# Very-Large-Scale-Integration Circuit Techniques in Internet-of-Things Applications

Jiangyi Li

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

# ABSTRACT

## Very-Large-Scale-Integration Circuit Techniques in Internet-of-Things Applications

Jiangyi Li

Heading towards the era of Internet-of-things (IoT) means both opportunity and challenge for the circuit-design community. In a system where billions of devices are equipped with the ability to sense, compute, communicate with each other and perform tasks in a coordinated manner, security and power management are among the most critical challenges.

Physically unclonable function (PUF) emerges as an important security primitive in hardware-security applications; it provides an object-specific physical identifier hidden within the intrinsic device variations, which is hard to expose and reproduce by adversaries. Yet, designing a compact PUF robust to noise, temperature and voltage remains a challenge.

This thesis presents a novel PUF design approach based on a pair of ultra-compact analog circuits whose output is proportional to absolute temperature. The proposed approach is demonstrated through two works: (1) an ultra-compact and robust PUF based on voltage-compensated proportional-to-absolute-temperature voltage generators that occupies 8.3× less area than the previous work with the similar robustness and twice the robustness of the previously most compact PUF design and (2) a technique to transform a 6T-SRAM array into a robust analog PUF with minimal overhead. In this work, similar circuit topology is used to transform a preexisting on-chip SRAM into a PUF, which further reduces the area in (1) with no robustness penalty.

In this thesis, we also explore techniques for power management circuit design.

Energy harvesting is an essential functionality in an IoT sensor node, where battery replacement is cost-prohibitive or impractical. Yet, existing energy-harvesting power management units (EH PMU) suffer from efficiency loss in the two-step voltage conversion: harvester-to-battery and battery-to-load. We propose an EH PMU architecture with hybrid energy storage, where a capacitor is introduced in addition to the battery to serve as an intermediate energy buffer to minimize the battery involvement in the system energy flow. Test-case measurements show as much as a 2.2× improvement in the end-to-end energy efficiency.

In contrast, with the drastically reduced power consumption of IoT nodes that operates in the sub-threshold regime, adaptive dynamic voltage scaling (DVS) for supply-voltage margin removal, fully on-chip integration and high power conversion efficiency (PCE) are required in PMU designs. We present a PMU–load co-design based on a fully integrated switched-capacitor DC-DC converter (SC-DC) and hybrid error/replica-based regulation for a fully digital PMU control. The PMU is integrated with a neural spike processor (NSP) that achieves a record-low power consumption of 0.61 µW for 96 channels. A tunable replica circuit is added to assist the error regulation and prevent loss of regulation. With automatic energy-robustness co-optimization, the PMU can set the SC-DC's optimal conversion ratio and switching frequency. The PMU achieves a PCE of 77.7% (72.2%) at $V_{IN}$ = 0.6 V (1 V) and at the NSP's margin-free operating point.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgment

It has been an exciting journey since I started my graduate studies 5 years ago at our group. Throughout this 5-year journey, there were too many unforgettable moments, with laughs and tears, joy and frustration. This long journey definitely shapes me from the inside out, not only because of the knowledge I learned from studies and from works, but also because of all the memories that accumulated through these moments with my colleagues, my friends, and my family.

Before all others, I am deeply grateful to Professor Mingoo Seok. This journey would not even have been possible without him; his advice guided me through many challenges and problems. I have learned from him, not just skills and knowledge, which of course will play a big part in my future, but more importantly his professionalism and passion for work and life, which I believe will help me in all my coming journeys. Also, I really appreciate all the suggestions and experiences that he shared, not only in terms of work and research but also in terms of life and future career.

I would also like to express my gratitude to my thesis committee, Professor Yannis Tsividis, Professor Peter Kinget, Professor Luca Carloni and Professor Martha Kim. Thank you all for your commitment and valuable feedback toward completing the thesis.

Then I come to my dear colleagues, some of whom have recently left our group. It is my great honor and pleasure to be able to sit in the same room and have all those exciting discussions on work and life. Some of my work would not be done without the help from you all: Seongjong Kim, Teng Yang, Doyun Kim, Joao Cerqueira, Zhewei Jiangyi, Minhao

Yang, Tianchan Guan, Pavan Chundi, Sung Kim and Dongkwun Kim. Thank you all for being a memorable part of my journey.

Outside of our group, I would like to first thank Professor Ioannis Kymissis, Professor Peter Kinget, and Professor Jae-sun Seo, for co-authoring of some of my papers. Your feedback was of great help to the work. Also, I would like to thank Dr. Mohammed Abu-Rahma at Apple, who advised me during my summer internship, as well as my colleagues at Apple. I had a great experience there, and I look forward to being there again soon.

Then, I would like to thank my old and new friends, both in the United States and in China, although I am sure most of you will not be reading this thesis. I have known some of you for more than 15 years and it is really great that we can chat and play along occasionally. Those experiences have been some of the best moments of my journey.

Finally, and most importantly, I would like to thank my family: my mom and dad for not only their support and love but also for all the life lessons I learned from you, my grandparents, uncles and aunts, sisters and brothers. I especially wish my grandfather was able to witness this moment.

To my family and friends

# Chapter 1 Introduction

## 1.1 Internet-of-things: Opportunities and Challenges

Initially enabled by the recent decade's advances in radio-frequency identification (RFID), sensor networks, communication techniques, and Internet protocols, Internet-of-things (IoT) has become a main information technology focus [1, 3]. The number of connected IoT nodes disruptively grew by 300% in the five years from 2009 to 2014 [64] and is projected to reach 200 billion by 2020 [65].

The deployment of billions of devices, each with the ability to sense, compute, communicate and perform specific tasks will enable and improve various daily-life applications. One of the examples is the smart home, where home appliances such as heating, ventilation, air conditioning, lighting systems, home security systems and more are integrated with monitoring, controlling and communicating capabilities to not only improve the quality of life for the homeowners but also yield economic benefits that come from more efficient use of energy.

However, with opportunity comes challenge; it is not easy to realize the IoT vision, and multiple key challenges must be tackled, including but not limited to availability, reliability, mobility, scalability, security, and privacy.

IoT's availability challenge exists for both software and hardware. For software, availability refers to the capability of the IoT application to provide access and service to the users anywhere, anytime and simultaneously. For hardware, this means the device must

be present when it is requested by the users and must have the capability (performance, energy, etc.) to handle the tasks.

IoT's reliability refers to the reliability of the service and devices. For instance, when the devices are deployed in-field, they must operate robustly under varying environmental conditions and aging effects of device elements. Also, the software must be verified thoroughly before deployment and updated quickly in response to known issues. IoT's reliability is especially critical when the application involves emergency response (such as leak detectors) or the tasks involve high potential financial risks (such as smart manufacturing).

Mobility refers to IoT's ability to continuously provide service while the users are moving. This is especially important when it comes to smart-car applications. IoT's scalability involves the ability to add and remove users, devices or services in the system. Specifically, when an IoT application is designed, it should be able to efficiently accommodate increasing numbers of users, devices, and services. This is a key requirement for smart buildings and cities; the cost of reorganizing the infrastructure can be prohibitively high.

Finally, security and privacy are the most critical challenges for IoT applications. An enormous amount of information will be transferred in an IoT system, which may include highly sensitive data that involves user privacy or public safety. Securing such information is critical. Moreover, given that many IoT nodes will have very stringent power budget as they will likely be powered through energy harvesting (especially for sensor nodes), the energy cost of conventional security techniques could easily dominate the system energy budget and thus drastically degrade the application's availability and reliability [2].

## 1.2 Emerging VLSI Techniques for IoT

To tackle these challenges, several emerging techniques are proposed in various layers of the IoT system architecture, from the object layer up to the application layer [1]. These techniques include but are not limited to ultra-low-power circuit design, lightweight authentication protocols, ultra-low-power hardware security circuits, energy-harvesting, ultra-low-power radio-frequency communication circuits and protocols, emerging networking protocols and more.

For the circuit community, these emerging techniques cover various fields of research and development. In this thesis, however, we focus mainly on two fields: physically unclonable function (PUF) for hardware security and fully-integrated power management circuits.

### 1.2.1 Physically Unclonable Function for Hardware Security

The most fundamental requirements for security and privacy in an IoT system are identification and authentication [2]. Identification involves creating and storing an object-specific identifier, which can be later used to identify the objects. Examples include chip anti-counterfeiting, RFID for supply chain management, and more. Authentication is more complicated as it is designed to not only identify an object but also to prevent the process from being compromised by impersonation attacks from potential adversaries.

The creation and storage of an object-specific identifier, a *physical root of trust* [4], is at the center of secure identification and authentication. Conventionally, these keys are created during the deployment phase of building an IoT system and stored in on- or off-chip nonvolatile memory (NVM). However, these keys can be exposed by various invasive

and noninvasive attacks, and it is easy to duplicate a malicious instance once the key of the original is exposed. Plus, the cost of the extra process to integrate on-chip NVM can be prohibitive in some cost-sensitive applications, especially those requiring many devices at relatively low cost.

As a result, the concept of PUF has been proposed, where the secret key is created and stored using the object's inherent hidden physical properties. As one of the early exploratory examples, [66] proposes to use a piece of three-dimensional inhomogeneous material to serve as an "optical PUF." A laser beam is blocked by the structure and the photon–material interferences will cast an instance-specific two-dimensional image.

Several "electrical PUFs" (for simplicity, all PUFs in this thesis are electrical) have been implemented by taking advantage of intrinsic device variations [14–19]. Various circuits are proposed to create instance-specific keys from random device variations. Some designs use random reset states of cross-coupled inverters [14–17], while other works use such analog circuits as current mirrors [18] and amplifiers [19].

Notably, in some cases, PUFs can also be used in an authentication protocol [6, 9–12]. These PUF designs, instead of generating a single, fixed key, generate instance-specific outputs based on certain input patterns. Each input-output pair forms a *challenge-response* pair, and by checking if the correct response is generated by an instance to a certain challenge, the instance is identified by the IoT system. Moreover, to avoid reproduction or impersonation, for this type of PUF, the challenge–response space is usually very large (more than $10^{10}$), to avoid replay attack where the adversary exhaustively reads all the challenge-response pairs.

One of the key challenges in PUF design is robustness. As the PUF output is generated from intrinsic device variations, it is susceptible to various environmental conditions, device aging, and temporal noise. A PUF's robustness refers to the stability of its output on repeated reads (where noise can change the output) and when temperature and supply voltage are changed. We detail robustness and other important design metrics of the prior art in Section 2.2.

### 1.2.2 Fully-Integrated Power Management Circuits

Another field of emerging techniques is the fully-integrated power management circuits in IoT devices. As is discussed in Section 1.1, the availability and reliability of devices are critical challenges in an IoT system. Both these challenges closely correlate to powering the IoT devices practically, efficiently and cost-effectively. For instance, for a smart building application, it can be cost-prohibitive and sometimes impractical to replace the batteries of the IoT devices. In such cases, energy harvesting is attractive [34, 36–43, 48] as it can enable energy-autonomous operations. On the other hand, fully-integrated power management circuit, which is based on switched-capacitor DC-DC converters (SC-DC), gains increasing interest in both academia and industry as it effectively reduces the system form factor and Bill-of-Materials (BoM) cost [34, 35, 37, 39, 40, 42, 46, 48, 49].

Photovoltaic (PV) energy harvesting is a highly attractive harvesting modality due to its high efficiency and low cost. Consequently, designing photovoltaic energy-harvesting power management units (EH PMU) has been an active research area [36–43].

Conventional EH PMU architecture involves two-step conversions in series: from a PV cell to a battery and from the battery to a load. However, the two-step conversion introduces

substantial energy loss, especially with large voltage differences between PV cell, battery, and load. Particularly, for SC-DC, these high-ratio conversions would result in substantial conversion losses.

An alternative architecture has been proposed [36–38, 40], where the two-step conversion is partially avoided by converting some of the energy directly from the PV cell voltage to the load voltage, while only the extra energy harvested by PV cell (or consumed by load) goes through the two-step conversion, improving the end-to-end energy efficiency. Yet, such improvement will diminish when the difference between the harvesting power and load power is large so that most of the energy will still go through the two-step conversion. Notably, for a load such as an IoT node, where the load power varies drastically, the efficiency improvement of such an architecture can be minimal.

On the other hand, near- and sub-threshold operation has gained popularity as it significantly improves a system's energy efficiency. However, the nature of near- and sub-threshold operation imposes further challenges in power management circuit design.

Due to process, voltage and temperature (PVT) variations, as well as other fast-varying variations (supply-voltage droop, coupling noise, etc.), robust near- or sub-threshold operation requires a prohibitive voltage margin. Thus, adaptive circuit techniques such as in-situ timing-error detection and correction (EDAC) [62], timing-error prediction [60, 61] and dynamic voltage scaling (DVS) have been used in the PMU–load co-design such that the supply voltage of the IoT node can be tuned dynamically according to the current PVT variations [59, 60, 63]. These works, however, either employ an off-chip regulator [59], which results in a slow transient response, increased system form factor, and extra power consumption for off-chip circuits, or utilize an on-chip low-dropout regulator (LDO) [63]

with substantial power conversion loss when the difference between LDO input and output becomes large. Moreover, both use voltage reference and comparators, which further increases quiescent power and lower overall power conversion efficiency (PCE).

It is proposed to use direct error regulation integrated with the SC-DC to enable a fully integrated and fully digital PMU design [58], where the timing error directly triggers the SC-DC for fast droop response and error statistics are used to scale the supply voltage ($V_{DD}$) up or down via the SC configuration. However, in this design, the loss of regulation due to inactivity or noncritical execution can cause critical failures. Also, the SC is designed to switch at a fixed frequency, thus its efficiency is degraded when the load power varies substantially.

## 1.3 Contribution of this Thesis

In this thesis, we present the work we have done in the field of PUF and fully-integrated power management circuit design.

In Chapter 2, we present a PUF design approach based on a pair of ultra-compact analog circuits whose output is proportional to absolute temperature (PTAT) [20]. The difference between the output voltages of a pair of PTAT voltage generators is sensitive to transistor threshold voltage ($V_{th}$) mismatch but robust against temperature and supply voltage ($V_{DD}$) variations. Moreover, comparing to the PUF design relying on bi-stability, the proposed analog PUF is more robust to temporal noises due to its non-switching operation. The measurement results from prototype chips show that the proposed PUF design achieves an 8.66× smaller per-bit area than the previous work with similar robustness, and the proposed

design is twice as robust as the previously most compact PUF design [15] in a robustness figure of merit defined Section 2.4.2.

In Chapter 3, we propose a technique to reuse existing System-on-Chip (SoC) resources (SRAM and ADC) to create a PUF. The work centers on configuring the access transistors of a 6T-SRAM bitcell to form a pair of threshold voltage ($V_{th}$) sensors with minimal area overhead. Compared to conventional digital PUFs using SRAM reset states, this work exhibits better robustness. Compared to the prior art using dedicated circuits [13–19, 31], the proposed design costs substantially less area: The per-bit area cost of the proposed design is 12× less than the most compact design with comparable robustness [7]. Such compactness allows 12× PUF bits to be integrated into the chip with minimal overhead.

In Chapter 4, we propose a novel EH PMU architecture with a hybrid capacitor–battery storage for end-to-end energy efficiency improvement under varying load and harvested power. Introducing a capacitor in the system as an intermediate energy buffer allows one to minimize the amount of energy that goes through two-step conversion and improves the end-to-end efficiency. Based on measurements in test cases with the chip prototype, the proposed architecture is able to achieve up to 2.2× better end-to-end efficiency. We analyze cases with a reasonably-simplified model and explore the trade-off between system design parameters, load/harvester conditions and expected efficiency improvements.

In Chapter 5, we present a PMU–load co-design based on a fully integrated SC-DC with a state-of-the-art Neural Spike Processor (NSP) performing motor intention decoding tasks with record-high power efficiency. In this work, we use direct error regulation to remove the excessive voltage margin and the need for a voltage reference and a comparator, making the control circuits fully digital. A tunable replica circuit is added to assist error

regulation, preventing the loss of regulation in the case of inactivity and noncritical execution, as well as supply voltage overshoot. Moreover, we propose a control scheme to automatically optimize the efficiency of the PMU while guaranteeing robust operation of the load, by searching for the DC-DC converter's optimal configuration and switching frequency.

In Chapter 6, we conclude this thesis with a summary of our work.

# Chapter 2 Ultra-compact and Robust Physically Unclonable Function Based on Voltage-Compensated Proportional-to-Absolute-Temperature Voltage Generators

## 2.1 Motivation

In the IoT era, physical objects will be integrated with electronics for sensing, computing, communication, and networking. The security of such systems is becoming one of the greatest challenges [1–3]. To tackle this challenge, both industry and academia have been making substantial efforts across layers spanning cryptographic algorithms, protocols, secure integrated circuits (IC), hardware architecture and many other areas.

As an important security primitive on the chip level, PUFs serve as *a physical root of trust* [4] to enable various higher-level security operations. PUFs produce unpredictable and unique responses to given challenge inputs, which can then be used for secure key generation and storage, or to perform security protocols such as hardware authentication [5]. Moreover, the unclonable nature of a PUF, which is achieved by hiding the challenge–response relationship in the PUF's intrinsic physical parameters, ensures that it is costly, though possible,[1] for the adversary to create an exact copy of a PUF instance.

[1] For example, Ref. [29] report the process of cloning an SRAM-based PUF through emission analysis and focused ion beam circuit edit.

Previously, several PUFs have been implemented by leveraging intrinsic device variations [6, 9–12, 14–19]. These implementations conventionally fall into two classes: strong and weak PUFs.[2] The main difference between those two types is the number of available challenge-response pairs (CRP). A strong PUF has an exponentially increasing number of CRPs in the PUF's building parameters (e.g., silicon area) while a weak PUF has linearly increasing number of CRPs in building parameters. An extreme example of a weak PUF has only one CRP, which can serve as a chip ID or a secret key. This type of PUFs is also referred to as a physically obfuscated key (POK).

We propose a novel design approach to a weak PUF based on a pair of ultra-compact analog circuits with PTAT output [20]. The difference between output voltages of a pair of PTAT voltage generators is sensitive to transistor threshold voltage ($V_{th}$) mismatch but robust to temperature and supply voltage ($V_{DD}$). We fabricate test chips in 65 nm CMOS, each containing two 256 b PUF arrays. The proposed PUF design achieves an 8.66× reduction in area/bit compared to previous work with the similar robustness against noise, temperature, and $V_{DD}$ variations [18]. The nature of static, non-switching operation, differing from bi-stability-based PUFs, substantially improves its robustness to temporal noise. Additionally, the proposed design—thanks to its inherent robustness to temporal noise, temperature and voltage, the symmetric circuit design, and the differential reading scheme—can achieve low bit instabilities of 2%, 3.5%, and 1.004% across temporal noise, temperature variation (0–80 °C), and $V_{DD}$ variation (0.6–1.2 V), respectively. In terms of a robustness figure-of-merit (RFoM) combining all these instabilities, the proposed design

[2] The name does not imply robustness to security attacks. It is solely based on the convention from [13].

achieves twice better robustness than the previously most compact PUF design [15]. In addition, we experiment with the static-masking method, which can reduce RFoM to less than $1.95 \times 10^{-4}$ with an area overhead of 25.4%. Finally, we use accelerated aging tests to confirm good robustness against device aging since the proposed PUF bitcell is based on analog circuits biased in the sub-threshold region.

## 2.2 Previous Work

In this section, we discuss several previous strong and weak PUF designs. We focus on six major requirements of the PUF design:

- Uniqueness: Each PUF instance should have unique CRPs.
- Unpredictability: It should be exceedingly hard to predict the outputs of a PUF based on a partial observation of the CRP space.
- Unclonability: It should be exceedingly hard to clone a PUF even if the circuit design is fully disclosed.
- Robustness: The output should be robust to noise, temperature, and $V_{DD}$ variations.
- Compactness: It should take a minimal amount of silicon area to control cost.
- Voltage scalability: It is desirable to have voltage scalability down to the near and sub-threshold regime for integration in an energy-constrained IC with neither separate supply distribution nor local regulation.

### 2.2.1 Strong PUFs

An arbiter PUF is one of the early delay-based strong PUFs [9]. The basic idea is to create two paths that have the same nominal delay. After manufacture, the two paths have slightly

different delays due to random process variations. Arbiter circuits can then find the faster path for producing a 1 b PUF output. However, for such a basic arbiter PUF, the uniqueness and unpredictability are found to be limited by the fact that multiple CRPs are partially based on the delays of the same paths. This creates correlations among CRPs. Inter-chip variation for arbiter PUFs can be only 23% [9]. Moreover, modeling attacks using machine-learning algorithms can successfully predict the outputs of the basic arbiter PUF [13] by reading only a small subset (hundreds or thousands of CRPs) of the whole CRP space. Also, the arbiter PUF is sensitive to $V_{DD}$ variation and exhibits the output flipping rate of 3.74% even with 2% $V_{DD}$ drop [9].

Several approaches have been proposed to improve unpredictability in arbiter PUFs [9, 10, 21, 22]. A feedforward PUF can improve uniqueness and unpredictability [9], where the configuration of the later stages in delay paths is determined by the racing results of the earlier stages. Multiple arbiter PUFs in parallel together with input and output networks can also improve unpredictability [21]. However, these approaches can degrade robustness since each arbiter PUF has limited robustness and combining outputs from multiple PUFs can worsen the overall robustness. Moreover, it is still feasible to perform the modeling attack with the help of side-channel power analysis [23].

Another type of delay-based strong PUF is based on ring oscillators (ROs). For example, a challenge input might select two ROs among $N$ ROs to compare their oscillating frequencies and generate a 1 b PUF output [10]. The number of CRPs is, therefore, $\binom{2}{N} \cong N^2/2$. Although RO PUFs have fewer CRPs than arbiter PUFs, making RO PUF's uniqueness higher, which is confirmed by the measured inter-chip variation of 46.15%, close to the ideal variation of 50%. However, uniqueness can be degraded, especially by

the layout impact, such as systematic gradient in the doping concentration, on RO frequencies. On the other hand, RO PUFs can also be vulnerable to modeling attacks [13]. The sequence of the frequencies of the ROs can be acquired by sorting algorithms and can be used to predict the output. Regarding robustness, the implementation in [10] exhibits an instability of 0.48% against the worst-case environmental variation (120 °C, 10% $V_{DD}$ drop) after employing a masking scheme that selects RO pairs with maximal frequency differences.

A bi-stable ring PUF, another design based on logic gate delay variations, was first presented in [11] for FPGA and [12] for ASIC. It exploits the bi-stability of a ring composed of an *even* number of inverters. The ring's steady state is determined by its inverters' delay variations. Assuming a ring of $N$ stages with each stage consisting of $k$ parallel paths, by selecting a path in each stage, we can create $k^N$ CRPs. The bi-stable ring PUFs are generally more robust against temporal noise than arbiter PUFs since noise is averaged out during the stabilization process. Moreover, as proposed in [11] and [12], the settling time can be used as an indicator to filter out unreliable CRPs, as the unreliable CRPs having less delay mismatch will take more time to stabilize. As reported in [12], fewer than $10^{-8}$ bit error rate can be achieved with this filtering scheme across −25–125 °C and 0.7–1.2 V. However, as multiple CRPs are based on the same delay variations, as with the arbiter and RO PUFs, the bi-stable ring PUF is vulnerable to a modeling attack.

### 2.2.2 Weak PUFs

As shown in the previous section, one of the major challenges in strong PUFs is the inherent lack of unpredictability. In contrast, weak PUFs can generate very unpredictable bits

because each bit is generated with a dedicated set of components, minimizing the correlation among CRPs. This can substantially reduce the risk of a modeling attack. However, weak PUFs do have their own drawbacks. One of the main drawbacks of weak PUFs is a limited CRP space, which makes it possible to read out all CRPs. This drawback mandates the removal of any direct access to PUF outputs. Moreover, SRAM-based PUF's risk of physical attacks, such as laser stimulation and emission analysis, is nontrivial and further countermeasures are needed to improve security [28, 29].

Several weak PUF designs have been proposed based on the bi-stability of cross-coupled inverters. The SRAM power-up state can be used as a fingerprint for an RFID chip as the value of an SRAM bitcell just after power-up is determined by the process mismatch between two cross-coupled inverters [14]. However, power-up states may not be consistent over noise and environmental variations. Some authors propose to use a latch with a reset as a PUF bitcell and implement a 128 b array in a 0.13 μm CMOS [17]. With ~3% unstable bits against noise, the impact of environmental variations is measured to be non-negligible. As much as 5.5% of the bits flip when the temperature rises from 20 °C to 80 °C. A buskeeper cell can be used for its smaller area overhead and lower design complexity [15]. However, its robustness against temperature variations is low. At 85 °C, the error ratio is as high as 20%. Finally, it is proposed in [16] to add delay variation in the reset path of a bitcell to prevent invasive power-up probing attack. An aging-hardening technique that increases the mismatch between two cross-coupled inverters, along with majority voting and bit masking, are used to improve robustness against noise and environmental variations [16].

Whereas most of the abovementioned PUF designs are based on digital circuits, a few designs are based on analog circuits. The input offset of a sense amplifier can be used to produce 1 b PUF output [19]. To enhance robustness against noise and environmental variations, the PUF design employs peripheral circuits to apply 3 V on one side of a differential pair to accelerate aging and thus increase the mismatch of the pair. Another PUF design leverages mismatches between NMOS and PMOS current mirrors [18]. The design takes a larger area than the digital designs [15], yet it achieves good robustness against noise and environmental variations, thanks to the large gain from the two current mirrors connected in series, as well as the non-switching nature of analog operation However, the impact of systematic variations and the non-tracking temperature sensitivity between NMOS and PMOS needs more investigation as these factors may skew the output and introduce temperature sensitivity.

## 2.3 The Proposed Design

This section introduces a weak PUF circuit based on ultra-compact voltage-compensated PTAT voltage generators. Figure 2.1(a) shows the overall architecture of the proposed design. It consists of a 16 × 16 bitcell array, 16 sets of top devices, each set shared by 16 bitcells in a column, an address decoder synthesized with standard cells, a 16-to-1 dual-ended multiplexer (MUX) built with NMOS pass-gates, and a comparator to digitize the MUX outputs.

The 16 × 16 array organization is determined by the leakage current flowing from unselected cells. As the top device is shared among cells in the same column to reduce area

overhead, the leakage current from the unselected cells affects $V_{OUTL}$ and $V_{OUTR}$. Based on simulations, we limit the number of cells sharing a top device and bitlines to 16.

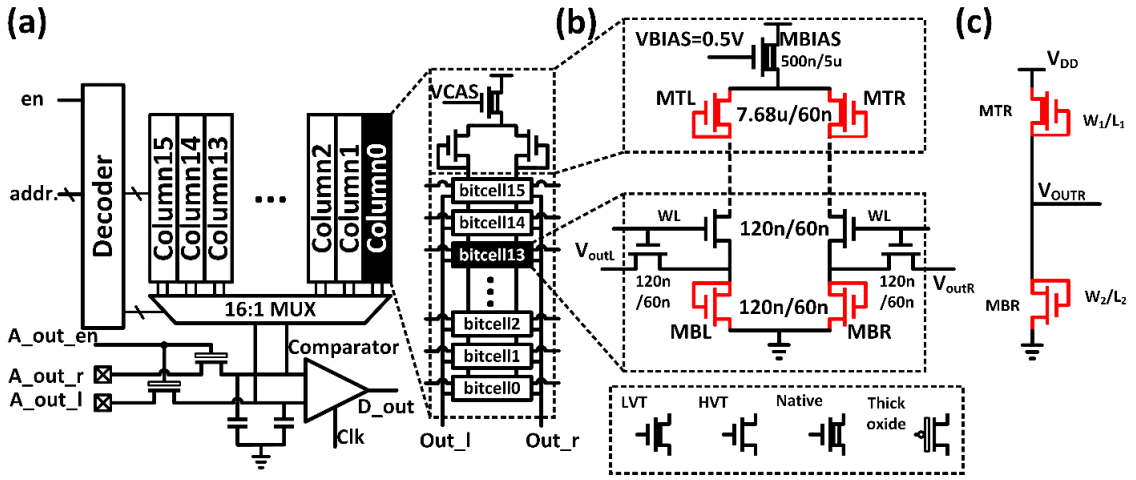



Figure 2.1 (a) The proposed 256 b PUF. It is composed of a bitcell array, an address decoder, an analog multiplexer, and a 1 b comparator. (b) A bitcell and a shared header in a column. (c) A PTAT voltage generator, two of which form a single PUF bitcell.

Our proposed design is based on a pair of low-current analog circuits working in the sub-threshold regime. Thus, two small capacitors (~30 fF) are added at the MUX outputs to reduce the impact of kick-back noise from the comparator without drastically degrading the speed of the output evaluation. The 30 fF capacitance is significantly larger than the bitline capacitance and mainly determines the evaluation time of $\Delta V_{out}$.

The comparator is a classic sense amplifier with a PMOS differential pair and an NMOS cross-coupled latch. Since the offset of the comparator can directly affect the predictability of the PUF outputs, we add a post-silicon calibration capability in the test chip. This can be upgraded with automatic and dynamic offset cancellation techniques [24] or an online calibration loop. On the other hand, these automatic calibration circuits need to be carefully protected to prevent an adversary from manipulating the offset and obtaining direct control of the PUF output.

The core of our proposed PUF design is a novel bitcell. As shown in Figure 2.1(b), it consists of six minimally sized NMOS transistors, which results in a footprint of 1.936 μm$^2$ (2.2 μm × 0.88 μm). Four of the six transistors are access transistors to select one of the 256 bitcells, while the remaining two (MBL and MBR) are configured as diodes to form a pair of PTAT voltage generators together with the shared top devices (MTL and MTR). High-threshold-voltage (HVT) devices are used for access transistors to reduce leakage current. The bottom devices are also HVT. The top devices are low-threshold-voltage (LVT) devices to meet the $V_{th}$-difference requirement of the top and bottom devices in the PTAT voltage generator design. This will be discussed further in the following derivations. A native biasing device (MBIAS) with low $V_{th}$ is added in each column to isolate the virtual $V_{DD}$ of PTAT voltage generators from the actual $V_{DD}$ and its variations.

We derive the output voltage of a PTAT voltage generator ($V_{out}$), similarly to the derivation in [18]. Figure 2.1(c) shows the simplified schematic of a single PTAT voltage generator ignoring the access transistors. The top devices are biased with $V_{gs}$ of 0 V. The bottom devices are diode-connected and biased in the sub-threshold regime as the top devices mainly determine the current. Thus, as both transistors (MTR and MBR) are biased in the sub-threshold region, we can start with the well-known sub-threshold current equation:

$$I_{sub} = \mu C_{ox} \frac{W}{L} (m-1) V_t^2 \exp\left(\frac{V_{gs}-V_{th}}{mV_t}\right)\left(1 - \exp\left(\frac{-V_{ds}}{V_t}\right)\right), \quad (2.1)$$

where $\mu$ is carrier mobility, $C_{ox}$ is sheet oxide-capacitance density, W and L are the width and length, $V_{th}$ is threshold voltage, $m$ is sub-threshold slope, $V_{gs}$/$V_{ds}$ is gate-source/drain-

source voltage and $V_t$ is the thermal voltage. Since MTR and MBR are connected in stack, the currents flowing through them are identical so that we can solve for $V_{OUTR}$:

$$V_{\mathrm{OUTR}} = \underbrace{V_{\mathrm{th2}} - \frac{m_2}{m_1}V_{\mathrm{th1}} - K_{V_{\mathrm{th}}}(T_0)}_{V_{\mathrm{th}}\text{ determined}} + \underbrace{K_{V_{\mathrm{th}}} \cdot T + m_2 \frac{kT}{q} \ln\left(\frac{\mu_1}{\mu_2} \cdot \frac{C_{\mathrm{ox1}}}{C_{\mathrm{ox2}}} \cdot \frac{W_1 L_2}{W_2 L_1} \cdot \frac{m_1 - 1}{m_2 - 1}\right)}_{\text{temperature dependent}}, \quad (2.2)$$

where the subscripts 1 and 2 represent MTR and MBR, respectively, $K_{V_{\mathrm{th}}}$ is a combined constant of the $V_{\mathrm{th}}$ temperature dependencies of MTR and MBR, $T_0$ is the reference temperature, $k$ is the Boltzmann constant, $T$ is temperature, and $q$ is an electron charge. In this derivation, we assume $V_{\mathrm{ds}}$ is sufficiently larger than $V_t$, allowing us to eliminate the second exponential term of Eq. (2.1). Noted that $V_{\mathrm{OUTR}}$ should be large enough such that $V_{\mathrm{ds}}$ of the bottom device satisfies the above assumption. We use the LVT top device to have sufficiently large difference between $V_{\mathrm{th1}}$ and $V_{\mathrm{th2}}$. Eq. (2.2) has temperature-dependent and -independent parts. The first term is a function of mainly $V_{\mathrm{th}}$ and is roughly insensitive to temperature. The second part, on the other hand, is proportional to temperature, and the slope is defined by the relative sizing of MTR and MBR and $K_{V_{\mathrm{th}}}$. Based on Eq. (2.2), we can derive the difference of the outputs of the PTAT voltage generator pair ($\Delta V_{\mathrm{out}}$):

$$\Delta V_{\mathrm{out}} = V_{\mathrm{outL}} - V_{\mathrm{outR}} = (V_{\mathrm{th2L}} - V_{\mathrm{th2R}}) - \left(\frac{m_{2\mathrm{L}}}{m_{1\mathrm{L}}}V_{\mathrm{th1L}} - \frac{m_{2\mathrm{R}}}{m_{1\mathrm{R}}}V_{\mathrm{th1R}}\right) + K_\Delta T, \quad (2.3)$$

where the subscripts L and R represent the left- and right-side outputs of a PUF bitcell, respectively, and $K_\Delta$ is the difference in the temperature slopes between the two PTAT voltage generators. $K_\Delta$ is supposedly very small as the generators are identically sized and symmetrically laid out.

Then, a PUF response bit can be produced by finding the polarity of $\Delta V_{\mathrm{out}}$ with a 1 b comparator. Note that $\Delta V_{\mathrm{out}}$ contains no $V_{\mathrm{DD}}$-related terms, implying that it is robust against

$V_{DD}$ variations. Also, as the operation of the proposed PUF is relying on non-switching analog voltage generators, it has significantly better robustness to temporal noise, as compared to those designs based on SRAM bi-stability. Furthermore, as the PTAT voltage generator is operating in the sub-threshold regime, and the proposed PUF relies on comparing those small current/voltage differences, it will be significantly harder, or practically impossible for emission analysis to obtain the PUF bits.

Equations (2.2) and (2.3) provide insights to design the proposed circuits.

First, while we design PTAT circuits in this work, by sizing the top and bottom devices (Eq. (2.2)), we could make $V_{out}$ to be complementary to absolute temperature (CTAT) or insensitive to temperature (i.e., $V_{out}$ becomes a voltage reference).

The reason that we choose PTAT over those two alternatives follows. First, if a CTAT is used, it tends to produce $V_{out}$ that is lower at higher temperatures. Together with a larger $V_t$ at high temperatures, this can reduce the $V_{ds}$ of the bottom transistor to the level where the second exponential term of Eq. (2.1) cannot be ignored. This is problematic since if that term appears in Eq. (2.3), it can create voltage dependency. Second, if the devices are sized to create a voltage reference, $V_{out}$ inevitably can have nonlinearity to the temperature, which can increase bit-flipping ratio across temperature variations. The root cause of the nonlinearity is $\mu$ and $m$ in the second term of Eq. (2.2), which are temperature dependent.

In addition, to maximize $\Delta V_{out}$'s randomness, the impact of the top devices (MTL and MTR, the second term of Eq. (2.3)), should be minimized. As MTL and MTR are shared by 16 bitcells in the same column, the second term in Eq. (2.3) can introduce correlations among the outputs of these bitcells. In the worst case, the $V_{th}$ mismatch between MTL and MTR is very large and can dominantly determine $\Delta V_{out}$, which causes the outputs of that

column to be all *1* or *0*. To avoid this, we enlarged the shared top devices and employed common-centroid layout style to minimize the second term in Eq. (2.3).

Finally, the third term in Eq. (2.3), the combined mismatches of various device parameters, can affect the robustness of PUF output if the first term is small, 1–2 mV (the second term is minimized as above), since the third term can dominantly determine the polarity of $\Delta V_{out}$. Upsizing the top devices and using the common-centroid layout style also helps to minimize the top devices' contribution to the third term. In contrast, we cannot simply reduce the mismatch between bottom devices by the same techniques since that can reduce the magnitude of the first term in Eq. (2.3) and thus hurt the robustness.

After minimizing the second and third terms of Eq. (2.3) and confirming they are negligibly small in most cases via simulation, we find that $\Delta V_{out}$ is mostly determined by the $V_{th}$ difference of the bottom devices. Since the local $V_{th}$ variations of two bottom transistors have the same normal distribution with zero mean and the same standard deviation, $\Delta V_{out}$ will also have a normal distribution with a zero mean. This ensures that a bitcell generates a *1* or *0* at close to 50%, improving inter-chip uniqueness.

## 2.4 Measurement Results

### 2.4.1 Test Chips

The test chips for the proposed PUF design are fabricated in a 65 nm general-purpose CMOS process. Figure 2.2 shows a test chip die photo. Each chip includes two PUF instances, each having a total area of 1,900 $\mu m^2$ (38 $\mu m \times 50 \mu m$). The area per bit is 3.07 $\mu m^2$ when bitcells, shared top devices, and biasing devices are considered, while the

total area per bit is 7.42 μm$^2$, including the bitcell array, decoder, comparator, and readout circuitry.

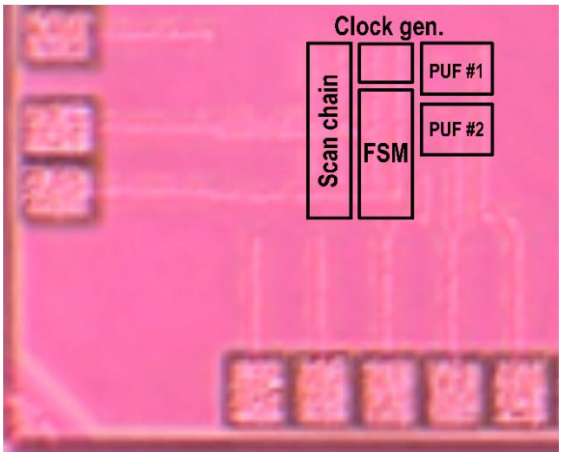


Figure 2.2 Test chip die photo

## 2.4.2 Robustness Measurement

First, we examine the $\Delta V_{out}$ distribution across all 20 PUFs from 10 chips. We use either the on-chip comparator or an off-chip ADC on a National Instruments data acquisition board to digitize or measure $\Delta V_{out}$, respectively. The effective accuracy specification of the ADC is 13 b with the input range configuration of ±1 V. As is shown in Figure 2.3, $\Delta V_{out}$ exhibits desirable characteristics that follow the normal distribution with $\mu$ = 0.09 mV and $\sigma$ = 31 mV.

We then perform the calibration to remove the comparator offsets. The measured offsets have a distribution with $\mu$ = 4.9 mV and $\sigma$ = 20 mV across 20 PUFs.

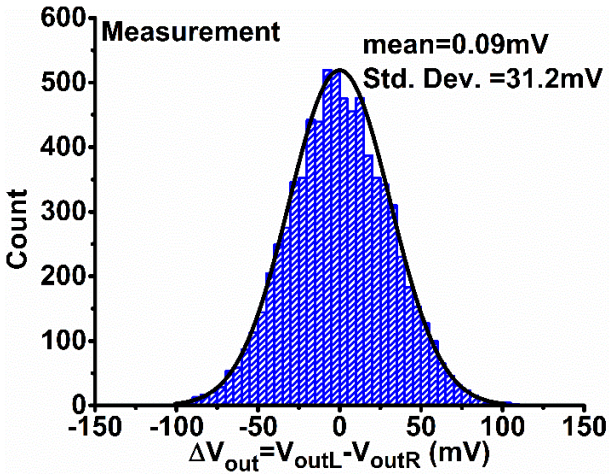



Figure 2.3 The differential output voltage ($\Delta V_{out}$) shows a normal distribution with a mean close to zero.

The digitized bits from the on-chip comparator are then read out and their distribution is investigated. Figure 2.4 shows the two-dimensional map of the averaged digital outputs at different locations of the arrays. No noticeable spatial artifact is observed, indicating that the proposed design has a negligible systematic bias.

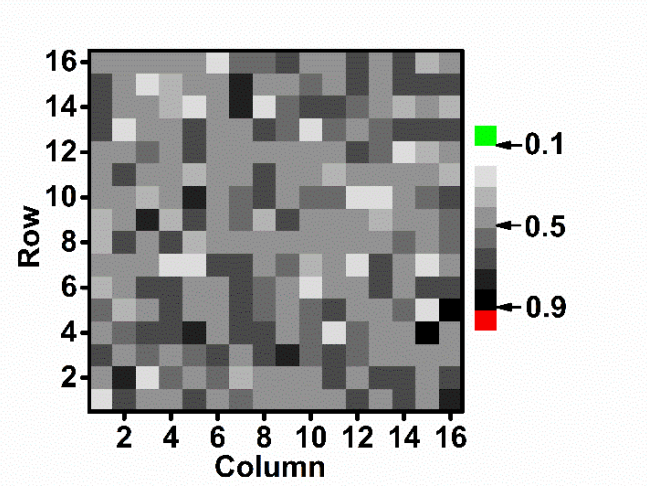



Figure 2.4 Spatial distribution of digital bits averaged across 20 PUF instances, showing no systematic patterns.

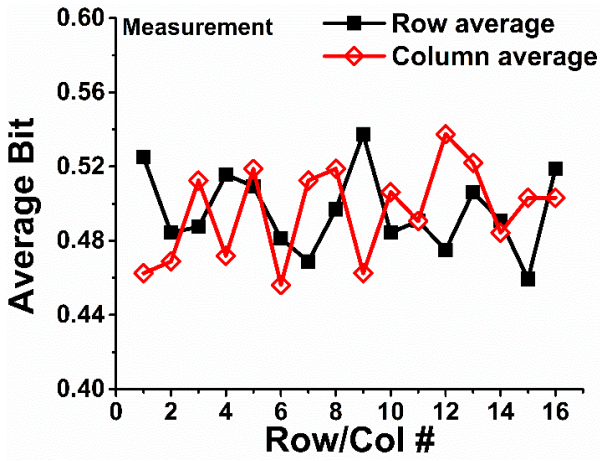



Figure 2.5 Row- and column-wise average values of 20 PUF instances showing no systematic bias.

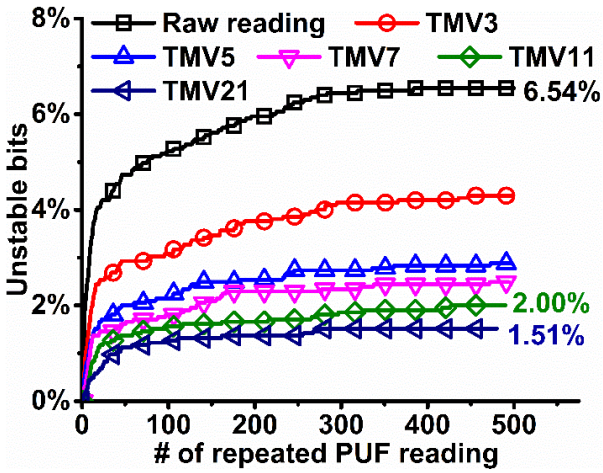



Figure 2.6 Temporal noise can cause instability in reading. Over 500 readings, 6.54% of 256 b is found to be unstable. Applying temporal majority voting with 11 and 21 readings (TMV11 and TMV21) reduces the unstable bit ratio to 2% and 1.5%, respectively.

Furthermore, Figure 2.5 shows the values of digital bits averaged across rows and columns. The average values show no systematic patterns or strong gradients, again suggesting the design can well mitigate systematic process variations. A fluctuation near 50% is observed, but this is mainly caused by the limited number of bits per column. For a binomial distribution with $p$ = 0.5, $n$ = 320 (20 PUFs, 16 b per row, 16 columns). The standard deviation can be calculated as $\sqrt{np(1-p)} = 8.944$, corresponding to 2.8% of 320 b. This

calculated variation is close to the variation found in our measurement, showing our PUF design exhibits only a small systematic bias.

We investigate the impact of temporal noise by repeatedly reading out the output of a PUF. As shown in Figure 2.3, some bitcells can have the small $\Delta V_{out}$ of just a few millivolts. Such small values can make the output bits sensitive to noise. As shown in **Error! Reference source not found.**, the measurements show that at most 6.54% of the 256 output bits can flip at least once while reading output 500 times. This probability is defined as an *unstable-bit ratio*. To improve robustness against noise, we employ a temporal majority voting (TMV) scheme [16]. We conduct the TMV off-chip, and the measurement shows that the TMVs using 11 and 21 readings (TMV11, TMV21) can reduce the unstable-bit ratio to 2% and 1.51%, respectively.

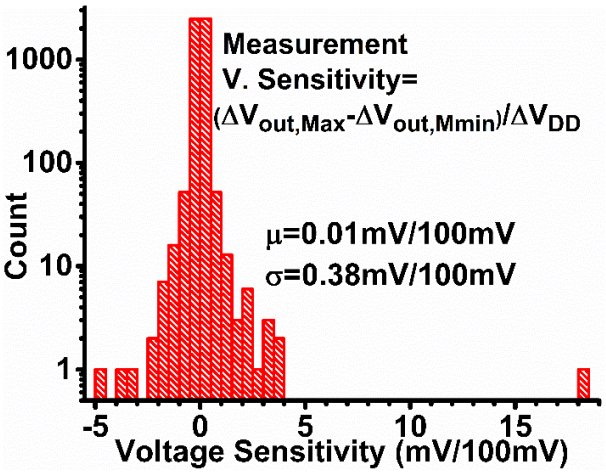



Figure 2.7 Distribution of the sensitivity of $\Delta V_{out}$ to $V_{DD}$ in 20 PUF instances. The mean and standard deviation are 0.01 mV/100 mV and 0.38 mV/100 mV, respectively.

We also investigate the robustness against $V_{DD}$ variations. We expect it is high based on Eq. (2.3). As shown in Figure 2.7, the sensitivity of $\Delta V_{out}$ to $V_{DD}$ variation exhibits $\mu$ = 0.01 mV/100 mV and $\sigma$ = 0.38 mV/100 mV across 20 256 b PUF instances. We also measure bit-flipping ratios of outputs across $V_{DD}$ = 0.6 to 1.2 V using the output generated

at 1 V as a reference. The bit-flipping is counted by comparing the PUF evaluations at various conditions against the reference. In this experiment, we use the off-chip ADC since the on-chip comparator is unstable when $V_{DD} \leq 0.9$ V due to design mistakes we made. As is shown in Figure 2.8, the bit-flipping ratio is measured to be less than 1.004% and 2.34% in the average and the worst case across 20 PUFs, indicating excellent robustness against $V_{DD}$ variation.

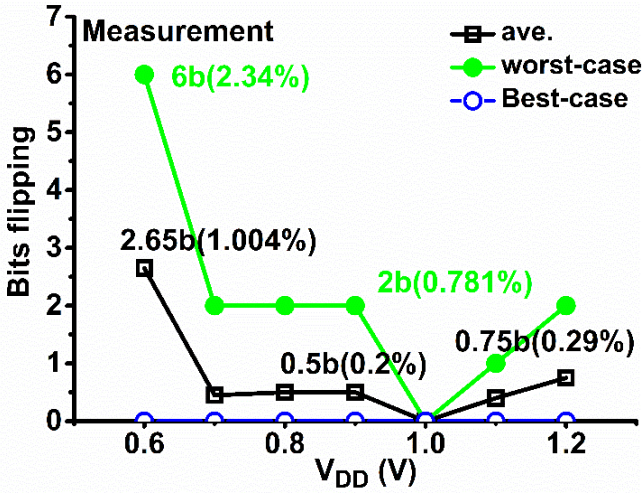



Figure 2.8 Bit-flipping ratios across $V_{DD}$ = 0.6–1.2 V. The digitization is done with the off-chip ADC.

We also measure the robustness of the output to temperature variations with the comparator recalibrated at each temperature. As is discussed in Section 2.3, automatic offset-cancellation techniques or on-chip calibration loops can be implemented to avoid manual calibration. TMV11 is used to remove the impact of noise. As shown in Figure 2.9, across 0–80 °C and 20 PUF instances, the bit-flipping ratio using the output at 20 °C as a reference is similar to its value in the $V_{DD}$ experiment, less than 3.5% and 6.64% in the average and the worst cases, respectively. Finally, we create a normal random-variable model based on two measured distributions: one for the difference of $\Delta V_{out}$ per 10 °C (defined as $\Delta(\Delta V_{out})$/10 °C) and the other for $\Delta V_{out}$ at 0 °C (Figure 2.10). The distribution of $\Delta(\Delta V_{out})$

is much narrower than that of $\Delta V_{out}$, implying good robustness against temperature variation. Using the model, we estimate the bit-flipping ratio against temperature variations. As shown in Figure 2.9, the bit-flipping ratios from the measurement and the random model match well, confirming the distribution is normal.

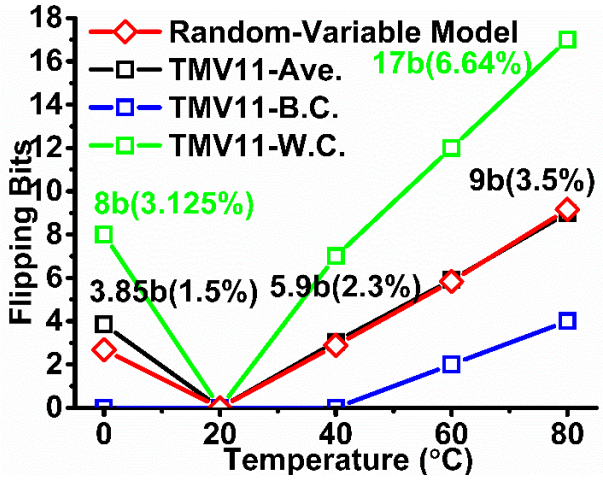



Figure 2.9 The number of flipped bits across temperature variations (0 to 80 °C). TMV11 is used to mitigate the impact of temporal noise. The measurements are well matched to the theoretical normal random-variable model based on these measurement results.

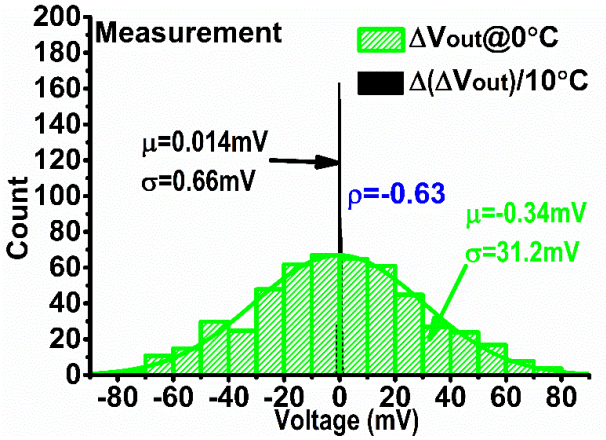



Figure 2.10 The distributions of $\Delta V_{out}$ and $\Delta(\Delta V_{out})$.

One of the promising ways to improve robustness against noise and environmental variations is to identify the unstable bits during manufacture tests and statically mask them. The masking information can be stored in nonvolatile memory. To identify potentially unstable bits, we can compare the $|\Delta V_{out}|$ of a bitcell with a certain threshold voltage ($V_{filter}$).

Those unstable bits can be discarded from the PUF output. Although we have no on-chip circuitry to perform this process, we can still analyze its feasibility based on the measured data. To evaluate robustness against noise and environmental variation at the same time, we define the RFoM as

$$RFoM = \sqrt{P_{\mathrm{nom}}^2 + P_{\mathrm{V}}^2 + P_{\mathrm{T}}^2}, \tag{2.4}$$

where $P_{\mathrm{nom}}$ is the ratio of unstable bits due to noise at a nominal environmental condition, $P_{\mathrm{V}}$ is the bit-flipping ratio against ±10% $V_{\mathrm{DD}}$ variations, and $P_{\mathrm{T}}$ is that against 0 to 80 °C temperature variations.

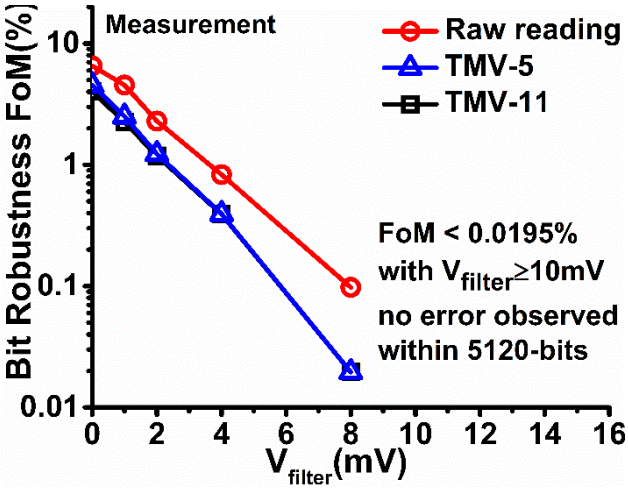



Figure 2.11 RFoM improves with larger $V_{\mathrm{filter}}$. At $V_{\mathrm{filter}}$ > 10 mV, the RFoM of a 256 b PUF instance is reduced to < 0.0195% (1/5120) even without TMV.

As shown in Figure 2.11, by setting $V_{\mathrm{filter}}$ = 10 mV, we can achieve a 0% bit-filliping ratio out of 20 256 b PUF instances even without TMV schemes. This implies that the RFoM is at most $1.95\times10^{-4}$ (i.e., 1/5120). The overhead of this filtering scheme, which is the proportion of discarded bits, is 25.4% on average and 30.4% in the worst case across 20 256 b PUF instances (Figure 2.12). Thanks to the proposed ultra-compact bitcell, the area overhead is manageable, making the filtering scheme attractive for improving robustness.

Also, the recent scaling of on-chip non-volatile memory [30] suggests that it would not introduce a substantial overhead to store the mask bits.

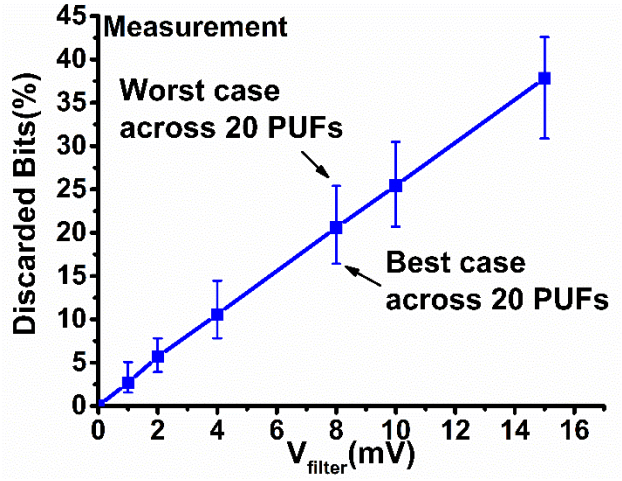



Figure 2.12 The use of a larger $V_{filter}$ requires discarding bits to produce output, causing area overhead. Thanks to the proposed bitcell's compact size, the overhead is manageable.

## 2.4.3 Unpredictability and Uniqueness Measurement

The uniqueness and unpredictability of the proposed PUF design are evaluated the through autocorrelation function (ACF) and NIST random tests [25]. A white-noise bitstream will have an ACF value close to zero at any lag value except zero. The ACFs for the 5120 b ($20 \times 256$) streams generated from 20 PUF instances shows a 95% confidence bound of $\sigma = 0.0188$ (Figure 2.13), suggesting remarkable randomness. In addition, as shown in Table 2.1, the 256 b outputs from 20 PUF instances pass all the NIST tests suitable to our data size (5120 b). We cannot conduct some of the tests—namely, Binary Matrix Rank, Linear Complexity, Overlapping Template Matching, Universal Statistical, Random Excursions, and Random Excursions Variant—due to the limited data size. The $\mu$ and $\sigma$ of the probability that a bitcell produces a *1* output are 49.3% (126.15 b) and 3% (7.6 b), respectively, across 20 PUF instances.

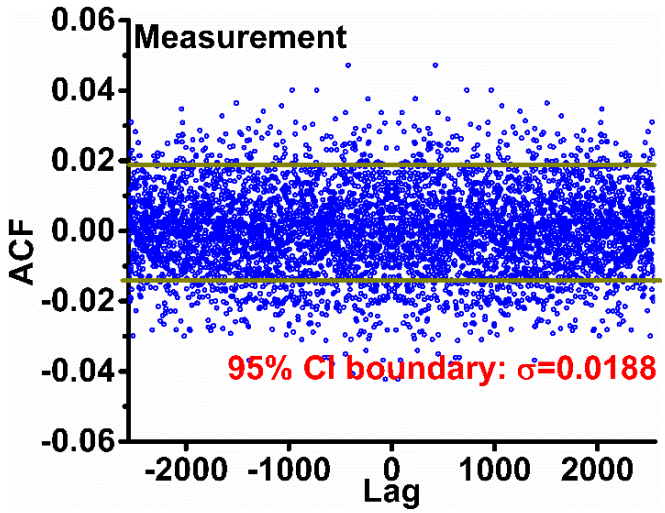



Figure 2.13 ACF of 5120 b generated from 20 PUF instances, showing an excellent randomness.

| Test Name | Stream Length | No. of runs | Pass % | Ave. P-value | Pass? |
|---|---|---|---|---|---|
| **Frequency** | 256 | 20 | 100% | 0.5276 | YES |
| **Block Frequency** | 256 | 20 | 100% | 0.4988 | YES |
| **Runs** | 256 | 20 | 100% | 0.5078 | YES |
| **Longest run of ones** | 256 | 20 | 100% | 0.5256 | YES |
| **Cumulative Sums** | 256 | 20 | 100% | 0.4889 | YES |
| **FFT** | 256 | 20 | 100% | 0.4256 | YES |
| **Non-overlapping template matching** | 256(m=4) | 20 | 100% | 0.5772 | YES |
| **Serial** | 256(m=3) | 20 | 100% | 0.4047 | YES |
| **Approximate Entropy** | 256(m=3) | 20 | 100% | 0.4846 | YES |

Table 2.1 The 256 b streams generated from 20 PUF instances pass all the NIST random tests.

We then evaluate the uniqueness of the proposed PUF outputs with Hamming distance (HD). For this, we calculate inter-PUF HD defined as the HD between the outputs (after TMV) of two different PUFs selected from 20 PUFs at reference temperature and voltage (27 °C, 1 V). The HD of two bitstreams is defined as the number of different bits. As shown in Figure 2.14, the inter-PUF HD has a mean value of 0.5, indicating high randomness in the outputs.

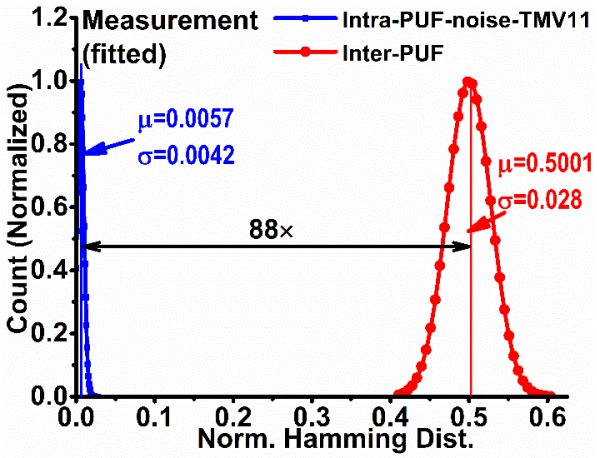



Figure 2.14 Inter- and Intra-PUF HDs. Intra-PUF HD is measured and calculated with TMV-11. The separation between the means of two HDs is measured to be 88×.

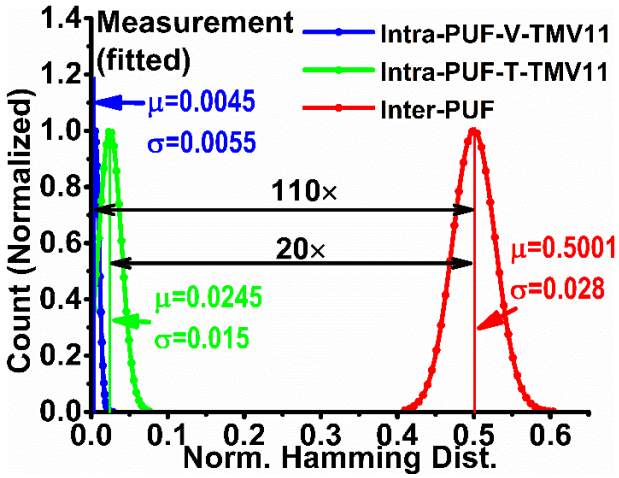



Figure 2.15 Intra-PUF HDs considering voltage (0.6–1.2 V) and temperature (0–80 °C) variations are measured. Those two HDs have 110× and 20× separations from the inter-PUF HD.

We also measure and calculate the intra-PUF HD to evaluate robustness. For the robustness against temporal noise, we first read a 256 b output from a PUF 500 times with TMV11 and calculate the HD of 256 b outputs from any two of the 500 readings. The distribution of the HDs in Figure 2.14 shows $\mu$ = 0.0057 and $\sigma$ = 0.0042, which exhibits an 88× separation from the inter-PUF HD distribution. Similarly, for the robustness against $V_{DD}$ and temperature variations, we measure and calculate the HD distributions with the same settings except changing $V_{DD}$ from 0.6 to 1.2 V and temperature from 0 to 80 °C,

respectively. The results in Figure 2.15 show the separations of 110× for $V_{DD}$ variations and 20× for temperature variations from the inter-PUF distribution. The HD measurements suggest that the proposed PUF has excellent randomness and robustness against noise and environmental variation.

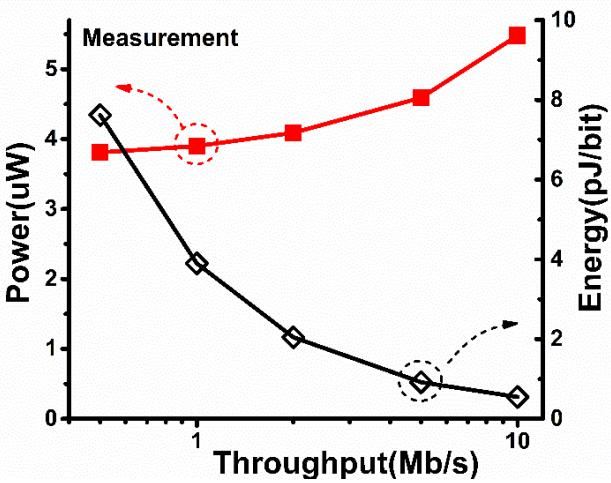



Figure 2.16 Power dissipation and energy per bit across throughputs. The maximum raw throughput is measured to be 10.2 Mb/s measured at 1 V $V_{DD}$, at which the circuits consume 0.548 pJ/bit.

## 2.4.4 Throughput and Power Consumption

We measure the proposed PUF design's maximum throughput while applying TMV11. The maximum operating frequency is defined as the frequency at which the bit-flipping ratio increases by a factor of two compared to the bit-flipping ratio at a very low frequency (500 kHz). The average maximum throughput across 20 PUFs is 10.2 Mb/s at $V_{DD}$ = 1 V. We also measure the power consumption and the energy per bit. As shown in **Error! Reference source not found.**, at 10 Mb/s, the average power dissipation and energy per bit are 5.48 µW and 0.548 pJ/bit, respectively. At 0.5 Mb/s, they are 3.81 µW and 7.62 pJ/bit. Those measurements at two different throughputs show that static power dominates total power dissipation. We can reduce the power consumption by activating

only one bitcell at a time whereas the current implementation unnecessarily activates 16 bitcells in a row.

## 2.4.5 Device Aging Measurement

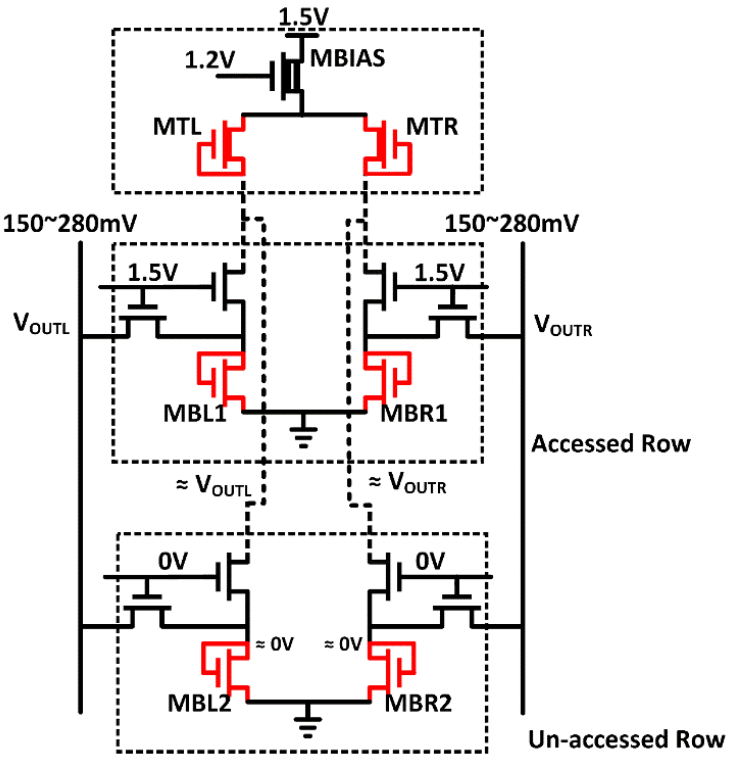



Figure 2.17 Nodal voltages during an accelerated aging test. While $V_{DD}$ and $V_{BIAS}$ are higher than nominal values, critical transistors (in red) are biased in the sub-threshold region.

Robustness against device aging emerges as an important challenge for designing PUFs [17, 19]. Aging effects such as bias-temperature instability (BTI) and hot carrier injection (HCI) can gradually change device parameters (e.g., $V_{th}$ and $\mu$) [26, 27]. Device aging effects can have a direct impact on PUF output over time; thus, it is essential to mitigate the impact of device aging. On the other hand, aging effects can be intentionally used to harden PUF circuits against noise and environmental variations. Different transistors in

PUF circuits can be selectively aged during manufacturing burn-in tests to improve robustness [16].

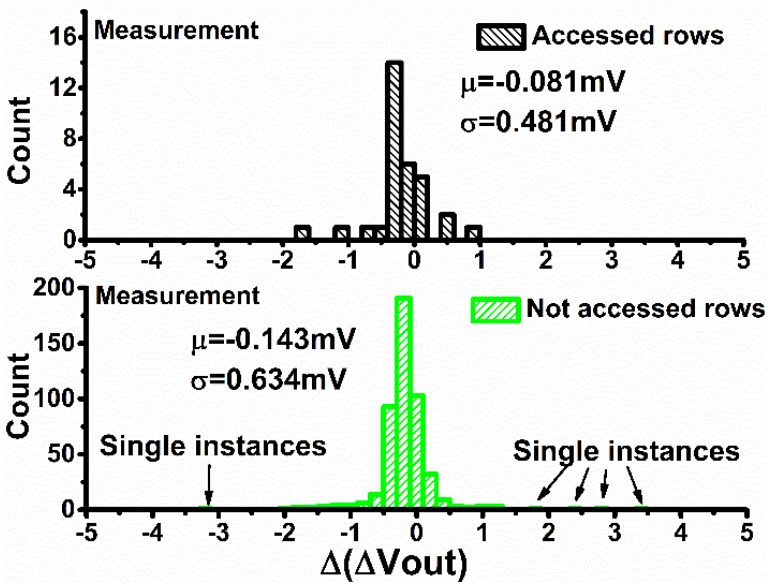



Figure 2.18 The measurement of $\Delta(\Delta V_{out})$ after the 82 h accelerated aging experiment at 1.5 V and 125 °C. The accessed row has smaller $\Delta(\Delta V_{out})$ because the aging effects of top and bottom devices cancel each other when creating $\Delta V_{out}$.

The proposed design exhibits a significant advantage regarding robustness against aging effects simply because all the critical transistors (MTL, MTR, MBL and MBR in Figure 2.1(b)) are always biased in the sub-threshold region across $V_{DD}$ = 0.6–1 V. In addition, their drain currents are less than 0.1 μA. Due to these small junction voltages and drain currents, therefore, the critical transistors experience little aging effects. Our measurements confirm the robustness of the PUF design against aging. Specifically, we conduct accelerated aging tests at a temperature of 125 °C and a $V_{DD}$ of 1.5 V for 82 h. Accelerated aging tests under the same condition for 16 h are able to induce up to 41.2 mV $V_{th}$ shift in the pull-up transistor in an SRAM bitcell [27], which can account for the aging of multi-year usage. During the aging test, we access one row of the array (i.e., the access transistors are turned on) while the remaining 15 rows are un-accessed. As shown in Figure 2.17, this

configuration stresses all the top devices (MTL and MTR) and the bottom diode-configured devices (MBL1 and MBR1) of the accessed rows but does not stress the diode-configured devices of the unaccessed rows (MBL2 and MBR2).

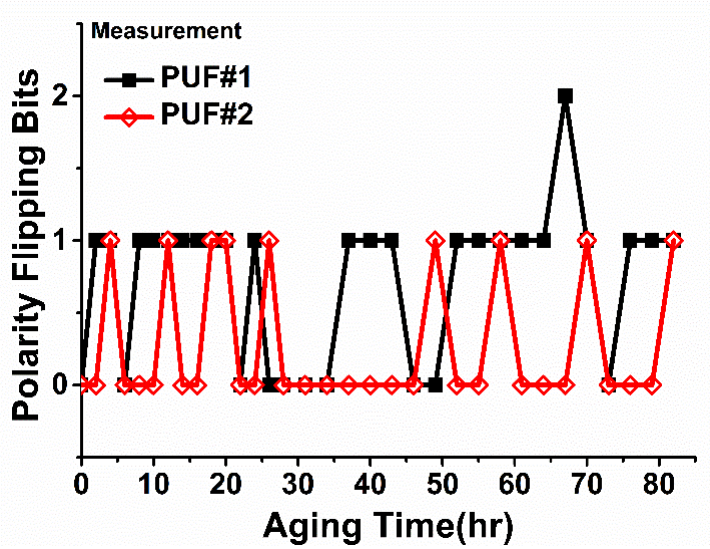



Figure 2.19 The number of flipped bits of two 256 b PUF instances during the 82 h accelerated aging test.

Figure 2.18 shows the distributions of the differences of bitcell $\Delta V_{out}$ before and after the accelerated aging test. We define the difference as $\Delta(\Delta V_{out})$ and plot the distributions separately for selected and unselected rows. The worst-case change is only 1.6 mV for selected and 3.5 mV for unselected rows in a 256 b PUF instance, indicating the aging has little impact on the output. The experiment shows that the unselected rows exhibit $\Delta(\Delta V_{out})$ which are actually larger than the selected row. This is because, for the selected row, the aging effects of the top (MTL and MTR in Figure 2.1(b)) and bottom (MBL and MBR in Figure 2.1(b)) devices cancel each other, whereas for the unselected rows the bottom devices are not aged and the aging effects of the top devices are translated to $\Delta V_{out}$. In this regard, the PUF could select all rows when not accessed to make it more robust against aging effects. We also measure the bit-flipping ratio before and after the accelerated aging

test. As shown in Figure 2.19, the bit-flipping ratio is less than 0.0078% (2/256) across two PUF instances. This low bit-flipping ratio can be further eliminated via the filtering scheme discussed in Section 2.4.2.

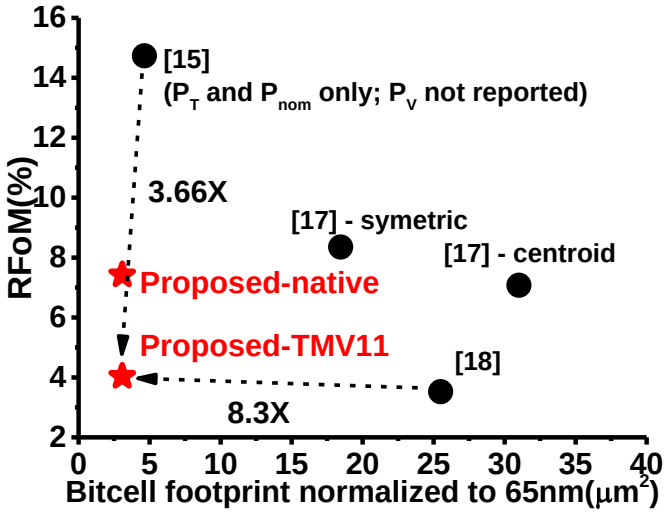



Figure 2.20 The proposed design achieves a significantly better trade-off of robustness and area efficiency compared to the state-of-the-art weak PUF designs. The measurements with and without TMV11 (native) are shown. The results of the previous designs are without TMV.

# 2.5 Comparison and Conclusion

## 2.5.1 Comparison

In this section, we compare the proposed PUF circuits with the state-of-the-art designs. Based on the RFoM we defined in Section 2.4.2, we are able to compare the trade-off between area and robustness to noise and environmental variations between the proposed and the previous PUF designs (Figure 2.20). The previous works show a clear trade-off between RFoM and area. The reason behind this is evident: To achieve higher robustness, we often need to introduce large, sophisticated circuits, which incur area overhead. The proposed design, thanks to our novel circuit design, can push the trade-off: It achieves a 2× to 3.66× better RFoM with native and TMV11 measurement, respectively, as compared

to the most compact yet still comparably sized work [15] and an 8.3× smaller footprint as compared to the most robust work [18].

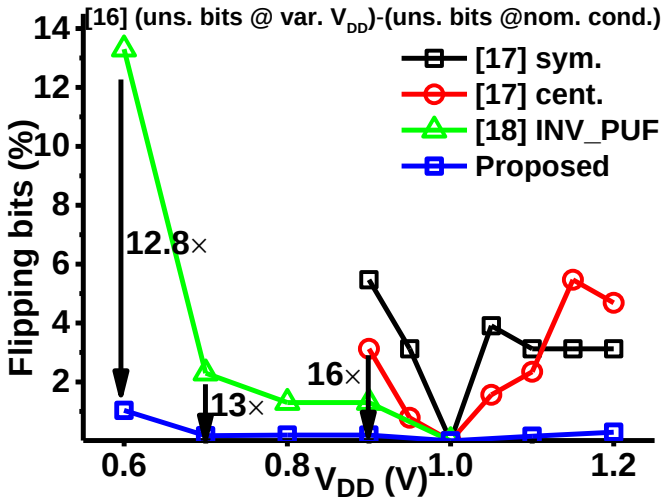



Figure 2.21 The proposed design achieves voltage scalability down to 0.6 V. The bit-flipping ratio at 0.6 V is more than 12.8× better than those of the previous designs. The off-chip ADC is used for digitizing $\Delta V_{out}$ in our proposed design.

We also compare the $V_{DD}$ scalability, which is an important feature if the PUF can be powered from the same power grids with other digital circuits, which actively seek to use deeply scaled $V_{DD}$ to reduce power dissipation. As shown in Figure 2.21, our design achieves an excellent $V_{DD}$ scalability down to 0.6 V with a 12.8× lower bit-flipping ratio seen in [18]. We consider Ref. [18] functional at $V_{DD} \geq 0.7$ V though it presents the results from 0.6 to 1 V; the error ratio increases significantly to 20% at 0.6 V (Table 2.2).

| | [16] | [17] Sym. | [17] Cent. | [18] SRAM | [15] Buskeeper | [18][3] | Proposed |
|---|---|---|---|---|---|---|---|
| **Technology** | 22nm | 0.13µm | 0.13µm | 65nm | 65nm | 65nm | **65nm** |
| **Total area/bit ($um^2$)** | N/A | 29.86 | 50.59 | N/A | N/A | N/A | **7.42** |
| **Norm. bitcell size ($um^2$)** | 40.68 | 18.46 | 31.01 | 3.42 | 4.6387 | 25.5 | **3.07** |
| **Unstable Bits at Norm. Cond.** | ~30%[4] ~3%[1] | 3.04%[4] | 3.78%[4] | 16.6%[4] | ~4%[7] | 1.73%[4] | **6.54%[4] 2.00%[2]** |
| **Temperature range(°C)** | 25-50 | 0-80 | 0-80 | 25-85 | -40~85 | 25-85 | **0-80** |
| **Bit flipping ratio per 10°C** | N/A | 0.68% | 0.635% | >6.67% | 1.14%[5] | 0.47%[5] | **0.44%** |
| **$V_{DD}$ range (V)** | 0.7-0.9 | 0.9-1.2 | 0.9-1.2 | 0.7-1 | N/A | 0.7-1 | **0.6-1.2** |
| **Bit flipping ratio per 0.1V** | N/A | 1.82% | 1.82% | >16.67% | N/A | 1.3%[5] | **0.13%[6]** |
| **Norm. Inter-PUF HD** | 0.4805 | 0.506 | 0.501 | 0.3321 | 0.491 | 0.5014 | **0.5001** |
| **Energy/bit (pJ)** | Native 0.013 TMV15 0.19 | 0.93 | 1.6 | 1.1 | N/A | 0.015 | **Native 0.548 TMV11 6.02** |

[1]: Bit-error rate after TMV, aging hardening and bit-masking are applied. (>12% before masking) [2]: TMV11 is used. [3]: Only INV_PUF is included. SA_PUF has the similar performance. [4]: Native read-out results. [5]: Conservatively estimated from the disclosed unstable bit ratio. [6]: Digitized with off-chip ADC [7]: only reported Bit-error rate

Table 2.2 Comparison of the state-of-the-art PUF circuits

## 2.5.2 Conclusion

This work presents compact and robust weak PUF circuits for security-oriented applications. The novel bitcell using a pair of voltage-compensated PTAT voltage generators consume a minimal footprint of 3.07 $\mu m^2$/bit, which is 8.66× smaller than the state-of-the-art design having comparable robustness [18]. The proposed design exhibits 2% bit instability under noise, 3.5% bit-flipping ratio across 0 to 80 °C, and 1.004% across 0.6 to 1.2 V. Its RFoM is 2× better than the previous work that has the smallest yet still comparable footprint [15]. Excellent unpredictability and uniqueness are verified with ACF, NIST randomness test and inter-PUF HD. The maximal throughput is 10.2 Mb/s on average and the proposed design consumes 0.548 pJ/bit at 10 Mb/s. The design also exhibits robustness against device aging effects. Compared to the state of the art, the proposed design improves the trade-off between robustness and area efficiency while

achieving $V_{DD}$ scalability down to 0.6 V, enabling it to be integrated into the energy-constrained systems that seek a cost-effective security measure.

# Chapter 3 Transforming a 6T-SRAM Array into a Robust Analog PUF with Minimal Overhead

## 3.1 Motivation

As the era of IoT approaches, secure interactions between devices become one of the biggest challenges [1–3]. PUFs have emerged to serve as a low-cost security primitive on the chip level by providing keys or physical hash functions from the uncontrolled random variations within circuit elements [13–19, 31]. Also, PUFs can be used for chip-ID generation for device anti-counterfeiting [17]. Several recent arts [13–19, 31] propose various techniques to create a PUF with either SRAM bi-stability [15–17] or mismatch across analog circuits [18, 19, 31], to achieve high robustness against noise, temperature or $V_{DD}$ variations and device aging at low cost.

Here, we propose a technique to reuse existing SoC resources—SRAM and ADC—to create a PUF. Unlike conventional SRAM-based PUFs utilizing reset states, our design centers on configuring the access transistors of a 6T-SRAM bitcell to form a pair of $V_{th}$ sensors with minimal area overhead. By digitizing and comparing the outputs of a pair of sensors, we can acquire a PUF bit. As the operation of $V_{th}$ sensor does not rely on transient switching activities, the proposed analog PUF has better robustness against temporal noise than conventional SRAM-based PUFs. Also, the proposed analog PUF design exhibits better robustness against temperature and $V_{DD}$ variations, as well as device aging. As

compared to prior arts using dedicated circuits [13–19, 31], our proposed design costs less area per bit, and this area savings is expected to increase with the number of PUF bits.

We prototype a test chip with 1 Kb of SRAM and circuitry for the proposed analog PUF transformation. Off-chip ADC and data processing are used to digitize, compare and generate the PUF bits from analog outputs, which is expected to be implemented on-chip in IoT devices in the future. The per-bit area cost of our design is 0.622 $\mu m^2$ for a 1 Kb PUF transformed from an SRAM array, marking 12× area reduction compared to the most compact design with comparable robustness [7]. The proposed design exhibits high robustness against noise, showing a Bit-Error-Ratio (BER) of less than 0.63% and an unstable-bit ratio of 2.15%. The bit-flipping ratio over −15~85 °C temperature variation is 4.88% and that over 0.5–1.2 V $V_{DD}$ variation is 5.3%. Hamming distances (HD) and NIST randomness test confirm the desirable randomness and uniqueness of the proposed PUF.

## 3.2 Proposed Technique

### 3.2.1 Analog PUF Architecture

Figure 3.1 shows the proposed analog PUF's architecture and operation. The entire system consists of a regular SRAM, extra circuits for PUF transformation, and a subsystem (dash-line box) for digitization and processing. By selecting and transforming the target bitcell at the input address (addr_key), the $V_{th}$s of its two access transistors (AL, AR) are sensed, digitized and compared to generate one PUF bit.

We build the SRAM with a 32 × 32 array of regular 6-T bitcells, a word-line (WL) decoder, and bit-line (BL) circuitries. We add a BL multiplexer (BLMUX), PUF switches (PSW), a PUF footer and a finite-state machine (FSM) to enable the transformation without

modifying the original circuits. We design the BLMUX with a one-layer multiplexing structure to minimize area overhead and use zero-$V_{th}$ thick-oxide NMOS for the PUF footer. Adding the BLMUX slightly increases the capacitance ($C_{db}$) on the bitline yet by less than 3% and ignorable leakage due to floating the BLS and BLBS nodes in normal SRAM operation. The two-to-one MUX inserted before the WL decoder delays the critical path only slightly.

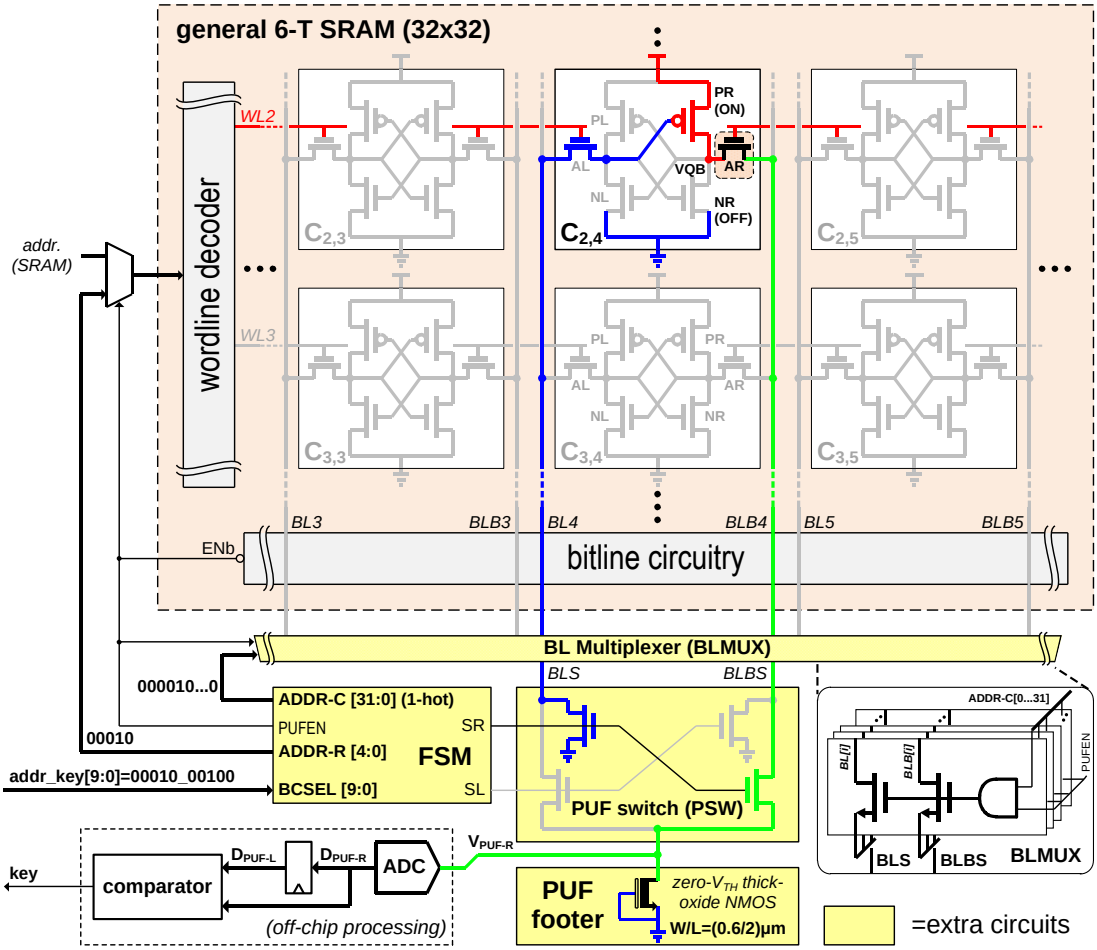



Figure 3.1 Proposed architecture to transform an SRAM array to a PUF.

Once triggered, the FSM disables normal SRAM operation, decodes the addr_key and controls other circuits through multiple signals: target WL at ADDR-R and the enable signal (PUFEN) are asserted; one of the 32 BL/BLB pairs is selected (BLS, BLBS) through

BLMUX by ADDR-C and PUFEN; PSW connects BLS and BLBS to either GND or the PUF footer at SL = 0 and SR = 1 (sense AR). The target bitcell is transformed into the effective analog PUF circuit as in Figure 3.2(a). The output voltage $V_{\text{PUF-R}}$ is proportional to the $V_{\text{th}}$ of AR, which is then digitized ($D_{\text{PUF-R}}$) by the ADC. Next, the FSM transforms it to Figure 3.2(b) to sense the $V_{\text{th}}$ of AL by changing the PSW control signals (SL = 1, SR = 0) while keeping other control signals. The output voltage $V_{\text{PUF-L}}$ is digitized ($D_{\text{PUF-L}}$) and compared with $D_{\text{PUF-R}}$ to generate a PUF bit. Those bits with same $D_{\text{PUF-L}}$ and $D_{\text{PUF-R}}$ values are set to fixed *1* or *0* based on the least-significant bit of the column address to balance the overall *1–0* ratio.

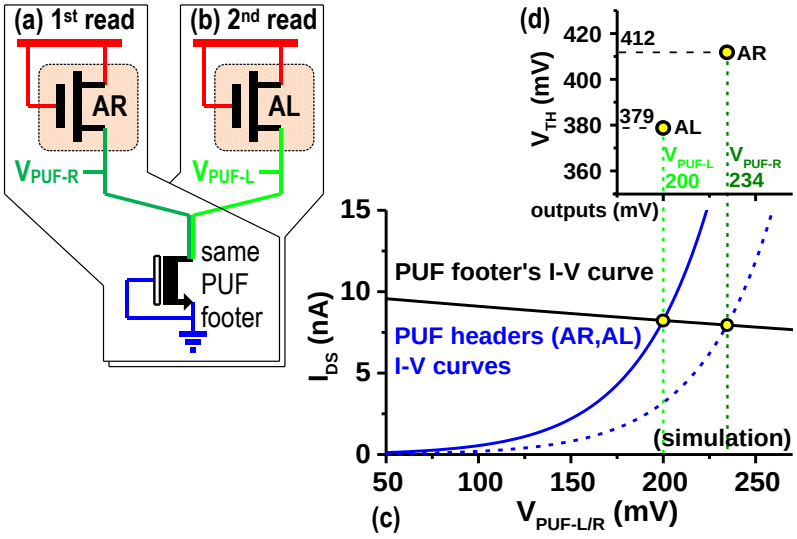



Figure 3.2 (a) (b) equivalent circuits for sensing $V_{\text{th}}$ of AR and AL, (c) operation principle, (d) physics ($V_{\text{th}}$) to outputs ($V_{\text{PUF-D}}$) conversion.

## 3.2.2 Principle of Operation

The $V_{\text{th}}$-sensing circuits used in the proposed PUF are based on two-transistor voltage-reference circuits [20]. $V_{\text{PUF,L}}$ is given by

$$V_{\text{PUF,L}} = V_{\text{DD}} + \frac{n_{\text{AL}}}{n_{\text{F}}} V_{\text{th,F}} - V_{\text{th,AL}} + \phi_{\text{t}} n_{\text{AL}} \ln\left(\frac{\beta_{\text{AL}}}{\beta_{\text{F}}} \cdot \frac{n_{\text{AL}}-1}{n_{\text{F}}-1}\right), \tag{3.1}$$

where $\phi_t$ is the thermal voltage and $n$ and $\beta$ are the sub-threshold slope and strength of the transistors. The expression of $V_{\mathrm{PUF,R}}$ has a similar form. The difference between $V_{\mathrm{PUF\text{-}L}}$ and $V_{\mathrm{PUF\text{-}R}}$ represents their $V_{\mathrm{th}}$ mismatch and shows reduced impacts from temperature and $V_{\mathrm{DD}}$ variations due to the removal of shared parameters, as given by

$$V_{\mathrm{PUF-D}} = V_{\mathrm{PUF,L}} - V_{\mathrm{PUF,R}} \approx V_{\mathrm{th,AR}} - V_{\mathrm{th,AL}}. \quad (3.2)$$

Figure 3.2(c) shows the circuit operation with the *I–V* curves of PUF footer and the access transistor (header). The PUF footer has almost constant current (black line) as it is a zero-$V_{\mathrm{GS}}$ device. The PUF headers' current are exponential functions of $V_{\mathrm{PUF,L/R}}$ (blue curves). The X-axis projection of the intersection point of the two currents is the output voltage. The $V_{\mathrm{th}}$ mismatch between two access transistors leads to different I–V curves, intersection points and, finally, output voltages with the same PUF footer. Figure 3.2 (d) shows a typical case: A 34 mV output-voltage difference represents a 33 mV $V_{\mathrm{th}}$ difference between the two headers (AL and AR).

## 3.3 Measurement results

We fabricate a test chip for the proposed analog PUF in 65 nm general-purpose CMOS (Figure 3.3). We use a National Instruments data acquisition card and LabVIEW as the digitization and data-processing system. The off-chip ADC has a 16 b precision for ±5 V input range. We truncate the ADC outputs to an effective number of bits (ENOB) of 9 for a range of 0–1 V. A lower ENOB will result in more bits set with the least significant bit of the column addresses as $D_{\mathrm{PUF\text{-}L}}$ and $D_{\mathrm{PUF\text{-}R}}$ are more likely to be the same, hurting the PUF's randomness, while a higher ENOB increases ADC cost significantly. We chose ENOB of 9 as a sweet spot between PUF randomness and hardware complexity.

First, we measure the distribution of analog outputs of $V_{PUF\text{-}L}$, $V_{PUF\text{-}R}$ and calculate $V_{PUF\text{-}D}$ (= $V_{PUF\text{-}L} - V_{PUF\text{-}R}$) across 6,144 b from six PUF instances under normal condition (27 °C, 1 V). Figure 3.4(a) shows the distributions of $V_{PUF\text{-}L}$ and $V_{PUF\text{-}R}$ with $\sigma$ and $\mu$ of 12.72 and 548.6 mV and 12.94 and 548.1 mV, respectively. The distribution of $V_{PUF\text{-}D}$ (Figure 3.4(b)) shows a mean of 0.536 mV, implying no bias in the PUF bits, and a standard deviation of 17.97 mV, close to $\sqrt{2}$ times the standard deviation of $V_{PUF\text{-}L}$ and $V_{PUF\text{-}R}$, suggesting independence between $V_{PUF\text{-}L}$ and $V_{PUF\text{-}R}$.

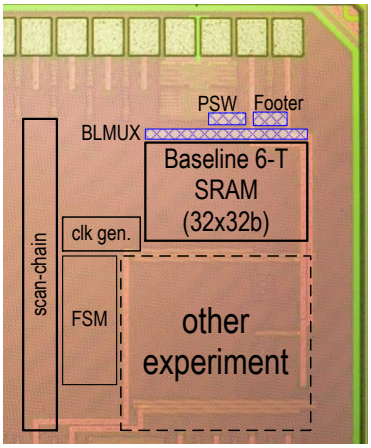



Figure 3.3 Chip Die Photo.

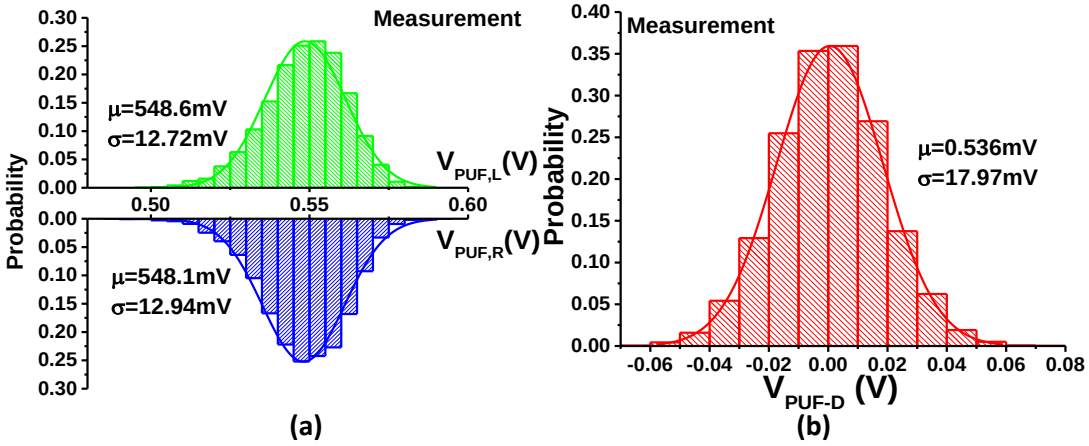



Figure 3.4 (a) Distributions of $V_{PUF\text{-}L}$ and $V_{PUF\text{-}R}$, (b) $V_{PUF\text{-}D}$ distribution.

We then evaluate the robustness of the proposed PUF design through various metrics. We first test the robustness of output against temporal noise. The unstable-bit ratio (number of

bits flipped at least once throughout all the readings) and BER (number of bit errors comparing each reading to a reference key) are evaluated by reading the PUFs 10,000 times. As is shown in Figure 3.5, with more and more readings, we observe that the unstable bit ratio saturates to 2.15%, with an upper boundary for BER of 0.63%. The average BER is measured to be 0.26%.

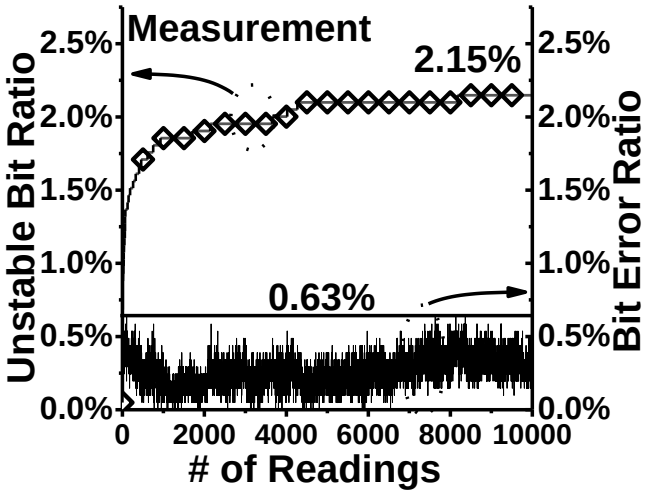



Figure 3.5 Unstable-bit ratio and bit-error ratio vs. the number of readings.

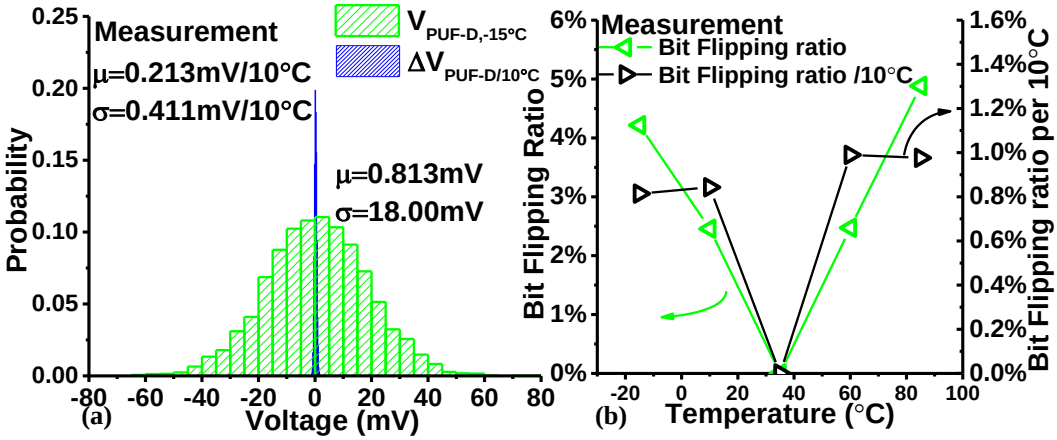



Figure 3.6 (a) Distributions of $\Delta V_{\text{PUF-D/10 °C}}$ and $\Delta V_{\text{PUF-D,−15 °C}}$, (b) bit-flipping ratio and the bit-flipping ratio per 10 °C vs. temperatures.

Second, we evaluated the robustness against temperature variation by sweeping temperature from −15 °C to 85 °C. $\Delta V_{\text{PUF-D}}$/10 °C, the average change in $V_{\text{PUF-D}}$ per 10 °C, is evaluated in Figure 3.6(a). Its distribution exhibits a mean of 0.213 mV/10 °C and a

standard deviation of 0.411 mV/10 °C, which is much narrower than the distribution of $V_{PUF\text{-}D}$ at −15 °C, implying desirable robustness against temperature variations. We also examine the robustness of digitized PUF bits by reading them at −15 °C, 10 °C, 35 °C, 60 °C and 85 °C and comparing them with the reference key produced at 35 °C. Shown in Figure 3.6(b), the maximal bit-flipping ratio across the temperature range is about 4.88%. The worst bit-flipping ratio per 10 °C temperature change is 0.99%/10 °C. This value is also almost constant across the temperature range.

Third, we sweep $V_{DD}$ from 0.5 V to 1.2 V and evaluate the output robustness, also for both the pre- and post-digitized cases. We measure $\Delta V_{PUF\text{-}D}$ across $V_{DD}$ at the normal temperature (27 °C) and use a similar robustness metric, $\Delta V_{PUF\text{-}D/0.1V}$, which is the average $V_{PUF\text{-}D}$ change per 0.1 V $V_{DD}$ variation. As shown in Figure 3.7(a), its distribution exhibits a mean of 0.05 mV/0.1 V and a standard deviation of 1.04 mV/0.1 V, which is much narrower than that of $V_{PUF\text{-}D}$ at $V_{DD}$ = 1 V. The digitized outputs are tested across a $V_{DD}$ range of 0.5 to 1.2 V with a step of 0.1 V and compared to the reference produced at 1 V. The maximal bit-flipping ratio across the $V_{DD}$ range is 5.3% and the worst bit-flipping ratio is 1.3%/0.1 V at $V_{DD}$ = 1.2 V (Figure 3.7(b)).

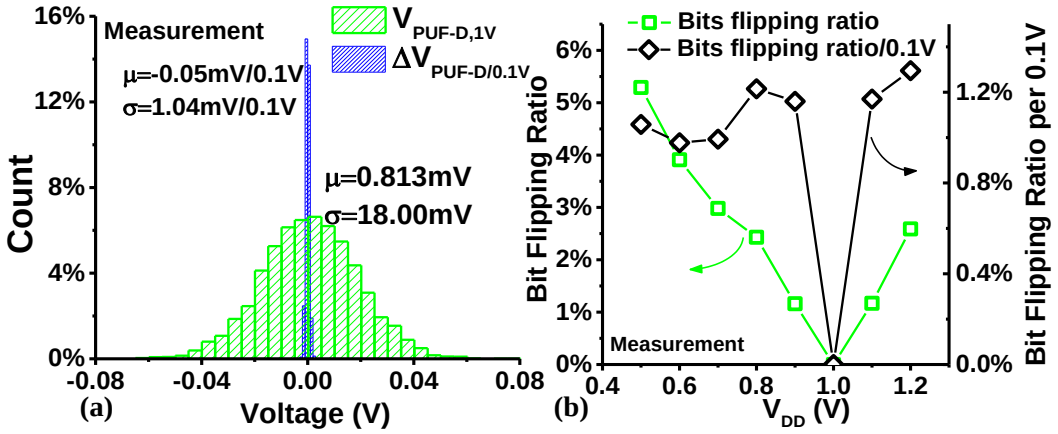



Figure 3.7 (a) Distributions of $\Delta V_{PUF\text{-}D/0.1V}$ and $V_{PUF\text{-}D,1V}$, (b) bit-flipping ratio and the bit-flipping ratio per 0.1 V vs. $V_{DD}$.

We then calculated both inter- and intra-PUF HDs to evaluate the uniqueness and robustness of our PUF design by dividing six PUF instances into 24 256 b keys. Figure 3.8 shows the inter-PUF HD mean of 127.5, close to the ideal 128. The intra-PUF HD mean is 0.731, 174.4× smaller than the average inter-PUF HD, great robustness to temporal noise.

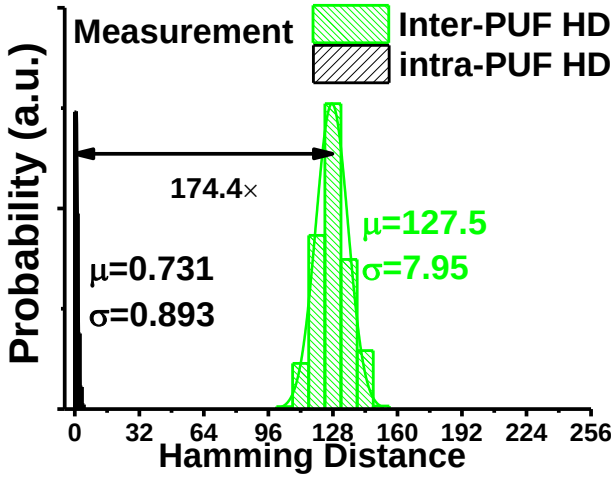



Figure 3.8 The inter- and intra-PUF HD measured under normal conditions.

The mismatches between PUs/PDs (PMOS/NMOS) in an SRAM bitcell can be modulated by aging effects (N/PBTI); thus, conventional SRAM-based PUFs' output can change over its lifetime. This prohibits reusing SRAMs in an SoC to create a reset-state-based PUF, as the device aging during normal operation can substantially modulate the PUF outputs. Yet, the proposed PUF design shows robustness against aging: It uses the mismatch between access transistors, instead of PUs/PDs. *The former experience much less aging than the latter as they are mostly OFF*.

We perform an accelerated aging test at 1.6 V, 125 °C for 16 h to induce an average 40 mV $V_{th}$ change in PUs [32], representing aging effects from multi-year use. Figure 3.9(a) shows $V_{PUF\text{-}D}$ changes as $\Delta V_{AGE}$ for a typical PUF pair with only minimal increase (0.35 mV) after 16 h. Figure 3.9(b) shows the distribution of $\Delta V_{AGE}$ across a chip with $\sigma$

and $\mu$ of 0.46 and 0.03 mV. The distribution of $\Delta V_{AGE}$ is much narrower than that of $V_{PUF\text{-}D}$ (Figure 3.9(b)), suggesting great robustness against aging. The change of the standard deviation of $\Delta V_{AGE}$ over aging time is shown in Figure 3.9(c), where $\sigma(\Delta V_{AGE})$ increases very slowly with stress time. Figure 3.9(d) shows a bit-flipping ratio due to aging effects of less than 1% over the stress time, suggesting desirable robustness against aging effects.

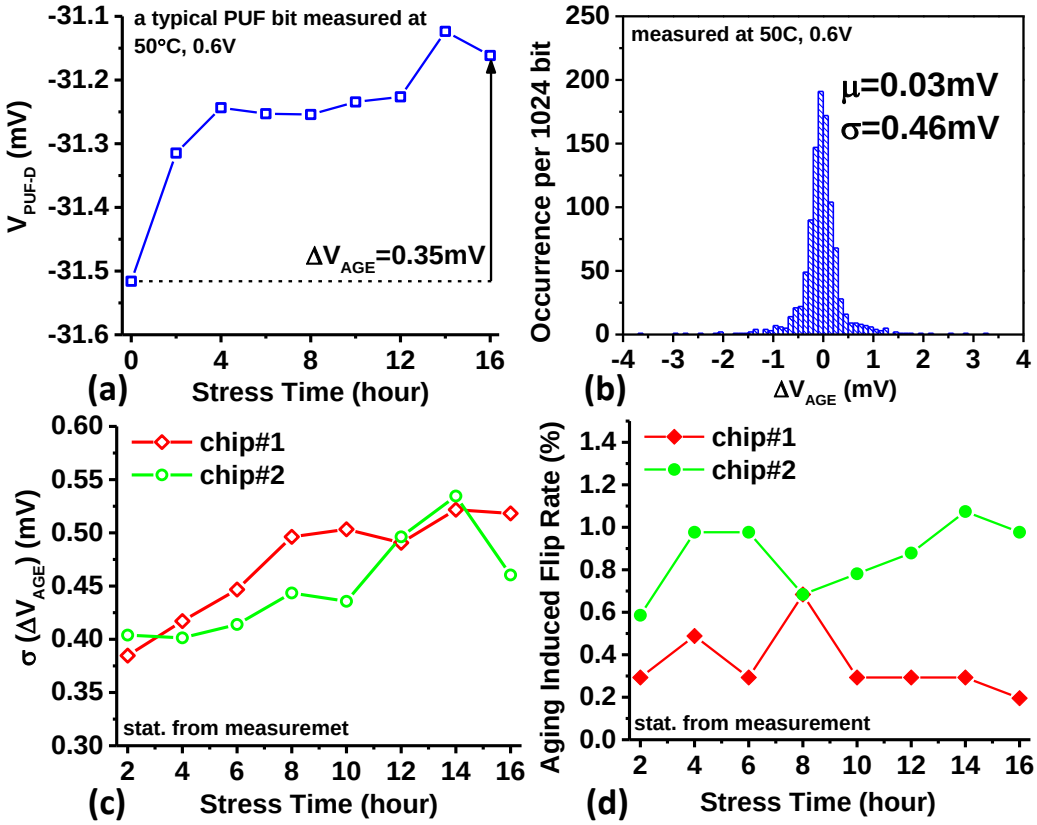



Figure 3.9 (a) A typical $\Delta V_{AGE}$ vs. accelerated aging time, (b) the distribution of $\Delta V_{AGE}$ of 1,024 PUF bits, (c) standard deviation of $\Delta V_{AGE}$ vs. accelerated aging time, (d) bit-flipping ratios during the aging test.

Last, we evaluate output randomness with NIST randomness test. Using a 6,000 b bitstream generated from six PUF instances, it passes all the NIST randomness tests applicable to the length of our data, including Frequency, Block Frequency, Runs, Longest Run of Ones, Cumulative Sum, FFT, Non-overlapping Template Matching, Serial and Approximate Entropy.

## 3.4 Comparison and Conclusion

### 3.4.1 Comparison

One of the benefits of reusing existing SRAM is lower area overhead as compared to designs with dedicated PUF cells. The total area overhead to transform a 1 kb SRAM array is 637 $\mu m^2$, or 12% of the SRAM area, assuming an on-chip ADC is available in an SoC. The per-bit-area cost is then 0.622 $\mu m^2$, marking the smallest per-bit area among the existing PUF designs with comparable robustness.

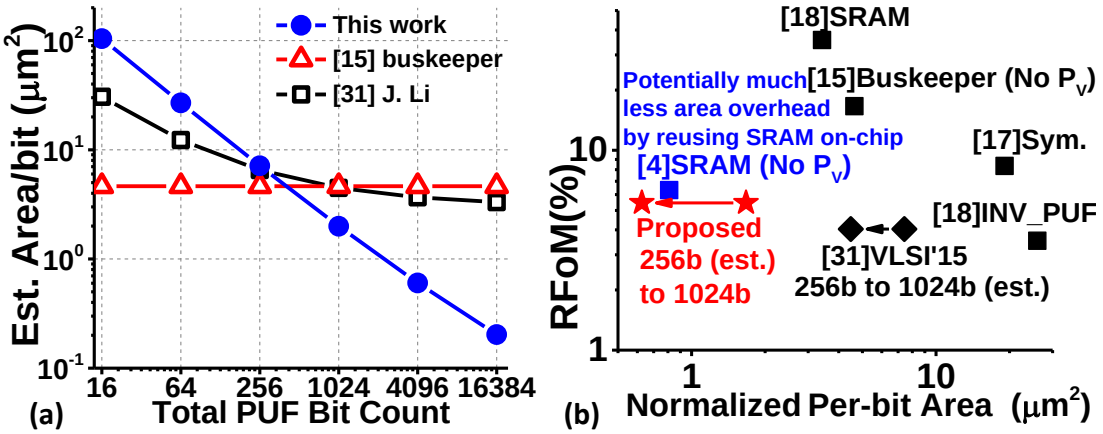



Figure 3.10 (a) Area/bit vs. PUF bit counts (b) RFoM: PUF robustness vs. area.

The area-per-bit overhead of the proposed PUF scales down with array size as the added peripheral circuits scale by a square root function of the array size (assuming the array has same row and column numbers). As shown in Figure 3.10(a), the estimated per-bit area of the proposed design is compared with [15] and [31]. As can be seen, the proposed design reduces per-bit area substantially, especially with larger arrays.

Figure 3.10(b) compares RFoMs (defined in Section 2.4.2) and areas of several prior arts. The proposed design improves the trade-off between RFoM and area. Compared to one of the state-of-the-art, it costs 3.9× (256 b) and 7.2× (1 kb) less area while exhibiting only slightly worse RFoM. One prior digital PUF using SRAM reset states, which can

potentially incur almost zero area overhead, achieves similar RFoM while its robustness against $V_{DD}$ variation and ramp-up time is not presented [15]. Interestingly, another digital PUF using SRAM reset states [18] shows substantially worse RFoM, implying that we should carefully examine such digital PUFs in term of robustness. Also, as described before, digital PUFs using SRAM reset states can be less robust against aging effects. Our comparisons are further summarized in Table 3.1. The proposed design achieves the lowest $V_{DD,min}$, comparable robustness to the state of the art while consuming much smaller area per bit.

| | Tech. (nm) | Norm. area per bit (um$^2$) | $V_{DD,min}$ (V) | Unstable Bits at Norm. Cond. | Temp. Range (°C) | Bit flipping ratio per 10°C | $V_{DD}$ range (V) | Bit flipping ratio per 0.1V | Norm. Inter-PUF HD |
|---|---|---|---|---|---|---|---|---|---|
| **[17] Sym.** | 130 | 29.86 | 0.9 | 3.04% | 0-80 | 0.68% | 0.9-1.2 | 1.82% | 0.506 |
| **[17] Cent.** | 130 | 50.59 | 0.9 | 3.78% | 0-80 | 0.635% | 0.9-1.2 | 1.82% | 0.501 |
| **[16]** | 22 | 40.68[5] | 0.7 | 0.97%[1] | 25-50 | N/A | 0.7-0.9 | N/A | 0.481 |
| **[15] Buskeeper** | 65 | 4.6387[5] | N/A | ~4.5% | -40~85 | 1.14%[4] | N/A | N/A | 0.491 |
| **[15] SRAM** | 65 | 0.81[5] | N/A | ~6% | -40~85 | 0.33%[4] | N/A | N/A | 0.497 |
| **[18] SRAM** | 65 | 3.42[5] | N/A | 16.6% | 25-85 | >6.67% | 0.7-1 | >16.67%[4] | 0.332 |
| **[18] INV.[3]** | 65 | 25.35[5] | 0.7 | 1.73% | 25-85 | 0.47%[4] | 0.7-1 | 1.3%[4] | 0.501 |
| **[31]** | 65 | 7.42 | 0.6 | 2.00%[2] | 0-80 | 0.44% | 0.6-1.2 | 0.13% | 0.500 |
| **Proposed** | **65** | **0.622[6]** | **0.5** | **2.15%** | **-15~85** | **0.99%** | **0.5-1.2** | **1.3%** | **0.498** |

[1]: TMV, aging hardening and bit-masking are applied. [2]: TMV11 is used. [3]: Only INV_PUF is included. SA_PUF has the similar performance.
[4]: Conservatively estimated from reported unstable bit ratio or BER. [5]: only bitcell area, not including peripherals [6]: Assuming on-chip ADC available

Table 3.1 Comparison to state-of-the-art PUF circuits.

## 3.4.2 Conclusion

In this work, we present a technique to transform a 6T-SRAM array into an analog PUF with minimal area overhead. The technique transforms a pair of access transistors in a bitcell into two $V_{th}$ sensors and produces a PUF bit by digitizing and comparing their outputs. We fabricated a silicon prototype in 65 nm CMOS and tested for randomness, uniqueness, and robustness against noise, temperature or $V_{DD}$ variations, and device aging. The measurements show that the proposed technique outperforms the prior arts in the trade-off between area and robustness and costs much less per-bit area with an increased number of PUF bits.

# Chapter 4 Triple-Mode, Hybrid-Storage, Energy-Harvesting Power Management Unit: Achieving High Efficiency against Harvesting and Load Power Variabilities

## 4.1 Motivation

The IoT vision projects deployment of trillions of devices virtually everywhere. Integrated with sensing, computation and communication capabilities, these devices will boost productivity and ultimately transform our everyday life experience [1, 3]. One of the biggest challenges in realizing the vision of IoT systems lies in powering all these devices practically and cost-effectively. Frequently recharging or replacing batteries of trillions of devices can significantly increase maintenance cost. In this aspect, energy harvesting garners a significant amount of attention [34, 36–43, 48] as it can enable energy-autonomous operations of IoT devices.

A PV cell, which converts light to electrical energy, is one of the most attractive harvesting modalities due to its high efficiency and low cost. Consequently, designing a PV EH PMU has been an active research area [36–43].

Figure 4.1(a) shows a conventional EH PMU architecture with two converters in series: one for transferring energy from a PV cell to a battery and the other from the battery to a load. This architecture, however, suffers from the fact those two converters always perform high-ratio voltage conversion. Typically, PV-cell output voltage ($V_{PV}$) is in the range of

0.3–0.6 V, Li-Ion battery charging voltage ($V_{Bat}$) is in the range of 3–4 V, and the load supply voltage ($V_{Load}$), considering energy-efficient near/sub-threshold circuits, is in the range of 0.3–0.6 V. Therefore, in the worst case, this conventional architecture can perform greater than 100× cumulative-ratio voltage conversion.

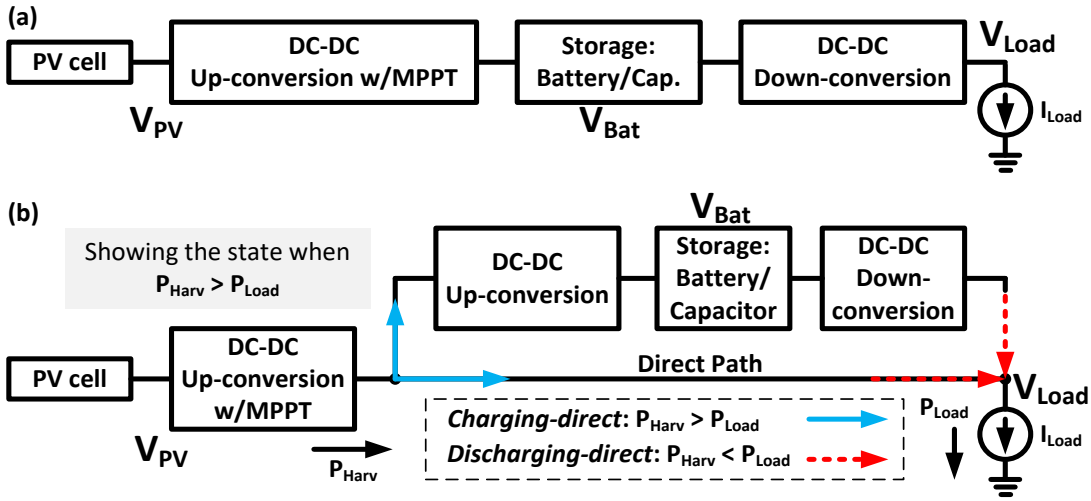



Figure 4.1 Conventional EH PMU architectures. (a) Single-mode architecture with two converters in series. (b) Dual-mode architecture with charging-direct and discharging-direct modes.

These high-ratio conversions can cause substantial loss in the end-to-end conversion efficiency. Particularly, SC-DC, which have gained increased popularity for the on-chip integration capability [34, 35, 37, 39, 40, 42, 46, 48, 49], often exhibit decreasing conversion efficiency as the conversion ratio increases [40, 49]. On the other hand, while inductor-based power converters are often more efficient in high-ratio voltage conversion, the aforementioned high conversion ratio can still introduce a considerable conversion loss in them [36].

To avoid such high-ratio conversion, an alternative architecture has been proposed [36–38, 40] (Figure 4.1(b)). This architecture has two modes of operation: charging-direct and discharging-direct modes. In the charging-direct mode, it establishes a direct path from a

PV cell to $V_{Load}$ while regulating $V_{Load}$ by storing excess energy in the battery. In the discharging-direct mode, the load receives power through the direct path and the discharging path from the battery to $V_{Load}$. This architecture can reduce the amount of energy that flows through high-conversion-ratio paths, i.e., the paths between $V_{Bat}$ and $V_{Load}$, improving the overall conversion efficiency. Note that Ref. [40] is slightly different from Ref. [36–38] in the sense that it uses a bidirectional converter for the converter between the battery and the load. Yet, their operations are roughly the same.

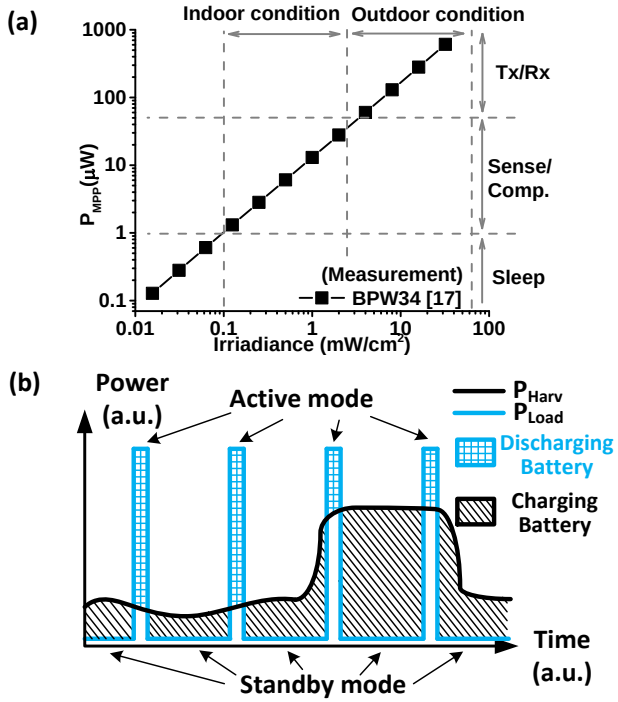



Figure 4.2 (a) Photovoltaic harvesting power values for different lighting conditions along with typical IoT node power dissipation. (b) Temporal variations of harvested power and load-dissipated power cause frequent battery charging and discharging, degrading end-to-end energy efficiency.

However, in this dual-mode architecture, the amount of energy that travels through the high-conversion-ratio paths between $V_{Bat}$ and $V_{Load}$ is dependent on the difference between harvested power ($P_{Harv}$) and dissipated power ($P_{Load}$). In practice, considering typical ranges of harvested power and IoT node power dissipation across operating modes (Figure

4.2(a)), there exists almost always a substantial temporal mismatch between $P_{Harv}$ and $P_{Load}$, due to the widely varying lighting conditions [51], as well as varying power dissipation in a load's different operating modes (e.g., data transmission, sensing/computation, sleep). In Figure 4.2(b), an IoT node with active/sleep mode is illustrated as an example, where the significant temporal variation of power dissipation forces a substantial portion of the energy to be either charged to or discharged from the battery, leading to high-conversion-ratio loss and lower end-to-end conversion efficiency. In such cases, the alternative architecture in Figure 4.1(b) only gains minimal efficiency improvement.

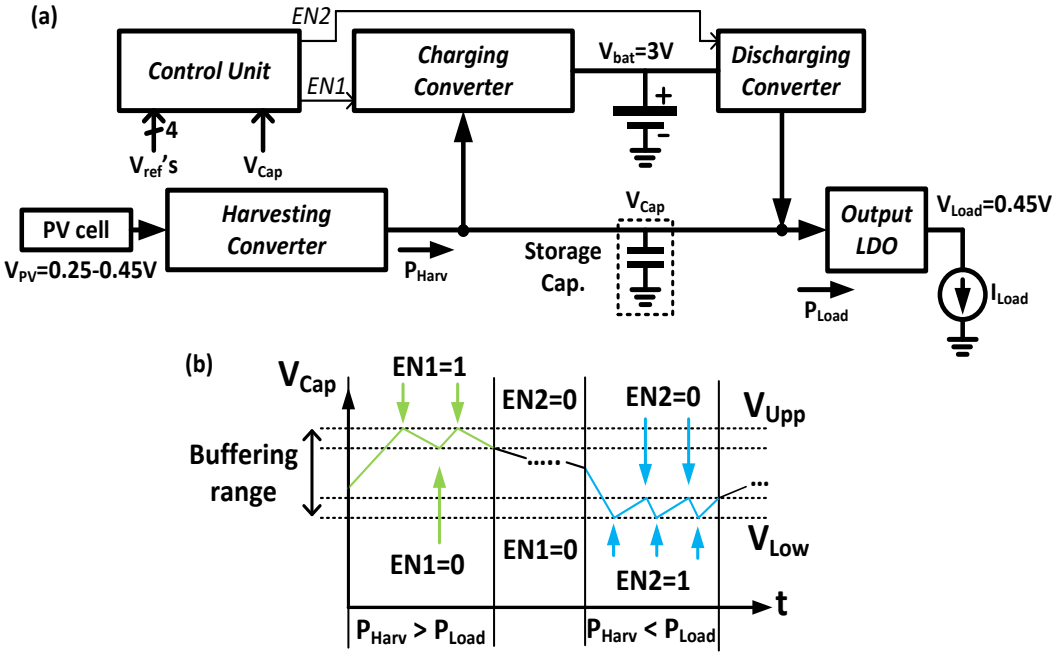



Figure 4.3 (a) The proposed triple mode EH PMU architecture with hybrid energy storage in battery and capacitor. (b) Hysteresis control scheme to switch among the three operating modes.

To address this challenge, we propose a novel EH PMU architecture that has three modes of operation, namely charging-direct, discharging-direct, and direct modes (Figure 4.3(a)). In the newly added direct mode, the proposed architecture only uses the direct path from a photovoltaic cell to a load, which involves a smaller conversion ratio. The path also

contains a capacitor, which serves as an intermediate energy storage and thus allows the voltage of the capacitor ($V_{Cap}$) to fluctuate within a set range. This enables the proposed architecture to use the direct path without battery involvement even under a temporal mismatch of $P_{Harv}$ and $P_{Load}$, improving the EH PMU's end-to-end energy efficiency. Although our architecture is based on SC-DC, it can be realized with inductive converters. Inductive converters typically have higher conversion efficiency, yet they can also suffer from low conversion efficiency for performing a large-ratio voltage conversion [36]. Maximal utilization of the low-conversion-ratio path under harvester and load power mismatches can improve end-to-end conversion efficiency.

We fabricated a test chip of the proposed PMU architecture in 65 nm CMOS. Our measurements show that the proposed architecture achieves up to 2.2× higher end-to-end conversion efficiency over the conventional dual-mode architecture under typical $P_{Harv}$ and $P_{Load}$ variation scenarios. We have also analyzed the effectiveness of the architecture and generated a framework for system design that guides capacitor sizing and capacitor voltage range selection for maximal efficiency improvement.

## 4.2 Proposed Design

### 4.2.1 System Architecture

Figure 4.3(a) shows the proposed EH PMU architecture. It consists of three SC-DC and one digital LDO, along with a battery and a capacitor. The first SC-DC, the *harvesting converter*, interfaces a PV cell to the capacitor, converting from $V_{PV}$ to $V_{Cap}$. The second converter, the *charging converter*, delivers the excess power to the battery (charging), converting from $V_{Cap}$ to $V_{Bat}$. The third converter, the *discharging converter*, supplies the

load the necessary power in case the power harvested by the PV cell is insufficient, converting from $V_{Bat}$ to $V_{Cap}$. Finally, a digital LDO regulates $V_{Load}$ to the desired voltage.

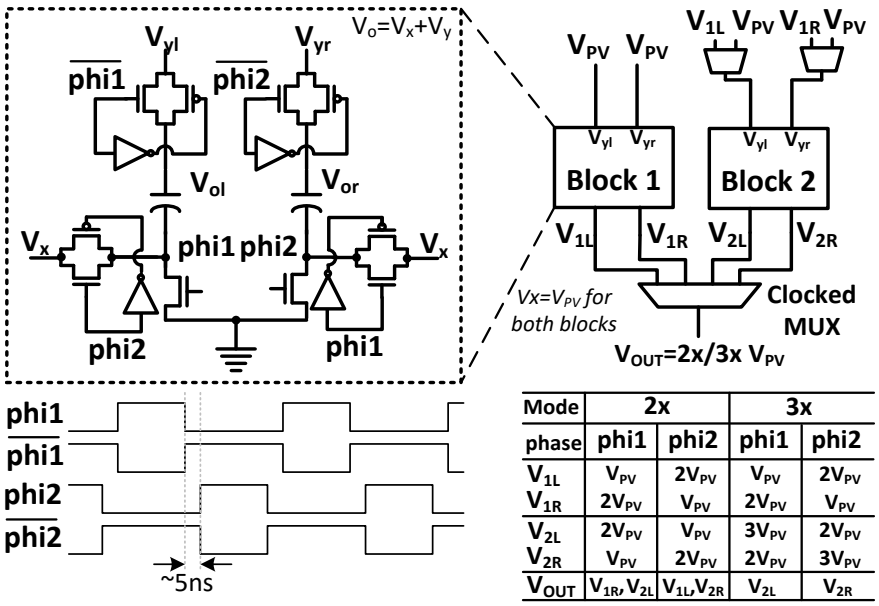
The proposed EH PMU operates in one of the three modes, namely *direct*, *charging-direct,* and *discharging-direct* modes. A control unit, built with comparators and digital logic circuits, switches among the three modes based on the level of $V_{Cap}$. Figure 4.3(b) shows the principle of the mode change control. If $P_{Harv} > P_{Load}$, the excess power charges the capacitor and raises $V_{Cap}$. If $V_{Cap}$ crosses an upper threshold ($V_{Upp}$), it asserts the enable signal (EN1) of the charging converter to store the excess harvested energy in the battery (i.e., charging-direct mode).

Similarly, if $P_{Harv} < P_{Load}$, the capacitor gets discharged, decreasing $V_{Cap}$. When $V_{Cap}$ crosses the lower threshold ($V_{Low}$), it asserts the enable signal (EN2) of the discharging converter to supply power from the battery to the load (discharging-direct mode).

Finally, If $P_{Harv} \approx P_{Load}$, both the charging and discharging converters are disabled, and the photovoltaic cell via the charging converter directly powers the load. Any temporal mismatch between $P_{Harv}$ and $P_{Load}$ is buffered by the intermediate storage capacitor without involving battery charging and discharging. In essence, this scheme regulates $V_{Cap}$ between $V_{Upp}$ and $V_{Low}$, creating a range of voltage as a buffering range. Note that we add hysteresis near $V_{Upp}$ and $V_{Low}$ to avoid excessive switching among the modes since the converters are switching at fixed frequencies. Other regulation schemes, such as pulse frequency modulation, can also be used to control the charging and discharging converters to improve the stability, as well as the efficiency of the system.

## 4.2.2 Circuit Implementations

Figure 4.4 shows the schematics of the harvesting converter. It is a step-up converter that can perform 2× or 3× voltage conversion from $V_{PV}$ to $V_{Cap}$. Specifically, it consists of two unit blocks that can be configured in series for 3× or in parallel for 2× voltage conversion. The settings for each conversion configuration are summarized in Figure 4.4 right bottom. Figure 4.4 left bottom also shows the waveforms of the non-overlapping clocks (phi1, phi2 and their inverted signals). The parallel configuration in 2× mode can maximize capacitor utilization and thus enable larger power-transfer capacity. To support low input voltage (~0.3 V), we used transmission gates in the power transfer path to reduce device ON resistance. Also, the converter generates and uses switching clock and other control signals that swing from 0 to $V_{Cap}$, further improving the ON resistance of the power transistors.




| Mode | 2x | | 3x | |
|---|---|---|---|---|
| phase | phi1 | phi2 | phi1 | phi2 |
| $V_{1L}$ | $V_{PV}$ | $2V_{PV}$ | $V_{PV}$ | $2V_{PV}$ |
| $V_{1R}$ | $2V_{PV}$ | $V_{PV}$ | $2V_{PV}$ | $V_{PV}$ |
| $V_{2L}$ | $2V_{PV}$ | $V_{PV}$ | $3V_{PV}$ | $2V_{PV}$ |
| $V_{2R}$ | $V_{PV}$ | $2V_{PV}$ | $2V_{PV}$ | $3V_{PV}$ |
| $V_{OUT}$ | $V_{1R}$, $V_{2L}$ | $V_{1L}$, $V_{2R}$ | $V_{2L}$ | $V_{2R}$ |

Figure 4.4 Schematics of the harvesting converter.

As shown in Figure 4.5, we designed the charging converter based on an eight-stage configurable-ratio charge pump topology. It can perform 6× to 9× step-up operation, converting $V_{Cap}$ to $V_{Bat}$. We can configure the conversion ratio by skipping some of the last

three stages. Figure 4.5 left shows the unit building block of each stage. High $V_i$ can incur reverse body bias, substantially increasing the ON resistance of the power transistors. To avoid such effects, using deep N-well, we connect the body of the NMOS power transistors to the input of the current unit stage ($V_i$) and the body of PMOS to the output ($V_{i+1}$).

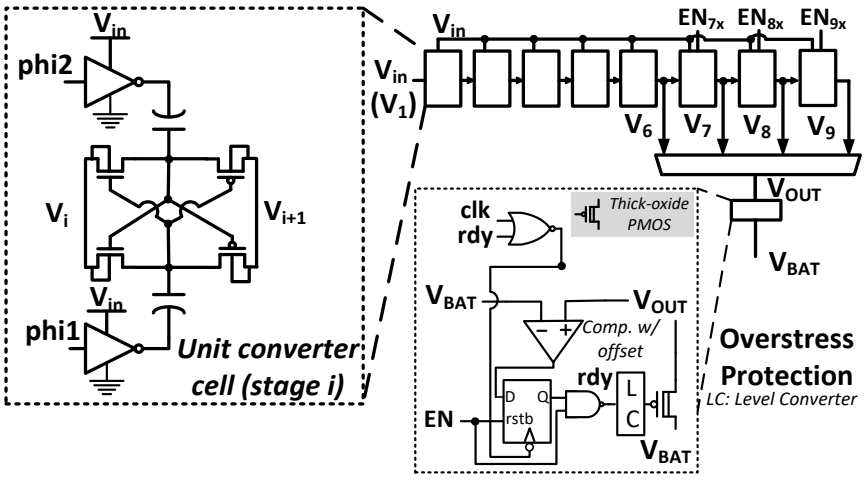



Figure 4.5 Schematics of the charging converter and overstress protection circuitry.

We used thin-oxide devices in the charging converter to reduce ON resistance while the output voltage of the converter can be as high as 3 V to 4 V. Therefore, it is critical to implement overstress protection. Particularly, while the converter is disabled—i.e., while the EH PMU is not charging the battery—the high battery voltage can stress the thin-oxide devices of the charging converters. As shown in Figure 4.5, we designed the overstress protection circuitry using a comparator with a predefined offset that compares the internal output of the converter and $V_{\text{Bat}}$. If the internal output is smaller than $V_{\text{Bat}}$ by a predefined amount, the protection circuitry asserts rdy, turning off the *thick*-oxide output PMOS to isolate $V_{\text{Bat}}$ from the converter's thin-oxide devices. The assertion of rdy gates the clock to the comparator and the flip-flop to save power dissipation. The level converter is designed based on the structure proposed in [52].

The discharging converter design is based on a two-stage 1/4-step-down architecture (Figure 4.6) to down-convert $V_{Bat}$ to $V_{Cap}$. Figure 4.6 shows the schematics of each stage. Similar to Figure 4.4, phi1 and phi2 are non-overlapping clock phases. The transistors on the high side of the stage receive clocks that swing from $V_o$ to $V_i$, instead of 0 to $V_i$, which reduces the power dissipation and allows the use of thin-oxide transistors. Similarly, the transistors on the low side receive clocks that swing from 0 to $V_o$. Finally, as shown in Figure 4.7, we based the output digital LDO on the shift-register topology [37]. This provides the regulated load supply voltage ($V_{Load}$) from $V_{Cap}$.

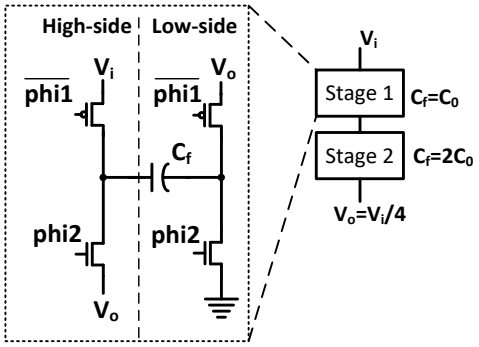



Figure 4.6 Schematics of the discharging converter.

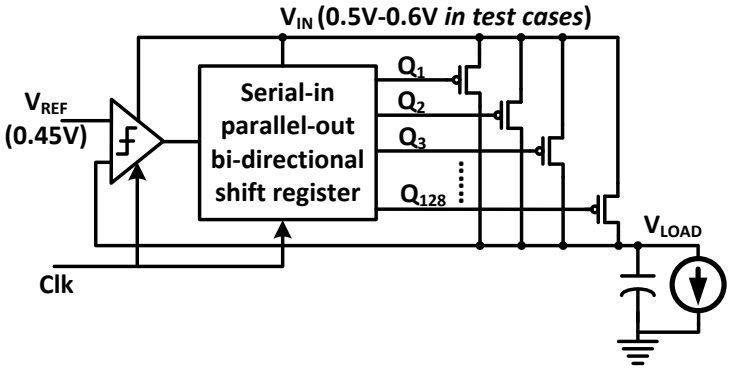



Figure 4.7 Schematics of the output digital LDO.

We include on-chip clock generators for the three SC-DCs and the output LDO. The clock generators for the harvesting and charging converters and the output LDO are designed

based on current-starved ring oscillators with tunable capacitive loads for frequency tuning (Figure 4.8(a)). These clocks swing from 0 to $V_{Cap}$. The clock generator for the discharging converter is based on ring-configured delay cells, each of which consists of a self-gated cross-coupled inverter and a leakage device (Figure 4.8 (b)). We can tune the oscillation frequency by modulating the gate bias of the leakage device [39]. This clock swings from 0 to $V_{Bat}$.

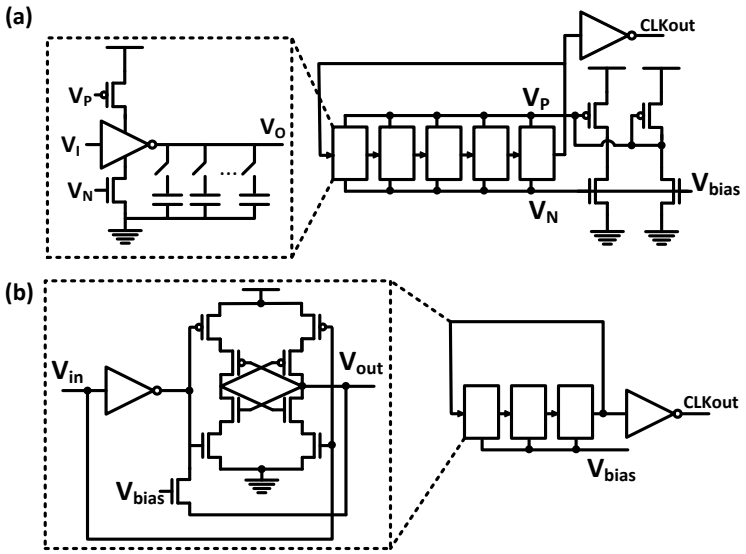



Figure 4.8 The clock generator design for the (a) harvesting converter, charging converter and LDO, and (b) discharging converter.

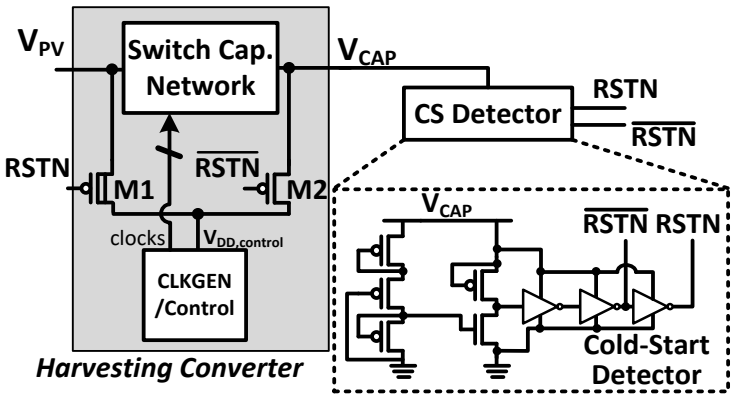



Figure 4.9 Self-start circuits including the cold-start detector.

Since the clock of the harvesting converter swings from 0 to $V_{Cap}$, if no charge is stored in the capacitor or battery, the clock generator cannot function, requiring the harvesting converter to have self-start capability. To implement the self-start circuitry, we added two PMOS transistors (M1 and M2, Figure 4.9) and a cold-start detector (similar to [45]). If $V_{Cap}$ is lower than a predefined threshold (i.e., too little charge resides in the capacitor), it makes $V_{PV}$ provide power to the clock generator for the cold start. As $V_{Cap}$ increases and crosses a predefined threshold of ~0.35 V, RSTN is asserted, which makes $V_{Cap}$ power the clock generator and other controls. While it is possible to keep using $V_{PV}$, the use of $V_{Cap}$ is desirable to reduce the ON resistance of the power transistors, thus improving the charging converter's conversion efficiency. Note that we used a thick-oxide device for the PMOS for $V_{PV}$ (M1) to reduce the leakage in the normal non-cold-start operation (Figure 4.9).

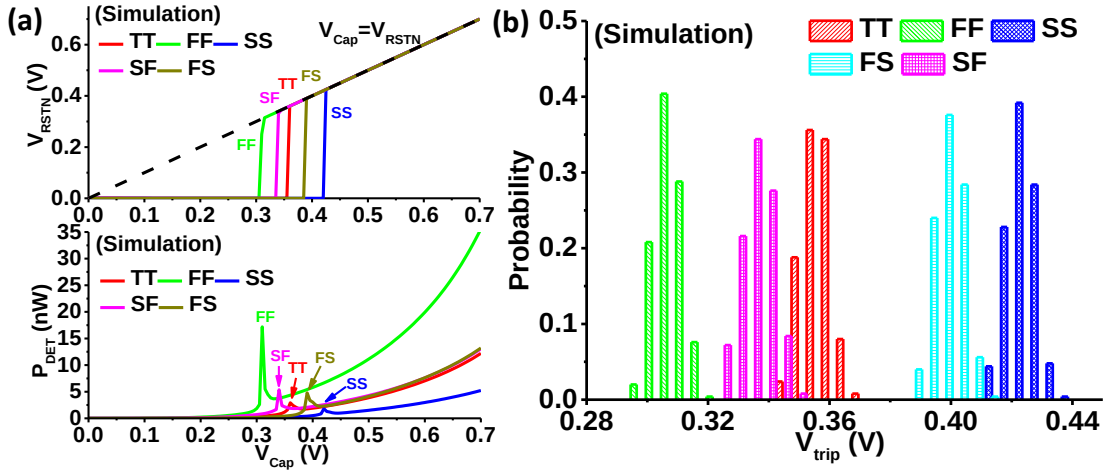



Figure 4.10 (a) Robustness of trip point voltage and power consumption of the cold-start detector across corners. (b) Monte-Carlo simulations of the trip-point voltage at each process corner.

We performed corner and Monte Carlo simulations to evaluate the cold-start technique's robustness. As shown in Figure 4.10(a), the predefined threshold ($V_{trip}$) of the cold-start detector varies by 110 mV across process corners. Its power consumption is in the range

of tens of nanowatts, small enough to have a minimal impact on the EH PMU efficiency. Figure 4.10(b) shows the results of the 250-point Monte Carlo simulation at each of the five process corners. As $V_{\text{trip}}$ varies in a reasonable range, the detector can robustly assert RSTN to start and stop the cold-start process. It is possible to implement tuning capability for the cold-start detector so that $\underline{V}_{\text{trip}}$ can be tuned based on process corner information. Yet, this is beyond the purpose of this work.

It is noteworthy that our proposed EH PMU architecture may frequently enable and disable the charging and discharging converters if $P_{\text{Load}}$ and $P_{\text{Harv}}$ vary widely. Frequent enabling and disabling of the converters can waste a considerable amount of energy since the charge stored in the flying capacitors while the converters are enabled can be lost via leakage while the converters are disabled. To save this energy, we can use a charge-retention technique such as one proposed in [46].

### 4.2.3 End-to-End Energy Efficiency Analysis of EH PMU

In this section, we analyze the end-to-end energy efficiency of the proposed EH PMU. To do so, we first define a metric for the efficiency ($Eff_{\text{ov}}$). The conventional PCE metric used for individual converters is insufficient for the evaluation of end-to-end efficiency: It cannot capture the impact of charging and discharging of energy storage devices. Therefore, we propose $Eff_{\text{ov}}$ as

$$Eff_{\text{ov}} = \begin{cases} \dfrac{E_{\text{Load}} + E_{\text{Bat}} \cdot PCE_{\text{discharge}}}{E_{\text{mpp}}}, if\ E_{\text{Bat}} > 0 \\ \dfrac{E_{\text{Load}}}{E_{\text{mpp}} + |E_{\text{Bat}}| / PCE_{\text{charge}} / PCE_{\text{Harvest}}}, if\ E_{\text{Bat}} < 0, \end{cases} \tag{4.1}$$

where $E_{mpp}$ is the total amount of energy that a PV cell harvests at its maximal power point (MPP); $E_{Load}$ is total load energy consumption; $PCE_{discharge}$ is the measured average PCE of battery-to-load conversion; $PCE_{charge}$ is the measured average PCE of capacitor-to-battery (proposed) or load-to-battery (conventional dual-mode architecture, Figure 4.1(b)) conversions; $PCE_{Harvest}$ is the efficiency of the harvester converter; $E_{Bat}$ is the amount of energy charged to or discharged from the battery during a time window (negative if a battery is discharged). Note that we do not include the LDO efficiency in the efficiency improvement analysis simply because it equally affects the proposed and the baseline architectures. In this regard, $Eff_{ov}$ is defined as the efficiency from harvester and battery to the LDO input.

Note that the energy drawn from the battery is what was previously harvested and charged to the battery. Therefore, we need to take the PCEs of harvesting and charging into account in defining $Eff_{ov}$. Similarly, the load will eventually consume the energy charged to the battery. Therefore, we include the impact of PCE for the discharging operation in defining $Eff_{ov}$. Also, note that $Eff_{ov}$ has no terms related to the change of energy stored in the capacitor. This is because in the test case studies, for simplicity, we assume that the EH PMU system operates in a steady state in the energy perspective: The amounts of energy flowing in and out of the capacitor are roughly the same. The capacitor energy storage could be incorporated into $Eff_{ov}$ for more generalized studies.

We constructed a simplified test case to emulate practical $P_{Harv}$ and $P_{Load}$ variations. We focus on short-term (a few seconds) temporal mismatch of $P_{Harv}$ and $P_{Load}$. Therefore, we assume $P_{Harv}$ is constant but $P_{Load}$ changes periodically between active and sleep modes. Figure 4.11(a) shows the $P_{Harv}$ and $P_{Load}$ profile of the test case. It can be modeled with

three parameters: (1) the active-to-sleep-mode power dissipation ratio of a load ($K_1$ = $P_{active}/P_{sleep}$), (2) the total load energy consumption to harvester energy generation ratio ($K_2$ = $E_{Load}/E_{Harv}$), and (3) the active-to-sleep-mode duration ratio ($K_3$ = $t_{active}/t_{sleep}$).

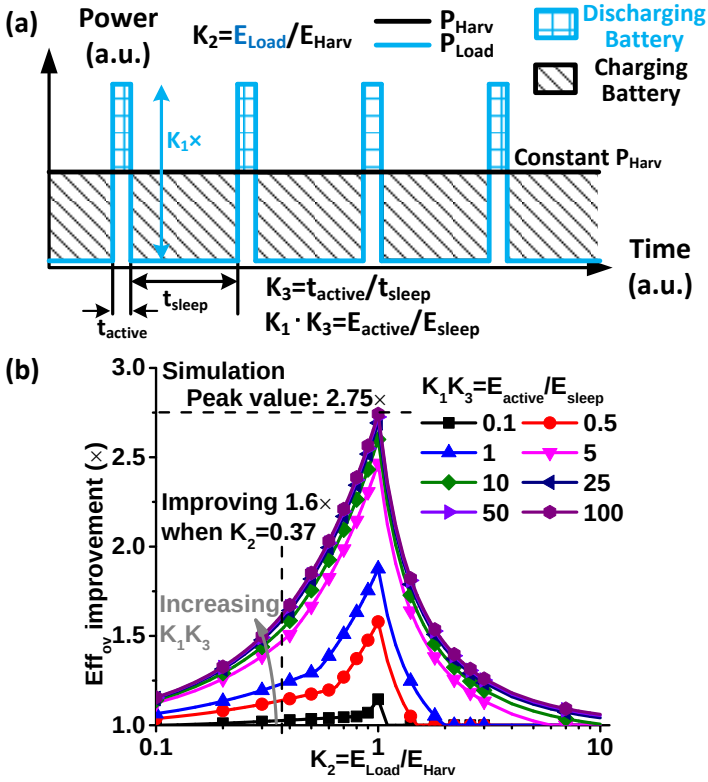



Figure 4.11 (a) Simplified test-case emulating $P_{Harv}$ and $P_{Load}$ variations. (b) End-to-end efficiency analysis of the proposed EH PMU architecture.

Using the test case with various parameter values for $K_1$, $K_2$ and $K_3$ representative for an IoT node system, we analyze and compare the $Eff_{ov}$ of the proposed EH PMU and the conventional dual-mode architecture. We use the representative values (60%) for $PCE_{charge}$ and $PCE_{discharge}$ based on the test chip measurement (see section 4.3). We do not need to specify $PCE_{Harvest}$ since it is the same between the proposed triple- and the conventional dual-mode architectures and thus has no impact in the comparisons. We also assume that the intermediate energy storage capacitor is sized large enough not to limit the power transfer over temporal power mismatch in the test-case duration. Therefore, for the

proposed architecture, only the net difference between $E_{Load}$ and $E_{Harv}$ over the test case duration is charged to or discharged from a battery.

Figure 4.11(b) shows the $Eff_{ov}$ improvement of the proposed EH PMU architecture over the conventional dual-mode architecture. Here we used the $V_{Upp}$ and the capacitor size of our design. We fixed $K_3$ at 1% while sweeping $K_2$ and $K_1 \bullet K_3$ (= $E_{active}/E_{sleep}$). This analysis shows that the $Eff_{ov}$ improvement of the proposed architecture peaks (2.75×) at $K_2 = 1$ and a high $K_1 \bullet K_3$ value. It reduces as $K_2$ deviates from 1 since $E_{Bat}$ dominantly degrades $Eff_{ov}$ in both the proposed and the conventional architectures (see Eq. (4.1)). Note that the worst case of our proposed architecture happens where it cannot use the storage capacitor and operates similarly with the conventional dual-mode architecture. Specifically, if $K_1 \bullet K_3$ is very small, for example, the variation of load-power dissipation becomes small and so does the energy charged to or discharged from the storage capacitor. This makes a large portion of energy flow to the battery, degrading $Eff_{ov}$.

## 4.3 Measurement Results

### 4.3.1 Chip Measurements

We fabricated the test chip for the proposed EH PMU in 65 nm CMOS. The target maximum load current of the EH PMU is 140 µA at 0.45 V load $V_{DD}$. We use ~2 fF/µm$^2$ metal-insulator-metal (MIM) capacitors to implement the capacitors. The total active area of the design is 0.48 mm$^2$. See the die photo in Figure 4.12(a).

We use a PV cell with a 7.5 mm$^2$ radiant-sensitive area for measurement. To facilitate the test procedure, instead of using a controllable light source and a real rechargeable battery, we use two source meter units (SMU) to emulate the behaviors of a PV cell and a

rechargeable battery. Figure 4.12(b) describes the setup, where the first SMU (SMU1) is connected in parallel with a PV cell that is fully covered in dark. This SMU-PV system can emulate PV cell behavior under various lighting conditions simply by adjusting the amount of the current that SMU1 sources. The other SMU (SMU2) is connected to $V_{Bat}$ and $V_{SS}$, which can emulate a rechargeable battery by operating in the voltage source mode. SMU2 can measure the current coming in and out of the unit, which are roughly equivalent to charging and discharging currents.

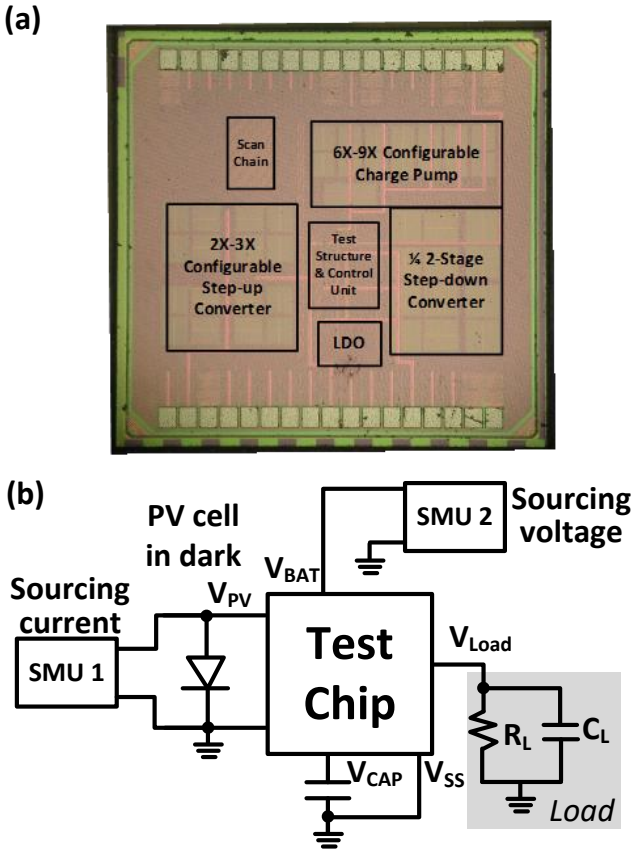



Figure 4.12 (a) Die photo. (b) Testing setup using SMU-based PV cell and rechargeable battery.

We measured the harvesting efficiency ($PCE_{Harvest}$) of the harvesting converter across PV cell currents ranging from 0.5 to 32 mW/cm$^2$. As shown in Figure 4.13(a), the converter's peak $PCE_{Harvest}$ is 60.9% for $V_{Cap}$ = 0.5 V and 61.9% for $V_{Cap}$ = 0.6 V with off-chip MPPT

control. (Our test chip has no on-chip MPPT control loop due to the limited design time.) The extraction efficiency, $P_{OUT}/P_{MPP}$, is shown in Figure 4.13(a), where the $P_{MPP}$ is the output power of the PV cell at the maximal power point. Figure 4.13(b) shows the measured $P_{MPP}$ across varying irradiance levels and the corresponding optimal $PCE_{Harvest}$ of the two operation modes. The maximal $P_{MPP}$ is ~600 µW and the corresponding $P_{Harv}$ is 196 µW for $V_{Cap}$ = 0.5 V and 298 µW for $V_{Cap}$ = 0.6 V.

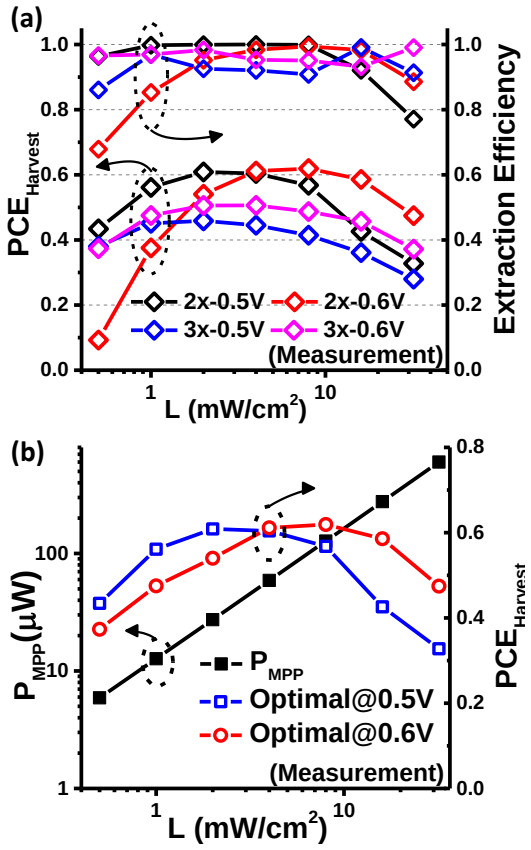



Figure 4.13 (a) Harvesting converter conversion efficiency, PV cell extraction efficiency. (b) The output power of PV cell at maximal power point and optimal harvesting efficiency of the harvesting converter.

We also measured the charging-converter PCE for variable-ratio conversions from 0.6 V to 3 V, where the charging converter is enabled when $V_{Cap}$ reaches $V_{Upp}$ (0.6 V in our test cases). As shown in Figure 4.14, the peak PCE is measured to be 63.8% at 6× conversion

and decreases at higher conversion ratios, which is able to convert higher input power. Then, we measured discharging-converter PCE (Figure 4.15). Converting 3 V $V_{Bat}$ to 0.6 V $V_{Cap}$, the converter achieves 59.1% peak PCE. When $V_{Cap}$ is 0.5 V, the peak PCE is 55.8%.

Finally, we measured the current efficiency of the output LDO. As shown in Figure 4.16, operating at 1 MHz clock, 0.5 V $V_{Cap}$, and 50 mV dropout voltage, the LDO consumes 2.7 µA quiescent current. Considering a typical load of 120 µA, the current and power efficiencies are 97.7% and 87.9%. Table 4.1 summarizes the measurement results and Table 4.2 compares the proposed design with some prior work.

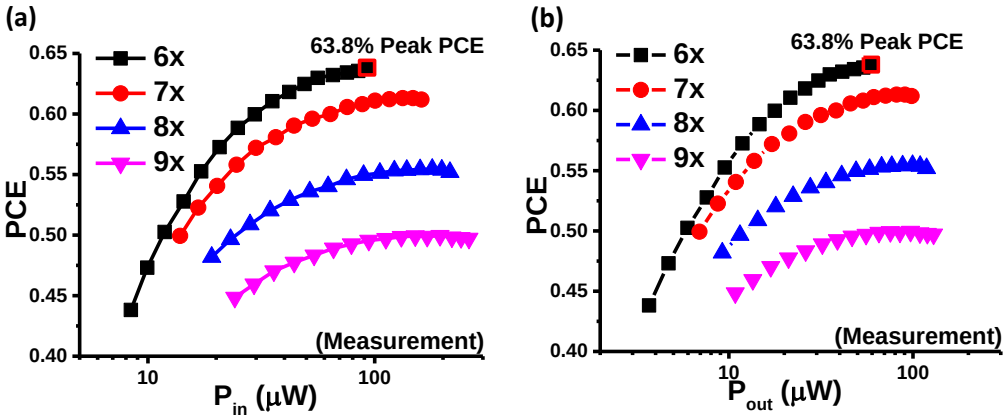



Figure 4.14 Power conversion efficiency of the charging converter.

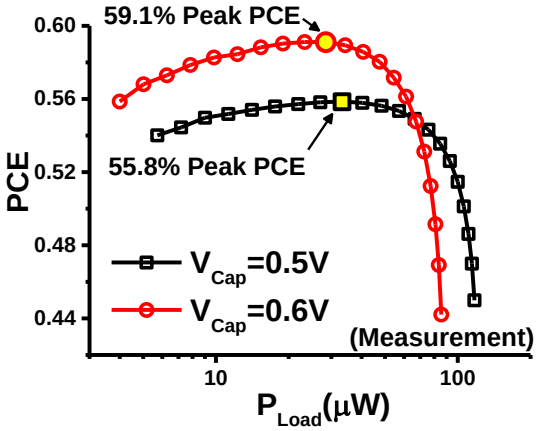



Figure 4.15 Power conversion efficiency of the discharging converter.

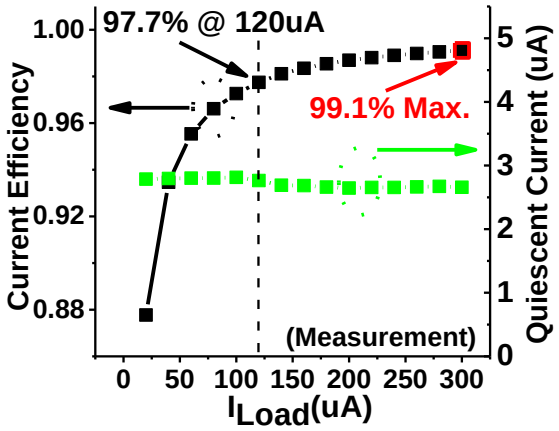



Figure 4.16 Current efficiency and quiescent current of the output LDO.

| | |
|---|---|
| **Process** | 65nm CMOS |
| **Area** | 0.48mm$^2$ |
| **$V_{Load}$** | 0.45V |
| **$V_{Bat}$** | 3V |
| **Converters** | SC DC-DC×3 + LDO |
| **Capacitors** | First step-up SC:200pF<br>Second step-up SC: 160pF<br>Step-down SC: 240pF<br>600pF in total (MIM) |
| **Harvesting power** | $P_{MPP}$=600µW Max. measured @ $V_{MPP}$=0.45V<br>$P_{Harv}/P_{MPP}$: 32% @VCap=0.5V, 50% @$V_{Cap}$=0.6V |
| **Load Power** | LDO: 63µW Max. @ 0.5V $V_{Cap}$<br>Step-down SC: 106µW with >0.5 PCE |
| **Peak Conv. Efficiencies** | Harvesting: 60.9% @$V_{Cap}$ =0.5V,<br>61.9% @$V_{Cap}$ =0.6V<br>Step-up SC: 63.8% @ 6×, 0.6V $V_{Cap}$<br>Step-down SC: 59.1% @ $V_{Cap}$ =0.6V,<br>55.8% @ $V_{Cap}$ =0.5V<br>LDO: 97.7% current efficiency @ 120µA |
| **End-to-End efficiency** | Direct path: 53.3% @$V_{Cap}$ =0.5V<br>Charging and discharging path: 19.3% |

Table 4.1 Chip summary.

| | [36] | [38] | [40] | This work |
|---|---|---|---|---|
| Architecture | Dual mode | Dual mode | Dual mode | Triple mode |
| Converters | Inductor based | SC+LDO | SC | SC+LDO |
| Off-chip components | 22µH inductor<br>15µF, 47µF decaps | No | No | <200µF storage cap.<br>@$V_{Upp}$=0.6V |
| Process | 0.35µm | 0.18µm | 65nm | 65nm |
| Area | ~2.5mm$^2$ | ~0.94mm$^2$ | 0.48mm$^2$ | 0.48mm$^2$ |
| Energy storage | Battery | Battery | Capacitor | Battery & capacitor |
| Voltage levels | Battery: 3.3V<br>Thermal: 20-160mV<br>PV: 0.15-0.75V<br>Piezo.: 1.5-5V<br>Load: 1.8V | Battery: 3.6V<br>PV: ~0.3V<br>Load: 0.4 and 0.5V | Storage cap: 0-3V<br>PV: 1V | Battery: 3V<br>PV: 0.3-0.5V<br>Load: 0.45V |
| End-to-end efficiency | 58%: Thermal; 83%: PV; 79%: Piezo. | 38% (Only down conversion reported) | 53%<br>(cycle average) | 53.3% (direct)<br>19.3% (up-down) |

Table 4.2 Comparison to prior works.

## 4.3.2 Test Case Studies

In the test case studies, we configure our proposed EH PMU architecture with the following system parameters: $V_{Upp}$ = 0.6 V, $V_{Low}$ = 0.5 V, $C_{Cap}$ = 0.47 mF. We choose the capacitor size that is sufficient to handle transient power mismatch between $P_{Harv}$ and $P_{Load}$. We will further discuss optimizing these system parameters in Section 4.4. We emulate the conventional dual-mode architecture by reusing our EH PMU and configuring it with the following system parameters: $V_{Upp}$ = 0.52 V, $V_{Low}$ = 0.5 V, $C_{Cap}$ = 1 µF. We cannot completely eliminate $C_{Cap}$ as it causes control stability issues, but this slightly improves the efficiency of the emulated dual-mode architecture, ensuring fair comparisons. Due to the greatly reduced buffering voltage range and a smaller capacitor, even with a slight mismatch of $P_{Harv}$ and $P_{Load}$, the EH PMU operates in either charging-direct or discharging-direct mode.

| Test | | $I_{load}$ Variations | $P_{PV}$ Variations | $V_{Low}$-$V_{Upp}$ | $C_{Cap}$ |
|---|---|---|---|---|---|
| **Exp. 1** | Proposed | Constant 50µA | 1Hz enabling/disabling | 0.5-0.6 | 0.47mF |
| | Baseline | Constant 50µA | 1Hz enabling/disabling | 0.5-0.52 | 1µF |
| **Exp. 2** | Proposed | 120µA, $T_{active}$=50ms, Period=1s | Always enabled | 0.5-0.54 | 0.47mF |
| | Baseline | 120µA, $T_{active}$=50ms, Period=1s | Always enabled | 0.5-0.52 | 1µF |
| **Exp. 3** | Proposed | 120µA, $T_{active}$=50ms, Period=1s | 1Hz enabling/disabling | 0.5-0.6 | 0.47mF |
| | Baseline | 120µA, $T_{active}$=50ms, Period=1s | 1Hz enabling/disabling | 0.5-0.52 | 1µF |

Table 4.3 Detailed setup of the experiments.

We operate and measure our proposed EH PMU under three test cases that emulate realistic $P_{Harv}$ and $P_{Load}$ variations. Table 4.3 summarizes the details of the test cases. In the first test case (Experiment 1), we periodically modulate $P_{Harv}$ by enabling and disabling the PV cell

output every second. This case can mimic a shading effect: The PV cell has some shade and thus produces substantially smaller $P_{Harv}$. In the second test case (Experiment 2), we use a $P_{Load}$ profile with a 50 ms 120 µA peak power dissipation every second. This mimics the load's active-sleep-mode transition. Finally, in the third test case (Experiment 3), we combine Experiments 1 and 2, both $P_{Harv}$ and $P_{Load}$ vary. In all three experiments, we consider the lighting level in the range of 0.5 to 32 mW/cm$^2$.

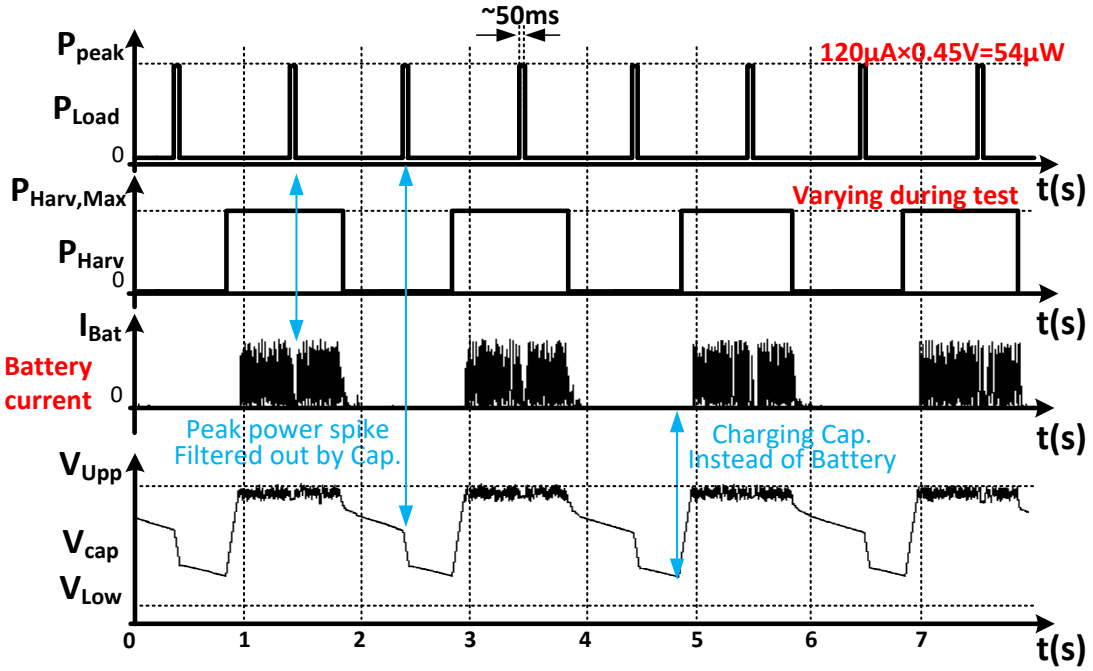



Figure 4.17 Example waveforms measured during Experiment 3. The proposed EH PMU can handle the mismatch between $P_{Harv}$ and $P_{Load}$, thus minimizing battery involvement.

Figure 4.17 shows several key waveforms measured during Experiment 3.

In the beginning, $P_{Harv}$ is set larger than $P_{Load}$, assuming the load is in the sleep mode. This increases $V_{Cap}$ toward $V_{Upp}$ as the capacitor is charged. We observe zero battery charging current ($I_{Bat}$) during this time. Once $V_{Cap}$ crosses $V_{Upp}$, our proposed PMU enables the charging converter that charges the battery with the excess harvested power, giving a

nonzero $I_{Bat}$. Note that $I_{Bat}$ and $V_{cap}$ fluctuate because of the hysteresis-based control (Figure 4.2(b)).

Specifically, if $V_{cap}$ reaches $V_{Upp}$, the controller enables the charging converter, and this makes $V_{cap}$ not rise too much beyond $V_{Upp}$, and eventually drop to below $V_{Upp} - \Delta V$, at which point the controller disables the charging converter. This makes both $I_{BAT}$ and $V_{cap}$ fluctuate.

Eventually, the load enters active mode and starts to draw current, directly from the harvester. This temporarily makes $I_{Bat}$ smaller, creating a notch in the $I_{Bat}$ waveform (denoted by the first blue arrow).

Then, $P_{Harv}$ is reduced, assuming the lighting conditions get worse. The load, although in the sleep mode, consumes some amount of power, and thus $V_{Cap}$ drops. As denoted by the second blue arrow, the load again enters the active mode, consuming the peak amount of power. This creates a steeper slope in the $V_{Cap}$ waveform. Still, the sufficiently sized capacitor delivers the necessary power and therefore no current is drawn from the battery. Note that $V_{Cap}$ does not cross $V_{Low}$.

After this, $P_{Harv}$ again becomes larger than $P_{Load}$, and the excessive power is charged into the buffering capacitor, increasing $V_{Cap}$ toward $V_{Upp}$ (denoted by the third blue arrow). Throughout this experiment, our proposed EH PMU operates either in the direct or the charging-direct mode. No battery discharge occurs.

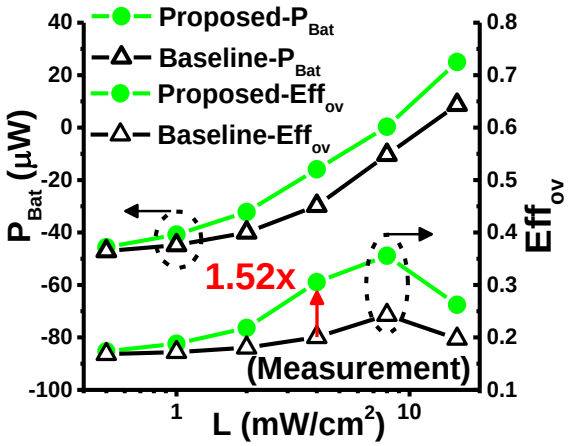



Figure 4.18 Battery charging and discharging power (average) and end-to-end efficiency measurement during Experiment 1. The proposed EH PMU achieves up to 1.52× higher $Eff_{ov}$ over the conventional dual-mode architecture.

Across all three test cases (Experiments 1, 2 and 3), the proposed EH PMU architecture can significantly improve the end-to-end efficiency.

In Experiment 1, as shown in Figure 4.18, we find that our proposed architecture achieves 1.52× higher $Eff_{ov}$ than the baseline that emulates the conventional dual-mode architecture. Notably, the $Eff_{ov}$ improvement is large at the light intensity that makes $E_{Harv}$ similar to $E_{Load}$. This is because any difference of $E_{Harv}$ and $E_{Load}$ requires charging to or discharging from the battery, worsening $Eff_{ov}$. This is indeed the same conclusion that we have in the $Eff_{ov}$ analysis in Section 4.2.3.

Similarly, in Experiments 2 and 3, the proposed EH PMU architecture exhibits 1.83× and 2.20× higher $Eff_{ov}$ over the emulated baseline (Figure 4.19 and 4.20). Again, we find that $Eff_{ov}$ improvement is large at the light intensities that make $E_{Harv}$ and $E_{Load}$ similar. In addition, larger $P_{Harv}$ and $P_{Load}$ variations (used in Experiment 3) benefit the proposed EH PMU architecture.

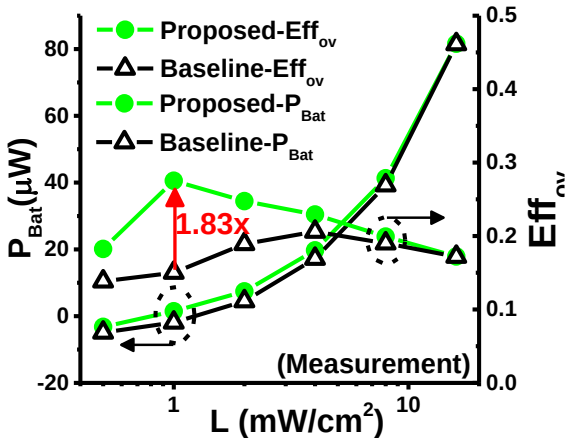



Figure 4.19 Battery charging and discharging power (average) and end-to-end efficiency measurement during Experiment 2. The proposed EH PMU achieves up to 1.83× higher $Eff_{ov}$.

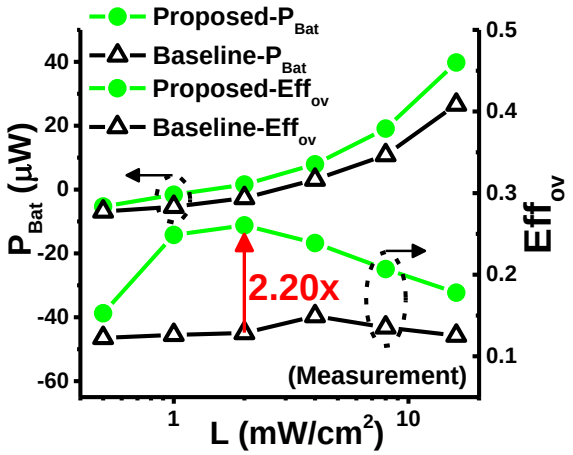



Figure 4.20 Battery charging and discharging power (average) and end-to-end efficiency measurement during Experiment 3. The proposed EH PMU achieves up to 2.2× higher $Eff_{ov}$.

## 4.4 System Design Tradeoffs

In this section, we further analyze the trade-off among capacitor size, buffering voltage range and end-to-end efficiency to provide a system design framework.

The size of the capacitor and buffering voltage range together determine the amount of buffered energy in the intermediate capacitor storage. The capacitor size proportionally increases the buffering capability while the buffering voltage range has a more complicated

relationship. In the first order, the amount of energy buffered in the capacitor increases quadratically with the buffering voltage range. However, as we increase the buffering voltage range, i.e., increasing $V_{Upp}$, the linear loss of the last output LDO proportionally increases. We can reduce such loss by replacing the output LDO with a more efficient SC-DC. However, as we discussed in the Introduction, SC-DCs indeed exhibit lower PCE with an increasing voltage conversion ratio. Furthermore, as $V_{Cap}$ gets higher, the harvesting converter also becomes less efficient, again due to the increasing conversion ratio. On the other hand, a higher $V_{Cap}$ requires smaller voltage conversion for the charging converter (i.e., $V_{Cap}$ to $V_{Bat}$), improving its conversion efficiency.

To understand these trade-offs, we created a framework to estimate the proposed EH PMU's end-to-end efficiency (Figure 4.21).

This framework takes the following inputs to predict $Eff_{ov}$:

- $P_{Harv}$: average harvesting power.
- $P_{active}$: load power dissipation during the active mode.
- $P_{sleep}$: load power dissipation during the sleep mode.
- $t_{active}$: active mode time.
- $t_{window}$: steady-state time window.
- $V_{Low}$: lower bound of buffering voltage range.
- $V_{Upp}$: upper bound of buffering voltage range.
- $PCE_0$, $PCE_1$, and $PCE_2$: PCE of the discharging, harvesting and charging converters, respectively.
- $PCE_3$: PCE of the last down-conversion from $V_{Cap}$ to $V_{Load}$, which the digital LDO performs in the current design.

```
1   // PCE0: PCEDischarge from VBat to VLoad; independent of VCap
2   // PCE1: PCEHarvest; function of VCap
3   // PCE2: PCECharge; function of VCap
4   // PCE3: Conversion PCE from VCap to VLoad; function of VCap
5   // We use Linear interpolation to find PCE1, PCE2, PCE3 as a function of VCap
6   // Eq1(EBat, Eharvest, PCE0, PCE1, PCE2): Equation (1) calculating Effov
7   // when cap is small, VCap fluctuate between upper/lower bound
8   PCE3previous = PCE3((VUpp + VLow)/2);  // initial PCE3, as VCap fluctuate btw. VUpp & VLow
9   PCE1previous = PCE1((VUpp + VLow)/2);  // initial PCE1, as VCap fluctuate btw. VUpp & VLow
10  For CCap= CCap,min: CCap,step: CCap,max {
11     Ecap-active = (Pactive/PCE3previous - Pharvest · PCE1previous) · Twindow · K3;
12     Ecap-sleep = (Pharvest · PCE1previous – Psleep) · Twindow · (1-K3);
13     Ebuffer = CCap · (VUpp2-VLow2);
14     if (Ecap-active>Ecap-buffer & Ecap-sleep>Ebuffer) { // VCap reach both boundaries
15        EBat = (Ebuffer - Ecap-active)/PCE0)/PCE0 + (Ecap-sleep - Ebuffer) · PCE2((VUpp + VLow)/2);
16        Effov = Eq1(EBat, Eharvest, PCE0, PCE2((VUpp + VLow)/2));
17     } else {
18           if (Ecap-active > Ecap-sleep)  {// VCap stays near VLow
19                 // x: actual higher bound of VCap during test
20              x = solve (CCap · (x2-VLow2) == Ecap-sleep);
21              PCE3previous = PCE3((x + VLow)/2);
22              PCE1previous = PCE1((x + VLow)/2);
23              EBat = (Ecap-sleep - Ecap-active)/PCE0; // EBat<0
24              Effov = Eq1(EBat, Eharvest, PCE0, PCE1((x+VLow)/2), PCE2((x+VLow)/2));
25           } else { // VCap stays near VUpp during test
26                 // x: actual lower bound of VCap during test
27              x = solve (CCap · (VUpp2-x2) == Ecap-active);
28              PCE3previous = PCE3((x + VUpp)/2);
29              PCE1previous = PCE1((x + VUpp)/2);
30              EBat = (Ecap-sleep - Ecap-active) · PCE2((x+VUpp)/2); // EBat>0
31              Effov = Eq1(EBat, Eharvest, PCE0, PCE1((x+VUpp)/2), PCE2((x+VUpp)/2));
32           }
33        }
34  }
```

Figure 4.21 Framework to evaluate the trade-off between capacitor size, buffering range selection and end-to-end efficiency of the proposed EH PMU architecture.

In this framework, we assume for higher $V_{Upp}$, an SC-DC will be used to perform this conversion. $E_{cap\text{-}active}$ and $E_{cap\text{-}sleep}$ represent the amount of energy discharged from and

charged to the capacitor during the active period and sleep period, respectively. $E_{buffer}$ is the amount of energy that $C_{Cap}$ can store (provide) while $V_{Cap}$ increases (decreases). Note that the power profile in Figure 4.21 is represented by $K_1$, $K_2$ and $K_3$, the same as in Section 4.2.3. We can evaluate the framework while increasing $C_{cap}$ (intermediate storage capacitor size) in a small step size. Comparing $Eff_{ov}$ from the framework evaluations yields the optimal system parameters.

Note that $V_{Cap}$ is a complex function of various parameters including power profile, capacitor size, $V_{Upp}$ and more. To precisely solve for $V_{Cap}$, therefore, we would need to perform an iterative process, increasing the complexity of the framework. Instead, we approximated the converter efficiencies (which are functions of $V_{Cap}$) using efficiency values found in the previous evaluation with one step smaller capacitor size, using $PCE_{1previous}$ and $PCE_{3previous}$ for $PCE_1$ and $PCE_3$ in calculating $E_{cap\text{-}active}$ and $E_{cap\text{-}sleep}$ (Lines 11 and 12 in Figure 4.21). As long as we use a fine-grained step in the capacitor size sweep, this approximation will introduce only minimal error.

As an example, we evaluated the framework with the following parameters: $P_{harvest}$ = 10 µW, $P_{active}$ = 360 µW, $P_{sleep}$ = 0.360 µW, $t_{active}$ = 10 ms, $T_{window}$ = 1 s, $V_{Low}$ = 0.5 V, $V_{Upp}$ = [0.6 V, 0.9 V, 1.2 V, 1.5 V]. These parameters result in $E_{Load}/E_{Harv}$ = 0.4, $t_{active}/t_{window}$ = 1%, and $P_{active}/P_{sleep}$ = 1000. Figure 4.22(a) shows the results of the framework evaluations: $Eff_{ov}$ improvement of the proposed EH PMU architecture over the conventional dual-mode architecture as a function of $C_{cap}$. Here we assume the PCEs of the converters are roughly constant across $V_{Cap}$ That is, the increased loss when converting from higher $V_{Cap}$ to $V_{Load}$ and from $V_{PV}$ to higher $V_{Cap}$, as well as a reduced loss when converting from higher $V_{Cap}$ to

$V_{Bat}$, are all neglected. In Figure 4.22(b), we considered the impact of $V_{Cap}$ on the PCEs in evaluating the framework.

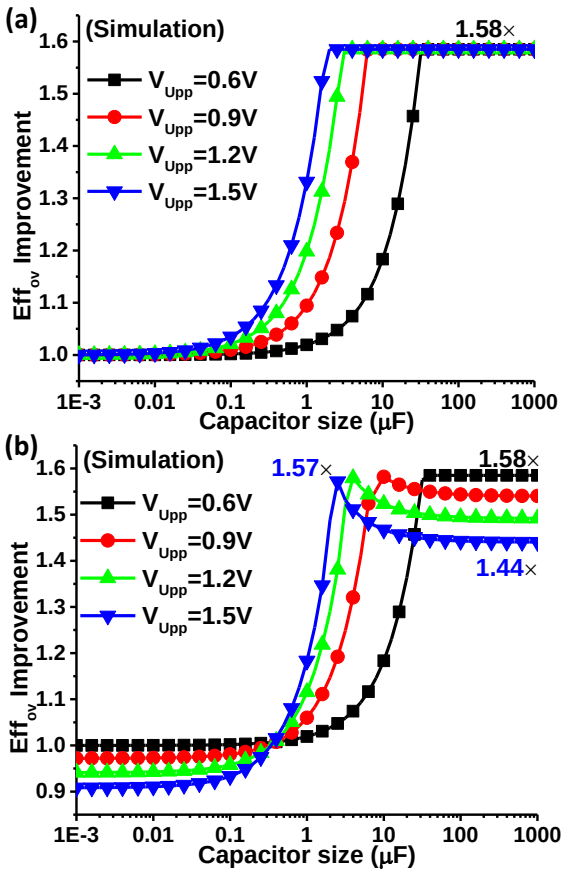



Figure 4.22 The trade-off among capacitor size, buffering voltage range ($V_{Upp}$) and $Eff_{ov}$ improvement over the conventional dual-mode architecture. (a) Results without considering converters' PCE dependencies on $V_{Cap}$. (b) Results with considering the dependencies. At $E_{Load}/E_{Harv}$ = 0.4, the maximal $Eff_{ov}$ improvement is ~1.58× for both (a) and (b).

Both results show that increasing $C_{cap}$ benefits the proposed EH PMU architecture with a maximal improvement of ~1.58×. It is also shown that increasing $V_{Upp}$ can help reduce $C_{cap}$ for the same amount of $Eff_{ov}$ improvement. Finally, as shown in Figure 4.22(b), the $V_{Upp}$-incurred PCE degradation of the converters are pronounced if $C_{cap}$ is too small or large: If the capacitor is too small, it can tolerate only a small amount of power mismatch between harvester and load, worsening $Eff_{ov}$. If the capacitor is too large, $V_{Cap}$ remains very close to

$V_{\text{Upp}}$ because $C_{cap}\left(V_{Upp}^2 - V_{Low}^2\right)$ is larger than $E_{\text{cap-active.}}$ In this case, high $V_{\text{Upp}}$ can hurt $PCE_1$ and $PCE_3$.

Finally, it is noteworthy that the model we used above can be extended to general power profiles, by dividing the time window in a more fine-grained way, adding the history effect on the $V_{\text{Cap}}$, and employing an iterative solving function for $V_{\text{Cap}}$ from the power and previous $V_{\text{Cap}}$ values. Also, the PCE calculation can be improved as the current PCEs in the model is using an average value calculated at average $V_{\text{Cap}}$, while the actual PCE should be averaged in the energy aspect. Finally, in the above analysis, leakage of the energy-storage capacitor is ignored as we use a ceramic capacitor with little leakage in the experiments. A leakier capacitor can waste a fixed amount of power, degrading end-to-end efficiency. Thus, we need to avoid the capacitor whose leakage is a good fraction of average harvesting and load power.

## 4.5 Conclusion

In this chapter, we present a triple-mode, hybrid-storage EH PMU architecture interfacing a PV cell, a 3 V battery and a 0.45 V load. It consists of three SC-DCs and a digital LDO for load–supply voltage regulation. The proposed EH PMU architecture can effectively cope with the temporal mismatch of harvested and load power while minimizing battery-charging and -discharging operations. Based on several test cases that emulate practical harvested and load power variations, our proposed architecture achieves up to 2.2× better end-to-end efficiency than the conventional dual-mode architecture. We also analyze the trade-off among capacitor sizing, capacitor-voltage-range selection, and end-to-end efficiency, providing system-level design guidelines.

# Chapter 5 Fully Integrated and Fully Digital Hybrid Error/Replica-based Nanowatt Power Management Unit and Neural Spike Processor Co-design with Energy-Robustness Co-optimization Control

## 5.1 Motivation

Enabled by the recent advances in ultra-low-power circuits, the drastically reducing power consumption of IoT nodes that scales down to the sub-microwatt range impose new challenges in PMU design, including compact form factor, ultra-low quiescent power, high PCE while delivering sub-microwatt, wide input voltage range and low input voltage support for energy-harvesting applications, among others.

Moreover, as near- or sub-threshold operation gains popularity for its significantly improved system energy efficiency, the nature of such operation imposes further challenges. Due to PVT variations, as well as other fast-varying variations (supply voltage droop, coupling noise, etc.), robust near- or sub-threshold operation requires a prohibitive voltage margin. As a result, IoT nodes must support adaptive circuit techniques such as in-situ timing-error detection and correction [62], timing-error prediction [60, 61], and DVS capabilities in the PMU such that the IoT node's supply voltage can be tuned dynamically according to the current PVT variations [59, 60, 63].

Previous adaptive DVS works (Figure 5.1) either employ an off-chip regulator [59], resulting in a slow transient response, increased system form factor, and extra power

consumption for off-chip circuits, or utilize on-chip LDO [63] that has substantial power-conversion loss when the difference between input and output becomes large. Both factors prohibit their use in a sub-microwatt system that operates in the deep sub-threshold regime. Moreover, both of them require a voltage reference and comparators, which further increases quiescent power and lowers overall PCE.

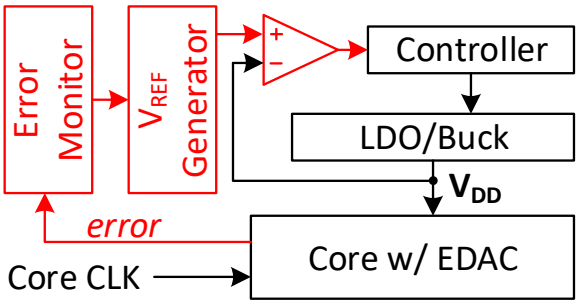



Figure 5.1 Previous adaptive DVS works either require off-chip components [59] or have an inefficient LDO [63]. Both of these require a voltage reference and a comparator.

Direct error regulation can be integrated with SC-DC to enable a fully integrated and fully digital PMU design [58]. In such a system, the timing error directly triggers switching of the SC-DC for fast droop response, and error statistics are used to scale the supply voltage ($V_{DD}$) up or down by changing the SC configuration. However, in this design, the loss of regulation due to inactivity or noncritical execution can cause critical failures. Also, the SC is designed to switch at a fixed frequency [58], so its efficiency is degraded when the load power varies substantially. This problem worsens significantly for IoT nodes, where the load power can vary by orders of magnitude (Chapter 4).

This work presents a PMU–load co-design that tackles the aforementioned challenges. The proposed PMU is designed based on a fully integrated SC-DC and integrated with a state-of-the-art Neural Spike Processor (NSP) performing motor intention decoding tasks. The NSP marks a record power efficiency of 0.61 µW for a 96-channel system

(6.35 nW/Channel), 27× better than the prior art [56]. Employing in-situ error detection and correction (EDAC), the NSP operates robustly in the deep sub-threshold ($V_{th}$) regime while the timing error is used by the PMU to modulate its output voltage for (1) robust operation across PVT variations and (2) high PCE at ultra-low load power dissipation through a hybrid error/replica-based, energy-robustness co-optimization control scheme.

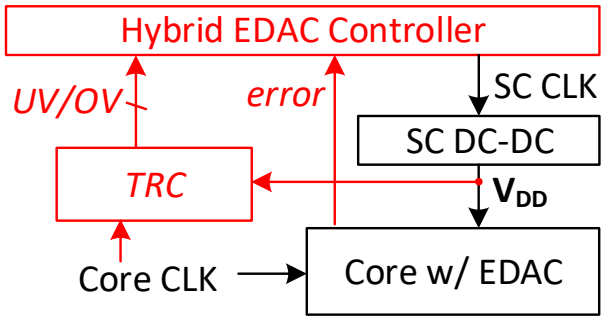



Figure 5.2 The proposed DVS architecture employs hybrid error/replica-based regulation integrated with an on-chip SC-DC to tackle the challenges in prior works.

Our hybrid error/replica-based controller (Figure 5.2) removes the need for voltage reference and comparator, making the control scheme fully digital and operable under low input voltage ($V_{IN}$), and allowing the SoC to be directly powered by capacitors and harvesters. This avoids efficiency degradation in conversions to/from battery level (~4 V) [57]. In the hybrid error/replica-based design, a tunable replica circuit (TRC) is added to assist the error regulation, preventing loss of regulation. Moreover, with automatic energy-robustness co-optimization, the PMU is able to set the optimal conversion ratio (*CR*) and switching frequency ($f_{SC}$) for the SC-DC. With the PMU achieving a PCE of 77.7% (72.2%), the NSP-PMU SoC consumes 0.77 µW (0.83 µW) at $V_{IN}$ of 0.6 V (1 V) at the margin-free operating point, respectively, marking a record-high power efficiency of 8.1 nW/Channel for the 96-channel system, 21× reduction from the prior art [56].

## 5.2 System Architecture and Implementations

### 5.2.1 Neural Spike Processor

Recent advances in neuroscience and integrated circuits offer possibilities for long-term brain-computer-interface (BCI) implants. Similar to IoT nodes, such implants have stringent power budgets as they are preferably powered with wireless energy-harvesting devices. However, with wireless communication often easily dominating overall power consumption, on-implant processing that reduces the wireless data rate substantially is highly desirable. In this work, an NSP that integrates spike detection, sorting and the first half of motor intention decoding is able to reduce the wireless data rate by more than four orders of magnitude (Figure 5.3).

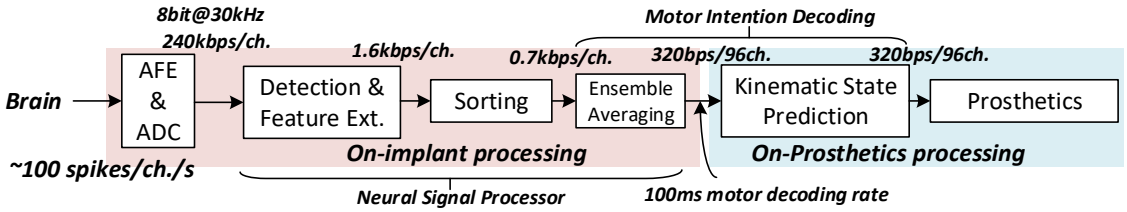



Figure 5.3 The proposed NSP with a high level of integration reduces the wireless data rate by more than four orders of magnitude.

Figure 5.4 shows the architecture of the NSP–PMU SoC. The NSP starts with 96 threshold-crossing spike detectors whose outputs feed three sorters via three 32-to-1 priority-encoded MUXs. The three sorter outputs then merge into the decoder via a queue. The sorter adopts our 1.5D Bayesian boundary sorting algorithm, where the decision is made based on partitions defined by orthogonal boundaries in the two-dimensional space of features (max and min values of a spike waveform) (Figure 5.5). This algorithm requires one to two orders of magnitude× less computation and achieves a comparable or better accuracy than the conventional distance-based algorithms for many datasets [54].

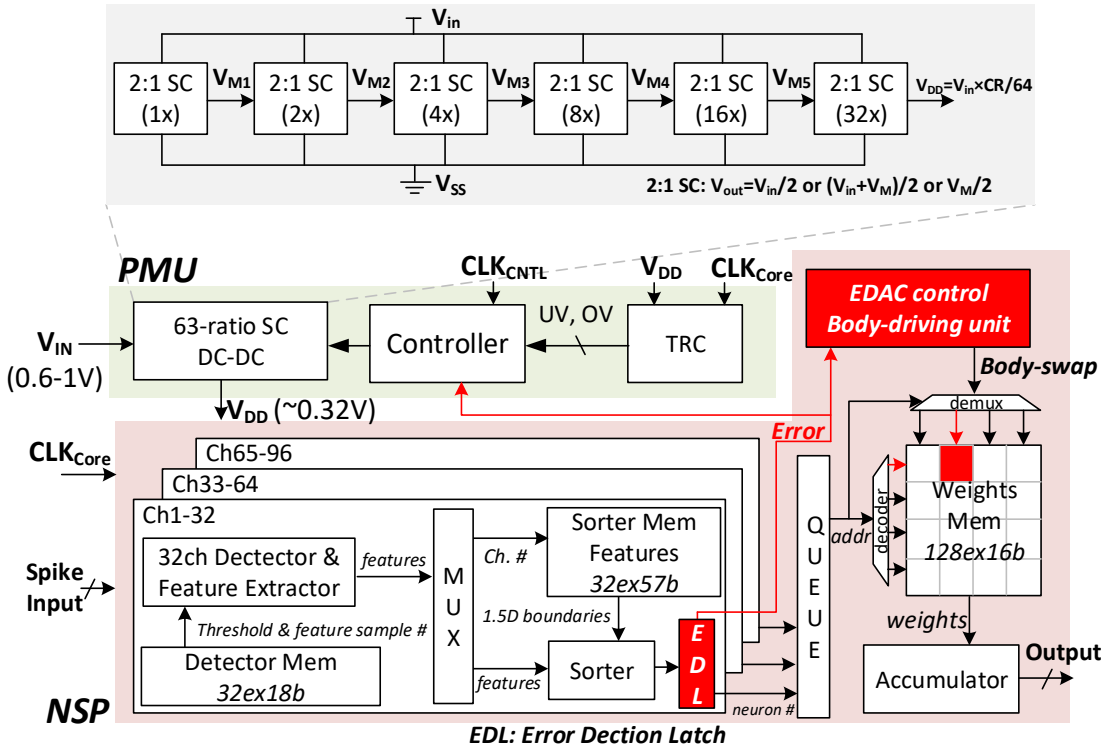



Figure 5.4 Schematic of the proposed NSP–PMU SoC.

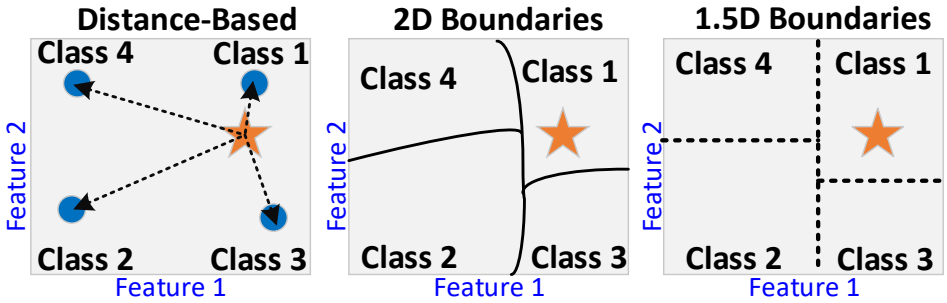



Figure 5.5 The proposed 1.5D-boundary-based sorting.

For decoding, we adopt our ensemble observation Kalman filter (EOKF) [54], which uses *regressed* spiking rates as state observation in a Kalman filter (KF) and saves 400× computation versus the standard KF (Table 5.1). For the DREAM dataset, the EOKF achieves better accuracy than the standard KF especially in the higher velocity regime (Figure 5.6) [54]. We mapped these algorithms to deep sub-$V_{th}$ circuits in a 0.18 µm that is

carefully selected for low leakage. The NSP marks a record-low power dissipation of 0.61 µW for a 96-channel system (6.35 nW/Ch.), a 27× reduction from the prior art [56].

| KF Operations | # of Calculation (Mult./Add/Div.) | |
|---|---|---|
| | Standard KF | EOKF |
| A Priori Estimate | 4/2/~ | 4/2/~ |
| Posterior Estimate | 80/80/~ | 46/46/~ |
| Kalman Gain | 32180/32060/1180 | 10/9/4 |
| Post. Error Cov. Matrix | 96/92/~ | 16/16/~ |
| Total | 32360/32234/1180 | 76/73/4 (***~400x less***) |

Table 5.1 EOKF reduces computational complexity by 400×.

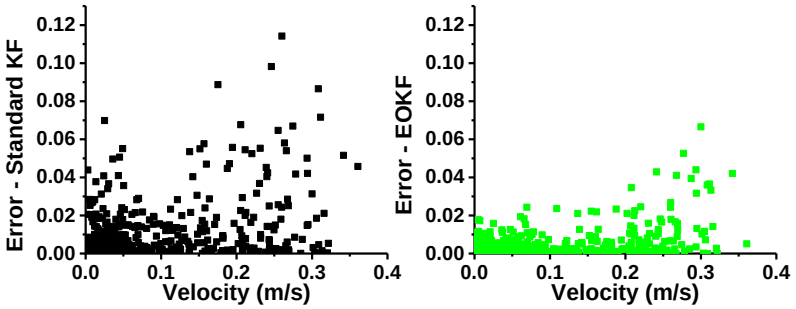



Figure 5.6 Performance comparison between standard KF and EOKF.

## 5.2.2 PMU and NSP Co-design

The PMU and NSP are co-designed to support two features: (1) modulating NSP supply voltage ($V_{DD}$) across PVT variations to remove the prohibitive safety margin for deep sub-$V_{th}$ circuits and (2) optimizing its PCE by automatically finding an optimal configuration for the DC-DC converter.

To enable the first feature, we propose *hybrid error/replica-based regulation*. This scheme uses *in-situ* EDAC [58] embedded in the NSP, which *directly* regulates $V_{DD}$ by controlling a 63-ratio configurable SC-DC. We added the EDAC capability to the NSP by leveraging recent advances [62]. First, following the sparse insertion scheme, we inserted error detection latches (EDL) only between the sorters and the queue (Figure 5.4; highlighted in red). Second, we employed body-swapping-based error correction in the

weight memory. To reduce the power overhead of body swapping, we designed it to swap only the body of the accessed memory bank (out of 16).

Figure 5.7 showcases the process of the error/replica-based regulation. Upon each timing error detected, the PMU controller sets the SC-DC to immediately start a new switching phase so as to recover its output voltage. Moreover, the $f_{SC}$ of SC-DC is instantly boosted for the rest of the core clock cycle so as to supply the extra power for the body-swapping error correction. If errors continue to occur, the controller reduces the SC-DC *CR*, thus raising $V_{DD}$.

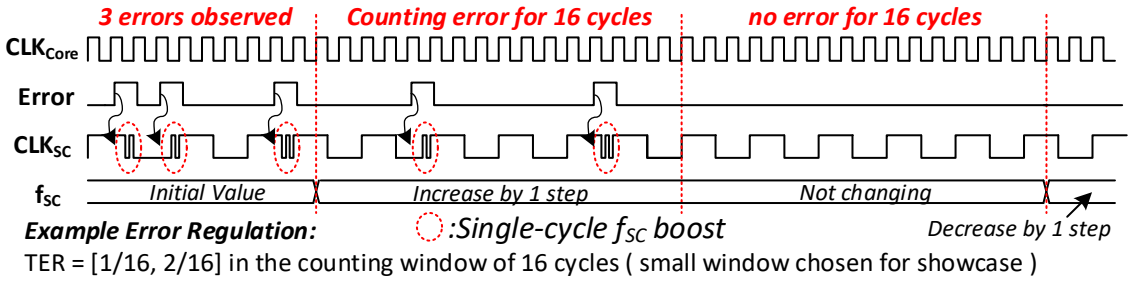



Figure 5.7 Showcase waveforms of the error-based regulation process.

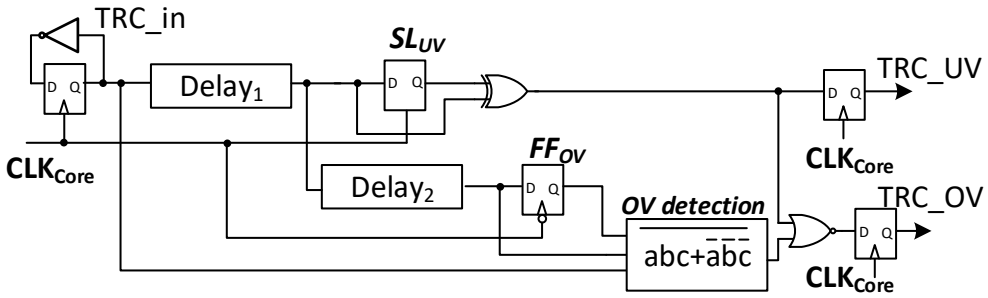



Figure 5.8 The tunable replica circuit schematic.

We added a TRC to enable *continuous* error-based regulation (Figure 5.8). The existing error-based works [58, 59] lose regulation if the critical paths are not executed due to e.g., input inactivity or $V_{DD}$ overshoot. The TRC sets the upper and lower $V_{DD}$ bounds, avoiding the issue. Since we pipelined the NSP with two-phase latches, the critical path per latch stage is $0.5 \cdot T_{CLK}$. In the TRC, $Delay_1$ is tuned to $0.4 \cdot T_{CLK}$ such that a timing violation is

detected by $SL_{UV}$ slightly before the actual critical path in NSP fails. On the other hand, if the NSP critical path becomes much shorter than $0.5 \cdot T_{CLK}$ under the $V_{DD}$ overshoot case, TRC will assert TRC_OV through the OV detection logic when $(Delay_1 + Delay_2) < 0.5 \cdot T_{CLK}$.

However, false OV detection can happen when $2 \cdot T_{CLK} < (Delay_1 + Delay_2) < 2.5 \cdot T_{CLK}$. This limits the maximal value of $Delay_2$. As a result, we chose $Delay_2$ as $0.8 \cdot T_{CLK}$ and added a NOR gate after the OV detection logic to mask TRC_OV with TRC_UV. In such cases, when a false OV detection happens, $Delay_1$ will reside in $[0.67,0.83] \cdot T_{CLK}$ and TRC_UV will be asserted. Note that TRC does not directly affect exact $V_{DD}$ settings, thus adding no margin.

To enable the second feature in the PMU, the automatic PCE optimization across operating conditions, we devised the $CR/f_{SC}$ search scheme in the context of the error-based regulation. It finds $f_{SC}$ and $CR$ of the SC-DC for the optimal PCE. Prior works on error-based regulation paid little attention to such a scheme [58, 59], while voltage look-up tables are often used in some of the conventional voltage-based regulations [48, 60].

The scheme is devised to first scale $V_{DD}$ to the point of the first failure (PoFF, $V_{DDmin}$) using EDAC. This minimizes the NSP's power dissipation since it operates at a fixed clock frequency (30 kHz). Then, it finds a PCE-optimal $CR/f_{SC}$ setting by setting a proper value of $\Delta V = V_{DD,OC} - V_{DD}$ ($V_{DD,OC}$ as the open-circuit SC-DC output voltage) which balances the SC-DC's switching and linear losses.

In more general application cases, where the clock frequency can vary, the same control scheme can still be applied. In these cases, the clock frequency can be determined either by the system performance requirement or by the energy management systems that sense

the energy availability. After that, the proposed control scheme can then be triggered and scale $V_{DD}$ to PoFF so that the power consumption of the load is minimized.

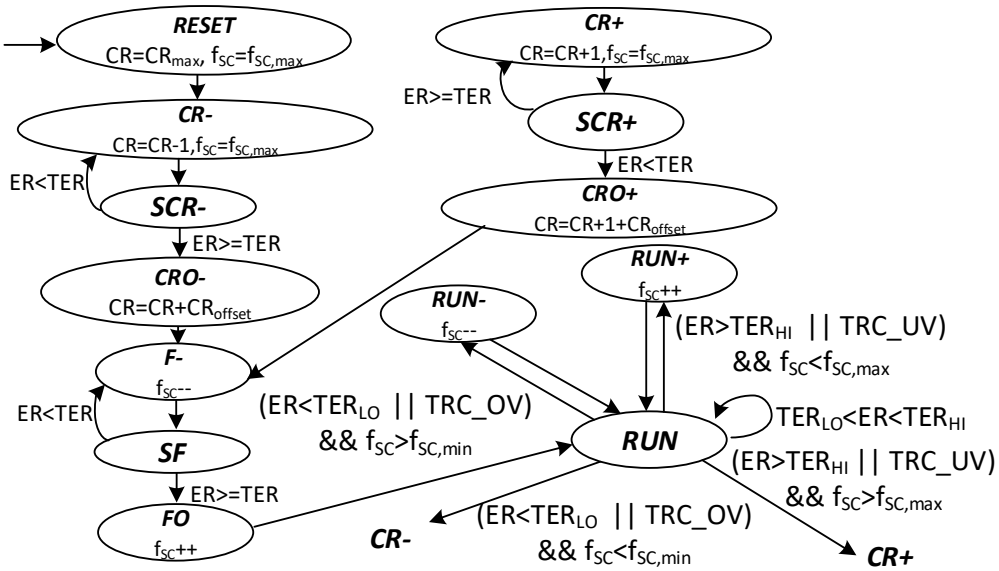



Figure 5.9 The state diagram of the proposed controls: the $CR/f_{SC}$ search and the error-based regulation.

Figure 5.9 shows the state diagram of the $CR/f_{SC}$ search.

First, the controller finds $V_{DDmin}$ by reducing $CR$ to the minimum ($CR_{min}$) where the error rate (ER) reaches a target error rate (TER), using the highest allowed $f_{SC}$ that minimizes $\Delta V$ (***CR-*** and ***SCR-)***.

Then, $CR$ is set to an optimal value: $CR_{opt} = CR_{min} + CR_{offset}$ at ***CRO-***. $CR_{offset}$ is proportional to the optimal $\Delta V$, which is a function of SC-DC design parameters and $V_{IN}$, but *insensitive to* load conditions. Setting $CR_{opt}$ raises $V_{DD}$, largely reducing the ER.

After that, the optimal $f_{SC}$ is found by reducing $f_{SC}$ until ER reaches TER again, at which point $\Delta V$ also reaches its optimal value. The system then goes into the aforementioned error/replica-based regulation (***RUN, RUN+,*** *and* ***RUN-***).

During the regulation, if $f_{SC}$ reaches predefined upper/lower bounds ($f_{SC,max}/f_{SC,min;}$ implying switching or linear loss becomes dominant), the controller reinitiates the above

*CR*/$f_{SC}$ search (***CR+*** or ***CR-***). ***CR+, SCR+,*** and ***CRO+*** states are for the *CR* search if the optimal *CR* is higher than the current *CR*. Measured example waveforms of the control scheme are shown in Figure 5.10, confirming its correct operation.

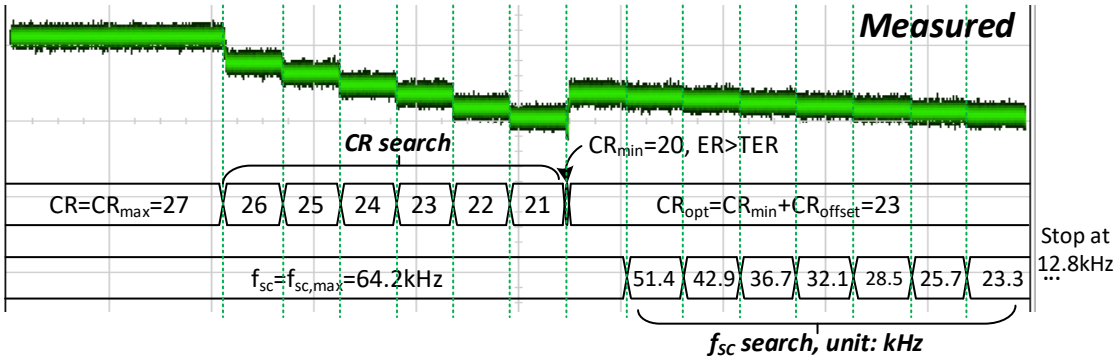



Figure 5.10 Waveforms of the *CR*/$f_{SC}$ search process.

We designed the fully digital controller performing the error/replica-based regulation and the *CR*/$f_{SC}$ search to avoid variable $V_{REF}$ generators and voltage comparators. These would impose power and delay overhead and also $V_{IN}$ scaling limits on the PMU. This allows the SoC to be powered directly by capacitors and energy harvesters and avoids the efficiency degradation in conversions to and from battery level (4 V) [57].

## 5.3 Measurement Results

We prototyped the SoC in a 0.18 µm process (Figure 5.11), which is carefully selected for low leakage, as the performance requirement of the system (both NSP and PMU) is relatively low. The NSP operates at the target frequency of 30 kHz at $V_{DD}$ of 0.32 V with 0.1% TER while consuming only 0.61 µW (Figure 5.11).

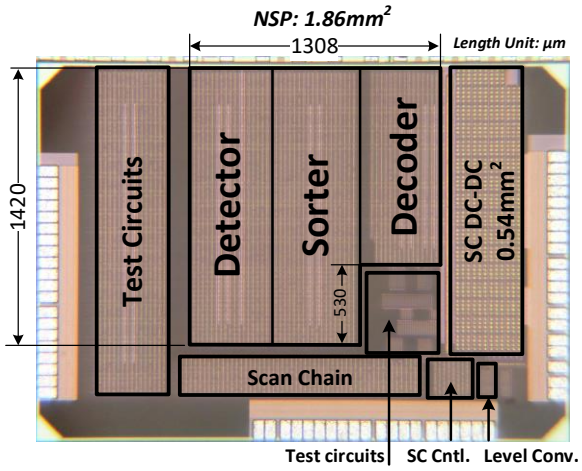



Figure 5.11 Chip Die Photo.

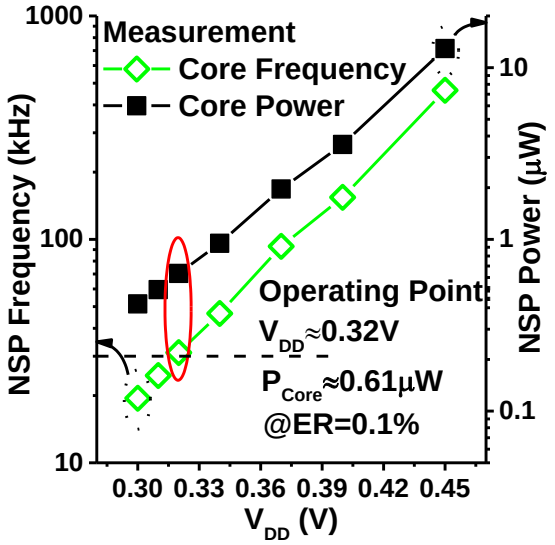



Figure 5.12 Power and performance of the NSP.

We also tested the PMU's full functionality by varying its output in a fine-grained manner and changing the *CR* (Figure 5.13(a)) at high PCEs (Figure 5.13(b)).

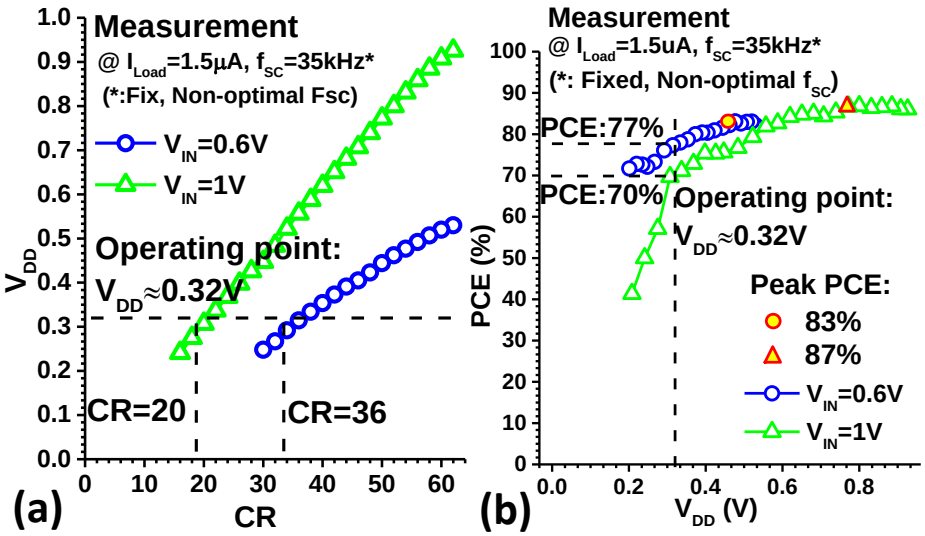



Figure 5.13 Performance of the integrated SC-DC. (a) Desirable $V_{DD}$ scaling behavior and (b) high PCE.

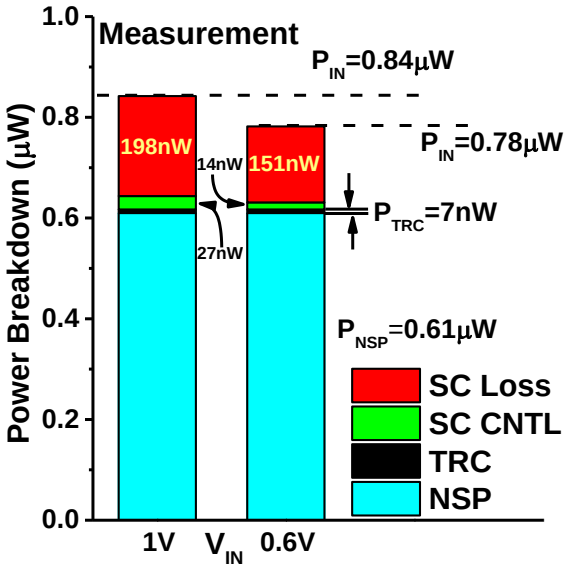



Figure 5.14 NSP-PMU SoC power breakdown.

The SoC consumes 0.78 µW at the setting found automatically via the proposed control scheme. Of this, the NSP consumes 0.61 µW and TRC uses 7 nW. The SC-DC losses (SC Loss) are 151 nW at 0.6 V $V_{IN}$ and 198 nW at 1 V. The PMU controller (SC CNTL) consumes 14 nW at 0.6 V $V_{IN}$ and 27 nW at 1 V (Figure 5.14).

We also validated the proposed $CR/f_{SC}$ search. We first validated that $CR_{offset}$ remains two to three robustly across ranges of $CR$ settings and load current levels (Figure 5.15).

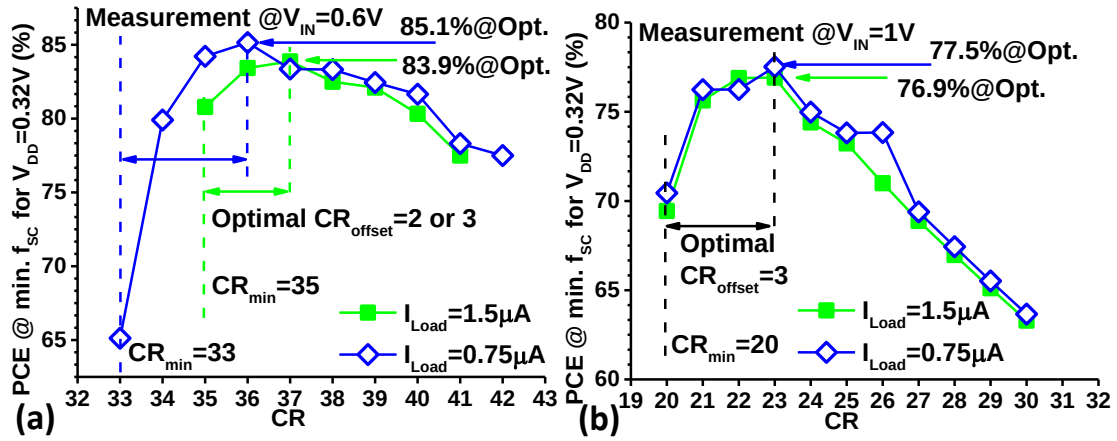



Figure 5.15 The measurements validate the optimal $CR_{offset}$ robustness across load and $V_{IN}$ conditions.

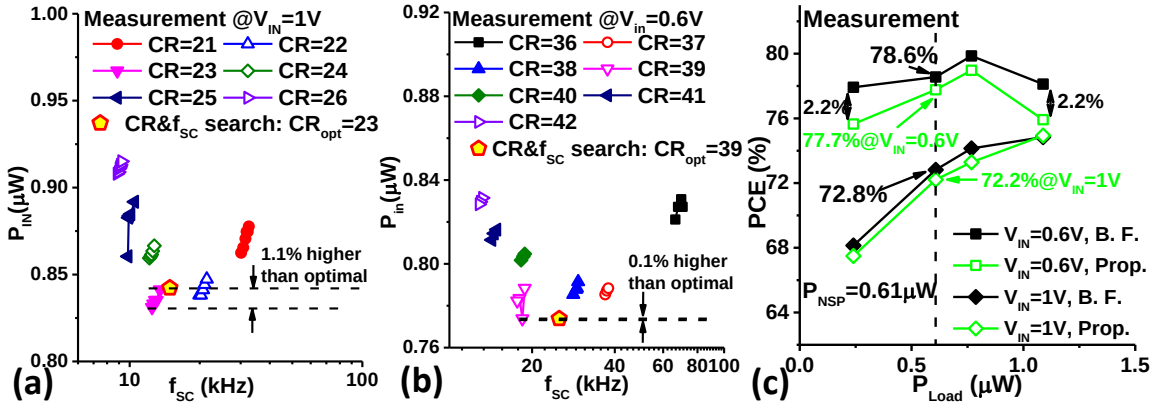



Figure 5.16 $P_{IN}$ comparison between the brute-force search and the proposed $CR/f_{SC}$ search at (a) $V_{IN}$ = 0.6 V and (b) $V_{IN}$ = 1 V. (c) PCE comparison between the brute-force search and the proposed $CR/f_{SC}$ search for the wider load power range.

Using $CR_{offset}$ = 3 and for 1 V $V_{IN}$, our scheme found $f_{SC}$ = 14.9 kHz and $CR$ = 23, at which the SoC draws $P_{IN}$ = 0.84 µW (Figure 5.16(a)). For 0.6 V $V_{IN}$, it found $f_{SC}$ = 18.4 kHz and $CR$ = 39, at which the SoC consumes 0.77 µW (Figure 5.16(b)). We then swept all the possible $CR$s and $f_{SC}$s, and found that the SoC consumes the optimal $P_{IN}$ of 0.83 µW at $f_{SC}$ = 12.9 kHz, $CR$ = 23 and $V_{IN}$ = 1 V, and $P_{IN}$ = 0.77 µW at $f_{SC}$ = 18.4 kHz, $CR$ = 39 and $V_{IN}$ = 0.6 V. Our search achieves PCEs within 1% of the brute-force PCE (Figure 5.16(a, b)). We further evaluated the efficacy of the scheme across a wider load-power range. It still achieves at most 2.2% worse PCE than the brute-force search (Figure 5.16(c)).

| | This work | [55] | [56] |
|---|---|---|---|
| **Process (nm)** | 180 | 65 | 65 |
| **No. of Channels** | 96 | 96 | 128 |
| **Detection & Sorting** | Y | Y | Y |
| **Partial Decoding** | **Y** | N | N |
| **Integrated PMU** | **Y** | N | N |
| **Core $V_{DD}$ (V)** | **0.32** | 0.6 | 0.54 |
| **Core Power/Ch. (nW)** | **6.3** | 1740 | 175 |
| **Core Area/Ch. ($mm^2$)** | 0.0194 | 0.12 | **0.003** |

Table 5.2 Comparison to the prior BCI processors.

| | This work | [58] | [60] | [48] |
|---|---|---|---|---|
| **Process (nm)** | 180 | 65 | 65 | 180 |
| **Converter Type** | 63-ratio SC | 63-ratio SC | LDO+2-ratio SC | Multi-SCs |
| **$P_{Load}$ Range (W)** | <1µ | ~25µ | 8-173m | 20n-500µ |
| **$V_{IN}$ (V)** | **0.6-1** | **0.6-1** | 0.65-1.05 | 1-4 |
| **Core $V_{DD}$ (V)** | ~0.32 | ~0.45 | 0.38-0.92 | 0.6/1.2/3.3 |
| **PVT Adaptive Scheme** | **EDAC** | **EDAC** | Replica | N.A. |
| **PVT margin** | **None** | **None** | YES | N.A. |
| **Regulation Method** | **Error/Replica** | Error | $V_{REF}$+CMP | $V_{REF}$+CMP |
| **Continuous Regulation** | **Yes** | No | Yes | Yes |
| **PCE optimazation** | **Error-based Search** | Fixed $f_{SC}$ | $V_{REF}$ Look-up Table | $V_{REF}$ look-up Table |
| **PCE Range** | 72-77% | 73-87% | ~52-73%(SC) | 60-68% |

Table 5.3 Comparison to the prior PMU–load co-designs.

Table 5.2 compares to the state-of-the-art BCI processors [55], [56]. The proposed SoC draws 21× less power of [56] even after including the PMU overhead. It also demonstrates the first end-to-end NSP integration at comparable or better accuracy over prior arts. As compared to the prior PMU–load co-designs (Table 5.3), our design demonstrates *continuous* regulation and optimal PCE search in error-based regulation, which has an advantage over voltage-based regulation designs [48], [60] in terms of PMU $V_{IN}$ scalability, fully digital control, and safety-margin reduction.

# Chapter 6 Conclusions

One of the key challenges in realizing the vision of IoT is security and privacy. The enormous amount of information transferred in the network between billions or trillions of IoT devices imposes harsh security requirements on the communication. PUF serves as an important primitive in circuit design for hardware security, as it provides an object-specific identifier that is hard to expose or impersonate for malicious adversaries. Such identifiers are critical in identification and authentication tasks, which are among the most fundamental requirements of secure IoT systems. However, building compact yet robust PUF circuits remains a challenge as there exists a direct tradeoff between these two metrics in a PUF design.

Chapter 2 presents a novel compact PUF design based on a pair of ultra-compact PTAT voltage generators. Measurements from chip prototypes demonstrate substantial improvement in the robustness–compactness tradeoff over the prior art. A similar idea is employed in Chapter 3, where instead of using dedicated circuits, we transform a preexisting on-chip SRAM into an array of PUFs. By reusing existing hardware resources while adding minimal peripheral circuits, we are able to further reduce the per-bit area of a PUF substantially without hurting its robustness.

Another important aspect in IoT systems is power management circuits. On one hand, energy harvesting is highly preferred for low-power IoT nodes as battery replacement for those nodes is cost prohibitive and impractical. However, previous EH PMU architectures

suffer from efficiency loss as most of the energy goes through two voltage conversions (harvester-to-battery and battery-to-load), resulting in substantial energy loss. We present an EH PMU in Chapter 4 that introduces a hybrid storage architecture using a capacitor as intermediate energy buffer and minimizing the amount of energy going through the two voltage conversions. We also analyze the design trade-offs between the capacitor size, load and harvester power profiles, and the efficiency improvement from the proposed architecture.

On the other hand, with the drastically reduced power consumption of IoT nodes, there come new challenges in PMU design: Adaptive DVS, fully on-chip integration and high PCE are required. In Chapter 5, we present a PMU–load co-design based on fully integrated SC-DC and hybrid error/replica-based regulation for fully digital PMU control. We also devise a control scheme to optimize both the $V_{DD}$ margin and the PMU PCE by automatically searching for the SC-DC's optimal configuration and switching frequency, at which point the loss components are balanced and the optimal PCE is achieved.

There is more to be studied for the research fields covered in this thesis. For PUFs, the recent art shows that although the robustness and compactness have substantially improved over the last decade, there is still much work to be done. For weak PUFs such as we are working on, the major challenge is that its bit instability can never be reduced to be close enough to zero. As a result, extra information must be stored on-chip to assist reliable PUF evaluation, such as error correction codes or bit mask. This allows various possible attacks, especially fault injection attack by manipulating the bit mask stored in NVM. Thus, it is yet to be explored on the architecture level how to minimize the amount of stored information. Also, it is noteworthy that multiple strong PUF structures use a circuit

topology similar to that used in weak PUFs. Thus, it is worth exploring if the compact voltage-generator-pair structure can also be used for strong PUF design.

For the EH PMU, the improvement of the proposed architecture in the context of a real-life harvester/load power consumption profile is yet to be investigated. Additionally, the possibility of implementing a single SC topology to transfer power between multiple voltage domains, instead of using multiple converters is worth investigating.

Lastly, for PMU–load co-design, one of the major challenges for further investigation when applying such adaptive design in a commercial product is timing closure. Specifically, when the supply voltage is designed to adapt to the critical path, preventing hold-time violations on all possible timing arcs becomes a critical issue.

# Bibliography

[1] Ala Al-Fuqaha, Mohsen Guizani, Mehdi Mohammadi, Mohammed Aledhari, and Moussa Ayyash, "Internet of things: A survey on enabling technologies, protocols, and applications," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376, 2015.

[2] Kaiyuan Yang, David Blaauw, Dennis Sylvester, "Hardware Designs for Security in Ultra-Low-Power IoT Systems: An Overview and Survey," *IEEE Micro*, vol. 37, no. 6, pp. 72–89, 2017.

[3] L. Atzori, A. Iera, G. Morabito, "The Internet of Things: A survey," *Computer Networks*, vol. 54, no. 15, pp. 2787–2805, 2010.

[4] Roel Maes, "Physically Unclonable Functions: Constructions, Properties and Applications," Dissertation, Katholieke Universiteit Leuven, 2012.

[5] Nathan Beckmann and Miodrag Potkonjak, "Hardware-Based Public-Key Cryptography with Public Physically Unclonable Functions," *Information Hiding* 2009, pp. 206–220, 2009.

[6] Joonho Kong, Farinaz Koushanfar, Praveen K. Pendyala, Ahmad-Reza Sadeghi, and Christian Wachsmann, "PUFatt: Embedded Platform Attestation Based on Novel Processor-Based PUFs," *ACM/EDAC/IEEE Design Automation Conference*, pp. 1–6, 2014.

[7] Masoud Rostami, Mehrdad Majzoobi, Farinaz Koushanfar, Dan S. Wallach, and Srinivas Devadas, "Robust and Reverse-Engineering Resilient PUF Authentication and Key-Exchange by Substring Matching," *IEEE Transactions on Emerging Topics in Computing*, vol. 2, no. 1, pp. 37–49, 2014.

[8] Pappu, R. S., Recht, B., Taylor, J., and Gershenfeld, "Physical One-Way Functions," *Science*, vol. 297, no. 5589, pp. 2026–2030, Sep. 2002.

[9] Jae W. Lee, Daihyun Lim, Blaise Gassend, G. Edward Suh, Marten van Dijk, and Srinivas Devadas, "A Technique to Build a Secret Key in Integrated Circuits for Identification and Authentication Applications," *IEEE Symposium on VLSI Circuits*, pp. 176–179, 2004.

[10] G. Edward Suh and Srinivas Devadas, "Physical Unclonable Functions for Device Authentication and Secret Key Generation," *ACM annual Design Automation Conference*, pp. 9–14, 2007.

[11] Qingqing Chen, György Csaba, Paolo Lugli, Ulf Schlichtmann, Ulrich Rührmair, "The bistable ring PUF: A new architecture for strong physical unclonable functions." *IEEE International Symposium on Hardware-Oriented Security and Trust*, pp. 134–141, 2011.

[12] Kaiyuan Yang, Qing Dong, David Blaauw, Dennis Sylvester, "A Physically Unclonable Function with BER < 10–8 for Robust Chip Authentication Using Oscillator Collapse in 40 nm CMOS," *IEEE International Solid-State Circuits Conference*, pp. 254–255, 2015.

[13] Ulrich Rührmair, Jan Solter, Frank Sehnke, Xiaolin Xu, Ahmed Mahmoud, Vera Stoyanova, Gideon Dror, Jurgen Schmidhuber, Wayne Burleson, and Srinivas

Devadas, “PUF Modeling Attacks on Simulated and Silicon Data,” *IEEE Transaction on Information Forensics and Security*, vol. 8, no. 11, pp. 1876–1891, 2013.

[14] Daniel E. Holcomb, Wayne P. Burleson, and Kevin Fu, “Initial SRAM State as a Fingerprint and Source of True Random Numbers for RFID Tags,” *Proceedings of the Conference on RFID Security*, vol. 7, 2007.

[15] Roel Maes, Vladimir Rozic, Ingrid Verbauwhede, Patrick Koeberl, Erik van der Sluis, and Vincent van der Leest, “Experimental evaluation of Physically Unclonable Functions in 65 nm CMOS,” *IEEE European Solid-State Circuits Conference*, pp. 486–489, 2012.

[16] Sanu K. Mathew, Sudhir K. Satpathy, Mark A. Anders, Himanshu Kaul, Steven K. Hsu, Amit Agarwal, Gregory K. Chen, Rachael J. Parker, Ram K. Krishnamurthy, and Vivek De, “A 0.19 pJ/b PVT-Variation-Tolerant Hybrid Physically Unclonable Function Circuit for 100% Stable Secure Key Generation in 22 nm CMOS,” *IEEE International Solid-State Circuits Conference*, pp. 278–279, 2014.

[17] Ying Su, Jeremy Holleman, and Brian P. Otis, “A Digital 1.6 pJ/bit Chip Identification Circuit Using Process Variations,” *IEEE Journal of Solid-State Circuits*, vol. 43, no. 1, pp. 69–77, 2008.

[18] Anastacia Alvarez, Wenfeng Zhao, and Massimo Alioto, “15 fJ/b Static Physically Unclonable Functions for Secure Chip Identification with <2% Native Bit Instability and 140× Inter/Intra PUF Hamming Distance Separation in 65 nm,” *IEEE International Solid-State Circuits Conference*, pp. 256–257, 2015.

[19] Mudit Bhargava and Ken Mai, "A High Reliability PUF Using Hot Carrier Injection Based Response Reinforcement," *Cryptographic Hardware and Embedded Systems*, pp. 90–106, 2013.

[20] M. Seok, G. Kim, D. Blaauw, and D. Sylvester, "A Portable 2-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5 V," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 10, pp. 2534–2545, 2012.

[21] Blaise Gassend, Dwaine Clarke, Marten van Dijk, and Srinivas Devadas, "Controlled Physical Random Functions*," IEEE 18th annual Computer Security Applications Conference*, pp. 149–160, 2002.

[22] Mehrdad Majzoobi, Farinaz Koushanfar, and Miodrag Potkonjak, "Lightweight Secure PUFs," *IEEE/ACM 2008 International Conference on Computer-Aided Design*, pp. 670–673, 2008.

[23] Ahmed Mahmoud, Ulrich Rührmair, Mehrdad Majzoobi, and Farinaz Koushanfar, "Combined Modeling and Side Channel Attacks on Strong PUFs," *IACR Cryptology ePrint Archive*, https://eprint.iacr.org/2013/632, 2013.

[24] Lingkai Kong, Yue Lu, and Elad Alon, "A Multi-GHz Area-Efficient Comparator with Dynamic Offset Cancellation," *IEEE Custom Integrated Circuits Conference*, pp. 1–4, 2011.

[25] A. Rukhin, et al. "A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications," *National Institute of Standards and Technology*, vol. 800–22, no. Rev 1a, 2010.

[26] J. Li and M. Seok, “Robust and In-Situ Self-Testing Technique for Monitoring Device Aging Effects in Pipeline Circuits,” *ACM annual Design Automation Conference*, pp. 1–6, 2014.

[27] T. Yang, D. Kim, P. Kinget, M. Seok, “In-situ techniques for in-field sensing of NBTI degradation in an SRAM register file,” *IEEE International Solid-State Circuits Conference*, pp. 264–265, 2015.

[28] Clemens Helfmeier, Christian Boit, Dmitry Nedospasov, Shahin Tajik, and Jean-Pierre Seifert, “Physical vulnerabilities of Physically Unclonable Functions,” *ACM Design, Automation & Test in Europe*, no. 350, 2014.

[29] C. Helfmeier, C. Boit, D. Nedospasov, and J.-P. Seifert, “Cloning physically unclonable functions,” *IEEE International Symposium on Hardware-Oriented Security and Trust (HOST)*, pp. 1–6, 2013.

[30] “ISSCC 2016 TRENDS, Page 31, Fig. 5,” in *the Proceedings of IEEE International Solid-State Circuits Conference*, 2016, San Francisco, United States, 2016. Available: http://isscc.org/doc/2016/ISSCC2016_TechTrends.pdf

[31] Jiangyi Li and Mingoo Seok, “A 3.07 $\mu m^2$/bitcell Physically Unclonable Function with 3.5% and 1% Bit-Instability across 0 to 80 °C and 0.6 to 1.2 V in a 65 nm CMOS,” *IEEE Symposium on VLSI Circuits*, pp. 250–251, 2015.

[32] Teng Yang, et. al., “Register File Circuits and Post-Deployment Framework to Monitor Aging Effects in Field,” *IEEE European Solid-State Circuits Conference*, pp. 425–428, 2016.

[33] Jiangyi Li, Teng Yang, and Mingoo Seok, "A Technique to Transform 6T-SRAM Arrays Into Robust Analog PUF with Minimal Overhead," *IEEE International Symposium on Circuits and Systems (ISCAS),* Baltimore, United States, May 2017.

[34] I. Doms, P. Merken, R. Mertens, and C. Van Hoof, "Integrated Capacitive Power-Management Circuit for Thermal Harvesters with Output Power 10 to 1000 μW," *IEEE International Solid-State Circuits Conference*, pp. 300–301, 2009.

[35] Hanh-Phuc Le, Seth R. Sanders, and Elad Alon, "Design Techniques for Fully Integrated Switched-Capacitor DC-DC Converters," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 9, pp. 2120–2131, 2011.

[36] Saurav Bandyopadhyay and Anantha P. Chandrakasan, "Platform Architecture for Solar, Thermal, and Vibration Energy Combining With MPPT and Single Inductor," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 9, pp. 2199–2215, 2012.

[37] Suyoung Bang, Yoonmyung Lee, Inhee Lee, Yejoong Kim, Gyouho Kim, David Blaauw, and Dennis Sylvester, "A Fully Integrated Switched-Capacitor Based PMU with Adaptive Energy Harvesting Technique for Ultra-Low Power Sensing Applications," *IEEE Symposium on Circuits and Systems*, pp. 709–712, 2013.

[38] Matthew Fojtik, Daeyeon Kim, Gregory Chen, Yu-Shiang Lin, David Fick, Junsun Park, Mingoo Seok, Mao-Ter Chen, Zhiyoong Foo, David Blaauw, and Dennis Sylvester, "A Millimeter-Scale Energy-Autonomous Sensor System with Stacked Battery and Solar Cells," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 3, pp. 801–813, 2013.

[39] Wanyeong Jung, Sechang Oh, Suyoung Bang, Yoonmyung Lee, Zhiyoong Foo, Gyouho Kim, Yiqun Zhang, Dennis Sylvester, and David Blaauw, "An Ultra-Low

Power Fully Integrated Energy Harvester Based on Self-Oscillating Switched-Capacitor Voltage Doubler," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 12, pp. 2800–2811, 2014.

[40] David Bol, El Hafed Boufouss, Denis Flandre, and Julien De Vos, " A 0.48 mm2 5 µW–10 mW Indoor/Outdoor PV Energy-Harvesting Management Unit in a 65 nm SoC based on a Single Bidirectional Multi-Gain/Multi-Mode Switched-Cap Converter with Supercap Storage," *IEEE European Solid-State Circuits Conference*, pp. 241–244, 2015.

[41] Inhee Lee, Wootaek Lim, Alan Teran, Jamie Phillips, Dennis Sylvester, and David Blaauw, "A >78%-Efficient Light Harvester over 100 to 100 Klux with Reconfigurable PV-Cell Network and MPPT Circuit," *IEEE International Solid-State Circuits Conference*, pp. 370–371, 2016.

[42] Xiaosen Liu, and Edgar Sanchez-Sinencio, "An 86% Efficiency 12 µW Self-Sustaining PV Energy Harvesting System With Hysteresis Regulation and Time-Domain MPPT for IOT Smart Nodes," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 6, pp. 1424–1437, 2015.

[43] Dina El-Damak and Anantha P. Chandrakasan, "Solar Energy Harvesting System with Integrated Battery Management and Startup using Single Inductor and 3.2 nW Quiescent Power," *IEEE Symposium on VLSI Circuits*, pp. 280–281, 2015.

[44] Yildiz Sinangil, Sabrina M. Neuman, Mahmut E. Sinangil, Nathan Ickes, George Bezerra, Eric Lau, Jason E. Miller, Henry C. Hoffmann, Srini Devadas, and Anantha P. Chandraksan, "A self-aware processor SoC using energy monitors integrated into

power converters for self-adaptation," *IEEE Symposium on VLSI Circuits*, pp. 1–2, 2014.

[45] Po-Hung Chen, Koichi Ishida, Katsuyuki Ikeuchi, Xin Zhang, Kentaro Honda, Yasuyuki Okuma, Yoshikatsu Ryu, Makoto Takamiya, Takayasu Sakurai, "A 95 mV-Startup Step-Up Converter with VTH-Tuned Oscillator by Fixed-Charge Programming and Capacitor Pass-On Scheme," *IEEE International Solid-State Circuits Conference*, pp. 216–218, 2011.

[46] Massimo Alioto, Elio Consoli, and, Jan M. Rabaey. "'EChO' Reconfigurable Power Management Unit for Energy Reduction in Sleep-Active Transitions," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 8, pp. 1921–1932, 2013.

[47] Yasuyuki Okuma, Koichi Ishida, Yoshikatsu Ryu, Xin Zhang, Po-Hung Chen, Kazunori Watanabe, Makoto Takamiya, and Takayasu Sakurai, "0.5-V input digital LDO with 98.7% current efficiency and 2.7-µA quiescent current in 65 nm CMOS," *IEEE Custom Integrated Circuits Conference*, pp. 1–4, 2010.

[48] Wanyeong Jung, Junhua Gu, Paul D. Myers, Minseob Shim, Seokhyeon Jeong, Kaiyuan Yang, Myungjoon Choi, ZhiYoong Foo, Suyoung Bang, Sechang Oh, Dennis Sylvester, and David Blaauw, "A 60%-Efficiency 20 nW–500 µW Tri-Output Fully Integrated Power Management Unit with Environmental Adaptation and Load-Proportional Biasing for IoT Systems," *IEEE International Solid-State Circuits Conference*, pp. 154–155, 2016.

[49] Loai G. Salem and Patrick P. Mercier, "A 45-Ratio Recursively Sliced Series-Parallel Switched-Capacitor DC-DC Converter Achieving 86% Efficiency," *IEEE Custom Integrated Circuits Conference*, pp. 1–4, 2014.

[50] Vishay Semiconductors, "Silicon PIN photodiode," *BPW34 datasheet*, Rev. 2.1, Aug. 2011.

[51] Maria Gorlatova, Aya Wallwater, Gil Zussman, "Networking Low-Power Energy Harvesting Devices: Measurements and Algorithms," *IEEE Transaction of Mobile Computing*, vol. 12, no. 9, pp. 1853–1865, 2013.

[52] Yejoong Kim, Yoonmyung Lee, Dennis Sylvester, and David Blaauw, "SLC: Split-Control Level Converter for Dense and Stable Wide-Range Voltage Conversion," *IEEE European Solid-State Circuits Conference*, pp. 478–481, 2012.

[53] Jiangyi Li, Pavan K. Chundi, et al., "A 0.77-µW 96-Ch. Deep Sub-Vt Neural Spike Processor Integrated with a Nanowatt Power Management Unit," submitted to *IEEE Symposium of VLSI Circuits*, 2018.

[54] Zhewei Jiang et al., "Microwatt End-to-End Digital Neural Signal Processing Systems for Motor Intention Decoding," *ACM Design, Automation & Test in Europe*, 2017.

[55] Zhewei Jiang et al., "1.74-µW/ch, 95.3%-accurate Spike-Sorting Hardware based on Bayesian Decision," *IEEE Symposium on VLSI Circuits*, 2016.

[56] S. M. A. Zeinolabedin et al., "A 128-Channel Spike Sorting Processor Featuring 0.175 µW and 0.0033 $mm^2$ per channel in 65-nm CMOS," *IEEE Symposium of VLSI Circuits*, 2016.

[57] Jiangyi Li et al., "Triple-Mode, Hybrid-Storage Energy Harvesting Power Management Unit: Achieving High Efficiency against Harvesting and Load Variabilities," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 10, pp. 2550–2562, 2017.

[58] SeongJong Kim, Mingoo Seok, "Ultra-Low-Power and Robust Power-Management/Microprocessor System Using Digital Error-based Regulation," *IEEE European Solid-State Circuits Conference*, 2017.

[59] David Bull, et al., "A Power-Efficient 32 bit ARM Processor Using Timing-Error Detection and Correction for Transient-Error Tolerance and Adaptation to PVT Variation," *IEEE Journal of Sold-State Circuits*, vol. 46, no. 1, pp. 18–31, 2011.

[60] Stepehn T. Kim, Yi-Chun Shih, et al., "Enabling Wide Autonomous DVFS in a 22 nm Graphics Execution Core Using a Digitally Controlled Hybrid LDO/Switched-Capacitor VR with Fast Droop Mitigation," *IEEE International Solid-State Circuits Conference*, 2015.

[61] Weiwei Shan, Xinchao Shang, Longxing Shi, Wentao Dai, and Jun Yang, "Timing error prediction AVFS with detection window tuning for wide-operating-range ICs," *IEEE Transaction on Circuit and Systems II: Brief Paper*, (Early access) 2017.

[62] Seongjong Kim, Joao Pedro Cerqueira, and Mingoo Seok, "A 450 mV timing-margin-free waveform sorter based on body swapping error correction," *IEEE Symposium of VLSI Circuits*, 2016.

[63] K. Hirairi, et al., "13% Power Reduction in 16 b Integer Unit in 40 nm CMOS by Adaptive Power Supply Voltage Control with Parity-Based Error Prediction and Detection (PEPD) and Fully Integrated Digital LDO," *IEEE International Solid-State Circuits Conference*, pp. 486–488, 2012.

[64] J. Manyika et al., "Disruptive Technologies: Advances that Will Transform Life, Business, and the Global Economy," San Francisco, CA, USA: McKinsey Global Institute, 2013.

[65] J. Gantz and D. Reinsel, “The digital universe in 2020: Big data, bigger digital shadows, and biggest growth in the far east,” *IDC iView: IDC Analyze the Future*, vol. 2007, pp. 1–16, 2012.

[66] Ravikanth Pappu, Ben Recht, Jason Taylor, and Neil Gershenfeld, “Physical one-way functions,” *Science*, vol. 297, no. 5589, pp. 2026–2030, 2002.