Integrated Circuit Components Design and Modeling for Ultra-Low Power Internet-of-Things Applications

---

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

---

in partial fulfillment
of the requirements for the degree

Doctor of Philosophy

by

Ningxi Liu

May 2019

APPROVAL SHEET

This Dissertation
is submitted in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

Author Signature: _____

This Dissertation has been read and approved by the examining committee:

Advisor: Benton H. Calhoun

Committee Member: Mircea R. Stan

Committee Member: Scott T. Acton

Committee Member: Steven M. Bowers

Committee Member: Brad Campbell

Committee Member: _____

Accepted for the School of Engineering and Applied Science:

Craig H. Benson, School of Engineering and Applied Science

May 2019

Abstract

The internet-of-things (IoT) is playing a role in revolutionizing human life by providing a network of devices with smart sensors, actuators, and network connectivity. Reliable systems with diverse functionality are in demand for a wide variety of IoT applications, for instance, healthcare, smart building, and computer vision. As billions of IoT devices are emerging in every corner of the world, there is a need for low-cost systems capable of sensing, processing, storing, and transmitting data. Ultra-low power (ULP) is a necessary feature for such systems because it is too costly and impractical to frequently replace or recharge the vast number of batteries to power these devices. In the quickly evolving battery-less systems, the power from energy harvesters also cannot reliably sustain high-power-consumption solutions. To address the dilemma between the growing need for greater functionality and lower power consumption, we propose to develop a full series of cutting-edge ULP integrated circuit (IC) components, such as subthreshold embedded static random access memory (SRAM), wake-up receivers (WURX), clock references, and deep neural network (DNN) hardware accelerators, to enable low-cost and ULP IoT systems. These four IC components are studied because they are critical circuit blocks for achieving the ULP operations of IoT system-on-chip (SoCs), and they can employ different low-power techniques to effectively reduce the system power consumption.

Circuit modeling plays a vital role in guaranteeing reliable operations in the subthreshold region, speeding up the design period, predicting the circuit and system performance, and guiding the direction for design improvements. The challenges of designing reliable and ULP IC components are dramatically different from the traditional performance-driven IC designs because the performance of CMOS devices is more sensitive to the process variation in the near

or subthreshold region.

SRAM could consume up to 50% of the power consumed by an IoT system, so it is desirable to operate in the subthreshold region to suppress the active and the standby power, where it also faces read and write reliability issues. The SRAM yield analysis model utilizes the normal distribution feature and the importance sampling of SRAM metrics to estimate the bit error rate (BER) of different failure types by speeding up the simulation time by 10,000x compared to the conventional Monte Carlo simulation. An SRAM bitcell auto-generation flow and an SRAM macro design exploration tool are proposed to smartly make design decisions for a 2 KB SRAM testchip in the 65 nm technology while satisfying user requirements and guaranteeing reliability in the subthreshold region.

WURXs could relieve the burden of milli-watt level power dissipation of the radio system by waking up the primary receiver from the idle mode rather than being active all the time. The challenges of designing a nano-watt level WURX are improving the sensitivity for remote wake-up signals and rejecting the false alarms. We propose a correlator wake-up code model to guide the WURX's baseband circuit design and improve the sensitivity by choosing appropriate comparator threshold voltages and correlator wake-up codes. Assisted by the robust baseband circuit, our WURX taped-out in the 130 nm technology achieves a -76 dBm sensitivity and less than one false wake-up (FWU) per hour at 10 nW of power consumption.

Commercial radio systems such as the Bluetooth low-energy (BLE) system require an off-chip crystal as the super-stable clock reference, which increases the cost and bulkiness of IoT devices. On-chip clock references design is extremely challenging because the requirement of frequency stability needs to be less than 150 ppm across the process, voltage, and temperature (PVT) corners. We propose two circuit models to improve the temperature and supply

voltage stability of RC relaxation oscillators (ROSC). An on-chip ROSC with both analog and digital frequency compensation is taped-out in the 65 nm technology, and it achieves a temperature coefficient (TC) of 2.5 ppm/^oC and an absolute variation of 100 ppm over the body-compatible range of 0 to 40^oC. The supply voltage stability is also improved by 30% with a simple outside capacitor. Power consumption of the ROSC is reduced from 69 μW to 100 nW by supporting the power gating technique.

DNN hardware accelerators as the artificial intelligence (AI) inference computing engine are appealing for supporting computer vision applications in IoT systems. In-memory computing (IMC) is a new DNN computing architecture that can relieve the data movement issue of von Neumann architectures, thus potentially achieving energy-efficient computation. However, the process variation of the on-chip memory bitcells and the noise in the mixed-signal computation introduce precision degradation of DNN inference. We propose an IMC accuracy model to guide the direction for choosing appropriate memory micro-architectures and to predict the impact of IMC accuracy loss on the DNN inference precision. A 30 fJ per multiplication and accumulation (MAC) SRAM-based IMC architecture with binary weights and 2-bit activations is predicted by the proposed accuracy loss model to achieve 97.7% precision in the hand-written digit recognition.

Acknowledgement

Thanks to all the committee members, Dr. Mircea R. Stan, Dr. Scott T. Acton, Dr Steve M. Bowers, and Dr. Brad Campbell for advising on my dissertation, and especially to my advisor Dr. Benton H. Calhoun for educating me with the valuable insights on both the technical skills and the methodologies during the four and a half years of my Ph.D. life. Also I really appreciate your patience during the dissertation defense, and it will be an important lesson in my life.

Thanks to my parents for supporting me to pursue knowledge in school for more than 20 years. Now I am ready to take over the responsibilities.

Thanks to my wife Muyang for the understandings and accompany after we met. You have brought me a lot of happiness and positive energy to go through all the challenges. I am confident that we can go through all the happiness and sorrows in the future.

Thanks to all the former and current students and research scientists, Yanqing, Aatmesh, Kyle, Jim, Alicia, Seyi, Yu, Patricia, Manula, Farah, Chris, Dilip, Kevin, He, Arijit, Divya, Abhishek, Harsh, Jacob, Shuo, Daniel, Rishika, Sumanth, Anjana, Henry, Peng, Nick, Chien-Hen, Shourya, Sudipta, Haoyi, Jesse, Pouyan, Divya, and etc. I am so glad to know and work with you.

Thanks to Tom, Nikola, Brian, Rangha, Angad, Miaorong, Sanquan, Xi, Sudhir, Sophia, John, Brucek, Bill, and other colleagues at Nvidia Research for the insightful discussions during my internship. The experience opened a new window of research for me.

Special thanks to Terry for your help and kindness. You make the Bengroup like a family.

List of Figures

1 (a) A 6T SRAM bitcell schematic diagram (b) SRAM butter fly curves, and SNM for read and hold 9

2 (a) A 6T SRAM bitcell schematic during WL sweeping (b) SRAM WM defined by the difference between the WL voltage and V_{DD} at the crossing point of Q and QB. 9

3 (a) RSNM under the impact of random process variations. RSNM is the minimum between RSNM0 and RSNM1. (b) WM under the impact of random process variations. WM is the minimum between WM0 and WM1. 10

4 (a) Readable bitcell timing waveform, and read critical time (RTcrit) defined by the offset voltage between BL and BLB. (b) Read failure waveform due to read data disturbance. (c) Write-able bitcell timing waveform, and write critical time (WRcrit) defined by the crossing point of Q and QB. (d) Write failure waveform due to non-write-able bitcell. 11

5 16K points Monte Carlo simulations of (a) the BERs calculated from the RSNM, the transient readability, the transient read stability, and the transient half-select (HS) stability, and (b) the BERs calculated from the WM and the transient write-ability. The clock period is 15 μs for the transient simulations. 13

6 SRAM bitcell auto-generation flow 15

7 Normalized NMOS turned-off current of HVT, RVT, and LVT type vs. channel width at 0.5V 18

8 Normalized NMOS and PMOS turned-on current of HVT, RVT, and LVT type vs. channel width at 0.5V 18
9 Monte carlo simulation results on the initial bitcell size. (a) Normally distributed RSNM; (b) Normally distributed WM 19
10 Monte carlo simulation results on the adjusted bitcell size. (a) Normally distributed RSNM; (b) Normally distributed WM; (c) Normally distributed WM with 20% of negative BL assist 20
11 Monte carlo simulation results on the (a) RTcrit, (b) 1/RTcrit, (c) WTcrit, and (d) 1/WTcrit. The RTcrit and WTcrit distributions have a long tail on the larger delay side. The 1/RTcrit and 1/WTcrit distributions are normal. 22
12 (a) Illustration of the two-dimension parameter space and the boundary between the pass region and failure region. P1 is the MPFP by only considering the dominating parameter x. P2 is the MPFP by considering both parameters x and y. (b) The sensitivity analysis of $f(x, y)$ on x and y. 23
13 Evolutions of the failure probability using one-dimensional importance sampling (IS1) based on P1, two-dimensional importance sampling (IS2) based on P2, and Monte Carlo simulations. 24
14 (a) Internal node Q is disturbed during the read operation. (b) Sensitivity analysis on the maximum voltage of Q with simulations.The maximum voltage of Q is sensitive to Vth variations in PDL and PGL. 25
15 (a) A read disturbance is related to the trip voltage of the right inverter in the bitcell. (b) Sensitivity analysis on the trip voltage with static simulations shows it is sensitive to Vth variations in PDR. 26
16 Chip measurement results of the 2KB SRAM 29

17 Delay and energy of CACTI, a commercial technology and the predictive technology model (PTM) of a gate chain using high performance (HP) and low power (LP) transistors [21] 31

18 8T 1R/1W port bitcell with (a) differential BL sensing scheme, (b) single-ended BL sensing scheme [21] 32

19 Hierarchical BL structures of both read and write operations. Global BL is divided into N local BLs, and each local is constituted with given number of bitcells. Global RBL is for read operation, and it is realized by AND of an upper global RBL and a lower global RBL to further reduce BL parasitic [21] 33

20 Trends of maximum data throughput under various memory capacities for 6T differential BL, 8T 1R/1W single-ended BL, 8T 1R/1W differential BL, 8T 1R/1W single-ended BL with local BL (16 bits/LBL and 32 bits/LBL), 10T 2R/1W single-ended BL and 10T 2R/1W differential BL schemes [21] 35

21 Trends of minimum energy consumption per operation under various memory capacities for 6T differential BL, 8T 1R/1W single-ended BL, 8T 1R/1W differential BL, 8T 1R/1W single-ended BL with local BL (16 bits/LBL and 32 bits/LBL), 10T 2R/1W single-ended BL and 10T 2R/1W differential BL schemes. [21] 37

22 Pareto curves of 8T 1R/1W bitcells with four different BL sensing scheme and one combined at 0.5 KB capacity. [21] 38

23 Pareto curves of 8T 1R/1W bitcells with four different BL sensing scheme and one combined at 8 KB capacity. [21] 38

24 Simulation results comparison between ViPro and full register file schematic. (a)Read energy, (b) Read delay, (c) Write energy, (d) Write delay. [21] 41

25 (a) Conventional on-chip RC relaxation oscillator with two sets of comparators and RC components. (b) Timing waveforms. [42] 46
26 (a) Conventional on-chip RC relaxation oscillator with a single comparator and RC components. (b) Timing waveforms [41]. 47
27 Concept of the proposed RC relaxation oscillator with one comparator and inverted capacitor bank. 48
28 Change in the clock period T_{OSC} vs. temperature with different first order and 2nd order TC subcomponents of td, toffset, and t_{RC} in (a) and (b). The flattest region of the T_{OSC} occurs at different temperatures for (a) and (b), indicating that selecting between the two different t_{RC} can improve the TC of T_{OSC} across the full range. 49
29 Supply voltage sensitivity with different values of C_{OUT}. 51
30 The relaxation oscillator system circuit diagram. [47] 52
31 Simulated relaxation oscillator frequency compensation with PTAT current reference. [47] 54
32 Measured relaxation oscillator frequency with different Rpdiff configurations. [47] 54
33 Measured ring oscillator frequency and RO counter values. [47] 55
34 Simulated timing waveforms during start-up. [47] 56
35 Measured ROSC frequency with DFC automatically tuning the capacitor and resistor bank based on the temperature sensor output. [47] . 57
36 Measured RMS clock period jitter. [47] 58
37 Measured ROSC frequency of 3 chips at different supply voltages with two different outside capacitor values. 58
38 Performance comparison. 59
39 Die photo and design area. [47] 59

40 WURXs block diagram, and two types of RF input signal. 63
41 Relation between false positive/negative rate and the comparator Vtrip (normalized to 0-1V) at the comparator output. 65
42 False wake-ups per hour at different Vtrip. 66
43 Missed detection rate at different Vtrip. 67
44 The sensitivity improvements with different code selections. 68
45 The sensitivity improvements with error tolerance. 69
46 Improvement of available code space with the correlator error tolerance algorithm. 69
47 The minimal RF transmission energy per wake-up and the sensitivity improvements at different RF turn-on time. 71
48 A 63-bit correlator with 4x oversampling of the comparator output and error tolerance. 71
49 Simulation versus measurement results on the number of false wake-ups in an hour. 72
50 Simulation versus measurement results on the MD rate. 73
51 A simple four layer DNN with an input layer, an output layer, and two hidden layers [60]. 76
52 The architecture of LeNet-5. Both convolutional layers and FC layer are used. [56] 77
53 The structure of a convolutional layer with a 3D input H·W·C, and a 3D output P·Q·K 78
54 The bottom-up design methodology of IMC circuit [56] 79
55 The top-down design methodology of IMC circuit [56] 79
56 (a) Weight value vs. the cell current of an memristor bitcell [67] (b) Memristor conductance vs. programming pulses [68] 81

57 Map a convolutional layer to memory 83
58 (a) The serial input activation IMC architecture with complementary memristors. The left column stores all the positive weights, and the right column stores all the negative weights. (b) BL and BLB discharge during analog computing. The partial dot product value equals to the voltage difference between the BL and BLB. 86
59 (a) The weight errors resulting from the random variation are added to the stored weight values. The memristor bitcell error follows an uniform distribution, and the SRAM bitcell error follows a normal distribution. (b) The BL and BLB discharging slopes are affected by the added errors. The partial dot product value includes an error item. 90
60 The ADC quantization error is introduced by rounding the partial dot products to the closest ADC readings. 91
61 IMC model with a dot product volume of 2048 and a WL chunk size of 4. (a) The histogram of the dot product (the standard deviation is 747 LSB). (b) The histogram of the dot product error (the standard deviation is 9.1 LSB), (c) the intuitive explanation of ENOB (6.1b) calculated from the two distributions 92
62 IMC model with a dot product volume of 2048 and a WL chunk size of 128. (a) The histogram of the dot product (the standard deviation is 736 LSB). (b) The histogram of the dot product error (the standard deviation is 297 LSB), (c) The intuitive explanation of ENOB (1b) calculated from the two distributions. 94
63 IMC ENOB model with varied WL chunk size and varied weight bit length 95

64 The injected error is sampled from a normal distribution decided by the SNR of the IMC micro-architecture 97
65 Histogram of the correlation coefficient between the dot products and the dot product errors. 97
66 MNIST experiment results of the FP32 pre-trained LeNet-5 with noise injection and quantization 98
67 MNIST experiment results of the LeNet-5 retrained with quantization errors 98
68 MNIST experiment results of the LeNet-5 retrained with quantization and noise injection 99
69 The ENOB of SRAM evaluated by the IMC error model by sweeping the WL chunk size 100

List of Tables

1 SRAM bitcell metrics interpreted from the user inputs 17

2 Initial bitcell sizes 19

3 Adjusted bitcell sizes for dynamic simulation 21

4 The most probable read disturbance failure point of Vth shifts 27

5 BERs of the auto-generated bitcell estimated by the proposed methods 28

6 BERs of the auto-generated bitcell calculated from $4.2 \cdot 10^5$ iterations of Monte Carlo simulation 28

7 Comparison of recently proposed on-chip oscillator designs [47] 60

8 Pros and Cons of input DAC and input bit serial 85

9 Iread mean and sigma of a SRAM bitcell at different voltages 95

10 Chips taped-out in each chapters 103

Contents

1 Introduction **1**

1.1 Motivation for ULP IC Components Design and Modeling 1

1.2 Thesis Statement 4

1.3 Goals 4

1.4 Dissertation Organization 5

2 ULP SRAM Design and Memory Design Exploration Tool Development **6**

2.1 Motivations 6

2.1.1 Motivation of the Subthreshold SRAM Bitcell Auto-Generation Flow 6

2.1.2 Motivation of the Memory Design Exploration Tool 7

2.2 Prior Arts 8

2.2.1 SRAM Static Metrics 9

2.2.2 SRAM Dynamic Metrics 11

2.2.3 SRAM Yield Analysis Modeling Methods 14

2.3 SRAM Bitcell Auto-Generation Flow 15

2.3.1 Technology Characterization and User Input Interpretation . . 17

2.3.2 BER Estimation with Normally Distributed Metrics 19

2.3.3 Rethinking the Importance Sampling 23

2.3.4 Importance Sampling for Yield Analysis 25

2.3.5 Summary of The Auto-Generated Bitcell 27

2.3.6 2KB SRAM Chip Design and Measurement Results 29

2.4 Multi-Port Register File Design Space Explorations 30

2.4.1 Comparison between ViPro and CACTI 31

2.4.2 New Features in ViPro 32
2.4.3 Design Explorations in the Maximum Data Throughput . . . 34
2.4.4 Design Explorations in the Minimum Energy Consumption . . 37
2.4.5 Trade-offs between the Energy and Delay 39
2.5 Conclusions 39
2.5.1 Roles of Circuit Models 39
2.5.2 Summary of The Bitcell Auto-Generation Flow 40
2.5.3 Summary of The Register File Design Exploration Tool 41

3 Temperature and Supply Variation Stable Clock Reference Design 43
3.1 Motivation & Prior Arts 43
3.2 Conventional Relaxation Oscillator Structures 45
3.2.1 Oscillator Structure with Two Comparators 45
3.2.2 Oscillator Structure with One Comparator 47
3.3 Temperature Stability Analysis 48
3.4 Supply Stability Analysis 51
3.5 Circuit Implementation for the Proposed Relaxation Oscillator 52
3.5.1 System Overview 52
3.5.2 PTAT Current Reference 53
3.5.3 Digital Frequency Compensation 55
3.5.4 Power-Gating for Rapid Duty Cycling 56
3.6 Chip Measurement Results 58
3.7 Conclusion 60

4 ULP Wake-Up Receiver (WURX) Base-band Design Exploration 62
4.1 Motivation 62
4.2 Background 64

4.3 WURX Baseband Design Explorations 65
4.3.1 Impact of Threshold Voltage on False Wake-up and Missed Detection Rate 67
4.3.2 Impact of the Wake-Up Code on the Sensitivity Improvements 67
4.3.3 Impact of the Correlator Error Tolerance on the Sensitivity Improvements 68
4.3.4 Trade-off between the RF Transmission Time and the Energy Consumption per Wake-Up 70
4.4 Wake Up Code Correlator Circuit Implementations 72
4.5 Measurement Results 72
4.6 Conclusion 73

5 Design Explorations of In-Memory Computing for Deep Neural Networks 75

5.1 Motivation 75
5.1.1 Introduction of Deep Neural Networks 75
5.1.2 Motivation of In-Memory Computing 78
5.1.3 Motivation of IMC Modeling and Top Down Design Methodology 79
5.2 Background & Prior Art 81
5.2.1 IMC with the Memristor 82
5.2.2 IMC with SRAMs 83
5.2.3 Other Types of Analog Computing 83
5.3 IMC Architecture 84
5.3.1 Discussion of the Input DAC 84
5.3.2 Memristor IMC Micro-Architecture 85
5.3.3 SRAM IMC Micro-Architecture 88

5.4 IMC Accuracy Loss Modeling 89
5.4.1 Bitcell Variations in IMC 89
5.4.2 ADC Quantization Error in IMC 90
5.4.3 Numerical Modeling for IMC 91
5.4.4 IMC Modeling with SRAM Bitcell Variations 95
5.4.5 Parameter Sweeping for IMC Modeling 95
5.5 IMC Experiments on Deep Neural Networks 96
5.5.1 Error Injection in DNN layers 96
5.5.2 MNIST Precision without DNN Retraining 97
5.5.3 DNN Retraining with Quantization Errors 99
5.5.4 DNN Retraining with Quantization Errors and Noise Injections 99
5.5.5 An SRAM Micro-Architecture for LeNet-5 100
5.6 Conclusions 101

6 Conclusion **103**

Appendices **107**

Appendix A List of Publications **107**

A.1 Publications 107
A.2 Pending Publications 107

Appendix B Glossary of Terms **109**

1 Introduction

1.1 Motivation for ULP IC Components Design and Modeling

The concept of the internet of things (IoT) is promising because it provides a new picture of everyday life in the future, with ubiquitous IoT nodes with smart sensing, data processing, and wireless connectivity to monitor the human body, the home, and the environment. However, due to the limited energy available from batteries and energy harvesters, the trend of sub-μW system-on-chip (SoC) design is compromising the performance and functionality for power reduction, for example, the embedded SRAM capacity is reduced to a few KB, and the clock frequency is scaled to 32 KHz [1]. For some simple applications like ECG signal monitoring and fall detection, the ULP SoC with the above specifications is sufficient, but more memory and processing performance is necessary for more complex tasks, for instance, object recognition applications. There is a trend towards designing IoT SoCs with more functionalities within the affordable power budget. Assuming the SoCs active power rises to 50 μW, it can be sustainable with a thermoelectric generator (TEG) or a piezo energy harvester [2], and it can also work for 1,000 hours with a 50 mAh button cell battery. The battery recharge cycle can be further prolonged by dynamic voltage and frequency scaling (DVFS), duty-cycling, and power gating techniques.

Circuit modeling plays an essential role in modern circuit design. The most commonly used circuit model is the SPICE model, which provides a comprehensive characterization of all the necessary components. The computation for large circuit simulations based on the SPICE model could be expensive because all the features are considered in the simulation, regardless of their impact on the critical metrics. More specific models based on a particular type of circuit can potentially be useful. For

developing a well-researched circuit, a model for speeding up the critical metric evaluation is helpful in reducing the design period. For exploring a new design field, an analytic model can guide the direction for improvement. For circuit blocks to be used in a more extensive system, a black box model can be used to predict the yield and performance of the system.

Static random access memories (SRAMs) are commonly used as on-chip memories for the IoT SoCs, and they contribute to a large proportion of the chip area and power consumption. The power gating technique does not usually apply to SRAM for data retention, and duty-cycling cannot deal with SRAMs high leakage power problem. To suppress both the active and leakage power, we aim to design SRAMs capable of operating at the subthreshold region with only the necessary assist techniques and other peripheral circuits. In the subthreshold region, SRAM faces serious read and write stability issues due to the random process variation, so bit error rate (BER) estimation is necessary for guaranteeing reliable operations. The BER obtained with the conventional Monte Carlo simulation usually requires millions of iterations, so a fast, accurate, and cost-efficient SRAM yield analysis model is in demand. Another reason for the long SRAM design period is the need to satisfy different user requirements, so we propose an SRAM bitcell auto-generation flow and an SRAM macro design space exploration tool to effectively reduce the engineering hours during the design period.

The near-zero power consumption wake-up receiver (WURX) is a potential game-changer in the field of IoT nodes communication because it has negligible power overhead and can eliminate the mW-level idle power of the primary receiver by waking up the main radio from the power gating mode [3]. The sensitivity of WURXs is defined by the minimal detectable radio frequency signal magnitude with a certain missed detection (MD) rate and false wake-up (FWU) rate. These two metrics affect

the WURX's reliability and SoC's unnecessary active power dissipation, which are both closely related to the WURX's baseband circuit design. We propose a wake-up code analysis model to deal with the challenges of finding appropriate comparator trip voltage, correlator code length, code weight, and error tolerance in the multi-dimensional design space.

To implement low cost IoT SoCs, on-chip RC relaxation oscillators (ROSC) are desired to replace the off-chip crystal oscillator (XO) as the stable clock references. The main challenge is to maintain the clock stability comparable to the XOs across the PVT corners. The power gating technique can suppress the active power of ROSCs, but it also deteriorates the clock stability because the added power gates in the oscillation loop introduce additional sources of variations. Two ROSC stability analysis models are employed to guide the direction for improving the temperature and supply sensitivity.

Deep neural networks (DNNs) have attracted a lot of attention due to their success in artificial intelligence (AI)-related field. In today's DNN computing systems based on the von Neumann architecture, the large volume of multiplication and accumulation (MAC) computations require a significant amount of data movements between the memory and processing units, resulting in high energy consumption and a degradation of throughput. The classical von Neumann computing architecture faces serious challenges regarding energy efficiency and chip area for enabling DNN in edge devices. In-memory computing (IMC) provides a new DNN computing architecture that potentially avoids the data movement issue, thus achieving energy-efficient MAC computations. However, the process variation of the on-chip memory bitcell and the noise of the mixed-signal calculation introduce DNN inference precision degradation. We propose an IMC accuracy loss model to guide the direction for choosing appropriate memory micro-architectures for certain DNN precision and predict the impact

of IMC accuracy loss on DNN performance.

1.2 Thesis Statement

To address the dilemma between the growing need for greater functionality and lower power consumption, a wide variety of cutting-edge ULP IC components, such as subthreshold embedded SRAMs, WURX, ULP and stable clock references, and DNN accelerators, are critical to enable low-cost and ULP IoT systems. The methodologies for designing reliable and ULP IC components are dramatically different from the traditional performance-driven IC designs because the behavior of CMOS devices is more sensitive to the process variation in the near or subthreshold region. We propose to use techniques, such as DVFS, duty-cycling, power gating, and several block-specified approaches to both lower the power consumption and maintain the functionality of IC components. Circuit modeling should be intensively utilized in making design decisions and guaranteeing reliable operations for all the IC components.

1.3 Goals

The major research goals of this dissertation include:

- Build a fast, accurate, and cost-efficient SRAM yield analysis model to guarantee the robustness of ULP subthreshold SRAM designs.
- Develop an SRAM bitcell auto-generate flow and an SRAM macro design exploration tool to make SRAM design decisions for a 0.5V 2 KB SRAM testchip in the 65 nm technology.
- Build the temperature and supply stability analysis models to guide on-chip ROSC designs.

- Design an on-chip ROSC with a stability comparable to XO-based clock references in the 65 nm technology.
- Model the impact of the correlator wake-up code and the comparator trip voltage on the sensitivity of WURXs.
- Design the WURX's baseband circuit in the 130 nm technology with appropriate wake-up code length, code weight, and error tolerance for 0.1% MD rate and less than one FWU per hour.
- Build an accuracy loss model for IMC micro-architectures and predict the DNN precision using the model.
- Design an IMC micro-architecture capable of recognizing hand-written digits with more than 97.5% precision.

1.4 Dissertation Organization

Chapter 2 demonstrates the SRAM bitcell auto-generation flow with the yield analysis model and the SRAM macro design space exploration tool (ViPro). Measurement results of a 2KB SRAM testchip designed using the proposed tools are also included in this section. Chapter 3 demonstrates the methodologies for designing good temperature and supply voltage stable ROSC as well as the measurement results of an ROSC taped-out in the 65 nm technology. Chapter 4 introduces the wake-up code analysis model for improving the sensitivity of WURXs. Measurement results of a WURX tape-out in the 130nm technology are presented at the end of the chapter. Chapter 5 demonstrates implementations of the IMC accuracy loss model and the impact of IMC accuracy loss on the DNN performance. Chapter 6 summarizes the dissertation.

2 ULP SRAM Design and Memory Design Exploration Tool Development

2.1 Motivations

Customized-memory macro designs are usually necessary and challenging for ultra-low power (ULP) internet of things (IoT) applications because the design decisions should be made based on specific process technologies and application scenarios. The long design period results from the complex design decisions on the multi-dimensional knob space, so tools for efficient memory design can enable the deployment of more IoT nodes by reducing the development cost and speeding up the time to market.

2.1.1 Motivation of the Subthreshold SRAM Bitcell Auto-Generation Flow

In the recently published IoT platforms, static random access memory (SRAM) is still the most commonly used instruction memory and data memory [1] [4] [5]. The demand for extended capabilities of the IoT nodes requires larger-capacity SRAMs, which easily take more than 30% of the total SoC chip area [5] [6]. SRAMs become one of the major contributors to the static power dissipation of the power-hungry IoT system-on-chip (SoCs) [4] [5]. As a part of the system, the SRAM should adapt to the system clock frequency and operating supply voltage (V_{DD}). In the circumstance of data transmission, the SoC and the radio module operate at a high frequency for fast communication, and the V_{DD} should also be high enough for high-speed operations. When the SoC enters the idle mode, the digital logic is usually switched to the subthreshold region for suppressing the leakage power. The SRAM minimum operating voltage (VMIN) is limited by the yield, so it cannot be as low as the digital

circuit VMIN [6]. As a consequence, both the digital circuit and SRAM have to work at the higher SRAM V_{DD}, which sacrifices part of the leakage reduction benefit.

Subthreshold SRAM designs are desired for achieving sub-micro watt operation in the power-hungry IoT SoCs [1]. The subthreshold design is challenging because the transistor turn-on current (Ion) is exponentially related to the transistor threshold voltage (Vth), and the random process variation of Vth can have a substantial adverse impact on the performance of SRAM [7]. As a result, SRAMs designed for superthreshold operations might not work in the subthreshold region without various assist techniques [8] [9]. The transistor size is regarded as a less effective knob in the subthreshold region due to the linear relationship between the Ion and the transistors channel width (W) and channel length (L). However, we find that resizing transistors can still significantly influence the SRAM yield because the Vth random variation is proportional to $WL^{-0.5}$. Also, the SRAM bitcell-size knob is also closely related to other important SRAM metrics, for example, the leakage power and the area. To design a subthreshold SRAM that satisfies user requirements for the capacity, the yield, the leakage power, the energy per operation, and the area, it is complicated to make decisions on all the available SRAM knobs, such as the bitcell type, the bitcell size, the assist techniques, and the micro-architecture. A tool that automatically generates the SRAM bitcell and selects the appropriate assist techniques based on a target specification can significantly improve the efficiency of the multi-dimensional design space exploration.

2.1.2 Motivation of the Memory Design Exploration Tool

A lot of emerging process technologies aiming at the IoT market enable low power circuit design by providing energy efficient devices, and the design effort spent on new technology is substantial. The device variability, leakage, and interconnect de-

lay make memory design even more labor-demanding than digital circuits. Another adverse fact is some of the useful assist techniques might not be effective anymore in different process technology. Previous work about the technology agnostic SRAM virtual prototyping tool can facilitate memory designers to make the design choice efficiently [10]. A Virtual Prototyping (ViPro) tool has been developed for memory systems, which enables early design space exploration by creating virtual prototypes of a complete 6T SRAM macro. Register files are multi-port SRAM bitcells capable of simultaneous read and write operations, and they are preferred in applications need high data throughput. One deficiency of the base version of ViPro is not supporting different memory bitcell types, which significantly limits ViPro for broader usages. In addition, ViPro only facilitates design space exploration, but it is not capable as a memory compiler. ViPro will be more useful if it can automatically make design decisions about the memory bitcell and assist techniques for satisfying different system requirements.

2.2 Prior Arts

SRAMs have attracted a lot of attention during the past 20 years, and previous research on the SRAM acts as the concrete foundation of appealing design automation. In addition, rapidly upgrading computing cluster speed and multi-threading technology also empower the automatic design and exploration of SRAM.

In this section, various metrics are presented as candidates for the SRAM yield analysis. The Monte Carlo simulation results quantitatively demonstrate deficiencies of the static metrics compared to transient simulation results of the SRAM read and write stability. As a result, dynamic metrics should be used to predict the SRAM yield in fulfilling the SRAM design space exploration, although it increases the simulation time.

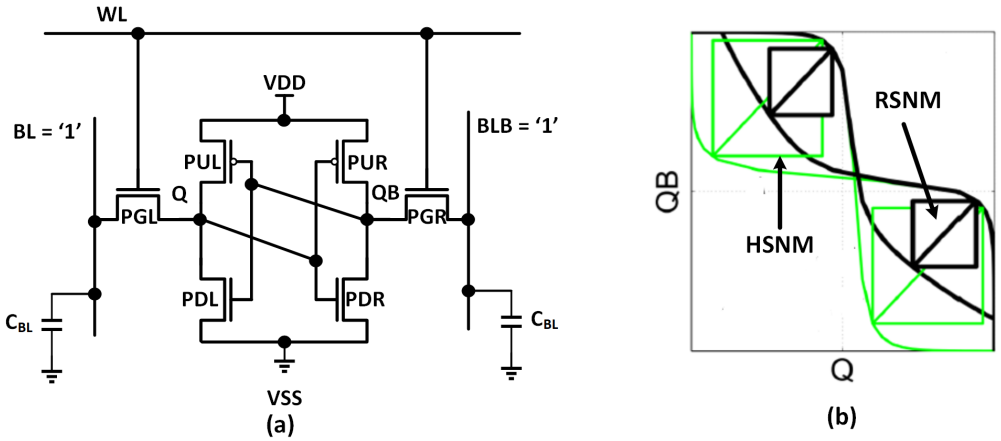


Figure 1: (a) A 6T SRAM bitcell schematic diagram (b) SRAM butter fly curves, and SNM for read and hold

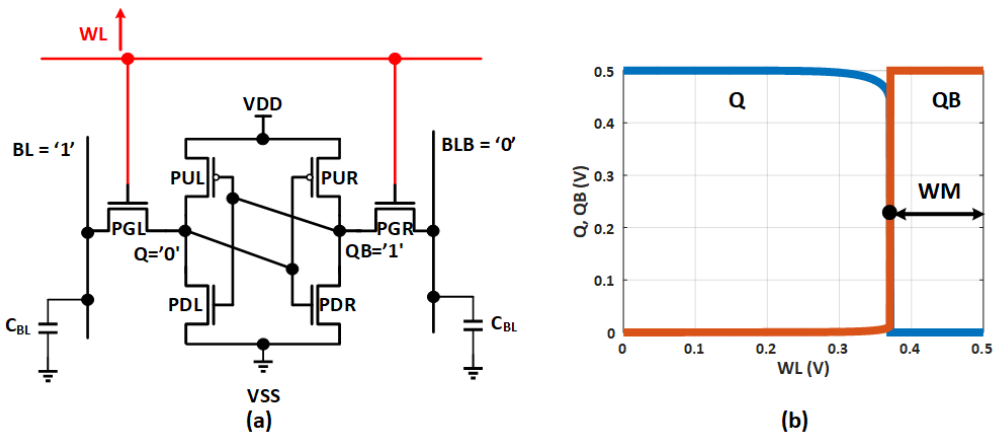


Figure 2: (a) A 6T SRAM bitcell schematic during WL sweeping (b) SRAM WM defined by the difference between the WL voltage and V_{DD} at the crossing point of Q and QB.

2.2.1 SRAM Static Metrics

The static noise margin (SNM) is introduced by [11] to represent the read and hold stability of an SRAM bitcell. The definition of SNM is the maximum window in the two voltage transfer curves (VTC) as demonstrated in Figure. 1. In an SRAM hold status, the WL is driven to ground, and the bitline (BL) voltage does not affect the two VTCs. During a SRAM read operation, the wordline (WL) voltage changes

to V_{DD}, and the high voltage on BLs increase the voltage of the internal nodes, Q and QB. The fact that the read SNM (RSNM) is smaller than the hold SNM (HSNM) can be intuitively demonstrated by the butterfly curve in Figure 1 (b), so the SRAM bitcell is more susceptible to be disturbed during the read operation. The remainder of the discussion focuses on the RSNM because it is the bottleneck in SRAM stability.

There are multiple methods to define the static write margin (WM), and Figure 2 illustrates the one determined by the difference between the WL voltage and V_{DD} at the crossing point of Q and QB. During a write operation, BL is driven to a low voltage, and BLB is driven to a high voltage. Internal nodes Q and QB will eventually flip during as the WL voltage increases. If the flip occurs when WL is lower than or equal to the V_{DD}, the SRAM bitcell is believed to be write-able; otherwise, the WM value is negative, and the write operation fails. This WM definition represents dynamic write-ability very well because it closely mimics an actual dynamic write operation. The research in [12] also indicates a similar observation.

The two static metrics RSNM and WM can reflect the impact of random process variations on the SRAM read and write stability. Due to the randomness of variation,

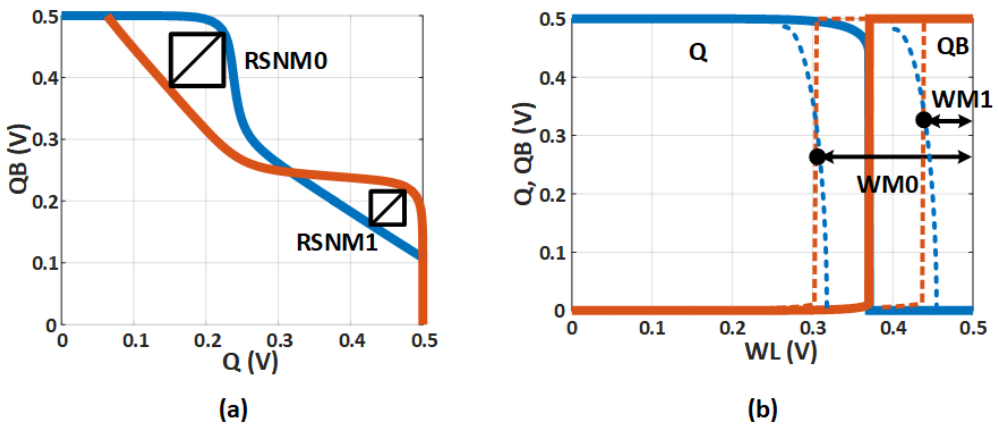


Figure 3: (a) RSNM under the impact of random process variations. RSNM is the minimum between RSNM0 and RSNM1. (b) WM under the impact of random process variations. WM is the minimum between WM0 and WM1.

the six transistors of an SRAM bitcell can deviate in different directions, so that RSNM windows of the butterfly curve are no longer symmetric as illustrated in Figure. 3 (a). The RSNM value should be the smaller one between RSNM0 and RSNM1, which reveals the significant read stability degradation that results from the random variation [13]. Similarly, the WM value is the smaller one between WM0 and WM1.

2.2.2 SRAM Dynamic Metrics

The static metrics have some limitations. For example, static metrics cannot reveal the timing information during transient read and write operations, and they

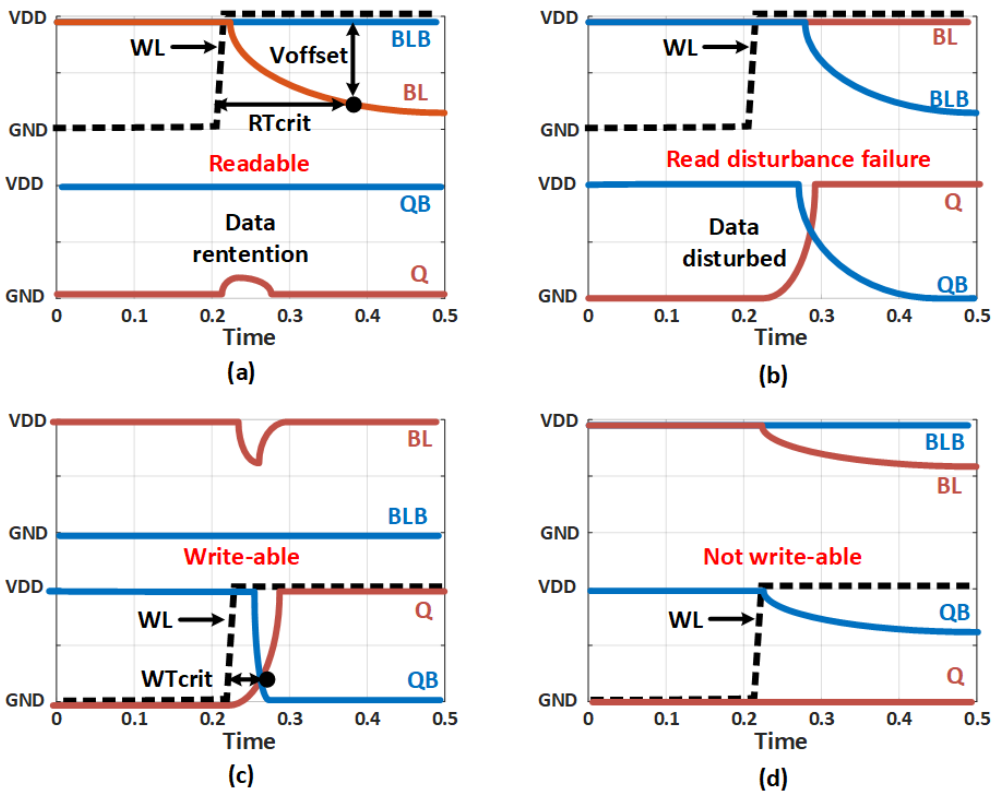


Figure 4: (a) Readable bitcell timing waveform, and read critical time (RTcrit) defined by the offset voltage between BL and BLB. (b) Read failure waveform due to read data disturbance. (c) Write-able bitcell timing waveform, and write critical time (WRcrit) defined by the crossing point of Q and QB. (d) Write failure waveform due to non-write-able bitcell.

cannot reflect the impact of parasitics on the read and write stability. Figure. 4 introduces a few dynamic metrics for SRAM characterization.

- Read critical time (RTcrit). The RTcrit metric represents the timing requirement of a successful read operation. It is defined by the interval between the time of WL rising and the time of an offset voltage developing between the BL and BLB. If the SRAM uses a differential sense amplifier (SA) during the read operation, the offset voltage is decided by the minimum voltage that can be amplified by the SA. If an SA is not included, the offset voltage is decided by the trip voltage of the BL buffer.

- Read stability and half-select (HS) stability. The read stability stands for the probability of data disturbance of the selected bitcell during a read operation. The HS stability defines the probability of data disturbance in bitcells sharing the same WL with the selected bitcell. The HS stability is usually better than the read stability because the parasitic capacitance is different in the two cases. It should be mentioned that the static metric RSNM cannot differentiate between the two types of stability.

- Readability. This is the overall probability of a successful read operation, and it considers the read failures resulting from both the timing violation and the read data disturbance.

- Write critical time (WTcrit). The WTcrit is a dynamic metric representing the dynamic timing requirement for a successful write operation.

- Write-ability. This is the overall probability of a successful write operation, and it considers the write failures resulting from both the timing violation and the non-write-able bitcells.

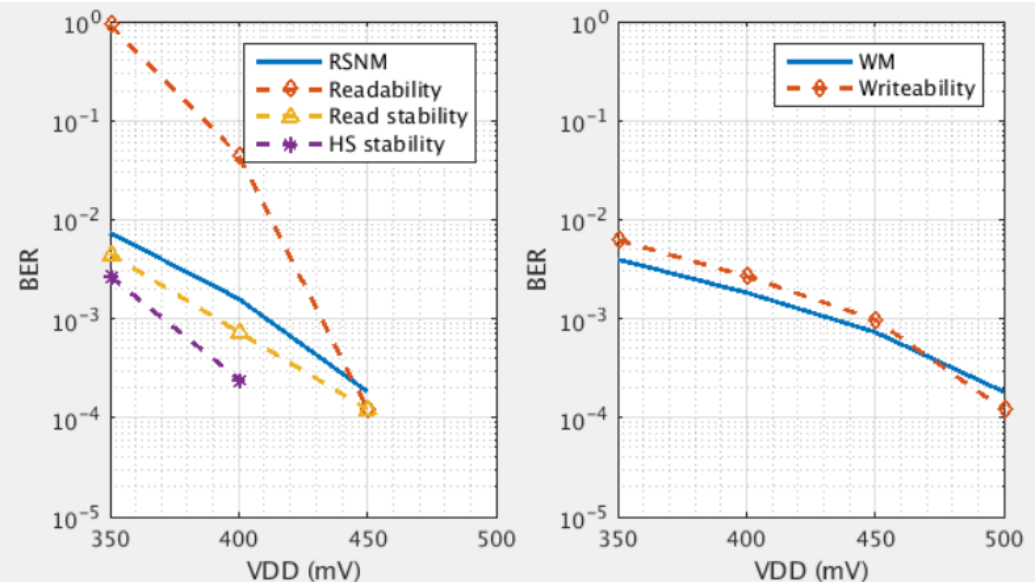


Figure 5: 16K points Monte Carlo simulations of (a) the BERs calculated from the RSNM, the transient readability, the transient read stability, and the transient half-select (HS) stability, and (b) the BERs calculated from the WM and the transient write-ability. The clock period is 15 μs for the transient simulations.

These dynamic metrics are useful for understanding different failure mechanisms and for predicting the SRAM's bit error rate (BER). Studies on the random process variations indicate that they follow the Gaussian distribution [7] [13]. To quantitatively explore to what extent the dopant fluctuation of transistors degrades the SRAM read and write stability, Monte Carlo simulation [14] can be employed to mimic the random threshold variation with a large number of iterations. Readability and write-ability are overall failures during dynamic SRAM read and write access, so they can be regarded as the golden reference of BER. Figure. 5 presents the Monte Carlo simulation results of 16,000 iterations. It illustrates that the BER calculated from RSNM is higher than that from the dynamic read stability, and that the HS stability BER is lower than the read stability BER because the BL parasitic capacitance seen by an HS bitcell is smaller than that seen by a selected bitcell. Below 400mV, the overall readability BERs are 100x larger than the RSNM BERs because the RTcrit

metric fails to meet the timing requirement of less than 15 μS. Figure. 5(b) illustrates that the WM BER is slightly smaller than the write-ability BER in the lower voltage range because of the timing violation, but they are very close to each other.

2.2.3 SRAM Yield Analysis Modeling Methods

To avoid stability problems in the subthreshold SRAMs, yield analysis should be employed to make design decisions on the bitcell sizing and assist techniques. For a small-capacity 2KB SRAM, the BER should be less than $6 \cdot 10^{-6}$ to guarantee 90% of yield, and $1.6 \cdot 10^{7}$ iterations of Monte Carlo simulation are required to obtain 90% confidence and 90% accuracy of the BER [15]. Static metrics have a low simulation time cost, but they cannot represent the SRAM failures due to timing violations. Dynamic metrics consider the timing violations, but the simulation time can be 10 times longer compared to the static simulations because of the extra circuits for simulating the real working condition of the SRAM.

Researchers in [16] employ RTcrit and WTcrit to predict the SRAM yield and VMIN, and they use sensitivity analysis to improve the simulation speed by 112x compared to the recursive statistical blockade [17] with only 3% average loss in accuracy. However, the RTcrit is not able to catch the readability failures resulting from read data disturbance, and the WTcrit also fails to predict the BER of non-write-able bitcells. Importance sampling [18] [19] offers the capability of directly using dynamic readability and write-ability as the metrics for calculating BER, but it requires a complicated algorithm to search for the most probable failure point (MPFP), which still requires nearly 10^4 simulations. The simulation speedup is encouraging for evaluating the BER of one SRAM operating condition, but it is not enough for our purpose of SRAM bitcell auto-generation and design space exploration.

2.3 SRAM Bitcell Auto-Generation Flow

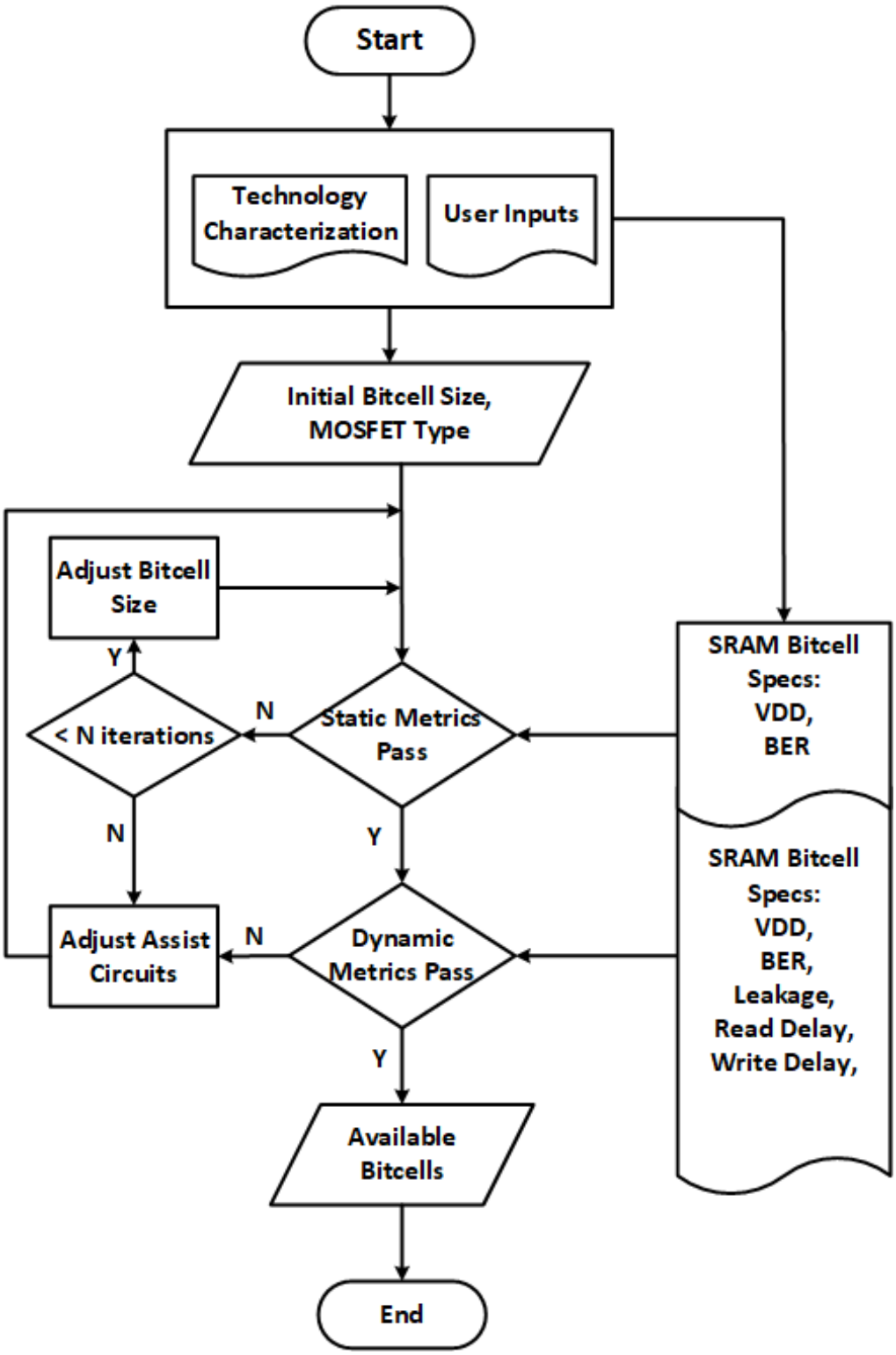


Figure 6: SRAM bitcell auto-generation flow

The goal of the SRAM bitcell auto-generation flow is to decide the bitcell level design knobs based on the user requirements. The user is expected to regard the SRAM as a black box and to provide only high-level design specifications. The flow should follow a predefined logic and use simulation results to make design decisions without human intervention. With the assistance of the auto-generation flow, the human engineer hours can be replaced by the machine computing time to obtain a pool of available SRAM bitcells. The auto-generated bitcells act as the input of the design space exploration tools like ViPro [20] [21], which makes design decisions on the SRAM-macro level knobs. The SRAM design knobs are provided in the below bullets.

- Bitcell sizing. SRAM comprises six different devices: two pull-down (PD) devices, two pull-up (PU) devices, and two pass-gate (PG) devices. The ratio of PD/PG determines the read reliability, and the ratio of PU/PG determines the write reliability. The three devices can be intentionally re-sized for reliable read or write operations. When auto-generated, the bitcell sizes are adjusted in the first N (i.e., 10) iterations.

- Peripheral assist techniques. Peripheral assist techniques are used to ensure reliable operation. Popular read assist techniques include V_{DD} boosting, V_{SS} lowering, and WL under-drive. Write assist techniques include V_{DD} lowering, Negative BL, and WL boosting. The read or write assist techniques are chosen based on the SRAM failure reason.

- Bitcell type. The most commonly used bitcell types are 6T with a shared read and write port and 8T with another single-ended decoupled read port to manage the half-select issue during a read operation. In the auto-generation flow, the default bitcell type is 6T because of less leakage current and area. The 8T

bitcell is considered only when the read data disturbance BER fails to meet the target, while the HS data disturbance BER happens to meet the target.

- Memory micro-architecture. With a given memory capacity, design knobs such as the number of rows, columns, and banks can affect the delay and energy of an SRAM macro. The array structure knobs impact the read and write delay, so they are used during the dynamic metric simulations.

Spec	Description	Example
V_{DD}	The bitcell V_{DD} is the same with the SRAM macro V_{DD}	0.5V
BER	The BER is calculated from the capacity and yield with equation BER=(1-yield)/capacity	$6 \cdot 10^{-6}$
P_{leak}	Assume the total bitcell leakage power accounts for a half of the maximum SRAM macro leakage power	2 pW
P_{act}	Assume the total bitcell active power accounts for a half of the maximum SRAM macro active power	200 nW
T_{dly}	Assume half of the clock period is for read and write operation, and the other half is for pre-charging	7.5 μS

Table 1: SRAM bitcell metrics interpreted from the user inputs

2.3.1 Technology Characterization and User Input Interpretation

Figure 6 demonstrates the SRAM bitcell auto-generation flowchart. The user inputs are anticipated to be SRAM Macro level specifications, for example, the capacity, the yield, the operating V_{DD}, the maximum leakage power, the maximum active power, and the clock period. Since the flow is developed for the SRAM bitcell auto-generation, the SRAM macro level specifications should be interpreted as bitcell level metrics as illustrated in Table 1. For example, if an user requires a 90% yield 2 KB SRAM with 64 nW leakage power and 400 nW active power at 0.5 V, the SRAM bitcell specification should be $6 \cdot 10^{-6}$ BER, 2 pW leakage power per bit, 200 nW

total active power, and 7.5 μS operation delay. All these specifications are used as the passing or failing criteria in the static or dynamic simulations.

The technology characterization is implemented by sweeping the on and off current of the available device types like HVT, RVT, and LVT, across different channel width

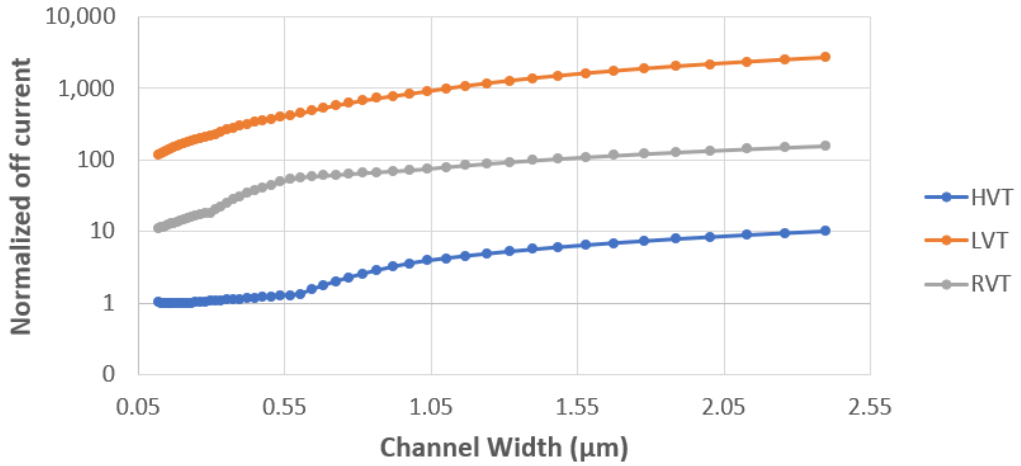


Figure 7: Normalized NMOS turned-off current of HVT, RVT, and LVT type vs. channel width at 0.5V

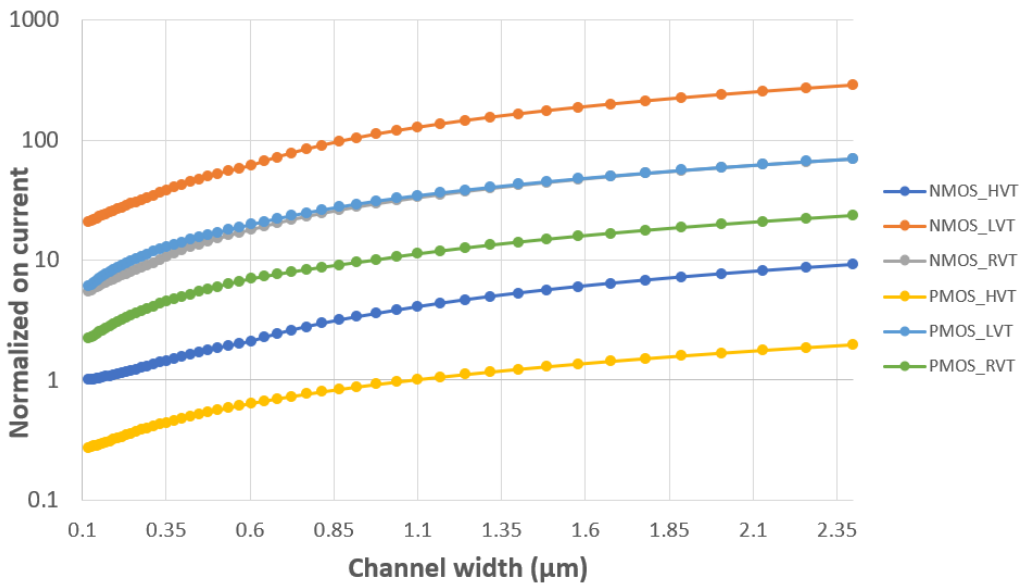


Figure 8: Normalized NMOS and PMOS turned-on current of HVT, RVT, and LVT type vs. channel width at 0.5V

and length. Figure 7 reveals the leakage current of LVT device is about 100 times more than the leakage of HVT device, so the HVT device is the first choice in building ULP SRAM bitcells. The RVT and LVT devices can also be an option if the flow could not find a valid HVT SRAM bitcell that meets the bitcell specifications defined in Table 1. The initial sizes of PD, PG and PU are calculated based on the ratio of 2:1.5:1, as provided in Table 2.

	PD	PG	PU
Width (nm)	200	120	120
Length (nm)	60	70	60

Table 2: Initial bitcell sizes

2.3.2 BER Estimation with Normally Distributed Metrics

According to the flow in Figure 6, static simulations are still used due to the quick simulation speed compared to dynamic metric simulations. The research in [13] found that the two RSNMs in the butterfly curve are identically distributed random variables that follow the same normal distribution, and the BER of SRAM can be

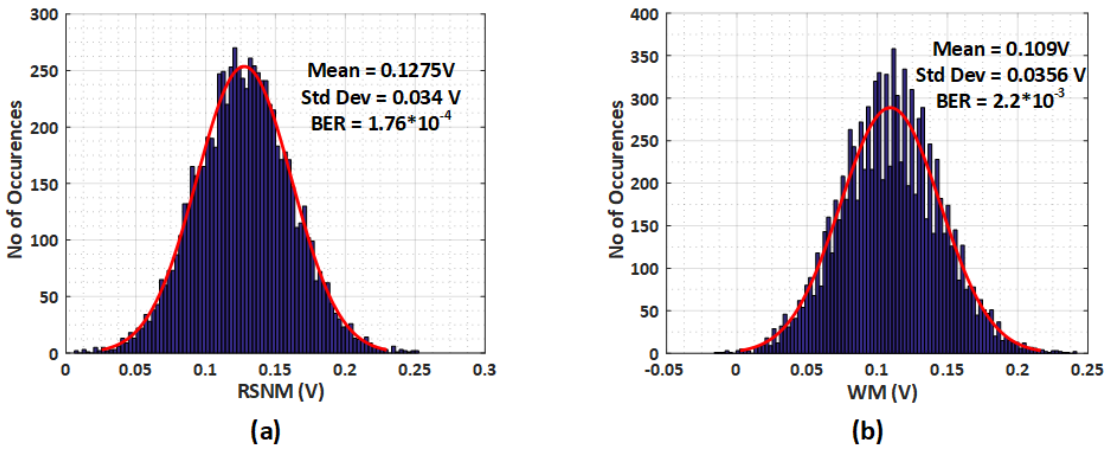


Figure 9: Monte carlo simulation results on the initial bitcell size. (a) Normally distributed RSNM; (b) Normally distributed WM

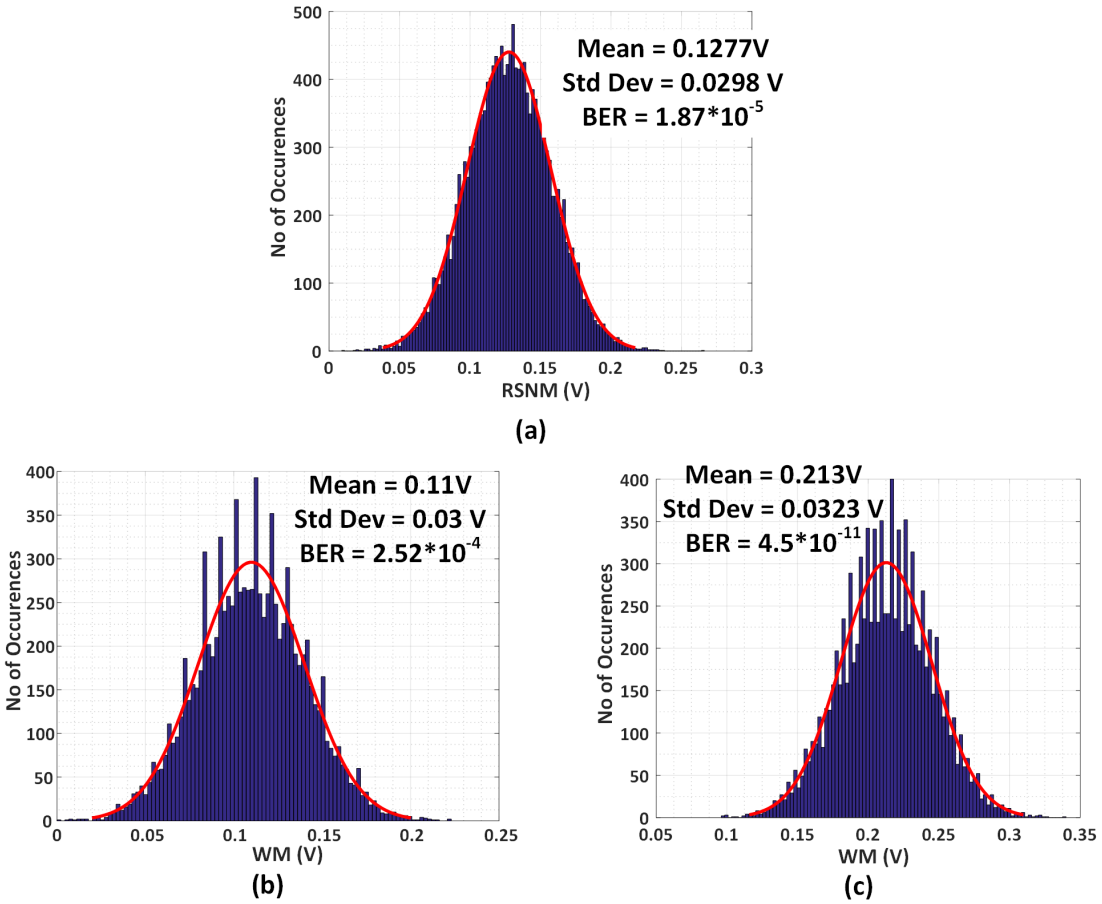


Figure 10: Monte carlo simulation results on the adjusted bitcell size. (a) Normally distributed RSNM; (b) Normally distributed WM; (c) Normally distributed WM with 20% of negative BL assist

estimated by the Cumulative Distribution Function (CDF). Figure 9 (a) demonstrates one of the RSNM distribution of the initial-size-bitcell at 0.5 V and the estimated BER of $1.76 \cdot 10^{-4}$. Figure 9 (b) illustrates the WM of initial-size-bitcell, which also follows a normal distribution. Similarly, the estimated BER is $2.2 \cdot 10^{-3}$.

Since the initial-size-bitcell cannot meet the BER requirement of less than $6 \cdot 10^{-6}$, the auto-generation flow will adjust the bitcell sizes in the first ten iterations. Based on knowledge about the process variation, increasing the channel width of transistors can effectively reduce the Vth random variation and produce more compact distribu-

tions of RSNM and WM. In the experiment, larger PD and PG widths are helpful in improving the RSNM, and larger PG and PU widths are useful in enhancing the WM. After 10 iterations of bitcell-size adjustment, SRAM assist techniques are explored to further lower the BER if necessary. For subthreshold SRAMs, V_{DD} boosting and negative BL are proved to be very effective in improving the read and write stability. Figure 10 (a) illustrates the RSNM distribution with updated bitcell sizes in Table 3. The BER is reduced to $1.87 \cdot 10^{-5}$ due to the 10% smaller standard deviation. No read assist is employed because the RSNM BER is usually a pessimistic estimation of the read disturbance compared to the read and HS BER, as presented in Figure 5. Figure 10 (b) and (c) illustrate that the WM BER of the adjusted bitcell is $2.52 \cdot 10^{-4}$, and the BER is drastically reduced to $4.5 \cdot 10^{-11}$ with 20 % of negative BL assist.

	PD	PG	PU
Width (nm)	230	180	150
Length (nm)	60	70	60

Table 3: Adjusted bitcell sizes for dynamic simulation

The updated bitcell sizes in Table 3 pass the initial screening for the BER of the static metric WM. Dynamic metrics like the RTcrit and WTcrit are not normally distributed, so we cannot use the CDF to estimate their BERs, as plotted in Figure 11 (a) and (c). Fortunately, Figure 11 (b) and (d) demonstrate the inverse of RTcrit and WTcrit both follow normal distributions, which can be explained in [22]. To meet the requirement of less than 7.5 μS delay, the worst case inverse delay should be larger than $1.3 \cdot 10^5 Hz$. Following the same approach used for calculating BERs of the RSNM and the WM, BERs of the RTcrit and the WTcrit are $2.4 \cdot 10^{-6}$ and $2.9 \cdot 10^{-10}$. It should be mentioned that, the write delay in Figure 11 (c) is simulated without the negative BL as a write assist technique, and the non-write-able bitcell BER counted from the dynamic simulation result is $1.2 \cdot 10^{-4}$, which is very close to

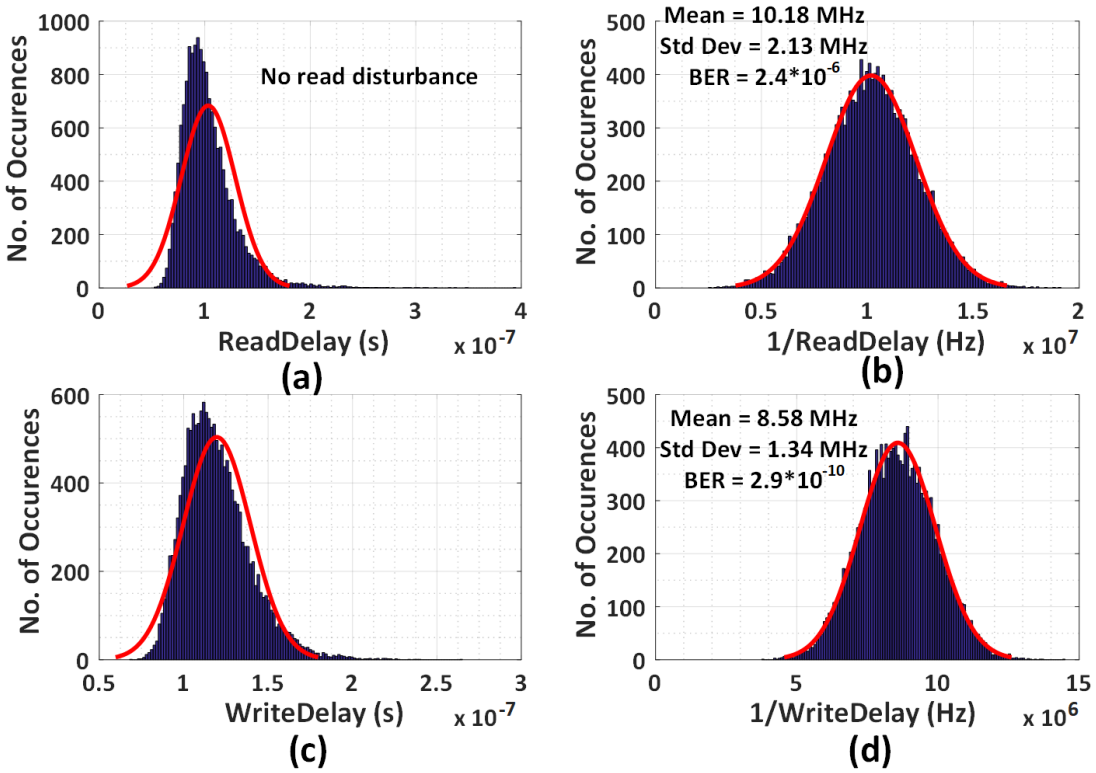


Figure 11: Monte carlo simulation results on the (a) RTcrit, (b) 1/RTcrit, (c) WTcrit, and (d) 1/WTcrit. The RTcrit and WTcrit distributions have a long tail on the larger delay side. The 1/RTcrit and 1/WTcrit distributions are normal.

the WM BER in Figure 10 (b). In addition, no read disturbance failure is observed in the 16,000 iterations of Monte Carlo simulations.

The overall SRAM readability BER consists of two parts, the data disturbance BER and the timing violation BER. The overall SRAM write-ability BER also has two sources, the BER of non-write-able bitcells and the BER resulting from write timing violations. The BER analysis of the above normally distributed metrics can be summarized in the below bullets.

- The RSNM BER is a pessimistic estimation of the actual read data disturbance, and the dynamic read disturbance metric is useful to get an accurate BER estimation.

- The inverse RTcrit and WTcrit both fit normal distributions, so their CDF can be used to calculate the BERs resulting from read and write timing failures.

- The static WM BER agrees with the BER of non-write-able bitcells counted from the 16,000-point dynamic simulation, so it is not necessary to use the more expensive dynamic metrics for estimating the BER of non-write-able bitcells.

2.3.3 Rethinking the Importance Sampling

Importance sampling is proposed to speed up the simulation time for obtaining a rarely happening event by shifting the sampling region of parameters to a place where the event is not rare, which is called the most probable failure point (MPFP). Previous works employing the importance sampling consider multiple parameters during searching the MPFP [15] [18] [19], even not all the parameters are strongly

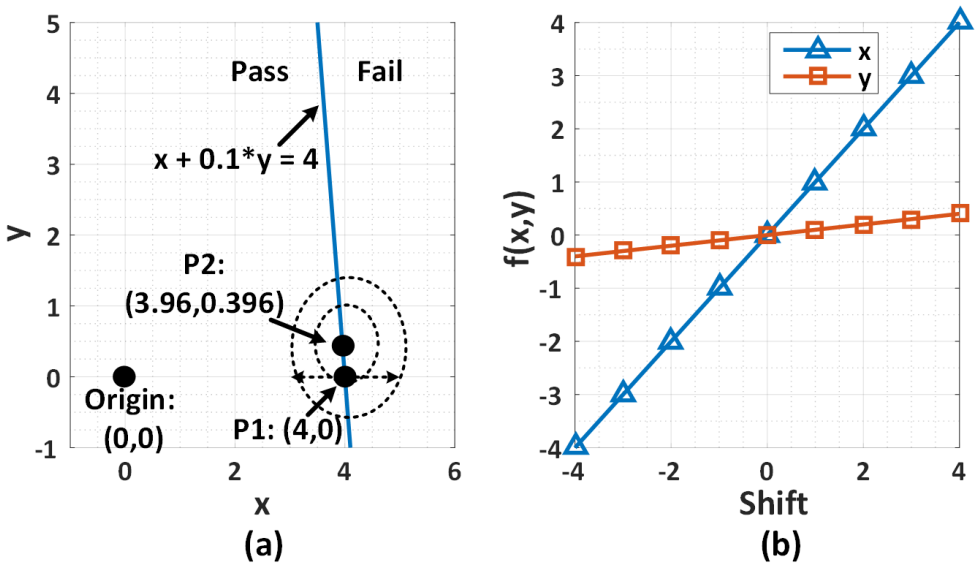


Figure 12: (a) Illustration of the two-dimension parameter space and the boundary between the pass region and failure region. P1 is the MPFP by only considering the dominating parameter x. P2 is the MPFP by considering both parameters x and y. (b) The sensitivity analysis of $f(x, y)$ on x and y.

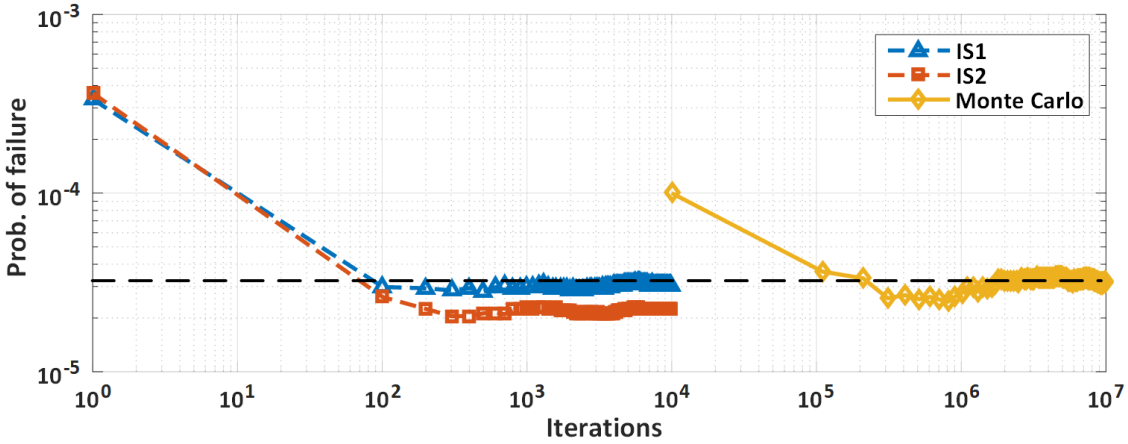


Figure 13: Evolutions of the failure probability using one-dimensional importance sampling (IS1) based on P1, two-dimensional importance sampling (IS2) based on P2, and Monte Carlo simulations.

related to the target event. One thought that could simplify the time-consuming searching procedure is to reduce the number of parameters to be searched.

To justify the above thought, we assume a simple function with two parameters $f(x, y) = x + 0.1 * y$, where both x and y follow the standard normal distribution. The goal is to find the failure probability of satisfying $f(x, y) <= 4$. Figure 12 (a) illustrates the pass and failure regions separated by a line defined by an equation $x + 0.1 * y = 4$. Definition of the MPFP is the failure point closest to the origin, so its coordinates are (3.96,0.396). Figure 12 (b) presents the sensitivity analysis of the parameters x and y, and it is obvious that x dominates the effect on $f(x, y)$. If we only consider the dominating parameter x, the MPFP will be P1; if we consider both x and y, the MPFP is P2, as illustrated in Figure 12 (a). After locating the two MPFPs, the importance sampling is implemented in the x dimension around P1 and in two dimensions around P2.

Figure 13 demonstrates evolutions of the failure probability obtained by the one-dimensional importance sampling (IS1), the two-dimensional importance sampling

(IS2), and the Monte Carlo simulation. The convergence value of the IS1 failure probability is closer to that of the Monte Carlo estimated value compared to the IS2, which means importance sampling on the dominating parameter is a better estimator of the actual failure probability. The convergence value of IS2 tends to be an estimator biased to a smaller failure probability because the deviation of parameter y increases the distance to the origin but not effectively changes the value of $f(x, y)$, so the impact of the sampled failure points is diluted.

The takeaway point is that importance sampling on the dominating parameters of a failure event not only reduces the searching time for the MPFP but also improves the accuracy of the estimated failure probability.

2.3.4 Importance Sampling for Yield Analysis

To prove that the read disturbance BER is less than the target BER of $6 \cdot 10^{-6}$, at least $1.6 \cdot 10^{7}$ Monte Carlo sample points should be simulated for 90% accuracy and 90% confidence [15], which takes roughly 4,500 hours with the dynamic simulation

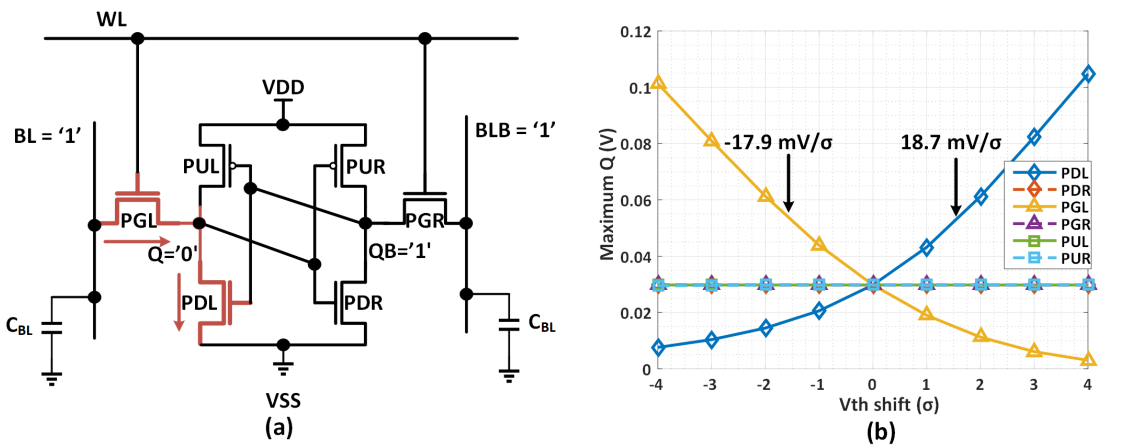


Figure 14: (a) Internal node Q is disturbed during the read operation. (b) Sensitivity analysis on the maximum voltage of Q with simulations.The maximum voltage of Q is sensitive to Vth variations in PDL and PGL.

setup using 30 threads of Intel Xeon Gold 6150 CPU. Importance sampling can speed up the simulation time by finding the MPFP, which works well for metrics that have a continuous change of value during the spherical searching for the MPFP. However, the dynamic read disturbance is a binary metric giving either a pass or a fail. This problem can be partially solved by running a small 100-point Monte Carlo simulation during the spherical searching, but the overhead is a 100x longer simulation time [19].

The read disturbance failure can be studied with two parameters. The first parameter is the maximum voltage of node Q, as illustrated in Figure 14 (a). It represents the severity of disturbance on the internal node during a read access. The other parameter is the trip voltage (Vtrip) of the right inverter of an SRAM bitcell, which can be defined by the point that Q equals QB in Figure 15. Vtrip reveals if the disturbance on Q can flip the other node QB and cause a read data disturbance. Sensitivity analysis results presented in Figure 14 (b) illustrate that both a positive PDL Vth shift and a negative PGL Vth shift increase the maximum Q voltage, while it is not sensitive to the Vth shift of the other transistors.

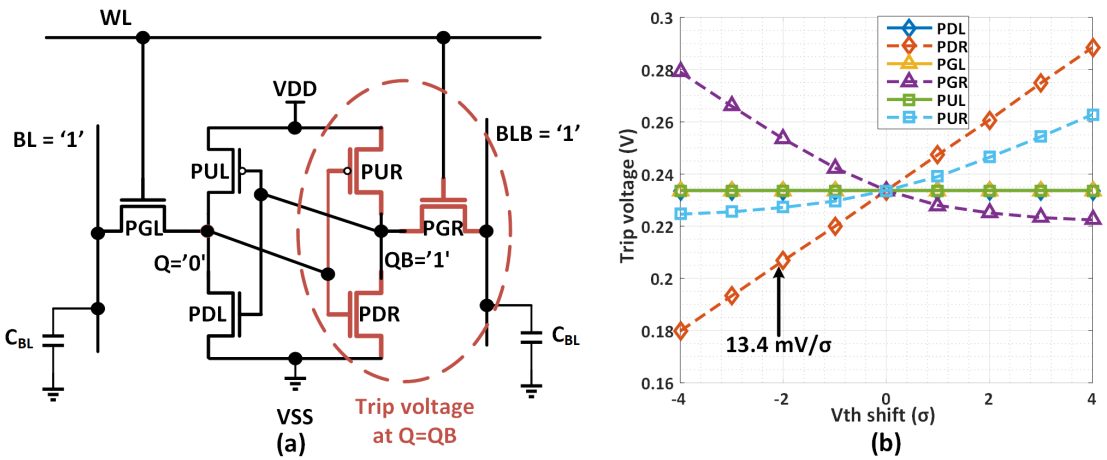


Figure 15: (a) A read disturbance is related to the trip voltage of the right inverter in the bitcell. (b) Sensitivity analysis on the trip voltage with static simulations shows it is sensitive to Vth variations in PDR.

Figure 15 (b) demonstrates that a negative PDR Vth shift can decrease Vtrip. We are only interested in the Vth shifts that increase the Q voltage and lower the Vtrip, where a read disturbance is more likely to occur. Because the two parameters are continuous value, their sensitivities on the Vth shift can also be used in finding the MPFP. During the spherical searching, only PDL, PGL, and PDR are considered because the impact of the other three devices is small. Table 4 illustrates that the MPFP is within 5 (σ) of distance from the origin, and a 1,500-point Monte Carlo simulation with the shifted Vth demonstrates that the BER of the importance sample is 0.72. Following the approach defined in [17], the estimated BER of read disturbance is about $2.1 \cdot 10^{-6}$.

The number of simulations to find the MPFP using the sensitivity analysis is 54, and the importance sampling requires 1,500 iterations. The total simulation time for estimating the BER of read disturbance is about 30 minutes, which is more than 10000 times faster than the conventional Monte Carlo simulation.

PDL shift (σ)	PGL shift (σ)	PDR shift (σ)	Distance (σ)	IS BER	Est. read dist. BER
3.2137	-3.0643	-2.2983	5	0.72	$2.1 \cdot 10^{-6}$

Table 4: The most probable read disturbance failure point of Vth shifts

2.3.5 Summary of The Auto-Generated Bitcell

Table 5 summarizes the BER results of read delay timing failure, read data disturbance failure, write delay timing failure, and non-write-able bitcell failure, using the adjusted bitcell sizes in Table 3, with and without 20 % of negative bitline as a write assist technique. The worst-case BER is less than $4.6 \cdot 10^{-6}$, assuming that different types of failures are happening in different SRAM bits, which still meets the required BER of $6 \cdot 10^{-6}$. The total simulation time for estimating all four BERs is

1.5 hours using 30 threads of Intel Xeon Gold 6150 CPU.

To verify the accuracy of the BER estimates, $4.2 \cdot 10^5$ iterations of Monte Carlo simulation on the dynamic metrics are provided in Table 6. The non-write-able BERs without negative BL have less than 5% difference in both cases. The read disturbance BERs are within the same order of magnitude, but a larger number of Monte Carlo simulations is needed because the BER is not very accurate with this sample size. No timing-related failure is observed in the large Monte Carlo simulation, which also agrees with the estimated small RTcrit BER and WTcrit BER. The total simulation time for the Monte Carlo simulation is 117 hours using 30 threads of Intel Xeon Gold 6150 CPU, and the simulation time will increase to 4,500 hours for 90% of confidence and 90% accuracy [15].

Other metrics, like the leakage power and the active power, are also simulated along with the dynamic simulation setup. The average leakage power per bit is 1 pW, the average read active power is 123 nW, and the average write active power is 157 nW. These metrics all satisfy the requirements in Table 1.

Target BER	RTcrit BER	Read dist. BER	WTcrit BER	non-write-able BER w/ and w/o NegBL	Sim. time (hrs)
$6 \cdot 10^{-6}$	$2.4 \cdot 10^{-6}$	$2.1 \cdot 10^{-6}$	$4.5 \cdot 10^{-11}$	$2.9 \cdot 10^{-10}$, $2.52 \cdot 10^{-4}$	1.5

Table 5: BERs of the auto-generated bitcell estimated by the proposed methods

Target BER	RTcrit BER	Read dist. BER	WTcrit BER	non-write-able BER w/o NegBL	Sim. time (hrs)
$6 \cdot 10^{-6}$	0	$4.76 \cdot 10^{-6}$	0	$2.62 \cdot 10^{-4}$	117

Table 6: BERs of the auto-generated bitcell calculated from $4.2 \cdot 10^5$ iterations of Monte Carlo simulation

2.3.6 2KB SRAM Chip Design and Measurement Results

A 2 KB 6T SRAM test chip is designed and taped-out based on the above BER analysis results. No read assist techniques are used because both the RTcrit BER and the read disturbance BER meet the target. Negative BL is utilized as the write assist technique. The SRAM consists of two 64x128 subarrays which are the same structure used in the RTcrit and WTcrit dynamic simulation setup. No sense amplifier is included in the design to reduce the leakage power because the RTcrit is simulated to be working at 15 μS clock period.

The simulation result is displayed in Figure 16. The 2KB SRAM works at 0.5 V and 100 KHz. Two out of four measured chips have a read disturbance failure only during reading '0'. The leakage power per bit is 1.4 pW.

Write 1 Read 1								
VDD/Fre	2.5KHz	5KHz	10KHz	25KHz	50KHz	100KHz	250 KHz	500 KHz
0.3 V	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail
0.4 V	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail
0.5 V	Pass	Pass	Pass	Pass	Pass	Pass	Fail	Fail
0.6 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
0.7 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
0.8 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
0.9 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass

Write 0 Read 0								
VDD/Fre	2.5KHz	5KHz	10KHz	25KHz	50KHz	100KHz	250 KHz	500 KHz
0.3 V	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail
0.4 V	Fail	Fail	Fail	Fail	Fail	Fail	Fail	Fail
0.5 V	Pass	Pass	Pass	Pass	Pass	Pass	Fail	Fail
0.6 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
0.7 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
0.8 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass
0.9 V	Pass	Pass	Pass	Pass	Pass	Pass	Pass	Pass

Figure 16: Chip measurement results of the 2KB SRAM

2.4 Multi-Port Register File Design Space Explorations

The bitcell auto-generation flow only accurately evaluates the SRAM bitcell metrics while ignoring other critical SRAM macro metrics such as the read and write access time and the energy per operation. Also, a lot of macro-level design knobs, like the array structure, the memory hierarchy, and the number of memory ports, also have a significant impact on the performance. The auto-generated bitcells can act as the input of ViPro for comprehensive SRAM macro design space explorations.

A register file is one of the widely used memories in processors that is usually implemented with a multiple port SRAM for fast and compact operation [23]. Register files take up a significant fraction of the power budget of processors [24], and they are also the critical path that constraints the clock cycle in processors [25]. Thus, a delicate Register file design for low power and high performance is necessary, which requires a significant design effort. Furthermore, it is common that more than 30 unique custom register files are employed in a single CPU or SoC [26] [27]. Different specifications of register files lead to a huge amount of effort in full custom circuit design. Register files are required instead of conventional 6T SRAM by the need for simultaneous read and write operations in high-performance processors [23]. The conventional 6T SRAM bitcell has one shared port for both read and write operations, so each operation has to occupy one separate clock cycle. A bitcell with multiple read and write ports can realize higher data throughput by allowing several accesses in one clock cycle. However, the benefit of a multi-port bitcell is accompanied by area overhead, which could eventually limit data throughput because of increasing interconnect delay, and we discuss details about this trade-off in the experimental results section.

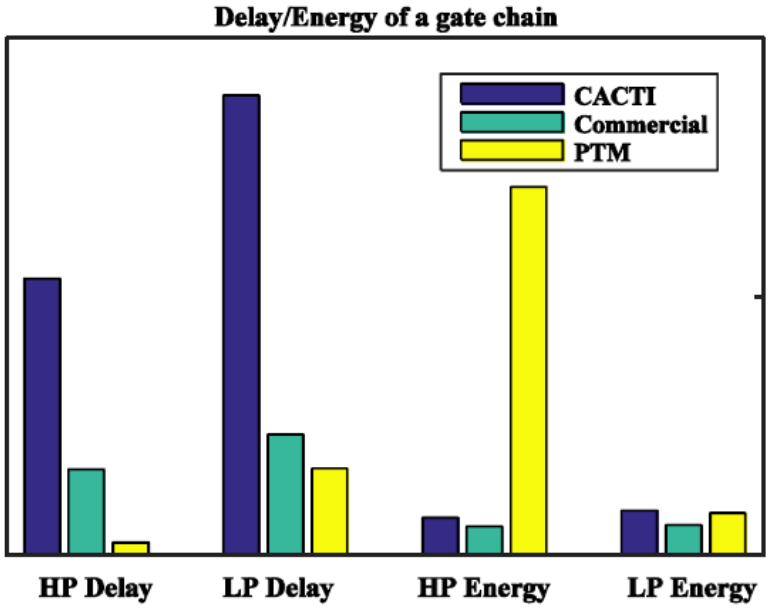


Figure 17: Delay and energy of CACTI, a commercial technology and the predictive technology model (PTM) of a gate chain using high performance (HP) and low power (LP) transistors [21]

2.4.1 Comparison between ViPro and CACTI

ViPro [28] is a virtual prototyping memory design tool that provides a good opportunity to assist register file design optimization, since it can rapidly evaluate different register file prototypes with built-in sub-circuits. The previous version of ViPro only supports 6T SRAM design, and it basically runs brute force simulation by enumerating design knobs like the number of rows, number of columns, and number of banks [28] [20].The outputs of ViPro are delay and energy of all of the evaluated prototypes which can inform the designer of the structures that satisfy the requirements. Similar tool likes CACTI developed by HP Laboratories also evaluates delay and energy of memories, but the results are extremely inaccurate due to using a mathematical circuit model [29]. Figure 17 illustrates the delay and energy of a gate chain which is a fundamental element of circuits, and results of CACTI using high

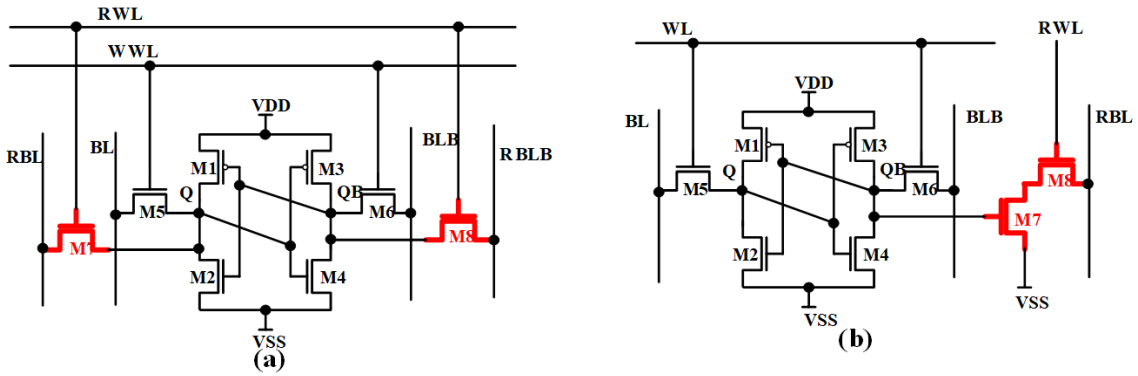


Figure 18: 8T 1R/1W port bitcell with (a) differential BL sensing scheme, (b) single-ended BL sensing scheme [21]

performance and low power transistors are both substantially different from SPICE simulation results of the commercial technology and the predictive technology model for the same gate chain. Fortunately, ViPro can overcome this issue by easily adapting to any selected technology.

2.4.2 New Features in ViPro

Two types of bitcell read port topology are commonly used in register files as shown in Figure 18. One employs the pass gate structure of conventional 6T SRAM which enables differential BL reading; the other utilizes a decoupled read buffer for single-ended BL sensing. The differential BLs is concerned to be more competitive than the single-ended BL in performance, since a sense amplifier (SA) is used to accelerate the differential voltage development. SA can further reduce the total read energy because BLs do not have to be discharged below inverters threshold voltage. However, the cost of SA includes area overhead and design effort for setting dedicated timing constraints. The benefits of differential BL reading are becoming more trivial with the shrinking of supply V_{DD} for two reasons. First, a non-scaling voltage difference VBL should be built on the two BLs to meet the resolution requirement

of the SA, so the reduction of delay is becoming less effective when VBL occupies a larger proportion of V_{DD}. Also, dynamic power consumed by SRAM reading can be calculated as $1/2 * VBL * CBL * V_{DD}$, which is linearly reduced with V_{DD} shrinking but not quadratic as in digital circuits. The second reason is variability in recent technologies. In the research of [30], variation degrades the access time dramatically at low voltage, because the delay distribution of the BL is much wider. Consequently, faster BLs should wait for the worst case BL so that the whole array can function properly, and extra energy is consumed while the slowest BL pair develops its differential. Based on the above reasons, the single-ended BL sensing scheme reduces design complexity in low V_{DD} applications, while the differential BL sensing scheme is better for high performance design.

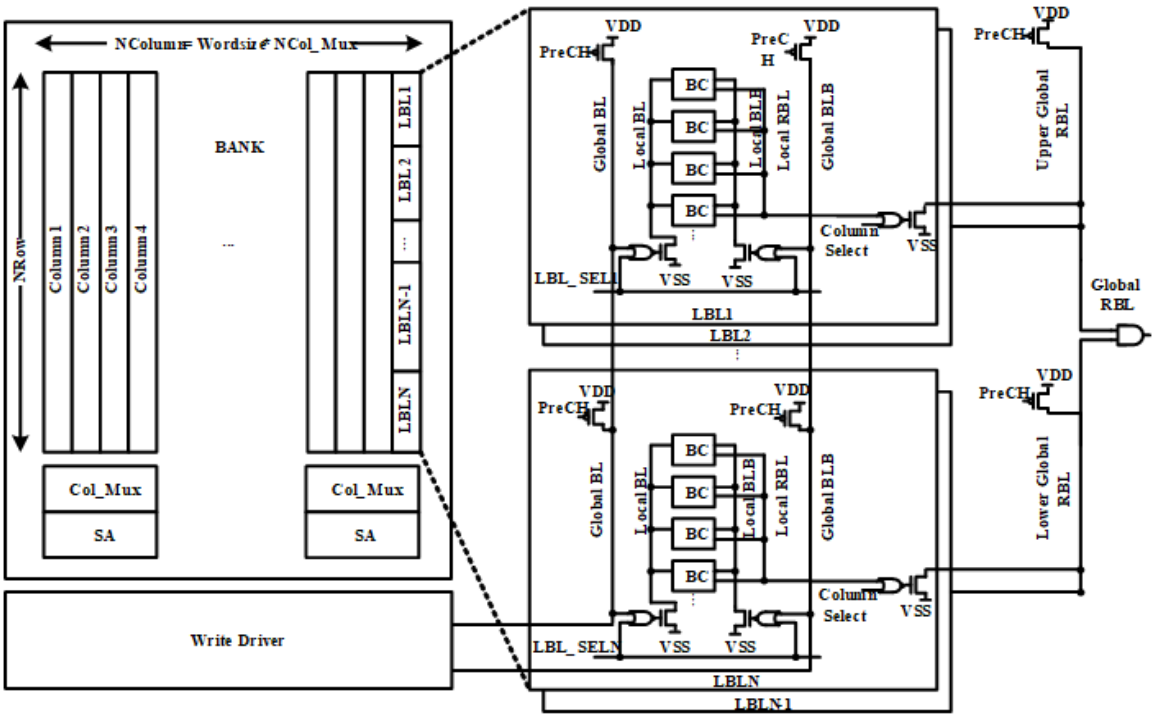


Figure 19: Hierarchical BL structures of both read and write operations. Global BL is divided into N local BLs, and each local is constituted with given number of bitcells. Global RBL is for read operation, and it is realized by AND of an upper global RBL and a lower global RBL to further reduce BL parasitic [21]

Hierarchical BL sensing [31] [32] with local BLs can potentially achieve lower energy consumption as well as faster read and write speed because BL parasitic capacitance CBL is lower for reduced BL length. Number of bitcells on one local BL also affects energy and performance, because hierarchical BL involves overhead. Each of the above three BL sensing schemes can dominate the others for certain applications, so comprehensive analysis of them is critical in designing register file circuit. Several types of multi-port register file bitcell and configurable BL structures have been incorporated into ViPro as new simulation templates.

Figure 19 shows the mechanisms of realizing a hierarchical BL structure for both read and write operations. The register file array has NRow rows and NColumn columns, and rows are divided into N LBLs, thus the number of bitcells in each LBL is NRow/N. Local BLs and global BLs are all precharged to V_{DD}. For read operations, the local RBL will be discharged to ground if the selected row is located in this LBL, and then the upper or lower global RBL will also be discharged by the local RBL based on the location of the selected bitcell. At last, the global RBL is driven to low voltage by AND of the upper global RBL and the lower global RBL. For write operation, the global BL or global BLB will be discharged to ground in accordance with the input data from write driver. LBL_SEL chooses which local BL or local BLB is to be discharged, and then data can be written into the selected bitcell. With the hierarchical BL structure, only a small fraction of total NRow bitcells contribute to the local BL parasitic capacitance, so that the speed of discharging local BLs is significantly improved.

2.4.3 Design Explorations in the Maximum Data Throughput

An example of register file design optimization with ViPro is implemented in the 45 nm technology. A set of simulation results are generated across all combination

of design knobs, different types of multi-ports bitcell, three BL sensing schemes, and memory capacity from 0.5 KB to 512 KB. Data throughput is calculated by the total amount of bytes can be read and write in one clock cycle, and energy is calculated by weighting the energy consumption of read and write operations. For 8T bitcell, we separately measure the read energy as EREAD and write energy as EWRITE, so the energy consumption per operation is 1/2*(EREAD + EWRITE).

As multi-port register file bitcells can realize read and write operations in one single clock cycle, they tend to have higher data throughput than 6T single port bitcells. For example, read and write operations share one port in 6T bitcell, so the data throughput is one half of a two port 8T 1R/1W bitcell if we assume the clock cycle to be the same.

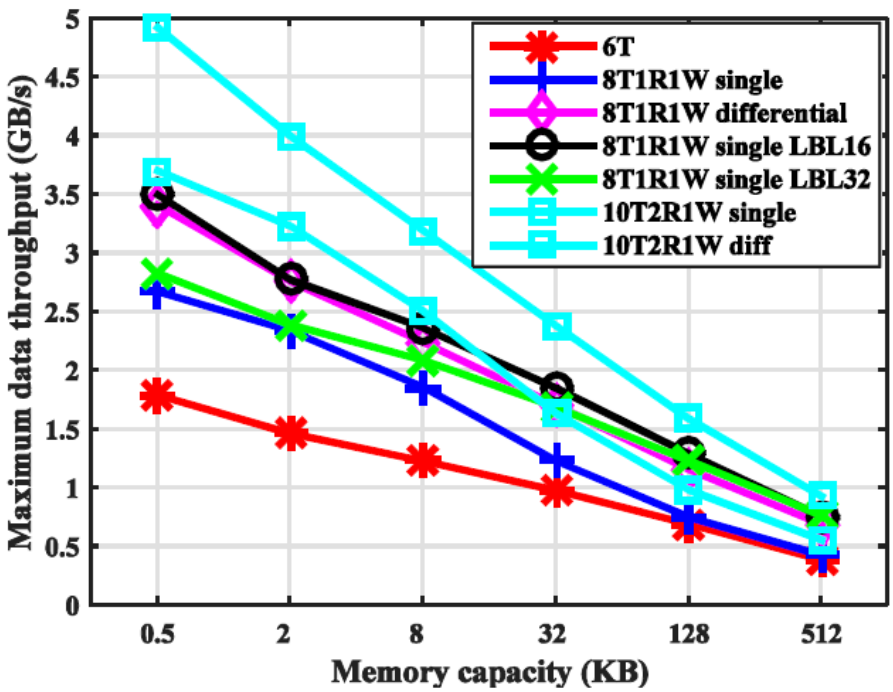


Figure 20: Trends of maximum data throughput under various memory capacities for 6T differential BL, 8T 1R/1W single-ended BL, 8T 1R/1W differential BL, 8T 1R/1W single-ended BL with local BL (16 bits/LBL and 32 bits/LBL), 10T 2R/1W single-ended BL and 10T 2R/1W differential BL schemes [21]

In Figure 20, each data point stands for the maximum data throughput of a bitcell type with certain BL sensing scheme under a given memory capacity, and it is chosen from brute force simulations of all possible combinations of design knobs such as number of rows, columns and banks. ViPro makes this comprehensive analysis easy to implement and fast to produce. Additionally, the tool allows the analysis to be redone in any new PDK very rapidly.

The results imply that the hierarchical BL structure is superior to the single-ended BL sensing scheme for throughput, and the improvement is more significant at larger memory capacity. At 0.5 KB capacity, the maximum data throughput of 8T 1R/1W bitcell is improved by 31% from 2.68 GB/s to 3.5 GB/s with using the hierarchical BL (16 bits/LBL) sensing scheme rather than the single-ended sensing scheme. At 128 KB capacity, the maximum data throughput of 8T 1R/1W bitcell is improved by 72% from 0.75 GB/s to 1.29 GB/s with the local BL structure. This is because hierarchical BL scheme with local BL structure can effectively reduce the bitline data building delay which constitutes a substantial portion of total delay when more bitcells are in one column. ViPro reveals how the best implementation for the 8T cells changes with capacity, which would be costly to discern through manual design and simulation.

With the increasing of memory capacity, the benefit of multiple ports for data throughput becomes more trivial. As Figure 20 shows, the data throughput of 10T 2R/1W ports bitcell with differential BL sensing scheme is about 2.75 times than 6T bitcell at 0.5 KB, while the ratio is only 2.33 times at 512 KB. Although data throughput is degraded with larger memory capacity for all bitcell types, but the decreasing speed of bitcell with more read and write ports is much faster than bitcell with fewer ports. The fact that degradation of performance is more severe in multi-ports bitcell can be explained by greater interconnect parasitic capacitance due to larger dimensions.

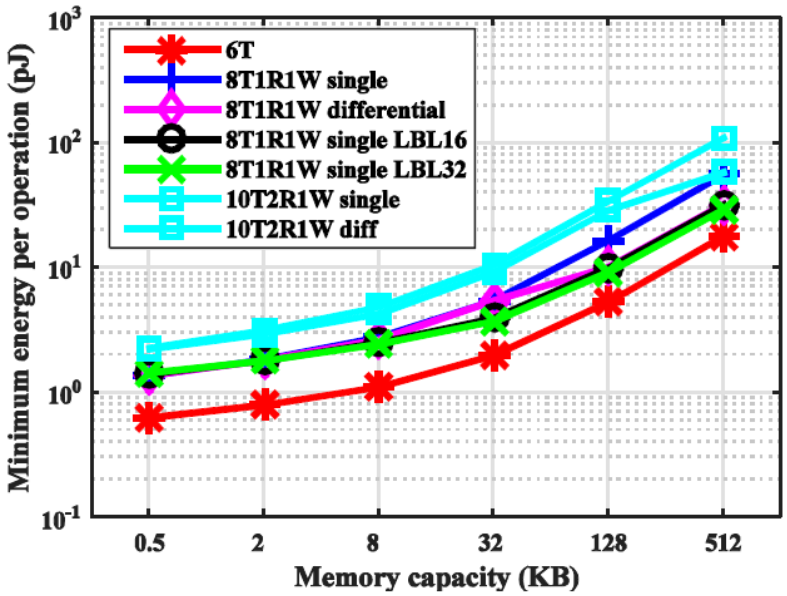


Figure 21: Trends of minimum energy consumption per operation under various memory capacities for 6T differential BL, 8T 1R/1W single-ended BL, 8T 1R/1W differential BL, 8T 1R/1W single-ended BL with local BL (16 bits/LBL and 32 bits/LBL), 10T 2R/1W single-ended BL and 10T 2R/1W differential BL schemes. [21]

2.4.4 Design Explorations in the Minimum Energy Consumption

In Figure 21, each point stands for the minimum energy consumption of a certain bitcell type under given memory capacity, and the minimum energy per operation is chosen from sweeping across all possible combinations of design knobs in ViPro. The take away points are below.

The optimum energy consumption is achieved by different BL sensing schemes as the memory capacity varies. At 0.5 KB, the lowest energy per operation of 8T 1R/1W bitcell is realized by single-ended BL sensing scheme, which is 7.5% lower than energy consumption per operation of differential BL sensing scheme. At 512 KB capacity, the 8T 1R/1W bitcell with the hierarchical BL sensing technique (16bits/LBL) outperforms the single-ended BL sensing scheme about 45% in minimum energy per operation. With ViPro, the most energy efficient register file prototype can be

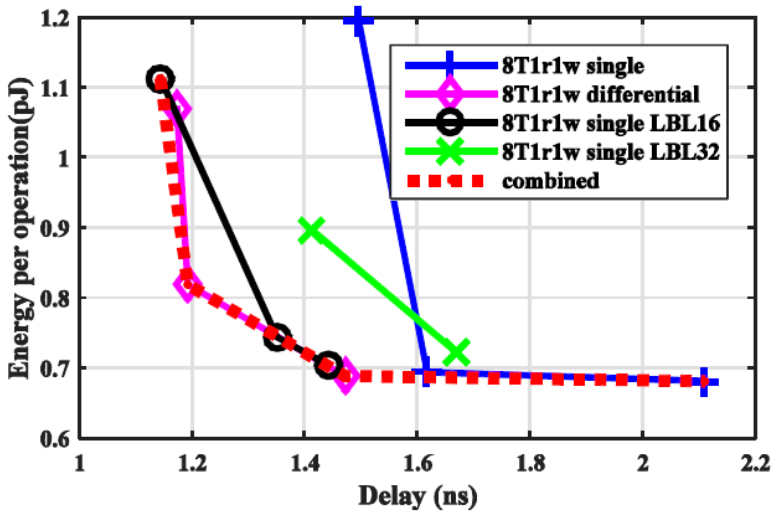


Figure 22: Pareto curves of 8T 1R/1W bitcells with four different BL sensing scheme and one combined at 0.5 KB capacity. [21]

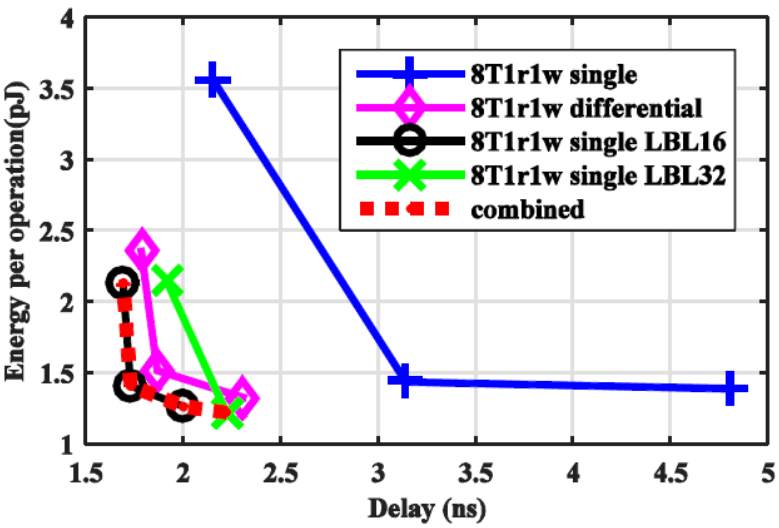


Figure 23: Pareto curves of 8T 1R/1W bitcells with four different BL sensing scheme and one combined at 8 KB capacity. [21]

quickly determined for every memory capacity.

The minimum energy consumption per operation increases with memory capacity, and the effect is increased for bitcell with more ports. At 0.5 KB, the 10T 2R/1W bitcell with the single-ended BL sensing scheme is 1.7X and 3.6X larger than the 8T

1R/1W differential BL and 6T bitcell respectively, while the ratios are 3.4X and 6X at 512KB. The energy consumption is related to interconnect parasitic capacitance, which is more significant in multi-port bitcells.

2.4.5 Trade-offs between the Energy and Delay

Figure 22 shows the Pareto curves of the 8T 1R/1W bitcell with single-ended BL sensing, differential BL sensing, and two hierarchical BL (16 bits/LBL 32 bits/LBL) sensing schemes at 0.5KB capacity. A combined Pareto curve with either lower energy consumption at certain delay or lower delay at certain energy consumption level is plotted as the dotted line, and it shows the design space limit for register files with existing design techniques. Points of the combined Pareto curve originate from different BL sensing schemes. At 0.5 KB capacitance, single-ended BL sensing, differential BL sensing, and hierarchical BL (16bits/LBL) sensing schemes all contribute to the combined Pareto curve; at 8KB capacity, two hierarchical BL sensing schemes contribute to the combined Pareto curve, as shown in Figure 23. Again, ViPro enables this thorough design space exploration that combines circuit level and architecture level knobs for controlling the design.

2.5 Conclusions

2.5.1 Roles of Circuit Models

The SRAM yield analysis model employs features of the normal distribution and the importance sampling to estimate the BER of different SRAM failure types. The simulation time of obtaining the BERs can be reduced by 10,000x compared to millions of iterations of the conventional Monte Carlo simulation. Based on the estimated BERs, the SRAM is guaranteed to have 90% yield in the subthreshold region with

90% confidence. The measurement results for the 2KB SRAM chip agree with the SRAM yield model.

The energy and delay models are defined using simulation results of all the necessary circuit blocks of a memory macro. The models enable accurate estimations of the energy per operation, the read and write delay, and the data throughput, which are the critical metrics used for design explorations in ViPro.

2.5.2 Summary of The Bitcell Auto-Generation Flow

The SRAM bitcell auto-generation flow is a perfect example to demonstrate how circuit modeling can be used to predict the performance of design and to choose appropriate techniques for satisfying various user requirements. Prior research has provided almost all the fundamental knowledge and advanced SRAM design techniques, which are significant advantages for designing ULP and reliable SRAMs. At the same time, it is time-consuming to choose the most effective design knobs and the best performance evaluation metrics from all the available options. For quick yield analysis, the SRAM bitcell auto-generation flow evaluates the BERs with CDFs of the known normally distributed metrics and employs the importance sampling for BERs without knowing the distribution of metrics.

The contributions of this work include:

- Proposes a technology-agnostic subthreshold SRAM bitcell auto-generation flow that explores design knobs in the hyperdimensional design space, for example, the bitcell sizes, bitcell types, and assist techniques.

- Categorizes the SRAM failure mechanisms into read data disturbance, HS data disturbance, read timing failure, write timing failure, and non-write-able bitcells.

- Utilizes appropriate metrics to evaluate the different failures accurately and efficiently. The inverse RTcrit and WTcrit can be used to calculate the BER of read and write timing failures. The static WM can be used to calculate the BER of non-write-able failures. The dynamic read disturbance with importance sampling can be used to calculate the BER of read data disturbance.

- Improves the importance sampling technique to estimate the read data disturbance BER by considering only the dominating parameters, with a simulation time reduction of 10,000x compared to the conventional Monte Carlo simulations.

2.5.3 Summary of The Register File Design Exploration Tool

The ViPro for register file expands the ViPro tool [28] to support fast design optimization for multi-port register files, and it also explored the methodologies of multi-port register file design with the built-in models of memory macro delay and energy. Besides the design effort reduction in building the register file schematics, the simulation time is also accelerated by 5 to 10 times. To improve the accuracy of

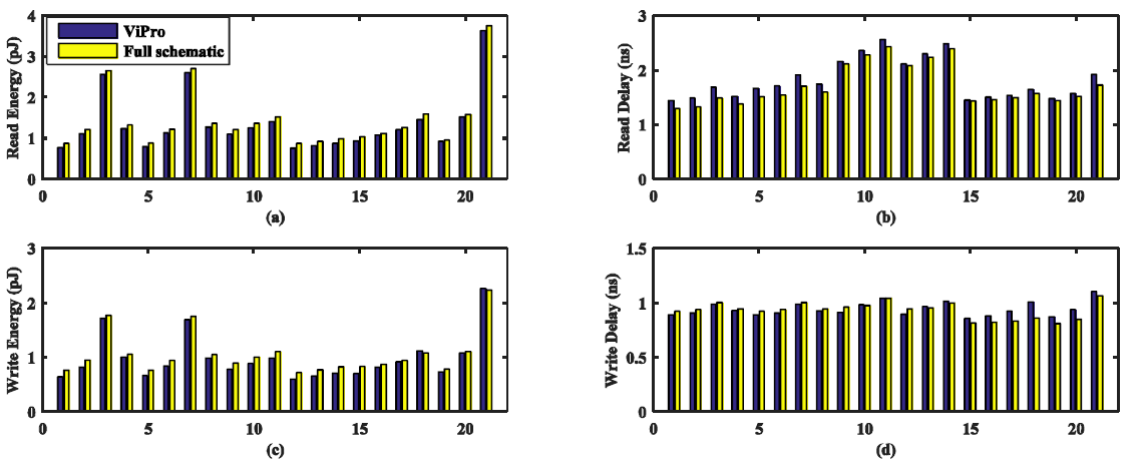


Figure 24: Simulation results comparison between ViPro and full register file schematic. (a)Read energy, (b) Read delay, (c) Write energy, (d) Write delay. [21]

the ViPro tool, several layouts of multi-port bitcells are designed and used to calculate interconnect parasitic capacitance. Comparisons of simulation results between ViPro and full register file schematics verify the accuracy of ViPro as shown in 24. Average errors of read energy, read delay, write energy, and write delay compared to the SPICE simulation are 7.4%, 6.5%, 8.6%, and 1.7% respectively.

In the example of register file design optimization based on a 45 nm technology, ViPro achieves 31% and 72% improvements on the maximum data throughput with a hierarchical BL sensing scheme at 0.5 KB and 512 KB capacities, respectively. The minimum energy consumption per operation decreases by 7.5% with a single-ended BL sensing scheme at 0.5 KB and 45% with a hierarchical BL sensing scheme at 512 KB. The two combined Pareto curves indicate that the optimized register file design technique should be adapted based on the specification, as points of the Pareto curves are from different BL sensing schemes. ViPro supports the analysis of the best option for a given memory capacity and specification requirement by rapidly sweeping through the full design space. Furthermore, the choice of design technique varies at different memory capacities, as the differential BL sensing scheme achieves the optimum balance between delay and energy at 0.5 KB capacity, while the hierarchical BL structure with 16 bits/LBL achieves the balance point at 8 KB.

In conclusion, the expanded version of ViPro for register file not only fills the blank of optimizing multi-port register file optimization, it also employs an additional hierarchical BL sensing technique which brings significant performance improvement and energy reduction to meet required specifications. Following the same method of this work, templates for cache memory have been included by adding a tag array and a comparator circuit, which enhances ViPro to be capable of optimizing the most commonly used SRAM-based scratchpad memory and cache design.

Part of the work in this chapter has been published in [21].

3 Temperature and Supply Variation Stable Clock Reference Design

This chapter presents a 1.05 MHz, on-chip RC relaxation oscillator (ROSC) with a temperature coefficient (TC) of 2.5 ppm/^oC and an absolute variation of 100 ppm over the body-compatible range of 0 to 40^oC. The TC increases to 4.3 ppm/^oC over the range from -15 to 55^oC. The high temperature stability is achieved using a PTAT current reference and a TC-tunable resistor bank for first-order frequency error compensation along with a digital frequency compensation (DFC) block using a single-bit temperature sensor for second-order compensation. A measured RMS period jitter of 160 ps is achieved with a high-speed comparator. The active power consumption of the ROSC is 69 μW with a 1 V supply, and the leakage power consumption is 110 nW while power-gated. The ROSC achieves a fast startup time of 8 μs by employing a voltage buffer to quickly stabilize the voltage reference.

3.1 Motivation & Prior Arts

Radio modules used for wireless networking in internet of things (IoT) systems-on-chip (SoCs) require a high clock stability, and the nature of IoT applications places increasing pressure to limit total radio system active power below 1 mW while also providing rapid on/off and low leakage power for duty cycled operation. For the clock implementation, off-chip components such as crystal oscillators are undesirable due to cost, physical volume, and a long start-up time. As a result, there are a growing number of works targeting stable on-chip clock generation. Among the on-chip clock reference solutions, RC relaxation oscillators (ROSC) are better than ring oscillators (RO) in supply stability and temperature stability. In the field of low-power MHz range clock reference designs, ROSCs are more attractive compared to LC oscillators

by avoiding the integration challenge of high-quality large inductances [33].

In on-chip ROSCs, frequency-locked loops (FLL) are employed with a digitally controlled oscillator (DCO) [34] or voltage-controlled oscillator (VCO) [35] to achieve a clock temperature coefficient (TC) close to that of a crystal oscillator. The FLL architecture provides a good TC by locking the clock frequency to the time constant of an RC filter, but the large power overhead [34] [35], complex start-up sequence [36], and non-negligible start-up time [37] of the FLL architecture prevent this oscillator structure from being used in rapidly duty-cycled ultra-low-power (ULP) radios. ROSCs without FLLs have much lower power consumption, but their frequency stability is limited by the offset and delay of the comparator. Techniques such as chopping and comparator offset cancellation have been used [38] - [41] to alleviate the frequency inaccuracy introduced by the comparator offset, however these techniques become less effective in the MHz range, where the comparator delay becomes a relatively large portion of the clock period and consequently begins to dominate the overall temperature stability. To deal with the frequency error induced by the comparator delay, digital delay compensation is proposed in [42] with a temperature insensitive reference pulse generation, a pulse width detector, and a loop delay tuning circuit, but the TC of the additional circuits and the granularity of the reference pulse limit the frequency error to be about 8000ppm. A wake-up timer [43] implements a constant charge subtraction technique to eliminate the temperature-dependent comparator delay from the clock period, whereas the accuracy of charge subtraction is affected by the amplifier gain. This oscillator consumes nano-watt power, but the frequency error is 4500ppm across temperature at 11 Hz. Other open-loop frequency-error compensation ROSC architectures like the feed-forward period control scheme [44] and the replica inverter switching point tracking technique [45] are also unable to achieve overall temperature stability close to that of a crystal oscillator (< 150 ppm) [46].

To maintain the clock stability comparable to the XOs across the PVT corners, two ROSC stability analysis models are developed to guide the direction for improving the temperature and supply voltage sensitivity. This section also presents a frequency-error compensated ROSC design with a TC of 2.5 ppm/^oC from 0 to 40^oC [47], targeting a wearable, body-compatible range. A high-speed comparator biased by a PTAT current reference reduces the first-order frequency error by stabilizing the comparator delay. A digital frequency compensation (DFC) block automatically tunes a capacitor bank and a TC-configurable resistor bank based on a sub-nW, single-bit temperature sensor to minimize the second-order frequency inaccuracy. To achieve an RMS period jitter of 160 ps, the flicker noise is alleviated by using large size transistors in the first stage of the comparator, and the thermal noise is suppressed with a large bias current obtained from the PTAT current reference. The oscillator consumes a total active power of 69 μW, which is compatible with mW-level radios, and can be power-gated to reduce the power by 627x to only 110 nW. The clock leverages a specially-designed power gating technique to enable quick start-up (within 9 clock cycles), allowing a system to take full advantage of power savings acquired from power-gating during rapidly duty-cycled operation for IoT applications without any significant impact on start-up latency.

3.2 Conventional Relaxation Oscillator Structures

3.2.1 Oscillator Structure with Two Comparators

Figure 25 shows the block diagram and timing diagram of a conventional relaxation oscillator with two sets of comparators and RC components for oscillation. A constant reference voltage V_{ref} is generated by a reference current I_B through a resistor R. Capacitor C_1 and C_2 are alternately charged from 0 to V_{ref} by the current source

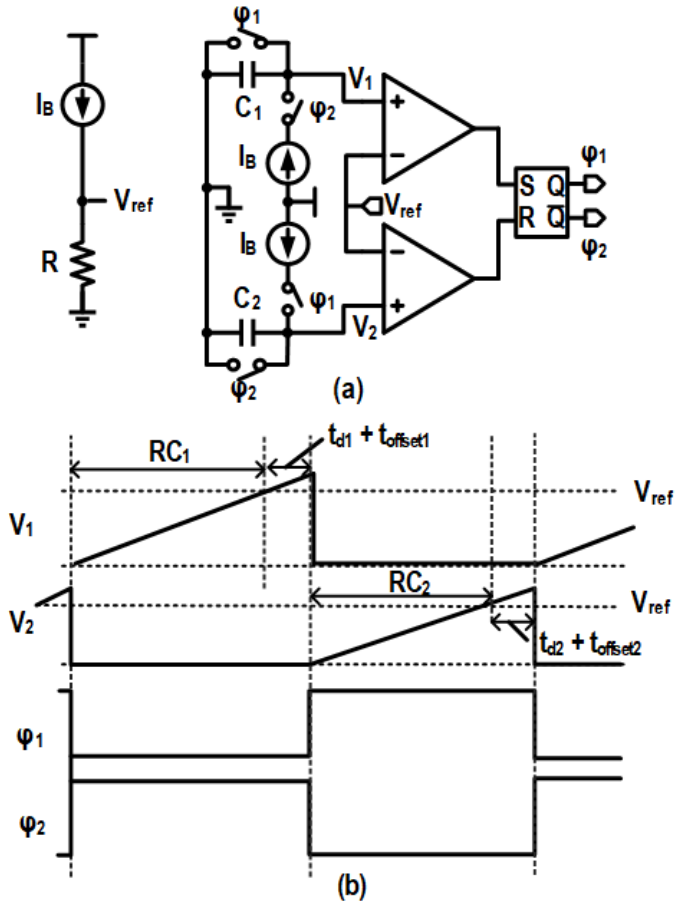


Figure 25: (a) Conventional on-chip RC relaxation oscillator with two sets of comparators and RC components. (b) Timing waveforms. [42]

I_B in every half clock period. The discharging process of C_1/C_2 is hidden in the charging half clock period of C_2/C_1, so the loop delay elements mainly consist of the comparator delay t_{d1} and t_{d2}, and the comparator offset voltage induced delay $t_{offset1}$ and $t_{offset2}$.

A key advantage of this scheme is that it doesnt rely on the absolute accuracy of the current source, thus nominally eliminating a major source of uncertainty from the system. The clock period can be expressed as

$$Tosc = RC_1 + RC_2 + t_{d1} + t_{d2} + C_1 \frac{V_{offset1}}{I_B} + C_2 \frac{V_{offset2}}{I_B} \quad (1)$$

, where $V_{offset1}$ and $V_{offset2}$ are the offset voltage of the two comparators. According to 1, the oscillation period is not sensitive to the supply variation, but it is sensitive to the temperature coefficients of R, the loop delay, and the comparator offset. In reality, the temperature coefficient of the clock period is even more complicated considering the mismatch effect between the two comparators.

3.2.2 Oscillator Structure with One Comparator

To eliminate the mismatch effect of comparators on the temperature stability of the clock period, another ROSC structure with a single comparator and RC components is shown in Figure 26. In this oscillator structure, capacitor C is charged from 0 to V_{ref} by a constant current reference IB, which is the major part of the clock period,

$$Tosc = RC + t_d + t_{offset} + 2t_{reset} \tag{2}$$

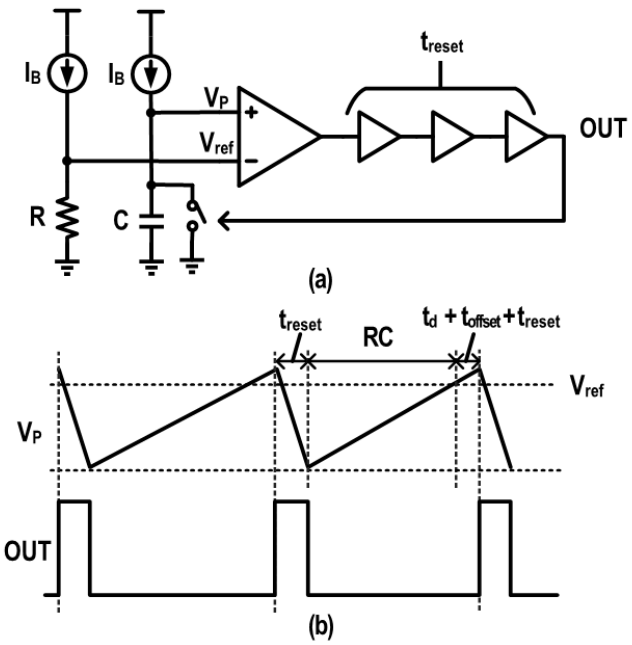


Figure 26: (a) Conventional on-chip RC relaxation oscillator with a single comparator and RC components. (b) Timing waveforms [41].

In equation 2, except for the comparator delay td and the comparator offset Voffset related delay toffset, the capacitor discharging time treset also contributes to the clock period, and it must be long enough for discharging C to ground. To generate treset, an extra number of inverters are used in addition to the switching logic. For a MHz oscillator, assuming R=250KOhm, C=4pF, IB=2μA, and the average discharging current of capacitor is 200μA, treset should be at least 1% (10000ppm) of the clock period, and it can be a large source of frequency variation at different temperatures, because the inverter chain usually has a large temperature coefficient.

3.3 Temperature Stability Analysis

In Figure 27, we propose a ROSC architecture without using two sets of comparators and RC components for decreased clock period variation. To enable oscillation, the upper and bottom plates (CAP PL+/-) of the capacitor bank are periodically swapped by control signals SW and SWB when node VN reaches the reference

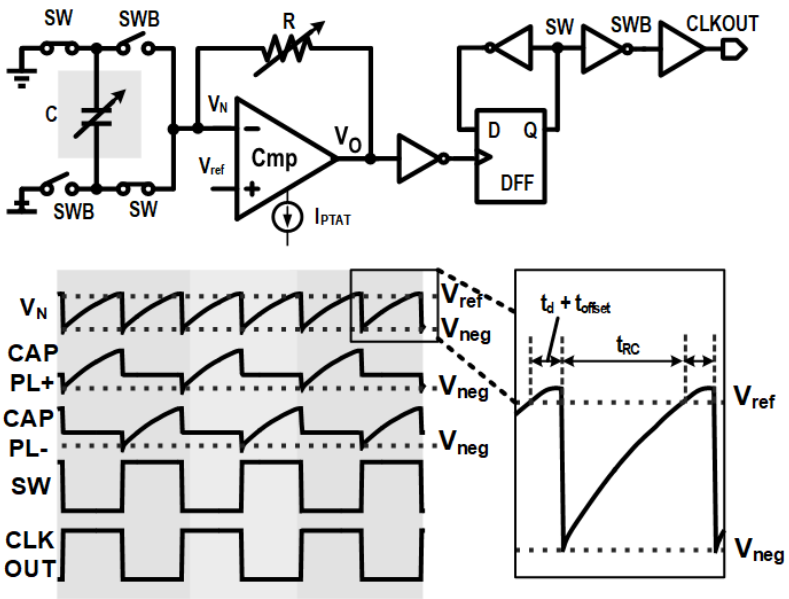


Figure 27: Concept of the proposed RC relaxation oscillator with one comparator and inverted capacitor bank.

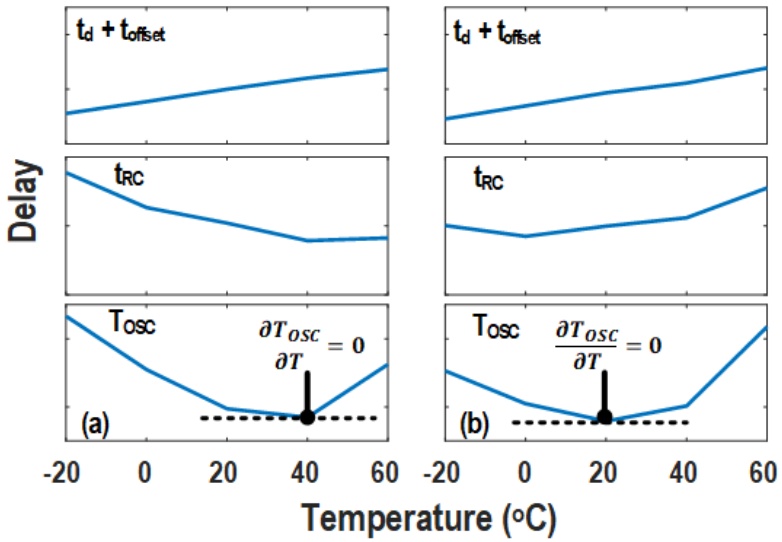


Figure 28: Change in the clock period T_{OSC} vs. temperature with different first order and 2nd order TC subcomponents of td, toffset, and t_{RC} in (a) and (b). The flattest region of the T_{OSC} occurs at different temperatures for (a) and (b), indicating that selecting between the two different t_{RC} can improve the TC of T_{OSC} across the full range.

voltage V_{ref}, and Vneg is the negative plate voltage during swapping. The inverting capacitor structure avoids the extra loop delay of the conventional ROSC in Figure 26 by eliminating the capacitor discharging time. The clock period T_{OSC} and its sub-components, the RC charging time t_{RC}, the comparator delay td, and the comparator offset related delay toffset, are provided in 3. The output stage of the comparator is biased with a PTAT current source to provide controllability over the temperature sensitivity of t_{RC} as shown in 4, where R is the resistance in the oscillation loop, R0 is the output impedance of the comparator, and C is the value of the tunable metal-oxide-metal capacitor bank. The comparator offset voltage Voffset ranges from 0.3mV at -20^oC to 0.7mV at 60^oC in simulation, which accounts for less than 0.1% of T_{OSC} according to 5, but its impact on temperature stability of the clock period is

not negligible for a targeted specification of 150ppm of total frequency change.

$$Tosc = 2(t_{RC} + t_d + t_{offset}) \tag{3}$$

$$t_{RC} = (R + R_0)C * \ln\left(\frac{V_{DD} - I_{PTAT} * R_0 - V_{neg}}{V_{DD} - I_{PTAT} * R_0 - V_{ref}}\right) \tag{4}$$

$$t_{offset} = (R + R_0)C * \frac{V_{offset}}{V_{DD} - V_{ref}} \tag{5}$$

Studies in [35] use a polynomial with temperature coefficients (TCs) to represent the clock frequency. Similarly, if we ignore higher orders of TC larger than 2, T_{OSC} can be expressed as

$$\begin{aligned} Tosc = T_{OSC0}[1 + \\ \left(\frac{t_{RC0}}{T_{OSC0}} TC1_{RC} + \frac{t_{d0}}{T_{OSC0}} TC1_d + \frac{t_{offset0}}{T_{OSC0}} TC1_{offset}\right)\Delta T + \\ \left(\frac{t_{RC0}}{T_{OSC0}} TC2_{RC} + \frac{t_{d0}}{T_{OSC0}} TC2_d + \frac{t_{offset0}}{T_{OSC0}} TC2_{offset}\right)\Delta T^2] \end{aligned} \tag{6}$$

$$TC1 = \frac{t_{RC0}}{T_{OSC0}} TC1_{RC} + \frac{t_{d0}}{T_{OSC0}} TC1_d + \frac{t_{offset0}}{T_{OSC0}} TC1_{offset} \tag{7}$$

$$TC2 = \frac{t_{RC0}}{T_{OSC0}} TC2_{RC} + \frac{t_{d0}}{T_{OSC0}} TC2_d + \frac{t_{offset0}}{T_{OSC0}} TC2_{offset} \tag{8}$$

, where TC1s and TC2s with their names in the subscripts are the first and second order TC of all the specific clock period subcomponents, the $T_{OSC}0$ is the clock period at the reference temperature T_0, and T is the temperature difference T-T_0. According to 6, we can get the first order temperature coefficient TC1 and the second-order temperature coefficient TC2 of the clock period with weighted additions of all the sub-component TC1s and TC2s as demonstrated in 7 and 8. TC1 of T_{OSC} can be

compensated by carefully choosing the $TC1_{RC}$ which is decided by the temperature coefficient of the resistance R and the PTAT current source I_{PTAT} in 4. TC2 is relatively hard to be negated, but the temperature stable region of T_{OSC} shifts to a different temperature during tuning the $TC1_{RC}$, as shown in Figure 28. Instead of burning a huge amount of power in the analog blocks to decrease TC2 of the clock period, we can employ the digital compensation technique described in Section III to improve the overall temperature stability of the ROSC.

3.4 Supply Stability Analysis

If we assume the on-chip voltage reference always generates $V_{ref} = V_{DD}/2$, Equation (4) can be represented as

$$t_{RC} = (R + R_0)C * \ln\left(\frac{V_{DD} - I_{PTAT} * R_0 - V_{neg}}{V_{DD}/2 - I_{PTAT} * R_0}\right) \tag{9}$$

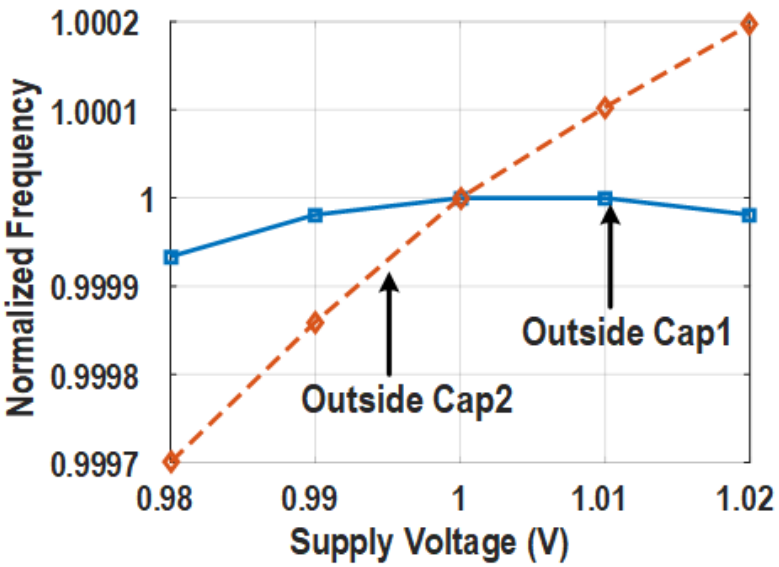


Figure 29: Supply voltage sensitivity with different values of C_{OUT}.

If I_{PTAT}*R0 = 0, then t_{RC} is not sensitive to the supply variation because Vneg is linearly related to V_{DD}. To cancel the positive supply coefficient induced by I_{PTAT}*R0, we can add a capacitor C_{OUT} in Figure 29 outside of the inverting capacitor bank to slightly adjust the value of Vneg at different voltages. Simulation results in Figure 29 illustrate the impact of varying the C_{OUT} value. The frequency error between 0.98V and 1.02V can be reduced to less than 100ppm by carefully tuning C_{OUT}.

3.5 Circuit Implementation for the Proposed Relaxation Oscillator

3.5.1 System Overview

The proposed ROSC system, shown in Figure 30 , consists of a high-speed comparator, a PTAT current reference, a DFC block with a ring oscillator as the on-chip temperature sensor, a resistor bank with 6-bit tunable p-type diffusion resistors (Rpdiff [5:0]), a tunable capacitor bank with 8-bit coarse capacitors and 12-bit fine capacitors, and a $V_{DD}/2$ voltage reference.

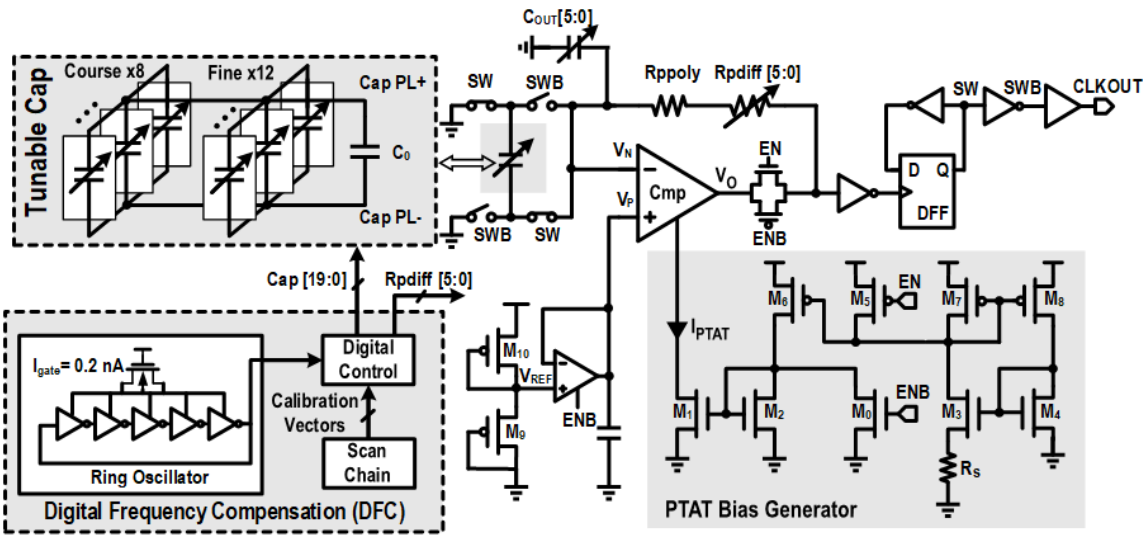


Figure 30: The relaxation oscillator system circuit diagram. [47]

Equation 6 shows that the overall TC of the clock period is determined by the temperature variation of R, Voffset, and td (the TC of C is negligible). To achieve first-order error cancellation, we bias the comparator with a PTAT current source to compensate for the increasing of td and toffset with temperature. This results in higher overall frequency stability than when biasing the comparator with a constant current source, as shown in Fig. 7. Finely tuning Rpdiff [5:0] allows the TC of the clock period to be dramatically reduced over the temperature range, as illustrated in Fig. 8. Because the first-order compensation already provides high temperature stability, only a small amount of second-order tuning is required for Rpdiff compensation across the target temperature range. As a result, we can implement a simple single-bit temperature sensor with only one calibration point such that Rpdiff can be switched between two configurations, corresponding to two different temperature regions. Along with the Rpdiff bits, the DFC also tunes the capacitor bank configuration bits Cap [19:0] to compensate for the potential oscillator frequency shift resulting from adjusting the Rpdiff. To achieve this, a current-starved ring oscillator (RO) is used to provide temperature information to the DFC.

3.5.2 PTAT Current Reference

Figure 30 illustrates the PTAT current reference circuit implemented with a constant-gm bias structure. The current reference value I_{REF} is proportional to $1/R_S^2$, therefore the TC of the I_{REF} is strongly dependent onR_S. By carefully selecting R_S with a proper TC, the first-order ROSC frequency inaccuracy can be significantly reduced. As Figure 31 shows, the clock frequency variation across temperature with a constant current bias is 0.5%, which reduces to 0.2% when using the PTAT current reference.

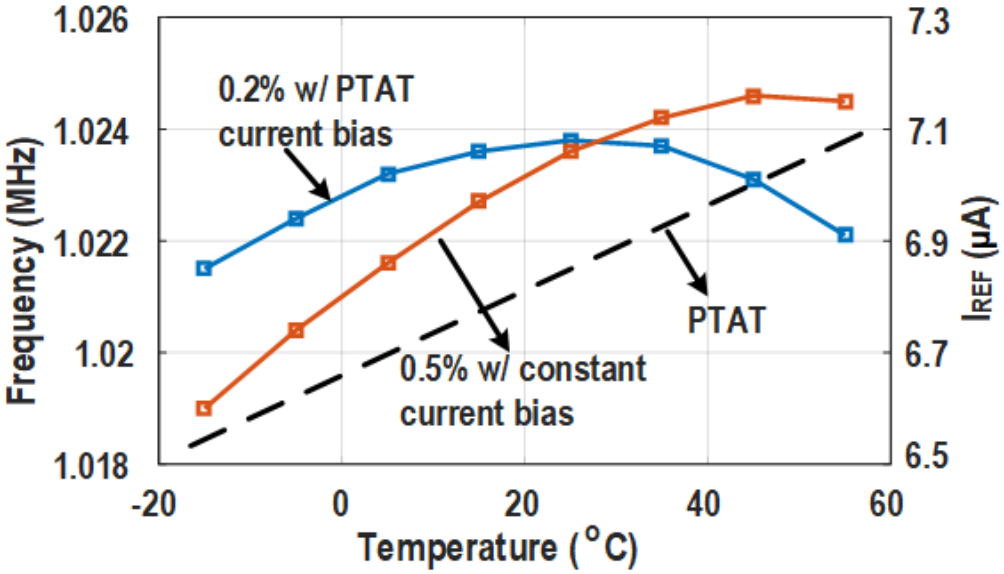


Figure 31: Simulated relaxation oscillator frequency compensation with PTAT current reference. [47]

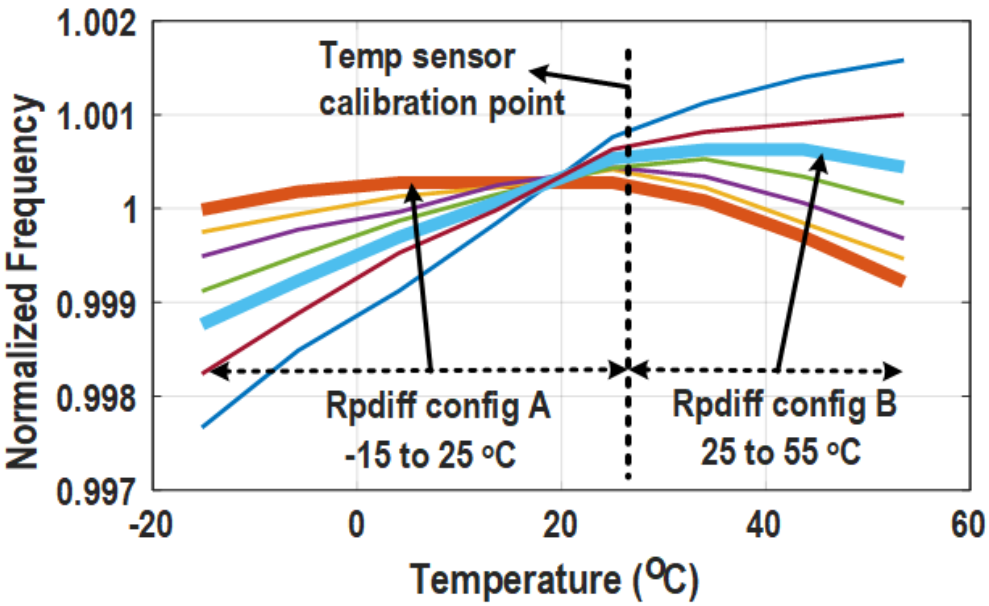
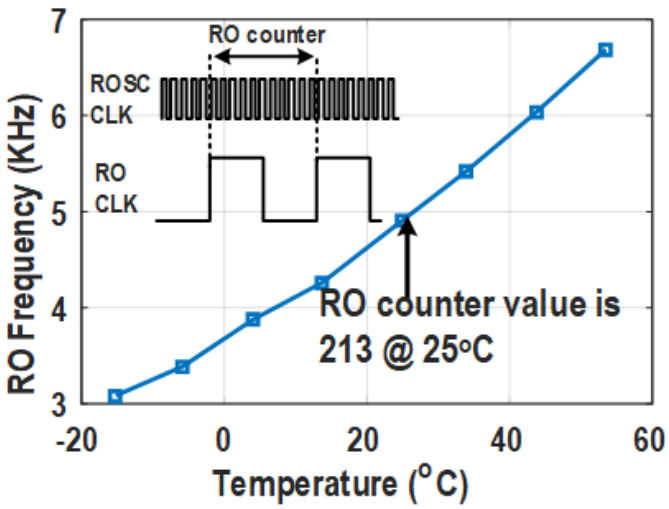


Figure 32: Measured relaxation oscillator frequency with different Rpdiff configurations. [47]

3.5.3 Digital Frequency Compensation

The frequency error compensation obtained from tuning the Rpdiff provides an additional temperature stability to the PTAT current reference, and it also offers the flexibility of post-silicon trimming of the TC of the clock frequency. Measurement results in Figure 32 demonstrate the fine control over the TC of the clock frequency via the Rpdiff and capacitor bank tuning bits. The frequency is normalized to 1 to clearly show the frequency slope change with temperature. Based on these measurements of our ROSC, the frequency error is further reduced to 0.1% (1000 ppm) by finely tuning Rpdiff.

Figure 33 shows the output frequency of the current-starved, leakage powered RO, which is in the low kHz range. Since the ROSC has less than 0.5% raw frequency error across temperature, it can double as a clock reference for temperature sensing. By counting the number of ROSC clock cycles in one RO clock cycle, an RO counter value is obtained that varies roughly linearly with temperature. A single-bit temperature sensor can be easily realized by comparing this counter value with a calibration



Temp (°C)	RO Counter
-15.2	339
-5.8	309
4.1	270
13.7	246
25	213
34	193
43.7	173
53.5	156

Figure 33: Measured ring oscillator frequency and RO counter values. [47]

value in order to distinguish between two temperature regions. Then, the Rpdiff bits are set separately for each region to minimize the total frequency error across the entire temperature range. For example, the Rpdiff configuration A provides the lowest frequency error from -15 to 25^oC, as illustrated in Figure 32 , while Rpdiff configuration B provides the lowest frequency error from 25 to 55^oC. If the temperature sensor calibration value is set to 213, which corresponds to 25^oC for this design (Figure 33), the DFC will automatically adjust the Rpdiff configurations above and below 25^oC to apply the second-order error compensation.

3.5.4 Power-Gating for Rapid Duty Cycling

Since the ROSC design must achieve high temperature stability and minimal period jitter, traditional power-gating header and footer transistors cannot be used since they can introduce new sources of instability that deteriorate performance.

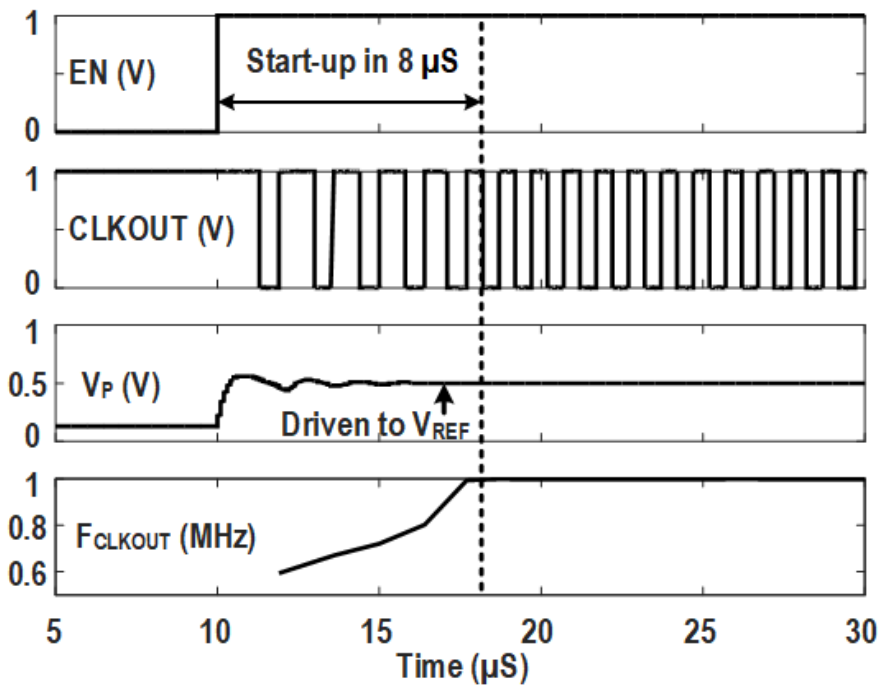


Figure 34: Simulated timing waveforms during start-up. [47]

Instead, transistor M0 is employed to ground the gate of M1, and the high-speed comparator is cut-off from the ground. Transistor M5 drives the gates of M6 M8 to V_{DD}, effectively power gating the PTAT current reference. In normal operation, M0 and M5 are off, so their effect on the frequency stability is negligible. M9 and M10 are a diode-connected stacked voltage reference, so no power-gate is added to them in order to ensure high temperature stability. Long channel length IO devices are used in the voltage reference structure to achieve low power consumption of 12.3 nW.

The start-up time of the proposed ROSC is limited by the time required to charge the internal capacitive node, VP. A large decoupling capacitor and a voltage buffer are added to the output of the voltage reference, as shown in Figure 30, to suppress kickback noise from the oscillator. The high driving strength of the voltage buffer allows the decoupling capacitor to quickly stabilize to V_{ref} so that the ROSC can reach steady state after only a few clock cycles. The start-up waveforms of the ROSC are shown in Figure 34.

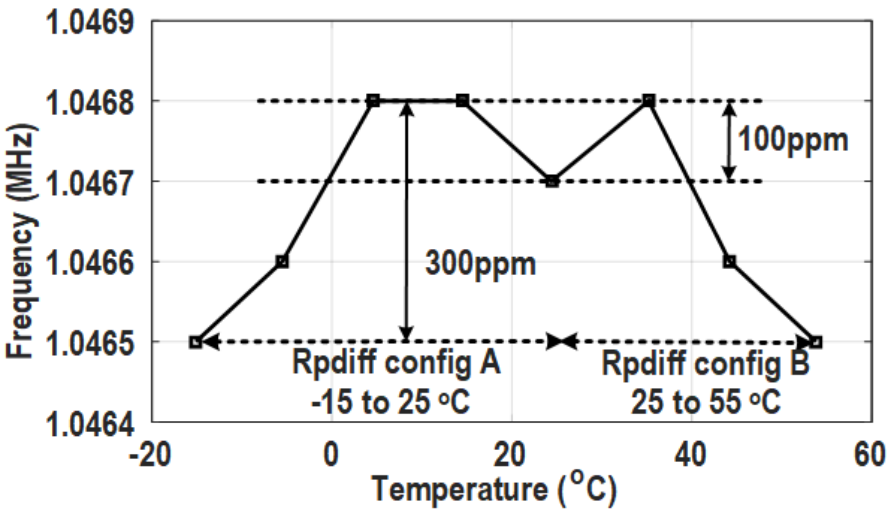


Figure 35: Measured ROSC frequency with DFC automatically tuning the capacitor and resistor bank based on the temperature sensor output. [47]

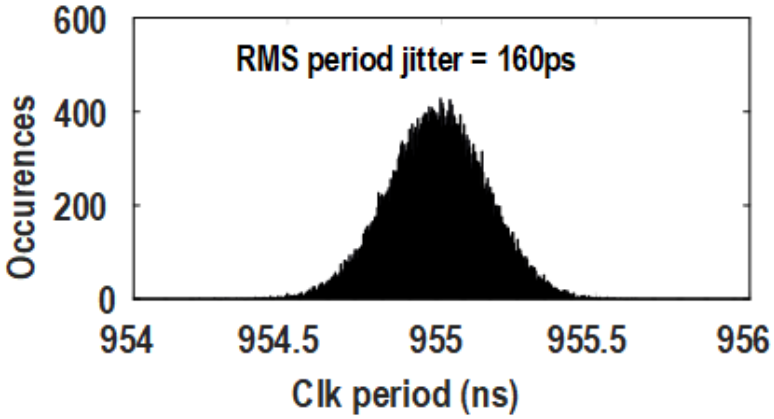


Figure 36: Measured RMS clock period jitter. [47]

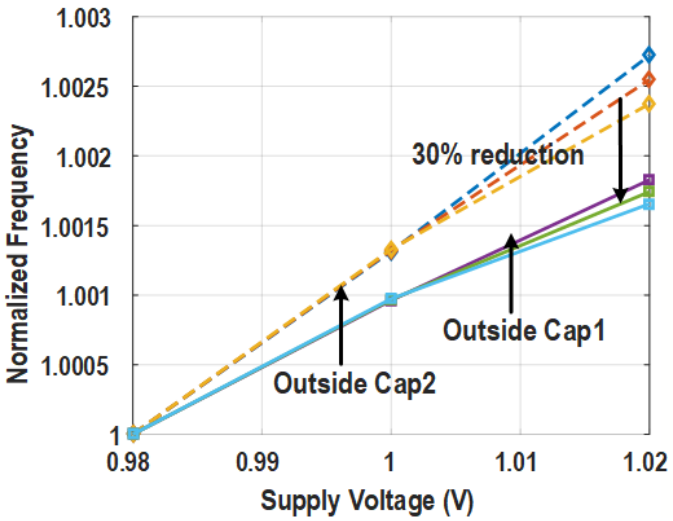


Figure 37: Measured ROSC frequency of 3 chips at different supply voltages with two different outside capacitor values.

3.6 Chip Measurement Results

The frequency output of the ROSC is shown in Figure 35, which demonstrates the transition from the DFC block automatically applying Rpdiff configuration A from -15 to 25^oC and configuration B from 25 to 55^oC. The frequency of each Rpdiff configuration reduces at the edges of the target temperature range, which is

reflected in Figure 32. The measured frequency variation is 300 ppm from -15 to 55^oC, equivalent to a TC of 4.3 ppm/^oC. From 0 to 40^oC, the frequency variation is just 100 ppm, corresponding to a TC of just 2.5 ppm/^oC. Figure 36 shows the measured RMS clock period jitter is 160 ps. Figure 37 demonstrates a consistent 30% improvement of supply stability for three chips between using two different outside capacitor values. The average frequency error with supply variation is 0.17% from 0.98 V to 1.02 V. The measured active power is 69 μW, and the leakage power consumption is 110 nW while power gated.

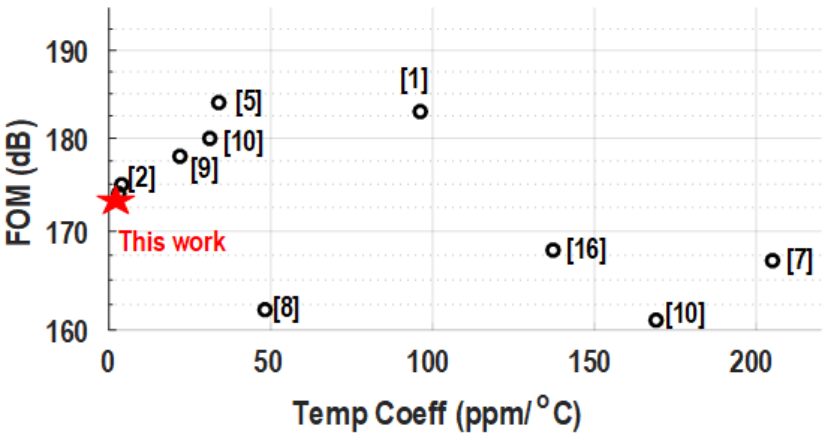


Figure 38: Performance comparison.

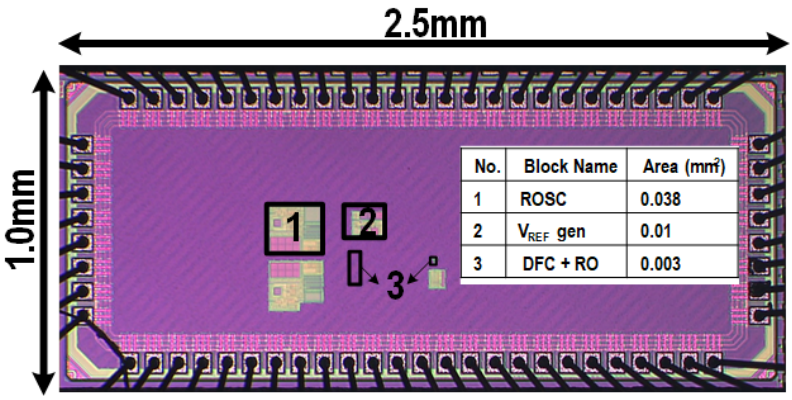


Figure 39: Die photo and design area. [47]

	This Work	[1] ISSCC 2017	[2] ISSCC 2018	[5] JSCC 2016	[7] ISSCC 2013	[8] ISSCC 2016	[9] JSSC 2016	[10] TCAS-I 2016	[16] ESSCIRC 2017	[17] ISSCC 2017
Tech. (nm)	65	65	180	180	65	180	65	180	350	180
Temp. Coeff (ppm/°C)	2.5 @ 0 - 40 4.3 @ -15 – 55	96 @ 0 - 140	3.85 @ -45 - 85	34.3 @ -40 - 80	205 @ 0 - 80	137 @ -40 - 125	22 @ 0 - 90	31 @ -30 - 120	48 @ -40 - 125	169 @ -20 - 100
RMS Period Jitter (ps)	160 (0.016%)	N.A.	22 (0.016%)	N.A.	N.A.	10 (0.01%)	N.A.	79 (0.1%)	N.A.	1060 (0.05%)
Area (mm^2)	0.051	0.005	0.015	0.26	0.01	0.015	0.032	0.012	0.04	0.058
Freq. (MHz)	1.05	1.35	7	0.074	12.6	10.5	0.0185	12.77	1	0.444
P_{DC} (µW)	69	0.92	750	0.11	98.4	219.8	0.12	56.2	210	21.3
Start-up time (s)	8µ	N.A.	N.A.	<2.5m	N.A.	N.A.	N.A.	N.A.	1µ	N.A.
FOM (dB)	174	183	175	184	167	168	178	180	162	161

$$FOM = 10 \log\left(\frac{Freq \cdot \Delta T}{P_{DC} \cdot T_{Coeff}}\right)$$

Table 7: Comparison of recently proposed on-chip oscillator designs [47]

3.7 Conclusion

Table 7 and Figure 38 compare this work with state of the art on-chip clock reference designs, highlighting the excellent temperature stability, low active power, and rapid start-up time. The FOM value, derived from the TC and the power consumption, is 174 dB. This work successfully demonstrates single-digit TC performance in a MHz-range ROSC architecture by employing multiple first and second order frequency error compensation techniques. The advantages of our ROSC architecture compared to FLL structures include lower power consumption [34] and a quick start-up time [37] [49] thanks to the open-loop frequency control. The 10 times power reduction compared to [34] really matters for sub-mW or power harvesting designs. We also leverage the digital frequency compensation technique with traditional analog approaches to make the frequency TC stands out from the ROSCs without FLL [38]-[45], [47], [48]. The high temperature stability of this work within the body-worn temperature range is comparable to the performance of crystal oscillators, indicating that healthcare-related IoT SoCs can potentially operate without an off-chip crystal

clock source, which is critical for reducing system volume in wearable applications. The supply voltage stability is also improved by 30% with an outside capacitor as indicated by the supply stability analysis model.

The main contributions of this chapter include:

- Achieves the XO comparable temperature stability of 100 ppm from 0 to 40^oC for IoT SoCs functioning in the temperature range of human body.
- Enables power gating for reducing the active power and 8 μS rapid start-up time without degrading stability.
- The ROSC chip taped-out in the 65 nm technology achieves 69 μW active power at 1.05 MHz and 1 V, and 110 nW leakage power in the power gating mode.
- Employes two circuit models to guide directions for improving the temperature and supply sensitivity.

Part of the work in this chapter has been published in [47].

4 ULP Wake-Up Receiver (WURX) Base-band Design Exploration

This chapter presents explorations of WURX backend design, which has a significant impact on the reliability and performance of WURXs. Mathematical models on the false wake-up and the missed detection are built for analyzing the sensitivity of the WURX under varying code-word structures. Base on the analysis, smaller length wake-up codes can have up to 6dB sensitivity improvement than longer codes without error tolerance, and error tolerance algorithms can reduce the sensitivity degradation of the longer codes with respect to the smaller codes to about 1dB, which significantly enhances the number of wake-up codes available for large-scale WSNs. An analysis of the radio frequency (RF) transmission energy per wake-up indicates higher power and higher bandwidth signal transmissions are more energy efficient than the lower power and lower bandwidth transmissions for quadratic receivers. Silicon measurement results demonstrate similar trends in the analysis by varying the comparator threshold voltages and correlator error tolerance numbers.

4.1 Motivation

The introduction of ultra-low power (ULP) wake-up receivers (WURX) can suppress the standby power of Wireless Sensor Network (WSN) nodes by duty cycling power hungry main radios [50]. The mesh network topology is promising for reliable and efficient data transmission because nodes are directly connected to each other. To wake up a node in the mesh network, other nodes should broadcast signal packets to be detectable by the target WURX. Unlike base stations in star networks which can afford high power transmitting, nodes in the mesh network are usually energy constrained on both transmission and reception. A higher sensitivity WURX allows

the nodes to expend less energy for a given signal transmission, which improves the battery life-time of WSN nodes or improves maximum transmission distance [50].

The lowest power WURXs demonstrated in literature have employed a detector first front end which has an inherent quadratic dependency between input signal power and output signal power. This work is primarily applied towards the optimization of such ULP receivers [51]. Most of the efforts towards improving WURX sensitivity in the prior arts are dedicated to the RF frontend designs, however, the receiver backend circuits including the comparator and correlator also have significant effects on the sensitivity [52]. In [53], the comparator threshold is automatically controlled to suppress interference and maintain the sensitivity. In [54], optimizing the number of tolerated errors in the correlator shows an enhancement of 5dB in sensitivity. However, none of these works has thoroughly explored the backend design knobs for improving the WURX sensitivity. This work focuses on exploring RF backend designs by evaluating the impact of backend design on WURXs system performance.

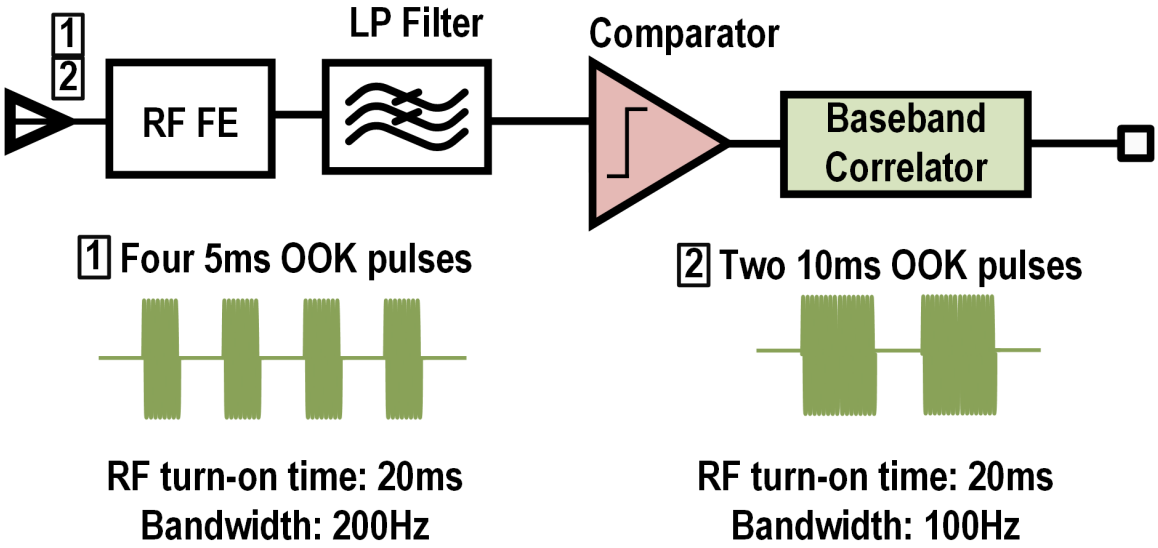


Figure 40: WURXs block diagram, and two types of RF input signal.

4.2 Background

Since the RF backend design is closely correlated to the frontend design, two assumptions are made for clarification. The first assumption is that RF transmission turn-on duration per wake up is fixed to a reference time of 20ms by default for fair comparisons. As shown in Figure 40, a total 20ms RF turn-one time can be used as four 5ms signal pulses or two 10ms pulses in an OOK scheme. The other assumption is that an ideal low pass filter exists at the output of RF frontend, which limits the bandwidth of the signals to the bandwidth required for the desired data rate. The frontend voltage noise levels follow the Gaussian distribution, whose variance varies with the square root of the signal bandwidth. In Figure 40, the integrated noise power of case 1 is about 1.41x higher than case 2.

Three metrics employed to evaluate the WURX baseband design are:

- The sensitivity improvement in dB.
- False wake-ups (FWU) per hour.
- Missed detection (MD) rate.

The sensitivity improvement is relative to a reference sensitivity, which is defined as the minimum input RF power level where a target MD rate of less than 2% and FWU rate of less than 0.5 per hour are achieved utilizing an 8-bit length wake-up code with 50% code weight.

Three knobs explored for baseband circuit design are:

- The comparator threshold voltage (Vtrip).
- The wake-up code length and code weight.
- The correlator error tolerance.

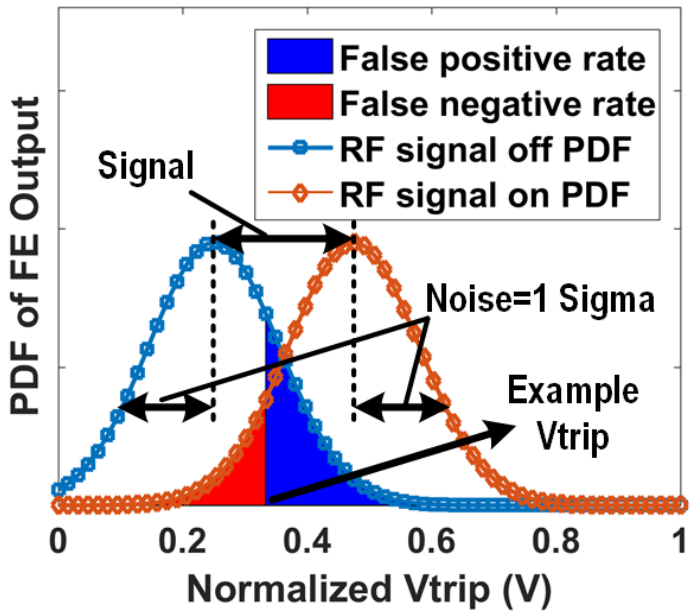


Figure 41: Relation between false positive/negative rate and the comparator Vtrip (normalized to 0-1V) at the comparator output.

4.3 WURX Baseband Design Explorations

In this section, we derive mathematical equations representing the number of false wake-ups and missed detection rate for baseband design analysis. Since noise at the RF frontend output typically follows a Gaussian distribution, the probability density functions (PDF) with RF signal on and off are shown in Figure 41. RF signal power is linearly related to the shift between the baseband analog output level in the presence of an RF input signal and without an RF input signal, and the noise power equals to the sigma of distributions. The bit level false positive rate (Pfp) when RF signal is off and the false negative rate (Pfn) when RF signal is on can be calculated by the blue and red area respectively, relative to a specific comparator threshold voltage and a frontend noise level.

Equation 10 calculates the number of FWUs in an hour using the wake-up code

length (N), the number of 1s (M), bandwidth (BW = M/RF transmission time), and Pfp. The equation doesnt count the false wake-ups resulting from the RF interference, and it assumes the RF signals are independent events. Equation 11 calculates the MD rate.

$$FWU_{1h} = Pfp^{M}(1-Pfp)^{N-M}BW \times 3600 \tag{10}$$

$$MD = 1 - (1-Pfp)^{N-M}(1-Pfn)^{M} \tag{11}$$

$$FWU\ w.err\ tol = \sum_{0}^{err\ thre} FWU(err\ tol) \tag{12}$$

$$MD\ w.err\ tol = 1 - \sum_{0}^{err\ thre} MD(err\ tol) \tag{13}$$

Correlator error tolerance is not shown in equation 10 and 11 for simplicity, and its effect on FWU and MD is considered in equation 12 and 13 by summing up all the possible error tolerance conditions.

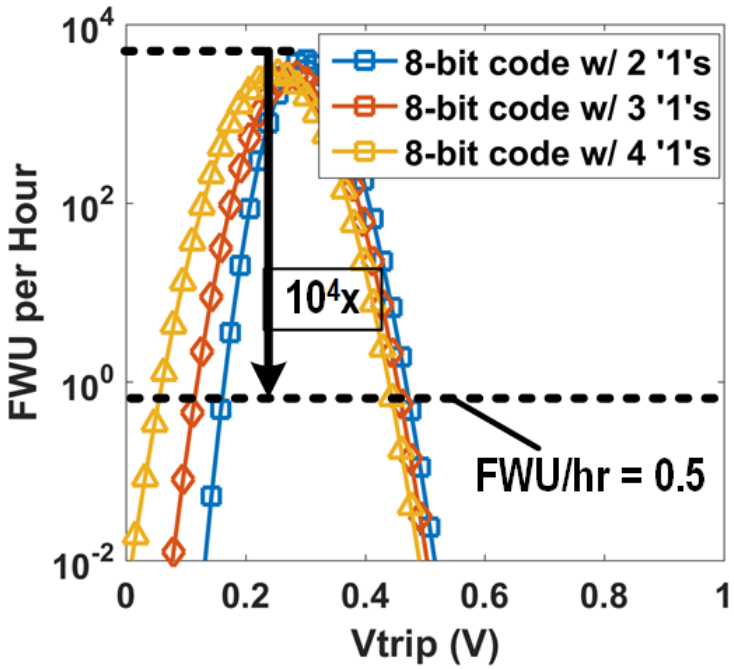


Figure 42: False wake-ups per hour at different Vtrip.

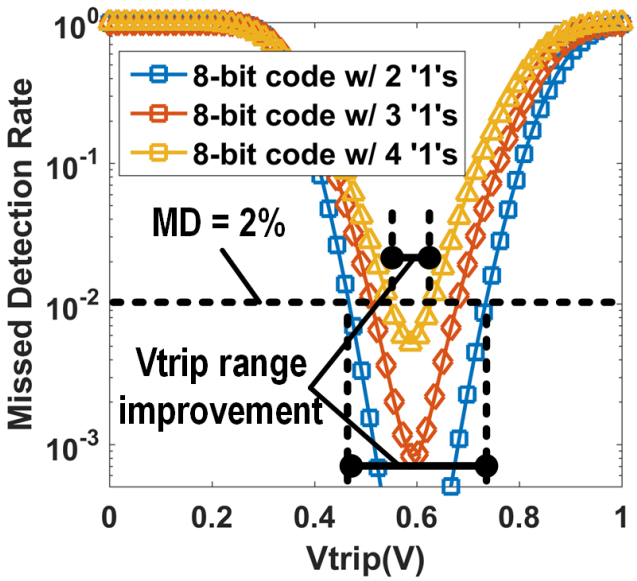


Figure 43: Missed detection rate at different Vtrip.

4.3.1 Impact of Threshold Voltage on False Wake-up and Missed Detection Rate

The comparator Vtrip affects FWU and MD by changing the bit level Pfp and Pfn. In Figure 42 and Figure 43, FWU rate and MD rate are plotted by sweeping the Vtrip at the reference sensitivity according to equation 10 and 11. Results indicate that both the FWU and MD number can vary orders of magnitude by setting different Vtrip. In addition, the range of Vtrip which obtains the target MD rate is important for the WURXs robustness, which also changes for different wake-up codes.

4.3.2 Impact of the Wake-Up Code on the Sensitivity Improvements

In section 4.3.1, the optimal threshold voltage that meets the given requirements on FWU and MD can always be found by looking at Figure 42 and Figure 43. In the sensitivity analysis, the optimal Vtrip is chosen by default to guarantee the anticipation of FWU in an hour is less than 0.5 and the MD rate is less than 2 percent.

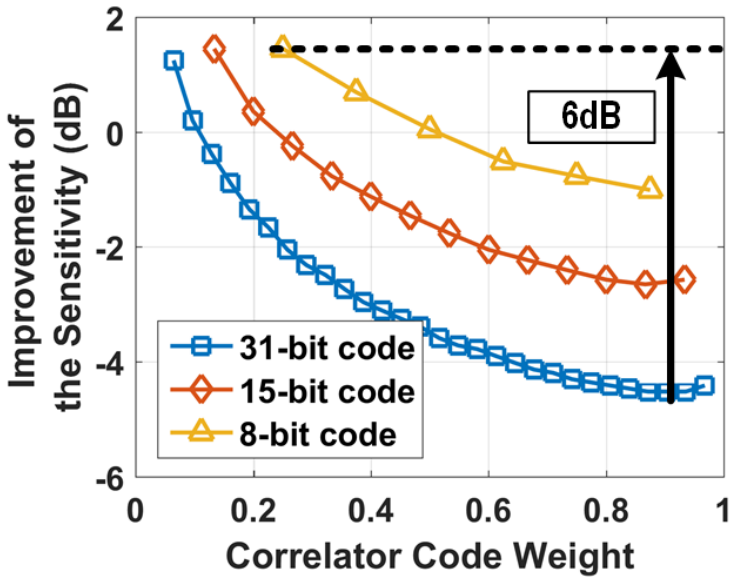


Figure 44: The sensitivity improvements with different code selections.

As equation 10 and 11 show, the wake-up code length N and the code weight of 1s M both affect the FWU and MD value. In addition, M is related to the RF bandwidth since the RF turn-on time is kept constant, thus a larger M requires the RF frontend to work in a higher bandwidth, which results in an increase in the frontend noise power. This relationship results from the fact that the transmitted RF zeros are lower energy compared to RF ones due to the OOKed nature of the code. Figure 44 reveals similar trends that the sensitivity decreases with a larger code weight. An improvement of 6dB in sensitivity is observed by using smaller length codes with less code weight than longer length code with larger code weight.

4.3.3 Impact of the Correlator Error Tolerance on the Sensitivity Improvements

In Figure 44, smaller length codes have better bit-wise SNR than larger ones because the bit-wise error probability rises with a larger number of independent events

due to an increased baseband bandwidth. The baseband correlator can deal with the errors by post-processing the comparator output with error tolerance. Conventional error tolerance algorithm treats the false positives (comparator output is 1 when it should be 0) and false negatives indifferently. As equation 12 and 13 shows, the FWU with error tolerance degrades, however, the MD rate is improved by employing error tolerance.

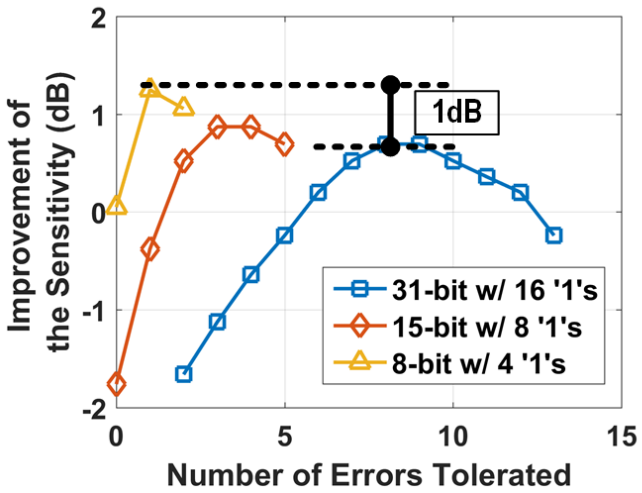


Figure 45: The sensitivity improvements with error tolerance.

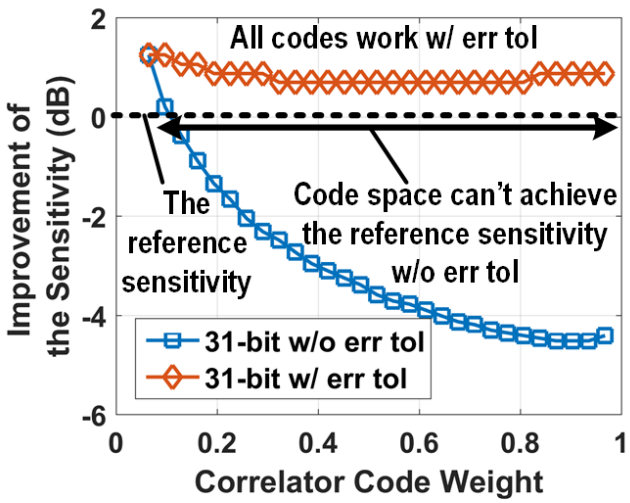


Figure 46: Improvement of available code space with the correlator error tolerance algorithm.

Wake-up codes with a balanced 1s and 0s are favorable because available code space is larger than extreme cases. With a larger code space, its much easier to find a set of wake-up codes with better cross-correlations. In Figure 45, three wake-up codes with different code length and about 50 percent code weight are selected to present the sensitivity improvements with error tolerance. The sensitivity of 31-bit code is still worse than the 8-bit code, but the difference decreases to 1 dB with proper error tolerance. Another observation is that a higher level of the error tolerance is not always better because it harms the FWU.

Figure 46 demonstrates significant improvements in the sensitivity of the 31-bit length codes with optimized error threshold. Given a target sensitivity is set to the reference sensitivity, the full code weight range of the 31-bit codes can successfully detect wake-ups using correlator error tolerance algorithms, while only a small range of code weight can successfully wake-up without the error tolerance at the reference sensitivity. The performance deviations of different code weight are also reduced to about 0.5dB with error tolerance.

4.3.4 Trade-off between the RF Transmission Time and the Energy Consumption per Wake-Up

The previous analysis assumes the RF transmission turn-on time is constant per wake-up event. With a longer RF transmission time, the WURXs sensitivity is improved, however, the transmission energy per wake-up also increases as shown in Figure 47. When keeping transmitted energy constant, it can be shown that shorter transmission times with higher transmission powers perform better than a lower power and longer transmission. This is due to the inherent quadratic nature of the receiver which favors higher power lower duration detection when compared to traditional linear receivers. For energy efficient wake-up, the signal should be transmitted at a

higher bandwidth with a larger broadcasting power.

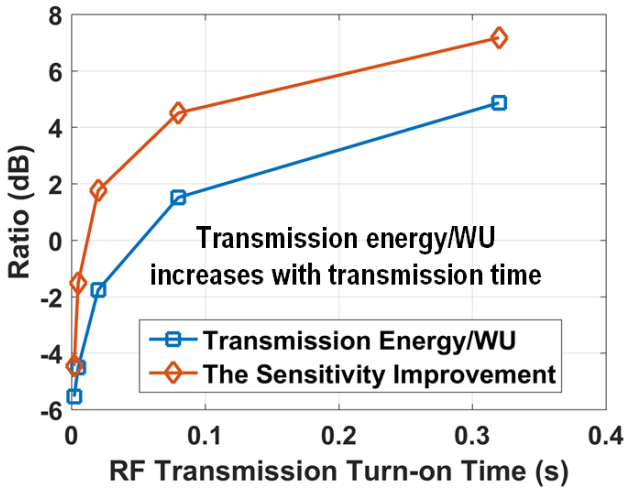


Figure 47: The minimal RF transmission energy per wake-up and the sensitivity improvements at different RF turn-on time.

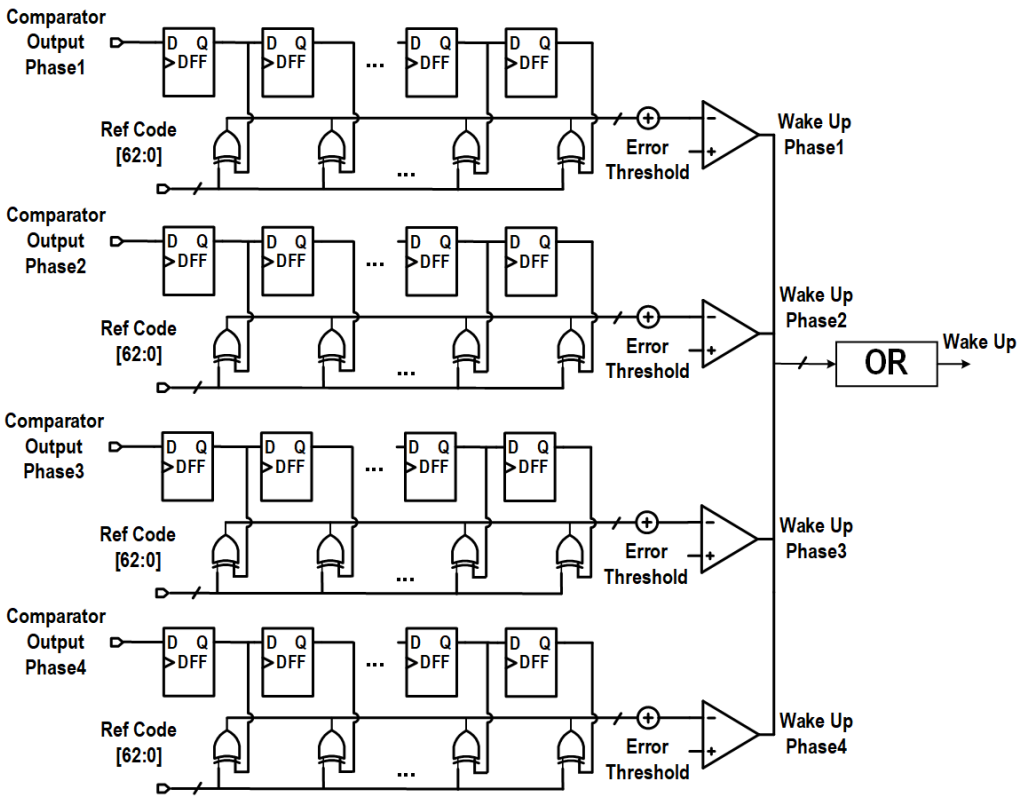


Figure 48: A 63-bit correlator with 4x oversampling of the comparator output and error tolerance.

4.4 Wake Up Code Correlator Circuit Implementations

Several different versions of wake-up code correlator are implemented for the design space exploration. Figure 48 shows a shift register based correlator with 63-bit of reference wake up code, and it enables 4x oversampling of the comparator output. It gives a wake-up high signal when the any of the four received codes has an error less than the error threshold. The reference code and the error threshold are both re-configurable. The other correlators differ from the one in Figure 48 by having different wake-up code length or different oversampling rates.

4.5 Measurement Results

To verify the proposed model of WURXs baseband, measurement results on false wake-ups and missed detections are collected from a ULP passive WURXs chip taped-out in the 130 nm technology similar to the block diagram in Figure 40. Measurement results are available for 4 different comparator threshold voltages and error tolerance number from 0 to 3 at the RF bandwidth of 200 Hz. The correlator error tolerance

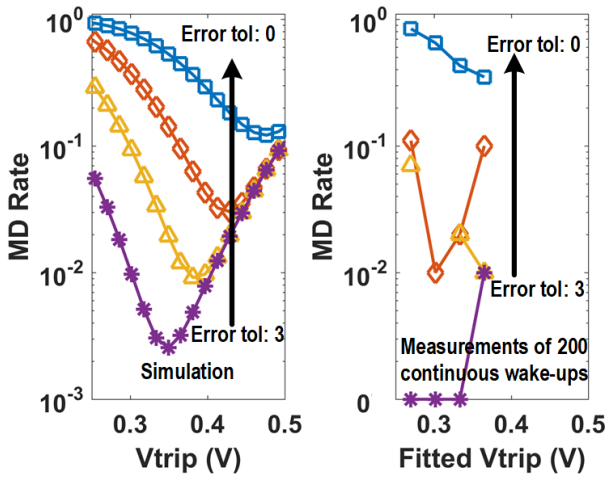


Figure 49: Simulation versus measurement results on the number of false wake-ups in an hour.

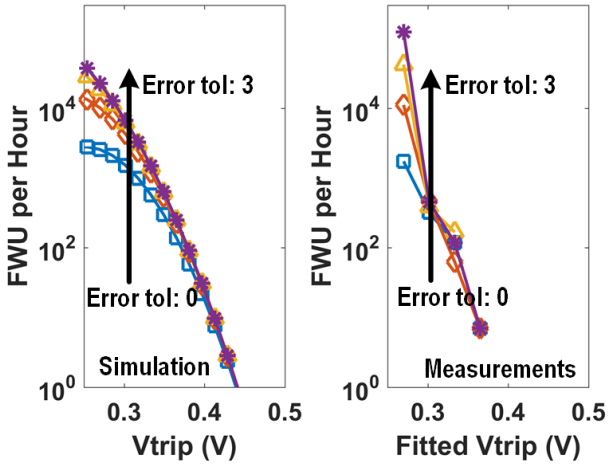


Figure 50: Simulation versus measurement results on the MD rate.

algorithm only tolerates false positive errors. The simulation results are all calculated using equation 10 to 13, and the parameters are set according to the measurement setup. The measurement threshold voltages are post-processed to fit with the simulation threshold voltages. Figure 49 and Figure 50 both demonstrate the measurement results follow the same trend with the simulation by varying two baseband design knobs, the comparator Vtrip and correlator error tolerance. The WURX chip achieves a -76 dBm sensitivity, less than 0.1% missed detection, and less than 1 FWU per hour using a baseband correlator circuit with 8-bit wake-up code and 1-bit error tolerance.

4.6 Conclusion

Near-zero power consumption WURXs can be integrated into IoT nodes to eliminate the mW-level idle power of the primary receiver by waking up the main radio from the power gating mode [3] with a negligible power overhead. The challenge of WURX design is to maintain a good sensitivity while satisfying requirements on the MD rate and FWU rate. This work explores the WURX baseband design by tuning

knobs such as the comparator Vtrip, the wake-up code selection, and the correlator error tolerance. These three knobs are observed to have significant effects on the WURX sensitivity and robustness. The mathematical equations for analyzing the FWU rate and the MD rate are proved to match with the silicon measurement results. According to the analysis, smaller wake-up codes have better sensitivity than the longer codes for a fixed RF transmission turn-on time. The error tolerance significantly increases the available code space by improving the sensitivity of longer codes. For energy efficient wake-up, the signal should be transmitted at a higher bandwidth with a larger broadcasting power.

The main contributions of this chapter include:

- The 7.4 nW WURX system [55] taped-out in the 130nm technology achieves a -76 dBm sensitivity, less than 0.1% missed detection, and less than 1 FWU per hour using a baseband correlator circuit with 8-bit wake-up code and 1-bit error tolerance.

- The wake-up code analysis model helps choosing the appropriate correlator code length, the code weight, and the error tolerance.

- The wake-up code analysis model predicts that a longer wake-up code (longer RF transmission time) improves the sensitivity and consumes more RF energy under the assumption of the same RF power.

- The available wake-up code set is larger with a longer wake-up code length, but it degrades the sensitivity with the same RF energy. Error tolerance of the correlator can restore the sensitivity of 31-bit codes to be less than 1 dB compared to 8-bit codes.

Part of the work in this chapter has been published in [55].

5 Design Explorations of In-Memory Computing for Deep Neural Networks

5.1 Motivation

5.1.1 Introduction of Deep Neural Networks

Deep neural network (DNN), or deep learning (DL), is an essential technology in the field of artificial intelligence (AI). In recent years, multiple breaking through projects, like the AlphaGo and the autonomous driving, are all developed with DNNs. However, the 2010s is not the first era when the DNN becomes a hot research topic. Back to the 1990s, the AI research community started paying attention to the DNN because of the success in hand-written digit recognition [56]. Limited by the performance of computers, researchers found DNN algorithms could not be adapted to solve larger scale problems. Thanks to Moore's law [57] [58], the fact that, the transistor's speed and the number of transistor per chip both double for every 18 months, empowers a new wave of today's AI research in fields like computer vision, natural language processing, and robotics. Several computing platforms other than the conventional CPU are emerging for high-performance DNN processing, for example, the graphics processing unit (GPU) and the tensor processing unit (TPU). Performance of these new processing units can be 1,000 petaFLOPS with an energy efficiency of 10s gigaFLOPS per watt [59]. The large power consumption prevents DNNs to be widely used in the energy-deficient edge computing devices. The upcoming ULP revolution in the computer hardware can potentially be another game changer for DNNs and AI applications.

Figure 51 illustrates the structure of a simple four-layer DNN [60]. The two hidden layers are the reason of deep for DNNs, in contrast to the shallow neural

networks which only have an input layer and an output layer. The input layer of a DNN has an activation vector X or the so-called neuron with a size of three, and it is fully connected (FC) to the first hidden layer with an activation Y by a weight matrix W_{4x3}. Equation 14 demonstrates the multiplication and accumulation (MAC) operations, or the matrix multiplication operation between the activation X and Y. In this case, 24 MAC operations are required for calculating the dot product, where one multiplication and one accumulation are defined as two MAC operations.

$$Y_j = \sum_{i=0}^{2} W_{ji} * X_i + bias \tag{14}$$

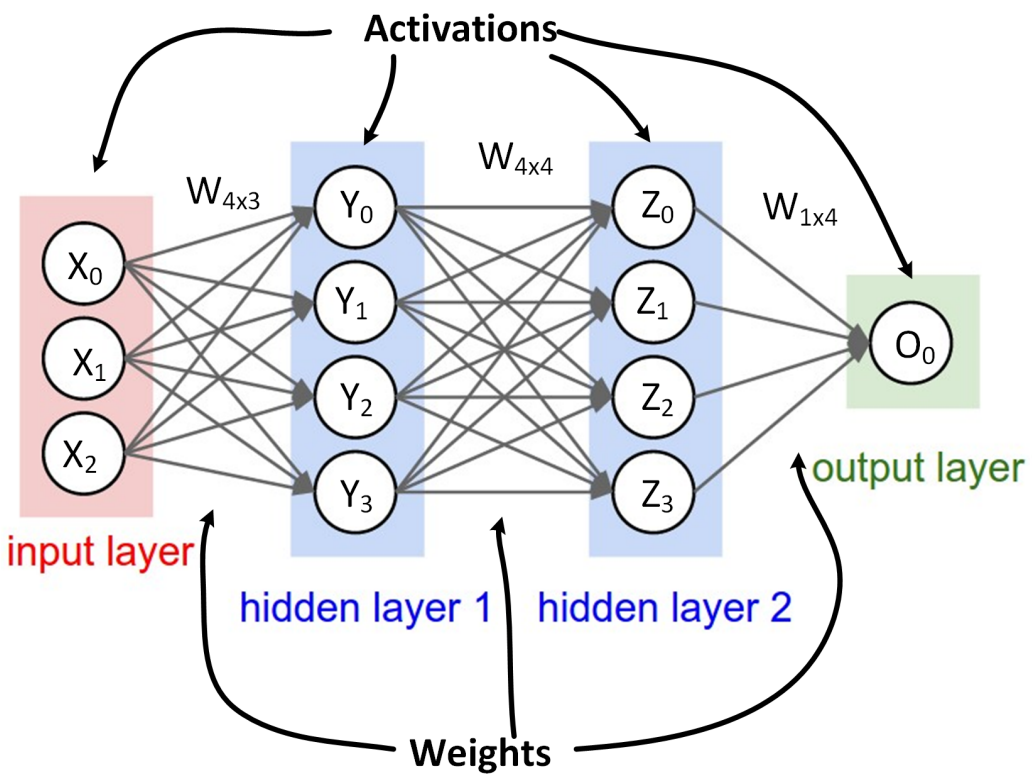


Figure 51: A simple four layer DNN with an input layer, an output layer, and two hidden layers [60].

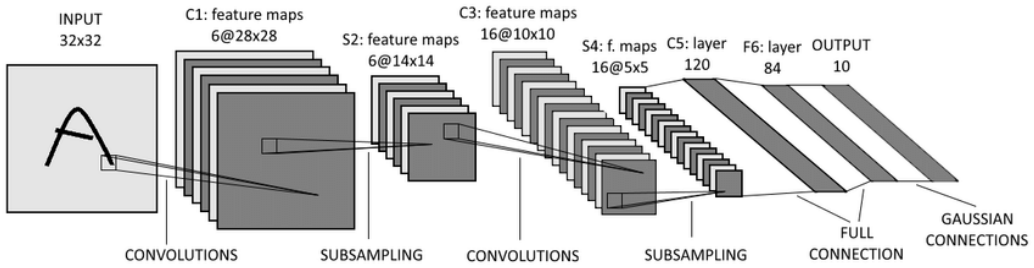


Figure 52: The architecture of LeNet-5. Both convolutional layers and FC layer are used. [56]

Unlike the simple network presented in Figure 51, the input of an actual DNN can be two or three dimensional. For the small hand-written digits database MNIST [56], the input image size is 32x32, so the total number of weights for a FC layer would be 32x32xN, where N is the size of output activations. The activation size does not scale in the FC structure, so convolutional layers are commonly used for the first few hidden layers as demonstrated in Figure 52. For a convolutional layer, the input activations are only connected to a local region of the output activations, so the weights behave like small filters for the input layer. Usually, multiple filters are employed to extract different features, which makes the convolutional layer outperforms the FC structure.

Figure 53 demonstrates a convolutional layer with three-dimensional (3D) shapes. The input activations are organized in width (W), height (H), and depth (C). The output activations are organized in width (Q), height (P), and depth (K). The calculation of one output activation can be described by equation 15 by integrating across all the dimensions of a filter. The computation complexity of the convolutional layer for one input frame is 6-dimension and it consists of P·Q·K·R·S·C MAC operations.

$$Output[p][q][k] = \sum_{r=0}^{R-1} \sum_{s=0}^{S-1} \sum_{c=0}^{C-1} Input[p+r][q+s][c] \cdot Weight[r][s][c] \tag{15}$$

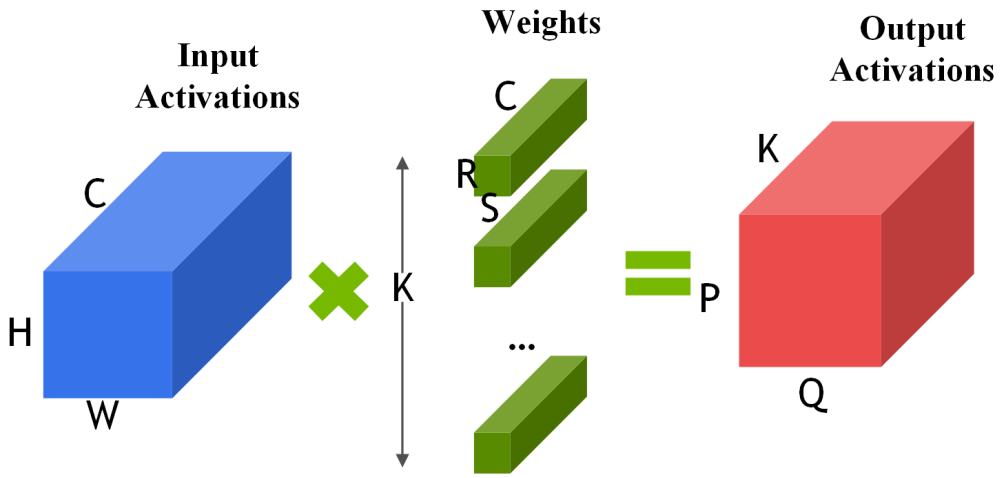


Figure 53: The structure of a convolutional layer with a 3D input H·W·C, and a 3D output P·Q·K

5.1.2 Motivation of In-Memory Computing

The most popular high-performance DNN hardware platforms are GPUs and TPUs, and both of them belong to the classical von Neumann computing architecture. In the architecture, the memory hierarchy and the processing element (PE) are separated. The energy breakdown of the AlexNet in [61] reveals that more than 60% of energy is consumed by on-chip SRAMs in convolutional layers, while the algorithmic logic unit (ALU) only consumes less than 20% of the total energy. The limited on-chip SRAM capacity requires frequent access to off-chip dynamic random access memories (DRAMs) for the FC layers because the high volume of weights could not always be cached on-chip. To make things worse, the energy per bit of the DRAM is more expensive than that of the SRAM.

In summary, issues of the classical von Neumann computing architecture include:

- Data movements between the memory and the processing unit result in high energy consumption and throughput degradation.
- The energy efficiency and the area efficiency are not good enough for supporting

DNNs in edge computing devices.

In-memory computing (IMC) provides a new DNN computing architecture for avoiding the data movement issue and achieving energy efficient computation. This chapter only evaluates the IMC for DNN inference with offline network retraining because online DNN training requires more complicated algorithms, which are not compatible with our interest in the ULP edge computing scenario.

5.1.3 Motivation of IMC Modeling and Top Down Design Methodology

The concept of IMC circuits is to implement MAC operations inside a memory array. The weights are stored in the memory, and the input activations are provided

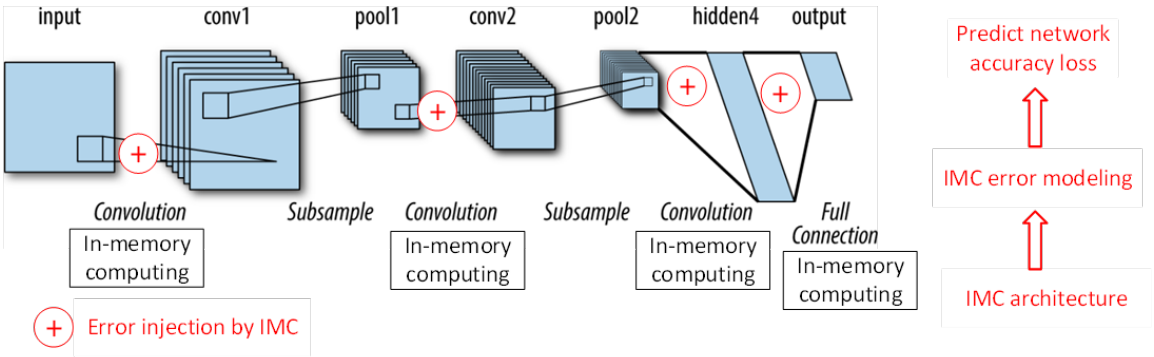


Figure 54: The bottom-up design methodology of IMC circuit [56]

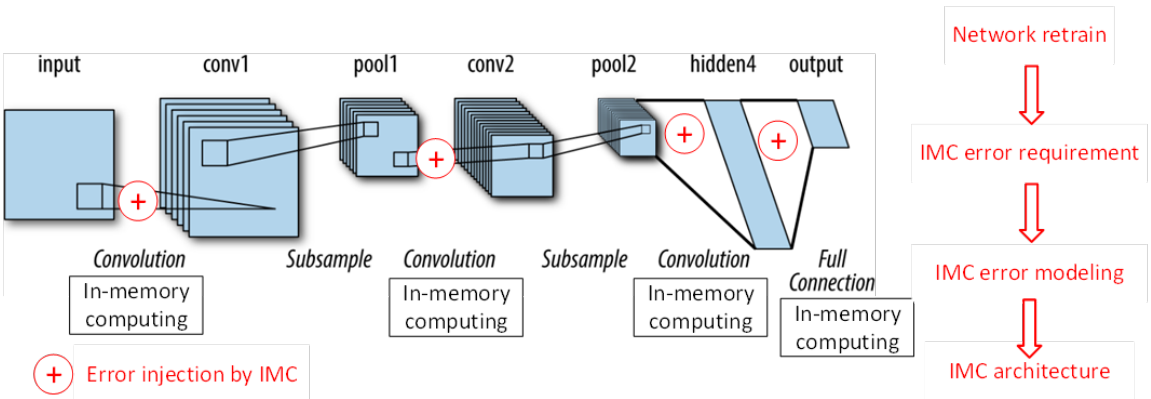


Figure 55: The top-down design methodology of IMC circuit [56]

from the last layer. Digital to analog converters (DACs) are usually required to convert the digital inputs to analog values for IMC. After the finish of MAC operations, the calculated analog dot product should be converted to digital values by analog to digital converters (ADCs) because reliable analog-value storage method is not available. The random process variation of the on-chip memory bitcells and errors of the mixed-signal DACs and ADCs inevitably introduce accuracy losses during the computation.

Figure 54 demonstrates the most commonly used bottom-up design methodology for IMC circuit designs. Designers usually come up with an IMC micro-architecture and use simulations to estimate the error of the IMC circuit. The DNN-level precision loss can also be estimated using the error model to be described in section 5.3. The bottom-up methodology is a passive design approach which could not guarantee to pick an appropriate IMC micro-architecture for better NN performance and energy efficiency.

The top-down design methodology is presented in Figure 55, which provides an interactive design approach between DNNs and IMC circuits. An IMC error model is developed with the effective number of bits (ENOB) as an important metric, which is derived from the signal to noise ratio (SNR) of different IMC circuits. DNNs can also be retrained to tolerate the unbiased random error of IMC circuits. The retraining is critical in maintain the DNN performance because ULP circuit design techniques are likely to introduce more noise during computations. Benefited by the IMC error model, designers can easily make design decisions by evaluating trade-offs between the DNN precision and energy efficiency.

This chapter mainly focuses on building the IMC error model considering the random process variation of memory bitcells and the ADC quantization error, and evaluating impact of the IMC errors on precision of an actual DNN. The off-line DNN

retraining is also studied by including the IMC error during forward propagation of the DNNs, and an example of retraining the LeNet [56] significantly improves the error-tolerance capability of low power IMC circuits.

5.2 Background & Prior Art

Among the recent publications about IMC, a great variety of memory types have been researched for overcoming the data movement issue of the classic computer architectures. The memory types include data non-volatile memories (NVM), such as the phase change RAM (PCRAM) [62], the memristive crossbar [63] - [69], and the floating gate storage [70], and data volatile memories, like the SRAM [71] - [73] and the dynamic RAM (DRAM) [74]. In this section, we focus on modeling IMC circuits based on the memristors (or RRAMs) and SRAMs, because they are compatible with the regular CMOS process technology and can be easily integrated on-chip.

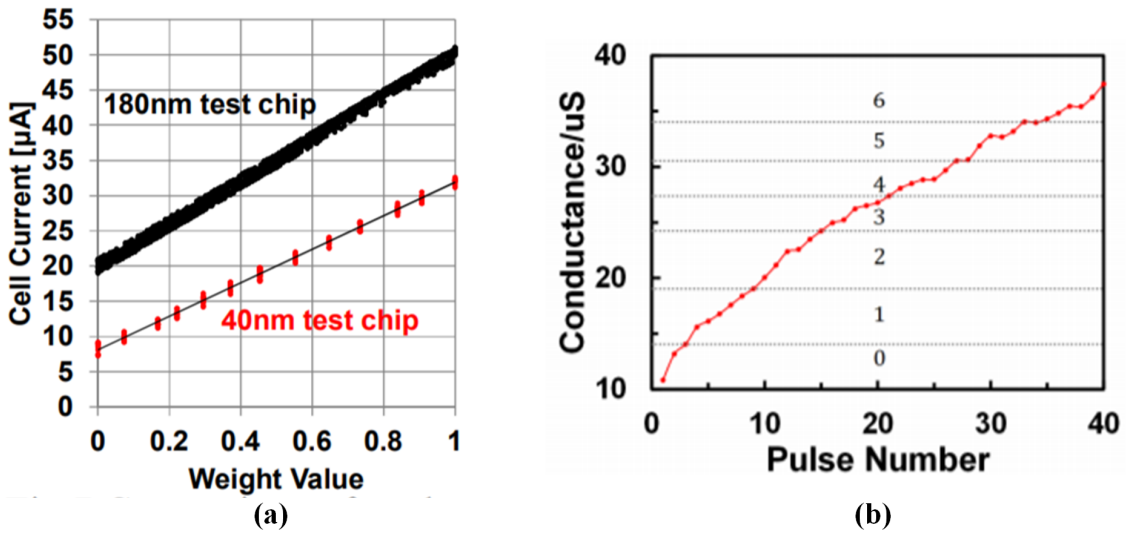


Figure 56: (a) Weight value vs. the cell current of an memristor bitcell [67] (b) Memristor conductance vs. programming pulses [68]

5.2.1 IMC with the Memristor

Several advantages of the memristor are promising for IMC:

- Dense weight storage. Figure 56 (a) presents that the stored weight value of a memristor bitcell can be 3-bit to 4-bit with a monotonic cell current [67]. The multi-level memristor enables more weights to be stored on-chip which potentially reduces the weight movement. The linearity of memristors can be improved by post-silicon calibration for storing even a larger number of bits in one bitcell.

- Small bitcell read current. Figure 56 illustrates the conductance of a memristor bitcell with different programming pulses [68]. The bitcell current is less than 100 nA at a 1 V supply, and it is promising for energy efficient MAC calculations.

- Data retention. The non-volatile feature enables complete shut down of memristor arrays during the idle mode, which is beneficial for the ULP IoT devices.

Even though the reported memristor bitcell levels differ a lot in the linearity and the cell current, we could roughly estimate the energy efficiency per MAC operation for IMC. Assume a situation that, all the memristor bitcells are calibrated to hold 4-bit weights, the input activations are also 4-bit, the average cell current is 1 μA, the V_{DD} is 1V, and the current integration time is 10 ns. The energy consumption per MAC can be estimated by

$$E_{MAC} = (1\text{V} * 1\mu A * 10ns + \frac{E_{ADC}}{N}) \tag{16}$$

, where N is the volume of vector multiplication, and E_{ADC} is about 2 pJ according to the ADC survey [75]. N is usually limited by the number of bitcells on a mem-

ory bitline and a nominal value is 128. An estimation of the E_{MAC} is less than 30 fJ regardless of the computation inaccuracy, and the ADC energy consumption is dominating compared to the memristor bitcell.

5.2.2 IMC with SRAMs

Unlike memristors, SRAMs are available on every CMOS process technology. However, an SRAM bitcell can only store 1 bit of weight value, so the IMC based on SRAMs is limited to DNNs with binary or ternary weights. Fortunately, the XNOR-Net proves that the performance of binary weighted networks is comparable to networks using larger weights [76]. A recent work utilizing SRAMs for IMC also demonstrated a good energy efficiency of 72 fJ/MAC [72].

5.2.3 Other Types of Analog Computing

IMC is one type of analog computing techniques. Other innovated analog computing techniques, i.e., computing in the switched capacitor arrays [77] and computing in ring oscillators [78], have shown good energy efficiency with chip measurement results.

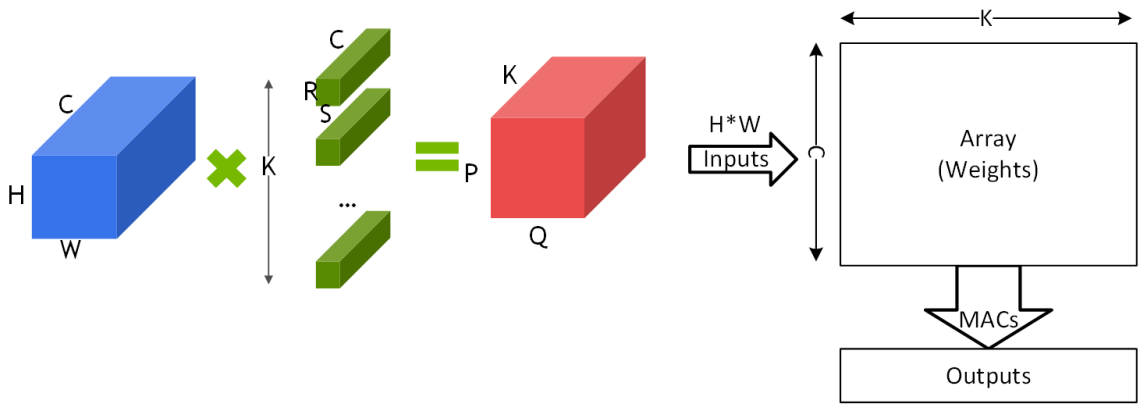


Figure 57: Map a convolutional layer to memory

5.3 IMC Architecture

Figure 57 illustrates the method to map a convolutional layer to memory arrays for IMC. The weights are stored in the memory arrays, the input activations are provided from the last layer, and the output activations are fed to the next layer. The weights are mapped with the following methods:

- Map the weight dimension C to different rows. The channel depth C is usually a large number, and the nominal number of rows of a memory array is likely to be smaller than C, so C needs to be mapped to multiple memory arrays. In this case, the vector dot product is divided into several partial dot products, and each of them is calculated on one of the memory bitlines. These analog partial products are quantized with ADCs and stored digitally in registers, and they are summed together outside of the memory arrays.

- Map the weight dimension K to different columns.

- Map the weight dimensions R and S to different arrays or different rows. In the case that C is not large enough to fill up all the rows, R*S can also be mapped to different rows.

5.3.1 Discussion of the Input DAC

Input activations are multiple-bit digital values for most of the DNNs, and they usually requires digital to analog conversion for computing the vector dot products on the memory bitlines (BLs) in an analog manner. The digital input activations can be converted to the amplitude or the slope of memory wordlines (WLs), which influence the bitcell current [74]. They can also be converted to the WL pulse width or the bitcell current integration time [72].

An alternative approach without using DACs is feeding the input activations serially in the binary form [79]. The serial binary inputs decide that if the WLs to be turned on or not, and IMC calculates the partial dot products with analog MAC computations and digital shift additions. The main benefit of using the serial input approach is that, no extra non-linearity and noise are introduced by DACs. Also a smaller ADC resolution is needed for quantizing the analog partial dot products because the input activations only contribute to 1-bit more required resolution instead of multiple bits. The detailed implements of the serial input approach will be described in section 5.3.2.

Advantages and disadvantages of both approaches are summarized in Table 8. This work only evaluates the serial input approach because it alleviates the resolution requirement on ADCs which can reduce the energy consumption as demonstrated in equation 16. Reasons of not using the input DACs are larger IMC error and higher energy consumption.

	Advantages	Disadvantages
Input bit serial	Ideal input activation, low ADC resolution required	Extra memory access energy
Input DAC	Less memory access times	Extra source of input activation error

Table 8: Pros and Cons of input DAC and input bit serial

5.3.2 Memristor IMC Micro-Architecture

Figure 58 (a) illustrates the IMC micro-architecture based on memristors. Assume that memristor bitcells have enough resolution to hold multi-level weights, the complementary bitcell structure is used to represent the sign of weights. If the weight is a positive value, it will be stored as the conductance of the left memristor bitcell. The right memristor bitcell will be programmed to the high resistance state (or '0'

state), so that it won't affect the BLB discharging during the analog computation. If the weight is a negative value, it will be stored as the conductance of the right bitcell while the left bitcell is in the high resistance state.

Input activations are fed in the bit-serial manner to different rows of the memristor array. Assume that the activations are four-bit values, they are assigned to the WLs of different rows in four continuous clock cycles. If the incoming serial bit is '0', the corresponding WL will stay at the turned-off state, and no discharging current

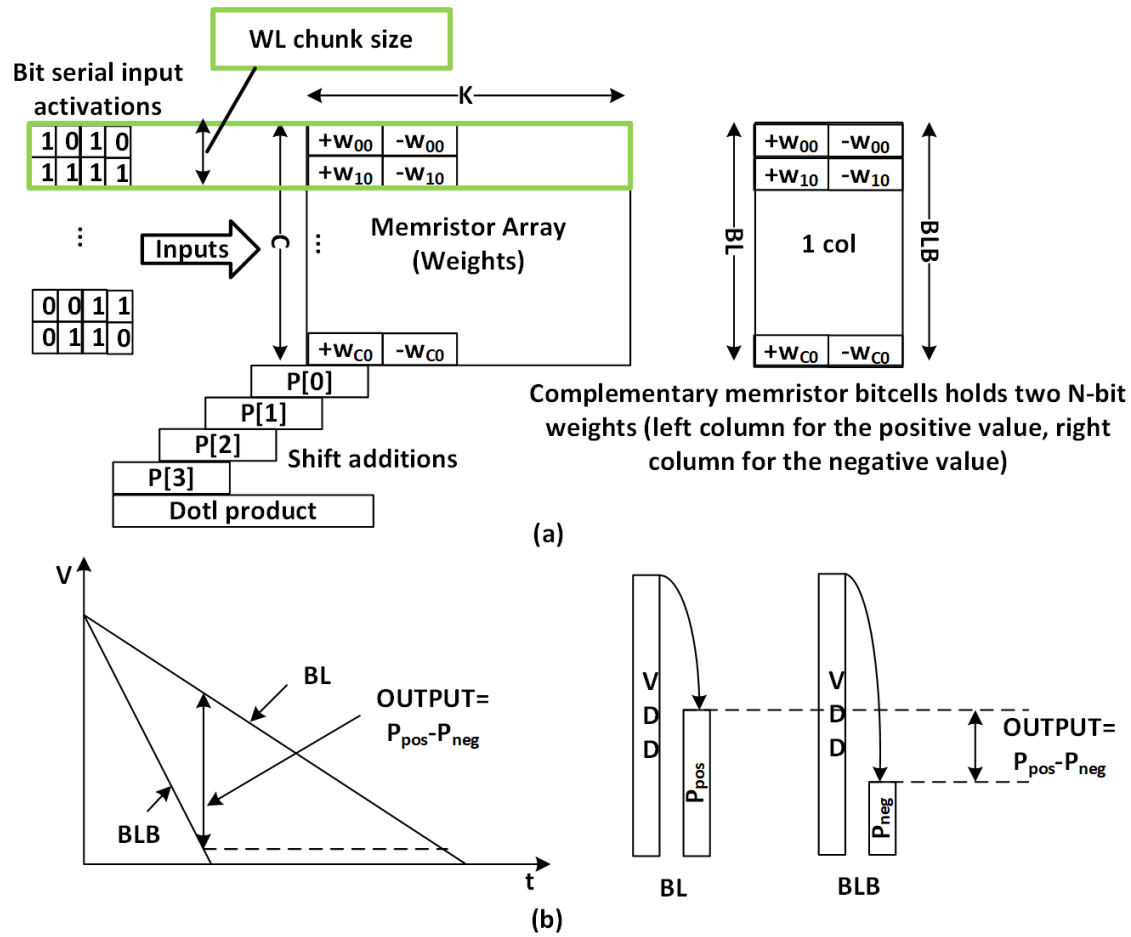


Figure 58: (a) The serial input activation IMC architecture with complementary memristors. The left column stores all the positive weights, and the right column stores all the negative weights. (b) BL and BLB discharge during analog computing. The partial dot product value equals to the voltage difference between the BL and BLB.

is contributed by memristor bitcells from this row. If the incoming serial bit is '1', the corresponding WL will be turned on, and one of the complementary memristor bitcell starts to discharge the BL or BLB based on the sign bit of the stored weight. Given that the discharging current is proportional to the bitcell conductance G_{BCL} or G_{BCR} and the WL acts like the turn on or turn off switch, the amount of charge discharged by a memristor pair is described in equation 17 and 18.

$$Q_{BL}[i] = \sum_{n=1}^{WL\ chunk} V * G_{BCL} * t * Activation[i][n] \tag{17}$$

$$Q_{BLB}[i] = \sum_{n=1}^{WL\ chunk} V * G_{BCR} * t * Activation[i][n] \tag{18}$$

The WL chunk size is defined as the number of WLs being turned-on at the same clock cycle, and it means bitcells in multiple rows can discharge the BLs at the same time. The WL chunk size can range from one to the channel depth C. This feature should be mentioned because it affects the computation noise and the required ADC resolution. In an extreme case that only one WL is turned on during the analog computation, the ADC resolution can be low because it only needs to differentiate different voltage levels resulting from one memristor bitcell. However, a larger WL chunk size is helpful to amortize the ADC energy consumption while introducing a larger quantization error.

Figure 58 (b) demonstrates the timing diagram of BL and BLB during discharging. The amount of voltage drop is decided by Q_{BL} and Q_{BLB} in equation 17 and 18. The voltage difference between BL and BLB is equivalent to the vector partial dot product P[i] of the input activations and the weights. The last step is shift additions of all the

partial dot products corresponding to the serial input bits, as illustrated in equation 20.

$$P[i] = P_{pos}[i] - P_{neg}[i] = ADC(Q_{BL}[i]) - ADC(Q_{BLB}[i]) \tag{19}$$

$$P = \sum_{i=1}^{len(act)} P[i] * 2^{i-1} \tag{20}$$

In summary, dot products of the input activation vectors and the weight vectors can be calculated by the serial input memristor IMC micro-architecture presented in Figure 58. The analog computing happens on the BLs and BLBs by discharging charge proportional to the dot product of the inputs and the weights. Analog partial dot products are quantized by ADCs and stored in the digital format. Shift additions of the digital partial dot products generate the final dot product value.

5.3.3 SRAM IMC Micro-Architecture

The SRAM IMC micro-architecture is very similar to that of the memristor, and the differences include:

- SRAM bitcells can only store binary or ternary weights.
- SRAM bitcells are organized using the complementary structure. If the weight is '+1', the node Q is in the high voltage state and QB is in the low voltage state. If the weight is '-1', the node QB is in the high voltage state and Q is in the low voltage state. In one extreme case that all the weights are '+1' in one column, only the BLB will be discharged, and the positive partial dot product can be represented by the voltage difference between the BL and the BLB.

5.4 IMC Accuracy Loss Modeling

Goals of IMC modeling are predicting the IMC accuracy loss and estimating the energy efficiency of IMC micro-architectures for both memristors and SRAMs. As described in section 5.3, two main sources of IMC noise are the memory bitcell variation which affects the accuracy of weights, and the ADC quantization error which affects the accuracy of the digitized partial dot products. The input activations do not contribute to the accuracy loss benefited from the serial input structure.

5.4.1 Bitcell Variations in IMC

Before introducing the details of IMC modeling, the following assumptions should be made:

- All memristor bitcells are monotonic with post-silicon calibration. The Random variation of memristor bitcells follows an uniform distribution with two parameters representing the lower and the upper boundaries.
- The random variation of SRAM bitcells follows a normal distribution.

Figure 59 (a) illustrates the approach of inserting errors into the memory bitcells. An array of randomly generated error values following the predefined distribution are added to the stored weights. Figure 59 (b) presents that the discharging traces of the BL and the BLB fall into a range defined by the two red dotted line because of the uncertainty introduced by the randomly generated errors. The final partial dot product deviates from the correct value by value of the accumulated random errors. For a larger WL chunk size, the analog MAC operation is affected by more bitcell variations, which results in a noisier partial dot product output.

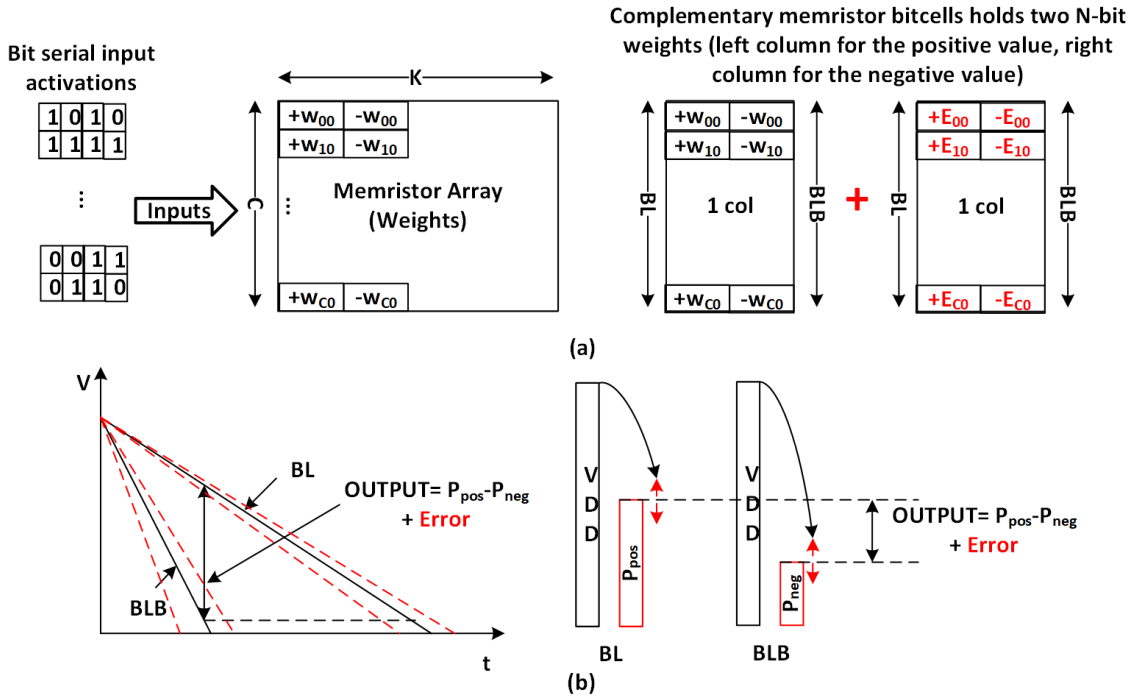


Figure 59: (a) The weight errors resulting from the random variation are added to the stored weight values. The memristor bitcell error follows an uniform distribution, and the SRAM bitcell error follows a normal distribution. (b) The BL and BLB discharging slopes are affected by the added errors. The partial dot product value includes an error item.

5.4.2 ADC Quantization Error in IMC

The ADCs are used to convert the analog partial dot products to digital values. Assume that the non-linearity of ADCs can be eliminated by calibration, the digitized partial dot product values cannot be the same as the analog values due to the limited resolution of ADCs. An ADC quantization error is added to the partial product by rounding the digitized value to the closest ADC reading, as presented in Figure 60. For the serial input scheme, the minimum ADC resolution for no quantization error can be described by $log2(WL\,chunk\,size) + Weight\,bit\,length - 1$, and a larger WL chunk size tends to require higher resolution ADCs.

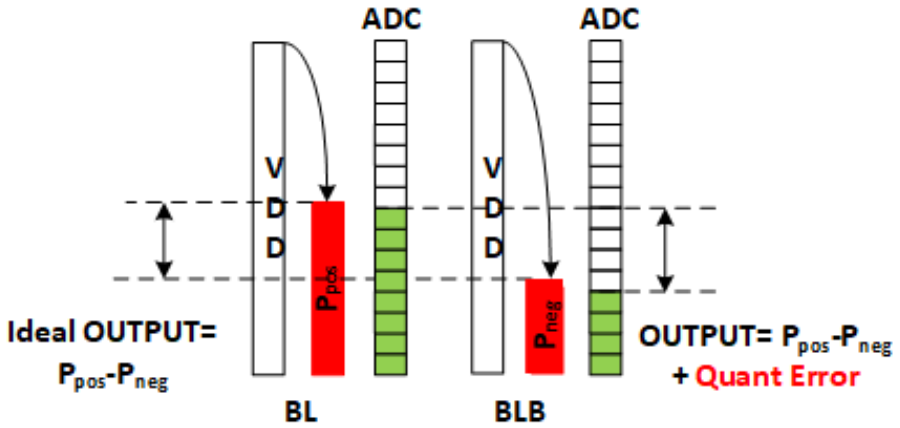


Figure 60: The ADC quantization error is introduced by rounding the partial dot products to the closest ADC readings.

5.4.3 Numerical Modeling for IMC

This work employs numerical models to learn the statistic behavior of the IMC accuracy loss. The key steps are:

- Use mathematical models to represent the input activations, the weights, and the IMC circuit related features.
- Randomly generate the input activations, the weights, and the bitcell errors with values sampled from the mathematical models.
- Calculate two dot products with the values generated in the previous step. One represents the ideal case without any non-ideal circuit behavior, and the other represents the actual situation considering effects of the bitcell error and the ADC quantization error.
- Repeat the previous steps for a large sample of the dot products and learn the statistical behavior.

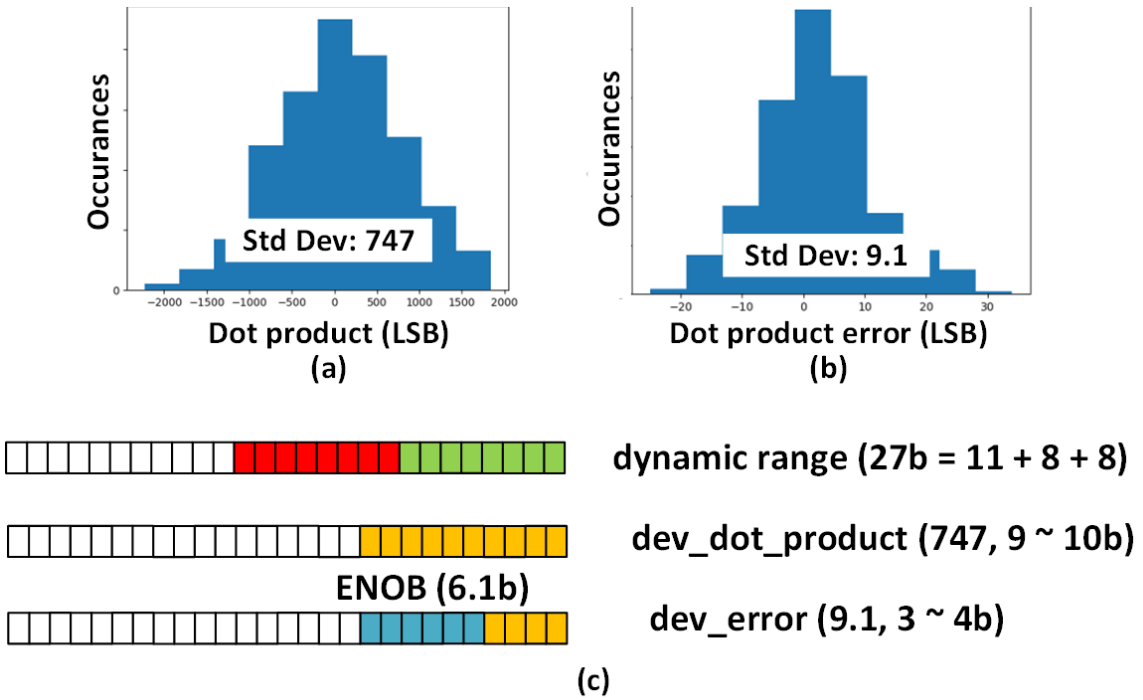


Figure 61: IMC model with a dot product volume of 2048 and a WL chunk size of 4. (a) The histogram of the dot product (the standard deviation is 747 LSB). (b) The histogram of the dot product error (the standard deviation is 9.1 LSB), (c) the intuitive explanation of ENOB (6.1b) calculated from the two distributions

For an actual DNN, the input activations usually follow a normal distribution before the rectified linear unit (ReLu) layer. Based on the observations of input activations, a good estimation is a normal distribution with the mean of 0 and the standard deviation of $2^{N-1}/3$, where N is the bit length of input activations. A ReLu filter is applied to replace all the negative values with zeros for the input activations. The weights are sampled from the same normal distribution, then a certain percent of weights are randomly replaced with zeros.

Figure 61 (a) and (b) demonstrate distributions of the dot products and the dot product error. Both of them are generated with the following assumption on the IMC micro-architecture:

- Input: 8-bit, normally distributed with the standard deviation of 128/3

- Weight: 8-bit, normally distributed with the standard deviation of 128/3
- Sparsity: 0.4
- WL chunk size: 4
- Bitcell error: +/-0.5 LSB with calibration
- ADC precision: 10b
- Dot product volume: 2048

The dot products and the dot product errors are both normally distributed, and the SNR can be calculated with their standard deviation as presented in equation 21. The ENOB is borrowed from the ADC design to represent that how much of information is left in the distribution of dot products. According to equation 22, the ENOB is 6.1 bit given that the standard deviation of dot products is 736 and the standard deviation of dot product errors is 7.9. Figure 61 (c) provides an intuitive approach to understand the ENOB of IMC circuits. The first row standards for the total dynamic range of the dot product, which is the maximum possible value. The yellow squares in the second row are the standard deviation of dot products in the binary format, which is roughly 9 to 10 bits. The yellow squares in the third row are the standard deviation of dot product errors in the binary format, which is roughly 3 to 4 bits. The valid information left in the dot product distribution is approximately the difference between the two yellow squares, which agrees with the ENOB of 6.1 bits.

$$\text{SNR} = 20*\log10(\text{dev\_dot\_product}/\text{dev\_error}) \tag{21}$$

$$\text{ENOB\_SNR} = (\text{SNR}-1.8)/6.02 = 6.1\text{b} \tag{22}$$

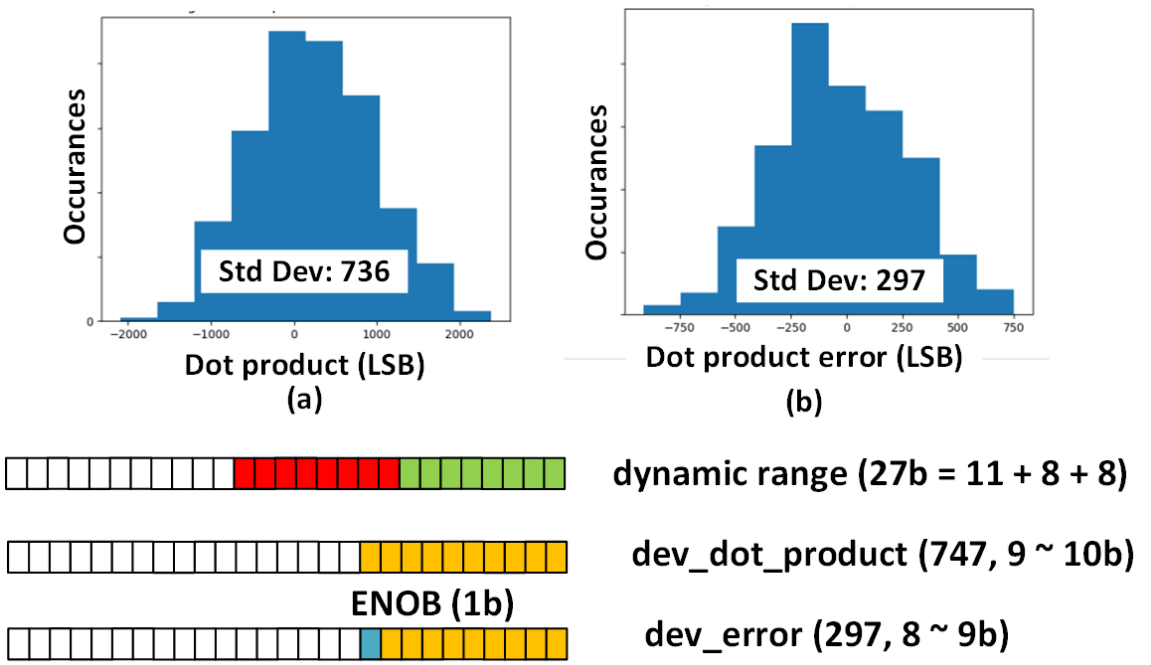


Figure 62: IMC model with a dot product volume of 2048 and a WL chunk size of 128. (a) The histogram of the dot product (the standard deviation is 736 LSB). (b) The histogram of the dot product error (the standard deviation is 297 LSB), (c) The intuitive explanation of ENOB (1b) calculated from the two distributions.

Figure 62 presents another example of the IMC model by changing the WL chunk size to be 128. As discussed before, an increase of the WL chunk size results in the accumulation of bitcell random variations and requires a higher ADC resolution. The standard deviation of dot products remains similar compared to that in Figure 62, but the standard deviation of errors dramatically increases to 297. The ENOB is 1b, which means only 1 bit of information in the dot product distribution is valid.

In summary, the ENOB is employed as a statistical metric to represent the accuracy loss of IMC. For one DNN layer, the dot products and the dot product errors are usually normal distributed, and the ENOB can be calculated from their standard deviations.

5.4.4 IMC Modeling with SRAM Bitcell Variations

As discussed in section 5.4.1, the SRAM bitcell current (Iread) follows a normal distribution, and the simulated means and standard deviations of Iread are provided in Table 9. The ratio of mean and sigma is utilized to randomly generate the weight errors in the SRAM bitcells. The same numerical modeling method in section 5.4.3 is employed for modeling the accuracy loss of SRAMs.

WL voltage (V)	Mean of Iread (μA)	Sigma of Iread (μA)	Mean/Sigma
0.72	43.6	2.0	21.8
0.62	30.6	2.0	15.3
0.52	17.4	1.8	9.7
0.42	6.6	1.2	5.5
0.32	1	0.4	2.5

Table 9: Iread mean and sigma of a SRAM bitcell at different voltages

5.4.5 Parameter Sweeping for IMC Modeling

Figure 63 demonstrates the ENOBs calculated from the accuracy loss model with sweeping the WL chunk size and the weight bit length. The ENOB of the 8-bit weight is larger than that of the 2-bit weight because more information is stored in a larger

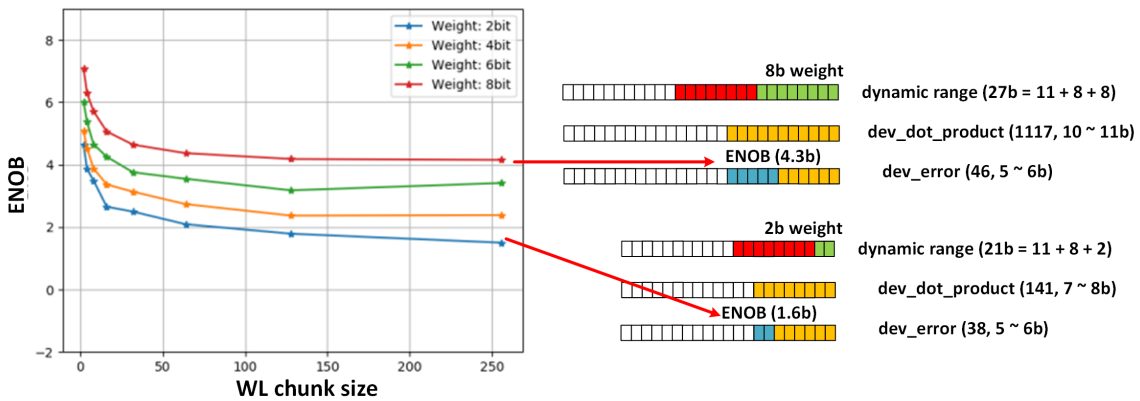


Figure 63: IMC ENOB model with varied WL chunk size and varied weight bit length

weight bit length. It is interesting that the ENOB degradation is not significant when the WL chunk size is larger than 64. The reason is that the accuracy loss is dominated by the accumulated bitcell variations of the entire dot product volume. Based on the ADC survey in [75], resolutions of the most energy efficient ADCs are between 8 bits and 10 bits, so a WL chunk size of 64 and above is desired for amortizing the ADC energy because the bitcell variation still dominates the accuracy loss compared to the ADC quantization error. The IMC parameters, like the bitcell variation value, the input activation length, the ADC resolution, and the dot product volume are swept to obtain the ENOB of each case. These ENOB values are stored in a look-up table for error injections in the following DNN experiments.

5.5 IMC Experiments on Deep Neural Networks

5.5.1 Error Injection in DNN layers

LeNet-5 and the MNIST database for hand-written digits recognition are the most frequently studied DNN in the previous IMC publications, so this work utilizes the same network for the apple to apple comparison.

In the DNN experiments, a lumped error is sampled from different distributions for each DNN layer, and it is injected into each of the output activations as Figure 64 presents. The error is sampled from a normal distribution with a mean of 0 and a standard deviation of the dot product errors. The standard deviation of errors equals to the standard deviation of the dot products divided by the SNR of the IMC micro-architecture. The method of deciding the error distribution is valid because the dot product errors are not correlated with the dot products, which is supported by the fact that their correlation coefficients are close to 0, as Figure 65 reveals.

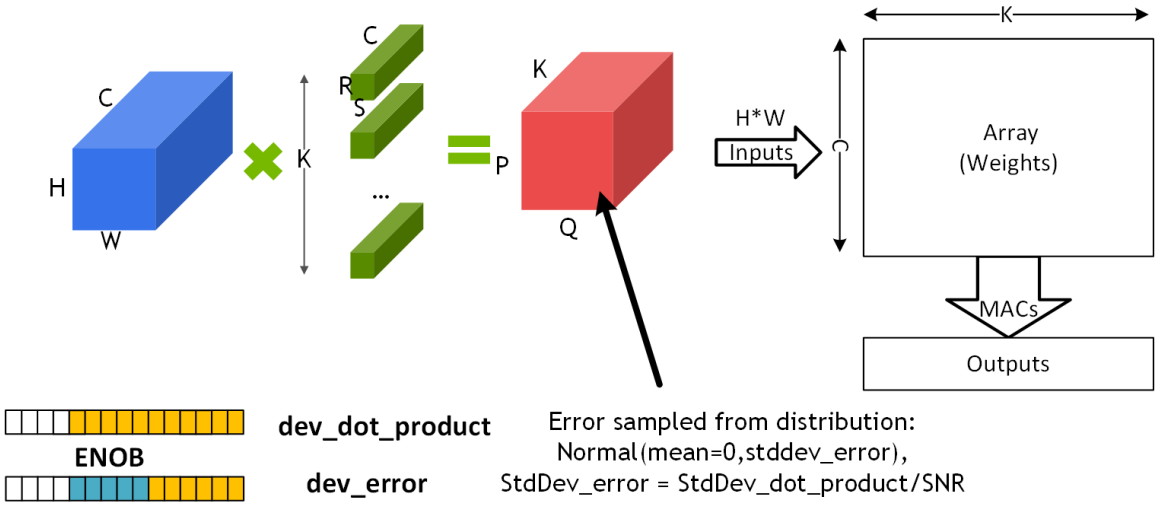


Figure 64: The injected error is sampled from a normal distribution decided by the SNR of the IMC micro-architecture

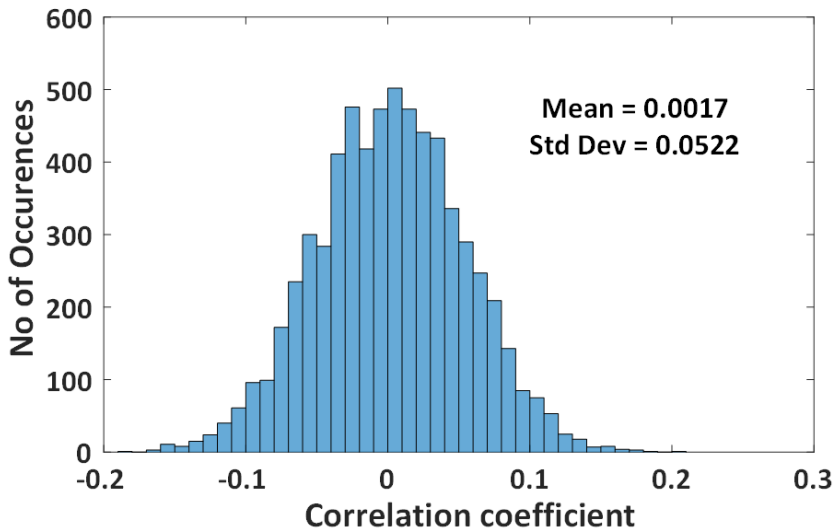


Figure 65: Histogram of the correlation coefficient between the dot products and the dot product errors.

5.5.2 MNIST Precision without DNN Retraining

In Figure 66, the weights are trained with the 32-bit floating point (FP32) resolution, and the recognition precision of hand-written digits is 98.65%. The output

activations are quantized to 4-bit during the DNN inference. The conclusion of the experiment without retraining is that 3-bit weight with error injection equivalent to an ENOB of 3-bit can guarantee less than 1% precision loss compared to the FP32 precision for the hand-written digits recognition on the MNIST dataset. The accuracy drops dramatically with weight less than 3-bit.

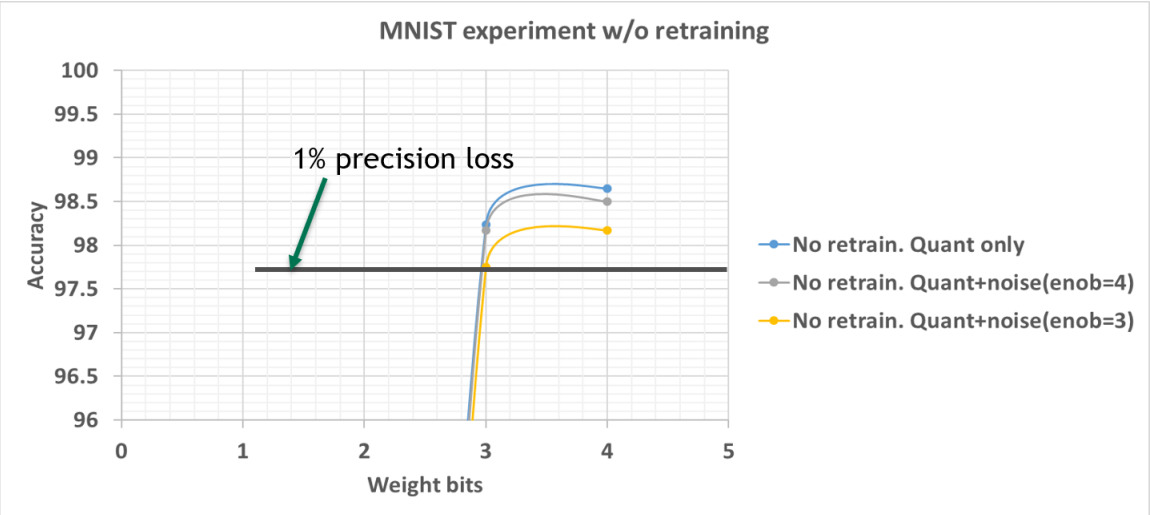


Figure 66: MNIST experiment results of the FP32 pre-trained LeNet-5 with noise injection and quantization

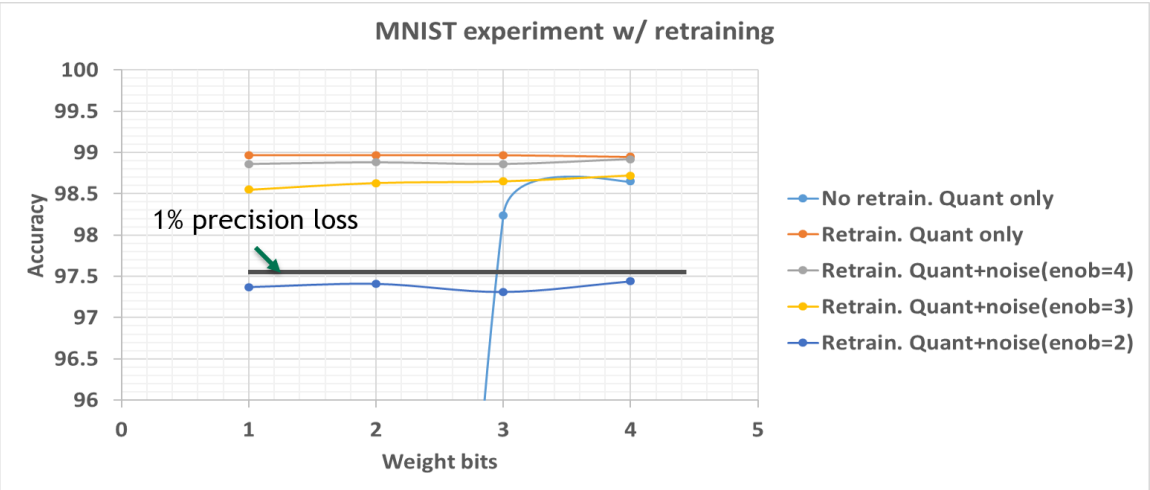


Figure 67: MNIST experiment results of the LeNet-5 retrained with quantization errors

5.5.3 DNN Retraining with Quantization Errors

In Figure 67, the DNN is retrained with 4-bit activations and 4-bit weights, but without noise injections. During inference, the output activations are quantized to 4 bit. The inference precision degradation is less than 1% with 1-bit weights, 4-bit activations with error injection equivalent to an ENOB of 3 bit. Compared to the experiment results without DNN retraining in section 5.5.2, the requirement on the weights is relaxed from 3-bit to 1-bit, which means IMC using SRAMs can meet the requirement.

5.5.4 DNN Retraining with Quantization Errors and Noise Injections

Figure 68 presents that the precision of inference is improved with noise injections during retraining, compared to the DNN retrained with only quantization errors. The requirement on the ENOB of IMC circuit is relaxed from 3-bit to 2-bit in achieving more than 97.5% of recognition precision. It means the IMC circuit can employ more low power techniques.

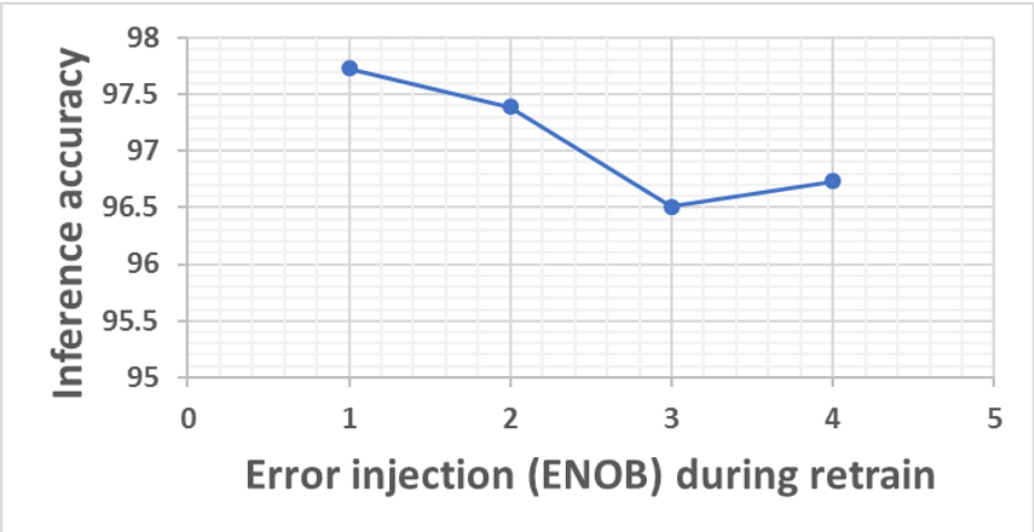


Figure 68: MNIST experiment results of the LeNet-5 retrained with quantization and noise injection

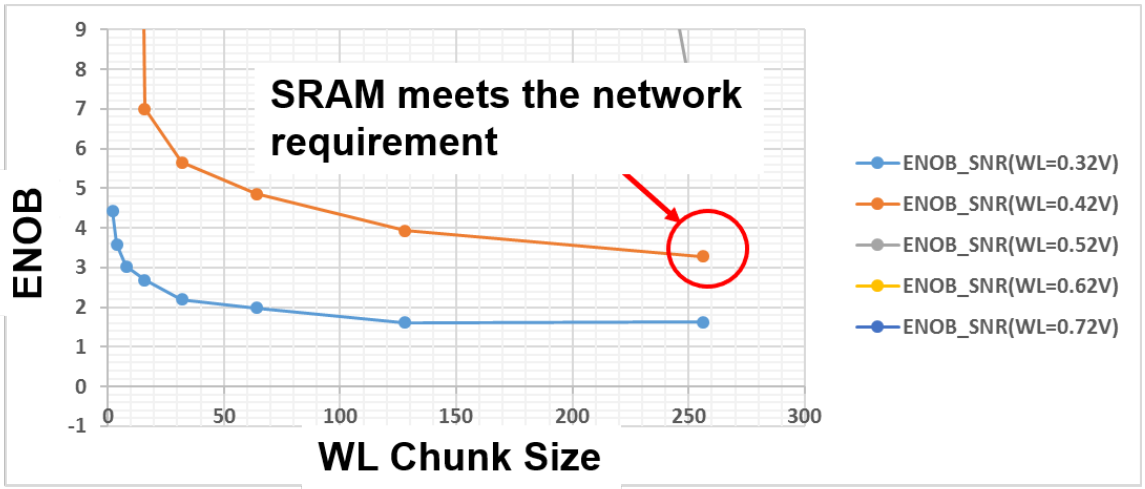


Figure 69: The ENOB of SRAM evaluated by the IMC error model by sweeping the WL chunk size

5.5.5 An SRAM Micro-Architecture for LeNet-5

Network retraining on LeNet demonstrates that the hand-written-digit recognition precision can be 97.7% using binary weights, 2-bit activations, and error injection equivalent to a 2-bit ENOB. According to the IMC error model in Figure 69, an SRAM working at 0.42 V with Iread of 6.2μA, an 8-bit ADC, and a WL chunk size of 256 meets the ENOB requirement of 2 bit. The estimated IMC energy per MAC using this SRAM is about E_{ADC}*4/256 + 0.42V*6.2A*1ns = E_{ADC}/64 + 2.5fJ. Based on the ADC survey in [75], the most energy efficient ADCs consumes 2 pJ per sample with the resolution between 8-bit and 10-bit, so the energy per MAC is roughly 30 fJ. In comparison, a recently published work [72] reports a lower recognition precision of 96% and a higher energy per MAC of 72 fJ using the same DNN LeNet on the same MNIST dataset.

5.6 Conclusions

DNNs are proved to be successful in many AI applications, but their computation power in the von Nuemann architectures are too high to be applicable in the energy deficient IoT SoCs. IMC provides an opportunity for ULP DNN computation by dramatically reducing the amount of data movements and by enabling energy efficient analog MAC computations. The challenge of IMC is to manage the DNN precision degradation resulting from the noise introduced by the memory bitcell variations and the ADC quantization errors. Fortunately, DNN is an error-tolerable algorithm which can minimize the impact of unbiased random noise by retraining the weights of neural networks. Design decisions, such as operating voltage and WL chunk size, provides possibilities to further lower the power consumption of IMC circuits, but their impact on the DNN performance is unclear. The IMC accuracy loss model fills the gap between the IMC circuit and the DNN performance. For IMC circuit designers, the model answers the question that to what extent the noise in IMC micro-architectures can be tolerated by the DNN with retraining. The model can also evaluate the available IMC circuits using ENOB, so that the DNN architects are able to choose an appropriate IMC micro-architecture for the required DNN precision.

The main contributions of this work include:

- Builds the IMC accuracy loss model to predict the ENOB of a given IMC micro-architecture, which considers the memory bitcell random noise and the ADC quantization error.

- Employs the top-down methodology to find the most relaxed requirement on memory. To guarantee a hand-written digit recognition precision of 97.7% for the MNIST dataset, a 16 nm SRAM working at 0.42 V with binary weights, 8-bit ADCs, and 2-bit input activations is required. The estimated energy per

MAC is $E_{ADC}/64 + 2.5fJ$.

- Proves DNN retraining with error injection and quantization in dot products can alleviate the impact of noise introduced by IMC.
- Separates the analysis of memory and the evaluation of DNN, so that the circuit designers and the DNN system architects can optimize the IMC circuits and architectures separately with considering the impacts between each other.

6 Conclusion

IoT devices capable of sensing, processing, storing, and transmitting data are appearing in every corner of our world. ULP and fully-integrated are appealing features to lower the cost of deploying billions of IoT devices. Greater functionality, such as larger on-chip storage capacity, shorter response latency, faster processing speed, and application specific accelerators are desired within the limited power budget of ULP IoT SoCs.

Chip description (V)	Technology node	Chapter
2 KB subthreshold SRAM	65 nm	Chapter 2
32 KB wide voltage range SRAM	130 nm	Chapter 2
1.05 MHz ROSC	65 nm	Chapter 3
-76 dBm 7.4 nW WURX	130 nm	Chapter 4
-106 dBm 33 nW WURX	65 nm	Chapter 4

Table 10: Chips taped-out in each chapters

This dissertation shows a broad interests in the field of ULP IC components design. Four IC components are studied because they are critical circuit blocks for achieving ULP operations of IoT SoCs and they can employ different low power techniques to effectively reduce the system power consumption. The chips taped-out related to this dissertation are listed in Table 10. SRAM is a significant source of leakage power, so subthreshold operation is useful. WURX is helpful to enable duty cycling of the IoT devices and wakes up the IoT node only when necessary. High-speed clock reference for radio block with a quick start-up time can be power-gated to reduce the idle current. DNN hardware accelerator implemented by IMC with proper DVFS can effectively reduce the active power consumption. Low power techniques like DVFS, power gating, and duty cycling are widely used in circuit designs, but they also possibly cause reliability issues.

To guarantee reliable operations, we intensively involve circuit modeling tech-

niques in this dissertation. The SRAM failure modeling enables the efficient BER analysis of all kinds of SRAM failures by reducing the simulation time by 10,000 times. The on-chip relaxation oscillator stability modeling guides the direction for improving the temperature and supply sensitivity, and it achieves one of the best temperature stability among all the ROSCs. The wake-up code modeling predicts the WURXs sensitivity improvements and helps with the selection of wake-up codes and error tolerance. The WURX system achieves the best sensitivity among all the sub-10 nW WURXs. The IMC accuracy loss modeling assists in evaluating the noise in IMC circuits with ENOB for the first time, choosing appropriate memory micro-architectures, and predicting the impact of IMC accuracy loss on the performance of DNNs.

The main contributions are summarized below:

SRAM bitcell auto-generation flow and design space exploration tool.

- Proposes a technology-agnostic subthreshold SRAM bitcell auto-generation flow that explores design knobs in the hyperdimensional design space, for example, the bitcell sizes, bitcell types, and assist techniques.

- Categorizes the SRAM failure mechanisms into read data disturbance, HS data disturbance, read timing failure, write timing failure, and non-write-able bitcells.

- Utilizes appropriate metrics to evaluate the different failures accurately and efficiently. The inverse RTcrit and WTcrit can be used to calculate the BER of read and write timing failures. The static WM can be used to calculate the BER of non-write-able failures. The dynamic read disturbance with importance sampling can be used to calculate the BER of read data disturbance.

- Improves the importance sampling technique to estimate the read data distur-

bance BER by considering only the dominating parameters, with a simulation time reduction of 10,000x compared to the conventional Monte Carlo simulations.

- The ViPro explored the methodologies of multi-port register file design with the built-in models of memory macro delay and energy.

On-chip relaxation oscillator

- Achieves the XO comparable temperature stability of 100 ppm from 0 to 40^oC for IoT SoCs functioning in the temperature range of human body.
- Enables power gating for reducing the active power and 8 μS rapid start-up time without degrading stability.
- The ROSC chip taped-out in the 65 nm technology achieves 69 μW active power at 1.05 MHz and 1 V, and 110 nW leakage power in the power gating mode.
- Employes two circuit models to guide directions for improving the temperature and supply sensitivity.

Wake-up code analysis

- The 7.4 nW WURX system [55] taped-out in the 130nm technology achieves a -76 dBm sensitivity, less than 0.1% missed detection, and less than 1 FWU per hour using a baseband correlator circuit with 8-bit wake-up code and 1-bit error tolerance.
- The wake-up code analysis model helps choosing the appropriate correlator code length, the code weight, and the error tolerance.

- The wake-up code analysis model predicts that a longer wake-up code (longer RF transmission time) improves the sensitivity and consumes more RF energy under the assumption of the same RF power.

- The available wake-up code set is larger with a longer wake-up code length, but it degrades the sensitivity with the same RF energy. Error tolerance of the correlator can restore the sensitivity of 31-bit codes to be less than 1 dB compared to 8-bit codes.

In-memory computing modeling and DNN study

- Builds the IMC accuracy loss model to predict the ENOB of a given IMC micro-architecture, which considers the memory bitcell random noise and the ADC quantization error.

- Employs the top-down methodology to find the most relaxed requirement on memory. To guarantee a hand-written digit recognition precision of 97.7% for the MNIST dataset, a 16 nm SRAM working at 0.42 V with binary weights, 8-bit ADCs, and 2-bit input activations is required. The estimated energy per MAC is $E_{ADC}/64$ + 2.5fJ.

- Proves DNN retraining with error injection and quantization in dot products can alleviate the impact of noise introduced by IMC.

- Separates the analysis of memory and the evaluation of DNN, so that the circuit designers and the DNN system architects can optimize the IMC circuits and architectures separately with considering the impacts between each other.

Appendix A List of Publications

A.1 Publications

- **N. Liu**, Agarwala, A. Dissanayake, D. S. Truesdell, S. Kamineni, X. Chen, D. D. Wentzloff, and B. H. Calhoun, "A 2.5 ppm/C 1.05 MHz Relaxation Oscillator with Dynamic Frequency-Error Compensation and 8 μs Start-up Time," ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC), Dresden, 2018, pp. 150-153.
- Kosari A, Breiholz J, **N. Liu**, et al. A 0.5 V 68 nW ECG Monitoring Analog Front-End for Arrhythmia Diagnosis[J]. Journal of Low Power Electronics and Applications, 2018, 8(3): 27.
- J. Moody, P. Bassirian, A. Roy, **N. Liu**, S. Pancrazio, N. S. Barker, B. H. Calhoun, S. M. Bowers, A -76dBm 7.4nW Wakeup Radio with Automatic Offset Compensation, ISSCC, 2018
- Yahya, F., C. J. Lukas, J. Breiholz, A. Roy, H. N. Patel, **N. Liu**, X. Chen, A. Kosari, S. Li, D. Akella, et al., "A battery-less 507nW SoC with integrated platform power manager and SiP interfaces", Symposium on VLSI Circuits, 2017
- Banerjee, A., **N. Liu**, H. N. Patel, and B. H. Calhoun, and etc. "A 256kb 6T self-tuning SRAM with extended 0.38V1.2V operating range using multiple read/write assists and VMIN tracking canary sensors", IEEE CICC, 2017
- Patel, H. N., A. Roy, F. B. Yahya, **N. Liu**, B. H. Calhoun, "A 55nm Ultra Low Leakage Deeply Depleted Channel Technology Optimized for Energy Minimization in Subthreshold SRAM and Logic", ESSCIRC, 2016
- **N. Liu**, and B. H. Calhoun, "Design Optimization of Register File Throughput and Energy using a Virtual Prototyping (ViPro) Tool", IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 2016

A.2 Pending Publications

- **N. Liu**, Agarwala, A. Dissanayake, D. S. Truesdell, S. Kamineni, and B. H. Calhoun, "A 2.5 ppm/C 1.05 MHz Relaxation Oscillator with Dynamic Frequency-Error Compensation and 8 μs Start-up Time," submitted to JSSC.

- **N. Liu**, S. Kamineni, and B. H. Calhoun, "A technology agnostic subthreshold SRAM bitcell auto-generation flow for ultra-low power applications, " to be submitted.
- **N. Liu**, B. H. Calhoun, and etc. "Design Exploration and Accuracy Loss Modeling of In-memory Computing for Deep Neural Networks, ", to be submitted.
- **N. Liu**, P. Bassirian, J. Moody, S. M. Bowers, and B. H. Calhoun, "Explorations on RF Backend Design and Coding for Ultra-Low Power WURXs," to be submitted

Appendix B Glossary of Terms

- AI - Artificial Intelligence
- BL - Bit Line
- BER - Bit error rate
- CMOS - complimentary metal oxide semiconductor
- CDF - Cumulative distribution function
- CPU - Central processing unit
- DNN - Deep Neural Network
- DRAM - Dynamic random access memory
- DVFS- Dynamic Voltage Frequency Scaling
- GPU - Graphics processing unit
- IC - Integrated Circuit
- IMC - In memory computing
- IoT - Internet-of-Things
- MPFP - Most probable failure point
- NMOS - N-type metal oxide semiconductor
- PMOS - P-type metal oxide semiconductor
- RF - Radio Frequency
- RO - Ring Oscillator
- ROSC - Relaxation osicllator
- RSNM - Read static noise margin
- RTcrit - Read critical time
- SA - sense amp
- SoC - System on chip
- SRAM - Static random access memory
- SNM - Static noise margin

- TPU - Tensor processing unit
- ULP - Ultra-Low-Power
- WL - Word Line
- WM - Write static noise margin
- WTcrit - Write critical time
- WURX - Wake-up Receiver
- XO - crystal oscillator
- V_{DD} - supply voltage
- Vipro - Virtual Prototyping
- VT - threshold voltage

References

[1] Yahya F, Lukas C J, Breiholz J, et al. A battery-less 507nW SoC with integrated platform power manager and SiP interfaces[C]//VLSI Circuits, 2017 Symposium on. IEEE, 2017: C338-C339.

[2] Lee Y, Blaauw D, Sylvester D. Ultralow power circuit design for wireless sensor nodes for structural health monitoring[J]. Proceedings of the IEEE, 2016, 104(8): 1529-1546.

[3] Piyare R, Murphy A L, Kiraly C, et al. Ultra Low Power Wake-Up Radios: A Hardware and Networking Survey[J]. IEEE Communications Surveys Tutorials, 2017, 19(4): 2117-2157.

[4] Klinefelter A, Roberts N E, Shakhsheer Y, et al. 21.3 A 6.45 μW self-powered IoT SoC with integrated energy-harvesting power management and ULP asymmetric radios[C]//Solid-State Circuits Conference-(ISSCC), 2015 IEEE International. IEEE, 2015: 1-3.

[5] Kim H, Kim S, Van Helleputte N, et al. A configurable and low-power mixed signal SoC for portable ECG monitoring applications[J]. IEEE transactions on biomedical circuits and systems, 2014, 8(2): 257-267.

[6] Chang M F, Chen C F, Chang T H, et al. 17.3 A 28nm 256kb 6T-SRAM with 280mV improvement in V MIN using a dual-split-control assist scheme[C]//Solid-State Circuits Conference-(ISSCC), 2015 IEEE International. IEEE, 2015: 1-3.

[7] Y. Tsukamoto, K. Nii, S. Imaoka, Y. Oda, S. Ohbayashi, T. Yoshizawa, H. Makino, K. Ishibashi, and H. Shinohara. 2005. Worst-case analysis to obtain stable read-/write DC margin of high density 6T-SRAM-array with local Vth variability. In

Proceedings of the 2005 IEEE/ACM International conference on Computer-aided design (ICCAD '05). IEEE Computer Society, Washington, DC, USA, 398-405

[8] F. B. Yahya, H. N. Patel, V. Chandra and B. H. Calhoun, "Combined SRAM read/write assist techniques for near/sub-threshold voltage operation," 2015 6th Asia Symposium on Quality Electronic Design (ASQED), Kula Lumpur, 2015, pp. 1-6.

[9] A. Banerjee, N. Liu, H. N. Patel, B. H. Calhoun, J. Poulton and C. T. Gray, "A 256kb 6T self-tuning SRAM with extended 0.38V1.2V operating range using multiple read/write assists and VMIN tracking canary sensors," 2017 IEEE Custom Integrated Circuits Conference (CICC), Austin, TX, 2017, pp. 1-4.

[10] Nalam S, Bhargava M, Mai K, et al. Virtual prototyper (ViPro): An early design space exploration and optimization tool for SRAM designers[C]//Proceedings of the 47th Design Automation Conference. ACM, 2010: 138-143.

[11] Seevinck E, List F J, Lohstroh J. Static-noise margin analysis of MOS SRAM cells[J]. IEEE Journal of solid-state circuits, 1987, 22(5): 748-754.

[12] J. Wang, S. Nalam and B. H. Calhoun, "Analyzing static and dynamic write margin for nanometer SRAMs," Proceeding of the 13th international symposium on Low power electronics and design (ISLPED '08), Bangalore, 2008, pp. 129-134.

[13] B. H. Calhoun and A. P. Chandrakasan, "Static noise margin variation for sub-threshold SRAM in 65-nm CMOS," in IEEE Journal of Solid-State Circuits, vol. 41, no. 7, pp. 1673-1679, July 2006.

[14] D.J. Frank, Y. Taur, M. Ieong, H.P. Wong, Monte Carlo Modeling of Threshold Variation due to Dopant Fluctuations, Symp. VLSI Technology,1999.

[15] Dolecek L, Qazi M, Shah D, et al. Breaking the simulation barrier: SRAM evaluation through norm minimization[C]//Proceedings of the 2008 IEEE/ACM International Conference on Computer-Aided Design. IEEE Press, 2008: 322-329.

[16] Boley, James, et al. "Leveraging sensitivity analysis for fast, accurate estimation of SRAM dynamic write V MIN." Proceedings of the Conference on Design, Automation and Test in Europe. EDA Consortium, 2013.

[17] A. Singhee and R. Rutenbar, Statistical blockade: a novel method for very fast Monte Carlo simulation of rare circuit events, and its application, DATE, 2007.

[18] B. Zimmer et al., "SRAM Assist Techniques for Operation in a Wide Voltage Range in 28-nm CMOS," in IEEE Transactions on Circuits and Systems II: Express Briefs, vol. 59, no. 12, pp. 853-857, Dec. 2012.

[19] M. Qazi, M. Tikekar, L. Dolecek, D. Shah and A. Chandrakasan, "Loop flattening spherical sampling: Highly efficient model reduction techniques for SRAM yield analysis," 2010 Design, Automation Test in Europe Conference Exhibition (DATE 2010), Dresden, 2010, pp. 801-806.

[20] J. Boley, P. Beshay and B. Calhoun, "Virtual Prototyper (ViPro): An SRAM Design Tool for Yield Constrained Optimization," in IEEE Transactions on Very Large Scale Integration (VLSI) Systems, vol. 23, no. 12, pp. 3109-3113, Dec. 2015.

[21] N. Liu and B. Calhoun, "Design Optimization of Register File Throughput and Energy Using a Virtual Prototyping (ViPro) Tool," 2016 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), Pittsburgh, PA, 2016, pp. 535-540.

[22] Agarwal K, Nassif S. Statistical analysis of SRAM cell stability[C]//Proceedings of the 43rd annual Design Automation Conference. ACM, 2006: 57-62.

[23] P. Chang; T. Lin; J. Wang; Y. Yu, "A 4R/2W Register File Design for UDVS Micro-processors in 65-nm CMOS," Circuits and Systems II: Express Briefs, IEEE Transactions on , vol.59, no.12, pp.908,912, Dec. 2012.

[24] Nalluri, R.; Garg, R.; Panda, P.R., "Customization of Register File Banking Architecture for Low Power," VLSI Design, 2007. Held jointly with 6th International Conference on Embedded Systems., 20th International Conference on, vol., no., pp.239,244, 6-10 Jan. 2007.

[25] Kulkarni, J.P.; Tokunaga, C.; Aseron, P.; Nguyen, T.; Augustine, C.; Tschanz, J.; De, V., "4.7 A 409GOPS/W adaptive and resilient domino register file in 22nm tri-gate CMOS featuring in-situ timing margin and error detection for tolerance to within-die variation, voltage droop, temperature and aging," Solid- State Circuits Conference - (ISSCC), 2015 IEEE International , vol., no., pp.1,3, 22-26 Feb. 2015.

[26] Kurd, N.A.; Bhamidipati, S.; Mozak, C.; Miller, J.L.; Wilson, T.M.; Nemani, M.; Chowdhury, M., "Westmere: A family of 32nm IA processors," Solid-State Circuits Conference Digest of Technical Papers (ISSCC), 2010 IEEE International , vol., no., pp.96,97, 7-11 Feb. 2010.

[27] Donkoh, E.; Lowery, A.; Shriver, E., "A hybrid and adaptive model for predicting register file and SRAM power using a reference design," Design Automation Conference (DAC), 2012 49th ACM/EDAC/IEEE , vol., no., pp.62,67, 3-7 June 2012.

[28] Nalam, S.; Bhargava, M.; Ken Mai; Calhoun, B.H., "Virtual prototyper (ViPro): An early design space exploration and optimization tool for SRAM designers," Design Automation Conference (DAC), 2010 47th ACM/IEEE , vol., no., pp.138,143, 13-18 June 2010

[29] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, Cacti 5.1, HP Laboratories

[30] Kawasumi, A.; Suzuki, T.; Moriwaki, S.; Miyano, S., "Energy efficiency degradation caused by random variation in low-voltage SRAM and 26% energy reduction by Bitline Amplitude Limiting (BAL) scheme," Solid State Circuits Conference (A-SSCC), 2011 IEEE Asian , vol., no., pp.165,168, 14-16 Nov. 2011.

[31] Patwary, A.R.; Greub, H.; Zhongfeng Wang; Geuskens, B.M., "Bit-Line Organization in Register Files for Low-Power and High-Performance Applications," Electrical and Computer Engineering, 2006. ICECE '06. International Conference on , vol., no., pp.505,508, 19-21 Dec. 2006.

[32] Ishikura, S.; Kurumada, M.; Terano, T.; Yamagami, Y.; Kotani, N.; Satomi, K.; Nii, K.; Yabuuchi, M.; Tsukamoto, Y.; Ohbayashi, S.; Oashi, T.; Makino, H.; Shinohara, H.; Akamatsu, H., "A 45nm 2port 8T-SRAM using hierarchical replica bitline technique with immunity from simultaneous R/W access issues," VLSI Circuits, 2007 IEEE Symposium on , vol., no., pp.254,255, 14-16 June 2007.

[33] A. Savanth, J. Myers, A. Weddell, D. Flynn and B. Al-Hashimi, "5.6 A 0.68nW/kHz supply-independent Relaxation Oscillator with 0.49%/V and 96ppm/C stability," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2017, pp. 96-97.

[34] . Grleyk, L. Pedala, F. Sebastiano and K. A. A. Makinwa, "A CMOS Dual-RC frequency reference with 250ppm inaccuracy from 45C to 85C," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, 2018, pp. 54-56.

[35] G. Zhang, K. Yayama, A. Katsushima and T. Miki, "A 3.2 ppm/C Second-Order Temperature Compensated CMOS On-Chip Oscillator Using Voltage Ratio Adjusting Technique," in IEEE Journal of Solid-State Circuits, vol. 53, no. 4, pp. 1184-1191, April 2018.

[36] Y. Tokunaga, S. Sakiyama, A. Matsumoto and S. Dosho, "An On-Chip CMOS Relaxation Oscillator With Voltage Averaging Feedback," in IEEE Journal of Solid-State Circuits, vol. 45, no. 6, pp. 1150-1158, June 2010.

[37] M. Choi, T. Jang, S. Bang, Y. Shi, D. Blaauw and D. Sylvester, "A 110 nW Resistive Frequency Locked On-Chip Oscillator with 34.3 ppm/C Temperature Stability for System-on-Chip Designs," in IEEE Journal of Solid-State Circuits, vol. 51, no. 9, pp. 2106-2118, Sept. 2016.

[38] K. Hsiao, "A 32.4 ppm/C 3.2-1.6V self-chopped relaxation oscillator with adaptive supply generation," 2012 Symposium on VLSI Circuits (VLSIC), Honolulu, HI, 2012, pp. 14-15.

[39] Y. Cao, P. Leroux, W. D. Cock and M. Steyaert, "A 63,000 Q-factor relaxation oscillator with switched-capacitor integrated error feedback," 2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers, San Francisco, CA, 2013, pp. 186-187.

[40] J. Lee, A. George and M. Je, "5.10 A 1.4V 10.5MHz swing-boosted differential relaxation oscillator with 162.1dBc/Hz FOM and 9.86psrms period jitter in 0.18μm CMOS," 2016 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2016, pp. 106-108.

[41] A. Paidimarri, D. Griffith, A. Wang, G. Burra and A. P. Chandrakasan, "An RC Oscillator With Comparator Offset Cancellation," in IEEE Journal of Solid-State Circuits, vol. 51, no. 8, pp. 1866-1877, Aug. 2016.

[42] J. Wang, W. L. Goh, X. Liu and J. Zhou, "A 12.77-MHz 31 ppm/C On-Chip RC Relaxation Oscillator With Digital Compensation Technique," in IEEE Transactions on Circuits and Systems I: Regular Papers, vol. 63, no. 11, pp. 1816-1824, Nov. 2016.

[43] S. Jeong, I. Lee, D. Blaauw and D. Sylvester, "A 5.8 nW CMOS Wake-Up Timer for Ultra-Low-Power Wireless Applications," in IEEE Journal of Solid-State Circuits, vol. 50, no. 8, pp. 1754-1763, Aug. 2015.

[44] T. Tokairin et al., "A 280nW, 100kHz, 1-cycle start-up time, on-chip CMOS relaxation oscillator employing a feedforward period control scheme," 2012 Symposium on VLSI Circuits (VLSIC), Honolulu, HI, 2012, pp. 16-17.

[45] D. Griffith, P. T. Rine, J. Murdock and R. Smith, "17.8 A 190nW 33kHz RC oscillator with 0.21% temperature stability and 4ppm long-term stability," 2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC), San Francisco, CA, 2014, pp. 300-301.

[46] J. Mikuli, G. Schatzberger and A. Bari, "A 1-MHz on-chip relaxation oscillator with comparator delay cancelation," ESSCIRC 2017 - 43rd IEEE European Solid State Circuits Conference, Leuven, 2017, pp. 95-98.

[47] N. Liu, Agarwala, A. Dissanayake, D. S. Truesdell, S. Kamineni, and B. H. Calhoun, "A 2.5 ppm/C 1.05 MHz Relaxation Oscillator with Dynamic Frequency-Error Compensation and 8 μs Start-up Time," ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC), Dresden, 2018, pp. 150-153.

[48] J. Koo, K. Moon, B. Kim, H. Park and J. Sim, "5.5 A quadrature relaxation oscillator with a process-induced frequency-error compensation loop," 2017 IEEE International Solid-State Circuits Conference (ISSCC), San Francisco, CA, 2017, pp. 94-95.

[49] Y. Tokunaga, S. Sakiyama, A. Matsumoto and S. Dosho, "An On-Chip CMOS Relaxation Oscillator With Voltage Averaging Feedback," in IEEE Journal of Solid-State Circuits, vol. 45, no. 6, pp. 1150-1158, June 2010.

[50] Piyare R, Murphy A L, Kiraly C, et al. Ultra Low Power Wake-Up Radios: A Hardware and Networking Survey[J]. IEEE Communications Surveys Tutorials, 2017, 19(4): 2117-2157.

[51] Jiang, Haowei, et al. "A 4.5 nW wake-up radio with 69dBm sensitivity." Solid-State Circuits Conference (ISSCC), 2017 IEEE International. IEEE, 2017.

[52] Zhang, Yan, et al. "A 3.72 W ultra-low power digital baseband for wake-up radios." VLSI Design, Automation and Test (VLSI-DAT), 2011 International Symposium on. IEEE, 2011.

[53] J. Moody, P. Bassirian, A. Roy, Y. Feng, S. Li, R. Costanzo, N. S. Barker, B. H. Calhoun, S. M. Bowers, An 8.3 nW -72 dBm Event Driven IoE Wake-up Receiver, 2017 European Microwave Integrated Circuits Conference (EuMIC), Nuremberg, Germany, 2017, pp. 1-4.

[54] Milosiu, Heinrich, et al. "A 3-W 868-MHz wake-up receiver with 83 dBm sensitivity and scalable data rate." ESSCIRC (ESSCIRC), Proceedings of the. IEEE, 2013.

[55] J. Moody et al., "A 76dBm 7.4nW wakeup radio with automatic offset compensation," 2018 IEEE International Solid - State Circuits Conference - (ISSCC), San Francisco, CA, 2018, pp. 452-454.

[56] LeCun Y, Jackel L D, Bottou L, et al. Comparison of learning algorithms for handwritten digit recognition[C]//International conference on artificial neural networks. 1995, 60: 53-60.

[57] Schaller R R. Moore's law: past, present and future[J]. IEEE spectrum, 1997, 34(6): 52-59.

[58] https://www.nature.com/news/the-chips-are-down-for-moore-s-law-1.19338

[59] Nvidia, https://www.nvidia.com/en-us/data-center/dgx-saturnv/

[60] F. F. Li et al, CS231n: Convolutional Neural Networks for Visual Recognition. , http://cs231n.github.io/convolutional-networks/.

[61] Y.-H. Chen, J. Emer, and V. Sze, Eyeriss: A Spatial Architecture for Energy-Efficient Dataflow for Convolutional Neural Networks, in ISCA, 2016.

[62] Ambrogio S, Narayanan P, Tsai H, et al. Equivalent-accuracy accelerated neural-network training using analogue memory[J]. Nature, 2018, 558(7708): 60.

[63] Shafiee A, Nag A, Muralimanohar N, et al. ISAAC: A convolutional neural network accelerator with in-situ analog arithmetic in crossbars[J]. ACM SIGARCH Computer Architecture News, 2016, 44(3): 14-26.

[64] Hu M, Strachan J P, Li Z, et al. Dot-product engine for neuromorphic computing: programming 1T1M crossbar to accelerate matrix-vector multiplication[C]//Proceedings of the 53rd annual design automation conference. ACM, 2016: 19.

[65] Yu S, Li Z, Chen P Y, et al. Binary neural network with 16 Mb RRAM macro chip for classification and online training[C]//Electron Devices Meeting (IEDM), 2016 IEEE International. IEEE, 2016: 16.2. 1-16.2. 4.

[66] Ankit A, Sengupta A, Panda P, et al. Resparc: A reconfigurable and energy-efficient architecture with memristive crossbars for deep spiking neural networks[C]//Proceedings of the 54th Annual Design Automation Conference 2017. ACM, 2017: 27.

[67] Mochida R, Kouno K, Hayata Y, et al. A 4M Synapses integrated Analog ReRAM based 66.5 TOPS/W Neural-Network Processor with Cell Current Controlled Writing and Flexible Network Architecture[C]//2018 IEEE Symposium on VLSI Technology. IEEE, 2018: 175-176.

[68] Y. Liao et al., "Novel In-Memory Matrix-Matrix Multiplication with Resistive Cross-Point Arrays," 2018 IEEE Symposium on VLSI Technology, Honolulu, HI, 2018, pp. 31-32.

[69] Lee S T, Lim S, Choi N, et al. Neuromorphic Technology Based on Charge Storage Memory Devices[C]//2018 IEEE Symposium on VLSI Technology. IEEE, 2018: 169-170.

[70] Lu J, Young S, Arel I, et al. A 1 TOPS/W analog deep machine-learning engine with floating-gate storage in 0.13 m CMOS[J]. IEEE Journal of Solid-State Circuits, 2015, 50(1): 270-281.

[71] Zhang J, Wang Z, Verma N. In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array[J]. J. Solid-State Circuits, 2017, 52(4): 915-924.

[72] Biswas A, Chandrakasan A P. Conv-RAM: An energy-efficient SRAM with embedded convolution computation for low-power CNN-based machine learning applications[C]//Solid-State Circuits Conference-(ISSCC), 2018 IEEE International. IEEE, 2018: 488-490.

[73] Liu R, Peng X, Sun X, et al. Parallelizing SRAM arrays with customized bit-cell for binary neural networks[C]//Proceedings of the 55th Annual Design Automation Conference. ACM, 2018: 21.

[74] Jiang L, Kim M, Wen W, et al. Xnor-pop: A processing-in-memory architecture for binary convolutional neural networks in wide-io2 drams[C]//Low Power Electronics and Design (ISLPED, 2017 IEEE/ACM International Symposium on. IEEE, 2017: 1-6.

[75] B. Murmann, "ADC Performance Survey 1997-2018," [Online]. Available: http://web.stanford.edu/ murmann/adcsurvey.html.

[76] Rastegari M, Ordonez V, Redmon J, et al. Xnor-net: Imagenet classification using binary convolutional neural networks[C]//European Conference on Computer Vision. Springer, Cham, 2016: 525-542.

[77] Bankman D, Yang L, Moons B, et al. An always-on 3.8 J/86% CIFAR-10 mixed-signal binary CNN processor with all memory on chip in 28nm CMOS[C]//Solid-State Circuits Conference-(ISSCC), 2018 IEEE International. IEEE, 2018: 222-224.

[78] Yoshioka K, Toyama Y, Ban K, et al. PhaseMAC: A 14 TOPS/W 8bit GRO based Phase Domain MAC Circuit for In-Sensor-Computed Deep Learning Accelerators[C]//2018 IEEE Symposium on VLSI Circuits. IEEE, 2018: 263-264.

[79] Ando K, Ueyoshi K, Orimo K, et al. BRein memory: A single-chip binary/ternary reconfigurable in-memory deep neural network accelerator achieving 1.4 TOPS at 0.6 W[J]. IEEE Journal of Solid-State Circuits, 2018, 53(4): 983-994.