

# Energy Efficient Circuits and System for Internet of Things and Hardware Accelerator Design for Genome Sequencing

by

Xiao Wu

A dissertation submitted in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy  
(Electrical and Computer Engineering)  
in The University of Michigan  
2019

Doctoral Committee:

Professor David Blaauw, Chair  
Assistant Professor Reetuparna Das  
Assistant Professor Hun-Seok Kim  
Professor Dennis Sylvester

Xiao Wu

lydiaxia@umich.edu

ORCID iD: 0000-0001-5731-1000

© Xiao Wu 2019

All Rights Reserved

To my loving husband Zhiyoong Foo,  
my lovely dog Ella,  
and my parents for all your love and care through these years.

# TABLE OF CONTENTS

<b>DEDICATION</b> . . . . .	ii
<b>LIST OF FIGURES</b> . . . . .	v
<b>LIST OF TABLES</b> . . . . .	ix
<b>LIST OF APPENDICES</b> . . . . .	x
<b>ABSTRACT</b> . . . . .	xi
<b>CHAPTER</b>	
<b>I. Introduction</b> . . . . .	1
1.1 Powering Miniaturized IoT Sensor Node . . . . .	2
1.2 Scaling Programmable Sensor Nodes to Sub-mm <sup>3</sup> . . . . .	5
1.3 Extending Portable Computation to Genomics . . . . .	6
1.4 Dissertation Outline . . . . .	7
<b>II. A 20pW Discontinuous Switched-Capacitor Energy Harvester         for Smart Sensor Applications</b> . . . . .	10
2.1 Introduction . . . . .	10
2.2 Proposed Technique: Discontinuous Harvesting . . . . .	14
2.2.1 Discontinuous Harvesting . . . . .	14
2.2.2 Energy efficiency trade-off analysis . . . . .	16
2.3 Implementation of Discontinuous Harvester . . . . .	20
2.3.1 Proposed Architecture . . . . .	20
2.3.2 Moving-Sum Charge Pump . . . . .	25
2.3.3 Automatic Conversion Ratio Modulator . . . . .	32
2.3.4 Low Power Mode Controller . . . . .	33
2.4 Measurements . . . . .	35
2.5 Conclusion . . . . .	39

<b>III. A Fully Integrated Counter Flow Energy Reservoir for Peak Power Delivery in Small Form-Factor Sensor Systems . . . . .</b>	<b>40</b>
3.1 Introduction . . . . .	40
3.2 Proposed Technique: Counter Flow Method . . . . .	44
3.2.1 Operation concept of counter flow method . . . . .	44
3.2.2 Energy efficiency analysis . . . . .	50
3.2.3 Voltage overshoot analysis . . . . .	57
3.3 Implementation of Counter Flow Energy Reservoir . . . . .	60
3.4 Measurements . . . . .	63
3.5 Conclusion . . . . .	70
<b>IV. A 0.04mm<sup>3</sup> 16nW Wireless and Batteryless Sensor System with Integrated Cortex-M0+ Processor and Optical Communication for Cellular Temperature Measurement . . . . .</b>	<b>71</b>
4.1 Introduction . . . . .	71
4.2 Cellular Temperature Sensing System . . . . .	73
4.3 Circuit Block Implementation . . . . .	74
4.4 Measurements . . . . .	77
<b>V. Pruning-based Pair Hidden Markov Model Accelerator for Whole Genome Sequencing . . . . .</b>	<b>81</b>
5.1 Introduction . . . . .	81
5.2 Pruning-Based Pair-HMM Algorithm . . . . .	85
5.2.1 Conventional Pair-HMM Algorithm . . . . .	85
5.2.2 Proposed Pruning-based Pair-HMM Algorithm . . . . .	86
5.3 Pruning-based Pair Hidden Markov Model Architecture . . . . .	92
5.3.1 PE Array . . . . .	92
5.3.2 Accelerator Architecture . . . . .	93
5.4 Conclusions . . . . .	95
<b>VI. Conclusions . . . . .</b>	<b>96</b>
6.1 Summary of Contributions . . . . .	96
6.2 Directions for Future Research . . . . .	97
<b>APPENDICES . . . . .</b>	<b>99</b>
<b>BIBLIOGRAPHY . . . . .</b>	<b>108</b>

## LIST OF FIGURES

### Figure

1.1	Evolution of computers shows volume reduction of computation devices [14] . . . . .	2
2.1	Recent advances in low power harvesting . . . . .	12
2.2	Conventional harvester . . . . .	13
2.3	Conceptual efficiency illustration of (a)traditional harvester efficiency (b)proposed harvester . . . . .	14
2.4	Concept of discontinuous harvester . . . . .	14
2.5	Conceptual operation of discontinuous harvester . . . . .	15
2.6	Dependency of efficiencies on $\Delta V_{sol}$ based on model prediction and simulation . . . . .	18
2.7	Dependency of $\Delta V_{sol, opt}$ on $C_{buf}$ based on calculation . . . . .	20
2.8	Proposed architecture . . . . .	21
2.9	Detailed architecture of discontinuous harvester . . . . .	22
2.10	Timing diagram of the discontinuous harvester . . . . .	22
2.11	Proposed reference voltages generation . . . . .	24
2.12	Simulated end-to-end efficiency with approximated $V_{ref\_H}$ and $V_{ref\_L}$	24
2.13	Structure of moving-sum charge pump . . . . .	26
2.14	3-phase operation of moving-sum charge pump . . . . .	27

2.15	Simulated startup energy comparison between moving-sum charge pump and Dickson charge pump . . . . .	28
2.16	Structure of binary charge pump . . . . .	30
2.17	2-phase operation of binary charge pump . . . . .	31
2.18	Charge pump efficiency comparison based on simulation . . . . .	32
2.19	Automatic Conversion Ratio Modulator . . . . .	33
2.20	(a) Circuit diagram of the mode controller and (b) timing diagram .	34
2.21	Die photo . . . . .	35
2.22	Automatic Conversion Ratio Modulator measurements . . . . .	36
2.23	Moving-sum charge pump measurements . . . . .	36
2.24	Harvester measurements . . . . .	37
2.25	Measured trade-off between transfer phase efficiency and solar efficiency . . . . .	38
2.26	Measured dependency of transfer phase efficiency on Cbuf size . . .	38
3.1	Voltage and current waveforms with direct battery connection . . .	41
3.2	Voltage and current waveforms with single capacitor method . . . .	42
3.3	Voltage and current waveforms with DC-DC converter . . . . .	42
3.4	Voltage and current waveforms with series parallel reconfiguration .	43
3.5	Voltage and current waveforms with charge sharing reconfiguration .	44
3.6	Operation concept of counter flow energy reservoir (split phase) . .	46
3.7	Operation concept of counter flow energy reservoir (1 <sup>st</sup> round of recombine phase) . . . . .	47
3.8	Operation concept of counter flow energy reservoir (2 <sup>nd</sup> round of recombine phase) . . . . .	48

3.9	Summary of the 2-phase operation of counter flow energy reservoir .	50
3.10	Illustration of steps in split phase for 16 unit capacitors . . . . .	55
3.11	Efficiency of counter flow energy reservoir over number of discrete capacitors . . . . .	56
3.12	Efficiency gain of the proposed method across relative allowable supply voltage drop . . . . .	56
3.13	Operation concept of time-spreading technique . . . . .	58
3.14	Maximum supply voltage overshoot wi/wo time-spreading across number of capacitors . . . . .	59
3.15	Top-level architecture of the implemented counter flow energy reservoir	60
3.16	Circuit implementation of the current limiter . . . . .	62
3.17	Illustration of switch connections in the energy reservoir . . . . .	62
3.18	Die photo . . . . .	63
3.19	Captured supply voltage waveform . . . . .	64
3.20	Measured energy breakdown . . . . .	64
3.21	Single shot energy delivered across allowable voltage drop (left) and load power (right) . . . . .	66
3.22	Single shot energy delivered at different temperatures . . . . .	66
3.23	End-to-end efficiency of the proposed energy reservoir . . . . .	67
3.24	Captured waveform showing counter flow charging . . . . .	68
3.25	Integration with radio . . . . .	69
3.26	Captured transmitter output pulse and supply voltage waveform . .	69
4.1	CTS encased with bio-compatible material and implanted in a cluster of homogeneously dispersed HS5 human bone marrow stromal cells .	72
4.2	Measured waveform with fully assembled CTS system . . . . .	74

4.3	System architecture of CTS . . . . .	75
4.4	Circuit implementation of optical transmitter subsystem . . . . .	76
4.5	Implementation of temperature sensor . . . . .	77
4.6	Measured temperature sensing performance . . . . .	78
4.7	Testing setup showing CTS stack in use with base station . . . . .	79
4.8	Sensing error and RMS resolution measured wirelessly with fully assembled CTS stack . . . . .	79
5.2	Compare and prune based on relative value of $f^I$ , $f^D$ and $f^M$ . . . . .	87
5.1	Data dependencies of (a) $f^M$ , (b) $f^I$ and (c) $f^D$ . . . . .	87
5.3	Illustration of proposed pruning-based Pair-HMM algorithm . . . . .	88
5.4	PE array structure used in Pair-HMM accelerator . . . . .	92
5.5	Hardware architecture of pruning-based Pair-HMM accelerator . . . . .	95
D.1	Dependency of simulated end-to-end efficiency on VH . . . . .	107

## LIST OF TABLES

### Table

2.1	Performance summary and comparison . . . . .	39
3.1	Chip characteristic summary . . . . .	69
4.1	System performance comparison . . . . .	80
5.1	Profiling result of HaplotypeCaller using chromosome 16-18 of sample HG00419 . . . . .	83
5.2	Computation reduction of Pruning-base Pair-HMM algorithm . . .	91
5.3	Performance comparison between floating point and fixed point PE arrays . . . . .	93

**LIST OF APPENDICES**

**Appendix**

A. Related Publications . . . . . 100

B. End-to-end efficiency of Energy Harvester . . . . . 102

C. Solar Efficiency . . . . . 104

D. Model Simplifications of Discontinuous Energy Harvester . . . . . 106

## ABSTRACT

The Internet of Things improves peoples lives and assists research advancement by collecting and converting data about the environment. With the reduction of volume of sensor nodes, data collection and computation can be achieved in a wider range of environments, enabling applications such as health monitoring, tamper detection and industrial sensing. Small form factor sensor nodes present unique challenges in circuit design and system integration due to size and power constraints. This thesis introduces circuit techniques and system designs featuring small form factor sensor nodes: (1) a discontinuous switched-capacitor energy harvester for ultra-low power energy harvesting, (2) a fully integrated energy reservoir unit that uses a counter flow method for peak power delivery in space-constrained sensor systems, and (3) a complete wireless sensor node for accurate cellular temperature measurement, enabling implantation in a cluster of cells or large egg cells for biological studies. Just as IoT systems provide convenience to peoples life and help with development in research and industry, genomics can transform precision health and enable tailored treatment plans for patients. This thesis introduces a pruning-based hardware accelerator for Pair Hidden Markov Model, which is a major time-consuming step in the secondary analysis of whole-genome sequencing.

# CHAPTER I

## Introduction

The Internet of Things (IoT) refers to sensing devices that collect and convert data about the environment. Since Kevin Ashton first defined the concept of the IoT in 1999, researchers have broadened its application space by enabling wireless communication, incorporating more sensing abilities, and reducing the volume of these sensing devices. With the reduction of volume of such sensor nodes, data collection and computation is now possible where it was previously impossible. These small form factor wireless sensor nodes enable a wide range of applications such as health monitoring, tamper detection and industrial sensing. Figure 1.1 illustrates the recent advances in computation platforms. Sensing devices have shrunk their volume to  $1 \text{ mm}^3$  or even sub- $\text{mm}^3$  scale. This trend of volume reduction has enabled more versatile sensor platforms but has also created unique design challenges and opportunities. This section will discuss the challenges in designing miniaturized sensor systems and introduce the solutions covered in this thesis.

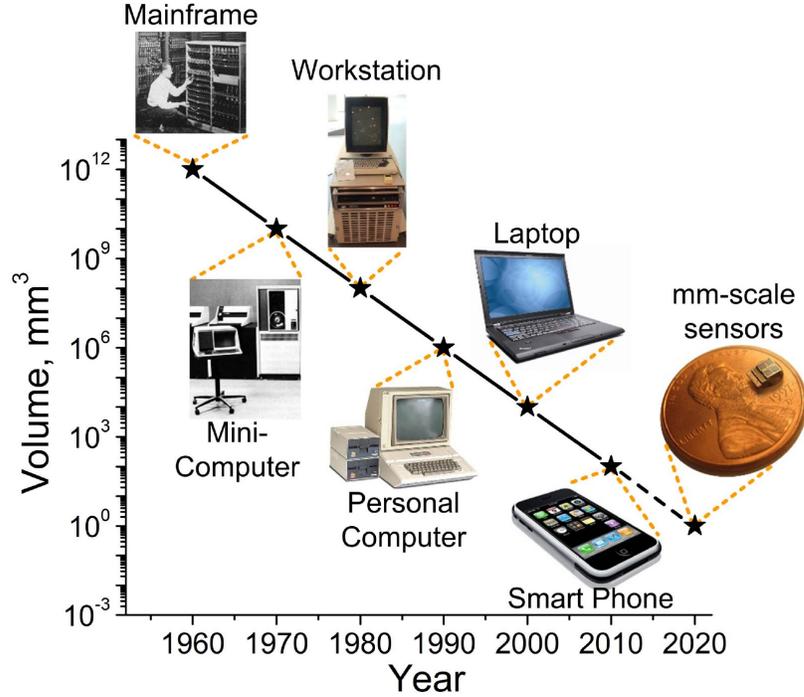


Figure 1.1: Evolution of computers shows volume reduction of computation devices [14]

## 1.1 Powering Miniaturized IoT Sensor Node

In order to work in a wide range of application spaces, IoT sensors often need to be adaptable to the environment. This means that the sensor nodes need to survive in locations where continuous power delivery from external sources is not available or unreliable. For example, industrial sensors used for oil pipeline infrastructure monitoring [1] or temperature and humidity monitoring inside concrete structures [2] are often located in harsh places, making it difficult to establish wire connections to such sensor nodes for power delivery. Localization sensors [3] are usually placed on moving subjects like cargo to track its movement, and it is impractical to provide a continuous source to power these sensing systems. Sensors for biomedical implants are isolated from the outside, making any wire connections intrusive. Continuous wireless power delivery is an option for such biomedical implants. A continuous wireless power

source can be integrated into a wearable device for patients [4], delivering energy through the skin to an implanted sensor node. A power source can also be installed in a lab environment for implanted sensors in animal-related experiments. However, this could interfere with communication and accurate measurements, adding design complexity to separate or cancel out noise from the wireless power source. In addition, wireless power sources may not be able to continuously and reliably deliver power to sensor nodes. A wearable power source for patients relies on a patient's discipline in wearing the device for continuous delivery. An installed wireless power source in a lab environment requires target animal to be in close proximity for maximum power delivery efficiency, which can sometimes be difficult to control. Therefore, sensor nodes need to handle the absence of direct power delivery or periods of power shortage even with wireless power sources.

For sensors located in hard-to-reach places where continuous power delivery is impractical, a battery is often included in the system to temporally store energy. Loading circuits in the sensor node can draw current directly from the battery if designed to operate at near battery voltage without regulation. As the size of the IoT sensor shrinks, the batteries integrated in the system must also shrink, resulting in higher internal resistance ( $\sim 10\text{ k}\Omega$ ). Although the average power of a small sensor node is usually small, the peak current of the system remains in the mA range due to periodic operations of high-power modules. Modules integrated in IoT systems such as radio, computation accelerators and non-volatile memory tend to draw high peak current (mA) during a short period of active time. Drawing such high current directly from the battery results in a high voltage drop on the supply voltage, which is impractical for load circuits designed at higher supply voltages. Therefore, an energy reservoir circuit is needed between the battery and the load circuit for short-term peak power delivery to high-power circuits in small IoT systems.

There are several design considerations for such energy reservoir circuits. First,

the energy reservoir is area-constrained because it needs to be integrated into the miniaturized sensor node. This area constraint directly limits the size of energy storage elements such as capacitors or inductors used in the energy reservoir design. Second, the energy delivered to the load circuit before it needs to be recharged must be sufficient for load circuits to complete one operation cycle. Load circuits such as a transmitter require uninterrupted peak power supply at least in one bit of transmission. Non-volatile memory in the sensor node often requires stable peak power delivery during one read or write operation. We refer to the amount of energy delivered by a reservoir before recharging as single-shot energy. The single-shot energy of an energy reservoir limits load circuit performance such as the maximum energy per bit available to transmitters. Third, the voltage droop on the supply voltage generated by the energy reservoir should be small enough for the desired circuit operation. Load circuits typically require a minimum supply voltage for correct operation. The energy reservoir should be able to maintain the supply voltage even when the remaining charge is decreasing during peak power delivery.

This thesis introduces an efficient energy reservoir that increases single-shot energy in peak power delivery. This solution aims to extract the maximum amount of energy from the reservoir with controlled voltage droop, leading to higher single-shot energy delivered or smaller area given the same energy requirement.

For some sensing applications, equipping sensor nodes with batteries only solves half of the problem. Sensor nodes generally benefit from a long life time and low maintenance. As discussed before, external power delivery requires power source installation, which can be unfeasible for some applications. Operating sensor nodes only on batteries requires periodic battery replacement or recharging for long-term monitoring. This involves retrieving sensor nodes physically from time to time, which is invasive for biomedical implants and difficult for sensor nodes on moving objects or in hard-to-reach places. Therefore, for these applications, it is crucial for sensor

nodes to be able to harvest energy from the ambient environment. There is a wide range of ambient energy sources available in the environment, such as light, heat and vibration. However, the power density can be very low depending on the location of the sensor node. In addition, the power density can also vary drastically. With low-power and wide-range harvesters, sensor nodes can survive in a wider range of ambient environments, including places where it was previously impossible due to the limited ambient energy. However, the lowest level of harvestable power of an energy harvester is limited by leakage power, and the harvesting range is often limited by the chosen topology. This work introduces a capacitive discontinuous energy harvester for a low-power and wide-input energy range.

## 1.2 Scaling Programmable Sensor Nodes to Sub-mm<sup>3</sup>

As industry and academia continue to push the limits of small form factor sensor nodes, more applications are becoming realistic. For example, small sensors can be placed on bumblebees [5] to track changes in colonies without interfering with their normal activities. A small intraocular pressure (IOP) sensor [6] can be placed in eyes for continuous pressure monitoring to track the progression of eye diseases. These advances create more design challenges for circuit designers. The first challenge is efficient power management. IoT sensors often include batteries for energy storage and use power management units (PMU) to extract energy from the battery and provide a regulated supply to the load circuits. Batteries scale very slowly in size compared to the scaling of CMOS feature size, making it extremely difficult for systems to scale beyond the smallest battery available. With strict size constraints, efficient PMU design becomes difficult due to the limited number of capacitors or inductors that can be fit into the system. Second, synchronized communication for programmability and data extraction is difficult due to the size constraints. The efficiency of a traditional RF antenna degrades with antenna size, forcing the use of high-power circuits and

resulting in mm-transmit distances. Third, accurate sensing often requires accurate current, voltage or frequency reference. However, the bandgap reference is too power hungry and crystal oscillators are too big for sub mm<sup>3</sup> systems.

This thesis introduces a 0.04-mm<sup>3</sup>, wireless and batteryless cellular temperature sensor system. Monitoring cellular temperature as an indicator of cellular metabolism is highly beneficial for disease study and drug discovery as many diseases (e.g., cancer) are characterized by abnormal metabolism. This work presents a fully programmable sensor node with a Cortex-M0+ processor, custom SRAM, optical energy harvesting, 2-way communication, and a subthreshold temperature sensor.

### **1.3 Extending Portable Computation to Genomics**

Just as IoT systems provide convenience to peoples lives and help with development in research and industry, genomics can transform precision health and enable tailored treatment plans for patients. Genome sequencing devices are being miniaturized, and companies such as Oxford Nanopore are selling palm-sized portable sequencers, making genomics a new application for the IoT. The advancement in genome sequencing techniques has reduced the cost of primary sequencing for one genome from ten million dollars [7] to one thousand dollars in the past decade. The speed and volume of sequencing machines have also improved greatly. This advancement can enable us to detect cancer without invasive biopsies, detect rare genetic disorders for early intervention, and identify pathogens for more accurate use of antibiotics. Advancements in the primary analysis has triggered a growing demand for computing power to speed up secondary analysis. Secondary analysis in whole-genome sequencing is a crucial but time-consuming step, taking hundreds to thousands of CPU hours [8] for one genome. As Moores Law tapers off, researchers have been developing customized accelerators using ASIC [9] or FPGA [10][11][12] to speed up the secondary analysis. This thesis also introduces a pruning-based hardware accelerator for a Pair

Hidden Markov Model (HMM) calculation, which is a major time-consuming step in the secondary analysis of whole-genome sequencing.

## 1.4 Dissertation Outline

This dissertation is composed of three chapters introducing circuit techniques and system design for small form factor sensor nodes and one chapter introducing a pruning-based Pair-HMM accelerator for genome sequencing.

Chapter II introduces a discontinuous switched-capacitor solar energy harvester that enables ultra-low power energy harvesting. Smart sensor applications rely on ultra-low power energy harvesters to scavenge energy across a wide range of ambient power levels and charge the battery. Based on the key observation that energy source efficiency is higher than charge pump efficiency, we present a discontinuous harvesting technique that decouples the two efficiencies for a better trade-off. By slowly accumulating charge on an input capacitor and then transferring it to a battery in burst-mode, DC-DC converter switching and leakage losses can be optimally traded off with the loss incurred by non-ideal MPPT operation. The harvester duty cycle is automatically modulated instead of the charge pump operating frequency to match with the energy source input power level. The harvester uses a hybrid structure called a moving sum charge pump for low startup energy upon a mode switch, an automatic conversion ratio modulator based on conduction loss optimization for fast conversion ratio increment and a  $<15$  pW asynchronous mode controller for ultra-low power operation. In 180-nm CMOS, the harvester achieves  $>40\%$  end-to-end efficiency from 113 pW to  $1.5 \mu\text{W}$  with 20 pW minimum harvestable input power.

Chapter III introduces a fully integrated energy reservoir unit using a counter flow method for peak power delivery in space-constrained sensor systems. Recent advances in circuits have enabled a significant reduction in the size of wireless systems such as implantable biomedical devices. As a consequence, the batteries integrated

in these systems have also shrunk, resulting in high internal resistances ( $\sim 10\text{ k}\Omega$ ). However, the peak current requirement of power-hungry components such as radios remains in the mW range and hence cannot be directly supplied from the battery. Therefore, an energy reservoir with high output power but small size is required. This chapter presents an efficient energy reservoir that dynamically reconfigures a storage capacitor array using a so-called counter flow approach. By creating a voltage gradient on capacitor arrays and moving the capacitors along the slope of the gradient, the supply voltage can be maintained while the energy stored in the reservoir is delivered efficiently to the load. The counter flow energy reservoir delivers 65% of stored energy before recharging is needed, which allows up to a  $12\times$  reduction in the overall capacitor size compared with our implementation of the previous method [13]. The design supplies up to 13.6 mW output power for  $1\mu\text{s}$ . This chapter demonstrates the proposed concept with a pulsed radio, showing an  $11.5\times$  increase in pulse length compared with the previous method [13].

Chapter IV introduces a complete wireless sensor node for accurate cellular temperature measurement, enabling implantation in a cluster of cells or large egg cells for biological studies. This is a complete wireless sensor node for accurate cellular temperature measurement that includes a fully programmable Cortex-M0+ processor, custom SRAM, optical energy harvesting, 2-way communication, and a subthreshold temperature sensor. The temperature resolution is  $0.034^\circ\text{C}$  RMS, and the transmit distance extends to 15.6 cm. The  $0.04\text{ mm}^3$  ( $\sim 500\times$  smaller than a grain of rice) fully assembled cellular temperature sensing system (CTS) is  $24\times$  smaller than prior programmable sensing systems[14], enabling implantation in a cluster of cells or large egg cells for biological studies.

Chapter V introduces a hardware accelerator using a pruning based algorithm for Pair-HMM in genome sequencing. In the primary analysis of whole-genome sequencing, sequencing machines generate billions of strings called reads, representing

fractions of DNA strands. In the secondary analysis, reads are first aligned to a previously sequenced genome using reference-guided assembly. Aligned reads are then processed to identify differences from the reference genome in the step called variant calling. Variant calling is complicated because the algorithm needs to identify real variants of the sequenced genome from errors introduced by sequencing machines. Pair-HMM is the most computationally intensive step in variant calling, taking 52% of the total run time. This chapter introduces an algorithm that explores the huge differences in values among the floating point numbers in Pair-HMM calculation so that the floating point calculation can be reduced dramatically for speed up. A hardware architecture is also introduced for the pruning-based Pair-HMM algorithm.

Chapter VI summarizes all the contributions of the presented work and discusses future directions.

## CHAPTER II

# A 20pW Discontinuous Switched-Capacitor Energy Harvester for Smart Sensor Applications

### 2.1 Introduction

Energy harvesting from the ambient environment is critical for self-sustaining IoT devices such as miniature-scale sensor nodes [16] and implantable medical systems [17] [30] [31] [32]. Energy sources including photovoltaic [16] [26], thermal [33], piezoelectric [34] [35] and RF energy [31] [32] are available for harvesters to scavenge to charge the batteries.

However, there are three main challenges in energy harvesting for IoT devices. First, power level varies dramatically with ambient conditions. Illuminance can range from 10 lux at twilight to 100K lux under direct sunlight. Under the illuminance range of 10 to 100K lux, a  $2.6 \times 3$  mm solar cell can produce 20nW to 200uW [18], marking a 10,000 $\times$  range, which is difficult for harvesters to efficiently scale across. Second, it is advantageous for harvesters to harvest from low ambient power level. Admittedly, there are applications where sufficient high input ambient power is available to harvesters and sufficient battery size to survive through periods of low ambient input power. However, there are also situations where the sensor nodes are supplied with limited maximum input power for long periods of time or with very limited

battery size or no battery at all. For some applications such as infrastructure monitoring, nodes may be placed in hidden or difficult-to-reach locations, often dark and possibly cold and quiet, providing extremely low ambient energy available for harvesting (e.g. 150pW for a 0.01mm<sup>2</sup> photovoltaic cell at 32 lux). Biological sensing, as another example, may require that the sensor nodes to be placed on moving animals, possibly restricting the level of maximum ambient energy available from pW to nW (from nanogenerators [38] [39] [40], from biofuel cell [41]). Therefore, harvesters which remain efficient with low ambient input energy may open up possibilities for wider choice of sensor node placements and energy scavenge sources. However, few harvester has been presented to date that can maintain reasonable efficiency with sub-nW input power. For convenience, we refer to the minimum harvestable power as the harvesting floor. As shown in Figure 2.1, the harvesting floor has decreased in recent publications, with some papers pushing the limit to 1nW at 30 to 50% efficiency. An inductor-based harvester was proposed in [21] which extends the harvesting floor to 1.2nW by reducing the leakage power of the harvester to 544pW, setting the harvesting floor to be near 500pW. As an alternative approach, a self-oscillating switched capacitor DC-DC converter was proposed [22] that extends harvesting floor by reducing clock generation overhead. Both these works sought to reduce the on-power of the harvesters, and thus pushed the harvesting floor down to near 500pW. This work is the first to our knowledge that can harvest below 500pW; it does so while maintaining at least 40% efficiency across an input power range of 13000 $\times$ . The third challenge for a harvester is that each energy source needs to be biased properly to produce maximum power; this process is called maximum power point tracking (MPPT). As shown in Figure 2.1, harvesters typically achieve  $\geq 90\%$  energy source efficiency when incorporating maximum power point tracking. In summary, we face three challenges: wide input power range, low ambient power, and maximum power point tracking.

To extract energy efficiently from an energy source, a proper bias condition is

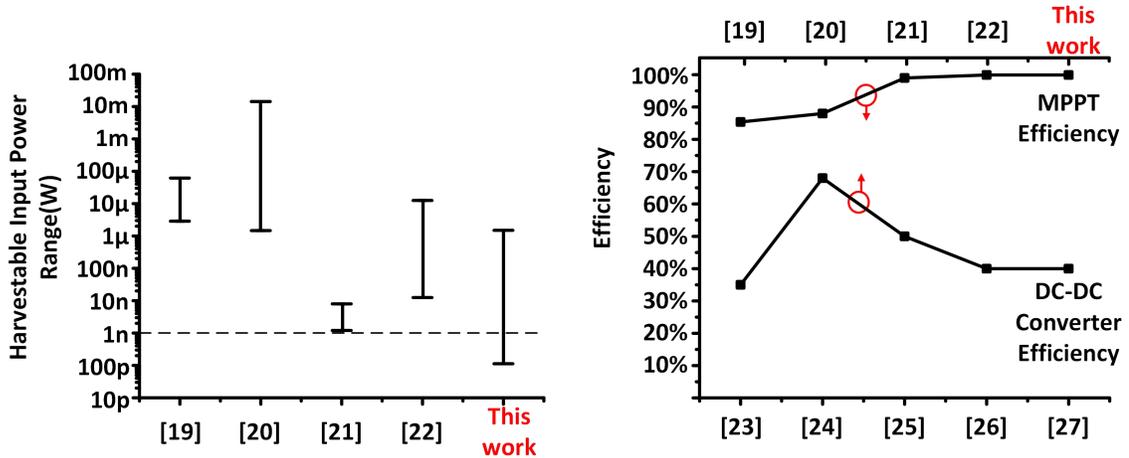


Figure 2.1: Recent advances in low power harvesting

required to match the ambient power levels (e.g.,  $V_{\text{MPPT}}$  depends on incident light level for photovoltaics). Fundamentally the ability to bias the energy source correctly for maximum power extraction is not limited by power levels, as long as the proper voltage or impedance is seen by the energy source. However, the efficiency of DC-DC converters is closely related to input power levels, and a DC-DC converter is usually only efficient for a certain power range [28], and limited by leakage for low input power. Hence, we observe that energy sources can offer much higher efficiency than DC-DC converters for low ambient power levels and across wide power ranges. Therefore, to extend the harvesting floor by increasing efficiency at low ambient power levels, this thesis proposed a new method called a discontinuous harvester, in which we intentionally trade off MPPT efficiency for DC-DC converter efficiency.

Conventionally, a harvester is a DC-DC converter, with one common topology being a switch-capacitor (SC) based charge pump as shown in Figure 2.2. This charge pump is continuously pumping charge from the energy source, which produces a low voltage, in order to charge the battery at a high voltage. DC-DC converter efficiency remains relatively flat for a certain range of input power as seen in Figure 2.3(a). As input power increases, the charge pump will increase its frequency to match the power

## Conventional Continuous Harvester

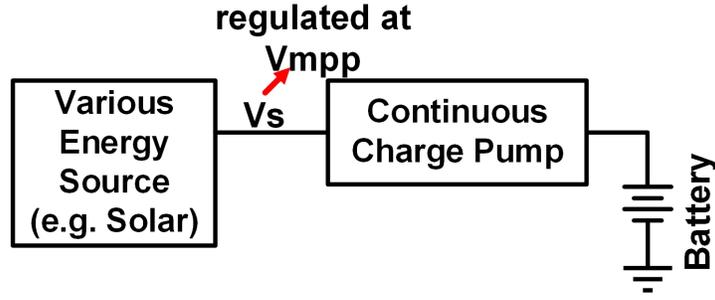


Figure 2.2: Conventional harvester

level. Eventually a point where efficiency flattens is reached where the efficiency is limited by the drive strength of the power switches. On the other hand, as input power decreases, the charge pump runs slower and becomes leakage dominated, leading to poor harvesting efficiency at low ambient power. Typically, reducing switch sizes can limit leakage. However, this approach concurrently reduces the maximum input power the system can harvest, resulting in a similar harvesting range. Therefore, size optimization cannot effectively extend the range of harvestable input power. In contrast, while charge pump power range is inherently limited, it is relatively easy to maintain MPPT efficiency across a wide range of input power. Put another way overall efficiency is given by MPPT efficiency multiplied by charge pump efficiency, and overall efficiency is limited by charge pump efficiency.

To extend the harvesting floor, the idea proposed in this chapter ([36] [43]) is to trade off MPPT efficiency to allow for higher charge pump efficiency at low input power levels. At the same time efficiency is maintained at high input power, so that an ultra-wide range harvester with low harvesting floor is achieved (Figure 2.3(b)).

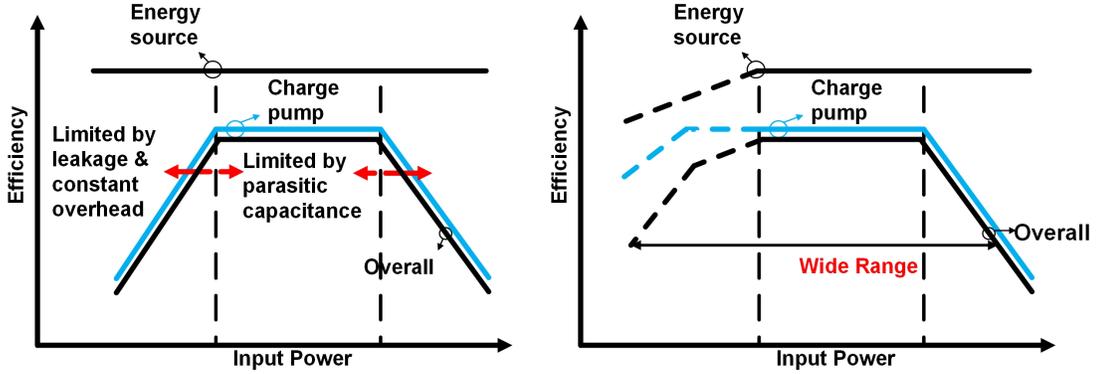


Figure 2.3: Conceptual efficiency illustration of (a) traditional harvester efficiency (b) proposed harvester

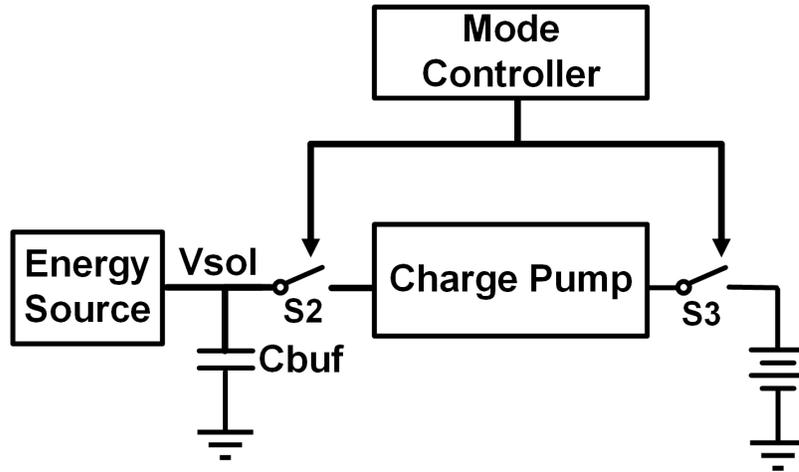


Figure 2.4: Concept of discontinuous harvester

## 2.2 Proposed Technique: Discontinuous Harvesting

### 2.2.1 Discontinuous Harvesting

The proposed work is a discontinuous harvester that operates in two phases (Figure 2.4). In these two phases, the bias voltage of the energy source,  $V_{sol}$ , deviates from  $V_{mppt}$ , which results in a slightly lower harvesting source efficiency. At the same time the charge pump is duty cycled to achieve a much higher CP efficiency. This work uses an off-chip capacitor controlled by on-chip switches S2 and S3 to isolate the charge pump. A mode controller enables the two phase discontinuous operation.

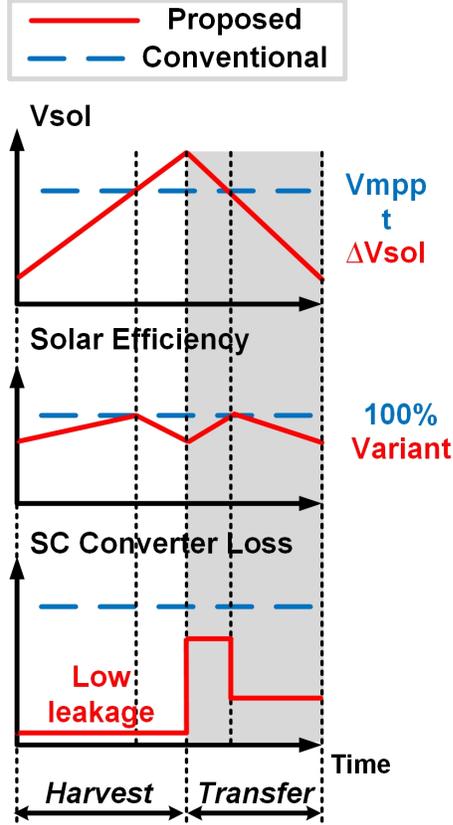


Figure 2.5: Conceptual operation of discontinuous harvester

It should be noted that this discontinuous burst-mode operation is only applicable when the ambient power accessible to the harvester is at the low end of its operating range. In this situation, the harvester efficiency is limited by leakage, and the discontinuous operation can effectively reduce the efficiency degradation due to leakage. When the ambient power accessible to the harvester is high enough for the harvester to operate efficiently and leakage is not dominant, the harvester is configured to operate continuously as a conventional charge pump, which matches its frequency to the given input power. In both scenario, we aim to extract maximum power from the input energy source.

Phase 1 is a harvest phase where S2 and S3 are open. In this phase, the energy source slowly accumulates charge on the capacitor. As shown in Figure 2.5, bias voltage  $V_{sol}$  increases from below  $V_{mpt}$  to above  $V_{mpt}$ . In contrast, a conventional

harvester attempts to hold the energy source output at a fixed voltage  $V_{mppt}$ . Hence as shown in the second plot, the proposed method sacrifices MPPT efficiency. In this phase, the charge pump is power gated reducing system leakage to below 15pW this value is critical as it sets the harvesting floor. In contrast, conventional continuous harvesters have a consistently high leakage, resulting in a low or even negative charge pump efficiency at extremely low input power levels (e.g., sub-nW).

When  $V_{sol}$  is sufficiently high, the harvester enters phase 2, which is a transfer phase. In this phase, S2 and S3 are closed to power on the charge pump, effectively transferring charge to the battery in a burst-mode. The charge pump goes through a startup mode and operates at its peak efficiency in steady state.  $V_{sol}$  quickly decreases in this phase, and at some point the harvester is reconfigured back to the harvest phase. It should be noted that when operating discontinuously (i.e. available input power is low), the charge pump always operates at its optimal frequency with peak efficiency, and when input power level is high enough for efficient continuous operation, the charge pump needs to adjust operating frequency for maximum power extraction. Therefore, this technique simplifies the charge pump design because optimizations (flying capacitor size, switch size, etc.) are only needed for high input power range. In this implementation, capacitor and switch sizes are optimized for input power  $>100nW$  for the given die area.

The resulting solar efficiency of the proposed harvester is lower because  $V_{sol}$  deviates from  $V_{mppt}$ , however a much higher charge pump efficiency is achieved due to the low leakage in harvest phase and peak efficiency in transfer phase. Therefore, the discontinuous harvester has much higher overall efficiency under low input power.

### 2.2.2 Energy efficiency trade-off analysis

The discontinuous harvester enters transfer phase when the capacitor  $C_{buf}$  is charged and returns to harvest phase when  $C_{buf}$  is depleted, resulting in a voltage

range seen at  $C_{buf}$ . We refer to the voltage range of this capacitor as  $\Delta V_{sol}$ . It is important to note that there is a trade-off between MPPT and DC-DC converter efficiencies that serves to limit  $\Delta V_{sol}$ .

$\Delta V_{sol}$  is an indicator of how often the system goes into transfer phase. Figure 2.6 shows the trade-off related to  $\Delta V_{sol}$  based on a mathematical derivation given later in this section. As  $\Delta V_{sol}$  decreases,  $V_{sol}$  becomes closer to  $V_{mppt}$  and solar efficiency accordingly rises. As  $\Delta V_{sol}$  decreases towards zero, the harvester becomes a conventional harvester operating continuously, biasing the solar cell at a fixed voltage where maximum power point tracking can be achieved for the given light condition, battery voltage and the implemented charge pump. However, in this latter case the harvester enters transfer phase more often, introducing extra losses. One cost associated with entering transfer phase very frequently includes a startup process in which the CP initializes the flying caps, requiring a large amount of energy. With infrequent entry to transfer phase (i.e., larger  $\Delta V_{sol}$ ) the startup loss gradually becomes negligible, and total charge pump efficiency approaches its peak. In summary a large  $\Delta V_{sol}$  limits solar efficiency while a small  $\Delta V_{sol}$  is limited by charge pump efficiency. Therefore, there is an optimal  $\Delta V_{sol}$  that achieves the highest overall efficiency.

To derive the optimal  $\Delta V_{sol}$  we define two voltages  $V_H$  and  $V_L$ , which indicate the high and low voltages seen at  $C_{buf}$  when the harvester enters transfer phase and harvest phase, respectively. Thus,  $\Delta V_{sol} = V_H - V_L$  by definition. The optimal pair of  $V_H$  and  $V_L$  results in the maximum end to end efficiency. End-to-end efficiency  $Eff_{tot}$  can be expressed as Equation 2.1, where  $Eff_{solar}$  is the solar efficiency in the harvest phase,  $Eff_{tran}$  is the overall charge pump efficiency in the transfer phase,  $P_{leak}$  is the leakage power in the harvest phase and  $P_{mppt}$  is the solar cell output power when biased at its maximum power point. (See Appendix B for details) Equivalent series resistance (ESR) of the capacitor  $C_{buf}$  (ESR=0.3 $\Omega$ , measured) can potentially limit the maximum current (e.g.  $I_{max}$ =30mA for 10mV voltage drop) supplied by  $C_{buf}$ ,

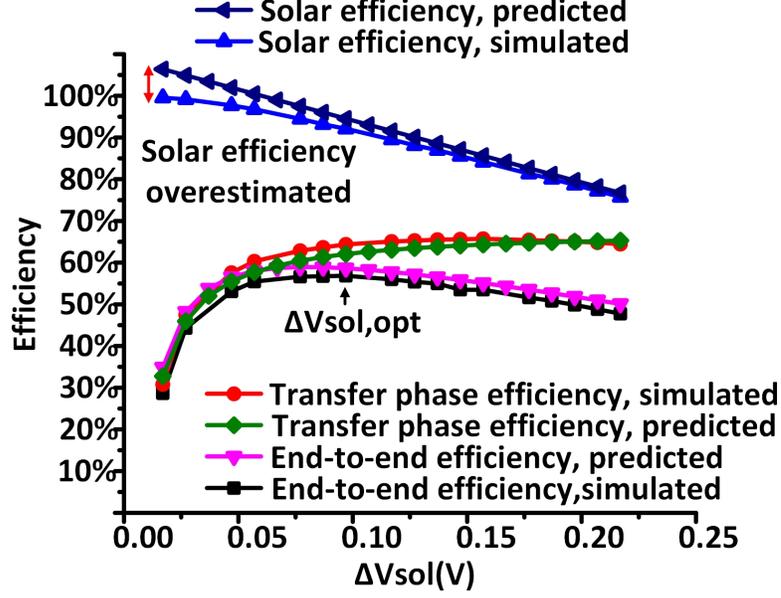


Figure 2.6: Dependency of efficiencies on  $\Delta V_{sol}$  based on model prediction and simulation

and thus sets an upper bound on input power in transfer phase for discontinuous operation. However, the charge pump implemented in this design operates at a much lower power level and therefore is not limited by ESR.

$$Eff_{tot} = Eff_{solar} * Eff_{tran} - \frac{P_{leak}}{P_{mppt}} \quad (2.1)$$

Overall charge pump efficiency in the transfer phase is given in Equation 2.2, where  $E_{out_{st}}$  and  $E_{in_{st}}$  are the energy drawn from the battery and Cbuf respectively during the startup step, and  $E_{out_{ss}}$  and  $E_{in_{ss}}$  are the steady state output and input energy, respectively.

$$Eff_{tran} = \frac{E_{out_{ss}} - E_{out_{st}}}{E_{in_{ss}} + E_{in_{st}}} \quad (2.2)$$

The transfer phase efficiency can be expressed in terms of VH and VL and charge pump efficiency in steady state,  $Eff_{ss}$  (Equation 2.3). For simplicity,  $Eff_{ss}$  is assumed to be independent of VH and VL for this derivation, since  $\Delta V_{sol}$  is only a few

hundreds of mV and charge pump efficiency is relatively insensitive to VH and VL compared to solar cell efficiency.  $Eout_{st}$  and  $Ein_{st}$  depend mostly on VH, and vary depending on the charge pump structure used. For simplicity,  $Eout_{st}$  and  $Ein_{st}$  are assumed to be independent of VH and VL.

$$Eff_{tran} = \frac{[\frac{1}{2} * Cbuf * (VH^2 - VL^2) - Ein_{st}] * Eff_{ss} - Eout_{st}}{\frac{1}{2} * Cbuf * (VH^2 - VL^2)} \quad (2.3)$$

When the harvester goes into harvest phase, solar cell outputs power to slowly charge Cbuf. Due to the voltage ripple  $\Delta V_{sol} = VH - VL$ , solar cell is not biased at its maximum power point, introducing a reduced solar efficiency. Therefore, solar efficiency can be expressed as in Equation 2.4, where VL and VH are the voltage across Cbuf at the beginning and at the end of the harvest phase respectively, and  $P(v)$  is the instantaneous output power of the solar cell when biased at voltage  $v$ . (See Appendix C for details.)

$$Eff_{solar} = \frac{\int_{VL}^{VH} 2vdv}{P_{mppt} \int_{VL}^{VH} \frac{2v}{P(v)} dv} \quad (2.4)$$

Here we set  $VH = V_{mppt}$ , which is the maximum power point of the solar cell. After simplification (see Appendix D for details), solar efficiency can be simplified as shown in Equation 2.5, where  $Isc$  is the short circuit current of the solar cell. Transfer phase efficiency can be rewritten in Equation 2.6 in terms of  $\Delta V_{sol}$ .

$$Eff_{solar} = \frac{(2V_{mppt} - \Delta V_{sol}) * Isc}{2P_{mppt}} \quad (2.5)$$

$$Eff_{tran} = \frac{[\frac{1}{2} * Cbuf * (V_{mppt}^2 - (V_{mppt} - \Delta V_{sol})^2) - Ein_{st}] * Eff_{ss} - Eout_{st}}{\frac{1}{2} * Cbuf * (V_{mppt}^2 - (V_{mppt} - \Delta V_{sol})^2)} \quad (2.6)$$

The optimal  $\Delta V_{sol}$  can be found by taking first order derivatives of Equation 2.1

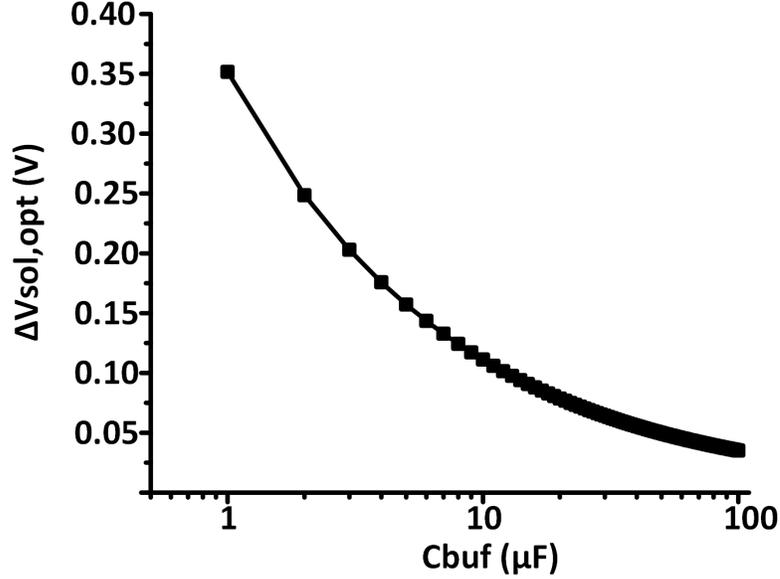


Figure 2.7: Dependency of  $\Delta V_{sol, opt}$  on  $C_{buf}$  based on calculation

assuming  $P_{leak}$  is independent of  $\Delta V_{sol}$ . Equation 2.7 shows that the optimal  $\Delta V_{sol}(\Delta V_{sol, opt})$  increases with startup energy and decreases with  $C_{buf}$  size. Figure 2.7 graphically shows the relationship between  $\Delta V_{sol, opt}$  and  $C_{buf}$ ; intuitively as  $C_{buf}$  grows the harvester should be able enter the transfer phase more often (startup costs are well amortized) and this improves total efficiency since the energy source operates closer to its MPP. One tradeoff here is in area and cost at the discrete component level.

$$\Delta V_{sol, opt} = \frac{\sqrt{2} * \sqrt{Eff_{ss} * Ein_{st} + Eout_{st}}}{\sqrt{C_{buf} * Eff_{ss}}} \quad (2.7)$$

## 2.3 Implementation of Discontinuous Harvester

### 2.3.1 Proposed Architecture

The proposed harvester (Figure 2.8) consists of an always-on power domain, shown in dashed lines, a gated power domain, an off-chip capacitor, and switches S1 through S3 used to enable the two phases. In harvest phase, the mode controller power gates the other circuits, while the solar cell charges the capacitor as discussed before. The

low power mode controller consists of a clocked comparator that monitors  $V_{sol}$  and triggers a transition to transfer phase if  $V_{sol}$  increases above  $V_{ref\_H}$ . The comparator is clocked by a leakage-based oscillator [37]. Mode transition is controlled using asynchronous logic to eliminate clock power that would otherwise be dominant.

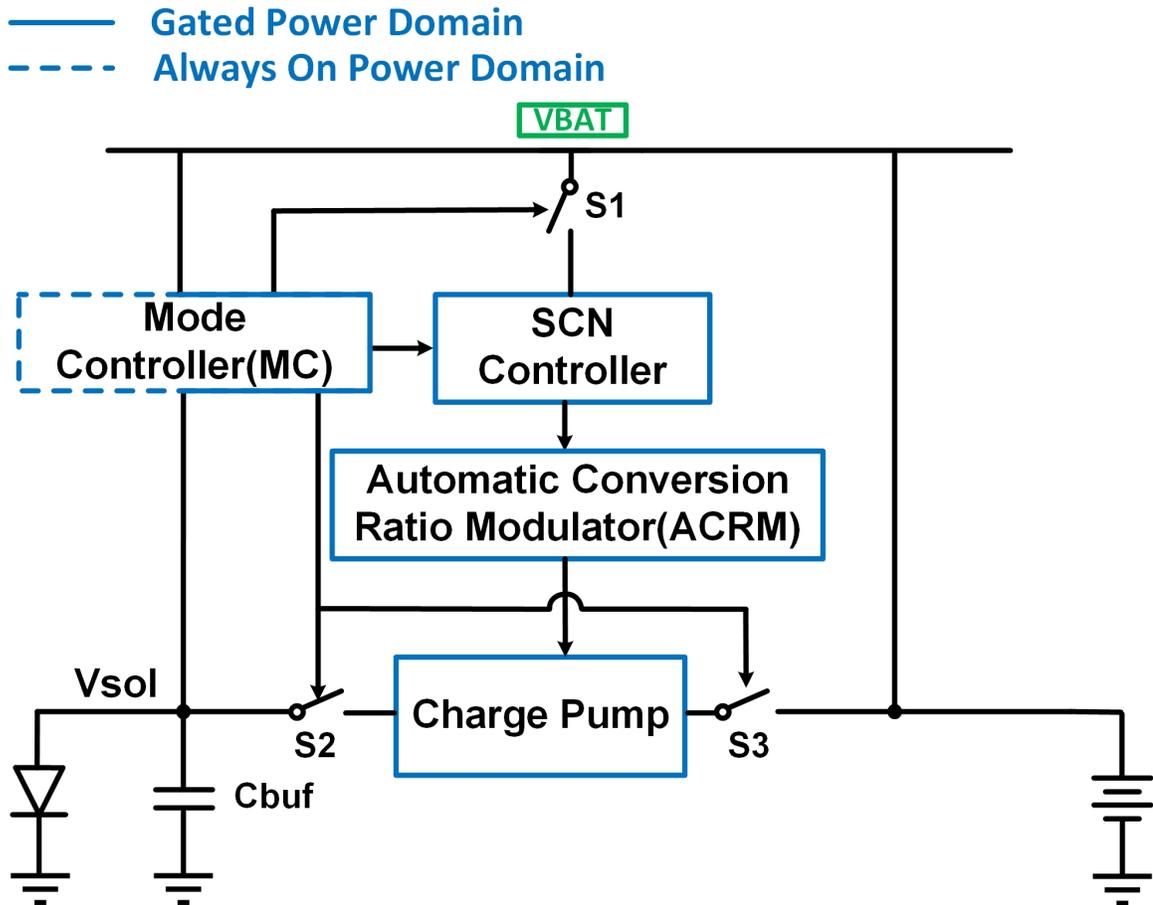


Figure 2.8: Proposed architecture

As shown in Figures 2.9 and 2.10, when the harvester enters the transfer phase,  $S1$ - $S3$  are enabled and the system is powered on. First, the charge pump needs to be powered up. The system controller is powered by battery voltage  $V_{BAT}$ , which is the only voltage available. It configures the charge pump to an initial conversion ratio, and begins counting cycles as the charge pump builds up its internal voltages. The



At this point the system controller switches to a slower clock to reduce dynamic power; a divided down version of the charge pump clock is generated and selected by the Clock CTRL module. Once the charge pump is stabilized it only requires occasional conversion ratio reconfiguration. As  $V_{sol}$  decreases during the transfer phase, the harvester automatically increases the conversion ratio. An automatic conversion ratio modulator (ACRM) monitors  $V_{sol}$  and determines whether conversion ratio should be changed or not, and increments the conversion ratio accordingly. Comparator C2 is a clocked comparator that controls the transition back to harvest phase. It fires when  $V_{sol}$  becomes lower than  $V_{ref\_L}$ , and returns the harvester to harvest phase. In this implementation,  $V_{ref\_H}$  and  $V_{ref\_L}$  are external references which vary with incident light conditions, and the light conditions here are sensed externally. A more complete system for future work should include the generation of the references, a light sensing module and a mapping module to map the lighting condition to the optimal reference voltages, which can be predetermined. The implementation of these modules may introduce extra power overhead. One possible method to generate the reference voltages is shown in Figure 2.11.  $V_{ref\_H}$  and  $V_{ref\_L}$  can be approximated as fractions of the open circuit voltage ( $V_{oc}$ ) of a solar cell, which can be generated using a dummy solar cell unit connected in parallel with a voltage divider. This provides a low power (simulated power consumption: 14fW typical, <100fW across corners) way of generating reference voltages that automatically tracks the lighting condition. Figure 2.12 compares the simulated optimal end-to-end efficiency with the efficiency when using the proposed circuit to generate  $V_{ref\_H}$  and  $V_{ref\_L}$  as fractions of  $V_{oc}$ . Optimal reference voltages are approximated with <10mV error, and the resulting efficiency degradation is within 2%.

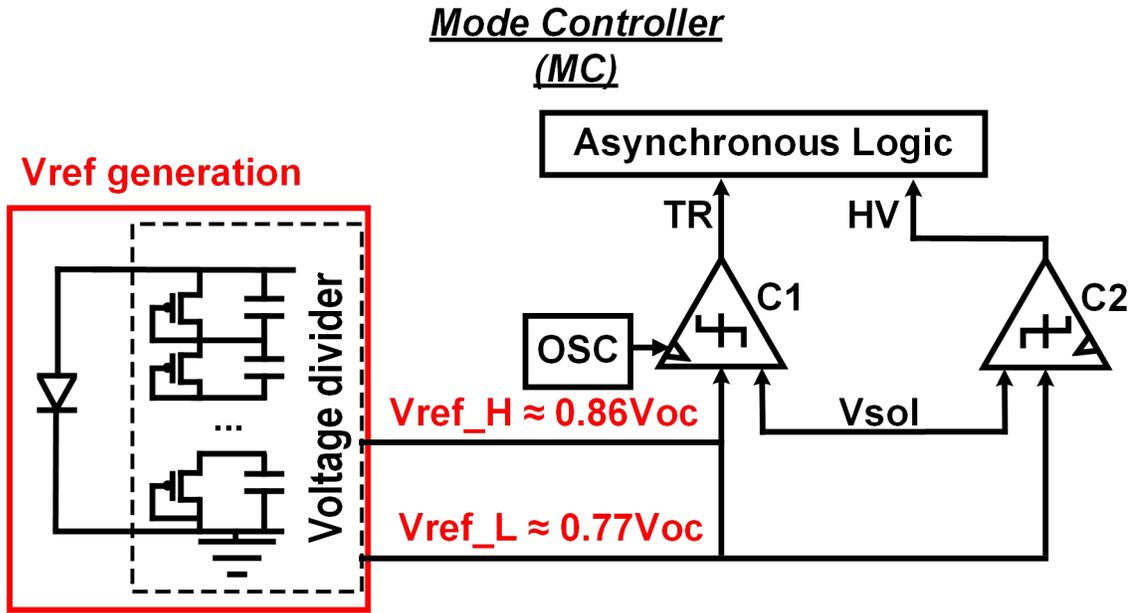


Figure 2.11: Proposed reference voltages generation

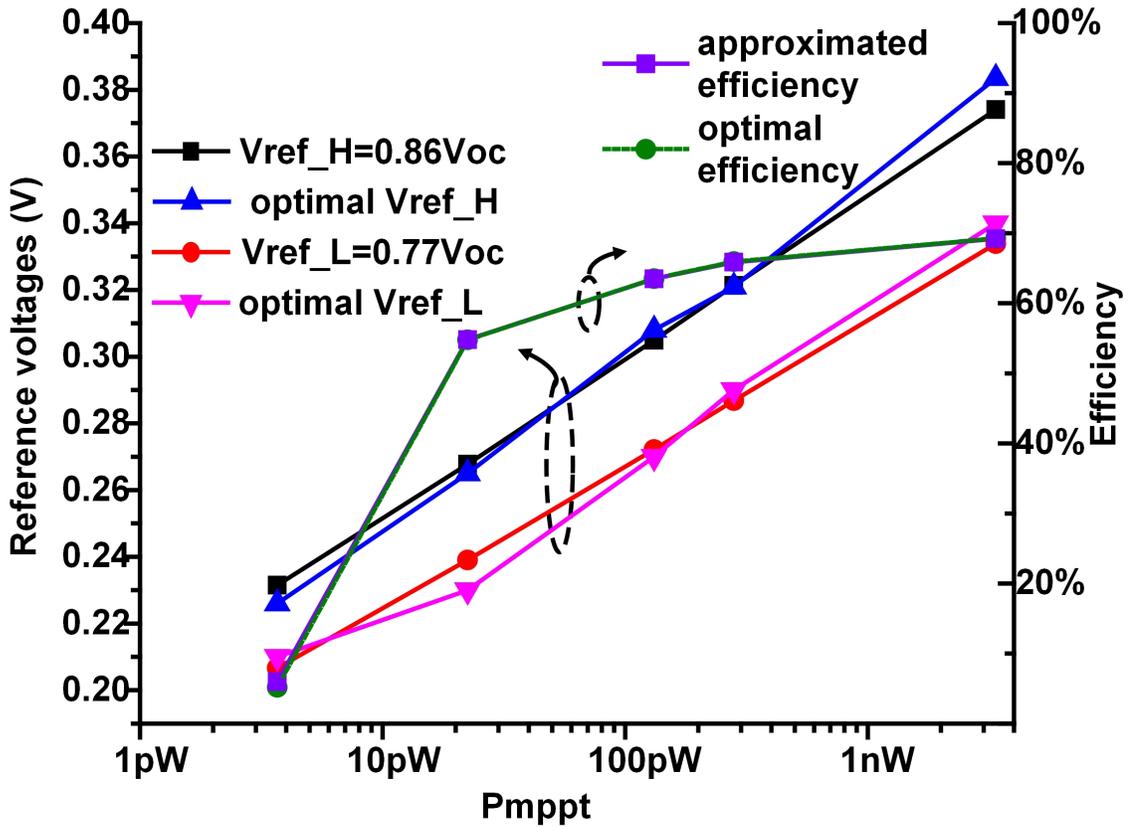


Figure 2.12: Simulated end-to-end efficiency with approximated  $V_{ref\_H}$  and  $V_{ref\_L}$

### 2.3.2 Moving-Sum Charge Pump

The DC-DC converter used in the harvester upconverts  $V_{sol}$  to the battery voltage in order to charge the battery, and it is only enabled during transfer phase. To accommodate solar voltage from 0.25–0.45V, we need 10–31 $\times$  variable conversion ratio. A standard approach would use a Dickson charge pump, which has high efficiency and offers fine-grained conversion ratios. However, Dickson charge pumps have drawbacks that are unique to the proposed discontinuous harvesting system. Dickson charge pumps have a large number of flying capacitors, and high voltage across each of them. For example, to obtain a 31 $\times$  conversion ratio, thirty flying capacitors are needed. Voltage across the capacitors range from  $1\times V_{IN}$  to  $30\times V_{IN}$ . This will result in large startup losses when initializing the flying caps. This is not a concern in always-on continuous harvesters, however these losses will greatly degrade efficiency in the proposed discontinuous harvester, as the harvester frequently starts and shuts down the charge pump.

In order to reduce the number of flying caps while maintaining all needed conversion ratios, this chapter introduces a new structure named moving sum charge pump, which is shown in Figure 2.13. It consists of a reduced Dickson charge pump to produce 2–9 $\times$  times  $V_{IN}$ , a voltage mux to select four voltages from 2–9 $\times$  according to the conversion ratio, and a summing series parallel stage where the selected voltages on the flying caps are placed in series and summed to charge  $V_{OUT}$ .

The operation has three phases as seen in Figure 2.14. In phases A and B, the reduced Dickson CP stage operates identically to a standard Dickson charge pump. Four different intermediate voltages are tapped out as  $V_{s1} - V_{s4}$ . Four flying caps in the summing stage are connected to the Dickson stage separately and charged to  $V_{s1} - V_{s4}$ . Charge is transferred from Dickson stage to summing stage. In phase C, the four flying caps in the summing stage are disconnected from Dickson stage, and then stacked together to produce  $V_{OUT}$ .

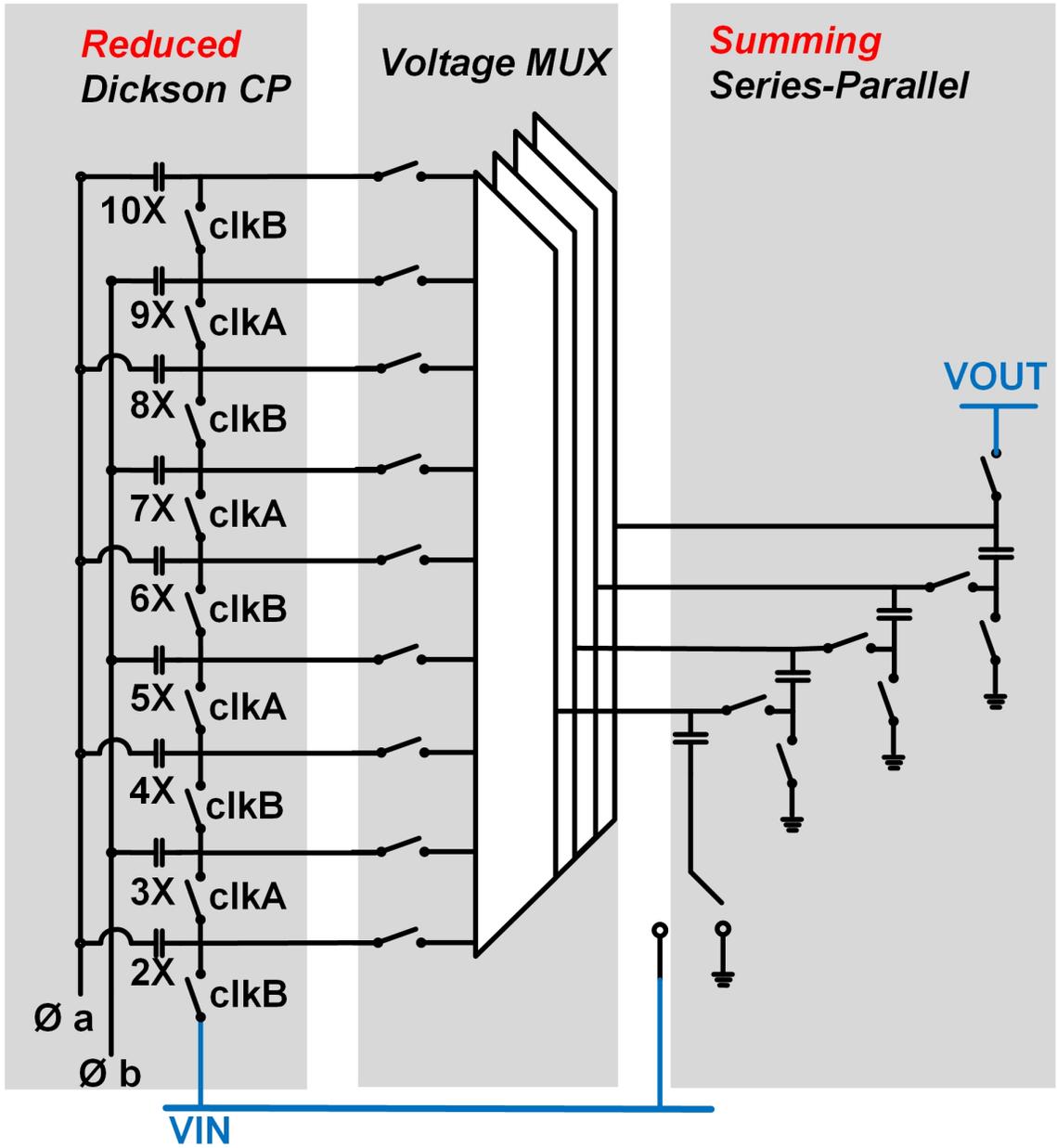


Figure 2.13: Structure of moving-sum charge pump

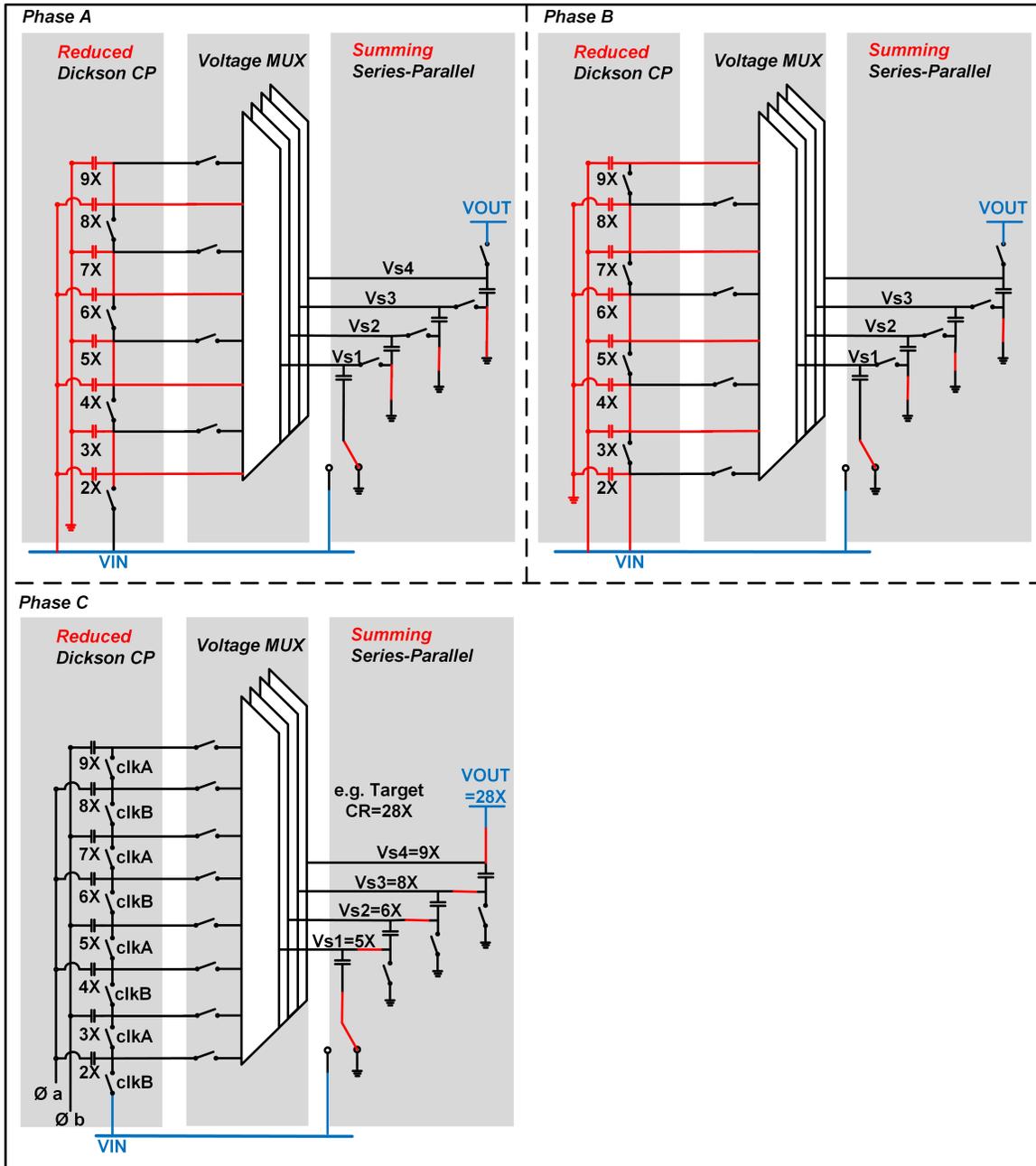


Figure 2.14: 3-phase operation of moving-sum charge pump

By selecting from  $2-9 \times V_{IN}$  and summing voltages, we are able to produce conversion ratios from  $10 \times$  to  $31 \times$  with only 12 fly caps instead of 30. For example, to produce a conversion ratio of  $28 \times$ , we need to select  $5 \times$ ,  $6 \times$ ,  $8 \times$ , and  $9 \times$  as  $V_{s1}-V_{s4}$ .

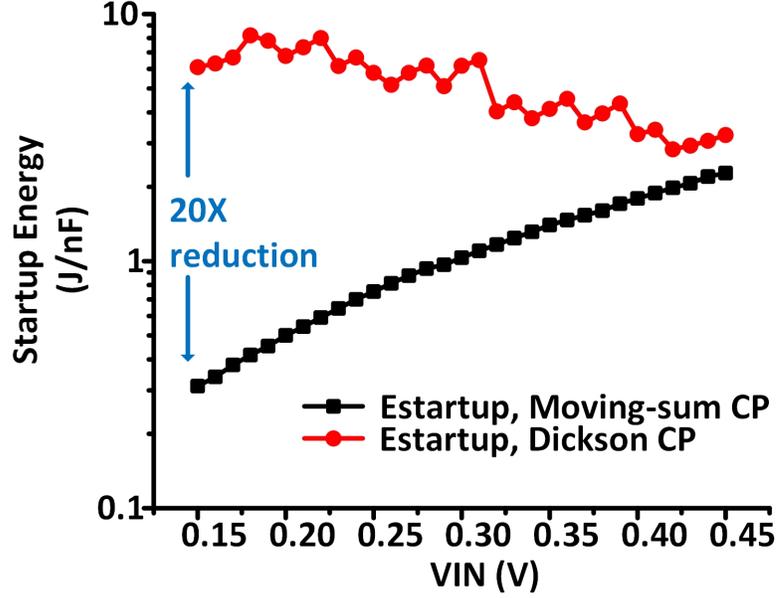


Figure 2.15: Simulated startup energy comparison between moving-sum charge pump and Dickson charge pump

Figure 2.15 shows the improvement in startup energy of the proposed moving-sum charge pump compared to the traditional Dickson charge pump, with both simulated and calculated values plotted. By reducing the number of flying caps and limiting the voltage across flying caps, the proposed structure reduces the startup energy by up to  $20\times$  compared to a Dickson charge pump. According to Equation 2.7, this can translate to an increase in MPPT efficiency through an allowable reduction of  $\Delta V_{sol, opt}$  by  $4.5\times$ .

In addition to the moving-sum charge pump, alternative hybrid charge pump structures can also be considered. The SAR DC-DC converter proposed in [29] is modified here for step-up conversion. This structure achieves fine-grain conversion ratios by reconfiguring 1:2 doublers. Similar to the moving-sum charge pump, a new hybrid structure, which is referred to as binary charge pump, is compared here as an alternative. As shown in Figure 2.16, the binary charge pump has two stages, a doubler chain stage that produces  $2\times$ ,  $4\times$ ,  $8\times$ , and  $16\times$   $V_{IN}$ , a voltage mux that selects four voltages ( $V_{s1}$  to  $V_{s4}$ ), and a summing stage that sums voltages  $V_{s1}$

to  $V_{s4}$  using four capacitors. The voltage selection is directly based on the binary representation of the conversion ratio, and this is where the name binary comes from. The operation requires two phases as shown in Figure 2.17. In phase A, the doublers chain stage is connected to the four capacitors from the summing stage in parallel, charging the capacitors to the desired voltages. In phase B, the four capacitors in the summing stage are disconnected from the doublers chain stage, and connected in series to produce output  $V_{OUT}$ .

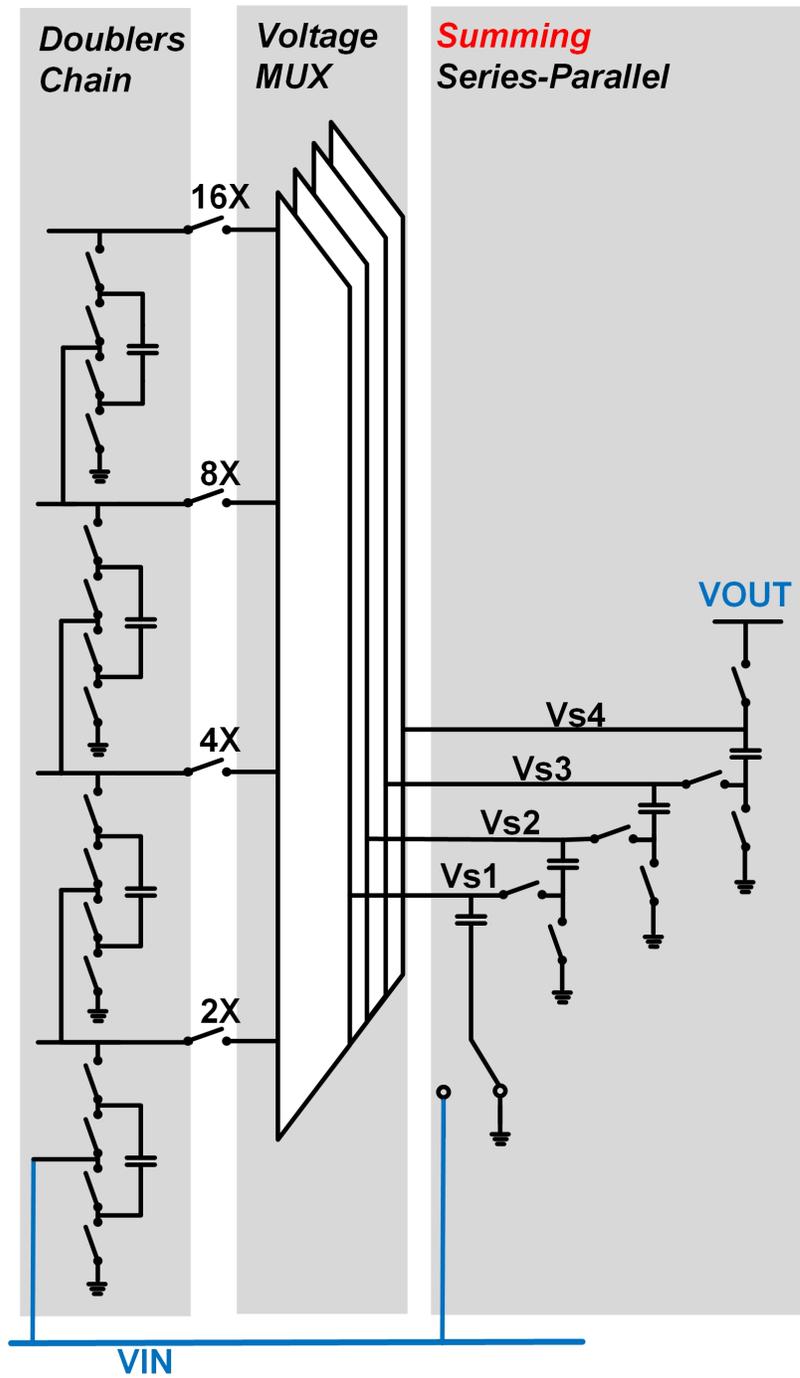


Figure 2.16: Structure of binary charge pump

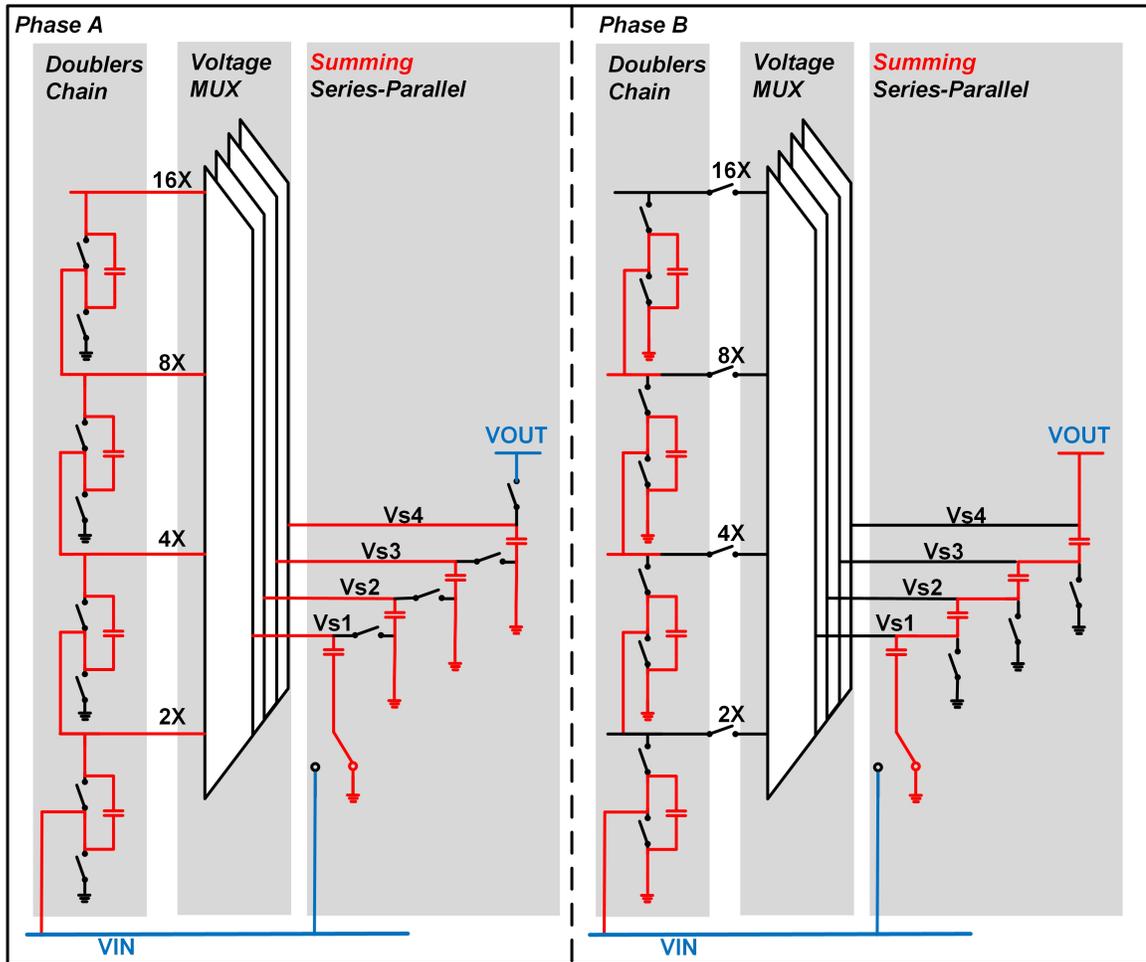


Figure 2.17: 2-phase operation of binary charge pump

In theory, SAR and binary charge pump can further reduce the startup energy by reducing the number of flying capacitors. However, these two doubler-based structures have lower steady state efficiency compared to Dickson-based structures such as the moving-sum charge pump. Figure 2.18 compares the simulated efficiencies of traditional Dickson, moving-sum, SAR, and binary charge pumps. Moving-sum charge pump maintains a higher efficiency over a wider range of input voltages than binary or SAR structures. Therefore, considering both startup and steady state losses, as well as a large desired input range, the moving-sum charge pump is implemented to

achieve better overall performance.

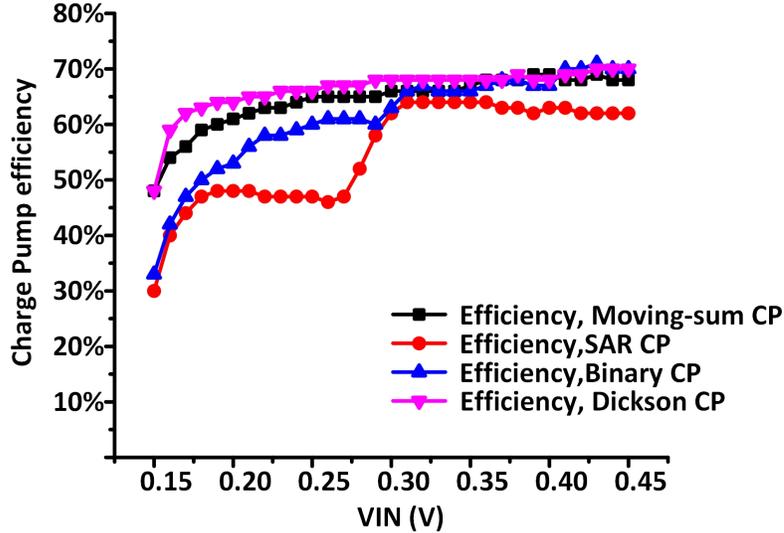


Figure 2.18: Charge pump efficiency comparison based on simulation

### 2.3.3 Automatic Conversion Ratio Modulator

ACRM (Figure 2.19) is only enabled in the transfer phase to automatically increase conversion ratio as  $V_{sol}$  decreases. For each input voltage, there is an optimal conversion ratio where the conduction loss is balanced with the switching loss. The harvester increases conversion ratio when the conduction loss is smaller than this balancing point. As an indicator of conduction loss, we use  $\Delta V$ , defined as conversion ratio  $CR$  multiplied by  $V_{IN}$  minus  $V_{BAT}$ , which is the difference between unloaded charge pump output and loaded output.

Conversion ratio is modulated by calculating  $\Delta V$  for the next conversion ratio,  $\Delta V_{CR+1}$ , and comparing it to the optimal  $\Delta V$ , which is approximated to be a fraction of  $V_{BAT}$ . We reduce all voltages here by half for easy implementation. After cleanup, the equation used for implementation is  $(CR + 1) * V_{in} * M < V_{BAT}$ , where  $M$  is a constant. The left hand side of the equation is defined to be  $V_{mult}$ , which is generated using a switch-capacitor amplifier in phase 1 and 2, then compared with

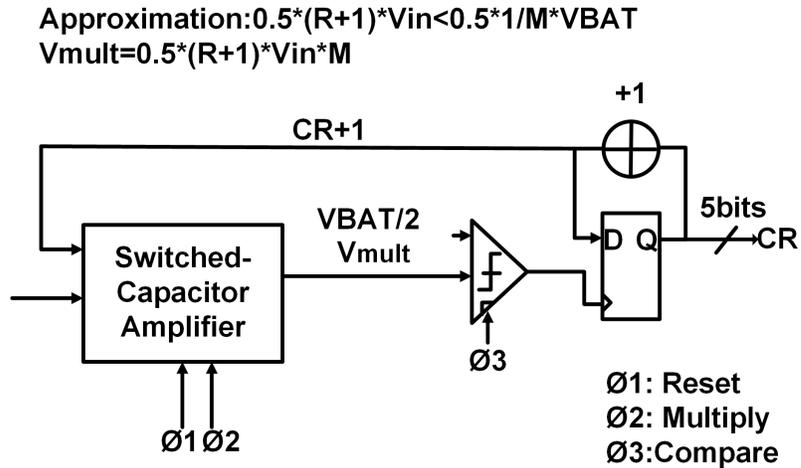


Figure 2.19: Automatic Conversion Ratio Modulator

half VBAT in phase 3. If  $V_{mult}$  is smaller than half VBAT, the comparator fires and conversion ratio increases by 1. This conversion ratio signal will be sent to a switch selection module to change the configuration of the moving-sum charge pump. Since  $V_{sol}$  is guaranteed to monotonically decrease during transfer phase, the logic for ratio modulation is simplified as it only needs to check for improved performance in one direction (i.e., towards a higher CR).

### 2.3.4 Low Power Mode Controller

The mode controller (Figure 2.20) controls the transition between harvest and transfer phases. It consists of a flip-flop to store the current state, a MUX, two pulse generators to clock the flip-flop at phase transitions, and delay cells to ensure correct timing. The complete controller has leakage power of less than 15pW, which is critical to enabling harvesting at ultra-low input power levels. Asynchronous logic is used to save clock power.

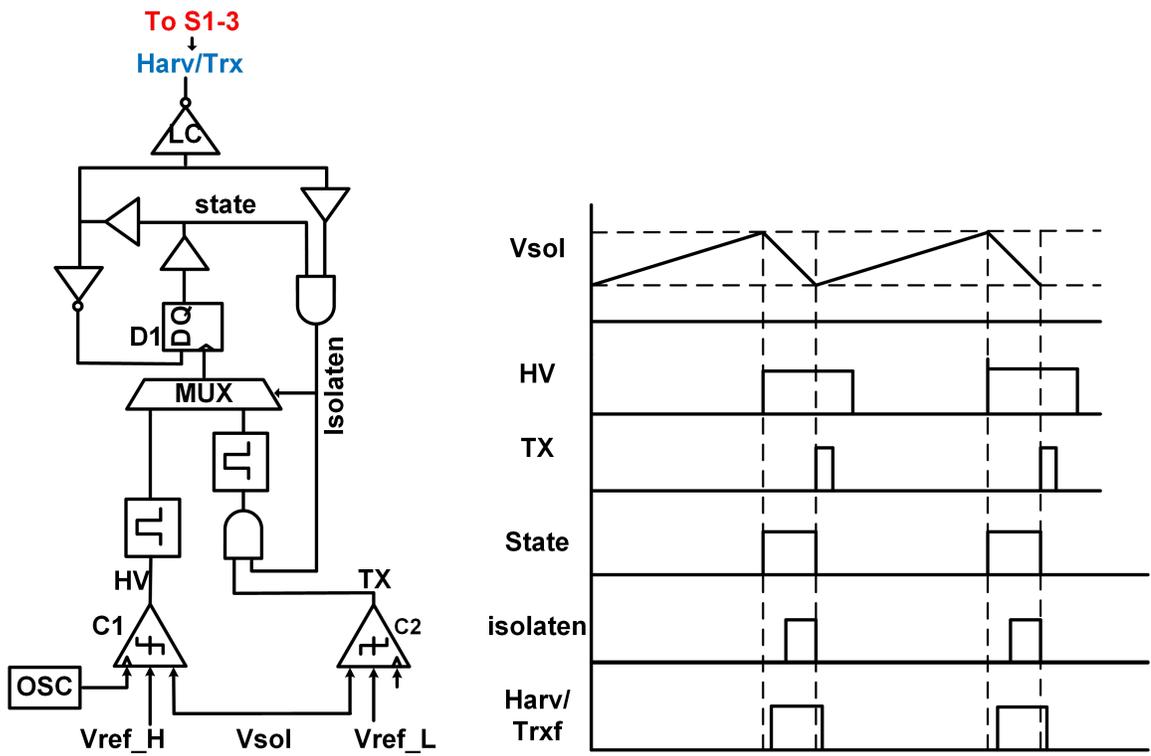


Figure 2.20: (a) Circuit diagram of the mode controller and (b) timing diagram

## 2.4 Measurements

The chip is fabricated in 180nm CMOS and occupies  $1.7\text{mm} \times 1.6\text{mm}$  (Figure 2.21). The design uses 12 flying capacitors with total cap size of  $1.5\text{nF}$ . The chip is tested with controlled lighting conditions using a  $0.01\text{mm}^2$  GaAs solar cell and two stacked CMOS solar cells, which are  $0.001\text{mm}^2$  and  $0.037\text{mm}^2$ , respectively. Harvester output energy is accumulated on a test capacitor, whose voltage is continuously monitored by electrometer.

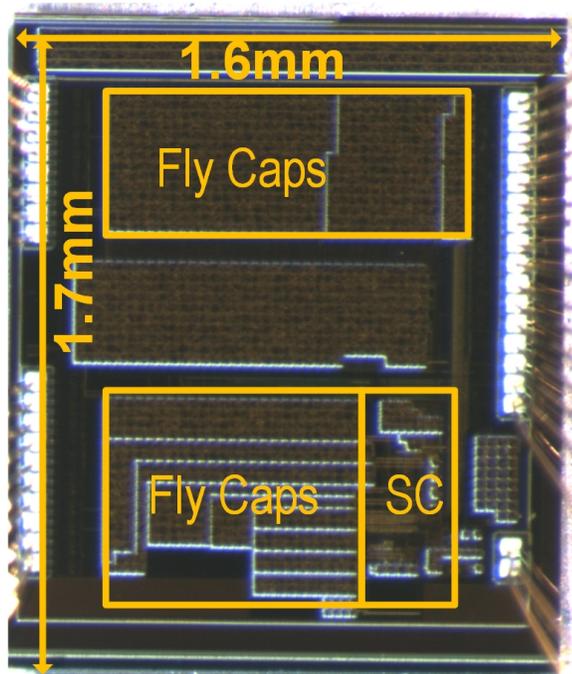


Figure 2.21: Die photo

Figure 2.22 shows the measurement of automatic conversion ratio modulator across  $V_{IN}$ , which is swept from 0.26 to 0.6V. The ACRM can select the correct conversion ratio within 2 codes from optimal, resulting in only a few percent efficiency degradation for most of the conversion ratios.

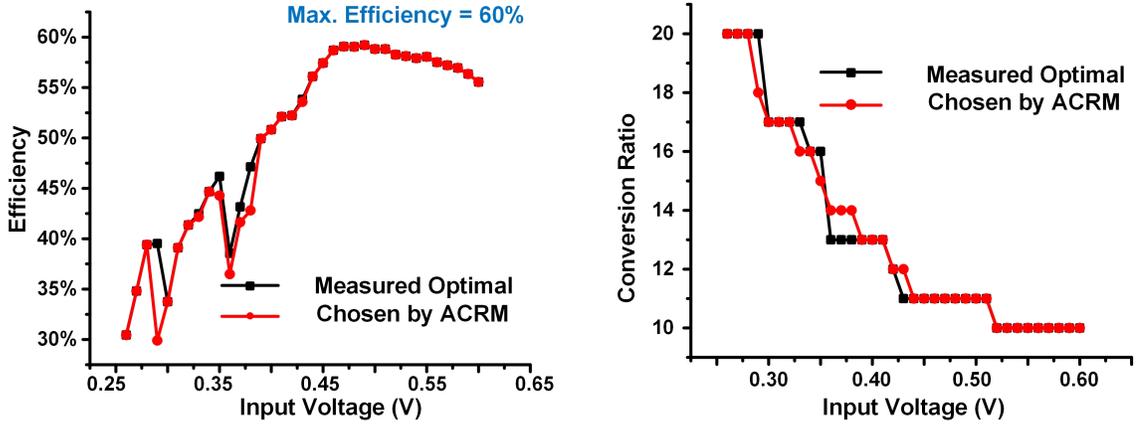


Figure 2.22: Automatic Conversion Ratio Modulator measurements

Figure 2.23 characterizes the moving-sum charge pump efficiency versus output power. It achieves 60% peak efficiency at 256nW output power when converting solar voltage to a 4V battery voltage, and maintains  $\geq 45\%$  efficiency over the 4nW to  $4\mu\text{W}$  output power range.

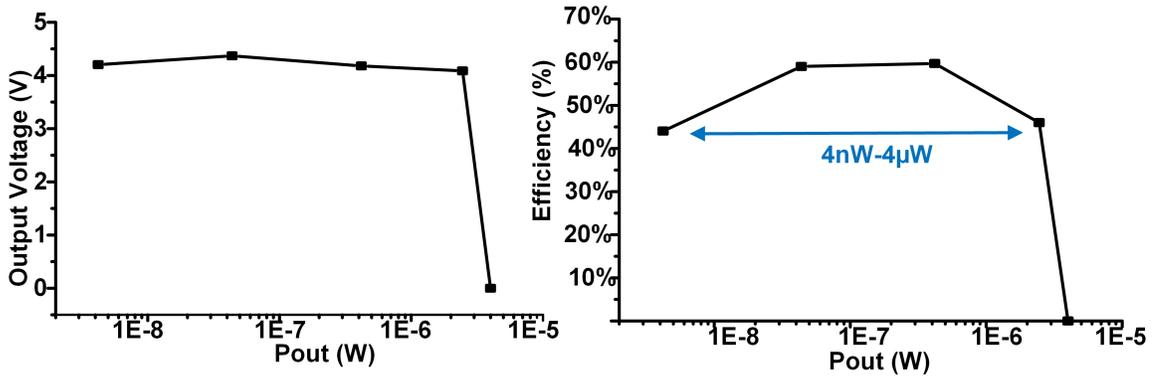


Figure 2.23: Moving-sum charge pump measurements

The efficiency improvement of the proposed discontinuous harvester over the conventional continuous harvester is compared in Figure 2.24. Data points with  $P_{mppt} > 66\text{pW}$  are taken using the GaAs solar cell, and data points with  $P_{mppt} < 66\text{pW}$

are taken using stacked CMOS solar cells to boost solar voltage at very low input light levels. End-to-end efficiency is calculated as harvester output power  $P_{out}$  divided by source power at its maximum power point. For the continuous harvester, its harvestable input power range is approximately 10nW to 1.5 $\mu$ W. The proposed discontinuous harvester efficiency can harvest from 113pW to 1.5 $\mu$ W with efficiency >40%. The discontinuous harvester also provides >20% efficiency at 20pW.

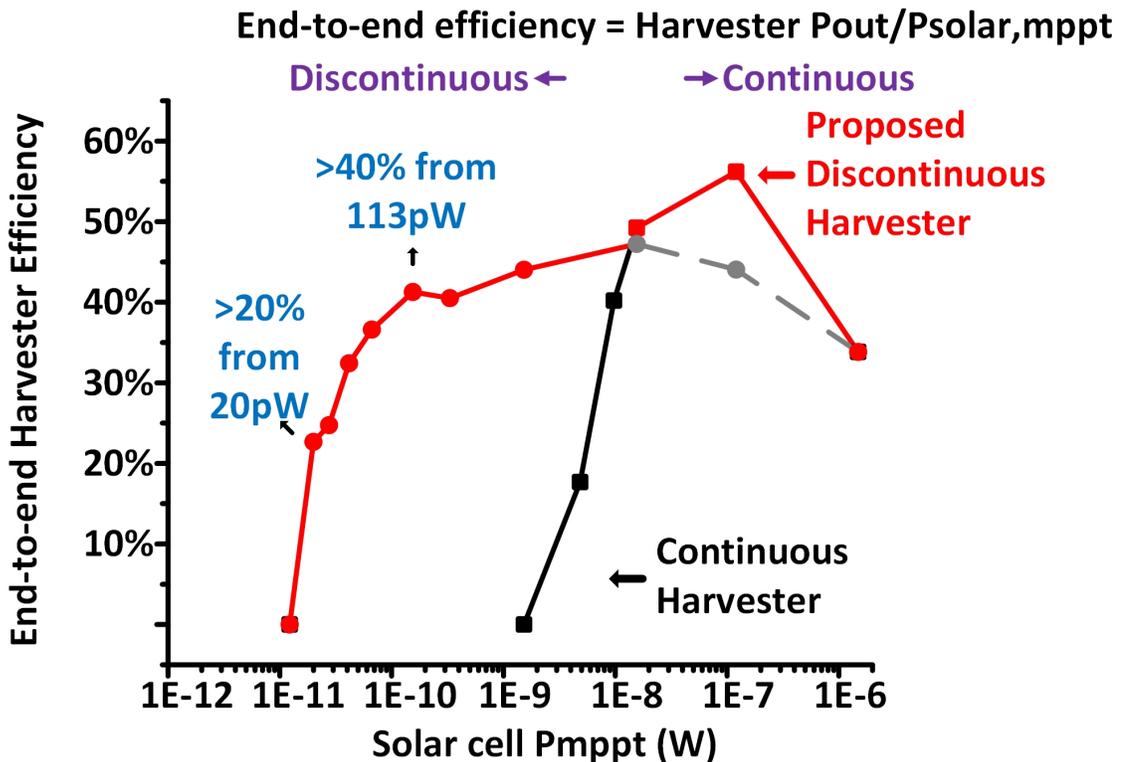


Figure 2.24: Harvester measurements

As described earlier, there is a trade-off between MPPT efficiency and charge pump efficiency that is quantified by  $\Delta V_{sol}$ . Measurements in Figure 2.25 show that as  $\Delta V_{sol}$  increases the solar efficiency decreases, while charge pump efficiency increases. This yields an optimal  $\Delta V_{sol}$  of 120mV in this case. This measurement confirms the previous efficiency analysis. Figure 2.26 provides measurements to demon-

strate the relationship between Cbuf size and transfer phase efficiency – this confirms that increased Cbuf size will initially improve transfer efficiency and then saturate at peak efficiency.

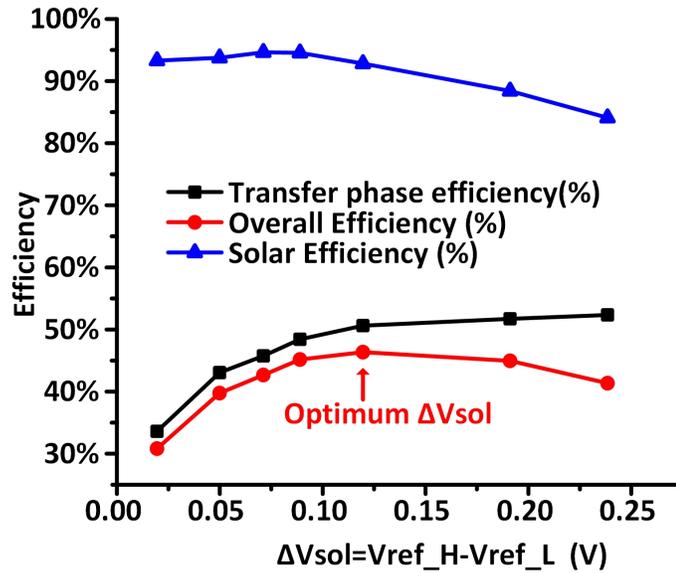


Figure 2.25: Measured trade-off between transfer phase efficiency and solar efficiency

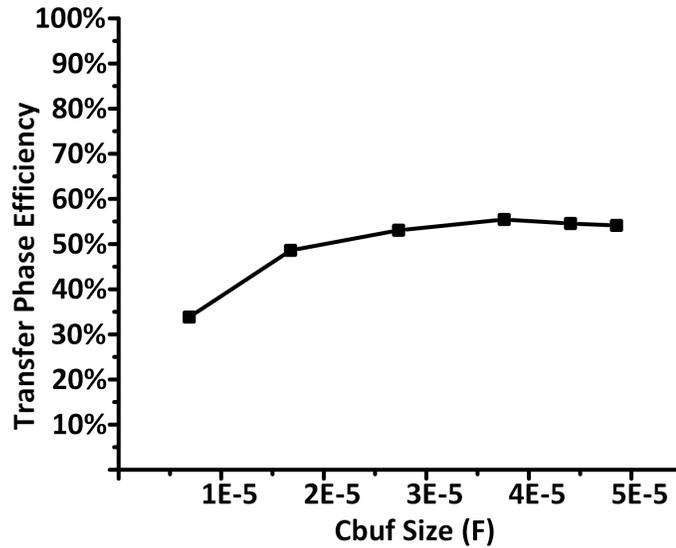


Figure 2.26: Measured dependency of transfer phase efficiency on Cbuf size

## 2.5 Conclusion

In conclusion, this chapter presented a discontinuous harvester where the solar efficiency and charge pump efficiency are separated and co-optimized to allow for a wider output power range and lower harvesting floor. The harvester achieves  $13,000\times$  input power range, 20pW harvesting floor, and less than 15pW idle power (Table 2.1). To optimize discontinuous harvesting, a new moving-sum charge pump topology is implemented to reduce startup energy. An automatic conversion ratio modulator increments conversion ratio to match decreasing input voltage while charge transfers to the battery. A low leakage mode controller is implemented to reduce idle power, lowering the harvesting floor.

Metric	[22] ISSCC,2014	[21] JSSC,2014	[20] JSSC,2015	This Work
Technology	0.18 $\mu$	0.18 $\mu$	0.18 $\mu$	0.18 $\mu$
Topology	Switched-Capacitor	Boost with Voltage Doubler	Buck boost	Switched-Capacitor
Input voltage	0.14-0.5V	20-70mV	N/R	0.25-0.65V
Output voltage	2.2-5.2V	1.5-1.9V	1V,1.8V and 3V	3.8-4V
CP Peak Efficiency	50% @ 0.45V	56% @ 0.1V	N/R	60% @ 0.5V
End-to-end Peak Efficiency	50% @ 100nW output power <sup>1</sup>	56% @ 0.9nW output power <sup>1</sup>	83% @ 90 $\mu$ W <sup>1</sup>	50% @ 8nW
Input Power Range	12.5nW-12.5 $\mu$ W w/ >40% efficiency <sup>1</sup>	1.2nW-8nW w/>50% efficiency <sup>1</sup>	1.47 $\mu$ W-14mW w/>68% efficiency <sup>1</sup>	113pW – 1.5 $\mu$ W w/ >40% efficiency
Efficiency at minimum input power	> 30% @ 4.5nW	53% @ 1.2nW	68% @ 1.47 $\mu$ W	37% @ 66pW 22% @ 20pW
Harvestable Power Range (Pin,max/Pin,min)	1000	7	9500	13000
Idle Power Consumption	3nW	544pW	400nW	<15pW

N/R: Not reported

<sup>1</sup>Estimated number from the paper

Table 2.1: Performance summary and comparison

## CHAPTER III

# A Fully Integrated Counter Flow Energy Reservoir for Peak Power Delivery in Small Form-Factor Sensor Systems

### 3.1 Introduction

Small form-factor systems are widely applicable in biomedical research and medical implants. Millimeter-scale implantable systems can monitor ECG signals [44] and intraocular pressure [46], stimulate the spine [45] and analyze blood samples [47]. To store energy, many of these small implantable systems use small form-factor batteries, which often have high internal resistance. For example, the commercial battery used in [48] has an internal resistance of up to 30 k $\Omega$ , which limits the direct current that can be drawn from the battery to 7 $\mu$ A with 200 mV voltage drop. Moreover, the internal resistance of batteries becomes worse with cycling, which further limits the output current.

However, the peak current requirement of power-hungry components, such as radios, remains in the hundreds of  $\mu$ A or even mA range. Therefore, if we directly connect the battery to the supply as shown in Figure 3.1, the battery voltage VBAT drops unacceptably when a burst of large current is pulled by the load circuits. One solution to this problem, illustrated in Figure 3.2, is to directly power this high

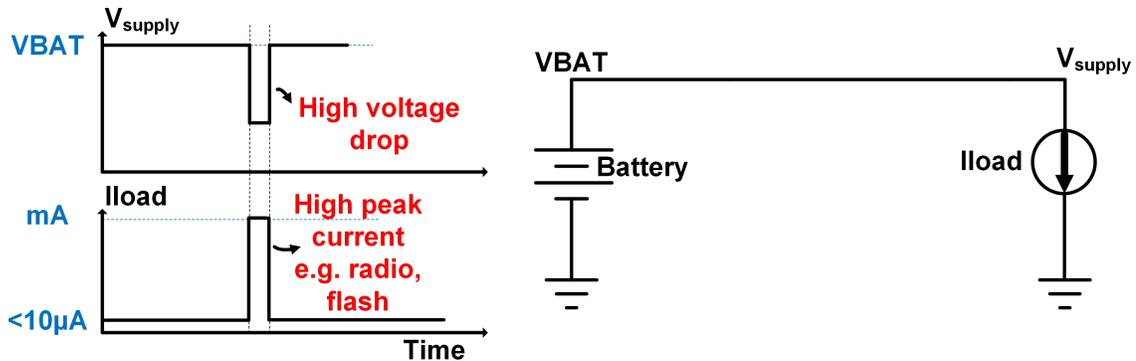


Figure 3.1: Voltage and current waveforms with direct battery connection

burst of current through a storage capacitor  $C_s$ , which is proposed for some pulse-based radios [13]. The capacitor is then recharged using a current limiter to protect the battery from excessive droop. This paradigm raises two challenges: 1) to supply sufficient energy, very large capacitance ( $>50$  nF) is often needed based on calculation (200 mV drop with battery voltage of 4V, for 10 mW and 5  $\mu s$  pulse duration), leading to a large die area or a bulky off-chip discrete component; and 2) only a small fraction ( 5% based on calculation) of energy stored in the capacitor is actually delivered to the high power components since the capacitor can only be discharged by a few hundred mV while maintaining proper circuit operation. In this section, we set 200mV drop (equivalent to 5% supply voltage drop) as a criteria to compare different design alternatives, because it is reasonable for many supply voltage sensitive circuits, such as amplifiers and memory. In real implementation, the proposed method can still operate beyond 5% supply voltage drop.

Alternatively, to extract more energy from the storage capacitor and reduce its size, a DC-DC converter can be used to more fully deplete the stored energy while maintaining the required supply voltage (Figure 3.3) [49]. However, such a high output power DC-DC converter requires either an off-chip discrete inductor or a large on-chip flying capacitor array with a total capacitance similar to or even larger than that of the mentioned storage capacitor, making this solution also unsuitable for small

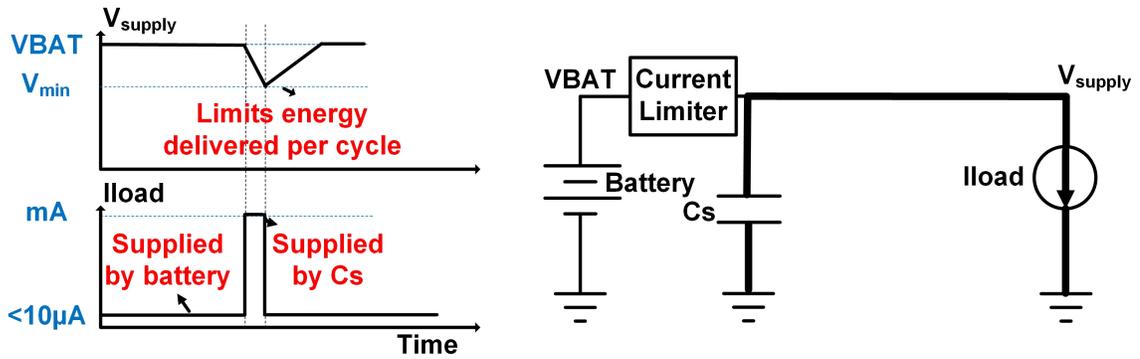


Figure 3.2: Voltage and current waveforms with single capacitor method

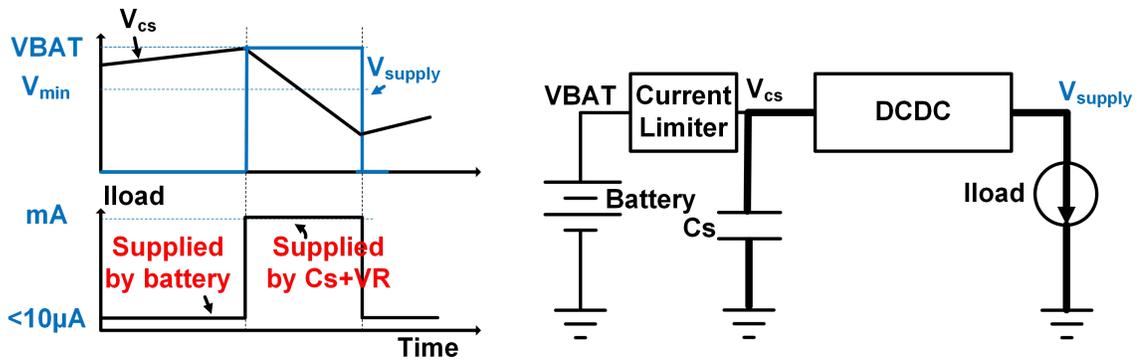


Figure 3.3: Voltage and current waveforms with DC-DC converter

form-factor sensors.

Another alternative solution is to decompose the large storage capacitor  $C_s$  into multiple small capacitors and reconfigure them to maintain supply (Figure 3.4). When the supply voltage drops below the minimum allowable voltage for the circuit, which we refer to as  $V_{\min}$ , the simplest reconfiguration scheme is to stack the capacitors in series to boost the voltage [50]. However, this leads to a  $2 \times V_{\min}$  supply voltage overshoot, which is not possible for many circuits.

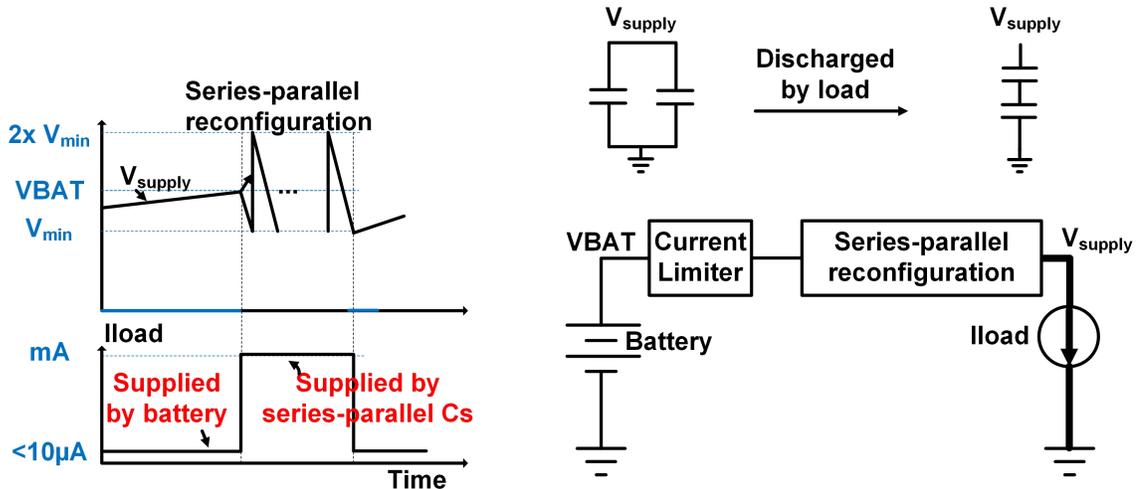


Figure 3.4: Voltage and current waveforms with series parallel reconfiguration

A feasible alternative is to stack a small portion of the capacitors in series and connect this stack in parallel with the rest of the capacitors (Figure 3.5). The supply voltage can then be boosted with acceptable overshoot. However, each reconfiguration creates charge sharing loss due to voltage drop across the switches. Therefore the energy extraction is only 41% for 16-unit capacitors based on calculation.

To deliver charge with minimum charge sharing loss and to extract a very high percentage of the total charge, this chapter introduces a counter flow reconfigurable energy reservoir ([52] [54]). This fully integrated energy reservoir unit dynamically reconfigures a storage capacitor array using a so-called counter flow approach. This design efficiently integrates the storage capacitor and DC-DC converter into one circuit, thereby maximizing the area efficiency and minimizing the charge sharing loss.

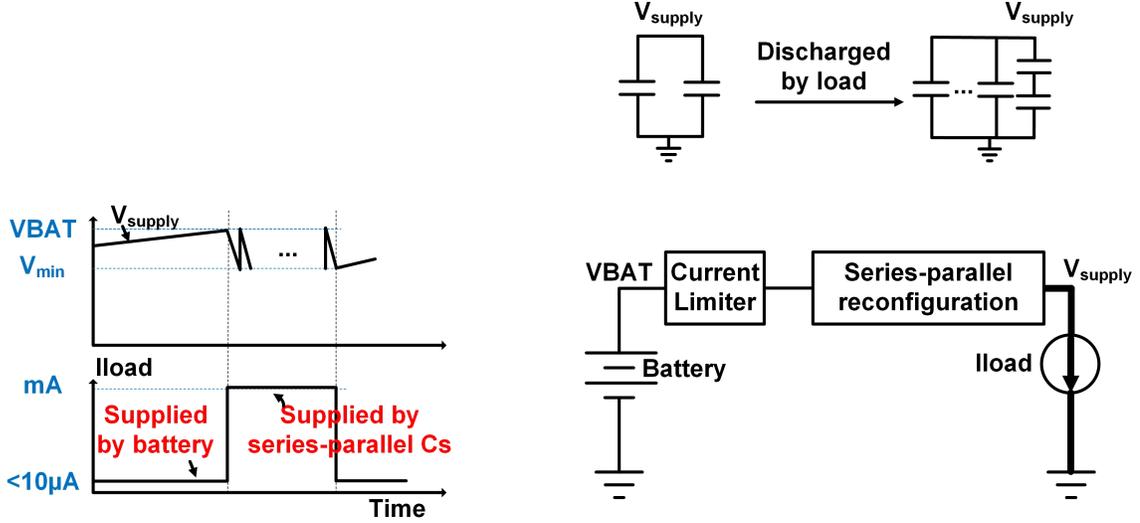


Figure 3.5: Voltage and current waveforms with charge sharing reconfiguration

## 3.2 Proposed Technique: Counter Flow Method

### 3.2.1 Operation concept of counter flow method

The key challenge in the reconfigurable energy reservoir is to deliver charge with minimum charge sharing loss (i.e., minimized intentional charge sharing steps inside capacitor array) and to extract a very high percentage of the total charge (i.e., minimum remaining voltage on capacitors). To accomplish this, we use an approach inspired by a biological phenomenon called counter flow, where oxygen and blood flow in opposite directions in fish gills, creating a slowly declining oxygen gradient for maximum gas exchange.

We use this idea and apply it to our problem of the efficient extraction of energy from a storage capacitor. This method is conceptually shown in Figure 3.6, 3.7, 3.8 with 8 unit capacitors, a battery voltage of 5 V and a minimum circuit operating voltage of 4 V. As shown in Figure 3.6(a), we start with all capacitors charged to 5 V in parallel, and they are then discharged by the load to 4 V. Then 2 capacitors are connected in series (Figure 3.6(b)) and subsequently connected in parallel with

the other 2 capacitors, boosting the supply voltage upon closing the switches. In a real implementation, a time-spreading technique (see Section 3.2.3) in which switches are closed sequentially is used to limit the voltage overshoot to 5%. As the load discharges the supply voltage to 4 V (Figure 3.6(c)), we obtain 2 capacitors at 2 V and 2 capacitors at 4 V. Next, we connect a capacitor at 2 V and a capacitor at 4 V in series (Figure 3.6(d)) and then connect them in parallel with a capacitor at 4 V, boosting the voltage again. Similarly, at the end of discharge to 4V by the load, we have formed a capacitor at 1 V and a capacitor at 3 V (Figure 3.6(e)). Along with the capacitor at 2 V and the capacitor at 4 V formed in the previous steps, we form a capacitor array with a trapezoid voltage gradient of 3 V (Figure 3.6(f)). In each round of reconfiguration, we stack the capacitor with the highest voltage with the lowest, the second highest with the second lowest, etc., and share charge with the capacitors when the load is at  $V_{\min}$ . This operation has two purposes: 1) The supply voltage is maintained with each reconfiguration. 2) Intermediate voltages are formed systematically at the end of discharge, and all previous formed voltages are conserved. In this way, the trapezoid capacitor array becomes more fine-grained with each round of operation. At the end of this process, each capacitor size is split in half, forming two identical sets of trapezoid capacitor arrays.

Then we stack the 2 set of capacitors in series as shown in Figure 3.7(a). The top 4 capacitors are charged to decreasing voltages that is represented by a trapezoidal shape, and the bottom 4 capacitors are charged to increasing voltages, creating a second trapezoid. The blue and red lines indicate the voltages of the different capacitors. Then we insert switches that allow us to reconfigure the capacitors. As the load discharges the supply voltage to 4 V (Figure 3.7(b)), we shift the two trapezoids in opposite directions (Figure 3.7(c)), boosting the supply voltage to 5 V again. This process is repeated. As the supply voltage is once again discharged to 4 V by the load (Figure 3.7(d)), we shift the two trapezoids in opposite directions, increasing the

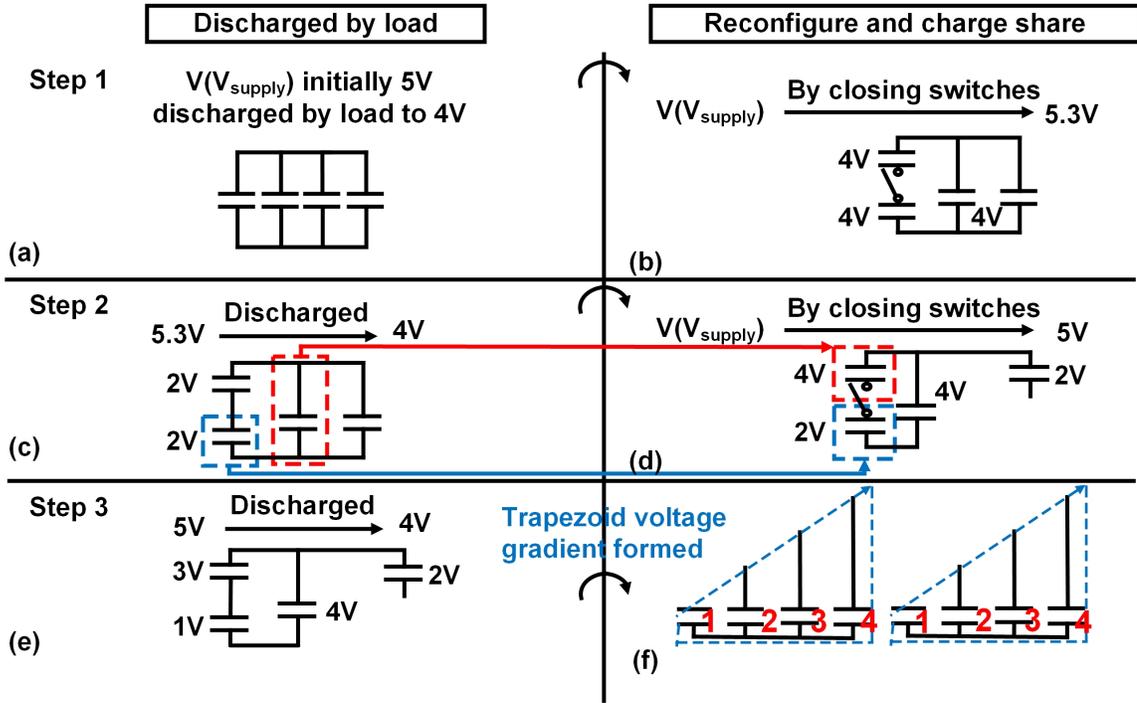


Figure 3.6: Operation concept of counter flow energy reservoir (split phase)

voltage by the slope of the gradient with each shift (Figure 3.7(e)). Since each shift operation simply increases the supply voltage and does not cause any charge sharing within the capacitor array, charge sharing loss is eliminated, resulting in highly efficient energy delivery. 78% energy is extracted based on theoretical calculation with 16 unit caps and 200mV voltage drop.

Figure 3.8(a) shows the final state of this process and indicates the charge still remaining on the capacitors. To extract this remaining energy from the reservoir, we fold the trapezoid and stack 4 capacitors in series to restore nominal supply voltage, forming 2 new trapezoids. As shown in Figure 3.8(b-e), when the voltage is discharged to  $V_{\text{min}}$ , we repeat the stack and shift operation as we previously did in the first round, resulting in 82% total energy extraction efficiency based on calculation with 16 unit caps and 200mV voltage drop. This second round of operation requires 13 more switches in a real implementation.

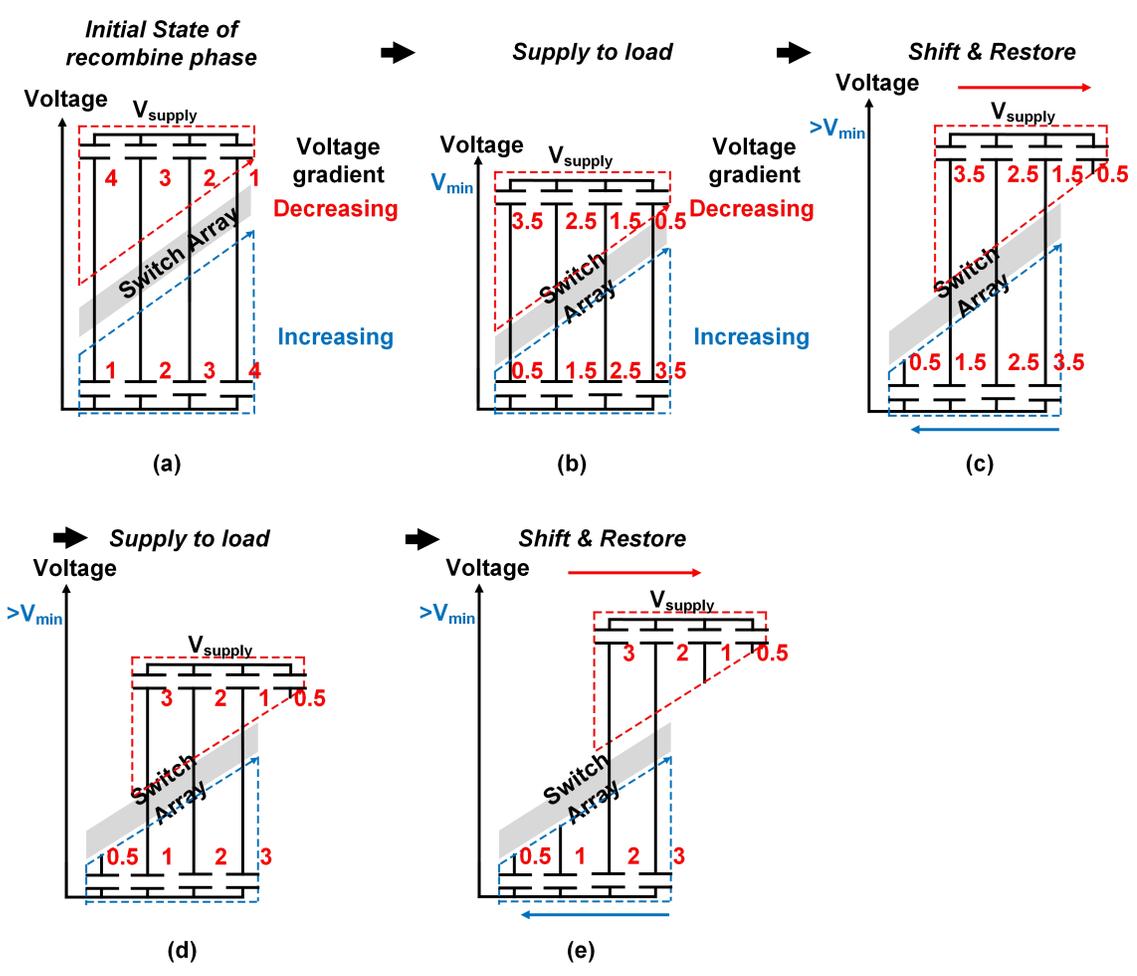


Figure 3.7: Operation concept of counter flow energy reservoir (1<sup>st</sup> round of recombine phase)

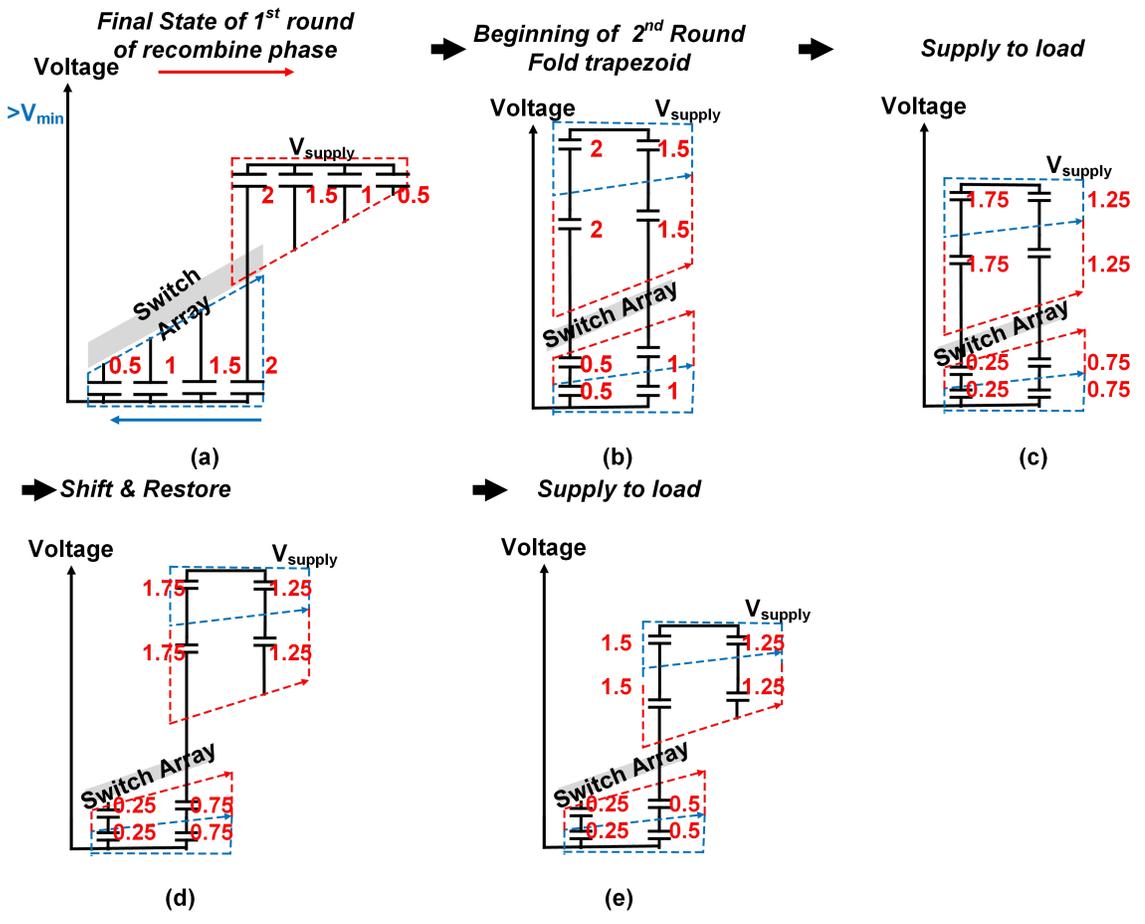


Figure 3.8: Operation concept of counter flow energy reservoir (2<sup>nd</sup> round of recombine phase)

In summary, the proposed counter flow energy reservoir has 2 phases (Figure 3.9). In phase 1, which is referred to as the split phase, a voltage gradient is created across two sets of capacitors with 13% charge sharing loss. In phase 2, which is referred to as the recombine phase, the two sets of capacitors are stacked together in reverse direction and repeatedly shifted in opposite directions as the load draws charge from the reservoir, increasing the voltage by the slope of the gradient with each shift. Since this shift operation simply increases the load voltage  $V_{\text{supply}}$  and does not cause any charge sharing within the capacitor array, charge sharing loss is avoided, resulting in highly efficient energy delivery. By repeatedly shifting, the vast majority of stored charge can be extracted, maximizing the total delivered charge. A second round of the recombine phase can be implemented by folding the trapezoids, as described previously, leaving only 5% energy not extracted based on calculation with 16 unit caps and 200mV voltage drop (See Section 3.2.2 for derivation). This process results in a total energy extraction efficiency of 82% for the entire process. It should be noted that we claimed no charge sharing loss in recombine phase assuming no decoupling capacitors at output. When decoupling capacitors are added at the output, there will be charge sharing loss every time the energy reservoir is reconfigured, including the recombine phase. This can degrade energy delivery efficiency depending on the capacitance of the decoupling capacitor, and we think this issue is hard to avoid for most switched-capacitor energy delivery schemes.

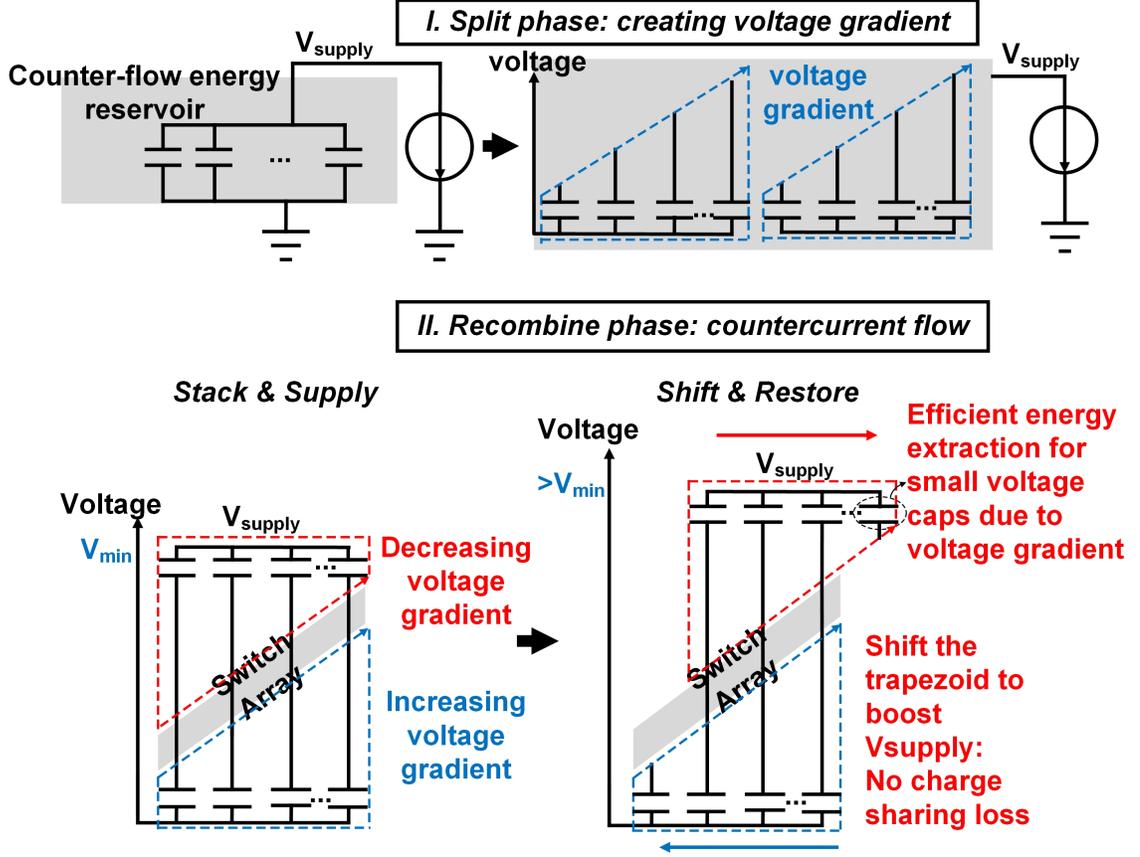


Figure 3.9: Summary of the 2-phase operation of counter flow energy reservoir

### 3.2.2 Energy efficiency analysis

To analyze the theoretical performance of the proposed counter flow energy reservoir, we define an efficiency metric named single shot energy delivery efficiency,  $Eff_{single-shot}$ . The efficiency is defined in Equation 3.1, where  $E_{deliver,split}$  is the energy delivered in split phase,  $E_{deliver,recombine}$  is the energy delivered in the recombine phase, and  $E_{stored}$  is the total energy originally stored in the energy reservoir. The single shot energy delivery efficiency indicates how efficiently energy is extracted before recharging.

$$Eff_{single-shot,proposed} = \frac{E_{deliver,split} + E_{deliver,recombine}}{E_{stored}} \quad (3.1)$$

In order to analyze the single shot energy delivery efficiency, we need to understand the two major sources of loss in the proposed design. The first source of loss is charge sharing. As discussed before, charge sharing loss only exists in the split phase due to explicit charge sharing in each round of reconfiguration to boost the supply voltage. During the recombine phase, the supply voltage is maintained by shifting the capacitors without explicit charge sharing. The second source of loss is the residual energy in the reservoir at the end of the entire reconfiguration process. This general analysis is performed with  $2^{(n+1)}$  unit capacitors, total capacitance  $C_{tot}$ , a battery voltage VBAT, and a minimum voltage  $V_{min}$  for proper circuit operation.

First, we analyze the charge sharing loss in the split phase. In each step of the split phase, a set of capacitors (with capacitance  $C_{left}$  and voltage  $V_{left1}$ ) is connected in series with a second set of capacitors (with capacitance  $C_{left}$  and voltage  $V_{left2}$ ), and they share charge with another set of capacitors (with capacitance  $C_{de}$  and voltage  $V_{min}$ ). Figure 3.10 shows the 3 reconfiguration steps, which we denote as  $j$ , and their sub-steps, which we denote as  $i$ , in the split phase for 16 unit capacitors ( $n=3$ ). Values of  $C_{left}$ ,  $C_{de}$ ,  $V_{left1}$ , and  $V_{left2}$  for each step  $j$  and sub step  $i$  are also shown in the figure. In general, these values are expressed in Equations 3.2, 3.3, 3.4, 3.5 for each step  $j$  and sub step  $i$ . In these equations,  $n$  is the number of steps in the split phase for  $2^{(n+1)}$  unit capacitors.

$$C_{(left,j)} = \begin{cases} \frac{C_{tot}}{4} & j = 1 \\ \frac{C_{tot}}{2^j} & 2 \leq j \leq n \end{cases} \quad (3.2)$$

$$C_{(de,j)} = \begin{cases} \frac{C_{tot}}{2} & j = 1 \\ \frac{C_{tot}}{2^j} & 2 \leq j \leq n \end{cases} \quad (3.3)$$

$$V_{(left1,ij)} = \begin{cases} V_{min} & j = 1 \\ \frac{V_{min}}{2^{j-1}} * i & 2 \leq j \leq n, 1 \leq i \leq 2^{j-2} \end{cases} \quad (3.4)$$

$$V_{(left2,ij)} = \begin{cases} V_{min} & j = 1 \\ (1 + \frac{1}{2^{j-1}}) * V_{min} - V_{left1,ij} & 2 \leq j \leq n, 1 \leq i \leq 2^{j-2} \end{cases} \quad (3.5)$$

The energy loss for the split phase is expressed in Equation 3.6, as the sum of  $E_{loss,step}$ , where  $E_{loss,step}$  (Equation 3.7) is the charge sharing loss at each sub step  $i$ . The total charge sharing loss in the split phase  $E_{loss,split}$  is then calculated in Equation 3.8. From the voltages across each of the capacitors at the end of the split phase (listed in Equation 3.9, each number represents the voltage across 2 unit capacitors), we can derive the energy remaining in the reservoir at the end of the split phase,  $E_{endstate,split}$  (Equation 3.10). Finally, the energy delivered to the load during the split phase  $E_{deliver,split}$  can be calculated using (Equation 3.11), where  $E_{stored}$  is the energy originally stored in the energy reservoir,  $E_{endstate,split}$  is the energy left at the end of split phase and  $E_{loss,split}$  is the charge sharing loss in the split phase.

$$E_{loss,split} = \sum_{j=1}^n \sum_{i=1}^{2^{j-2}} E_{loss,step}(C_{left,j}, C_{de,j}, V_{left1,ij}, V_{left2,ij}) \quad (3.6)$$

$$E_{loss,step}(C_{left,j}, C_{de,j}, V_{left1,ij}, V_{left2,ij}) = \frac{2^{1-3j}}{3} C_{tot} V_{min}^2 \quad (3.7)$$

$$E_{loss,split} = 0.05 C_{tot} V_{min}^2 + \frac{2^{-3-2n}}{9} (-4 + 4^n) C_{tot} V_{min}^2 \quad (3.8)$$

$$\left\{ \frac{V_{min}}{2^n}, \frac{2V_{min}}{2^n}, \frac{3V_{min}}{2^n}, \dots, V_{min} \right\} \quad (3.9)$$

$$E_{endstate,split} = \frac{1}{2} \frac{C_{tot}}{2^n} \sum_{m=1}^{2^n} \left(m * \frac{V_{min}}{2^n}\right)^2 \quad (3.10)$$

$$\begin{aligned} E_{deliver,split} &= E_{stored} - E_{endstate,split} - E_{loss,split} \\ &= \left(\frac{1}{2}C_{tot}V_{bat}^2\right) - \frac{1}{2} \frac{C_{tot}}{2^n} \sum_{m=1}^{2^n} \left(m * \frac{V_{min}}{2^n}\right)^2 - [0.05C_{tot}V_{min}^2 + \frac{2^{-3-2n}}{9}(-4 + 4^n)C_{tot}V_{min}^2] \end{aligned} \quad (3.11)$$

Next, we will analyze the energy non-extracted in the 1<sup>st</sup> round of the recombine phase. In this phase, the trapezoid capacitor arrays are stacked in reverse direction and repeatedly shifted in opposite directions. The capacitors with original voltage  $m \frac{V_{min}}{2^n}$  are shifted  $m$  times in total. The voltage across each capacitor is discharged by  $\frac{V_{min}}{2^{n+1}}$  with each shift. Therefore, the voltage across each capacitor at the end of the 1<sup>st</sup> round of the recombine phase is  $m \frac{V_{min}}{2^n} - m \frac{V_{min}}{2^{n+1}} = \frac{1}{2} m \frac{V_{min}}{2^n}$ , which is half of the original voltage, as shown in Equation 3.12 (each number represents the voltage across 2 unit capacitors).

$$\left\{ \frac{1}{2} \frac{V_{min}}{2^n}, \frac{1}{2} \frac{2V_{min}}{2^n}, \frac{1}{2} \frac{3V_{min}}{2^n}, \dots, \frac{1}{2} V_{min} \right\} \quad (3.12)$$

By folding the trapezoids in the 2<sup>nd</sup> round of the recombine phase, more energy can be extracted. The voltage across each capacitor at the end of the 2<sup>nd</sup> round of recombine phase (Equation 3.13) can be derived by writing out the start and end state of each step. The energy remaining at the end of this phase is expressed in Equation 3.14. This is the final step in the counter flow energy reservoir. Therefore,  $E_{endstate,2^{nd}recombine}$  shown in Equation 3.14 is the energy left non-extracted at the end of the process, which is 5% for 16 unit capacitors and 200mV voltage drop. Since there is no charge sharing loss in the recombine phase, this residual energy

$E_{endstate,2^{nd}recombine}$  is the only loss in the recombine phase. The energy delivered for 2 rounds of the recombine phase can be easily calculated in Equation 3.15, as the difference between the energy left at the end of split phase  $E_{endstate,split}$  and the energy left at the end of the 2<sup>nd</sup> round of recombine phase  $E_{endstate,2^{nd}recombine}$ .

$$\left\{ \frac{1}{4} \frac{V_{min}}{2^n}, \frac{1}{4} \frac{2V_{min}}{2^n}, \frac{1}{4} \frac{3V_{min}}{2^n}, \dots, \frac{1}{2} (2^{n-1} + 1) \frac{V_{min}}{2^n} - \frac{1}{4} \frac{V_{min}}{2^n}, \frac{1}{2} (2^{n-1} + 2) \frac{V_{min}}{2^n} - \frac{2}{4} \frac{V_{min}}{2^n}, \dots, \frac{1}{2} 2^n \frac{V_{min}}{2^n} - \frac{1}{4} 2^{n-1} \frac{V_{min}}{2^n} \right\} \quad (3.13)$$

$$E_{endstate,2^{nd}recombine} = \frac{1}{2} \frac{C_{tot}}{2^n} \left[ \sum_{m=1}^{2^{n-1}} \left( \frac{1}{4} m \frac{V_{min}}{2^n} \right)^2 + \sum_{m=2^{n-1}+1}^{2^n} \left( \frac{1}{2} m \frac{V_{min}}{2^n} - \frac{1}{4} (m - 2^{n-1}) \frac{V_{min}}{2^n} \right)^2 \right] \quad (3.14)$$

$$\begin{aligned} E_{deliver,recombine} &= E_{endstate,split} - E_{endstate,2^{nd}recombine} \\ &= \frac{1}{2} \frac{C_{tot}}{2^n} \left[ \sum_{m=1}^{2^n} \left( m \frac{V_{min}}{2^n} \right)^2 - \sum_{m=1}^{2^{n-1}} \left( \frac{1}{4} m \frac{V_{min}}{2^n} \right)^2 \right. \\ &\quad \left. - \sum_{m=2^{n-1}+1}^{2^n} \left( \frac{1}{2} m \frac{V_{min}}{2^n} - \frac{1}{4} (m - 2^{n-1}) \frac{V_{min}}{2^n} \right)^2 \right] \end{aligned} \quad (3.15)$$

Finally, the single shot energy delivery efficiency for the entire reconfiguration process is expressed in Equation 3.16, where a, b, and c are positive, topology-dependent constants; c is approximately 0.18. As shown in Equation 3.16, c is a topology-dependent scaling factor which indicates how sensitive efficiency is to  $\frac{V_{min}}{V_{BAT}}$ <sup>2</sup>. One major contribution of c comes from the charge sharing loss in the 1<sup>st</sup> step of split phase (step j=1 in Figure 3.10), due to the lack of reconfiguration choices. As shown in this equation and Figure 3.11, the efficiency of the proposed energy reservoir is relatively independent (for n>1) of the total number of discrete capacitors, 2<sup>n+1</sup>. However

efficiency is more dependent on  $V_{min}$  than on number of capacitors. To compare the efficiency gain of the proposed reservoir over the conventional single capacitor method, we calculate the efficiency using the conventional method in Equation 3.17. The efficiency gain of the proposed method is expressed in Equation 3.18 and plotted in Figure 3.12 for 16 unit capacitors over  $\% \Delta V$ , which is the relative voltage drop between  $V_{min}$  and  $V_{BAT}$ . The efficiency gain is  $16.7\times$  with 2.5% voltage drop (100 mV with  $V_{BAT} = 4V$ ).

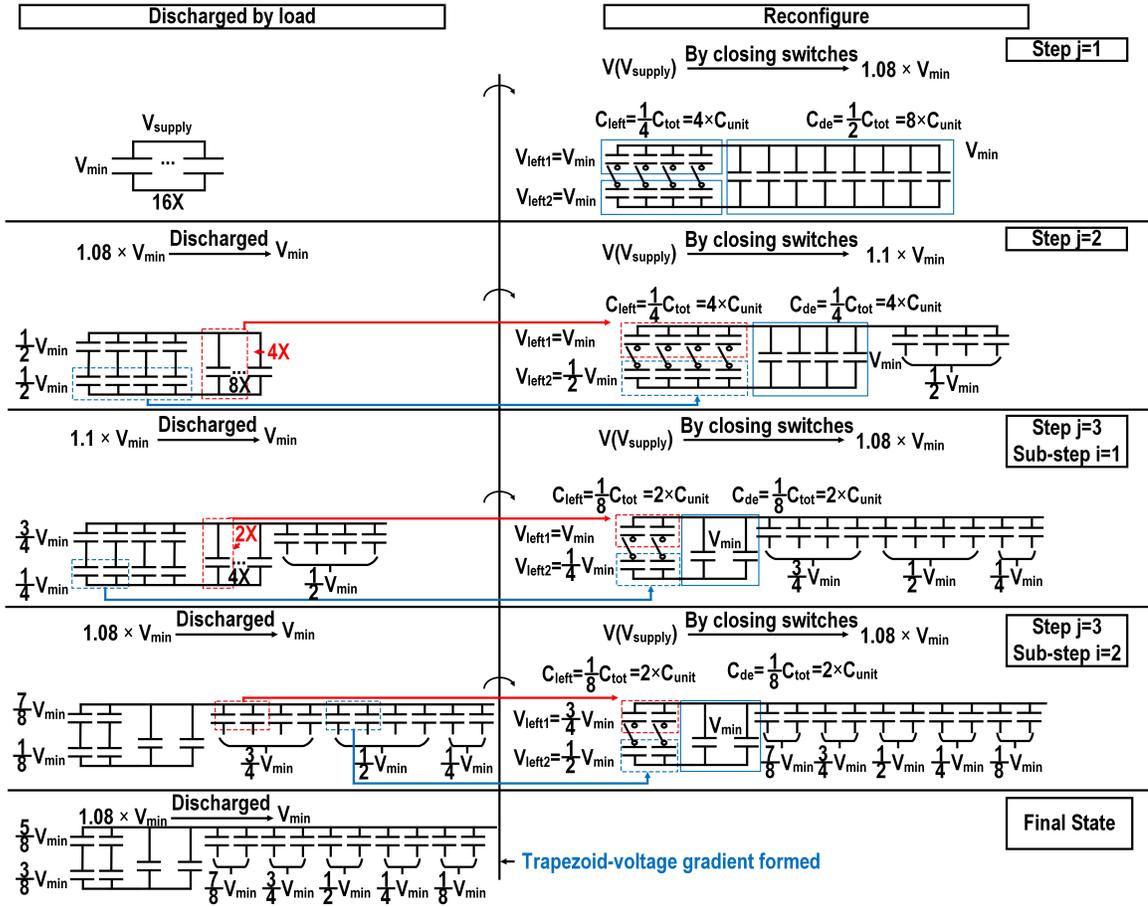


Figure 3.10: Illustration of steps in split phase for 16 unit capacitors

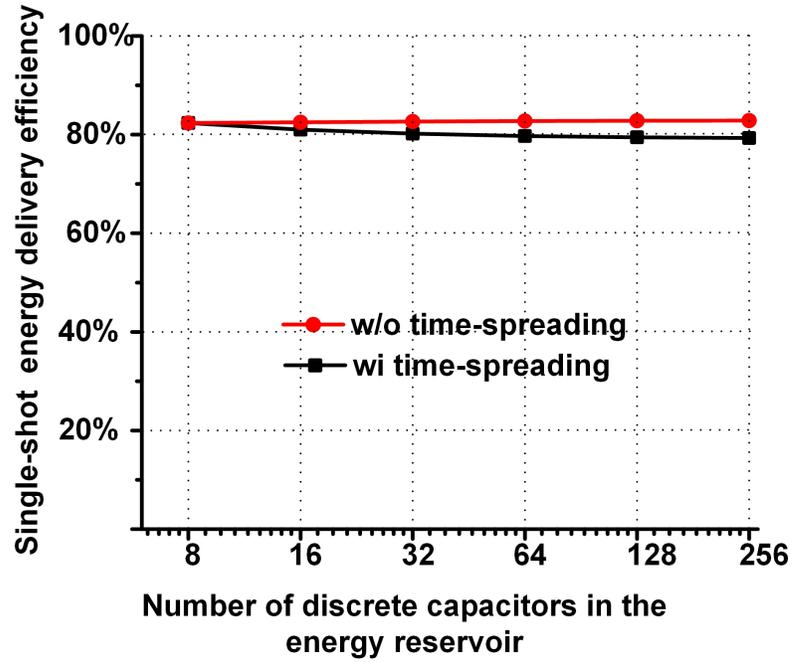


Figure 3.11: Efficiency of counter flow energy reservoir over number of discrete capacitors

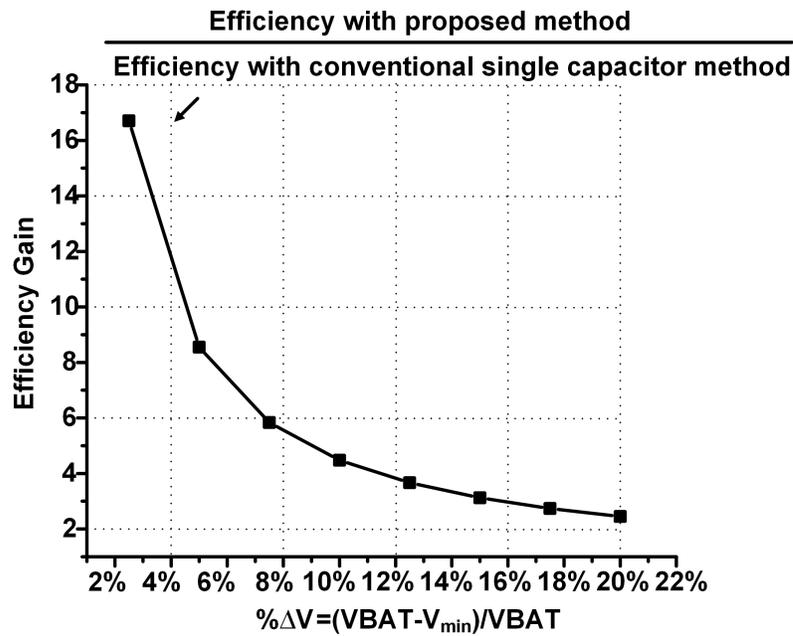


Figure 3.12: Efficiency gain of the proposed method across relative allowable supply voltage drop

$$\begin{aligned}
Eff_{single-shot,proposed} &= \frac{E_{deliver,split} + E_{deliver,recombine}}{E_{stored}} \\
&= 1 - \left(\frac{V_{min}}{VBAT}\right)^2 * (c + b * 2^{-n} - a * 4^{-n}) \\
&\simeq 1 - c\left(\frac{V_{min}}{VBAT}\right)^2
\end{aligned} \tag{3.16}$$

$$Eff_{single-shot,conventional} = \frac{\frac{1}{2}C_{tot}(VBAT^2 - V_{min}^2)}{\frac{1}{2}C_{tot}VBAT^2} = 1 - \frac{V_{min}^2}{VBAT^2} \tag{3.17}$$

$$Gain \simeq \frac{1 - c\left(\frac{V_{min}}{VBAT}\right)^2}{1 - \frac{V_{min}^2}{VBAT^2}} \tag{3.18}$$

### 3.2.3 Voltage overshoot analysis

After each reconfiguration step in the proposed energy reservoir, the supply voltage is boosted. In the split phase, the boosted voltage in each step is denoted as  $V_{boost,j}$ , where  $j$  is the step number. Without any regulation, the maximum supply voltage overshoot occurs in the 1<sup>st</sup> step and is denoted as  $V_{max,unregulated,split} = \frac{6V_{min}}{5}$ . To reduce  $V_{max,unregulated,split}$ , time-spreading is used. As shown in Figure 3.13, instead of stacking 2 sets of big capacitors and over boosting the supply voltage in one step, we stack smaller sets of capacitors to boost the supply voltage in multiple steps. In this way, smaller capacitor decks  $C_{left}$  are charge shared with  $C_{de}$  to reduce the resulting voltage overshoot after charge sharing. The maximum boosted supply voltage using time-spreading is denoted as  $V_{max,ts,split}$  and shown in Equation 3.19.

In the recombine phase, the maximum supply voltage overshoot  $V_{max,recombine}$  is shown in Equation 3.20. Figure 3.14 shows  $V_{max,ts,split}$ ,  $V_{max,recombine}$  and  $V_{max,unregulated,split}$  over the total number of capacitors  $2^{1+n}$ , and voltage overshoot is reduced from 20% to <6% with time-spreading (when the total number of capacitors is 32).

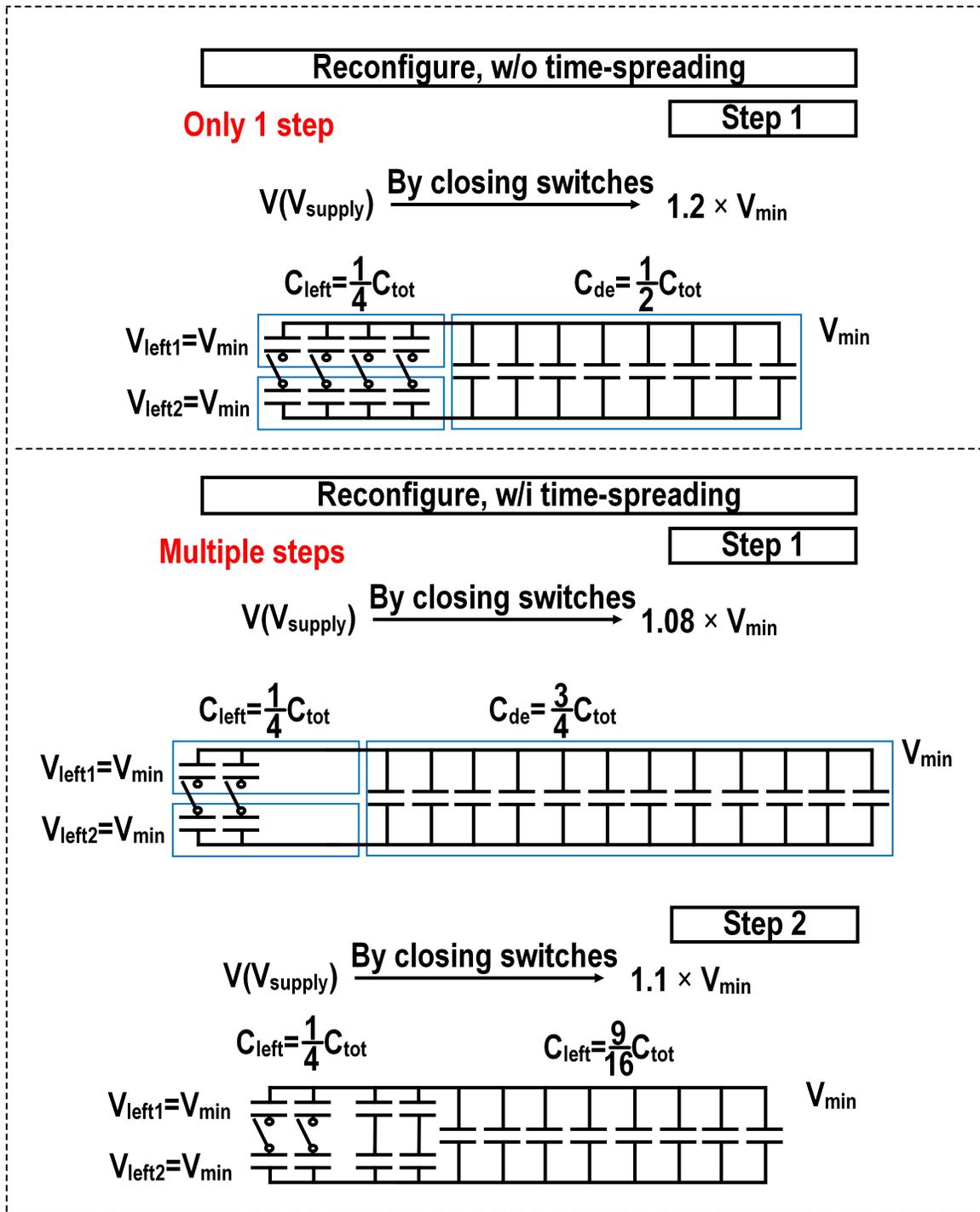


Figure 3.13: Operation concept of time-spreading technique

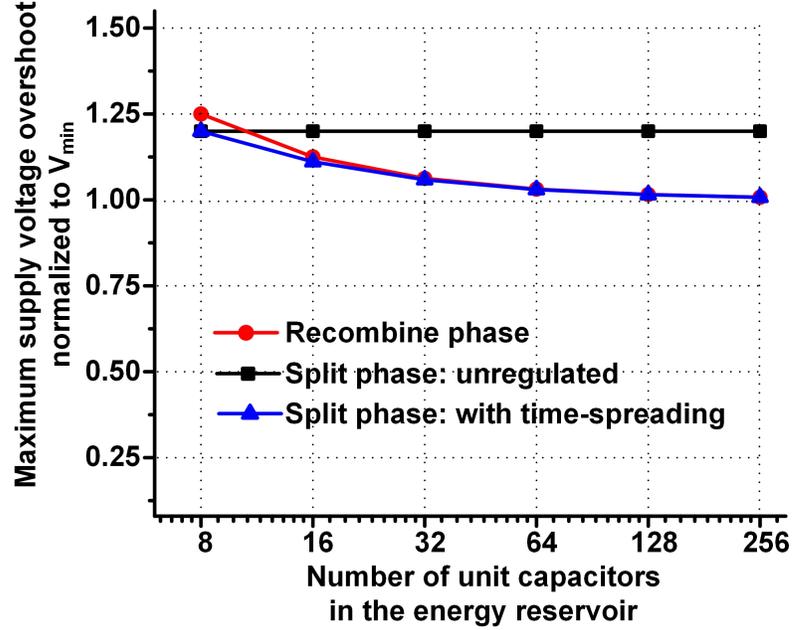


Figure 3.14: Maximum supply voltage overshoot wi/wo time-spreading across number of capacitors

$$V_{max,ts,split} = \left(1 + \frac{1}{1 + 2^n}\right)V_{min} \quad (3.19)$$

$$V_{max,recombine} = \left(1 + \frac{1}{2^n}\right)V_{min} \quad (3.20)$$

It is also important to note that time-spreading lowers the supply voltage overshoot at the cost of slightly higher charge sharing loss. As shown in Figure 3.11, the resulting efficiency is still  $>79\%$ , only 2% lower compared with the efficiency achieved without using time-spreading. Efficiency degradation is slightly overestimated here because time-spreading is applied to all of the steps of the split phase. In a real implementation, only the first few steps require time-spreading.

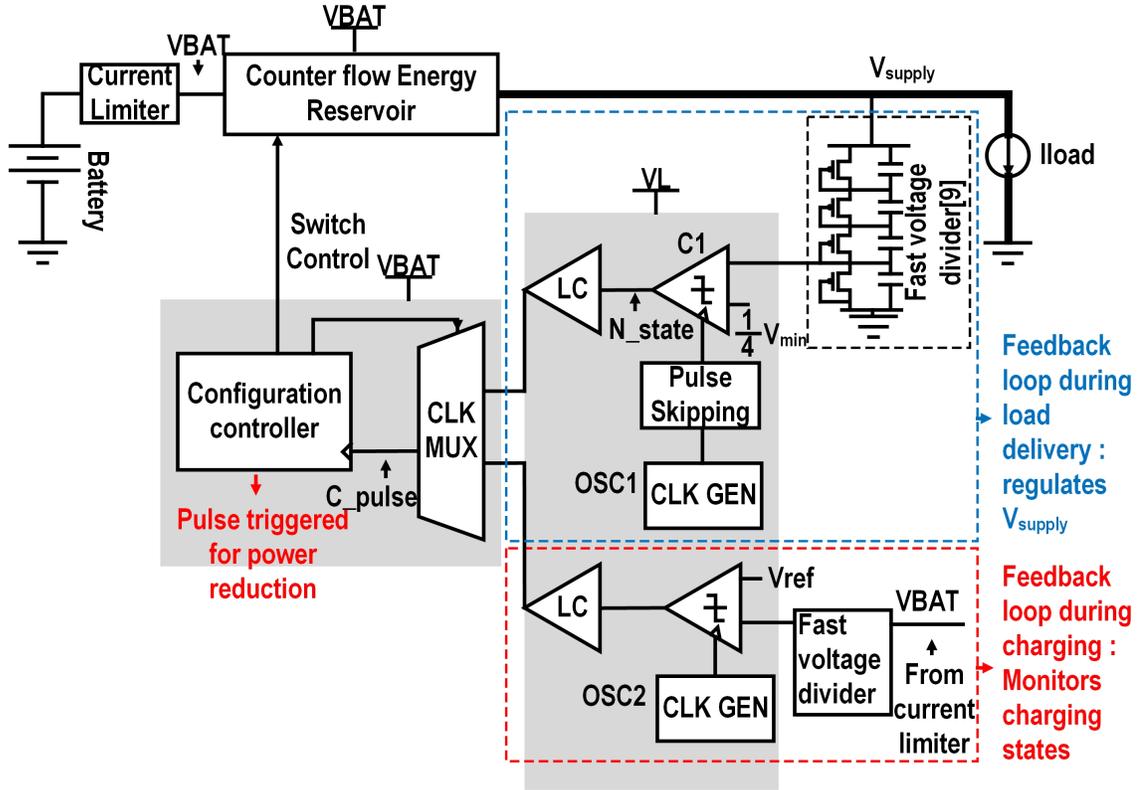


Figure 3.15: Top-level architecture of the implemented counter flow energy reservoir

### 3.3 Implementation of Counter Flow Energy Reservoir

Figure 3.15 shows the top level architecture of the design, consisting of the counter flow energy reservoir, a feedback loop for delivery modulation, a feedback loop for charging modulation, and a configuration controller. When the energy reservoir is in delivery mode, the load is enabled, and  $V_{supply}$  is monitored using the fast voltage divider, which combines a resistive and capacitive voltage divider for fast response time. Capacitors in the fast voltage divider are sized to mitigate the effect of parasitic capacitance at the output node. Leakage power of this structure is 1.9nW in simulation. When  $V_{supply}$  drops below  $V_{min}$ , the comparator C1, clocked by clock generator OSC1, triggers a pulse  $N\_state$ . The configuration controller is an unconditional pulse-based state machine, which proceeds through pre-programmed states

on each rising edge of  $N\_state$  and generates the reconfiguration control signals. In charging mode, the reservoir energy is restored by reversing the steps of the recombine phase, which reduces the voltage difference seen by the current limiter to achieve a much lower charging loss ( $3\times$  lower) than that resulting from directly charging all of the capacitors. The charging state is again monitored by a clocked comparator. The comparators and clock generation operate at 1.2V to reduce dynamic power (measured  $5.2\mu W$  for  $700\mu W$  output power). Static power consumption of the proposed energy reservoir is assumed to be negligible compared to  $>100\mu W$  designed output power. A pulse-skipping module skips clock cycles immediately after  $C\_pulse$  triggers a configuration change, allowing time for the energy reservoir to restore  $V_{supply}$  and thus avoiding false  $C\_pulse$  edges.

It should be noted that the 1.2V supply in this implementation is generated off-chip for prototype verification. In a more realistic system implementation, this voltage can be generated on-chip by power management unit (PMU). Since many sensor systems require PMU to generate voltages lower than the battery voltage for efficient operation [45] [53], the efficiency and area degradation of generating an extra voltage may be mitigated by sharing the 1.2V supply with already existing voltage domains in the system.

The topology used for current limiter is shown in Figure 3.16, which composed with a resistor array and 8 selection switches for tuning. It should be noted that this structure cannot eliminate reverse current from the energy reservoir to the battery when  $V_{supply}$  is higher than battery voltage (simulated peak current is 60nA when  $V_{supply}$  is 300mV higher than battery voltage). A reverse current protection unit can be implemented to completely turn off PMOS selection switches during charge delivery.

There are 16 unit capacitors used in the implemented energy reservoir, and each of them is 0.197nF. All capacitors are implemented using MIM capacitors. Figure 3.17

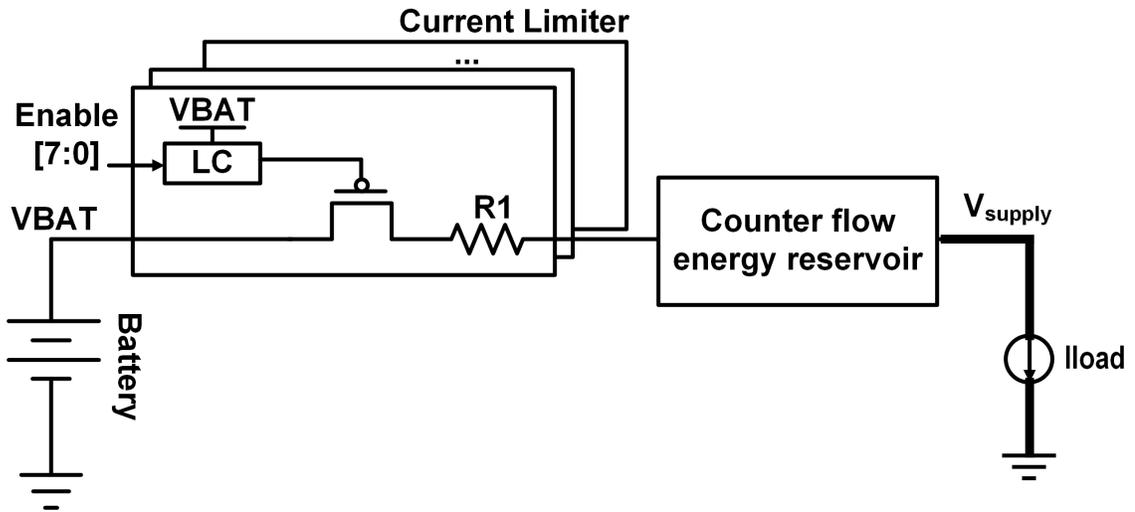


Figure 3.16: Circuit implementation of the current limiter

illustrate the switch connections. There are in total 119 switches used to configure the capacitor array. There are 22 switches used in split phase, 60 switches used in 1<sup>st</sup> round of recombine phase, 13 switches used in 2<sup>nd</sup> round of recombine phase, and 24 switches used for supply and ground connections.

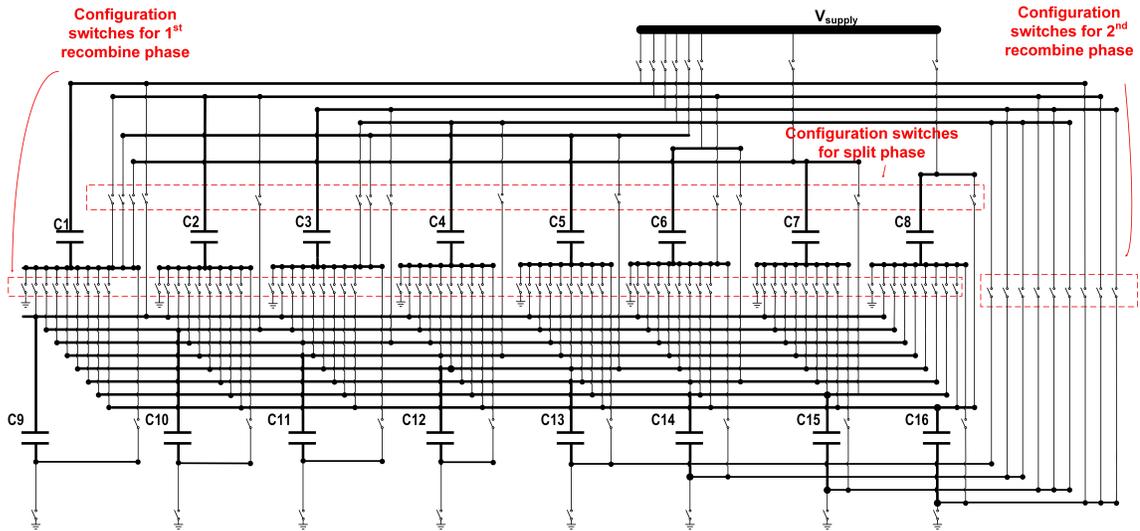


Figure 3.17: Illustration of switch connections in the energy reservoir

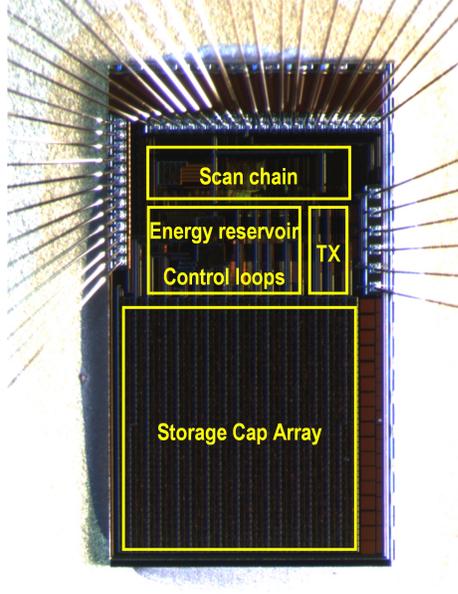


Figure 3.18: Die photo

### 3.4 Measurements

The test chip shown in Figure 3.18 is fabricated in 180-nm CMOS, and the die area is 3.8 mm<sup>2</sup>. The total capacitor is 3.15 nF with 16 unit capacitors, and the control loop area overhead is 18%. We implemented the idea proposed in [13] to compare the performance gain of the proposed method over the single storage capacitor method [13].

Figure 3.19 shows the captured supply voltage waveform for a load power of 1.4 mW,  $V_{\min}$  setting of 3.6V and VBAT of 3.8V. The waveform shows 8× longer high-current delivery time compared with the conventional single storage capacitor method (i.e. the 1<sup>st</sup> spike on the waveform labeled as conventional single storage capacitor method). It should be noted that the dip of supply voltage in 2<sup>nd</sup> round of recombine phase is due to the inability of the energy reservoir to be reconfigured fast enough at heavy load, causing a degradation of energy delivery efficiency. In the 2<sup>nd</sup> round of recombine phase, 4 capacitors are connected in series, making the equivalent capacitance the smallest, and therefore the time allowed for reconfiguration is the smallest.

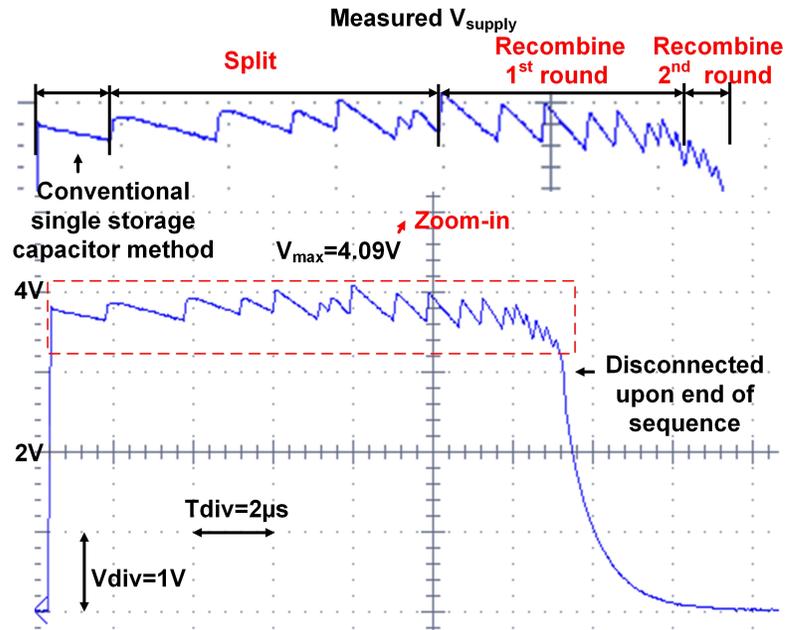


Figure 3.19: Captured supply voltage waveform

### Energy breakdown of energy reservoir

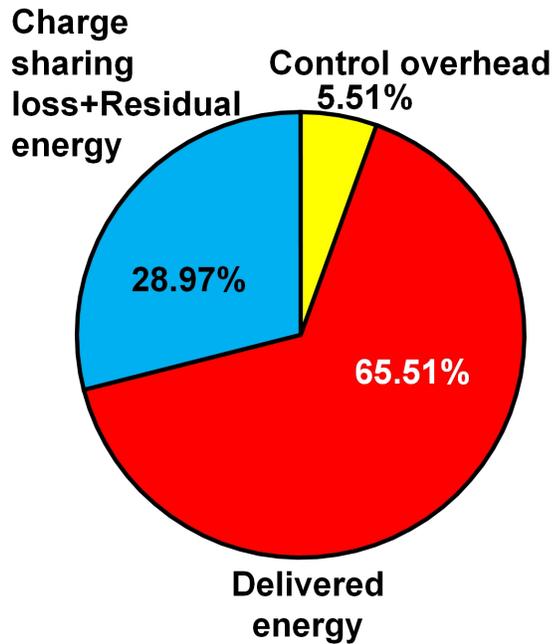


Figure 3.20: Measured energy breakdown

Figure 3.20 depicts the measured energy breakdown, which shows that comparator and control overhead is 5.5% (power overhead is  $59\mu\text{W}$  when delivering  $700\mu\text{W}$  output power, at average output voltage of 3.71V), and the measured charge sharing loss and residual energy loss is 28.97%.

In Figure 3.21, we quantify the performance of the reservoir using single shot energy delivery efficiency, as defined in Section II. We measure it over the allowable voltage drop  $\Delta V$  on the left. The 1<sup>st</sup> line from the top is the energy stored originally. The 2<sup>nd</sup> and 3<sup>rd</sup> lines are the energy extracted using the proposed method, and the 4<sup>th</sup> and 5<sup>th</sup> lines are the energy extracted using the conventional single storage capacitor method. The 3<sup>rd</sup> line shows that the reservoir extracts 17.5 nJ before recharging, representing an up to  $12\times$  improvement over the conventional single storage capacitor method. We also measure the single shot energy delivered over load power on the right. The 3<sup>rd</sup> line shows that the reservoir maintains 62% efficiency across  $45\mu\text{W}$  to 8 mW load power. The discrepancy between the measured and theoretical calculated energy extraction is caused by factors including decoupling capacitance at supply voltage, switching loss of power switches and controllers, parasitic bottom capacitance of the capacitor array, and possible incorrect reconfiguration timing. It should be noted that the single shot efficiency drops at heavy load power. There are 2 main reasons. 1) As load power goes up,  $V_{\text{supply}}$  drops faster than the reservoir controller can catch up, causing the capacitors to be reconfigured at non-optimal time. This inaccurate timing leads to insufficient energy extraction. 2) On-resistance of the reconfiguration switches causes large voltage drop at heavy load, and therefore capacitors connected to the switches cannot be fully discharged to the intended voltages. Hence the resulting efficiency is lower than expected. This degradation may be improved by using a more advanced technology. Figure 3.22 shows the measured single shot energy delivered at different temperatures.

In Figure 3.23, we measure the end-to-end efficiency, which is the ratio of the

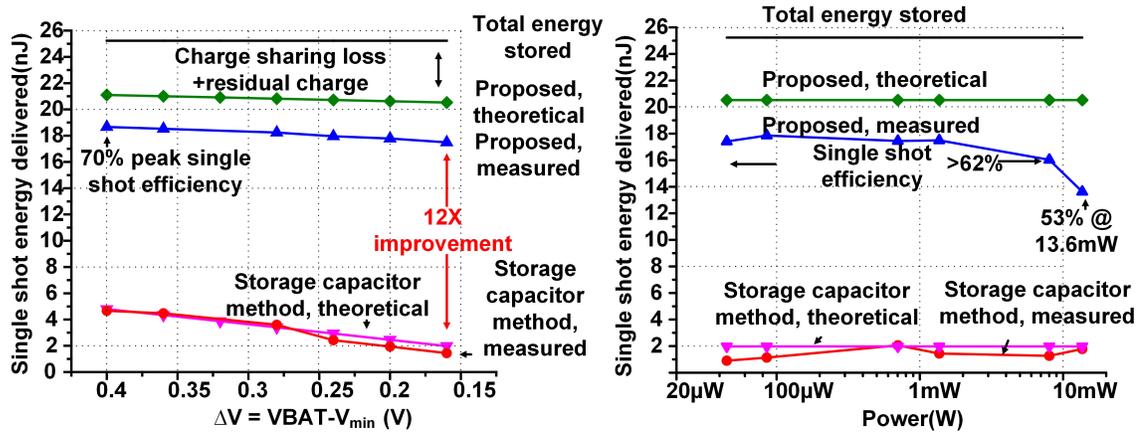


Figure 3.21: Single shot energy delivered across allowable voltage drop (left) and load power (right)

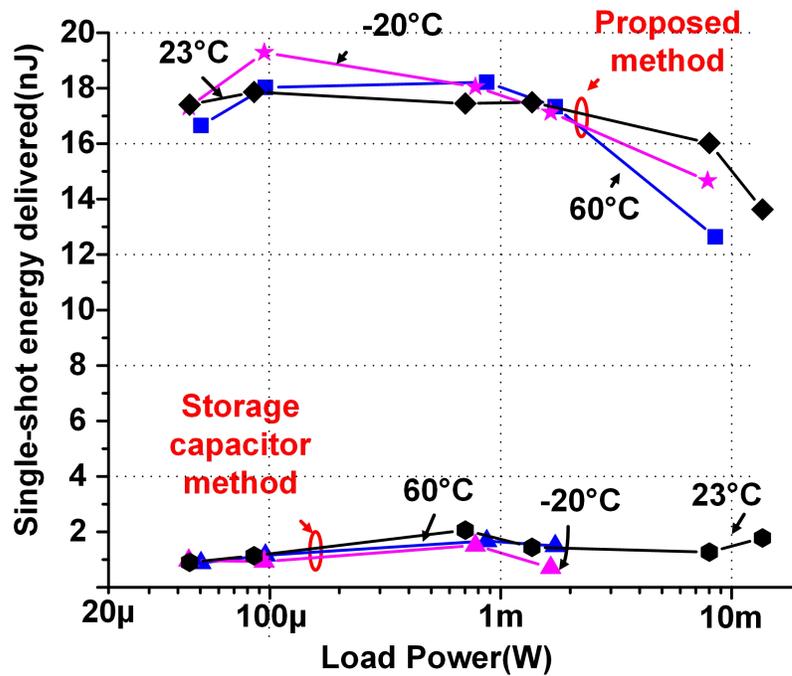


Figure 3.22: Single shot energy delivered at different temperatures

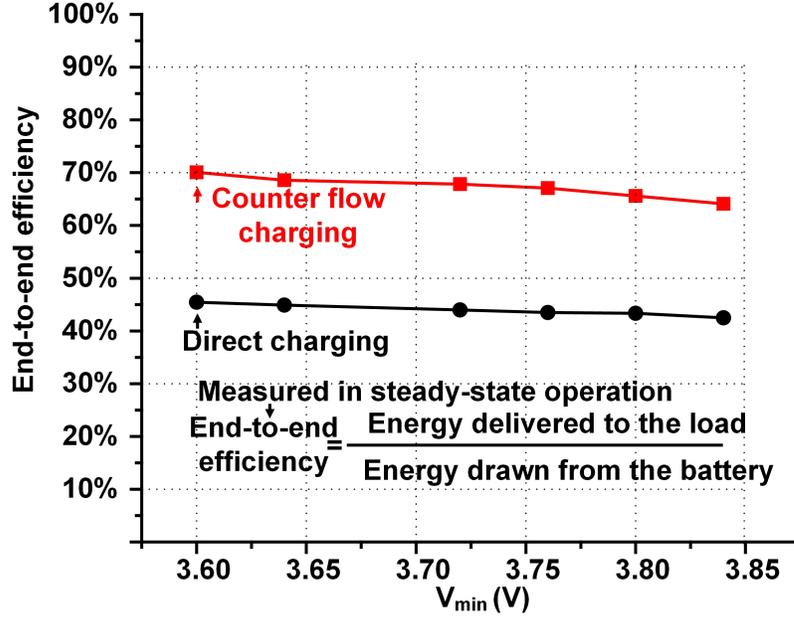


Figure 3.23: End-to-end efficiency of the proposed energy reservoir

energy delivered to the load to the energy supplied by the battery. The top line shows the end-to-end efficiency achieved by reversing the steps in the recombine phase during charging. The bottom line shows the efficiency achieved by connecting the reservoir to the battery directly. The proposed counter flow charging method improves the end-to-end efficiency from 45% to 70%.

Figure 3.24 shows the captured waveforms using counter flow charging and discharging. By reversing the steps in recombine phase, capacitors with trapezoid voltage gradients are stacked in reverse directions, and shift in opposite directions whenever the stacked voltage  $V_{\text{supply}}$  is charged to battery voltage  $V_{\text{BAT}}$ . The top waveform shows the charging process of one of the unit capacitors in the energy reservoir. Each small steps represent a shift of trapezoid stacks in opposite directions. By charging in small steps like walking up a ladder, voltage difference seen across the power switches in each step is reduced, and therefore the resulting charging efficiency is higher than directly connect all capacitors in parallel with the battery.

In Figure 3.25, we integrate the test chip with a transmitter as load, which is

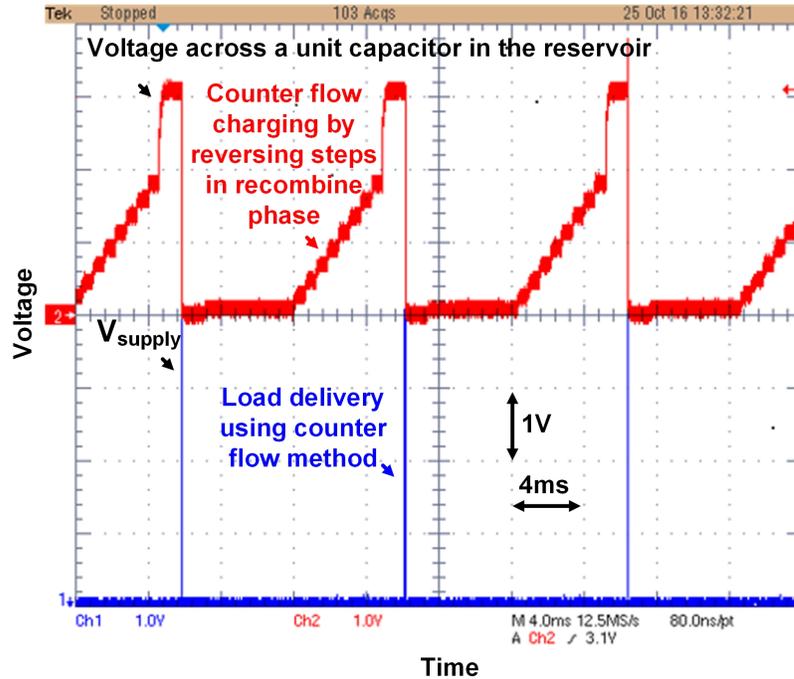


Figure 3.24: Captured waveform showing counter flow charging

connected with an inductive antenna. In Figure 3.26, the captured transmitter output pulse, shown on the top, demonstrates  $11.5\times$  longer continuous transmission than conventional single capacitor method (i.e. 1<sup>st</sup> spike representing energy delivered without configurability) with radio power 2 mW and duration of  $12.4\mu\text{s}$ . The captured supply voltage waveform on the bottom shows supply voltage ramping up at radio power-on. A zoom-in of the supply voltage waveform is shown on the right, with each spikes in split phase labeled with step numbers 1–3 corresponding to step  $j=1-3$  in Figure 3.10. Step 0 is when all capacitors are connected in parallel and discharged by the load as a single capacitor. This correspond to the conventional single storage capacitor method. For step 1 and 2, two small spikes are seen in each step because time-spreading technique is used.

Table 3.1 summarizes that the proposed reservoir can deliver 18.7 nJ with 10% supply voltage drop.

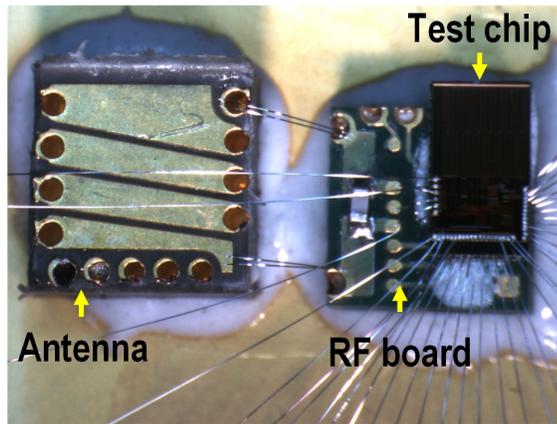


Figure 3.25: Integration with radio

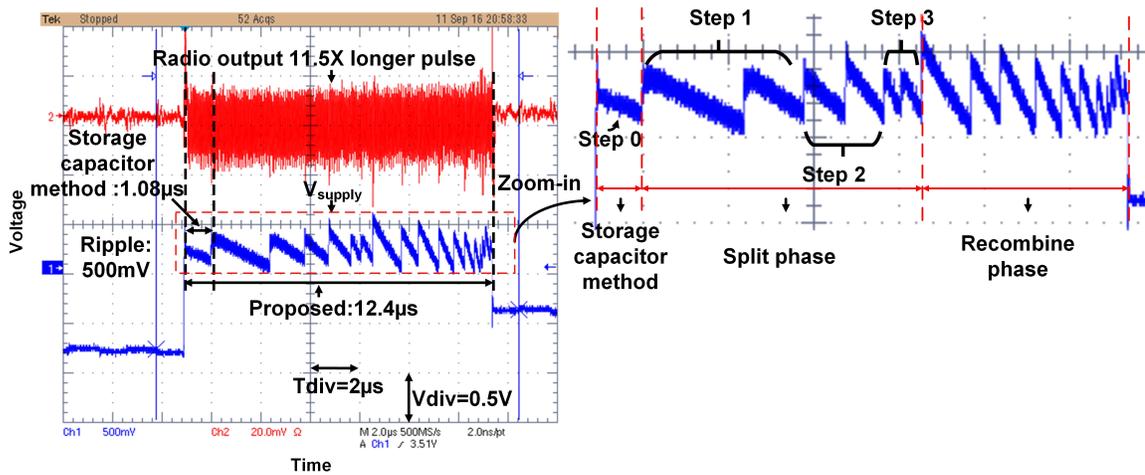


Figure 3.26: Captured transmitter output pulse and supply voltage waveform

<b>Technology</b>	<b>0.18µm CMOS</b>
<b>Chip area</b>	<b>3.8mm<sup>2</sup></b>
<b>Fully integrated</b>	<b>Yes</b>
<b>Capacitor size</b>	<b>3.15nF</b>
<b>Output power range</b>	<b>45µW – 13.6mW</b>
<b>Control overhead</b>	<b>5.5%</b>
<b>Single shot energy delivered</b>	<b>18.7nJ @ 10% V<sub>supply</sub> drop</b>

Table 3.1: Chip characteristic summary

### 3.5 Conclusion

In conclusion, we presented an energy reservoir that dynamically reconfigures a storage capacitor array using a so-called counter flow approach for large single shot energy output at high power. The reservoir achieves  $45\mu\text{W}$  to  $13.6\text{mW}$  output power range and 70% peak single shot energy delivery efficiency with 10% voltage drop (Table 3.1). The proposed method consists of a split phase, where a trapezoid voltage gradient is formed across capacitor arrays, and a recombine phase, where the capacitor arrays are stacked in series and shifted in opposite directions to achieve energy extraction with no charge sharing loss.

## CHAPTER IV

# A $0.04\text{mm}^3$ $16\text{nW}$ Wireless and Batteryless Sensor System with Integrated Cortex-M0+ Processor and Optical Communication for Cellular Temperature Measurement

### 4.1 Introduction

Monitoring cellular temperature, as an indicator of cellular metabolism, is highly beneficial for disease study and drug discovery, as many diseases (e.g., cancer) are characterized by abnormal metabolism. Recently scientists have achieved passive temperature mapping inside living cells using fluorescent materials [55] with limited accuracy of  $1.3^\circ\text{C}$  and  $0.58^\circ\text{C}$  resolution. This method nevertheless leads to the discovery that mitochondria are  $10^\circ\text{C}$  higher than in other parts of a cell [56]. Silicon implementation of accurate, autonomous sensor systems for cell cluster temperature measurement is lacking and can facilitate further biological discoveries. Direct measurement of such cellular temperatures is extremely challenging since it requires highly localized measurements. Cellular sensor size cannot exceed  $0.1\text{ mm}^3$  to achieve good spatial resolution, making prior miniature implantable sensor systems (typically several  $\text{mm}^3$ ) [14] [57] [58] impractically large.

**Proposed programmable Cellular Temperature Sensing System (CTS) –500 × smaller than rice**

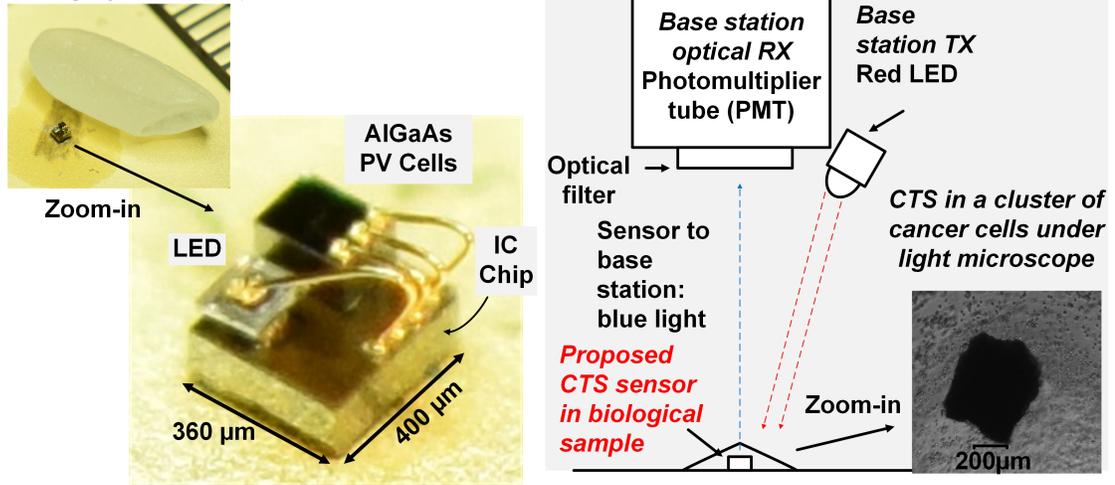


Figure 4.1: CTS encased with bio-compatible material and implanted in a cluster of homogeneously dispersed HS5 human bone marrow stromal cells

This aggressive size constraint for cellular sensor systems ( $<0.1\text{mm}^3$ ) creates two major design challenges: 1) Efficient wireless communication to program the processor and retrieve data is very difficult given the sub-mm area constraint. RF antenna efficiency degrades quickly with antenna size, forcing very high carrier frequencies (and correspondingly high power circuits and mm TX distance [59]). The proposed CTS [62] uses optical communication since transmitter and receiver elements (LED and PV diodes) readily scale to tens of  $\mu\text{m}$  without efficiency loss. 2) Temperature-independent frequency and voltage references are critical for communication synchronization and high accuracy temperature sensing. However, crystals are far too large and bandgaps too power hungry for a sub-mm sensor. Hence, CTS uses a base-station generated clock reference encoded with the optical link, enabling reliable communication over 15.6cm and temperature measurement using a subthreshold oscillator to achieve a high accuracy of  $+0.11/-0.08^\circ\text{C}$  and  $0.034^\circ\text{C}$  RMS resolution.

## 4.2 Cellular Temperature Sensing System

Figure 4.1 shows the CTS, which integrates a commercial Cree LED for optical transmission, custom  $50 \times 50 \mu\text{m}$  AlGaAs diode for optical reception, and  $180 \times 230 \mu\text{m}$  custom AlGaAs diode for power harvesting on the top layer. The bottom layer of CTS is a custom chip ( $360 \times 400 \times 150 \mu\text{m}$ ) in 55nm CMOS (MIFS C55DDC) including a M0+ processor with full programmability, subthreshold oscillation based temperature sensor [60], TX and RX circuits, LED drivers, and custom SRAM. A Photomultiplier Tube (PMT) in the base station senses transmitted data from the sensor node (Figure 4.1) and includes an optical filter to remove self-interference. Since cellular-level temperature measurement is typically performed in a controlled laboratory environment, lighting conditions can be restricted to wavelengths that limit interference. The always-on base station supplies modulated light (615nm) to power the battery-less sensor node and supply an accurate clock. CTS operates at 3klux with 16nW system power consumption (including TX and temperature sensor). We verified full autonomous, wireless system operation with the complete stack shown in Figure 4.1. Its measured system operation (Figure 4.2) shows boot-up, default program operation, wireless programming by the base station, temperature measurement with on-chip recovered accurate clock, transmission of temperature codes through sensor node LED, and successful demodulation of the correct packet at the base station (Figure 4.2).

Figure ?? 4.2 show the CTS architecture and captured operation sequence. When the base station sends only DC light, CTS enters a power-on mode in which it executes a default program stored in a register file. To program the CTS, the base station sends Manchester-coded modulated light, which is received by the integrated photodiode, canceled for ambient light, and demodulated [61]. Once CTS recognizes the password, it shifts its system clock source to the recovered accurate base station clock. The system then stores the received program in a 4Kb SRAM optimized for static power

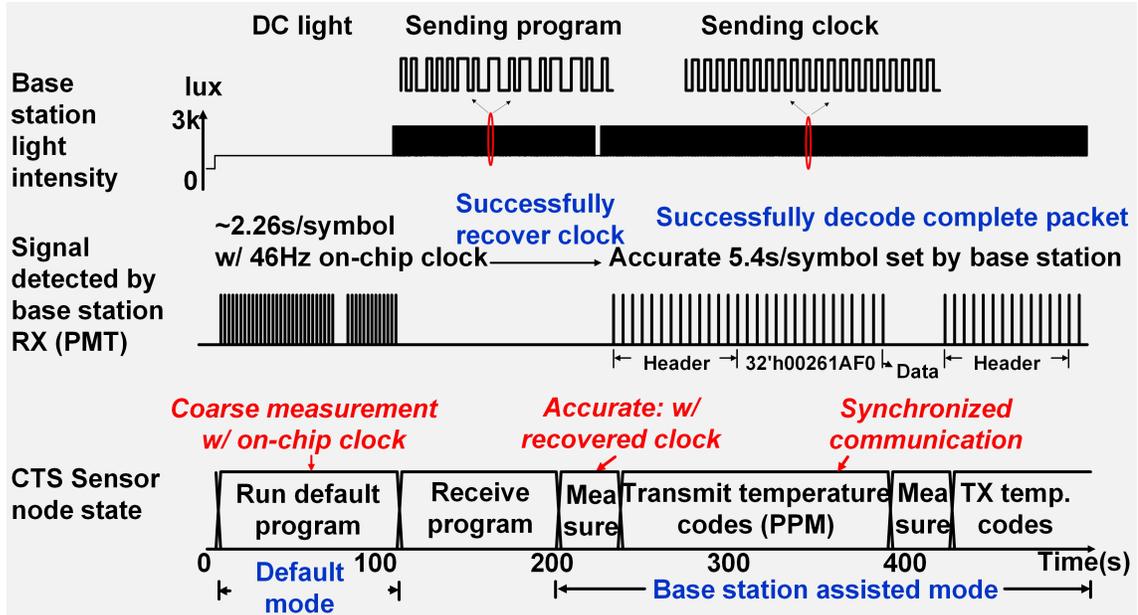


Figure 4.2: Measured waveform with fully assembled CTS system

reduction and activates the M0+ processor for program execution. Our sensors were programmed to take temperature measurements, store them and then transmits data with pulse position modulated light signals via the integrated LED at 180pJ/bit (simulated) using energy accumulated on 100pF on-chip capacitor C1.

### 4.3 Circuit Block Implementation

Figure 4.4 shows the transmitter circuit implementation. A charge pump accumulates charge harvested from the photovoltaic (PV) cell on the on-chip capacitor C1, which then supplies energy to the LED with regulated current and accurate timing dictated by a PPM modulator. Each LED flash sends out a 2-bit symbol. Regulated LED current is optimized for minimum energy per bit. A voltage regulation loop controls VLED\_Anode on C1 to prevent voltage overshoot. As shown at the bottom of Figure 4.4, the regulation loop divides the voltage on VLED\_Anode with a charge-sharing voltage divider and compares the divided voltage  $V_{cs}$  with the on-chip

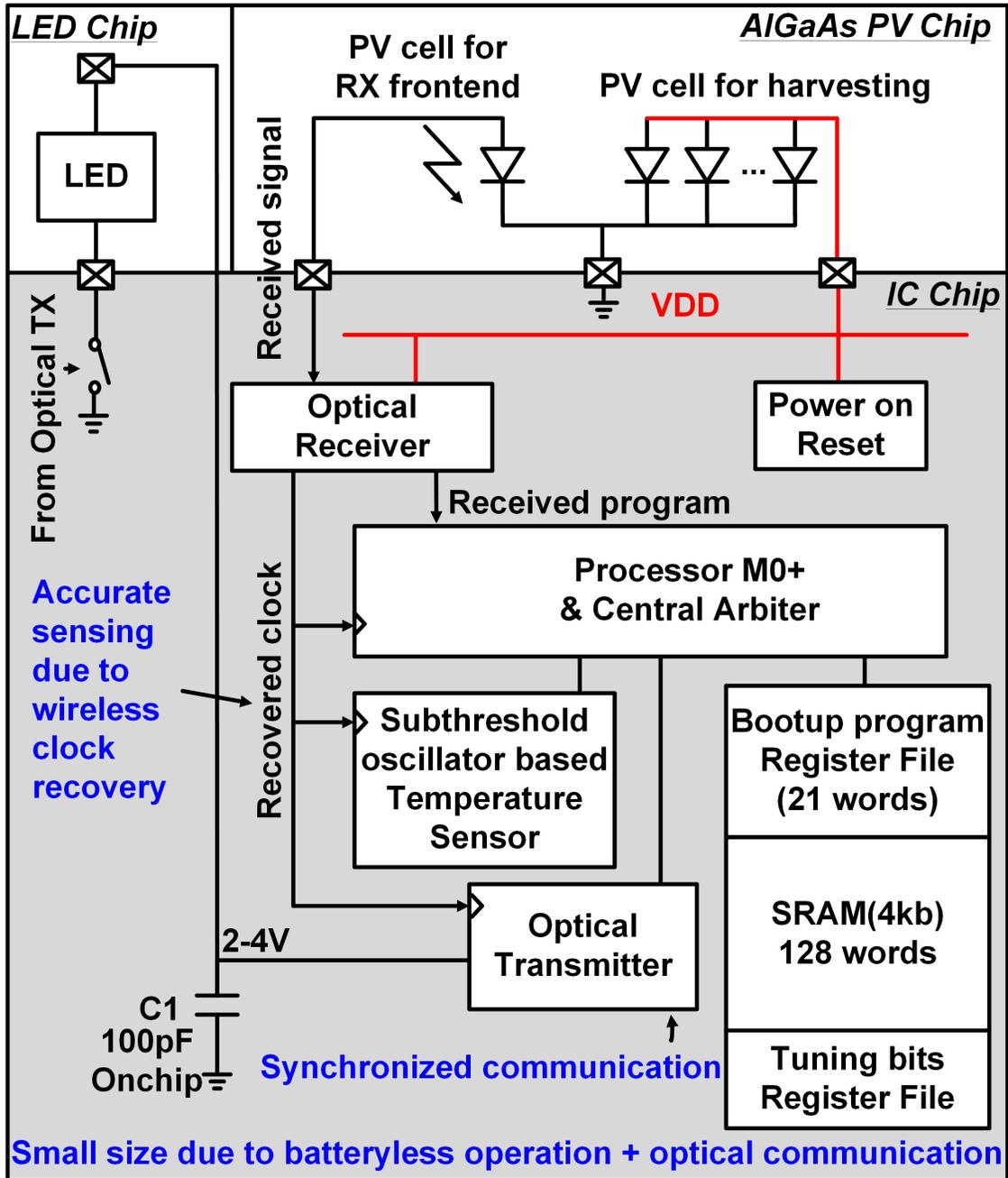


Figure 4.3: System architecture of CTS

generated reference voltage  $V_{ref}$ . The charge pump is clock gated when  $V_{cs} > V_{ref}$ .

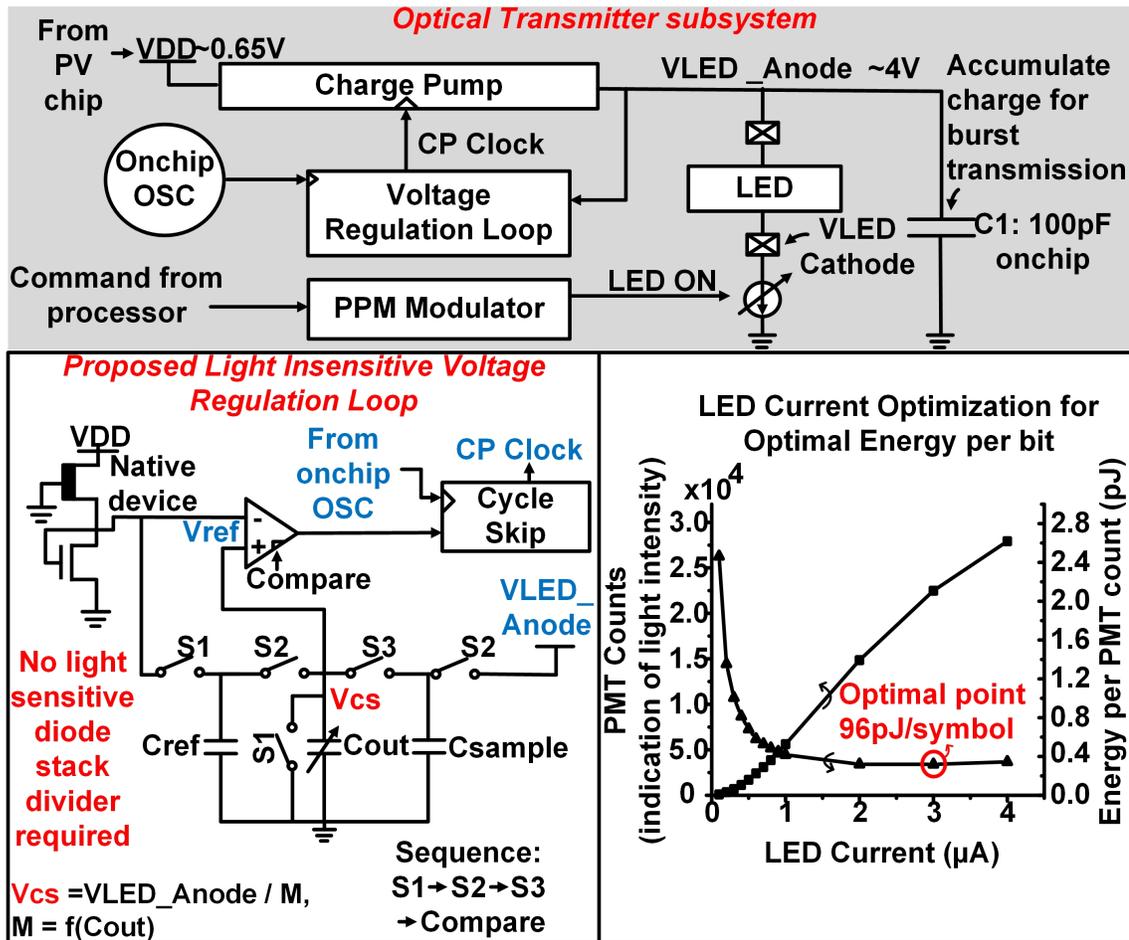


Figure 4.4: Circuit implementation of optical transmitter subsystem

A key design consideration for the IC layer is light exposure as coating with a light blocking epoxy is not feasible in the required form factor. This led to different design decisions than in other ultra-low power systems, e.g., the voltage divider in Figure 4.4 uses capacitive charge sharing instead of a conventional diode stack divider to avoid inaccuracies introduced by photo-generated current from parasitic P-N junctions in diode stacks under light exposure. Similarly, the voltage reference providing  $V_{ref}$  is sized to have a bias current larger than the photogenerated currents to ensure robust

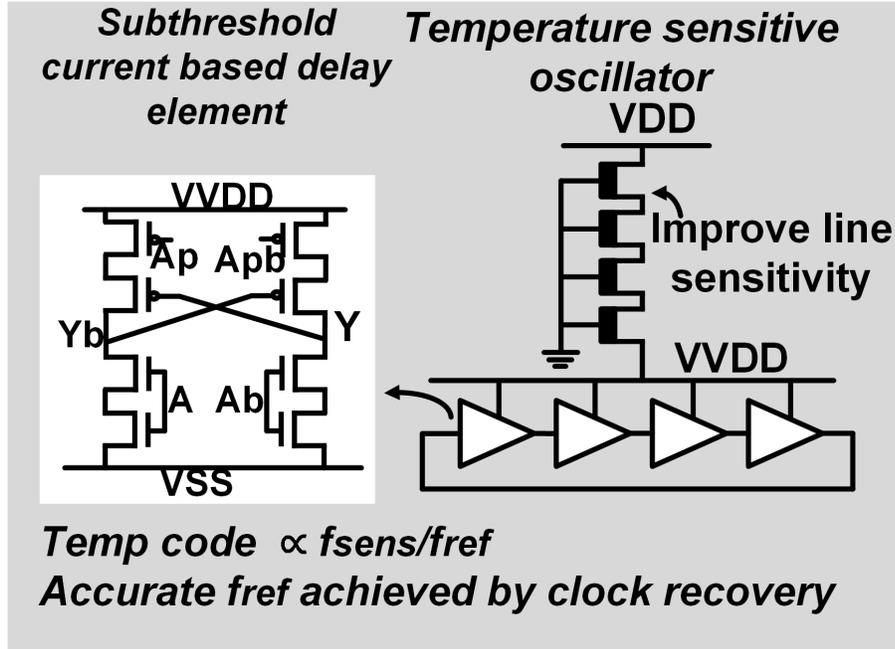


Figure 4.5: Implementation of temperature sensor

operation under light.

CTS senses temperature (Figure 4.5) by converting subthreshold current, which is exponentially dependent on temperature, to frequency, which is measured relative to the accurate reference clock. We employ a sensing oscillator structure similar to [60] due to its low line sensitivity created by a stacked native NMOS header that serves as a supply voltage regulator. This supply voltage invariant temperature sensor greatly relaxes supply regulation requirements in the system, enabling batteryless operation without voltage regulation even under modulated light intensities, improving power and area efficiency.

#### 4.4 Measurements

The proposed CTS circuit exhibits  $+0.38/-0.33^\circ\text{C}$  average error (2-point calibration) for five chips across  $10-60^\circ\text{C}$  (Figure 4.6), which is a wider range than required for biological measurements. Line sensitivity is  $0.6\%/V$ , corresponding to  $0.17^\circ\text{C}/V$ .

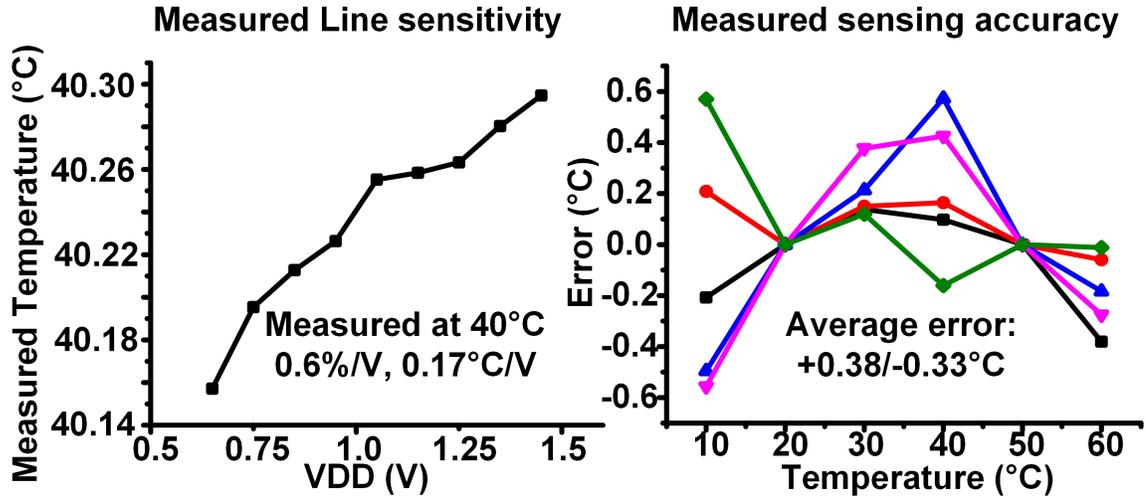


Figure 4.6: Measured temperature sensing performance

Heating effect of the base station on the sensor was measured to be negligible ( $<0.1^{\circ}\text{C}$  in 3hrs). In addition, heating effect from sensor LED can be mitigated by delayed read-out after experiment, thanks to the integrated processor and memory. A fully assembled CTS stack is measured using the setup in Figure 4.7, demonstrating successful wireless programming and accurate sensing using clock recovery (Figure 4.2). Figure 4.8 shows temperature readings received wirelessly from a fully assembled CTS stack across  $10\text{--}50^{\circ}\text{C}$ , showing  $0.034^{\circ}\text{C}$  RMS resolution and  $+0.11/-0.08^{\circ}\text{C}$  error. Table 4.1 compares this work to other small sensing platforms [14] [57] [58] [59].

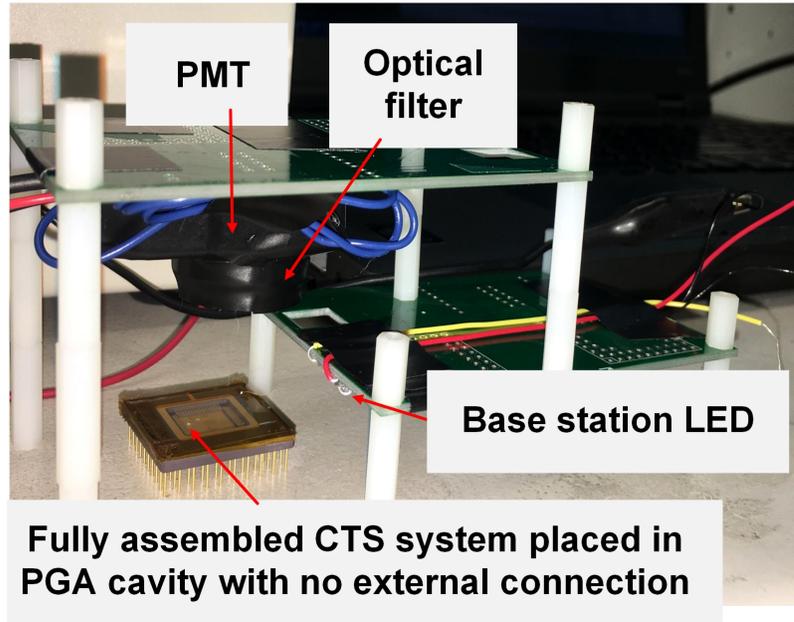


Figure 4.7: Testing setup showing CTS stack in use with base station

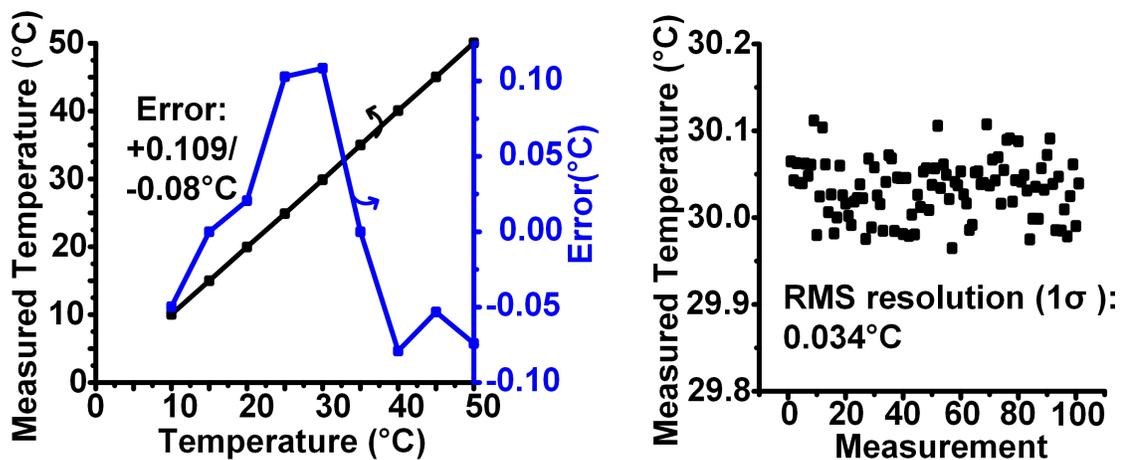


Figure 4.8: Sensing error and RMS resolution measured wirelessly with fully assembled CTS stack

	This work	[14] JSSC-2013	[57] ISSCC 2017	[58] Nature Bio. 2012	[59] RFIC 2017
Technology	55nm	0.18 $\mu$ m,0.13 $\mu$ m	0.18 $\mu$ m	0.18 $\mu$ m	65nm
System Dimension	360 x 400 x 280 $\mu$ m	1.1 x 2.21 x 0.4mm	2.8mm diameter, 200 $\mu$ m thick	11 x 9 x 0.2mm *	200 x 200 x 100 $\mu$ m
System Volume	0.04mm <sup>3</sup>	0.97mm <sup>3</sup>	0.38mm <sup>3</sup> **	19.8mm <sup>3</sup>	0.004mm <sup>3</sup>
System Power	16nW	11nW (standby) 20 $\mu$ W (active)	48.9 $\mu$ W	1.12nW	63nW
Integrated Processor	Yes	Yes	No	No	No
Sensor	Temperature	Temperature	Pressure	Endocochlear potential	Glucose concentration
Sensor Performance	Error +0.38/-0.33 $^{\circ}$ C Line sensitivity 0.17 $^{\circ}$ C/V RMS resolution 0.034 $^{\circ}$ C	RMS resolution: 0.51 $^{\circ}$ C	Pressure sensitivity: 0.67mmHg	RMS error: 0.45mV	--
Communication	Optical	RF/Optical	RF	RF	RF/Optical
TX/RX Area	0.07mm <sup>2</sup>	0.168mm <sup>2</sup>	4.52mm <sup>2</sup>	12mm <sup>2</sup>	0.04mm <sup>2</sup>
Transmit Distance	15.6cm	10cm	20cm	1m	2mm

\*System thickness is estimated from paper

\*\*Not including volume enclosed by powering coil

Table 4.1: System performance comparison

## CHAPTER V

# Pruning-based Pair Hidden Markov Model Accelerator for Whole Genome Sequencing

### 5.1 Introduction

Over the past decade, genomics has developed rapidly and transformed precision health to provide tailored treatment plans to patients. While it costed \$3 billion to sequence the first human genome in 2001, the cost has been reduced to one thousand dollars over the past decade. Speed and volume of sequencing machines have also improved greatly. This advancement can enable us to detect cancer without invasive biopsies, detect rare genetic disorders for early intervention, and identify pathogens for more accurate use of antibiotics. Biologist George Church predicted a future where people can use hand-held sequencer to sequence droplets of the person sneezes nearby and get real-time pathogen identification.

While the speed and cost of primary analysis using sequencing machines have been improved, performance growth of general-purpose processor has slowed down compared to Moores law. Advancement in the primary analysis brings growing demand for computing power to speed up secondary analysis. Secondary analysis in the whole genome sequencing is a crucial but time-consuming step, taking hundreds to thousands of CPU hours [11] for one genome. As Moores Law tapering off, researchers

have been developing customized accelerators using ASIC or FPGA to speed up the secondary analysis.

A genome can be viewed as a long (3.08Gbp for a human genome) string composed of DNA base-pairs (bp) A, G, T, C. A sequencing machine chops a DNA into billions of small fragments with 30-50 $\times$  coverage to reduce error and generates short string fragments called reads to be passed to secondary analysis. In secondary analysis, short reads are first aligned to a previously sequenced reference genome. Aligned reads are then processed to identify differences from reference genome in the step called variant calling. Variant calling is complicated because the algorithm needs to identify real variants of the sequenced genome from errors introduced by sequencing machines.

GATK's HaplotypeCaller is one of the most widely used variant calling tool today [63]. The tool first identifies active regions where reads are likely to be different from reference genome. Second, each of the active regions are reassembled using De-Bruijn graph. Top candidates called haplotypes assembled from De-Bruijn graph represent possible composition of this active region given evidence in reads. These haplotypes contain different composition from the reference genome. The tool will evaluate evidence of a set of most probable haplotypes based on some probabilistic model, and finally decide if there is enough evidence showing that the patient's genome contains certain variants.

HaplotypeCaller's algorithm assumes that a read and haplotype pair follows a Pair Hidden Markov Model (HMM). Pair-HMM [64] is a probabilistic model to evaluate pairwise alignments between two sequences. In variant calling, Pair-HMM is used to find out how much each haplotype is supported by related reads. Two algorithms of Pair-HMM are widely used to infer different probabilistic features. Viterbi algorithm looks for optimal alignment of the two sequence. Forward algorithm calculates overall alignment probability of the two sequence by computing the summation of

likelihoods of all alignments. The forward algorithm is used in GATK’s haplotypewriter to evaluate the similarity between each read and haplotype. Table 5.1 shows run time profile of the major steps in GATK’s HaplotypeCaller (version 4.0.11) using chromosome 16,17 and 18 of HG00419 from 1000 Genomes database. The program is run on Intel Xeon CPU E5 with single thread and AVX support for Pair-HMM step. Pair-HMM takes 53% of total execution time, making this one of the bottleneck in the pipeline that needs to be accelerated. This chapter introduces methods and hardware implementation to speed up Pair-HMM calculation.

	<b>Run Time (hour)</b>	<b>Percentage</b>
<b>Assembly</b>	<b>1.22</b>	<b>25%</b>
<b>Pair-HMM (forward algorithm)</b>	<b>2.6</b>	<b>52%</b>
<b>Genotyping</b>	<b>0.2</b>	<b>4%</b>
<b>Other</b>	<b>0.85</b>	<b>18%</b>

Table 5.1: Profiling result of HaplotypeCaller using chromosome 16-18 of sample HG00419

Pair-HMM is a complex dynamic programming problem. The unit operation includes floating point summation and multiplication. Operation involves calculation of three matrices which depend on each other. Unlike many other dynamic programming problems which only involves minimum, maximum or integer arithmetic, Pair-HMM used here requires at least single precision floating point operation to avoid overflow or underflow. One Pair-HMM calculation is required for each read and each haplotype pair. Each Pair-HMM calculation involves  $L_h * L_r$  unit operation with several floating point summation and multiplication, where  $L_h$  is the length of haplotype, and  $L_r$  is the length of read. Big dynamic range in data involved and three mutual dependent matrices make it difficult to accelerate Pair-HMM.

There has been several work on accelerating Pair-HMM calculation using different computational platforms. An GPU implementation [65] deploys inter-job and between job parallelism to achieve a higher throughput. An FPGA implementation by Altera OpenCL [66] maps Pair-HMM matrix to a systolic array of processing elements (PE). However, read length and haplotype length can vary from 10s of base pair (bp) to >100bp, leading to Pair-HMM matrices with various sizes. This makes inefficient utilization of systolic array with fixed size which can reduce overall throughput given fixed hardware area. An ASIC implementation of dynamic programming [67] uses propagation delay of circuits to represent alignment scores to achieve speedup. However, this work only applies to dynamic programming problems with min, max operations. In addition, converting data from digital domain to time domain requires timing circuits with matching resolution, which is more difficult to achieve across process, voltage and temperature variations, and thus can reduce yield of chips meeting accuracy requirements.

To further speed up Pair-HMM calculation, this chapter introduces 1) A pruning-based Pair-HMM algorithm which uses fixed point calculation in log domain to prune out sections in Pair-HMM matrix which contribute little to final result, limiting demand for floating point operation only to un-pruned sections, and therefore reduce area for large floating point units and achieve higher throughput given fixed area. This approach takes advantage of the wide dynamic range of data in Pair-HMM matrices. It decouples the trade-off between speed and accuracy by recognizing and accurately calculating the critical portion of Pair-HMM matrices that actually requires floating point operations, and providing the rest with mathematical bounds for downstream processing. 2) An efficient ASIC architecture to implement the pruning-based Pair-HMM accelerator.

## 5.2 Pruning-Based Pair-HMM Algorithm

### 5.2.1 Conventional Pair-HMM Algorithm

Pair-HMM is a statistical model which allows us to draw inference about the alignment quality between read and haplotype. It helps determine the real DNA expression of an individual given the possibly incorrect reads. Forward algorithm is used in , which efficiently calculates the overall probability of all possible alignments between read and haplotype.

Pair-HMM model alignments using three hidden states insertion (I), deletion (D) and match (M). All alignments of a read-haplotype pair can be expressed using an alignment matrix of size  $L_r * L_h$ , where  $L_r$  is the length of read and  $L_h$  is the length of haplotype. Each cell  $(i,j)$  indicates how base pair  $i$  in the read is aligned to base pair  $j$  in haplotype using one of the three states insertion, deletion and match. Each path in the alignment matrix can be thought of a series of state transitions, and this is one alignment between read and haplotype. Probabilities are associated with each state transition depending on state, base pair and quality scores of base pair. Probability of each path can therefore be inferred by calculating state transition probabilities. The forward algorithm used for Pair-HMM aims to infer the overall probability of all alignments. This is done using dynamic programming. Each cell in the alignment matrix now contains three matrices  $f^M, f^I, f^D$ .  $f^k(i, j)$  corresponds to the combined probability of all alignments up to position  $(i, j)$  of read and haplotype that ends in state  $k$ .  $k$  can be I (insertion), D (deletion) and M (match). For each position  $(i, j)$ ,  $f^M, f^I, f^D$  are calculated as below, where  $p_{mm}, p_{im}, p_{dm}, a_{mi}, a_{ii}, a_{md}$ , and  $a_{dd}$  are probabilities related to state transition and read quality score.

Final output of forward algorithm is sum of insertion and match probabilities in the final row:  $\sum_{j=1}^{L_h} (f^M(L_r, j) + f^I(L_r, j))$ , where  $L_r$  and  $L_h$  are the length of read and haplotype.

$$f^M(i, j) = p_{mm}f^M(i - 1, j - 1) + p_{im}f^I(i - 1, j - 1) + p_{dm}f^D(i - 1, j - 1) \quad (5.1)$$

$$f^I(i, j) = a_{mi}f^M(i - 1, j) + a_{ii}f^I(i - 1, j) \quad (5.2)$$

$$f^D(i, j) = a_{md}f^M(i, j - 1) + a_{dd}f^D(i, j - 1) \quad (5.3)$$

As can be seen above, forward algorithm is based on probabilities which can get very small quickly. Therefore requires computational intensive floating point calculation.

## 5.2.2 Proposed Pruning-based Pair-HMM Algorithm

In order to speed up Pair-HMM calculation, this chapter introduces a pruning-based algorithm to reduce the amount of floating point operation by using upper bound estimations of the result. Reducing floating point operations can reduce area costly floating point units required for ASIC and FPGA acceleration, and therefore achieves a higher throughput given fixed area.

### 5.2.2.1 Cell Level Pruning

At each position  $(i, j)$ , the floating point calculation involves summing  $f^M$ ,  $f^I$  and  $f^D$  from adjacent cells in order to calculate overall alignment probability of read-haplotype pair. Figure 5.1 illustrates data dependencies for  $f^M(i, j)$ ,  $f^I(i, j)$  and  $f^D(i, j)$  according to equation 5.1, 5.2, 5.3. For example,  $f^M$  in each square (indexed  $(i, j)$ ) depends on weighted sum of  $f^M$ ,  $f^I$  and  $f^D$  from the square before it along the diagonal line (indexed  $(i - 1, j - 1)$ ).

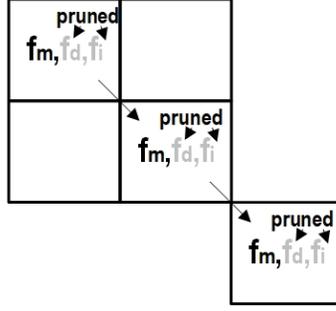


Figure 5.2: Compare and prune based on relative value of  $f^I$ ,  $f^D$  and  $f^M$

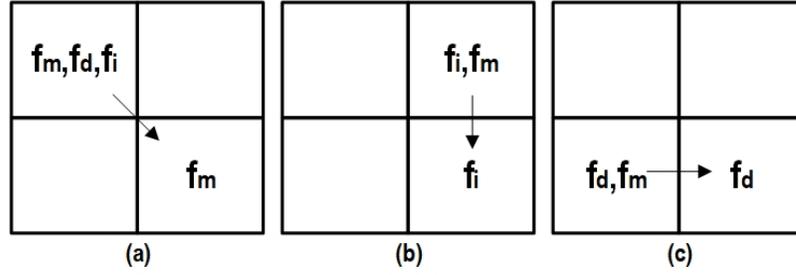


Figure 5.1: Data dependencies of (a)  $f^M$ , (b)  $f^I$  and (c)  $f^D$

We made the key observation that in many cases, weighted  $f^I(i-1, j-1)$  and  $f^D(i-1, j-1)$  are much smaller than  $f^M(i-1, j-1)$ , which means that setting  $f^I(i-1, j-1)$  and  $f^D(i-1, j-1)$  to zero (i.e. prune them) could have given us negligible loss in the result  $f^M(i, j)$ . As we continue to calculate the Pair-HMM matrix shown in Figure 5.2, if  $f^I(i, j)$  and  $f^D(i, j)$  are significantly smaller compare to  $f^M(i, j)$ , we can continue to prune  $f^I(i, j)$  and  $f^D(i, j)$  without sacrificing the accuracy of  $f^M(i+1, j+1)$  too much. We continue this compare and prune method for each square in the Pair-HMM matrix, and finally we can identify segments of diagonal lines.  $f^I$  and  $f^D$  in all cells along this line can be pruned because they are significantly smaller than  $f^M$  in the same cell, and all  $f^M$  along this line need to be accurately computed because they play an important role in the final result.

The goal of proposed pruning-based Pair-HMM is to identify one diagonal line among all the red line segments in the matrix that can represent a dominant align-

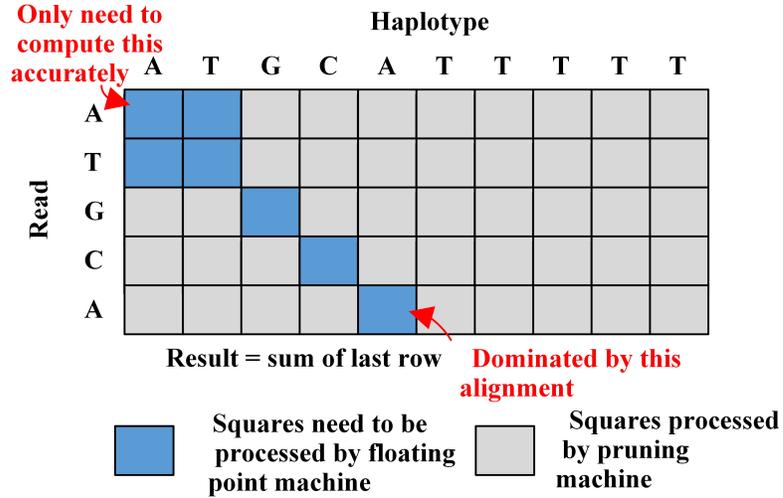


Figure 5.3: Illustration of proposed pruning-based Pair-HMM algorithm

ment. This line is determined by choosing the red line whose end cell contains the largest number in the final row. The only red line which ends with maximum number in the final row is selected. The final un-pruned computations are all  $f^M$  in the diagonal line selected, and all cells in the rectangle which contribute to the  $f^M$  of the start cell in the diagonal line. As illustrated in Figure 5.3, using this pruning method, we found that for most read-haplotype pairs, the overall alignment probability is dominated by only a few, or even one alignment, due to limited number of mismatch positions and high read quality scores. In other words, when summing  $f^M$ ,  $f^I$  and  $f^D$  from adjacent cells, result is dominated by only one of the inputs.

The proposed pruning-based algorithm calculates a Pair-HMM matrix in two passes as illustrated in Figure 5.3. In the first pass, the entire matrix is rapidly calculated using fixed point approximation in log domain. The calculation can be done using approximation, including fixed point calculation with fewer bits to optimize speed. Log-sum is substituted with fast table lookup. Based on approximate values, the accelerator prunes squares in the matrix whose values contribute insignificantly to overall probabilities using the method introduced previously. The first round of approximate calculation can be implemented by rounding up in each approximation

steps, yielding an upper bound of exact result using conventional method. We refer to this round of calculation as upper bound round.

In the second pass, precise calculation using floating point operation is carried out only in un-pruned subsection of the alignment matrix. Since only a subset of alignments are calculated, results is naturally a lower bound of exact result. We refer to this round of calculation as lower bound round. Speed up is achieved by substituting area costly floating point operation with fixed point operation as much as possible, and therefore increase throughput given fixed area.

Unlike conventional method where output of Pair-HMM is one exact value computed on the entire alignment matrix using floating point, the proposed pruning-based Pair-HMM outputs a lower bound and an upper bound of exact result. Upper bound result comes from the first round of calculation on the entire alignment matrix using fixed points in log domain. Upper bound is generated by rounding up in each approximation. Lower bound result comes from the second round of calculation on the unpruned section of alignment matrix using floating point. Since only a subset of alignments are calculated, results is naturally a lower bound of exact result. In downstream processing of HaplotypeCaller, output of Pair-HMM is filtered and used to infer probabilities of genotypes, and genotypes of highest probability will be selected as final output. In this process, upper bound and lower bound can be combined to determine the best genotype. For example, if the lower bound of probability of selected genotype is higher than upper bound of all the unselected genotypes, then we can guarantee that the selected genotype is of the highest probability. This methodology can be used in downstream processing steps involving filtering and comparison as long as the operation preserves bounds. If bound check fails, we cannot infer a guaranteed result based on current lower and upper bounds. In this case, re-computation of the original Pair-HMM matrix is required. Re-computation is done in a read-by-read basis, and only haplotypes contributing to final results are selected for re-computation.

Once a read-haplotype pair is selected for re-computation, conventional Pair-HMM method using all floating point operations is used to obtain exact result.

### 5.2.2.2 Matrix Level Pruning

As discussed before, the upper bound round can be implemented by rounding up in each approximation steps, yielding an upper bound of exact result using conventional method. If upper bound result from fixed point calculation is too small, the entire Pair-HMM matrix will be pruned and floating point calculation will be skipped entirely. The reason behind this optimization is that extremely low upper bound value indicates low similarity between read and haplotype, and is very likely to be ignored during marginalization in downstream processing. In downstream processing, read-haplotype probabilities are marginalized to read-allele probability, where haplotypes containing the same allele competes with each other, and the haplotype with the highest likelihood score gets to represent the allele. In other words, only the highest read-haplotype score will be picked and passed onto the next step. Low probability scores will likely lose in marginalization to the highest score. Therefore, skipping the entire lower bound calculation based on upper bound estimation can further reduce floating point operations with minimal impact on final result.

### 5.2.2.3 Early Termination in Upper Bound Round

To further reduce computation workload of the algorithm, we can reduce work required in the upper bound round. During upper bound calculation, values in the matrix can get extremely small depending on similarity of the two input sequences. As discussed in Section 5.2.2.2, low output probability from Pair-HMM calculation is likely to be ignored in marginalization during downstream processing. To further reduce computation in proposed pruning-based Pair-HMM, early termination is implemented in upper bound round. As processing units propagates through the

alignment matrix and estimate upper bound of values in each cell, maximum value of all the cells in each row  $f_{max}(i)$  is computed and compared to a threshold value  $f_{th}$ . If  $f_{max}(i)$  is smaller than  $f_{th}$ , upper bound calculation is terminated.

Table 5.2 summarized characteristic of the proposed pruning-based Pair-HMM and the computation reduction based on chromosome 1 of sample HG00419 from 1000 Genome database. Compared to baseline algorithm, the pruning-based algorithm can save 99% floating point operation when re-computation is not considered, and it can save 97.8% floating point operation after all bound check fails are handled by re-computation, leading to a  $45\times$  reduction in floating point execution time. By using early termination technique in upper bound round of calculation, 19.4% fixed point calculation can further be saved, leading to an extra  $1.24\times$  increase in upper bound round throughput.

	<b>Pair-HMM cells computed with floating point</b>	<b>Pair-HMM cells computed with fixed point</b>
<b>Baseline</b>	<b><math>2.5 * 10^{12}</math></b>	<b>0</b>
<b>Pruning-based</b>	<b><math>21.8 * 10^9</math> (exclude re-computation) 99% reduction</b>	<b><math>2.05 * 10^{12}</math> 19% reduction due to early termination</b>
	<b><math>58.5 * 10^9</math> (include re-computation) 97.7% reduction</b>	
<b>Workload Reduction</b>	<b>43x</b>	

Table 5.2: Computation reduction of Pruning-base Pair-HMM algorithm

## 5.3 Pruning-based Pair Hidden Markov Model Architecture

### 5.3.1 PE Array

In the previous sections, we only discuss speedup due to reduction in floating point operations assuming fixed point operation is "free". To understand the overall speedup of the proposed pruning-based algorithm, we designed PE arrays in hardware for floating point and fixed point operations. In this design, we use a series of processing elements (PE) as illustrated in Figure 5.4. All PEs propagate in a diagonal line for maximum parallelism. We compare their key performance after place and route in Table 5.3. Both fixed point and floating point PE arrays are implemented using TSMC 40nm LP technology, and their standard cell density after place and route is 85% and 82% respectively. From this comparison table, a fixed point PE is  $4.6\times$  smaller on average compared to a floating point PE, and it is  $2.1\times$  faster than a floating point PE, leading to  $9.3\times$  maximum throughput improvement compared to baseline floating point implementation.

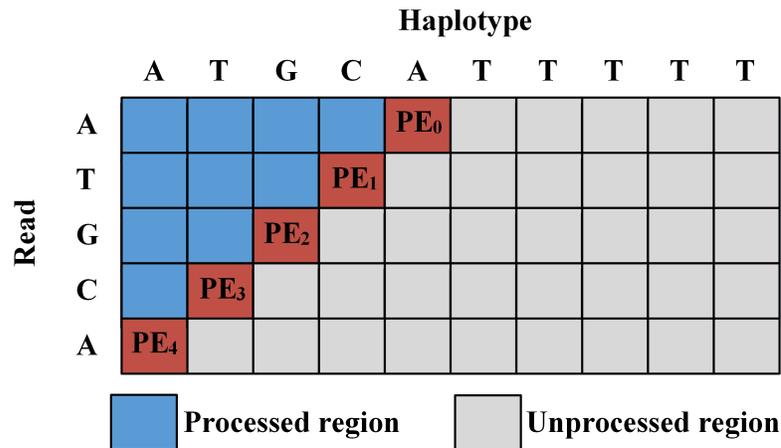


Figure 5.4: PE array structure used in Pair-HMM accelerator

	Number of PE/Array	Area (um <sup>2</sup> )	Fmax (Mhz)	Maximum Throughput (GCUP/s/mm <sup>2</sup> )*	Gain
Fixed point PE array	16	420311	473	18	9.3
Floating point PE array	4	478109	230	1.9	1

\*CUP: number of Pair-HMM cell update

Table 5.3: Performance comparison between floating point and fixed point PE arrays

### 5.3.2 Accelerator Architecture

To effectively implement the proposed pruning-based Pair-HMM algorithm, this chapter introduces the hardware architecture shown in Figure 5.5. The accelerator has 1) a pruning machine with fixed point processing elements which produces upper bound of overall alignment probability and an index for unpruned area, 2) a precise machine which works on un-pruned subsection of Pair-HMM matrix and generates a lower bound result, and 3) a on-demand job scheduler which issues correct Pair-HMM jobs to pruning machine and precise machine, collect results and schedule memory accesses from both machines effectively.

The pruning machine consists of 10 fixed point PE arrays, each of which has 16 PEs for fixed point calculation and pruning logic. A PE array first takes read-haplotype pairs, store them in local register files to reduce memory bandwidth requirement. The stored read and haplotype pair is then fed into processing elements for log domain  $f^M$ ,  $f^I$  and  $f^D$  calculation as mentioned in Section 5.2.2.1. Each PE calculates log domain with fixed point adders (equivalent to multiplication in real domain) and log-sum table lookup (equivalent to addition in real domain). PE prunes each cell it has processed if its output contributes insignificantly to final alignment probabilities. As discussed in Section 5.2.2.2, if the resulting upper bound of Pair-HMM probability is smaller than a threshold, the entire matrix is skipped for floating point calculation. As discussed in Section 5.2.2.3, when PE arrays propagate horizontally through the matrix, maximum

value is computed for each row processed by the tailing PE. If the maximum value is smaller than a threshold, this PE array will terminate this job early to reduce workload in upper bound round calculation. As shown in equations 5.1, 5.2, 5.3, each cell at position  $(i, j)$  only depends on cells at position  $(i - 1, j - 1)$ ,  $(i - 1, j)$  and  $(i, j - 1)$ . Therefore all cells in the same diagonal is independent and can be parallelized. PEs are designed to progress alignment matrix in a waterfront form to exploit intra job parallelism. Pruning machine outputs an upper bound of overall alignment probability and an index for un-pruned region for accurate machine to process later.

The accurate machine has two floating point PE arrays each with 4 PEs for jobs with larger un-pruned section, and two standalone single PE for jobs with very small un-pruned section. The accurate machine takes in read-haplotype pairs and an un-pruned region, performs baseline floating point calculation only in the un-pruned region, and outputs a lower bound of overall alignment probability.

The proposed accelerator is implemented and fabricated in TSMC 40nm LP technology. Chip area is  $7\text{mm}^2$ . From pre-silicon verification, pruning-based accelerator achieves 71 GCUP/s average throughput, which is  $8.3\times$  higher than baseline accelerator with equal area.

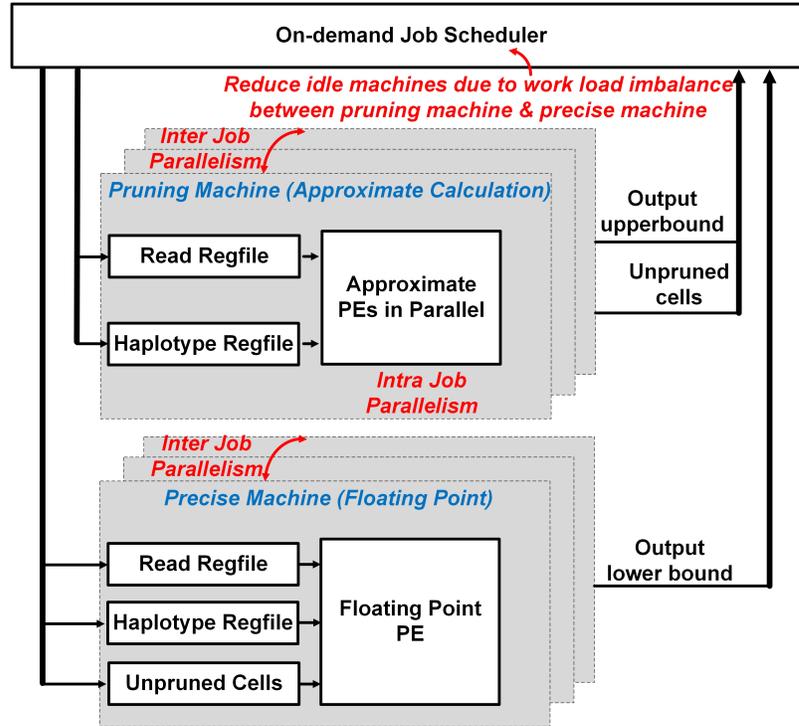


Figure 5.5: Hardware architecture of pruning-based Pair-HMM accelerator

## 5.4 Conclusions

In summary, this chapter introduces 1) a pruning-based algorithm for Pair-HMM calculation with  $43\times$  floating point operation reduction; 2) an efficient ASIC architecture for pruning-based Pair-HMM accelerator with  $8.3\times$  throughput improvement compared to ASIC accelerator of baseline algorithm.

## CHAPTER VI

### Conclusions

#### 6.1 Summary of Contributions

Recent research has been pushing the limits of the Internet of Things in terms of system volume and power consumption, enabling applications such as biomedical sensing, localization of small objects and industrial sensing. Small wireless sensor nodes are able to survive in places where it was previously impossible. One example is intracellular temperature sensing for cancer studies [62]. This thesis discussed challenges in powering small form factor sensor nodes and scaling system volume to a sub-mm<sup>3</sup> level. This thesis introduced energy harvesting and power management circuit techniques as well as system design for miniaturized wireless sensor nodes. This thesis also expanded the discussion to include accelerating computations for portable DNA sequencing devices. A pruning-based Pair-HMM algorithm for whole-genome sequencing and its hardware accelerator design was introduced.

In Chapter II, we discussed a discontinuous switched-capacitor solar energy harvester that enables ultra-low power energy harvesting. The harvester uses a hybrid structure called a moving sum charge pump for low startup energy upon a mode switch, an automatic conversion ratio modulator based on conduction loss optimization for fast conversion ratio increment, and a <15pW asynchronous mode controller for ultra-low power operation. In 180-nm CMOS, the harvester achieves >40% end-

to-end efficiency from 113 pW to 1.5  $\mu$ W with 20 pW minimum harvestable input power.

In Chapter III, we discussed a fully integrated energy reservoir unit using a counter flow method for peak power delivery in space-constrained sensor systems. The counter flow energy reservoir delivers 65% of stored energy and supplies up to 13.6 mW output power for 1  $\mu$ s before recharging is needed.

In Chapter IV, we discussed a complete wireless sensor node for accurate cellular temperature measurement with a fully programmable Cortex M0+ processor, custom SRAM, optical energy harvesting, 2-way communication, and a subthreshold temperature sensor. The 0.04 mm<sup>3</sup> fully assembled sensor node temperature resolution is 0.034°C RMS, and the transmit distance extends to 15.6 cm.

In Chapter V, we discussed a Pair-HMM hardware accelerator using a pruning-based algorithm. The algorithm explores the huge differences in values among the floating point numbers in a Pair-HMM calculation, so that the floating point calculation can be reduced dramatically for speed up.

## 6.2 Directions for Future Research

The techniques introduced in this thesis open up opportunities for future improvements to further relieve the design challenges discussed before. Circuit techniques introduced in Chapter II and Chapter III are based on capacitors. The energy transfer efficiency of these techniques depends on the quality and density of capacitors in the design. Metal-insulator-metal (MIM) capacitors are mostly used in these works. However, using high density capacitors such as trench capacitors can improve overall circuit performance.

Chapter IV demonstrated a sub-mm<sup>3</sup> wireless system for intracellular temperature measurement. The discussion focused on circuit design challenges and techniques. However, the assembly of these tiny systems is a crucial part of system design and

impacts the yield of systems significantly. As the system volume keeps shrinking, robust assembly techniques will become as important as circuit design techniques. Future designers should consider assembly plans (for example pad locations) when making system level decisions such as discrete components choices, communication techniques and overall power budgets.

Chapter V introduced a hardware accelerator for a Pair-HMM in whole-genome sequencing. The work presented is only on step in the whole-genome sequencing pipeline. The performance benchmark focus on acceleration due to improved algorithm and hardware architecture assumes ideal software interfaces. A real-world whole-genome sequencing pipeline is often a heterogeneous computing system that includes both software running on CPU and hardware accelerated kernels. Future research can focus on further designing an end-to-end system and investigate the trade-offs introduced due to these interfaces.

## APPENDICES

## APPENDIX A

### Related Publications

X. Wu, Y. Shi, S. Jeloka, K. Yang, I. Lee, D. Sylvester and D. Blaauw. A 66pW discontinuous switch-capacitor energy harvester for self-sustaining sensor application. *2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, pages 1–2, 2016.

X. Wu, Y. Shi, S. Jeloka, K. Yang, I. Lee, Y. Lee, D. Sylvester and D. Blaauw. A 20-pW discontinuous switched-capacitor energy harvester for smart sensor applications. *IEEE Journal of Solid-State Circuits*, 52(4):972–984, 2017.

X. Wu, K. Choo, Y. Shi, L. Chuo, D. Sylvester and D. Blaauw. A fully integrated counter-flow energy reservoir for 70%-efficient peak-power delivery in ultra-low-power systems. *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 380–381, 2017.

X. Wu, K. Choo, Y. Shi, L. Chuo, D. Sylvester and D. Blaauw. A fully integrated counter flow energy reservoir for peak power delivery in small form-factor sensor systems. *IEEE Journal of Solid-State Circuits*, 52(120):3155–3167, 2017.

X. Wu, I. Lee, Q. Dong, K. Yang, D. Kim, J. Wang, Y. Peng, Y. Zhang, M. Sali-ganc, M. Yasuda, K. Kumeno, F. Ohno, S. Miyoshi, M. Kawaminami, D. Sylvester and D. Blaauw. A 0.04mm<sup>3</sup> 16nW Wireless and Batteryless Sensor System with Integrated Cortex-M0+ Processor and Optical Communication for Cellular Temperature

Measurement. *2018 IEEE Symposium on VLSI Circuits*, pages 191–192, 2018.

## APPENDIX B

### End-to-end efficiency of Energy Harvester

End-to-end efficiency is defined as the total output energy from the harvester  $E_{out}$  divided by energy generated by the solar cell  $E_{mppt}$  when biased at its maximum power point (Equation B.1). Total output energy is the output energy generated in transfer phase  $E_{out,tran}$  minus the total leakage energy from the battery in the harvest phase  $E_{leak}$  (Equation B.2). Leakage power in the transfer phase is small (1.2nW, simulated) compared to the steady state output power in the transfer phase (>170nW, measured), and is accounted for in  $E_{out,tran}$  ( $E_{out,tran}$  is the output energy of the charge pump minus the leakage energy in the transfer phase).  $E_{out,tran}$  can be expressed as the product of solar efficiency  $Eff_{solar} = \frac{EnergyAccumulatedonCbuf}{E_{mppt}}$  in the harvest phase and overall charge pump efficiency in the transfer phase  $Eff_{tran} = \frac{E_{out,tran}}{EnergyAccumulatedonCbuf}$  (Equation B.3) Therefore, total efficiency can be expressed in Equation 2.1 in Section 2.2.2 where  $P_{leak}$  is the leakage power in harvest phase.

$$Eff_{tot} = \frac{E_{out}}{E_{mppt}} \quad (\text{B.1})$$

$$Eff_{tot} = \frac{E_{out,tran} - E_{leak}}{E_{mppt}} \quad (\text{B.2})$$

$$Eff_{tot} = \frac{E_{mppt} * Eff_{solar} * Eff_{tran} - E_{leak}}{E_{mppt}} \quad (\text{B.3})$$

## APPENDIX C

### Solar Efficiency

The solar efficiency in the harvest phase is defined as the ratio of the average power  $P_{solar,avg}$  accumulated on Cbuf in the harvest phase, and the maximum power point of the solar cell  $P_{mppt}$ , as shown in Equation C.1.

$$Eff_{solar} = \frac{P_{solar,avg}}{P_{mppt}} \quad (C.1)$$

$P_{solar,avg}$  is the average power harvested over duration of harvest phase (Equation C.2). By definition, it can be expressed as the integral of  $P(v)$  (instantaneous output power of the solar cell when biased at voltage  $v$ ) from  $t_0$  to  $t_1$  divided by the duration of harvest phase where  $t_0$  and  $t_1$  are the start and end times of the harvest phase, respectively.

$$P_{solar,avg} = \frac{\int_{t_0}^{t_1} P(v) dt}{\int_{t_0}^{t_1} dt} \quad (C.2)$$

Here  $dt$  can be calculated by Equation C.3 and simplified to Equation C.4.

$$dt = \frac{1}{2} * Cbuf * \frac{(v + dv)^2 - v^2}{P(v)} \quad (C.3)$$

$$dt = \frac{1}{2} * C_{buf} * \frac{2v dv}{P(v)} \quad (C.4)$$

Therefore, solar efficiency can be expressed in Equation 2.4 in Section 2.2.2.

## APPENDIX D

# Model Simplifications of Discontinuous Energy Harvester

$P(v)$ , which is defined as the solar cell output power when biased at  $v$ , is the product of  $v$  and  $I_{solar}(v)$ .  $I_{solar}(v)$  is supposed to be modeled as Equation D.1 [42], where  $I_0$ ,  $I_L$ ,  $R_s$ ,  $k$  and  $R_p$  are variables related to solar cell characteristics. Unfortunately, there are no analytical solution to Equation A.8. To simplify the calculation, two assumptions are made here. First, we assume  $I_{solar}(v) = I_{sc}$  for  $v < V_{mppt}$ , where  $I_{sc}$  is the short circuit current of the solar cell. Second, we set  $VH = V_{mppt}$  to limit the voltage range in this calculation to  $v \in [0, V_{mppt}]$ , where VH is the voltage on Cbuf at the end of harvest phase. By assuming  $I_{solar}(v) = I_{sc}$ , we overestimate solar output current  $I_{solar}(v)$  and therefore overestimate solar efficiency  $Eff_{solar}$  in the harvest phase. The resulting error is shown in Figure 2.6. By limiting VH to  $V_{mppt}$ , we could potentially miss the global optimal pair of VH and VL. The error compared to the optimal point found without setting  $VH = V_{mppt}$  is shown in Figure D.1. Practically, the optimal VH can be close to but slightly higher than  $V_{mppt}$  for a better trade-off between solar efficiency and overall charge pump efficiency. With the two assumptions,  $VH = V_{mppt}$  and  $VL = V_{mppt} - \Delta V_{sol}$ , The problem of finding

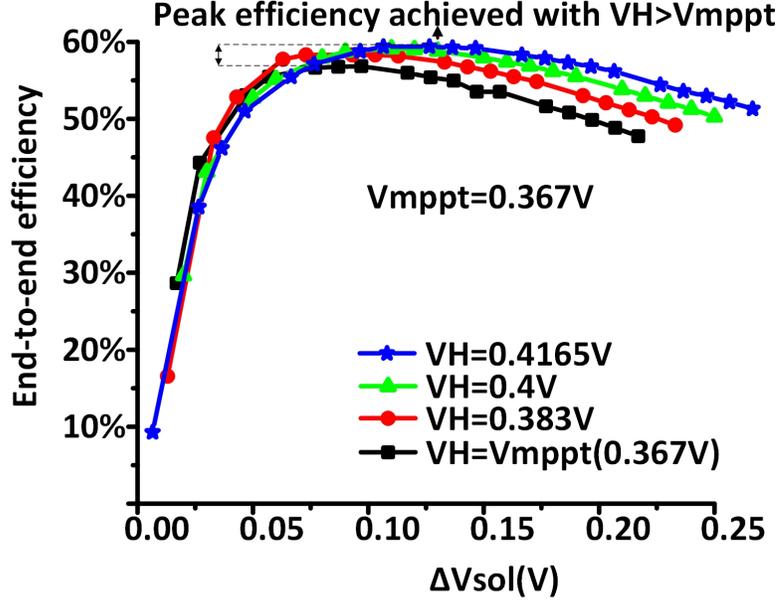


Figure D.1: Dependency of simulated end-to-end efficiency on VH

the optimal pair of VH and VL is simplified to finding the optimal  $\Delta V_{sol}$ .  $P(v)$  is simplified to Equation D.2.

$$I_{solar}(v) = I_L - I_0 * \left( e^{\frac{v + I_{solar}(v) * R_s}{k}} - 1 \right) - \frac{v + I_{solar}(v) * R_s}{R_p} \quad (D.1)$$

$$P(v) = v * I_{SC} \quad (D.2)$$

Therefore, solar efficiency can be simplified as shown in Equation 2.5 (Section 2.2.2), and transfer phase efficiency can be rewritten in Equation 2.6 (Section 2.2.2 II.B) in terms of  $\Delta V_{sol}$ .

## BIBLIOGRAPHY

## BIBLIOGRAPHY

- [1] N. Mohamed and I. Jawhar. A fault tolerant wired/wireless sensor network architecture for monitoring pipeline infrastructures. In *2008 Second International Conference on Sensor Technologies and Applications (sensorcomm 2008)*, pages 179–184, Aug 2008.
- [2] Norberto Barroca, Luis Borges, Fernando Velez, Filipe Monteiro, Marcin Grski, and Joo Castro-Gomes. Wireless sensor networks for temperature and humidity monitoring within concrete structures. *Construction and Building Materials*, 40:1156–1166, 03 2013.
- [3] Li-Xuan Chuo, Zhihong Luo, Dennis Sylvester, David Blaauw, and Hun-Seok Kim. Rf-echo: A non-line-of-sight indoor localization system using a low-power active rf reflector asic tag. pages 222–234, 10 2017.
- [4] Hugo Dinis and Paulo M. Mendes. *Recent Advances on Implantable Wireless Sensor Networks*. 10 2017.
- [5] Vikram Iyer, Rajalakshmi Nandakumar, Anran Wang, Sawyer Fuller, and Shyamnath Gollakota. Living iot: A flying wireless platform on live insects, 12 2018.
- [6] Yoonmyung Lee, Gregory K. Chen, Scott Hanson, Dennis Sylvester, and David Blaauw. Ultra-low power circuit techniques for a new class of sub-mm<sup>3</sup> sensor nodes. *IEEE Custom Integrated Circuits Conference 2010*, pages 1–8, 2010.
- [7] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, 2001.
- [8] H. Li. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 2013.
- [9] Y. Wu, J. Hung, and C. Yang. 14.8 A 135mW fully integrated data processor for next-generation sequencing. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 252–253. IEEE, 2017.
- [10] Y. Turakhia, K. Zheng, G. Bejerano, and W. J. Dally. Darwin : A hardware-acceleration framework for genomic sequence alignment. 2017.

- [11] D. Fujiki, A. Subramaniyan, T. Zhang, Y. Zeng, R. Das, D. Blaauw, and S. Narayanasamy. GenAx: a genome sequencing accelerator. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pages 69–82, 2018.
- [12] S. S. Banerjee, M. El-Hadedy, C. Y. Tan, Z. T. Kalbarczyk, S. S. Lumetta, and R. K. Iyer. On accelerating pair-HMM computations in programmable hardware. *2017 27th International Conference on Field Programmable Logic and Applications (FPL)*, pages 1–8, 2017.
- [13] Y. Shi, M. Choi, Z. Li, G. Kim, Z. Foo, H. Kim, D. Wentzloff, and D. Blaauw. A 10mm<sup>3</sup> syringe-implantable near-field radio system on glass substrate. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 448–449, 2016.
- [14] Y. Lee, S. Bang, I. Lee, Y. Kim, G. Kim, M. H. Ghaed, P. Pannuto, P. Dutta, D. Sylvester, and D. Blaauw. A modular 1 mm<sup>3</sup> die-stacked sensing platform with low power I<sup>2</sup>C inter-die communication and multi-modal energy harvesting. *IEEE Journal of Solid-State Circuits*, 48(1):229–243, 2013.
- [15] S. Huang, G. J. Manikandan, A. Ramachandran, K. Rupnow, W. Hwu, and C. Deming. Hardware acceleration of the pair-HMM algorithm for DNA variant calling. In *Proceedings of the 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays - FPGA '17*, pages 275–284, 2017.
- [16] G. Chen, M. Fojtik, D. Kim, D. Fick, J. Park, M. Seok, M. Chen, Z. Foo, D. Sylvester, and D. Blaauw. Millimeter-scale nearly perpetual sensor system with stacked battery and solar cells. In *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 288–289, 2010.
- [17] D. Hodgins, A. Bertsch, N. Post, M. Frischholz, B. Volckaerts, J. Spensley, J. M. Wasikiewicz, H. Higgins, F. von Stetten, and L. Kenney. Healthy aims: developing new medical implants and diagnostic equipment. *IEEE Pervasive Computing*, 7(1):14–21, 2008.
- [18] I. Lee, W. Lim, A. Teran, J. Phillips, D. Sylvester, and D. Blaauw. A >78%-efficient light harvester over 100-to-100klux with reconfigurable PV-cell network and MPPT circuit. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 370–371, 2016.
- [19] E. Carlson, K. Strunz, and B. Otis. 20mV input boost converter for thermoelectric energy harvesting. In *2009 Symposium on VLSI Circuits*, pages 162–163, 2009.
- [20] G. Yu, K. W. R. Chew, Z. C. Sun, H. Tang, and L. Siek. A 400 nW single-inductor dual-inputtri-output DCDC buckboost converter with maximum power point tracking for indoor photovoltaic energy harvesting. *IEEE Journal of Solid-State Circuits*, 50(11):2758–2772, 2015.

- [21] S. Bandyopadhyay, P. P. Mercier, A. C. Lysaght, K. M. Stankovic, and A. P. Chandrakasan. A 1.1 nW energy-harvesting system with 544 pW quiescent power for next-generation implants. *IEEE Journal of Solid-State Circuits*, 49(12):2812–2824, 2014.
- [22] W. Jung, S. Oh, S. Bang, Y. Lee, D. Sylvester, and D. Blaauw. A 3nW fully integrated energy harvester based on self-oscillating switched-capacitor DC-DC converter. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 398–399, 2014.
- [23] I. Doms, P. Merken, R. Mertens, and C. Van Hoof. Integrated capacitive power-management circuit for thermal harvesters with output power 10 to 1000 $\mu$ W. In *2009 IEEE International Solid-State Circuits Conference - Digest of Technical Papers*, pages 300–301,301a, 2009.
- [24] Y. Qiu, C. Van Liempd, B. O. het Veld, P. G. Blanken, and C. Van Hoof. 5 $\mu$ W-to-10mW input power range inductive boost converter for indoor photovoltaic energy harvesting with integrated maximum power point tracking algorithm. In *2011 IEEE International Solid-State Circuits Conference*, pages 118–120, 2011.
- [25] R. Enne, M. Nikolic, and H. Zimmermann. A maximum power-point tracker without digital signal processing in 0.35 $\mu$ m CMOS for automotive applications. In *2012 IEEE International Solid-State Circuits Conference*, pages 102–104, 2012.
- [26] W. Liu, Y. Wang, and T. Kuo. An adaptive load-line tuning IC for photovoltaic module integrated mobile device with 470 $\mu$ s transient time, over 99% steady-state accuracy and 94% power conversion efficiency. In *2013 IEEE International Solid-State Circuits Conference Digest of Technical Papers*, pages 70–71, 2013.
- [27] S. Uprety and H. Lee. A 43V 400mW-to-21W global-search-based photovoltaic energy harvester with 350 $\mu$ s transient time, 99.9% MPPT efficiency, and 94% power efficiency. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 404–405, 2014.
- [28] Yogesh Kumar Ramadass. Energy processing circuits for low-power applications. *Ph.D. dissertations*, 2009.
- [29] Suyoung Bang, David Blaauw, and Dennis Sylvester. A successive-approximation switched-capacitor DCDC converter with resolution of  $\frac{V_{IN}}{2^N}$  for a wide range of input and output voltages. *IEEE Journal of Solid-State Circuits*, 51:543–556, 2016.
- [30] D. Jeon, Y. Chen, Y. Lee, Y. Kim, Z. Foo, G. Kruger, H. Oral, O. Berenfeld, Z. Zhang, D. Blaauw, and D. Sylvester. An implantable 64nW ECG-monitoring mixed-signal SoC for arrhythmia diagnosis. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 416–417, 2014.

- [31] A. Donida, G. Di Dato, P. Cunzolo, M. Sala, F. Piffaretti, P. Orsatti, and D. Barrettino. A 0.036mbar circadian and cardiac intraocular pressure sensor for smart implantable lens. In *2015 IEEE International Solid-State Circuits Conference - (ISSCC) Digest of Technical Papers*, pages 1–3, 2015.
- [32] K. Paralikar, P. Cong, W. Santa, D. Dinsmoor, B. Hocken, G. Munns, J. Giftakis, and T. Denison. An implantable 5mW/channel dual-wavelength optogenetic stimulator for therapeutic neuromodulation research. In *2010 IEEE International Solid-State Circuits Conference - (ISSCC)*, pages 238–239, 2010.
- [33] X. Liu and E. Sanchez-Sinencio. A single-cycle MPPT charge-pump energy harvester using a thyristor-based VCO without storage capacitor. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 364–365, 2016.
- [34] M. Shim, J. Kim, J. Jung, and C. Kim. Self-powered 30W-to-10mW Piezoelectric energy-harvesting system with 9.09ms/V maximum power point tracking time. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 406–407, 2014.
- [35] Y. Yuk, S. Jung, H. Gwon, S. Choi, S. D. Sung, T. Kong, S. Hong, J. Choi, M. Jeong, J. Im, S. Ryu, and G. Cho. An energy pile-up resonance circuit extracting maximum 422% energy from piezoelectric material in a dual-source energy-harvesting interface. In *2014 IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*, pages 402–403, 2014.
- [36] Xiao Wu, Yao Shi, Supreet Jeloka, Kaiyuan Yang, Inhee Lee, Dennis Sylvester, and David Blaauw. A 66pw discontinuous switch-capacitor energy harvester for self-sustaining sensor applications. *Symposium on VLSI Circuits : [proceedings]. Symposium on VLSI Circuits*, 2016, 06 2016.
- [37] M. Wieckowski, G. K. Chen, M. Seok, D. Blaauw, and D. Sylvester. A hybrid DC-DC converter for sub-microwatt sub-1V implantable applications. In *2009 Symposium on VLSI Circuits*, pages 166–167, 2009.
- [38] Xudong Wang, Jing Liu, Jinhui Song, and Zhong Lin Wang. Integrated nanogenerators in biofluid. *Nano letters*, 7,8:2475–9, 2007.
- [39] Xudong Wang, Jinhui Song, Jin Liu, and Zhong Wang. Direct-current nanogenerator driven by ultrasonic waves. *Science (New York, N.Y.)*, 316:102–5, 05 2007.
- [40] Yong Qin, Xudong Wang, and Zhong Lin Wang. Microfibrenanowire hybrid structure for energy scavenging. *Nature*, 451:809–813, 2008.
- [41] Michelle Rasmussen, Roy E. Ritzmann, Irene A. Lee, Alan J. Pollack, and Daniel A. Scherson. An implantable biofuel cell for a live insect. *Journal of the American Chemical Society*, 134 3:1458–60, 2012.

- [42] Jangwoo Park, Hong-Geun Kim, Yongyun Cho, and Chang-Sun Shin. Simple modeling and simulation of photovoltaic panels using Matlab/Simulink. 2014.
- [43] Xiao Wu, Yao Shi, Supreet Jeloka, Kaiyuan Yang, Inhee Lee, Yoonmyung Lee, Dennis Sylvester, and David Blaauw. A 20-pw discontinuous switched-capacitor energy harvester for smart sensor applications. *IEEE Journal of Solid-State Circuits*, 52:972–984, 2017.
- [44] Y. P. Chen, D. Jeon, Y. Lee, Y. Kim, Z. Foo, I. Lee, N. Langhals, G. Kruger, H. Oral, O. Berenfeld, Z. Zhang, D. Blaauw, and D. Sylvester. An injectable 64 nW ECG mixed-Signal SoC in 65 nm for arrhythmia monitoring. *Solid-State Circuits, IEEE Journal of*, 50:375–390, 01 2015.
- [45] Y. K. Lo, C. W. Chang, Y. C. Kuan, S. Culaclii, B. Kim, K. Chen, P. Gad, V. Edgerton, and W. Liu. A 176-channel 0.5cm<sup>3</sup> 0.7g wireless implant for motor function recovery after spinal cord injury. volume 2016, pages 382–383, 02 2016.
- [46] A. Donida, G. Di Dato, P. Cunzolo, M. Sala, F. Piffaretti, P. Orsatti, and D. Barrettino. A Circadian and Cardiac Intraocular Pressure Sensor for Smart Implantable Lens. *IEEE Transactions on Biomedical Circuits and Systems*, 9(6):777–789, 2015.
- [47] P. Kuo, J. Kuo, H. Hsueh, J. Hsieh, Y. Huang, T. Wang, Y. Lin, C. Lin, Y. Yang, and S. Lu. A smart CMOS assay SoC for rapid blood screening test of risk prediction. *IEEE Transactions on Biomedical Circuits and Systems*, 9(6):790–800, 2015.
- [48] Cymbet. *Rechargeable solid state bare die batteries*, 2016.
- [49] J. Yang, M. Lee, M. Park, S. Jung, and J. Kim. A 2.5-V, 160- $\mu$ J-output piezoelectric energy harvester and power management IC for batteryless wireless switch (BWS) applications. In *2015 Symposium on VLSI Circuits (VLSI Circuits)*, pages C282–C283, 2015.
- [50] M. Ang, R. Salem, and A. Taylor. An on-chip voltage regulator using switched decoupling capacitors. In *2000 IEEE International Solid-State Circuits Conference. Digest of Technical Papers (Cat. No.00CH37056)*, pages 438–439, 2000.
- [51] W. Jung, S. Oh, S. Bang, Y. Lee, Z. Foo, G. Kim, Y. Zhang, D. Sylvester, and D. Blaauw. An ultra-low power fully integrated energy harvester based on self-oscillating switched-capacitor voltage doubler. *IEEE Journal of Solid-State Circuits*, 49(12):2800–2811, 2014.
- [52] Xiao Wu, Kyojin Choo, Yao Shi, Li-Xuan Chuo, Dennis Sylvester, and David Blaauw. 22.6 a fully integrated counter-flow energy reservoir for 70%-efficient peak-power delivery in ultra-low-power systems. pages 380–381, 02 2017.

- [53] M. Konijnenburg, S. Stanzione, L. Yan, D. Jee, J. Pettine, R. van Wegberg, H. Kim, C. van Liempd, R. Fish, J. Schluessler, H. de Groot, C. van Hoof, R. F. Yazicioglu, and N. van Helleputte. A battery-powered efficient multi-sensor acquisition system with simultaneous ECG, BIO-Z, GSR, and PPG. In *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 480–481, 2016.
- [54] Xiao Wu, K. Choo, Y. Shi, L. Chuo, D. Sylvester, and D. Blaauw. A fully integrated counter flow energy reservoir for peak power delivery in small form-factor sensor systems. *IEEE Journal of Solid-State Circuits*, 52(12):3155–3167, 2017.
- [55] K. Okabe, N. Inada, C. Gota, Y. Harada, T. Funatsu, and S. Uchiyama. Intracellular temperature mapping with a fluorescent polymeric thermometer and fluorescence lifetime imaging microscopy. *Nature Communications*, 3(705), 2012.
- [56] D. Chrtien, P. Bnit, H. H. Ha, S. Keipert, R. El-Khoury, Y. Chang, M. Jastroch, H. T. Jacobs, P. Rustin, and M. Rak. Mitochondria are physiologically maintained at close to 50 C. *PLoS biology*, 16(1)(e2003992), 2018.
- [57] H. Bhamra, J. Tsai, Y. Huang, Q. Yuan, and P. Irazoqui. A sub-mm<sup>3</sup> wireless implantable intraocular pressure monitor microsystem. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 356–357, 2017.
- [58] P. P. Mercier, A. C. Lysaght, S. Bandyopadhyay, A. Chandrakasan, and K. M. Stankovic. Energy extraction from the biologic battery in the inner ear. *Nature Biotechnology*, 30:1240–1243, 2012.
- [59] S. O’Driscoll, S. Korhummel, P. C. Cong, Y. Zou, K. A. Sankaragomathi, J. G. Zhu, T. Deyle, A. M. Dastgheib, B. Jian Lu, M. Tierney, J. Shao, C. Gutierrez, S. L. Jones, and H. Yao. A 200 $\mu\text{m}$  200 $\mu\text{m}$  100 $\mu\text{m}$ , 63nW, 2.4GHz injectable fully-monolithic wireless bio-sensing system. *2017 IEEE Radio Frequency Integrated Circuits Symposium (RFIC)*, pages 256–259, 2017.
- [60] K. Yang, Q. Dong, W. Jung, Y. Zhang, M. Choi, D. Blaauw, and D. Sylvester. A 0.6nJ  $-0.22/+0.19^\circ\text{C}$  inaccuracy temperature sensor using exponential sub-threshold oscillation dependence. In *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pages 160–161, 2017.
- [61] W. Lim, T. Jang, I. Lee, H. S. Kim, D. Sylvester, and D. Blaauw. A 380pW dual mode optical wake-up receiver with ambient noise cancellation. In *2016 IEEE Symposium on VLSI Circuits (VLSI-Circuits)*, pages 1–2, 2016.
- [62] Xiao Wu, Inhee Lee, Qing Dong, Kaiyuan Yang, Dongkwun Kim, Jingcheng Wang, Yimai Peng, Yiqun Zhang, Mehdi Saligane, Makoto Yasuda, Kazuyuki Kumeno, Fumitaka Ohno, Satoru Miyoshi, Masaru Kawaminami, Dennis

- Sylvester, and David Blaauw. A 0.04mm<sup>3</sup> wireless and batteryless sensor system with integrated cortex-m0+ processor and optical communication for cellular temperature measurement. *2018 IEEE Symposium on VLSI Circuits*, pages 191–192, 2018.
- [63] Aaron McKenna, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, David Altshuler, Stacey Gabriel, Mark Daly, and Mark A DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research*, 20:1297–303, 09 2010.
- [64] Richard Durbin, Sean R. Eddy, Anders Krogh, and Graeme J. Mitchison. Biological sequence analysis: Probabilistic models of proteins and nucleic acids. 1998.
- [65] Shanshan Ren, Koen Bertels, and Zaid Al-Ars. Efficient Acceleration of the Pair-HMMs Forward Algorithm for GATK HaplotypeCaller on Graphics Processing Units. *Evolutionary Bioinformatics*, 14:117693431876054, 03 2018.
- [66] Altera. Accelerating genomics research with openCL and FPGAs. 2016.
- [67] A. Madhavan, T. Sherwood, and D. Strukov. Race logic: A hardware acceleration for dynamic programming algorithms. In *2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, pages 517–528, June 2014.