TV-PUF : A Fast Lightweight Aging-Resistant Threshold Voltage PUF

Tanujay Saha* and Vikash Sehwag*

Dept. of Department of Electronics and Electrical Communication Engineering,
Indian Institute of Technology,
Kharagpur, West Bengal, INDIA-721302
tanujay.saha@gmail.com, sehwag.vikash@gmail.com

Abstract. Physical Unclonable Function (PUF) is the hardware analog of a one-way function which can address hardware security issues such as device authentication, generating secret keys, producing seeds for Random Number Generators, etc. Traditional silicon PUFs are based on delay (Ring Oscillator PUFs and Arbiter PUFs) or memory structures (e.g., SRAM PUFs). In this paper, we propose the design of an aging resistant, lightweight and low-power analog PUF that exploits the susceptibility of Threshold Voltage ($V_{th}$) of MOSFETs to process variations. Analysis shows improvement in power consumption, reliability over device aging along with quality metrics like uniformity, reliability and uniqueness for a 64-bit key generation. For 1 GHz clock input, this design consumes 0.18µW/bit power with 50 % uniqueness and 51% uniformity along with the independence of these metrics on technology nodes. Experimental results suggest 4% variation in reliability under temperature variation from -55°C to 125°C and 20% variation in supply voltage. Aging analysis further projects the independence of reliability over device aging.

Keywords: Hardware security, Physical unclonable function (PUF), Process variation, Aging resistant PUF, Low power embedded systems

1 Introduction

Physical Unclonable Functions (PUFs) are now widely used for generating cryptographic keys. It has the ability to create unique cryptographic keys specific to devices properties which exploit the process variation in device manufacturing [1]. Secure key generation is possible even from biased PUFs [2]. PUFs also find use in error detection methods in Finite State Machines [3]. Nowadays, Physically Unclonable Function (PUF) [4,5] is a major weapon against IC counterfeiting. The hard-coded key in the IC can be replaced with PUF circuits which take challenges as inputs and provides responses as outputs. The challenge to response mapping should strictly be a one-to-one mapping, even under changing environmental conditions like fluctuations in temperature and supply voltage.

* these authors contributed equally to this work
The responses depend on the process variations of the components involved in the circuit. The security of the PUF is based on the fact that practically any circuit cannot be mimicked exactly due to the process variations present in its components. Hence, it is almost impossible for the adversary to predict the responses. The more random the process variations in the PUF are, more difficult it is for the adversary to clone the PUF. Still adversaries have attacked many PUF designs using Machine Learning prediction techniques like Support Vector Machines [6]. In [7], it is also shown that PUF enabled cryptographic primitives are vulnerable to advanced side channel attacks like Differential Template attacks. PUFs can be broadly classified into two design categories: i). Delay based PUFs (Arbiter PUF, Ring oscillator PUF, Glitch PUF and Schmitt trigger PUF) ii). Memory based PUFs (SRAM-based PUF, butterfly PUF, Latch PUF). A technique for IC identification based on the unique resistance values in its power supply distribution system was introduced in [8]. In [9], irregular current-voltage characteristics of diodes packed in a crossbar memory are exploited to construct a PUF. Some PUFs have tried to capture the randomness in the subthreshold operating region of the FET [10]. New technologies have emerged which uses memristors as the basic building blocks of Arbiter PUFs [11]. Previously, another threshold voltage based PUF called ICID [12] had been proposed. It is shown that our Threshold Voltage PUF (TV-PUF) is more efficient than ICID in terms of both challenge-response performances (security parameters) as well as area and power consumption (VLSI design parameters). Another important aspect in PUF characteristics is its reliability degradation over aging of the device. Previous works in this field [13–15] discuss the effect and mitigation of degradation of PUF reliability due to aging. In semiconductor devices, performance degradation due to the aging of devices is a crucial factor in design consideration as it affects the reliability of device over time. Most of the previous PUF designs employs CMOS technology which incorporates both PMOS and NMOS. In MOSFETs, the key reliability issues are Bias Temperature Instability (BTI) and Hot Carrier Injection (HCI). Negative bias temperature instability (NBTI) is effective in PMOS while Positive-bias temperature instability has an impact in NMOSs. Both the effects causes an increase in Threshold voltage of the transistor (\(V_{th}\)) over time. In our model, we have used only NMOS transistors which are supported by the result that NBTI is the major reliability concern as compared to HCI [16]. Use of only NMOSs in PUF design exclude the impact of NBTI on the reliability of PUF as in NMOS, only HCI and PBTI concern the aging issue. However, PBTI only plays a role when high-k gate oxide materials are used [17,18].

In this paper, we propose a Threshold voltage (\(V_{th}\)) based PUF (TV-PUF) which captures the effects of process variations of \(V_{th}\). Threshold voltage has been chosen as the primary basis for the operation of the PUF as it accumulates the process variations in many factors like doping of the n and p regions, gate length, width and oxide thickness. This makes it much more difficult to clone a particular instance of the TV-PUF. TV-PUF takes a n-bit challenge and produces a 1-bit response similar to the Ring Oscillator PUF (RO-PUF). A sense amplifier
is used at the end of the circuitry to convert the analog response to a digital response. TV-PUF will find use in various lightweight applications such as device authentication, countering IC counterfeiting, etc. Similar to the operation of a Ring-Oscillator PUF, a unique part of the circuit is employed to operate for a particular challenge. This avoids any dependency between the responses of any two challenges. The rest of the paper is organized as follows. Section 2 briefly discusses the contribution of this design where section 3 focuses on the design and working aspects. In Section 4 we evaluate the performance of TV-PUF based on standard puf metrics. In section 5 we discuss the effecting of device aging on the reliability of TV-PUF. In section 6 we compare our design with previous Puf design in terms of power consumption and number of transistors. Finally, we conclude the paper in section 7 with a discussion of future work in section 8.

2 Our Contributions

TV-PUF contributes to the literature of PUFs in the following dimensions:

- TV-PUF requires only three NMOS transistors for the implementation of a block. Additionally, it consists of a decoder and a sense amplifier. Due to lightweight, it can be embedded on any IC chip with minimum circuit overhead.
- TV-PUF can operate at very high frequency due to the less critical path. Such reduction in critical path is caused due to only three NMOS in the critical path which provides a delay in the order of 100ps at 1V supply voltage. Due to very less latency, it can be used in a wide variety of day-to-day devices like smart cards for authentication and security processes.
- Improvement in PUF properties such as uniformity, uniqueness, reliability and bit aliasing have also been observed for TV-PUF. A detailed comparison with the most widely used PUFs, i.e., Ring Oscillator PUF and Arbiter PUF, have been given in the section-4 of this paper. Further, it is also shown that these characteristic are independent on technology node.
- TV-PUF has also employed aging resistant nature. In section-5 we have discussed the effect of aging over a span of 5-years.

3 Design of TV-PUF

The threshold voltage of a MOSFET, being highly susceptible to process variations, is being used as a differentiating factor in the design of TV-PUF. For NMOS pass transistor (fig.1), \( V_x \) is initially zeros. As \( V_x \) is increased, current-drive of the transistor(\( V_{gs} \)) reduces significantly. This reflects in the long tail in \( V_x \) as it approaches \( V_{dd} - V_{th} \). We use only NMOS pass transistors (fig.1) in the design of TV-PUF. The motivation is to compare source voltages, which is output at the terminal(x) of two independent NMOSs. There will be slightly different due to the difference in the \( V_{th} \) of the two FETs caused by process...
Example 6.9 Voltage swing for pass transistors circuits

Assuming a power supply voltage of 2.5V, the transient response of Figure 6.34 shows the output of a NMOS charging up (where the drain voltage is at $V_{dd}$ and the gate voltage is ramped from 0V to $V_{dd}$). Assume that node $x$ was initially 0. Also notice that if $IN$ is low, node $x$ is in a high impedance state (not driven to one of the rails using a low resistance path).

Extra transistors can be added to provide a path to GND, but for this discussion, the simplified circuit is sufficient. Notice that the output charges up quickly initially, but has slow tail. This is attributed to the fact that the drive (gate to source voltage) reduces significantly as the output approaches $V_{dd} - V_{tn}$ and the current available to charge up node $x$ reduces drastically. Hand calculation using Eq. (6.24), results in an output voltage of 1.8V, which comes close to the simulated value.

WARNING: The above example demonstrates that pass-transistor gates cannot be cascaded by connecting the output of a pass gate to the gate input of another pass transistor. This is illustrated in Figure 6.35a, where the output of $M_1$ (node $x$) drives the gate of another MOS device. Node $x$ can charge up to $V_{dd} - V_{tn1}$. If node $C$ has a rail to rail swing, node $Y$ only charges up to the voltage on node $x - V_{tn2}$, which works out to $V_{dd} - V_{tn1} - V_{tn2}$. Figure 6.35b on the other hand has the output of $M_1$ (node $x$) driving the junction of $M_2$, and there is only one threshold drop. This is the proper way of cascading pass gates.

$\begin{align*}
Response &= 1, \quad if \quad V_{x0} > V_{x1} \\
&= 0, \quad if \quad V_{x0} < V_{x1}
\end{align*}$

where, $x_0$ and $x_1$ represent the source terminal of two different pass transistors.

3.1 Block Diagrams and Working of the design

If PUF response is produced solely on the basis of the differences of $V_{th}$ of a single transistor, the response will be less robust. This is due to less voltage difference which depends on the variance of $V_{th}$ process variations. To increase robustness 'n' NMOS pass transistors can be cascaded creating a multiplying factor such that output voltage is $V_{DD} - nV_{th}$. In this design we are cascading two pass transistors i.e. $n = 2$ (fig. 2).

Block level design of TV-PUF is projected in fig.3. Input which is a $n$-bit
challenge passes to the \( n - to - 2^n \) line decoder (Active High). This cause one of the output pins to level HIGH \((V_{dd})\) while all the others are LOW \((0V)\). Each output pin of decoder acts as input to two blocks of puf. These pins are connected to the gate terminals \((\text{IN})\) of the first MOSFET in each block. For a particular challenge, only one of the decoder outputs is HIGH, so only two blocks out of 128 have \( V_g = V_{dd} \). Rest of the blocks serves as the high-Z state as they have gate voltage \( = 0V \) implies cut-off region. As explained earlier, the output of the blocks with \( V_g = V_{dd} \) approaches its limiting voltage \( V_{out} = V_{dd} - 2V_{th} \). Due to process variation effects, \( V_1 \) and \( V_2 \) will be different. These two voltages further act as inputs to the sense amplifier. The response \((R)\) of the PUF is given as:

\[
R = 1, \quad V_1 \geq V_2 \\
R = 0, \quad V_1 < V_2
\]

All these operations do not affect the other transistors in the other blocks because they are in High Z condition, which means they are not operational.

All the above operations occur when the enable \((en)\) signal is HIGH. After

![Fig. 3. Block Diagram of proposed TV-PUF with 6 - bit challenge](image)

the output is obtained, the enable signal changes its state to LOW which activates transistors \( F_a \) and \( F_b \) thus causes \( V_1 = V_2 = 0V \). Transistors \( F_a \) and \( F_b \) are named as flush transistors. These force the voltages at their drain to become zero, thus re-establishing the initial condition of the TV-PUF. Fig. 4 demonstrates the simulation of output \( V_1 = V_2 \) when en=low and en=high. The motivation behind using the flush transistor is to increase the reliability of PUF by keeping the system close to ideal because in long run \( V_{GS} < V_{th} \) due to sub-threshold conduction. It implies that even when \( V_{GS} < V_{th} \) i.e, cut-off region NMOS transistor will conduct following the exponential curve of sub-threshold conduction. To generate a 64-bit key 64 different pair of blocks are implemented. After generating one bit, enable(en) signal is used to reset the output voltages to zero using transistor \( F_a \) and \( F_b \). As the time required to reset (fig. 4) the voltages \( V_1 = V_2 = 0V \) is approx. 35ps to 40ps for 65nm technology node which is much less than the time of conduction. Thus, the en signal need not have 50% duty cycle. In contrast to the ICID [12] and [20] in which current flows through all
the transistors for generation of every bit. The selective current flow mechanism reduces the power consumption of the TV-PUF in comparison to that of ICID.

4 Performance Evaluation

In this section, we evaluate our PUF by testing it for well-known performance metrics. The results are equivalent to other PUF designs close to the ideal values. The uniqueness of responses, reliability of the PUF with variations in temperature and supply voltage, uniformity, bit-aliasing and correlation values for TV-PUF are reported. We have further evaluated its performance on the basis of its frequency of operation and power consumption. We have also reported the performance metrics of TV-PUF on different silicon technology nodes, namely 45nm, 65nm and 90nm using BSIM level=54 PTM models.

4.1 Uniqueness, Uniformity, Bit Aliasing and Correlation Analysis

The response of an instance of a PUF for a particular challenge should be independent of the response of another instance of the PUF for the same challenge. The value of the uniqueness metric should ideally be 50%. It is measured by calculating the inter-die Hamming Distance of the different keys. The inter-die Hamming Distance follows a Normal distribution \( N(\mu, \sigma) \). The ideal values of \( \mu \) and \( \sigma \) are 50% and 0 respectively. For our experiment, Monte Carlo simulation is used to capture process variation in 100 different chips. Fig.5 shows the simulation results for inter-die hamming distance for all pairs of two chips.

\[
R_{XX}(j) = \sum_n x_n x_{n-j}
\]

It is desired that the response of the PUF is random, hence unpredictable. For ideal PUF response, the number of 1’s and 0’s should be equal. This metric of
**Table 1. Comparison of PUF characteristics with different proposed intrinsic PUF construction [21]**

<table>
<thead>
<tr>
<th>PUF Construction</th>
<th>( \mu_{\text{inter}} \pm \sigma_{\text{inter}} )</th>
<th>( \mu_{\text{intra}} \pm \sigma_{\text{intra}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feed-forward Arbiter PUF</td>
<td>38% ± 9.8%</td>
<td></td>
</tr>
<tr>
<td>Subthreshold Arbiter PUF</td>
<td>( \approx 50% ) ± 9.8%</td>
<td>&lt;5%</td>
</tr>
<tr>
<td>Ring Oscillator PUF</td>
<td>46.15% ± 4.8%</td>
<td>0.48%</td>
</tr>
<tr>
<td>Glitch PUF</td>
<td>41.5% ± 6.6%</td>
<td>&lt;6.6%</td>
</tr>
<tr>
<td>SRAM PUF</td>
<td>49.97% ± 0.3%</td>
<td>&lt;12%</td>
</tr>
<tr>
<td>Latch PUF</td>
<td>50.55% ± 3.04%</td>
<td>&lt;4%</td>
</tr>
<tr>
<td>Flip Flop PUF</td>
<td>36% ± 2.9%</td>
<td>&lt;13%</td>
</tr>
<tr>
<td>Butterfly PUF</td>
<td>( \approx 50% )</td>
<td>&lt;4%</td>
</tr>
<tr>
<td><strong>Proposed TV-PUF</strong></td>
<td>( \approx 51% )</td>
<td>&lt;4%</td>
</tr>
</tbody>
</table>

*the results encompass environmental fluctuations*

the PUF is measured by Uniformity. Similar to the inter-die distance, uniformity follows a normal distribution \( N(\mu, \sigma) \) and ideal values of \( \mu \) and \( \sigma \) are 50% and 0 respectively. If bit-aliasing happens, different chips may produce nearly identical PUF responses which is an undesirable effect. We estimate bit-aliasing of the \( l^{th} \) bit in the PUF identifier as the percentage Hamming Weight(HW) of the \( l^{th} \) bit of the identifier across \( k \) devices [22]. Also, neighboring bits must not influence each other. Each bit should be independent otherwise the PUF will be threatened by modelling attacks. In order to check if correlation exists in the test chip, the autocorrelation function is used: PUF reliability captures how efficient a PUF is in reproducing the response bits. We employ intra-chip HD among several samples of PUF response bits to evaluate this metric. To estimate the intra-chip HD, a \( n \) -bit reference response \( R_i \) is extracted from the chip \( i \) at normal operating condition (at room temperature using the normal supply...
Fig. 6. a) Intra-chip Hamming Distances with Temperature variation (-55 to 125°C) b) Intra-chip Hamming Distances with Supply voltage.

The same n-bit response is extracted at a different operating conditions (different ambient temperature or different supply voltage) with a value $R'_i$. For our experiment table-1 contains results of uniformity and reliability with the comparison of previous designs.

4.2 Comparison of performance for various Technology Nodes

For the PUF design to be robust, it must have promising performances on all platforms and technology nodes. With the exponential growth in semiconductor device modeling, we demonstrate that TV-PUF maintains its high performance across different semiconductor technology nodes in table-2.

<table>
<thead>
<tr>
<th>Technology Node</th>
<th>45nm</th>
<th>65nm</th>
<th>90nm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniqueness</td>
<td>50.02%</td>
<td>50.03%</td>
<td>50.10%</td>
</tr>
<tr>
<td>Uniformity</td>
<td>49.70%</td>
<td>49.84%</td>
<td>49.96%</td>
</tr>
<tr>
<td>Reliability</td>
<td>96%</td>
<td>96%</td>
<td>97%</td>
</tr>
<tr>
<td>Bit-aliasing</td>
<td>49.7%</td>
<td>49.84%</td>
<td>49.96%</td>
</tr>
<tr>
<td>Autocorrelation(1,2)</td>
<td>15.3, 15.45%</td>
<td>15.48, 15.58%</td>
<td>17.31, 16.9%</td>
</tr>
</tbody>
</table>

Table 2. Comparison of PUF characteristics on different technology nodes

5 Effect of Aging on Reliability

In MOSFETs, the key reliability issues are Bias temperature instability (BTI) and Hot carrier injection (HCI). In our model we have used only NMOS transistors. Use of only NMOSs in PUF design exclude the NBTI effect in our design. In devices performance NBTI is the major reliability concern as compared to HCI [16]. So this exclude the concern over NBTI. The following analytical model [23] can
For better device robustness, one should expect no variation in output bit for a particular challenge. Suppose for a challenge initially $V_1 > V_2$. Over the device aging, one should expect that the same relation will hold. It implies that $\Delta V_{th}$ of different pass transistors should be independent of $V_{th}$. This argument is supported by the fig.7, which shows relation output voltage $V_1$ and $V_2$ for a particular challenge vs. time for one challenge. Fig.7 shows the independent of $\Delta V_{th}$ over initial process variation in $V_{th}$ which is also reflect by equation 1 if $V_{gs} \approx V_{th}$. These analyses are performed using MOSRA [24] for 45nm, 65nm and 90nm technology nodes accounting HCI effect. Simulation results for 10 different chips shows that impact of HCI is negligible on reliability. Although the $\Delta V_{th}$ increases approx. 50% for each pass transistors in span of 5 years, voltage difference at output($V_1 and V_2$) hardly changes only by 0.2-0.3mV. This argument is also supported by fig.7 as it shows that due to increase in $V_{th}$ over time, output voltages $V_1$ and $V_2$ start decreasing but the relative difference remain constant. Major hurdle in reliability concern is the sensitivity of sense amplifier. Fig.8 project the gray zone with a voltage difference in $\pm dV$. $dV$ will further depend on the sensitivity of sense amplifier. Voltage difference lower this level demand a design of sense amplifier with better sensitivity which again cause increase in power consumption.

\[
\Delta V_{th} = \frac{q}{C_{ox}} K \sqrt{C_{ox}(V_{gs} - V_{th})e^{\frac{E_{ox}}{E_0}}e^{\frac{-\phi_{it}}{q\lambda E_m}}}t^n
\]  

where, $E_m = \frac{V_{ds} - V_{dsat}}{l}$, $V_{dsat} = \frac{(V_{gs} - V_{th} + 2V_i)L_{eff}E_{sat}}{V_{gs} - V_{th} + 2V_i + A_{bulk}L_{eff}E_{sat}}$

$E_{ox} = \frac{V_{gs} - V_{th}}{T_{ox}}$, $C_{ox} = \frac{\varepsilon_o x}{T_{ox}}$, $V_i = \frac{kT}{q}$

Fig. 7. Output Voltage vs. aging time duration
Fig. 8. Voltage difference across PUF output terminals

6 Comparison of TV-PUF with Sub-threshold PUF, Super-threshold PUF and other existing PUF schemes

The clock frequency of the sense amplifier in this design will be affected by the delay in measuring two responses. As fig.4 demonstrates, after measuring one responses we change input of decoder and flush the output terminal of PUF using en signal. For next decoder input, different blocks will get selected which further result in a different output(0,1) from amplifier output terminal. The voltage difference between $V_1$ and $V_2$ increases with time and approaches the limiting value caused by threshold voltage mismatch. Hence the input can be sampled any time in the rising period of voltage at $V_1$ and $V_2$ depending on the desired voltage difference. In previous works [10], the design of a lightweight PUFs operating in the subthreshold region is proposed. The operation of devices in the sub-threshold or super-threshold region invokes huge delay and consequently, it works at a much lower clock frequency. Also, the power consumption of proposed design is comparable to previous ones due to less circuit complexity. The following table compares power consumption and clock frequency of proposed design with previous works. However stated power consumption and no. of transistors of our design doesn’t include decoder and sense amplifier designs.

<table>
<thead>
<tr>
<th></th>
<th>Sub-threshold</th>
<th>Super-threshold</th>
<th>$TV-$ PUF</th>
<th>HUI $D$</th>
<th>Thermal [25]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Power (µW @ 1 MHz)</td>
<td>0.047 pJ</td>
<td>0.136 pJ</td>
<td>0.181 x 10^-3 pJ</td>
<td>500 pJ</td>
<td>0.14 pJ</td>
</tr>
<tr>
<td>Power (µW @ 1 GHz)</td>
<td>0.047 pJ</td>
<td>0.136 pJ</td>
<td>1.81 x 10^-3 pJ</td>
<td>NA</td>
<td>NA</td>
</tr>
<tr>
<td>Energy/cycle</td>
<td>0.047 µW</td>
<td>0.136 µW</td>
<td>1.81 µW</td>
<td>500 µW</td>
<td>0.14 µW</td>
</tr>
<tr>
<td>No. of Transistors</td>
<td>1672</td>
<td>1672</td>
<td>586</td>
<td>NA</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 3. Comparison with Sub-threshold and Super-threshold PUF(for 1-bit generation)
7 Conclusion

The TV-PUF requires voltage comparison. This has certain advantages over delay-based PUFs. In delay-based PUFs, the delay has to be substantial for the delays between the two blocks/paths to be detectable by the sensor (for both RO-PUF and Arbiter PUF). On one side arbiter, PUF suffers from the requirement of the symmetric path in the circuit and hold time of arbiter. Similarly, at the inverters in Ring Oscillator PUFs, PMOS is used which is susceptible to NBTI which causes degradation in reliability with time. Delay PUFs have a large path (e.g., Arbiter PUF) or it has to wait for many cycles of operation before producing the response (for RO-PUF) which makes the effective path quite long for RO-PUFs. This causes the time of operation to be huge and throughput is very low. TV-PUF, on the other hand, has a very low critical path delay which makes its clock frequency much larger than the delay based PUFs. The design of TV-PUF is quite similar to that of a Ring Oscillator PUF (RO-PUF), except the fact that the Ring Oscillator PUF compares the frequencies of two blocks whereas the TV-PUF compares the cumulative threshold voltages of two cascaded MOSFETs in a block. It is shown that the RO-PUF is impossible to be modeled [26] and is safe from modeling attacks.

Similar to the RO-PUF, the TV-PUF also requires an exponential number of blocks, that is, for a n-bit challenge the number of blocks required is $2^n$. If it is required to generate more responses from the TV-PUF to strengthen the security of a system, it becomes an obstacle. Increasing the number of challenge bits by one doubles the hardware requirement of the TV-PUF. This causes a disruption in the environment it is placed in. A future direction of research may be to investigate various combinations of the blocks to reduce the exponential dependency of hardware requirement on the size of the response.

8 Future Work

A significant improvement will be to reduce the exponential hardware requirement of the TV-PUF. As its key generation mechanism correlate with RO-PUF, those techniques can be employed in this design. Reliability of this design can be further increased by using a control circuit which neglects the inputs challenges which generate output voltage difference less than a specified threshold depending on the sensitivity of sense amplifier. Attempts can be made to bring about a better optimization between the sensitivity of the sense amplifier and reliability.

References


