ChipSecure: A Reconfigurable Analog eFlash-Based PUF with Machine Learning Attack Resiliency in 55nm CMOS

Mohammad Mahmoodi, Hussein Nili, Shabnam Larimian, Xinjie Guo, Dmitri Strukov
ECE Department, UC Santa Barbara, Santa Barbara, CA, 93106-5630, USA
{mrmahmoodi,strukov}@ucsb.edu

ABSTRACT

We exploit randomness in static I - V characteristics and reconfigurability of embedded flash memories to design very efficient physically unclonable function. Leakage current and subthreshold slope variations, nonlinearity, nondeterministic tuning error, and sneak path current in the redesigned commercial flash memory arrays are exploited to create a unique digital fingerprint. A time-multiplexed architecture is designed to enhance the security and expand the challenge-response pair space to 10^{21} . Experimental results demonstrate 50.3% average uniformity, 49.99% average diffuseness, and native <5% bit error rate. The analysis of the measured data also shows strong resilience against machine learning attacks and possibility for extremely energy efficient, 0.56 pJ/b operation.

KEYWORDS

Floating Gate Memory, PUF, Security Primitive

ACM Reference format:

M. R. Mahmoodi, H. Nili, S. Larimian, X. Guo, and D. B. Strukov 2019. ChipSecure: A Reconfigurable Analog eFlash-based PUF with Machine Learning Attack Resiliency in 55nm CMOS. In *Proceedings of DAC '19, June 2–6, 2019, Las Vegas, NV, USA*.

1 Introduction

Physically unclonable function (PUF) is a new family of security primitives based on physical hardware roots of trust that utilizes the underlying unique and random process-induced variation properties of a chip (that are unpredictable, inimitable, and impossible to learn, extrapolate, or reproduce) [1]. Silicon PUFs are physically identical at design time, but, after fabrication, produce distinct, die-specific responses to input requests [2]. Security primitives based on unique process-induced variations have been utilized in authentication and key generation applications. For the former, a unique and unclonable fingerprint,

while for the latter, an unpredictable reproducible random number generation are needed [3].

Variety of PUFs have been proposed so far, most relying on CMOS circuits and techniques including current mirrors [4], SRAM and pre-charged cross-coupled inverters [5,6], ring oscillators [7] and subthreshold biased MOSFETs [8]. Emerging technologies such as resistive random-access memories [9,10] and spin-torque transfer magnetic random-access memories [3] have also offered excellent functional and physical performance prospects for the next generation of hardware security primitives.

Most existing PUFs utilize linear mapping function and/or a single-layer architecture [5], which make them vulnerable to machine learning attacks. In addition, many proposed approaches, e.g., SRAM-based PUFs, exhibit a large bit error rate (BER) requiring extensive post-processing and error correction [4]. A number of more advanced designs have been recently proposed to address these issues [1, 3, 7, 8]. The general idea behind the improvement is to follow Shannon's security paradigms [11], i.e., to utilize nonlinear elements to ensure better robustness and response diffuseness.

In this paper, we introduce an architecture that exploits intrinsic process-induced variations in analog properties of flash memories and inevitable uncertainty at memory state tuning process, which creates an explicit randomness in their extremely nonlinear subthreshold characteristics. Furthermore, to improve diffuseness, our circuit is specifically designed to minimize correlation between input bits of challenges and responses. Finally, in our approach, programmability of flash memories allows for implementing different unique PUF instances using the same hardware, which could be an important factor in many applications.

2 Entropy Sources in Flash Memory

Fig. 1 shows the cross-section of a split-gate NOR flash memory device, which has been recently proposed for low-power analog computing as a promising replacement for bulky synaptic transistors [12, 13]. Commercial NOR flash memories have a compact footprint ($\sim 25 F^2$) and are highly optimized in terms of endurance and retention for digital applications. The redesigned routing of erase-gate lines facilitated precise individual cell programming (Fig. 1b) and enabled an extremely energy-efficient analog-domain vector-by-matrix multiplication [14], the most common operation in many signal processing and neuromorphic algorithms.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

DAC.COM '19, June 2–6, 2019, Las Vegas, NV, USA

© 2019 Copyright is held by the owner/author(s).

ACM ISBN 978-1-4503-6737-0/19/06.

DOI: 10.1145/3316781.3324890

Fig. 1: (a) TEM image of SST's 55nm flash supercell used as the main entropy source in the proposed design. (b) I - V characteristics of the device in weak inversion (at $V_{WL}=1.2$ V, $V_{BL}=1.2$ V, and $V_{SL}=0$ V) tuned with 1% accuracy at different states.

Process variations in purely CMOS analog circuits often limit computation accuracy and result in large performance overheads, e.g., due to over-designing and calibration techniques. On the other hand, analog circuits with integrated dense embedded NOR flash memories allow to dramatically reduce such overhead, and, e.g., implement very efficient analog and mixed-signal neuromorphic circuits [12,15]. In the proposed PUF design, we capitalize on the superior performance of analog circuits enriched with nonvolatile memories, and directly harness memory cells' variations to ensure unique digital fingerprints.

Uncontrollable etching process leads to variations in gate oxides' thicknesses and line-edge roughness, which in turn result in two major types of device-to-device variations in I - V characteristics, namely variations in subthreshold drain current with respect to gate voltage and drain voltage (due to drain-induced barrier lowering). The former can be conveniently decomposed into three separate currents due to gate voltage independent leakage, weak inversion, and higher order effects. For example, Fig. 2a shows these three components by fitting $\log(I)$ expression of the measured subthreshold currents for 150 devices to a quadratic function of V_{CG} . Fig. 2b,c shows more details on distribution of leakage currents (at $V_{CG}=0$ V) for different values of V_{DS} , in particular highlighting a significant spread in leakage current distribution. (For all these experiments, memory cells have been tuned with better than 1% accuracy to eliminate the impact of tuning error in the measurements.)

Programming the state of analog-grade nonvolatile memories is typically based the write-verify algorithm, which has limited tuning accuracy (e.g., due to noise floor of the readout circuitry). Such tuning error can be used as an additional source of randomness. Fig. 2d, for example, shows the distribution of tuning error for 100 devices programmed randomly between 30 nA and 5 μ A with a specified targeted accuracy.

Fig. 2 provides extensive evidence of the variations and randomness in an array of programmed analog-grade flash memories. We exploit these entropy sources to design "ChipSecure" - a low-power and dense PUF instance with excellent uniformity and security. We first discuss the primitive block structure and detailed measurements, and then follow with the top-level architecture and its results.

Fig. 2: Major entropy contributors in ChipSecure: (a) the results of fitting $\log(I)$ of the subthreshold drain currents to $C_1 V_{CG}^2 + C_2 V_{CG} + C_3$, highlighting significant variations of leakage component (C_3), weak-inversion slope (C_2), and higher order nonlinearities (C_1). Box plots show 25%, 50%, and 75% quantiles. (b, c) Leakage current variations for the same 150 devices as a function of V_{DS} for two representative states (1 μ A and 5 μ A @ $V_{CG}=V_{WL}=V_{DS}=1$ V), showing the nonlinear semi-exponential dependence of the leakage current on V_{DS} , with more prominent variations at larger biases. (d, e) Distribution of tuning accuracy for 100 devices programmed with (d) 5% and (e) 1% targeted accuracy to random states within representative dynamic range. (Note that the shape of the distribution could be adjusted, if needed, with a more rigorous tuning procedure.)

3 Primitive Block

The primitive building block is designed to exploit randomness in a network of nonlinear elements (Fig. 3). Guided by Shannon's security paradigm, the devices are operated at deep subthreshold with exponential drain-source dependency so that the network is comprised of nonlinear computational elements, with all of them contributing to the output response. The idea is similar to the one in [1], in which sneak path currents are circulated in a crossbar of passively integrated memristors to build a compact security primitive. However, memristor fabrication technology is still in

need of improvement to enable building practical security systems. Commercial flash memories, on the other hand, are already embedded in high-end CMOS process technologies and are excellent candidate for low-power operation.

A. Operation Principle

A primitive block consists of $N \times (M-N)$ array of floating gate cells and peripheral (switching) circuitry for selecting cells and reading differential current (Fig. 3). The block is fed with L -bit input challenge to generate 1-bit output response. The input bits are partitioned into three groups: the first N bits of the challenge encode the selected word- (WLs) and control-gate (CG) lines; the next $M-N$ bits are used to specify the selected bit-lines (BLs); while the remaining bits encode the selected source lines (SLs). (Note that each SL is shared between two row of cells.) The selected CGs, WLs, and SLs are biased with $V_{CG,SEL}$, $V_{WL,SEL}$, and V_{SL} , respectively, while the selected BLs are grounded. Unselected CGs, BLs, and SLs are floated, while the unselected WLs are biased with $V_{WL,US}$. The response bit is generated by comparing output currents.

For the considered approach, cells in the array are categorized into four groups (inset of Fig. 3). For selected devices and half-selected type-A devices, SL is effectively a drain, while BL is a source. For the half-selected type-B devices, the current flows from BL to SL. The unselected devices can conduct current at either direction enabling circulation of sneak-path current in the array.

In all cases, 1s/0s of an input bit-vector directly specify position of the selected/unselected lines. This simplifies implementation of peripheral circuitry by requiring only one MOS transistor switch per line. Current sensing and comparison are implemented similar to [16]. It is also worth mentioning that differential scheme allows for better noise immunity and power supply rejection ratio as compared to single-ended designs.

In the balance of the paper, we show that such design features are essential for implementing high performance, robust, and resilient PUF.

B. Block Measurement Results

Our focus is on 10×10 primitive block (i.e. with $N = 10$, $M = 20$, and $L = 25$), in which each challenge selects 5 WLs/CGs, 5 BLs, and 2 SLs. Furthermore, we consider Gaussian-distributed states (currents) of memory cells in the array [1]. (Finding the optimum distribution for cell currents, during tuning and under nominal biasing conditions, is an important future work.)

To characterize block performance, we have fabricated 10×10 memory arrays in 55 nm embedded CMOS based on redesigned layout [14]. Keysight B1500A and B1530A tools and a custom-made switch matrix were utilized for characterization, programming, and measurements.

Fig. 4a shows an example of a current map after tuning the array with 10% accuracy to the randomly generated distribution with $\mu=500$ nA and $\sigma=150$ nA. Due to reconfigurability of our approach, a completely different map, i.e., a new fingerprint, is obtained after re-tuning the same physical array to a new

Fig. 3: ChipSecure primitive PUF block.

Fig. 4: Primitive block measurements at room temperature: (a,b) two examples of resultant map of conductance states in 10×10 array of cells, (c) the distribution of read-out currents for 3000 cases, and (d) the corresponding distribution of differential current.

distribution with $\mu=7.5$ μ A and $\sigma=1.5$ μ A (Fig. 4b). Fig. 4c and d shows, respectively, the measured read-out current distribution (I_R and I_L) and their difference for the PUF instance (with $V_{WL,SEL}=1.25$ V, $V_{WL,US}=1.35$ V, $V_{CG,SEL}=0.3$ V, and $V_{SL}=0.1$ V) corresponding to Fig. 4a. The similar shapes of distributions indicate that there is no explicit bias in the output. The corresponding uniformity is 52.6%, which is very close to the ideal, 50% value. The lack of bias is also confirmed by the data in Fig. 5, which shows that the output response is balanced with respect to the selected line in the array, i.e. value of '1' at certain position in the challenge bit-vector.

Furthermore, we have measured the response uniformity of 12 different primitive block using 4 different silicon chips. For each primitive, we employed the same tuning procedure (Gaussian distribution with 10% targeted accuracy) but with different common-mode current, using mean values randomly picked from 200 nA to 5 μ A range. Also, we have studied sensitivity of uniformity metric to biasing condition. This was done by selecting appropriate V_{WL} from 0.65 V to 1.35 V, V_{SL} from 0.1 V to 0.5 V, and V_{CG} from 0.1 V to 0.5 V to match the selected common-mode currents for each instance. For each block, 4K randomly selected challenges have been applied and the response was measured at

room temperature. The experimental results show again close to 50% uniformity for majority of the considered instances.

Perhaps, the most interesting result is the dependence of BER on the utilized common-mode current (Fig. 7). This experiment was performed using 5 different block instances with specified current-mode currents. Each primitive block was characterized by measuring responses to 1K challenges at different ambient temperatures (with $\pm 5^\circ\text{C}$ accuracy) and nominal voltage deviations. The results show that increasing temperature above the nominal 25°C , at which devices were tuned, results in a semi-quadratic increase of BER, while the reliability is always improved by operating at higher bias currents. This is most likely due to the weaker temperature dependency at larger subthreshold currents. Indeed, the currents are almost independent of the temperature in strong inversion (which can be used to build a temperature insensitive current-reference [17]). Therefore, there is a clear trade-off between power consumption and BER, and, e.g., the desired operating point could be determined based on the power budget and BER requirements of the PUF application. The same trend in BER is also observed with respect to the variations on the biased SL voltage (Fig. 7b), though the dependence is weaker.

Fig. 5: Measured correlation (fraction of 1s in the response when particular bit at the input is selected) based on 4K random challenge-response pairs.

Fig. 6: Response uniformity of 12 different primitive blocks obtained from reprogramming 4 different chips.

Fig. 7: Measured BER as function of (a) temperature at nominal SL readout voltage for several common-mode readout currents and (b) bias voltage deviation for different common-mode currents at room temperature.

4 ChipSecure Architecture

The ChipSecure architecture consists of two layers of primitive blocks connected via hidden shift register (HSR) (Fig. 8). An input challenge C is a 1010-bit vector. It is partitioned into subvectors $C_1 C_2 \dots C_{42}$, such that C_i s with $1 \leq i \leq 40$ are 25-bit long, while C_i s with $41 \leq i \leq 42$ are comprised of 5 bits. The C_i s from the first group are used as an input to five primitive blocks (P_{1-5}) in the first layer. The remaining two subvectors are used to select SLs of the two primitive blocks ($P_{6,7}$) in the second layer.

ChipSecure operation is time multiplexed with the help of control signals ϕ_{L1} , ϕ_{L2} , and ϕ_C . Specifically, ϕ_{L1} controls the first layer to generate and fill in HSR with 40 response bits over 8 cycles of operation, with a total of 5 bits produced in each cycle by five corresponding primitive blocks (e.g., D_{1-5} in the first cycle in Fig. 8). Using ϕ_{L1} , HSR bits are applied to WL/CG and BL of P_{6-7} to generate outputs R_1 and R_2 . These outputs are then XORed to generate the final response (R_F) and reduce any undesirable bias.

Assuming the delay of each block is t_d , a simple pipelining of the design would result in $1/8t_d$ (bit per second) throughput. Based on prior work [16], a comparator consumes 1.625 fJ per conversion with 1 nA sensitivity in 55 nm CMOS, which is equivalent to 192.3 Mbps throughput.

The proposed time-multiplexing approach enables low area implementation with enormous number of CRPs, which is a necessary condition for a very secure PUF, at the expense of lower throughput and decreased energy-efficiency. Indeed, for the proposed scheme, the maximum number of distinct selections for P_{1-6} is $S_P = \binom{10}{5} \binom{10}{5} \binom{5}{2}$. Note that all primitive blocks in the first layer contribute equally (4 bits) to the 20 bits needed in $P_{6,7}$. The total number of distinct selections is given by

$$\binom{S_P}{8} \binom{5}{2}^2 \approx 10^{211}.$$

It is worth mentioning the benefits of using smaller arrays, which are essential for time-multiplexed implementation. First, smaller arrays are more efficient for utilizing the sneak path currents, because leakage currents in larger arrays would be mostly controlled by a relatively small fraction of semi-selected

Fig. 8: ChipSecure architecture and example of timing diagram.

cells. Second, undesired stuck-on device in an array can bias the PUF response and potentially make the circuit vulnerable to probing attacks. Clearly, such stuck-on devices can be mitigated more efficiently when using smaller arrays.

5 ChipSecure Results

The ChipSecure was characterized using primitive blocks reported in Section 3. The measured currents were used in post-layout simulation setup of CMOS circuits. The fully integrated design occupies $1.3 \times 1.0 \text{ mm}^2$. It is dominated by low-voltage (0.3 mm^2) and high-voltage (0.1 mm^2) IOs and unused silicon ($\sim 0.9 \text{ mm}^2$). Active circuits, including programming circuitry ($4475 \mu\text{m}^2$), flash memory array ($235 \mu\text{m}^2$), registers ($19,250 \mu\text{m}^2$), comparators ($150 \mu\text{m}^2$), and logic ($110 \mu\text{m}^2$) are very compact (total of $24,216 \mu\text{m}^2$).

Fig. 9a shows that the measured uniformity of HSR bits is near-ideal for $P_{1,2,4}$ blocks, though there is visible bias in $P_{3,5}$ responses. Despite that, the differential current distribution of $P_{6,7}$ looks symmetrical as shown in Fig. 9b for P_6 . (Here, P_6 was tuned using 500 nA average state current and operated at $V_{WL,SEL}=0.85 \text{ V}$, $V_{WL,US}=0.9 \text{ V}$, $V_{CG,SEL}=0.3 \text{ V}$, and $V_{SL}=0.3 \text{ V}$.) Interestingly, the measured correlations, based on 100K challenge response pairs, are much weaker, as compared to those for single primitive block. The randomness in the output response is also highlighted by 2D visual representation of 1K randomly selected 128-bit keys (Fig. 11).

Fig. 12a shows the fractional Hamming weight distribution for 5K 64-bit and 128-bit keys generated based on R_1 , R_2 , and R_F . Based on these results, ChipSecure offers the average uniformity of 50.3%. (The measured uniformity for a larger set with 100K responses was 50.9%, 52.0%, and 50.3% for R_1 , R_2 , and R_F , respectively.) Figure 12b shows near optimal results for diffuseness - the other important metric that evaluates the difference (Hamming distance) between unique keys generated by the same PUF under different challenges [18].

As mentioned before, we can tune the devices to achieve a certain common-mode current and subsequently, a desirable BER. Based on Fig. 7, the native BER of $<5\%$ can be achieved with $\sim 30 \mu\text{A}$. In this case, the energy efficiency is 0.56 pJ per bit, with $88\% / 12\%$ contributed by array /comparators. Several previously proposed post-processing and error correction methods [5] can be utilized in our design to further improve reliability.

Machine learning attack resiliency was assessed with a multi-layer perceptron network ($1010 \times 100 \times 100 \times 100 \times 1$) with rectified linear activation function in the hidden layers and a sigmoid activation function in the output layer. We used RMSprop as optimizer and manually-found semi-found optimal learning rate of 0.001 . 64% and 16% of CRPs were used for training and validating the network, respectively. The classifier was trained with a specific size subset of the observed challenge-response pairs and then tested on another mutually exclusive data. The test accuracy as a function of training set size is provided in Fig. 13, which indicates close to ideal 50% prediction accuracy. We have also tried both networks (LIBSVM and LIBLINEAR) introduced in [6] and for both, the validation accuracy was also close to 50% .

NIST randomness test suite has been also utilized to evaluate the statistical properties of 100K measured data (Fig. 14). The results show that the bitstreams generated by ChipSecure pass all relevant tests. Finally, Table I shows a favorable quantitative comparison between ChipSecure and the state-of-the-art implementations.

Fig. 9: (a) Normalized Hamming weight of HSR bits over 100K applied challenges. (b) Differential current distribution of P_6 , with the inset showing the corresponding CDF.

Fig. 10: Measured correlation based on 100K random challenge-response pairs.

Fig. 11: 2D representation of 1K 128-bit keys (black='1').

Fig. 12: (a) Fractional Hamming weight and (b) fractional Hamming distance distribution of R_1 , R_2 , and R_F . The results were computed based on (a) 5K and (b) 1K randomly generated 64-bit and 128-bit keys.

6 Conclusion

In this paper, we have presented a design of physically unclonable function that offers average uniformity of 50.3%, average diffusivity of 49.99% and features 0.56 pJ/b energy efficiency and <5% BER. Using a time-multiplexing technique in 55nm eFlash CMOS process, we proposed a PUF with 10^{211} challenge-response pair space in only 24,216 μm^2 . ChipSecure passed all relevant NIST randomness tests and showed strong resiliency toward machine learning attacks.

ACKNOWLEDGEMENTS

This work was supported by SRC JUMP CRISP program and DARPA's UPSIDE program via BAE Systems, Inc.

Fig. 13: Modeling attack by an MLP network as a function of training set size (shown in bits). Training is performed with RMSdrop optimizer with the learning rate of 0.001. For all data points, the validation set size is 25% of training size. The inset shows the learning curve for 2K training data over 150 epochs.

Fig. 14: Result of NIST randomness test using 100K measured responses based on 50 2K-length bitstreams.

Fig. 15: (a) The die photo of a primitive block fabricated in GF's 55nm CMOS process and (b) the measurement setup.

Table. I: Comparison with previous works.

	VLSI'17 [6]	VLSI'17 [8]	ISSCC'15 [7]	ISSCC'15 [4]	ISSCC'14 [5]	This work
Technology	FDSOI (28nm)	CMOS (130nm)	CMOS (40nm)	CMOS (65nm)-SA	CMOS (22nm)	eFlash CMOS (55nm)
Number of CRPs	~10 ¹¹	~10 ¹⁹	~10 ²⁸	NA	~10 ⁶	~10 ²¹¹
Worst-case BER*	~11%	9%	9%	~21%	~30%	5%
Area (μm ²) **	~1.1 k	~44 k	~0.85 k	~623 k	~611 k	~24.2 k
Energy Efficiency (pJ/ b)	0.097	11	17.7	0.163	0.190	0.56
Temperature Range (°C)	0~80	-20~80	-25~125	25~85	25~50	25~85
NIST Test		NA	NA	PASS	FAIL	PASS
ML Attack	Prediction Error	15%	40%	NA	NA	~50%

(*without applying correction methods, ** Active PUF area, excluding IOs and BIST.)

REFERENCES

- [1] H. Nili *et al.*, "Hardware-intrinsic security primitives enabled by analogue state and nonlinear conductance variations in integrated memristors," *Nature Electronics*, vol.3, pp. 197-202, 2018.
- [2] A. Neale and M. Sachdev, "A low energy SRAM-based physically unclonable function primitive in 28 nm CMOS," in: *Proc. CICC'15*, San Jose, CA, Sept. 2015, pp. 1-4.
- [3] D. Y. Wang *et al.*, "Hardware implementation of physically unclonable function (PUF) in perpendicular STT MRAM," in: *Proc. VLSI-TSA'17*, 2017, pp. 1-2.
- [4] A. Alvarez *et al.*, "15fJ/b static physically unclonable functions for secure chip identification with <2% native bit instability and 140x inter/intra PUF Hamming distance separation in 65nm," in: *Proc. ISSCC'15*, San Francisco, CA, Feb. 2015, pp. 1-3.
- [5] S. Mathew *et al.*, "A 0.19 pJ/b PVT-variation-tolerant hybrid physically unclonable function circuit for 100% stable secure key generation in 22nm CMOS," in: *Proc. ISSCC'14*, San Francisco, CA, 2014, pp. 278-279.
- [6] S. Jeloka *et al.*, "A sequence dependent challenge-response PUF using 28nm SRAM 6T bit cell," in: *Proc. VLSI-Symp'17*, Kyoto, Japan, Jun. 2017, pp. 270-271.
- [7] K. Yang *et al.*, "A physically unclonable function with BER < 10^{-8} for robust chip authentication using oscillator collapse in 40nm CMOS," in: *Proc. ISSCC'15*, San Francisco, CA, Feb. 2015, pp. 1-3.
- [8] X. Xi *et al.*, "Strong subthreshold current array PUF with 265 challenge-response pairs resilient to machine learning attacks in 130nm CMOS," in: *Proc. VLSI-Symp'17*, Kyoto, Japan, Jun. 2017, pp. 268-269.
- [9] Y. Yoshimoto, *et al.*, "A ReRAM-based physically unclonable function with bit error rate <0.5% after 10 years at 125 $^{\circ}$ C for 40nm embedded application," in: *Proc. VLSITech'16*, Honolulu, HI, 2016, pp. 1-2.
- [10] H. Jiang *et al.*, "A provable key destruction scheme based on memristive crossbar arrays," *Nature Electronics*, vol.10, pp. 548, 2018.
- [11] M. Rostami *et al.*, "Quo vadis, PUF?: Trends and challenges of emerging physical-disorder based security," in: *Proc. DATE'14*, Dresden, Germany, Mar. 2014, pp. 1-6.
- [12] M. R. Mahmoodi, and D. B. Strukov, "An ultra-low energy internally analog, externally digital vector-matrix multiplier based on NOR flash memory technology," in: *Proc. DAC'18*, San Francisco, CA, Jun. 2018, art. 22.
- [13] X. Guo *et al.*, "Fast, energy-efficient, robust, and reproducible mixed-signal neuromorphic classifier based on embedded NOR flash memory technology," in: *Proc. IEDM'17*, San Francisco, CA, Dec. 2017, pp. 6.5.1-6.5.4.
- [14] X. Guo *et al.*, "Temperature-insensitive analog vector-by-matrix multiplier based on 55 nm NOR flash memory cells," in: *Proc. CICC'17*, Austin, TX, Apr. 2017, pp. 1-4.
- [15] M. R. Mahmoodi and D. B. Strukov, "Breaking Pop/J barrier with analog multiplier circuits based on nonvolatile memories," in: *Proc. ISLPED'18*, Bellevue, WA, Jul. 2018, art. 39.
- [16] M. R. Mahmoodi, H. Nili and D. B. Strukov, "RX-PUF: Low power, dense, reliable, and resilient physically unclonable functions based on analog passive RRAM crossbar arrays," in: *Proc. VLSI-Symp'18*, Honolulu, HI, Jan. 2018, pp. 99-100.
- [17] A. Bendali, and Y. Audet, "A 1-V CMOS current reference with temperature and process compensation," *IEEE TCAS-I*, vol. 54, pp.1424-1429, 2007.
- [18] Y. Hori *et al.*, "Quantitative and statistical performance evaluation of arbiter physical unclonable functions on FPGAs," in: *Proc. Reconfigurable Computing and FPGAs (ReConFig'10)*, Quintana Roo, Mexico, Dec. 2010, pp. 298-303.