Understanding Sources of Variations in Flash Memory for Physical Unclonable Functions

Sarah Q. Xu, Wing-kei Yu, G. Edward Suh, Edwin C. Kan
School of Electrical and Computer Engineering
Cornell University
Ithaca, NY 14853 USA
{qx33, wsy5, gs272, eck5}@cornell.edu

***Abstract*—This paper provides detailed characterizations of physical sources behind Flash memory based Physical Unclonable Functions (FPUFs). Universal process variations in Flash physical systems are identified and decomposed into layout, intrinsic, stress and bit-wise fluctuation sources. The study shows the understanding of systematic variations and noise sources are essential for improving the security and reliability of FPUFs. Bit-wise variations are proven to be originated mainly from random dopant fluctuation, which is indeed truly random and impossible to clone. Overall, this paper provides a theoretical foundation for the security of FPUFs whereas previous PUF studies rely only on experimental evidence for its security and entropy.**

***Index Terms*—Physical Unclonable Function, Flash Memory, Physical Modeling, Variation Sources**

I. Introduction

Physical Unclonable Functions (PUFs) is a physical one-way function that provides unique challenge-response pairs based on the intrinsic, uncontrollable but reproducible randomness of the implementing device. So far, studies on PUFs have been largely focused on experiments to demonstrate that there exist enough variations to distinguish individual chip fingerprints [1-4]. Few detailed characterizations of the physical mechanisms behind the variations have been incorporated into these studies [5].

Unfortunately, without concrete definition and modeling of the physical variation sources, it is almost impossible to generalize the experimental conclusions to a variety of different devices, circuits and systems. In particular, technologies used to implement silicon based PUFs change very quickly due to the drastic scaling of the feature size [6, 7]. Only by physical modeling of the PUF physical origins, we can guarantee that proper PUF characteristics can be extended to different technology nodes and other manufacturers.

This paper presents a semiconductor device level modeling and analysis to understand underlying physical mechanisms behind variations in Flash-memory PUFs (FPUFs), and discusses their implications for designing secure and reliable protocols. To the best of our knowledge, this study is the first to underpin the physical mechanisms of FPUFs.

II. Overview of The Flash PUF

Unlike prior PUFs, FPUFs [2, 8] do not require any custom hardware circuits, and the Open NAND Flash interface (ONFi) [9] sufficiently provides universal extraction methods for most commercial Flash chips. Experiments on producing FPUFs have been successfully carried out on commercial off-the-shelf (COTS) components [2].

FPUFs can be extracted based on a technique called partial or aborted programming. The initial and after-erase V_{th} for a Flash device are different from cell to cell due to process variations. This difference also induces changes in Fowler-Nordheim tunneling currents (I_{FN}) during programming operations. Therefore, each cell will require different number of partial program pulses hence the specific program time to change state. This partial program number is sufficiently consistent in program/erase cycles for the same cell, but varies distinctively from bit to bit, page to page and chip to chip, and therefore can be considered as a unique PUF function [2, 10].

FPUFs have several practical advantages over conventional PUF implementations. In addition to the wide applicability, FPUF extractions do not require a power cycle compared to the PUFs based on bi-stable elements [5, 10]. Since Flash is one of the most aggressively scaled technologies, FPUF is also superior in bit capacity compared to other PUFs [11].

III. Manufacturing Variations

A. Layout Variations

Average page FPUF is obtained by averaging the bit-level partial program pulse numbers for a particular page. This average page PUF is then plotted for every page across the same block for several blocks on various chips. Results from

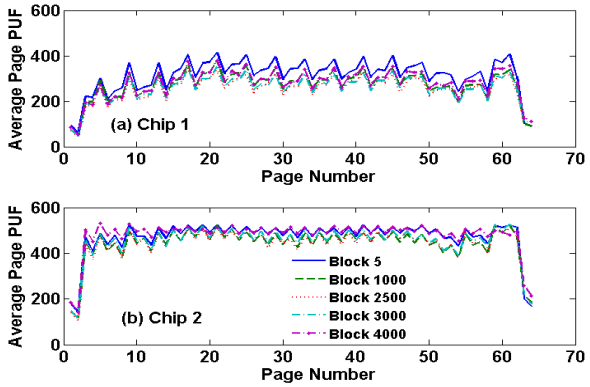

Figure 1. Layout variations in Flash chips introduce systematic fluctuations in average page FPUFs

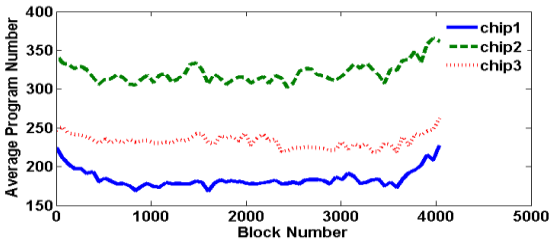

Figure 2. Spatial variations result in off-sets between various chips. Layout variations can be seen from the systematic fluctuations.

two identical Hynix 50nm chips are presented in Fig. 1.

A consistent systematic variation can be seen among average page FPUFs for all the blocks. A cyclic fluctuation is highly correlated among blocks in the same chip with a Pearson correlation coefficient [8] as high as 0.99. Similar fluctuation patterns have been found in other chips with the same part number, and the correlation coefficient is around 0.88.

Because this page-wise fluctuation is consistent with all the blocks and also highly correlated for various chips, the contributor of this systematic variation has to come from a universal variation source for all blocks and chips fabricated in the same manufacturing process. Since it cannot come from spatial variation alone that varies from chip to chip, it is reasonable to attribute this prominent systematic effect to the layout design.

B. Spatial Variations

In order to observe the spatial variation components [12] of the FPUFs, block averages across three similar chips are plotted in Fig. 2. Each curve represents the average partial program pulse number for over 4000 blocks throughout the chip. The fluctuation in the block average FPUF across the chip is highly correlated among all three chips, with an average correlation coefficient of 0.76. Because it is very unlikely that all three chips come from the same spatial wafer location, the high correlation should be additionally attributed to the systematic layout-induced variations.

However, there is a clear offset in the block average among three chips. This offset is not a layout systematic component and very likely comes from the die spatial location difference, which can be attributed to a spatial variation component of the manufacturing variations.

C. Intrinsic Random Variations

Flash page-level fingerprints are unique and robust enough to be used to authenticate individual chips [2, 8]. It has been reported that the average correlation coefficient for the same page is on the order of 0.97, and fingerprints extracted from different pages, either the same page from different chips or different pages from the same chip, have an average correlation coefficient around 0.0076 [2].

Previous studies [2, 8] all simply contributed this randomness to general intrinsic process variations of Flash memory, but no detailed characterization and modeling have been performed. It is crucial to understand this source of random variation in order to determine if the uniqueness and

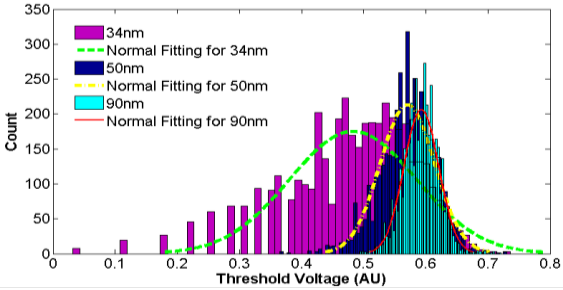

Figure 3. Scaling effect of the extracted threshold voltage distributions from three different technology generations.

robustness of the FPUF are indeed universally applicable, and not just a phenomenon presented in the limited selection.

The predominant intrinsic variation sources in sub-100nm MOS devices include random dopant fluctuation (RDF) [6, 7, 13] and line edge roughness (LER) [7]. Due to the additional floating gate and control dielectric, Flash memory can suffer more severely from RDF and less from LER compared to conventional logic devices due to the larger effective oxide thickness (EOT). In order to analyze the physical origins of the randomness, FPUF distributions are translated to threshold voltage distributions via tunneling current during programming, and then fitted to RDF analytical models [6, 13].

Since RDF is more severe as device scales further, a device with a smaller feature size should result in a larger standard deviation, which is observed in Fig 3 for chips from three technologies of 34nm, 50nm and 90nm. According to theoretical RDF predictions [7, 13], standard deviation ratio between 90nm and 50nm devices should be around 0.7, while the standard deviation ratio between 50nm and 34nm devices is roughly 0.58. Our experimental extractions yield average ratios of 0.68 and 0.55, respectively.

D. Implications for PUF Designs

TABLE 1. DIEHARD TESTS ON FPUFS

Test Type	**P values for FPUFs with and without systematic variations**			
	FPUFs from same chip		Concatenated pages from multiple chips	
	With	Without	With	Without
OQSO	1.0000	.9231	1.0000	.5414
	1.0000	.5695	1.0000	.9045
	1.0000	.9539	1.0000	.8659
DNA	1.0000	.5906	1.0000	.8440
	1.0000	.7040	1.0000	.3938
	1.0000	.2569	1.0000	.0250
	1.0000	.6446	1.0000	.0507
	1.0000	.4831	1.0000	.8076
	1.0000	.9977	1.0000	.8180

Systematic variations can degrade the uniqueness and entropy of FPUFs. In order to assess the effect of the layout and design induced variations, Diehard randomness tests [14] are performed for FPUFs containing systematic components and for FPUFs with systematic component removed from their page average. The tests are performed between FPUFs from the

same chip and FPUFs created by concatenating same-page FPUFs from multiple chips with highly correlated systematic variations. OQSO (Overlapping-Quadruples-Sparse-Occupancy) and DNA are two randomness tests that sequentially sample 10 and 2consecutive bits from the 32-bit integer data respectively, and p values close to one suggesting data fail the specific category. Sample p-values are shown in Table 1, where each row represents test result from a different selected group of bits. By removing the systematic components from the FPUF bits, improvement on the randomness is evident in both cases of OQSO and DNA tests.

On the other hand, determining RDF as the major source of FPUF variations proves that FPUFs can be extracted from any Flash processes, since it is a universal phenomenon and cannot be fully controlled or cloned by today's fabrication technology. Extensive modeling attempts have also been conducted to recreate RDF effects [6, 15], but today's computing power is still insufficient for carrying out 3-D "atomistic" simulations on a large statistical scale to accurately depict the RDF behavior.

IV. Variations in the Field

A. Stress-induced Variation

Previous studies [2, 3] have suggested that repetitive program and erase (P/E) cycles can alter the partial program time of Flash cells due to cyclic endurance aging effects. This stress effect can be isolated from the manufacturing variations because RDF and LER should not be directly affected by P/E stress. Fig 4(a) illustrates how page average changes as P/E cycles increase. Correlation coefficients of fresh and stressed FPUFs from the same pages are plotted in Fig 4(b). The decrease in both correlation coefficients and average partial program numbers suggest that FPUFs are becoming increasingly different as the P/E stress level rises, which is coherent with stress induced leakage current (SILC) and bias-temperature instability (BTI) [16].

B. Random Telegraph Noise

When the same PUF bits are measured multiple times, the bit-wise partial program times have non-negligible fluctuations

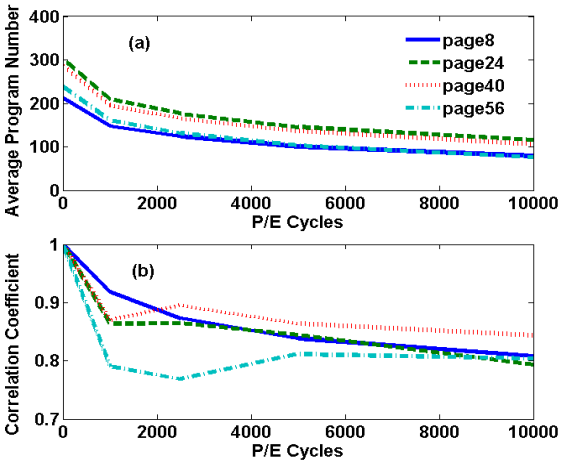

Figure 4. (a) P/E stress effect on average partial program numbers and (b) FPUF correlation coefficients.

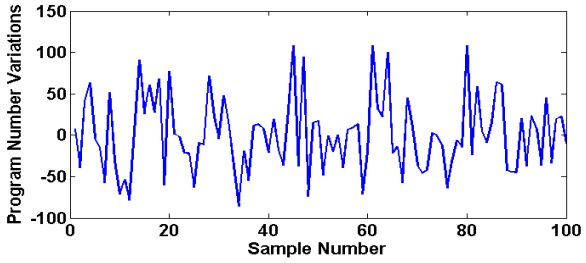

Figure 5. Bit-wise fluctuations for multiple FPUF measurements.

as depicted in Fig 5. This can affect the reproducibility of FPUFs when used as crypto keys or authentication directly, but can be useful for statistical query models with noises [17].

An example of bit-wise fluctuation is analyzed via power spectral density; results are plotted in Fig 6(a). A clear $1/f^x$ relationship can be observed, and line fitting yields an x value around 1.8. These power coefficients were then extracted for multiple bits within a page and the results are shown in Fig 6(b). The average x is around 1.7 and most of the values are within 1 to 2. This proves that bit-wise fluctuations in general display shot noise behavior. In addition, this relationship closely resembles a $1/f^2$ characteristic, which corresponds to the random telegraph noise (RTN) behavior very well, especially for low frequencies [18].

C. Block-level Erase Effect

During the P/E experiments, significant number of bits fluctuates together in the same P/E cycle. Sample correlation coefficients on how these bits fluctuate among 5000 FPUF measurements are plotted in Fig 7. A substantial percentage of bits have fluctuation correlation coefficients around 0.5.

This suggests that bit-wise fluctuation is not purely due to RTN. Observation from the same experiments have confirmed that conventional full erase operation dynamically adjusts the erase time during each block erase, and therefore can add an undesirable global bias to the bit-wise fluctuations. An erase operation with fixed erase time was performed in order to remove the block erase bias. The resulting correlation drops significantly compared to using dynamically adjusted full erase operations, as shown in Fig 7.

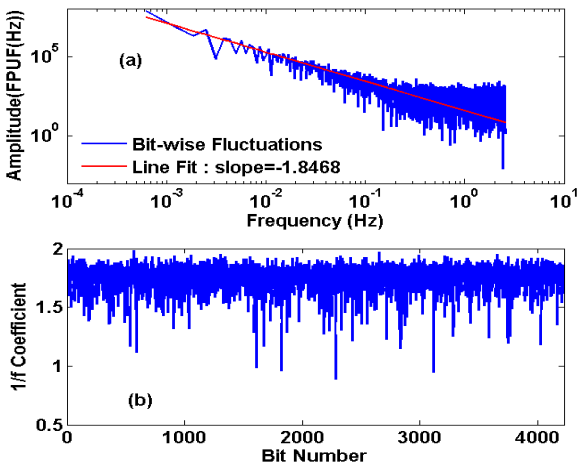

Figure 6. (a) Power spectral density of bit-wise fluctuations and (b) corresponding $1/f^x$ coefficient distribution.

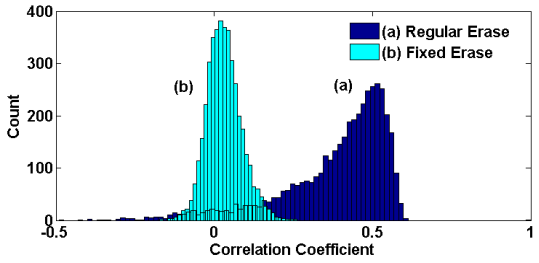

Figure 7. Bit-wise fluctuation correlation distributions obtained with (a) regular erase or (b) fixed erase.

D. Implications for FPUF Designs

Bit-wise fluctuation originating from RTN is globally presented in all Flash memory devices. Therefore, reduction of this variation over time will be crucial to improving the reliability and reproducibility of FPUFs.

The fluctuations are generally split into two levels caused by capture and emission of the trapped carriers [18], which provides means of reducing the FPUF variations through averaging several measurements from the same physical system. This RTN effect also illustrates the improper assumption of independence when low correlation coefficients are extracted, as the fluctuations from true randomness can mask other systematic components.

Fig 8 illustrates the reproducibility as a function of number of averaging measurements to extract FPUF. The percentage of the partial program number variation of individual bits is drastically reduced by averaging 10 measurements. Global erase, on the other hand, can add undesirable bias. Fig 9 illustrates how erase biasing can reduce the consistency of FPUF due to its dynamic nature.

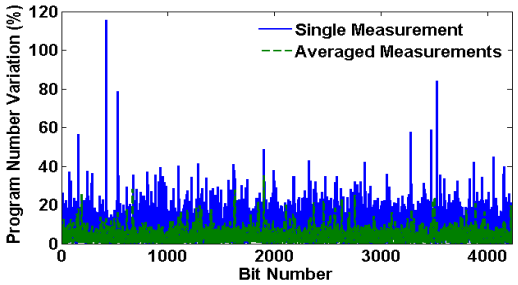

Figure 8. Percentage of partial program number variations of FPUF responses obtained between single measurements and between averaged measurements.

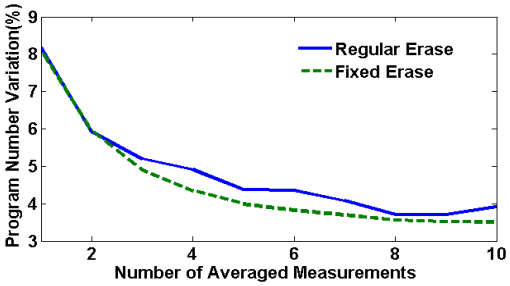

Figure 9. Average percentage bit-wise fluctuations between FPUFs with full erase and fixed erase against number of averaged measurements.

V. Conclusions

This study illustrates the importance of characterizing and understanding the physical mechanisms behind FPUFs. Strong systematic variations can severely degrade the uniqueness and entropy of the FPUF bits, and should be properly removed in appropriate security applications. Bit-wise FPUF fluctuation is caused by inherent RTN during operations, and ways to improve FPUF designs in both uniqueness and reliability are discussed. Erase fluctuation can add undesired global biases, but can be alleviated by fixing the erase time.

References

[1] G. E. Suh et al., "Physical unclonable functions for device authentication and secret key generation," *DAC*, 2007.

[2] Y. Wang et al., "Flash memory for ubiquitous hardware aecurity functions: True random number generation and device fingerprints," *Proc. of the IEEE Symp. on Sec. and Privacy,* 2012.

[3] Y. Wang et al., "Hiding information in flash memory," *Proc. of the IEEE Symp. on Sec. and Privacy*, 2013.

[4] B. Gassend et al., "Silicon Physical Unknown Functions," *ACM CCS*, 2002.

[5] D. E. Holcomb et al., "Initial SRAM state as a fingerprint and source of true random numbers for RFID Tags," *Proc. of the Conf. on RFID Security*, Jul. 2007.

[6] A. Asenov, "Random dopant induced threshold voltage lowering and fluctuations in sub 0.1 micron MOSFETs: A 3-D "atomistic" simulation study," *IEEE Trans. Elec. Dev*., vol. 45, pp.2505 -2513 1998.

[7] Y. Ye et al., "Statistical modeling and simulation of threshold variation under dopant fluctuations and line-edge roughness," *DAC*, June, 2008.

[8] P. Prabhu et al., "Extracting device fingerprints from flash memory by exploiting physical variations," *Trust and Trustworthy Computing, Lecture Notes in Computer Science*. Springer Berlin / Heidelberg, 2011.

[9] Open NAND Flash Interface Specification. Hynix Semiconductor, Micron Technology.

[10] R. Maes et al., "Intrinsic PUFs from flip-flops on reconfigurable devices," *Workshop on Information and System Security*, 2008.

[11] P. Koeberl et al., "Experimental evaluation of physically unclonable functions in 65 nm CMOS," *IEEE Euro. Solid-State Circ. Conf.*, Sep. 2012.

[12] E. Chang et al, "Using a statistical metrology framework to identify systematic and random sources of die- and wafer-level ILD thickness variation in CMP processes," *IEDM Tech. Dig*., Dec. 1995.

[13] H. S. Wong et al., "Three dimensional 'atomistic' simulation of discrete random dopant distribution effects in sub-0.1 mm MOSFET's," *IEDM Tech. Dig*., 1993.

[14] G. Marsaglia, "The Marsaglia random number CDROM including the Diehard battery of tests of randomness", Florida State University. 1995.

[15] Y. Su et al., "A 1.6pJ/bit 96% stable chip-ID generating circuit using process variations," *IEEE Int. Solid-State Circ. Conf*., Feb. 2007.

[16] S. Kamohara et al., "Deep-trap SILC model for nominal and weak oxides," *Proc. IRPS*, 1998.

[17] A. Blum, et al, "Noise-tolerant learning, the parity problem, and the statistical query model," *J. ACM*, vol. 50, no. 4, 506–519, 2003.

[18] M. J. Uren et al., "1/f and random telegraph noise in silicon metal-oxide-semiconductor field-effect transistors," *Appl. Phys. Lett*., vol. 47, no. 11, pp.1195 -1197, 1985.