in Fig. 2(b). Note that the $CK\tau$ signal may not explicitly exist in the real circuit implementation, which means the time difference between CK_1 and CK_2 can be detected in another manner [47] without generating $CK\tau$, but illustrating the $CK\tau$ waveform here can help the explanation. Also, during the Δt-to-τ conversion-process, the dynamic range of the time-interval measurement, Δt_{MAX}, is equivalently converted to the period of $CK\tau$, T, so the dynamic ranges for different time-interval measurements can be set by the tunability of the whole conversion-factor, $K_{TAC} \cdot K_{VGA} \cdot K_{DL}$, for a certain T. Changing the value of T is also possible but would affect the design specifications of the DCO and digital logics. It is important to note that this time-interval measurement system does not require extra hardware to calibrate its circuit non-idealities, including offset and non-linearity from the analog circuits and Δt-to-τ conversion-process, etc., since the RSA measurement

mechanism itself can be used for executing these calibration tasks [48]; for example, the calibration process for TAC circuit non-idealities is elaborated in the Appendix section.

At this point, the rest of the RSA process is realized by exploiting an independent clock signal, CK_{DCO}, from a free-running DCO with randomized frequency modulations to sample $CK\tau$ through a single D Flip-Flop, which is basically a "time to 1-bit digital" conversion process per sample. Under the assumptions of the uncorrelation [51], [52] between $CK\tau$ and CK_{DCO} and maximum one sample per T, any time instant within the period of $CK\tau$ can have an equal probability to be sampled by the rising edges of CK_{DCO} to form a one-dimensional geometric probability density function [51]. That is, CK_{DCO} can uniformly sample the waveform of $CK\tau$ within a T duration to generate either a Logic-1 or Logic-0 digital output, Y, based on the voltage of $CK\tau$ at each sampling instant. As shown in Fig. 2(b), the red dots on the $CK\tau$ waveform represents the sampled points in terms of their voltages and timings along the absolute time-domain axis. Because of the repeatability of $CK\tau$, all the sampled points (red dots) can be treated as being sampled and recorded within a single T, which is equivalent to a modulo-T operation of each absolute sampling instant. After many samples, the recorded samples are all accumulated within a single T duration as shown at the bottom (i.e., accumulated Y_n, "n" is the index of samples) of Fig. 2(b).

This whole RSA process can be described in three different aspects. First, if the number of samples is large enough, the PDF of the DCO sampling edge (red arrow) should be uniformly distributed across one period of the $CK\tau$ waveform, i.e., $f_{DCO}(t) = 1/T$ and $0 \leq t < T$ as shown in Fig. 2(b), while the $CK\tau$ waveform, y(t), behaves in Logic-1 and Logic-0 within and outside of the τ duration, respectively, with negligible rise/fall times of $CK\tau$; therefore, the probability of obtaining a Y_n as Logic-1, P_1, is exactly the ratio of τ to T (τ/T), which is a one-dimensional geometric probability in the time domain, i.e., a 1-bit TDC process. Second, in the voltage domain, the probability function of Y shown in Fig. 2(b) is a Bernoulli distribution [51] owning only two possible outcomes and the corresponding probability values, P_1 (= τ/T) and P_0 (= $1 - P_1$). Third, the Monte Carlo estimate, $\bar{Y}$, can be obtained by finding the ratio between the accumulated number of Logic-1s, N_Y, and the accumulated number of samples, N_{DCO}, as shown in both Fig. 1 and 2(b). The mathematical expressions of these three aspects are respectively summarized in the 1st, 2nd, and 3rd lines of (6):

$$
\begin{aligned}
E[Y] &= \int_0^T y(t) \cdot f_{DCO}(t) \cdot dt = \int_0^\tau 1 \cdot \frac{1}{T} \cdot dt = \frac{\tau}{T} \\
&= \int_{-\infty}^{\infty} y \cdot f(y) \cdot dy = 1 \cdot P_1 + 0 \cdot P_0 = P_1 \\
&= E[\bar{Y}] = \lim_{N_{DCO} \to \infty} \frac{\sum_{n=1}^{N_{DCO}} Y_n}{N_{DCO}} = \lim_{N_{DCO} \to \infty} \frac{N_Y}{N_{DCO}}
\end{aligned}
\tag{6}
$$

$$
\begin{aligned}
Var[\bar{Y}] &= \sigma_{\bar{Y}}^2 = \frac{1}{N_{DCO}} \cdot \sigma_Y^2 \\
&= \frac{1}{N_{DCO}} \cdot \left(E[Y^2] - E^2[Y] \right) = \frac{1}{N_{DCO}} \cdot P_1 \cdot P_0
\end{aligned}
\tag{7}
$$

In agreement with the weak law of larger numbers, the theoretical variance, $Var[\bar{Y}]$, i.e., the power of the estimation error or quantization noise, reciprocally degrades with the sampling number of N_{DCO}, and additionally it is a function of P_1 and P_0 as shown in (7), which matches the variance of a Bernoulli random variable. Finally, the time-interval under the RSA measurement can be obtained by

$$
\begin{aligned}
\Delta t &= \frac{\tau}{K_{TAC} \cdot K_{VGA} \cdot K_{DL}} \\
&\approx \frac{1}{K_{TAC} \cdot K_{VGA} \cdot K_{DL}} \cdot \frac{N_Y}{N_{DCO}} \cdot T
\end{aligned}
\tag{8}
$$

As mentioned, this RSA is basically a TDC process by measuring the clock duty-cycle with a 1-bit ADC (i.e., a single DFF), and its accuracy is improved with the number of samples. Therefore, the performance merit can be presented by the resolution and conversion-rate of each measurement result, $\bar{Y}$. For the resolution, the quantization noise, Q, of this TDC process is also a random variable, which equals the delta between the digital output Y and ideal signal information $E[Y] = \tau/T$. Since Y_n is either Logic-1 or Logic-0, Q_n only has two possible outcomes, $(1 - \tau/T)$ and $(0 - \tau/T)$, with the same PDF profile of Y as shown in Fig. 2(b) as well. It can be easily proven that $Var[Y] = Var[Q]$, and $Var[\bar{Y}] = Var[\bar{Q}] = P_1 \cdot P_0 / N_{DCO}$ as expressed in (7). When only the quantization noise power (i.e., $Var[\bar{Q}]$) is considered, the effective number of binary bits (ENOB) of the RSA technique can be expressed as

$$
\begin{aligned}
&ENOB \\
&= \frac{SNR}{6.02} \\
&= \begin{cases}
\dfrac{log_{10}\left(\dfrac{P_0^2}{Var[\bar{Y}]}\right)}{0.602} = \dfrac{log_{10}\left(\dfrac{P_0}{P_1} \cdot N_{DCO}\right)}{0.602}, & P_1 \leq \dfrac{1}{2} \\[4mm]
\dfrac{log_{10}\left(\dfrac{P_1^2}{Var[\bar{Y}]}\right)}{0.602} = \dfrac{log_{10}\left(\dfrac{P_1}{P_0} \cdot N_{DCO}\right)}{0.602}, & P_1 > \dfrac{1}{2}
\end{cases}
\end{aligned}
\tag{9}
$$

where the signal power is either P_1^2 or P_0^2 based on the magnitude of P_1 compared to 0.5 due to the symmetric and signal-dependent variance property of the Bernoulli distribution, which is further elaborated by the simulation results in Section IV. Conceptually, Equation (9) can be examined by considering when $N_{DCO} = 1$, then $Var[\bar{Y}] = P_1 \cdot P_0$, which is exactly the variance of a single sample experiment of flipping a coin (i.e., either a single "head" or "tail"). If the coin is fair, then the SNR = 0 dB and ENOB = 0 bits because of the equal powers between the signal and quantization noise. With increasing N_{DCO}, the quantization noise power, $Var[\bar{Y}]$, degrades 3 dBW per octave of N_{DCO}. To reach 12-ENOB accuracy, for example, N_{DCO} has to reach at least 2^{24} ($\approx$ 16.8 million) samples, which is actually very low cost in hardware since the RSA technique only needs the accumulation result of the 1-bit samples for averaging. For the conversion-rate, it is essentially determined by the sampling frequency from DCO and the required number of samples. In the same example, an average 4-GHz DCO sampling frequency with $N_{DCO} = 2^{24}$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

WU *et al.*: RSA TECHNIQUES FOR SINGLE-PHOTON ARRIVAL-TIME DETECTIONS IN QUANTUM APPLICATIONS 5
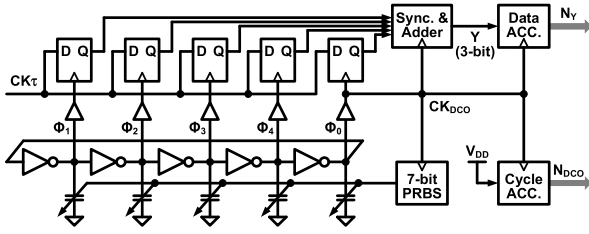

Fig. 3. The conversion-rate of the RSA technique can be linearly enhanced by utilizing the inherent multi-phases of the DCO to simultaneously sample $CK\tau$.

requires about 4.2 ms to complete a single RSA measurement, which is not superior in terms of the deadtime and frame-rate requirements in quantum applications. One of the conversion-rate enhancement approaches is done by utilizing the inherent multi-phases of the DCO to simultaneously sample $CK\tau$ and accumulate the multiple-bit data per DCO cycle as shown in Fig. 3 [48]. In addition to having five-stage outputs in this example, a pseudo-differential DCO architecture overall can linearly improve the number of sampling phases by $(5\cdot2)\times$ per cycle, which can moderately push the conversion-rate up to 2.4 kHz ($\approx 10/4.2$-ms) for a 12-ENOB RSA-based TDC. For the sake of simplicity, the theoretical analysis in the remainder of this paper assumes only using a single phase from the DCO. Also, note that ENOB here only considers quantization noise because the measurement quantity τ/T itself can be treated as a constant DC signal. Any zero-mean noise induced jitter on $CK\tau$ can be filtered out by the averaging process (i.e., data accumulation); any phase-noise on CK_{DCO} can help to generate continuous (i.e., zero step size) random sampling PDFs (discussed in Section III-B and Appendix). Therefore, this RSA implementation can possess extremely high noise immunity and measurement accuracy.

A. Synchronous Random Sampling-and-Averaging

To elaborate on the connection between the probability theory and realistic implementation, this paper splits the RSA technique into two categories, synchronous and asynchronous RSA, based on the relationship between $CK\tau$ and CK_{DCO}. The definition of synchronous RSA is illustrated in Fig. 4(a) associated with multiple different values of the oversampling ratio (OSR) defined by the number of samples per $CK\tau$ period, T, which is not the same definition in the sampling theorem for anti-aliasing; and note that the intent of RSA is to extract the duty cycle of $CK\tau$, not to recover the $CK\tau$ signal or its frequency. Therefore, synchronous RSA here means the number of samples per T across the entire sampling process is consistently set by OSR even though each sampling PDF is still I.I.D. As shown in Fig. 4(a), when OSR = 1/2, $CK\tau$ is always uniformly sampled by one CK_{DCO} edge (i.e., arrows) within every 2·T; one sample per T for OSR = 1; one sample per T/2 for OSR = 2; and so forth. Consistently, the probability of each sampling edge occurs uniformly within the region of T/OSR; in other words, when OSR = 4, for example, each $CK\tau$ period is sampled by four independent CK_{DCO} edges, and the occurrence of each sampling edge has a uniform PDF bounded within its own T/4 region. Under these criteria, $CK\tau$ can be seamlessly and uniformly sampled regardless of the values of OSR, but OSR becomes the main factor affecting

performance under a certain value of τ/T. Note that the circuit realization of generating these well-bounded random sampling edges is actually very expensive; and the resolution or step size of the random sampling edges has to be always higher than the target ENOB, which is almost impossible for the sub-picosecond accuracy requirement. These practical concerns can be resolved by asynchronous RSA discussed later, but the concept of synchronous RSA is more friendly for the purposes of initial instruction and theoretical analysis. Therefore, the assumption here is that the resolution of each sampling PDF for synchronous RSA is high enough to behave like a "continuous" probability density function [51] within its own distribution boundary.

Once synchronous RSA is well defined, the theoretical variances with respect to different OSRs can be derived. In the case of OSR = 1 as shown in Fig. 4(a), the probability of obtaining a Logic-1 per sample is exactly $\tau/T = P_1$, so the expectation and theoretical variance of each RSA measurement are equal to the results shown in (6) and (7), respectively. In the cases of OSR < 1 (i.e., subsampling), for example, OSR = 1/2 as shown in Fig. 4(a), although the sampling region is extended to 2·T for each sample, the probability of obtaining a Logic-1 maintains the same as $\tau/T = (2\cdot\tau)/(2\cdot T) = P_1$ because the high-voltage level duration is also doubled. In sum, when OSR ≤ 1, the expectation and theoretical variance of synchronous RSA can be always represented by (6) and (7), and note that the OSR has to be the reciprocal of a positive integer to maintain a constant $CK\tau$ duty-cycle within each uniformly distributed sampling region.

In the cases of OSR > 1, each uniformly distributed sampling region has been labeled by an index, k, which is a positive integer from 1 to OSR, as shown in Fig. 4(a). Also, one of the equally divided regions within each T is shaded to highlight that the sampling outcomes within these regions are possibly to be Logic-1 or 0, i.e., the $CK\tau$ high-to-low voltage transitions occur within these regions. For example, when OSR = 2 and $\tau/T < 0.5$, each T is equally split into two sampling regions, and only the first-half region (i.e., k = 1) can possibly have a Logic-1 or 0 outcome, and the probability of obtaining Logic-1 is $\tau/(T/2) = 2\cdot P_1$, where P_1 ($= \tau/T$) is the probability of obtaining Logic-1 when OSR ≤ 1, and the outcome of the second half (i.e., k = 2) is deterministically a Logic-0. Therefore, the equivalent outcome per T should be represented by the average of the first- and second-half sampling outcomes. This example indicates the variance of each RSA measurement result, Var[$\bar{Y}$], is the function of OSR and one of the region-index values, where the $CK\tau$ high-to-low voltage transition occurs. Intuitively, the oversampling process improves the resolution or reduces the quantization noise of the RSA measurement since those unshaded regions designate the deterministic coarse information (i.e., MSBs) of the $CK\tau$ duty-cycle; then the shaded regions offer the fine information (LSB) determined by the high-vs-low voltage duration ratio within them. Higher OSRs confine the uncertain outcomes within finer shaded regions to improve the resolution like the quantization process of an ADC. Alternatively speaking, the equivalent outcome per T can be grouped by averaging all sub-region outcomes within each T; and only one of these

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

6

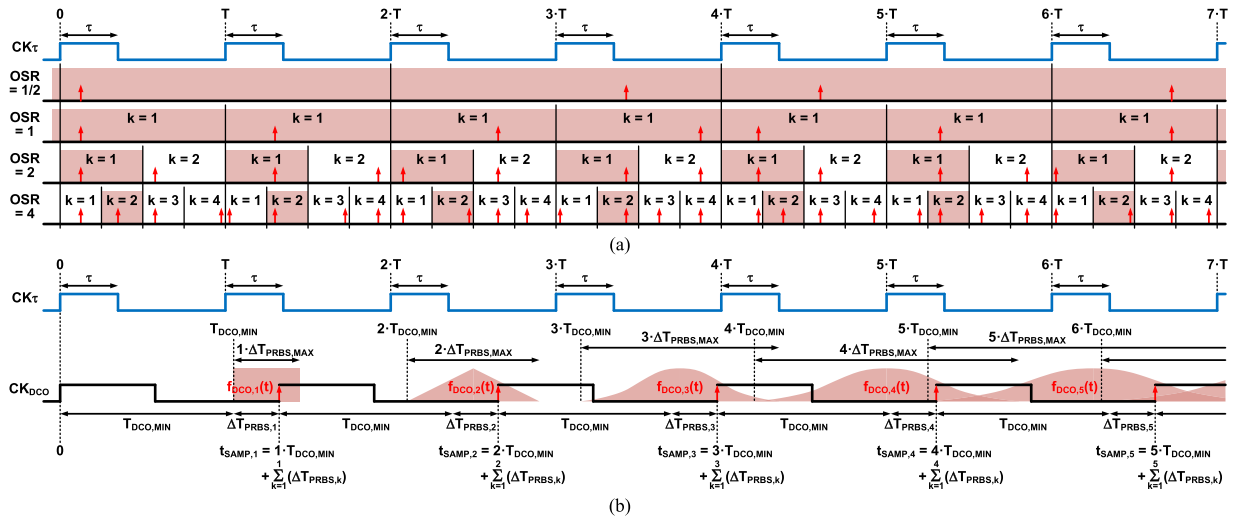IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—I: REGULAR PAPERS

Fig. 4. (a) The relationship between $CK\tau$ and CK_{DCO} random sampling PDFs of the synchronous RSA technique with different OSRs and their corresponding equally divided sampling regions. (b) The relationship between $CK\tau$ and CK_{DCO} random sampling PDFs of the asynchronous RSA technique with the nature of the DCO phase-noise accumulation property.

sub-regions can have an uncertain Logic-1 or 0 outcome. Therefore, when OSR > 1, the equivalent outcome per T, $Y_{OSR,n}$, has a value between 0 and 1, which represents a higher ENOB per T.

From the perspective of probability theory, evenly segmenting each T by OSR and then averaging all sub-region results to form the equivalent outcome per T are actually creating correlations among the samples within each T. Also, based on the theoretical variances in (11) and simulation results demonstrated in Section IV, the correlations due to the oversampling are negative and effectively reduce the overall variances. Overall, for the cases of OSR $\geq$ 1, the theoretical variance in (11) is obtained by finding the total covariance sum associated with the joint PDFs among all samples within each T. Thus, the expectation and theoretical variance of the synchronous RSA measurement can be generalized as (10) and (11), respectively, shown at the bottom of the page. "k" in (10) simply represents the sweeping index of the summation operator to find $Y_{OSR,n}$ per T, but "k" in (11) is a specific integer number within 1 to OSR based upon the transition of $CK\tau$ as shown in Fig. 4(a). That is, for a certain τ/T under the measurement, only a certain k represents the shaded region and can be obtained in the 1st line of (11). Compared to the cases of OSR $\leq$ 1, the weak law of large numbers is still valid in (11), and ENOB also follows the same definition in (9) with the variance replaced by (11). The expectation of $\bar{Y}$ in (10) stays the same as (6) regardless of the OSR value.

B. Asynchronous Random Sampling-and-Averaging

By definition in this paper, the asynchronous RSA technique can equivalently perform I.I.D. random sampling within a $CK\tau$ period, T, without implementing strict sampling PDF boundaries and frequency relationships between $CK\tau$ and CK_{DCO}, which can be practically realized by low-cost and power/area efficient integrated circuits. Thus, the implementation related content, like the circuit/system overview in Section II and the beginning of Section III, is all based on the concept of asynchronous RSA. This sub-section describes the theory of how I.I.D. random sampling can be performed in an asynchronous manner for realistic applications.

First of all, a few parameters have to be introduced: $T_{DCO,n}$ is the n-th period of the free-running DCO; $t_{SAMP,n}$ is the n-th absolute sampling time at the n-th rising edge of the DCO; $\Delta T_{PRBS,n}$ is the n-th DCO period extension controlled by a digital pseudo-random-binary-sequence generator (i.e., PRBS Gen. in Fig. 1), which dynamically modulates the DCO period, ($T_{DCO,n} = T_{DCO,MIN} + \Delta T_{PRBS,n}$), between ($T_{DCO,MIN} + 0$) and ($T_{DCO,MIN} + \Delta T_{PRBS,MAX}$) [48]. Both $T_{DCO,MIN}$ and $\Delta T_{PRBS,MAX}$ can be coarsely adjusted by their own static controls.

One example of the asynchronous RSA sampling process is shown in Fig. 4(b), where the waveforms of $CK\tau$ and CK_{DCO} are assumed to have coincident rising edges at t = 0, for the sake of simplicity, though this is not a required condition. One of the most important properties of a free-running ring oscillator is the phase-noise accumulation behavior under

$$E[\bar{Y}] = \lim_{N_{DCO} \to \infty} \frac{1}{N_{DCO}} \cdot \sum_{n=1}^{N_{DCO}} Y_n = \lim_{N_{DCO} \to \infty} \frac{1}{\frac{N_{DCO}}{OSR}} \cdot \sum_{n=1}^{\frac{N_{DCO}}{OSR}} \frac{\sum_{k=1}^{OSR} Y_k}{OSR} = \lim_{N_{DCO} \to \infty} \frac{1}{\frac{N_{DCO}}{OSR}} \cdot \sum_{n=1}^{\frac{N_{DCO}}{OSR}} Y_{OSR,n} \quad (10)$$

$$\frac{k-1}{OSR} < \left(\frac{\tau}{T} = P_1 = 1 - P_0\right) \leq \frac{k}{OSR}, k \in 1, 2, \ldots, OSR$$

$$Var[\bar{Y}] = Var[\bar{Q}] = \sigma_{\bar{Y}}^2 = \begin{cases} \frac{(OSR \cdot P_1 - k + 1)}{N_{DCO}} \cdot \left(\frac{k}{OSR} - P_1\right), & \frac{k}{OSR} \leq \frac{1}{2} \\ \frac{(OSR \cdot P_0 + k - OSR)}{N_{DCO}} \cdot \left(\frac{OSR - k + 1}{OSR} - P_0\right), & \frac{k}{OSR} > \frac{1}{2} \end{cases} \quad (11)$$

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

WU *et al.*: RSA TECHNIQUES FOR SINGLE-PHOTON ARRIVAL-TIME DETECTIONS IN QUANTUM APPLICATIONS
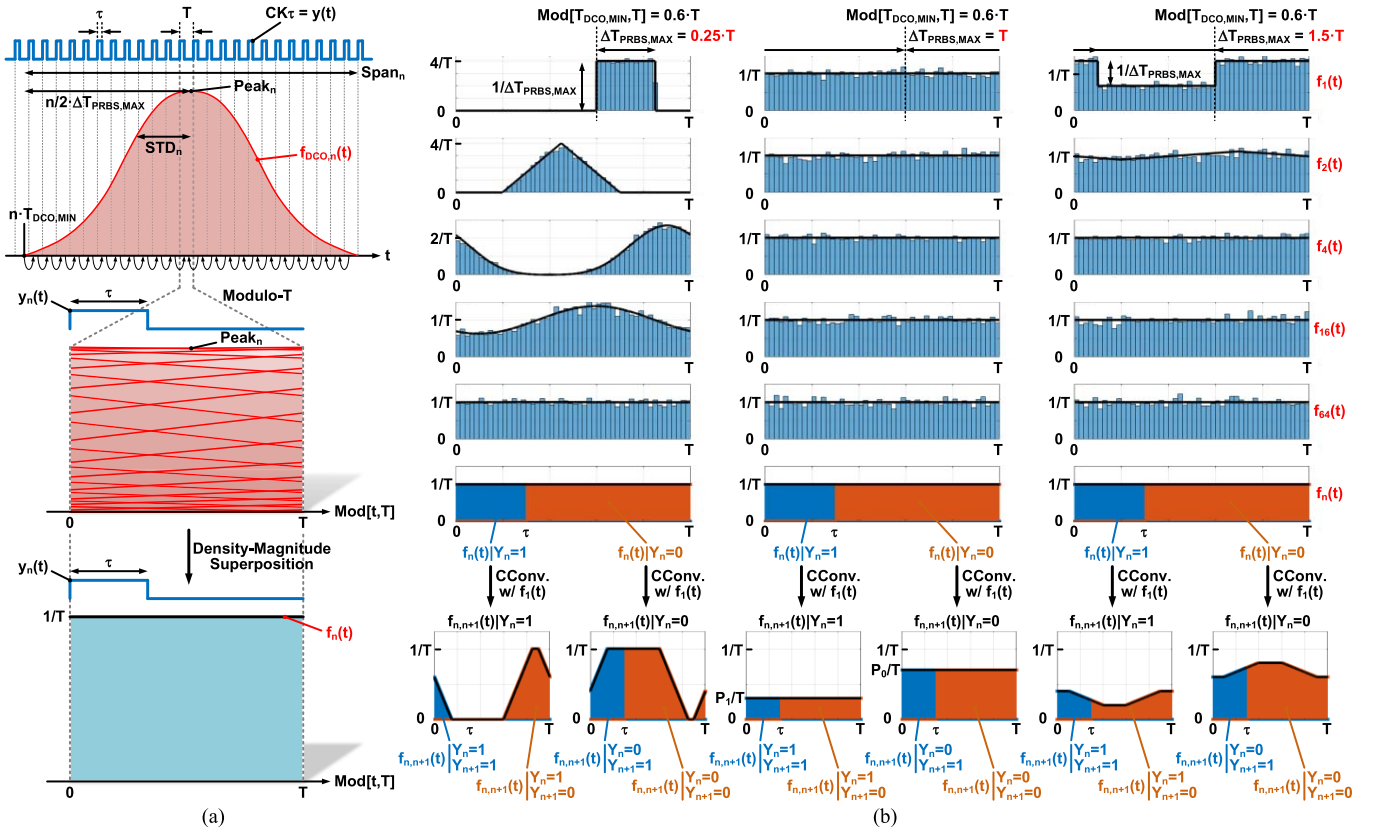
7

Fig. 5. (a) The concept of the modulo-T operation and density-magnitude convergence of a random sampling PDF in the asynchronous RSA implementation. (b) Upper-half: the theoretical and statistical ($N_{EXP} = 2^{13}$) results of the modulo-T random sampling PDFs, $f_1(t)$, $f_2(t)$, $f_4(t)$, $f_{16}(t)$, $f_{64}(t)$ and $f_n(t)$ under the three $\Delta T_{PRBS,MAX}$ scenarios. Lower-half: the process of deriving the conditional joint PDFs for the theoretical covariance calculation of the adjacent samples, Y_n and Y_{n+1}.

the presence of any noise source regardless of the artificial noise from the PRBS generator, device thermal/flicker noise, or power-supply noise. A strong phase-noise accumulation tends to form a widely distributed random sampling PDF with a minimal number of samples, so a noise-energy dominated PRBS noise source is preferrable in the asynchronous RSA technique. In other words, theoretically any I.I.D. noise source or combinational effect of multiple I.I.D. noise sources are all applicable to the following analysis; this paper focuses on artificial PRBS noise mainly because of its controllable noise energy and dominant noise accumulation effect for asynchronous RSA. However, though the natural noise sources have relatively low energy, it is very important to emphasize that they offer arbitrarily small phase-noise accumulations to fill the gaps of finite PRBS noise step sizes. In other words, the accumulations of natural phase-noise help asynchronous RSA to possess true "continuous" sampling PDFs. This concept can be more comprehensive at the end of this sub-section.

Under the parameters and noise definitions, the n-th absolute sampling time can be represented by (12), and a few of $t_{SAMP,n}$ are also labeled in Fig. 4(b).

$$
\begin{aligned}
t_{SAMP,n} &= \sum_{k=1}^{n} T_{DCO,k} \\
&= \sum_{k=1}^{n} \left(T_{DCO,MIN} + \Delta T_{PRBS,k} \right) \\
&= n \cdot T_{DCO,MIN} + \sum_{k=1}^{n} \Delta T_{PRBS,k}
\end{aligned} \quad (12)
$$

Each $t_{SAMP,n}$ contains two components: the deterministic term (i.e., $n \cdot T_{DCO,MIN}$) and stochastic term due to the phase-noise accumulation, which describes the uncertainty of each sampling instant and can only be represented by a PDF, $f_{DCO,n}(t)$. Therefore, the n-th CK_{DCO} rising edge, i.e., the red arrows in Fig. 4(b), occurs randomly but is confined within the distribution span and density magnitude of its own PDF, i.e., pink shaded areas in Fig. 4(b). More importantly, the stochastic term of each $t_{SAMP,n}$ is the accumulation of "n" I.I.D. random variables (i.e., $\Delta T_{PRBS,k}$, k = 1 to n) created by the PRBS generator for "n" times as shown in the second term of (12); equivalently, $f_{DCO,n}(t)$ is the convolution result of total "n" uniformly distributed PDFs from the PRBS generator. When n = 1, the stochastic term of $t_{SAMP,1}$ only has one random variable from the PRBS generator without any accumulation, so its PDF, $f_{DCO,1}(t)$, is the fundamental uniform distribution with a $\Delta T_{PRBS,MAX}$ distribution span and constant $1/\Delta T_{PRBS,MAX}$ density magnitude as shown in Fig. 4(b). When n = 2, the stochastic term of $t_{SAMP,2}$ is the summation of two I.I.D. random variables, which are independent from each other and have an identically distributed PDF though they are sequentially created by the PRBS generator. Based on the convolution theorem [51], the PDF, $f_{DCO,2}(t)$, of the sum of these two I.I.D. random variables is the convolution of their individual PDFs, which becomes an isosceles triangular distribution with a $2 \cdot \Delta T_{PRBS,MAX}$ distribution span and $1/\Delta T_{PRBS,MAX}$ peak

density magnitude. With increasing "n", the central limit theorem guarantees that the PDF of the stochastic term of $t_{SAMP,n}$ converges to a Gaussian distribution regardless of the PDF of the single random variable from the PRBS generator. Succinctly, the mean, standard deviation, span, and peak of $f_{DCO,n}(t)$ are well defined in (13) and illustrated at the top of Fig. 5(a):

$$Mean_n = n \cdot T_{DCO,MIN} + \frac{n}{2} \cdot \Delta T_{PRBS,MAX}$$

$$Span_n = n \cdot \Delta T_{PRBS,MAX}$$

$$STD_n = \sqrt{\frac{n}{12}} \cdot \Delta T_{PRBS,MAX}$$

$$Peak_n = \sqrt{\frac{6}{\pi \cdot n}} \cdot \frac{1}{\Delta T_{PRBS,MAX}}$$

$$f_{DCO,n}(t) = Conv\left[f_{DCO,n-1}(t), f_{DCO,1}(t)\right], n > 1$$

$$\approx \frac{1}{STD_n \cdot \sqrt{2 \cdot \pi}} \cdot exp\left[\frac{-(t - Mean_n)^2}{2 \cdot STD_n^2}\right], n \gg 1$$

(13)

Note that the time-domain variable "t" represents absolute time values referenced to t = 0 as shown in Fig. 4(b). Since the deterministic term (i.e., $n \cdot T_{DCO,MIN}$) of $t_{SAMP,n}$ sets the left distribution bound of $f_{DCO,n}(t)$, the PDFs of adjacent samples quickly exhibit a large amount of distribution overlaps along with the growth of $Span_n$ as the examples shown in Fig. 4(b). It is very important to note that the overlaps among multiple PDFs in the absolute time-domain do not mean the order of the $t_{SAMP,n}$ occurrences would change its monotonic ascending behavior; for example, $t_{SAMP,4}$ always occurs before $t_{SAMP,5}$ even though $f_{DCO,4}(t)$ and $f_{DCO,5}(t)$ have a distribution overlap. This indicates the correlation among all $t_{SAMP,n}$ (or $f_{DCO,n}(t)$) because of the phase-noise accumulation since the very first sample to the n-th sample as shown in (12).

So far, the generation of these DCO sampling PDFs for asynchronous RSA seems to have two major violations (i.e., not uniform and not independent) from the assumption described in (6), (7) and Fig. 2(b): first, $f_{DCO,n}(t)$ turns out to be a Gaussian, not a uniform distribution regardless of the initial PDF of the noise source; second, all of the sampling instants, $t_{SAMP,n}$, become correlated due to the phase-noise accumulation property of the DCO. Actually, these concerns have been effectively resolved by a time-domain "modulo operation" with proper noise-energy setups (i.e., $\Delta T_{PRBS,MAX}$), so the Gaussian and correlated characteristics of all $f_{DCO,n}(t)$ can be turned into uniform and independent sampling PDFs, $f_n(t)$.

The modulo operation is automatically accomplished by the effort of converting Δt into the duty-cycle, τ/T, of the periodic signal $CK\tau$ as described in Section II. The concept of this modulo-T operation is shown in Fig. 5(a): when $n \gg 1$, the distribution span of the Gaussian PDF, $f_{DCO,n}(t)$, covers multiple $CK\tau$ periods, and this periodicity basically slices the entire Gaussian PDF into multiple segments in the absolute time-domain. Thus, each PDF segment has its own density distribution, of course, but all have the same distribution span, T, and all see (or sample) the same one-period waveform of $CK\tau$, $y_n(t)$. Equivalently, all segmented PDFs are strictly

distributed within a modulo-T time-interval between 0 and T as shown in the middle of Fig. 5(a), where the net density-magnitude at any time instant is contributed by all segmented PDFs. That is, the equivalent PDF of the n-th DCO sampling instant, $f_n(t)$, is the superposition of the all segmented PDFs from $f_{DCO,n}(t)$:

$$S = Ceil[0.5 \cdot Span_n / T]$$

$$Mean_n' = Mod[Mean_n, T]$$

$$f_n(t) \approx \frac{1}{STD_n \cdot \sqrt{2 \cdot \pi}}$$

$$\cdot \sum_{k=-S}^{S} exp\left[\frac{-(t - Mean_n' - k \cdot T)^2}{2 \cdot STD_n^2}\right], n \gg 1$$

(14)

where the time-domain variable "t" is confined within [0, T]; $(2 \cdot S + 1)$ is the number of segments set by the $Span_n$; $Mean_n'$ and $k \cdot T$ are used to shift these $(2 \cdot S + 1)$ segments to the modulo-T time-interval, [0, T). When $n \gg 1$, $f_n(t)$ converges to a constant 1/T across the single T span as shown at the bottom of Fig. 5(a); this fact can be proven by both mathematical calculation of (14) and statistical simulations shown in the upper-half of Fig. 5(b), which demonstrates $f_1(t)$, $f_2(t)$, $f_4(t)$, $f_{16}(t)$, and $f_{64}(t)$ for three representative $\Delta T_{PRBS,MAX}$ scenarios (i.e., <, =, and > T). The black curves are the theoretical PDFs from the mathematical calculations while the light-blue histograms are obtained from the statistical simulations with 2^{13} experiments. Clearly, the theoretical and experimental results align pretty well across all PDFs along with different settings of noise-energy from the PRBS generator. More importantly, with increasing "n", all sampling PDFs converge to a uniformly distributed PDF with a constant density magnitude 1/T across the [0, T) distribution span, independent from the parameters of $T_{DCO,MIN}$, $\Delta T_{PRBS,MAX}$, and even "n" when $n \gg 1$ as illustrated at the bottom of Fig. 5(a). In other words, for all "n" $\gg 1$, $f_n(t)$ becomes an "identically distributed" PDF, which satisfies the "second" criterion of the I.I.D. random variable and can be implemented by the low-cost circuitry described in Section II. This convergence of the uniform distribution also guarantees the convergence of the asynchronous RSA measurement result:

$$E[\bar{Y}] = E\left[\frac{1}{N_{DCO}} \cdot \sum_{n=1}^{N_{DCO}} Y_n\right]$$

$$= \frac{1}{N_{DCO}} \cdot \sum_{n=1}^{N_{DCO}} E[Y_n] = E[Y_n]$$

$$= \int_0^T y_n(t) \cdot f_n(t) \cdot dt = \int_0^\tau \frac{1}{T} \cdot dt + \int_\tau^T \frac{0}{T} \cdot dt$$

$$= \frac{\tau}{T} = P_1$$

(15)

which exactly matches to the expectation in (6).

One more practical concern may be raised for those PDFs which exhibit a non-uniform distribution since their values of "n" are not large enough. For example, in the case of $\Delta T_{PRBS,MAX} = 0.25 \cdot T$, $f_n(t)$ obviously is not a uniformly distributed PDF until n = 64. The concern is whether these non-uniform PDFs affect the overall RSA measurement result,

i.e., (15). The short answer is negative: these non-uniform PDFs can only be observed when all of the experiments have the same absolute time reference at t = 0 (i.e., the initial condition) as the assumption for (12) and Fig. 4(b). In reality, the asynchronous relationship between $CK\tau$ and CK_{DCO} randomizes the initial condition of the sampling process, especially for $f_{DCO,1}(t)$, so all $f_n(t)$ actually exhibit a uniform distributed PDF within [0, T] even though the value of "n" is one or close to one. This can be verified by consolidating the randomized initial condition, t_{INT}, into the unit impulse, $\delta(t - t_{INT})$, discussed in (16) and the following paragraphs.

Similar to $f_{DCO,n}(t)$ in (13), the exact form of $f_n(t)$ can be expressed in the format based on the convolution theorem [51] as well, but now it becomes a modulo-T circular convolution process [54] due to the combined operations of the linear convolution and modulus:

$$f_1(t) = CConv[f_{DCO,1}(t), \delta(t), T]$$
$$f_n(t) = CConv[f_{n-1}(t), f_1(t), T], \quad n > 1 \quad or$$
$$f_{n+1}(t) = CConv[f_n(t), f_1(t), T], \quad n \geq 1 \quad (16)$$

where $\delta(t)$ is the unit impulse; the "t" of $f_1(t)$, $f_{n-1}(t)$, $f_n(t)$ and $f_{n+1}(t)$ is the modulo-T time-domain variable within [0, T], but the "t" of $f_{DCO,1}(t)$ and $\delta(t)$ is the absolute time-domain variable referenced to t = 0. Based on (16), $f_1(t)$ plays as not only the PDF of the first sampling instant but also the fundamental PDF element to obtain any $f_n(t)$ from $f_{n-1}(t)$.

Here summarizes the important attributes of $f_1(t)$. The distribution of $f_1(t)$ always starts at the remainder of a single $T_{DCO,min}$ divided by T, e.g., Mod[$T_{DCO,MIN}$, T] = 0.6·T in Fig. 5(b). Whenever the PDF reaches t = T, it circulates back to t = 0 and then continues its distribution toward t = T as shown in all PDF plots of Fig. 5(b), but the top-row plots demonstrate this modulo-T circular behavior more obviously since $f_1(t)$ has the minimal distribution span among all $f_n(t)$. In the case of $\Delta T_{PRBS,MAX} < T$, for example $\Delta T_{PRBS,MAX} = 0.25 \cdot T$, intuitively $f_1(t)$ has non-zero values from 0.6·T to 0.85·T. When $\Delta T_{PRBS,MAX} = T$ or Mod[$\Delta T_{PRBS,MAX}$, T] = 0, $f_1(t)$ circulates multiple integer cycles uniformly from 0.6·T and then back to 0.6·T. In the case of $\Delta T_{PRBS,MAX} > T$ and Mod[$\Delta T_{PRBS,MAX}$, T] $\neq$ 0, $f_1(t)$ exhibits two non-zero density magnitudes because $f_{DCO,1}(t)$ circulates within [0, T] multiple times with a non-zero remainder, and the delta between these two density magnitudes is $1/\Delta T_{PRBS,MAX}$ as the example of $f_1(t)$ shown in Fig. 5(b) for $\Delta T_{PRBS,MAX} = 1.5 \cdot T$. This indicates $f_1(t)$ itself can be directly converged to a uniform distribution by having a $\Delta T_{PRBS,MAX} \gg T$ to level its density magnitude. The attributes of $f_1(t)$ discussed so far seem unimportant because anyway $f_n(t)$ converges to a uniform distribution as verified. However, from the perspective of the correlations among all sampling PDFs, i.e., $f_n(t)$ where n = 1 to N_{DCO}, this fundamental PDF element, $f_1(t)$, plays a significant role even though all $f_n(t)$ are identical when n $\gg$ 1.

To examine the correlation among all sampling PDFs, the process can start with calculating the covariance of the adjacent samples, Y_n and Y_{n+1}:

$$Cov[Y_n, Y_{n+1}]$$
$$= \iint_R (y_n - E[Y]) \cdot (y_{n+1} - E[Y]) \cdot f(y_n, y_{n+1}) \cdot dy_n \cdot dy_{n+1}$$

$$= \int_0^T (y_n(t) - P_1) \cdot (y_{n+1}(t) - P_1) \cdot f_{n,n+1}(t) \cdot dt$$
$$= (1 - P_1) \cdot (1 - P_1) \cdot \int_0^\tau \left(f_{n,n+1}(t) \,\middle|\, \begin{matrix} Y_n = 1 \\ Y_{n+1} = 1 \end{matrix} \right) \cdot dt$$
$$+ (1 - P_1) \cdot (0 - P_1) \cdot \int_\tau^T \left(f_{n,n+1}(t) \,\middle|\, \begin{matrix} Y_n = 1 \\ Y_{n+1} = 0 \end{matrix} \right) \cdot dt$$
$$+ (0 - P_1) \cdot (1 - P_1) \cdot \int_0^\tau \left(f_{n,n+1}(t) \,\middle|\, \begin{matrix} Y_n = 0 \\ Y_{n+1} = 1 \end{matrix} \right) \cdot dt$$
$$+ (0 - P_1) \cdot (0 - P_1) \cdot \int_\tau^T \left(f_{n,n+1}(t) \,\middle|\, \begin{matrix} Y_n = 0 \\ Y_{n+1} = 0 \end{matrix} \right) \cdot dt$$
$$(17)$$

where the 2nd line of (17) is based on the fundamental covariance definition of two random variables, Y_n and Y_{n+1}, on the same sample space, R, with their joint PDF, $f(y_n, y_{n+1})$, and PDF variables, y_n and y_{n+1}. Similar to (6) and (15), the 3rd line of (17) uses the one-dimensional geometric probability format to represent the covariance with the modulo-T time-domain PDF variables, $y_n(t)$ and $y_{n+1}(t)$, and joint PDF, $f_{n,n+1}(t)$, for asynchronous RSA. $y_n(t)$ and $y_{n+1}(t)$ are identical due to the $CK\tau$ periodicity as shown in Fig. 5(a), and the possible outcomes of Y_n and Y_{n+1} are either Logic-1 or Logic-0 both with the expectation E[Y] = τ/T = P_1 verified in (15). By taking advantage of simple binary values, Cov[Y_n, Y_{n+1}] can be expanded into the summation of four conditional covariances based on the total four possible combinations of Y_n and Y_{n+1} with their corresponding conditional joint PDFs as shown in the 4th to 7th lines of (17). The approach of finding these four conditional joint PDFs is illustrated at the lower-half of Fig. 5(b). As discussed, the generic case (i.e., n $\gg$ 1) of $f_n(t)$ (black curves) is a uniform distribution for all $\Delta T_{PRBS,MAX}$ scenarios, so the conditional PDFs, $f_n(t)|Y_n = 1$ (blue regions) and $f_n(t)|Y_n = 0$ (orange regions), are simply set by the value of $y_n(t)$ across the modulo-T time-interval, [0, T), and all scenarios have the same conditional PDFs, $f_n(t)|Y_n$. To further obtain the conditional joint PDFs, $f_{n,n+1}(t)|Y_n$ (black curves), each $f_n(t)|Y_n$ has to circularly convolute with the fundamental PDF element, $f_1(t)$, which is a function of $\Delta T_{PRBS,MAX}$, so different $f_1(t)$ generate their corresponding $f_{n,n+1}(t)|Y_n = 1$ and $f_{n,n+1}(t)|Y_n = 0$ as shown in Fig. 5(b) and below:

$$(f_{n,n+1}(t)|Y_n = 1) = CConv[(f_n(t)|Y_n = 1), f_1(t), T]$$
$$(f_{n,n+1}(t)|Y_n = 0) = CConv[(f_n(t)|Y_n = 0), f_1(t), T]$$
$$(18)$$

It is important to note the difference between $f_{n+1}(t)$ in (16) and $f_{n,n+1}(t)|Y_n$ in (18): $f_{n+1}(t)$ is obtained from $f_n(t)$, and it turns out to be independent from $f_1(t)$; but, $f_{n,n+1}(t)|Y_n$ is from $f_n(t)|Y_n$, and their correlation is determined by $f_1(t)$. Finally, the four conditional joint PDFs of each scenario are obtained by including the value of $y_{n+1}(t)$ across the modulo-T time-interval, [0, T) to fully cover all possible conditions of Y_n and Y_{n+1}.

In the case of Mod[$\Delta T_{PRBS,MAX}$, T] = 0, i.e., the middle column of Fig. 5(b), the four conditional joint PDFs all maintain constant density magnitudes within their own integral time-intervals, [0, τ) and [τ, T), so the covariance of the

adjacent samples can be further derived for this scenario easily:

If $Mod\left[\Delta T_{PRBS,MAX}, T\right]$ equals 0,

$$Cov\left[Y_n, Y_{n+1}\right] = P_0 \cdot P_0 \cdot \frac{P_1}{T} \cdot \tau - P_0 \cdot P_1 \cdot \frac{P_1}{T} \cdot (T - \tau)$$
$$- P_1 \cdot P_0 \cdot \frac{P_0}{T} \cdot \tau + P_1 \cdot P_1 \cdot \frac{P_0}{T} \cdot (T - \tau) = 0 \quad (19)$$

Based on the result of (19), i.e., a zero covariance between any adjacent samples, Y_n and Y_{n+1}, and the accumulated relation from $f_1(t)$ to $f_{n+1}(t)$ shown in (16), Mod[$\Delta T_{PRBS,MAX}$, T] = 0 is the necessary condition for all, not just adjacent, $f_n(t)$ to be "independent." By consolidating the identicality and independency of $f_n(t)$ for all $\Delta T_{PRBS,MAX}$ scenarios illustrated in Fig. 5(b), Mod[$\Delta T_{PRBS,MAX}$, T] = 0 is the requirement to form an asynchronous RSA measurement with "independent and identically distributed (I.I.D.)" random sampling PDFs. This conclusion equivalently matches the requirement of synchronous RSA with OSR = 1 where the PDF of the CK_{DCO} sampling edge is uniformly distributed across one $CK\tau$ cycle as discussed in Section III-A; the only difference between the two is that asynchronous RSA takes more time per sample due to the deterministic time-domain offset, $T_{DCO,MIN}$. On the other hand, this constant time-offset per sample makes asynchronous RSA more practical for the circuit implementation.

One concern here is about the implementation of Mod[$\Delta T_{PRBS,MAX}$, T] = 0, which may increase the cost of asynchronous RSA. To resolve this issue, the case of $\Delta T_{PRBS,MAX} > T$ shown in the right column of Fig. 5(b) could be considered though the result of (19) indicates that the non-uniform conditional joint PDFs induce non-zero covariances. These non-uniform conditional joint PDFs are mainly caused by the non-uniform $f_1(t)$ as shown at the bottom-right of Fig. 5(b): the hollow and bump of $f_{n,n+1}(t)|Y_n = 1$ and $f_{n,n+1}(t)|Y_n = 0$ are due to the non-uniform $f_1(t)$ circularly convoluting with the uniform conditional PDFs, $f_n(t)|Y_n = 1$ and $f_n(t)|Y_n = 0$, respectively. Also, the locations of the hollow and bump are functions of Mod[$T_{DCO,MIN}$, T]. As mentioned, the flatness of $f_1(t)$ can actually be achieved by having $\Delta T_{PRBS,MAX} \gg T$, and then all of the conditional joint PDFs can become approximately uniform. That is, as long as $\Delta T_{PRBS,MAX} \gg T$, asynchronous RSA with I.I.D. sampling PDFs can be easily implemented without considering the exact relationship between $\Delta T_{PRBS,MAX}$ and T [48], [49]. However, the downside is having an even lower conversion-rate because $T_{DCO,MIN}$ inevitably increases with $\Delta T_{PRBS,MAX}$ to reach the requirement for $\Delta T_{PRBS,MAX} \gg T$. This issue of compromising between the circuit cost and conversion-rate further assures the necessity of additional conversion-rate enhancement techniques for asynchronous RSA.

IV. SIMULATION RESULTS AND SUMMARY

To verify the theoretical derivations of the RSA technique, an additional parameter, N_{EXP}, has to be introduced; it is the number of $\bar{Y}$ obtained from the experiments or simulations to statistically estimate the ideal standard deviation or theoretical variance. Based on the theoretical definitions of the variances shown in (5), (7) and (11), variance is a quantity used to represent the power of error, but variance itself is also an expectation. Therefore, to experimentally approach "this expectation" (i.e., theoretical variance), it also requires a Monte Carlo estimate:

$$Var\left[\bar{Y}\right] = E\left[\left(\bar{Y} - E\left[\bar{Y}\right]\right)^2\right]$$
$$= \lim_{N_{EXP} \to \infty} \frac{1}{N_{EXP}} \cdot \sum_{m=1}^{N_{EXP}} \left(\bar{Y}_m - \frac{\tau}{T}\right)^2$$
$$\approx \frac{1}{N_{EXP} - 1}$$
$$\cdot \sum_{m=1}^{N_{EXP}} \left(\bar{Y}_m - \frac{1}{N_{EXP} - 1} \cdot \sum_{p=1}^{N_{EXP}} \bar{Y}_m\right)^2 \quad (20)$$

As shown in (20), the verification accuracy depends on the number of $\bar{Y}$, N_{EXP}, obtained from simulations. For a finite but at least moderately large N_{EXP}, the calculation in (20) can take the mean of $\bar{Y}$ in place of $E[\bar{Y}]$ along with the Bessel's correction, ($N_{EXP} - 1$), for an unbiased Var[$\bar{Y}$] simulation estimate [50], [53]. Note that, the purpose of introducing N_{EXP} and (20) is mainly for verification; in real RSA measurements, under a certain ENOB requirement along with the settings of OSR, N_{DCO}, $T_{DCO,MIN}$, $\Delta T_{PRBS,MAX}$ and T, only a single $\bar{Y}$ is required to represent one measurement result.

The theoretical variances and the corresponding simulation results are demonstrated in Fig. 6. In Fig. 6(a), the blue curves are the theoretical variances plotted as functions of τ/T from (11) with a constant $N_{DCO} = 2^{14}$ and four different OSR configurations for the synchronous RSA measurements. With higher OSR values, the variance curves show more periodic lobes along the τ/T axis and lower peak quantization noise powers. The red dots are the behavioral model simulation results of the synchronous RSA system following the same N_{DCO}, OSR, and τ/T range with 0.1 step size. The theoretical and simulation results are well aligned as shown in Fig. 6(a), and again the simulation error can be reduced by increasing N_{EXP}, which has been set as 2^{11} for each red-dot generation. In Fig. 6(b), the variances of asynchronous RSA are also plotted as functions of τ/T with $N_{DCO} = 2^{14}$ and four different $\Delta T_{PRBS,MAX}$ configurations. As proven by (19), when Mod[$\Delta T_{PRBS,MAX}$, T] approaches zero or $\Delta T_{PRBS,MAX}$ itself increases, the variance converges to the case of the I.I.D. random sampling condition, which is the blue curve shown in Fig. 6(b) and obtained from (7). In Fig. 6(c), the variances are plotted (on the dBW scale) as functions of N_{DCO}, including some cases from synchronous and asynchronous RSA measurements, where the value of τ/T for each curve is chosen to exhibit the worst-case quantization noise in Fig. 6(a) and 6(b). There are 2^8 (= N_{EXP}) grey dots for each N_{DCO} to denote the experimental variances under the I.I.D. random sampling condition for both synchronous and asynchronous RSA, so each red dot for a certain N_{DCO} approximately on the I.I.D. theoretical variance line is obtained from the average of the 2^8 grey dots as shown in Fig. 6(c). All theoretical (i.e., blue lines) and simulation (i.e., red dots) results are well aligned and all follow the weak law of large number to perform a consistent variance degradation in -3 dBW per octave of N_{DCO} or equivalently -6 dBW per octave of ENOB as predicted by (11) and (7). Also, the worst-case variances

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

WU *et al.*: RSA TECHNIQUES FOR SINGLE-PHOTON ARRIVAL-TIME DETECTIONS IN QUANTUM APPLICATIONS 11
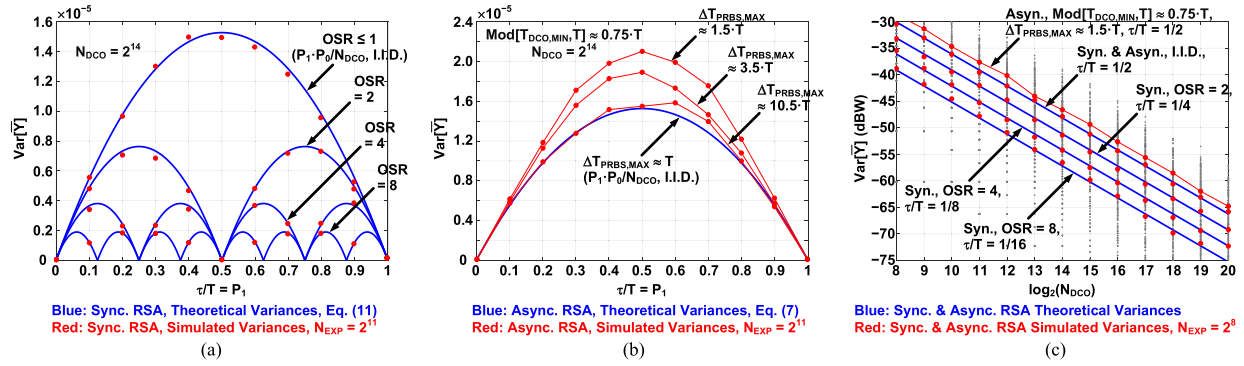

Fig. 6. (a) The theoretical and simulated variances of synchronous RSA plotted as functions of τ/T under four OSR settings with $N_{DCO} = 2^{14}$ and $N_{EXP} = 2^{11}$. (b) The theoretical and simulated variances of asynchronous RSA plotted as functions of τ/T under four $\Delta T_{PRBS,MAX}$ settings with $Mod[T_{DCO,MIN}, T] \approx 0.75 \cdot T$, $N_{DCO} = 2^{14}$, and $N_{EXP} = 2^{11}$. (c) The worst-case theoretical and simulated variances of synchronous and asynchronous RSA plotted as functions of N_{DCO} with $N_{EXP} = 2^8$.

TABLE I

TIME-INTERVAL MEASUREMENT TECHNIQUE COMPARISON & SUMMARY

Work	This Work	This Work	[55]	[56]	[57]	[3]	[21]	[4]
Sim. or Meas. Result	Sim.	Sim.	Meas.	Meas.	Meas.	Meas.	Meas.	Meas.
Application	High-Res. TCSPC	High-Res. TCSPC	High-Res. TDC	High-Res. TDC	High-Res. TDC	TOF, 3D-Ranging	TOF, LiDAR	TOF, LiDAR
Technology	22 nm	22 nm	130 nm	90 nm	40 nm	350 nm	110 nm	130 nm
Technique	Synchronous RSA	Asynchronous RSA	MASH $\Delta\Sigma$	Vernier & GRO	Flash $\Delta\Sigma$	Counter & Delay Line	Counter & Delay Line	Folded Flash
Central Clock Generator & Power	PLL & PI 25 mW	DCO 3 mW	No Info.	No Info.	No Info.	DLL & PI	DLL	PLL
ENOB	10	> 14 (Scalable)	11	< 14 *	< 8 *	14	No Info.	No Info.
Resolution	15 ps	< 1 ps (Scalable)	5.6 ps	3.2 ps	1.6 ps	10 ps	156.25 ps	71 ps
Dynamic Range	16 ns	> 16 ns (Scalable)	20 ns	40 ns	320 ps	160 ns	320 ns	18.8 ns
Sampling Frequency	4 GS/s	4 GS/s	50 MS/s	25 MS/s	50 MS/s	No Info.	100 MS/s	14 GS/s
Conversion-Rate	2.4 kHz @ 12 ENOB	2.4 kHz @ 12 ENOB	200 kHz	1.56 MHz	5 MHz	3 MHz	No Info.	14 GHz
Power	1.3 mW	1.3 mW	1.7 mW	3.6 mW	1.32 mW	< 15 mW	5 mW **	14.1 mW
Area	0.01 mm^2	0.01 mm^2	0.11 mm^2	0.03 mm^2	0.08 mm^2	0.062 mm^2	No Info.	0.03 mm^2
Inherent Calibration	RSA Mechanism	RSA Mechanism						No Need
Inherent Noise Filtering	Averaging Process	Averaging Process					Interference Suppression	
Theoretical Expectation	Eq. (10)	Eq. (15)						
Theoretical Variance	Eq. (11)	Eq. (7) for I.I.D.						
Signal & Circuit Parameters	τ, T, OSR, N_{DCO}	τ, T, $T_{DCO,MIN}$, $\Delta T_{PRBS,MAX}$, N_{DCO}						

* Estimated based on the reported effective resolutions. ** Estimated based on the reported power of 36 channels.
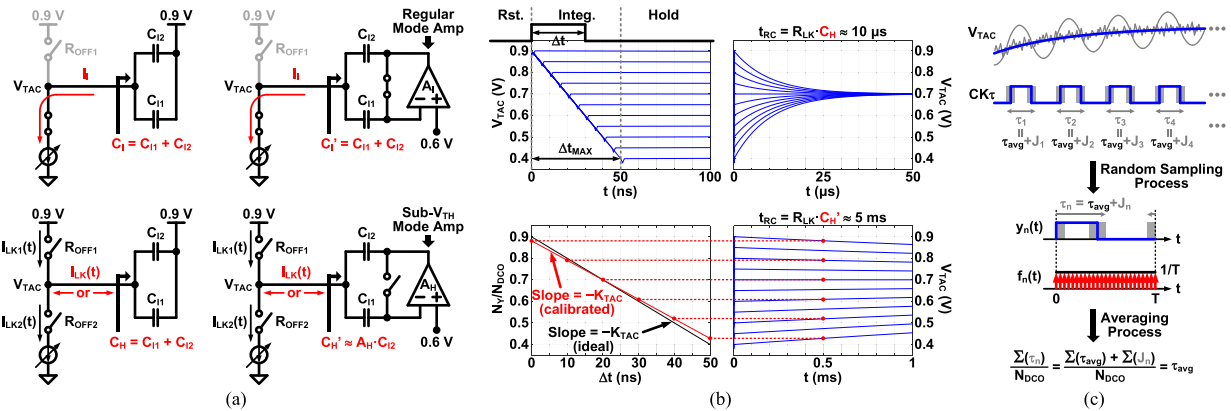

Fig. 7. (a) Left: a simple TAC circuit with push-pull leakage paths. Right: a TAC circuit with push-pull leakage paths and Miller-hold capacitance scheme. (b) Top-left: V_{TAC} waveforms under different Δt inputs in the integration- and hold-phases. Right: V_{TAC} waveforms in the hold-phase without and with the leakage time-constant enhancement by Miller-hold capacitance. Bottom-left: the analog-circuit (TAC only) conversion gain (input-to-output transfer function) calibration process by measuring the RSA outcomes with respect to injected known Δt inputs. (c) The conceptual illustration of low-pass filtering non-zero frequency noise and crosstalk in the RSA measurement.

of synchronous RSA exhibit -3 dBW per octave of OSR for a certain N_{DCO} in both Fig. 6(a) and 6(c).

The comparison between synchronous and asynchronous RSA is summarized in Table I based on the circuit simulations

along with the silicon measurement results of multiple state-of-the-art TDC implementations in both high-speed and high-resolution TCSPC applications. Although asynchronous RSA has superior power/area efficiency, scalable dynamic ranges, high noise immunity, and almost unlimited measurement accuracy, its slow conversion-rate is the primary downside limiting the broadness of RSA in quantum applications. Meanwhile, synchronous RSA can improve its conversion-rate by increasing its OSR, but this brute-force approach would quickly hit a brick wall due to the limitations of clock bandwidth and resolutions of phase-interpolators (PI) and PRBS generators. However, the analysis and simulation results demonstrated in this paper indicates that there is a possibility to generate variances lower than that of the I.I.D. random sampling condition, so the conversion-rate could be enhanced by trading in the reduced variance for a shorter measurement time, which is called the variance reduction technique.

V. Conclusion

This paper evaluates the feasibility of using the RSA technique for high-resolution and low-power TCSPC systems in emerging quantum applications. The theoretical analysis and circuit realization assessment conclude that variance reduction techniques are required to fully boost the value and benefit of the RSA technique. In addition, the theoretical findings verified and demonstrated in this paper will be the foundations of realizing practical variance reduction techniques with a minimum amount of circuit/power overhead in the forthcoming research.

Appendix

This section addresses the non-idealities in the TAC circuit and the methodology of its practical low-power implementation with the calibration process offered by the RSA technique itself. Here lists the circuit/device parameters and simulation conditions in a 22-nm digital CMOS process technology: supply voltage = 0.9 V; temperature = 100 °C; regular-V_{TH} switch OFF-resistance, R_{OFF} = 3 MΩ with 600 kΩ standard deviation; tunable TAC integration current, I_I = 10 μA; TAC integration capacitance, C_I = C_{I1} + C_{I2} = 1 pF (including the parasitic capacitance, C_P) mainly implemented by a 7-layer metal-finger capacitor with a 12 μm × 12 μm silicon area; time-interval dynamic range, Δt_{MAX} = 50 ns; TAC voltage dynamic range, V_{TAC} = 0.4 to 0.9 V; RSA measurement time = 1 ms.

First of all, a simple TAC circuit shown at the left of Fig. 7(a) is considered by assuming the top switch was on to set the initial V_{TAC} to 0.9 V during the reset-phase. In the integration-phase, I_I discharges V_{TAC} with a slew-rate of $-I_I/C_I$ = -10 V/μs during the time-interval, Δt, as the waveforms shown at the top-left of Fig. 7(b). In the hold-phase, the switches are all off, and the TAC relies on the floating hold-capacitance, $C_H = C_I$, to maintain V_{TAC} for the 1-ms long RSA measurement, which is the major performance concern of this TAC circuit. As shown at the bottom-left of Fig. 7(a), the primary leakage paths contain both pull-up and push-down currents, $I_{LK1}(t)$ and $I_{LK2}(t)$, due to the OFF-resistances of the switches, R_{OFF1} and R_{OFF2}, respectively. Though R_{OFF1} and R_{OFF2} are roughly time invariant, $I_{LK1}(t)$

and $I_{LK2}(t)$ are functions of $V_{TAC}(t)$ along the time in the hold-phase. More importantly, the equivalent leakage current, $I_{LK}(t)$, is actually the delta between the push and pull currents, $|I_{LK}(t)| = |I_{LK1}(t) - I_{LK2}(t)|$, but its direction depends on the V_{TAC} voltage at the end of the integration-phase (determined by Δt) and the V_{TAC} destination voltage when $I_{LK}(t)$ vanishes to zero, i.e., $I_{LK1}(t) = I_{LK2}(t)$ at t → ∞. This V_{TAC} destination voltage is determined by $R_{OFF2}/(R_{OFF1}+ R_{OFF2})$·0.9 V and is highly sensitive to PVT variations. The only requirement here is to roughly skew the OFF-resistance ratio so V_{TAC} can eventually settle to somewhere within the dynamic range, i.e., 0.4 to 0.9 V. As the example shown at the top-right of Fig. 7(b), all different V_{TAC} voltages from the integration-phase eventually settle to 0.7 V only based on the values of R_{OFF1}, R_{OFF2}, and the supply. Also, the DC voltage at the other terminal of C_I (= C_H) does not affect the charge leaking process as long as it is a low-impedance AC ground, which is the supply in this case (the supply noise coupling is discussed at the end of this section). The key point here is that the first-order current cancellation between the push and pull leakage paths increases total leakage resistance, R_{LK}, up to 10 MΩ. However, the equivalent RC time-constant, 10 μs, is still far from enough to maintain V_{TAC} as shown at the top-right of Fig. 7(b).

The technique of Miller-hold capacitance [58] as shown at the right of Fig. 7(a) offers a low-power solution especially for maintaining a DC voltage. In the integration-phase, C_I stays the same as 1 pF, so the power consumption of I_I can also stay the same. The extra power cost is from the unity-gain amplifier to maintain the low-impedance AC ground, which is 0.6-V DC in this example, for the -10 V/μs integration slew-rate. More importantly, in the hold-phase, the equivalent hold-capacitance now has been effectively enhanced by the DC gain of the amplifier, A_H, during the hold-phase to achieve $C_H' \approx A_H \cdot C_{I2}$ = 500 pF, and therefore the equivalent RC time-constant is improved up to 5 ms as shown at the bottom-right of Fig. 7(b). Also, since the hold-phase occupies the majority of RSA measurement time, the amplifier can be switched to a sub-threshold operation mode [59] to main a high DC gain and low-power consumption for the Miller-hold capacitance scheme.

Overall, the non-idealities of the TAC circuit at least include non-constant integration slew-rate due to device channel-length modulations, glitches and offsets due to switch charge-injections and control-feedthroughs, and finite RC time-constants due to leaky switch OFF-resistances. Fortunately, all non-idealities from the entire analog front-ends and circuits (though here only TAC is considered) can be pre-calibrated by the RSA technique itself in terms of a combinational and average effect without extra hardware. As shown in Fig. 7(b), by injecting known Δt pulses into the RSA system, the averages of the TAC non-idealities (red dots) in the bottom-right figure can be reflected by the corresponding RSA measurement results, N_Y/N_{DCO}, with respect to these known Δt inputs as shown in the bottom-left figure. After this calibration process, the new transfer curve (red curve) can be obtained and readily used for the corrections of the realistic measurements [48]. Note that the calibrated transfer function could be a high-order monotonic curve due to the

non-linearities of the analog circuits; for example, VCDL can induce a certain amount of non-linearity from converting its control voltage to its time-domain delay.

The noise immunity of the RSA system is relatively high [48] even though it uses simple analog circuitry and floating capacitance to hold the TAC output voltage, V_{TAC}, for a long measurement time as well as in a digital-circuit dominated silicon environment. As illustrated in Fig. 7(c), the reason is that the averaging process of RSA is an integrator-based low-pass filter, so it basically can filter out all sampled jitters, J_n, converted from the zero-mean high-frequency uncorrelated noise and low-frequency correlated crosstalk (e.g., supply noise) on V_{TAC}, so only the DC components from the signal under the measurement, systematic offset, and circuit non-linearity can be preserved, i.e., $\tau_{AVG} = \tau + \tau_{OFF} + \tau_{NON}$; the offset and non-linearity can be removed by the calibration process as discussed.

REFERENCES

[1] W. Becker, *Advanced Time-Correlated Single Photon Counting Techniques*. Berlin, Germany: Springer, 2005.

[2] W. Becker, *The bh TCSPC Handbook*, 7th ed. Berlin, Germany: Becker & Hickl GmbH, 2017.

[3] B. Markovic, S. Tisa, F. A. Villa, A. Tosi, and F. Zappa, "A high-linearity, 17 ps precision time-to-digital converter based on a single-stage Vernier delay loop fine interpolation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 60, no. 3, pp. 557–569, Mar. 2013.

[4] N. A. W. Dutton *et al.*, "A time-correlated single-photon-counting sensor with 14 GS/s histogramming time-to-digital converter," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2015, pp. 204–205.

[5] F. Villa *et al.*, "CMOS imager with 1024 SPADs and TDCs for single-photon timing and 3-D time-of-flight," *IEEE J. Sel. Topics Quantum Electron.*, vol. 20, no. 6, pp. 364–373, Nov. 2014.

[6] C. Niclass, C. Favi, T. Kluter, M. Gersbach, and E. Charbon, "A 128 × 128 single-photon image sensor with column-level 10-bit time-to-digital converter array," *IEEE J. Solid-State Circuits*, vol. 43, no. 12, pp. 2977–2989, Dec. 2008.

[7] D. Tamborini, M. Buttafava, A. Ruggeri, and F. Zappa, "Compact, low-power and fully reconfigurable 10 ps resolution, 160 μs range, time-resolved single-photon counting system," *IEEE Sensors J.*, vol. 16, no. 10, pp. 3827–3833, May 2016.

[8] M. Beck, "Comparing measurements of $g^{(2)}(0)$ performed with different coincidence detection techniques," *J. Opt. Soc. Amer. B, Opt. Phys.*, vol. 24, no. 12, pp. 2972–2978, Dec. 2007.

[9] M. Beck, *Quantum Mechanics—Theory and Experiment*. New York, NY, USA: Oxford Univ. Press, 2012.

[10] U. Leonhard, *Measuring the Quantum State of Light*. Cambridge, U.K.: Cambridge Univ. Press, 1997.

[11] C. Spee, J. I. de Vicente, and B. Kraus, "Remote entanglement preparation," *Phys. Rev. A, Gen. Phys.*, vol. 88, no. 1, Jul. 2013, Art. no. 010305.

[12] C. H. Bennett, "Quantum cryptography using any two nonorthogonal states," *Phys. Rev. Lett.*, vol. 68, no. 21, pp. 3121–3124, May 1992.

[13] B. A. Slutsky, R. Rao, P.-C. Sun, and Y. Fainman, "Security of quantum cryptography against individual attacks," *Phys. Rev. A, Gen. Phys.*, vol. 57, no. 4, pp. 2383–2398, Apr. 1998.

[14] M. Hillery, V. Bužek, and A. Berthiaume, "Quantum secret sharing," *Phys. Rev. A, Gen. Phys.*, vol. 59, no. 3, pp. 1829–1834, Mar. 1999.

[15] M. Unternährer, B. Bessire, L. Gasparini, D. Stoppa, and A. Stefanov, "Coincidence detection of spatially correlated photon pairs with a monolithic time-resolving detector array," *Opt. Exp.*, vol. 24, no. 25, pp. 28829–28841, Dec. 2016.

[16] L. H. C. Braga *et al.*, "A fully digital 8 × 16 SiPM array for PET applications with per-pixel TDCs and real-time energy output," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 301–314, Jan. 2014.

[17] L. Gasparini *et al.*, "A 32×32-pixel time-resolved single-photon image sensor with 44.64 μm pitch and 19.48% fill-factor with on-chip row/frame skipping features reaching 800 kHz observation rate for quantum physics applications," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2018, pp. 98–99.

[18] C. Veerappan *et al.*, "A 160×128 single-photon image sensor with on-pixel 55 ps 10 b time-to-digital converter," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2011, pp. 312–313.

[19] R. M. Field, S. Realov, and K. L. Shepard, "A 100 fps, time-correlated single-photon-counting-based fluorescence-lifetime imager in 130 nm CMOS," *IEEE J. Solid-State Circuits*, vol. 49, no. 4, pp. 867–880, Apr. 2014.

[20] M. Perenzoni, D. Perenzoni, and D. Stoppa, "A 64 × 64-pixels digital silicon photomultiplier direct TOF sensor with 100-MPhotons/s/pixel background rejection and imaging/altimeter mode with 0.14% precision up to 6 km for spacecraft navigation and landing," *IEEE J. Solid-State Circuits*, vol. 52, no. 1, pp. 151–160, Jan. 2017.

[21] H. Seo *et al.*, "A 36-channel SPAD-integrated scanning LiDAR sensor with multi-event histogramming TDC and embedded interference filter," in *Proc. IEEE Symp. VLSI Circuits*, Honolulu, HI, USA, Jun. 2020, pp. 1–2.

[22] T. E. Rahkonen and J. T. Kostamovaara, "The use of stabilized CMOS delay lines for the digitization of short time intervals," *IEEE J. Solid-State Circuits*, vol. 28, no. 8, pp. 887–894, Aug. 1993.

[23] K. Karadamoglou, N. P. Paschalidis, E. Sarris, N. Stamatopoulos, G. Kottaras, and V. Paschalidis, "An 11-bit high-resolution and adjustable-range CMOS time-to-digital converter for space science instruments," *IEEE J. Solid-State Circuits*, vol. 39, no. 1, pp. 214–222, Jan. 2004.

[24] Y. Liu *et al.*, "Multi-stage pulse shrinking time-to-digital converter for time interval measurements," in *Proc. Eur. Microw. Integr. Circuits Conf.*, Munich, Germany, 2007, pp. 267–270.

[25] C. Ljuslin, J. Christiansen, A. Marchioro, and O. Klingsheim, "An integrated 16-channel CMOS time to digital converter," *IEEE Trans. Nucl. Sci.*, vol. 41, no. 4, pp. 1104–1108, Aug. 1994.

[26] R. B. Staszewski, S. Vemulapalli, P. Vallur, J. Wallberg, and P. T. Balsara, "1.3 V 20 ps time-to-digital converter for frequency synthesis in 90-nm CMOS," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 3, pp. 220–224, Mar. 2006.

[27] Y. Arai and M. Ikeno, "A time digitizer CMOS gate-array with a 250 ps time resolution," *IEEE J. Solid-State Circuits*, vol. 31, no. 2, pp. 212–220, Feb. 1996.

[28] L. Vercesi, A. Liscidini, and R. Castello, "Two-dimensions Vernier time-to-digital converter," *IEEE J. Solid-State Circuits*, vol. 45, no. 8, pp. 1504–1512, Aug. 2010.

[29] C.-S. Hwang, P. Chen, and H.-W. Tsao, "A high-precision time-to-digital converter using a two-level conversion scheme," *IEEE Trans. Nucl. Sci.*, vol. 51, no. 4, pp. 1349–1352, Aug. 2004.

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

[31] M. Zanuso, P. Madoglio, S. Levantino, C. Samori, and A. L. Lacaita, "Time-to-digital converter for frequency synthesis based on a digital bang-bang DLL," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 3, pp. 548–555, Mar. 2010.

[32] A. Mantyniemi, T. E. Rahkonen, and J. Kostamovaara, "A CMOS time-to-digital converter (TDC) based on a cyclic time domain successive approximation interpolation method," *IEEE J. Solid-State Circuits*, vol. 44, no. 11, pp. 3067–3078, Nov. 2009.

[33] J.-P. Jansson, A. Mäntyniemi, and J. Kostamovaara, "A CMOS time-to-digital converter with better than 10 ps single-shot precision," *IEEE J. Solid-State Circuits*, vol. 41, no. 6, pp. 1286–1296, Jun. 2006.

[34] J. P. Jansson, A. Mäntyniemi, and J. Kostamovaara, "Synchronization in a multilevel CMOS time-to-digital converter," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 56, no. 8, pp. 1622–1634, Aug. 2009.

[35] E. Räisänen-Ruotsalainen, T. E. Rahkonen, and J. Kostamovaara, "An integrated time-to-digital converter with 30-ps single-shot precision," *IEEE J. Solid-State Circuits*, vol. 35, no. 10, pp. 1507–1510, Oct. 2000.

[36] B. K. Swann *et al.*, "A 100-ps time-resolution CMOS time-to-digital converter for positron emission tomography imaging applications," *IEEE J. Solid-State Circuits*, vol. 39, no. 11, pp. 1839–1852, Nov. 2004.

[37] E. Temporiti, C. Weltin-Wu, D. Baldi, M. Cusmai, and F. Svelto, "A 3.5 GHz wideband ADPLL with fractional spur suppression through TDC dithering and feedforward compensation," *IEEE J. Solid-State Circuits*, vol. 45, no. 12, pp. 2723–2736, Dec. 2010.

[38] M. A. Abas, G. Russell, and D. J. Kinniment, "Embedded high-resolution delay measurement system using time amplification," *IET Comput. Digit. Techn.*, vol. 1, no. 2, pp. 77–86, Mar. 2007.

[39] M. Lee and A. A. Abidi, "A 9 b, 1.25 ps resolution coarse–fine time-to-digital converter in 90 nm CMOS that amplifies a time residue," *IEEE J. Solid-State Circuits*, vol. 43, no. 4, pp. 769–777, Apr. 2008.

[40] B. Tong, W. Yan, and X. Zhou, "A constant-gain time-amplifier with digital self-calibration," in *Proc. IEEE 8th Int. Conf. ASIC*, Changsha, China, Oct. 2009, pp. 1133–1136.

[41] J. Kalisz, "Review of methods for time interval measurements with picosecond resolution," *Metrologia*, vol. 41, no. 1, pp. 17–32, Feb. 2004.

[42] S.-K. Shin *et al.*, "A 12 bit 200 MS/s zero-crossing-based pipelined ADC with early sub-ADC decision and output residue background calibration," *IEEE J. Solid-State Circuits*, vol. 49, no. 6, pp. 1366–1382, Jun. 2014.

[43] S.-H. W. Chiang, H. Sun, and B. Razavi, "A 10-bit 800-MHz 19-mW CMOS ADC," *IEEE J. Solid State Circuits*, vol. 49, no. 4, pp. 935–949, Aug. 2014.

[44] M. Wahl, "Time-correlated single photon counting," PicoQuant GmbH, Berlin, Germany, Tech. Rep., 2014. [Online]. Available: https://www.picoquant.com/images/uploads/page/files/7253/technote_tcspc.pdf

[45] E. Alon, V. Stojanovic, and M. A. Horowitz, "Circuits and techniques for high-resolution measurement of on-chip power supply noise," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 820–828, Apr. 2005.

[46] E. Alon, V. Abramzon, B. Nezamfar, and M. Horowitz, "On-die power supply noise measurement techniques," *IEEE Trans. Adv. Packag.*, vol. 32, no. 2, pp. 248–259, May 2009.

[47] M. Mansuri, B. Casper, and F. O'Mahony, "An on-die all-digital delay measurement circuit with 250 fs accuracy," in *Proc. Symp. VLSI Circuits (VLSIC)*, Honolulu, HI, USA, Jun. 2012, pp. 98–99.

[48] T.-C. Hsueh, F. O'Mahony, M. Mansuri, and B. Casper, "An on-die all-digital power supply noise analyzer with enhanced spectrum measurements," *IEEE J. Solid-State Circuits*, vol. 50, no. 7, pp. 1711–1721, Jul. 2015.

[49] T.-C. Hsueh *et al.*, "A 25.6 Gb/s differential and DDR4/GDDR5 dual-mode transmitter with digital clock calibration in 22 nm CMOS," in *IEEE ISSCC Dig. Tech. Papers*, San Francisco, CA, USA, Feb. 2014, pp. 444–445.

[50] P. Glasserman, *Monte Carlo Methods in Financial Engineering*. New York, NY, USA: Springer, 2003.

[51] S. Ghahramani, *Fundamentals of Probability*. Upper Saddle River, NJ, USA: Prentice-Hall, 1996.

[52] S. Haykin, *Communication Systems*, 4th ed. New York, NY, USA: Wiley, 2001.

[53] G. Upton and I. Cook, *A Dictionary of Statistics*, 3rd ed. Oxford, U.K.: Oxford, 2014.

[54] A. V. Oppenheim, R. W. Schafer, and J. R. Buck, *Discrete-Time Signal Processing*, 2nd ed. Upper Saddle River, NJ, USA: Prentice-Hall, 1999.

[55] Y. Cao, W. De Cock, M. Steyaert, and P. Leroux, "1-1-1 MASH Δ Σ time-to-digital converters with 6 ps resolution and third-order noise-shaping," *IEEE J. Solid-State Circuits*, vol. 47, no. 9, pp. 2093–2106, Sep. 2012.

[56] P. Lu, A. Liscidini, and P. Andreani, "A 3.6 mW, 90 nm CMOS gated-Vernier time-to-digital converter with an equivalent resolution of 3.2 ps," *IEEE J. Solid-State Circuits*, vol. 47, no. 7, pp. 1626–1635, Jul. 2012.

[57] Y. Wu, P. Lu, and R. B. Staszewski, "A time-domain 147 fs$_{rms}$ 2.5-MHz bandwidth two-step flash-MASH 1-1-1 time-to-digital converter with third-order noise-shaping and mismatch correction," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 67, no. 8, pp. 2532–2545, Aug. 2020.

[58] P. J. Lim and B. A. Wooley, "A high-speed sample-and-hold technique using a Miller hold capacitance," *IEEE J. Solid-State Circuits*, vol. 26, no. 4, pp. 643–651, Apr. 1991.

[59] P. E. Allen and D. R. Holberg, *CMOS Analog Circuit Design*, 3rd ed. New York, NY, USA: Oxford Univ., 2011.

Tony Wu (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electrical and computer engineering from the University of California, San Diego, CA, USA, in 2019 and 2021, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering.

In 2019, he was a Research Intern with the BIOEE Laboratory. Since 2020, he has been a Research and Teaching Assistant with the Integrated Communication Circuits Laboratory (ICCL), University of California, San Diego. His research interests include mixed-signal circuits and systems with an emphasis on high-speed and low-power digital designs.
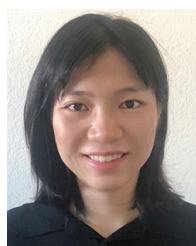
Ruoman Yang (Graduate Student Member, IEEE) received the B.S. degree in microelectronics from the University of Electronic Science and Technology of China, Chengdu, China, in 2017, and the M.S. degree in electrical and computer engineering from the University of California, San Diego, CA, USA, in 2019, where she is currently pursuing the Ph.D. degree.

Her research interests include analog/mixed-signal integrated circuits, especially electrical/optical wireline transceivers.

Tzu-Chien Hsueh (Senior Member, IEEE) received the B.S. and M.S. degrees in electrical engineering from National Taiwan University, Taiwan, in 1999 and 2001, respectively, and the Ph.D. degree in electrical and computer engineering from the University of California, Los Angeles, CA, USA, in 2010.

From 2001 to 2006, he was a Mixed-Signal Circuit Design Engineer in Hsinchu, Taiwan. From 2010 to 2018, he was a Research Scientist at Intel Lab Signaling Research and an Analog Engineer at Intel I/O Circuit Technology, Hillsboro, Oregon. Since 2018, he has been an Assistant Professor in electrical and computer engineering at the University of California, San Diego. His research interests include wireline electrical/optical transceivers, clock-and-data recovery, data-conversion circuits, on-chip performance measurements/analyzers, and digital/mixed signal processing techniques.

Dr. Hsueh was a recipient of the Multiple Intel Division and Academy Awards from 2012 to 2018, the 2015 IEEE JOURNAL OF SOLID-STATE CIRCUITS (JSSC) Best Paper Award, and the 2020 National Science Foundation (NSF) CAREER Award. He served on the Patent Committee for Intel Intellectual Property (Intel IP) and the Technical Committee for Intel Design and Test Technology Conference (DTTC) from 2016 to 2018. Since 2018, he has served on the Technical Program Committee for IEEE Custom Integrated Circuits Conference (CICC) and the Guest Associate Editor for IEEE SOLID-STATE CIRCUITS LETTERS (SSC-L).