Applications of Resistive Switching Memory as Hardware Security Primitive

Roberto Carboni and Daniele Ielmini
Dipartimento di Elettronica, Informazione e Bioingegneria
Politecnico di Milano
Piazza L. da Vinci 32, 20133 Milano, Italy
daniele.ielmini@polimi.it

May 18, 2019

Abstract

With the widespread diffusion of ubiquitous mobile computing and internet of things (IoT), secured communication and chip authentication become key requirements. Hardware-based security concepts generally provide the best performance in terms of good security standard, low power consumption, and large area density. In these concepts, the stochastic properties of the device, such as the physical and geometrical variations of the process, are harnessed to generate random bits and functions. This is the basis for the true-random number generator (TRNG), where true random numbers are generated by exploiting the physics and randomness of nanoscale devices. The same random variations can also be used to implement physical unclonable function (PUF) for the authentication of individual hardware chips. Emerging memory devices rely on unique physical mechanisms for transport and switching, thus appear as the ideal source of entropy for hardware TRNG and PUF. These novel memory concepts include resistive switching memory (RRAM), phase change memory (PCM) and spin-transfer torque magnetic memory (STT-MRAM) devices. As these devices are increasingly adopted as memory and computing elements in several applications, exploiting their intrinsic stochastic variations for TRNG and PUF becomes an attractive solution for low cost, high performance security primitives.

This chapter provides an overview of TRNG and PUF adopting emerging memory devices as the fundamental entropy source. TRNG concepts are classified by the microscopic stochastic variation that is adopted as entropy source, namely, current noise, switching delay time, or switching voltage. While most TRNG concepts rely on RRAM devices, we also show novel concepts using STT-MRAM devices which take advantage of the excellent endurance and speed of switching. The TRNG schemes are discussed in terms of the simplicity of the design, e.g., the ability to generate random bits without a probability tracking by adopting a differential circuit scheme. Finally, the status of PUF implementations using RRAM and their array circuits are presented and discussed.

1 Introduction

Information security has been a topic of intense research since the mid 1970s, when the main purpose was to guarantee the confidentiality and integrity of data within mainframe computers [1]. In more recent times, as mobile computers, internet of things (IoT) and cloud computing are becoming ubiquitous, there is an ever-increasing need for secure communication among them [2]. Portable devices such as smartphones and tablets can now enable financial transactions and act as the primary authentication token for the user. Therefore, there is a need for electronic chips to (1) securely authenticate and be authenticated by other parties, (2) securely handle private/sensitive information, and (3) operate in an untrusted environment where the adversary might have physical access to the system [3]. These tasks must be implemented in mobile devices at the level of integrated circuits (IC), featuring at the same time both low power consumption and small area occupation. For application in large scale IoT [4] and cyber-physical systems (CPS) [5], security methodologies must also feature high speed, low cost and robustness to physical and side-channel attacks [6].

Hardware-intrinsic security primitives such as true random number generators (TRNG) and physical unclonable functions (PUF) are gaining interest towards low-cost and high-performance security tools [7]. On the one hand, TRNG can conveniently and efficiently generate the random bitstreams required by most of cryptographic and security applications [8, 9]. On the other hand, PUF can securely store a secret key in the random characteristics of an IC, by e.g. exploiting the random process fluctuations, and enabling fast and low-cost authentication and secure key storage [3]. Nano-devices are currently considered as the most promising approach for TRNG and PUF thanks to the small area, the low power consumption, the scalability, the 3D integration, and the ability to offer intrinsic stochastic phenomena via the inherent physical transport and switching mechanisms. These properties are all extremely beneficial for portable and IoT applications. Nanoelectronics can provide scalable device concepts via either the well-established complementary metal oxide semiconductor (CMOS) technology, or via alternative memory concepts based on resistive, phase change, magnetic and ferroelectric materials [10], sometimes referred to as the 'memristive' concepts [11]. CMOS-based TRNG [8] and PUF [12] were first introduced thanks to the strong integration capabilities and technological maturity. Nevertheless, they soon demonstrated a limited entropy quality and the need for increased area and power overhead to improve randomness [13]. On the other hand, memristive devices are currently gaining increasing interest for hardware security thanks to their intrinsic stochastic behavior that can be harnessed for high-performance and low cost, low energy on chip entropy sources.

This chapter provides an overview of the current state-of-the art for both TRNG and PUF implementations with resistive (memristive) switching devices.

The focus is on the applications of emerging memory technologies such as resistive switching memory (RRAM) and spin-transfer torque magnetic memory (STT-MRAM), combining binary stochastic switching, good endurance and scalable device area. The chapter is organized as follows: Section II describes the general framework of hardware security primitives such as TRNG and strong PUF. Section III is a short overview of the RRAM device, including the device structure and the switching mechanisms. Sections IV to VIII presents the possible approaches toward RRAM-based TRNG, based on the stochastic phenomena in RRAM devices such as current noise (Sec. IV), switching time variability (Sec. V) and switching voltage variability (Sec. VI). Sec. VII presents TRNG schemes based on differential pairs of RRAM, while Sec. VIII illustrates STT-MRAM-based TRNG concepts. Section IX reviews recently presented PUF concepts based on the RRAM technology. Finally, Section X provides a summary and an outlook for the research and development of hardware security using RRAM devices.

2 Hardware Security Primitives

2.1 PRNG and TRNG

Security of internet-based data transmission usually requires the generation of random keys [8, 9] via an on-chip random number generator (RNG). In the era of IoT, the need for compact and reliable RNG circuits with high entropy and high throughput has been considerably increased [14]. Other applications of RNG include the emerging computing paradigms, such as stochastic [15, 16, 17] and brain-inspired neuromorphic computing [18, 19], which inherently rely on large streams of random analog/digital signals for their operation . Within this scenario (Fig. 1), RNG circuits providing reliable random numbers with small circuit area, low energy consumption, and high throughput become essential.

A classical method for generating random bits is the pseudo-random number generator (PRNG) which can generate a random-looking bitstream according to a deterministic algorithm initialized by a seed (e.g. interrupt events, kernel calls, incoming TCP/IP request, etc.) [8, 20]. For example, a linear-feedback shift register (LFSR) is a digital circuit that, after being initialized with a seed, can generate a deterministic sequence of pseudo-random numbers [21]. However, LFSR is a finite-state machine, hence its output will be periodic, namely non-random over a sufficiently large time scale. Also, the seed might be manipulated since it is derived from user activity, or the knowledge of internal state and feedback tap structure of the LFSR might allow operation monitoring. These limitations arise from the fact, already recognized by Von Neumann [22], that random numbers cannot originate from a deterministic, arithmetical algorithm. These are all critical issues that make the PRNG output exposed to crypto-analysis [8, 23]. Due to the limited randomness and the high vulnerability, these systems are generally unsuitable for integration in IoT devices [24].

The data protection against cyber-attacks can be improved with the true RNG

(TRNG), where the output bitstream is obtained via an inherently stochastic physical process [25]. It has been demonstrated that the high unpredictability of hardware-based TRNG makes them more reliable with respect to software-based PRNG systems [26, 27]. In recent years, various physical entropy sources were proposed for TRNG, like the random telegraph noise (RTN) in dielectrics [28, 29] , stochastic quantum processes [25], stochastic spintronic phenomena [30, 31] and memristive transport and switching [32, 33, 34]. Several stochastic entropy sources were identified in both CMOS technologies and emerging memristive devices.

CMOS-based TRNGs were demonstrated by exploiting the noise in scaled MOSFET [28], the metastability at turn-on of cross-coupled inverter pair (namely, SRAM core) [8], or the increased noise of dual drain MOSFET driving a voltage-controlled oscillator (VCO) [35]. All these concepts take advantage from the mature integration capability of CMOS logic chips. However, CMOS-type TRNGs suffer from various drawbacks: for instance, a colored noise spectrum, e.g., due to capture/emission events originating 1/f noise, results in a biased output bitstream, requiring considerable post-processing and a consequent circuit overhead. Noise in CMOS devices also critically depends on environmental/process fluctuations, whose impact can be minimized only with entropy-tracking feedback loops [8], thus resulting in additional power consumption, circuit area and added complexity. On the other hand, memristive concepts such as RRAM and STT-MRAM enable ultra-small entropy source with high quality randomness, which makes these technologies very promising for TRNG.

2.2 Strong and Weak PUF

Secret information transmission with classical mathematical cryptography has relied on sufficiently hard-to-break algorithms (i.e. the “lock”) and secret keys since its inception [36]. Typically, the secret key is stored in a nonvolatile electrically erasable read-only memory (EEPROM) or a battery-backed static random-access memory (SRAM) which results in a relatively large area occupation and power consumption. For low-power IoT devices, storing secret keys at low energy consumption in a secure way is becoming an increasingly difficult task [37], especially considering emerging attack techniques such as the side-channel attacks [38]. This has led to an intense research interest for hardware-intrinsic security primitives that do not require secret key storage in the digital memory.

In this scenario, the physical unclonable function (PUF) is a promising solution. A PUF is a physical system that statistically maps an input challenge to an output response through a secret key controlled by a stochastic property of the chip, e.g. the silicon process variations or the intrinsic physical variability of device parameters [3]. In general, PUF security is guaranteed by the extreme difficulty of accessing the physical features of the hardware, and by the negligible probability that two chips are manufactured with the same or similar set of parameters. These properties make PUF an excellent scheme to uniquely identify a component or a circuit, thus enabling hardware authentication and preventing IC counterfeiting [12].

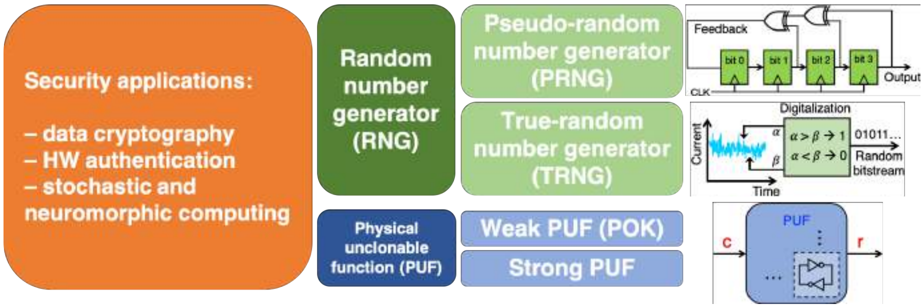

Figure 1: Schematic representation of the current information security scenario with some applications of hardware security primitives, comprising data cryptography, hardware (HW) authentication and stochastic/neuromorphic computing. The main building blocks are the random number generator (RNG) and the physical unclonable function (PUF). The RNG can be implemented either as a pseudo-random number generator (PRNG), such as the linear-feedback shift register (LFSR) in the figure, or as a true-random number generator (TRNG), which harnesses the stochasticity of physical phenomena (like the noise in the current trace in the figure) to generate a random bitstream. A PUF systems introduces a challenge (c) response (r) relation, where r $= f(\text{c})$ and $f(\cdot)$ is given by the physical details defining that specific PUF instance. In the figure, a typical SRAM-based PUF circuit is schematically shown. Adapted from [32].

A PUF system can be represented as a black box that for each input challenge c returns an output response r = f(c), with f describing the unique internal physical characteristics of the PUF (Fig. 1). The set of possible challenge-response pairs (CRPs) defines a particular PUF instance.

Depending on the number of unique CRPs, there are two main categories of PUFs: the weak PUF, which can only support a relatively small number of challenges, and the strong PUF, with an extremely large set of CRPs [3]. More specifically, in a weak PUF the number of CRPs increases linearly or polynomially with the number of basic cells, *i.e.*, the building blocks forming the PUF system [39], while the number of CRPs increases exponentially in a strong PUF [12]. The weak PUF is often referred to as physical obfuscated key (POK), since its primary task is the generation or storage of a cryptographic key [40, 41]. The most popular implementation of the PUF circuit is based on the digital static random access memory (SRAM), and exploits the metastable states of cross-couple inverters [42]. In each inverter pair, the response bit is determined by which of the two nominally equal-sized inverters of the memory cell addressed by the challenge reaches the tri-state point faster. Memory-based PUFs (POKs) are relatively easy to design even with low area overhead. Such memory-based systems are essentially weak PUFs since the set of CRP is limited by the available memory capacity [43]. As a result, their CRP set can be completely explored within polynomial time, compromising their use as identification tools. On the other hand, given their large CRPs, strong PUFs are practically immune from brute-force attacks [12] and are therefore suitable for low-cost authentication.

Although there is no general metric to certify a PUF system in terms of security properties, the following characteristics can be considered as the best figures of merit (FOM) for PUF [12]:

- *Reliability*: A PUF should always give the same response to a given challenge over a wide range of operating conditions (voltage, temperature etc.)
- *Unpredictability*: The PUF response to an arbitrary challenge should not be predicted based on the CRPs of another PUF or from the previous CRPs of the same PUF.
- *Unclonability*: The CRP mapping of a PUF cannot be physically or mathematically cloned, even for the original manufacturer of the PUF.
- *Physical Unbreakability*: Any physical attempt to maliciously modify the PUF should result in a malfunction or a permanent damage of the chip.

The practical evaluation of these FOMs for a specific PUF system is not straightforward in general, as discussed in Sec. IX. Although extremely promising for low-cost chip authentication, the PUF should be strong enough against attacks aiming at building a model for the PUF. This kind of attacks try to develop a model of a PUF instance by looking at a subset of its input-output pairs. Among these, the machine-learning attacks have been demonstrated to be particularly successful [44, 45]. The careful co-design of the stochastic memory and the

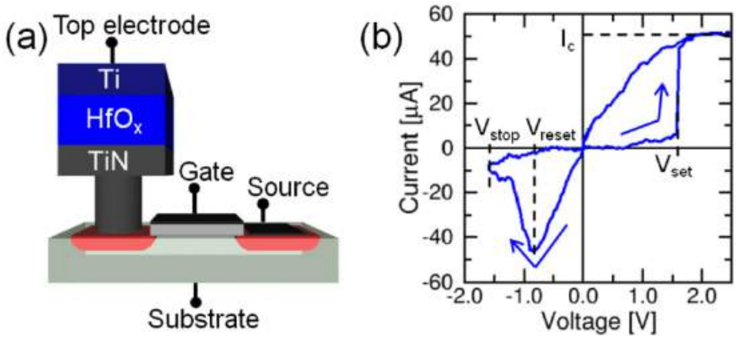

Figure 2: Schematic of a typical 1T1R structure comprising a RRAM cell integrated on top of the drain of an integrated MOSFET (a). In this example, the RRAM stack includes a Si-doped HfO_x switching layer, a Ti top electrode (TE) and a TiN bottom electrode (BE). The corresponding I-V characteristics shows the definition of the set voltage V_{set}, the compliance current I_C, the reset voltage V_{reset}, and the stop voltage V_{stop}. Reprinted with permission from [34]. Copyright (2016) IEEE.

circuit-dependent function is therefore essential for developing strong PUFs for hardware security.

3 RRAM Devices for Hardware Security

Among the emerging memory technologies, RRAM is one of the most promising due to its non-volatile retention, fast switching, low power and CMOS compatibility [46, 47, 48, 49]. The RRAM integration in cross-point arrays, in the back-end-of-line (BEOL), and adopting 3D structures allows for increased density and easy of integration [50, 51, 52, 53]. Fig. 2a shows a bipolar RRAM device, comprising a HfO_2 switching layer sandwiched between a TiN bottom electrode (BE) and Ti top electrode (TE). The Ti layer at the TE acts as an oxygen exchange layer inducing the generation of oxygen vacancies in the oxide layer, which thus become HfO_x with $x < 2$ [34]. RRAM devices are usually integrated in a one-transistor/one-resistor (1T1R) structure to enable the control of the resistance level by limiting the current flowing in the select transistor during the set transition. Fig. 2b shows the current voltage (I-V) characteristics of the RRAM device, where the application of a positive voltage to the TE causes a set transition from the high resistance state (HRS) to the low-resistance state (LRS) in correspondence of the set voltage V_{set}. The application of a negative voltage to the TE induces instead a reset transition from the LRS to the HRS in correspondence of the reset voltage V_{reset}. The resistance window between the LRS and the HRS is at least one order of magnitude, but can reach 5 orders of magnitude by the adoption of high band gap dielectrics such as SiO_x [54]. A relatively small gate voltage is applied during the set transition to limit the compliance current IC across the device, thus allowing to control the LRS resistance according to $R = V_C/I_C$, where V_C is a characteristics voltage generally lower than 1 V [55]. The LRS resistance can be thus controlled by the parameter I_C, while the HRS resistance can be controlled by the maximum negative voltage along the reset sweep, namely the stop voltage V_{stop} [54].

RRAM switching is caused by ionic migration induced by the voltage and the local Joule heating [56]. Because of the atomistic nature of the switching and local impact of microstructure, such as crystalline grain boundary and orientation, the set and reset transitions are characterized by a significant random variation [57]. The local conduction path does not only change during set and reset operations, but is also prone to stochastic atomistic fluctuations such as defect relaxation and diffusion which can cause a significant variation in the read current after the programming pulse [58, 59]. RRAM variations thus affect both the device-to-device consistency within a memory array [60], and the cycle-to-cycle variations within the same device because of the several different defect configurations. Variations can affect all RRAM parameters, including the LRS and HRS resistance, the set voltage V_{set} and the reset voltage V_{reset}. While stochastic variations are critical in hindering memory and computing application of RRAM [60], they offer the physical source of entropy that is needed for hardware security primitives.

Stochastic phenomena in RRAM devices can be exploited as entropy source for TRNG. RRAM schemes for TRNG can be grouped in three classes according to Fig. 3, where the sources of entropy are (a) stochastic noise, (b) stochastic switching time, or (c) stochastic switching voltage [10].

4 TRNG based on Stochastic Noise

The fluctuation of a bistable defect within the RRAM conduction path in either the LRS or HRS can lead to a relatively large fluctuation of the current between two levels called random telegraph noise (RTN, see Fig. 3a) [61]. RTN induces a random change in the read current between a low value I0 and a high value I1. By sampling the current trace in Fig. 3a, one obtains a bimodal distribution of currents in Fig. 3b which can be used to assign the random bits "0" and "1".

The current fluctuation in RTN can be ascribed to the modification of the charge state of a bistable defect close to the conductive path, due to e.g. electron trapping and detrapping combined by a structural relaxation of the defect. The charge state affects the carrier concentration close to the defect, thus resulting in a macroscopic change of the measured current [58]. As the filamentary path diameter of the LRS becomes smaller, the impact of the individual defect increases markedly, which is generally evidenced by the difference between the two resistance values ΔR increasing with the square of the average resistance ($\Delta R \sim R^2$) [61, 58]. This is similar to the RTN affecting the channel current in a MOS transistor, resulting from a bistable fluctuation of the charge state of an oxide defect. As the RTN amplitude can be quite significant, it can be exploited

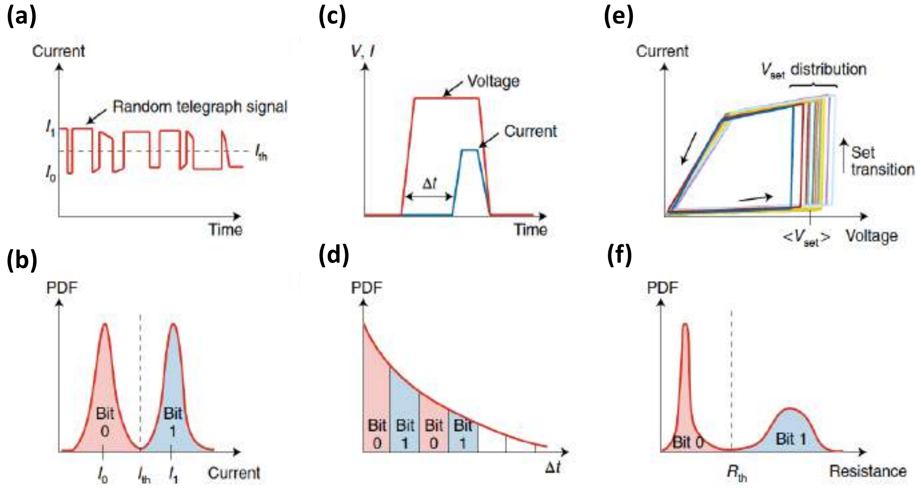

Figure 3: Random telegraph noise current fluctuations (a) and corresponding probabilistic distribution function (PDF) (b). In (c) the applied voltage pulse and its corresponding current response evidencing the random delay time Δt, and PDF of Δt (d) with an equally spaced time window to uniformly attribute bit values 0 and 1. Measured I-V curves evidencing cycle-to-cycle variation of V_{set} (e), and PDF of the resistance measured after a stochastic set (f), where sub-distributions of the high resistance state and the low resistance state are attributed to bits 0 and 1, respectively. Reprinted with permission from Macmillan Publishers Ltd: Nature Electronics [10]. Copyright (2018).

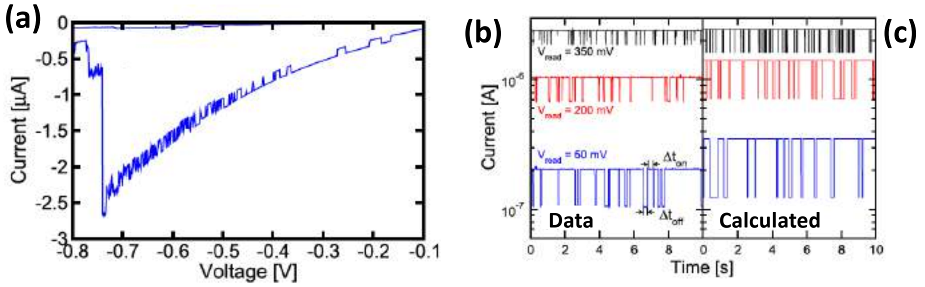

Figure 4: (a) Measured I-V characteristics for negative voltage showing RTN. (b) Measured read current as a function of time for read voltage $V_{read} = 50$, 200 and 350 mV and (b) corresponding simulations. The RTN switching times Δ t_{ON} and Δ t_{OFF} decrease with V_{read}. Reprinted with permission from [58]. Copyright (2014) IEEE.

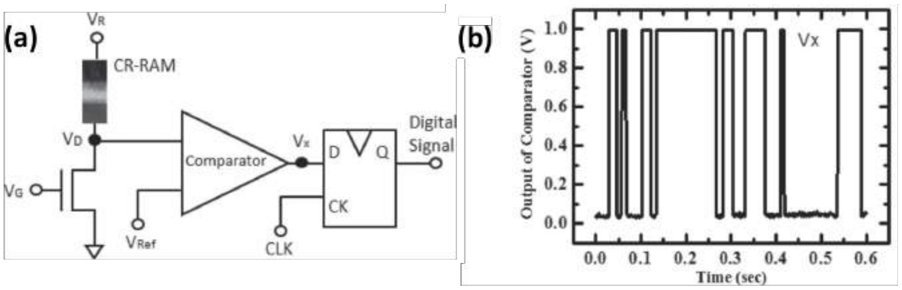

Figure 5: (a) Schematic representation of the TRNG block diagram, including CRRAM, comparator and clock control circuit. (b) Comparator output, showing binary random digital behavior. Reprinted with permission from [29]. Copyright (2012) IEEE.

as an entropy source in RRAM devices.

To understand the impact of RTN on device behavior, Fig. 4a shows the measured current voltage (I-V) characteristics for a RRAM device with HfO_x switching layer. The current trace was measured at negative voltage in the LRS state and clearly evidences discrete transitions, typical of RTN. Data show that RTN transition rate increases at higher V_{read}, which can be better understood by constant-voltage measurements of current as a function of time in Fig. 4b. Here, the average rate of switching between the two RTN states increases with the read voltage for V_{read} = 50, 200 and 350 mV. Conversely, the average time for which the current remains high (Δt_{ON}) and the time for which the current remains low (Δt_{OFF}) both decrease at increasing V_{read}. The same behavior can be seen in the numerical simulations of Fig. 4c obtained with a finite-element method (FEM) numerical model of RTN [58]. The voltage dependence of RTN can be understood as the acceleration of RTN fluctuation kinetics due to the voltage induced Joule heating. Similarly, RTN can be accelerated at high ambient temperature [58]. Fig. 5 shows the architecture of a TRNG circuit exploiting RTN in RRAM [29]. The TRNG is based on a contact-resistive random-access memory (CRRAM), integrated on the drain contact of a MOS transistor with a 1T1R structure. The 1T1R structure is biased with a voltage V_R, thus any RRAM fluctuation due to RTN results in a fluctuation of the voltage V_D at the transistor drain. The drain potential V_D is compared to a reference voltage V_{ref} by an integrated comparator (Fig. 5a), leading to a binary random digital output as shown in Fig. 5b. Sampling the digital output at increasing times with a clock frequency f_{CK} leads to a sequence of random bits provided $f_{CK} \ll f_{RTN}$, where f_{RTN} is the average rate of RTN fluctuations.

Although the scheme is very simple, the TRNG of Fig. 5 has few issues related to both the physical concept and the circuit. The circuit has been reported to have a relatively large area, namely 2400 F^2 in 65 nm technology, *i.e.* 10 μm^2 [29]. Practical TRNG based on RTN phenomena are also affected by a difficult control of amplitude, rate and uniformity of the physical RTN. In fact, an unbiased RNG with equal 50% probability of generating either a “0” or a “1” is obtained only if the I_1 and I_1 sub-distributions in Fig. 3b have the same area. Also, as previously described, RTN is affected by temperature and voltage, leading to instabilities of the RTN entropy source. The amplitude of the RTN should be large enough to be distinguishable by the comparator stage, while the reference voltage V_{ref} needs to be adjusted carefully depending on the specific level of the resistance and its fluctuation. The non-uniformity of the “0”/“1” balance in the output bitstream can be compensated by a digital post-processing such as the von Neumann algorithm, however this comes to the expense of an additional circuit area and power overhead.

Most recently, to compensate for the area occupation and other issues of the TRNG circuit of Fig. 5, the current difference in the $1/f^{\beta}$ noise of the RRAM device was used as the entropy source [32]. The RRAM noise is associated to multi-trap capture and emission events in defects (*e.g.* oxygen vacancies) along the conductive filament (CF) in LRS and the localized conductive path in HRS [59]. Fig. 6a shows the read current I_{read} measured for a RRAM in the LRS with an average R = 10 kΩ, biased with a read voltage $V_{\mathrm{read}} = 10$ mV. Current fluctuations due to the $1/f$ noise result in an increasing relative standard deviation σI/I, where I is the average value of I_{read} at any time t, while σI is its standard deviation (Fig. 6b). The increasing value of σI/I with the time is due to the increasing noise contributions at low frequency, which is typical of $1/f$ behavior of noise. The simulation results by a numerical Monte Carlo model of $1/f$ noise in Fig. 6c and d show similar behavior [59]. Fig. 6e shows the measure and calculated power spectral density (PSD) SI, evidencing a clear -1 slope, typical of the $1/f$ noise. The $1/f$ noise can be harvested for TRNG by the circuit shown in Fig. 7 [32]. Here, the noisy current is sampled at subsequent times t and t+Δt, then the two sampled currents are subtracted leading and the difference ΔI is compared to 0. Finally, the random bit value is assigned to 0 or 1 depending on ΔI being positive or negative, respectively. With respect to the RTN scheme of TRNG, the differential scheme allows both for a reduced area of 0.256 μm^2 (or 160 F^2 in 40 nm technology) and a reduced bias in the probability of extracting a “0” or a “1” bit. In fact, the differential current ΔI (Fig. 7a) follows a Gaussian distribution, thus ensuring that “0” and “1” have exactly the same probability of 50%. The circuit design (Fig. 7b) allows for a precise current value extraction using a timing sense amplifier (TSA) and a resistance-to-time converter (RTC) [62], while the parallel configuration of multiple devices enables up to 32 Mbps operation, with a 0.04 nJ/bit energy efficiency. Test results are finally reported by showing a minimum entropy higher than 0.999 over a broad range of temperature ($-40 < \mathrm{T}[°\mathrm{C}] < 120$) and with different voltages ($V_{\mathrm{DD}} = 0.1$ V) (Fig. 7c). The high performance of the scheme is further demonstrated by the P-value, *i.e.*, a FOM for randomness of the random bit stream, of 1000 groups of 1 Mb bitstream for the frequency NIST 800-22 test [63] (Fig. 7d).

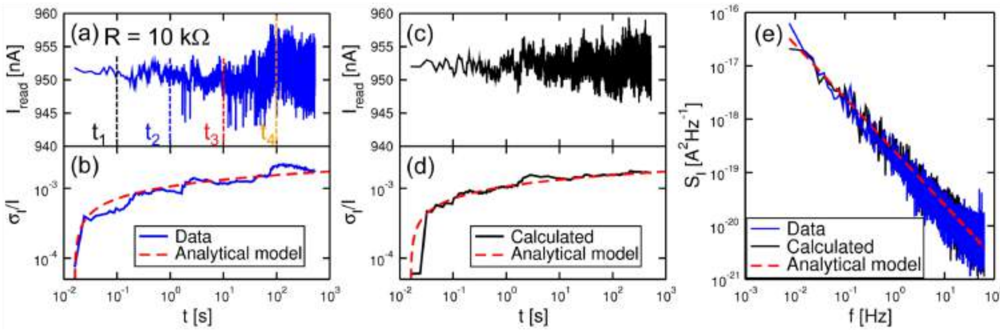

Figure 6: (a) Measured read current as a function of time for a device in LRS with R = 10 kΩ and (b) corresponding relative standard deviation σI/I. (c) and (d) Calculated current versus time and corresponding relative stadnard deviation. (e) PSD of experimental and calculated noise, showing a $1/f$ behavior. Results from the analytical model of [59] are reported in (b), (c) and (e). Reprinted with permission from [59]. Copyright (2015) IEEE.

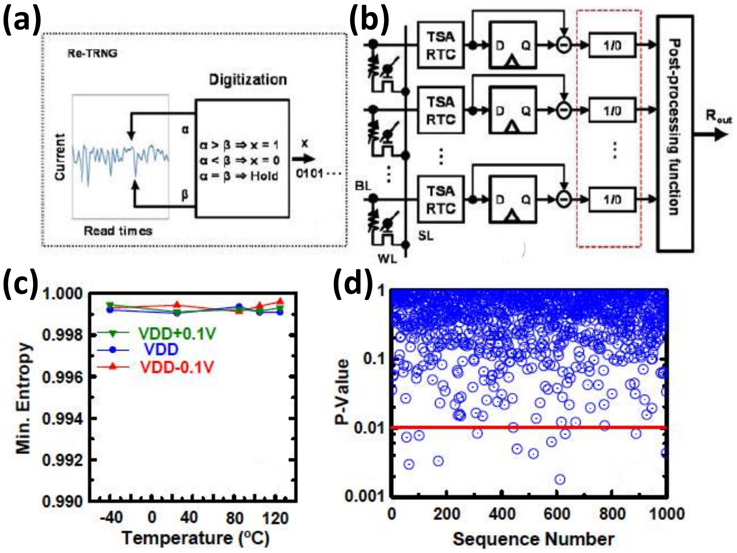

Figure 7: (a) Conceptual representation of the entropy harvesting algorithm for TRNG. (b) Block diagram of the parallel TRNG circuit, which allows for a 32 Mbps random bitstream. (c) Minimum entropies are higher than 0.999 over broad range of operative temperature and voltages. (d) P-value of 1000 sequences of 1 Mb bitstreams for the frequency test. Reprinted with permission from [32]. Copyright (2016) IEEE.

5 TRNG based on Stochastic Time

A key limitation of noise as a source of entropy is the unpredictable amplitude and frequency dependence. To better control the generation of random bit, TRNG can rely on the stochastic properties of switching, namely time and voltage.

Fig. 3c shows the basic concept for exploiting the variation of the stochastic delay time (Δt) for the set transition. Assuming that a voltage V slightly larger or comparable to V_{set} is applied to a RRAM device in the HRS, set transition occurs after a certain delay time Δt, where Δt decreases as the applied voltage is increased [56]. Most importantly, Δt is subject to relatively large variation from cycle to cycle, due to the ion migration being dependent on the local microstructure and atomistic migration of ions [59]. The resulting probabilistic distribution of Δt is exponentially decreasing as shown in Fig. 3d. The exponential distribution can be understood by the set transition being described by thermally-driven process to overcome a given energy barrier E_A [64]. As a result, the delay time Δt follows a Poissonian distribution $P(\Delta t) = 1/\tau$ exp(-$\Delta t/\tau$), where τ is the characteristic time constant given by $\tau = \tau_0 \exp(E_A/kT)$, where τ_0 is a constant, k is the Boltzmann constant and T is the local temperature [65].

For every set pulse in Fig. 3d, a random bit equal to 0 or 1 can be assigned based on the set transition taking place in even or odd time window controlled by a given constant frequency f_{CLK}. By repeating the set transition several times, a random bitstream can be generated. By using this scheme, an improved randomness quality of the generated bitstream can be demonstrated, provided that Δt is sufficiently large compared to the selected time window TCK and sufficiently small compared to the overall width t_P of the applied pulse ($T_{CLK} < \Delta t < t_P$) [66]. Therefore, the inherent randomness in the stochastic switching time received a big deal of interest as the fundamental entropy source for stochastic computing [67], neuromorphic circuits [68] and TRNG [10, 66]. Note that the sensitivity for switching is set by the window between HRS and LRS, thus providing a more robust TRNG scheme compared to the poorly predictable resistance change of RTN or 1/f noise.

Fig. 8 shows the measured distributions of Δt for increasing voltages V = 2.6 V (a), 3.2 V (b) and 3.6 V (c) [69]. The constant voltage was applied to the device in the HRS state, while the delay time Δt was measured at the onset of the set transition. The device was then reset with a negative voltage pulse, to allow for a repeated set transition [69]. The results confirm the exponential Poissonian distribution of the delay time Δt in Fig. 3d. Most importantly, the average $\Delta t = \tau$ in Fig. 8d decreases exponentially with the applied voltage, thus reflecting the decrease of the effective energy barrier E_A with the applied voltage [64, 56]. Data in Fig. 8d highlights that, although the single switching event is stochastic, the overall distribution of switching times can be predicted and controlled by the applied voltage [69, 68].

The stochastic delay time was adopted as the entropy source for TRNG by the circuit shown in Fig. 9a [66]. The proposed TRNG consists of a volatile

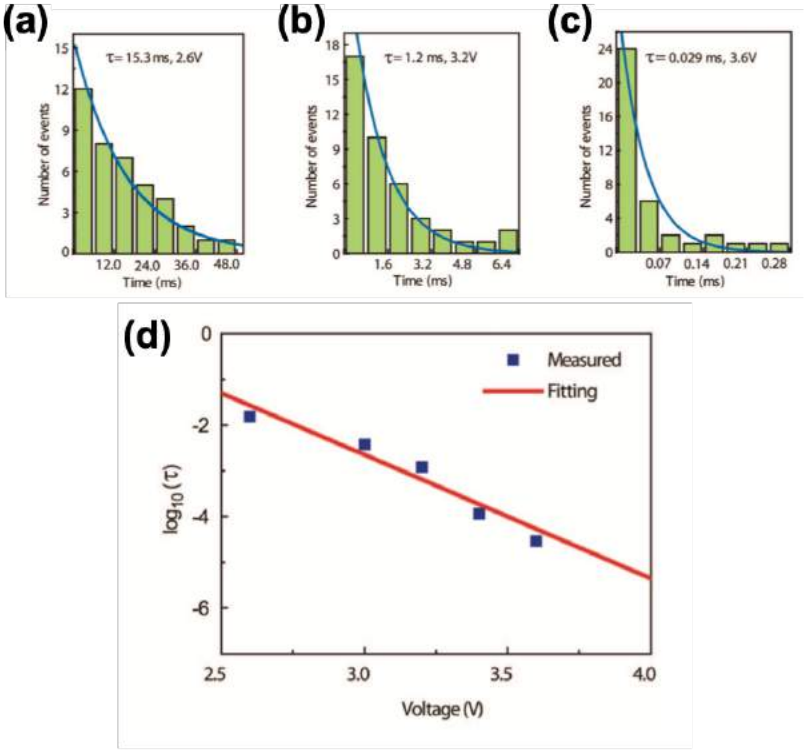

Figure 8: Distributions of switching time delay for applied voltage of 2.6 V (a), 3.2 V (b), and 3.6 V (c), with their corresponding fitting with the Poisson distribution. The only fitting parameter was τ = 15.3 ms, 1.2 ms and 0.029 ms for figure (a), (b) and (c), respectively. (d) shows the voltage dependence of τ. Reprinted with permission from [69]. Copyright (2008) American Chemical Society.

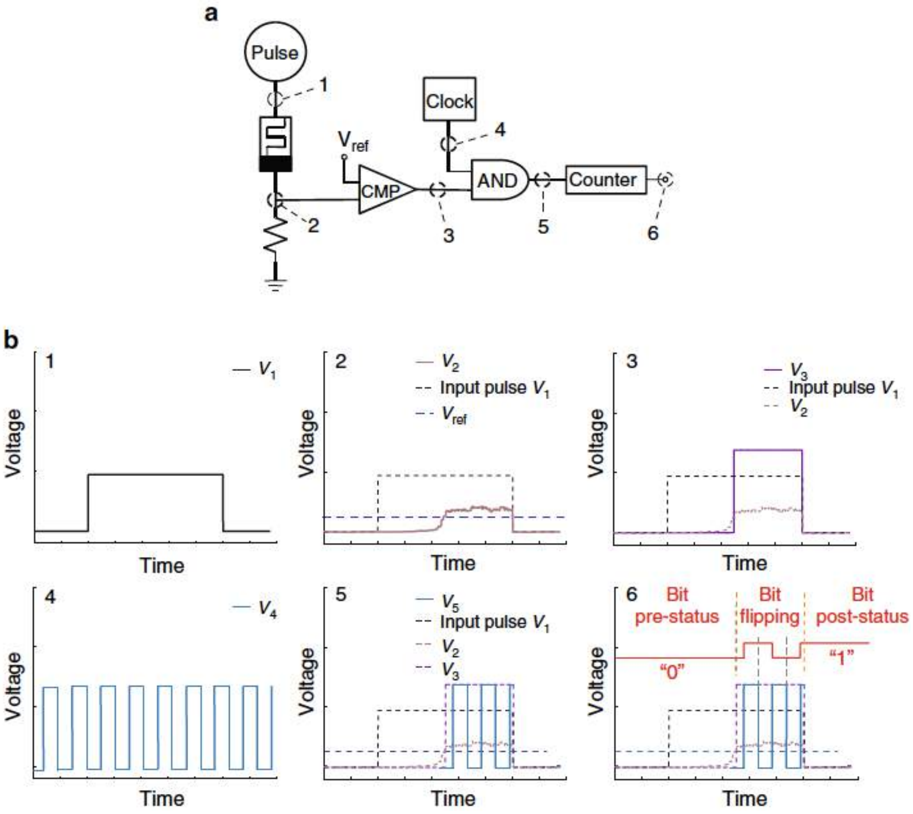

Figure 9: (a) Schematic representation of the TRNG circuit block diagram, comprising a memristive device, a comparator, an AND gate, and a counter. (b) Pulsed waveforms at each stage of the circuit, explaining the working principle of the TRNG. Reprinted from [66]. Creative Commons (2017).

RRAM device with Ag TE and Ag-doped SiO_2 dielectric layer. In this type of devices, the Ag migration from the TE results in the formation of an unstable CF, which decays soon after the set transition with a retention time ranging from few μs to few ms [70, 71, 72, 73]. The volatile behavior is due to the large diffusivity of Ag combined with the mechanical compressive stress in the dieletric layer [74] and the tendency to minimize the surface to volume ratio of the CF [71]. In the TRNG circuit of Fig. 9a, the volatile RRAM device is connected with a series resistance in a voltage divider configuration. The potential V_2 of the intermediate node of the voltage divider serves as the input of a voltage comparator. The comparator output and a clock pulse serve as the input of an AND gate, and a counter reads the AND output. A TRNG cycle is shown in Fig. 9b: the application of a voltage pulse V_1 (1) causes a set transition in the RRAM device after a stochastic Δt, which causes V_2 to suddenly increase above the reference V_{ref} (2), thus making the comparator output go to a high logic level V_3 (3). Due to the stochastic Δt, the V_3 pulse has a random duration, which is measured by the counter in units of the clock period T_{CLK}. Note that the binary bit (6) flips between 0 and 1 for the whole duration of the V_3 pulse, eventually resulting in a random bit. Note that a nonvolatile RRAM could be adopted in this scheme as well, however a reset pulse would be needed to re-initiate the device for a new cycle. The use of volatile RRAM in this case makes the TRNG algorithm easier and more energy efficient, as no reset pulse is needed.

To match the time window $T_{CLK} < \Delta t < t_P$, the pulse voltage V_1 should be carefully tuned, which usually requires complicated probability tracking techniques [75]. Also, extracting entropy from the stochastic switching time can be difficult due to its sensitivity on device parameters and process variations, requiring a probability tracking of the applied voltage for every TRNG on the same chip, or in separate chips [76].

6 TRNG based on Stochastic Voltage

A promising and more robust TRNG relies on the exploitation of the stochastic switching voltage. Namely, instead of measuring the delay time Δt for switching, one can monitor the device for a given amount of time, where the switching probability becomes the stochastic entropy source. This approach is schematically depicted in Fig. 3e, where various current-voltage characteristics measured on the same RRAM device demonstrate a distribution of set voltage Vset, due to the cycle-to-cycle variation. The application of a voltage equal to the average transition voltage <Vset> to the device in the HRS will then lead to a set transition with 50% probability. As a result, the measured resistance of the device after the applied voltage pulse then shows a bimodal distribution as indicated in Fig. 3f, where the two sub-distributions correspond to LRS and HRS. The random bitstream can thus be generated by associating the LRS and HRS to bit values “0” and “1”, respectively [33]. A similar scheme can be extended to stochastic computing, where an analogue value can be obtained as the sequence of stochastic bimodal resistance values obtained from the same

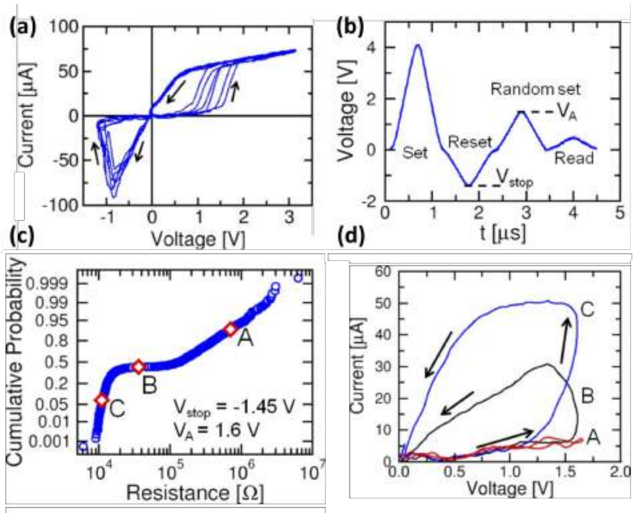

Figure 10: (a) Measured I-V characteristics for six cycles on the same 1T1R structure, evidencing stochastic switching. (b) Sequence of applied pulses for TRNG, with (c) the cumulative distribution of read resistance. Random set process is highlighted in the three I-V curves (d), corresponding to states A, B and C in (c). Reprinted with permission from [33]. Copyright (2015) IEEE.

device [68]. To illustrate the voltage-based TRNG concept, Fig. 10a shows the measured I-V curves for the same RRAM device with 1T1R configuration for six successive set/reset cycles [33]. The switching parameters, such as set and reset voltages, and the HRS and LRS resistance values show a large variability from cycle to cycle, which can be explained by considering the physics of the random formation and disruption of the conductive filament [57]. Fig. 10b shows the pulse sequence for characterizing the random set transition process, including: 1) a positive set pulse to deterministically initialize the device in LRS, 2) a negative reset pulse with a stop voltage V_{stop} to induce transition to the HRS, 3) a positive set pulse with voltage VA close to $<V_{set}>$ to stochastically induce a set transition event, and 4) a read pulse to measure the resistance in the final state. Fig. 10c shows the resulting resistance distribution for a random set experiment with $V_A = 1.6$ V. Data shows a bimodal distribution, corresponding to LRS sub-distribution with $R \approx 12$ kΩ and HRS sub-distribution above 100 kΩ. The origin of the bimodal distribution is clarified in Fig. 10d, which shows three characteristic I-V curves for various stochastic events, corresponding to state A, B and C in Fig. 10c. Case A corresponds to a cycle where V_{set} was higher than the applied V_A, due to a relatively high HRS after the reset pulse. As a result, no set process took place in this case, thus the resistance was found in the HRS sub-distribution (Fig. 10b). Case C corresponds to V_{set} being smaller than V_A, thus resulting in a set transition with a compliance current $I_C = 50$ μA controlled

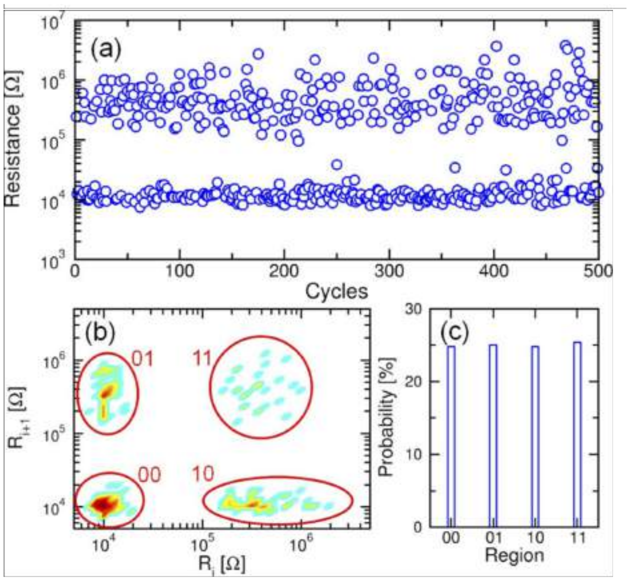

Figure 11: Measured resistance for 500 random set cycles with $V_{stop} = -1.45$ V and $V_A = -1.6$ V (a), correlation of R in cycle i+1 as a function of R in cycle i (b) and (c) population of the four regions in (b). Reprinted with permission from [33]. Copyright (2015) IEEE.

by the MOS transistor connected in series with the device. After the set event the resistance falls within the LRS sub-distribution in Fig. 10b. Finally, the intermediate case B corresponds to an applied V_A close to V_{set}. In this case, the device undergoes set transition, however cannot complete the transition within the pulse time. This results in an intermediate resistance between LRS and HRS, constituting the flat region of the bimodal distribution between HRS and LRS sub-distributions in Fig. 10b. It has been shown that this flat region, *i.e.* the occurrence of intermediate cases of type B, can be minimized by using a reduced pulse width or a proper shape (*e.g.* a saw-tooth shape with abrupt drop after reaching V_A) [33]. A key requirement for the TRNG in Fig. 9 is the absence of memory effects in the entropy harvesting process. To support this point, Fig. 11a shows the measured resistances during 500 successive random set cycles, clearly evidencing a bimodal distribution [33]. The absence of memory effect is further demonstrated in Fig. 11b, showing the correlation plot of R in cycle i + 1 as a function of the R in cycle i for all the cycles showed in Fig. 11a. We can identify four different regions, corresponding to the cell being in the same state (LRS or HRS) or different states in the two consecutive cycles. Fig. 11c shows the histogram representation of the probability for the four regions, showing comparable values around 25%. This demonstrates the lack of correlation across two consecutive cycles, which is consistent with true randomness of the bit stream.

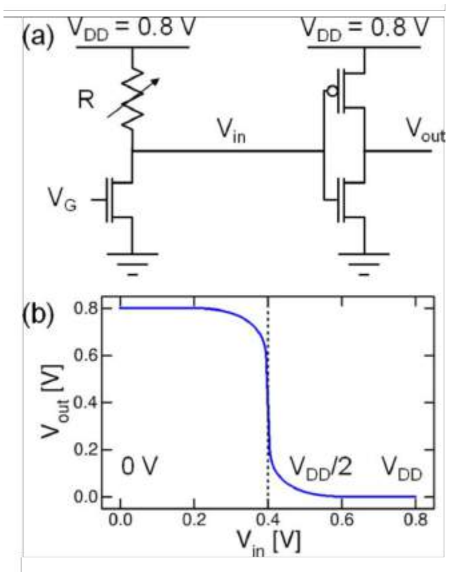

Figure 12: Regeneration circuit (a), comprising the 1T1R RRAM device and a CMOS inverter. (b) V_{in}–V_{out} characteristic of the inverter. Reprinted with permission from [33]. Copyright (2015) IEEE.

To guarantee proper RNG operation, a positive-feedback regeneration of the analogue output values might be required. Fig. 12a shows a compact regeneration circuit [33], comprising a RRAM device in 1T1R structure as the first stage and a CMOS inverter as the second stage. This scheme takes advantage of the relatively large resistance window between LRS and HRS, thus allowing the use of a small CMOS inverter instead of the larger analogue comparator, which is instead typically required for recovering the small signal in RTN-based RNG [29]. Fig. 12b shows the V_{in}–V_{out} characteristics of the CMOS inverter, evidencing the high gain in the transition region (with a threshold voltage $V_T = 0.4$ V) which allows for digital restoration. The impact of this regeneration circuit on the random bit distribution is illustrated in Fig. 13, showing measured and simulated bimodal resistance distributions (a), the simulated digital bimodal distribution of the inverter output V_{out} (b) and the sequence of the output voltage V_{out} for $2x10^5$ cycles (c).

To achieve a sufficient uniformity of the generated random bits, the applied voltage should be finely tuned to match the exact value $<V_{set}>$. This requires a preliminary probability tracking procedure [76], which results in a certain overhead in terms of complexity, area and power consumption

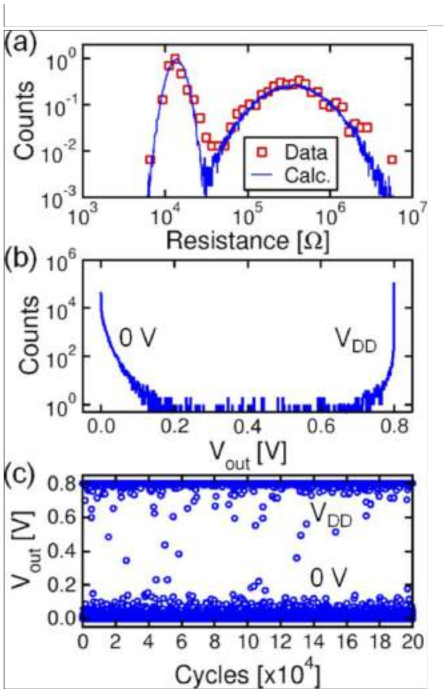

Figure 13: Measured and calculated distribution of the RRAM resistance (a), simulated distribution of the inverter output voltage V_{out} (b) and measured V_{out} for 2x10^5 cycles. Reprinted with permission from [33]. Copyright (2015) IEEE.

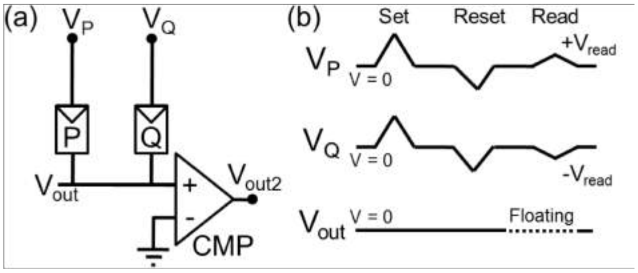

Figure 14: (a) Parallel-reset differential scheme for TRNG and (b) sequence of applied signals. Both P and Q start in HRS and are independently set, then reset and finally read using a voltage-divider configuration. The analogue comparator (CMP) digitally restores the output signal. Reprinted with permission from [34]. Copyright (2016) IEEE.

7 Differential TRNG Schemes

To overcome the need for a probability tracking in voltage-based TRNG, various differential schemes have been recently developed [34, 75]. In these TRNGs, either the competition between two RRAM devices [34] or the comparison between consecutive cycles on the same device [75] yields high-quality entropy without probability tracking, thus with a relatively simple circuit layout. A typical differential scheme relies on the coupling of two RRAM devices in either series or parallel configurations with the entropy source being the variability of set or reset transitions [34]. Three different schemes were proposed, namely: (a) parallel reset, (b) series reset and (c) parallel set, as detailed in the following [34]. Fig. 14a shows the parallel-reset TRNG circuit, comprising two RRAM cells, referred to as P and Q, connected in parallel. The common BE is connected to a comparator for the differential read. Fig. 14b shows the waveform applied to the TE of devices P and Q, *i.e.* V_P and V_Q, respectively, and the voltage V_{out} of the common BE node between P and Q. During a TRNG cycle, the applied waveforms include three phases, namely, 1) a positive voltage is applied across both P and Q in parallel, inducing set transition at both devices, 2) a negative voltage is applied across P and Q in parallel, inducing reset transition in both devices, 3) a differential read phase where $+V_{read}$ and $-V_{read}$ are applied at P and Q with floating BE to test the voltage divider between P and Q. Depending on the resistance values of P and Q, namely R_P and R_Q, respectively, the output voltage is found to be positive or negative, thus dictating the value of the output random bit. Given the relatively large variability of the HRS resistance [33, 57], Vout varies stochastically from cycle to cycle, thus constituting the basis for random bit generation. In this first approach, HRS resistance variation acts as the entropy source. Note that the bit value probability is automatically set to 50% by the uniform cycle-to-cycle distributions of HRS resistance of P and Q, as

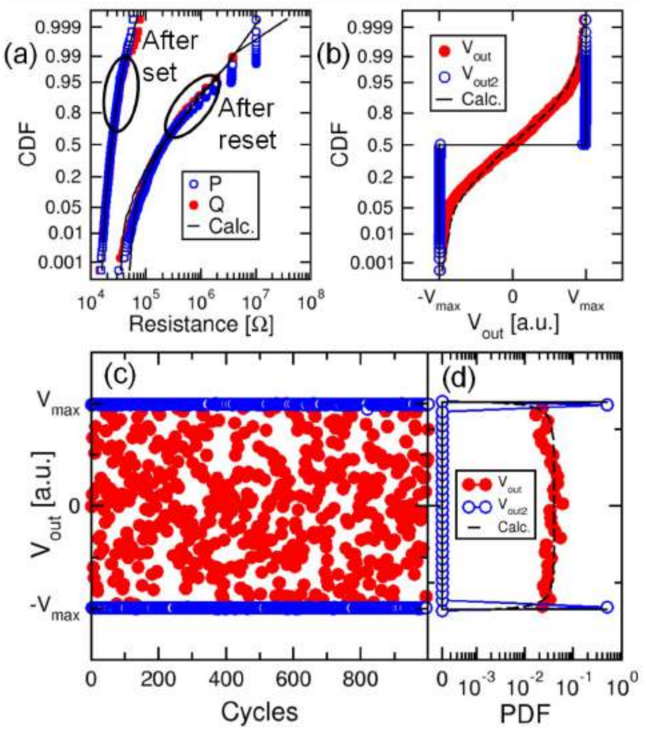

Figure 15: Cumulative distributions of resistance after set and after reset for cell P and Q (a). (b) Distributions of the output voltage V_{out} and V_{out2}, before and after the CMP, respectively. (c) Measured V_{out} and V_{out2} for 1000 RNG cycles with the corresponding PDFs (d). Reprinted with permission from [34]. Copyright (2016) IEEE.

the cycle-to-cycle variation in RRAM is comparable to the cell-to-cell variation [77].

Fig. 15a shows the cumulative distributions of measured and calculated R_P and R_Q, both after set and after reset. The read V_{out} distributions are shown in Fig. 15b for experimental and calculated data, indicating a bimodal shape with 50% transition probability. By reading the voltage V_{out} with an analogue comparator (Fig. 14a), the bimodal distribution can be improved, as shown by the distribution of the comparator output V_{out2} in Fig. 15b. The bulky comparator may be replaced by a CMOS inverter, thus reducing the on-chip area occupation [33]. To demonstrate the cycle-by-cycle operation of the parallel-reset scheme, Fig. 15c shows V_{out} and $_{\mathrm{Vout2}}$ for 1000 consecutive cycles, while Fig. 15d shows their corresponding probability density function. The TRNG does not require any probability tracking thanks cycle-to-cycle variability being comparable to the cell-to-cell variability [77]. Fig. 16a shows an alternative

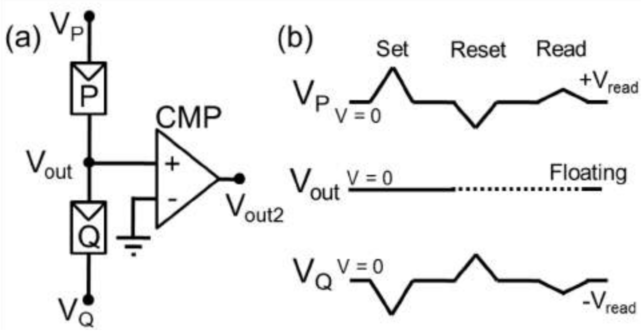

Figure 16: (a) Series reset differential scheme for TRNG and (b) sequence of applied signals. From the HRS, the cells are independently set, then they undergo a random reset, during which only one can switch, and finally they are read in voltage-divider configuration. Reprinted with permission from [34]. Copyright (2016) IEEE.

differential TRNG scheme, namely the serial reset configuration. This comprises two RRAM devices connected in series with $\mathrm{V_P}$ and $\mathrm{V_Q}$ as supply voltages and the intermediate node of potential $\mathrm{V_{out}}$ connected to an output comparator. Fig. 16b shows the applied waveform of $\mathrm{V_P}$, $\mathrm{V_Q}$ and $\mathrm{V_{out}}$ during a TRNG cycle, consisting of 1) independent set of P and Q, 2) random reset of either P or Q, 3) differential read of $\mathrm{V_{out}}$. For simplicity, we assumed $\mathrm{V_Q} = -\mathrm{V_P}$ in the figure. During the random reset event, a negative voltage $\mathrm{V_P} - \mathrm{V_Q} < 0$ is applied to the two devices in series, while the common node is left floating. A total applied voltage $|\mathrm{V_P} - \mathrm{V_Q}| > 2\ \mathrm{V_{reset}}$ drops across the devices, thus inducing reset transition in one of the two devices. In fact, once the transition begins in one of the two cells, the voltage across it increases because of the voltage divider effect, while the voltage drop across the other device decreases, thus preventing the two devices to both undergo reset transition. This configuration thus realizes a positive feedback, resulting in a self-accelerated reset event that takes place randomly in one device only. Specifically, the reset transition takes place in the device with the smallest $\mathrm{V_{reset}}$. Because of the cycle-to-cycle variability of $\mathrm{V_{reset}}$, the probability for one device to reset is ideally 50% [57]. Fig. 17a shows the cumulative distribution of $\mathrm{R_P}$ and $\mathrm{R_Q}$ after set and reset pulses in Fig. 16b [34]. After the random reset pulse, both P and Q show the same bimodal distribution with transition point at 50% probability, thus demonstrating unbiased TRNG with no need for probability tracking. To gain further insight on the random reset process, Fig. 17b shows the correlation plot of $\mathrm{R_Q}$ as a function of RP after either set or reset. $\mathrm{R_P}$ and $\mathrm{R_Q}$ appear to be anti-correlated after the reset phase, namely $\mathrm{R_P}$ is high for low $\mathrm{R_Q}$ and vice versa, which thus reveals a conditional reset of one RRAM device only. Fig. 17c shows the distributions of experimental and calculated $\mathrm{V_{out}}$, indicating a bipolar mode with transition point at 50% probability. Similar to other TRNG schemes, a digital regeneration can be

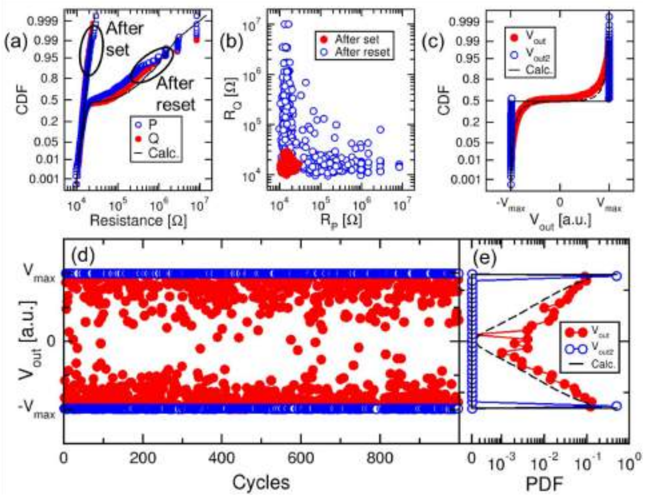

Figure 17: (a) Cumulative distributions of R after set and after reset for both cells P and Q. (b) Correlation plot of R_Q as a function of R_P. (c) Cumulative distributions of V_{out} and V_{out2}. (d) Measured V_{out} and V_{out2} during RNG cycling and (e) corresponding PDF. Reprinted with permission from [34]. Copyright (2016) IEEE.

obtained by a comparator or a CMOS inverter. Fig. 17d shows the cycle-to-cycle values of V_{out} and V_{out2} during the application of the RNG pulse scheme of Fig. 16b. Note that after each differential read phase, a final deterministic reset pulse was applied to ensure equal HRS conditions in P and Q before the application of the set pulse. Fig. 17e shows the corresponding distributions of V_{out} and V_{out2} for both data and calculations [34]. Fig. 18a shows the parallel set scheme [34], where the two RRAM devices in parallel configuration are connected to a common select transistor, with the drain terminal connected to the input node of a comparator. Fig. 18b shows the applied waveform cycle, including 1) an independent reset of P and Q, 2) a random set pulse of P and Q, and 3) a differential read by the application of a voltage 2 V_{read} across the two devices, while the transistor is biased in the off state. This TRNG scheme is based on the one-transistor/two-resistor (1T2R) structure in Fig. 18a, where the application of a positive voltage across the devices causes set transition to take place randomly in one of the two devices first. As a result of the transition to LRS and the voltage divider effect with the transistor, the voltage drops across both devices, which prevents any set transition to take place in the second RRAM device. In this TRNG scheme, the cycle-to-cycle variability of V_{set} plays the role of entropy source. Fig. 19a shows the read resistance distributions for P and Q, evidencing the expected bimodal shape with HRS/LRS transition at 50%. In order to verify that the random set happens stochastically in either one

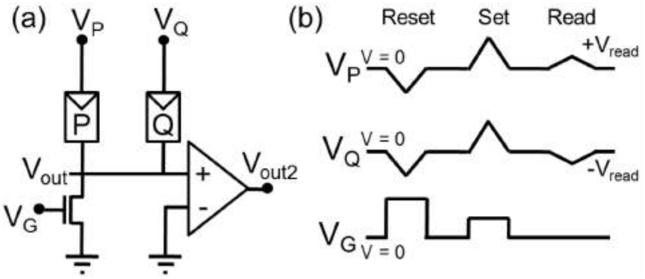

Figure 18: (a) Parallel set differential scheme and (b) sequence of applied signals. From the LRS, the cells are first independently reset, the subjected to parallel set, and finally read with voltage-divider configuration. Reprinted with permission from [34]. Copyright (2016) IEEE.

of the devices, Fig. 19b shows the correlation plot of R_Q as a function of R_P, again indicating an anti-correlation where P is in HRS for Q in LRS, and vice versa. Finally, Fig. 19c shows the cycle-to-cycle output values of V_{out} and ${}_{Vout2}$, while Fig. 19d shows their corresponding probability distributions.

Comparing these solutions for entropy harvesting, different performances are apparent in terms of bimodal distribution of R and V_{out}. For instance, the parallel-set TRNG (Fig. 19) shows improved results with respect to the parallel-reset TRNG (Fig. 15). This can be understood considering the abrupt set transition in the parallel set process as opposed to the more gradual reset event in the parallel reset process. The abrupt set transition is explained by the physical positive feedback where the first initiation of the filament causes an increase of the local Joule heating, thus accelerating the further growth of the filament [57]. This highlights the key role of the physics of the entropy-generating process has in controlling the quality of the TRNG circuit.

A general drawback of the differential pair approach is the assumption that cycle-to-cycle variation dominates over the cell-to-cell variation. In presence of a large mismatch between the two cells in the differential pairs, *e.g.*, where one cell systematically displays a lower V_{set} than the other cell, the TRNG might show deviations from the uniform behavior. Although this might be acceptable for PUF applications, where the random unique key has to be generated only once in the lifetime of the device, it might cause non-acceptable non-uniformities in TRNG [34].

8 STT Magnetic Memory for TRNG

The presented TRNG schemes can be adopted for all stochastic memory devices, *e.g.*, the phase change memory (PCM) or the STT-MRAM. In particular, STT-MRAM offers improved cycling endurance [78] and fast switching [79] which might

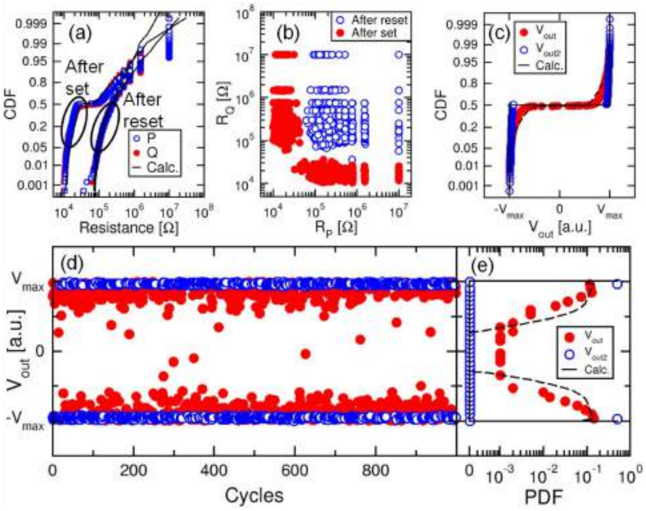

Figure 19: (a) Cumulative distributions of R after set and after reset for P and Q. (b) Correlation plot of R_Q as a function of R_Q after set and reset. (c) Cumulative distributions of V_{out} and V_{out2} during RNG cycling, and (e) corresponding measured V_{out} and V_{out2} PDF. Reprinted with permission from [34]. Copyright (2016) IEEE.

benefit the TRNG operation by providing an extended lifetime and throughput. Fig. 20a shows a typical state-of-the-art STT-MRAM device, consisting of a magneto-tunnel junction (MTJ) with perpendicular magnetic anisotropy (PMA) [78]. The MTJ consists of a pinned layer (PL) and a free layer (FL), acting as bottom electrode (BE) and top electrode (TE), respectively, and both made of ferromagnetic CoFeB. Between the two electrodes, a dielectric layer made of crystalline MgO serves as the tunneling barrier to induce the MTJ effect [80]. As schematically shown in Fig. 20b, this memory device has two stable states, where the magnetic polarization of the FL can be either parallel (P) or antiparallel (AP) to the magnetization of the PL, resulting in low or high resistance of the MTJ, respectively [78, 80]. Fig. 20c shows the measured current-voltage (I-V) characteristics, while the corresponding resistance-voltage (R-V) characteristics is shown in Fig. 20d. Set transition from AP to P, and reset transition from P to AP, take place at the positive voltage V_{set} and at the negative voltage V_{reset}, respectively.

As for the RRAM device, set and reset transitions in STT-MRAM are affected by stochastic switching, thus introducing a randomness causing a voltage-dependent bit error rate (BER) in memory applications [79]. The inherent stochastic switching causes cycle-to-cycle variations of both V_{set} and V_{reset} [81]. Although showing apparently similar variability, the physical origin of the stochastic switching voltage is quite different in STT-MRAM and RRAM. In

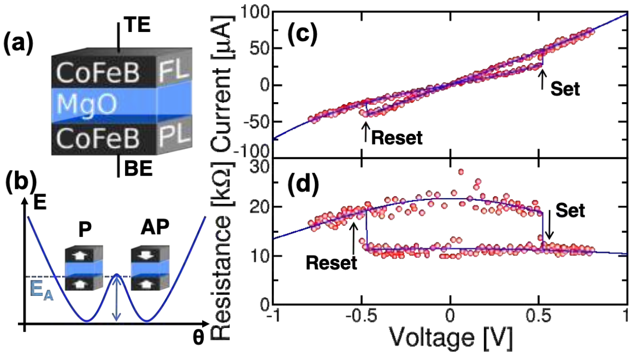

Figure 20: (a) Typical STT-MRAM device , consisting in a magnetic tunnel junction (MTJ) stack. (b) Energy as a function of the FL magnetic polarization direction with respect to the PL, showing P and AP states. (c) Measured and calculated I-V and (d) R-V pulsed characteristics with 1 μs pulse width. Reprinted with permission from [75]. Copyright (2018) IEEE.

fact, the statistical variations in STT-MRAM switching can be explained by the thermally-assisted magnetization reversal [82], where the transition from AP to P and vice versa are induced by a random thermal fluctuation within the potential well of Fig. 20b, and a stochastic transition over the energy barrier E_A between the two states. As a result, for each applied positive or negative voltage V_A, there is a statistical distribution of set time t_{set} or reset time t_{reset}, respectively.

The stochastic switching in STT-MRAM has been used for various TRNG concepts, either based on the time variation [31, 83] or the voltage variation [76, 30]. In particular, in the work from Vodenicarevic et al. [83] the stochastic switching time was exploited through an MTJ stack engineering. Namely, a low-stability (*i.e.* characterized by a reduced magnetic stability) free-layer was introduced instead of relatively high-stability nanomagnet used in memory applications. This structure is referred to as superparamagnetic tunnel junction [84] and shows spontaneous stochastic switching between the two stable states due to low stability relative to thermal fluctuations.

However, all these schemes necessarily rely on a careful biasing configuration, thus requiring a probability tracking approach to ensure the TRNG uniformity. Probability tracking can be avoided by using differential concepts, however, the differential pair approach is affected by the cell-to-cell mismatch within the pair. To solve these issues, a novel differential concept was presented, where the consequent switching cycles are compared in the same device, instead of two coupled devices [75]. Fig. 21a shows the applied voltage and the device current response over two consecutive set/reset cycles. In each cycle, a stochastic pulse with positive V_+ is applied, followed by a deterministic pulse with negative V_-. Both pulses have a pulse duration of 1 μs, although the concept can be

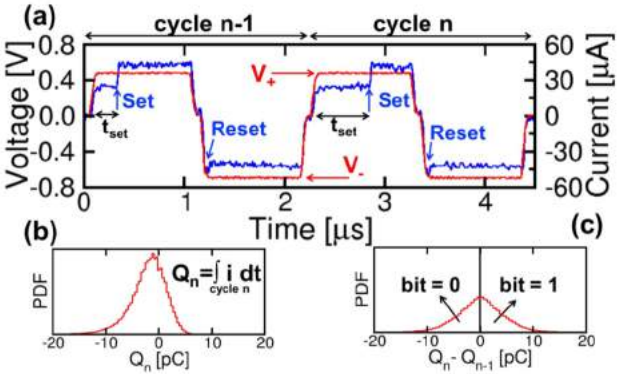

Figure 21: (a) Measured rectangular voltage pulses and current response for 2 consecutive cycles n–1 and n, (b) PDF of the integrated current $\mathrm{Q_n}$ and (c) PDF of differential charge $\Delta\mathrm{Q_n} = \mathrm{Q_n} - \mathrm{Q_{n-1}}$. The pulse sequence includes positive and negative rectangular pulses for stochastic set and reset transitions, respectively, as evidenced by the abrupt steps in the current response. The random bit is assigned from the value of $\Delta\mathrm{Q_n}$ in (c). Reprinted with permission from [75]. Copyright (2018) IEEE.

easily scaled to a shorter pulse-width thanks to the high speed of the switching process in the STT-MRAM. The stochastic switching is evidenced in Fig. 21a, where a shorter delay time $\mathrm{t_{set}}$ is observed during cycle n–1 with respect to cycle n. the TRNG relies on the comparison between the current responses between two consecutive cycles of the same STT-MRAM device. Fig. 21b shows the probability distribution of the integrated current $Q_n = \int idt$ while Fig. 21c shows the corresponding difference $\Delta Q_n = Q_n - Q_{n-1}$. Given the highly symmetric distribution of ΔQ_n, the latter is chosen as the statistical variable for random bit generation, where a random bit value 0 or 1 is assigned for $\Delta Q_n < 0$ or $\Delta Q_n > 0$, respectively [75].

Fig. 22a shows the same concept for TRNG applied to the case of a triangular waveform. Both positive and negative triangular pulses are applied for stochastic set and deterministic reset, respectively. In this case, the stochastic switching is evidenced by the different set and reset voltage in cycles n–1 and n, resulting in different current waveform during the two consecutive cycles. Fig. 22b shows the distribution of the integrated current over a single cycle $Q_n = \int idt$ while Fig. 21c shows the difference $\Delta Q_n = Q_n - Q_{n-1}$ over two consecutive cycles, serving as the stochastic variable for bit generation. In the TRNG concepts illustrated in Fig. 21 and Fig. 22, the entropy source is either the stochastic distribution of switching time, or the stochastic distribution of switching voltage, respectively [75].

Generally, TRNG concepts require further whitening algorithm, such as the Von Neumann correction [76] or the XOR operation [83], to achieve a truly

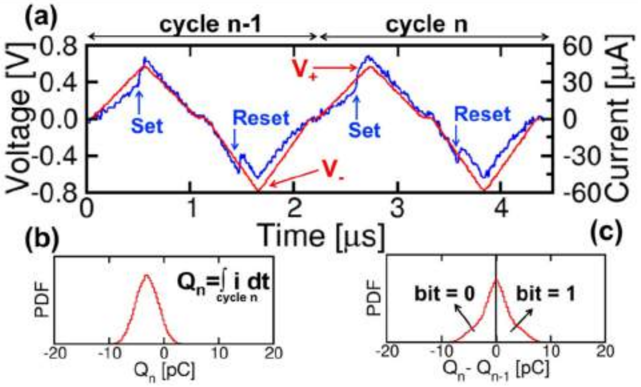

Figure 22: (a) Measured triangular voltage pulses and current response for 2 consecutive cycles n–1 and n, (b) PDF of the integrated current $\mathrm{Q_n}$ and (c) PDF of differential charge $\Delta\mathrm{Q_n} = \mathrm{Q_n} - \mathrm{Q_{n-1}}$. The pulse sequence includes positive and negative triangular pulses for stochastic set and reset transitions, respectively, as evidenced by the abrupt steps in the current response. The random bit is assigned from the value of $\Delta\mathrm{Q_n}$ in (c). Reprinted with permission from [75]. Copyright (2018) IEEE.

unbiased bitstream. However, the scheme of Figs. 21 and 22 can pass the standard test of the National Institute for Standards and Technology (NIST) [63] without any post-processing, thus enabling a reduced energy and area overhead of the TRNG circuit [75]. Fig. 23 reports the pass rate for the non-overlapping template test in the NIST criteria as a function of pulse voltage for rectangular and triangular pulses. The TRNG with rectangular pulse shows an acceptable accuracy only in correspondence of a narrow window of voltage, with a randomness degradation for both high and low voltages. On the other hand, the TRNG with the triangular pulse shows high pass rate over the whole test range, demonstrating a high voltage-independent randomness. These results can be explained by considering the applied voltage ($\mathrm{V_A}$) dependence of the switching parameters $\mathrm{t_{set}}$ and $\mathrm{V_{set}}$ (or $\mathrm{t_{reset}}$ and $\mathrm{V_{reset}}$) for rectangular and triangular pulses [75]. Considering a rectangular pulse, the set time $\mathrm{t_{set}}$ can be written as [85]:

$$t_{set} = \tau_0 \exp\left(\Delta(1 - \frac{V}{V_0})\right), \tag{1}$$

where V_0 and τ_0 are constants, V is the applied voltage, and Δ is the thermal stability factor. Given the exponential dependence in Eq. 1, there is only a narrow window of voltages where the switching time $\mathrm{t_{set}}$ is comparable to the applied pulse width (Fig. 21a). On the other hand, the set voltage under a triangular pulse, where the applied voltage is ramped according to $V(t) = 2V_A t/t_P$, can be estimated from the switching integrated probability reaching one, namely $\int 1/t_{set}dt = 1$, with $\mathrm{t_{set}}$ defined by Eq. 1. Thus, the set voltage along a triangular

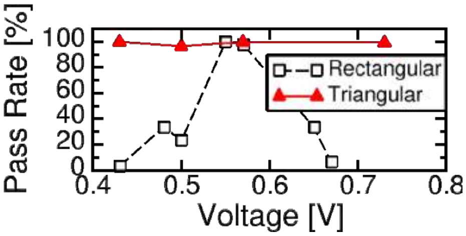

Figure 23: Pass rate of the non-overlapping template NIST test as a function of pulse voltage for rectangular and triangular pulses. The pass rate is referred to a total of 148 tests. Rectangular pulses show an operation window around 0.6 V, whereas triangular pulses show voltage-independent high randomness. Reprinted with permission from [75]. Copyright (2018) IEEE.

pulse is given by [82, 64]:

$$V_{set} \approx V_0 \ln \frac{t_0 V_A}{V_0 t_P}, \tag{2}$$

suggesting a logarithmic dependence of $\mathrm{V_{set}}$ on the maximum applied voltage $\mathrm{V_A}$. This explains the voltage-independent high entropy for the triangular pulse scheme with respect to the rectangular pulse in Fig. 23. Owing to this different dependence, the time-based scheme (rectangular pulse) might still require some probability tracking to find the correct $\mathrm{V_A}$ for optimal performance. In general, differential reading schemes based on stochastic voltage look more promising with respect to schemes based on stochastic time thanks to a lower sensitivity to the external biasing. For example, the application of an external magnetic field or change in temperature would only affect the switching threshold of the triangular pulse scheme, but not its cycle-to-cycle variability, which acts as the entropy source. On the other hand, for the rectangular pulse scheme, an external bias would change the voltage window for maximum entropy, requiring a re-tuning of the applied voltage.

9 PUF Implementations

The RRAM device variability sources discussed for TRNG can in principle be adopted for PUF systems, thus enabling a small area, low power consumption, and high PUF performance in terms of uniqueness and reliability. For instance, the stochastic resistance variation in RRAM was proposed for a reconfigurable PUF [86]. Fig. 24a shows the calculated lognormal distributions of RRAM resistance for LRS and HRS. Fig. 24b is a sketch of a PUF circuit consisting of an RRAM array where each cell represents a single bit and can be initialized in either LRS or HRS. The challenge consists of the address of two n-bit data, while the response is the bit-wise comparison of the RRAM resistance of the two data. In this PUF concept, the stochastic switching allows for the reconfiguration of the

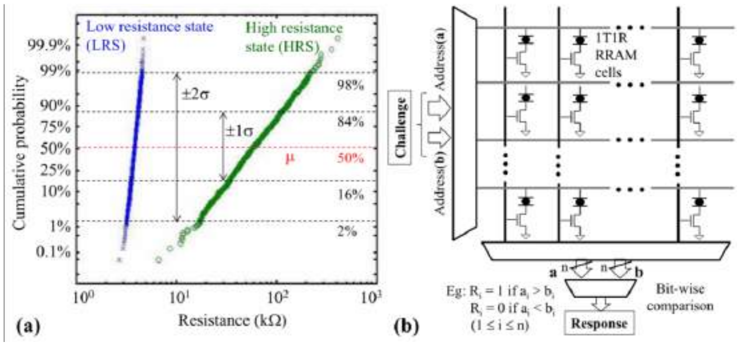

Figure 24: (a) Simulated resistance distributions for LRS and HRS, following normal and lognormal distributions, respectively. (b) Schematic illustration of a PUF implementation exploiting RRAM resistance variability. Reprinted with permission from [86]. Copyright (2014) IEEE.

PUF by reprogramming the RRAM array, in strike contrast with systems based on fixed manufacturing variations. PUF reconfigurability significantly enhances security protocols based on authentication [87], since it allows to overcome the limitations due to device degradation or small CRP set. Fig. 25 shows the characterization of the PUF against three of the performance parameters in Sec. II, namely unpredictability, unclonability and reliability. First, the unpredictability of the PUF response can be measured by studying the output bit uniformity. Fig. 25 shows the characterization of the PUF against three of the performance parameters in Sec. II, namely unpredictability, unclonability and reliability. First, the unpredictability of the PUF response can be measured by studying the output bit uniformity. Fig. 25a shows "1" bias distributions of 256-bit responses, thus supporting a uniform output, also confirmed by the almost equal probabilities of 3-bit responses in Fig. 25b. Second, the unclonability requires that the physical (or mathematical) CRP mapping cannot be replicated, which in turn requires a strong uniqueness of PUF to distinguish a specific chip from another. This property can be assessed as the Hamming distance (HD) between the responses of two different PUFs to the same challenge. It is also referred to as the inter-chip HD ($\mathrm{HD_{inter}}$), which should be ideally 50%. Fig. 25c shows the calculated $\mathrm{HD_{inter}}$ for 100 PUF samples of 256 kb RRAM arrays, demonstrating an ideal $\mathrm{HD_{inter}}$ close to 50%. Finally, reliability refers to the ability of a PUF of giving always the same response to a given challenge. To evaluate the PUF reliability, the intra-chip HD ($\mathrm{HD_{intra}}$) can be calculated in this case among different responses to the same challenge for the same PUF under different conditions (such as temperature). The $\mathrm{HD_{intra}}$ should be 0% for an ideal PUF, and a large separation between $\mathrm{HD_{inter}}$ and $\mathrm{HD_{intra}}$ reduces false identification rate [86]. $\mathrm{HD_{intra}}$ might be affected by the dependence of RRAM resistance on temperature and voltage. For instance, Fig. 25d shows the

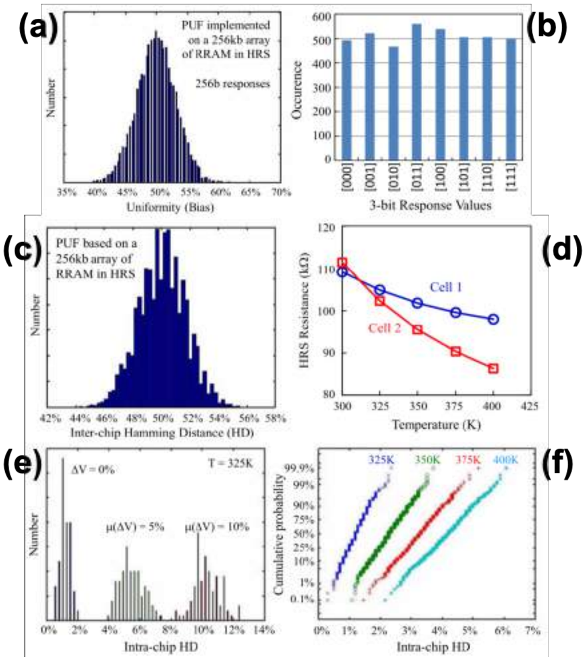

Figure 25: (a) Distribution of the uniformity measured by "1" bias of a PUF implemented on a 256 kbit array. The relatively uniform output is demonstrated by the uniform occurrence of the 3-bit responses (b). (c) Uniqueness measured by $\mathrm{HD}_{\mathrm{inter}}$ distribution. (d) A resistance crossing event between two different cells at increasing temperature, which causes a bit flipping and consequently a reliability degradation. (e) Effect on $\mathrm{HD}_{\mathrm{intra}}$ distributions under different voltage fluctuations. (f) $\mathrm{HD}_{\mathrm{intra}}$ distributions at different temperatures. Reprinted with permission from [86]. Copyright (2014) IEEE.

resistance as a function of temperature for two RRAM cells with two different activation energies [86]. Note the crossing between the two resistance values at high temperature, thus resulting in a bit flip and a consequent reduction of the reliability. Figs. 25e-f shows the impact of voltage and temperature on reliability, described by the parameter $\mathrm{HD}_{\mathrm{intra}}$. In general, PUF implementation with RRAMs requires that the spatial (*i.e.*, cell to cell) variability dominates over temporal variability (*i.e.*, noise) [86]. As a result, particular attention should be paid on device retention properties to minimize possible aging effects that might reduce the window between $\mathrm{HD}_{\mathrm{intra}}$ and $\mathrm{HD}_{\mathrm{inter}}$. To develop a strong PUF, not only the RRAM randomness and reliability, but also the circuit implementation of the response function should be robust enough. Fig. 26a shows a possible PUF implementation based on a crosspoint RRAM array [43]. Here, the entropy source is provided by the large analogue resistance distribution of the RRAM.

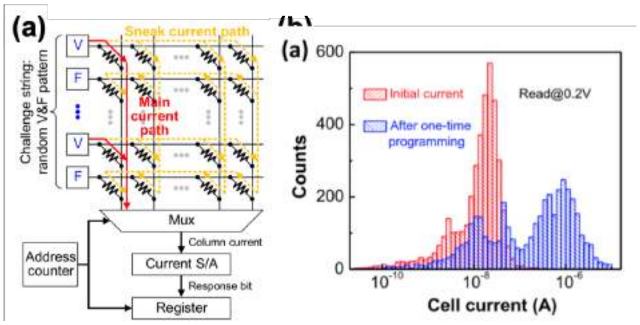

Figure 26: (a) Schematic illustration of the resistive crosspoint array, which implements a strong PUF by exploiting the sneak paths. (b) Distributions of cell current before and after the one-time programming, showing quite large analogue distribution. Reprinted with permission from [43]. Copyright (2016) IEEE.

The current sneak path is then exploited to go beyond the typical limitation of memory-based PUF, which have a limited set of CRPs. Note that for memory applications the sneak path effect is detrimental for cell read-out margin [88]. On the other hand, sneak path provides the unclonable function in this case, enabling an exponential scaling of the CRP set, which is required for a strong PUF. In the NxN crosspoint PUF of Fig. 26a, the challenge consists of a N-bit vector applied to the N rows, where an input bit value of 1 corresponds to an applied voltage equal to V_{DD}, while the row is left floating for a bit value of 0. The current from the N columns is then read and converted to an N-bit response by a sense amplifier. Theoretically, the maximum number of CRPs is 2^N, since each row may be either floating or with an applied voltage. The actual number of CRPs is reduced since 50% of the rows are required to be biased in order to generate a comparable range of column currents for different challenges [43]. It is estimated that CRP set is around 5 x 10^{75} for an array of 256 x 256 bits. RRAM devices in the array are initialized only at the beginning of the PUF operation, resulting in large cell current variability thanks to the variation in switching dynamics (Fig. 26b).

The performance of the crosspoint PUF in Fig. 26 is evaluated in terms of the experimental HD_{inter} and HD_{intra}. In particular, the uniqueness is evaluated by HD_{inter} by comparing the responses across 28 PUF instances. Fig. 27a shows HD_{inter} distributions for 11 different challenges. The average $HD_{inter} \approx 46.2\%$ is sufficiently close to the ideal 50%, thus demonstrating a good uniqueness. In addition, a good PUF reliability requires a sufficient retention of the array cell resistance state. To this purpose, the output currents (*i.e.* the responses) were measured as a function of time for increasing temperature. Fig. 27b reports the results of an annealing experiment for T = 120°C as a function of time, underlining the RRAM variation with time as already demonstrated

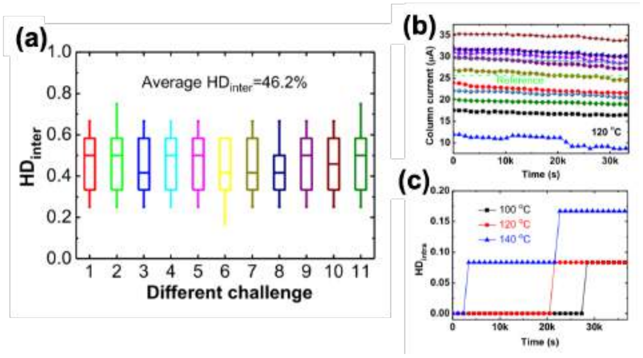

Figure 27: (a) Distributions of $\mathrm{HD}_{\mathrm{inter}}$ of 12-bit responses for 11 different input vectors. (b) Measured read current for 12 column as a function of time at T = 120°C. (c) $\mathrm{HD}_{\mathrm{intra}}$ of 12-bit responses to the same challenge as a function of time for three different temperatures T = 100, 120 and 140 °C. Reprinted with permission from [43]. Copyright (2016) IEEE.

for $\mathrm{HfO_x}$ RRAM [89]. The results are summarized in Fig. 27c as $\mathrm{HD}_{\mathrm{intra}}$ for increasing temperature T = 100, 120 and 140 °C, showing an increasing value of $\mathrm{HD}_{\mathrm{intra}}$, from 0% to 8%. Note that $\mathrm{HD}_{\mathrm{intra}}$ and $\mathrm{HD}_{\mathrm{inter}}$ distributions do not overlap, as the minimum for $\mathrm{HD}_{\mathrm{inter}}$ is around 17% (see Fig. 27a), demonstrating the feasibility of the crosspoint PUF concept as hardware security primitive. Embedding resistive devices in security primitives allows for their hardware reconfigurability, which opens new possibilities for secret keys management. A key-based permission granting system requires eventual key erasure, after the permissions have been revoked. This system allows for logic locking [90], which is used against intellectual property (IP) theft and circuit counterfeiting. However, proving that the digital key has been erased is a difficult task. More in general, a security protocol with erasable PUF responses is desirable [44].

Recently, a provable key destruction scheme based on memristive devices was demonstrated [91] with a 128x64 $\mathrm{Ta/HfO_2}$ crosspoint array, shown in Fig. 28a. The unclonable fingerprint is derived by comparing the conductance value of neighboring cell pairs in the array, after initializing all of them in the LRS. The random bit identifying each pair is set to "1" if $\mathrm{G}_{\mathrm{LRS,left}} \geq \mathrm{G}_{\mathrm{LRS,right}}$, to "0" otherwise. Owing to the intrinsic variability of LRS, a random pattern (*i.e.* the fingerprint)is generated to identify uniquely the device, as shown in (Fig. 28b). Fig. 29 shows the experimental demonstration of provable key destruction. Here, an initial fingerprint ($\mathrm{FP}_{\mathrm{chip}}$, Fig. 29a) is generated and securely stored in a trusted database. Then, a random key ($\mathrm{K}_{\mathrm{chip}}$, Fig. 29b) is written in the array, thus preventing the re-writing of $\mathrm{FP}_{\mathrm{chip}}$ without losing $\mathrm{K}_{\mathrm{chip}}$. $\mathrm{K}_{\mathrm{chip}}$ is also sent to the trusted party, so that it can be used for unlocking features of the specific chip instance storing $\mathrm{K}_{\mathrm{chip}}$. When a key erasure is necessary, the user simply reinitialises the array to the LRS, therefore destroying the key

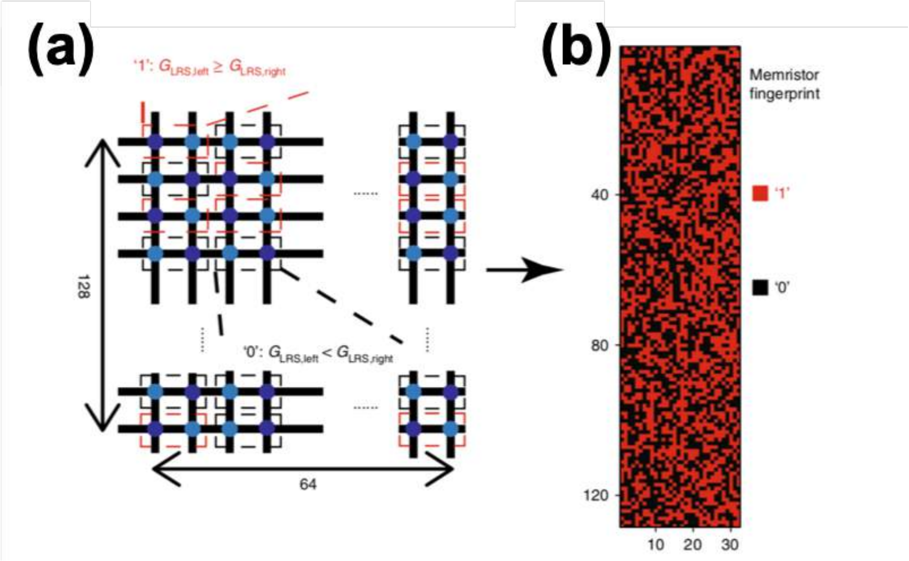
(a)
'1': $G_{LRS,left} \geq G_{LRS,right}$
'0': $G_{LRS,left} < G_{LRS,right}$
128
64
(b)
Memristor fingerprint
'1'
'0'
40
80
120
10
20
30

Figure 28: (a) Schematic of the crosspoint array enabling secure fingerprint extraction only after provable key erasure, where the fingerprint is given by the comparison of LRS conductance between two neighboring memristor cells. (b) Typical 128x32 fingerprint that can be generated from a 128x64 memristor array. Reprinted with permission from Macmillan Publishers Ltd: Nature Electronics [91]. Copyright (2018).

K_{chip} and generating a new fingerprint (FP'$_{chip}$, Fig. 29c), which constitutes the demonstration of key erasure. The new fingerprint FP'$_{chip}$ is finally sent to the trusted party for comparison with the previously stored FP_{chip}. If the HD between the two fingerprints is compatible with the expected distance between fingerprints of the same chip, then the chip can be authenticated by the trusted party. . In addition, the trusted party also gets confirmation that K_{chip} has been erased, since it is required for generating a valid FP'$_{chip}$. The practical feasibility of the described concept is demonstrated in Fig. 29d, showing that the distribution of HD for the same chip is clearly separated from the distribution of HD for different chips. Fig. 29e shows the same distributions for 256-bit fingerprint, where the improved separation between the two distribution supports the need for a large number of bits in the fingerprint.

10 Summary and Conclusions

The exponential increase of internet-based communication devices is raising the demand for data/hardware security. A severe challenge is the limited area and power for IoT devices, which spurs the research on low power, high performance

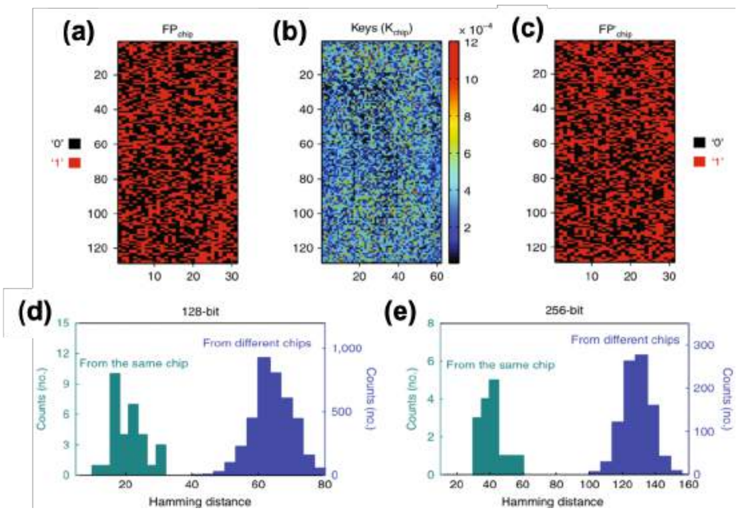

Figure 29: (a) Initial fingerprint $\mathrm{FP}_{\mathrm{chip}}$ stored by the trusted party. (b) Digital key $\mathrm{K}_{\mathrm{chip}}$ written in the memristor array. (c) A second fingerprint $\mathrm{FP'}_{\mathrm{chip}}$ generated by the same array, thus destroying the key. (d) HD distributions of 128-bit fingerprints from same chip and different chips, showing sufficient separation, hence demonstrating the feasibility of the scheme. (e) The same comparison is given for 256-bit fingerprints. Reprinted with permission from Macmillan Publishers Ltd: Nature Electronics [91]. Copyright (2018).

hardware security blocks such as TRNG and PUF. While TRNGs are essential for encryption adopted in data and transmission security, PUFs are becoming the preferred solution for hardware authentication and verification.

The chapter provides an overview of TRNGs and PUFs based on emerging resistive switching memory technology. We review the various schemes for using a nanoscale device as entropy source, including stochastic noise, stochastic switching delay time and stochastic switching voltage. The various implementations are discussed in terms of simplicity of the concept and the stability over various operating condition, such as process, voltage and temperature. The effectiveness of differential schemes for TRNGs, which do not require any probability tracking to tune the operating voltage and/or time, is also discussed and emphasized.

While the status of memory-based security primitives is already encouraging, there are still many challenges toward a practical implementation of these concepts in IoT and other integrated systems. In particular, device optimization needs to be focused on high-frequency operation ($>$ 1 Gbit/s), low-energy per bit (tens of fJ range), aggressive area scalability (1x nm node) and infinite endurance. Most importantly, a CMOS-compatible technology is paramount for an easy integration capability. The device should also engineered toward enhancing the stochastic behavior, which is generally unwanted and intentionally suppressed in memory applications. A differentiation of the device geometry, materials, and operation algorithms toward optimized random performance might be needed for TRNG and PUFs. From the circuital point-of-view, the research effort should focus on design solutions which minimize the area, power and circuit overhead. Clearly, this means that TRNG schemes which do not require any post-processing algorithm or entropy tracking feedback should be preferred. In general, a thorough device/circuit co-design methodology is extremely important and should be carefully explored. Finally, a fascinating direction of research is the hardware reconfigurability, where the same fundamental structure (*e.g.* a cross-point memory array) is used for either memory, computing (*e.g.* as a hardware primitive for stochastic/neuromorphic computing), or hardware-security. This offers new possibilities for ultra-small/low-power IoT-devices, which would be able to perform a wide range of tasks (*e.g.* pattern recognition and classification, fast/low-power analog computation, authentication, etc.) within a single hardware chip.

11 Acknowledgment

This work was supported by the European Research Council under Grant ERC-2014-CoG-648635-RESCUE.

References

[1] Jeyavijayan Rajendran, Ramesh Karri, James Bradley Wendt, Miodrag Potkonjak, Nathan R McDonald, Garrett S Rose, and Bryant T Wysocki.

Nanoelectronic solutions for hardware security. *IACR Cryptology ePrint Archive*, 2012:575, 2012.

[2] Christos Stergiou, Kostas E Psannis, Byung-Gyu Kim, and Brij Gupta. Secure integration of iot and cloud computing. *Future Generation Computer Systems*, 78:964–975, 2018.

[3] Charles Herder, Meng-Day Yu, Farinaz Koushanfar, and Srinivas Devadas. Physical unclonable functions and applications: A tutorial. *Proceedings of the IEEE*, 102(8):1126–1141, 2014.

[4] Min-Woo Ryu, Jaeho Kim, Sang-Shin Lee, and Min-Hwan Song. Survey on internet of things. *SmartCR*, 2(3):195–202, 2012.

[5] Kim-Kwang Raymond Choo, Mehran Mozaffari Kermani, Reza Azarderakhsh, and Manimaran Govindarasu. Emerging embedded and cyber physical system security challenges and innovations. *IEEE Transactions on Dependable and Secure Computing*, (3):235–236, 2017.

[6] Fatemeh Tehranipoor. Towards implementation of robust and low-cost security primitives for resource-constrained iot devices. *arXiv preprint arXiv:1806.05332*, 2018.

[7] Hussein Nili, Gina C Adam, Brian Hoskins, Mirko Prezioso, Jeeson Kim, M Reza Mahmoodi, Farnood Merrikh Bayat, Omid Kavehei, and Dmitri B Strukov. Hardware-intrinsic security primitives enabled by analogue state and nonlinear conductance variations in integrated memristors. *Nature Electronics*, 1(3):197, 2018.

[8] Sanu K Mathew, Suresh Srinivasan, Mark A Anders, Himanshu Kaul, Steven K Hsu, Farhana Sheikh, Amit Agarwal, Sudhir Satpathy, and Ram K Krishnamurthy. 2.4 gbps, 7 mw all-digital pvt-variation tolerant true random number generator for 45 nm cmos high-performance microprocessors. *IEEE Journal of Solid-State Circuits*, 47(11):2807–2821, 2012.

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

[10] Daniele Ielmini and H-S Philip Wong. In-memory computing with resistive switching devices. *Nature Electronics*, 1(6):333, 2018.

[11] J Joshua Yang, Dmitri B Strukov, and Duncan R Stewart. Memristive devices for computing. *Nature nanotechnology*, 8(1):13, 2013.

[12] Chip-Hong Chang, Yue Zheng, and Le Zhang. A retrospective and a look forward: Fifteen years of physical unclonable function advancement. *IEEE Circuits and Systems Magazine*, 17(3):32–62, 2017.

[13] Garrett S Rose. Security meets nanoelectronics for internet of things applications. In *Proceedings of the 26th edition on Great Lakes Symposium on VLSI*, pages 181–183. ACM, 2016.

[14] Swaroop Ghosh. Spintronics and security: Prospects, vulnerabilities, attack models, and preventions. *Proceedings of the IEEE*, 104(10):1864–1893, 2016.

[15] Armin Alaghi and John P Hayes. Survey of stochastic computing. *ACM Transactions on Embedded computing systems (TECS)*, 12(2s):92, 2013.

[16] Joseph S Friedman, Laurie E Calvet, Pierre Bessière, Jacques Droulez, and Damien Querlioz. Bayesian inference with Müller C-elements. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 63(6):895–904, 2016.

[17] Wolfgang Maass. Noise as a resource for computation and learning in networks of spiking neurons. *Proceedings of the IEEE*, 102(5):860–880, 2014.

[18] Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014.

[19] Giacomo Pedretti, Valerio Milo, Stefano Ambrogio, Roberto Carboni, Stefano Bianchi, Alessandro Calderoni, Nirmal Ramaswamy, Alessandro S Spinelli, and Daniele Ielmini. Stochastic learning in neuromorphic hardware via spike timing dependent plasticity with rram synapses. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 8(1):77–85, 2018.

[20] Gonzalo Alvarez and Shujun Li. Some basic cryptographic requirements for chaos-based cryptosystems. *International journal of bifurcation and chaos*, 16(08):2129–2151, 2006.

[21] Maxim Integrated. Pseudo random number generation using linear feedback shift registers, 2010, Retrieved from Maxim Integrated website: http://www.maximintegrated.com/an4400.

[22] John Von Neumann. Various techniques used in connection with random digits. *Appl. Math Ser*, 12(36-38):5, 1951.

[23] John Kelsey, Bruce Schneier, David Wagner, and Chris Hall. Cryptanalytic attacks on pseudorandom number generators. In *International Workshop on Fast Software Encryption*, pages 168–188. Springer, 1998.

[24] Suresh Chari, Charanjit Jutla, Josyula R Rao, and Pankaj Rohatgi. A cautionary note regarding evaluation of aes candidates on smart-cards. In *Second Advanced Encryption Standard Candidate Conference*, pages 133–147. Citeseer, 1999.

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

[26] Benjamin Jun and Paul Kocher. The intel random number generator. *Cryptography Research Inc. white paper*, 27:1–8, 1999.

[27] Shubham Sahay and Manan Suri. Recent trends in hardware security exploiting hybrid cmos-resistive memory circuits. *Semiconductor Science and Technology*, 32(12):123001, 2017.

[28] Ralf Brederlow, Ramesh Prakash, Christian Paulus, and Roland Thewes. A low-power true random number generator using random telegraph noise of single oxide-traps. In *Solid-State Circuits Conference, 2006. ISSCC 2006. Digest of Technical Papers. IEEE International*, pages 1666–1675. IEEE, 2006.

[29] Chien-Yuan Huang, Wen Chao Shen, Yuan-Heng Tseng, Ya-Chin King, and Chrong-Jung Lin. A contact-resistive-random-access-memory-based true-random-number generator. *IEEE Electron Device Letters*, 33(8):1108, 2012.

[30] Akio Fukushima, Takayuki Seki, Kay Yakushiji, Hitoshi Kubota, Hiroshi Imamura, Shinji Yuasa, and Koji Ando. Spin dice: A scalable truly random number generator based on spintronics. *Applied Physics Express*, 7(8):083001, 2014.

[31] Sungwoo Chun, Seung-Beck Lee, Masahiko Hara, Wanjun Park, and Song-Ju Kim. High-density physical random number generator using spin signals in multidomain ferromagnetic layer. *Advances in Condensed Matter Physics*, 2015, 2015.

[32] Z Wei, Y Katoh, S Ogasahara, Y Yoshimoto, K Kawai, Y Ikeda, K Eriguchi, K Ohmori, and S Yoneda. True random number generator using current difference based on a fractional stochastic model in 40-nm embedded reram. In *Electron Devices Meeting (IEDM), 2016 IEEE International*, pages 4–8. IEEE, 2016.

[33] Simone Balatti, Stefano Ambrogio, Zhongqiang Wang, and Daniele Ielmini. True random number generation by variability of resistive switching in oxide-based devices. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 5(2):214–221, 2015.

[34] Simone Balatti, Stefano Ambrogio, Roberto Carboni, Valerio Milo, Zhongqiang Wang, Alessandro Calderoni, Nirmal Ramaswamy, and Daniele Ielmini. Physical unbiased generation of random numbers with coupled resistive switching devices. *IEEE Transactions on Electron Devices*, 63(5):2029–2035, 2016.

[35] Sheng-hua Zhou, Wancheng Zhang, and Nan-Jian Wu. An ultra-low power cmos random number generator. *Solid-State Electronics*, 52(2):233–238, 2008.

[36] Eric Diehl. *Ten Laws for Security*. Springer, 2016.

[37] Jimson Mathew, Rajat Subhra Chakraborty, Durga Prasad Sahoo, Yuanfan Yang, and Dhiraj K Pradhan. A novel memristor-based hardware security primitive. *ACM Transactions on Embedded Computing Systems (TECS)*, 14(3):60, 2015.

[38] Paul Kocher, Joshua Jaffe, and Benjamin Jun. Differential power analysis. In *Annual International Cryptology Conference*, pages 388–397. Springer, 1999.

[39] Ravikanth Pappu, Ben Recht, Jason Taylor, and Neil Gershenfeld. Physical one-way functions. *Science*, 297(5589):2026–2030, 2002.

[40] Meng-Day Yu, Richard Sowell, Alok Singh, David M'Raïhi, and Srinivas Devadas. Performance metrics and empirical results of a puf cryptographic key generation asic. In *Hardware-Oriented Security and Trust (HOST), 2012 IEEE International Symposium on*, pages 108–115. IEEE, 2012.

[41] Le Zhang, Zhi Hui Kong, Chip-Hong Chang, Alessandro Cabrini, and Guido Torelli. Exploiting process variations and programming sensitivity of phase change memory for reconfigurable physical unclonable functions. *IEEE Transactions on Information Forensics and Security*, 9(6):921–932, 2014.

[42] Daniel E Holcomb, Wayne P Burleson, and Kevin Fu. Power-up sram state as an identifying fingerprint and source of true random numbers. *IEEE Transactions on Computers*, 58(9):1198–1210, 2009.

[43] Ligang Gao, Pai-Yu Chen, Rui Liu, and Shimeng Yu. Physical unclonable function exploiting sneak paths in resistive cross-point array. *IEEE Transactions on Electron Devices*, 63(8):3109–3115, 2016.

[44] U. Rührmair, J. Sölter, F. Sehnke, X. Xu, A. Mahmoud, V. Stoyanova, G. Dror, J. Schmidhuber, W. Burleson, and S. Devadas. Puf modeling attacks on simulated and silicon data. *IEEE Transactions on Information Forensics and Security*, 8(11):1976–1891, 2013.

[45] Arunkumar Vijayakumar and Sandip Kundu. A novel modeling attack resistant puf design based on non-linear voltage transfer characteristics. In *Proceedings of the 2015 Design, Automation & Test in Europe Conference & Exhibition*, pages 653–658. EDA Consortium, 2015.

[46] Rainer Waser and Masakazu Aono. Nanoionics-based resistive switching memories. *Nature materials*, 6(11):833, 2007.

[47] Hiroyuki Akinaga and Hisashi Shima. Resistive random access memory (reram) based on metal oxides. *Proceedings of the IEEE*, 98(12):2237–2251, 2010.

[48] H-S Philip Wong, Heng-Yuan Lee, Shimeng Yu, Yu-Sheng Chen, Yi Wu, Pang-Shiu Chen, Byoungil Lee, Frederick T Chen, and Ming-Jinn Tsai. Metal–oxide rram. *Proceedings of the IEEE*, 100(6):1951–1970, 2012.

[49] Daniele Ielmini. Resistive switching memories based on metal oxides: mechanisms, reliability and scaling. *Semiconductor Science and Technology*, 31(6):063002, 2016.

[50] Shimeng Yu, Hong-Yu Chen, Bin Gao, Jinfeng Kang, and H-S Philip Wong. Hfox-based vertical resistive switching random access memory suitable for bit-cost-effective three-dimensional cross-point architecture. *ACS nano*, 7(3):2320–2325, 2013.

[51] Haitong Li, Tony F Wu, Subhasish Mitra, and H-S Philip Wong. Resistive ram-centric computing: Design and modeling methodology. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 64(9):2263–2273, 2017.

[52] Seong-Geon Park, Min Kyu Yang, Hyunsu Ju, Dong-Jun Seong, Jung Moo Lee, Eunmi Kim, Seungjae Jung, Lijie Zhang, Yoo Cheol Shin, In-Gyu Baek, et al. A non-linear reram cell with sub-1μa ultralow operating current for high density vertical resistive memory (vrram). In *Electron Devices Meeting (IEDM), 2012 IEEE International*, pages 20–8. IEEE, 2012.

[53] Jun Yeong Seok, Seul Ji Song, Jung Ho Yoon, Kyung Jean Yoon, Tae Hyung Park, Dae Eun Kwon, Hyungkwang Lim, Gun Hwan Kim, Doo Seok Jeong, and Cheol Seong Hwang. A review of three-dimensional resistive switching cross-bar array memories from the integration and materials property points of view. *Advanced Functional Materials*, 24(34):5316–5339, 2014.

[54] Alessandro Bricalli, Elia Ambrosi, Mario Laudato, M Maestro, R Rodriguez, and Daniele Ielmini. Siox-based resistive switching memory (rram) for crossbar storage/select elements with high on/off ratio. In *Electron Devices Meeting (IEDM), 2016 IEEE International*, pages 4–3. IEEE, 2016.

[55] Daniele Ielmini. Modeling the universal set/reset characteristics of bipolar rram by field-and temperature-driven filament growth. *IEEE Transactions on Electron Devices*, 58(12):4309–4317, 2011.

[56] Stefano Larentis, Federico Nardi, Simone Balatti, David C Gilmer, and Daniele Ielmini. Resistive switching by voltage-driven ion migration in bipolar rram—part ii: Modeling. *IEEE Transactions on Electron Devices*, 59(9):2468–2475, 2012.

[57] Stefano Ambrogio, Simone Balatti, Antonio Cubeta, Alessandro Calderoni, Nirmal Ramaswamy, and Daniele Ielmini. Statistical fluctuations in hfo x resistive-switching memory: Part i-set/reset variability. *IEEE Transactions on electron devices*, 61(8):2912–2919, 2014.

[58] Stefano Ambrogio, Simone Balatti, Antonio Cubeta, Alessandro Calderoni, Nirmal Ramaswamy, and Daniele Ielmini. Statistical fluctuations in hfo x resistive-switching memory: Part ii—random telegraph noise. *IEEE Transactions on Electron Devices*, 61(8):2920–2927, 2014.

[59] Stefano Ambrogio, Simone Balatti, Vincent McCaffrey, Daniel C Wang, and Daniele Ielmini. Noise-induced resistance broadening in resistive switching memory—part i: Intrinsic cell behavior. *IEEE Transactions on Electron Devices*, 62(11):3805–3811, 2015.

[60] Stefano Ambrogio, Simone Balatti, Vincent McCaffrey, Daniel C Wang, and Daniele Ielmini. Noise-induced resistance broadening in resistive switching memory—part ii: Array statistics. *IEEE Transactions on Electron Devices*, 62(11):3812–3819, 2015.

[61] Daniele Ielmini, Federico Nardi, and Carlo Cagli. Resistance-dependent amplitude of random telegraph-signal noise in resistive switching memories. *Applied Physics Letters*, 96(5):053503, 2010.

[62] Y Yoshimoto, Y Katoh, S Ogasahara, Z Wei, and K Kouno. A reram-based physically unclonable function with bit error rate$<$ 0.5% after 10 years at 125° c for 40nm embedded application. In *VLSI Technology, 2016 IEEE Symposium on*, pages 1–2. IEEE, 2016.

[63] STS NIST. Special publication 800-22. *A statistical test suite for random and pseudorandom number generators for cryptographic applications*, 2010.

[64] Carlo Cagli, Federico Nardi, and Daniele Ielmini. Modeling of set/reset operations in nio-based resistive-switching memory devices. *IEEE Transactions on electron devices*, 56(8):1712–1720, 2009.

[65] Sung Hyun Jo, Ting Chang, Kuk-Hwan Kim, Siddharth Gaba, and Wei Lu. Experimental, modeling and simulation studies of nanoscale resistance switching devices. In *Nanotechnology, 2009. IEEE-NANO 2009. 9th IEEE Conference on*, pages 493–495. IEEE, 2009.

[66] Hao Jiang, Daniel Belkin, Sergey E Savel'ev, Siyan Lin, Zhongrui Wang, Yunning Li, Saumil Joshi, Rivu Midya, Can Li, Mingyi Rao, et al. A novel true random number generator based on a stochastic diffusive memristor. *Nature communications*, 8(1):882, 2017.

[67] Siddharth Gaba, Phil Knag, Zhengya Zhang, and Wei Lu. Memristive devices for stochastic computing. In *Circuits and Systems (ISCAS), 2014 IEEE International Symposium on*, pages 2592–2595. IEEE, 2014.

[68] Siddharth Gaba, Patrick Sheridan, Jiantao Zhou, Shinhyun Choi, and Wei Lu. Stochastic memristive devices for computing and neuromorphic applications. *Nanoscale*, 5(13):5872–5878, 2013.

[69] Sung Hyun Jo, Kuk-Hwan Kim, and Wei Lu. Programmable resistance switching in nanoscale two-terminal devices. *Nano letters*, 9(1):496–500, 2008.

[70] Takeo Ohno, Tsuyoshi Hasegawa, Tohru Tsuruoka, Kazuya Terabe, James K Gimzewski, and Masakazu Aono. Short-term plasticity and long-term potentiation mimicked in single inorganic synapses. *Nature materials*, 10(8):591, 2011.

[71] Zhongrui Wang, Saumil Joshi, Sergey E Savel'ev, Hao Jiang, Rivu Midya, Peng Lin, Miao Hu, Ning Ge, John Paul Strachan, Zhiyong Li, et al. Memristors with diffusive dynamics as synaptic emulators for neuromorphic computing. *Nature materials*, 16(1):101, 2017.

[72] Alessandro Bricalli, Elia Ambrosi, Mario Laudato, Marcos Maestro, Rosana Rodriguez, and Daniele Ielmini. Resistive switching device technology based on silicon oxide for improved on–off ratio—part ii: Select devices. *IEEE Transactions on Electron Devices*, 65(1):122–128, 2018.

[73] Rivu Midya, Zhongrui Wang, Jiaming Zhang, Sergey E Savel'ev, Can Li, Mingyi Rao, Moon Hyung Jang, Saumil Joshi, Hao Jiang, Peng Lin, et al. Anatomy of ag/hafnia-based selectors with 1010 nonlinearity. *Advanced Materials*, 29(12):1604457, 2017.

[74] Stefano Ambrogio, Simone Balatti, Seol Choi, and Daniele Ielmini. Impact of the mechanical stress on switching characteristics of electrochemical resistive memory. *Advanced Materials*, 26(23):3885–3892, 2014.

[75] Roberto Carboni, Wei Chen, Manzar Siddik, Jon Harms, Andy Lyle, Witold Kula, Gurtej Sandhu, and Daniele Ielmini. Random number generation by differential read of stochastic switching in spin-transfer torque memory. *IEEE Electron Device Letters*, 2018.

[76] Won Ho Choi, Yang Lv, Jongyeon Kim, Abhishek Deshpande, Gyuseong Kang, Jian-Ping Wang, and Chris H Kim. A magnetic tunnel junction based true random number generator with conditional perturb and real-time output probability tracking. In *Electron Devices Meeting (IEDM), 2014 IEEE International*, pages 12–5. IEEE, 2014.

[77] Andrea Fantini, Ludovic Goux, Robin Degraeve, DJ Wouters, N Raghavan, G Kar, Attilio Belmonte, Y-Y Chen, Bogdan Govoreanu, and Malgorzata Jurczak. Intrinsic switching variability in hfo 2 rram. In *Memory Workshop (IMW), 2013 5th IEEE International*, pages 30–33. IEEE, 2013.

[78] Roberto Carboni, Stefano Ambrogio, W Chen, M Siddik, J Harms, A Lyle, W Kula, G Sandhu, and Daniele Ielmini. Understanding cycling endurance in perpendicular spin-transfer torque (p-stt) magnetic memory. In *Electron Devices Meeting (IEDM), 2016 IEEE International*, pages 21–6. IEEE, 2016.

[79] Janusz J Nowak, Ray P Robertazzi, Jonathan Z Sun, Guohan Hu, Jeong-Heon Park, JungHyuk Lee, Anthony J Annunziata, Gen P Lauer, Raman Kothandaraman, Eugene J O'Sullivan, et al. Dependence of voltage and size on write error rates in spin-transfer torque magnetic random-access memory. *IEEE Magnetics Letters*, 7:1–4, 2016.

[80] Dmytro Apalkov, Bernard Dieny, and JM Slaughter. Magnetoresistive random access memory. *Proceedings of the IEEE*, 104(10):1796–1830, 2016.

[81] Adrien F Vincent, Nicolas Locatelli, Jacques-Olivier Klein, Weisheng S Zhao, Sylvie Galdin-Retailleau, and Damien Querlioz. Analytical macrospin modeling of the stochastic switching time of spin-transfer torque devices. *IEEE Transactions on Electron Devices*, 62(1):164–170, 2015.

[82] Z Li and S Zhang. Thermally assisted magnetization reversal in the presence of a spin-transfer torque. *Physical Review B*, 69(13):134416, 2004.

[83] Damir Vodenicarevic, Nicolas Locatelli, Alice Mizrahi, Joseph S Friedman, Adrien F Vincent, Miguel Romera, Akio Fukushima, Kay Yakushiji, Hitoshi Kubota, Shinji Yuasa, et al. Low-energy truly random number generation with superparamagnetic tunnel junctions for unconventional computing. *Physical Review Applied*, 8(5):054045, 2017.

[84] Alice Mizrahi, Nicolas Locatelli, Romain Lebrun, Vincent Cros, Akio Fukushima, Hitoshi Kubota, Shinji Yuasa, Damien Querlioz, and Julie Grollier. Controlling the phase locking of stochastic magnetic bits for ultra-low power computation. *Scientific reports*, 6:30535, 2016.

[85] R Heindl, William H Rippard, Stephen E Russek, Matthew R Pufall, and Anthony B Kos. Validity of the thermal activation model for spin-transfer torque switching in magnetic tunnel junctions. *Journal of Applied Physics*, 109(7):073910, 2011.

[86] An Chen. Utilizing the variability of resistive random access memory to implement reconfigurable physical unclonable functions. *IEEE Electron Device Letters*, 36(2):138–140, 2015.

[87] Klaus Kursawe, Ahmad-Reza Sadeghi, Dries Schellekens, Boris Skoric, and Pim Tuyls. Reconfigurable physical unclonable functions-enabling technology for tamper-resistant storage. 2009.

[88] Jiantao Zhou, Kuk-Hwan Kim, and Wei Lu. Crossbar rram arrays: Selector device requirements during read operation. *IEEE Transactions on Electron Devices*, 61(5):1369–1376, 2014.

[89] Y. Y. Chen, M. Komura, R. Degraeve, B. Govoreanu, L. Goux, A. Fantini, N. Raghavan, S. Clima, L. Zhang, A. Belmonte, A. Redolfi, G. S. Kar, G. Groeseneken, D. J. Wouters, and M. Jurczak. Improvement of data retention in hfo2/hf 1t1r rram cell under low operating current.

[90] Yang Xie and Ankur Srivastava. Mitigating sat attack on logic locking. In Benedikt Gierlichs and Axel Y. Poschmann, editors, *Cryptographic Hardware and Embedded Systems – CHES 2016*, pages 127–146, Berlin, Heidelberg, 2016. Springer Berlin Heidelberg.

[91] Hao Jiang, Can Li, Rui Zhang, Peng Yan, Peng Lin, Yunning Li, J. Joshua Yang, Daniel Holcomb, and Qiangfei Xia. A provable key destruction scheme based on memristive crossbar arrays. *Nature Electronics*, 1(10):548–554, Oct 2018.