## Weak PUF vs Strong PUF

The distinction is rooted in the security properties of their challenge-response pairs One definition of a Strong PUF:


Even after giving a adversary access to the PUF instance for a prolonged period of time, it is still possible to come up with a challenge that with high probability, the adversary does not know the response

This implies that exponential in



- The PUF has a very large challenge space, otherwise the adversary can simply query the PUF with all challenges to learn its complete CRP behavior


PUFs which do not meet these requirements are called Weak PUTs
In the limit, some PUFs have only a single challenge and are called physically obfuscated key or OK
We discussed the SRAM PUF earlier that has only one challenge
It is infeasible to build an accurate model of the PUF using only a subset of CRPs to 'train' the model, as a means of learning its complete CRP behavior



## PUF Usage Scenarios

## Identification

The PUF can be used to generate a 'serial number' to identify and/or track parts through manufacturing (the original proposed use by Keith Loftstrom in 1999!)

For manufacturing, uniqueness is the most important metric
A weak PUF is sufficient for this type of low security application

Reliability is not a concern as long as

- Bit flip errors are infrequent, i.e., $\mathrm{HD}_{\text {intra }}$ is relatively small, otherwise the probability of 'aliasing' gets unacceptably large
- It is possible to use a fuzzy match criteria after the identifier is generated
- Authentication

The PUF is used to securely identify the chip in which it is embedded to an authority through corroborative evidence

As we will see when we discuss authentication scenarios, a strong PUF is best, particularly when the device is resource-constrained

## PUF Usage Scenarios

Also, the challenge-response form of authentication implemented by strong PUFs is considered strong, in contrast to weak forms of authentication, e.g., passwords

Note that in contrast to encryption discussed below, the PUF inputs and outputs are exposed (to different degrees depending on the authentication scheme)

This makes the PUF more accessible (and vulnerable) to adversaries, and enables model-building attacks

There is a rapidly growing need for hardware-based authentication, e.g., in the supply chain, in the field (electronic voting machines) and for IoT devices

For the supply chain, the PUF is an important new security primitive that can address threats related to

- IC theft
- IC reuse
- Malicious substitution (hardware Trojans)
- Reverse engineering and cloning


## PUF Usage Scenarios

The same is true for 'in the field' authentication, particularly with IoT devices which are vulnerable to physical attacks and are resource-constrained

All three statistical metrics, i.e., uniqueness, randomness and reliability, are important for authentication

Some simple schemes relax the reliability metric as we will see

Why use PUFs for authentication?

- They can eliminate the requirement for NVM, a real cost benefit for resource-con- strained devices
- They can potentially provide a very large number of CRPs, i.e., a much larger source of entropy when compared to an NVM
- They are tamper-evident, making it more difficult for adversaries to physically probe the device to steal the secrets
-They can be designed to never eveal their secrets, i.e., even the manufacturer does not have knowledge of the embedded secrets
- They can be used to provide a stronger challenge-response form of authentication


## PUF Usage Scenarios

## - Encryption

The PUF is used to generate

- A key for symmetric encryption algorithms
- A random nonce that can be used to select a specific public-private key pair for asymmetric encryption

In typical encryption applications, the key is not revealed outside the chip and therefore, a weak PUF can be used (although a strong PUF is better here too)

The inaccessability of the PUF responses makes model-building impossible However, recent work shows that power analysis attacks can be used to enable model-building, which argues in favor of using strong PUFs for encryption too

Unfortunately, in contrast to authentication schemes, tolerance to bit flip errors is 0 Even a difference of 1 bit in a 256 -bit key completely wrecks communication between parties because of the avalanche effect

## PUF Usage Scenarios

In summary

- All three applications require uniqueness
- Identification:

PUF bitstrings must be large enough to suit the \# of chips in the population
$\mathrm{HD}_{\text {infra }}$ can be $>0$ but bear in mind, this reduces the number of unique IDs that can be generated and used

- Authentication: Add randomness as a critical metric

Having a very large CRP space prevents adversaries from reading them all out and building a clone, and prevents them from succeeding at model-building

- Encryption: Adds both randomness and reliability as critical metrics

Having a large number of CRPs is not necessary in cases where only a single key (or small number of keys) need to be generated over lifetime of chip
$H D_{\text {infra }}$ must be zero, which requires error correction or error avoidance

## PUF Implementations

There are MANY PUF implementations that have been proposed

A rough characterization is as follows:

- Delay-based PUFs:

Delays along 'matched' paths (Arbiter)
Ring Oscillator frequencies
Glitches produced along paths within a functional unit
Delays along glitch-free paths within a functional unit (HELP)

- Bi-stable PUFs: SRAM

Butterfly, Buskeepers
FFs and Latches

- Mixed-Signal PUFs: (These require a specialized analog-to-digital converter: ADC) Transistor threshold voltage/transconductance
- Dynamic/leakage current
- Resistance/Capacitance


## Arbiter PUF



A specialized structure implements two paths, each of which can be individually configured using a set of challenge bits

Each of the challenge bits controls a 'Switch box', that can be configured in either pass mode and switch mode

Pass mode connects the upper and lower path inputs to the corresponding upper and lower path outputs, while switch mode flips the connections

A stimulus, represented as a rising edge, cause two edges to propagate along the two paths configured by the challenge bits

## Arbiter PUF

challenge


The faster path controls the value stored in the Arbiter located on the right side of the figure

If the propagating rising edge on the upper input to the Arbiter arrives first, the response bit output becomes a ' 0 ', otherwise a ' 1 '

The switch boxes are designed identically as a means of avoiding any type of systematic bias in the delays of the two paths

Within-die process variations change the delay through the switch boxes, which makes each instance of the Arbiter PUF unique


It is clear that the arbiter PUF has an exponential number of input challenges
In particular, $2^{n}$ with $r$ representing the number of switch boxes

However, the total amount of entropy is relatively small
For $n$ equal to 128 , the total number of path segments that can vary individually from one instance to another is $4 * 128=512$

$$
4
$$

The exponential number of challenges simply gombine the entropy in different ways Although the arbitor DUF idconsldered a strong PUF, researchers have 'broken' it using model building many times

## Arbiter PUF

challenge

rising
edge
1

## Model Building

The number of individual sources of entropy in the Arbiter is only linear with $n$
Therefore, dependencies must exist among the $2^{\mathrm{n}}$ challenges and response bits
For example, if it were possible for the adversary to learn the individual path segment delays, then the PUF is no longer needed to predict the responses

Modeling attacks leverage a simple additive delay model where the delay of the entire path is equal to the sum of the individual segment delays

$$
r_{2}<r_{1}<r_{1}<r_{3}
$$

By strategically selecting CRPs, machine-learning techniques can quickly determine the relative delays through each switch box

Machine-learning techniques include artificial neural networks (ANNs), supportvector machines (SUMs), genetic algorithms and decision trees

Goal is deduce the relationship of segment delays using as few CPs as possible
A PUF is $\left(\boldsymbol{p}_{\text {model }}, \boldsymbol{q}_{\text {train }}\right)$-modelable if known modeling attacks exist which have a successful prediction rate of $p_{\text {model }}$ after training with $q_{\text {train }}$ RPs

## Arbiter PUF Evolution

Early examples in the literature on ASIC implementations show
$-\mathrm{HD}_{\text {indra }}$ of $4.82 \%$ with a temperature range of $25^{\circ} \mathrm{C}$ to $67^{\circ} \mathrm{C}$

- $\mathrm{HD}_{\text {inter }}$ of $23 \%$ not very good $\Rightarrow$ not sPry ramdoun
- SVM-based machine learning attack produced $\left(p_{\text {model }}=96.45 \%, q_{\text {train }}=5000\right)$, which indicates the implementation is not secure

All subsequent work attempt to make model-building attacks more difficult by:

- Introducing non-linearities, ie., feed-forward and XOR-mixed versions
- Obfuscating the challenges to the PUF and the responses from the PUF $6 / \mathrm{em}$.

XOR-mixed version


## Ring Oscillator PUF

The RO PUF is also a delay-based PUF but the configuration and measurement technique are different from the Arbiter PUF

- An odd number of inverters are connected in a ring, which causes an edge to circu- late continuously
- The Arbiter is replaced by a counter an odd \# of,h vorters


By enabling the RO for a fixed $\Delta t$, the frequency of the RO is reflected in the count, and is given by count $/ \Delta \mathrm{t}$

But since $\Delta t$ is constant for all RO testing, the digital coynt value can be used $\begin{array}{lll}\text { instead } & \text { speed } \sim \text { Temp. Templ spoary } \\ & \text { speed } \sim \text { Volt. } & \text { JuHT } A \text { spard }\end{array}$
Similar to the Arbiter PUF, a differential frequency post-processing scheme is typically used to compensate for temperature/supply voltage variations



## Ring Oscillator PUF



Lehmer-Gray encoding has been proposed to optimize entropy and nearly achieves the maximum $\log _{2}(n!)$ number of independent response bits The cost is increased processing complexity


0 . 4 , comparison
ROT?
This strategy is not optimal, however, in utilizing the available entropy, reducing the number of generated response bits to $n / 2=10 / 2=56$

## Metal Resistance PUF

The metal PUF measures voltage drops across polysilicon wires, metal wires and vias as the source of entropy


An SMC cell from a larger array is selected using column and row select signals
Once selected, a Stimulus-Measure-Circuit (SMC) enables a shorting transistor (stimulus) which creates a voltage drop across the poly-metal-via stack

Two 'pass gates' are also enabled that allow voltages to be sensed and measured


## Metal Resistance PUF

Voltages generated by an element in the SMC are digitized by a VDC
SMC array of 2048 elements Voltage-to-digital-converter (VDC)


## Metal Resistance PUF

Similar to the RO bit generation method, the algorithm used for the metal PUF create TC differences (TCDs) by randomly selecting pairs of TVs from the distribution


J. Tu, R. Chakraborty, C. Lamech and J. Plusquellic,
"Stability Analysis of a Physical Unclonable Function based on Metal Resistance Variations", HOST, 2013.

An error avoidance scheme is proposed that creates two thresholds around the mean of the TCD distribution

TCDs around the mean are unstable and are not permitted to generate a bit in the bitstring/key
Variations in poly metal/Via

The red and blue TCDs illustrate that TV-noise-related variations duffing regeneration are small enough to prevent bit flip errors


## Metal Resistance PUF

Statistical analysis of bitstrings generated from 7343 TCDs $n d 63$ chips


We developed a reliability-enhancing technique called XMR, which creates redundan copies of the bitstring

Majority voting is then used to 'correct' bit-flip errors

$$
10^{-6}
$$

Typical reliability standards target $1 \mathrm{e}^{-6}$ ( 1 in a million) to $1 \mathrm{e}^{-9}$ ( 1 in a billion) 3MR (TMR) and 5MR provide reliability in this range



## The Billion Dollar PUF Question

- Can the underlying physics of a PUF be harnessed to provide the following
- An exponentially large (as opposed to polynomial) challenge-response space
- Statistically reliable responses which can be utilized for cryptography
- Authentication
- Encryption
- Sufficient sizes of " $n$ " such that an adversary cannot carry out brute-force attacks successfully
- Physical characteristics that do not allow model building or machine learning techniques to reduce the search space from exponential to polynomial


## Hardware Embedded Delay PUF (HELP)

HELP measures path delays in an on-chip functional unit, e.g., AES, and leverages random within-die variations in propagation delay as a source of entropy


HELP can be described entirely in an HDL, and therefore can be implemented on FPGAs

The functional unit (entropy source) is implemented using a specialized logic style that is hazard-free ( l.e., no slitches, i.e., for each laquch

This ensures paths remain stable, and can be timed accurately, as TV conditionsc pptare,


## Hardware Embedded Delay PUF (HELP)

HELP uses a launch-capture timing mechanism to obtain high-resolution path delay values for combinational logic paths


Path delays can be measured using a clock strobing method
Or using an alternative flash $A D \overline{C \text { method that also works well }}$
The fine phase shift feature within modern digital clock managers (DCMs)) can be used to incrementally tune a capture clock, $C l k_{2}$, in a series of launch-capture tests

The integer-based fine phase shift value is used as the digitized path delay

## HELP Experiments and Features

We implemented HELP on a Xilinx Zynq 7020 and tested 20 chips, with 25 copies of HELP implemented in different locations (but 'fixed') on each of the chips


The total number of paths in the AES functional unit is approx. 8 million ( 4 million rising paths and 4 million falling paths)


This large \# is the first important characteristic that makes HELP a strong PUF

Other features are related to its multi-dimensional CRP space which includes:

- Parameters including two LFSR seeds, $\mu_{\text {ref }}$ and Rngref, a Modulus and Margin
- The full set of two vector sequences, Path-Select masks and Distribution Effect

$$
\begin{align*}
& n=64 \text { ontprits } 28  \tag{2/7/18}\\
& \text { of AES func.unit }
\end{align*}
$$

## HELP Processing Steps

STEP 1: Apply a set of challenges to generate 2048 rising path delays (calle $\mathbf{P N R}$ ) and 2048 falling path delays (called PNF), with PN for PUFNumber


Changes in TV conditions shift and scale the digitized path delays

These digitized path delays are processed as a group, NOT individually as is true of all other PUFs, i.e., no bits are generated until all group processing is complete

## HELP Processing Steps

STEP 2: Create unique pairing of rising and falling path delays using two 11-bit LFSRs, to create PN Differences or PND


Shifting and scaling of entire distribution is exacerbated, but TV variations are reduced (partially compensated for) in the individual PND b/c of common mode refection Teed. Shift Ry.

LFSR seeds expand the response space of HELP and allow up to $\boldsymbol{n}^{2}$ bits to be generted from $n$ PNR and $n$ PDF

$$
(20+0)^{2}
$$

As we will see later, a Modulus operation nearly eliminates the classical dependencies that exist when PN are reused

$$
\left(2^{11}\right)^{2}=2^{22}
$$

(i) consider that there ane 4096 phys. measurement
(i) One could choose two $=(40$ $\frac{4096(4095)}{2}=\left(\begin{array}{l}\text { for } \\ n=2048) \\ n\end{array}=\frac{2 n(2 n-1)}{2}=n(2 n-1)=\right.$
(iii) By choosing from a bucket of 2041 . $n^{2}$ N and another bucket of 2048 PNFs, aug id esultirs in $2043^{2}$ comb (Lequer Gate
 and sequences through the full statespoce
(y) details regarding how the LFSR pick from the PNF, PANR buckets ore skipped

## HELP Processing Steps

Illustration of one PNR and one PNF, collected across 12 TV corners (x-axis) and 500 chips-instances (y-axis)


Single PNR/PNF illustrate that shifting and scaling is significant, while PND in right plot show reduced jig-saw pattern

Goal is to have flat horizontal lines, i.e., all TV corners produce same PND

The data from the 25 instances from Chip $_{20}$ are highlighted in red to illustrate performance similarities

The large spread along $y$-axis is largely due to chip-to-chip variations

## Hardware Embedded Delay PUF (HELP)

HELP measures path delays in an on-chip functional unit, e.g., AES, and leverages random within-die variations in propagation delay as a source of entropy


HELP can be described entirely in an HDL, and therefore can be implemented on FPGAs

The functional unit (entropy source) is implemented using a specialized logic style that is hazard-free

This ensures paths remain stable, and can be timed accurately, as TV conditions vary

HELP is a STRONG PUF and is capable of generating a large \# of random bitstrings

## HELP Processing Steps

Its clear that the difference operation is NOT able to remove all of the path delay farlation introduced by TV-noise

STEP 3: Apply TVCompensation (TVComp) to remove remaining TV-noise

Th d $\mu_{\text {chip }}$ and Rng $_{\text {chip }}$ are computed from a histogram distribution
The ref values are user-specified parameters
TVComp creates a histogram distribution of PND, and then scales and shifts the path delay distribution to a reference distribution



The reference distribution values expand the response space of HELP in a similar fashion to the 2 LFSR seeds used to create the PND from the PNR and PNF


## HELP Processing Steps

TVComp ELIMINATES all chip-to-chip variations, but preserves within-die variatons Or iginal) no no $_{\text {eND }}$ TVipozo $\operatorname{comp}$





12Temperature-Voltage Corner


This fact is illustrated on the right with $\mathbf{P N D}_{\mathbf{c}}$, which show the data from the 25 instances from Chip ${ }_{20}$ now distributed across entire range of $y$-axis

In contrast to the grouping of $\mathrm{Chip}_{20}$ data on the left, which shows similar performance among the different instances, as expected $\mathrm{b} / \mathrm{c}$ data is from same chip

## HELP Processing Steps

The $\mathrm{PND}_{\mathrm{c}}$, although compensated for TV variations, still possess path length bias

Bias is delt with in two ways, first by optionally applying an Offset (for fine tuning) and then using a coarse-grained Modulus operation


STEP 4: Add server-computed Offsets (computed using enrollment data) and then apply a Modulus operation to remove path length bias




Offsets are computed from the median of the chip population and are added to each $\mathrm{PND}_{\mathrm{c}}$, which stiffs pop. to a multiple of 10 and then a Modulus of 20 is applied


The $\mathbf{P N D}_{\mathbf{c}}$ with offsets are called $\mathbf{P} \mathbf{N D}_{\mathbf{c o}}^{\prime}$ and the final values are called modPND ${ }_{\text {co }}$

## HELP Processing Steps

STEP 5: Bitstring generation uses a Margin parameter, that implements a bit-flip avoidance reliability-enhancing scheme tued avoidance reliability-enhancing scheme


We call this the Single Helper Data scheme b/c the Margin scheme is run only by the token during enrollment weak zerofones - not

We also have a Dual Helper Data scheme that combines helper data generated by both the token and server

We have a suite of reliability-enhancing schemes for stand-alone (no server) applications, e.g., key-encryption-key (KEK) mode

## HELP Statistical Results

Statistics using the Offset method



Probability of Failure $e$ ( o rs

$0^{\left(10^{\circ}\right)}$
0
-1
-2
-3
-4
-5
-6
-7
-8
-9
-10




NIST Statistical Results
 These statistical results indicate the bitstrings generated by HELP are of cryptographic quality


Note that this implementation of HELP includes all four functions, including token authentication, verifier authentication, session encryption and KEK

Versions dedicated to one function would be smaller in size

