In-situ and **In-field** temperature and transistor BTI sensing techniques with microprocessor-level implementation

Teng Yang

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
under the Executive Committee
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2022

ABSTRACT

In-situ and **In-field** temperature and transistor BTI sensing techniques with microprocessor-level implementation

Teng Yang

In modern deep-scaled CMOS technologies, various silicon-related pitfalls present challenges to the long-term performance of microprocessors. Such challenges include (1) local hot spots, which breach the thermal limitations of a microprocessor, and (2) transistor aging, especially NBTI, which degrades transistor threshold voltage, ultimately threatening the reliability of the entire memory block. In previous systems, the dummy circuit was placed next to the subject, where the dummy was frequently analyzed, and the readout was used to infer the condition of the target. Due to rapidly changing ambient conditions (e.g., temperature and voltage) and the potential scale of the target dimensions, such metrics may not accurately represent the condition of the target. Moreover, such temperature sensors and canary circuits occupy significant area.

Therefore, it would be highly preferable to monitor the target circuit in-situ, i.e., to sense the precise transistor at operation. It is also important to achieve an accurate sensing metric. When the temperature is analyzed, the readout should account for voltage and process variations. While sensing the aging degradation, the readout should account for voltage and temperature fluctuations. This would allow testing during in-field operation, while the circuits achieve area-efficiency.

This research had two stages. One result of the first stage was a silicon test chip that was a compact temperature sensor. It involved a family of PTAT+CTAT sensor front-ends that unitized only 6 to 8 conventional CMOS logic devices, yielding a smaller sized chip.

The sensor demonstrates accuracy within the target and achieves a 14.3x smaller foot print than preceding published designs. The second product of the first stage was a PMOS aging sensor used in 6T SRAM circuits. The test chip has a real SRAM array, integrated with the proposed PMOS NBTI sensor. It can sense real PMOS NBTI effects in any bit cell (in-situ) and provide robust readings of temperature and voltage (in-field). Intensive aging tests validated the proposed sensing technique.

The second stage was focused on implementing the in-situ and in-field sensing techniques in a real processor. The MIPS microprocessor had a modified instruction cache (I$) and instruction set architecture. With the addition of new instruction aging sensing and minor modification of the circuits, the processor can execute aging sensing opportunistically to evaluate the aging level of its instruction cache. A software framework was developed and verified to estimate the retention voltage of the instruction cache over the lifetime of the chip.

An area-efficient SoC was developed that could transform the instruction cache to an ambient temperature sensor. It had a physically unclonable function (PUF), and it was built with an area-saving technique similar to the earlier work.

This thesis has four chapters. They are presented in the chronological and they are aligned with the research described above.

Table of Contents

List of Figures

List of Tables

Acknowledgments

As I recall seven years of life at Columbia University in New York City, a few vivid memories come to mind. To me, they are important, wonderful, and will never be forgotten.

After completing a bachelor's degree at BUPT, I came to Columbia University to pursue an EE master's degree in 2010. In this fantastic city, I found my true passion and future career path. The process was arduous and not straightforward. Unlike others, I chose to study a variety of fields in the first semester; I took courses in the fields of EE, BME, and CS, and then I changed to neuroscience. I started a research project with Prof. A.A. Lazar and studied related courses.

However, I ultimately realized that I was not sufficiently motivated, as the project required a greater understanding of mathematics and theory. I am more excited about practical and experimental subjects, such as VLSI design. I had some basic knowledge and skills accumulated in the VLSI field, but I hesitated as it was late for a master's student to switch the panel during the last semester. I talked to my father about my thoughts, and he said that he would support my decision unconditionally, even if I were to re-apply for another EE degree program in the VLSI track. My roommate and schoolmate, Jianxun Zhu, who studies radiofrequency circuit design, introduced me to his research field and encouraged me to talk to Prof. Kinget about my dilemma. Prof. Kinget nicely provided his suggestions for my next steps, approved my request to extend my master's degree study, and received me as a research student. Fortunately, with my determination and assistance from my mentor, Karthik Tripurary, and Prof. Kinget, I joined Prof. Mingoo Seok's group as a Ph.D. candidate to start the next page of my study.

No words can describe my sincere gratitude to Prof. Kinget, Jianxun Zhu, my father, and Karthik Tripurary for your kind help, which lights my future path.

During the 3rd year of my Ph.D. research, we enhanced the previous NBTI sensor to a general aging sensor and investigated a more delicate circuit to embed into real SRAM. On the technical side, it went smoother than the previous year, as I accumulated knowledge and skills. With the improved structure and the promising measurement results, I expected another fruitful year; however, my papers were all rejected. Because of the readers' critical feedback, I felt distraught, and I doubted whether the research topic I was pursuing followed industry trends and met specific requirements. That was the darkest period during my Ph.D. Imagine a person who had an encouraging start with ISSCC and JSSC publications, only to lose motivation in the third year. Mingoo detected my frustration and offered help in many ways. On the research side, we went through readers' comments carefully, modified the flows of the experiment accordingly, and demonstrated results systematically. Further, he advised me to investigate the implementation of the technique at the block or system level. On the personal side, he continued to encourage me to persist with my research goal. He also brought me opportunities to interact with peers and referred me to IBM Research for a summer internship. There, I was able to learn about industry trends and talk to experts. Gradually, life became better. We revised our papers, and they were accepted. In the following year, we build a real pipeline processor with aging sensors integrated. Eventually, we proudly demonstrated our novel processor in the CICC. I am highly confident and satisfied with my latest silicon chip from both the technical and spiritual perspectives.

I have received a great deal of support and assistance throughout my Ph.D. study. Without you, I would not have become an independent researcher and understood the power of

critical reasoning.

First, I would like to express my sincere gratitude to Professor Mingoo Seok, the principal advisor throughout my Ph.D. study. You offered valuable technical guidance for the research, and you provided considerable spiritual mentoring. I would not have been motivated to finish the degree without your help and encouragement. Your rich knowledge and charming personality inspired me, and it will continue to affect the rest of my life. I also want to extend my thanks to Professor Peter Kinget, my co-advisor. You kindly accepted me as your research student and brought me into the world of VLSI. You also generously provided technical and life advice in the past few years.

Second, I would like to thank other committee members, Professor Martha Kim, Professor Harish Krishnaswamy, and Professor Ioannis Kymissis. Thank you all for your commitment and valuable time to help me toward graduating.

Third, I am grateful to my group colleagues, Seongjiong Kim, Doyun Kim, Jiangyi Li, and Joao Pedro Cerqueira. Thank you all for assisting with my research, especially during the busy tape-out season. I am grateful to CISL peers from neighbor research groups, Jianxun Zhu, Yang Xu, and Ning Guo, for their generous help with CAD tools and chip integration.

Finally, I express my profound thanks to my family for their continuous love and support. I am grateful to my parents for always listening to me, relieving my pressure, and providing emotional and financial support. I am also grateful to my grandma, with whom I spent most of my childhood, for your unconditional trust and timely encouragement.

Chapter 1

Introduction

1.1 Background

In modern deep-scaled CMOS technologies, various silicon-related pitfalls present long-term challenges to microprocessor performance. Such challenges include but are not limited to: (1) local hot spots, which breach the thermal limitations of a microprocessor, and (2) transistor aging, especially NBTI that degrades transistor threshold voltage, ultimately threatening the reliability of the entire memory block. In previous systems, the dummy circuit was placed next to the subject, where the dummy was frequently analyzed and the readout was used to infer the condition of the target. Due to rapidly changing ambient conditions (e.g., temperature and voltage), and the potential scale of the target dimensions, such metrics may not accurately represent the condition of the target; additionally, such temperature sensors and canary circuits occupy a significant amount of area.

Therefore, it would be highly preferable to monitor the target circuit in-situ, i.e., to sense the exact transistor at operation. It is also important to achieve an accurate sensing metric: when the temperature is analyzed, the readout should account for voltage and process variations; while sense the aging degradation, the readout should account for voltage

and temperature fluctuations. Such a feature is called in-field operation. Additionally, the circuits should achieve area-efficiency.

During my Ph.D. research period, I built 4 test chips (one in collaboration with Jiangyi Li) to achieve my research goals.

1.2 Compact Voltage-scalable on-chip temperature sensor

Compact temperature sensors are critical to implement dynamic thermal management (DTM) techniques in high-performance microprocessors (μPs) and systems on chips (SoC). Those sensors are embedded at multiple locations on a die, and provide fine-grained temperature information to DTM engine, which to maintain the μPs operate efficiently within thermal budget.

There are three requirements for those temperature sensors. First, sensor front-ends need to be very area efficient. So we opt-out BJT structure. Second, sensor readings need to have a low calibration cost while achieving target accuracy (e.g. absolute $<8°$C and a relative $<3°$C). So we avoid to use expensive 2-temperature-point calibration. Third, the sensors should be able to operate at a low supply voltage and robust to VDD fluctuations. This also eliminates the possibility of any BJT-based structures.

Based on three requirements, we designed a new type of temperature sensors, as shown in fig. 1.1 and fig. 1.2. To achieve small area cost, the sensor only costs 6 or 8 NMOS normal-sized transistors. To achieve low voltage operation, we set the sensors to operate at the sub-threshold region. To achieve good accuracy at the low-cost one temperature point calibration (OPC), we differentiate the readouts of VDD-compensated proportional-to-absolute-temperature (PTAT) and complementary-to-absolute-temperature (CTAT) volt-

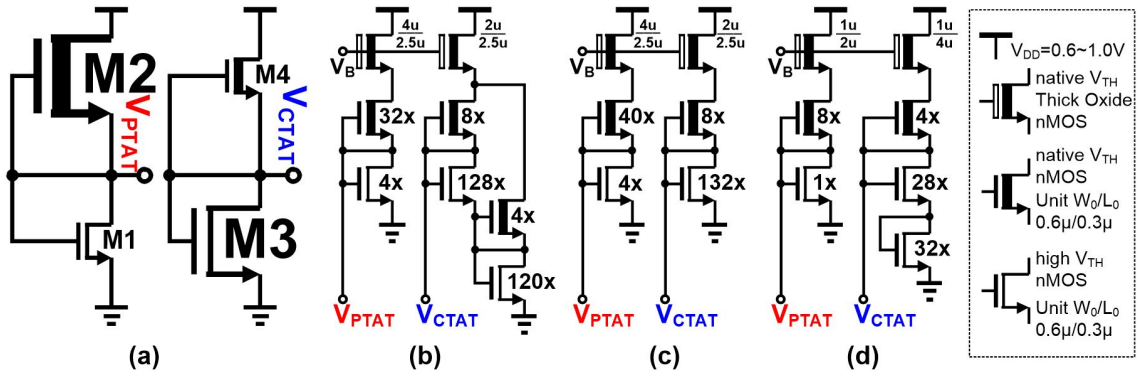


Figure 1.1: Structures of proposed sensor front end

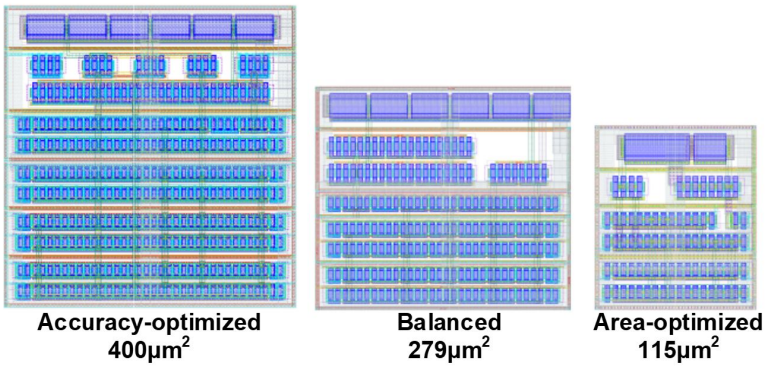


Figure 1.2: Layouts of proposed sensor front end

age generators [23]. We implemented three sensor frontend designs on the 65nm CMOS. Three sensor front-ends are presented with different sizes and accuracies to meet different performance-area requirements (area-optimized, balanced, and accuracy-optimized).

The balanced front end achieves a 14.3× smaller footprint than the 22nm design in [13], while it exhibits a worst error of less than 7.0°C (-3.4°C < error < 3.6°C), after calibration, across 64 sensors in 8 chips. The worst-case measured error among 8 sensors in a chip across 8 chips is 5.2°C. Our sensor can operate at VDDs from 0.6 to 1V, whereas none of the other designs in the comparison can operate below 1V. The average error incurred by the voltage scaling is 0.3°C. The area-optimized front end has a footprint of $115\mu m^2$, and a worst-case error of 8.8°C (-2.0 < error < 6.8°C) across 64 sensors in 8 chips after calibration. The

accuracy-optimized front end exhibits the per-front-end area of $400\mu m^2$ which is still 10x smaller than [13] and a worst-case error of 5.4°C (-0.7 < error < 4.7°C).

The compact footprint and the large voltage-scalability of the designs enable the integration of order-of-magnitude more sensor front ends on a chip at a small additional overhead, enabling dense thermal monitoring in modern VLSI systems.

1.3 In-situ and In-field BTI sensors for register file

In modern deeply-scaled CMOS technologies, transistor aging effects such as bias temperature instability (BTI), hot carrier injection (HCI), and time-dependent dielectric breakdown (TDDB) have been one of the major challenges for maintaining long-term reliability of computing systems [26]. In particular, negative bias temperature instability (NBTI) is one of the most critical aging mechanisms, which can increase PMOS threshold voltage (V_{TH}) at high temperature when PMOS is negatively biased [27, 29, 57]. Such V_{TH} degradation causes digital circuit delay to increase, compromising the maximum clock frequency over chip's lifetime. Recent studies also show that NBTI can be worse for technology scaling [59], confirming its importance in future microelectronics.

NBTI can also degrade robustness of embedded memory circuits [30, 59]. Particularly an 6-transistor (6T) register file (RF) is one of the most vulnerable blocks since it often experiences high temperature due to high switching activities and the heat generated by other digital gates around it. Furthermore, the 6T bitcell typically used in an RF has one of its PMOSs negatively biased (i.e., stressed) during an RF is powered on. Last but not the least, if a bitcell is not written frequently, one of the PMOSs in the bitcell can receive DC stress, which is more detrimental than the AC one [31]. The degradation of V_{TH} of PMOSs

in bitcells can hurt robustness and performance of an RF. It can reduce the static noise margin (SNM), thus worsening data retention voltage (DRV), read V_{MIN} and read access time. Note that the bitcell that undergoes the worst degradation determines the robustness and performance of the entire RF it belongs to.

Conventional approeaches to mitigate NBTI impacts is to set design margins based on a coarsed block-level estimation. Those margins including (1) Upsizing device size to reduce random variations and (2) Operate RF block at a certain amount of higher VDD voltage than pre-silicon block V_{MIN}, a.k.a, a voltage margin. The coarse estimation leads to conservative margin that degrades the area effeciency as well as increase power consumptions. Such margin is pessimistic and unnecessary becuase the only one or tow worst-case bitcells limit the perofrmance of the entire RF block [32].

To tackle this challenge, we propose techniques to dynamically monitor and decelerate NBTI degradation in an RF with a focus to enable/enhance three critical abilities, namely **in-situ** monitoring, **in-field** monitoring, and **post-deployment** NBTI management [55]. First, we devised an in-situ monitoring technique, i.e., directly sensing V_{TH} of a target PMOS in a bitcell instead of using of replica/canary circuits [33, 35]. Second, we have pursued to enhance robustness in monitoring so as to enable in-field monitoring (post-deployment). Robustness against temperature and power supply voltage (VDD) variations is paramount since it is non-trivial to control such parameters in field as shown in fig. 1.3.

Finally, we developed a software framework for dynamic reliability management (DRM). During chip's lifetime, we can execute the framework routinely (e.g., every several months) in the maintenance mode, which monitors V_{TH} degradation, evaluates the degree and the progressing rate of NBTI degradation, and analyzes the skew of V_{TH} degradations between

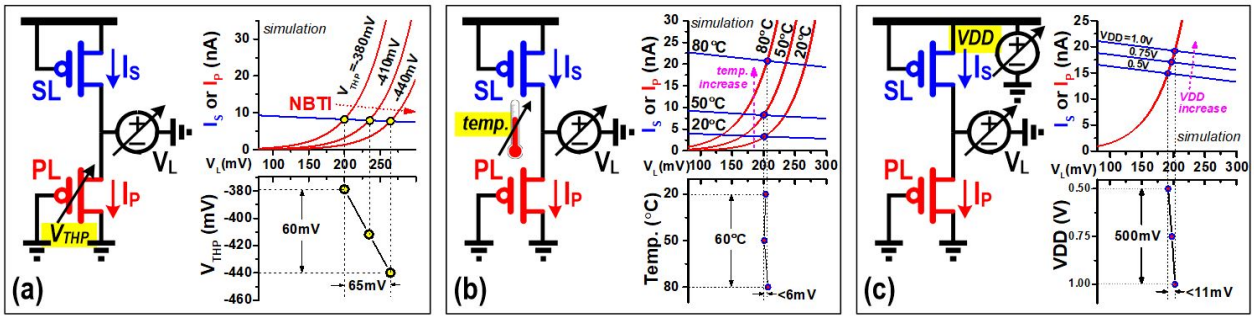


Figure 1.3: Demonstration of V_{TH} sensing capability and robustness

two PMOSs in a bitcell. In addition to those monitoring, the framework can also create recover vectors (RV) using the skew information of bitcells. Written into bitcells opportunistically, RVs can partially recover the more-aged one of two PMOSs in each bitcell and thereby decelerating SNM and DRV degradation.

We prototype test chips, each of which includes a 1-Kb RF with the proposed techniques, in a 65nm CMOS. The measurement results confirm the in-situ and in-field capability with the average error of 19% against temperature variation (20-80C) and that of 21.8% against VDD variations (0.5-1V) in monitoring an NBTI-induced V_{TH} degradation larger than 30mV. Those errors are respectively 4.4X and 3.4X smaller than the estimation of previous work [39]. We also confirm that RVs created based on our monitoring technique can successfully slow down DRV degradation: in our 16-hour accelerated aging experiments, the RFs that store the RVs from the proposed technique exhibit 30mV to 70mV less DRV degradation in average than RFs storing random fixed values. The area overhead of the proposed technique is 27% for a 1-kb RF and 21% for a 4-kb RF.

1.4 Circuit, architecture and run-time framework for memory reliability management in a microprocessor

In deeply scaled VLSI systems, device aging effects, such as bias temperature instability (BTI) have been identified as a key reliability challenge. Especially, embedded caches ($) and register-files (RF) are highly vulnerable to NBTI since they use intrinsically sensitive circuits and become hot as nearby logic circuits are actively switching and dissipating heat. What's worse is that the single worst-case bitcell can determine the reliability of the entire memory block. It is paramount to manage the reliability of the embedded memory over the chip's lifetime.

To manage reliability it is cost-prohibitive to disassemble working μP and send the chip to a laboratory. Thus, a key requirement is to perform the reliability management without disassembly, i.e., post-deployment and in-field. In this work, we propose such a solution, including circuits, a microarchitecture and a runtime software framework to implement a post-deployment in-field memory reliability management.

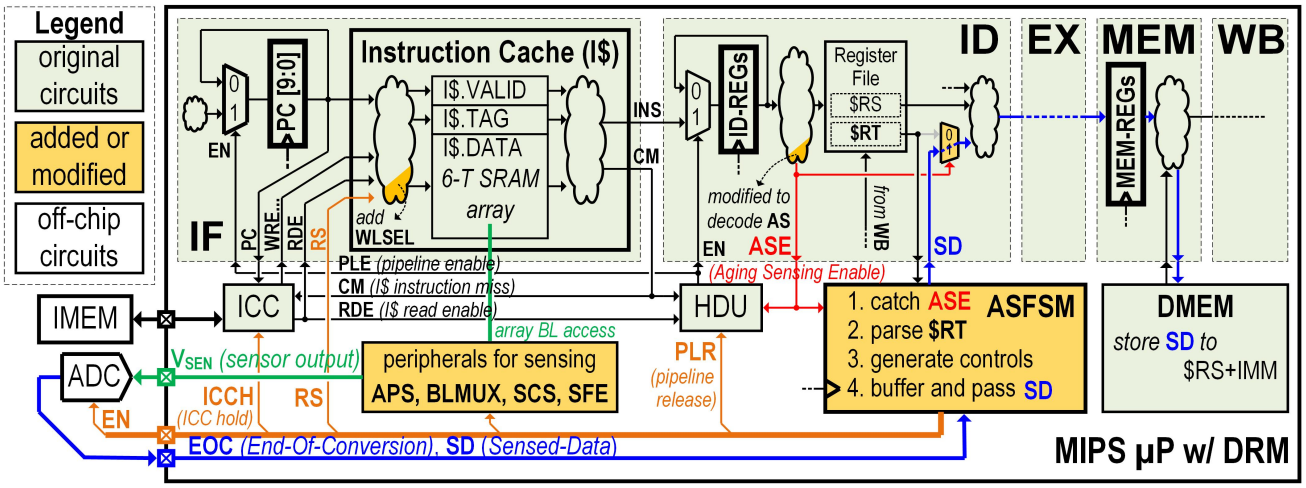


Figure 1.4: Microprocessor architecture of proposed DRM technique

We prototyped a test chip with the implementation of previous developed NBTI sensing circuits into the circuit of instruction cache (I$) of a standard 5-stage MIPS μP (fig. 1.4). we devised a RISC microarchitecture having a new instruction (called AS) to trigger the reliability management. The μP can execute the AS instruction opportunistically during its regular operation to sense the V_ts of its bitcells, thereby evaluate its aging effects. The sensor is able to *in-situ* sense the V_t of any one of six transistors in any bitcell. The sensing readout is robust against temperature variations and thus very suitable to the on-the-fly sensing, which it is impractical to regulate temperature rapidly.

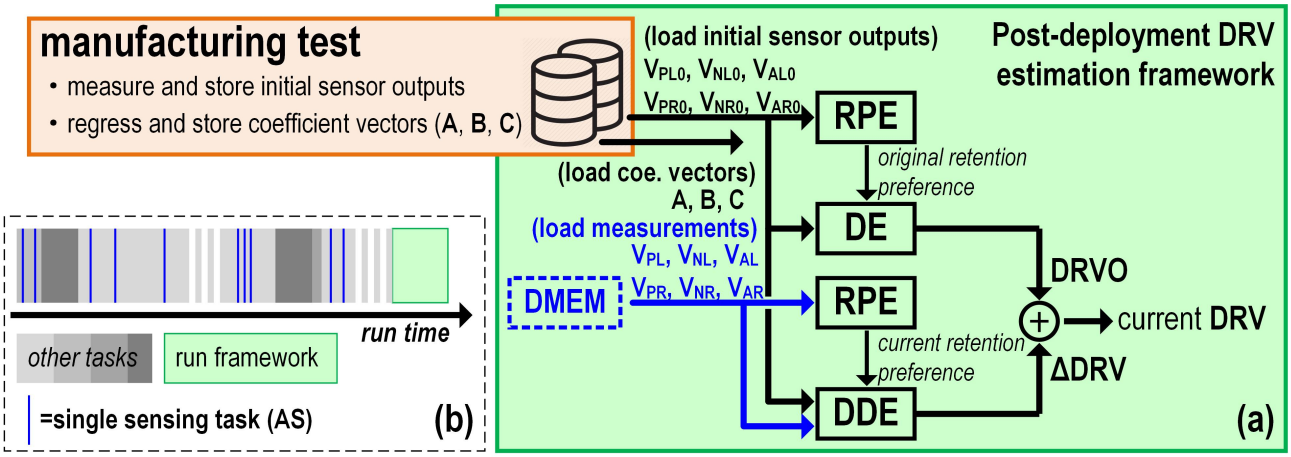


Figure 1.5: (a) Sensing task assignment (b) Framework to estimate current DRV

We also proposed a runtime software framework (fig. 1.5) to convert the low-level V_t measurements to circuit-/architecture-level metrics, i.e., the data retention voltage (DRV). The framework is composed of basic operations (+, $\times$) and can be accomplished by original instructions of the μP ISA. Such extraction is critical for operating systems and firmware as they need high-level metrics to run dynamic reliability managements (DRM), such as to decelerate aging in bitcells [65], to reduce the guard band in dynamic voltage scaling of memory during the standby mode, and to balance aging degradations among memory

banks [63].

1.5 An area-efficient SoC with instruction cache transformable to a temperature sensor and a PUF

Heading towards the era of Internet of Things (IoT), it is critical for integrated-circuit research and development to deliver compact, low-cost, and dependable edge devices with various capabilities, e.g., sensing, computing, communication, and security [67]. This challenge has motivated to integrate an increasing number of components and function blocks into a Microprocessor-based System-on-Chip (μP-SoC) to shrink system footprint and associated cost [82, 88, 89]. However, such integration often incurs silicon area increase since most of analog, mixed-signal, and digital circuits require substantial amounts of silicon area to implement fast, accurate, and robust functions.

An ambient temperature sensor (T-sensor) and a Physically Unclonable Function (PUF) are two widely used components in IoT devices. The former is a critical building block for environmental monitoring; the latter is a notable security macro used for secret key generation for cryptography and chip-ID generation for authentication. However, implementing dedicated circuits for those functions requires non-negligible silicon area, especially when they are designed for high accuracy and robustness [68–77, 79, 84, 86–89].

It is noteworthy that in many applications, T-sensors and PUFs exhibit low duty cycle, making the approach of dedicated hardware further inefficient in area. As shown in fig. 5.1(a), for example, a T-sensor can be only active every several seconds (or even longer) since ambient temperature changes rather slowly [82,88]. A PUF also needs to be active only upon a request for e.g., encrypting and decrypting messages, and chip authentication processes

[85, 86]. Therefore, the dedicated hardware can be idle for most of the time.

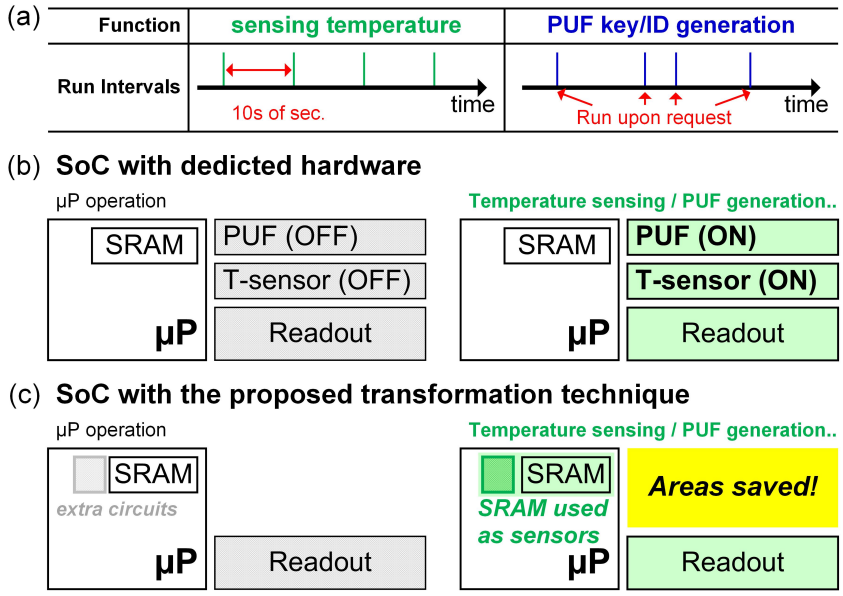


Figure 1.6: (a) Low duty cycle operation such as ambient temperature sensing and PUF. (b) Dedicate hardware implementation of those functions compared with conventional design. (c) Proposed transformation approach can save area.

Therefore, we aim to address such area inefficiency, and propose a novel technique to transform the existing SRAM in the instruction cache (I$) of a μP into a T-sensor or a PUF (fig. 1.6(c)). This hardware recycling approach can reduce silicon footprint while integrating more features on a chip. To enable such transformation, we made a minimal amount of change in the SRAM circuits, Instruction Set Architecture (ISA), and pipeline control logic. The outputs of the transformed T-sensor and PUF operations are stored in the data memory of the μP for post digital processing.

We prototyped a μP-based SoC with the proposed technique in a 65nm general-purpose CMOS, fig. 5.1. The μP can operate at 320MHz at 1V supply voltage (V_{DD}) and consumes 10.6 pJ/cycle. The transformed T-sensor achieves an error of -0.5/+1.5°C after One-temperature-Point Calibration (OPC) across 26 instances. It achieves low V_{DD} sensitivity,

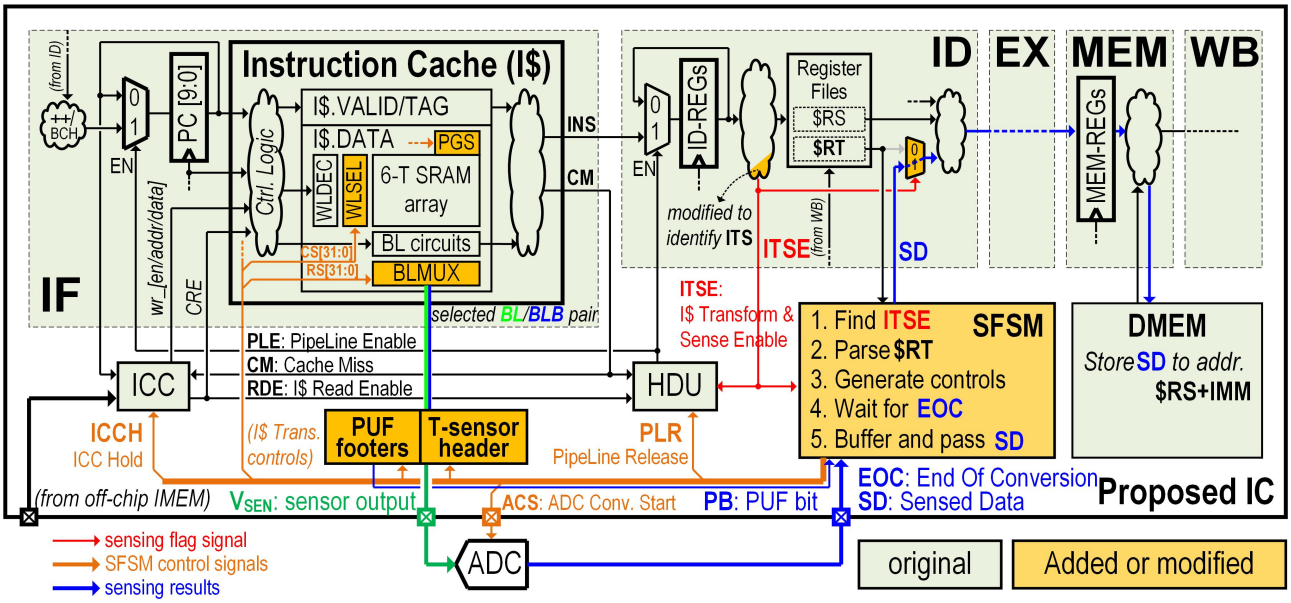


Figure 1.7: The proposed µp-SoC microarchitecture. The modified and added portions are highlighted in yellow.

exhibiting only 0.46°C error for 100mV V_{DD} variation from 1V to 0.5V. The transformed PUF also achieves a desirable randomness: the analog differential output shows a normal distribution with μ=-1.3mV and σ=31.2mV; the digitized bitstream passes all the applicable NIST tests and achieves 0.502 inter-PUF Fractional Hamming Distance (FHD). It also achieves robustness comparable to the state of art: 0.027% unstable bit ratio and 1.97×10^{-5} Bit Error Ratio (BER) after Temporal Majority Voting (TMV11) and Comparator Input Swapping (CIS) based masking.

The proposed transformation capability increases the area of the baseline μP by 12.9% (9.2% only for the T-sensor and 9.1% only for the PUF). The first 6.3% is for the update in the SRAM circuits and the next 6.6% is for the microarchitecture modification. The standalone T-sensor [73] and PUF [76] circuits achieving the similar accuracy and robustness would consume more silicon area, that would be 62.9% of the baseline μP area.

Chapter 2

Compact, Voltage-scalable on-chip Temperature Sensor

2.1 Motivation

Compact temperature sensors are critical to implement dynamic thermal management (DTM) techniques in high-performance microprocessors (μPs) and systems on chips (SoC). Those sensors are embedded at multiple locations on a die, and the temperature information sensed is used to maintain the chip operation within thermal constraints. While existing sensor designs [13–18] achieve small area and high accuracy, emerging technology trends such as multi-core architectures, 3D-integration, fin-fet devices, and low-voltage operation require the development of sensors of better performance that meet stricter requirements.

There are three requirements for those emerging applications.

First, sensors need to be very area efficient. Recently, the number of sensors embedded on a digital VLSI system has rapidly increased (fig. 2.1). Meanwhile, the continuing increasing the level of integration (3D integration, SiP, etc.) and consequently the number of thermal hot spots, future digital VLSI systems will have more locations that require thermal monitoring. In order to reduce the overhead while monitoring all of those locations, the sensor footprint needs to be minimized. Additionally, a compact footprint ensures design flexibility

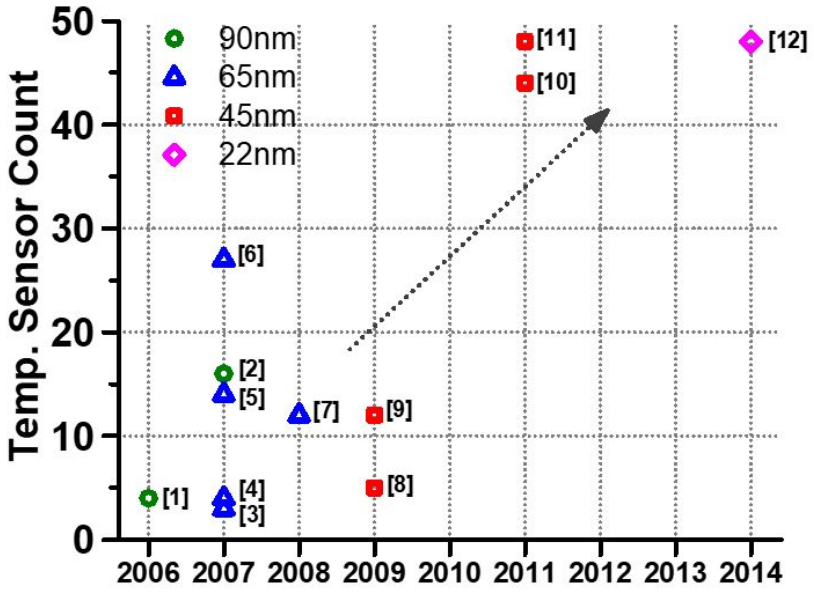


Figure 2.1: Trends in numbers of on-chip temperature sensors in μPs and SoCs

as sensor's location is often determined at the later stages of the design process [13].

Second, sensors need to have a low calibration cost while achieving target accuracy (e.g. absolute <8°C and a relative <3°C).

Third, the sensors need to be able to operate at a low supply voltage. Sub-1V operation for digital VLSI systems is being extensively explored, as part of the ongoing request to reduce power consumption. The conventional sensors hardly operate below 1V, thereby requiring additional power distribution or local power regulation. The ability to operate below 1V eliminates those overheads.

Existing published temperature sensor designs [13–18] hardly meet above requirements simultaneously. BJT-based sensors [13, 14] achieve good accuracies but cost large area and require above 1V VDD. A lateral CMOS diode based sensor [15] and CMOS V_t-based design [17] achieve small foot print but require expensive two temperature point calibration (TPC) to satisfy accuracy target.

Therefore, we are motivated to propose a new type of temperature sensor design. To achieve small area cost, the sensor only costs 6 or 8 NMOS normal-sized transistors. To achieve low voltage operation, we set the sensors to operate at the sub-threshold region. To achieve good accuracy at the low-cost one temperature point calibration (OPC), we differentiate the readouts of VDD-compensated proportional-to-absolute-temperature (PTAT) and complementary-to-absolute-temperature (CTAT) voltage generators [23]. To meet different performance-area requirements, we implemented three sensor frontend designs on the 65nm CMOS. Three sensor front ends with different sizes and accuracies, namely area-optimized, balanced, and accuracy-optimized.

2.2 Operation Principle

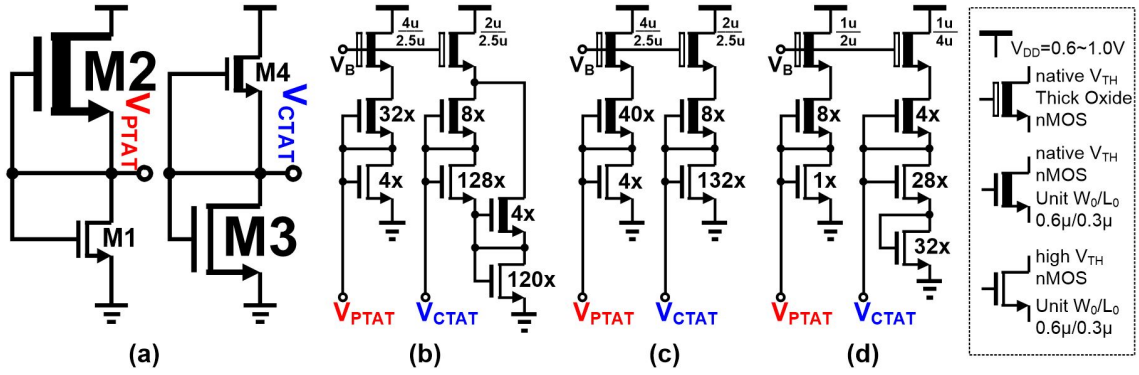


Figure 2.2: Structures of proposed sensor front end

fig. 2.2(a) shows the basic structure of the proposed sensor front-end, which contains a PTAT and a CTAT voltage generator with outputs V_{PTAT} and V_{CTAT}. Transistor M2, M4 are native-V_t NMOS and transistor M1, M3 are thin-oxide, high-V_t NMOS at diode-connection. To generate PTAT voltage, M2 is sized larger than M1; similarly, M4 is sized smaller than M3 to generate CTAT voltage. The V_{PTAT} and V_{CTAT} are sensed, differentiated and digitized

at the back-end read-out circuits. The read-out circuits are shared across all sensor front ends. fig. 2.2(b), (c) and (d) show the actual implementations of the three sensor front ends. Several supplementary devices are added to the basic structure to improve various metrics. fig. 2.2(b) shows the design for optimal accuracy at the larger footprint (referred to as accuracy-optimized design); fig. 2.2(c) shows the design for balancing area and accuracy (balanced design); and fig. 2.2(d) shows the design for minimizing area (area-optimized design). fig. 2.3 shows the layouts of the front ends. The areas of the three sensor front ends are $400\mu m^2$, $279\mu m^2$ and $115\mu m^2$.

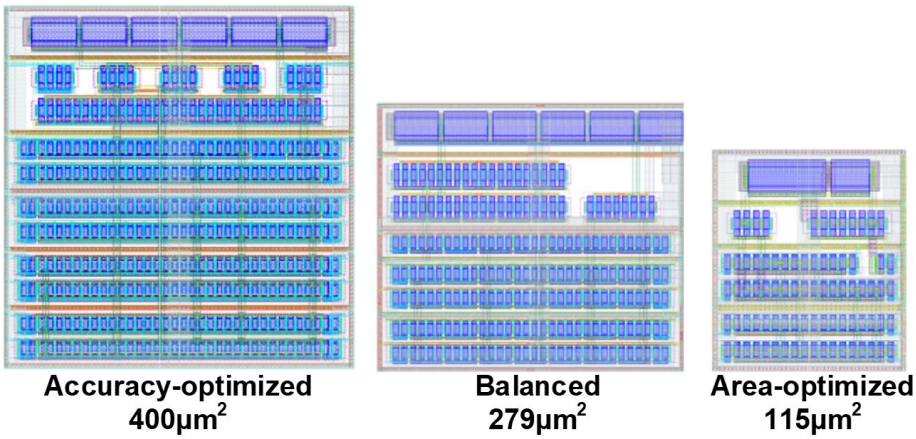


Figure 2.3: Layout snapshots of three proposed sensor front ends

The operation of the PTAT or CTAT generator can be explained through the device current-voltage (I-V) characteristics between the top and bottom transistors. fig. 2.4(a) shows the model that we used: either PTAT or CTAT generator contains a top transistor (M2 or M4) with gate-source shorted (a.k.a. a single-transistor current source with zero V_{gs}) and a bottom transistor (M1 or M3) configured as diode. Then we split the structure into top and bottom parts from the output node of the sensor. Each breaking node is connected to an ideal voltage source and force the voltage to V_{OUT}. To demonstrate different mechanisms

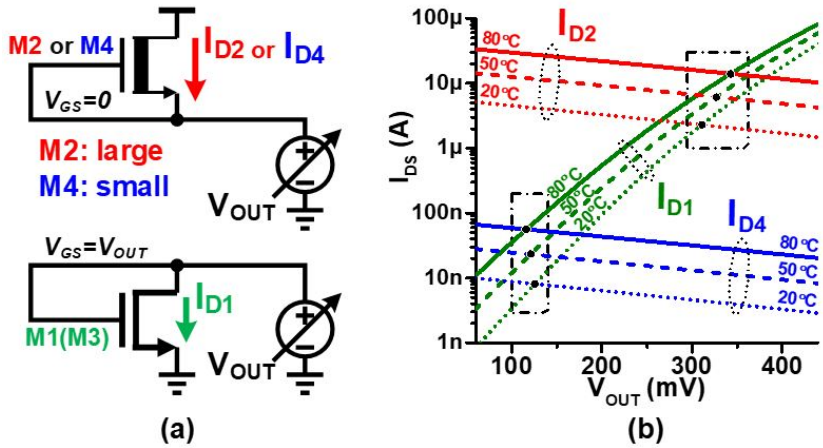


Figure 2.4: Simulation setups for the I-V characterizations

between PTAT and CTAT, we chose two sizes for top transistor (M2 is bigger than M4). To ensure same reference, we use only one size for bottom transistor (M1). Then we swept V_{OUT} from 50 to 450mV at three different temperatures. The channel currents (Id) of transistors are measured and plotted in fig. 2.4(b). Green lines represent Id of bottom transistor, increasing with V_{OUT} and temperature. Blue lines represent Id of the smaller top transistor (M4), decreasing with V_{OUT} but increasing with temperature. Red lines represent Id of the bigger top transistor (M2) with similar trend, however, the absolute current value is nearly three magnitude bigger than that of M4.

The intersect points between blue and green curves are the output of CTAT generator (M4 with M1). The intersect points between red and green curves are the output of PTAT generator (M2 with M1). Both intersections are at deep sub-threshold region. fig. 2.5(a) shows the zoom-in of Id intersection between M4 (smaller top) and M1, the V_{OUT} is CTAT. fig. 2.5(b) shows the zoom-in of Id intersection between M2 (bigger top) and M1, the V_{OUT} is PTAT.

The analytical equation for V_{PTAT} and V_{CTAT} can be derived similarly as for the compact

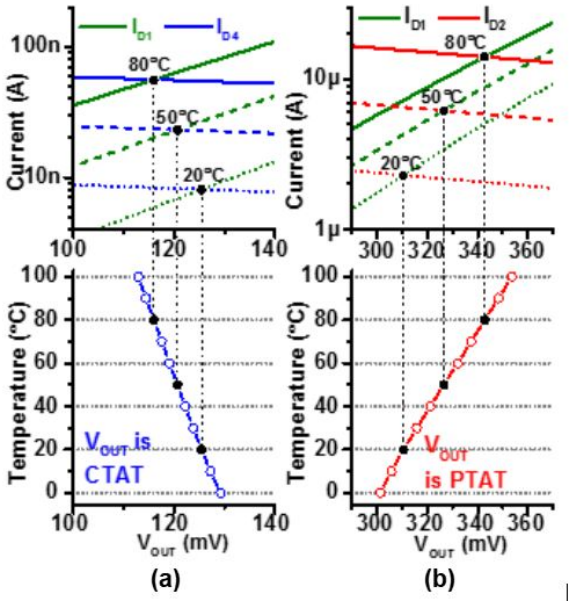


Figure 2.5: The zoomed-in temperature characteristics of the CTAT and PTAT generator

2-transistor voltage reference circuits [19]. Here, we focus only on V_{PTAT} since the equations for V_{CTAT} are the same except for the different transistor notations. The sub-threshold current is Equation (1)

$$I_D = \mu C'_{ox} \frac{W}{L} (n-1) \phi_t^2 e^{\frac{V_{gs}-V_t}{n\phi_t}} (1 - e^{-\frac{V_{ds}}{\phi_t}}) \tag{1}$$

where μ is the carrier mobility, C'_{ox} is sheet oxide-capacitance density, W, L are the width and length of the transistor, V_t is threshold voltage, n is subthreshold slope, V_{GS} is gate-source voltage, V_{DS} is drain-source voltage, and ϕ_t is the thermal voltage. Based on Equation (1), the current equations for M1 and M2 can be derived as Equation (2)

$$\begin{cases} I_{D1} = \mu_1 C'_{ox1} \frac{W_1}{L_1} (n_1 - 1) \phi_t^2 e^{\frac{V_{PTAT}-V_{t1}}{n_1 \phi_t}} \\ I_{D2} = \mu_2 C'_{ox2} \frac{W_2}{L_2} (n_2 - 1) \phi_t^2 e^{-\frac{V_{t2}}{n_2 \phi_t}} \end{cases} \tag{2}$$

Since M1 and M2 are connected in series, I_{D1} and I_{D2} in Equation (2) are identical so that we can solve the V_{PTAT} as Equation (3)

$$V_{PTAT} = \underbrace{n_2 \ln\left(\frac{\mu_1}{\mu_2} \cdot \frac{C'_{ox1}}{C'_{ox2}} \cdot \frac{W_1 L_1}{W_2 L_2} \cdot \frac{n_1 - 1}{n_2 - 1}\right) \frac{k}{q}}_{\text{slope}} \cdot T + \underbrace{V_{t2} - \frac{n_2}{n_1} V_{t1}}_{\text{offset}} \tag{3}$$

where k is the Boltzmann constant and q is the electron charge. The temperature sensitivity (slope of V_{PTAT}) is, to the first order, determined by the size ratio between M1 and M2. The offset of V_{PTAT} is a function of process parameters V_t and n.

2.3 Accuracy Improvements

2.3.1 Output Range Tuning

We optimized the sensor front end to ensure that all transistors to operate in subthreshold and in saturation (i.e., V_{DS} is several times larger than ϕ_t) across temperature and process variations. This is critical because a high V_{DS} ensures that the last term in Equation (1) to be negligible, thereby reducing the V_{DD} dependency and improving the linearity of V_{PTAT} and V_{CTAT} over temperatures.

In order to ensure sufficiently high V_{DS}s for transistors, we connect the gates of the top transistors to the outputs, as shown in fig. 2.2(a). This is a modification of the original topology in [19], and raises the level of V_{PTAT} by approximately $1/2(V_{t2} - V_{t1})$ and that of V_{CTAT} by approximately $1/2(V_{t4} - V_{t3})$, contributing to sufficiently higher V_{DS}s for the bottom transistors (M1, M3). In addition, we carefully chose the device types to give V_{PTAT} and V_{CTAT} optimal offsets in their output voltage, i.e. the second term in Equation (3). The optimal offset is $1/2VDD$ since it ensures largest V_{DS}s for both top and bottom transistors

of the PTAT or CTAT generators. However, as the offset is roughly the difference of Vts of the top and the bottom transistors, it tends to be less than $1/2VDD$.

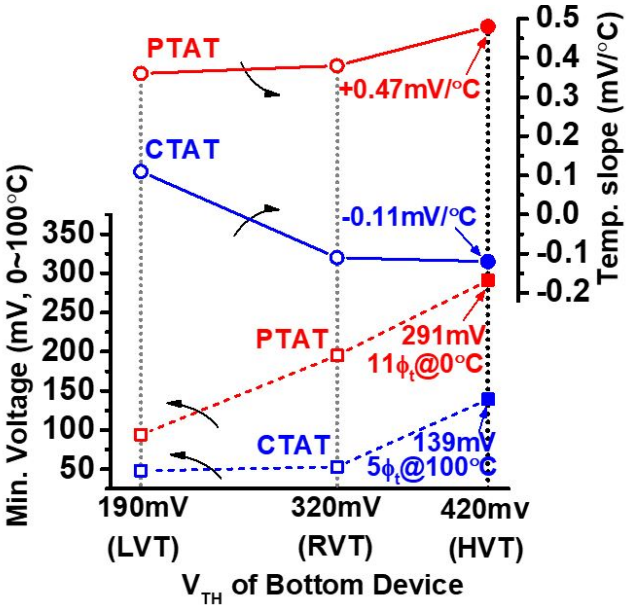


Figure 2.6: The output range and slope of PTAT and CTAT generator

Therefore, we explored the use of three different types of devices and found the lowest values of the V_{PTAT} (at 0°C) and V_{CTAT} (at 100°C). As shown in fig. 2.6, using high-V_t NMOSs for the bottom devices provides higher offsets ($V_{PTAT,MIN}$ = 291mV = $11\phi_t$ at 0°C and $V_{CTAT,MIN}$ = 139mV = $5\phi_t$ at 100°C). This device choice further enables a sufficient temperature sensitivity of 0.47mV/°C for V_{PTAT}, and -0.11mV/°C for V_{CTAT}.

For the accuracy-optimized and area-optimized front ends, we added a diode and a temperature- and voltage-compensated voltage reference circuit [19], respectively, as a footer. The footers can increase the offset of V_{CTAT} which is typically smaller than V_{PTAT} and limits the accuracy of sensor front ends.

2.3.2 Differential Read-out

In order to improve linearity over temperature even in the presence of process variations, we propose a differential read-out scheme where the difference between V_{PTAT} and V_{CTAT} is used to measure temperature. Good linearity over temperature is critical to reduce errors after OPC. Using either V_{PTAT} or V_{CTAT} alone causes poor linearity since some of the parameters determining the temperature-slope of V_{PTAT} and V_{CTAT}, namely n, μ, and V_t, are temperature dependent [20, 21]. Moreover, those parameters vary with process, further degrading linearity.

The differential reading scheme mitigates the impact of those nonlinear parameters on the temperature-slope of V_{PTAT} or V_{CTAT}. Now the temperature-slope is mainly determined by transistor sizing ratios after canceling process-dependent parameters. Both V_{PTAT} and V_{CTAT} have form of Equation (3), we can find the expression for $V_{DIFF} = V_{PTAT} - V_{CTAT}$ as Equation (4)

$$V_{DIFF} \approx \underbrace{n_1 \ln\left(\frac{W_1 W_4 L_2 L_3}{W_2 W_3 L_1 L_4}\right)\frac{k}{q}}_{\text{slope}} \cdot T - \Delta V_{t1,3} + \frac{n_1}{n_2}\Delta V_{t2,4} \tag{4}$$

where $\Delta V_{ti,j}$ is $V_{ti} - V_{tj}$. Since transistors {M1,M3} and {M2,M4} are the same type, it is reasonable to assume that the T_{NOM} values and temperature-dependencies of the pairs {n1,n3}, {n2,n4}, {μ_1,μ_3}, {μ_2,μ_4}, {V_{t1},V_{t3}}, {V_{t2},V_{t4}}, {C'_{ox1},C'_{ox3}} and {C'_{ox2},C'_{ox4}} are tracking each other across temperatures and process variations. Equation (4) clearly shows that the sources of nonlinearity, i.e. n, μ, and V_t, are mostly canceled out.

In addition to reduce nonlinearity, the proposed differential reading scheme can mitigate

the impact of systematic process variations. Systematic variations modulate the parameters of the same transistor type in similar directions. As shown in Equation (4), the use of V_{DIFF} cancels out most of the dependencies of systematic process variations, comparing with the single V_{PTAT} (Equation (3)) or V_{CTAT}.

Although removing most of the temperature and process-dependent parameters (μ, n) in the slope term, the slope of V_{DIFF} still suffers from local mismatches, which hurt the accuracy. The expression of V_{DIFF} can be expanded further to investigate the impacts from local mismatches. The temperature dependency of threshold voltage (V_t) is given by Equation (5)

$$V_t = V_{t0} + (K_1 + K_2 V_{BS}) \left(\frac{T}{T_{NOM}} - 1 \right) \tag{5}$$

where K_1 and K_2 are temperature coefficients of V_t, T_{NOM} is the reference temperature. By plugging this into Equation (5), we can derive the explicit expression of V_{DIFF}, as shown in Equation (6)

$$\begin{aligned} V_{DIFF} \approx & \underbrace{\left[n_1 \ln \left(\frac{W_1 W_4 L_2 L_3}{W_2 W_3 L_1 L_4} \right) \frac{k}{q} - \frac{\Delta K_{1,3}}{T_{NOM}} + \frac{n_1}{n_2} \frac{\Delta K_{2,4}}{T_{NOM}} \right]}_{\text{slope}} \cdot T \\ & + \underbrace{\frac{n_1}{n_2} (\Delta V_{t0-2,4} - \Delta K_{2,4}) - (\Delta V_{t0-1,3} - \Delta K_{1,3})}_{\text{offset}} \end{aligned} \tag{6}$$

where $\Delta K_{1,3}$ is the mismatch of K_1 between transistor {M1,M3}; $\Delta K_{2,4}$ is the mismatch of K_1 between transistor {M2,M4}. Body effect coefficients K_2 is ignored since the value is much smaller than K_1.

As shown in Equation (6), the offset term, which is insensitive to temperature, can be removed after OPC. The slope term contains more parameters than Equation (4) after expansion. The slope variation of V_{DIFF} is caused by (1) process variation of n1, (2) mismatches between four transistors' dimension, (3) mismatch of K_1 across four transistors, (4) mismatch of n between transistor {M1,M2}. The straight forward approach to reduce slope variation is to upsize W and L, which reduces the variability of such parameters as W, L, n and ΔK_1. This is pursued in this paper. Another way to further improve the variability might be to use the common-centroid layout technique between PTAT and CTAT generators. This can mitigate the parameter mismatches due to gradient variation.

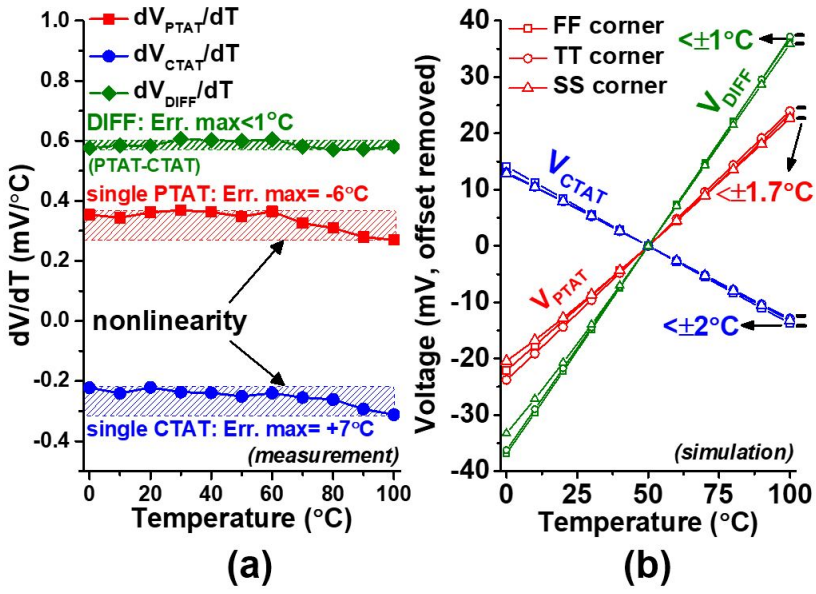


Figure 2.7: Differential read improves linearity

Measurement results confirm the reduced impact of the temperature dependency of those parameters on the linearity of V_{DIFF}. As shown in fig. 2.7(a), when using either V_{PTAT} or V_{CTAT} only, the poor linearity over the temperature range of 0 to 100°C can cause -6°C or +7°C error after OPC, respectively. The use of V_{DIFF} achieves drastically better linearity,

resulting in an error less than 1°C after OPC. This is also confirmed by the process-corner simulations. As shown in fig. 2.7(b), the use of V_{DIFF} has errors of only ±1°C across FF, TT, and SS process corners, while the use of either V_{PTAT} or V_{CTAT} can cause 1.7 to 2X larger errors.

2.3.3 Supply-voltage Scalability

Dynamic voltage scaling (DVS) is a popular technique in today's digital VLSI systems to reduce power consumption. The VDD is dynamically modulated to less than 1V when lower performance demands allow to opportunistically save energy. Temperature sensors that can work below 1V can share power grids with the digital circuits. The existing sensors, however, cannot operate at sub-1V supplies; they need additional power distribution and local regulation, causing a significant area overhead. In order to make the sensors to use digital power rails, it is further critical to achieve a good power-supply reject ratio (PSRR). In the proposed design the low-frequency PSRR (LF-PSRR) is particularly important since the low-pass-filter like VDD-to-output behavior of PTAT and CTAT generators has a good high-frequency PSRR [19].

The proposed differential reading can improve LF-PSRR, by canceling the error caused by the common-mode change in V_{PTAT} and V_{CTAT} due to VDD scaling. The total output change in V_{PTAT} and V_{CTAT} over VDD-scaling can be decomposed into a differential mode error ($DME = \Delta V_{PTAT} - \Delta V_{PTAT}$) and a common mode error ($CME = (\Delta V_{PTAT} + \Delta V_{PTAT})/2$), where ΔV_{PTAT} and ΔV_{CTAT} are the change of V_{PTAT} and V_{CTAT} for a given VDD change.

In order to remove the remaining DME, as shown in fig. 2.2(b)(c) and (d), we added cascode devices on the top and biased them at VB. Thick-oxide native-Vt NMOSs were chosen

for the cascode devices for two reasons. First, these devices have only a small amount of gate-leakage. Second, their Vt is close to zero, allowing a bias voltage (VB) lower than VDD which can be generated by an on-chip low-voltage low-power voltage reference circuit like in [19]. fig. 2.8(a) and (b) show the error-reduction achieved by using the differential-reading scheme and employing the cascode devices, respectively. The former achieves approximately 10dB improvement, and the latter can achieve additional 26dB in LF-PSRR. The resultant measured error is 0.8°C for a VDD scaling from 1V down to 0.6V in a typical chip. Besides using cascode devices, increasing transistor length could be another way to improve LF-PSRR, but this would cause an area penalty.

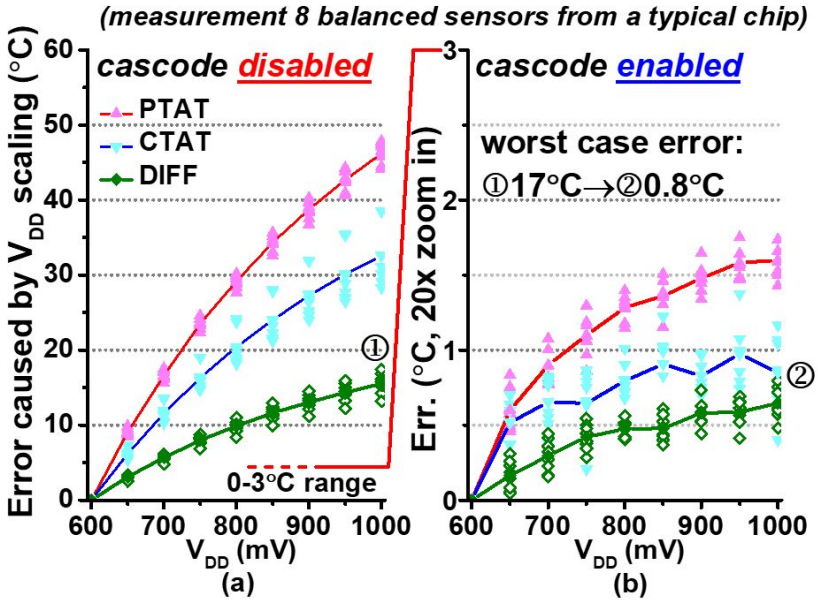


Figure 2.8: Differential reading achieves smaller error across the VDD scaling

2.3.4 Noise

The noise performance of the sensor front ends have been investigated with simulations. fig. 2.9(a) shows the output noise spectrum of the balanced front end design at 100°C.

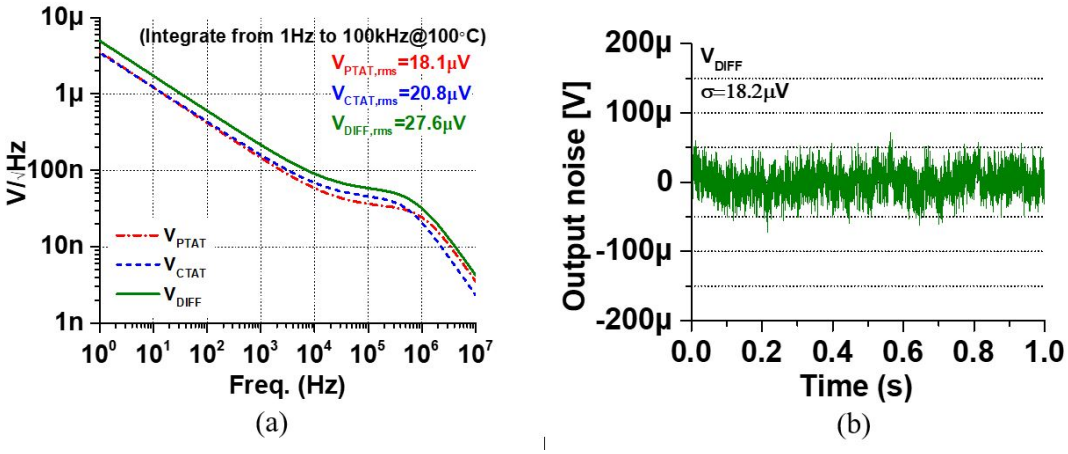


Figure 2.9: Noise simulation

Flicker noise with a corner frequency of about 10 kHz dominates. The differential reading scheme slightly increases the noise output compared to reading either V_{PTAT} or V_{CTAT}. The integration of the V_{DIFF} noise spectrum from 1Hz to 100kHz is 27.6μVRMS, which is only 14.5ppm of the nominal V_{DIFF} value of 190 mV at 100°C. As shown in fig. 2.9(b), we also performed transient noise simulations. The root-mean-square (rms) value of the noise output of V_{DIFF} is found to be 18.2μVRMS. This corresponds to a worst-case $\pm 3\sigma$ error of 0.19°C.

2.4 Read-out conditioning circuit design

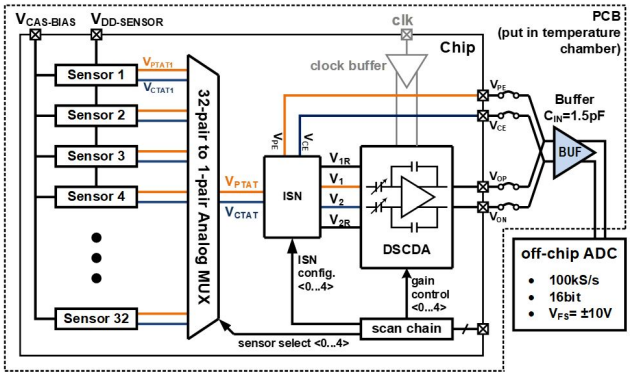


Figure 2.10: Test-chip block diagram

As shown in fig. 2.10, we designed a test chip that has 32 sensor front ends with a

shared back end. The back end comprises of a 32-to-1 two-channel analog multiplexer, an input switch network (ISN), and an on-chip differential switched-capacitor difference amplifier (DSCDA). The analog multiplexer takes 32 output pairs (the V_{PTAT} and V_{CTAT} of 32 sensor front ends) and passes one such pair to the ISN. The ISN then convey the inputs to the DSCDA. The analog voltage from the amplifier (V_{OP}, V_{ON}) is digitized by an off-chip analog-to-digital converter (ADC). Alternatively, the on-chip amplifier can be bypassed via the ISN. The ISN the produces V_{PE} and V_{CE} which are the outputs of the selected sensor front end. V_{PE} and V_{CE} can be sensed with an off-chip buffer and then digitized by the ADC. Both the analog multiplexer and the ISN use thick-oxide devices in order to eliminate gate-leakage. Since the V_{PTAT} and V_{CTAT} are less than 0.4V across 0 to 100°C, NMOS-only switches can be used instead of transmission gates, which is more area efficient.

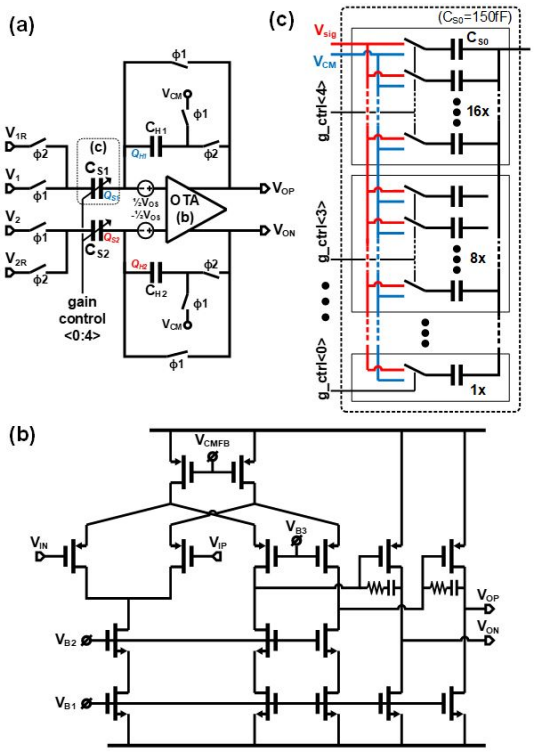


Figure 2.11: DSCDA schematics

fig. 2.11(a) shows the schematic of the DSCDA, which consists of a classical 2-stage folded-cascode fully differential amplifier (fig. 2.11(b)) and two symmetric programmable switched-capacitor banks (CS1 and CS2). In order to minimize the impact of input-offset (V_{OS}) and low-frequency noise, auto-zero double-sampling [21] is used in the amplifier. As shown in fig. 2.11(c), C_{S1} and C_{S2} can be configured from $1 \cdot C_{S0}$ to $(16+8+4+2+1) \cdot C_{S0}$, where C_{S0} is 150 fF. The holding-capacitors (C_{H1}, C_{H2}) are 150 fF each. MIM capacitors are used to implement these capacitors.

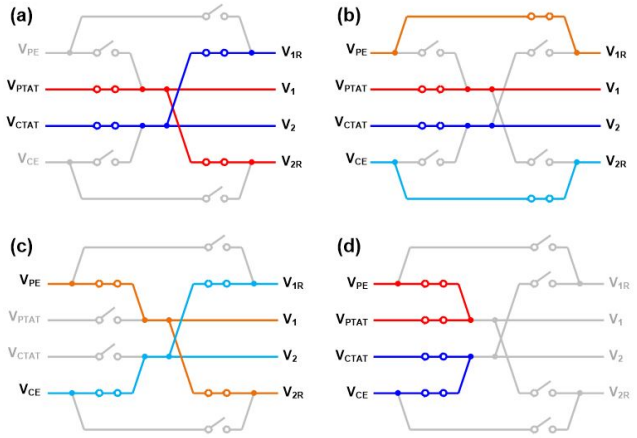


Figure 2.12: Four modes of DSCDA

The ISN and the DSCDA can support four different modes to produce four analog outputs, namely V_{OP}, V_{ON}, V_{PE}, and V_{CE}. In the self-reference (SR) mode, shown in fig. 2.12(a), the V_{PTAT} and V_{CTAT} not only act as the two input signals, but also serve as each other's reference voltage for the DSCDA. The second mode is the external-reference (ER) mode, where V_{PTAT} and V_{CTAT} become signal inputs and two off-chip signals (V_{PE} and V_{CE}) are used for reference inputs for the DSCDA (fig. 2.12(b)). The third mode is the amplifier calibration mode, where two off-chip signals from V_{PE} and V_{CE}, respectively, are connected to both the signal (V_1, V_2) and the reference (V_{1R}, V_{2R}) nodes of the amplifier (fig. 2.12(c)). The fourth

operation mode is the off-chip readout mode, where the V_{PTAT} and V_{CTAT} of a sensor front end are fed to the off-chip buffer via the V_{PE} and V_{CE} nodes (fig. 2.12(d)). In this work we mainly used the SR mode since it supports the differential reading scheme without external reference voltages and it can also provide higher robustness to the mismatch between C_Ss and C_Hs than the ER mode. The output function of SR and ER modes is Equation (7)

$$\begin{aligned} ER: V_{OD} &= \frac{C_S}{C_H}[(V_{PTAT} - V_{PE}) - (V_{CTAT} - V_{CE})] \\ SR: V_{OD} &= 2\frac{C_S}{C_H}(V_{PTAT} - V_{CTAT}) \end{aligned} \tag{7}$$

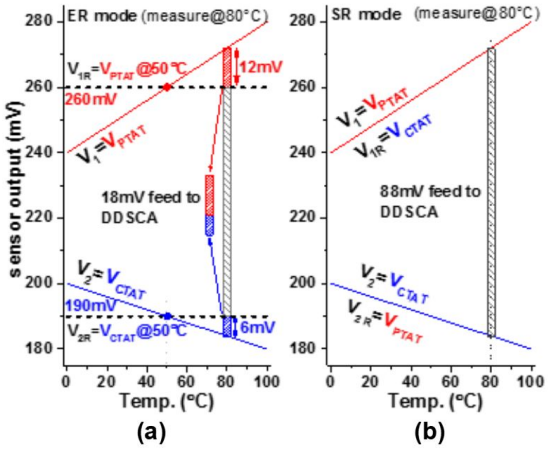


Figure 2.13: The available amplification rooms of the ER and the SR mode

While we mainly use the SR mode, it is noteworthy that the ER mode can provide higher voltage gain which may relax the precision requirement of the ADC. In fig. 2.13(a), at 50°C, V_{2R} can be set as V_{PTAT} (the middle point of the temperature range) and likewise, V_{1R} as V_{CTAT}. The full difference between V_{PTAT} and V_{CTAT} (e.g., 18mV at 80°C) can now be amplified at a higher (e.g., 30X) gain. For the same case, the SR mode limits the maximum gain to 9X, since the input voltage, $V_{PTAT} - V_{CTAT}$, contains a large temperature-independent

offset (e.g., 70mV at 80°C as seen in fig. 2.13(b)).

2.5 Silicon Implementation

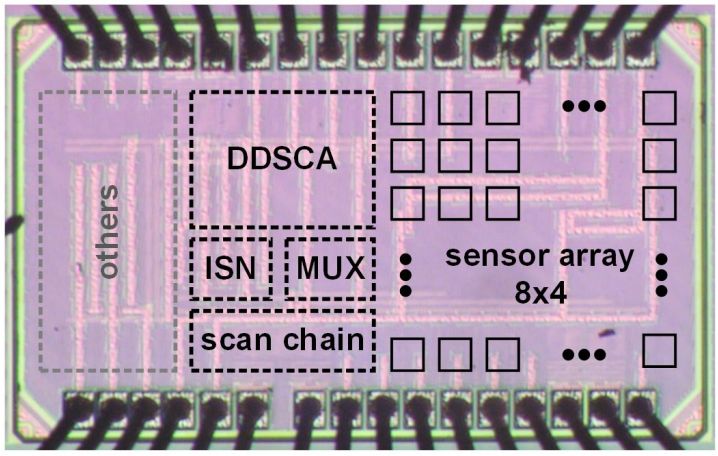


Figure 2.14: Test chip die photo

The test chips for the proposed temperature sensors have been fabricated in a 65nm General-Purpose CMOS process. fig. 2.14 shows the die photo of the test chip. The three types of sensor front ends are configured in a 4-by-8 array. The area of the chip, including I/O pads, is 0.9x0.72mm^2.

2.6 Measurements

2.6.1 Sensor Accuracy

We measured multiple sensor front ends with two scenarios (off-chip amplifier with off-chip ADC; on-chip DSCDA with off-chip ADC). Then the worst case errors are reported based on statistics of the data.

First, we measured the performance of sensor front ends using an off-chip amplifier and ADC. The off-chip ADC has a 16-bit resolution for an input range of ±10V, a sampling

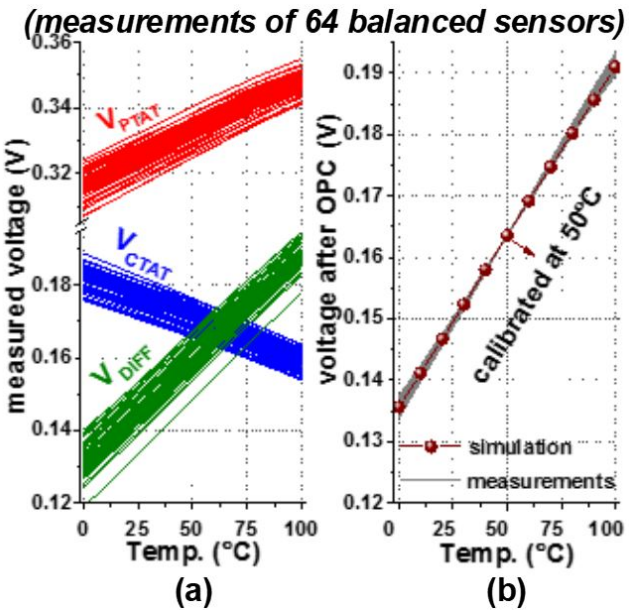


Figure 2.15: The measurements PTAT, CTAT, PTAT-CTAT and after OPC

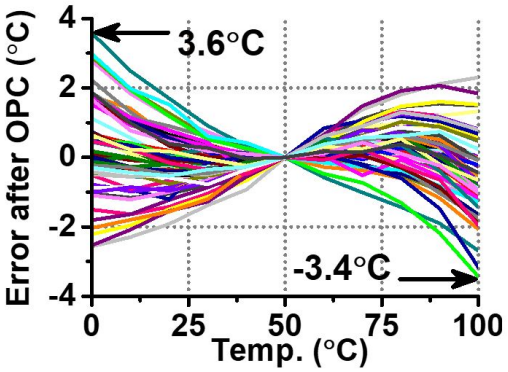


Figure 2.16: The errors of 64 balanced front-end circuits after OPC

rate of up-to 100 kS/s, and an input common-mode voltage of 1.25V. We operated the ADC at 20 kS/s. We swept the temperature from 0 to 100°C with a step-size of 10°C, while measuring the V_{PTAT} and V_{CTAT} of the 64 balanced-sensor front-end designs across 8 chips. fig. 2.15(a) shows the V_{PTAT}, V_{CTAT}, and V_{DIFF}, with the V_{DIFF} being calculated only after we digitized and subtracted the V_{PTAT} and V_{CTAT}. Then, as shown in fig. 2.15(b), we perform OPC, where the slope of V_{DIFF} across temperature is extracted from the SPICE simulations. The OPC-ed V_{DIFF} are then converted to temperatures, and now the accuracy

of the sensor front end, shown in fig. 2.16, can be found. The sensor front ends exhibit an acceptable worst-case error of 7°C (+3.6°C < error < -3.4°C). Worst-case errors of the accuracy-optimized and area-optimized front end designs also are measured and found to be 5.4°C (-0.7 < error < +4.7°C) and 8.8°C (-2.0 < error < +6.8°C), respectively.

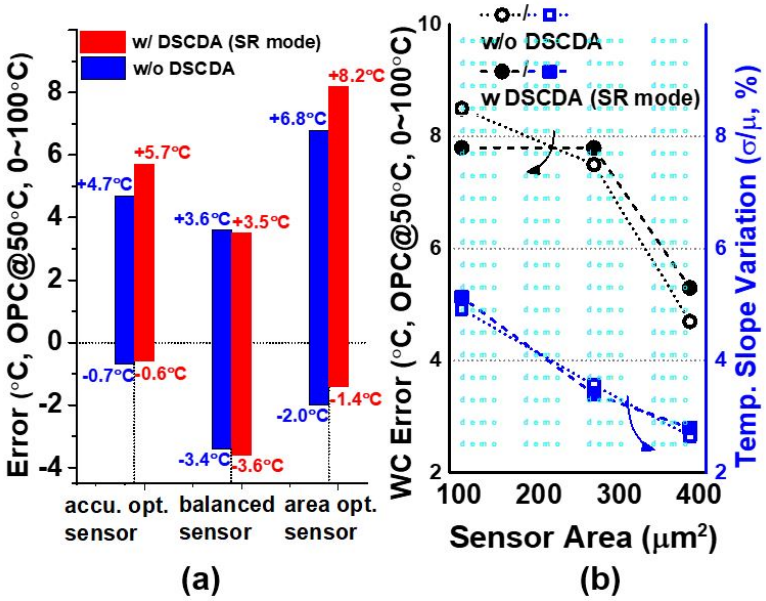


Figure 2.17: The summary of the error performance of three front-end designs after OPC

The sensor front ends are also measured with the on-chip DSCDA and the off-chip ADC. The ISN is configured for the SR mode without external reference voltages. For the different sensor front-end designs we use the different gains in the amplifier - 10x, 12x, and 16x for the accuracy-optimized, balanced, and area-optimized designs, respectively – to amplify the outputs to better match the input-range of the off-chip ADC. In contrast to the case using the off-chip amplifier, the on-chip DSCDA amplifier directly generates V_{DIFF}, which is digitized by the ADC. fig. 2.17(a) shows the measured error of the three types of sensor front ends. When the on-chip amplifier in the SR mode is used worst-case errors are 6.3°C (-0.6 < error < +5.7°C), 7.1°C (-3.6 < error < +3.5°C), and 9.6°C (-1.4 < error < +8.2°C)

for the accuracy-optimized, balanced, and area-optimized designs, respectively.

The measurements show that the sensor front ends with larger footprints can achieve higher accuracy. Increasing the footprint of the sensor reduces worst-case error. As shown by the black curves (black circles) in fig. 2.17(b), the error-spread decreases from 8.8°C to 5.4°C when the footprint of a front end increases from $115\mu m^2$ to $400\mu m^2$. The larger area reduces the variations on temperature slope, represented by the blue curves (blue squares) seen in fig. 2.17(b).

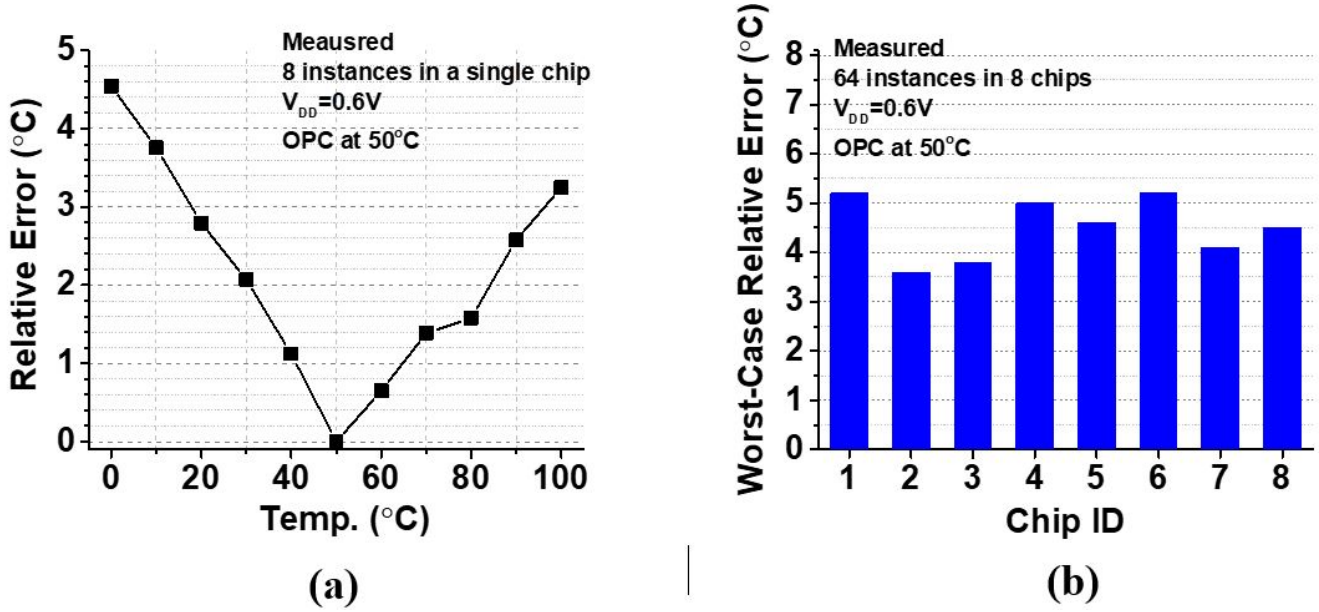


Figure 2.18: The relative error from measurements

We also measure the errors of the front ends relative to one another in a single chip (defined as relative errors [14]). As shown in fig. 2.18(a), the worst-case relative error in a chip is measured to be 4.5°C. We also find the relative errors across eight chips. As shown in fig. 2.18(b), the relative errors are measured to be from 3.6°C to 5.2°C. The accuracy can be improved further when two-temperature-point (20 and 80°C) calibration (TPC) is used. Worst-case errors with the TPC are measured to be 4.7°C, 3.8°C, and 7.8°C for the

accuracy-optimized, balanced, and area-optimized designs, respectively.

2.6.2 VDD Scalability

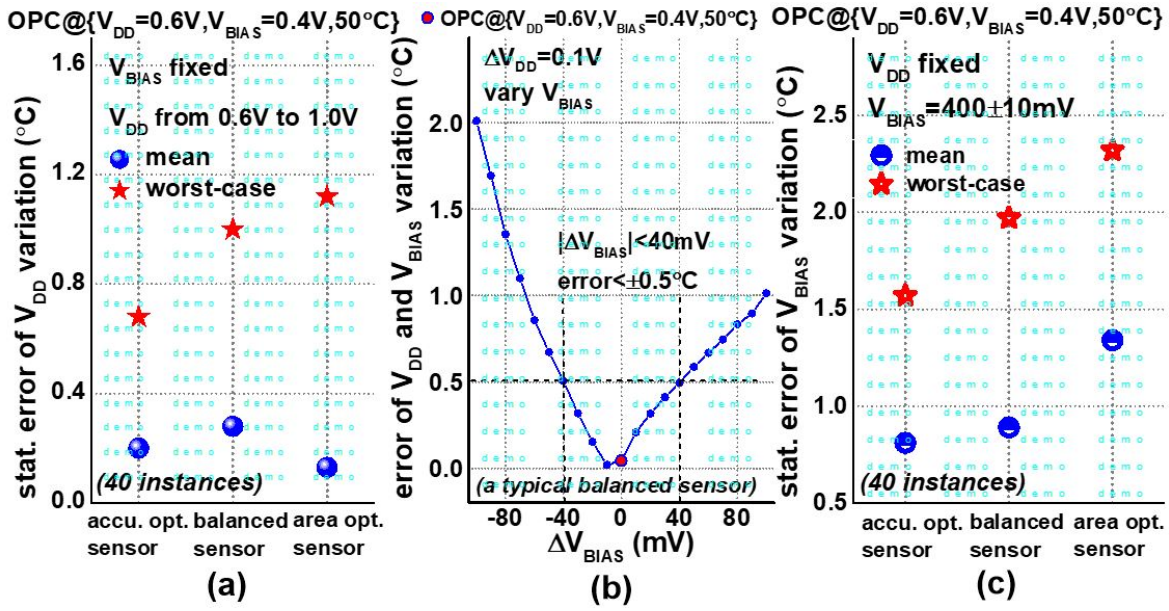


Figure 2.19: VCC scalability measurements

We measured the VDD scalability of the proposed sensor front-end designs. For each front end type we measured 40 instances across 5 chips, while sweeping the VDD from 0.6 to 1V. The OPC is performed at 0.6V, and the calibration settings are used across the entire voltage range. fig. 2.19(a) depicts the error measurements of three types of designs. The average- and worst errors of the balanced front end are 0.3°C and 1.0°C, respectively, across 0 to 100°C and 0.6 to 1V, when they are calibrated at VDD=0.6V. The VB used is 0.4V. fig. 2.19(a) also shows the errors of the other types of front ends. The average- and worst-case errors of the accuracy-optimized front ends are measured to be 0.2°C and 0.7°C, respectively.

We also considered the case where the VB has variations. As shown in fig. 2.19(b), a

typical balanced sensor with VB=0.4V has the LF-PSRR of -79dB and an error of 0.1°C over the voltage scaling from 0.6 to 1V. If VB fluctuates within a range of ±40mV from 0.4V, the LF-PSRR is measured to be <-55dB, which corresponds to an error of 0.5°C for the same voltage scaling. The measurement results for the area, power dissipation, temperature slope, LF-PSRR, and errors of the front ends are summarized in table 2.1. fig. 2.19(c) depicts the error measurements of three types of designs. The average- and worst errors of the balanced front end are 0.7°C and 1.6°C, respectively, across 0 to 100°C and VB varies from 390mV to 410mV, when they are calibrated at VDD=0.6V. fig. 2.19(c) also shows the errors of the other types of front ends. The average- and worst-case errors of the accuracy-optimized front ends, area-optimized front ends, are measured to be 0.8°C and 2.0°C, 1.4°C and 2.3°C respectively.

2.7 Conclusion

We compare the proposed sensor circuits to the state-of-the-art designs. As shown in table 2.2, the balanced front end achieves a 14.3x smaller footprint than the 22nm design in [13], while it exhibits a worst error of less than 7.0°C (-3.4°C < error < 3.6°C), after OPC, across 64 sensors in 8 chips. The worst-case measured error among 8 sensors in a chip across 8 chips is 5.2°C. Our sensor can operate at VDDs from 0.6 to 1V, whereas none of the other designs in the comparison can operate below 1V. The average error incurred by the voltage scaling is 0.3°C. The area-optimized front end has a footprint of $115\mu m^2$, and a worst-case error of 8.8°C (-2.0 < error < 6.8°C) across 64 sensors in 8 chips after OPC. The accuracy-optimized front end exhibits the per-front-end area of $400\mu m^2$ which is still 10x smaller than [13] and a worst-case error of 5.4°C (-0.7 < error < 4.7°C).

Table I. Summary of three sensor front-end circuits

Type	Area (μm^2)	Power[1] (μW) min/max	TC (mV/ºC) Var. (σ/μ, %)	LF-PSRR[2] (dB) DC-Sen. (ºC/V)	Operating scenario	Error (ºC) min./max.
Accuracy optimized	400	0.03/0.85	0.74 2.7%	71 0.4	OPC w/o. Amp	-0.7/4.7
					OPC w. SR Amp	-0.6/5.7
					TPC w/o. Amp	-1.1/2.1
					TPC w. SR Amp	-1.6/3.1
Accuracy-area balanced	279	0.04/0.92	0.57 3.6%	67 0.8	OPC w/o. Amp	-3.4/3.6
					OPC w. SR Amp	-3.6/3.5
					TPC w/o. Amp	-2.4/1.5
					TPC w. SR Amp	-1.9/1.9
Area optimized	115	0.01/0.21	0.72 5.1%	65 0.8	OPC w/o. Amp	-2.0/6.8
					OPC w. SR Amp	-1.4/8.2
					TPC w/o. Amp	-1.6/3.2
					TPC w. SR Amp	-2.3/5.5

[1]The minimum and maximum powers are simulated at 0ºC and 100ºC, respectively; [2]Measurement results without the DSCDA;

Table 2.1: Summary table of three sensors

Table II. Comparisons of temperature sensors for dynamic thermal management techniques

	[13]		[14]	[15]	[17]	[18]	Accu.opt.	Balanced	Area-opt.
Technology	32nm	22nm	32nm	90nm	90nm	160nm	**65nm**		
V_{DD} (V)	1.4-1.8	1.35	1.05	1	1	1.35/1.8V	**0.6~1.0**		
Sensor core	BJT	BJT				TD	**NMOS**		
Power (total, mW)	3.78	1.35	1.6	/	/	3.6mW	**0.36**[6]		
Power (front end, μW)	/	/	/	25	25	/	**0.85**[7]	**0.92**[7]	**0.21**[7]
Area[1] (μm^2)	20000	6100	20000	/	/	4600	/	/	/
Area[2] (μm^2)	11000*	4000*	4000*	48	48	/	**400**	**279**	**115**
Output temperature slope	4.23 [counts/°C]	3.82 [counts/°C]	/	1.8 [mV/°C]	1.8 [mV/°C]	/	**0.74** [mV/°C]	**0.57** [mV/°C]	**0.72** [mV/°C]
Range (ºC)	20~110	-10~110	-10~110	50~125	50~125	-10~125	**0~100**		
Resolution (ºC)	0.19	0.25				0.6	/		
Error[3] (ºC)	/	/	<5	/	/	±6.5	/		
Error[4] (ºC)	<4.5	/	/	/	/	-1.5/+1.5	**-0.7/+4.7**	**-3.4/+3.6**	**-2.0/+6.8**
Error[5] (ºC)	<0.59	<1.5	/	-1~0.8	-1~0.8	/	**-1.1/+2.1**	**-2.4/+1.5**	**-1.6/+3.2**
LF-PSRR (dB)	50	/	/	/	/	/	**71 (typical)**	**67 (typical)**	**65 (typical)**
DC-Line Sensitivity (°C/V)	0.7	/	/	2.2	2.2	/	**0.4 (typical)**	**0.8 (typical)**	**0.8 (typical)**
speed	2kS/s	1.4kS/s	/	/	/	0.9kS/s	**20kS/s****	**20kS/s****	**20kS/s****

[1]Area including read-out circuitry; [2]Area per front end; [3]Error without calibration; [4]Error after OPC; [5]Error after TPC; [6]SR mode; [7]the worst-case power consumption at 100ºC; *Estimated from die photo; **measurement setup; [3,4,5]The error numbers are based on statistics from multiple measurements from recent state-of-the-art works

Table 2.2: Comparison table to the state-of-the-art

We presents three ultra-compact and voltage-scalable on-chip temperature sensor designs to support dynamic on-chip thermal management. The compact footprint and the large voltage-scalability of the designs enable the integration of order-of-magnitude more sensor front ends on a chip at a small additional overhead, enabling dense thermal monitoring in modern VLSI systems.

Chapter 3

In-situ and In-field NBTI sensor for Register Files

3.1 Motivation

In modern deeply-scaled CMOS technologies, transistor aging effects such as bias temperature instability (BTI), hot carrier injection (HCI), and time-dependent dielectric breakdown (TDDB) have been one of the major challenges for maintaining long-term reliability of computing systems [26]. In particular, negative bias temperature instability (NBTI) is one of the most critical aging mechanisms, which can increase PMOS threshold voltage (V_{TH}) at high temperature when PMOS is negatively biased [27, 29, 57]. Such V_{TH} degradation causes digital circuit delay to increase, compromising the maximum clock frequency over chip's lifetime. Recent studies also show that NBTI can be worse for technology scaling [59], confirming its importance in future microelectronics.

NBTI can also degrade robustness of embedded memory circuits [30, 59]. Particularly an SRAM-based register file (RF) is one of the most vulnerable blocks since it often experiences high temperature due to high switching activities and the heat generated by other digital gates around it. Furthermore, the 6- transistor (6T) bitcell typically used in an RF has one of its PMOSs negatively biased (i.e., stressed) during an RF is powered on. Last but not

the least, if a bitcell is not written frequently, one of the PMOSs in the bitcell can receive DC stress, which is more detrimental than the AC one [31].

The degradation of V_{TH} of PMOSs in bitcells can hurt robustness and performance of an RF. It can reduce the static noise margin (SNM), thus worsening data retention voltage (DRV), read VMIN and read access time. Note that the bitcell that undergoes the worst degradation determines the robustness and performance of the entire RF it belongs to.

One of the conventional approaches to mitigate NBTI effects is to increase device sizes. Upsizing a bitcell helps reduce random variations thus renders a narrower pre-aging V_{TH} distribution. However, upsizing can achieve only a limited amount of improvement since it cannot completely eliminate the asymmetric degradations between two pull-up PMOSs in a bitcell (i.e., $\Delta V_{TH,PL}$ vs. $\Delta V_{TH,PR}$). Note that asymmetric degradation largely affects the bitcell's stability. In addition, the amount of upsizing may need to be determined to address the worst-case aging effects, which may result in pessimistically large bitcells. Finally, recent studies show that V_{TH} degradations caused by NBTI is relatively insensitive to initial V_{TH} distribution [32].

To tackle this challenge, we propose techniques to dynamically monitor and decelerate NBTI degradation in an RF with a focus to enable/enhance three critical abilities, namely $in-situ$ monitoring, $in-field$ monitoring, and $post-deployment$ NBTI management [55].

First, we devised an in-situ monitoring technique, i.e., directly sensing V_{TH} of a target PMOS in a bitcell instead of using of replica/canary circuits [33, 35]. Canary circuits can be embedded to tracks the reliability of the target circuits. It is, however, not in-situ, and therefore various mismatches between the target and the canary circuits can degrade monitoring accuracy. V_{TH} degradation caused by NBTI effects (defined as $\Delta V_{TH,stress}$) is a

complex function of various parameters such as process, voltage, and temperature (PVT) variations. Ref. [36] derived the equation for V_{TH} degradation (eq. (1)).

$$\Delta V_{TH-stress} = \left[K_V \sqrt{t_{stress}} + \sqrt[2n]{\Delta V_{TH0}}\right]^{2n} \tag{1}$$

where t_{stress} is stress time, K_V is a parameter that depends on an electrical field and temperature, n is time exponent, and V_{TH0} is an initial V_{TH} process variation from the nominal value. If a replica sensor experiences the different stress condition from target circuits, it cannot accurately track target circuits. Due to this unavoidable mismatch, a designer has to add the margin to replica sensor output for the worst-case mismatch. This indeed renders the results of replica circuits largely pessimistic.

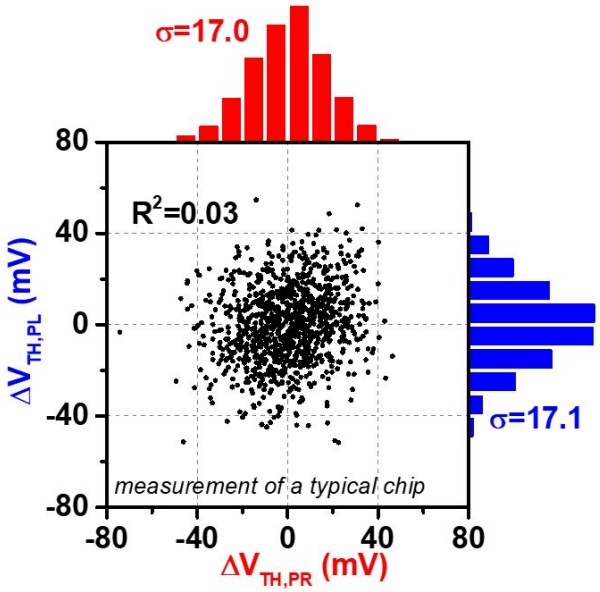


Figure 3.1: Measured V_{TH}s between left and right PMOS in the bitcells from 1Kb RF

In fact, the mismatch between replica circuits and target circuits can be larger than the typical amount of NBTI degradation. fig. 3.1 shows the measured pre-aging V_{TH} variations of right and left PMOSs in each bitcell in a 1-kb 6T-SRAM register file in 65nm CMOS. The

variations of both left and right V_{TH}s, after removing the mean values, vary from -50mV to 50mV with the sigma values of 17.0 and 17.1mV, respectively. This makes the existing replica sensors less accurate in monitoring NBTI in SRAM bitcells.

Second, we have pursued to enhance robustness in monitoring so as to enable in-field monitoring (post-deployment). Robustness against temperature and power supply voltage (VDD) variations is paramount since it is non-trivial to control such parameters in field. Previous works have proposed in-situ sensing techniques [**?**]. However, they are not in-field, sensitive to temperature and VDD variations. This is because they use ring-oscillator frequencies [37–39], bitline currents [**?**, 40] or logic gate delays [42], to monitor NBTI degradation. One may consider to use a built-in self-test (BIST) technique post deployment. For example, we can embed a BIST that can store test patterns, sweep VDD, and check bit flips to find DRV. This requires, however, high-precision voltage regulators to modulate VDD in a fine-grained manner. In our experiment, as a baseline we swept VDD from 0.2V to 0.5V at step of 0.01V. It also requires a considerable number of test iterations to find 0-to-1 and 1-to-0 bit flipping. Another problem of this approach is that the results may be sensitive to temperature.

Finally, we developed a software framework for dynamic reliability management (DRM). During chip's lifetime, we can execute the framework routinely (e.g., every several months) in the maintenance mode, which monitors V_{TH} degradation, evaluates the degree and the progressing rate of NBTI degradation, and analyzes the skew of V_{TH} degradations between two PMOSs in a bitcell. In addition to those monitoring, the framework can also create recover vectors (RV) using the skew information of bitcells. Written into bitcells opportunistically, RVs can partially recover the more-aged one of two PMOSs in each bitcell and thereby decel-

erating SNM and DRV degradation. We prototype test chips, each of which includes a 1-Kb RF with the proposed techniques, in a 65nm CMOS. The measurement results confirm the in-situ and in-field capability with the average error of 19% against temperature variation (20-80C) and that of 21.8% against VDD variations (0.5-1V) in monitoring an NBTI-induced V_{TH} degradation larger than 30mV. Those errors are respectively 4.4X and 3.4X smaller than the estimation of previous work [39]. We also confirm that RVs created based on our monitoring technique can successfully slow down DRV degradation: in our 16-hour accelerated aging experiments, the RFs that store the RVs from the proposed technique exhibit 30mV to 70mV less DRV degradation in average than RFs storing random fixed values. The area overhead of the proposed technique is 27% for a 1-kb RF and 21% for a 4-kb RF.

3.2 Register File Architecture

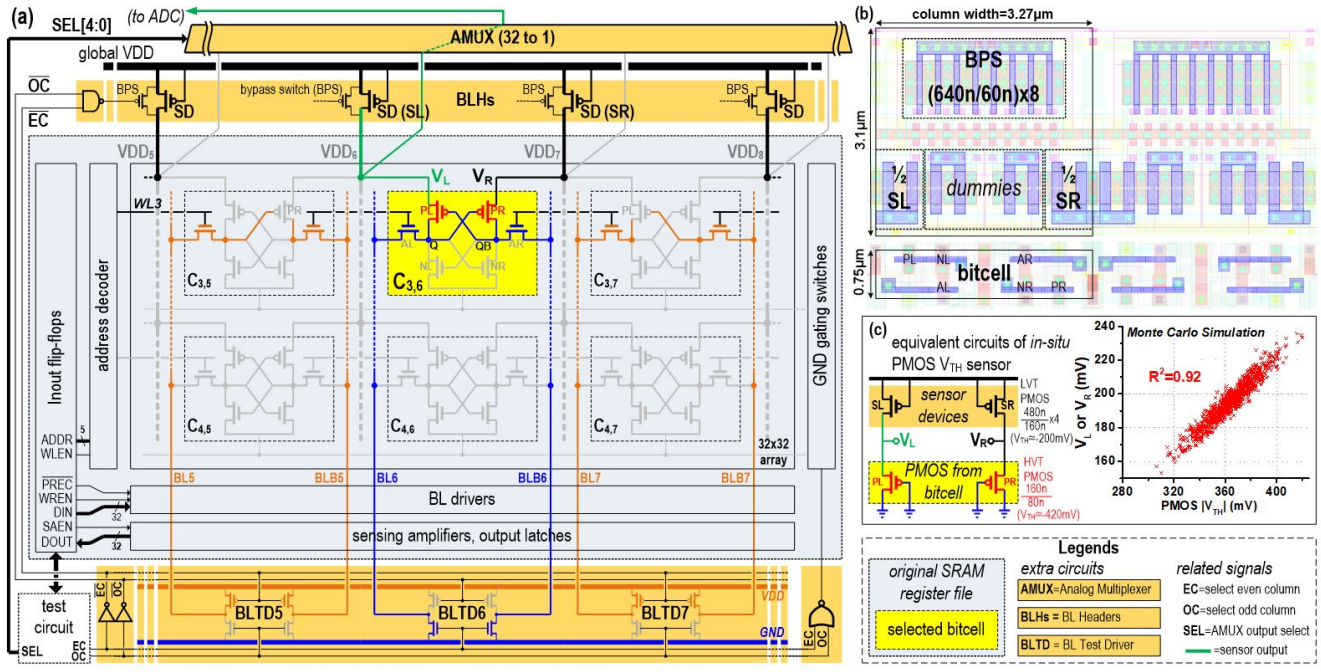


Figure 3.2: RF configured to sensing mode supports in-situ monitor PMOS V_{TH}

fig. 3.2(a) shows the schematic of the RF designed with the proposed monitoring technique. It consists of a baseline RF in the light blue box and circuits added for implementing the proposed technique in the orange boxes. The baseline RF is based on the regular 6T SRAM bitcell. It has a 32x32 bitcell array, inout flip-flops, an address decoder, bitline (BL) circuits (i.e., BL drivers, sensing amplifiers, and output latches) and ground (GND) gating switches to support low-leakage sleep mode [43]. The inout flip-flops hold/generate address (ADDR), word-line enable (WLEN), bitline pre-charge enable (PREC), input data (DIN), and output data (DOUT).

To implement our sensing scheme, we add a bitline header (BLH) and a BL test driver (BLTD) to each column, and one 32-to-1 analog multiplexer (AMUX) shared across all columns. We also add several control signals: EC to select the even-indexed columns of the array, OC to select the odd-indexed columns of the array, and SEL[4:0] to select one of the 32 inputs in the AMUX. In the normal SRAM operation, EC and OC are set to 0.

The top part of fig. 3.2(a) shows the schematics of the BLHs. Each BLH contains a sensor device (SD), dummy devices and a bypass switch (BPS). The BPS bypasses SD at normal SRAM operation. The bottom part of fig. 3.2(a) shows the schematics of the BLTDs. It can connect the BL and BLB of a column to either V_{DD} or GND based on the states of EC and OC. GND gating switches on the right side of fig. 3.2(a) are turned off upon the assertion of EC or OC. The AMUX on the top of fig. 3.2(a) can select one of the sensor outputs, i.e., one of the vertical 32 V_{DD} rails denoted as V_{DD1} to V_{DD32}, and feed it to an off-chip analog-to-digital converter (ADC).

By exercising those additional circuits, we can transform a selected 6T SRAM bitcell into a pair of sensor circuits whose outputs track V_{TH}s of PMOSs in a bitcell. fig. 3.2(c)

schematically shows the transformation of a bitcell $C_{3,6}$ as an example (the bitcell is highlighted in yellow in fig. 3.2(a)). The first step of the transformation is to assert EC and de-assert OC (because the bitcell $C_{3,6}$ is located at an even-indexed column). This disables the BPS and thus connects the power supply rails of the $C_{3,6}$ (V_{DD6} and V_{DD7}) to the main supply rail (V_{DD}) only through the SDs (use SL and SR for identification). This also makes the BLTDs to pull the BLs and BLBs of the even-indexed columns to GND and pull those of the odd-indexed columns to V_{DD}. The EC assertion (or OC assertion) also makes the GND gating switches to float the GND of bitcells.

The second step is to assert WL3 so as to turn on the access transistors (AL and AR) of the $C_{3,6}$. As the BL and BLB of the $C_{3,6}$ is connected to GND through BLTD6, the drain and gate nodes of the PL and the PR in the $C_{3,6}$ also become GND. It is noteworthy that this turns off the PR of the adjacent bitcell C3,5 and the PL of the another adjacent bitcell C3,7 since their gates receives V_{DD} from BLTD5 and BLTD7. This reduces the leakage from the bitcells C3,5 and C3,7 to the V_{DD6} and V_{DD7} rails and improve monitoring accuracy (further discussed in Sec. II. E). Nodes Q and QB can be successfully pull down to GND since AL and AR are much stronger turned on than PL and PR. Monte Carlo simulations at 80°C, 0.6V V_{DD} indicated that the maximum voltage (3σ) at Q or QB is 0.5mV at TT corner and 1.7mV at FS corner. NMOS access transistors are turned on only during read and write operations and thus undergo little BTI aging.

With these two steps, the effective circuits of the bitcell $C_{3,6}$ become a pair of sensor circuits in fig. 3.2(c). The vertical supply rails V_{DD6} and V_{DD7} become the output of the circuits, denoted as V_L and V_R. The outputs V_L and V_R become proportional to the V_{TH} of the DUT PMOSs: PL and PR, respectively. In the next two subsections, we will discuss the

circuit design and operation of this effective circuit.

3.3 In-situ PMOS V_{TH} Sensor

3.3.1 Sensor Circuits

The post-transformed effective circuit (fig. 3.2(c)) consists of a low-V_{TH} PMOS SL (or SR) biased at zero-V_{GS} and a high-V_{TH} PMOS PL (or PR) configured as a diode. The PL (or PR) is minimum-sized as in a typical SRAM bitcell design. We set the width and the length of SL (or SR) to 1.92μm and 0.16μm, respectively, to optimize temperature sensitivity.

fig. 3.2(b) shows the layouts of BLHs and bitcells of two adjacent columns. BLHs and bitcells have identical column pitch of 3.27μm. The BPS, located in the top of BLH, is a high-V_{TH} PMOS whose length is 60nm and width is 5.12μm. We size it for a balance between two competing objectives, namely minimizing IR drop during regular SRAM operation and minimizing its leakage during NBTI sensing operation.

Each BLH contains a half of SL (1/2 SL), a half of SR (1/2 SR), and dummies. The other halves of SL and SR are located in the SHs in the adjacent columns. The dummies can improve parameter matching between 1/2 SL and 1/2 SR. Two V_{DD} rails run vertically on both sides of a BLH. The area of a BLH is $10.1\mu m^2$.

Our bitcell follows the NMOS-centered layout for separating the VDD rails of PMOSs in a bitcell. Ref. [44] shows that the NMOS-centered layout has little area penalty as compared to the conventional PMOS-centered layout. The area of the bitcell is $2.45\mu m^2$ drawn under the logic rule.

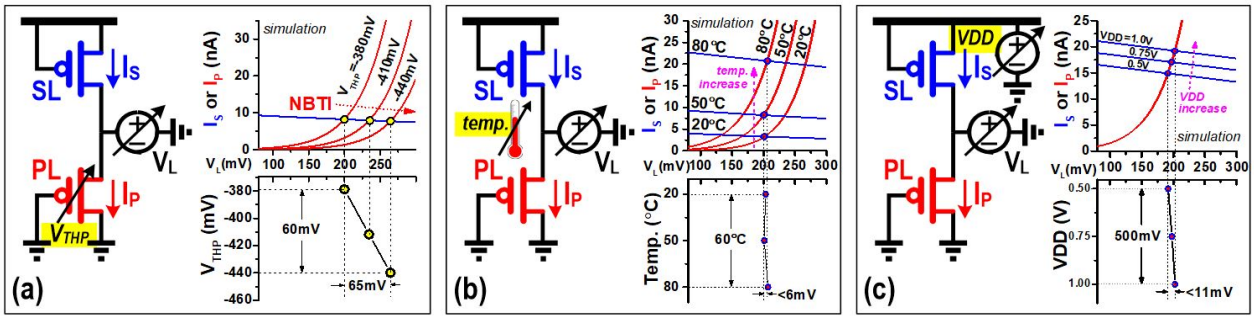


Figure 3.3: Demonstration of V_{TH} sensing capability and robustness

3.3.2 Sensor Operation Principle

In this section, we explain the operation of the post-transformed effective circuit (fig. 3.2(c)). First, we investigate the sensor's ability to track PL's (or PR's) V_{TH} change. We first characterize the current-voltage (I-V) curves of both transistors PL and SL at a nominal condition (20°C, 0.5V, no aging). By modulating V_L from 50 to 300mV, as shown in fig. 3.3(a), we can plot the drain current of SL (IS, blue) and PL (IP, red) as a function of V_L. Since SL and PL are connected in series, the intersection point of the two currents (i.e., I_S=I_P) represents the output of the sensor. Now, to emulate NBTI degradation, we change the V_{TH} of PL to -380mV, -410mV and -440mV and re-plot IPs. As shown in fig. 3.3(a), the V_{TH} modulations move the intersection of IH and IP. To be clearer we plot V_L as a function of V_{THP}, where we can observe that V_L tracks V_{THP} almost linearly. We also perform 1-k Monte-Carlo simulations to find the correlation between V_L (or V_R) and the V_{TH} of the PL (or PR). As shown in fig. 3.2(c), they are strongly correlated with 0.92 R^2 value, implying that the sensor outputs can well track NBTI incurred V_{TH} degradation.

Second, we investigate the output robustness against temperature variations. We re-plot both IS and IP across three temperatures (20, 50 and 80°C). As shown in fig. 3.3(b), both I_S and I_P increase with temperature. This makes the intersection point of I_S and I_P move

upward, but their projections to X-axis, i.e., sensor output voltage, change little, only by 6mV.

Finally, we investigate the output robustness against supply voltage variation. We sweep V_{DD} from 0.5V to 1.0V and re-plot I_S and I_P. While I_S and I_P change, as shown in fig. 3.3(c), again V_L changes little, by 11mV.

The size of PU transistors (PL, PR) is determined firstly according to bitcell implementation. For example, high-density bitcells have minimize size of PU (e.g. 120n/60n) while low leakage bitcells have a larger length for PU (e.g. 120n/80n). The sensor (SD) is tuned at second step to the optimal size to achieve good sensor gain, robustness against temperature and V_{DD} variations. Several combinations between PU-SD with simulated specs are listed in table 3.1.

Table I. Different PU-SD combinations

PU size (nm)	SD size (nm)	ΔV_{OUT} due to temperature (20~80°C)	ΔV_{OUT} due to V_{DD} (0.6~1.0V)	sensor gain (ΔV_{TH}/ ΔV_{OUT})
160/80	[480/160]x4	5.2mV	17.5mV	0.88
120/60	[380/160]x4	5.3mV	14.9mV	0.83
120/80	[440/160]x4	5.5mV	16.6mV	0.85

Table 3.1: Different PU-SD combinations

In addition to the above simulation-based investigation, we also derive an analytical expression for V_L. Based on well-know sub-threshold channel current equation, we can calculate I_H with V_{GS}=0 and I_P with V_{GS}=-V_L. By equating I_S and I_P, we can derive V_L as eq. (2)

$$\begin{cases} \mathrm{V_L} = |\mathrm{V_{TH,PL}}| - \left[\frac{\mathrm{n_{PL}}}{\mathrm{n_{SL}}}|\mathrm{V_{TH,SL}}| + \mathrm{T}\frac{\mathrm{n_{PL}k}}{\mathrm{q}}\ln\left(\frac{\beta_{\mathrm{SL}}}{\beta_{\mathrm{PL}}}\cdot\frac{\mathrm{n_{SL}}-1}{\mathrm{n_{PL}}-1}\right)\right] \\ \beta_{\mathrm{i}} = \mu_{\mathrm{i}}\mathrm{C'_{oxi}}\frac{\mathrm{W_i}}{\mathrm{L_i}} \end{cases} \tag{2}$$

, where k is the Boltzmann constant, q is the electron charge, and i is an index to represent either PL or SL. In addition, based on [46], V_{TH} can be modeled as eq. (3)

$$\mathrm{V_{TH}} = \mathrm{V_{TH0}} + (\mathrm{K_T} + \mathrm{K_{TBS}V_{BS}})\left(\frac{\mathrm{T}}{\mathrm{TNOM}} - 1\right) \tag{3}$$

, where V_{TH0} is the threshold voltage at T=T_{NOM}, K_T is the temperature coefficient, K_{TBS} is the body-source voltage related temperature coefficient, and T_{NOM} is nominal temperature. For both SL and PL, VBS equals zero. Then, from eq. (2) and eq. (3), we can derive the equation for V_L as below.

$$\begin{aligned} \mathrm{V_L} = |\mathrm{V_{TH0,PL}}| + \mathrm{T}\left[\frac{\mathrm{K_{T_{T,PL}}}}{\mathrm{T_{NOM}}} - \frac{\mathrm{n_{PL}K_{T,SL}}}{\mathrm{n_{SL}T_{NOM}}} - \frac{\mathrm{n_{PL}k}}{\mathrm{q}}\ln\left(\frac{\beta_{\mathrm{SL}}}{\beta_{\mathrm{PL}}}\cdot\frac{\mathrm{n_{SL}}-1}{\mathrm{n_{PL}}-1}\right)\right] \\ - \left(\mathrm{K_{T,PL}} - \frac{\mathrm{n_{PL}}}{\mathrm{n_{SL}}}\mathrm{K_{T,SL}} + \frac{\mathrm{n_{PL}}}{\mathrm{n_{SL}}}|\mathrm{V_{TH0,SL}}|\right) \end{aligned} \tag{4}$$

Above equation well describes the characteristic of the output V_L. First, V_L is a linear function of V_{THPL}. Second, we can reduce V_L's temperature dependency by optimizing βSL values; in our design we can size SL as PL is fixed as a part of a bitcell. Third, the lack of V_{DD} term implies that V_L is robust to V_{DD} variations for the first order. Note that V_L can include the effects from various aging mechanism (e.g., HCI that degrades mobility) other than NBTI. Still the NBTI is considered the leading mechanism in SRAM circuits in the modern highly-scaled technologies [57]. Thus, we treat the operation of our circuits for

monitoring NBTI for the simplicity.

3.3.3 Differential Reading

Although eq. (4) implies that the output of the V_{TH} sensors is robust against temperature and V_{DD} variations, various second-order effects can compromise the robustness. Process variation can further deviate the output (V_L) of the proposed V_{TH} sensor. However, we are mostly interested in the differences of outputs. For example, to evaluate the amount of NBTI degradation, we are interested in the difference of sensor output (ΔV_L) between pre- and post-aging moments. Additionally, to create RV (see Sec. V for details) we are interested in the difference of the V_{TH} degradations of PL and PR in a bitcell (ΔV_{L-R}). Using such differences of values can significantly relax the matching requirements between various devices, for example, between SL and PL, between SR and PR, and between SL and SR.

To evaluate the robustness improvement of such differential readings over process, voltage, and temperature variations, we derive ΔV_L as eq. (5).

$$\Delta V_L = |\Delta V_{\mathrm{TH0,PL}}| + \Delta \mathrm{T} \left[\frac{\mathrm{K_{T,PL}}}{\mathrm{T_{NOM}}} - \frac{\mathrm{n_{PL}K_{T,SL}}}{\mathrm{n_{SL}T_{NOM}}} - \frac{\mathrm{n_{PL}k}}{\mathrm{q}} \ln \left(\frac{\beta_{\mathrm{SL}}}{\beta_{\mathrm{PL}}} \cdot \frac{\mathrm{n_{SL}} - 1}{\mathrm{n_{PL}} - 1} \right) \right] \tag{5}$$

We assume that NBTI increases $V_{TH0,PL}$ by $\Delta V_{TH0,PL}$. This equation shows that the differential reading can remove the impacts of some of process variations. As shown in eq. (4), V_L has the third term that is sensitive to various device parameters. ΔV_L does not have this term. Note that eq. (5) contains the device parameters of SL but those parameters, e.g., $V_{TH0,SL}$, $K_{T,x}$ and n_x, barely change between pre- and post-aging moments since SL

is active only during the sensing operation and thus hardly undergoes NBTI degradation. This equation eq. (5) also shows that the differential reading can improve robustness against temperature variation since the temperature difference between pre- and post-aging measurements (ΔT) is smaller than absolute temperature (T) in K by 5.8X (=(273+80)/(80-20)). We also derive ΔV_{L-R} as eq. (6).

$$\Delta \mathrm{V_{L-R}} = \Delta \mathrm{V_L} - \Delta \mathrm{V_R} = |\Delta \mathrm{V_{TH,PL}}| - |\Delta \mathrm{V_{TH,PR}}| + \underbrace{\frac{\Delta \mathrm{T}}{\mathrm{T_{NOM}}} \cdot \mathrm{o}\left(\Delta \mathrm{K_{T,LR}}, \Delta \mathrm{n_{LR}}\right)}_{\text{sensing error}} \qquad (6)$$

As SL and SR are identical and large, their device parameters, e.g., K_T and n, can be cancelled each other in eq. (6), improving robustness. However, PL and PR are small and their mismatch, as denoted as ΔK_T and Δn, remains.

3.3.4 Sensor Gain (V_{TH} Sensitivity)

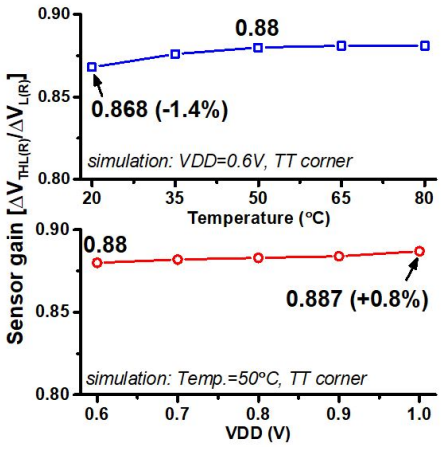


Figure 3.4: The gain of the V_{TH} sensor across temperature and supply voltage variations

To achieve accurate monitoring, we need sensor gain (defined as the ratio of NBTI-induced

Table II. Sensor gain ($\Delta V_{TH}/\Delta V_{OUT}$) across different process corners

DUT P (high V_{TH}) \ sensor SD (low V_{TH})	Fast	Typical	Slow
Fast	0.85 (-3.4%)	0.87	0.88
Typical	0.86	0.88 (100%)	0.88
Slow	0.88	0.89	0.90 (+2.2%)

Table 3.2: Sensor gain across different process corners

V_{TH} change to the sensor output voltage change) that is stable across process, temperature, and voltage variations. As shown in fig. 3.4, the sensor gain is found to be 0.88 at a nominal condition (50°C, 0.6V) and exhibits a small amount of variability across temperature and V_{DD} variations. In the worst-case temperature condition (20°C), the sensor gain is 0.868, which is 1.4% smaller than that in the nominal condition. In the worst-case V_{DD} condition (1.0V), the gain is 0.887, which is 0.8% larger than that in the nominal condition. table 3.2 shows the simulation results across different process corner combinations for sensor (SD) and DUT (PL or PR). The worst-case sensor gain variability is found to be 5.6% between FF and SS.

3.3.5 Leakage Reduction

In fig. 3.2(a), the net V_L (or V_R) is connected to 2x32 PMOSs from 2x32 bitcells across column5 and column6. The key challenges in designing the proposed sensor circuits is to suppress leakages from adjacent bitcells, e.g., bitcells at column5 and bitcells at column6 except the selected one ($C_{3,6}$). The leakage of those PMOSs can degrade the robustness against temperature and V_{DD} variations as well as the tracking ability of NBTI degradation of the proposed sensor circuits.

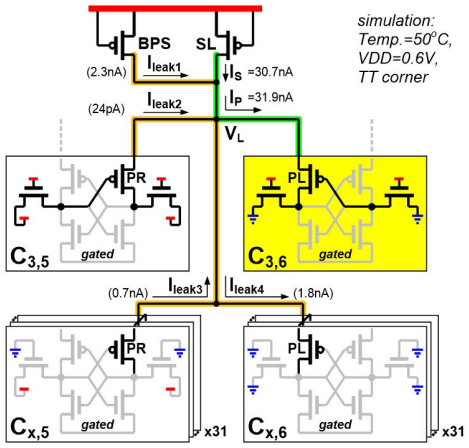


Figure 3.5: Sensor leakage snapshots

fig. 3.5 describes the leakages from 64 PMOSs when the bitcell $C_{3,6}$ is transformed to sensor circuits. In the ideal condition, the current of SL (I_S) should be identical to the current of PL (I_P). However, the other PMOSs inject leakage to or sink leakage from the node V_L, making I_S and I_P different (30.7nA and 31.9nA, respectively). The difference is attributed to: (i) the subthreshold and gate leakages of off-state BPS (I_{leak1}= 2.3nA); (ii) the leakage of the PMOS of the neighbor bitcell $C_{3,5}$ (I_{leak2}=24pA).

The remaining 62 PMOSs in the bitcells in the two columns (i.e., $C_{x,5}$ and $C_{x,6}$, where x is from 0 to 31 except 3) can also sink leakage (I_{leak3} and I_{leak4}) from V_L. To reduce them, we can utilize GND gating switches. Simulations show that this can reduce I_{leak3} and I_{leak4} down to 0.7nA and 1.8nA, respectively. In summary, the total leakage that perturbs V_L (i.e., I_{leak1}+I_{leak2}+I_{leak3}+I_{leak4}) is 1.2nA at 0.6V and 50°C. This total leakage is 3.8% of I_P and thus their impact on V_L is minimal.

We also investigate the ratio of the perturbing leakage to I_P across temperatures and V_{DD}s. fig. 3.6(a) indicates the ratio becomes the smallest at high temperature and low V_{DD} and the largest at low temperature and high V_{DD} (worst-case condition). fig. 3.6(b) shows

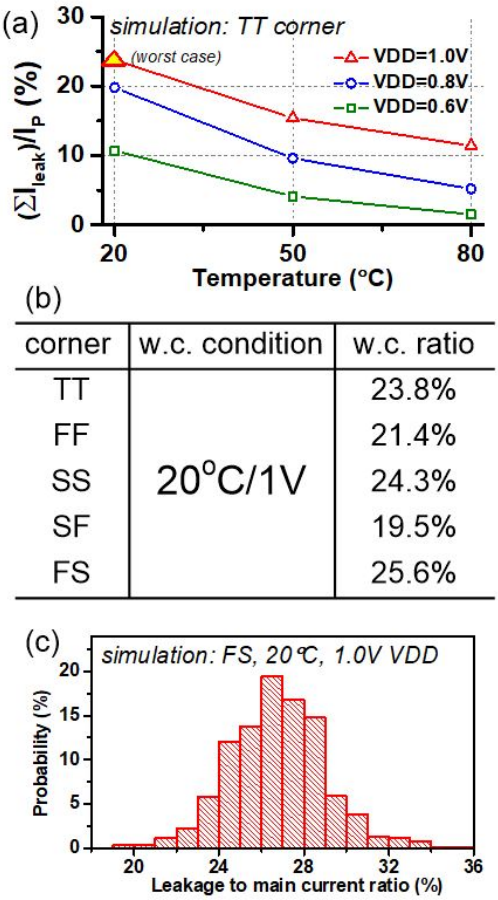


Figure 3.6: Sensor leakage simulation

worst-case ratio at different corners and at FS, the ratio is 25.6%. fig. 3.6(c) shows the Monte Carlo simulation at this corner, which provides a pessimistic estimation of the ratio.

3.3.6 Noise

Various noise sources, e.g., random telegraph noise (RTN), can affect the voltage output of the proposed circuits. Ref. [56] introduced a comprehensive study of the relations between RTN and BTI. In our work, we rely on the averaging technique to mitigate the impact of various noise (in addition to the differential reading for slow and static noise sources); for example we took 1-k samples and then use the mean value of the samples in most of our

measurement. This is because we envision relatively slow monitoring in our work.

3.4 Silicon Prototyping

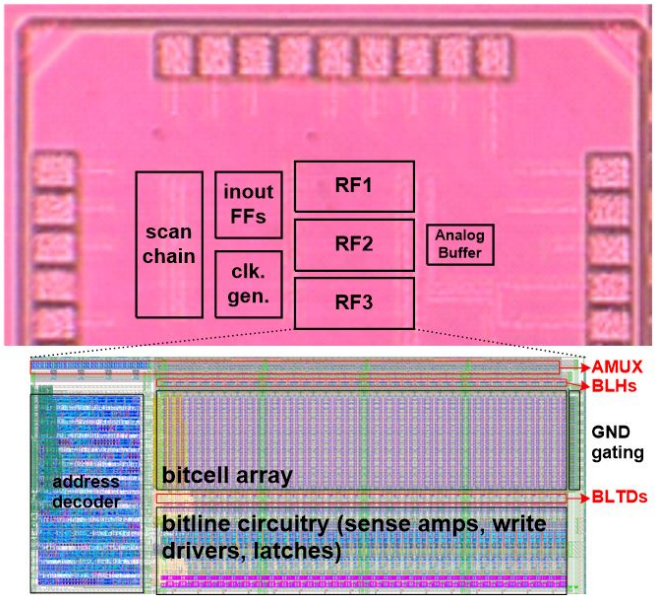


Figure 3.7: Chip microphotograph and the layout of a register file with the proposed technique

We prototyped test chips, each of which contains three 32X32b RFs with the proposed technique in a 65nm CMOS. fig. 3.7 shows the chip photo and the pre-silicon layout snapshot of one of the RFs. The total area of the RF is 7,070μm^2. The area of the baseline RF is 5,580μm^2. As shown in fig. 3.8, the additional circuits for the proposed sensing technique include an analog MUX, BL test drivers, BL headers, and the area increase of the controller due to the additional control signals, and take an additional silicon area of 1,490μm, marking an area overhead of 27%. When implement our technique to a larger RF, the sensing circuits (SD, BLPD and AMUX) only increase in one direction (along wordline). This makes the overhead increasing in square-root of memory size. For example, we estimate the overhead reduced to 21% (2,280μm^2 out of 10,660μm^2) for a larger 64X64b RF.

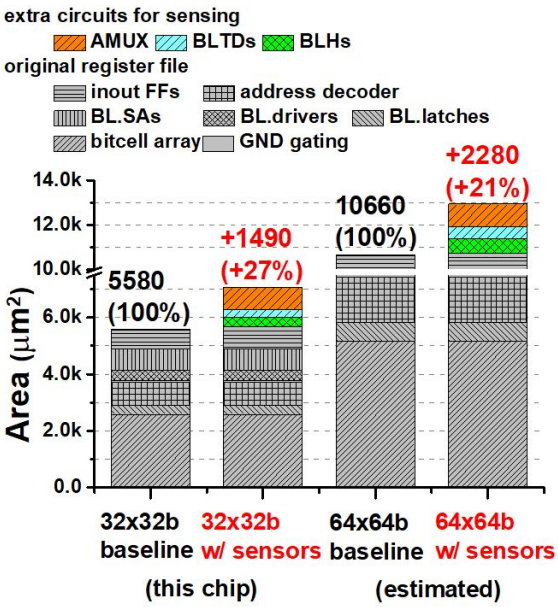


Figure 3.8: Area overhead of the proposed technique

3.5 Measurements

3.5.1 Monitoring NBTI Degradation

We first examine the ability and evaluate the performance of tracking NBTI degradation of the proposed sensors. For this, a chip undergoes an accelerated aging test (AAT). The setup of the testing is as follow. The chips are packaged in QFN and each of them is mounted on an FR4 PCB. We placed the PCB and the chip in a temperature chamber (TestEquity 107), which is communicated to a computer running National Instruments LabVIEW software, via the high-temperature tolerant cables and connectors. Then, we wrote the RF with a fixed, 0/1 randomly distributed pattern and then stressed at 1.6V and 125°C for 16 hours. The AAT aims to mimic the degradations after long-time use of chips.

In the measurement we have used an off-chip ADC which meet the relatively low to moderate bandwidth, throughput, and resolution (0.25 1mV V_{LSB}). The sensor output

(V_{SEN}) needs to travel through a potentially long wire. As our scheme does not envision high-speed aging monitoring, the wire RC delay is tolerable. Also, the sensor produces voltage output and drives a capacitive load. Therefore the current draw is low, making it immune to the large resistance of long wires. Noise from nearby digital systems, e.g., coupling noise, can be mitigated by shutting down their operation during the maintenance period. Shielding would be also helpful to mitigate digital noise. Finally, we sample signal multiple times and use the average value, further mitigating the noise impact.

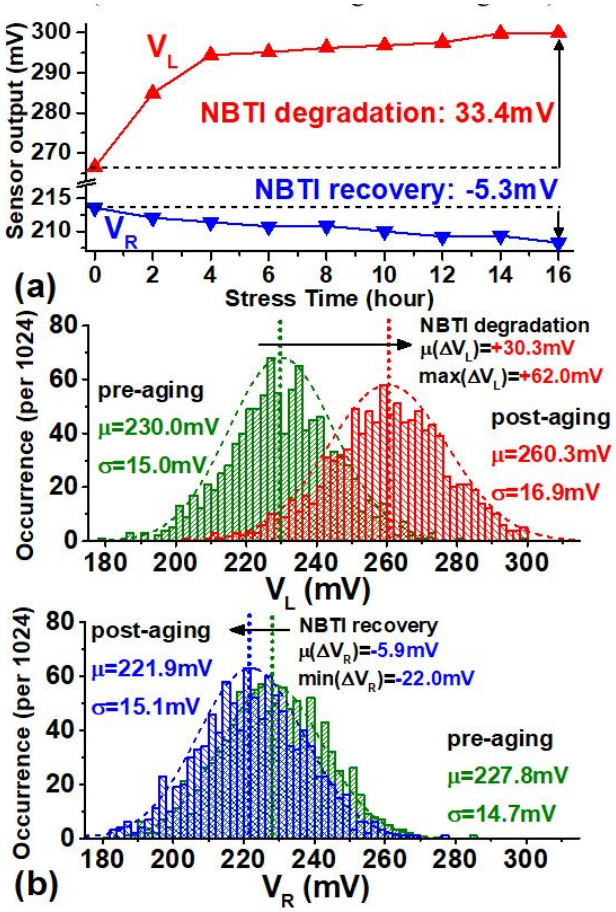


Figure 3.9: Measurements of NBTI-related V_{TH} degradation/recovery

We measured the sensor outputs every two hours. fig. 3.9(a) shows the measured sensor outputs for a typical bitcell which stores Q=1 and QB=0 along AAT. The sensor outputs

(V_L and V_R) can well track V_{TH} degradation and recovery for two PMOSs due to NBTI. QB=0 makes the PL in the bitcell to undergo NBTI degradation, which is tracked by the change of V_L of 33.4mV. On the other hand, the PR in the bitcell experiences a V_{TH} recovery of 5.3mV, indicated by the change of V_R. fig. 3.9(b) shows the measured the distributions of V_L and V_R values across 1024 bitcells in an RF after first step of AAT. Because the chip is fresh, the amount of recovery (shift between blue and green histograms) is smaller than the amount of degradation (shift between red and green histograms).

3.5.2 Robustness in Monitoring

We also experimentally verify the robustness of sensor outputs against temperature and V_{DD} variations. We measure the pre- and post-aging sensor outputs while sweeping temperature from 20 to 80°C across nine RF instances. We define a metric called reading error as eq. (7).

$$ERROR = \frac{V_{ERR}}{V_{ERR} + V_{NBTI}} \times 100\% \tag{7}$$

, where V_{ERR} is the worst-case output voltage change across temperature variations during post-aging measurement and V_{NBTI} is the difference of output voltages of pre- and post-aging measurements at a nominal temperature (50°C). fig. 3.10(a) shows the measurements of a typical sensor, exhibiting V_{ERR} is 7.7mV and V_{NBTI} is 33mV. This causes a reading error of 18.9%. We then repeat this measurement for 1024 sensors in a typical RF instance across V_{NBTI}'s of 10mV to 50mV and show the statistical result in fig. 3.10(b). The reading errors decrease with larger NBTI degradations: if monitoring NBTI degradation larger than 30mV,

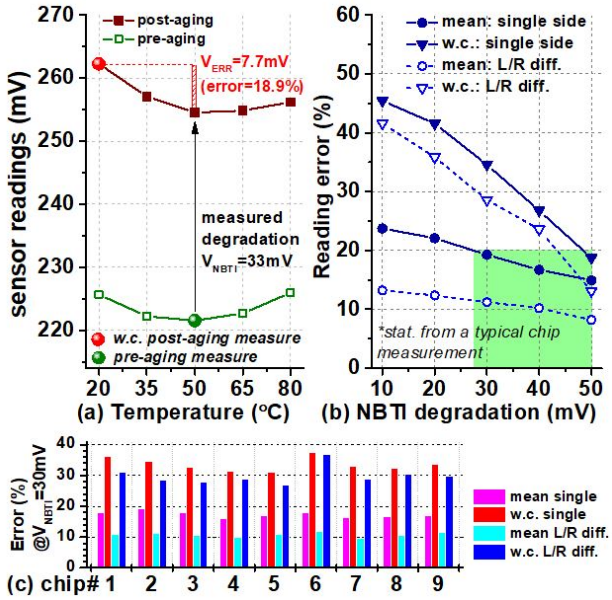


Figure 3.10: Robustness measurements against temperature variations

the proposed sensors exhibit the mean and the worst-case sensing errors of <19% and <35%, respectively.

Additionally, we investigate the output robustness if taking the difference between two sensor outputs of the same bitcell, which we use to create reliable recovery vector. If NBTI degradation is larger than 30mV, the L-R differential reading can achieve the mean and the worst-case sensing errors of <11% and <28%, respectively. We also repeat the similar experiments across nine RF instances. As shown in fig. 3.10(c), the mean sensing error is <20% for monitoring a single-ended output and <12% for monitoring the difference of two outputs from a bitcell, for monitoring NBTI degradation larger than 30mV.

Finally, we investigate the output robustness against V_{DD} variations. We perform the similar experiments by sweeping V_{DD} from 0.6V to 1.0V. As shown in fig. 3.11(a), a typical sensor exhibits a reading error of 20% with V_{ERR} =7.8mV and VNBTI=34.8mV under the worst-case 0.5V V_{DD} variation. Measurements with a typical RF instance show that the

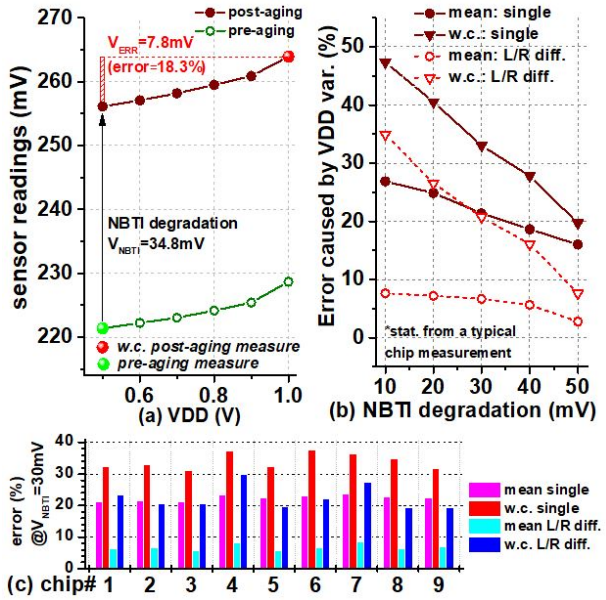


Figure 3.11: Robustness measurements against supply voltage variations

mean and the worst-case errors are <21.8% and <33.6%, respectively, for monitoring NBTI degradation greater than 30mV (fig. 3.11(b)). Again, sensing the difference of V_L and V_R reduces the mean and the worst-case errors down to <7.2% and <20.5%. As shown in fig. 3.11(c), across nine RF instances, we find that the mean error is <23% for a single-ended output and <9% for the difference of two outputs from a bitcell.

3.6 Aging Deceleration Experiment

Previously, techniques such as power gating and bit inversing have been proposed to decelerate aging in SRAM circuits. Power gating is an efficient way to decelerate BTI aging [60], which could recover the V_{TH} of each transistor to its pre-aging value that is set by random and systematic process variation. Bit-inversing proposed to periodically change the logic mapping of a memory block, which can conceptually balance the times that a bitcell stores 0 or 1 [61]. Whereas it is effective for the average aging behavior, it may not be able

to cover the bitcell that undergoes the worst-case aging.

In this section, we will outline our framework that can accurately estimate the polarity of DRV of a bitcell (e.g., 0 if a bitcell can hold 0 more robustly than 1 at low V_{DD}) from the sensor readings. With those polarity estimations, we will show that our framework can create RVs. By storing them in the RF, we can slow down aging and the skew of the left and right side of a bitcell. Since our proposed technique creates a recovery bit for each bitcell, it can target the very worst-case aged bitcell that dictates array-level reliability. The RV does not do this at the cost of the average-case aged bitcells because each of the average bitcells would also receive a recovery bit designed for each of them. The goal of the framework and experiment are to demonstrate the circuit-level feasibility of $in-situ$ aging deceleration (partial recovery), rather than the system-level study nor complete recovery.

3.6.1 Monitoring the Polarity of Data Retention Voltage

To characterize the DRV of bitcells, we use four metrics: DRV0, DRV1, DRVD and ΔDRVD. DRV0 is defined as the minimum V_{DD} that the bitcell can hold '0' robustly; similarly, DRV1 as the minimum V_{DD} for holding '1' robustly. DRVD is defined as (DRV0 - DRV1). Finally, we define ΔDRVD as eq. (8).

$$\Delta DRVD = DRVD_{post} - DRVD_{pre} = (DRV0_{post} - DRV1_{post}) - (DRV0_{pre} - DRV1_{pre}) \quad (8)$$

, where the pre- and post- subscripts represent the moment of bitcell measurement before and after chip's deployment. Presumably, ΔDRVD represents the shift of DRV caused by

NBTI degradation. To estimate ΔDRVD from measurement, we also define ΔV_D as eq. (9).

$$\Delta V_D = (V_{R,post} - V_{R,pre}) - (V_{L,post} - V_{L,pre}) \tag{9}$$

, and verify its relation to ΔDRVD through measurements. Specifically, we write random data (the same 0 and 1 probabilities) in an RF and perform an accelerated aging test at 1.6V, 125°C for 16 hours. We measure DRV0, DRV1, V_L, and V_R before and after the aging test. As shown in fig. 3.12, ΔDRVD and ΔV_D exhibit a good linear relationship. In particular, those bitcells that exhibit the change of DRV polarity (e.g., DRV0pre > DRV1pre but DRV0post > DRV1post) show the 0.667 R^2 value. Those bitcells experiencing small change in DRV show the 0.59 R^2 value.

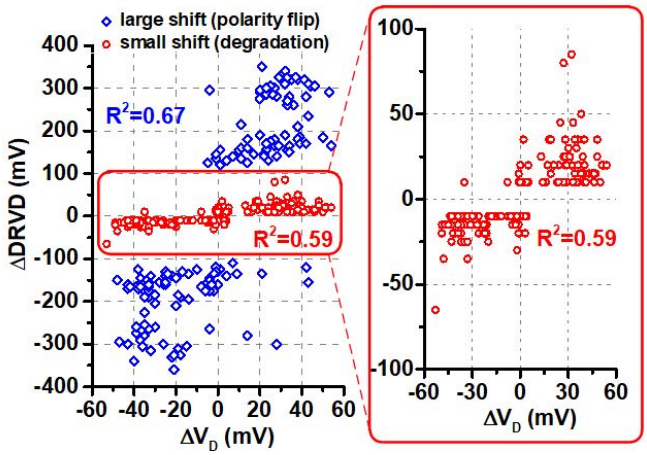


Figure 3.12: Measured ΔV_D as a function of Δ DRVD after AAT

3.6.2 Recovery Vector

It has been demonstrated in previous literature [47–50] that removing the stress can recover a part of wearout in transistors. At circuit level, recover vector (RV) is a word generated by algorithm that stays in an RF and decelerates aging of more-aged PMOSs in

the bitcells. Several previous works proposed techniques to generate RVs based on data-transaction statistics of memory blocks [51–53] and based on power-up state measurement during burn-in tests [54]. However, as an indirect sensing method, the transaction-statistics approach may not decelerate the aging of the worst-case bitcell. This is a critical problem as the worst-case bitcell determines the reliability of the entire RF. On the other hand, the power-up measurement approach does not target in-field operation, and thus may not be robust against temperature and voltage variations.

Indeed, the most accurate way to create RVs is to measure DRV by sweeping V_{DD} until each bit flips, which we call voltage-sweep based RV generation (VRV). However, this approach is also impractical for in-field operation. This approach requires a power supply that can modulate its output from super-threshold to sub-threshold levels in a fine-grained manner. Adding and using such power supply increases hardware and timing overhead sources and consumes very long time (Multilevel V_{DD}s sweeping for both '0' and '1' holding cases). Also, we prefer to use the nominal level of V_{DD} for the other part of systems while exercising VRV, which requires us to create an isolated power grid for the RF.

In this work, we investigate a novel in-situ method, called sensor-based RV generation (SRV), to create RVs based on ΔV_D monitoring. As shown in previous section and fig. 3.12, the sign of ΔV_D strongly relates to that of ΔDRVD. Using this relationship, therefore, we can create RVs.

To measure the performance of our proposed SRV framework, we performed aging experiments. Specifically, we compare three methods to generate RVs, namely our proposed SRV, the most accurate VRV, and the one that uses a fixed random pattern RV. As shown in fig. 3.13, the experiment steps are as follow:

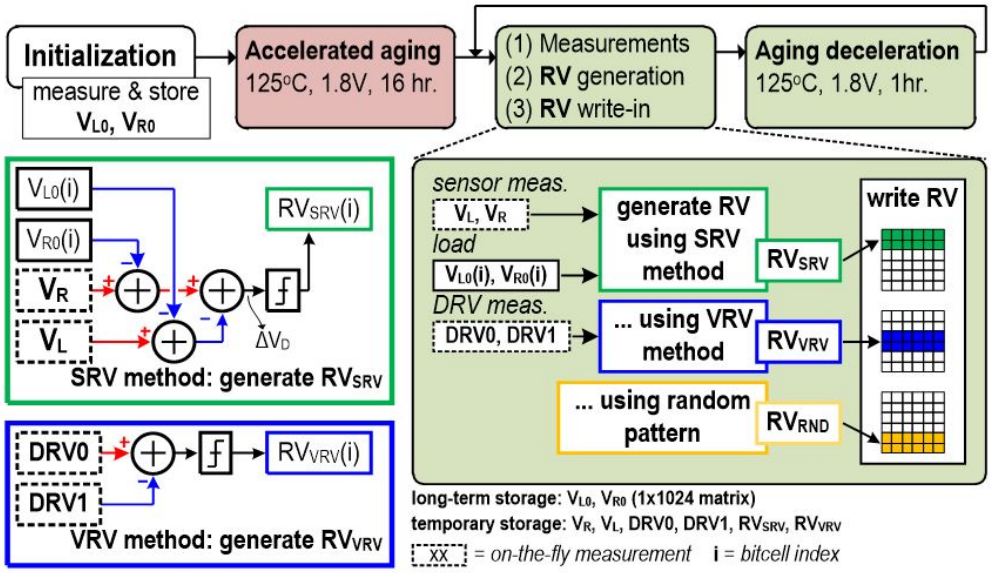


Figure 3.13: The sequence of recovery vector experiment

1. We measure the sensor outputs of bitcells in an RF instance at the nominal condition (50°C, 0.6V). The outputs are stored in two 1024-by-1 vectors V_{L0}(i) and V_{R0}(i), where i is a bitcell index. The text file that stores the vectors is 20kB and it is stored in an off-chip non-volatile memory (e.g. HDD).

2. We write random data in the RF and stress it at 125oC, 1.8V for 16 hours.

3. We generate RVs using the SRV method (defined as RV_{SRV}). We transform bitcells into the proposed sensors and read the outputs. To generate the current RV_{SRV}(i), we compare the current (V_L, V_R) and the initial (V_{L0}(i), V_{R0}(i)) sensor outputs. After finding RV_{SRV}, only (V_{L0}(i), V_{R0}(i)) are retained; others are discarded.

4. We generate RVs using the VRV method (defined as RV_{VRV}). We first measure DRV0 and DRV1 of bitcells by reducing V_{DD} from 1V to 50mV in the step of 5mV. Every step, we check if each bitcell flips. We calculate RV_{VRV} with the found DRV0 and

DRV1.

We write the RV_{SRV}, RV_{VRV}, and a random RV (RV_{RND}) in the three different sections of the same RF. Then we have the chip for 1 hour at 125oC and 1.8V. We repeat above steps 3 to 5 for eight times.

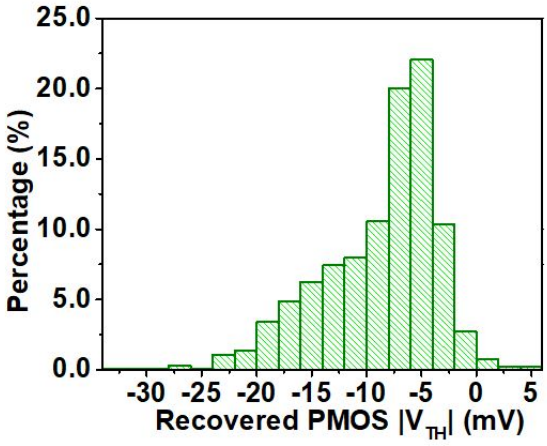


Figure 3.14: PMOS V_{TH} is recovered through RV_{SRV}

We first examine the effectiveness of recovering PMOS V_{TH} by decelerating with RV_{SRV}. We calculated and plotted the difference between sensor outputs that measured at the end step of AAT and after the first step of applying RV_{SRV}. fig. 3.14 shows the histograms of both V_L and V_R shift.

fig. 3.15(a) shows V_L-V_{L0} and V_R-V_{R0} measurements of a bitcell during the experiment. During the initial aging phase (yellow background), the bitcell stores 0. This makes the PR of the bitcell stressed and thus increases V_R by 16mV. On the other hand, PL recovers and V_L slightly decreases by 3mV. After the initial aging phase (i.e., at step 4), we start to calculate RV_{SRV}, which is 1 at the step 4 (fig. 3.15(b)). This RV_{SRV} is written on the bitcell. At the step 6, we find that the RV_{SRV} significantly recovers DRV degradation. Then we continue to generate and use RV_{SRV}, which roughly alternates between 0 and 1.

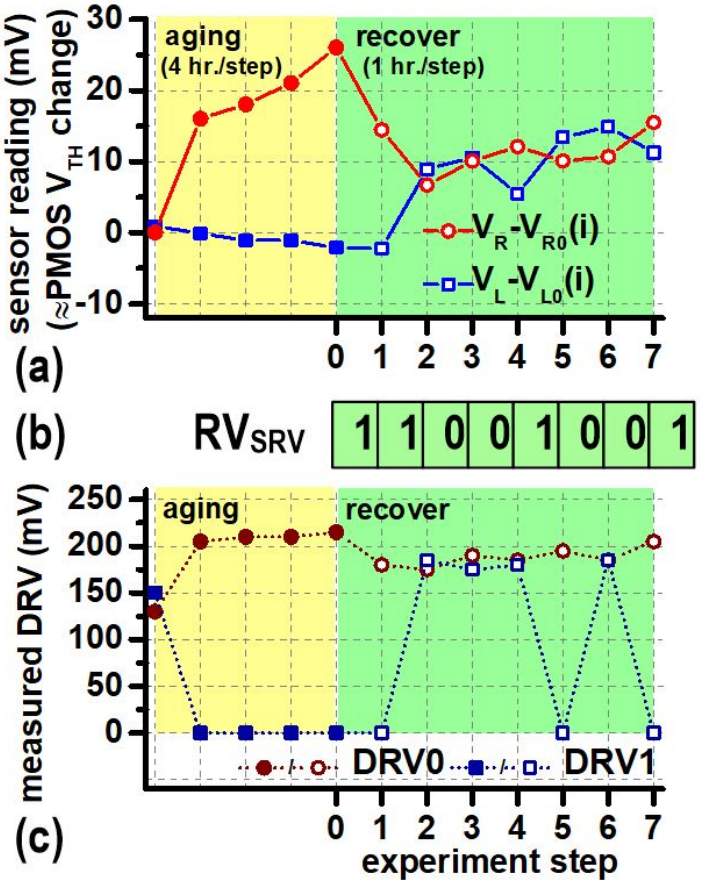


Figure 3.15: Measurements during aging deceleration experiment

This implies the RV_{SRV} keeps balancing the strengths of PL and PR. fig. 3.15(c) shows the measured DRV0 and DRV1 of the bitcell during this experiment. During the aging phase (yellow background), DRV0 increases by 80mV. The RV_{SRV} loaded at step 4 reduces DRV0 by 30mV after two steps. We confirm the similar results across all the bitcells in an RF. As shown in fig. 3.16, our proposed RV_{SRV} reduces V_{L-R} of bitcells. The number of bitcells with $|V_{L-R}| > 40\text{mV}$ reduces from 35 to 4.

We experimented on the RVs and the DRV of an entire RF, i.e., the DRV of the worst-case bitcell. In the similar aging experiment, we generate and use RV_{SRV}, RV_{VRV}, and RV_{RND}, each for one third of an RF. As shown in fig. 3.17(a), the bitcells receive RV_{SRV} exhibits

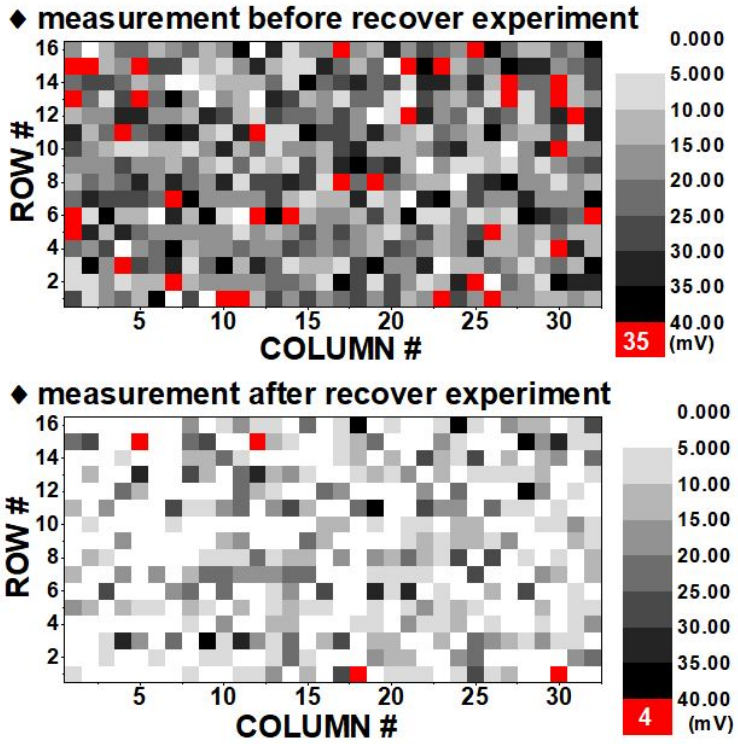


Figure 3.16: PMOS aging skews of bitcells before and after experiment

210mV RF-wide DRV, which is 10mV worse than the DRV of the group that receives RV_{VRV} and 30mV better than that of the group receiving RV_{RND}. We repeat this experiment for six RF instances. As shown in fig. 3.17(b), the proposed RV_{SRV} enables 30-70 mV less DRV degradation than RV_{RND}. It also performs competitively with RV_{VRV}.

We also estimated time needed for generating RVs, most of which is spent in sensing and digitizing. If we assume to use a 10-MSPS ADC and the average value of 1,000 samples, it would take 0.2 ms for sensing a bitcell. For a 32x32 array, the time to generate RVs is 0.2 s. The sensing/digitizing time is proportional to the size of the array. We can make the time more scalable with advanced sampling techniques. For example, we could monitor only the top 10% worst-case bitcells found in the first thorough sensing, which are likely to remain as the most-aged bitcells.

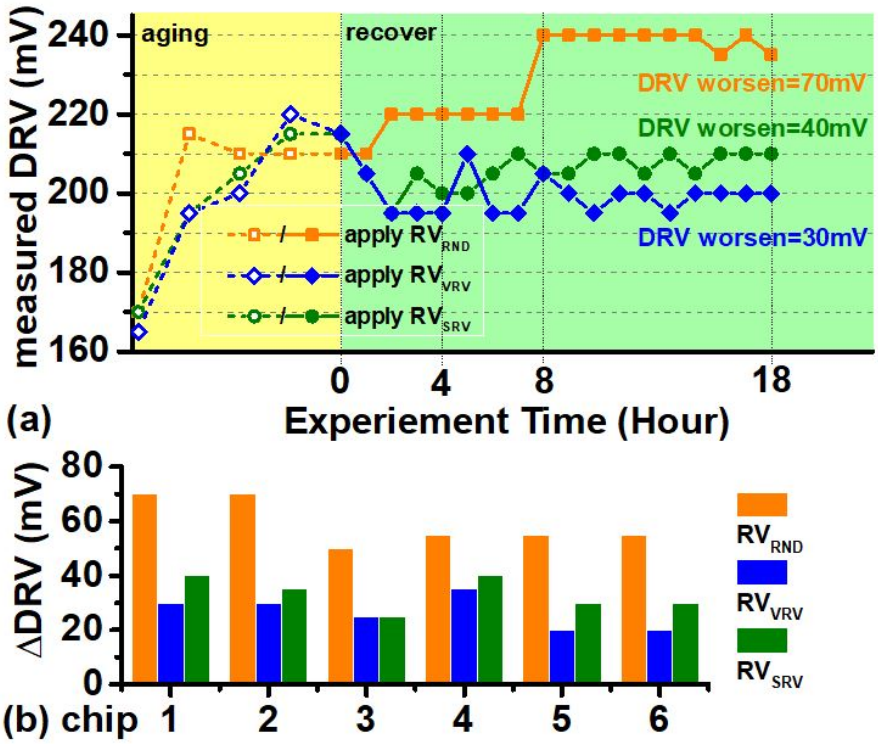


Figure 3.17: Measured DRVs during deceleration experiments of a typical array and multiple arrays

The results from those aging deceleration experiments confirm the feasibility of our proposed framework for aging management. It requires further system and architecture level research on long-term chip operation, aging dynamics, and available recovery time to be applied to a commercial processor.

3.7 Comparison and Conclusion

The closest work for in-situ monitoring of robustness of SRAM bitcells is Ref. [39] although it does not target to monitor NBTI degradation. The underlying idea of this work is to monitor the frequency of a ring oscillator that includes a bitcell of interest in it. However, the use of ring-oscillator frequency makes this technique not suitable for in-field operation since the frequency is a strong function of temperature and V_{DD}. As the work did not in-

vestigate the robustness against temperature and V_{DD} variations, we simulated the similar circuits in a 65nm. As summarized in table 3.3, the ring-oscillator based technique exhibits 3.4X more error over the same 60oC temperature variation and 4.6X larger error over the same 0.4V V_{DD} variation.

Table III. Comparisons to existing sensing techniques

	This work	[14] J. Tsai	[15] F. Ahmed	[33] Z. C. Lee
Technology	65nm	45nm	Simulation	28nm FDSOI
In-situ operation	Yes	Yes	No	No
Measured parameters	PMOS V_{TH}	ROSC freq.	bitcell current	Bit flip
VDD operation range	0.6-1.0V	0.7-1.1V	N/A	0.6-1.0V
Temperature operation range	20-80°C	N/A	N/A	N/A
Mean error by voltage variation	<21%	96%	N/A	N/A
Mean error by temp. variation	<19%	64%	N/A	N/A
Area overhead: 32x32b	27%	N/A	N/A	N/A
Area overhead: 64x64b	21% (est.)	N/A	N/A	N/A
Area overhead: 128x128b	15% (est.)	N/A	N/A	10%

Table 3.3: Comparisons to existing sensing techniques

table 3.3.

We introduce an *in-situ* technique to in-field monitor NBTI degradation in a 6T SRAM RF. The technique can transform each bitcell into a pair of V_{TH} sensors. The sensor outputs track the V_{TH} of PMOSs in a bitcell robustly against temperature and V_{DD} variations. Measurements confirm that the proposed technique exhibits 3.4X and 4.6X higher robustness in monitoring NBTI-induced V_{TH} degradation than the conventional technique. We can also use the outputs of the proposed sensors to generate RVs, recover PMOS NBTI aging and decelerate DRV degradation in an RF.

Chapter 4

Circuits, Architecture and Run-Time Framework for Memory Reliability Management in a Microprocesosr in the Field

4.1 Motivation

In deeply scaled VLSI systems, device aging effects, such as bias temperature instability (BTI) have been identified as a key reliability challenge. Especially, embedded caches ($) and register-files (RF) are highly vulnerable since they use intrinsically sensitive circuits, like 6-T SRAM structure, and become hot as nearby logic circuits are actively switching and dissipating heat. What's worse is that, unlike in logic circuits, the single worst-case bitcell can determine the reliability of the entire memory block. It is, therefore, paramount to manage the reliability of the embedded memory over the chip's lifetime.

To manage reliability it is cost-prohibitive to disassemble working μP and send the chip to a laboratory. Thus, a key requirement is to perform the reliability management without disassembly, i.e., post-deployment and in-field. In this work, we propose such a solution, including circuits, a microarchitecture and a runtime software framework to implement a

post-deployment in-field memory reliability management.

Towards this goal,

1. We devised a circuit technique that can *in-situ* sense the V_t of any one of six transistors in any bitcell in an array. The sensing readout is robust against temperature variations and thus can be used *in-field* where it is impractical to regulate temperature when measuring.

2. We applied this technique on a typical L1-cache (instruction cache, I$) design. In addition, we devised a RISC microarchitecture having a new instruction (called AS) to trigger the reliability management. This μP can execute the AS instruction opportunistically during its regular operation to sense the V_ts of its bitcells, thereby evaluate its aging effects.

3. We proposed a runtime software framework to convert the low-level V_t measurements to circuit-/architecture-level metrics, i.e., the data retention voltage (DRV). The framework is composed of basic operations (+, $\times$) and can be accomplished by original instructions of the μP ISA. Such extraction is critical for operating systems and firmware as they need high-level metrics to run dynamic reliability managements (DRM), such as to decelerate aging in bitcells [65], to reduce the guard band in dynamic voltage scaling of memory during the standby mode, and to balance aging degradations among memory banks [63].

We prototyped a μP test chip with the proposed techniques. Measurements show that the proposed sensing circuits can track aging-induced V_t modulation with 16.7% average error across 0 to 100°C. Also, the devised microarchitecture can sense the V_ts of the six transistors

of each of the 1-k bitcells in the I$ by executing 4-k instructions and taking less than 1s using a 10-bit 10-kSPS off-chip ADC. Assuming we perform such sensing every month, it takes only $7.8 \times 10^{-7}\%$ of the total operation time. Finally, the proposed framework can estimate I$'s DRV and its degradation within error of 10% and 12%.

4.2 Circuits and Micro-architecture Design

The circuit for sensing V_t of any of six transistors in a bitcell is based on the transformation technique [64] and summarized in fig. 4.2. By asserting and deserting proper WLs and BLs, we can transform a bitcell into a sensor that can produce voltages proportional to V_ts of pull-up (PU: PL, PR), pull-down (PD: NL, NR), or access (AX: AL, AR) transistors.

fig. 4.1 shows the configuration of the circuits during V_t sensing of **PL** in a bitcell $C_{5,7}$: V_{DDA} is connected to GND through APS by setting P[1:0]=11; WL[5] is asserted through WLDEC since RI[4:0]=5; with CI[4:0]=7, SN=0, SP=1, SL=1, and SR=0, BL7 and BLB7 are connected to SCS through BLMUX; one of five transistors (T0) in SFE is selected using 1-hot SC[4:0] to achieve better temperature robustness.

Finally, the target transistor PL in the selected bitcell is connected to SFE (sensor front-end) to form a circuit that produces a temperature-robust voltage (V_{SEN}) that is proportional to its V_t. Configurations for sensing PU, PD, and AX are summarized in fig. 4.2. V_{SEN} is then digitized by the ADC as SD.

To utilize the above sensing circuits and monitor its own memory reliability, we modified the microarchitecture of the 16-bit MIPS-ISA processor by adding the sensing instruction AS (fig. 4.3). The instruction contains OP, RS, RT and IMM domains. It is executed in a way similar to the memory-store instruction (ST) which stores the value of a target register

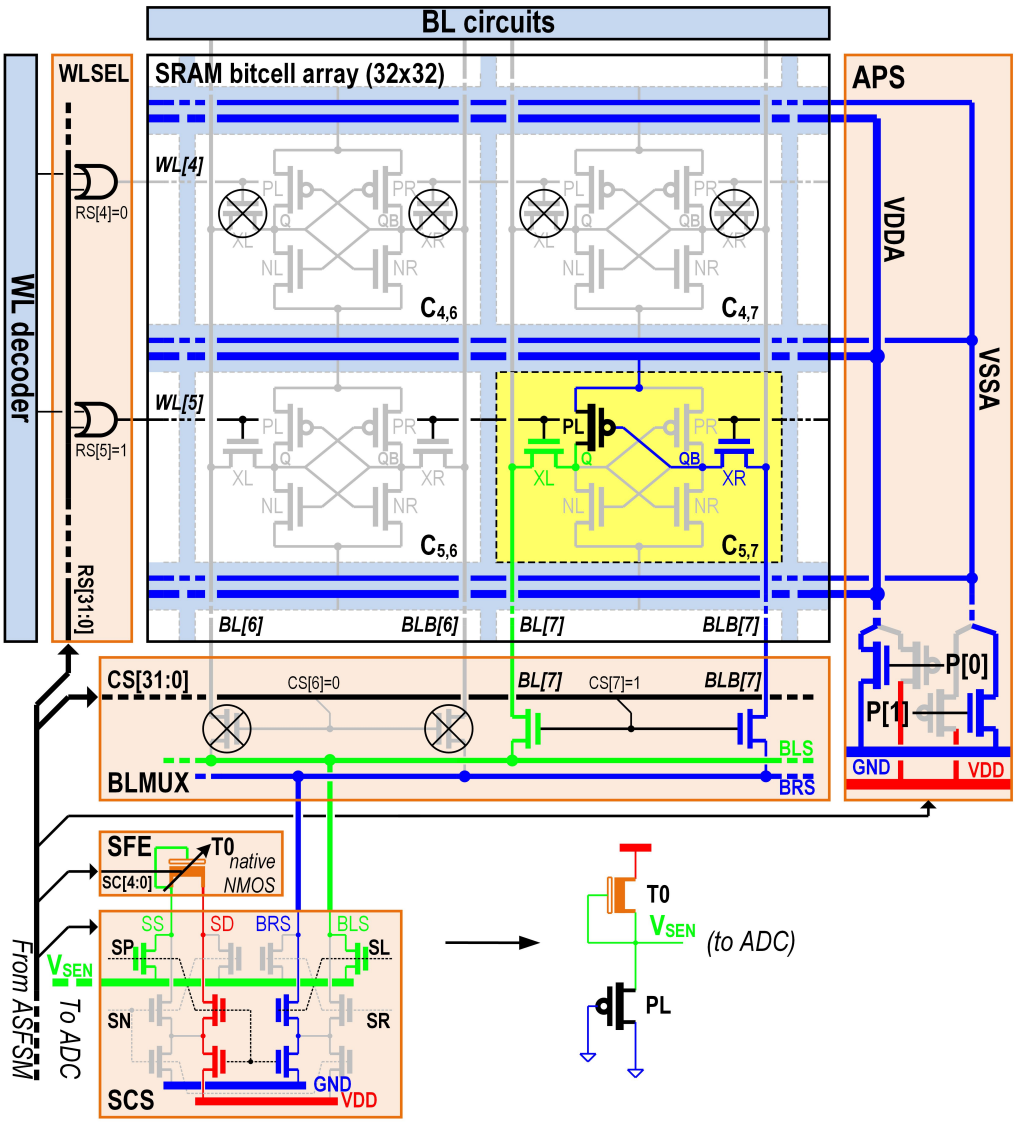


Figure 4.1: Peripherals that converting a bitcell PMOS *in-situ* into a V_t sensor

($RT) to the data memory (DMEM) at address $RS+IMM. The difference is that AS stores the sensing result (SD) to DMEM. The stored data can be used by other instructions for the DRV estimation framework or other post-sensing data processing. AS refers $RT for sensing configuration such as the row and column addresses of the bitcell to be sensed (RI[4:0], CI[4:0]) and the type of transistors (PU, PD, AX) to be sensed.

To support the **AS** execution, we modified pipeline stages IF and ID and added several

Block Name	Signal Name	value at different modes				
		Norm.	Sensing	PU / PD / AX		
APS	P[1:0]	2'b10		2'b11	2'b00	2'b10
	VDDA	VDD		GND	VDD	VDD
	VSSA	GND		GND	VDD	GND
WLSEL	WL[31:0]	1<<PC[5:1]		RS[31:0] (1<<RI[4:0])		
BLMUX	CS[31:0]	32'b0		CS[31:0] (1<<CI[4:0])		
SCS	SL/SR	0/0		$\overline{\text{L0R1}}$/L0R1*		
	SN/SP	0/0		0/1	1/0	1/0

Figure 4.2: Configurations to enable converting 6 transistors into V_t sensors

blocks (fig. 4.4 and fig. 4.1). In the IF stage, we added wordline selector (WLSEL) between original wordline decoder (WLDEC) and bitcell array (I$.DATA) to assert a target WL during sensing. We also added peripherals to I$: an array power switch (APS) to control power grids of the array (V_{DDA}, V_{SSA}); a bitline multiplexer (BLMUX) to select the bitline pair of the targeted bitcell. We use thick-oxide transistors for a larger on to off current ratio in the BLMUX to minimize leakage. The APS is carefully designed and verified to avoid IR drop across it.

In the ID stage, we modified it such that it can identify AS and generate a trigger signal (ASE) to activate the Hazard Detection Unit (HDU) and a FSM for sensing operation (ASFSM). We also added a MUX to pass either $RT (ASE=0) or SD (ASE=1) to the next stage EX.

We also added sensor configure switch (SCS) to change connections between the selected bitline pair and the sensor frontend (SFE) for sensing PU, PD, or AX of a bitcell.

The operation of the **AS** instruction is shown in fig. 4.5. The ASFSM (**AS** finite-state-machine, fig. 4.6) starts with the initial state S0. Once ID detects AS, it asserts ASE. HDU

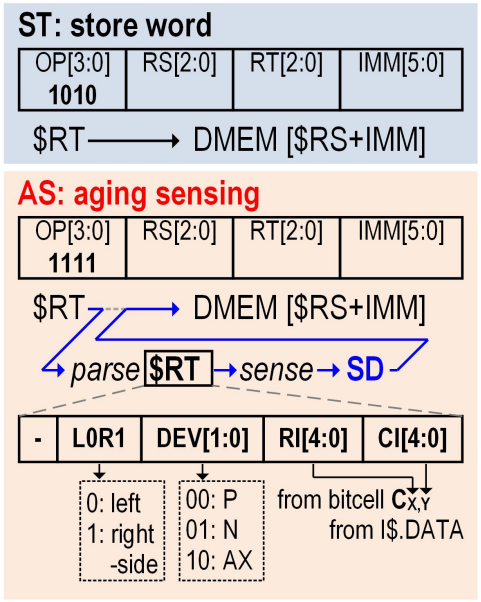


Figure 4.3: Formats comparison between existing instruction **ST** and added instruction **AS**

catches ASE and treats this as a structural hazard and thus immediately disables/stalls the pipeline. ASFSM catches ASE at the next clock edge and enters the state S1 from S0. At S1, it latches $RT, asserts ICCH to hold ICC and enters S2. At S2, it parses $RT, generates the control signals WLSEL, BLMUX, APS, SCS, SFE, and waits for tens of clock cycles during which a bitcell of the I$ is transformed to a sensor and produce a stable V_{SEN}. The V_{SEN} settling time is less than 20µs with an approximate 1pF capacitance load at room temperature. Due to the transformation, the cache loses the contents and therefore asserts CM (cache miss) to notify ICC to initiate an instruction loading. However, the assertion of ICCH delays this loading until the sensing is completed. At S3, ASFSM notifies the off-chip ADC to start conversion. Upon completion of the conversion, the ADC puts data (SD) to the bus and asserts EOC. ASFSM detects EOC, enters S4, latches SD and configures I$ back to the normal mode within a few cycles. Then ASFSM enters S5, which it releases ICC by de-asserting ICCH. This makes ICC to load instructions on the I$, and after that

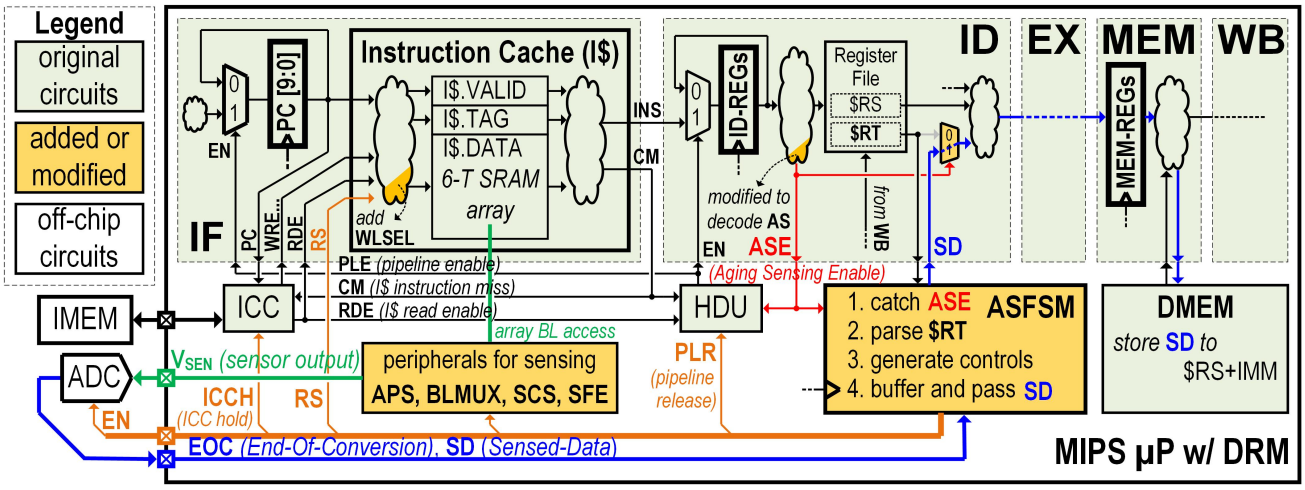


Figure 4.4: Microprocessor architecture of proposed DRM technique

assert RDE. ASFSM catches RDE, enters SE and releases the pipeline by asserting PLR in one cycle. At the next clock cycle, the pipeline stall is released (PLE:0→1), SD goes to the EX stage and ASFSM goes back to S0. Two cycles later, SD is stored to DMEM at address $RS+IMM.

4.3 Testchip and Measurements

The proposed μP was fabricated in 65nm CMOS. fig. 4.7 shows the die photo and area breakdown. The area overhead of the architecture change for µP with 1K-b L1 cache is 8.9% and expect to be scaled down with larger memories since only BLMUX increases in square root of memory size. We expect an on-chip ADC is available in most of the systems for other purposes and can be reused for our proposed application. We tested and verified that the proposed µP can successfully execute the newly-added and existing instructions. At 1V, it achieves 250MHz clock frequency.

We then performed an accelerated aging test (AAT) under 125°C and 1.8V. Every 15min, we change the random data (with 30% '1' and 70% '0') stored in the I$; every 2 hours we

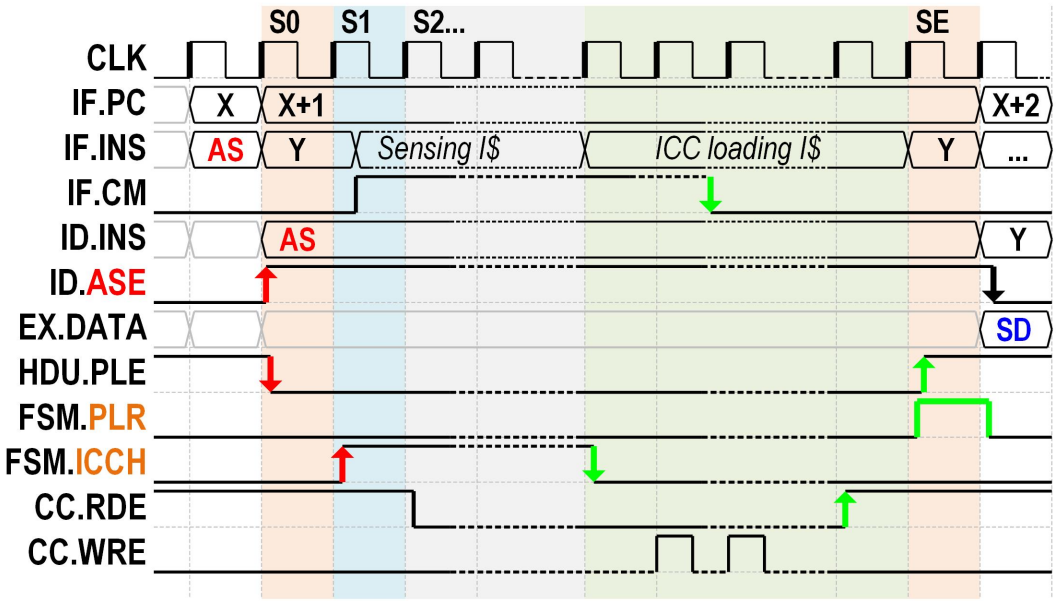


Figure 4.5: Timing diagram when executing instruction **AS**

measure both sensor outputs and DRVs. fig. 4.8(a) shows the aging of PU, PD, and AX of a typical bitcell. NBTI is measured stronger than PBTI; AX undergoes little aging as it is turned off for most of the time. fig. 4.8(b) shows the statistics of PMOS degradation. fig. 4.8(c) shows the temperature sensitivity of the sensor outputs: 130ppm/°C and 650ppm/°C for the average and the 3.5σ worst case, respectively.

4.4 DRV Estimation Framework

Using the prototyped μP chip, we devised a framework to estimate pre-aging DRV (called DRVO) and its degradation (called ΔDRV) across chip's lifetime. The framework consists of three modules: retention preference estimator (RPE), DRV estimator (DE), and DRV degradation estimator (DDE).

The RPE (fig. 4.9(a)) predicts if a bitcell is better at retaining '0' or '1' at low VDD. This asymmetry is developed as device mismatch, both from process variation and aging, shifts

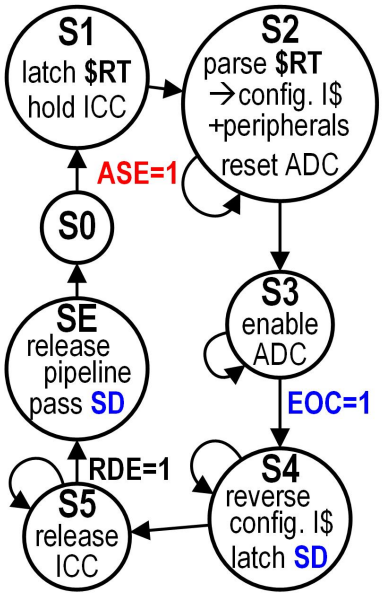


Figure 4.6: ASFSM state transfer graph

the voltage transfer curves (VTC) of the left and right inverters of a bitcell. The input to the RPE, a 3-by-1 vector, is the differences of sensor outputs of the left and right side of devices of a bitcell. We find this helps to remove the common-mode offset in sensing measurement. We then trained a 3-by-1 coefficient vector A using the pre-aging measurements from 3k bitcells of 3 chips. Finally, the multiplication of the input and coefficient vector A is threshold-ed to 1 or 0. fig. 4.10(a) shows the RPE output accuracy across 7 chips with the same A. For the critical bitcells that DRV>0.18V, it achieves zero error.

The DE (fig. 4.9(b)) estimates the DRV amplitude of a bitcell, which is the larger value between DRV0 (DRV for holding '0') and DRV1 (DRV for holding '1'). The input to the DE is a 6-by-1 vector, whose first three elements are the sensing results of the PU, PD, and AX of the stronger one of the two inverters in a bitcell. The second three elements are that of the weaker inverter. We can choose the stronger and weaker inverter based on the output of the RPE. Then we generate the coefficient vector B based on the pre-

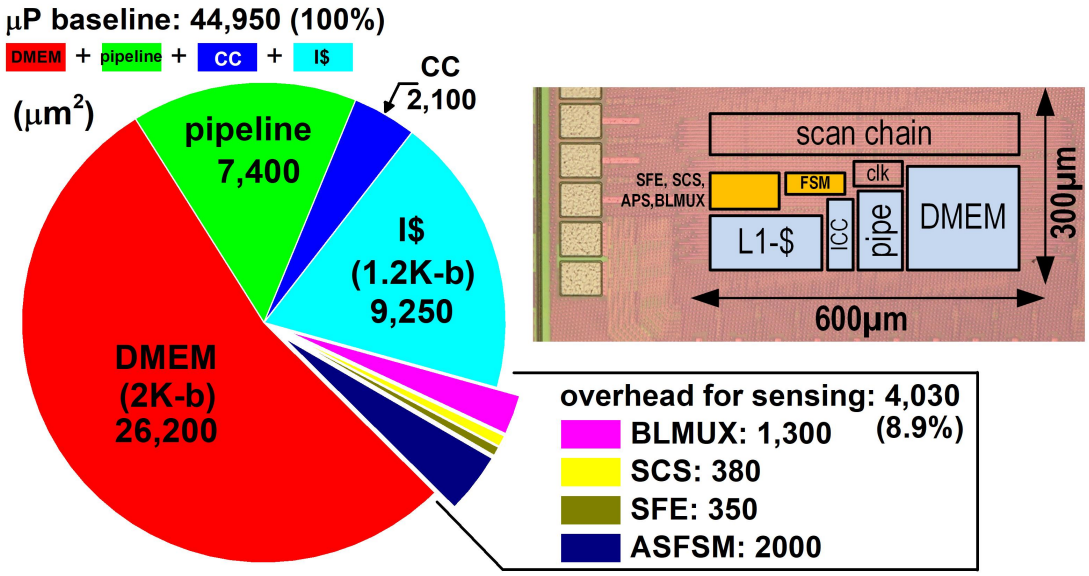


Figure 4.7: Area breakdown and chip die photo

aging measurement data from 3k bitcells from 3 chips. A bias term (b0) of the B is used to compensate systematic variations. fig. 4.10(b) shows the strong relationship between the measured DRV and estimated DRV. fig. 4.10(c) shows the accuracy of DRV estimation across 7 chips. For the DRV values>0.18V, the DE achieves =10% error.

Finally, the DDE (fig. 4.9(c)) estimates DRV degradation due to BTI aging. The input is a 2-by-1 vector, whose elements are the sensor output changes from the initial output (e.g., $\Delta V_{PL} = V_{PL} - V_{PL0}$) of the more aged transistor pair (PL, NR or PR, NL) of a bitcell. Here, we also use the RPE output to select the more aged transistor pair. We also find the coefficient vector C by regressing data from 2k bitcells from 2 chips before and after 12.5-hr AAT. The multiplication of the input and C vectors yields ΔDRV.

fig. 4.11 shows the foundation of DDE: strong relationships between ΔV_{PL} and ΔDRV1 (left); ΔV_{PR} and ΔDRV0 (right). fig. 4.12 shows the estimated and measured ΔDRV as a function of ΔV_P and ΔV_N. fig. 4.13(b) shows the histogram of estimation error. fig. 4.13(c)

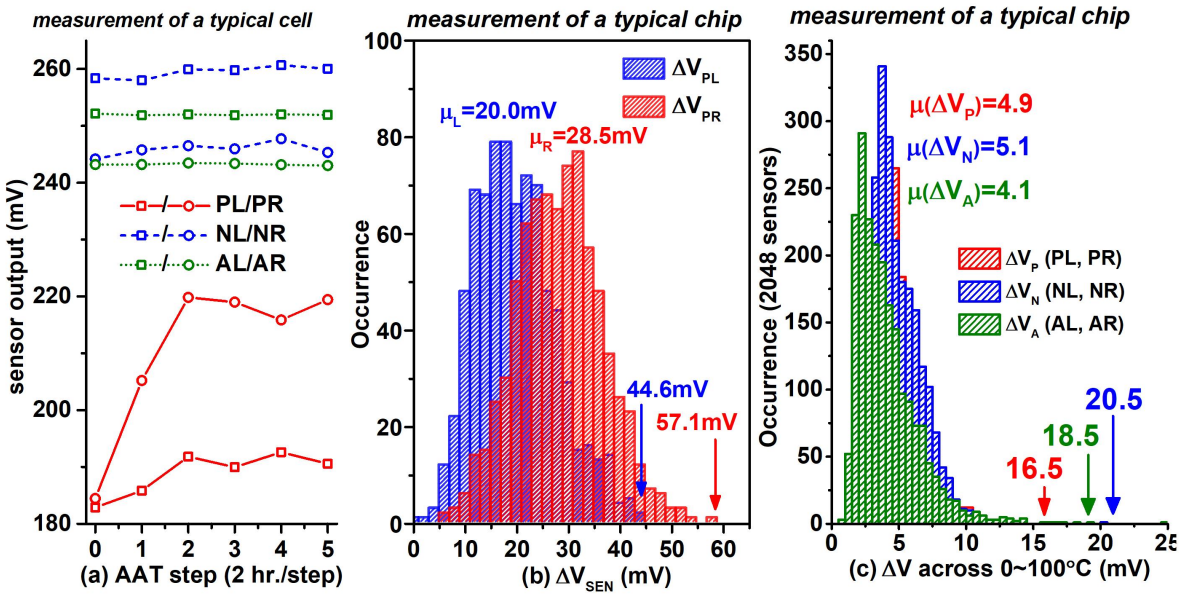


Figure 4.8: Sensor measurement (a) a cell case (b) a chip case (c) sensitivity to temperature

shows that the RMS and the worst-case error in the 12.5-hr AAT.

fig. 4.14(a) shows our framework combining the above three modules. During the manufacturing test (pre-aging), the initial sensor outputs (e.g., V_{PL0}) are measured and stored at non-volatile memory (CSV file with size of 306kB used in our experiment; the size can be smaller if data compression is used). Since aging develops slowly, the processor can perform sensing and DRV estimation infrequently and only when a processor is idle. After sensing the cache, the processor runs the framework to estimate the DRV: first it loads the initial sensor outputs to the RPE and DE modules to estimate the original DRV (DRVO); then it loads the current measurements to the RPE and DDE modules to estimate the DRV degradation (ΔDRV). The current DRV value is then DRVO+ΔDRV.

fig. 4.15(a) shows DRV estimation error: -12 to 7% for the critical bitcells (DRV>0.2V). fig. 4.15(b) shows the measured and estimated array DRVs (worst case across 1024 bitcells) of two chips as AAT goes, showing our framework estimates DRV well over long-term reliability

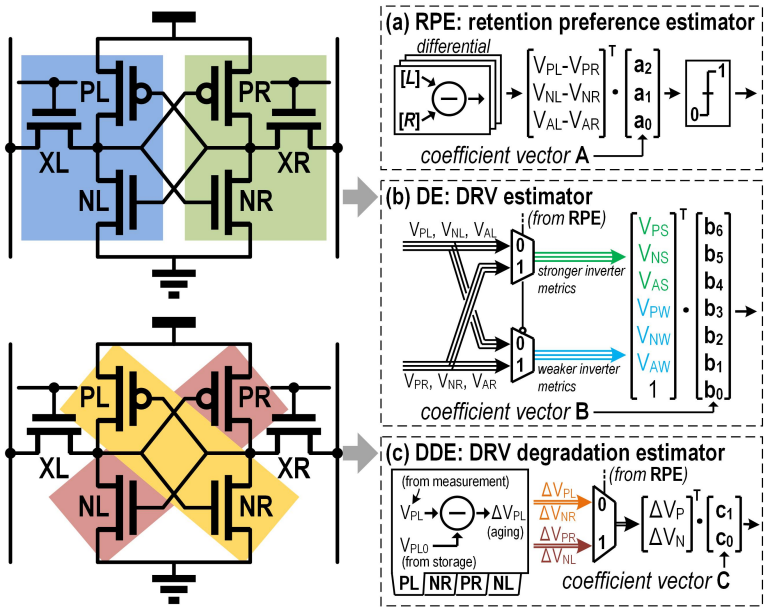


Figure 4.9: Framework modules to estimate (a) retention preference (b) DRV (c) DRV degradation

degradation.

		This chip	[1]	[2]	[3]	[4]
Technology		65nm			45nm	65nm
Measured physics		in-situ, V_t (P+N+A)		in-situ, V_t (P)	in-situ, f_{OSC}	in-situ, I_{BL}
Oper. Temp. Range (°C)		0~100	20~80		*Not reported*	
Error by Temp. [avg., w.c.]		17%, 55%	24%, 56%[1]	25%, 53%[1]		
On-the-fly DRV Estimate		Yes, w/ 4 V_t	Yes, w/ 6 V_t	Only DRV polarity	*No DRV estimation*	
DRV est. error (critical cell)	Preference	0%	3%	-		
	Amplitude	12%	13%	-		
Demonstration		µP & Mem	Mem only		Mem only	

[1]normalized to 0~100°C, based on an assumed 30mV BTI degradation

Table 4.1: Comparison table

Finally, we compare our results with the previous works (table 4.1) that in-situ monitor memory reliability [39, 64–66]. Existing work mainly focuses on memory design [39, 64–66], whereas our work demonstrates the holistic integration of memory, μP, and software framework. Compared to [65], the accuracy of our sensing circuits and DRV estimation

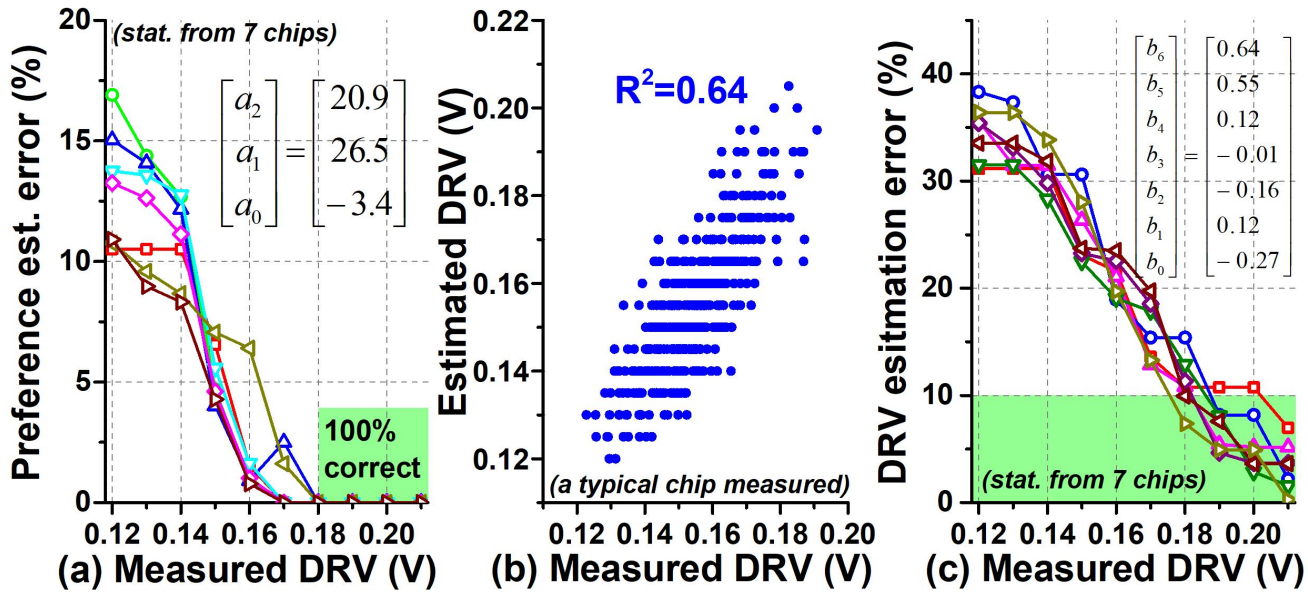


Figure 4.10: (a) RPE: coefficient A and accuracy (b) Estimated and measured original DRV (c) DE: coefficient B and accuracy

framework is better while reducing sensing time by 33% since it requires only to sense four Vts per bitcell, The techniques presented in [39, 66] are not designed for in-field operation and thus the sensing results vary across temperature.

4.5 Conclusion

In this work, we propose circuits, a microarchitecture, and a framework to dynamically sense the reliability of on-chip memory of a μP in the field. The sensing circuits can robustly monitor V_ts of transistors in bitcells *in-situ*, achieving a small error over a wide temperature range. The μP is expanded with a new instruction that exercises the circuit transformation for sensing with minimal hardware overhead. Finally, we developed and verified a software framework that estimates the DRV of the L1 cache over a chip's lifetime. The proposed technique can be used to reduce the guard-band in setting the standby-mode V_{DD} for memory. Also, it can be used for recovery vector generation [65] and aging management among

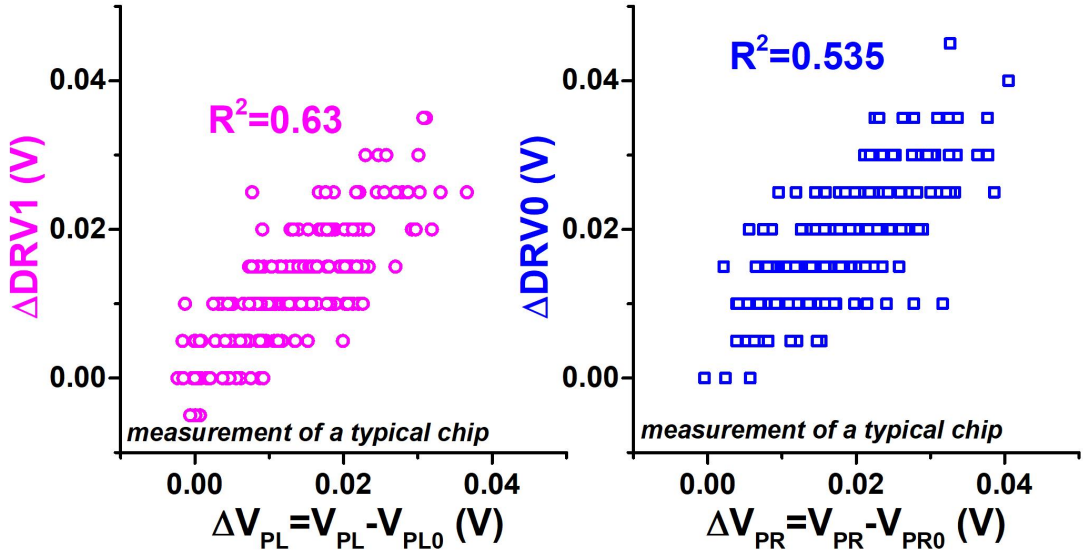


Figure 4.11: ΔDRV and PMOS V_t sensor output correlation

multiple memory blocks [63].

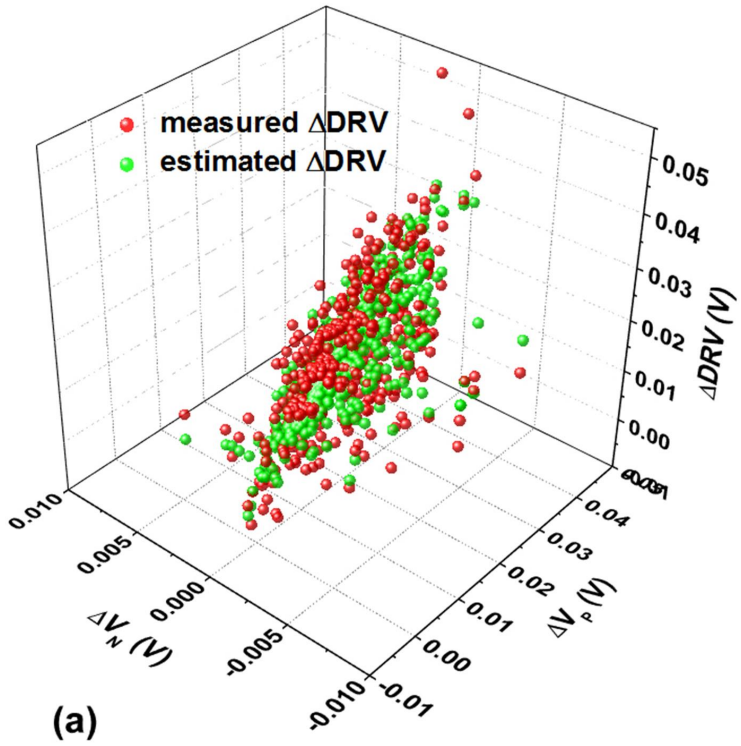


Figure 4.12: Estimated and measured ΔDRV correlation

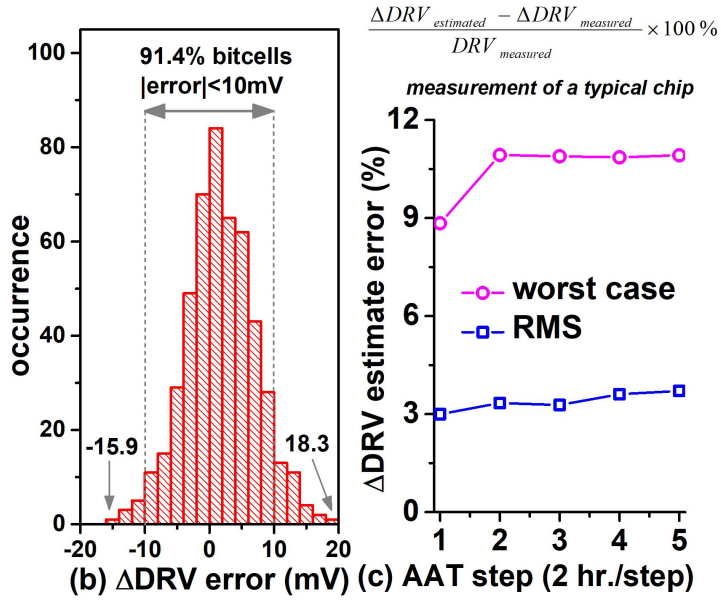


Figure 4.13: (b) error statistics (c) errors across AAT

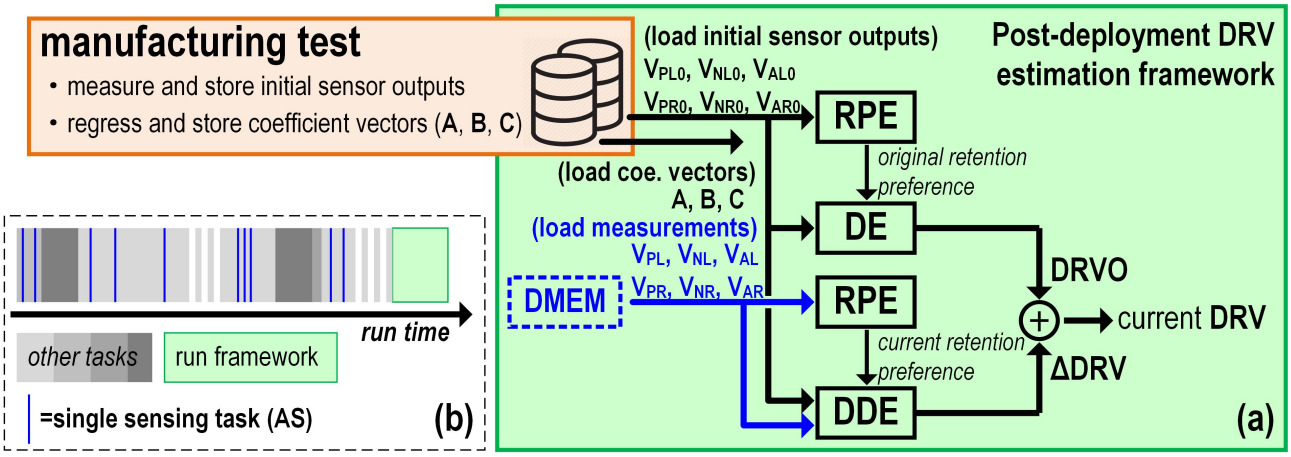


Figure 4.14: (a) Sensing task assignment (b) Framework to estimate current DRV

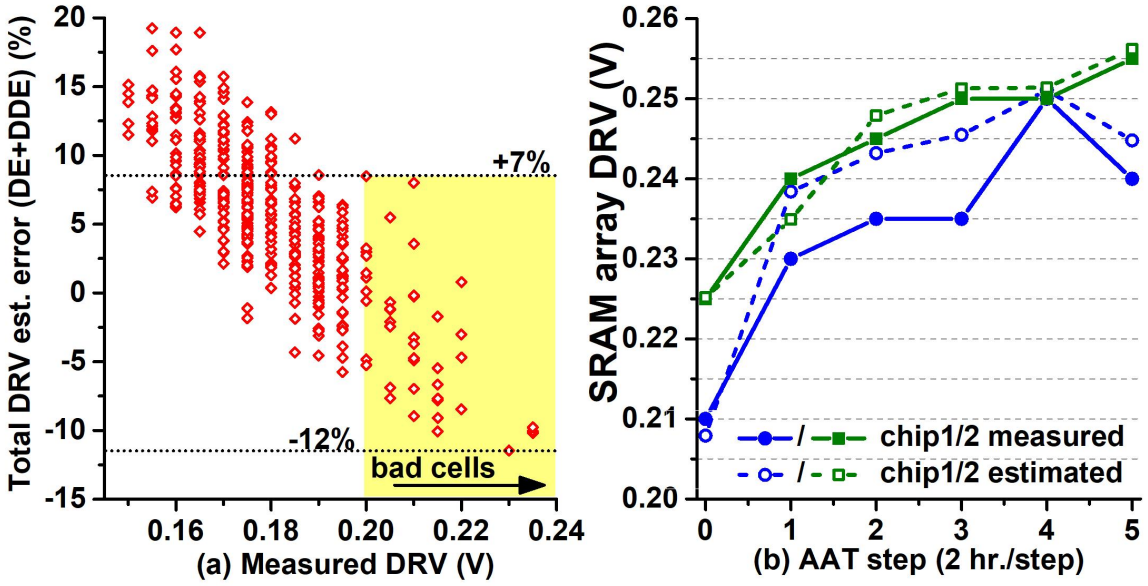


Figure 4.15: DRV estimation error for (a) bitcells, (b) L1 Cache

Chapter 5

An Area-Efficient SoC with an Instruction-Cache Transformable to an Ambient Temperature Sensor and a Physically Unclonable Function

5.1 Motivation

Heading towards the era of Internet of Things (IoT), it is critical for integrated-circuit research and development to deliver compact, low-cost, and dependable edge devices with various capabilities, e.g., sensing, computing, communication, and security [67]. This challenge has motivated to integrate an increasing number of components and function blocks into a Microprocessor-based System-on-Chip (μP-SoC) to shrink system footprint and associated cost [82, 88, 89]. However, such integration often incurs silicon area increase since most of analog, mixed-signal, and digital circuits require substantial amounts of silicon area to implement fast, accurate, and robust functions.

An ambient temperature sensor (T-sensor) and a Physically Unclonable Function (PUF) are two widely used components in IoT devices. The former is a critical building block for environmental monitoring; the latter is a notable security macro used for secret key gener-

ation for cryptography and chip-ID generation for authentication. However, implementing dedicated circuits for those functions requires non-negligible silicon area, especially when they are designed for high accuracy and robustness [68–77, 79, 84, 86–89].

It is noteworthy that in many applications, T-sensors and PUFs exhibit low duty cycle, making the approach of dedicated hardware further inefficient in area. As shown in fig. 5.1(a), for example, a T-sensor can be only active every several seconds (or even longer) since ambient temperature changes rather slowly [82,88]. A PUF also needs to be active only upon a request for e.g., encrypting and decrypting messages, and chip authentication processes [85, 86]. Therefore, as shown in fig. 5.1(b), dedicated hardware can be idle for most of the time.

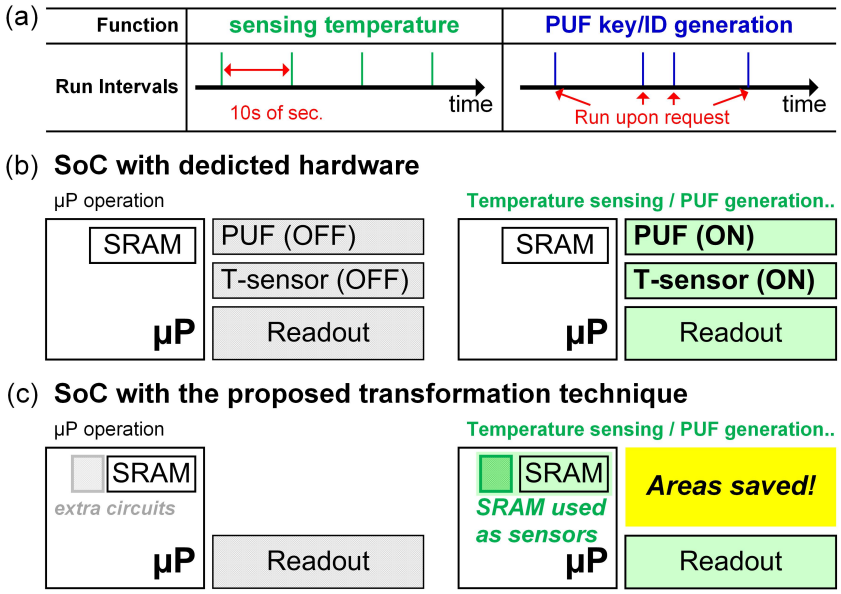


Figure 5.1: (a) Low duty cycle operation such as ambient temperature sensing and PUF. (b) Dedicate hardware implementation of those functions compared with conventional design. (c) Proposed transformation approach can save area.

Therefore, we aim to address such area inefficiency, and propose a novel technique to transform the existing SRAM in the instruction cache (I$) of a μP into a T-sensor or a PUF

(fig. 5.1(c)). This hardware recycling approach can reduce silicon footprint while integrating more features on a chip. To enable such transformation, we made a minimal amount of change in the SRAM circuits, Instruction Set Architecture (ISA), and pipeline control logic. The outputs of the transformed T-sensor and PUF operations are stored in the data memory of the μP for post digital processing.

We prototyped a μP-based SoC with the proposed technique in a 65nm general-purpose CMOS. The μP can operate at 320MHz at 1V supply voltage (V_{DD}) and consumes 10.6 pJ/cycle. The transformed T-sensor achieves an error of -0.5/+1.5°C after One-temperature-Point Calibration (OPC) across 26 instances. It achieves low V_{DD} sensitivity, exhibiting only 0.46°C error for 100mV V_{DD} variation from 1V to 0.5V. The transformed PUF also achieves a desirable randomness: the analog differential output shows a normal distribution with μ=-1.3mV and σ=31.2mV; the digitized bitstream passes all the applicable NIST tests and achieves 0.502 inter-PUF Fractional Hamming Distance (FHD). It also achieves robustness comparable to the state of art: 0.027% unstable bit ratio and 1.97×10^{-5} Bit Error Ratio (BER) after Temporal Majority Voting (TMV11) and Comparator Input Swapping (CIS) based masking.

The proposed transformation capability increases the area of the baseline μP by 12.9% (9.2% only for the T-sensor and 9.1% only for the PUF). The first 6.3% is for the update in the SRAM circuits and the next 6.6% is for the microarchitecture modification. The standalone T-sensor [73] and PUF [76] circuits achieving the similar accuracy and robustness would consume more silicon area, that would be 62.9% of the baseline μP area.

5.2 Circuit Design and Transformation

5.2.1 T-sensor Transformation

The key idea in the proposed transformation is convert SRAM bitcells, peripherals and addition devices into target analog circuits by applying certain logic values on bitlines (BL), wordlines (WL), etc. In the T-sensor transformation, the target analog circuit topology is a compact Complementary-To-Absolute-Temperature (CTAT) voltage generator (Refs. [68, 83]).

To support such transformation, as shown in Figs. 2(a, b), we updated the peripherals of an SRAM block (in the I$) in mainly four ways:

1. We inserted a wordline selector (WLSEL) between the WL decoder and the WLs such that it can assert multiple WLs based on the control signal RS[31:0].

2. We added a bitline multiplexer (BLMUX) in parallel with existing BL circuits. The BLMUX can connect multiple BL and BLB pairs into a pair of BLS and BRS (.e.g. BL[30:1] to BLS, BLB[30:1] to BRS) based on the control signal CS[31:0].

3. We added Power-Gating Switches (PGS), which can change the V_{DD} nodes of bitcells (V_{VDD}) to the ground (GND) level with control signal PS[1:0]. Note that the arrays shares a common V_{DD} and V_{SS}, although multiple PGS are drawn to match the physical layout.

4. We added Sensor Configuration Switches (SCS) and a T-sensor header (CT), where we used thick-oxide IO devices for part of the circuit to suppress subthreshold leakage and thus achieve better circuit isolation.

fig. 5.2(c) summarizes the settings for those peripherals, stored in and accessed as a register $RT.

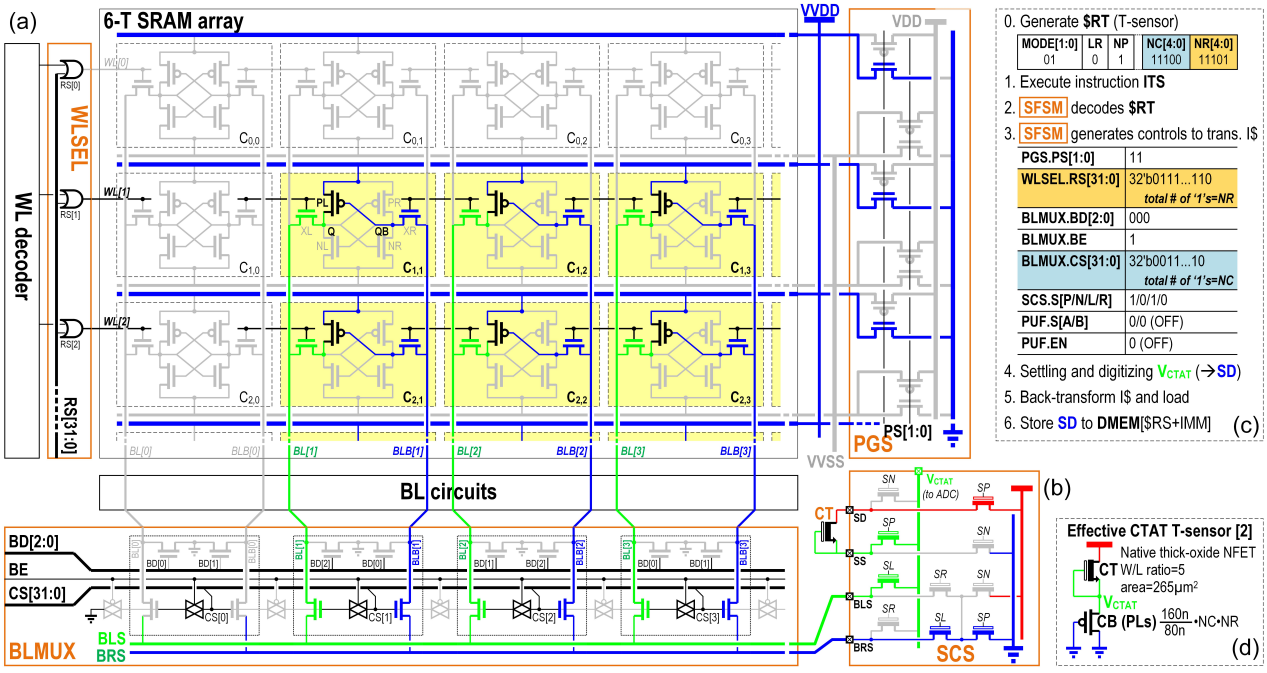


Figure 5.2: (a) The SRAM with the added peripherals showing the configuration for T-sensor transformation. (b) The schematics of SCS and CT. (c) $RT and control signal values. (d) The effective circuits of the transformed T-sensor.

The CTAT generator formed via the transformation mainly consists of two transistors: (i) parallel-connected Pull-Up (PU) PMOSs of selected 6-T SRAM bitcells and (ii) a native device added for the transformation. The equivalent circuits are shown in fig. 5.2(d). The bottom transistor CB is formed by aggregating the left pull-up PMOSs (PLs) of the selected bitcells. The top device CT is a native device added for the transformation. The gate voltage of the CB (i.e., the BRS node from the BLMUX) is connected to GND. We can sense the output V_{CTAT} to measure ambient temperature. Note that we can use the right PU PMOSs of bitcells for the same transformation with the complimentary settings for the SCS. Both CT and CB operate in the subthreshold region, thus via the similar steps in Ref. [68], the

expression of V_{CTAT} can be derived as:

$$V_{\mathrm{CTAT}} = \underbrace{\left[\frac{\mathrm{k}}{\mathrm{q}}\mathrm{n_{CB}}\ln\left(\frac{\beta_{\mathrm{CT}}}{\boldsymbol{\beta}_{\mathrm{CB}}}\cdot\frac{\mathrm{n_{CT}}-1}{\mathrm{n_{CB}}-1}\right)+\left(\mathrm{K_{T1,CB}}-\frac{\mathrm{n_{CB}}}{\mathrm{n_{CT}}}\mathrm{K_{T1,CT}}\right)\right]}_{\text{temperature coefficient}}\mathrm{T}+\underbrace{\mathrm{V_{th0,CB}}-\frac{\mathrm{n_{CB}}}{\mathrm{n_{CT}}}\mathrm{V_{th0,CT}}}_{\text{offset}} \tag{1}$$

, where k is the Boltzmann constant; q is the electron charge; $\beta_x = \mu_x C'_{ox}\frac{W_x}{L_x}$ is the transistor strength; n_x is the subthreshold slope; μ_x is carrier mobility; C'_{ox} is unit area oxide capacitance; W_x and L_x are channel width and length; K_{T1} is the temperature dependency of threshold voltage; V_{t0} is the threshold voltage at nominal temperature.

The temperature coefficient of V_{CTAT} (the first term of Equation (1)) is sensitive to process variations, and exhibiting non-linearity, yet can be calibrated by modulating β_{CB}, which is derived as:

$$\beta_{\mathrm{CB}} = \mathrm{NR}\cdot\mathrm{NC}\cdot\beta_{\mathrm{P}} = \mathrm{NR}\cdot\mathrm{NC}\cdot\mu_{\mathrm{P}}C'_{\mathrm{ox,P}}\frac{\mathrm{W_P}}{\mathrm{L_P}} \tag{2}$$

As shown in Equation (2), the control signals NC and NR, which respectively selects the number of columns and rows of the bitcells to be combined in the transformation, can modulate β_{CB}. fig. 5.3 shows the linearity-optimal NC and NR settings across process corners (simulation). The optimal NC and NR helps improve linearity of V_{CTAT} vs temperature curves, and can be found from a batch of chips in testing. We can then perform OPC to compensate the variation of the offset, which presents in the second term of Equation (1).

We also made several design choices to mitigate random process variation. First, we reduce the random process variation in CB by combining $NR\cdot NC$ of PU PMOS transistors

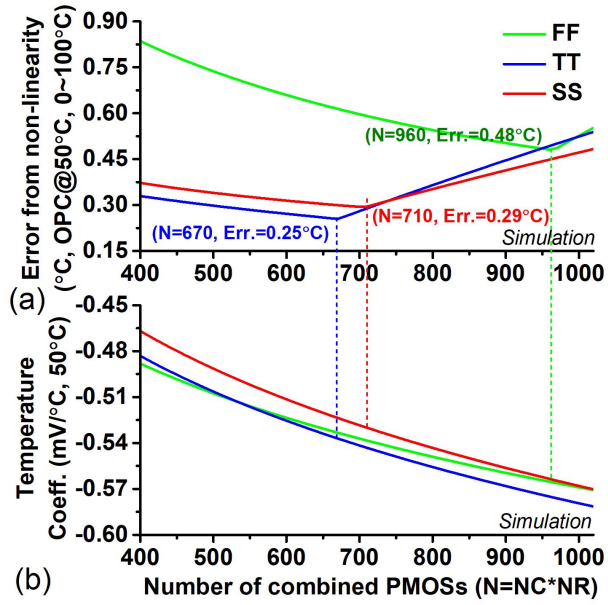


Figure 5.3: (a) Accuracy-optimal NC and NR combinations across process corners and (b) the corresponding temperature coefficient

to form a the total transistor area of $NR{\cdot}NC{\cdot}W_P{\cdot}L_P$, where W_P (=160nm) and L_P (=80nm) are the PU PMOS dimensions. Moreover, to minimize the edge effects, we designed the rows and columns to be selected from the center to the edge of the bitcell array.

We made several efforts to make the sensor output V_{CTAT} robust against V_{DD} variation. Similarly from the Ref. [68], we first set the V_{GS} of the CT to be 0V and increased the channel length of the CT so as to reduce the impact of Drain Induced Barrier Lowering (DIBL) and Channel Length Modulation (CLM). To digitize the outputs, an off-chip ADC is used for test flexibility. It communicates with the microcontroller via a simple hand-shaking protocol.

5.2.2 PUF Transformation

fig. 5.4(a) shows the circuits for the PUF transformation. The key idea is to form a pair of 2-Transistor threshold-voltage (V_t)-based temperature-compensated voltage generators and compare their outputs to produce one PUF bit using a voltage comparator. This is similar

to the bitcell proposed in Ref. [69].

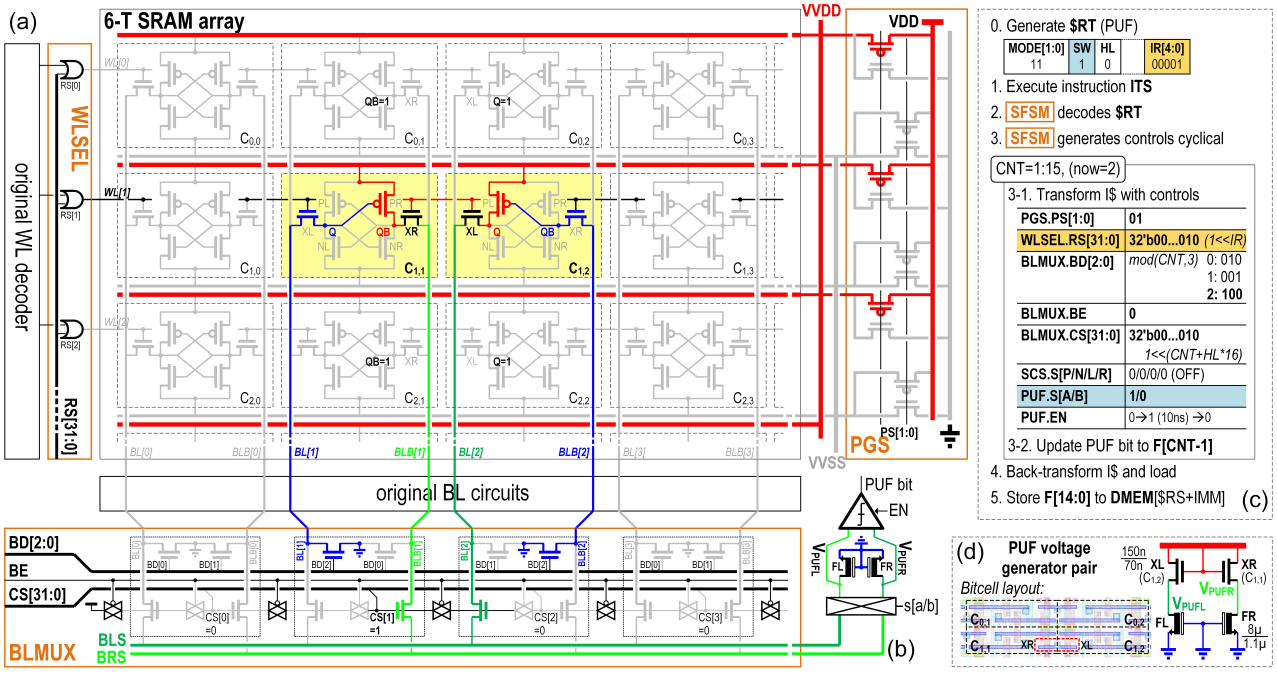


Figure 5.4: (a) Circuits configurations for PUF transformation. (b) The schematics of the PUF peripherals that contains PUF footers, a comparator and an input swapper. (c) $RT and control signals. (d) The effective circuits of the transformed PUF bitcell.

To form a pair of voltage generators, we connect two access transistors of two adjacent bitcells (XR of $C_{1,1}$ and XL of $C_{1,2}$) to a pair of footer devices (FR and FL) in the PUF peripherals (fig. 5.5) through WLSEL and BLMUX. fig. 5.4(c) summarizes the control signals of the WLSEL and the BLMUX and the settings for $RT for PUF transformation. These make the BLMUX to connect BLB[i] to BRS and BL[i+1] to BLS, where i stands for the selected column index. It also pulls BL[i] and BLB[i+1] down to the GND level by setting two nodes, QB in the column[i] and Q in the column[i+1], to the V_{DD} level. The two access transistors (XR of $C_{1,1}$ and XL of $C_{1,2}$) and two footers (FL and FR) now form a PUF bitcell (fig. 5.4(d)). All the devices in the effective PUF circuit operate again in the subthreshold region. Thus the output voltage (V_{PUFL} and V_{PUFR}) can be derived as:

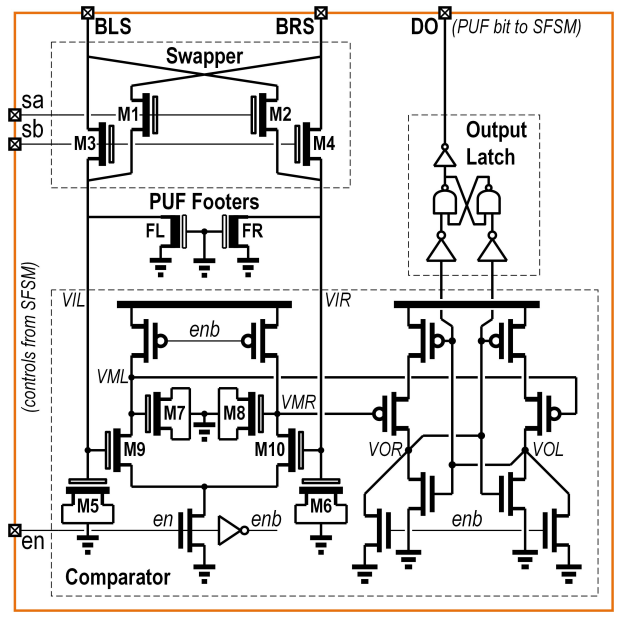


Figure 5.5: Schematics of the PUF peripherals

$$\mathrm{V_{PUFL}} = -V_{\mathrm{th,XL}} + \mathrm{V_{DD}} - \mathrm{V_{th,FL}} + \phi_{\mathrm{t}} \ln\left(\frac{\beta_{\mathrm{XL}}}{\beta_{\mathrm{FL}}} \cdot \frac{\mathrm{n_{XL}} - 1}{\mathrm{n_{FL}} - 1}\right) \tag{3}$$

The difference of the output voltages thus can be derived as:

$$\mathrm{V_{PUFD}} = \mathrm{V_{PUFL}} - \mathrm{V_{PUFR}} \approx \mathrm{V_{th,XR}} - \mathrm{V_{th,XL}} \tag{4}$$

, which shows that the V_{PUFD} is random since it is a strong function of random V_t mismatch of XR and XL. Note that FR and FL are significantly larger devices and thus exhibit much less random V_t variations. In addition, V_{PUFD} exhibits good robustness against temperature variations since we sized the PUF footer to minimize temperature dependency and the differential operation removes a good shared portion of the remaining temperature dependencies.

Then, by digitizing V_{PUFD} with a comparator (CMP), we can generate a PUF bit, namely

F[i]. After this, the BLMUX is configured to connect the PUF footers to the next pair of bitlines (i.e., BLB[i+1] and BL[i+2]) to generate the next PUF bit (F[i+1]). This process continues till a PUF word (15 bits) is generated from the bitcells in the first half of the selected row in the SRAM. The process continues for the second half of the row and the other rows and produces total 960 (32×30) PUF bits.

In the proposed transformation circuits, the leakage from the unselected rows could affect PUF output voltages. For example, the bitcells sharing the bitlines with the target bitcells can contribute leakage to the bitlines. Thus, to minimize such leakage, in the beginning of the PUF transformation, we first set all WLs high by asserting RS[31:0] and then select the target row. This ensures QBs and Qs of the unselected bitcells in the selected columns become high, creating a negative V_{GS} for the access transistors of those unselsected bitcells, significantly reducing the leakage current.

The Pull-Up (PU) or Pull-Down (PD) transistors could also be used to form the similar structure, but we choose to use the access transistors since they undergo less transistor aging effects, allowing better bit stability over chip's lifetime [78]. In fact, Negative Bias Temperature Instability (NBTI) can modulate the V_t of PUs and PDs by as high as several tens of mV [81]. This makes it difficult to use them in our PUF transformation.

We also chose one access transistor from each of two adjacent bitcells since they are placed in proximity (fig. 5.4(d)) and thus share the similar systematic variation. As comparing to Ref. [78] which digitizes two output voltages sequentially via an off-chip ADC, this design compares V_{PUFL} and V_{PUFR} directly with a comparator. This can improve the throughput, energy consumption, and reduce quantization error.

Last but not the least, we propose a method to identify unstable bitcells and generate a

mask to remove them in PUF evaluation. One of the critical problems in robustness is that a PUF bitcell whose XL and XR have small V_t mismatch can be sensitive to noise, temperature and V_{DD} variations. Those unstable bitcells could produce the same digital output even if we swap the inputs of the comparator, i.e. connecting BLB[i] to FR and BL[i+1] to FL, due to the offset of the comparator circuits, V_t mismatch of the PUF footers, or temporal noise.

Thus, we can perform such swapping multiple times to identify and create a mask for theses unstable bitcells. When implementing this, we leverage the configurability of the BLMUX, and therefore it incurs little additional overhead. Note that introduction of bit-masking involves storing the mask in non-volatile memory, and results in extra area overhead. However, such overhead scales with the miniaturization of non-volatile memories in the recent technologies [69].

We compare our proposed CIS based mask generation to the conventional Repetitive Readout (RR) method [77, 84]. In the RR-based method, a large number (N) of repetitive PUF readings, e.g. samples, are compared to identify if any bit has different reading between the samples. Similarly, in the CIS-based method, we use the total number of N samples, but with N/2 input-swapped and N/2 non-swapped. For each bit, if its readings with and without swapping are the same in any of the N/2 pairs, this bit is considered unstable. As shown in fig. 5.6(a), the CIS method can identify more unstable bits with a less number of samples. It also exhibits the smaller worst-case error. Here the error is defined against a reference mask generated based on roughly 10× more samples (500).

We test the performance of the masks. fig. 5.6(b) shows the unstable bit ratio post mask application. The results are from the worst-case mask among the total 400 masks that we generated across different sample sizes. It is shown that the CIS based method outperforms

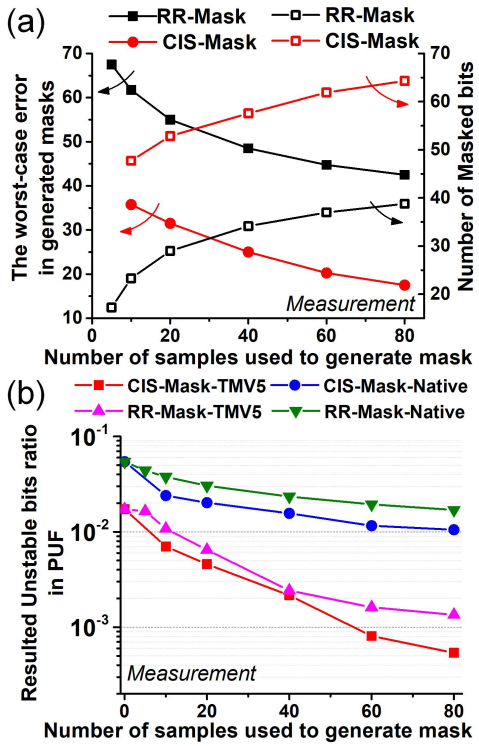


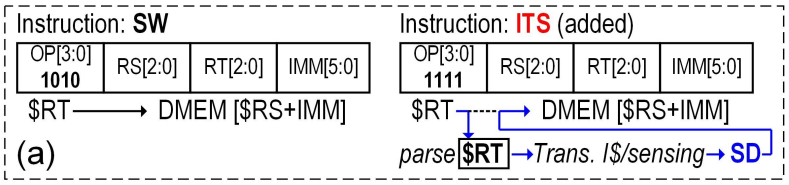
Figure 5.6: (a) The accuracies of masks generated by the proposed CIS and the conventional RR techniques. (b) The unstable bit ratios post mask applications.

the RR method, particularly if a good number of samples are used to generate a mask. Further improvement of the mask generation can be made through body-biasing the SRAM array to track the temperature dependencies of V_t [84]. Also, CIS and RR methods can be combined.

5.3 Micro-architecture Design

To perform the above-mentioned transformations and store the outputs of T-sensor and PUF to the Data MEMory (DMEM), we add a new instruction called ITS (fig. 5.7(a)). This instruction has the similar format as the Store Word (SW) instruction in the original MIPS ISA [90], which stores $RT to DMEM at the address $RS+IMM. Instead, the 16-bit register $RT in the ITS specifies the configuration of transformation, whose value needs to

be updated accordingly before the execution of ITS. The definition of the $RT is summarized in fig. 5.7(b). ITS also stores Sensor output Data (SD), instead of $RT, to DMEM.



[15:14]	**MODE:** 01=T-sensor	11=PUF
[13]	**LR:** use left (0) or right (1) transistor of the bitcell to make T-sensor	**SW:** swap PUF comparator's input
[12]	**NP*:** use NMOS (0) or PMOS (1) in the bitcell to make T-sensor	**HL:** output low (0) or high (1) 15 bits out of 30 possible PUF bits
[9:5]	**NC:** number of columns combined	N/A
[4:0]	**NR:** number of rows combined	**IR:** row index to generate PUF bits

(b) *use NMOS: transform I$ to proportional to absolute temperature (PTAT) T-sensor
use PMOS: transform I$ to complementary to absolute temperature (CTAT) T-sensor

Figure 5.7: (a) ITS and (b) $RT formats for transformations.

We modified the microarchitecture to support the added instruction, as is shown in fig. 5.8. The baseline μP has a standard 5-stage RISC pipeline and support a subset of the MIPS ISA [90]. The I$ in the Instruction Fetch (IF) stage is direct-mapped and can store up to 64 cache lines. The data memory of the I$ (I$.DATA) is made of conventional 6-T SRAM. We also updated the ID stage to support the newly-added ITS and added a 2-to-1 MUX for routing $RT. Finally, we updated the cache controller (CC) and the Hazard Detection Unit (HDU) such that the μP handles the transformation and related operation conformal to the existing microarchitecture control.

These modification increases the area of the μP roughly by 6.5% (3,000μm^2). The modifications also make a moderate impact on the critical path delay, increasing it by 12%, which comes from extra logic before the WL's and extra capacitance on the BL's, although we did not optimize the critical path delay post microarchitecture modification. On the other hand, as the SRAM bitcells are not affected in the modification, the read/write stability and

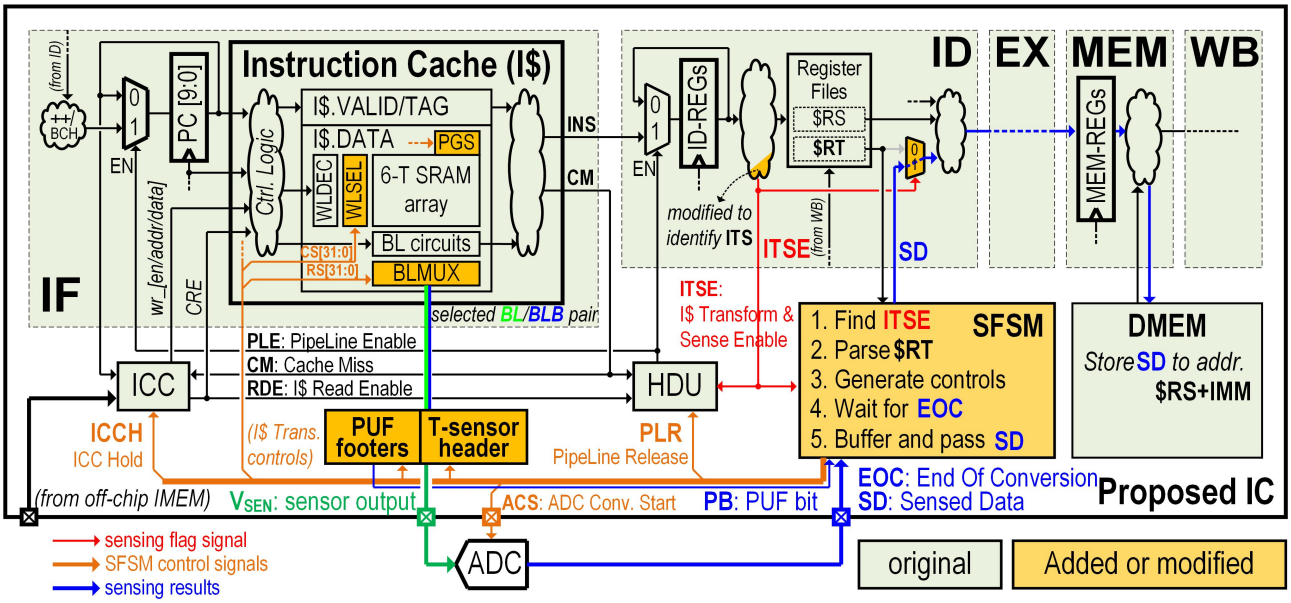


Figure 5.8: The proposed µp-SoC microarchitecture. The modified and added portions are highlighted in yellow.

noise margin are not impacted. Regarding the power overhead, as most part of the added circuitry are not active during normal operation, the power overhead is only 5μW.

The execution of ITS performs through four main steps; (i) it transforms the I$ to either a T-sensor or a PUF; (ii) the transformed T-sensor or PUF produces output, which is digitized and stored in DMEM; (iii) it transforms the T-sensor or PUF back to the I$; (iv) the I$ loads the next instructions to execute. During this process, it stalls the pipeline for several cycles, which is considered as a structural hazard in the microarchitecture. The T-sensor and PUF results, stored in DMEM, can be further processed via software in the μP.

fig. 5.9 describes the ITS execution in detail. First, in the S0 cycle, an ITS instruction enters the ID stage, asserting a signal called ITSE. The HDU considers this as one of the structural hazards, stalling the pipeline by de-asserting PLE. In the next cycle (S1), ITSE initiates the SFSM, which decodes $RT and transforms the I$ into either a T-sensor or a PUF based on the configurations specified in $RT. Since the circuit transformation invalidates all

the data stored in the I$ it asserts a Cache Miss flag (CM). However, the cache controller (CC) is notified by the cache-controller-hold (ICCH) signal from the SFSM in advance and thus ignores the asserted CM flag.

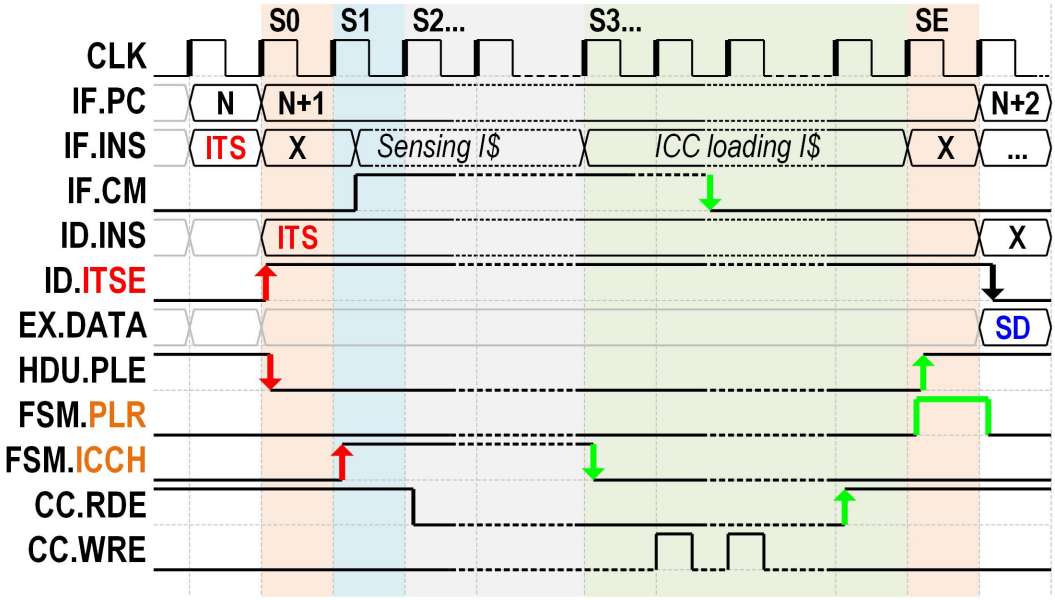


Figure 5.9: The sequences of the μP-SoC of executing one ITS instruction.

In the following cycle (S2), the sFSM waits for tens of micro-second until the analog output of the T-sensor or the PUF settles. The output is then digitized to SDATA by an ADC (in T-sensor) or by a CMP (in PUF). Finally, in the cycle S3, the SFSM transforms the T-sensor or the PUF back to the I$ in 1μs and it releases the ICCH. This makes the CC load the next set of instructions into the I$ from the main memory. Once this loading de-asserts the CM, the SFSM has the HDU to escape from the structural hazard state via a signal called Pipeline ReLease (PLR). Now the ITS instruction enters the EXecution (EX) and then MEMory (MEM) stages, where it stores SDATA in the DMEM.

5.4 Testchip and Measurement

Our μP-SoC with the proposed transformation capability was prototyped in a general purpose 65nm CMOS. fig. 5.10 shows the chip die photo. fig. 5.11 shows the detailed area breakdown. The additional hardware for the T-Sensor and PUF transformation takes 5,795μm^2, which is 12.9% of the 16-bit 5-stage RISC μP having 1.2kb I$ and 2kb DMEM. If only considering sensor frontends, i.e. the modifications made in the SRAM, the area overhead is 2,845μm^2, or 6.3% of the original μP area. When counting separately for T-sensor and PUF, the area overheads are 9.2% and 9.1%, respectively.

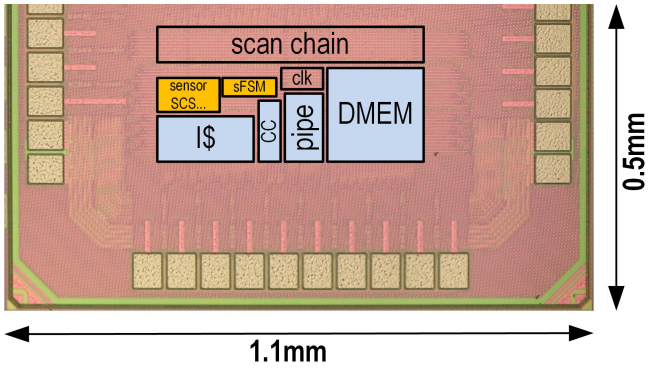


Figure 5.10: The die photo of the prototyped μP-SoC.

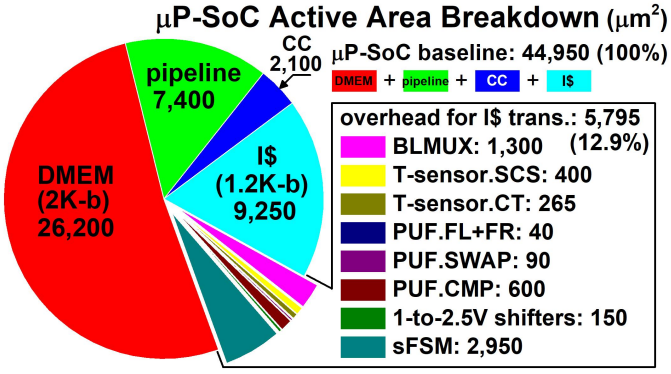


Figure 5.11: Detailed area breakdown.

fig. 5.12 shows the performance and power dissipation measurements of the μP-SoC. At

V_{DD}=1V, μP-SoC can operate at the clock frequency as high as 320MHz. It consumes 10.6 pJ/cycle performing a compute-intensive task (bubble sorting).

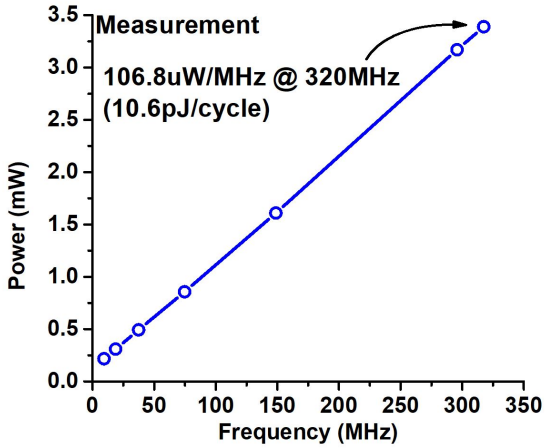


Figure 5.12: Clock frequency and power dissipation of the *mu*P-SoC.

We characterized the transformed T-sensor. Across 0-100°C, it achieves the temperature sensitivity of -0.62mV/°C. The post-OPC error is measured to be -0.53/+1.46°C across 26 instances at V_{DD}=0.6V (fig. 5.13(a)) after batch trimming for NC and NR. During batch trimming, we perform OPC while sweeping NC and NR on the first 10 instances and search for the optimal combination (fig. 5.13(b)). The optimal parameters are then applied to other instances. We also measured the sensitivity to V_{DD} variation across six T-sensor instances. Calibrated at 0.6V, the T-sensors achieve a worst-case error of 1.86°C across the V_{DD} variations of 0.5 to 1V and across the temperature range of 0-100°C (fig. 5.13(c)). As shown in fig. 5.14, we also tested Two-temperature-Point Calibration (TPC) at 10 and 90°C, which can reduce the error down to -0.52/+0.6°C across 26 instances. The power consumption of the T-sensor is simulated and shown in fig. 5.13(d).

For the transformed PUF, we measured the randomness, uniqueness, and robustness. The differential outputs (V_{PUFD}) show a normal distribution with μ=-1.3mV and σ=31.2mV. The

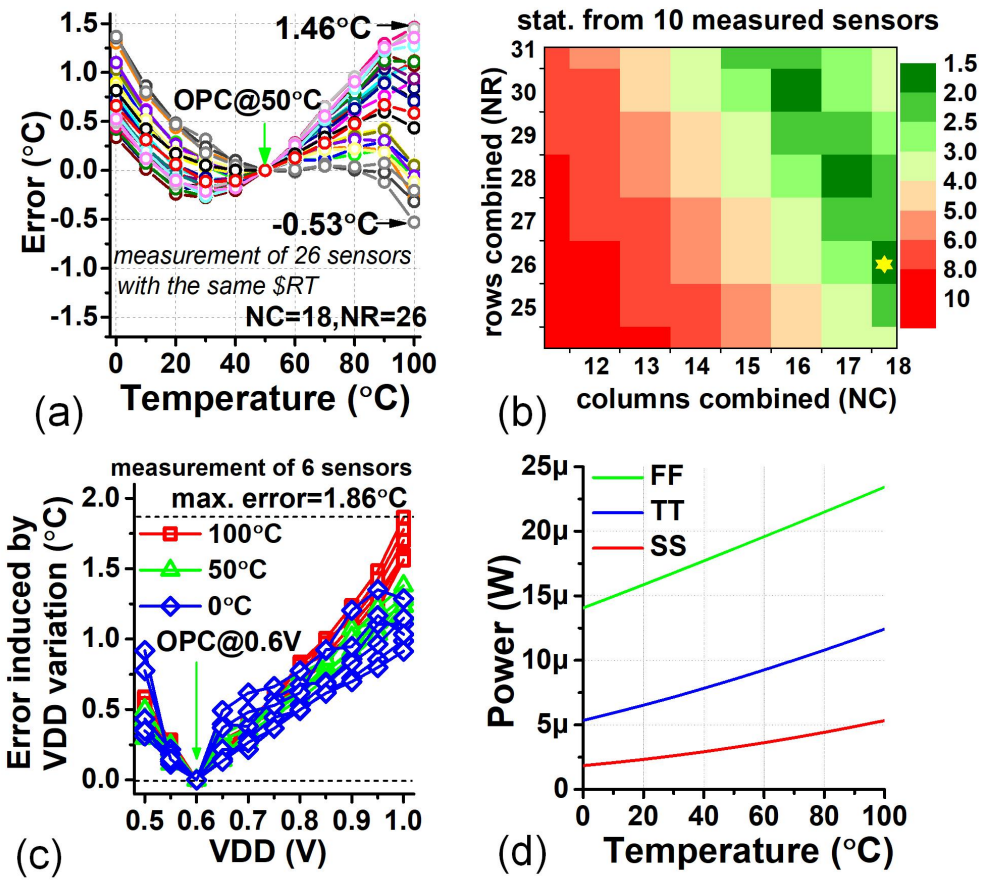


Figure 5.13: T-sensor measurement results: (a) Post-OPC accuracy. (b) The post-OPC worst-case error across NC and NR combinations. (c) Post-OPC accuracy across VDDs. (d) The power dissipation across corners and temperatures.

PUF bits passed all the applicable NIST random tests (fig. 5.15(a)).

We performed 500 PUF bit readings and found that the unstable bit ratio is 5.39% at the nominal condition (1V, 27°C). We also investigate the unstable bit ratios using TMV and CIS based masking techniques. fig. 5.15(b) shows that TMV11 can reduce the ratio down to 1.7%. The CIS based masking technique can reduce the ratio down to 0.027% with TMV11. We tested a mask generated with 80 samples and based on the proposed CIS based technique. It can reduce the unstable bit ratio down to 1.1% without TMV.

We also measured the BER at the nominal condition (1V, 27°C) across several TMV

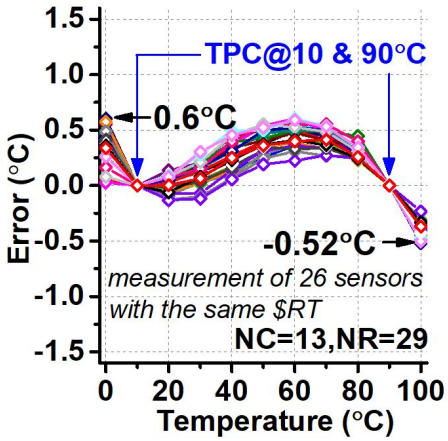


Figure 5.14: T-sensor measurement results: The error of the transformed T-sensors after TPC.

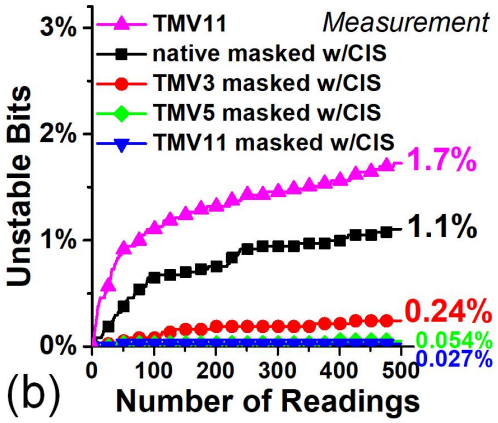
window sizes (fig. 5.15(c)). As we increase the sizes, the BER reduces. Specifically, TMV11 can scale BER down to 0.754%. The CIS based masking (80 samples), combined with TMV11, can reduce the BER down to 0.11% and 0.00197% for the worse and average case, respectively. Native readings exhibit 0.33% and 0.0153% BERs for the worse and average case, respectively. The power consumption of the PUF across temperature and process corners is simulated and shown in fig. 5.15(d).

As shown in fig. 5.16, we also characterized the uniqueness of the PUF outputs by the means of fractional Hamming distance (FHD). The inter-PUF FHD is measured to be 0.5017 (mean). This is close to the ideal value of 0.5, confirming that the PUF codes are highly unique. Due to the limited number of chips measured, we divided the PUF codes into 4 sub-codes for FHD evaluation. We also measured inter- and intra-PUF FHD distributions with TMV11 applied, which show the mean separation of 332×, exhibiting high robustness against temporal noise.

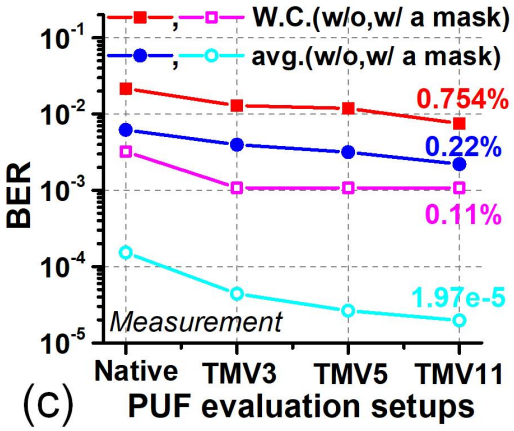
At the corner temperatures of -15°C and 85°C, the proposed PUF instances respectively exhibit the maximum BERs of 6.79% and 7.33% with TMV11. The BERs are measured by

NIST Test Name	Avg. P-val.	Pass?
Frequency	0.1310	YES
Block Frequency	0.7866	YES
Runs	0.9701	YES
Longest run of ones	0.0946	YES
Cumulative Sums	0.1855	YES
FFT	0.0468	YES
non-overlapping template matching	0.6270	YES
Serial	0.1299	YES
Approximate Entropy	0.1232	YES

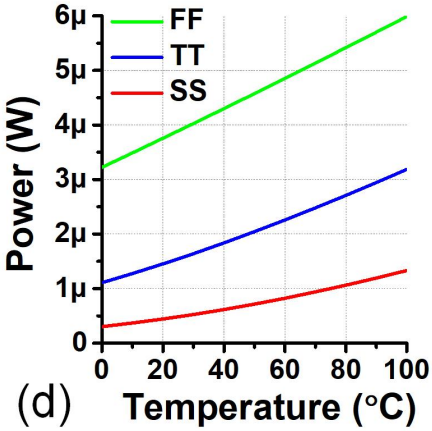
(a)



(b)



(c)



(d)

Figure 5.15: PUF measurement results: (a) Applicable NIST test results on the 3712-bit PUF output. (b) the unstable bit ratios of the PUF with the TMV and CIS. (c) The BER with the TMV and CIS. (d) The power dissipation across corners and temperatures.

comparing the reference PUF outputs generated at the corner temperatures and at 27°C. The temperature-induced bit-flipping ratio per 10°C is measured to be 0.86%. This is calculated from the increase in average BER to rule out the impact of temporal noise. The CIS-generated mask helps reduce BERs. With a TMV11 scheme, as shown in fig. 5.17, it is 0.002% at the nominal temperature, 5.28 % at 15°C and 5.82% at 85°C. Average bit-flipping ratio per 10°C variation is 0.5%. Note that we calibrated CMP's input offset voltage at 27°C only and thus we expect the results can be improved if automatic calibration is used across temperatures.

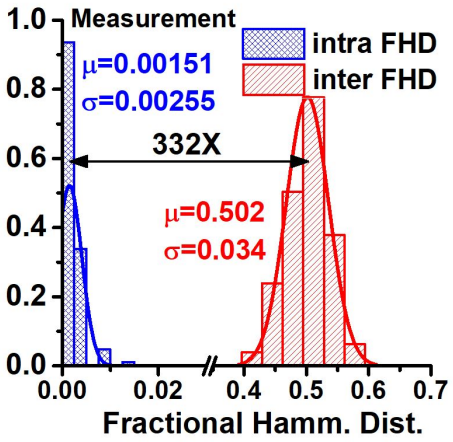


Figure 5.16: PUF measurement: Distributions of the inter-PUF and intra-PUF FHDs

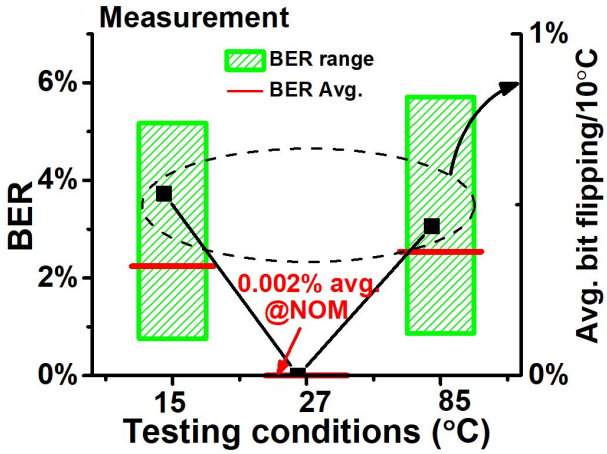


Figure 5.17: PUF measurement: BER across temperature variations

Finally, we compare our transformed T-Sensor and PUF to other recent works. table 5.1 summarizes the recent state-of-the-art temperature sensor circuits that report the similar V_{DD} and temperature operating range. One of the designs that achieves comparable performance and robustness is Ref. [73]. It achieves the post-OPC error of -1.4/+1.4°C, and its front end takes 2,700μm^2 in a 65nm.

table 5.2 summarizes the comparison with recent state-of—the-art PUF circuits. Our proposed transformed PUF achieves substantially better robustness than the previous works based on the power-up reset states of SRAM [76, 80]. Note that Ref. [80] uses the indus-

	Tech. (nm)	VDD range (V)	OPC error (°C)[1]	TPC error (°C)[1]	VDD sensitivity (°C/100mV)	Front-end area (μm²)[2]
This work	**65**	**0.5~1**	**-0.5/+1.5**	**-0.6/+0.5**	**0.46**	**1965**[3]
[4]	180	1.2~2	-0.5/+0.5	-	0.06	72527
[5]	180	1.2	-	-1.4/+1.5	-	48000
[6]	160	0.85~1.2	-0.2/+0.2	-	0.05	4800
[7]	**65**	1	-1.4/+1.4	-	-	2700
[8]	**65**	0.85~1.05	-	-2.3/+2.3	3.4	1180
[2]	**65**	0.6~1	-0.7/+4.7	-1.1/+2.1	0.04	400
[9]	16	0.7	-3/+3	-1/+1	-	5100

Table 5.1: Comparison table for T-sensor

	Tech. (nm)	Unstable Bits %, native/post-processed	BER	Area/Bit (F^2)[1]	Flip Rate per 10°C	Inter-PUF FHD	Intra-PUF FHD	Energy[4] pJ/Bit
This	**65**	**5.39/0.97**	**2.16/0.62**	**0.6k**[2]	**1.1**	**0.502**	**332x**	**0.38**[7]
[10] INV	**65**	2.34/-	-	6k	0.68[3]	0.501	140x	0.015
[10] SA	**65**	1.88/-	-	12k	0.62[3]	0.501	161x	0.163
[11]	22	30/3	8.5/0.97	9.6k	-	0.481	86x[6]	0.013
[12]	65	2.15/-	0.63/-	0.15k[5]	0.99	0.498	174x	-
[13] sym.	130	3.04/-	-	7.1k	0.68	0.506	-	0.93
[14]	65	-	4.5/-	1.1k	1.14[3]	0.491	-	-
[10] SRAM	65	16.6/-	-	0.81k	6.7	0.332	5.5x	1.1
[14] SRAM	65	-	6/-	0.19k	0.33[3]	0.497	-	-

[1]bitcell only [2](BLMUX+PUF+level-shifters)/992 [3]estimated [4]comparing energy w/ native read
[5]off-chip ADC not included [6]with dark-bits [7]simulation

Table 5.2: Comparison table for PUF

trial SRAM bitcells with push rules, which results in much smaller area. However, the use of SRAM power-up states for PUF in general has low robustness. Refs [78, 80] proposed techniques to improve the robustness. Some other weak PUF circuits having the similar robustness against temperature and V_{DD} variations take silicon footprints of 25,296μm^2 [76] to 40,374μm^2 [77] (scaled to 65nm) for the same code length of 928 bits. As shown in fig. 5.18(a), if a SoC integrates dedicated hardware for temperature sensing and PUF, for

example Refs. [73] and [76], the area overhead is 27,996 μm^2, which is 9.8× larger than that of our proposed transformation approach (2,845 μm^2). Note that the dedicated hardware approach also needs microarchitecture modification to have the interface to the T-sensor and the PUF. Thus we consider the overhead of microarchitecture modification to be common in both approaches.

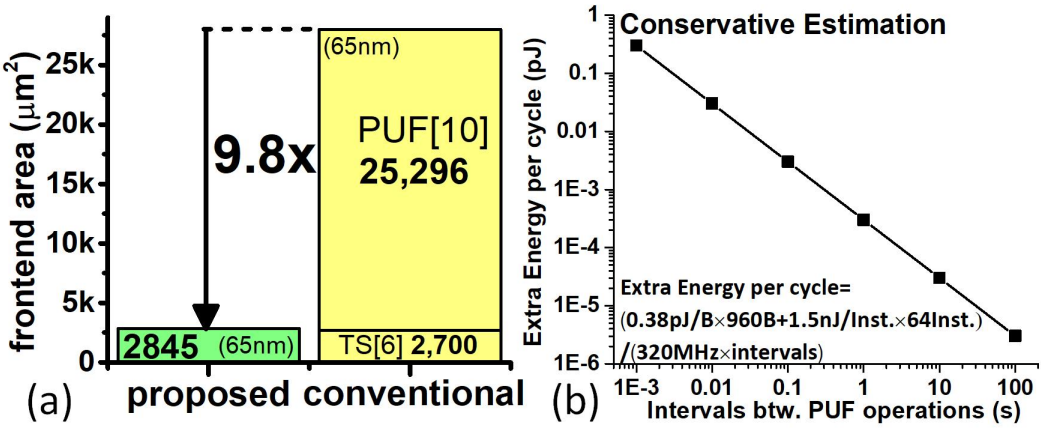


Figure 5.18: (a) The area overhead comparisons. (b) Extra energy per cycle conservatively estimated with PUF operation.

It is noteworthy that asides from energy consumption of T-sensor/PUF operation itself, the proposed transformation technique involves flushing and reloading I$ for each T-sensor/PUF operation. Assuming 1.5nJ per 16Bit instruction energy cost for reloading I$, and every time 2 instructions are reloaded as the array width is 32Bit, there will be 1.5nJ×2/960B=3.13pJ/bit extra energy cost for each PUF bit. Such energy cost dominates the system-level energy consumption of T-sensor and PUF operation. However, as T-sensor and PUF operations are relatively low-frequency tasks, the extra energy cost in such operations will show limited impact on the total system energy efficiency especially considering that the processor will also load instructions into I$ in normal operation, mitigating the relative energy overhead from the I$ reloading.

As is shown in fig. 5.18(b), considering the extra energy consumed per cycle, we conservatively estimated the impact of PUF operation by varying the time intervals between each PUF operation. The PUF operation is chosen as it consumes more energy than the temperature sensor since all 960 bits needs to be evaluated. As can be seen, the extra energy cost is 0.3pJ/cycle (2.8%) even when the PUF operation occurs frequently at every 1ms. Note that here we're conservatively assuming the processor do not access off-chip memory at all.

Meanwhile, the energy cost of masking and TMV can also be estimated. For masking, the extra energy cost can be roughly estimated as an extra instruction for bit-wise masking. Assuming 15 PUF bits are processed per instruction, which is the number of bits generated per PUF operation, the extra energy cost of bit-wise masking will approximately 0.71pJ/bit. Note that the energy cost can be minimized with extra logic hardware, instead of extra instructions assumed in the above estimation. For the TMV, its energy cost can be estimated by multiplying the energy cost in native reading scheme by the number of TMV iterations. For instance, the TMV11 scheme will cost 4.18 (=0.38×11) pJ/bit.

5.5 Conclusions

In this prototype work, we present a μP-SoC prototype that integrates temperature sensing and PUF features for area-constraint IoT devices. We propose a transformation approach which can recycle the I$ temporarily for ambient temperature sensing or PUF code generation. The area overhead of the hardware for the transformation is 9.8× smaller than the overhead incurred by the conventional approach that integrates dedicated hardware for each feature. Measurement results of the prototyped chips show that the transformed

T-sensor and the PUF achieves accuracy, robustness, and area-efficiency comparable to the state of the art.

Conclusion

As transistors continue to scale down in size and there are ever-increasing demands for performance, the chip die area is expected to keep growing. To allow for thermal management, more distributed temperature sensors are required. It is desirable to build such sensor front-ends from standard logic devices, give them small footprints and low calibration cost, and make them robust to variations in processes and voltage.

Meanwhile: transistor aging, especially NBTI in PMOS, has become a main concern in memory reliability. As NBTI gradually increases the PMOS threshold voltage, it affects the bit cells by lowering data retention voltage and decreasing the node charging speed. It can also damage peripheral circuits, such as weakening the retention capability of a domino keeper and increasing the droop of a power gate. It is vital to monitor PMOS aging throughout the lifetime of a chip. We believe in-situ monitoring will become popular if chips can be devised that have low overhead and low integration cost. Such in-situ monitoring must be stable against fluctuations of temperature and voltage to produce accurate readouts during in-field sensing.

In my research, we designed a family of compact temperature sensor front-ends to accommodate various SoC application scenarios (accuracy vs. area). We also investigated how to mitigate the impact of process and voltage variations. The test chip showed that our

proposed sensor front-ends could achieve target accuracy, and they had sufficient tolerance to voltage variations. Their key features were (1) small footprint, (2) using standard logic devices, and (3) low calibration cost.

We then explored an area-efficient PMOS NBTI sensor and developed a technique to integrate it into standard 6T SRAM bit cells. We built two test chips with real SRAM arrays to demonstrate the sensor, in addition to the sensing techniques that can capture PMOS NBTI effects during aging tests. The sensing can be conducted both in-situ and in-field. We also developed a model that takes measurements to estimate the data retention voltage of SRAM. Finally, we put the SRAM, integrated with the aging sensor, into a MIPS microprocessor as the I$ and tapeout in the silicon. We devised the instruction set with an additional AS to enable the sensing of aging on the fly. Based on the novel hardware structure, we also developed and verified a framework to estimate aging in the L1 cache. It can also generate on-time recover vectors to mitigate impacts. We believe our processor is a milestone of in-situ aging sensing technique.

The concepts of SRAM transformation and sensor integration are also used in my last test chip, which is a collaborative project with Jiangyi Li. we build a MIPS microprocessor which can transform I$ into a high-accuracy ambient temperature sensor and stable PUF. This test chip has demonstrated a new approach to reuse on-chip hardware resources.

Bibliography

[1] R. McGowen, C.A. Poirier, C. Bostak, J. Ignowski, M. Millican, W.H. Parks, S. Naffziger, "Power and temperature control on a 90-nm Itanium family processor," *IEEE Journal of Solid-State Circuits*, vol.41, no.1, pp.229-237, Jan. 2006

[2] M. Nakajima, H. Kondo, N. Okumura, N. Masui, Y. Takata, T. Nasu, H. Takata, T. Higuchi, M. Sakugawa, H. Yoneda, H. Fujiwara, K. Ishida, K. Ishimi, S. Kaneko, T. Itoh, M. Sato, O. Yamamoto, K. Arimoto, "Design of a Multi-Core SoC with Configurable Heterogeneous 9 CPUs and 2 Matrix Processors," *IEEE Symposium on VLSI Circuits*, pp.14-15, 2007

[3] D.E. Duarte, G. Geannopoulos, U. Mughal, K.L. Wong, G. Taylor, "Temperature Sensor Design in a High Volume Manufacturing 65nm CMOS Digital Process," *IEEE Custom Integrated Circuits Conference*, pp.221-224, 2007

[4] N. Sakran, M. Yuffe, M. Mehalel, J. Doweck, E. Knoll, A. Kovacs, "The Implementation of the 65nm Dual-Core 64b Merom Processor," *IEEE International Solid-State Circuits Conference*, pp.106,590, 2007

[5] J. Dorsey, S. Searles, M. Ciraula, S. Johnson, N. Bujanos, D. Wu, M. Braganza, S. Meyers, E. Fang, R. Kumar, "An Integrated Quad-Core Opteron Processor," *IEEE International Solid-State Circuits Conference*, pp.102-103, 2007

[6] M.S. Floyd, S. Ghiasi, T.W. Keller, K. Rajamani, F.L. Rawson, J.C. Rubio, M.S. Ware, "System power management support in the IBM POWER6 microprocessor," *IBM Journal of Research and Development*, vol.51, no.6, pp.733-746, Nov. 2007

[7] E. Saneyoshi, K. Nose, M. Kajita, M. Mizuno, "A 1.1V 35m × 35m thermal sensor with supply voltage sensitivity of 2°C/10%-supply for thermal management on the SX-9 supercomputer," *IEEE Symposium on VLSI Circuits*, pp.152-153, 2008

[8] R. Kumar, G. Hinton, "A family of 45nm IA processors," *IEEE International Solid-State Circuits Conference*, pp.58-59, 2009

[9] R. Kuppuswamy, S.R. Sawant, S. Balasubramanian, P. Kaushik, N. Natarajan, J.D. Gilbert, "Over one million TPCC with a 45nm 6-core Xeon® CPU," *IEEE International Solid-State Circuits Conference*, pp.70-71, 2009

[10] M. Floyd, M. Allen-Ware, K. Rajamani, B. Brock, C. Lefurgy, A.J. Drake, L. Pesantez, T. Gloekler, J.A. Tierno, P. Bose, A. Buyuktosunoglu, "Introducing the Adaptive Energy Management Features of the Power7 Chip," *IEEE Micro*, vol.31, no.2, pp.60-75, March-April 2011

[11] S. Dighe, S. Gupta, V. De, S. Vangal, N. Borkar, S. Borkar, K. Roy, "A 45nm 48-core IA processor with variation-aware scheduling and optimal core mapping," *IEEE Symposium on VLSI Circuits*, pp.250-251, 2011

[12] E.J. Fluhr, J. Friedrich, D. Dreps, V. Zyuban, G. Still, C. Gonzalez, A. Hall, D. Hogenmiller, F. Malgioglio, R. Nett, J. Paredes, J. Pille, D. Plass, R. Puri, P. Restle, D. Shan, K. Stawiasz, Z.T. Deniz, D. Wendel, M. Ziegler, "5.1 POWER8TM: A 12-core server-class processor in 22nm SOI with 7.6Tb/s off-chip bandwidth," *IEEE International Solid-State Circuits Conference*, pp.96-97, 2014

[13] J.S. Shor, K. Luria, "Miniaturized BJT-Based Thermal Sensor for Microprocessors in 32- and 22-nm Technologies," *IEEE Journal of Solid-State Circuits*, vol.48, no.11, pp.2860-2867, Nov. 2013

[14] H. Lakdawala, Y.W. Li, A. Raychowdhury, G. Taylor, K. Soumyanath, "A 1.05V 1.6mW, 0.45°C 3 Resolution Based Temperature Sensor With Parasitic Resistance Compensation in 32 nm Digital CMOS Process," *IEEE Journal of Solid-State Circuits*, vol.44, no.12, pp.3621-3630, Dec. 2009

[15] G.R. Chowdhury, A. Hassibi, "A 0.001mm2 100µW on-chip temperature sensor with ±1.95°C (3) Inaccuracy in 32nm SOI CMOS," *IEEE International Symposium on Circuits and Systems*, pp.1999-2002, 2012

[16] S. Paek, W. Shin, J. Lee, H.E. Kim, J.S. Park, L.S. Kim, "All-digital hybrid temperature sensor network for dense thermal monitoring," *IEEE International Solid-State Circuits Conference*, pp.260-261, 2013

[17] M. Sasaki, M. Ikeda, K. Asada, "A Temperature Sensor With an Inaccuracy of -1/+0.8°C Using 90-nm 1-V CMOS for Online Thermal Monitoring of VLSI Circuits," *IEEE Transactions on Semiconductor Manufacturing*, vol.21, no.2, pp.201-208, May 2008

[18] R. Quan, U. Sonmez, F. Sebastiano, K.A.A. Makinwa, "A 4600m2 1.5°C (3) 0.9kS/s Thermal-Diffusivity Temperature Sensor with VCO-Based Readout," *IEEE International Solid-State Circuits Conference*, pp.488-489, 2015

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

[20] I.M. Filanovsky, A. Allam, "Mutual compensation of mobility and threshold voltage temperature effects with applications in CMOS circuits," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol.48, no.7, pp.876-884, Jul 2001

[21] C.C. Enz, G.C. Temes, "Circuit techniques for reducing the effects of op-amp imperfections: autozeroing, correlated double sampling, and chopper stabilization," *Proceedings of the IEEE*, vol.84, no.11, pp.1584-1614, Nov 1996

[22] Y. Tsividis, C. McAndrew, Operation and Modeling of the MOS Transistors, third edition. Oxford University Press, 2011.

[23] T. Yang, S. Kim, P.R. Kinget, M. Seok, "0.6-to-1.0V 279m2, 0.92W temperature sensor with less than +3.2/-3.4°C error for on-chip dense thermal monitoring," *IEEE International Solid-State Circuits Conference*, pp.282-283, 2014

[24] http://isscc.org/doc/2015/isscc2015_trends.pdf

[25] http://web.stanford.edu/∼murmann/adcsurvey.html

[26] X. Li, J. Qin and J. B. Bernstein, "Compact Modeling of MOSFET Wearout Mechanisms for Circuit-Reliability Simulation," in *IEEE Transactions on Device and Materials Reliability*, vol. 8, no. 1, pp. 98-111, March 2008.

[27] B. C. Paul, Kunhyuk Kang, H. Kufluoglu, M. A. Alam and K. Roy, "Impact of NBTI on the temporal performance degradation of digital circuits," in *IEEE Electron Device Letters*, vol. 16, no. 8, pp. 560-561, Aug. 2005.

[28] K. Bernstein, D. J. Frank, A. E. Gattiker, W. Haensch, B. L. Ji, S. R. Nassif, E. J. Nowak, D. J. Pearson and N. J. Rohrer, "High-performance CMOS variability in the 65-nm regime and beyond," in *IBM Journal of Research and Development*, vol. 50, no. 4.5, pp. 433-449, July 2006.

[29] W. Wang, V. Reddy, A. T. Krishnan, R. Vattikonda, S. Krishnan and Y. Cao, "Compact Modeling and Simulation of Circuit Reliability for 65-nm CMOS Technology," in *IEEE Transactions on Device and Materials Reliability*, vol. 7, no. 4, pp. 509-517, Dec. 2007.

[30] V. Reddy, A. T. Krishnan, A. Marshall, J. Rodriguez, S. Natarajan, T. Rost and S. Krishnan, "Impact of negative bias temperature instability on digital circuit reliability," *IEEE International Reliability Physics Symposium*. 2001, pp. 148-154.

[31] A. Haggag, G. Anderson, S. Parihar, D. Burnett, G. Abeln, J. Higman and M. Moosa, "Understanding SRAM High-Temperature-Operating-Life NBTI: Statistics and Permanent vs Recoverable Damage," *IEEE International Reliability Physics Symposium Proceedings*. 45th Annual, Phoenix, AZ, 2007, pp. 451-456.

[32] T. Fischer, E. Amirante, K. Hofmann, M. Ostermayr, P. Huber and D. Schmitt-Landsiedel, "A 65nm test structure for the analysis of NBTI induced statistical variation in SRAM transistors," *European Solid-State Device Research Conference*, Edinburgh, 2008, pp. 51-54.

[33] P. Kolar, E. Karl, U. Bhattacharya, F. Hamzaoglu, H. Nho, Y. Ng, Y. Wang and K. Zhang, "A 31 nm High-k Metal Gate SRAM With Adaptive Dynamic Stability Enhancement for Low-Voltage Operation," in *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 76-84, Jan. 2011.

[34] M. Qazi, K. Stawiasz, L. Chang and A. P. Chandrakasan, "A 511kb 8T SRAM Macro Operating Down to 0.57V With an AC-Coupled Sense Amplifier and Embedded Data-Retention-Voltage Sensor in 45nm SOI CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 85-96, Jan. 2011.

[35] J. Wang and B. H. Calhoun, "Techniques to Extend Canary-Based Standby VDD Scaling for SRAMs to 45 nm and Beyond," in *IEEE Journal of Solid-State Circuits*, vol. 43, no. 11, pp. 1514-1513, Nov. 2008.

[36] M. Namaki-Shoushtari, A. Rahimi, N. Dutt, P. Gupta and R. K. Gupta, "ARGO: Aging-aware GPGPU register file allocation," 2013 *International Conference on Hardware/Software Codesign and System Synthesis* (CODES+ISSS), Montreal, QC, 2013, pp. 1-9.

[37] A. J. Bhavnagarwala, S. Kosonocky, C. Radens, Y. Chan, K. Stawiasz, U. Srinivasan, S. P. Kowalczyk and M. M. Ziegler, "A Sub-600-mV, Fluctuation Tolerant 65-nm CMOS SRAM Array With Dynamic Cell Biasing," in *IEEE Journal of Solid-State Circuits*, vol. 43, no. 4, pp. 946-955, April 2008.

[38] Z. Guo, A. Carlson, L. T. Pang, K. T. Duong, T. J. K. Liu and B. Nikolic, "Large-Scale SRAM Variability Characterization in 45 nm CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 44, no. 11, pp. 3174-3191, Nov. 2009.

[39] J. Tsai, S. O. Toh, Z. Guo, L. T. Pang, T. J. K. Liu and B. Nikolic, "SRAM stability characterization using tunable ring oscillators in 45nm CMOS," 2010 *IEEE International Solid-State Circuits Conference*, San Francisco, CA, 2010, pp. 354-355.

[40] F. Ahmed and L. Milor, "Reliable cache design with on-chip monitoring of NBTI degradation in SRAM cells using BIST," *VLSI Test Symposium*, Santa Cruz, CA, 2010, pp. 63-68.

[41] H. Park and C. K. K. Yang, "In Situ SRAM Static Stability Estimation in 65-nm CMOS," in *IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 1541-1549, Oct. 2013.

[42] S. O. Toh, Z. Guo and B. Nikolić, "Dynamic SRAM stability characterization in 45nm CMOS," *IEEE Symposium on VLSI Circuits*, Honolulu, HI, 2010, pp. 35-36.

[43] K. Zhang, U. Bhattacharya, Z. Chen, F. Hamzaoglu, D. Murray, N. Vallepalli, Y. Wang, B. Zheng and M. Bohr, "SRAM design on 65-nm CMOS technology with dynamic sleep transistor for leakage reduction," in *IEEE Journal of Solid-State Circuits*, vol. 40, no. 4, pp. 895-901, April 2005.

[44] S. Yoshimoto, T. Amashita, S. Okumura, K. Nii, H. Kawaguchi and M. Yoshimoto, "NMOS-inside 6T SRAM layout reducing neutron-induced multiple cell upsets," *IEEE International Reliability Physics Symposium (IRPS)*, Anaheim, CA, 2011, pp. 5B.5.1-5B.5.5.

[45] M. Seok, G. Kim, D. Blaauw and D. Sylvester, "A Portable 1-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5 V," in *IEEE Journal of Solid-State Circuits*, vol. 47, no. 10, pp. 1534-1545, Oct. 2011.

[46] UC Berkeley Device Group. BSIM4 Model, Dept. EECS, Univ. of California, Berkeley, CA, USA, Nov. 1, 2013 [Online]. Available: www-device.eecs.berkeley.edu/bsim/?page=BSIM4

[47] A. E. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra and M. A. Alam, "Recent Issues in Negative-Bias Temperature Instability: Initial Degradation, Field Dependence of Interface Trap Generation, Hole Trapping Effects, and Relaxation," in *IEEE Transactions on Electron Devices*, vol. 54, no. 9, pp. 1143-1154, Sept. 2007.

[48] S. Chakravarthi, A. Krishnan, V. Reddy, C. F. Machala and S. Krishnan, "A comprehensive framework for predictive modeling of negative bias temperature instability," 2004 *IEEE International Reliability Physics Symposium. Proceedings*, 2004, pp. 173-181.

[49] S. Rangan, N. Mielke and E. C. C. Yeh, "Universal recovery behavior of negative bias temperature instability [PMOSFETs]," *IEEE International Electron Devices Meeting* 2003, Washington, DC, USA, 2003, pp. 14.3.1-14.3.4.

[50] H. Reisinger, O. Blank, W. Heinrigs, A. Muhlhoff, W. Gustin and C. Schlunder, "Analysis of NBTI Degradation- and Recovery-Behavior Based on Ultra Fast VT-Measurements," 2006 *IEEE International Reliability Physics Symposium Proceedings*, San Jose, CA, 2006, pp. 448-453.

[51] S. V. Kumar, K. H. Kim and S. S. Sapatnekar, "Impact of NBTI on SRAM read stability and design for reliability," *International Symposium on Quality Electronic Design (ISQED'06)*, San Jose, CA, 2006, pp. 6 pp.-118.

[52] D. Papagiannopoulou, P. Prasertsom and I. Bahar, "Flexible data allocation for scratchpad memories to reduce NBTI effects," *International Symposium on Quality Electronic Design (ISQED)*, Santa Clara, CA, 2013, pp. 60-67.

[53] A. Calimera, M. Loghi, E. Macii and M. Poncino, "Dynamic indexing: Concurrent leakage and aging optimization for caches," *ACM/IEEE International Symposium on Low-Power Electronics and Design (ISLPED)*, Austin, TX, USA, 2010, pp. 343-348.

[54] J. Wang, S. Nalam, Zhenyu Qi, R. W. Mann, M. Stan and B. H. Calhoun, "Improving SRAM Vmin and yield by using variation-aware BTI stress," *IEEE Custom Integrated Circuits Conference 2010*, San Jose, CA, 2010, pp. 1-4.

[55] T. Yang, D. Kim, P. R. Kinget, M. Seok, "In-situ Techniques for In-field sensing of NBTI Degradation in an SRAM Register File," *IEEE International Solid-State Circuits Conference (ISSCC)*, 2015

[56] V. M. van Santen, J. Martin-Martinez, H. Amrouch, M. M. Nafria and J. Henkel, "Reliability in Super- and Near-Threshold Computing: A Unified Model of RTN, BTI,

and PV," in *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 1, pp. 293-306, Jan. 2018.

[57] H. Amrouch, V. M. van Santen, T. Ebi, V. Wenzel and J. Henkel, "Towards interdependencies of aging mechanisms," 2014 *IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, San Jose, CA, 2014, pp. 478-485.

[58] Z. C. Lee, M. S. M. Siddiqui, Z. H. Kong and T. T. H. Kim, "An 8T SRAM with BTI-Aware Stability Monitor and two-phase write operation for cell stability improvement in 28-nm FDSOI," *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, Lausanne, 2016, pp. 437-440.

[59] N. Goel, P. Dubey, J. Kawa and S. Mahapatra, "Impact of time-zero and NBTI variability on sub-20nm FinFET based SRAM at low voltages," 2015 *IEEE International Reliability Physics Symposium*, Monterey, CA, 2015, pp. CA.5.1-CA.5.7.

[60] A. Calimera, M. Loghi, E. Macii and M. Poncino, "Partitioned cache architectures for reduced NBTI-induced aging," 2011 *Design, Automation & Test in Europe*, Grenoble, 2011, pp. 1-6.

[61] A. Gebregiorgis, M. Ebrahimi, S. Kiamehr, F. Oboril, S. Hamdioui and M. B. Tahoori, "Aging mitigation in memory arrays using self-controlled bit-flipping technique," The 20th *Asia and South Pacific Design Automation Conference*, Chiba, 2015, pp. 231-236.

[62] T. Yang, P. R. Kinget and M. Seok, "Register file circuits and post-deployment framework to monitor aging effects in field," *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, Lausanne, 2016, pp. 425-428

[63] M. Namaki-Shoushtari, A. Rahimi, N. Dutt, P. Gupta and R. K. Gupta, "ARGO: Aging-aware GPGPU register file allocation," *2013 International Conference on Hardware/Software Codesign and System Synthesis (CODES+ISSS)*, Montreal, QC, 2013, pp. 1-9

[64] T. Yang, P. R. Kinget and M. Seok, "Register file circuits and post-deployment framework to monitor aging effects in field," *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, 2016, pp. 425-428

[65] T. Yang, D. Kim, J. Li, P. R. Kinget and M. Seok, "$In-Situ$ and In-Field Technique for Monitoring and Decelerating NBTI in 6T-SRAM Register Files," in *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 11, pp. 2241-2253, Nov. 2018

[66] F. Ahmed and L. Milor, "Reliable cache design with on-chip monitoring of NBTI degradation in SRAM cells using BIST," *2010 28th VLSI Test Symposium (VTS)*, 2010, pp. 63-68

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

[68] T. Yang, S. Kim, P. R. Kinget and M. Seok, "Compact and Supply-Voltage-Scalable Temperature Sensors for Dense On-Chip Thermal Monitoring," in IEEE Journal of Solid-State Circuits, vol. 50, no. 11, pp. 2773-2785, Nov. 2015, doi: 10.1109/JSSC.2015.2476815.

[69] J. Li and M. Seok, "Ultra-Compact and Robust Physically Unclonable Function Based on Voltage-Compensated Proportional-to-Absolute-Temperature Voltage Generators," in IEEE Journal of Solid-State Circuits, vol. 51, no. 9, pp. 2192-2202, Sept. 2016, doi: 10.1109/JSSC.2016.2586498.

[70] C. Wu, W. Chan and T. Lin, "A 80kS/s 36µW resistor-based temperature sensor using BGR-free SAR ADC with a unevenly-weighted resistor string in 0.18µm CMOS," 2011 Symposium on VLSI Circuits - Digest of Technical Papers, 2011, pp. 222-223.

[71] S. Jeong, Z. Foo, Y. Lee, J. Sim, D. Blaauw and D. Sylvester, "A Fully-Integrated 71 nW CMOS Temperature Sensor for Low Power Wireless Sensor Nodes," in IEEE Journal of Solid-State Circuits, vol. 49, no. 8, pp. 1682-1693, Aug. 2014, doi: 10.1109/JSSC.2014.2325574.

[72] K. Souri, Y. Chae, F. Thus and K. Makinwa, "12.7 A 0.85V 600nW all-CMOS temperature sensor with an inaccuracy of ±0.4°C (3) from -40 to 125°C," 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014, pp. 222-223, doi: 10.1109/ISSCC.2014.6757409.

[73] S. Hwang, J. Koo, K. Kim, H. Lee and C. Kim, "A 0.008 mm^2 500 /spl mu/W 469 kS/s Frequency-to-Digital Converter Based CMOS Temperature Sensor With Process Variation Compensation," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 60, no. 9, pp. 2241-2248, Sept. 2013, doi: 10.1109/TCSI.2013.2246254.

[74] T. Anand, K. A. A. Makinwa and P. K. Hanumolu, "A self-referenced VCO-based temperature sensor with 0.034°C/mV supply sensitivity in 65nm CMOS," 2015 Symposium on VLSI Circuits (VLSI Circuits), 2015, pp. C200-C201, doi: 10.1109/VLSIC.2015.7231257.

[75] J. Horng et al., "A 0.7V resistive sensor with temperature/voltage detection function in 16nm FinFET technologies," 2014 Symposium on VLSI Circuits Digest of Technical Papers, 2014, pp. 1-2, doi: 10.1109/VLSIC.2014.6858376.

[76] A. B. Alvarez, W. Zhao and M. Alioto, "Static Physically Unclonable Functions for Secure Chip Identification With 1.9–5.8% Native Bit Instability at 0.6–1 V and 15 fJ/bit in 65 nm," in IEEE Journal of Solid-State Circuits, vol. 51, no. 3, pp. 763-775, March 2016, doi: 10.1109/JSSC.2015.2506641.

[77] S. K. Mathew et al., "16.2 A 0.19pJ/b PVT-variation-tolerant hybrid physically unclonable function circuit for 100% stable secure key generation in 22nm CMOS," 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2014, pp. 278-279, doi: 10.1109/ISSCC.2014.6757433.

[78] J. Li, T. Yang and M. Seok, "A technique to transform 6T-SRAM arrays into robust analog PUF with minimal overhead," 2017 IEEE International Symposium on Circuits and Systems (ISCAS), 2017, pp. 1-4, doi: 10.1109/ISCAS.2017.8050630.

[79] Y. Su, J. Holleman and B. P. Otis, "A Digital 1.6 pJ/bit Chip Identification Circuit Using Process Variations," in IEEE Journal of Solid-State Circuits, vol. 43, no. 1, pp. 69-77, Jan. 2008, doi: 10.1109/JSSC.2007.910961.

[80] R. Maes, V. Rozic, I. Verbauwhede, P. Koeberl, E. van der Sluis and V. van der Leest, "Experimental evaluation of Physically Unclonable Functions in 65 nm CMOS," 2012 Proceedings of the ESSCIRC (ESSCIRC), 2012, pp. 486-489, doi: 10.1109/ESSCIRC.2012.6341361.

[81] T. Yang, P. R. Kinget and M. Seok, "Register file circuits and post-deployment framework to monitor aging effects in field," ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference, 2016, pp. 425-428, doi: 10.1109/ESSCIRC.2016.7598332.

[82] M. Fojtik et al., "A Millimeter-Scale Energy-Autonomous Sensor System With Stacked Battery and Solar Cells," in IEEE Journal of Solid-State Circuits, vol. 48, no. 3, pp. 801-813, March 2013, doi: 10.1109/JSSC.2012.2233352.

[83] M. Seok, G. Kim, D. Blaauw and D. Sylvester, "A Portable 2-Transistor Picowatt Temperature-Compensated Voltage Reference Operating at 0.5 V," in IEEE Journal of Solid-State Circuits, vol. 47, no. 10, pp. 2534-2545, Oct. 2012, doi: 10.1109/JSSC.2012.2206683.

[84] K. Yang, Q. Dong, D. Blaauw and D. Sylvester, "8.3 A 553F2 2-transistor amplifier-based Physically Unclonable Function (PUF) with 1.67% native instability," 2017 IEEE International Solid-State Circuits Conference (ISSCC), 2017, pp. 146-147, doi: 10.1109/ISSCC.2017.7870303.

[85] Van Herrewege, Anthony, et al., "Reverse Fuzzy Extractors: Enabling Lightweight Mutual Authentication for PUF-Enabled RFIDs," Financial Cryptography, Vol. 7397, pp. 374-389, 2012.

[86] S. Devadas, E. Suh, S. Paral, R. Sowell, T. Ziola and V. Khandelwal, "Design and Implementation of PUF-Based "Unclonable" RFID ICs for Anti-Counterfeiting and Security Applications," 2008 IEEE International Conference on RFID, 2008, pp. 58-64, doi: 10.1109/RFID.2008.4519377.

[87] B. Karpinskyy, Y. Lee, Y. Choi, Y. Kim, M. Noh and S. Lee, "8.7 Physically unclonable function for secure key generation with a key error rate of 2E-38 in 45nm smart-card chips," 2016 IEEE International Solid-State Circuits Conference (ISSCC), 2016, pp. 158-160, doi: 10.1109/ISSCC.2016.7417955.

[88] Mingoo Seok et al., "The Phoenix Processor: A 30pW platform for sensor applications," 2008 IEEE Symposium on VLSI Circuits, 2008, pp. 188-189, doi: 10.1109/VLSIC.2008.4586001.

[89] G. Chen et al., "Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells," 2010 IEEE International Solid-State Circuits Conference - (ISSCC), 2010, pp. 288-289, doi: 10.1109/ISSCC.2010.5433921.

[90] Opencores.org, Educational 16-bit MIPS Processor