Invited Paper: Ultra-low Energy Security Circuit Primitives for IoT Platforms

Sanu Mathew, Sudhir Satpathy, Vikram Suresh, Ram Krishnamurthy
Circuits Research Lab, Intel Corporation, Hillsboro, USA
sanu.k.mathew@intel.com

Abstract— **Low-area energy-efficient security primitives are key building blocks for enabling end-to-end content protection, user authentication in IoT platforms. This paper describes 3 designs that employ energy-efficient circuit techniques with optimal hardware-friendly arithmetic for seamless integration into area/battery constrained IoT systems: 1) A 2040-gate AES accelerator achieving 289Gbps/W efficiency in 22nm CMOS, 2) Hardened hybrid Physically Unclonable Function (PUF) circuit to generate a 100% stable encryption key. 3) All-digital TRNG to achieve >0.99 min-entropy with 3pJ/bit energy-efficiency.**

I. Introduction

The emergence of high-volume, low-cost Internet of Things (IoT) devices has motivated the development of compact, energy-efficient connected computing platforms. Since these devices depend on secure exchange of data between the sensors and the cloud, information security and privacy are critical features of SoCs that power these devices. Advanced Encryption System (AES) hardware accelerators, True Random Number Generator (TRNG) and Physically Unclonable Functions (PUFs) are primary circuit primitives that enable energy-efficient content protection, signature generation and attestation applications.

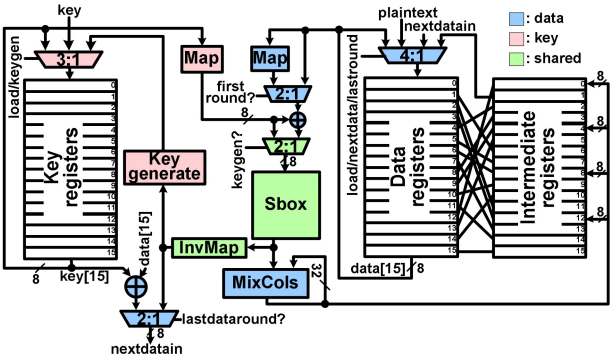
Security accelerators for high-performance microprocessors achieve high throughput at the expense of large die-area and power dissipation, rendering them unsuitable for use in area/power constrained mobile and wearable systems [1]. The energy and thermal constraints of such systems motivate lightweight low-cost hardware implementations of these security primitives with compact layout footprint and ultra-low leakage energy consumption. This paper describes three circuit design techniques that provide significant energy-efficiency and die-area reduction, while still providing adequate performance for IoT applications.

II. nanoAES

Advanced encryption standard (AES) is the de-facto symmetric-key cipher providing the security foundation of media content protection and memory encryption [1,2]. Conventional hardware accelerators use 128bit datapath, organized as 16 byte-slices with 64 wiring tracks for shift-row permutation to achieve single cycle round latency at the cost of high area and energy consumption of parallel round units, rendering them unsuitable for IoT systems.

A. Accelerator Organization

In contrast to conventional 16 byte-slice datapath designs, the nanoAES accelerator is organized around a 1byte Sbox circuit that computes the performance critical Galois-field inversion along with affine and inverse affine transformations (Fig. 1). Plain-text and key bytes are mapped from the standard-specified prime-field of $GF(2^8)$ to a composite-field of $GF(2^4)^2$ prior to first round iteration, allowing all downstream Sbox and Mixcolumns computations occur in the composite-field, thus amortizing the cost of field-transformation over 10 rounds [3]. The circuit processes 1byte of plain/cipher text every cycle, generating 128bit output after 16 cycles that is stored in shift-row permuted format in a shift register. Following round processing, the accelerator switches to key generation mode that uses existing Sbox and other datapath logic to compute the key for the subsequent round. This organization enables reuse of the dominant combinational circuits in the accelerator, resulting in 18% area reduction over a separate datapath design.



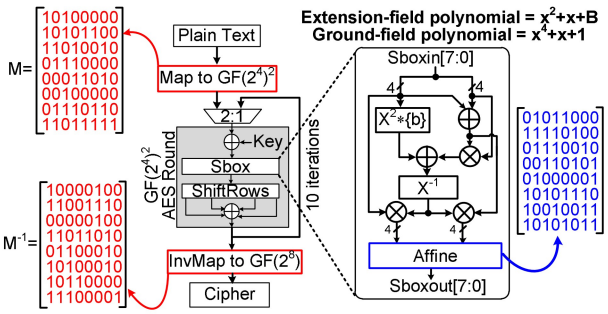


Fig. 2: Composite-Field $GF(2^4)^2$ Sbox

The performance critical Sbox circuit is optimized by avoiding direct computation of inverse in $GF(2^8)$. Instead, plain-text data and keys are mapped to a composite-field of $GF(2^4)^2$ using mapping transformations in the first round. All subsequent iterations are computed in this composite-field where elementary operations like squaring, additions and inverse operations involve simpler 4-bit logic (Fig. 2). The native Sbox design removes the mapping and inverse-mapping transformations from the performance critical AES round logic, this reducing critical-path delay by 12%.

978-1-5090-6023-8/17/$31.00 ©2017 IEEE

B. *Polynomial Optimization*

Although operations in the prime-field are defined by the Rjindael AES polynomial ($x^8+x^4+x^3+x+1$), the arithmetic complexity of the composite-field accelerator datapath is determined by a pair of polynomials (ground-field and extension-field) selected at designer's discretion. These polynomials not only determine the mapping/inverse-mapping transforms but also the logic complexity of the Sbox and MixColumns circuits. A framework to exhaustively evaluate the entire design space of all 2880 valid polynomial pairs was developed to further optimize the accelerator datapath.

Results of the polynomial-based area optimization using 22nm tri-gate high-k/metal-gate CMOS standard-cell library shows accelerator layout area grouped into three regions representing the 3 valid ground-field polynomials (Fig. 3). Within each region, the extension-field coefficients are swept from 0x0 to 0xF, with a 1.3× spread between the largest and smallest area polynomial-pair. The lowest area design occupying 2200µm^2 was obtained for a ground-field polynomial of x^4+x^3+1 and extension-field of x^2+6x+9, representing a 9% area reduction over previously-reported polynomial choices [1]. Decrypt datapath optimization results show a 1.4× spread in area, with 5% area reduction compared to [1] for the lowest-area design occupying 2736µm^2 obtained for a polynomial-pair of x^4+x+1 and x^2+2x+E.

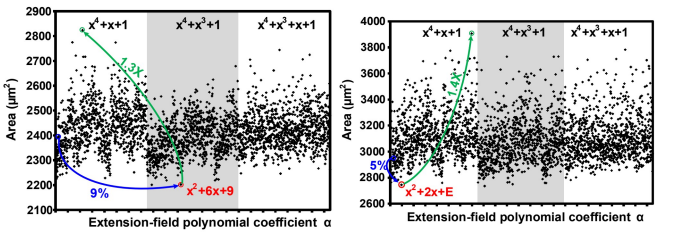


Fig. 3: Polynomial vs. Area trade-off

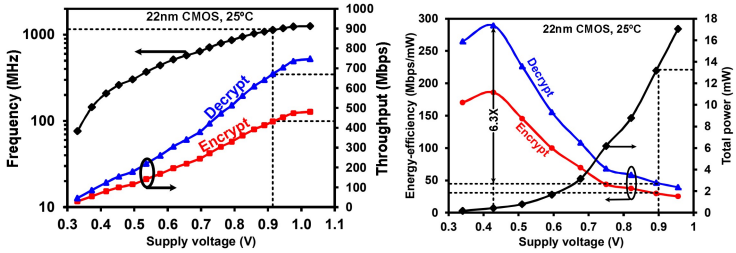


Fig. 4: NanoAES 22nm CMOS measurements

The arithmetically optimized composite-field Sbox circuit based AES accelerator with unified round compute and key expansion datapath was fabricated in a 22nm tri-gate high-K/metal-gate CMOS technology with encrypt/decrypt designs occupying 2200/2736µm^2 (1940/2090 equivalent gate count) respectively. At nominal supply of 0.9V, the design operates at 1.133GHz resulting in 432/631Mbps AES-128 encrypt/decrypt throughput (Fig. 4). Ultra-low voltage circuit optimizations enable reliable operation over a wide dynamic supply voltage down to 340mV, with peak energy-efficiency of 289Gpps/W measured at near-threshold supply of 430mV.

III. PUF

Physically Unclonable Functions (PUFs) are low-cost cryptographic primitives used for generation of a stable, repeatable device-specific secure key [4, 5, 6]. They represent a paradigm shift in physical security by transitioning from the conventional approach of explicitly programming digital IDs into fuses post manufacturing, to a new approach of generating the IDs by exploiting intrinsic characteristic of devices on the die. However, the static ID derived from PUF circuits by harnessing process and manufacturing induced variations is not inherently stable, and cannot be used in its raw form for security applications that require a unique ID that can be accurately generated across voltage and temperature fluctuations and long-term device aging. Hence, techniques that improve PUF stability are critical for high volume production targeting IoT applications.

A. *PUF based Secure Key Generation*

Fig. 5 shows an overview of a PUF based key generation platform. In the first evaluation of the array during tester-time operation, a golden key is generated. An ECC signature computed from this golden key and stored on-die using fuses can be used to recreate the golden key with 100% accuracy in subsequent evaluations. A PUF circuit that produces an output bit with a high degree of stability is necessary to minimize ECC overheads. The PUF entropy source is based off of a hybrid circuit that consists of a pair of cross-coupled inverters that are pre-charged into an unstable state. During the positive phase of the clock, the circuit evaluates to one of the 2 stable states, determined by the relative strengths of variation-impacted minimum size inverters in the cross-couple. Additionally, random variations in precharge transistors and clock delay inverters produce mismatches in clock rise and arrival times. These introduce a transient dimension of uncertainty into PUF resolution dynamics. This hybrid approach combines the stability and compactness properties of SRAM-based PUFs, with the resistance to invasive probing attacks that we get with delay-based PUFs.

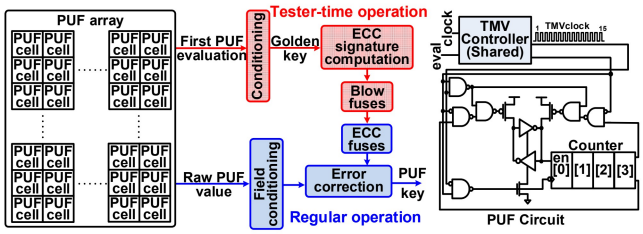


Fig. 5: Secure key generation system and PUF circuit

PUF cells that have insufficient net random variation generate unstable bits. These unstable bits resolve to either 0 or 1 based on thermal noise or voltage and temperature conditions. Such bits undermine PUF reliability and need correction for reliable key generation. Repeated evaluations of a 250Kbit PUF array fabricated in a 22nm tri-gate high-K metal-gate CMOS technology across 0.7V-0.9V shows a worst case bit-error rate (BER) of 8.5%. The overhead for storing an ECC signature is prohibitively expensive at such BER, and hence techniques to stabilize the PUF array are needed for practical applications.

B. *Conditioning Tecniques for Stability Improvement*

PUF cells contributing towards BER are handled using 3 conditioning schemes: temporal-majority-voting (TMV), burn-in hardening, and soft dark-bit masking. TMV stabilizes noisy cells that have a low, non-zero probability of becoming unstable by computing the quantized mean of responses within a voting window. This is accomplished using a counter that is selectively incremented when the PUF cell evaluates to "1" within a window of consecutive evaluation cycles (Fig. 5). A threshold set midway is then used to bin the total counts into "0" or "1". Although this approach effectively corrects mildly unstable cells, the need for exponentially larger voting windows and higher sensitivity of PUF circuit resolution towards thermal noise reduces its efficacy for highly unstable cells. Besides, a larger voting window also increases key generation latency and

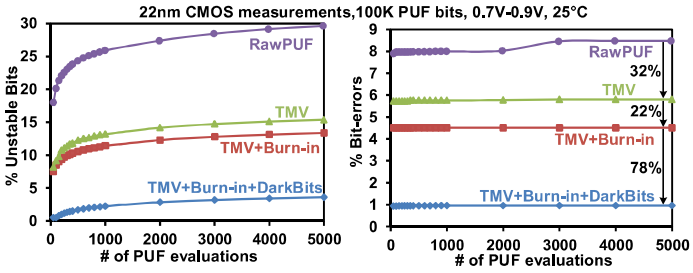


Fig. 6: Unstable bit and Bit-error measurements

reduces energy-efficiency because of multiple evaluations. 22nm CMOS measurements at 0.7V-0.9V, 25^{0}C indicate that a 4b counter based TMV circuit reduces BER of the 250Kbit PUF array to 6%. Further reduction in BER can be accomplished using burn-in hardening. Burn-in is a standard test procedure of subjecting dies to elevated supply voltage at high temperature in a controlled environment to identify chips that are prone to early failure in field. Although conventional burn-in process degrades devices leading to performance loss, we use a technique that selectively ages devices in the hybrid PUF circuit leading to overall improvement of array stability. The PUF circuit features a write-back scheme that enables biasing the devices in either cross-coupled inverters to a complementary state during tester operation. This directed accelerated aging reinforces pre-existing bias in the PUF cell by leveraging NBTI/PBTI degradation. Furthermore, it also injects a bias into the clock delay-path by aging the buffers and pre-charge transistors in a direction favoring PUF stability, thus reducing BER by 22% to 4.6% (Fig. 6).

TMV and burn-in hardening handle most of the mildly unstable cells, leaving behind the highly unstable ones that contribute towards the remaining BER. These bits manifest themselves as unstable within a few cycles of evaluation and are identified as dark-bits during regular operation, and can be excluded from participating in key generation. As opposed to storing dark-bit locations in NVM, a soft-masking scheme is used that recreates the mask at every PUF start-up. This not-only improves security by removing a potential tamper-point, but also reduces cost by eliminating extra storage requirement. Dark-bit measurements with a window of 100 evaluation cycles identifies 11% of the PUF array as unstable, reducing overall BER to 0.97% (Fig. 6) with 190fJ/bit energy-efficiency at 0.75V, 1GHz. ECC circuits use BCH coding to correct this residue bit-error resulting in 100% stable key generation.

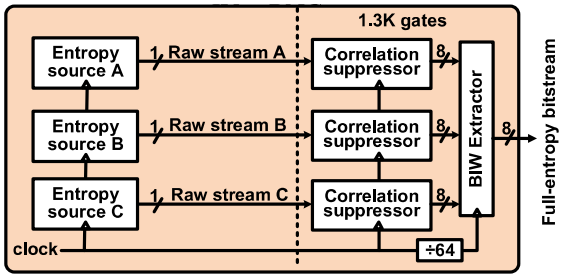


Fig. 7: Full-entropy µRNG organization

TRNGs exploit non-deterministic physical phenomenon as source of randomness to generate high entropy bitstreams that are used in many security applications for generating keys, process IDs, initialization vectors, and nonces [7,8,9]. Conventional TRNGs overcome non-idealities and serial-correlations exhibited in raw random streams using keyed functions like HMAC, CMAC, CMC-MAC-AES for post processing. The large area (<30k gates) and energy-consumption of such approach makes it unsuitable for use in IoT platforms. As a solution, an ultra-lightweight full entropy TRNG is presented that combines the entropy of three independent self-calibrating entropy sources using compact extractor circuits to generate cryptographic quality random bitstream that is indistinguishable from an ideal unbiased random source (Fig. 7).

A. *Self-Calibrating Entropy Source*

µRNG harvests randomness from metastability resolution uncertainty of a matched cross-coupled inverter pair that is precharged to an unstable high-gain state, and evaluated to generate a bit every cycle (Fig. 8a). Although, the direction of resolution depends on the relative magnitudes of thermal noise at the inverter nodes, manufacturing and process induced systematic mismatches or voltage/temperature fluctuations

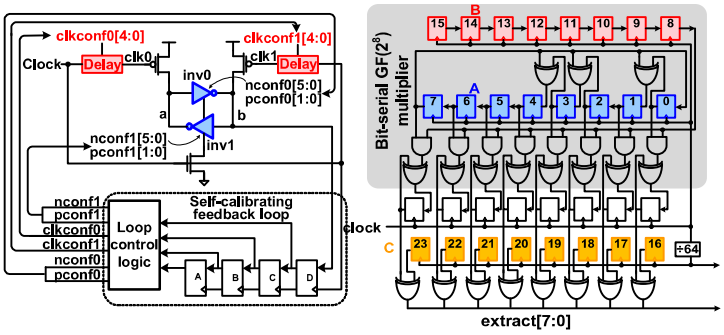


Fig. 8: (a) Entropy source (b) BIW Extractor

during run-time can disrupt ideal operation by biasing the bitstream towards a "1" or "0". A self-calibrating control loop examines a window of 4 consecutive output bits to detect any bias, and configures the device strengths and clock arrival delays on either side of the cross-couple to tune the circuit towards a state of higher entropy. Coarse grained tuning occurs at power-up where the calibrating FSM updates one of the two "nconf"

978-1-5090-6023-8/17/$31.00 ©2017 IEEE

counters to modulate NMOS strength until a transition is detected at the output. To account for the possibility of over compensation, the FSM then updates the "pconf" counters until the circuit generates a transition in the opposite direction. Following this, the circuit enters fine grained tuning where the "clkconf" counter values are constantly adjusted to keep the system dithering in the high-entropy zone. In rare events of counter saturation, the control is reverted back for coarse grained tuning, and calibrating steps are repeated. The all-digital design results in a compact layout spanning 1088µm^2 in a 14nm tri-gate CMOS process, enabling usage of simple extractor circuits that rely on multiple entropy sources to generate a random stream.

B. Lightweight Entropy Extraction Circuits

In contrast to traditional area expensive AES cipher based extractors (30k gates), µRNG uses a BIW extractor (1.3k gates) that extracts entropy from 3 independent sources using simple Galois-Field arithmetic operations (Fig. 8b). Statistical independence of the 3 sources is ensured by use of correlation suppressors, implemented as under sampled 29-deep XOR feedback shift-registers. Non-overlapping bits from 3 correlators (A, B, C) undergo multiply-add operation in BIW extractor using a bit-serial GF(2^8) multiplier. The all-digital compact design achieves 3pJ/bit energy-efficiency at 0.75V, 1.3GHz, and operates over a wide supply voltage down to 300mV generating random stream with lower-bound min-entropy >0.99.

V. SUMMARY

Lightweight energy-efficient security primitives are essential to enable content protection and user privacy in SoCs targeted for area/battery constrained IoT devices. A single Sbox based arithmetically optimized composite field nanoAES accelerator, secure key generation using a hybrid PUF circuit that leverages burn-in induced aging for stability improvement, and an all-digital variation tolerant compact TRNG with low area extractor circuits have been demonstrated (Fig. 9) in 22nm and 14nm tri-gate CMOS. Ultra-low voltage circuit techniques, arithmetic optimizations, and micro-architectural enhancements result in measured energy-efficiency of 289Gbps/W, 190fJ/bit, and 3pJ/bit for AES-128, PUF and µRNG respectively.

ACKNOWLEDGMENT

The authors thank M. Haycock, M. Mayberry, Vivek De, J. Tschanz, S. Iyengar, A. Rajan, and R. Parker for valuable discussions and encouragement.

REFERENCES

[1] S. Mathew et al. "53Gbps Native GF(2^4)^2 Composite-Field AES-Encrypt/Decrypt Accelerator for Content-Protection in 45nm High-Performance Microprocessors", *IEEE Journal of Solid-State Circuit*, pp. 767-776, v. 46, no.4, Apr 2011.

[2] C. Tokunaga and D. Blaauw, "Secure AES engine with a local switch-capacitor current equalizer," *ISSCC Digest of Technical Papers*, pp. 64-65, Feb. 2009.

[3] S. Mathew et al. "340 mV–1.1 V, 289 Gbps/W, 2090-Gate NanoAES Hardware Accelerator With Area-Optimized Encrypt/Decrypt GF(2^4)2 Polynomials in 22 nm Tri-Gate CMOS", *IEEE Journal of Solid-State Circuit*, pp. 1048-1058, v. 50, no.4, Apr 2015.

[4] J. Li, and M. Seok, " A 3.07µm^2/bitcell Physically Unclonable Function with 3.5% and 1% Bit-Instability across 0 to 80°C and 0.6 to 1.2V in a 65nm CMOS," *Symposium on VLSI Circuits Digest of Technical Papers*, pp. 250-251, June 2015.

[5] S. Mathew et al. "A 0.19pJ/b PVT-variation-tolerant hybrid Physically Unclonable Function circuit for 100% stable secure key generation in 22nm CMOS," *ISSCC Digest of Technical Papers*, pp. 278-279, Feb. 2014.

[6] S. Satpathy et al. "A 13fJ/bit probing-resilient 250K PUF array with soft dark-bit masking for 1.94% bit-error in 22nm Tri-gate CMOS," *IEEE Proc. of the ESSCIRC*, pp. 239-242, Sept. 2014.

[7] S. Mathew et al., "µRNG: A 300-950mV 323Gbps/W all-digital full-entropy TRNG in 14nm FinFET CMOS", *IEEE Proc. of the ESSCIRC*, pp. 116-119, Sep., 2015

[8] S. Mathew et al. "2.4Gbps, 7mW All-digital PVT-Variation Tolerant True Random Number Generator for 45nm CMOS High-Performance Microprocessors," *IEEE Journal of Solid-State Circuit*, v.47, no.11, pp.2807-2821, Nov. 2012.

[9] K. Yang et al., "A Robust -40 to 120°C All-Digital True Random Number Generator in 40nm CMOS", *Symposium on VLSI Circuits Dig. Tech. Papers*, pp. 248-249, June 2015.

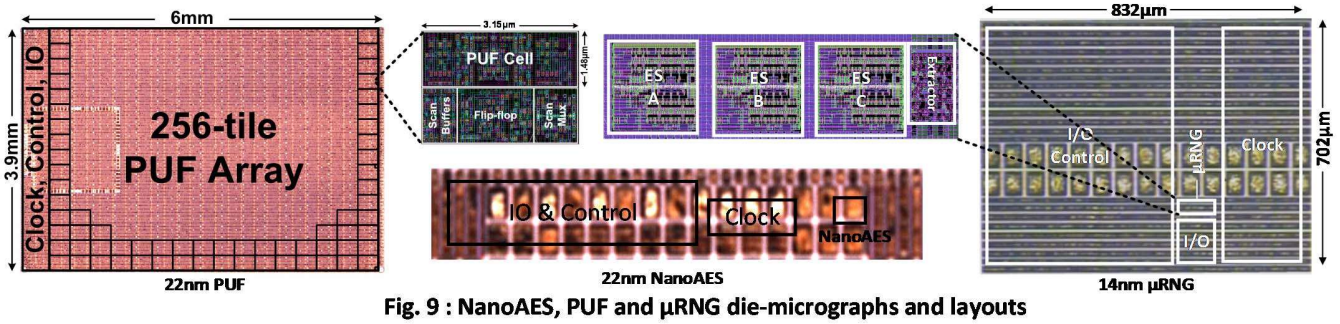


Fig. 9 : NanoAES, PUF and µRNG die-micrographs and layouts