Journal of
Low Power Electronics and Applications

MDPI

Article

Sizing of SRAM Cell with Voltage Biasing Techniques for Reliability Enhancement of Memory and PUF Functions †

Chip-Hong Chang *, Chao Qun Liu, Le Zhang and Zhi Hui Kong

School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798, Singapore; cliu016@e.ntu.edu.sg (C.Q.L.); lzhang15@e.ntu.edu.sg (L.Z.); zhkong@ntu.edu.sg (Z.H.K.)

* Correspondence: ECHChang@ntu.edu.sg; Tel.: +65-6790-5873

† This paper is an extended version of our paper published in 2015 IEEE International Symposium on Circuits and Systems (ISCAS 2015).

Academic Editors: Osnat Keren, Ilia Polian and Sanu Mathew
Received: 11 July 2016; Accepted: 18 August 2016; Published: 24 August 2016

Abstract: Static Random Access Memory (SRAM) has recently been developed into a physical unclonable function (PUF) for generating chip-unique signatures for hardware cryptography. The most compelling issue in designing a good SRAM-based PUF (SPUF) is that while maximizing the mismatches between the transistors in the cross-coupled inverters improves the quality of the SPUF, this ironically also gives rise to increased memory read/write failures. For this reason, the memory cells of existing SPUFs cannot be reused as storage elements, which increases the overheads of cryptographic system where long signatures and high-density storage are both required. This paper presents a novel design methodology for dual-mode SRAM cell optimization. The design conflicts are resolved by using word-line voltage modulation, dynamic voltage scaling, negative bit-line and adaptive body bias techniques to compensate for reliability degradation due to transistor downsizing. The augmented circuit-level techniques expand the design space to achieve a good solution to fulfill several otherwise contradicting key design qualities for both modes of operation, as evinced by our statistical analysis and simulation results based on complementary metal–oxide–semiconductor (CMOS) 45 nm bulk Predictive Technology Model.

Keywords: physical unclonable function (PUF); hardware security; Static Random Access Memory (SRAM); process variation; memory failures

1. Introduction

Going with the trend of increasing connectivity and services offered by computing devices, the amount of sensitive information processed by and stored on computing devices is growing rapidly. Recently, Physical Unclonable Function (PUF) has sprouted up as a promising primitive to enforce data privacy and access control to electronic devices. Among the PUF implementations [1–5], SRAM-based PUF (SPUF) has attracted tremendous attention. This is because SRAM, being an integral part of computer memory sub-system, plays a pivotal role in trusted computing platforms. The ability to use the storage cells of SRAM inseparably as PUF will replace or augment memory curtaining as a stronger fortification to the roots of trust for memory authentication. Unfortunately, existing SPUF cells cannot be doubled as regular storage elements. The exploitation of process-variation induced device mismatches in the cross-coupled inverter cell for random, unique and reliable response bit generation is detrimental to the regular memory operation, as it will result in increased parametric failures due principally to destructive read and unsuccessful write operations. Archived literatures reported wide varieties of design approaches that either improve the qualities of SPUF or harness the readability and writability of SRAM cells as data storage elements, albeit independently [5,6]. However, no design strategy has been proposed to tackle the conflicting quality requirements to unify the two different modes of operation in an SRAM cell.

In this paper, we expand the design space exploration of our preliminary proposal in [7], by taking into consideration circuit-level techniques that can be used to mitigate conflicting design criteria of both PUF response reliability and memory data stability. The proposed design procedure provides greater opportunity to meet tighter functional and performance specifications of both modes of operation at a small cost of area, delay and power overheads.

The remainder of this paper is organized as follows. Section 2 analyzes the impacts of process variations (PV) and transistor sizing on the two different working modes of SRAM. Section 3 presents the proposed design methodology to expand the design space for PUF and memory stability optimization. Section 4 discusses the comparison results for seven different design configurations. The paper is concluded in Section 5.

2. Impact of Process Variations (PV) on Different Operation Modes

The block structure of a complete SPUF is shown in Figure 1a. The core of the SPUF is a two-dimensional array of storage cells, where the horizontal rows are referred to as word-lines (WLs) and the two vertical lines that run through each cell are called the bit-line (BL) and complement bit-line (BLB). An individual cell can be selected by the row and column decoder circuit with an external input address, addr. One sense amplifier is attached to each bit-line pair of the column to accelerate read operation while the write driver circuit improves the access time in write operation. When the SPUF is set to PUF mode by the mode_sel input, the SPUF is reset and powered up. Due to mismatches and noise between the cross-coupled inverters formed by the transistor pairs, M1–M2 and M4–M5, a *response* bit is generated in each of the SRAM cells. An input *challenge* is applied to the WL of the addressed cells to turn on the access transistors, M3 and M6. The *response* bits are then read from the bit-lines [2]. In the memory mode, regular read and write operations are performed on the same SRAM array after the SPUF has been fully powered up. For a read operation, the small voltage difference between the true and complement bit-lines is amplified by the sense amplifier and the data stored in the addressed cell is transferred to the output. For a write operation, the bit-lines are initialized with the data that is to be written into the cell with its word-line asserted. These two different modes of operation on a 6-transistor (6T) SRAM cell are shown in Figure 1b.

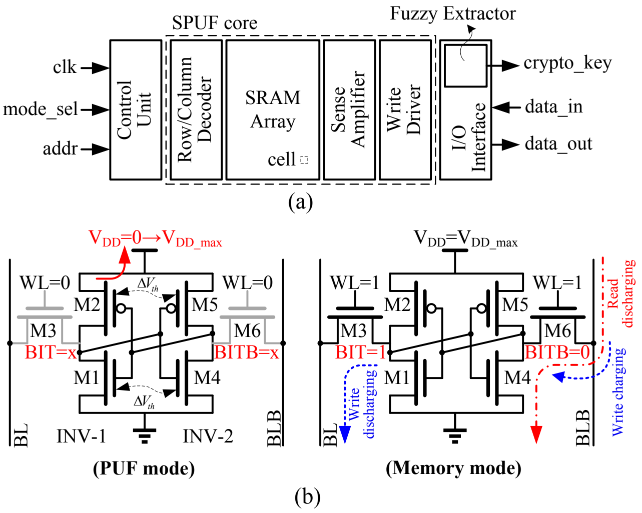


Figure 1. (**a**) Architectural diagram of an SRAM-based PUF (SPUF); (**b**) PUF and memory modes of operation of a 6-transistor Static Random Access Memory (SRAM) cell in an SPUF.

While the two modes of operation seem independent, they cannot be designed independently because the PV influences on them occur simultaneously. As shown in Figure 2, when the threshold voltage variation $\sigma_{V_{th}}$ increases, the failure rate of memory operations increases, whereas the reliability of response bits for PUF operation improves.

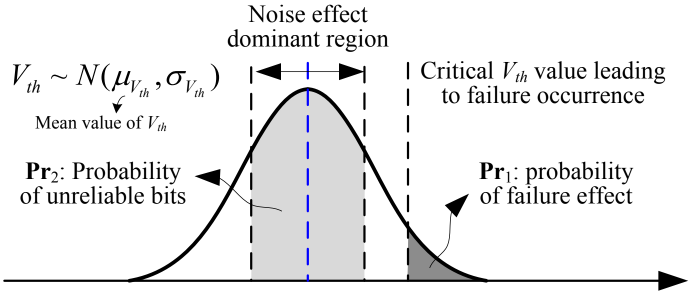


Figure 2. Design conflict of a SRAM cell for dual-mode applications. Increasing the variation of V_{th} improves the quality of PUF at the expense of higher failure rates in regular memory read/write operations.

2.1. Impact of PV on Physical Unclonable Function (PUF) Quality

During the power-up process, all transistors in an SRAM cell operate in the sub-threshold region. The currents that drive the two output nodes, BIT and BITB, in Figure 1b are determined by the strength of the two inverters in the cell. The currents flowing through M1–M2 and M4–M5 will simultaneously pull up the voltage at BIT and BITB, respectively. Owing to the transistor mismatches between the two inverters, one of the two nodes may reach the "trip-point" (V_{trip}) of the inverter faster than the other. If the voltage at BITB increases to V_{trip} first, the output node BIT of INV-1 will be pulled down to ground. As the voltage at BIT descends, M5 will be fully turned on eventually and pull the voltage at BITB to the supply voltage V_{DD}. This process is shown in Figure 3a. Figure 3b shows the state evolution of the SRAM in the phase plane [8]. During the initial phase of power-up, BIT and BITB are approaching the meta-stable point (V_{trip}) together. Due to mismatches of the two inverters, the SRAM cell will stabilize to one of the two stable states (V_{BIT} = "1" or "0").

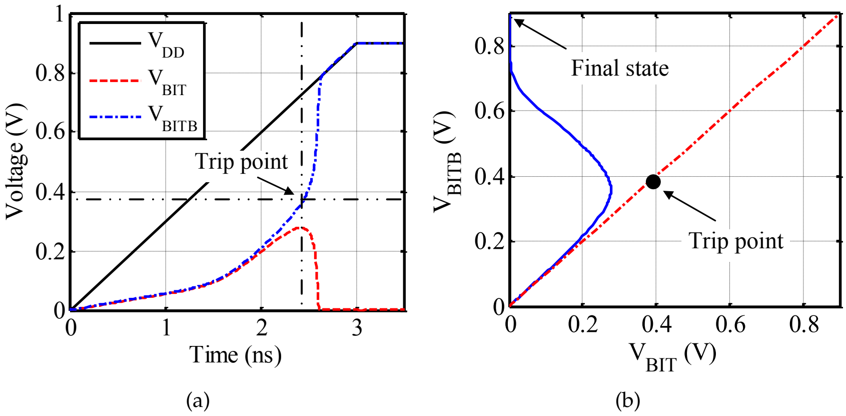


Figure 3. (**a**) BIT and BITB node voltages of SRAM cell during power-up process. The two curves set apart when one of them reaches the trip point of the inverter; (**b**) state evolution at power-up.

Since each response bit is independently generated from an SRAM cell and dominated by independent intra-die variations such as random dopant fluctuation and line edge roughness, the randomness among bits generated is guaranteed. While the response bits produced from individual SRAM cells are generally uncorrelated, their reliability may be jeopardized when the noise effects overwhelm the mismatches between the two inverters in an SRAM cell. In addition, temporal post-manufacturing variations such as unexpected change of environmental conditions in which the devices operate will also impact the reliability.

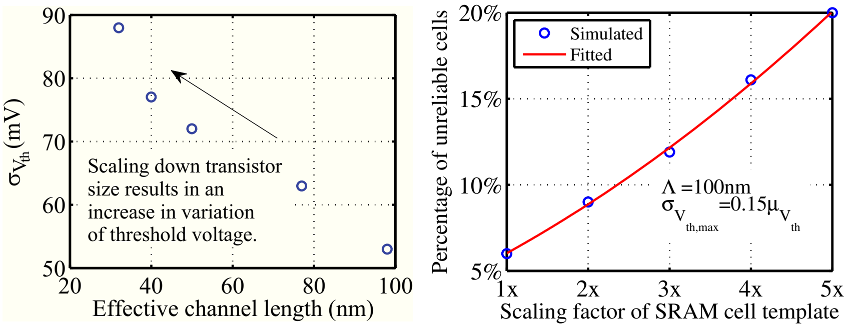
2.1.1. Mismatch of INV-1 and INV-2

The reliability of a PUF is a measure of the reproducibility of its response to the same challenge at different times and conditions [1]. It is the complement of the bit error rate (BER) of its responses, which can be measured by calculating the intra-chip Hamming distance of the responses generated from a single PUF chip with the same challenge. Ideally, BER should be 0% and reliability should be 100%. The reliability of the start-up state of an SRAM cell is highly dependent on the degree of mismatch between INV-1 and INV-2. The mismatch is dominated by the intra-die variation where the variation of V_{th} is the most prevalent. During power-up process, a small deviation of V_{th}, denoted as ΔV_{th}, between the transistors of INV-1 and INV-2 may change the drive currents and shift the intersection of their voltage transfer curves (VTC) [9], causing asymmetric output sensitivity to external stress. The magnitude of ΔV_{th} is primarily determined by $\sigma_{V_{th}}$, which has been proven empirically to be related to the transistor geometry as follows [10]:

$$\sigma_{V_{th}} = \frac{\Lambda \cdot \sigma_{V_{th,\max}}}{\sqrt{(W \cdot L)}}, \tag{1}$$

where Λ is a technology-dependent parameter and $\sigma_{V_{th,\max}}$ is the maximum value of $\sigma_{V_{th}}$.

Based on Equation (1), $\sigma_{V_{th}}$ can be reduced by sizing up the transistors in an SRAM cell (see Figure 4a). Depending on the transistor sizes, the mismatch between the inverters of a cell may be reduced to an extent that its output state becomes highly susceptible to variations of external condition or in the worst case, completely determined by the noise present in the SPUF. This can be explained in Figure 5. Based on the propensity of the output to settle at one of the two states upon power-up, SRAM cells can generally be categorized as 1-skewed, 0-skewed or neutral cells [2]. Highly 1-skewed (0-skewed) cells have larger mismatches between its two inverters and have a high tendency to power up to "1" ("0"). Neutral cells have equal probability to power up to either "1" or "0", depending on the noise.



(a) (b)

Figure 4. (**a**) Changes in $\sigma_{V_{th}}$ with transistor effective channel length [10]; (**b**) relationship between the percentage of unreliable bits and the scaling factor of SRAM cell template, where the widths of pull-down, pull-up and access transistors of a unit-size SRAM cell are $W_{M1} = 180$ nm, $W_{M2} = 90$ nm and $W_{M3} = 135$ nm, respectively, and all transistors have $L = 45$ nm.

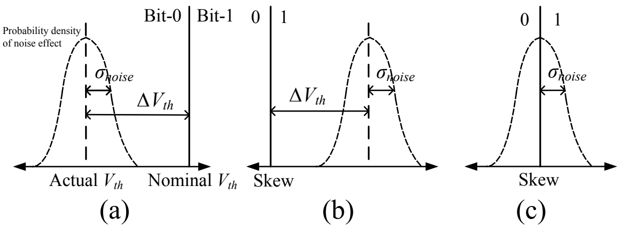


Figure 5. Three types of SRAM cells: (**a**) 0-skewed cells, (**b**) 1-skewed cells and (**c**) neutral cells.

Figure 4b depicts the simulation results of the response bits generated by powering up the same SRAM cell 10 times in the presence of noise, whose parameters are set as $f_{min} = 10^6$ and $f_{max} = 10^{11}$, where f_{min} and f_{max} are, respectively, the base and maximum frequencies used for modeling frequency dependent noise sources [11]. An SRAM cell that does not generate the same start-up value, i.e., "1" or "0", is consistently regarded as unreliable. The percentage of unreliable bits obtained from 500 cells of an SRAM array is plotted against the size of SRAM cell. The data points can be fitted by a second order polynomial, which indicates that the percentage of unreliable bits increases almost linearly with the transistor size.

2.1.2. Loop-Gain at Trip Point

The "Loop-gain" at the metastable trip point of a cross-coupled inverter pair is an important factor that affects the reliability of SPUF [12]. The node voltages of SRAM cells stabilize faster if the loop-gain is higher. According to [13], the loop-gain is formulated as

$$LG(V_{BIT}) = \frac{\partial \mathrm{VTC}_2}{\partial V_{BIT}} \cdot \frac{\partial \mathrm{VTC}_1}{\partial V_{BITB}}, \tag{2}$$

where $V_{BIT} = \mathrm{VTC}_1(V_{BITB})$ or $V_{BITB} = \mathrm{VTC}_2(V_{BIT})$. VTC refers to the voltage transfer characteristic function and can be formulated as follows [9]

VTC for INV-2:

$$\begin{aligned} V_{BITB} &= V_{DD} + V_T \ln\left\{\frac{1}{2}\left[1 - G + \sqrt{(G-1)^2 + 4\exp\left(-\frac{V_{DD}}{V_T}\right)G}\right]\right\}, \\ G &= \exp\left(\frac{2V_{BIT}}{nV_T} - \ln\frac{I_{S5}}{I_{S4}} - \frac{V_{DD}}{nV_T} - \frac{V_{th,4} - V_{th,5}}{nV_T}\right), \end{aligned} \tag{3}$$

where I_{Si} and $V_{th,i}$ refer to the saturation current and threshold voltages of the transistor Mi in Figure 1. VTC for INV-1 can be derived similarly.

As pointed out in [12], large loop-gain degrades the reliability of SPUF as the amplified noise in the power-up process overwhelms the transistor mismatches due to process variation. From Equations (2) and (3), the loop-gain is mainly governed by V_{DD}, V_{th} and V_T.

2.2. Impact of PV on SRAM Read/Write Failures

While the strength mismatches between different transistors of an SRAM cell are good to skew its power-up value to resist SPUF response bit flipping under temporal variations in operational condition, they can increase the read failure (RF) and write failure (WF) rates in the memory-mode. RF occurs when the voltage (V_{read}) sensed at BIT (or BITB) rises above the inverter read trip point (V_{tripRD}) to cause the bit stored in the SRAM cell to flip accidentally (see Figure 6a). As shown in Figure 1b, when the bit-line BL is precharged, the access transistor M6 and the pull-down transistor M4 act as a voltage divider. The PV induced voltage deviation at BITB may trigger a flip of the data bit stored in the SRAM cell [14]. To avoid RF, the relative strength of M4 and M6 (known as beta ratio ($\beta_{npd-nax}$)) should be designed larger, where $BR_{npd-nax}$ is given by

$$BR_{npd-nax} = \frac{\beta_{npd}}{\beta_{nax}} = \frac{\beta_{M4}}{\beta_{M6}} = \frac{\frac{\mu_{eff} C_{ox} W_{M4}}{L_{M4}}}{\frac{\mu_{eff} C_{ox} W_{M6}}{L_{M6}}}, \tag{4}$$

where μ_{eff} is the effective mobility and C_{ox} is the oxide capacitance. WF refers to the failure of SRAM cells to respond to the data bit written into it within an acceptable time (see Figure 6b) [14]. Consider the SRAM cell in Figure 1b, if "0" is to be written into the cell that stores logic "1", the voltage at BIT will be discharged to a low value. If this value is higher than the trip point of INV-2, the write operation fails and the stored data bit remains unchanged. To reduce the probability of WF, the beta ratio of the access transistor and pull-up p-channel metal-oxide-semiconductor field-effect transistor (PMOS) ($BR_{nax-pup} = \beta_{nax}/\beta_{pup} = \beta_{M3}/\beta_{M2}$) should be designed such that the write time is less than the word-line turn-on time.

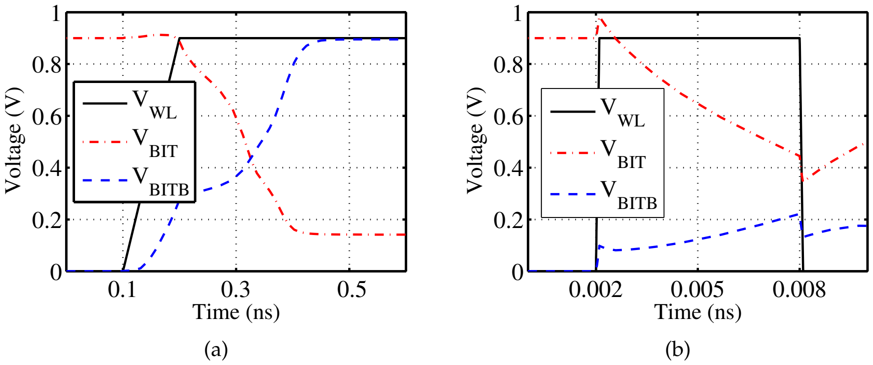


Figure 6. (**a**) Read and (**b**) write failure effects in an SRAM cell. The original bit preserved in the cell is 1.

It is possible to reduce the probability **Pr**(RF,WF) of memory access failures by manipulating the transistor sizes [14]. The relationship between **Pr**(RF,WF) and the sizing of each transistor in an SRAM cell is shown in Figure 7. Curve fitting by polynomial functions shows that sizing up the transistors helps to reduce failure probability. This is expected since $\sigma_{V_{th}}$ decreases when the size of transistors increases according to Equation (1).

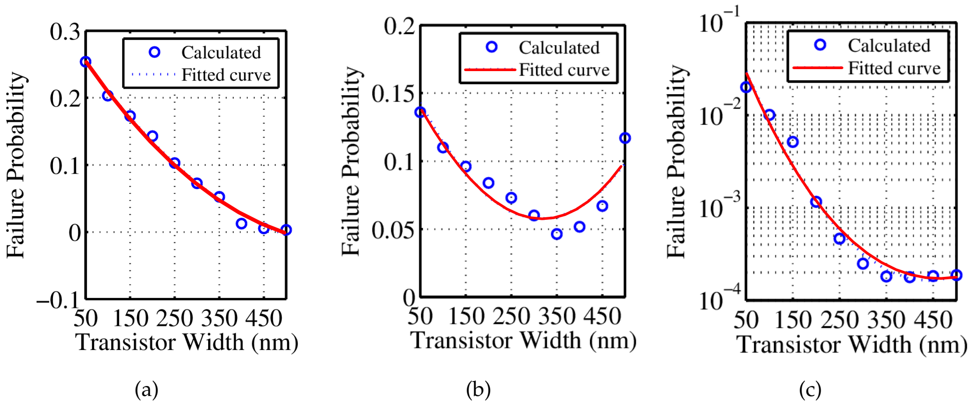


Figure 7. The relationship between the failure probability and the transistor widths (**a**) M1; (**b**) M2; (**c**) M3.

3. Proposed Design Method for Dual-Mode PUF

To design a dual-mode SPUF, the transistors in the SRAM cell need to be sized such that the quality expectations can be met for both modes. For memory mode, the memory failure probability should be kept as low as possible, whereas for PUF mode, the randomness, uniqueness and reliability of response bits should be made as high as possible. As randomness and uniqueness can be easily enhanced by fuzzy extractor at a smaller cost, this paper focuses on the reliability, which is more critical and difficult to improve for SPUF. In addition, failure probability and PUF reliability, area, leakage current and dynamic power constraints using a targeted process technology will also be considered.

3.1. Circuit-Level Techniques

In addition to transistor sizing, circuit-level techniques such as Word-Line Voltage Modulation (WLM), Dynamic Voltage Scaling (DVS), Negative Bit-Line (NBL) and Adaptive Body Bias (ABB) can also be applied to mitigate SRAM failures [15–18]. For ease of exposition, the nominal supply voltage is denoted as "SRAMVDD". In addition, since static noise margin (SNM) is more conveniently used in practice to calibrate the SRAM cell stability against DC noise [9], it will be used as a measure to compare the efficiency of different adaptive techniques.

3.1.1. WLM

The WLM technique proposed in [15] is shown in Figure 8. This technique increases the read stability by lowing the word-line voltage to increase $BR_{npd-nax}$. The WL selector adapts different supply voltage to the WL (V_{DD_WL}). Its control signal is determined by the requirements of speed and stability. By lowering the WL voltage and keeping the cell voltage supply at SRAMVDD, the strength of the access transistors will be lowered to improve $BR_{npd-nax}$. However, lowering WL voltage will also degrade the write ability due to the reduced $BR_{nax-pup}$. This problem can be solved by applying different V_{WL} for the read and write operations. This scheme has a drawback as it can only improve read stability but not write ability, as improving V_{WL} during write operation will degrade the SNM of the half-selected cells (which are equivalent to performing a read operation on these cells) even though they are not accessed.

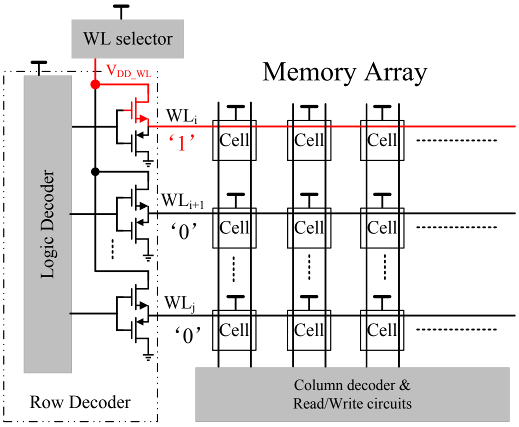


Figure 8. SRAM array with word-line power supply selector.

3.1.2. DVS

A column-based DVS technique was proposed in [16] to improve both read and write margins. A differential voltage is created between the SRAM cell and the WL nodes to optimize the read and write margins separately without compromising each other. The mechanism is shown in Figure 9. Each column of memory cells has a common voltage supply, which is either V_{DD_hi} or V_{DD_lo}, depending on the cell operation mode. During the read operation, the column of accessed cells switches its V_{DD} to V_{DD_hi}, which is higher than the word-line voltage SRAMVDD. During the write operation, the column of accessed cell switches its V_{DD} to V_{DD_lo}, which is lower than SRAMVDD. For the columns that are not accessed, their V_{DD} will always be V_{DD_hi}, as they are always experiencing "read stress" even though they are not selected to be read or written. When a cell is in standby mode, its V_{DD} is switched to V_{DD_lo} to minimize leakage power. In this way, both read and write stability will be increased.

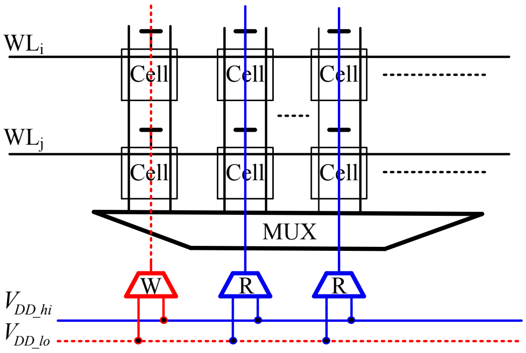


Figure 9. Dynamic power supply based on the selection of read or write operation.

3.1.3. NBL

A write-ability enhancement technique that combines negative bit-line and V_{DD} collapse scheme with high configurability was proposed in [17]. Like DVS, the V_{DD} collapse technique improves $BR_{nax-pup}$ by weakening the pull-up PMOS. The circuit implementation of NBL is shown in Figure 10. During write operation, the bit-line is driven to a negative voltage to write "0". This will strengthen the driving capability of pass transistors, thus improving $BR_{nax-pup}$ and the write ability. As this technique is only applied during the write operation, it has no influence on the read stability.

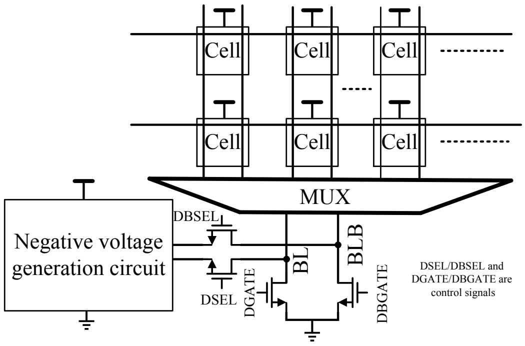

Cell
MUX
Negative voltage generation circuit
DBSEL
DSEL
DGATE
BL
BLB
DBGATE
DSEL/DBSEL and DGATE/DBGATE are control signals

Figure 10. Implementation of adaptive negative bit-line circuit.

3.1.4. ABB

The SRAM cell failures are not only affected by intra-die variations, but also by inter-die variations. This can be understood by analysing the change in V_{read} and V_{tripRD} with an inter-die V_{th} shift ($\Delta V_{thINTER}$) of the cell transistors [18]. As shown in Figure 11, when V_{BIT} = "1",

$$\begin{aligned} V_{tripRD} &= V_{trip}(0) + \Delta V_{thINTER}\frac{\sqrt{\beta_{M1}/\beta_{M2}}-1}{\sqrt{\beta_{M1}/\beta_{M2}}+1}, \\ V_{read} &= V_{trip}(0) - \Delta V_{thINTER}\left(1 \pm \frac{\sqrt{\beta_{M4}/\beta_{M5}}}{\sqrt{1+\beta_{M4}/\beta_{M5}}}\right). \end{aligned} \tag{5}$$

With a negative $\Delta V_{thINTER}$, V_{read} increases but V_{tripRD} decreases [18]. This will reduce ($V_{tripRD} - V_{read}$), thus degrading the read stability. When $\Delta V_{thINTER}$ is positive, write failure will increase due to the weakening of access transistors. Such stability degradation can be eliminated by employing the ABB technique proposed in [18]. Take n-channel metal-oxide-semiconductor field-effect transistor (NMOS) as an example: the body bias (BB) can be utilized to adjust V_{th} as

$$V_{th} = V_{th0} + \gamma(\sqrt{2\varphi_B + V_{SB}} - \sqrt{2\varphi_B}), \quad (6)$$

where γ is the body effect parameter, φ_B is a physical parameter and V_{SB} is the voltage between the source and body.

V_{th} of M1 to M6 can be adjusted by their respective body bias voltages V_{B1}–V_{B6}. For both NMOS and PMOS in general, reverse body bias (RBB) increases V_{th} whereas forward body bias (FBB) reduces V_{th} [19]. According to Equation (5), the increase of V_{th} with RBB leads to higher ($V_{tripRD} - V_{read}$), which, in turn, reduces the read failure probability. Conversely, the reduction of V_{th} with FBB increases the driving capability of the access transistors, which leads to reduced write time and improved write ability.

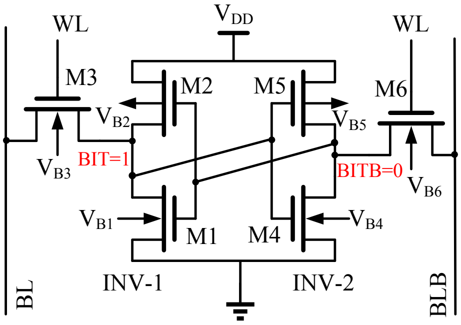


Figure 11. SRAM cell with adaptive body bias.

Even though ABB can be applied to reduce the read and write failures, it is not without limitations. The ideal case of biasing each transistor independently is not practical due to a large amount of voltage supplies and area overhead [20]. As a practical expedient, the same body bias can be applied to all cells to mitigate the degraded cell stability against inter-die variations. In addition, unlike PMOS BB, which can be applied in a standard dual-well process, the application of BB to NMOS requires a triple-well process [18]. If cost permits, it is preferable to adjust the BBs of both PMOS and NMOS transistors to mitigate the inter-die variations. Under the constraint of manufacturing cost, BB can only be applied to PMOS transistors to reduce the read and write failures despite less efficiency [18]. Since the inter-die V_{th}s of PMOS and NMOS may move in different directions, it is also desirable to control the BB of the two types of transistors individually.

3.1.5. Comparison of Different Reliability Enhancement Techniques

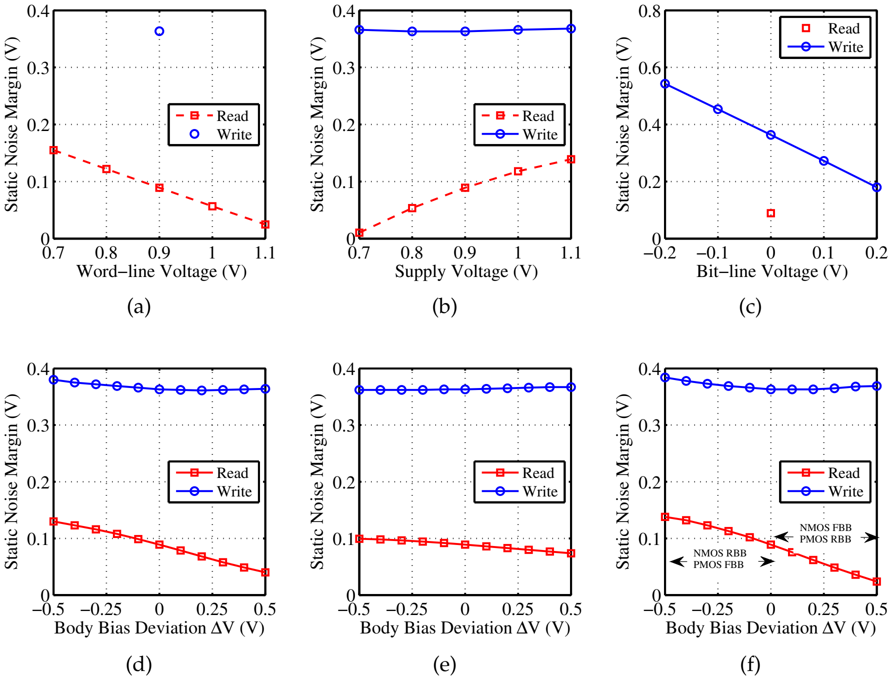
The SRAM cells incorporated with the four aforementioned techniques were simulated to compare their effectiveness in improving the SNM. For WLM and DVS, the word-line and supply voltages were swept by ±20% from 0.7 V to 1.1 V. For NBL, the BL voltage is swept from -0.2 V to 0.2 V. For ABB, the SNMs were simulated with BB applied to only NMOS, only PMOS and both NMOS and PMOS transistors, respectively. Due to the non-linear relationship between V_{th} and BB voltage, the BBs were swept for a large range from -0.5 V to 0.5 V. By representing the deviation of BB by ΔV, the BB voltages, V_{bn} and V_{bp}, for NMOS and PMOS, respectively, are given by:

$$\begin{aligned} V_{bn} &= \Delta V, \\ V_{bp} &= \text{SRAMVDD} + \Delta V. \end{aligned} \quad (7)$$

The SNM obtained by applying each circuit-level enhancement technique is shown in Figure 12. The efficiency is measured by the SNM sensitivity S as follows:

$$S = \left| \frac{\partial V_{SNM}}{\partial V_X} \right|, \tag{8}$$

where V_X is the word-line voltage for WLM, supply voltage for DVS, bit-line voltage for NBL, or body bias voltage for ABB.



(a) (b) (c)

(d) (e) (f)

Figure 12. Static noise margins by applying different circuit-level techniques: (**a**) Word-Line Voltage Modulation (WLM); (**b**) Dynamic Voltage Scaling (DVS); (**c**) Negative Bit-Line (NBL); (**d**) Adaptive Body Bias (ABB) (NMOS only) (**e**) ABB (PMOS only); (**f**) ABB (both NMOS and PMOS).

It can be seen that all these techniques have improved the cell stability by strengthening or weakening certain transistors. Even though the strengthening or weakening effect can be applied in the row (by modulating WL voltage) or column (by modulating bit-line or cell power supply voltages) direction, column-based approaches are preferred because they can blend well with column multiplexing and write masking [17]. A comparison of the functionalities of different adaptive techniques is shown in Table 1.

Table 1. Comparison of different adaptive techniques.

Technique	Column-Based	*S*(Read)	*S*(Write)	Other Drawbacks
WLM	×	0.325	0	-
DVS	√	0.323	0.006	-
NBL	√	0	0.908	-
ABB (NMOS only)	×	0.009	0.016	Triple-well process, Large area overhead
ABB (PMOS only)	×	0.026	0.006	
ABB (Both MOSFETs)	×	0.114	0.015	

WLM, DVS, NBL and ABB refer to Word-Line Voltage Modulation, Dynamic Voltage Scaling, Negative Bit-Line and Adaptive Body Bias, respectively. *S* refers to the SNM sensitivity.

From the table, it can be seen that the four adaptive techniques improve the SNM in their own ways. WLM is not column-based, but it is the most efficient in increasing the read SNM. DVS is efficient in increasing the read SNM but is inefficient in increasing the write SNM. NBL cannot increase read SNM, but it is the most efficient in increasing the write SNM. ABB is not efficient in increasing either the read SNM or write SNM, but it can be used to mitigate inter-die variations. Although DVS and ABB may not improve write SNM efficiently, they have the potential to reduce write failures. This is because write SNM is not a strong indicator of dynamic write ability [21]. The latter is better analyzed by the relative strength of access transistors and pull-up PMOS transistors.

3.2. Problem Formulation

Based on the above analysis, the design space of an SPUF is shown in Figure 13. The objective of the design problem is to find an optimal point $(\mathbf{X}, \mathbf{M})$ that combines the key parameters $\mathbf{X}$ (e.g., W of its transistors) associated with SRAM cell dimension and the aggressiveness $\mathbf{M}$ (e.g., by changing the voltage V_X) of WLM, DVS, NBL adaptation, to minimize

$$F(\mathbf{X}, \mathbf{M}) = \min_{\mathbf{X},\mathbf{M}} \mathbf{Pr}(\mathrm{RF},\mathrm{WF}), \tag{9}$$

subject to the following constraints:

$$\begin{cases} \text{reliability}(\text{ECC}[n,k,t]) \geq \text{reliability}_{\text{min}}, \\ \text{area\_cell} \leq \text{area\_cell}_{\text{max}}, \\ \text{leakage\_cell} \leq \text{leakage\_cell}_{\text{max}}, \\ \text{power\_cell} \leq \text{power\_cell}_{\text{max}}, \end{cases} \tag{10}$$

where $\text{ECC}[n,k,t]$ is the error correction code specified by the code word length n, number of information bits k and number of correctable errors t. The PUF reliability is measured by the complement of BER. The area_cell, leakage_cell and power_cell refer to the area, standby power and dynamic power dissipated during the read and write operations of the SRAM cell.

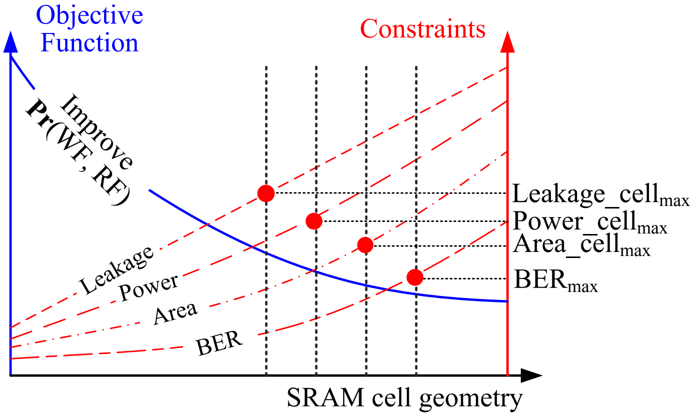


Figure 13. Design space exploration of SRAM cell for dual-mode SPUF design.

The design problem can be solved by the design flow described in Figure 14. It should be noted that the ABB technique is not considered at design time, but it is applied in the post-silicon tuning to calibrate against inter-die variation to reduce degradation of cell reliability [18]. Its application alongside with other techniques is limited mainly by the constraint of hardware resources and target fabrication process. In applying circuit-level techniques, combinations of multiple techniques should be considered if one technique alone is inadequate to reduce $\mathbf{Pr}(\mathrm{WF},\mathrm{RF})$ to meet the target. If $\mathbf{Pr}(\mathrm{RF})$

dominates **Pr**(WF,RF), priority should be given to WLM. Otherwise, if **Pr**(WF) dominates, NLB should be included primarily. If **Pr**(RF) and **Pr**(WF) are comparable, then DVS should be chosen. The aggressiveness of the three techniques are increased by tuning their control knobs differently. For WLM, it means decreasing V_{WL} in read operation. For DVS, it means increasing V_{DD} in read operation and decreasing V_{DD} in write operation. For NBL, it means decreasing the bit-line voltage.

The variables and constraints involved can be evaluated as follows.

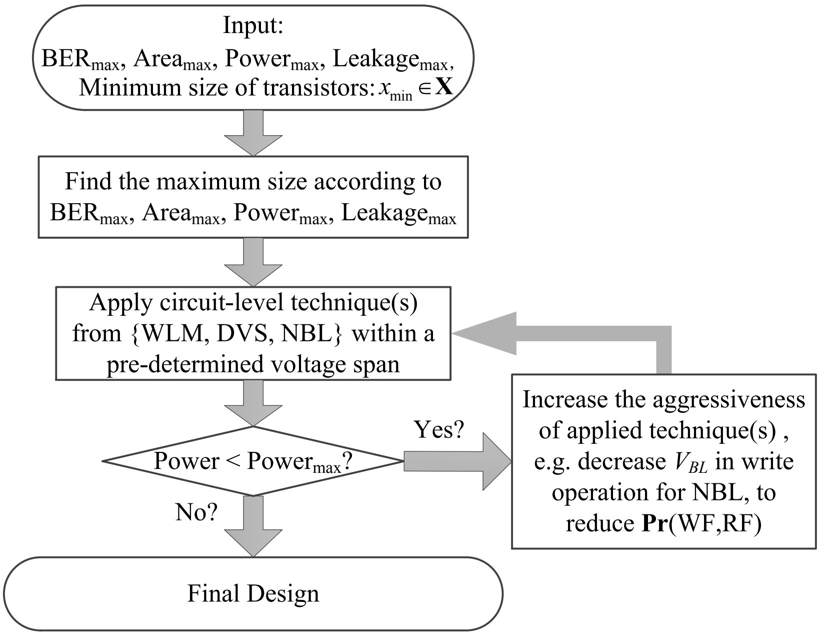

Input:
BER$_{max}$, Area$_{max}$, Power$_{max}$, Leakage$_{max}$,
Minimum size of transistors: $x_{min} \in \mathbf{X}$
Find the maximum size according to BER$_{max}$, Area$_{max}$, Power$_{max}$, Leakage$_{max}$
Apply circuit-level technique(s) from {WLM, DVS, NBL} within a pre-determined voltage span
Power < Power$_{max}$?
Yes?
No?
Increase the aggressiveness of applied technique(s) , e.g. decrease V_{BL} in write operation for NBL, to reduce **Pr**(WF,RF)
Final Design

Figure 14. Design flow for dual-mode SPUF.

3.2.1. Failure Probability

The overall failure probability of an SRAM cell can be estimated by [14]

$$\mathbf{Pr}(\mathrm{RF,WF}) = \mathbf{Pr}(\mathrm{RF}) + \mathbf{Pr}(\mathrm{WF}) - \mathbf{Pr}(\mathrm{RF} \cdot \mathrm{WF}). \tag{11}$$

The failure probabilities **Pr**(RF) and **Pr**(WF) can be efficiently estimated using the method proposed in [14].

3.2.2. Reliability

The raw reliability of the raw PUF response can be estimated by:

$$\mathrm{reliability}_{\mathrm{raw}} = 1 - \mathrm{BER}_{\mathrm{raw}} = 1 - \frac{1}{m}\sum_{i=1}^{m} \frac{\mathrm{HD}(R, R')}{n}, \tag{12}$$

where m is the number of responses generated by the same challenge, n is the bit-length of the response, and $\mathrm{HD}(R, R')$ is the Hamming distance between two responses, R and R', to the same challenge generated by the same PUF at different time or different environmental conditions. With $\mathrm{ECC}[n, k, t]$, the reliability can be improved to

$$\mathrm{reliability}_{\mathrm{final}} = \sum_{i=0}^{t} \binom{n}{t} (\mathrm{BER}_{\mathrm{raw}})^{i} (1 - \mathrm{BER}_{\mathrm{raw}})^{n-i}. \tag{13}$$

The final bit error rate is

$$\mathrm{BER} = 1 - \mathrm{reliability}_{\mathrm{final}}. \tag{14}$$

3.2.3. Area

The area of an SRAM array is estimated by

$$\text{area}_{\text{array}} \approx N \times \text{area}_{\text{cell}}, \tag{15}$$

where N is the number of SRAM cells and $\text{area}_{\text{cell}} = \text{X}_{\text{cell}} \times \text{Y}_{\text{cell}}$. Based on the layout of [14],

$$\begin{aligned} \text{X}_{\text{cell}} &= 5\lambda + 2\max(3\lambda, W_{M3}) + 2L, \\ \text{Y}_{\text{cell}} &= 9\lambda + \max(3\lambda, W_{M2}) + \max(3\lambda, W_{M1}) + L, \end{aligned} \tag{16}$$

where λ is half of the feature size. W and L are the channel width and length of the transistor, respectively. The area overhead incurred by any of the abovementioned circuit-level adaptive techniques is negligible compared to $\text{area}_{\text{array}}$.

3.2.4. Power Consumption

The power consumption of the SRAM array can be calculated by

$$\text{power}_{\text{array}} = N \times \text{power}_{\text{cell}}, \tag{17}$$

where $\text{power}_{\text{cell}}$ is the total power consumption of each SRAM cell. Its dynamic and leakage components can be obtained by the built-in function of the SPICE simulator. During the read and write operations, dynamic power dominates. Otherwise, leakage current is the major contributor of $\text{power}_{\text{array}}$ [14]. The power consumption of circuit-level techniques is insignificant when compared to the total power consumption of the SRAM core [17]. For computational efficiency without compromising the quality of optimization, a small fraction of the total power consumption can be set aside as a guard band in the power consumption constraint specifications to account for the power consumption overheads of circuit-level techniques.

4. Simulation Results and Discussions

To demonstrate the efficiency of our optimization strategy, seven different designs were compared. Designs 1 and 2 were optimized for single-mode PUF and single-mode SRAM with the parameters of **X** set to the minimum and maximum values, respectively. The transistors of Design 3 were sized for the lowest **Pr**(RF,WF) under the constraints of reliability ($1 - \text{BER}$), area and leakage without applying any circuit-level techniques. Design 4 uses WLM to improve the cell stability, with $V_{WL} = 0.8$ V in read operation and $V_{WL} = 0.9$ V in write operation. Design 5 uses DVS, with $V_{DD} = 1$ V in read operation and $V_{DD} = 0.8$ V in write operation. Design 6 uses NBL, with bit-line voltage -0.1 V applied in write operation. Design 7 is a ring oscillator (RO) PUF used for dual-mode application, i.e., as a PUF and a clock. The designs were simulated by HSPICE with CMOS 45 nm bulk Predictive Technology Model. The maximum standard deviation of V_{th} and Λ are assumed to be 15% of its nominal value and 100 nm, respectively. The channel length is set to 45 nm for all transistors. The ECC used is BCH-[127, 15, 27] [22]. The following design constraints were set: $\text{BER}_{\text{max}} = 10^{-6}$, $\text{area\_cell}_{\text{max}} = 1\ \mu m^2$, $\text{leakage\_cell}_{\text{max}} = 50\ nA$ and $\text{power\_cell}_{\text{max}} = 200\ \mu W$.

The results are tabulated in Table 2. Design 1 has a low BER, which is good for PUF, but bad for memory mode due to its high **Pr**(RF,WF). Comparing to Design 1, Design 2 has a lower **Pr**(RF,WF) and a higher BER, which makes it a better choice for memory but a poor choice for PUF. Design 3 has a lower **Pr**(RF,WF) than Design 1. Design 4 to Design 6 meet the power constraint, but Design 5 has the lowest failure rate, which is due to the fact that both **Pr**(RF) and **Pr**(WF) are high, and they can only be reduced simultaneously by DVS. However, if we look at **Pr**(RF) and **Pr**(WF) separately, WLM is the most efficient for **Pr**(RF) reduction and NBL is the most efficient for **Pr**(WF) reduction, which attests to our postulation in Section 3.2. Compared with the dual-mode SPUF, RO PUF in

Design 7 has a lower BER but a much higher functional failure rate. The optimal design (Design 5) reduces the failure rate in memory mode from the baseline design (Design 1) by 75%, while reducing the leakage by 21.4% and keeping the BER at an acceptable range of $\leq 10^{-6}$ in PUF mode at the cost of a reasonably small area and power overheads of $\sim$ 8.9% and $\sim$ 2.1%, respectively. It is worth noting that the bit error rate and code rate can both be improved by concatenating two ECCs at the expense of decoder complexity [22,23].

Table 2. Simulation results @$V_{DD} = 0.9$ V, temperature = 25 °C.

Description	Design	X	M	Pr(RF)	Pr(WF)	Pr(RF,WF)	BER	Area	Leakage	Power
Single-mode PUF	1	(180, 90, 135)	*N.A.*	0.131	0.136	0.249	1.37×10^{-7}	0.45	25.2	140
Single-mode memory	2	(500, 250, 375)	*N.A.*	0.022	0.047	0.068	3.29×10^{-2}	1.28	73.2	405
Dual-mode SPUF	3	(202, 101, 152)	*N.A.*	0.118	0.125	0.229	9.73×10^{-7}	0.49	28.5	158
	4	(202, 101, 152)	*WLM*	0.021	0.125	0.143	9.73×10^{-7}	0.49	28.5	154
	5	(202, 101, 152)	*DVS*	0.027	0.036	**0.062**	9.73×10^{-7}	0.49	19.8	143
	6	(202, 101, 152)	*NBL*	0.118	0.004	0.122	9.73×10^{-7}	0.49	28.5	155
Dual-mode RO PUF	7	–	–			0.51	1.46×10^{-10}	–	–	–

The transistor width is searched in the range of $[45, 500]$ nm, and the ratio $W_{\mathrm{M1}} : W_{\mathrm{M2}} : W_{\mathrm{M3}}$ is the same for the three designs so that no memory failure occurs with nominal parameter values; The power refers to the sum of dynamic power during read and write operations; The data of RO PUF was from [24]. The failure probability was calculated with 0.5% frequency tolerance and the BER was obtained with the same ECC as SPUF; The units for area, leakage and power are μm^2, nA and μW, respectively; Reliability can be computed from BER by $1 - \mathrm{BER}$; WLM, DVS, NBL and ABB refer to Word-Line Voltage Modulation, Dynamic Voltage Scaling, Negative Bit-Line and Adaptive Body Bias, respectively. RF and WF refer to read failure and write failure, respectively. PUF refers to physical unclonable function. SPUF refers SRAM-based PUF. RO refers to ring oscillator.

4.1. Uniqueness

Another important metric for PUF is uniqueness, which measures the capability of a PUF instance to distinguish itself from other PUF instances. The uniqueness can be measured by calculating the inter-die Harming Distance (HD). Ideally, the responses generated from two PUF instances by applying the same challenge should have half of the bits being different. Thus, the uniqueness should be 50%. Let R_u and R_v be the n-bit responses of two different chips, u and v, to the same challenge C, and the uniqueness U for m chips is expressed as [25]

$$U = \frac{2}{m(m-1)} \sum_{u=1}^{m-1} \sum_{v=u+1}^{m} \frac{\mathrm{HD}(R_u, R_v)}{n} \times 100\%. \tag{18}$$

To measure the uniqueness of the designed SPUF, two hundred 100-bit-long responses are generated and their fractional HDs are calculated. The distribution is shown in Figure 15. The uniqueness is calculated to be 0.4999, which closely approaches the ideal value of 0.5. The fitted Gaussian curve of the histogram for fractional HDs has a mean value of $\mu = 0.4999$ and a standard deviation of $\sigma = 0.0503$.

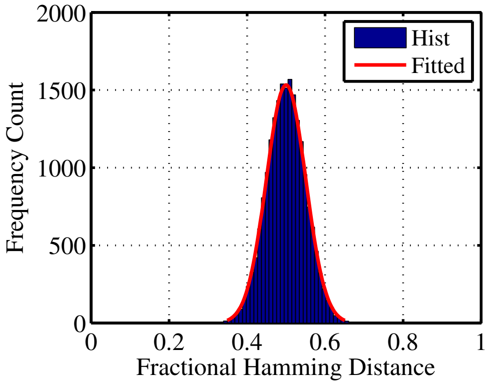


Figure 15. Frequency distribution of the simulated fractional inter-die Harming Distance (HD).

4.2. Randomness

Randomness refers to the ability of a PUF to generate response bits with an equal probability of being "1" and "0" and every bit is uncorrelated with any other bits. Ideally, the probability to obtain a "1" or a "0" bit is 50%. To measure the randomness of the augmented SPUF design, one million response bits were generated and split into 100 blocks with 10,000 bits each. The National Institute of Standards and Technology (NIST) randomness test results is shown in Table 3, which shows good randomness of the PUF responses.

Table 3. National Institute of Standards and Technology (NIST) test results.

Test Description	Passed/Total	P-value	Pass?
Frequency	98/100	0.145	√
Block Frequency ($m = 128$)	100/100	0.262	√
Cusum-Forward	98/100	0.249	√
Cusum-Reverse	99/100	0.817	√
Runs	97/100	0.102	√
Longest Run of Ones	98/100	0.868	√
Rank	100/100	0.015	√
Spectral DFT	100/100	0.024	√
Non-overlapping Templates ($m = 9, B = 000000001$)	99/100	0.367	√
Overlapping Templates ($m = 9$)	99/100	0.898	√
Approximate Entropy ($m = 4$)	98/100	0.475	√
Linear Complexity ($m = 1000$)	96/100	0.035	√
Serial ($m = 16, \bigtriangledown\Psi^2_m$)	100/100	0.038	√

Cusum refers to Cumulative Sums. DFT refers to Discrete Fourier Transform.

5. Conclusions

A novel methodology for designing SPUF for dual application modes has been proposed. Our analysis shows that the memory and PUF modes of operation are affected in a contradicting way by the device geometry of SRAM cells. Four circuit-level techniques are applied to alleviate the design constraints and expand the design space. With the proposed systematic approach to optimize the SRAM cell design for both memory and PUF requirements, the memory failure probability is reduced by 75% while the BER of PUF response is maintained below 10^{-6}. Meanwhile, the area, leakage current and dynamic power are kept to an acceptable level.

Acknowledgments: This research was supported by the Singapore Ministry of Education MOE AcRF Tier I grant no. MOE 2014-T1-002-141(RG186/14) and Research Scholarship by MediaTek Singapore Pte Ltd.

Author Contributions: C. H. Chang conceived the initial idea, directed this research, and improved the technical presentation; C. Q. Liu and L. Zhang contributed equally to the problem formulation and the optimization of its solutions, performed the experiments and analysed the data; Z. H. Kong contributed as a co-advisor to C. Q. Liu and L. Zhang, and improved the English and proofread the paper.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Suh, G.; Devadas, S. Physical Unclonable Functions for Device Authentication and Secret Key Generation. In Proceedings of the 44th ACM/IEEE Design Automation Conference, San Diego, CA, USA, 4–8 June 2007; pp. 9–14.
2. Holcomb, D.; Burleson, W.; Fu, K. Power-Up SRAM State as an Identifying Fingerprint and Source of True Random Numbers. *IEEE Trans. Comput.* **2009**, *58*, 1198–1210.
3. Zhang, L.; Kong, Z.H.; Chang, C.H. PCKGen: A Phase Change Memory based cryptographic key generator. In Proceedings of the 2013 IEEE International Symposium on Circuits and Systems (ISCAS2013), Beijing, China, 19–23 May 2013; pp. 1444–1447.

4. Van Aubel, P.; Bernstein, D.J.; Niederhagen, R. Investigating SRAM PUFs in large CPUs and GPUs. In *Security, Privacy, and Applied Cryptography Engineering*; Springer International Publishing: Cham, Switzerland, 2015; pp. 228–247.
5. Chellappa, S.; Clark, L.T. SRAM-Based Unique Chip Identifier Techniques. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2016**, *24*, 1213–1222.
6. Zimmer, B.; Toh, S.O.; Vo, H.; Lee, Y.; Thomas, O.; Asanovic, K.; Nikolic, B. SRAM assist techniques for operation in a wide voltage range in 28-nm CMOS. *IEEE Trans. Circuits Syst. II: Express Briefs* **2012**, *59*, 853–857.
7. Zhang, L.; Chang, C.H.; Kong, Z.H.; Liu, C.Q. Statistical analysis and design of 6T SRAM cell for physical unclonable function with dual application modes. In Proceedings of the 2015 IEEE International Symposium on Circuits and Systems (ISCAS2015), Lisbon, Portugal, 24–27 May 2015; pp. 1410–1413.
8. Vatajelu, E.I.; Natale, G.D.; Prinetto, P. Towards a highly reliable SRAM-based PUFs. In Proceedings of the Design, Automation and Test Europe Conference (DATE), Dresden, Germany, 14–18 March 2016; pp. 273–276.
9. Calhoun, B.H.; Chandrakasan, A.P. Static noise margin variation for sub-threshold SRAM in 65-nm CMOS. *IEEE J. Solid-State Circuits* **2006**, *41*, 1673–1679.
10. Orshansky, M.; Nassif, S.; Boning, D. *Design for Manufacturability and Statistical Design: A Constructive Approach*; Springer Science & Business Media: New York, NY, USA, 2008.
11. Synopsys, Inc. *HSPICE® User Guide: RF Analysis*; Mountain View, CA, USA, 2010.
12. Chellappa, S.; Dey, A.; Clark, L. Improved circuits for microchip identification using SRAM mismatch. In Proceedings of the 2011 IEEE Custom Integrated Circuits Conference (CICC), San Jose, CA, USA, 19–21 September 2011; pp. 1–4.
13. Agarwal, K.; Nassif, S. Statistical analysis of SRAM cell stability. In Proceedings of the 43rd ACM/IEEE Design Automation Conference, San Francisco, CA, USA, 24–28 July 2006; pp. 57–62.
14. Mukhopadhyay, S.; Mahmoodi, H.; Roy, K. Modeling of failure probability and statistical design of SRAM array for yield enhancement in nanoscaled CMOS. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2005**, *24*, 1859–1880.
15. Alorda, B.; Carmona, C.; Bota, S. Word-line power supply selector for stability improvement of embedded SRAMs in high reliability applications. In Proceedings of the 2014 Design, Automation Test in Europe Conference Exhibition (DATE), Dresden, Germany, 24–28 March 2014; pp. 1–6.
16. Zhang, K.; Bhattacharya, U.; Chen, Z.; Hamzaoglu, F.; Murray, D.; Vallepalli, N.; Wang, Y.; Zheng, B.; Bohr, M. A 3-GHz 70-Mb SRAM in 65-nm CMOS technology with integrated column-based dynamic power supply. *IEEE J. Solid-State Circuits* **2006**, *41*, 146–151.
17. Sinangil, M.E.; Poulton, J.W.; Fojtik, M.R.; Greer, T.H., III; Tell, S.G.; Gotterba, A.J.; Wang, J.; Golbus, J.; Zimmer, B.; Dally, W.J.; et al. A 28 nm 2 Mbit 6T SRAM With Highly Configurable Low-Voltage Write-Ability Assist Implementation and Capacitor-Based Sense-Amplifier Input Offset Compensation. *IEEE J. Solid-State Circuits* **2016**, *51*, 557–567.
18. Mukhopadhyay, S.; Mahmoodi, H.; Roy, K. Reduction of Parametric Failures in Sub-100-nm SRAM Array Using Body Bias. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2008**, *27*, 174–183.
19. Hokazono, A.; Balasubramanian, S.; Ishimaru, K.; Ishiuchi, H.; Liu, T.J.K.; Hu, C. MOSFET design for forward body biasing scheme. *IEEE Electron. Dev. Lett.* **2006**, *27*, 387–389.
20. Mostafa, H.; Anis, M.; Elmasry, M. Adaptive Body Bias for Reducing the Impacts of NBTI and Process Variations on 6T SRAM Cells. *IEEE Trans. Circuits Syst. I Reg. Pap.* **2011**, *58*, 2859–2871.
21. Wang, J.; Nalam, S.; Calhoun, B.H. Analyzing static and dynamic write margin for nanometer SRAMs. In Proceedings of the 2008 ACM/IEEE International Symposium on Low Power Electronics and Design (ISLPED), Bangalore, India, 11–13 August 2008; pp. 129–134.
22. Lin, S.; Costello, D. *Error Control Coding*; Prentice-Hall: Englewood Cliffs, NJ, USA, 2004.
23. Bösch, C.; Guajardo, J.; Sadeghi, A.R.; Shokrollahi, J.; Tuyls, P. Efficient helper data key extractor on FPGAs. In Proceedings of the Workshop on Cryptographic Hardware and Embedded Systems, Washington, DC, USA, 10–13 August 2008; pp. 181–197.

24. Maiti, A.; Casarona, J.; McHale, L.; Schaumont, P. A large scale characterization of RO-PUF. In Proceedings of the 2010 IEEE International Symposium on Hardware-Oriented Security and Trust (HOST), Anaheim, CA, USA, 13–14 June 2010; pp. 94–99.
25. Cao, Y.; Zhang, L.; Chang, C.H.; Chen, S. A Low-Power Hybrid RO PUF With Improved Thermal Stability for Lightweight Applications. *IEEE Trans. Comput.-Aided Des. Integr. Circuits Syst.* **2015**, *34*, 1143–1147.