In typical cryptographic applications, the secret keys are stored in volatile or non-volatile memory. In the latter case they remain in-memory and can be retrieved even when power is turned off. Even volatile memory is vulnerable to attacks if one has physical access to it. Thus the traditional approaches to key storage are not favored, especially in high security applications. A new concept, known as Physically Unclonable Functions (PUFs) has been recently investigated to mitigate this problem. The principal idea in a PUF is that the cryptographic key is not stored in memory. Instead, the binary data comprising the key are hidden in the form of unique analog identifiers within the hardware and are available only upon execution of the decrypting software on a designated authorized integrated circuit. This ensures that the key is completely unavailable when the circuit is powered down and conversely, when the circuit is turned on these analog values are converted into a binary code and are used as if they were a key stored in memory. In this article, the basic concepts behind PUFs are analyzed, highlights of important results from the recent literature are summarized, and application perspectives are presented.

I. PHYSICALLY UNCLONABLE FUNCTIONS: BACKGROUND

Modern society has an insatiable need for technology. Virtually every task is performed based on some sort of technology whether it is communications, business deals, financial transactions, personal identification, security and so on. It has always been necessary to have a sense of security when performing such tasks. Not only sensitive applications such as military operations, defense systems, and banking transactions require it but also for normal applications, people want privacy and information to be confined to the intended recipient only. Even before modern technologies had evolved, people were using some sort of cryptography to protect their communications. Cryptography is the process of converting a message, information, content or data into some unreadable form so that unintended users may not be able to extract any piece of useful information from it. Only the intended user should be able to extract the original form by decrypting it. The first such form of written cryptography has been found to be used by an Egyptian scribe in around 1900 B.C. Thereafter it has developed in various ways in terms of applications, purpose, types and security levels.

The main purpose of cryptography is to protect intellectual property, personal identity, and sensitive information from an intruder or unintended user. In 1975, the National Bureau of Standards (NBS, now National Institute of Standards and Technology (NIST)) proposed the Data Encryption Standard (DES) as the preferred secure public encryption algorithm. This lasted for around 20 years. When the security of DES proved to be not sufficient, the Advanced Encryption Standard (AES) was announced by the NIST in 2001. Today AES is the most popular and secure encryption algorithm used in almost every application that requires some sort of security. For real time implementation of the encryption, the algorithm needs to be implemented in hardware. The traditional approach of storing the secret key is in Non-Volatile Memory (NVM) which suffers from side channel information leakage and the key can be recovered even after the device is disconnected from the power supply. Often the algorithms and protocols which are theoretically secure are vulnerable to attacks during the implementation and various reports of vulnerability on AES have been reported. It is estimated that the US semiconductor industry is losing over $7.5 billion per year due to counterfeit electronic components. The security of the system should not depend on the computational power of the adversary. Thus, given an adversary with unlimited computational capabilities, the system should still remain secure and this is the basic idea for the development of a new secure concept called a Physically Unclonable Function (PUF), also known as physical unclonable function. In this article the first form will be used. The concept of a PUF is depicted graphically in Fig. 1. In a traditional cryptographic approach (shown as a hard disk in Fig. 1), the key is stored in the medium. A given challenge will produce a response based on the key and if the key becomes compromised, then any set of challenges can be converted to valid responses. On the other hand, in the case of a PUF, the key is the medium itself. Even duplicating the medium will not produce valid challenge response pairs as the variability in the physical world ensures that the PUF cannot be exactly duplicated, i.e. it is unclonable.

II. PHYSICALLY UNCLONABLE FUNCTIONS: FUNDAMENTALS AND APPLICATIONS

A. The Physically Uncloneable Function: What is it?

A PUF is a random function which maps intrinsic properties of hardware devices (as a set of inputs or challenges) to a unique bit stream of information (as a set of output or responses). In other words, a PUF takes advantage of inevitable variability in the physical world, such as process
variation during device fabrication. Since in nanometer technology it is practically impossible to have two identical Integrated Circuits (ICs), PUFs exploit these variations in a unique way for different applications where power, security and device size are the key requirements.

During IC fabrication, process variation may be introduced in the various fabrication steps such as ion implantation, chemical mechanical polishing, chemical vapor deposition, sub-wavelength lithography, gas flow, thermal processes, spin processes, microscopic processes, and photo processes. A device fabricated from the same fab, same process, same lot, same wafer and same die will vary from its neighbors. These process variations induce in turn variations in threshold voltage, dielectric thickness, channel length, channel width, gate doping concentration, channel doping concentration and source/drain doping concentration. Even a very carefully designed/fabricated circuit cannot completely avoid these variations. As a result, one transistor will not be completely identical with another transistor. There would be slight differences in propagation delay, leakage current, voltage drop between them, and performance changes in different rates with temperature and time (aging).

B. Properties of PUFs

Since PUFs are essentially random functions, they have to satisfy some properties depending upon the application and its requirements. Some of these properties are briefly described below:

1) The Challenge Response Pair (CPR) should be random and no two challenges should produce the same response.
2) The challenge response pair should be easily generated in a short amount of time.
3) It should not be possible to characterize or model the PUF from a set of challenge response pairs.
4) A PUF based implementation should have low attack multiplicity, i.e. even if the attacker is able to extract the CRP in one instance, it should not be possible to extract CRPs at other time instances.
5) It needs to satisfy the Strict Avalanche Condition (SAC) to achieve maximum security, i.e. the probability of bit-flip at the output should be 0.5 when a single input bit is flipped.

C. Classification of PUFs

PUFs can be classified based on fabrication method and security strength, as shown in Fig. 2.

1) Based on fabrication: The main applications of PUFs are security based and thus the foremost requirement for such types of PUFs is that they should be tolerant to different types of attacks, both invasive and semi-invasive. Thus, it is advantageous to fabricate PUFs within the IC they are meant to protect and based on this, PUFs are categorized as follows:
   a) Silicon PUFs: These types of PUFs are interfaced with other ICs and are fabricated on the same die as part of the circuit. The process variation during fabrication is captured as a challenge and the difference in circuit characteristics (timing and delay information) is produced as a unique response for a given challenge.
   b) Non-Silicon PUFs: PUFs that are not classified as silicon PUFs are referred as non-silicon PUFs. They are fabricated in silicon systems but require special fabrication techniques which are not part of generic CMOS fabrication technology. The response is generated from the challenge set obtained from the random physical variation in the physical system rather than the ICs.

2) Based on security: This classification category is based on the number of challenges and accessibility of the response by the outside world.
   a) Strong PUFs: This class of PUFs supports very large number of CRPs. Since access to the PUF’s CRPs is available to the outside world, protection is provided through very complex input-output relations and exponential number of CRPs. This type of PUF is frequently used for device authentication purposes and operates in the range of few MHz, so even a short duration of eavesdropping can collect many CRPs. However, all known strong PUFs are vulnerable to one or more type of attack.
   b) Controlled PUFs: This class of PUF has a strong PUF as an underlying core additionally protected by control logic. Since the challenges are pre-processed first before passing to the strong PUF and the responses are pre-processed before coming out of the strong PUF, the CRPs of the strong PUF are never accessed directly. This pre-processing of challenges and responses significantly improves the security of controlled PUFs. A controlled PUF is only accessed via an algorithm which is fabricated in the same chip as the PUF such that the PUF and the algorithm are inseparable. This implies that the actual PUF outputs are never given to the outside world directly and the outputs that are given to the outside world are the processed outputs from the algorithm.
   c) Weak PUFs: This type of PUF has very few (worst case can be one), fixed challenges and hence its responses are never given to the public. They are mainly used to derive a secret key in cryptographic algorithms and are the least susceptible to modeling attacks. They are also known as Physically Obfuscated Keys (POKs).
D. Implementation of PUFs

In practice three main types of PUF circuits have been implemented:

1) Cover based PUFs: They utilize special materials as coating to produce a random response. For example, in an optical PUF a special layer of light scattering material is randomly distributed. When a laser beam is incident over the layer, it produces a random interference pattern which is used as a unique signature. Similarly, in an electric PUF, a coating layer with random dielectric coefficient is distributed and the capacitance variation over that coating generates the random response.

2) Delay based PUFs: They utilize the propagation delay between identical circuits in order to derive a response. Ring oscillator (RO) PUFs are based on frequency variation while switch based/arbiter PUFs are based on propagation delay. A tristate buffer PUF is another delay based PUF.

3) Memory based PUFs: They produce an output response based on the unpredictable startup state of feedback-based CMOS memory structures such as latches, flip flops, and Static RAM (SRAM) cells. Most of these structures use cross coupled topologies with positive feedback to store the required logic. When the circuit is turned on, these structures settle at one of the stable states which is then used as the PUF’s response.

Cover based PUFs are the most secure among these three designs. However their design complexity limits their applications.

E. Key Metrics for PUFs

The quality of a PUF is determined based on metrics which can be used to verify the applicability of the PUF to a specific application. Since different types of applications have different sets of requirements, not all of these metrics are of equal importance. A taxonomy of such key metrics is shown in Fig. 3.

Cover based PUFs are the most secure among these three designs. However their design complexity limits their applications.

Common metrics of the PUFs are the following:

1) Uniqueness: It is a measure of the average inter-chip Hamming Distance (HD) of the response obtained from a group of chips. The HD of two strings of bits is simply the number of bits in which the strings differ. It quantifies how different is one chip from another. When the process variation is large, the value of uniqueness will also be large. An ideal PUF has a uniqueness value of 50% which means that approximately half the bits in the responses of the PUFs (for the same input) should be different.

2) Reliability: It is a measure of how much reliable is the CRP under noise and environmental variations. For the given challenge, the PUF should give the same response under varying operating conditions. The ideal value for reliability is 100% which means that the PUF should produce an identical response under widely varying noise and environmental interference.

3) Randomness: It is a measure of balance between “0”s and “1”s in the response bits of the PUF and measures the randomness. The ideal value is 100% (i.e. perfect balance).

4) Correctness: It is a measure of correctness of the response under different operating conditions. The ideal value is 100%.

5) Bit Aliasing: It is a measure of biasness of a particular response bit across several chips. The ideal value is 50%.

6) Uniformity: It is a measure of how random is the CRP. For a response to be random, the number of “0”s and “1”s in the response should occur with equal probability (i.e. 50%).

7) Steadiness: The measure of biasness of a response bit for a given number of “0”s and “1”s over a total number of samples gives the steadiness. The ideal value is 100%.

Along with these metrics, design cost in terms of area, power consumption, design complexity, cost and delay always plays a key role and should be considered. Similarly metrics like false positive rate and false negative rate are also important in PUFs for the identification of a particular chip. The false positive rate is the probability of identifying any given chip as some other chip whereas the false negative rate is the probability that a correct chip is identified as an incorrect chip. This information can be obtained from inter-chip and intra-chip variations. These probabilities should be very small (ideally zero).

F. Selected Applications of PUFs

PUF applications are very broad as shown in Fig. 4. PUFs can be used in applications that require some sort of randomness during their operation. PUFs seem to be an elegant solution in applications such as random number generators, Radio-Frequency Identification (RFID) tags, secret key generation, and in device authentication where the required randomness property is obtained from process variation. PUFs have also been used in consumer devices for low-cost authentication purposes.
In cryptographic applications, random number generators are used to generate encryption keys, create initial seed values and to introduce randomness into protocol and padding schemes. Random numbers generated from Pseudo-Random Number Generators (PRNGs) are not truly random in the sense that after a certain value, the pattern repeats itself. Each number is generated following some complex non-reversible algorithm. In addition, the PRNG needs to keep the initial seed value secret as the output from the PRNG is generated from this secret key. To avoid these problems, Hardware Random Number Generators (HRNGs) are used to generate a true random output without the need of an initial seed. A silicon PUF delay circuit is used where a rising edge signal is given as an input and a set of challenge bits as a selector input to a set of multiplexers. Depending on relative delays of the various paths, a particular set of challenge bits will produce a truly random bit at the output.

III. RECONFIGURABLE OR DYNAMIC PUFs

In order to share secrets between a physical device and a remote user, a protocol using controlled PUFs is useful. The user wants to make use of the computing capabilities of the chip in an insecure communication channel. For ensuring the user’s confidence in the authenticity of the result, namely that the computation has been performed in the intended processor, the user relies on authenticity from the device manufacturer. The PUF acts as an interface between the communication channel and the main chip (or processor). First the user needs to have the private list of CRPs. The response obtained from the PUF for the given challenge is compared with the corresponding response in the list. The uniqueness of the proposed protocol is the way challenges are used compared with the SAC condition is still satisfied.

1) CRP Reconfigurable PUF: The CRPs are made reconfigurable by adding additional circuits to the design but without changing the main PUF structure. It can be done using different ways:
   a) Pre-processing the challenge bits. Instead of giving directly the challenge bits to the PUF, they are given to this new added circuit whose output is then given as a challenge bit to the PUF. These new added circuits can be either Linear Feedback Shift Registers (LFSR) or hash functions.
   b) Pre-processing the response bits. Instead of taking the PUF output directly, it is first given to the newly added circuitry whose output is then treated as the PUF response. This newly added circuit is mostly a hash function.
   c) Output recombination, i.e. adding extra reconfigurable components to preprocess the output of the PUF.

2) Logic Reconfigurable circuits: The logic of the main PUF is reconfigured which automatically reconfigures the CRP. It is usually accomplished with some combination of multiplexers.

In a dynamic PUF (DPUF) device aging is used to alter the delay characteristics of the circuit. The advantages of the DPUF are:
   1) Increased security against reverse engineering and emulation attacks.
   2) DPUFs are customized to the user’s specifications. The user can introduce aging in the device using suitable input patterns and can introduce delay.
   3) Easy integration with other device components as they are entirely composed of gates.

A lightweight secure PUF with improved security and robustness is shown in Fig. 5. It consists of a parallel PUF structure with input network at the challenge side and output network at the response side. There is a wire interconnect network in order to connect the challenge bits across the parallel PUFs. A single PUF satisfies the SAC criterion but the parallel PUF structure does not satisfy it and this decreases the randomness and security. Hence an input network is defined. However the adversary can introduce failure in the design by applying the inverse transformation operation of the input network. To avoid this, a wire interconnect network is introduced which physically binds the input of the parallel PUFs. In the wire interconnect network, one challenge bit on each row is connected to another challenge bit from a different row, so that the SAC condition is still satisfied.

![Fig. 4: Typical applications of PUFs.](image)

![Fig. 5: Lightweight PUF.](image)
IV. DIODE BASED PUFs

Crossbar arrays have some distinct advantages over other circuit designs so they are becoming an attractive option for PUF design. These advantages include a regular structure, high packing density, low energy consumption and simple implementation. However, sneak current path and crosstalk problems need to be addressed in crossbar arrays with arbitrary array size. Sneak current mainly depends on the current through the reverse junction.

A crossbar array consists of two sets of conductive parallel wires, one horizontally placed and another vertically placed such that they intersect each other, as shown in Fig. 6. The intersections between these two wires are separated by either a thin film or by a molecular compound, whose properties (such as resistance) can be varied by a voltage applied between individual wires from the first and second wire set. Thus low and high resistive states can be obtained at the intersection, which can be used to store digital data. The horizontal set of wires is called “word lines” while the vertical set of wires is called “bit lines”. For both reading and writing, the input voltage is applied to the accessed word line and the current flow is measured in the accessed word line. The input voltage for writing is generally higher than that for reading.

V. MEMRISTOR BASED PUFs

Memristor based PUFs are compatible with CMOS fabrication standards. The memristor, also called Resistive RAM (RRAM), is a two terminal passive electric component whose resistance depends on the magnitude, direction and duration of voltage applied at its two terminals. Thus a memristive device can have multiple discrete resistance states or a continuous variable resistance and can switch its states from one resistive state to another. The change in its resistance depends on the past history of the device, i.e. previously applied voltage/current across/through the device. Thus, a memristor combines the behavior of memory and a resistor. The important feature of memristor is that it provides easy tamper detection, which ensures the security of a PUF’s response.

Two memristive PUF circuits have been proposed: Memristive memory based PUF and Lateral switching PUF cell. The memristive memory based PUF is based on the variability in the write time of the device which depends on the device thickness. The circuit configuration is shown in Fig. 7.

Two control signals are used. The first control signal ($R/W$) determines the read/write operation of the memristor while the second control signal, NEG, determines the polarity of the write operation. The circuit operates with the RESET operation where $R/W=1$ and NEG=1. With the R/W signal, the memristor goes to the Highly Resistive State (HRS). Then the SET pulse is applied to write either “0” or “1” by applying $R/W=1$ and NEG=0. Once the bit is stored in the memristor, the stored value can be read by applying $R/W=0$. The challenge bit is XORed with the output from the memory and the result from the XOR is the final response of the PUF.

The idea is to use the variation in the device thickness to have varying read and write times of the memristor. When the memristor is SET from HRS to Low Resistive State (LRS), a logical “1” can be written into the memristor provided that the SET time is longer than the minimum write time, $T_{WR,min}$; otherwise a logical “0” will be written into the memristor. If this SET time is close to $T_{WR,min}$, then the probability of having logical “1” is the same as that of having logical “0”. This uncertainty is what makes the memristor suitable as a PUF. Once the data has been set in the memristor, it can be read by applying $R/W = 0$. This output from the memristor can then be XORed with the challenge set to get the response.

VI. CARBON BASED PUFs

Currently most PUF circuits are based on silicon devices, and are mainly focused on improving randomness, reliability and robustness without giving much attention to power consumption. Since the semiconductor industry is scaling logic devices for high speed, low voltage and low power consumption, scaling down the gate length in the current
silicon devices has some limitations. Thus beyond a certain limit, it cannot be reduced further due to degradation of gate control on the channel. The addition of doping density in the channel can counter the effect but it reduces the carrier mobility. Very thin channel structure field effect transistors such as FinFETs can be used as another option but it becomes extremely difficult to scale these devices due to physical limitations. Thus a new material needs to be explored to meet the future demands of small size, high performance and low power consumption. Carbon derivative options like the carbon nanotube (CNT) and graphene are promising candidates.

A. CNT based PUF

Carbon Nanotube FET (CNTFET) based PUFs aim to achieve better reliability, low energy and power consumption compared to that of silicon based PUFs. In CNTFET transistors, CNTs are used in the channel instead of bulk silicon. The common variations in the CNT are chirality, diameter, growth density, misalignment, and doping concentration. The chirality variation can be used to generate secret digital bits. Chirality refers to the direction in which the graphene sheet is rolled. Thus it defines the behavior as metallic or semiconductor. In the metallic CNT, there is a direct shorting of drain-to-source, thus even when the CNTFET is turned off, metallic CNTs continue conducting. This contributes the off-current I_{off}. When the CNTFET is turned on, both metallic and semiconducting CNTs conduct to contribute to the on-current I_{on}. A serial connection of CNT-PUF parallel elements (CNTPUF-PEs), as shown in Fig. 8, can be used, where each CNTPUF-PE consists of parallel CNTFETs sharing a common gate voltage. Each CNTFET consists of a large number of semiconducting CNTs and a small number of metallic CNTs, typically in the ratio between 10% - 33% such that in the on condition the current due to semiconducting CNTs dominates the metallic CNTs and during the off condition, metallic CNTs dominate the semiconducting CNTs. As shown in Fig. 8, each CNTPUF-PE consists of one distinct state, each for high input gate voltage and low input gate voltage. However due to process variations, these states vary in each CNTPUF-PE, which is then compared at the comparator to give the final response.

![Fig. 8: Carbon nanotube based PUF design.](image)

B. Graphene based PUF

Graphene has very high electron mobility and thus can offer advantages as channel material in FETs for very high speed circuit operation. Graphene has a two-dimensional carbon atom network arranged on a honeycomb lattice. The layers of graphene form a graphite. In other words, a single layer of graphite structure is a graphene which is interconnected in a hexagonal lattice structure. Graphene and graphite have significantly different properties. Graphene is strong in almost all directions while graphite is strong in one direction because it consists of stacks of graphene. Graphene is the mother of all graphitic materials. There are some very interesting properties of graphene that make it suitable as a substitute for silicon in future technology.

PUF based security devices mostly contain a measurement circuit that measures some (or at least one) property of the graphene device and/or graphene layer (such as resistance, capacitance, voltage response) so as to derive the PUF output to be used as the cryptographic key.

VII. CONCLUSION AND FUTURE DIRECTIONS

The concept of a PUF was proposed in 2001 as the physical one-way function. Since then the concept has been explored for practical circuits to enhance security. The initially proposed models were not able to provide much security as they were successfully broken down through the use of mathematical modeling and other non-invasive methods. As a result, several other structures were proposed to enhance security, by considering other critical aspects like area, reliability, randomness and so on. Most of these designs are fabricated up to 45nm technology mostly for secret key generation and authentication purposes. Since the PUF technology has untapped potential, much research needs to be done. Especially in technologies beyond 30nm, it needs to be explored how these technologies can be used in the most effective way.

Much of the current research has concentrated on the circuit level implementation of the PUF. There is not much work done on the system level security model of PUFs. Since the PUF circuit needs to facilitate other circuits in the systems, a thorough analysis of the overall system must be provided and not just PUF circuitry.

PUFs are meant to complement or replace other hardware authentication techniques such as biometric authentication (when the hardware is tied to a specific person), smart cards and hardware one-time password (OTP) tokens. Of these methods, PUFs can only complement but not replace biometric authentication by providing a dual layer of protection: identification of a specific person (biometrics) possessing a specific piece of hardware (PUFs). Smart cards on the other hand can be completely replace by PUFs since they are much more secure and not amenable to theft (as a smart card is). OTP tokens can be either replaced or complemented by PUFs, depending on whether a single layer or protection (OTP token of PUF) is sufficient or two layers are preferred (OTP token and PUF).

Similarly, designs involving memristors which have the potential to scale down to 3nm need to be explored. Since CMOS technologies beyond 18nm face many challenges, other alternative options based on graphene or carbon nanotube are being considered. As the memristor based PUF is a very recent topic and not much work has been done on it,
research needs to be done on the performance of such devices based on the sources of variations in memristors along with other standard variations. Furthermore, technology advancements have opened new PUF sources like phase change devices, spin-torque transfer devices and devices manufactured with quantum dots, graphene and nanotubes. A lot of exploration needs to be done to come up with a solution that best suits the PUF purpose.

Read More About It


About the Authors

Shital Joshi (ShitalJoshi@my.unt.edu) is currently a Ph.D. candidate at the Department of Computer Science and Engineering, University of North Texas. He obtained his masters from Institute of Technology, Varanasi in 2011. He is an author of 5 peer-reviewed publications.

Saraju P. Mohanty (saraju.mohanty@unt.edu) is a Professor at the Department of Computer Science and Engineering, University of North Texas, and the director of the NanoSystem Design Laboratory. He obtained his Ph.D. in computer science and engineering from the University of South Florida in 2003, his master’s degree in systems science and automation from the Indian Institute of Science, Bangalore, India, in 1999. He is an author of 180 peer-reviewed journal and conference publications and 3 books. He is an inventor of 4 US patents. He is a Senior Member of IEEE and ACM.

Elias Kougianos (eliask@unt.edu) is currently an associate professor in Electrical Engineering Technology at the University of North Texas. He obtained his Ph.D. in electrical engineering from Louisiana State University in 1997. He is author or co-author of over 90 peer-reviewed journal and conference publications. He is a Senior Member of IEEE.