A Highly Reliable and Anti-Cloning STT-MRAM Reconfigurable Physical Unclonable Function

Hongrui Meng[1], Yajun Wu[1], Shengchao Zhou[1], Teng Zou[1], Tai Min[2], Shaohao Wang[3] and Yufeng Xie[1*]
[1]Fudan University, Shanghai, China; [2]Xi'An Jiaotong University, Xi'An, China; [3]Fuzhou University, Fuzhou, China.
Email: xieyf@fudan.edu.cn

***Abstract*—In this paper, we propose a highly reliable and anti-cloning STT-MRAM reconfigurable physical unclonable function (PUF) chip. To improve reliability, a charge-injection offset-compensation sense amplifier (CIOC-SA) with a low input offset of 5.6mV is proposed, which significantly reduces the worst-case Bit Error Rate (BER) from 0.146% to 0.0025%. To enhance anti-clonality, two innovations are proposed, including a novel STT-MRAM-based PUF reconfiguration scheme and an ultra-low-cost invasive-attack detection scheme. The reconfiguration scheme achieves a near-ideal reconfiguration hamming distance of 0.456. Featuring a low wordline (WL) voltage and a two-stage hysteresis comparator configured from the CIOC-SA, the proposed detection scheme achieves high detection accuracies of 90.4%, 88.6%, and 99.3% for three test PUF arrays emulating different levels of invasive attacks, demonstrating a notably enhanced anti-clonality with ultra-low cost in latency and area.**

***Keywords*—Physical Unclonable Function, cloning attack, STT-MRAM, stochastic switching, sense amplifier, reconfigurability.**

I. Introduction

The exponential growth of Internet of Things (IoTs) has been intensifying demands for more reliable hardware security solutions, which motivates the widespread adoption of Physical Unclonable Functions (PUFs). By utilizing process variations during fabrication, a PUF can generate unique responses for given challenges, forming a fingerprint-like Challenge-Response Pair (CRP) space and enhancing resistance to attacks.

Among all PUFs, memory-based PUFs (MemPUFs) are particularly well-suited for IoT devices that are typically resource-constrained. They can be implemented directly in existing on-chip memory, minimizing hardware overhead and design complexity. In recent years, many studies on MemPUFs have been conducted. As one of the earliest PUFs, SRAM PUFs require additional stabilization techniques [1] to mitigate their high native bit error rate (BER), increasing area and power consumption. To improve cost-efficiency, non-volatile MemPUFs [2]-[5] utilizing post-process entropy sources with higher randomness are gradually being applied. Among them, STT-MRAM PUFs emerge as a promising candidate for critical IoT nodes due to STT-MRAM's high speed, long retention time, high endurance, and compact cell area [6].

However, STT-MRAM PUFs face two challenges: 1) Limited reliability due to the small tunnel magnetoresistance (TMR) ratio [6] of the magnetic tunnel junction (MTJ). 2) The non-volatility of STT-MRAM increases the risk of CRP leakage compared to SRAM PUFs, which elevates the possibility of PUF cloning. In 2013, the first successful invasive cloning attack [7] on MemPUFs was conducted, where a 2KB SRAM PUF was characterized and cloned by invasive cloning attack, and the cloned PUF exhibits identical power-on state to the target PUF. Indeed, this cloning attack is also considered feasible for STT-MRAM PUFs [8]. However, existing studies on STT-MRAM PUFs almost rely on simulations, lacking real chip implementations that simultaneously optimize reliability, anti-clonality, and cost-efficiency.

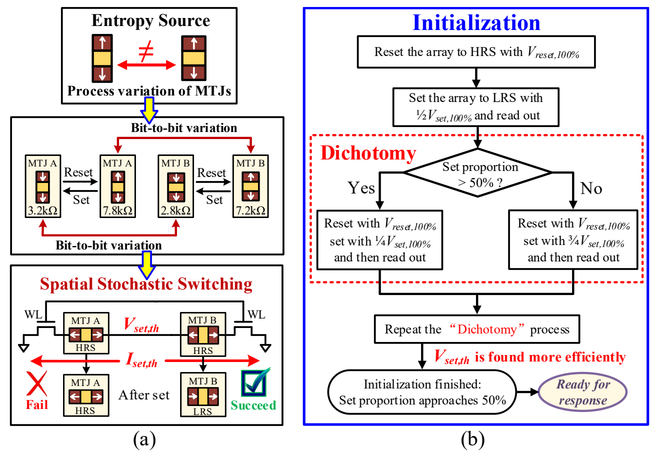

Fig. 1(a) Entropy source of the proposed PUF. (b) Workflow of initialization.

In this paper, we propose a highly reliable and anti-cloning STT-MRAM reconfigurable PUF chip. To improve reliability, a charge-injection offset-compensation sense amplifier is proposed to extract responses more reliably. Then, to enhance anti-clonality, two innovations are proposed: 1) A novel STT-MRAM-based PUF reconfiguration scheme to increase the cost and complexity of cloning. 2) An ultra-low-cost invasive-attack detection scheme to verify the authenticity of the PUF.

II. PUF Implementation

A. Entropy source, initialization, and chip architecture

The entropy source utilized in our PUF is the process variation. As illustrated in Fig. 1(a), the process variation induces the bit-to-bit (B2B) variation [6] of MTJs. Owing to this variation, a set voltage $V_{set,th}$, where half of the MTJs in high-resistance state (HRS) are set to low-resistance state (LRS), can be found. This can be defined as spatial stochastic switching, which is employed to initialize a random and uniform 0-1 distribution when our PUF is used for the first time and upon each reconfiguration. However, this initialization operation limits the operational efficiency of the PUF as it suffers from excessive latency. Therefore, we introduce a "Dichotomy" initialization scheme to find $V_{set,th}$ more efficiently. As shown in Fig. 1(b), the workflow of the proposed scheme begins with a 100% reset to HRS using a reset voltage $V_{reset,100\%}$. Next, ½$V_{set,100\%}$ ($V_{set,100\%}$ ensures 100% set) is applied

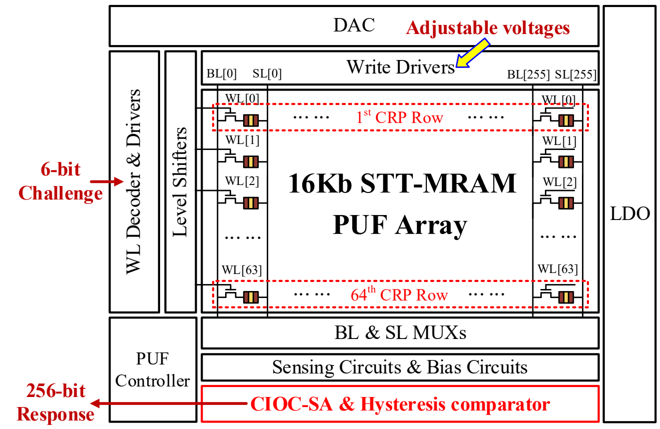

Fig. 2. Chip architecture of the proposed STT-MRAM PUF.

to set the array to LRS, and the array is read out. Then, a "Dichotomy" process is performed: If the set proportion exceeds 50%, which indicates the set current is too high, the array is reset and then set again with ¼$V_{set,100\%}$. Conversely, if the set proportion is below 50%, which indicates the set current is too low, the array is reset and then set with ¾$V_{set,100\%}$. By repeating this "Dichotomy" process, $V_{set,th}$ where set proportion approaches 50%, can be found more efficiently, significantly simplifying the initialization operation.

Fig. 2 depicts the chip architecture of the proposed PUF, featuring a 16kb 1T1M STT-MRAM PUF array and a CIOC-SA & hysteresis comparator group. The PUF array is organized into 64 rows by 256 columns, where each row functions as a 256-bit CRP, resulting in 64 groups of CRPs in total. While challenging, a 6-bit challenge is input to select the corresponding CRP row, which is subsequently extracted by 16 CIOC-SAs in parallel, ultimately generating a 256-bit response. The chip incorporates complete peripheral circuitry such as write drivers, WL decoder and drivers, BL & SL MUXs, level shifters, and sensing circuits & bias circuits. In addition, the chip integrates a PUF controller, a low-dropout regulator (LDO) for power management, and a digital-to-analog converter (DAC) to provide adjustable operational voltages.

B. Charge-injection offset-compensation sense amplifier

In this work, the resistance values of MTJs in LRS and HRS, denoted as R_P and R_{AP}, are 3kΩ and 7.5kΩ, resulting in a small TMR ratio of 150%, limiting the reliability during response extraction. To overcome this limitation, we propose a configurable charge-injection offset compensation sense amplifier (CIOC-SA) with a low input offset. As illustrated in Fig. 3, the CIOC-SA consists of a switched-capacitor preamplifier and a strongarm comparator. Fig. 4(a) details the workflow of the CIOC-SA. In Phase 0, S1, RST, and OS are turned on. P0, N0 and P1, N1 form two inverters to reset and equalize the sampling capacitors C_{SP} and C_{SN}. In Phase 1, RST is turned off. The voltage differences between VDD and the threshold voltages of P0 and P1 are sampled onto C_{SP} and C_{SN} for input offset compensation. When Phase 2 starts, OS is turned off, and PRE and AMP_EN are set to 1. Nodes ON and OP are predischarged to VSS by N0 and N1. When Phase 3 starts, S1 is turned off and S2 is turned on. Input voltages V_{BL} and V_{REF} are coupled to nodes INP and INN through C_{SP} and C_{SN}. Before the end of this phase, S2 is turned off early. Due to the asymmetric charge injection effects, as presented by (1) and (2), the voltage difference between INP and INN, namely sense

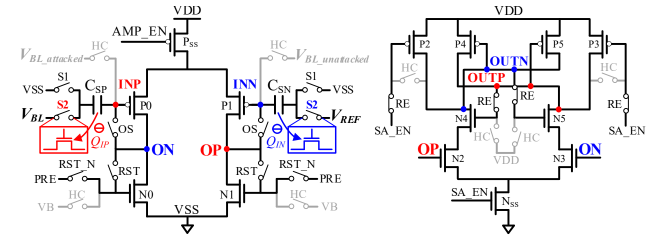

Fig. 3. Schematic of the configurable CIOC-SA.

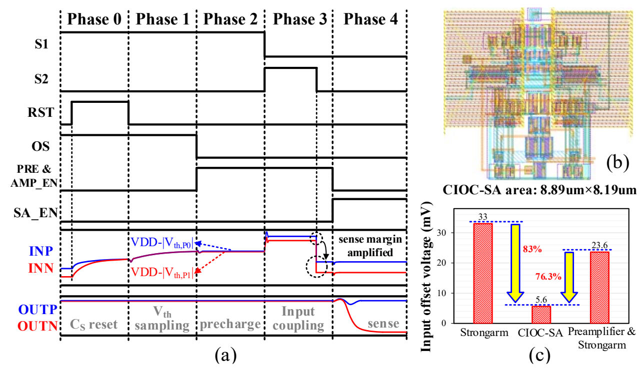

Fig. 4(a) Workflow of the CIOC-SA. (b) Layout of the CIOC-SA. (c) Input offset voltage within ±3σ from 10k-point Monte Carlo post simulation.

margin, is notably amplified. Before Phase 4, SA_EN remains 0, keeping output nodes OUTP and OUTN precharged to VDD by P2 and P3. When Phase 4 starts, PRE and AMP_EN are set to 0, and SA_EN is set to 1. If $V_{INP} > V_{INN}$, the driving capability of P1 is higher than that of P0, causing N2 to turn on earlier than N3. As a result of the positive feedback, OUTN is pulled down to 0 while OUTP remains at VDD, thus finishing a highly reliable response extraction.

$$Q_{IP} = C_{SP} \cdot \Delta V_{INP} = (WL)_{S2} C_{ox} \cdot [VDD - V_{BL} - (V_{th})_{S2}] \tag{1}$$

$$Q_{IN} = C_{SN} \cdot \Delta V_{INN} = (WL)_{S2} C_{ox} \cdot [VDD - V_{REF} - (V_{th})_{S2}] \tag{2}$$

Fig. 4(b) shows the layout of the CIOC-SA, which occupies an area of 72.8um^2. The inclusion of two sampling capacitors, each 20fF, incurs only a 25% increase in area. 10k-point Monte Carlo post simulation demonstrates an input offset voltage of 5.6mV within ±3σ range. As shown in Fig. 4(c), compared to the strongarm SA and the combination of preamplifier and strongarm SA (These two structures share the same transistor dimensions as CIOC-SA), the input offset voltage of the proposed CIOC-SA is reduced by 83% and 76.3%, respectively.

C. Innovations for enhancing anti-clonality

Even if an attacker fully characterizes the CRP space, the complexity and failure rate of cloning would be much higher with controllable reconfiguration. Therefore, a novel STT-MRAM-based PUF reconfiguration scheme is first proposed to enhance anti-clonality. As illustrated in Fig. 5(a), in addition to B2B variation, MTJ also exhibits cycle-to-cycle (C2C) variation [9], meaning that the resistance of a MTJ randomly fluctuates within a certain range after each switching. So, under a fixed $V_{set,th}$, the corresponding set current $I_{set,th}$ fluctuates in each cycle, leading to temporal stochastic switching. Although the C2C variation of STT-MRAM is relatively modest, even a slight variation can affect switching outcomes when the set current is $I_{set,th}$. For example, suppose the R_{AP} of MTJ A is 7.4

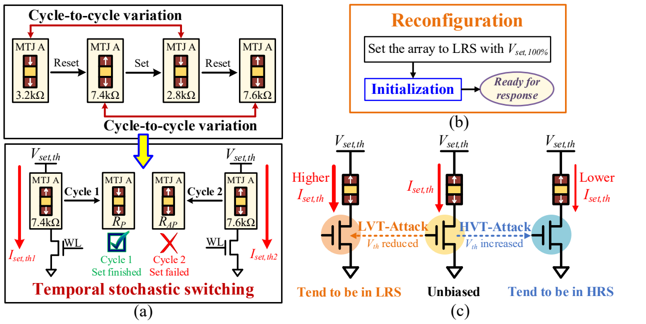

Fig. 5(a) Principle of the novel STT-MRAM-based reconfiguration scheme. (b) Workflow of the reconfiguration scheme. (c) Invasive-attack model on the proposed PUF.

kΩ in cycle 1, and is 7.6kΩ in cycle 2. The higher set current in cycle 1 increases the possibility of MTJ A switching to LRS, while the lower set current in cycle 2 makes switching less probable. Therefore, the C2C variation can be exploited to establish reconfigurable 0-1 distributions. The reconfiguration workflow is shown in Fig. 5(b), which begins by setting the array to LRS, followed by the initialization operation. Notably, if not pre-set, the resistance values of the PUF bits in HRS after last initialization remain unchanged. Hence, these bits still tend not to switch during next initialization, compromising the effectiveness of the reconfiguration operation.

However, for IoT nodes with high-value security assets, such as smart city management systems, cryptocurrency hardware wallets and medical life-support equipment, even a single successful attack could lead to catastrophic consequences. So, attackers have enough incentive to launch costly invasive attacks to directly manipulate the reconfiguration logic, thereby enabling much more effective cloning. In this case, designers, vendors or users must be able to verify whether the PUF has been cloned by invasive attacks. For STT-MRAM PUFs, the most susceptible target of invasive attacks is the access transistor in PUF cells, which can be characterized as the modification of its threshold voltage (V_{th}). As Fig. 5(c) presented, for LVT-attacks, which reduce the V_{th}, the $I_{set,th}$ is higher. Consequently, the set possibility exceeds 50%, indicating cells under LVT-attacks tend to be in LRS. For HVT-attacks, which increases the V_{th}, the $I_{set,th}$ is lower and the set possibility is below 50%, indicating cells under HVT-attacks tend to be in HRS. During response extraction, due to the B2B variation and C2C variation, the BL voltages of cells fluctuate within a certain range which overlaps with the BL voltage range varying due to invasive attacks. This is because the access transistors are large in size to provide enough switching current, making their on-resistance (R_{on}) much lower than R_{AP} and R_P. As a result, directly detecting invasive attacks through response extraction is impractical.

A scheme that extends the rise time of the WL signal [9] is proposed to detect invasive attacks, which is time-consuming and needs a complex ramp voltage circuit, notably increasing hardware overhead. Therefore, we propose an ultra-low-cost invasive-attack detection scheme, with its principle illustrated in Fig. 6(a). By applying a low gate voltage to the access transistor, its R_{on} is set to a much higher value. As a result, the BL voltages of the attacked cells distribute over a much wider range, which is about 100mV and enables detection with very low cost in latency. In fact, only attacks that can affect switching outcomes are worth detecting. Accordingly, the BL

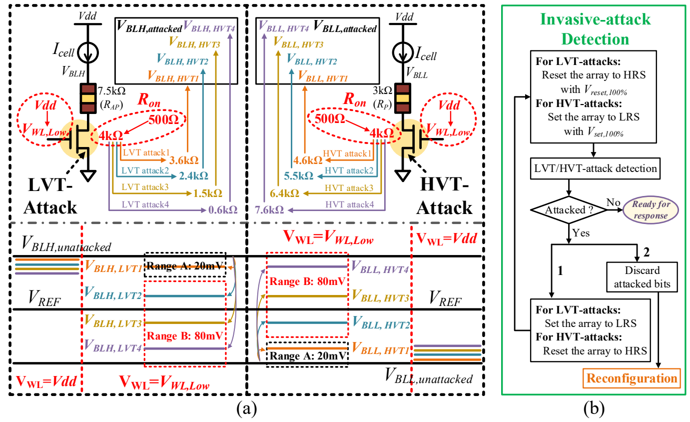

Fig. 6(a) Principle of the invasive-attack detection scheme. (b) Workflow of the invasive-attack detection scheme.

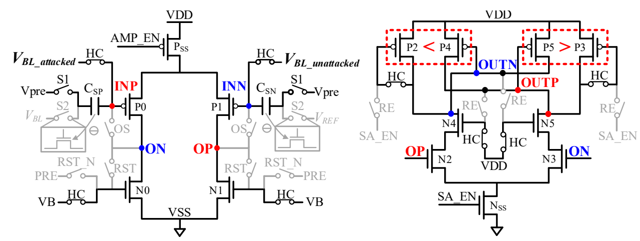

Fig. 7. Schematic of the hysteresis comparator configured from the CIOC-SA.

voltage range is divided into Range A and Range B, with voltage spans of 20mV and 80mV. BL voltages falling within range A, such as $V_{BLH,LVT1}$ and $V_{BLL,HVT1}$, are considered unattacked, while BL voltages within Range B, such as $V_{BLH,LVT2\text{-}4}$ and $V_{BLL,HVT2\text{-}4}$, are considered under LVT-attacks and HVT-attacks. The workflow of the detection scheme is presented in Fig. 6(b). For LVT (HVT)-attacks, firstly reset (set) the array to HRS (LRS), then detect the entire array. If any attacked bit is detected, the PUF supports two countermeasures. The first is to repeat the operation: set the array (reset) to LRS (HRS), then reset back to HRS (LRS), and detect again; The second is to reuse the PUF: discard the attacked bits and then reconfigure the PUF to update the leaked CRP space.

To detect whether any cell falls within Range B at minimal cost, we propose a two-stage hysteresis comparator which is directly configured from the CIOC-SA without extra area overhead, as shown in Fig. 7. Its workflow is as follows: when BL begins to precharge, S1 is turned on to reset the voltages on C_{SP} and C_{SN}. After enabling the low WL signal, the BL voltage begins to develop and S1 is turned off. By designing an appropriate ratio of $(W/L)_{P4,5}$ and $(W/L)_{P2,3}$, the hysteresis comparator is well-tuned such that when the input difference between $V_{BL,attacked}$ and $V_{BL,unattacked}$ exceeds the voltage span of Range A, output nodes OUTP and OUTN generate complementary step signals. These signals are captured by the subsequent D-FFs, which then produce rail-to-rail outputs, indicating the detection result.

IV. Performance Evaluation

Fig. 8(a) shows the chip occupies an area of 0.059mm^2, with each cell occupying only 0.1554um^2 (97.13F^2). As shown in Fig. 8(b), the μ/σ of the inter hamming distance (inter-HD) and intra hamming distance (intra-HD) are 0.5026/0.0217 and 0/0, demonstrating high reliability and uniqueness under nominal conditions. Fig. 8(c) illustrates the autocorrelation test result,

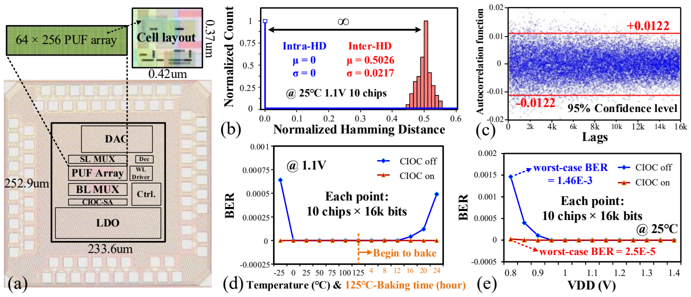

Fig. 8(a) Chip photo and cell layout. (b) Measured Inter-HD and Intra-HD. (c) Autocorrelation test result. (d) Measured BER under varying temperature and varying baking time at 125°C. (e) Measured BER under varying supply voltage.

Test	Stream length	Runs No.	Pass rate	Avg. P-value	Pass?
Frequency	1024	50	100%	0.4944	Yes
Block frequency	1024	50	100%	0.7399	Yes
Runs	1024	50	98%	0.4475	Yes
Longest run	1024	50	98%	0.5306	Yes
Rank	1024	50	100%	0.4732	Yes
FFT	1024	50	98%	0.4695	Yes
Approximate entropy	1024	50	100%	0.685	Yes
Linear complexity	1024	50	96%	0.4138	Yes

(a)

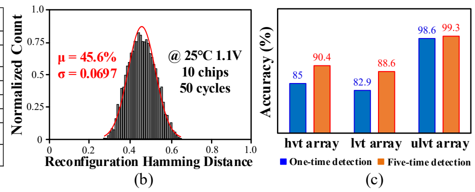

Fig. 9(a) NIST SP 800-22 randomness test result. (b) Measured reconfiguration hamming distance. (c) Measured one-time and five-time detection accuracies for three test arrays emulating different levels of invasive attacks.

where most data points fall within the 95% confidence level (±0.0122), indicating no apparent correlation and periodicity in PUF responses. Fig. 8(d) and Fig. 8(e) present the measured BER under varying temperature, baking time at 125°C, and supply voltage (VDD). When the configurable CIOC technique is disenabled, BER starts to exceed 0 as the temperature drops below 0°C and as VDD decreases below 0.9V, and the worst-case BER reaches 0.146% at 0.8V. During the 125°C-baking, BER begins to exceed 0 from the 16th hour. In contrast, with CIOC enabled, BER remains at 0 throughout the entire temperature range and 24-hour 125°C-baking. Even as VDD varies from 0.8V to 1.4V, BER remains almost constant, with the worst-case BER reduced from 0.146% to 0.0025% at 0.8V, demonstrating that the proposed configurable CIOC-SA significantly improves the reliability of the PUF.

The result of NIST SP 800-22 randomness test is shown in Fig. 9(a), where all test pass rates exceed 94%, indicating the high randomness and uniformity. Fig. 9(b) illustrates the reconfiguration hamming distance over 50 reconfiguration cycles, with a μ/σ of 0.456/0.0697, demonstrating a near-ideal reconfigurability. To experimentally evaluate the performance of the proposed invasive-attack detection scheme, three test PUF arrays, which are composed of high-threshold (hvt), low-threshold (lvt), and ultra-low-threshold (ulvt) MOS transistors, are added to emulate different levels of invasive attacks. During detection, the BL voltages of the hvt, lvt, ulvt arrays all fall within Range B and exceed Range A by about 12mV, 11.1mV, and 27mV. As shown in Fig. 9(c), the one-time detection accuracies for these three arrays are 85%, 82.9%, and 98.6%, while the five-time detection mitigates the miss-detection caused by C2C variations, achieving the accuracies of 90.4%, 88.6%, and 99.3% and marking a notable improvement. Finally, a summary of the measurement results and a comparison with state-of-the-art work are presented in Table I.

V. Conclusion

In conclusion, this paper presents a highly reliable and anti-cloning STT-MRAM reconfigurable PUF chip. The proposed configurable CIOC-SA reduces the worst-case BER from 0.146% to 0.0025%, significantly improving the reliability. Additionally, the proposed reconfiguration scheme achieves a near-ideal reconfiguration hamming distance of 0.456, while the proposed ultra-low-cost invasive-attack detection scheme achieves high detection accuracies for different levels of invasive attacks, demonstrating a notably enhanced anti-clonality with ultra-low cost in latency and area.

TABLE I. Comparison With State-of-the-Art Work

	JSSC'21 [2]	TCASII'22 [4]	VLSI'23 [1]	VLSI'24 [3]	TCASII'24 [5]	**This work**
Technology	130nm	28nm	28nm	28nm	28nm	**40nm**
PUF Type	RRAM	STT-MRAM	SRAM	RRAM	STT-MRAM	**STT-MRAM**
Cell area (F^2/bit)	169	-	442	67	-	**97.13**
Native BER	6.1E-6	<6.9E-2 (worst case)	4.2E-3	2.1E-4	-	**<1.46E-3 (worst case)**
Stabilization technique	-	-	Masking	Auto-extract	-	**CIOC-SA**
Stabilization overhead	-	-	High	High	-	**Low**
Stabilized BER	-	-	<1.5E-7	<5E-9	-	**<2.5E-5 (worst case)**
Temperature (°C)	25-150	-25-100	0-70	-40-125	0-100	**-25-125**
VDD (V)	1.4-2.2	0.5-1.0	0.6-0.7	0.7-1.3	0.9	**0.8-1.4**
Inter-HD	49.99%	49.76%	49.23%	50.07%	49.87%	**50.26%**
Intra-HD	0	0.447%	0.03%	0.04%	<0.09%	**0**
Energy (fJ/bit)	3028	870	75	23	-	**178**
Anti-clonality enhancement	No	Yes	No	No	No	**Yes**
Tape out	Yes	No	Yes	Yes	No	**Yes**

References

[1] J. Basu, S. Taneja, V. K. Rajanna, T. Wang and M. Alioto, "ECC-Less Multi-Level SRAM Physically Unclonable Function and 127% PUF-to-Memory Capacity Ratio with No Bitcell Modification in 28nm," *2023 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Kyoto, Japan, 2023, pp. 1-2.

[2] B. Lin *et al*., "A Highly Reliable RRAM Physically Unclonable Function Utilizing Post-Process Randomness Source," in *IEEE Journal of Solid-State Circuits*, vol. 56, no. 5, pp. 1641-1650, May 2021.

[3] Y. Cao, H. Yang, J. Yang, Q. Liu and M. Liu, "A 67F2 Reconfigurable PUF Using 1T2R RRAM Switching Competition in 28nm CMOS with 5e-9 Bit Error Rate," *2024 IEEE Symposium on VLSI Technology and Circuits (VLSI Technology and Circuits)*, Honolulu, HI, USA, 2024, pp. 1-2.

[4] R. Ali, D. Zhang, H. Cai, W. Zhao and Y. Wang, "A Machine Learning Attack-Resilient Strong PUF Leveraging the Process Variation of MRAM," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 6, pp. 2712-2716, June 2022.

[5] H. Du, Y. Wang, J. Yang and H. Cai, "Intrinsic MRAM Properties Enable Security Circuits," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 3, pp. 1695-1700, March 2024.

[6] T. Na, S. H. Kang and S. -O. Jung, "STT-MRAM Sensing: A Review," in *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 68, no. 1, pp. 12-18, Jan. 2021.

[7] C. Helfmeier, C. Boit, D. Nedospasov and J. -P. Seifert, "Cloning Physically Unclonable Functions," *2013 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, Austin, TX, USA, 2013, pp. 1-6.

[8] S. Ben Dodo, R. Bishnoi, S. Mohanachandran Nair and M. B. Tahoori, "A Spintronics Memory PUF for Resilience Against Cloning Counterfeit," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 11, pp. 2511-2522, Nov. 2019.

[9] R. Carboni *et al.*, "A Physics-Based Compact Model of Stochastic Switching in Spin-Transfer Torque Magnetic Memory," in *IEEE Transactions on Electron Devices*, vol. 66, no. 10, pp. 4176-4182, Oct. 2019.