36.1 Unified In-Memory Dynamic TRNG and Multi-Bit Static PUF Entropy Generation for Ubiquitous Hardware Security

Sachin Taneja, Viveka Konandur Rajanna, Massimo Alioto

National University of Singapore, Singapore, Singapore

2021 IEEE International Solid- State Circuits Conference (ISSCC) | 978-1-7281-9549-0/20/$31.00 ©2021 IEEE | DOI: 10.1109/ISSCC42613.2021.9366019

Secure integrated systems routinely require the generation of keys in the form of dynamic entropy from True Random Number Generators (TRNGs), and static entropy from Physically Unclonable Functions (weak PUFs) as in Fig. 36.1.1, to support the execution of security protocols [1]-[8] (e.g., for system authentication, data confidentiality). In low-cost systems such as sensor nodes, unified implementations of TRNGs and/or PUFs have been demonstrated to reduce their cost and area, thanks to circuit reuse across two different functions, such as a TRNG with a PUF in a single standalone macro [1], and a TRNG with a data converter [2]. Similarly, SRAM-based PUFs have been widely explored and commercially used to exploit their omnipresent availability, and their high PUF key density per unit area (e.g., [3], [4]). However, existing SRAMs are not able to assume the function of a TRNG, and their PUF operation is limited to one bit per bitcell at most, whereas multi-bit/cell operation is currently restricted to non-SRAM PUFs [5].

This work introduces an SRAM with unified a TRNG and multi-bit PUF for complete in-memory dynamic and static entropy generation for low-cost security, both in terms of area and design (e.g., reduced system integration effort). In both cases, the randomness generated by the SRAM array is extracted through low-area column periphery augmentation, while reusing the baseline SRAM circuitry and retaining compiler-based automated design. In-memory TRNG operation is enabled by digitization of the jitter accumulated in leakage-driven bitline discharge (Fig. 36.1.2). PUF entropy is based on bitcell read-current digitization (Fig. 36.1.2), which enables: 1) a multi-bit/bitcell PUF key for improved density; and 2) uninterrupted SRAM bank read and PUF access without intermediate data flushing (as opposed to conventional power-up state-based or unified SRAM PUFs [3], [4]).

Dynamic entropy is generated from bitline capacitance discharge time t_d under low current (all bitcells leaking on column, all wordlines disabled), accumulating noise into time jitter as in a Wiener process [9] by utilizing the existing SRAM infrastructure (Fig. 36.1.2). The adoption of leakage as discharge current has the benefit of: 1) magnifying the variance of t_d (i.e., randomness) since it is inversely proportional to the discharge current; 2) further magnifying randomness through concurrent (uncorrelated) noise contribution of all bitcells in a column; and 3) making transistor current flicker noise negligible compared to white noise [9] (i.e., statistical "coloring" is inherently eliminated). Accumulated jitter on bitline discharge is converted into a random pulsewidth t_w as the time interval between 60% and 40% crossing, as detected by simple skewed inverters. t_w is then digitized by counting the oscillation periods of a gated ring oscillator (GRO, Fig. 36.1.3), using truncated modulo-15 counters to: 1) reduce column area overhead while capturing the inherent randomness in LSBs [8]; 2) translate the Gaussian distribution of t_w (Wiener process) into a uniform one [8]; and 3) enable resilience against mismatch (local variations) as required in TRNGs (LSBs are affected by noise, not mean discharge current). After t_w, the skewed inverters are shut down by power gating, and the 4b count per column represents the TRNG output. To suppress unnecessary GRO energy increase at lower temperatures (i.e., lower leakage) due to longer t_w, the GRO current-starved inverter delay is set by a self-tuning current-starving voltage loop (activated infrequently). The loop keeps the average GRO oscillation count around the nominal target.

The proposed PUF digitizes the bitcell read current I_{read} to extract n>1 bits/bitcell, converting the difference (t_A-t_B) of the bitline discharge time of any two bitcells in adjacent bitlines (A and B in Fig. 36.1.2). To emphasize I_{read} mismatch over the dynamic entropy source (noise), the wordline is under-driven by 20% as a widely-available assist technique [10]. Bitcells in PUF SRAM rows store a "0", whereas unreserved rows in the same or different bank are used as normal address space. The bitcell pair is taken from bitlines within the same column MUX, allowing fruitful reuse of the energy conventionally wasted in pseudo-reads of such unselected bitlines. (t_A-t_B) is digitized through time-to-digital conversion, mapping (t_A-t_B) to one of 2^n time bins (see example with n=2 in Fig. 36.1.2). The PUF LSB output PUF[0] is derived by comparing (t_A-t_B) with a zero-time threshold via the NAND-based arbiter in Fig. 36.1.3, setting it to 1 (0) if (t_A-t_B) is negative (positive). The PUF MSB output PUF[1] is derived by comparing (t_A-t_B) with non-zero thresholds defining four equiprobable time bins (25% probability each), allocating a 1 if (t_A-t_B) falls in the outside lobe of its Gaussian distribution (0 otherwise) in Fig. 36.1.3. The delay thresholds are simply set at design time for 25% per bin split (i.e., $\pm 0.68\sigma$ via Monte Carlo simulations) at nominal conditions and corner, and equally applied to all dice with no added testing/calibration effort. Higher number of bits per bitcell (e.g., n=3) can be derived at higher area cost. Conventional masking suppresses instability, including PUF responses lying at bin boundaries (affecting masking only marginally, see BER below).

In Fig. 36.1.4, the TRNG was confirmed to have consistent measured output quality across very different data patterns (all 0's for minimum jitter vs. random data), 0.8-to-1V supply and -10 to 75°C temperature. The min-entropy is always greater than 0.99, all NIST tests passed (p-value>0.01), autocorrelation function (ACF) at 95% confidence is within 0.002, and the phi coefficient between simultaneous streams is near-zero (0.001 on average), confirming cryptographic-grade randomness. Under 1b Von Neumann extraction (6000F², off-chip) and 1 dropped LSB, ~2.25 random bits are generated by every column at 36,000F² area overhead. TRNG operation maintains nearly-constant energy across temperatures thanks to the GRO frequency tuning loop, reducing energy variability from 5× to 2.3× (Fig. 36.1.4).

From Fig. 36.1.5, the PUF LSB has a maximum native bit error rate (BER) of 1.84% and 11.9% unstable bits (UB) at 0.9V, 25°C (golden key). The MSB has maximum BER of 3.6%, and UB of 36.5%. Temperature variations from -10 to 75°C degrade the BER of the LSB (MSB) to 2.5% (7.5%), whereas voltage variations from 0.8-1V degrade it to 4.4% (12.6%). Combined temperature and voltage variations degrade the BER of LSB (MSB) to 4.8% (17.6%), and drastically different data patterns in unreserved rows (50% Hamming distance on adjacent bitlines) marginally degrade BER by another <0.2%. The statistical independence of multiple PUF output bits is confirmed by the near-zero measured phi coefficient (0.003). PUF output randomness is confirmed by ACF within 0.007, inter-die Hamming distance of 50.3%, HD separation of >14×, Shannon entropy >0.9997, and passing all applicable NIST tests (80kb data).

Compared to prior art (Fig. 36.1.6), the proposed SRAM enables both dynamic and static entropy within the same array. The digital nature of its periphery augmentation is suitable for low-cost systems (low area, low design and integration effort, digital-like scaling). From Fig. 36.1.6, the TRNG 16,000F² area overhead per output stream is 8.7-to-150× lower than prior art [1], [2], [7], [8]. The PUF exhibits 1.3-to-4.6× lower area/bit thanks to its multi-bit/bitcell capability, and the compact periphery with no area-hungry analog circuitry [4] (Fig. 36.1.3). This also leads to >50× higher throughput at 7% area overhead over a conventional SRAM, and an energy/bit of the same order of magnitude as [3], while not requiring a custom bitcell, and at least 4.9× lower than [4], [6]. Finally, the inherent data locality enforcement in key generation can be combined with secure SRAM techniques (e.g., internally scrambled, encrypted) to enable a higher level of security against physical attacks, in view of the resulting physical confinement of confidential data within the memory, and the elimination of obvious attack points on key generation.

Acknowledgement:
This work was supported by the Singapore National Research Foundation (grant NRF2018NCR-NCR002-0001), and by TSMC for chip fabrication support.

References:

[1] S. K. Satpathy et al., "An All-Digital Unified Physically Unclonable Function and True Random Number Generator Featuring Self-Calibrating Hierarchical Von Neumann Extraction in 14-nm Tri-gate CMOS," *IEEE JSSC*, vol. 54, no. 4, pp. 1074-1085, 2019.
[2] M. Kim et al., "A 82-nW Chaotic Map True Random Number Generator Based on a Sub-Ranging SAR ADC," *IEEE JSSC*, vol. 52, no. 7, pp. 1953-1965, 2017.
[3] K. Liu et al., "A 0.5-V 2.07-fJ/b 497-F2 EE/CMOS Hybrid SRAM Physically Unclonable Function with < 1E-7 Bit Error Rate Achieved through Hot Carrier Injection Burn-in," *IEEE CICC*, 2020.
[4] J. Li et al., "An Area-Efficient Microprocessor-Based SoC With an Instruction-Cache Transformable to an Ambient Temperature Sensor and a Physically Unclonable Function," *IEEE JSSC*, vol. 53, no. 3, pp. 728-737, 2018.
[5] K. H. Chuang et al., "A Multi-bit/Cell PUF Using Analog Breakdown Positions in CMOS," *IEEE IRPS*, pp. P-CR.2-1-P-CR.2-5, 2018.
[6] Y. Choi et al., "Physically Unclonable Function in 28nm FDSOI Technology Achieving High Reliability for AEC-Q100 Grade 1 and ISO26262 ASIL-B," *ISSCC*, pp. 426-427, 2020.
[7] V. R. Pamula et al., "An All-Digital True-Random-Number Generator with Integrated De-correlation and Bias Correction at 3.2-to-86 MB/S, 2.58 PJ/Bit in 65-NM CMOS," *IEEE Symp. VLSI Circuits, 2018*.
[8] E. Kim et al., "8.2 8Mb/s 28Mb/mJ Robust True-Random-Number Generator in 65nm CMOS Based on Differential Ring Oscillator with Feedback Resistors," *ISSCC*, pp. 144-145, 2017.
[9] A. A. Abidi, "Phase Noise and Jitter in CMOS Ring Oscillators," *IEEE JSSC*, vol. 41, no. 8, pp. 1803-1816, 2006.
[10] Z. Guo et al., "A 23.6-Mb/mm² SRAM in 10-nm FinFET Technology With Pulsed-pMOS TVC and Stepped-WL for Low-Voltage Applications," *IEEE JSSC*, vol. 54, no. 1, pp. 210-216, 2019.

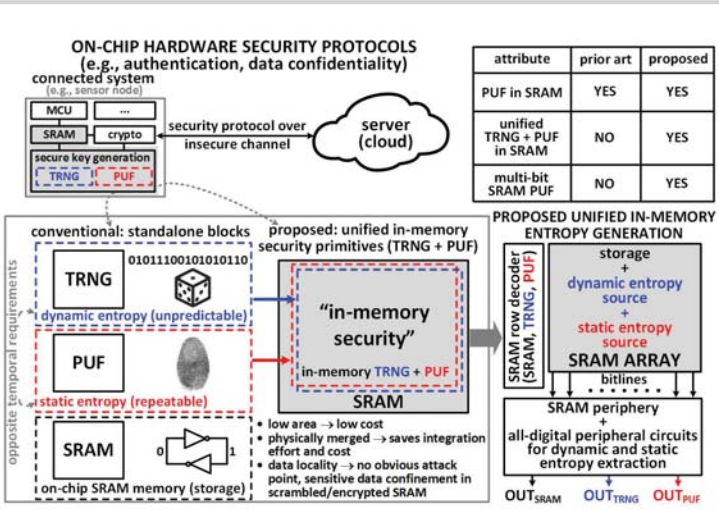


Figure 36.1.1: Proposed in-memory secure key generation within the same SRAM array. Both dynamic (True Random Number Generator, TRNG) and static entropy (Physically Unclonable Function, PUF) are generated within the same SRAM array for low-area and low-cost key generation for hardware security.

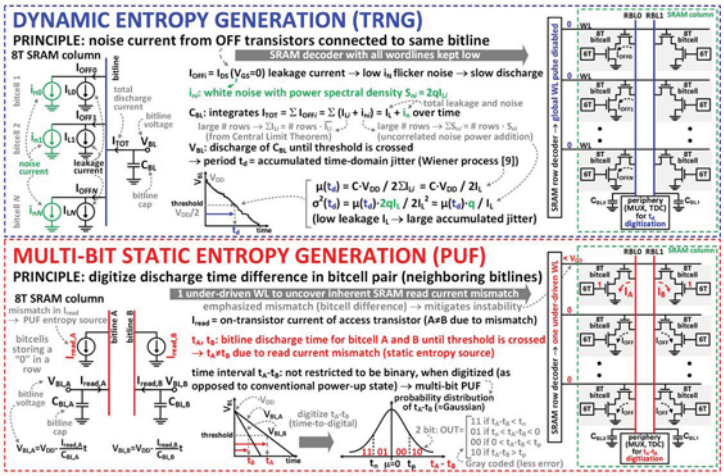


Figure 36.1.2: Proposed approach for in-memory dynamic (top) and static entropy (bottom), as exemplified for the 8T bitcell array used in this work (same for 6T and other bitcells).

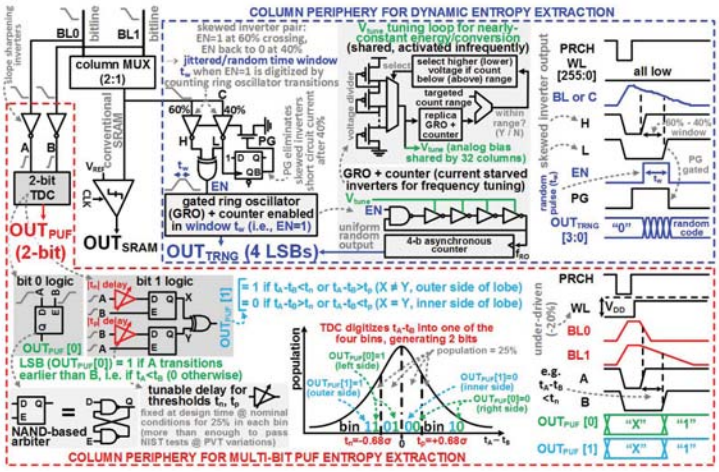


Figure 36.1.3: Column peripheral circuits for in-memory TRNG (top) and PUF (bottom).

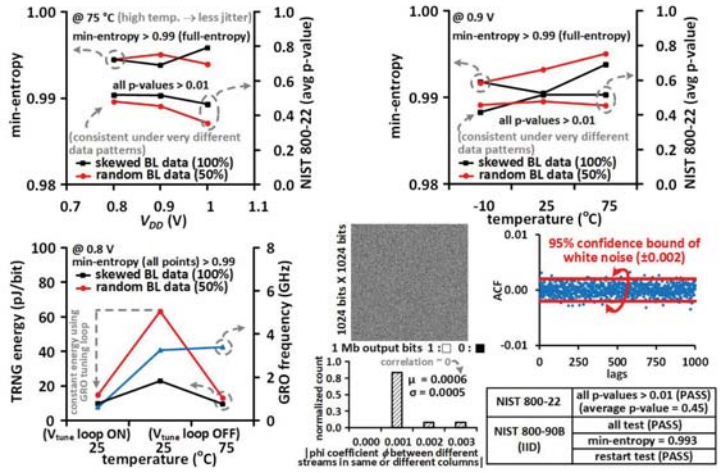


Figure 36.1.4: Measured dynamic entropy (TRNG) output characteristics with supply voltage (top-left), temperature (top-right), nearly-constant (minimum) energy operation across temperature (bottom-left) and randomness evaluation results (bottom-right).

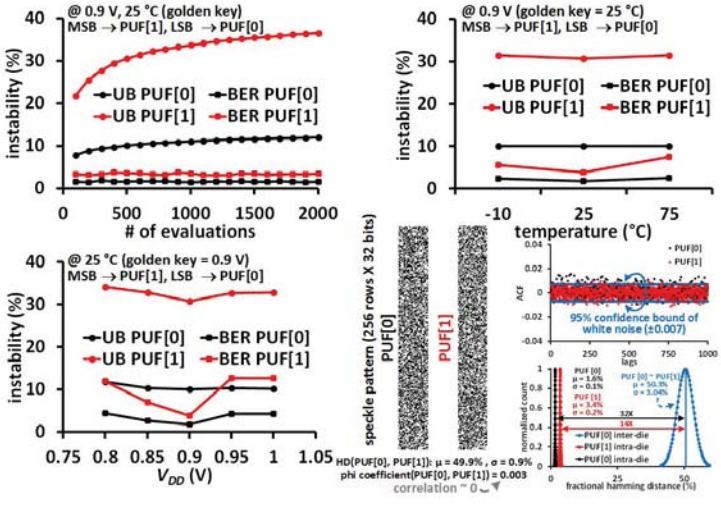


Figure 36.1.5: Instability vs. PUF evaluations (top-left), temperature (top-right), supply voltage (bottom-left) and statistics (bottom-right).

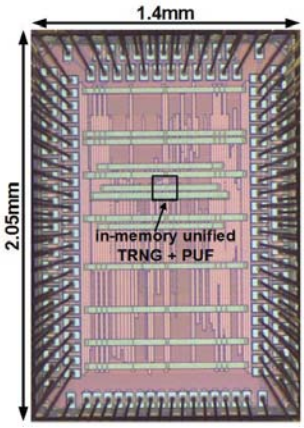
	This work (TRNG)	JSSC 2019 [1]	VLSI 2018 [7]	JSSC 2017 [2]	ISSCC 2017 [8]
technology (nm)	28	14	65	180	65
entropy source	SRAM BL leakage noise jitter	metastability	metastability	chaotic map	jitter
unified functions	TRNG + PUF + SRAM	TRNG + PUF	TRNG	TRNG + ADC	TRNG
calibration (functional)	NO	YES	YES	NO	NO
all-digital / in-memory	YES / YES	YES / NO	NO	NO	NO
area (10^4·F²)	0.016 [a]	0.3 [a]	2.4	0.14 [a]	0.22
supply voltage V_{DD} (V)	0.8 - 1.0	0.55 - 0.75	0.53 - 1.0	0.6 - 0.9	1.08 - 1.2
temperature (°C)	-10 - 75	70	-20 - 100	-	-
min energy (pJ/bit)	9.6	2.5	2.58	0.3	35.5
max throughput (Mbps)	3.6 [b]	1,480	86	0.27	9.9

[a] TRNG area overhead per random output stream
[b] Measured at 1 V, -10 °C, for 32 I/O SRAM and 1 bank (up-scalable)

	This work (PUF)	CICC 2020 [3]	ISSCC 2020 [6]	JSSC 2018 [4]	IRPS 2018 [5]
technology (nm)	28	130	28	65	40
entropy source type	SRAM bitcell read current	EE/CMOS hybrid SRAM	monostable NAND	SRAM PTAT diode	CMOS gate breakdown
unified functions	PUF + TRNG + SRAM	PUF + SRAM	PUF	PUF + SRAM	PUF
PUF area/bit (F²)	1,125 [a]	2,307 [a]	3,700	5,280 [a]	1,515
custom SRAM bitcell	NO	YES	-	NO	-
max # of bits/PUF cell	2	1	1	1	2
readout circuit	YES	YES	YES	YES	NO (array)
visual attack resilience	high	high	high	high	low
V_{DD} (V)	0.8 - 1.0	0.5 - 0.7	0.81 - 0.99	1.0	1.5
temperature (°C)	-10 - 75	-40 - 120	-40 - 150	15 - 85	25 - 125
native unstable bits (%) @ # of evaluations	11.9 (LSB), 36.5 (MSB) @ 2,000	2.71 @ 1,000	25 @ 8	5.39 @ 500	1.56, 4.98 @ 5
BER (%) @ nominal V/T	1.78 (LSB), 3.84 (MSB)	0.29	10.5	2.16	-
max throughput (Mbps)	10,160	96	128	0.01	DC sweep
PUF energy (fJ/bit)	78 @ 0.8 V	15.4 @ 0.6 V	2,960 @ 0.9 V	380 @ 1.0 V	-

[a] Considering total SRAM macro area (including peripherals for SRAM and PUF)

Figure 36.1.6: Comparison with state-of-the-art TRNGs (top) and PUFs (bottom) with best performance or feature in bold.



technology	28 nm
SRAM architecture	16 kb (8T 1R1W, logic rules)
bank organization	256 rows X 32 bits (2:1 MUX)
area (μm^2)	15,400 (SRAM + TRNG + PUF)
SRAM write throughput (Gbps)	17.6 @ 0.9 V
SRAM write energy (fJ/bit)	85 @ 0.9 V
SRAM read throughput (Gbps)	5.6 @ 0.9 V (3.2 → 20% WL underdrive)
SRAM read energy (fJ/bit)	92 @ 0.9 V
TRNG or PUF supply voltage (V)	0.8 - 1.0
TRNG throughput (Mbps)	2.2 - 3.6 @ -10 °C
TRNG energy (pJ/bit)	9.6 - 17.2
PUF throughput (Gbps)	3.2 - 10.16 (20% WL underdrive)
PUF energy (fJ/bit)	78 - 128

Figure 36.1.7: Die micrograph and measurement summary.

978-1-7281-9549-0/21/$31.00 ©2021 IEEE