Title

CMOS Single-Photon Avalanche Diode Circuits for Probabilistic Computing

Authors

Whitehead, William

Oh, Wonsik

Theogarajan, Luke

Publication Date

2024

DOI

10.1109/jxcdc.2024.3452030

Peer reviewed

HHS Public Access

Author manuscript

IEEE J Explor Solid State Comput Devices Circuits. Author manuscript; available in PMC 2024 November 01.

Published in final edited form as:

IEEE J Explor Solid State Comput Devices Circuits. 2024 ; 10: 49–57. doi:10.1109/jxcdc.2024.3452030.

CMOS Single-Photon Avalanche Diode Circuits for Probabilistic Computing

WILLIAM WHITEHEAD,
WONSIK OH,
LUKE THEOGARAJAN

Department of Electrical and Computer Engineering, UCSB, Santa Barbara, CA 93106 USA

Abstract

Intrinsically random hardware devices are increasingly attracting attention for their potential use in probabilistic computing architectures. One such device is the single-photon avalanche diode (SPAD) and an associated functional unit, the variable-rate SPAD circuit (VRSC), recently proposed by us as a source of randomness for sampling and annealing Ising and Potts models. This work develops a more advanced understanding of these VRSCs by introducing several VRSC design options and studying their tradeoffs as implemented in a 65-nm CMOS process. Each VRSC is composed of a SPAD and a processing circuit. Combinations of three different SPAD designs and three different types of processing circuits were evaluated on several metrics such as area, speed, and variability. Measured results from the SPAD design space show that even extremely small SPADs are suitable for probabilistic computing purposes, and that high dark count rates are not detrimental either, so SPADs for probabilistic computing are actually easier to integrate in standard CMOS processes. Results from the circuit design space show that the time-to-analog-based designs introduced in this work can produce highly exponential and analytical transfer functions, but that the less analytically tractable output of the previously proposed filter-based designs can achieve less variability in a smaller footprint. Probabilistic bits (P-bits) composed of the fabricated VRSCs achieve bit flip rates of 50 MHz and allow at least one order of magnitude of control over their simulated annealing temperature.

INDEX TERMS

Ising; Optimization; Potts; probabilistic; probabilistic bit (P-bit); single-photon avalanche diode (SPAD)

I. INTRODUCTION

Recent interest in probabilistic computing has been driven by the confluence of new intrinsically stochastic hardware devices and a growing understanding that natural neural processing—long held as an aspirational benchmark in terms of performance and energy

CORRESPONDING AUTHOR: W. WHITEHEAD (williamwhitehead@ucsb.edu).

efficiency—might not just be noisy, but fundamentally stochastic [2], [3], [4]. Various bistable devices and circuits have been proposed as probabilistic bits (P-bits) [5], [6], especially for the calculation of Ising models. However, other forms of randomness are also possible.

One alternate form of randomness is the nonhomogeneous Poisson process (NHPP). An NHPP circuit produces a Poisson distribution of random events (rather than random states, as a P-bit does) with a time-varying Poisson rate λ set by some external control signal. NHPPs can be the driving force in stochastic transition circuits [7], [8], which may be used to emulate Ising and Potts models [1] and more. In a proposal for probabilistic computing using resonant energy transfer (RET) dynamics [9], the RET chromophores function as an NHPP.

A more practical NHPP implementation based on single-photon avalanche diodes (SPADs) was recently developed by us and applied to combinatorial optimization [1]. Discrete SPADs were coupled with additional signal conditioning circuits to form variable-rate SPAD circuits (VRSCs), which were applied to graph coloring and traveling salesman optimization problems. Although the VRSCs handily solved these optimization problems [1], the workings of the VRSC units needed further development, especially for integrated CMOS implementations.

Each VRSC is a combination of an SPAD and a processing circuit; the design of both these components is developed in this article. Using avalanche multiplication, the SPAD amplifies single photon or thermally generated carriers into large voltage pulses, which naturally occur with fixed-frequency Poisson statistics under constant illumination. Although SPADs are frequently integrated in CMOS processes [10], [11], most CMOS process design kits lack any SPAD characterization or models, thus requiring experimental evaluation. The processing circuit shapes this raw stream of photon-triggered pulses from the SPAD into a random analog signal. The nature of the random analog signal allows the creation of an NHPP using a threshold. The threshold is the control signal and the positive or negative threshold crossings are Poisson events. In practice, this implies the analog signal contains a series of pulses of variable amplitude, Fig. 1(c).

In this work, nine VRSC designs, each a combination of one of three SPADs and one of three processing circuits, were fabricated in a 65-nm CMOS process and evaluated for their suitability in P-bit and Potts hardware accelerators. The breadth of tested SPAD designs validates the idea that small, low-quality SPADs are just as suitable for VRSC design as larger, more ideal SPADs. The circuit designs include one implementation of the filtering concept proposed in [1], and two versions operating on a new principle which is less reliant on the qualities of the SPAD itself. At the core of the new VRSC designs is time-to-amplitude integration following an exponential distribution of peak voltages, since photon arrival intervals follow an exponential distribution.

II. BACKGROUND: λ CURVES AND PROBABILISTIC COMPUTING

To understand what features are important in a VRSC (generically, an NHPP), the defining NHPP characteristic, the λ curve, must be related to the end application, probabilistic computing. The λ curve relates the instantaneous Poisson rate to an input which would be a voltage or current threshold in the physical domain, a cost C in the conceptual domain of optimization problems, or energy E from the perspective of statistical mechanics. In a probabilistic computer, the cost C is a continuously changing value capturing a component of the cost or loss of some objective function, such as an optimization problem. This cost is found across all probabilistic computing systems, usually as a simple sum of products [5], [6]. In a full system, P-bit or Potts units are connected in a recurrent topology, affecting and being affected by these cost terms to produce a useful computation, Fig. 1(b).

A P-bit or Potts unit is created by tying several NHPPs together in a latching circuit. Since the events from an NHPP are coupled together into static states using digital logic, the transformation from Poisson events to P-bits or Potts units is deterministic. Once λ curves are known, state probabilities can be derived without building the full circuit. In our earlier work [1], we show that P-bit and Potts circuits can easily be constructed to sample exactly from a Boltzmann distribution, as long as the $\lambda(C)$ curves are exponential. For a P-bit, the probability function is

$$\Pr(m_i = +1) = \frac{\lambda(C)}{\lambda(C) + \lambda(-C)} \approx \frac{e^{-C/T}}{e^{-C/T} + e^{C/T}} \tag{1}$$

where we have adopted the standard notation of referring to the P-bit output as the magnetization m_i which takes on values of +1 or −1.

In addition to log-linearity, two more desirable properties of $\lambda(C)$ curves are listed in Table 1. One is that the curve should not just be log-linear, but should have an adjustable log-linear relationship. Adjusting the slope of $\lambda(C)$ is equivalent to adjusting the temperature T of the Boltzmann distribution, giving a control knob for affecting simulated annealing. In VRSCs, various control biases do allow the slope to be adjusted over a wide range.

Second, the flips-per-second (FPS), of the circuit can also be readily predicted. Each time the P-bit is in state $m_i = +1$ and the NHPP for state $m_i = -1$ pulses, the P-bit state flips, and vice versa

$$\mathrm{FPS}(C) = 2\lambda(-C)\Pr(+1) = \frac{2\lambda(C)\lambda(-C)}{\lambda(C) + \lambda(-C)}. \tag{2}$$

The FPS actually depends on where along the tanh probability curve the P-bit is operating, and is highest at the midpoint where $C = 0$ and $\Pr(+1) = 0.5$. Thus, the value of $\lambda(0)$ (in the cost domain) should be maximized.

The Potts circuit can comprise an arbitrary number of NHPPs driven by an arbitrary set of costs C_i, each one attracting the Potts node to its ith state. The probability of each state can be determined in the same manner as the probability of the P-bit state

$$\Pr(\text{state}_i) = \frac{\lambda(C_i)}{\sum_{j=1}^{N} \lambda(C_j)} \approx \frac{e^{-C_i/T}}{\sum_{j=1}^{N} e^{-C_j/T}}. \tag{3}$$

Again, yielding a Boltzmann distribution when $\lambda(C)$ is approximately exponential. However, in this case the Boltzmann distribution is no longer easily characterized as a tanh probability, instead taking on the form of a SoftMax. Due to the increased flexibility of this circuit and its distribution, the FPS of the system is not as well-defined as in the P-bit case. It may be very slow if all the costs C_i are high, or very fast if all the costs are low.

III. PROCESSING CIRCUIT ALTERNATIVES

Two alternative operational principles are tested in this work. Both start with the assumption that each SPAD is a fixed-rate Poisson process, generating voltage/current avalanche pulses of uniform magnitude. Both the methods leverage the variable interval between pulses to generate randomly fluctuating analog signals, which are then compared with a reference to yield an actual detected pulse rate (which will be equal to or less than the SPAD's intrinsic pulse rate). The first VRSC principle studied here uses a time-to-amplitude converter (TAC) circuit at its core, while the second principle relies on a GmC filter. Other operational principles are also plausible: a VRSC could instead take advantage of intrinsic variation in the magnitude of avalanche pulses [12], [13].

A. TIME-TO-AMPLITUDE THEORY

The principle of the time-to-amplitude design is readily analyzed. In the circuits in Fig. 2(a) and 2(b), the time t between photon detections (which follows an exponential distribution) is converted into an analog voltage by integrating a fixed current on a capacitor, and the integrated voltage is compared to an adjustable threshold voltage V_{th}. All the interpulse intervals long enough to integrate above the threshold yield a detected pulse; those that do not are rejected. With the current I integrating onto the capacitance C for a time Δt (the time between two consecutive SPAD avalanches), the condition for detection is

$$\Delta t > (V_{\text{th}} - V_b)\frac{C}{I}. \tag{4}$$

With an additional voltage V_b added for the practical purpose of shifting this comparison around to lie within the range of an integrated circuit's supply rails. Given that the probability $P(\Delta t > t_{\text{th}}) = \int_{t_{\text{th}}}^{\infty} \lambda_s e^{-t\lambda_s} dt$, the transfer function of a time-to-amplitude-based circuit is

$$\lambda(V_{\text{th}}) = \lambda_s e^{-\lambda_S (V_{\text{th}} - V_b) C / I} \tag{5}$$

which is exactly an exponential as desired. The variable λ_s has been introduced to indicate the intrinsic rate of the SPAD, which sets the upper limit of λ when $V_{\text{th}} = V_b$.

The parameters V_b, I, and λ_s can all be concurrently adjusted in a circuit to alter the properties of the exponential. Measured data in Fig. 3(a) and (f) show how increasing the integration current lessens the exponential slope of $\lambda(V)$, providing a primary control for the ‘‘temperature’’ of a probabilistic computation. Increasing λ_s not only increases the maximum value of a λ curve but also makes it more steep since the time between avalanches decreases. This can be achieved physically by either increasing illumination [Fig. 3(c) and (h)] or increasing the SPAD’s reset current, which increases the DCR [Fig. 3(b) and (g)]. However Fig. 3 also shows real behaviors that are not described by (5), such as the prevalence of slight log-nonlinearities, the differences between the MOSCAP-TAC in Fig. 3(a) and the Diode-TAC in Fig. 3(f) and the importance of other physical conditions such as SPAD bias [Fig. 3(i)].

B. MOSCAP TAC

The MOS capacitor (MOSCAP)-TAC variety in Fig. 2(a) integrates a constant current set by $\text{Pb}_{\text{integrate}}$ onto a MOSCAP. The MOSCAP is reset to the baseline voltage V_{baseline} whenever the SPAD avalanches, Fig. 2(d), and is otherwise actively integrating the bias current. The circuit is entirely asynchronous and lacks the synchronization signals commonly found in TAC-based time-correlated single-photon sensor arrays [14], [15]. Out of the three tested circuit designs, the MOSCAP-TAC design produces λ curves with the highest log-linearity, Fig. 3(a) and (b), since the SPAD avalanches are highly decoupled from the time-to-amplitude component of the circuit.

The SPAD reset current controlled by the bias Pb_{reset} is intended to recharge the SPAD much faster than the integration current fills the MOSCAP. However, the speed of the SPAD reset cannot be increased too much, or the MOSCAP might not discharge correctly, forming the reverse-looking λ curve in Fig. 3(b). Incomplete discharge after each avalanche can also be seen in the increasingly rounded λ curve around the baseline voltage in Fig. 3(a) as the integration current competes with the discharge current. The second time-to-amplitude design, the Diode-TAC, does not suffer from these partial discharge problems but faces other issues instead.

C. DIODE TAC

The diode TAC circuit in Fig. 2(b) uses the SPAD itself as the integration capacitor. When the SPAD triggers in this design, the integration node automatically resets by discharging through the SPAD. While this guarantees that the integration node discharges on every SPAD avalanche, the disadvantage of this method is that the TAC circuit is now linked to

the reverse bias of the SPAD. Thus, the avalanche probability changes over the course of the integration period, affecting the overall statistics.

To mitigate this, the diode TAC circuit was designed with a precharge stage, resulting in the voltage profile in Fig. 2(e). At each avalanche, the SPAD discharges to a possibly below-rail voltage determined by the SPAD's breakdown voltage and negative applied bias, $V_{SPAD} + V_{br}$. Immediately after the avalanche, the SPAD is rapidly precharged back up to a baseline bias that serves as the starting point for normal integration. Although this precharge current plays a similar role as the reset current in the MOSCAP TAC, Fig. 3(g) shows how far greater precharge currents can be used to increase the overall Poisson rate without encountering the partial discharge issues of the MOSCAP TAC design.

By starting regular integration significantly above the breakdown voltage, there is less of a ''dead period'' during which integration occurs but the SPAD is unlikely to trigger, regularizing the SPAD's behavior over the normal integration range. In addition, the deep discharges result in avalanches continuing to be counted even as the threshold voltage is set below the baseline voltage, rather than going entirely undetected as in the MOSCAP design. We have found that the knee-shaped λ curves this produces [Fig. 3(c) and (f)–(i)] are in fact incredibly valuable for forming high-speed P-bits.

D. GmC-FILTERED CIRCUIT

In the GmC-filtered circuit of Fig. 2(c), the SPAD reset current is supplied by a diode-connected PMOS transistor, which also serves to mirror the current of each avalanche into a GmC-filter stage. An externally biased reset shunt current controlled by Pb_{reset} is also included to provide an extra tuning knob. Each time the SPAD avalanches and sends a pulse of current into the filter, the filter's peak discharge current will depend on its prior state of charge, again converting the Poisson timing of the SPAD into a random amplitude [Fig. 2(f)], with higher amplitudes expected to be exponentially less frequent. The random amplitude is taken to be the current leaving the filter, rather than the voltage in the filter, since it was found to have better distributions in simulation; the current is directly thresholded by a current comparator composed of a push–pull stage.

However, in this case, analysis is not easy since the filter the circuit relies on is nonlinear. Under certain conditions, the circuit does not produce exponential λ curves, but does for particular bias conditions, Fig. 3(d), (e), and (j), in which case we can use it as an NHPP in P-bit and Potts circuits. Without a guiding theory, we cannot say what the best adjustment mechanism for this circuit is, but we found that the SPAD bias was particularly useful. This is in sharp contrast to the effect of bias in the diode-TAC circuit, Fig. 3(i), where the maximum bias always yields the best results; lowering the bias on the diode TAC circuit only shifts the λ curve and reduces λ_{max}.

The reset shunt current gave mixed results. Although the initial expectation was that the shunt would reduce the amount of current passed through the current mirror at each avalanche, the actual effect was the opposite: more current went to the GmC filter as the reset shunt current was increased, Fig. 3(e) and (j). This was more pronounced for the smaller SPAD, Fig. 3(e), than for the medium SPAD, Fig. 3(j). The likely cause is that the

SPAD avalanches take longer to quench when a greater current supply is present, drawing current from both the shunt and the mirror for a longer period of time.

Although this circuit is less analytically tractable and is not as reliably exponential as the time-to-amplitude method, it does have a few benefits. The first is its small size, coming in at only 16 μm^2 in our implementation. Since it operates directly in the current domain, it is also easier to integrate into systems where weighted sums are calculated in the current domain, since no transimpedance amplifier would be necessary. A third benefit is that it requires fewer biases than the TAC circuits, which reduces variability introduced by the distribution of bias signals.

IV. SPAD ALTERNATIVES

The experimental SPAD options differ by junction type and area. Two different junction types were used, a P+ n-well (P+ NW) and a p-well deep-n-well (PW DNW), with cross sections shown in Fig. 4. Two P+ NW SPADs with small (sm.) and medium (med.) sizes and one PW DNW SPAD with large (lg.) size were selected for focused study. Although the three SPAD options have widely differing properties such as dark count rate (DCR), area, and junction capacitance, the slope, offset, and peak rates of the λ curves are ultimately adjustable over similar ranges. However, the differing properties means that the bias conditions required to achieve this behavioral uniformity are themselves highly varied, so the choice of SPAD in future applications may be driven less by what the SPAD can do than by what conditions are needed to drive the SPAD.

The two SPAD junctions represent different philosophies of SPAD design. The PW DNW junction has fewer defects and lower DCR, a measure of how frequently a SPAD avalanches even in the absence of illumination. The PW DNW exemplifies the type of SPAD used for imaging applications [10], where the goal is to detect as many photons as possible, with as little noise as possible. The P+ NW SPADs, on the other hand, are designed without paying any attention to minimizing DCR or maximizing photon detection probability (PDP); they lack guard rings or any other features commonly used in CMOS imaging SPAD design [16]. Fortunately, DCR and PDP are not critical to the construction of VRSCs: the thermally initiated and photon initiated avalanches are equally Poisson in nature. The benefit of using P+ NW SPADs is that they can be more compact at small sizes since they do not need any special well implants and operate at lower voltages since their p-n junctions have higher doping; see Table 2 for a comparison of SPAD types. The characteristics of the two junctions are similar to the results for 65 nm published elsewhere [11].

The junction choice does affect how much control illumination intensity has on the VRSC λ curves though. Because the Poisson rate in a P+ NW junction is dominated by its dark count, increasing illumination has only a small effect on the peak rate λ_{max} and the slope $d\lambda/dV$, seen in Fig. 3(h). The effect of illumination on a PW DNW SPAD in Fig. 3(c) has a more canonical response: λ_{max} clearly increases with illumination over a wide range. The magnitude of $d\lambda/dV$ increases as well, since less integration time between photons translates to a lower distribution of voltage amplitude peaks. Nonetheless, λ_{max} is similar for the two SPADs, on the order of 100 MHz, and other control mechanisms are available. The ability

to achieve high Poisson rates without illumination may even be in the favor of the P+ NW junction when designing a system without an illumination component is desirable.

SPAD size also affects VRSC properties through the mechanisms of photosensitivity and junction capacitance. Photosensitivity is a minor concern in principle since illumination can be scaled to compensate, but in practice we found that the illumination required could be quite high, especially for small PW DNW SPADs with low intrinsic DCR—one fabricated small PW DNW SPAD was excluded from further study due to the inability to provide enough illumination to reach 10 s of MHz Poisson rates. Similarly, SPAD capacitance differences require scaling reset and integration currents, the practical difficulties of which led to further exclusion of two P+ NW SPADs with larger areas. As the total charge of each avalanche changes with SPAD size as well, the behavior of the GmC-filtered circuit was also found to be highly dependent on SPAD size, even between SPADs of the same junction type, such as the response to the reset shunt current, Fig. 3(e) and (j).

V. RESULTS

The VRSC designs are ultimately evaluated on their suitability for P-bit and Potts nodes. For each of these two applications, an optimal operating condition is selected for each VRSC design (Fig. 5) and the designs are compared on speed and variability metrics shown in Fig. 6. Many data points have been identified as outliers; we believe these outliers likely reflect the handling of the individual die samples, rather than true tails to the distribution.

A. λCURVE OPTIMIZATION

Since each circuit can produce a wide variety of $\lambda(C)$ curves as their bias conditions are varied, an optimization routine was used to select the combination of bias conditions (illumination, SPAD bias, reset current, etc.) that would best meet the needs of P-bit and Potts applications. In addition to biasing conditions, additional degrees of freedom are introduced by allowing only a subset of each physical λ curve to be used. This windowing corresponds to the flexibility of global weight scaling and per-circuit biasing that should be available in probabilistic computing implementations [1]. For our purposes, we assume that the window width must be shared between all the devices (global scaling of weight magnitude) but that the center of each window can be tailored to each circuit (since each circuit will have a unique bias weight).

Two separate parameter selections were performed, one for P-bit operation and one for Potts operation. The constraints on the P-bit λ curves are relaxed to not require $\lambda(C) \propto \exp(C/T)$, instead optimizing for the tanh probability curve [Fig. 5(b)] and FPS [Fig. 5(d)] metrics which are readily derived from λ curves. Using this somewhat relaxed optimization target, λ curves with a bend around $C = 0$ [Fig. 5(a)] can be chosen so that that the FPS remains high [Fig. 5(d)] while maintaining a tanh-shaped probability function [Fig. 5(b)]. In contrast, since Potts nodes operate in a more complex probability space, Potts λ curves are selected for $\lambda(C) \propto \exp(C/T)$ to stick closer to the theoretical ideal, Fig. 5(c). For comparison to other proposed P-bit technologies, the P-bit optimized curves should be used, while both the

Potts-optimized and P-bit optimized curves are used to compare the different VRSC designs to each other.

B. PERFORMANCE

The performance of the optimized λ curve selections is presented in Fig. 6. The P-bit performance of the VRSCs is evaluated using the FPS metric, with Fig. 5(d) visualizing the distribution of device FPS ($n = 24$) for a single VRSC design while Fig. 6 summarizes the distribution of peak FPS for all the VRSC designs. Once optimized, all the VRSC designs were able to produce near-ideal tanh probability curves, Fig. 5(b) for example. The Potts performance of the VRSCs is evaluated based on the mean squared error (MSE) of each selected λ curve with respect to a target exponential, calculated in the log-linear domain. Thus, the Potts metric distributions in Fig. 6 summarize the variability and log-linearity displayed in Fig. 5(c). For the Potts metric, low error is desirable, while for the P-bit metric, high FPS indicates higher performance.

On the P-bit FPS metric, the MOSCAP-TAC circuit comes in with the worst FPS, since it cannot produce the knee-shaped λ curves critical to maintaining high FPS at lower temperatures. This metric also exposes that within the GmC-filtered category, performance depends heavily on which SPAD was chosen, with FPS equal to that of the Diode-TAC clearly being possible, but not guaranteed. Overall, the diode-TAC and filtered processing circuits yielded P-bit flip rates between 10 and 50 MHz. On the Potts MSE metric, the highly accurate exponential λ curves of the MOSCAP-TAC circuits finally outperform the less ideal curves of the diode-TAC circuits, although only by a small margin. However, much to our surprise, the optimized selections of the GmC-filtered λ curves also perform well, and sometimes better than the diode-TAC circuits, even though their λ curves are not analytically exponential.

C. VARIABILITY

The impact of device-to-device variations depends on how the VRSCs are being used, including whether they are part of a P-bit circuit or a Potts-model circuit. For the Potts model, it is important that the λ curves match each other as closely as possible at all points: if two VRSCs correspond to equal improvements in a combinatorial problem (say, reducing the overall cost by 1), then the two VRSCs should produce pulses at the same rate so that the two $\Delta E = 1$ options have the same probability.

In contrast, since P-bits can be made from single VRSCs in a feedback configuration [1], P-bits are extremely robust to VRSC device-to-device variation, in contrast to the sensitivity of MTJs [17]. No matter what the shape of the λ curve is, self-referential differential operation ensures that the midpoint of the probability curve is always 0.5 [Fig. 5(b)]. Variability in the probability curves still exists, but only affects the computational temperature (slope) of each probability curve. The FPS of each P-bit will also vary [Figs. 5(d) and 6], but this will only have a small effect on how fast the overall simulated annealing circuit works—it will not skew the probability distribution away from the true optimal state. Thus, for the purpose of evaluating variability, we use the Potts MSE metric from Fig. 6.

The VRSC circuits with the least variability are the GmC-filtered circuits. Although the overall MSE is no better than the Moscap-TAC MSE, the distributions in Fig. 6 are tighter—indicating that while the GmC circuits may not be as perfectly exponential, they are less variable. A key reason for this is that the GmC circuits do not require as many external biases as the TAC-based designs, removing many sources of variability. SPAD size is also a clear driver of variability on the Potts metric, with the small P+ NW SPADs leading to much greater variability than the medium-sized P+ NW SPADs across all three processing circuit designs.

D. COMPARISON TO P-BIT ALTERNATIVES

Although other device-based P-bit technologies have been projected to run faster and in even smaller chip areas than our VRSC-based circuits, to our knowledge none have reported such high performance as a completed, functional unit. Table 3 summarizes the performance of VRSC P-bits compared with alternatives.

For example, a report of 500-MHz magnetic tunnel junctions [18] did not demonstrate an ability to bias the P-bit distribution, and a study that did report biasing effects only reached 10 MHz [19]. Neither MTJ characterization implemented a complete P-bit; a study that did only reached an FPS of 1 kHz [21]. Complete P-bit implementations based on other bistable technologies have similarly reported very low flip rates [22], [23]. Similarly, most P-bit implementations do not report any intrinsic support for simulated thermal annealing (T_{max}/T_{min}), with the exception of [22].

One of the main draws of stochastic nanomagnets and other bistable electronic devices is their very small size. However, their reported areas (such as 0.16 μm^2 [20]), do not include the area of support circuitry or the effects of variability—using larger devices to reduce variability is a key aspect of analog design, making it difficult to know whether the currently reported areas can be maintained as MTJ and memristor P-bits mature into more complete systems.

Reports on energy consumption are similarly hard to pin down, with much lower consumption achieved in some experiments [18] but not others [19], [22], [23].

On the other hand, CMOS digital annealers and Ising solvers, which are more mature than all device-based P-bits (including ours), routinely report higher realized FPS. They also report lower power consumption. Their main drawback that VRSC P-bits address is their inability to accurately sample from Boltzmann distributions without consuming a very large logic area [25].

VI. CONCLUSION

Although a new time-to-amplitude-based circuit design was proposed and demonstrated, the previously suggested GmC-filtering based VRSC design may be a more promising avenue for further research and development. The time-to-amplitude circuits largely behaved as expected, but the GmC-filtered circuit outperformed on several critical dimensions, most notably variability. It is also the smallest of the tested circuits, at only 25 μm^2 in 65-nm

CMOS and can be paired with a current-based analog multiply–add engine without needing any additional functional blocks. However, care must be taken while porting this design to a different process since we have not yet fully elucidated how this circuit works exactly the way it does. Future experimental and modeling research can address this shortcoming and further improve the performance of the GmC-filtered VRSC.

REFERENCES

[1]. Whitehead W, Nelson Z, Camsari KY, and Theogarajan L, ‘‘CMOS-compatible Ising and Potts annealing using single-photon avalanche diodes,’’ Nature Electron., vol. 6, no. 12, pp. 1009–1019, Nov. 2023, doi: 10.1038/s41928-023-01065-0.

[2]. Maass W, ‘‘Noise as a resource for computation and learning in networks of spiking neurons,’’ Proc. IEEE, vol. 102, no. 5, pp. 860–880, May 2014.

[3]. Leng L et al. , ‘‘Spiking neurons with short-term synaptic plasticity form superior generative networks,’’ Sci. Rep, vol. 8, no. 1, Jul. 2018, Art. no. 10651, doi: 10.1038/s41598-018-28999-2. [PubMed: 30006554]

[4]. Zheng Y, Jia S, Yu Z, Huang T, Liu JK, and Tian Y, ‘‘Probabilistic inference of binary Markov random fields in spiking neural networks through mean-field approximation,’’ Neural Netw, vol. 126, pp. 42–51, Jun. 2020, doi: 10.1016/j.neunet.2020.03.003. [PubMed: 32197212]

[5]. Chowdhury S et al. , ‘‘A full-stack view of probabilistic computing with *p*-bits: Devices, architectures and algorithms,’’ IEEE J. Explor. Solid-State Comput. Devices Circuits, vol. 9, no. 1, pp. 1–11, Jun. 2023, doi: 10.1109/JXCDC.2023.3256981.

[6]. Mohseni N, Mcmahon PL, and Byrnes T, ‘‘Ising machines as hardware solvers of combinatorial optimization problems,’’ Nature Rev. Phys, vol. 4, no. 6, pp. 363–379, May 2022, doi: 10.1038/s42254-022-00440-8.

[7]. Jonas E, ‘‘Stochastic architectures for probabilistic computation,’’ Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, 2014.

[8]. Mansinghka V and Jonas E, ‘‘Building fast Bayesian computing machines out of intentionally stochastic, digital parts,’’ 2014, arXiv:1402.4914.

[9]. Wang S, Lebeck AR, and Dwyer C, ‘‘Nanoscale resonance energy transfer-based devices for probabilistic computing,’’ IEEE Micro, vol. 35, no. 5, pp. 72–84, Sep. 2015.

[10]. Bruschini C, Homulle H, Antolovic IM, Burri S, and Charbon E, ‘‘Single-photon avalanche diode imagers in biophotonics: Review and outlook,’’ Light Sci. Appl, vol. 8, no. 1, p. 87, Sep. 2019. [PubMed: 31645931]

[11]. Jiang W, Chalich Y, Scott R, and Deen MJ, ‘‘Time-gated and multijunction SPADs in standard 65 nm CMOS technology,’’ IEEE Sensors J, vol. 21, no. 10, pp. 12092–12103, May 2021.

[12]. Hayat MM, Ramirez DA, Rees GJ, and Itzler MA, ‘‘Modeling negative feedback in single-photon avalanche diodes,’’ Proc. SPIE, vol. 7681, p. 76810, Apr. 2010, Art. no. 76810W.

[13]. Giustolisi G, Mita R, and Palumbo G, ‘‘Verilog-A modeling of SPAD statistical phenomena,’’ in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), May 2011, pp. 773–776.

[14]. Luca P et al., ‘‘A 256 × 256 spad array with in-pixel time to amplitude conversion for fluorescence lifetime imaging microscopy,’’ in Proc. 2015 Int. Image Sensor Workshop (IISW), 2015. [Online]. Available: https://imagesensors.org/papers/10.60928/rtc5-akao/

[15]. Acconcia G, Crotti M, Antonioli S, Rech I, and Ghioni M, ‘‘High performance time-to-amplitude converter array,’’ in Proc. IEEE Nordic-Mediterranean Workshop Time-to-Digital Converters (NoMe TDC), Oct. 2013, pp. 1–5.

[16]. Palubiak DP and Deen MJ, ‘‘CMOS SPADs: Design issues and research challenges for detectors, circuits, and arrays,’’ IEEE J. Sel. Topics Quantum Electron, vol. 20, no. 6, pp. 409–426, Nov. 2014.

[17]. Abeed MA and Bandyopadhyay S, ‘‘Low energy barrier nanomagnet design for binary stochastic neurons: Design challenges for real nanomagnets with fabrication defects,’’ IEEE Magn. Lett, vol. 10, pp. 1–5, 2019.

[18]. Safranski C, Kaiser J, Trouilloud P, Hashemi P, Hu G, and Sun JZ, ‘‘Demonstration of nanosecond operation in stochastic magnetic tunnel junctions,’’ Nano Lett, vol. 21, no. 5, pp. 2040–2045, Mar. 2021. [PubMed: 33630604]

[19]. Zink BR and Wang J-P, ‘‘Influence of intrinsic thermal stability on switching rate and tunability of dual-biased magnetic tunnel junctions for probabilistic bits,’’ IEEE Magn. Lett, vol. 12, pp. 1–5, 2021.

[20]. Debashis P, Faria R, Camsari KY, and Chen Z, ‘‘Design of stochastic nanomagnets for probabilistic spin logic,’’ IEEE Magn. Lett, vol. 9, pp. 1–5, 2018.

[21]. Borders WA, Pervaiz AZ, Fukami S, Camsari KY, Ohno H, and Datta S, ‘‘Integer factorization using stochastic magnetic tunnel junctions,’’ Nature, vol. 573, no. 7774, pp. 390–393, Sep. 2019, doi: 10.1038/s41586-019-1557-9. [PubMed: 31534247]

[22]. Yan X et al. , ‘‘Reconfigurable stochastic neurons based on tin oxide/MoS_2 hetero-memristors for simulated annealing and the Boltzmann machine,’’ Nature Commun, vol. 12, no. 1, Sep. 2021, Art. no. 5710, doi: 10.1038/s41467-021-26012-5. [PubMed: 34588444]

[23]. Woo KS, Kim J, Han J, Kim W, Jang YH, and Hwang CS, ‘‘Probabilistic computing using $Cu_{0.1}Te_{0.9}/HfO_2$/Pt diffusive memristors,’’ Nature Commun, vol. 13, no. 1, Sep. 2022, Art. no. 5762, doi: 10.1038/s41467-022-33455-x. [PubMed: 36180426]

[24]. Pervaiz AZ, Sutton BM, Ghantasala LA, and Camsari KY, ‘‘Weighted *p*-bits for FPGA implementation of probabilistic circuits,’’ IEEE Trans. Neural Netw. Learn. Syst, vol. 30, no. 6, pp. 1920–1926, Jun. 2019. [PubMed: 30387748]

[25]. Smithson SC, Onizawa N, Meyer BH, Gross WJ, and Hanyu T, ‘‘Efficient CMOS invertible logic using stochastic computing,’’ IEEE Trans. Circuits Syst. I, Reg. Papers, vol. 66, no. 6, pp. 2263–2274, Jun. 2019.

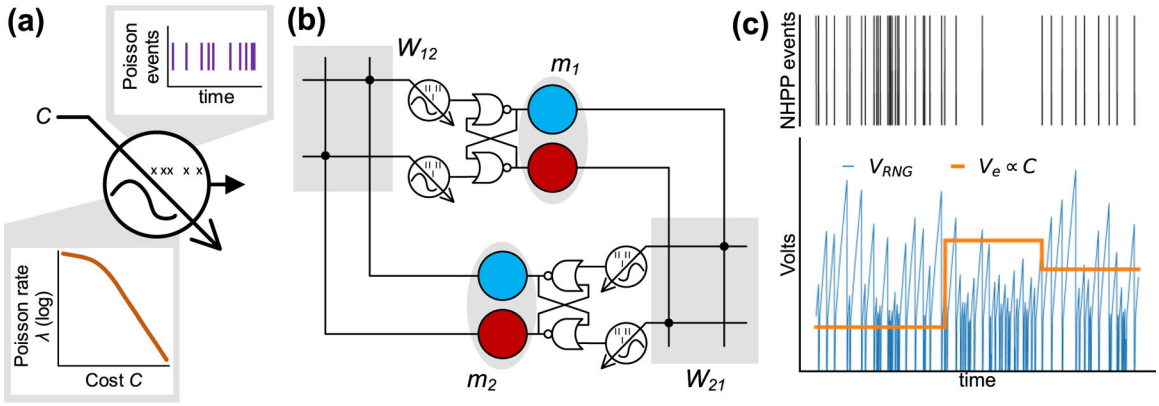

FIGURE 1.

VRSCs can be used for combinatorial optimization, such as graph coloring [1]. (a) VRSC is a CMOS realization of an NHPP. VRSCs/NHPPs are variable sources characterized by a latent Poisson rate $\lambda(C)$ and a resulting observable stochastic output of pulses. (b) Example of solving a (trivial) graph coloring problem using NHPPs. Pairs of NHPPs couple to RS latches to form P-bits, with the color of each vertex indicated by the state of the RS latch. Weighted connections link the states of the two P-bits such that they select different colors. (c) VRSC designs presented here all rely on a carefully designed random voltage, which is then compared with a threshold. NHPP events are observed when the random voltage crosses the threshold. The blue trace is a waveform of one of the time-to-amplitude converter circuits described later in the text.

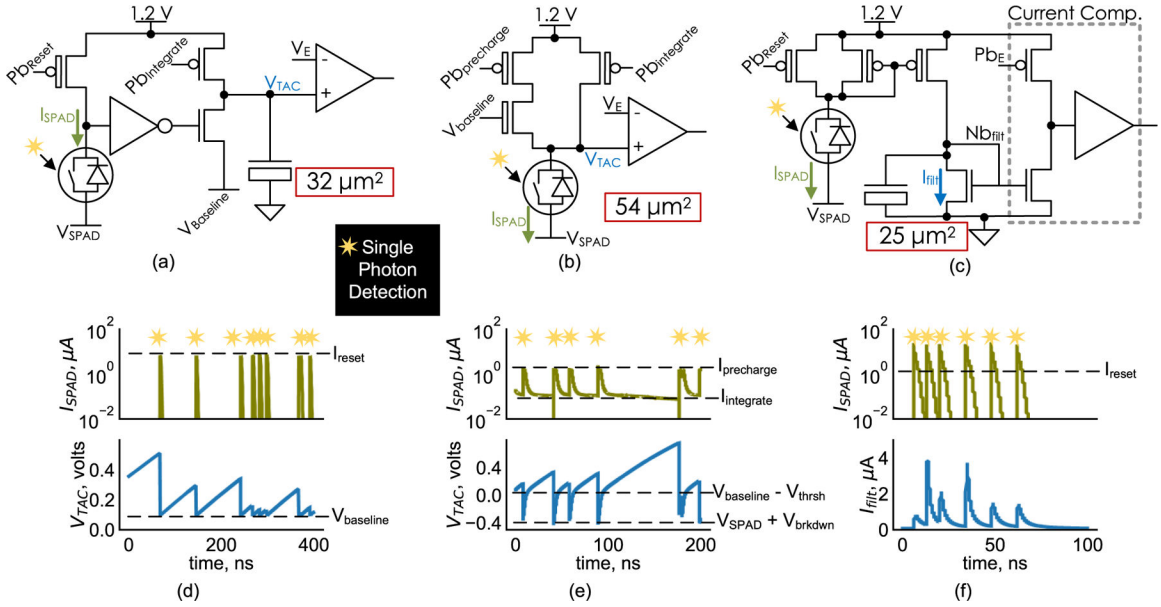

FIGURE 2.

Three tested types of VRSC processing circuits. The inputs Pb_{xx}, V_{SPAD}, and $V_{baseline}$ are constant biases; the Pb_{xx} voltages set the maximum current in their corresponding transistors. The inputs V_e and Pb_e are the analog thresholds for controlling the output rate of the VRSCs. Each design uses the random timing of SPAD avalanches to generate voltage (a) and (b) or current (c) ‘‘pulses’’ of random magnitude. After filtering these pulses with a comparator, the circuits become Poisson processes with adjustable mean rate λ. In the TAC designs (a) and (b), an integrated voltage is reset each time the SPAD avalanches; in the GmC-filtered design, the natural variability of each pulse and the prior state of a filtering capacitor result in variable current pulses each time the SPAD avalanches. (d) MOSCAP TAC simulated voltage (e) Diode TAC simulated voltage. (f) GmC-filtered simulated current.

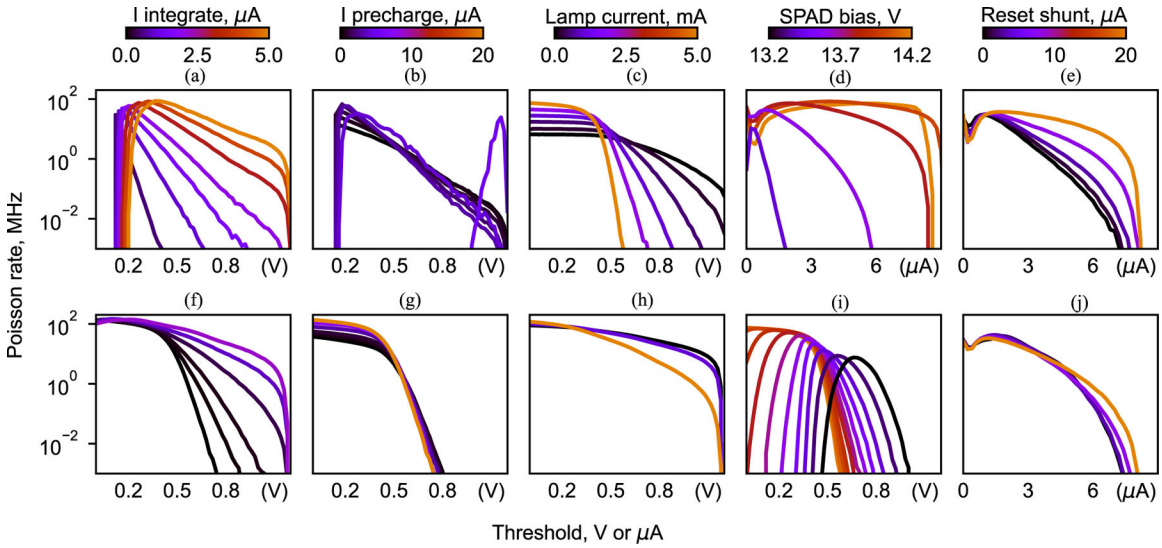

FIGURE 3.

A sampling of measured λ curves. Each vertical pair of subfigures compares the behavior of two different VRSC designs as a single test condition is adjusted, illustrating some of the key differences between designs. (a) MOSCAP TAC/sm. (b) MOSCAP TAC/sm. (c) Diode TAC/lg. (d) GmC filt/lg. (e) GmC filt/sm. (f) Diode TAC/sm. (g) Diode TAC/sm. (h) Diode TAC/med. (i) Diode TAC/lg. (j) GmC filt/med.

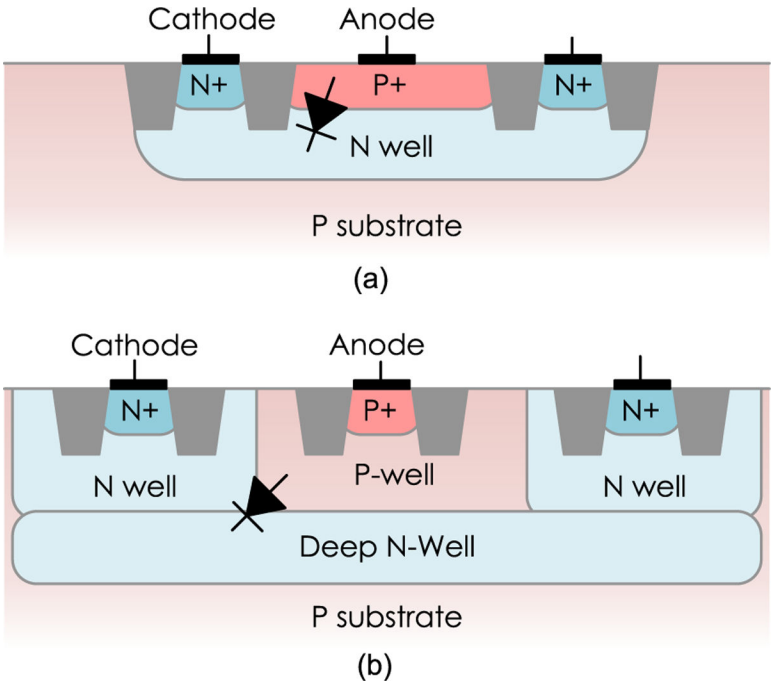

FIGURE 4.

Two different SPAD junction designs were used. (a) Testing included two sizes of a P+ NW SPAD, exhibiting very high DCRs. (b) One PW DNW SPAD was tested, which had much lower DCRs.

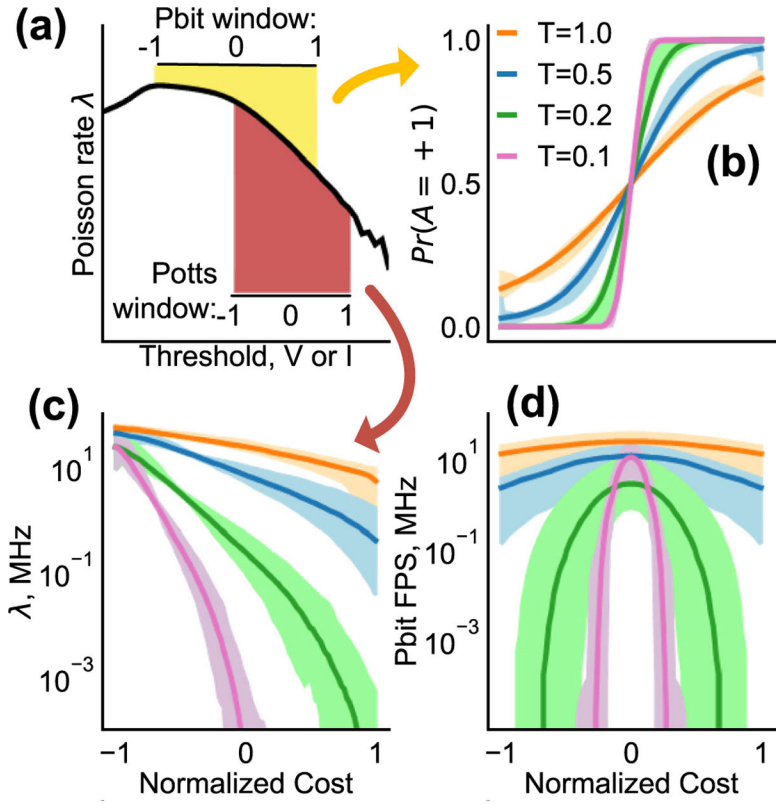

FIGURE 5.

For an application, the operating biases (I_{reset}, V_{SPAD}, etc.) need to be optimally selected. Selections can be tuned for specific temperatures. (a) In an application circuit, the VRSC's operating range can be restricted so that the application only sees the most functional portion of the λ (C) curve. (b) An example high-quality selection, optimized to match the tanh-shaped probability function required of P-bits. (c) A different selection, directly optimizing for the log-linearity of the λ curves, which is more critical for Potts model operation. (d) P-bit FPS, corresponding to the optimized condition in (a). The median and range of 24 samples are shown.

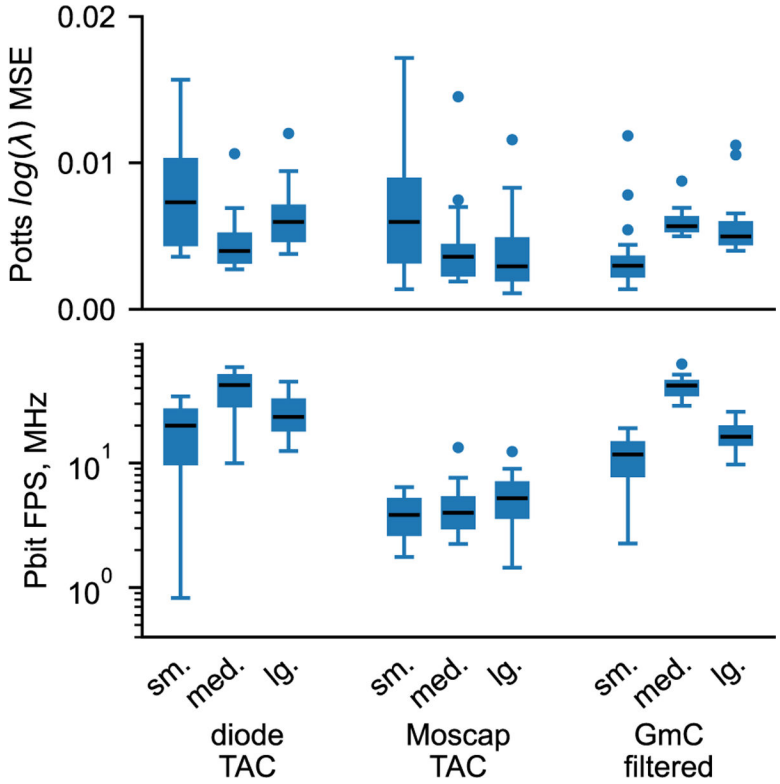

FIGURE 6.

For each VRSC design, the optimized performance and variability are shown. The Potts exponential cost (top) is desired to be a small as possible, with low variability; the P-bit flip rate (middle) should be as high as possible and does not need to be uniform.

TABLE 1.

Ideal VRSC characteristics.

VRSC Property	Ideal Case	Effect on Computation
Value of $\lambda(C=0)$	$\lambda(0) \Rightarrow \infty$	P-bit FPS / Speed of computation
Log-linearity of $\lambda(C)$	$\lambda(C) \propto \exp^{-C/T}$	Guaranteed Boltzmann Distribution Sampling
Slope of $\lambda(C)$	$T_{max}/T_{min} \Rightarrow \infty$	Thermal annealing schedule flexibility

TABLE 2.

Properties of tested SPADs.

Label	Junction	Breakdown	active area	cell area	DCR
sm.	P+ NW	9.7 V	$2\mu m^2$	$16\mu m^2$	1–10 MHz
med.	P+ NW	9.7 V	$10\mu m^2$	$56\mu m^2$	1–10 MHz
lg.	PW DNW	11 V	$20\mu m^2$	$64\mu m^2$	100 kHz

TABLE 3.

Comparison to other P-bit instances.

Type	Full Pbit?	Max FPS	T_{max}/T_{min}	Area	Energy/Flip
This work	Yes	50 MHz	10	81 μm^2	1 pJ
MTJ [18]	No	500 MHz	-	-	0.2 pJ
MTJ [19]	No	10 MHz	-	-	34 pJ
MTJ [20]	No	-	1	0.16 μm^2	-
MTJ [21]	Yes	1 KHz	1	-	-
Memristor [22]	Yes	0.5 Hz	5	-	1 μJ
Memrristor [23]	Yes	1KHz	1	-	154 pJ
FPGA [24]	Yes	100 MHz	1	42 LUTs	-
CMOS [25]	Yes	200 MHz	-	1600 μm^2	0.9 pJ