2023-11-09

Analog Artificial Neurons and Digital Amplifiers: Challenging the Roles of Analog and Digital Circuit Architectures in Modern CMOS Processes

Taylor S. Barton
Brigham Young University

Analog Artificial Neurons and Digital Amplifiers: Challenging the Roles of Analog and Digital Circuit Architectures in Modern CMOS Processes

Taylor Scott Barton

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Shiuh-hua Wood Chiang, Chair
Karl Warnick
Nancy Fulda

Department of Electrical and Computer Engineering

Brigham Young University

Analog Artificial Neurons and Digital Amplifiers: Challenging the Roles of Analog and Digital Circuit Architectures in Modern CMOS Processes

Taylor Scott Barton
Department of Electrical and Computer Engineering
Masters of Science

BYU Engineering

Abstract

As complimentary metal-oxide semiconductor (CMOS) technologies scale and field-effect transistor (FET) architectures change, the factors in deciding to utilize analog or digital transistor behaviors evolve. This thesis examines three case studies where traditionally analog or digital circuitry has dominated published works but I show that the opposite regime has significant benefits in scaled CMOS technologies. I present a highly digital operational amplifier (traditionally analog) and two artificial neurons (traditionally digital).

In Chapters 2 and 3, I present a highly-digital five-stage zero-crossing-based amplifier which breaks the trade-off between slew rate and settling accuracy. I investigate the optimal charge pump design by analyzing the effects of the current scaling factor, number of current sources, maximum current value, and input amplitude on the settling performance including overshoot and settling time. I find that there exists an optimal number of stages that yields the fastest settling for a given total current and load capacitance. The proposed amplifier achieves a signal-to-noise ratio of 57 dB at a sampling rate of 40 MHz and consumes 1.45 mW under a 1V supply.

In Chapters 4 and 5, I propose two novel analog artificial spiking neurons, operating in the voltage domain and phase domain respectively. The voltage domain neuron presented in Chapter 4 implements a novel fine-tuning method called neuromodulatory tuning which reduced the number of parameters to be tuned by four orders of magnitude as compared with traditional fine-tuning methods. Chapter 5 presents the design of a novel phase-domain neuron. Voltage domain neurons mimic biological neurons by integrating charge on a capacitor. I instead integrate phase in a voltage-controlled ring oscillator (VCO). I also propose a novel bidirectional switched-capacitor synapse which saves significant area compared to bidirectional current based synapses. The proposed neuron, synapse and weight memory occupy only 21x27μm, and consume 134fJ/spike under a 0.35V supply

Keywords: analog neuron, artificial neuron, spiking neuron network, time-domain computing, operational amplifier, charge pump

Acknowledgments

I would like to thank the many people who have assisted in the completion of this thesis. I owe much of my development as an analog designer to my advisor Dr. Shiuh-hua Wood Chiang. Thank you to my labmates, especially Shea Smith, Jared Marchant, Daniel Tebbs, Sharisse Poff, and Yixin Song, who were always willing to discuss ideas, brainstorm, and help solve difficult problems. Thank you to my colleagues in the computer science and neuroscience departments who assisted with the biological foundations, algorithm design and software development: Jordan Yorgason, Nancy Fulda, Karl Warnick, Hao Yu, Kyle Rogers, and Ryan Watson. Thank you to my wife and son for supporting me in my education and research endeavors. Thank you to Paul Ferguson and Stephen Fairbanks for providing valuable insight and feedback during chip testing.

Table of Contents

List of Figures

List of Tables

1 Introduction

1.1 Research Motivation and Description

Complimentary metal-oxide semiconductor (CMOS) field effect transistor (FET) manufacturing has been the driving force in technological innovation for the last several decades. Most recently, deep sub-nanometer planar FETs, FinFET, and GAAFET technologies have enabled ultra-high speed processors, machine learning and artificial intelligence, and supercomputing. As CMOS devices continue to shrink, so do device properties such as intrinsic gain and output impedance. This is a boon to high-speed digital circuits, but hampers the performance of analog circuits. Conversely, lower supply voltages have made subthreshold analog circuit architectures highly attractive in some applications. In some cases, these evolving device properties have made designers consider analog versions of traditionally digital circuits and vice versa. This thesis explores two such circuits circuits: an operational amplifier and an artificial neuron.

Operational amplifiers (op-amps) are usually implemented using analog circuit typologies like the folded cascode or telescopic op-amp. Recently, highly-digital zero-crossing-based amplifiers (ZCBA) have emerged as alternatives because of their lower power consumption and scalability to deep submicron CMOS processes. ZCBAs have found application in pipelined analog-to-digital converters (ADC), $\Delta\Sigma$ ADCs and other circuits [1]–[11]. This thesis presents an optimization algorithm for the ZCBA charge pump design. Then, a thorough analysis of the design of a ZCBA is provided. The ZCBA is simulated using Cadence Virtuoso, and the results verify the optimization algorithm and behavioral model. The proposed ZCBA achieves a signal-to-noise-and-distortion ratio of 57.4dB under a 1V supply at a 40MHz sample rate.

Artificial neurons circuits are used in neural networks to perform machine learning and artificial intelligence tasks. Traditionally, these neurons are implemented in digital circuitry using field-programmable gate arrays or graphics processing units (GPU). While computationally robust, these neurons consume high amounts of power and are therefore not suitable for edge computing applications. Spiking neural networks are neural networks which attempt to mimic the time-based neuronal spiking behavior observed in the human brain. Analog artificial neurons have been shown to be highly efficient at implementing spiking neurons [12]–[18].

This work proposes two low-power analog artificial spiking neurons. The design of a voltage-domain spiking neuron (VDSN) is presented first,

followed by the design and fabrication of a novel phase-domain spiking neuron (PDSN). While many voltage-domain neurons have been proposed to date, the proposed VDSN implements a custom fine-tuning algorithm which significantly reduces the number of weight updates required for a network to learn a new task.

Time-domain and phase-domain computing have been successfully utilized in analog-to-digital converters, time-to-digital converters and other circuits [19]–[24]. Time domain computing takes advantage of fast-switching devices, and experiences less performance degradation from processes scaling as compared to many analog circuits. Further, digital circuits can be ported between processes faster, reducing design costs in commercial applications [25]. I present a novel phase domain neuron that overcomes many of the challenges faced by voltage-domain neurons in deep sub-micron processes. I also propose a switched-capacitor-synapse that is particularly suited to a phase-domain spiking neuron. The VDSN reaches an output spike rate of 3.3Mhz achieving 1.08pF/spike under a 0.4V supply. The PDSN reaches an output spike rate of up to 5.8Mhz and consumes only 134fJ/spike under a 0.35V supply.

1.2 Outline

Chapters 2 and 3 discuss the design of a zero-crossing-based amplifier (ZCBA). Chapter 2 presents an algorithm to optimize the charge pump design for a ZCBA. I show that there exists an ideal current scaling factor for a given set of design constraints. Chapter 3 presents the design and simulation of a ZCBA based on our optimization algorithm. The design of each sub-circuit in a ZCBA is thoroughly analyzed, including a novel two-stage background-calibrated floating-inverter amplifier.

Chapters 4 and 5 present two artificial neuron designs. Chapter 4 presents a voltage-domain neuron specifically design to implement a novel fine-tuning algorithm called neuromodulatory tuning. Chapter 5 presents the design and fabrication of a novel phase-domain spiking neuron, including simulation and measurement results.

2 Multi-Stage Charge Pump Design Optimization for Zero-Crossing-Based Amplifiers

This chapter is composed from a paper entitled "Multi-Stage Charge Pump Design Methodologies for Zero-Crossing-Based Amplifiers" which is under review from the journal "IEEE Transactions on Computer-Aided Design of Integrated Circuits and System." I hereby confirm that the use of this article is compliant with all publishing agreements. The authors on this work are myself as lead author, Shea Smith, Yixin Song, Yen-Chen Kuan, and Shiuh-hua Wood Chiang. With support and advice from the other authors, I designed all the circuits and developed all the software presented in this chapter.

2.1 Introduction

Zero-crossing-based amplifiers (ZCBAs) have been proposed as a replacement for conventional operational amplifiers in applications such as pipelined analog-to-digital converters (ADCs) and delta-sigma ADCs [1], [2], [5]–[11], [26], [27]. Whereas traditional amplifiers implemented in scaled CMOS processes suffer from limited intrinsic gain and headroom, ZCBAs utilizes comparators, charge pumps, and digital logic that directly benefit from technology scaling [26].

The general topology of a ZCBA is shown in Fig. 3.1. The amplifier samples the input on C_{in} and a comparator senses the polarity of the sampled voltage V_X. The comparator decision then controls a charge pump to charge or discharge the load capacitor C_L to drive V_X toward zero through the feedback capacitor C_f. The comparator turns off the charge pump when it detects the zero crossing of V_X to complete the signal amplification. Due to the non-zero comparator decision time, the output of a ZCBA suffers from overshoot V_{OS} as shown in Fig. 2.2(a). With a comparator period T_{comp} and a comparator threshold $V_{th,comp}$, the comparator exhibits a worst-case overshoot $V_{OS} = T_{comp} I_{CP} / C'_L$, where I_{CP} is the charge pump current and C'_L the effective load capacitance. To decrease V_{OS} for a fixed C'_L, T_{comp} and I_{CP} must be reduced. But the former is limited by the technology speed and the latter directly trades off with the amplifier speed (i.e. a smaller I_{CP} gives a smaller V_{OS} but a longer settling time). Therefore, prior works have proposed a 2-stage charge pump to break the speed-precision trade-off [9] [10]. The 2-stage charge pump activates a large current (I_{CP}) until the first

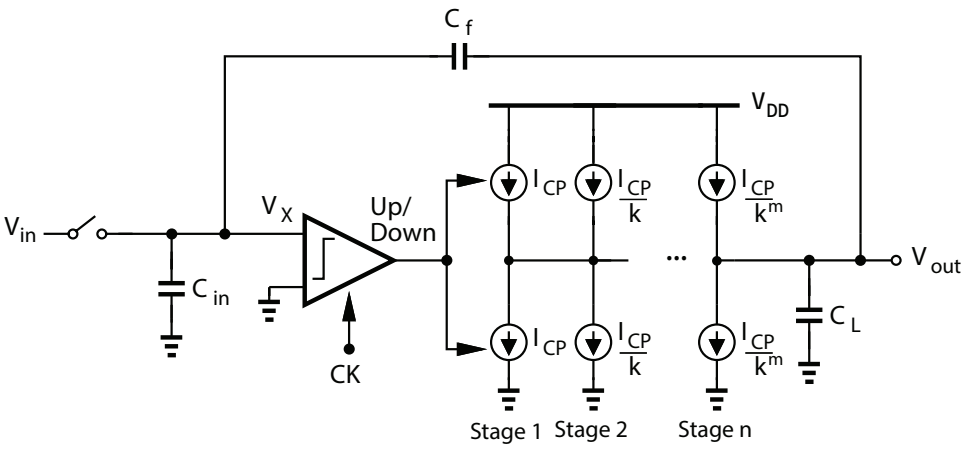

Figure 2.1: Zero-crossing-based amplifier utilizing multiple charge pump current stages with current scaling factor k.

zero-crossing, then a small current (I_{CP}/k) until the second zero-crossing as shown in Fig. 2.2(b). The large current provides fast settling and the small current small overshoot. The amplifier in [1] extends this idea further to six stages. Fig. 2.2(c) shows the output waveform of a generalized m-stage ZCBA where each successive charge pump current is reduced by a factor of k.

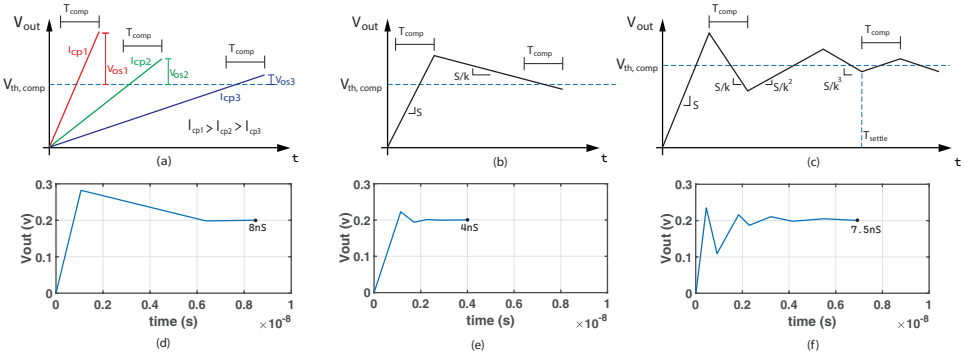

Figure 2.2: (a) Conceptual illustration of ZCBA settling behavior with 1-stage charge pump for different current values. (b) ZCBA settling with 2-stage charge pump. (c) ZCBA settling with n-stage charge pump. Behavioral simulations of ZCBA with (d) 3-stage, (e) 5-stage, and (f) 8-stage charge pump.

While the multi-stage charge pump for ZCBAs has been demonstrated in prior works, the exact scaling factor and number of stages that lead to an optimal settling time have not been investigated. For instance, a small number of stages suffers from a severe trade-off between settling time and overshoot, while a large number of stages suffers from the slow scaling of the current sources, thus resulting in a longer settling time. To understand the complex trade-offs in a ZCBA, this brief investigates the effects of the various circuit parameters including the current scaling factor, number of current sources, current value, and input amplitude on the settling performance.

We develop a generalized behavioral model of the ZCBA that allows us to vary the design parameters to analyze the circuit. Finally, we validate our behavioral model with transistor-level simulations of a complete multi-stage ZCBA. The numerical tool presented allows us to predict the optimal charge pump design for ZCBAs from the circuit parameters.

2.2 Mathematical and Behavioral Model

The overshoot error during each stage of amplification can be predicated mathematically. Due to the input-dependent nature of the overshoot error, we use a statistical model that predicts settling time with good accuracy. Fig. 2.3 shows a graphical representation of our analysis. Fig. 2.4 (a) compares the results of the analysis with our behavioral model.

Due to uniform comparator clock periods and the approximately linear behavior of the current sources on a small time scale, The overshoot error V_{OS} is uniform distributed as $V_{OS} = (I_{cp}/C_L) \times U(0, T_{comp})$

It follows that across many zero crossings, the average overshoot $V_{OS,avg}$ is $I_{cp,n}T_{comp}/2C_L$ Using this simplification, we can write the settling time for an m stage ZCBA as:

$$T_{settle} = T_{comp} \sum_{n=1}^{m} \lceil \frac{V_{x,n}}{\Delta V_n} \rceil \tag{2.1}$$

Where $V_{x,n}$ is the virtual ground voltage before the n^{th} current stage and $\Delta V_n = I_{cp,n}T_{comp}/C_L$ for the n^{th} current stage. This equation is not continuous and therefore not differentiable. Numerical methods are required to find a minimum.

We develop a MATLAB program that models the behavior of the amplifier with a generalized multi-stage charge pump. The program simulates the amplification cycle for different values of $I_{CP,total}$ (total charge pump current), m (number of current stages), k (current scaling factor), and V_{in}, and gives the settling waveform and settling time T_{settle}. The goal in developing the behavioral model was to gain intuition into the effects of circuit parameters on settling time. In this model, I_{CP} (first stage current), m, k, and $I_{CP,total}$ are related by $\sum_{n=0}^{m-1} I_{CP}/k^n = I_{CP,total}$. Algorithm 1 shows the pseudo-code of our program.

Fig. 2.2(d)-(f) show the simulation results of the ZCBA with three different charge pump designs (3-stage, 5-stage, and 8-stage). In these simulations, we set T_{comp} = 500 ps, C_L = 1 pF, and I_o = 1 uA (final stage current). Our choice of I_o assumes an output swing of 1.6 V and a 10-bit settling error target (1.5 mV), yielding a worst-case overshoot of ±500 μV. Simulations show that the respective T_{settle}'s are 8, 4, and 7.5 ns for the 3, 5, and 8-stage designs, indicating a strong dependency of the settling time on the number of stages as postulated earlier. Next, we analyze the effect of V_{in} on T_{settle} by sweeping the input. Fig. 2.4(a) shows T_{settle} versus V_{in} of the 5-stage ZCBA over its input range. We observe that while a larger V_{in} will generally take longer to settle, the characteristic is not monotonic. This is due to the nonlinear settling behavior introduced by the discrete current stages.

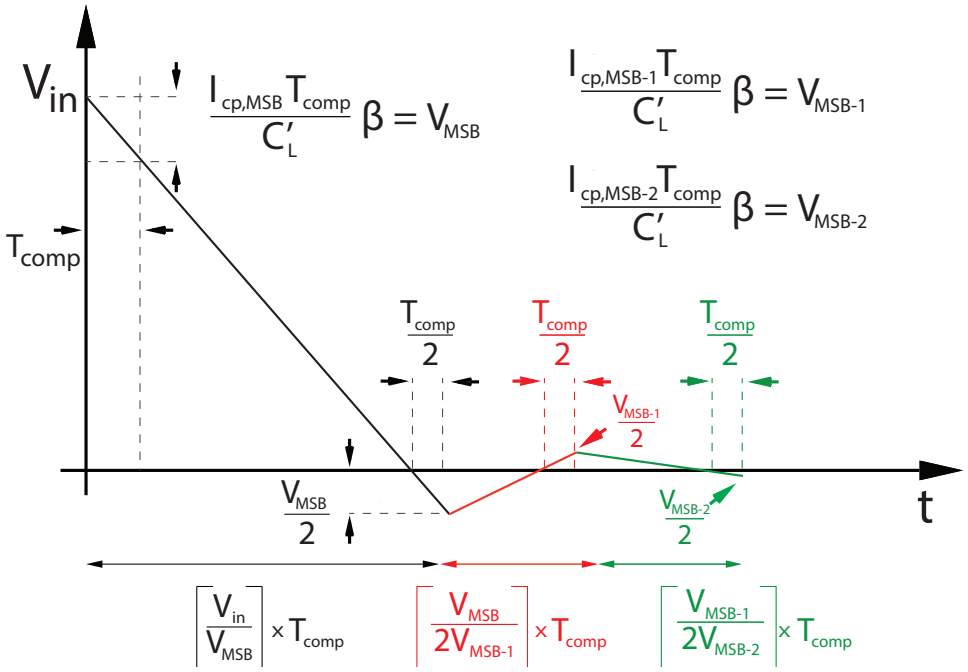

Figure 2.3: Visualisation of the ZCBA settling time analysis. C'_L is the equivalent load capacitor, and β is the amplifier's feedback factor.

Therefore, we numerically compute the *average* settling time $T_{settle,avg}$ over the input range for a given design. We then compare $T_{settle,avg}$ across designs for a fair evaluation. Fig. 2.4(c) shows the simulated settling time of the corresponding transistor-level design, whose details will be described in Section III."

Fig. 2.5 shows the average settling time $T_{settle,avg}$ of a ZCBA as we vary the number of current stages m. In this simulation, we keep $I_{CP,total}$ and I_o constant across designs so as to maintain the same total current and settling error for a fair comparison. The results show that $T_{settle,avg}$ decreases rapidly as m increases from 2, reaching a minimum at 5, then increases thereafter.

Algorithm 2.1 Generalized Zero-Crossing-Based Amplifier Behavioral Model.

```
FOR (I_tot = 10uA; I_tot < 500uA; I_tot += 10uA)
    FOR (m = 2; m < 10; m += 1)
        FOR (V_in = 0; V_in < V_in,max; V_in += 1mV )
            WHILE (zero-crossings < m)
                I_CP = computeNextCurrent(I_tot, m, V_in)
                incrementV_out(I_CP)
                IF V_X crossed V_th,comp
                    scaleI_CP()
                    ++zero-crossings
                saveSettlingTimeData()
```

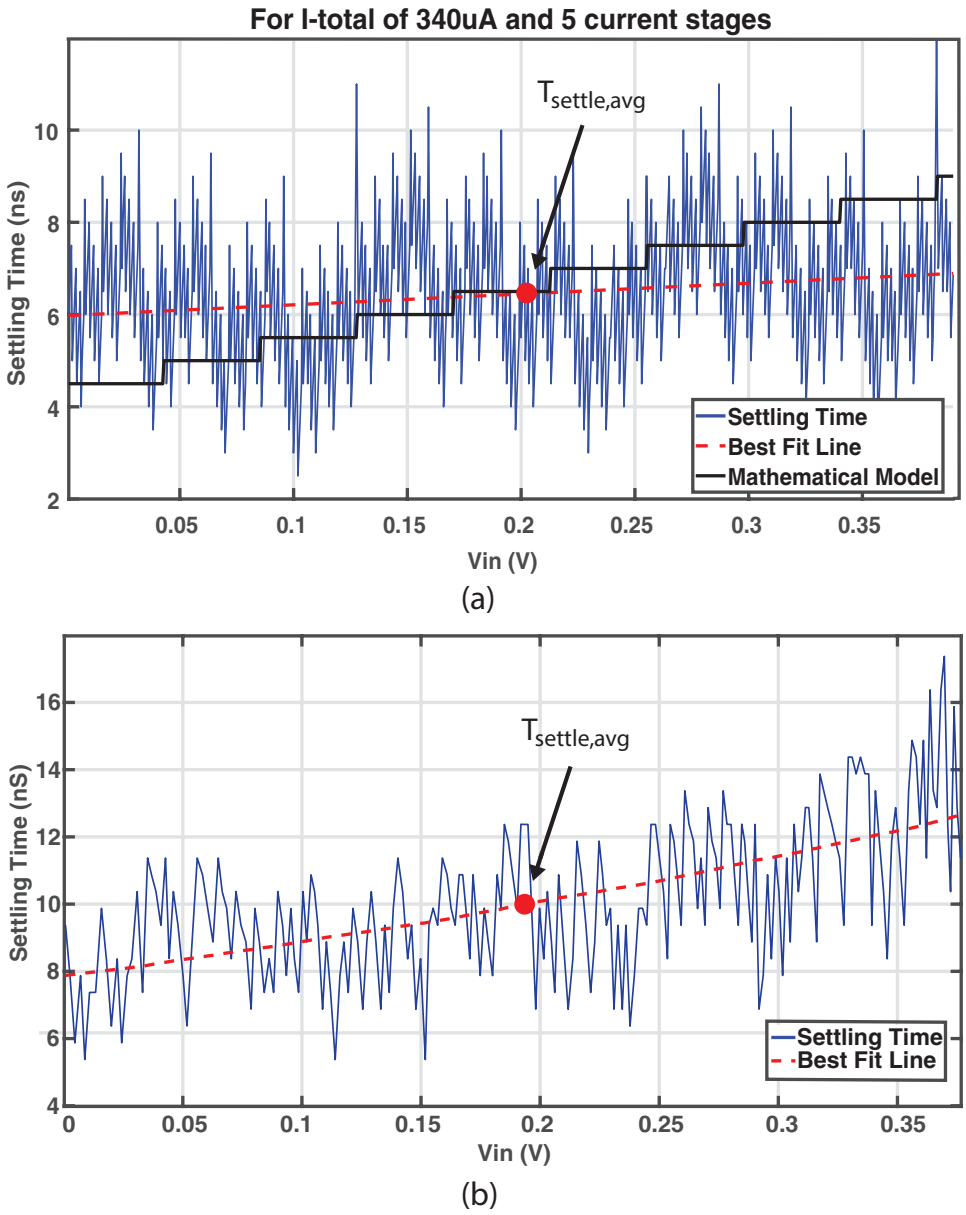

Figure 2.4: T_{settle} versus V_{in} from (a) behavioral simulation and mathematical model. (b) transistor-level simulations.

The large $T_{settle,avg}$ for low m is due to the large overshoot from the first current stage and the need to correct the overshoot with small currents in the subsequent stages. $T_{settle,avg}$ is large for high m because of the slower current scaling, i.e. more time is spent waiting for the comparator to make decisions for a larger number of stages.

The foregoing analysis suggests that an optimal number of stages exists to produce a minimum settling time, and that a 5-stage design is optimal for the particular circuit parameters chosen in the above simulations. For a more general case, we sweep both the total current $I_{CP,total}$ and number of stages m and observe $T_{settle,avg}$. Fig. 2.2 shows $T_{settle,avg}$ versus $I_{CP,total}$

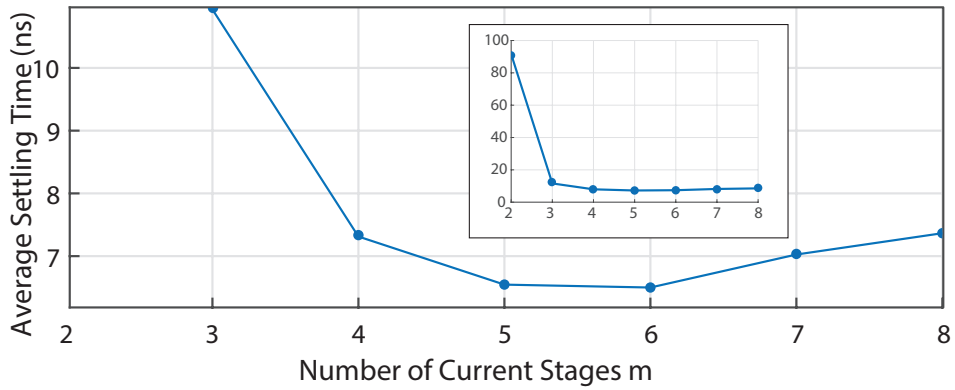

Figure 2.5: Average settling time versus number of stages m for $I_{CP,total}$ = 340 μA including charge pump nonlinearity.

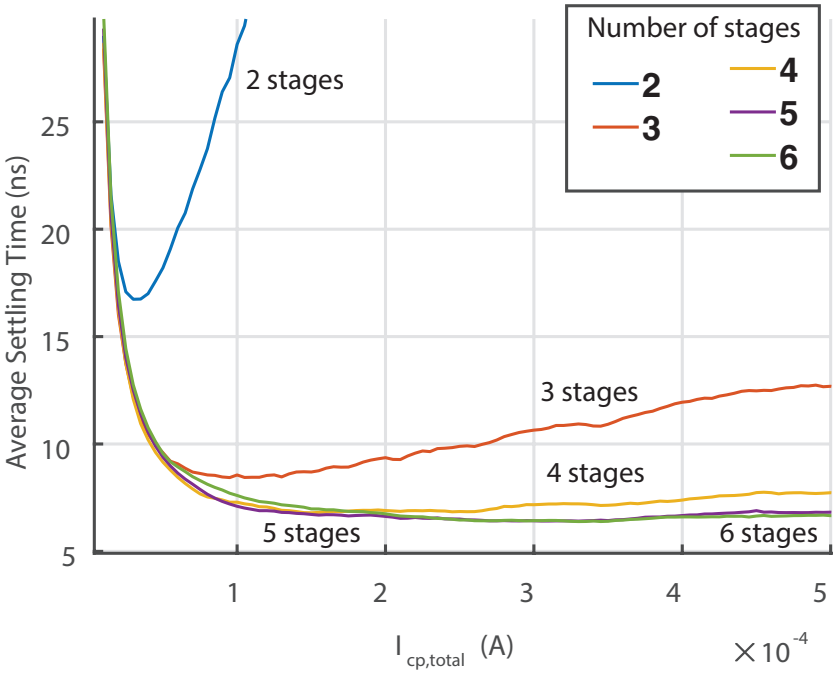

Figure 2.6: Average settling time versus $I_{CP,total}$ for different number of stages.

for different m's and C_L = 1 pF. We observe that as $I_{CP,total}$ increases, the optimal m increases. $T_{settle,avg}$ decreases with $I_{CP,total}$ but it becomes a weak function of $I_{CP,total}$ after about 400 μA, suggesting that using more current after that point will face diminishing returns. With the above results, we can predict the optimal total current and number of current stages for any load by multiplying both C_L and $I_{CP,total}$ by a constant α. Figure 2.8 is a 3D plot showing the asymptotic behavior of the settling time versus $I_{CP,total}$ and m. For our transistor-level design in the next section, C_L = 1 pF and we choose $I_{CP,total}$ = 340 μA and m = 5 for $T_{settle,avg}$2 ns. This design choice strikes a good balance by achieving a near-minimum settling time while

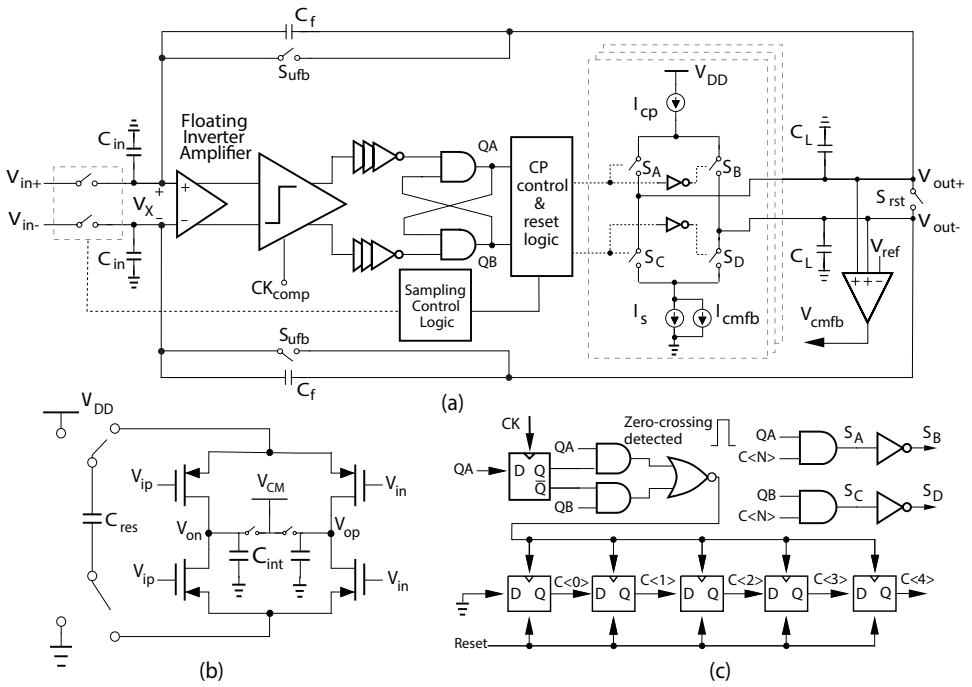

Figure 2.7: (a) Full circuit schematic diagram of the proposed ZCBA. (b) Schematic of floating inverter amplifier. (c) Schematic of charge pump current control and reset circuit.

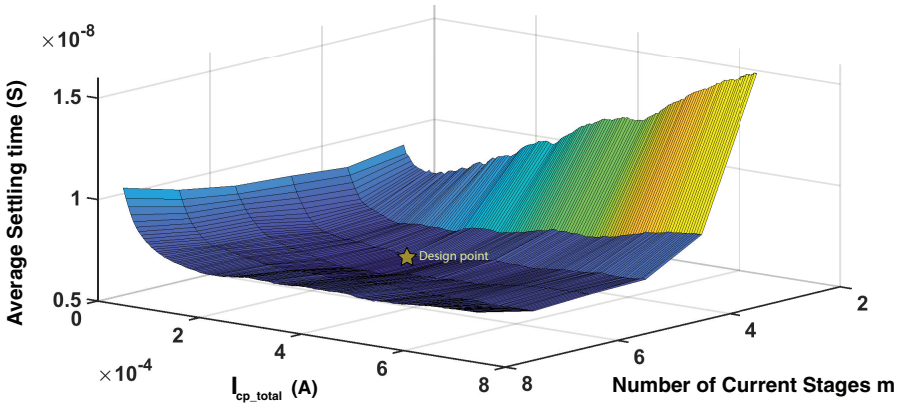

Figure 2.8: Surface plot of average settling time versus $I_{CP,total}$ and number of stages m. We choose the design point at a minimum of the surface: 5 stages with 340 μA total current.

keeping the current consumption and number of stages low as shown by the design point indicated in Fig. 2.8. Performance levels are similar beyond our chosen design point, but require either a higher current or more stages, which increase power, chip area, and circuit complexity. Our numerical tool allows us to predict the optimal charge pump design for ZCBAs. Noise adds random variations to the trends predicted by our analysis, and can be

suppressed by increasing power consumption. We use our behavioral tool in combination with transistor-level simulations of charge pump linearity to analyze the effect of charge pump nonlinearity on settling time. The simulation shows the same trend as the ideal case. Our tool predicts that nonlinearity increases the average settling time from 6.4 ns to 7.12 ns, an 11.25% increase.

The above simulations are based on an amplifier with a closed-loop gain (A_{CL}) of 2 V/V, but the optimal scaling factor is the same regardless of A_{CL}. This is because the optimal scaling factor is determined by the *average* settling time across all valid V_{in} (i.e. V_{in}'s within the input range). Since the maximum value for V_{in} is set by the amplifier's output swing divided by A_{CL}, it follows that if A_{CL} increases, the maximum V_{in} value must decrease by the same factor to maintain the same output swing, and vice versa. Simulations confirm that the optimal scaling factor across all valid V_{in} for a given A_{CL} is the same for any A_{CL}.

	This Work	[1]	[2]	[7]	[10]
Architecture	**ZCBA**	MASH ADC	ΔΣ ADC	Pipeline ADC	Pipieline ADC
Process (nm)	**28**	65	65	65	90
Supply (V)	**1**	1.2	1	1.2	1.2
F_s (MHz)	**40**	40	0.4**	26	50
SNDR (dB) *	**57.4**	70.4	61.1	54.3	62
SFDR (dB) *	**71.5**	90	--	70.4	68
ENOB *	**9.25**	11.4	9.85	8.7	10
Power (mW)*	**1.45**	3.73	0.948	1.78	4.5

*Reported values are for [1], [2]. [7] and [10] are for the ADC not the ZCBA only
** Signal Bandwidth

2.3 Transistor-Level Design and Simulations

We design a transistor-level ZCBA in a 28-nm CMOS process using parameters determined from our optimization. We then compare the transistor-level simulations with behavioral simulations to validate our analysis. . Fig. 2.7 shows the schematic of the ZCBA, which consists of a comparator, control logic, and 5-stage charge pump. The charge pumps send a differential current I_{CP} onto two load capacitors C_L which ramps the output voltage V_{out} up or down. The capacitor C_f creates a negative feedback to force V_X towards zero. In this design, C_f = 500 fF, C_L = 1 pF, $I_{CP,total}$ = 340 μA, and T_{comp} = 500 ps. The size of C_L is dictated by the magnitude of the smallest charge pump current and the desired overshoot error. The smallest current source must be sized for tolerable mismatch. Similar to the phase offset in a phase-locked loop, charge pump mismatch creates an output offset which reduces the output swing of the amplifier

To minimize the input-referred noise and kickback of the comparator, we utilize a floating-inverter pre-amplifier (FIA) before the comparator. Compared to a conventional differential common-source amplifier, the FIA features lower power consumption and more robustness against common-mode variations [28], [29]. The comparator assumes the StrongARM topology to sense the virtual ground voltage V_X to make a decision. An SR latch and combinational logic monitor the comparator decisions to detect the zero crossing, and control a bank of current stages in the charge pump. The logic detects a zero crossing when the previous decision is different from the current decision. The current stages are all active at the beginning of the amplification phase, and the charge pump logic turns off one stage at a time by advancing the "off" signal in a chain of registers as it detects zero crossings. At each step, the charge pump current is scaled by a factor $k = 4$ and the current polarity is reversed. These steps are repeated until the last current stage. A common-mode feedback (CMFB) amplifier senses the output common mode and adjusts the charge pump current to set the common mode to V_{ref}. The ZCBA forces V_X toward zero by alternating the current direction and successively scaling down the currents, producing the amplified V_{out}. Fig. 2.4(b) shows the transistor-level simulation results of T_{settle} versus V_{in}, which exhibits a similar profile as our behavioral simulations, Fig. 2.4(a), with $T_{settle,avg} = 10.1$ ns. The good agreement between the behavioral- and transistor-level simulations validates our behavioral model and analysis. Our ZCBA model and analysis allows the designer to rapidly select optimal system-level parameters for a given set of constraints, reducing the time-consuming transistor-level design iterations.

The trend and shape of the V_{in} vs T_{settle} plots are the same, albeit with a small offset. We acknowledge the limits of our mathematical tool, which is not designed to simulate process dependant parameters and nuanced transistor level behavior. The main effect not modeled by our mathematical tool is charge sharing error V_{CS} between C_L and the charge pump parasitic capacitance. Each time the charge pump switches directions, an error V_{CS} is induced on the output. For example, on the final current stage, V_{CS} ranges between $1 - 2$mV. With $I_o = 1$uA, it takes $1 - 2$nS for the output to return to it's value before the switching-induced error.

2.4 Conclusion

The multi-stage charge pump in a ZCBA presents complex trade-offs. This brief investigates the optimal design of the charge pump by analyzing the effects of the current scaling factor, number of current sources, current value, and input amplitude on the ZCBA's settling performance with the aid of a generalized behavioral model. We find that there exists an optimal number of stages that gives the fastest settling time for a given total current and load capacitance. We also find that the settling time approaches a limit as the total current increases. We validate our behavioral analysis with transistor-level simulations of a ZCBA and confirm that the two models are in good agreement with each other. The model developed in the brief allows us to predict the optimal charge pump design for ZCBAs.

3 A Multi-Stage Zero-Crossing-Based Amplifier Using Floating-Inverter Amplifier With Background Offset Calibration and Self-Timed Loop

This chapter is composed from a paper entitled "A Multi-Stage Zero-Crossing-Based Amplifier Using Floating-Inverter Amplifier With Background Offset Calibration and Self-Timed Loop" which will be published in the proceedings of the "66th IEEE International Midwest Symposium of Circuits and Systems". I hereby confirm that the use of this article is compliant with all publishing agreements. The authors on this work are myself as lead author, Yen-Chen Kuan, and Shiuh-hua Wood Chiang. With support and advice from the other authors, I designed all the circuits presented in this chapter.

3.1 Introduction

Zero-crossing-based amplifiers (ZCBAs) have been proposed as an alternative to the conventional operational amplifier in applications such as pipelined analog-to-digital converters (ADCs) and delta-sigma ADCs [1]–[11]. The idea of these ZCBAs is to replace the classic op amp structures with more digital- and scaling-friendly circuits, such as dynamic comparators and switched current sources. For instance, the works in [6], [9] propose a ZCBA using an inverter-based comparator to improve power efficiency. However, inverter-based comparators have ill-defined bias currents and feature single-ended signaling. Additionally, their threshold voltages vary over process-voltage-temperature (PVT) corners. The ZCBA in [30] employs a voltage-controlled oscillator (VCO) as the comparator. However, this VCO must operate at a high frequency to obtain a 10 MHz amplifier bandwidth. A subthreshold version of this VCO-based ZCBA faces a similar challenge [31]. In addition, both works [30], [31] require a loop filter for stability.

This paper proposes a fully dynamic ZCBA using mostly digital blocks and switched current sources that are amenable to technology scaling. The ZCBA employs a fully differential two-stage floating inverter amplifier (FIA) with background offset calibration using the feedback loop from the zero-crossing detection comparator. A novel self-timed loop controls the FIA to increase the time available for amplification to improve gain. Additionally, a five-stage charge pump breaks the trade-off between slewing and overshoot

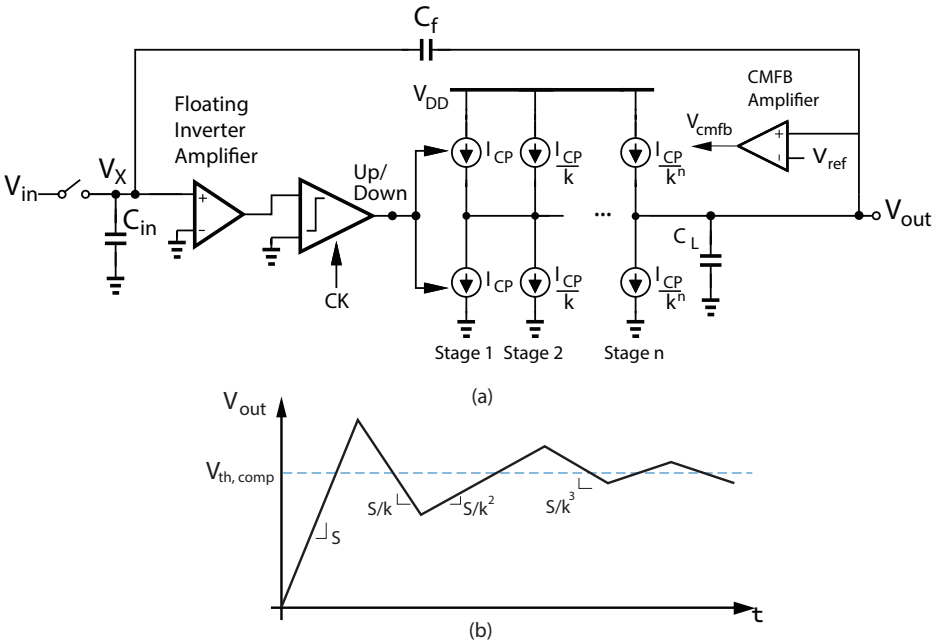

Figure 3.1: (a) Simplified block diagram of an n-stage radix-k ZCBA with a floating-inverter preamplifier and (b) its output settling waveform.

to minimize the settling time. Simulations of the proposed ZCBA in a 28-nm CMOS process show a signal-to-noise ratio (SNR) of 57.4 dB at a sample rate of 40 MHz while consuming 1.45 mW.

3.2 Circuit Design

Fig. 3.1(a) shows the block diagram of our proposed ZCBA. The amplifier samples the input on C_{in} and the FIA amplifies the sampled voltage V_X. The comparator determines the FIA output polarity and controls the current sources in a charge pump to charge or discharge the load capacitor C_L , driving V_X toward zero through the feedback capacitor C_f. The comparator turns off the most-significant-bit (MSB) current source when it detects a zero-crossing of V_X and activates the MSB-1 current source to drive V_X in the opposite direction. This process repeats until the least-significant-bit (LSB) current source is reached. Due to the non-zero comparator decision time, the output of the ZCBA suffers from an overshoot V_{OS}. With a comparator period T_{comp}, the ZCBA exhibits a worst-case overshoot $V_{OS} = T_{comp}I_{CP}/C'_L$, where I_{CP} is the charge pump current and C'_L is the effective load capacitance. Prior works have proposed a two-stage charge pump to break the speed-precision trade-off [9], [10], [26]. The two-stage charge pump activates a large current (I_{CP}) until the first zero-crossing, then a small current (I_{CP}/k) until the second zero-crossing. The large current provides fast slewing and the small current small V_{OS}. The amplifier in [1] extends this idea further to six stages. In our design, we utilize five stages with an optimized radix for the fastest settling based on our behavioral simulations. Fig. 2(a) shows the architecture of the proposed ZCBA, which includes a two-stage FIA preamplifier, dynamic

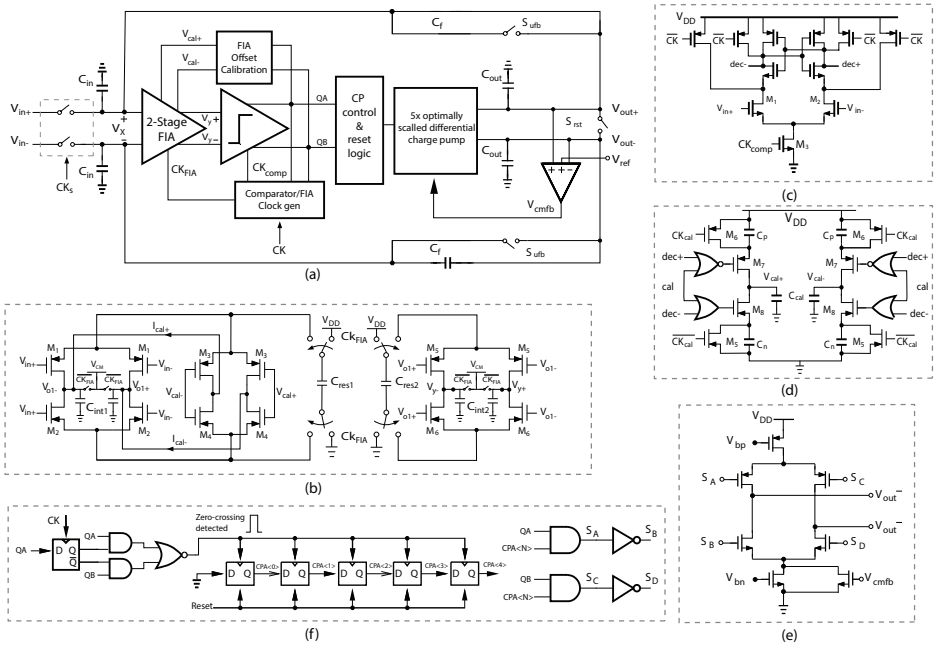

Figure 3.2: (a) Architecture of proposed ZCBA. Schematic of (b) two-stage floating-inverter amplifier with offset calibration, (c) StrongARM comparator, (d) offset calibration circuit, (e) one of the five charge pump stages, and (f) charge pump control and reset logic.

comparator, charge pump, and switched current sources. In addition, a common-mode feedback circuit defines the output common mode and clock generator generates the internal clocks from a master clock. The closed-loop gain is set by the ratio of C_{in} (1 pF) to C_f (500 fF) and the amplifier drives the load C_{out} (1 pF).

3.2.1 Two-Stage Floating-Inverter Amplifier and Comparator

Our design implements a novel background-calibrated FIA as the comparator preamplifier to suppress the comparator noise and kickback. The schematic of our FIA is shown in Fig. 3.2(b). The FIAs in [28], [29] offer several benefits over a traditional inverter amplifier. They consume no static power and tolerate input common-mode variations by virtue of the floating supply. Our two-stage FIA achieves a gain of approximately 28 dB, which is comparable to the three-stage common-source preamplifier reported in [1]. A two-stage FIA was also proposed in [32] as a residue amplifier in a pipelined ADC. Compared to the FIA in [32], our FIA has 5 dB less open-loop gain (28 dB versus 33 dB) but uses 80× smaller C_{res1} (150 fF), and 7.5× smaller C_{res2} (47 fF). Further, our FIA operates at a sampling frequency of 2 GHz, which is much higher than that of [32] (10 MHz). We implement the comparator using the StrongARM topology for its zero static power consumption and high speed operation (Fig. 3.2(c)). The combined input-referred noise of the proposed FIA and comparator combination is 175 μV, sufficient to meet the noise level requirement of the ZCBA.

The FIA and comparator are clocked by a self-timed loop, similar to the asynchronous successive-approximation-register ADC. The idea is to initiate the FIA amplification as soon as the previous comparison is done, thus maximizing the FIA amplification time. The schematic and timing diagram of our self-timed circuits are shown in Fig. 3.3. At each comparator cycle, and external clock source CK clocks a reset-low D flop-flop (DFF) $D1$, whose input is tied to V_{DD}, which pulls CK_{comp} high. A NAND gate at the comparator output is used to detect the completion of the comparison. The NAND gate output resets $D1$ to V_{SS} when the comparator decision is complete. This pulls CK_{comp} and CK_{FIA} low resetting the comparator and FIA, respectively. Once the comparator is reset, CK_{FIA} is pulled high to start amplification. This loop scheme increases the FIA amplification time from $1/(2f_{ck}) = 250$ ps for uniform clocking to approximately 415 ps on average over multiple cycles.

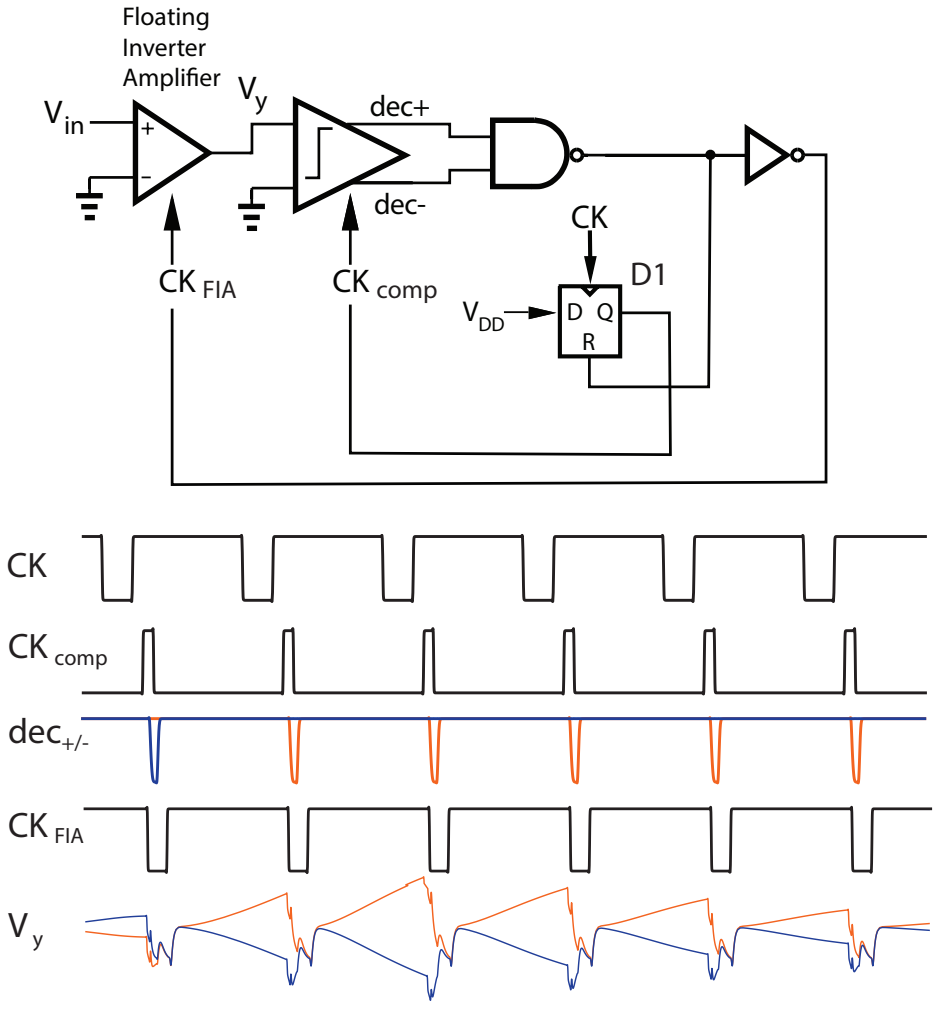

Figure 3.3: Schematic and timing diagram of the self-timed logic for the FIA and comparator loop.

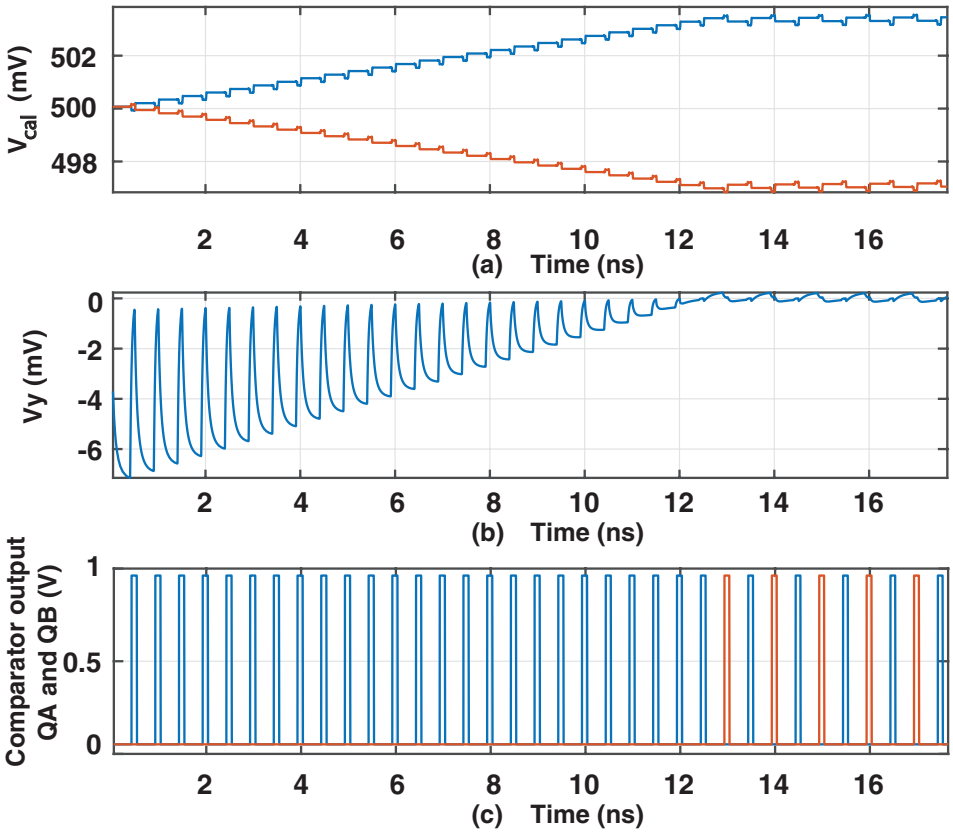

Figure 3.4: Operation of the FIA offset calibration circuit. (a) Calibration circuit outputs V_{cal+} and V_{cal-}, (b) FIA differential output, (c) comparator outputs.

3.2.2 Background Calibration for FIA Offset

In our particular topology, if the combination of the FIA and comparator have an effective input-referred offset of V_{OS}, and the ZCBA has a closed-loop gain of A_{CL}, then the ZCBA output will settle to a value of $V_{in}A_{CL} + V_{OS}A_{CL}$. The term due to the offset can significantly reduce the output swing and may even saturate the amplifier. In this regard, we propose calibrating this offset in the background by sensing the comparator output and correcting the offset via the FIA.

We implement a charge-sharing-based calibration circuit similar to [33]. Fig. 3.2(d) depicts the schematic of the calibration circuit. C_p and C_n are minimum-sized MOS capacitors and C_{cal} is a large capacitor of 1 pF. When CK_{cal} is low, C_p and C_n are reset to V_{DD} and V_{SS}, respectively through M_5 and M_6. When CK_{cal} goes high, the comparator, with its inputs shorted, makes a decision. Depending on the decision, it turns on either M_7 or M_8. This causes charge sharing between C_p (or C_n) and C_{cal}, raising (or lowering) $V_{cal} = V_{cal+} - V_{cal-}$. V_{cal+} and V_{cal-} connect to a second pair of inverters in the FIA, M_3 and M_4, which create a differential current I_{cal+} and I_{cal-} to cancel the offset. We size M_3 and M_4 significantly smaller than M_1 and M_2 to minimize the noise contribution. Fig. 3.4 shows the simulated waveforms of the proposed calibration. The calibration loop ramps the calibration voltages

V_{cal+} and V_{cal-} to null the differential FIA output V_y. In steady-state, the comparator output toggles between positive and negative decisions to show that the offset has been removed. The proposed calibration runs in the background during the reset phase to track any changes in the offset.

3.2.3 Five-Stage Differential Charge Pump

The proposed ZCBA employs a five-stage differential charge pump with an optimized current source scaling factor. We use a ZCBA behavioral model to predict the settling time versus current source scaling factor of the ZCBA. We chose our charge pump currents based on this optimization to minimize the settling time. The currents in stages 1 to 5 are 300 μA, 75 μA, 20 μA, 5 μA, and 1 μA, respectively. The schematic of one of the charge pump stages is shown in Fig. 3.2(e). Fig. 3.2(f) illustrates the charge pump control logic, which consists of a cascade of DFFs and a zero-crossing detector. The DFFs are reset to high at the start of each amplification cycle, activating all of the charge pump stages. When the n^{th} decision and the $(n-1)^{th}$ decision are different, a zero-crossing is detected and the control logic clocks all the DFFs. This shifts a low signal into the first DFF output, deactivating the MSB charge pump to reduce the current. Subsequent zero-crossings sequentially deactivate additional charge pumps until the LSB+1 unit. The LSB charge pump always remains active except during reset.

The LSB charge pump includes a common-mode feedback (CMFB) circuit to set the ZCBA's output common mode. A low-power CMFB amplifier senses the output common mode and compares it against a reference of $V_{DD}/2$. The CMFB amplifier controls a pull-down transistor in the LSB charge pump, driving the output CM toward $V_{DD}/2$. To ensure stability, the CMFB loop gain is kept low by including the pull-down transistor only in the LSB charge pump.

3.3 Simulation Results

We simulated the proposed ZCBA in a 28-nm CMOS technology. The ZCBA has a closed-loop gain A_{CL} of 2 V/V and a sampling rate f_s of 40 MHz while consuming 1.45 mW under a 1-V supply. The power is dominated by the FIA and comparator including the self-timed loop and calibration. They consume 1.1 mW, accounting for 77% of the total power. Fig. 3.5 shows the simulated waveforms of the ZCBA. The output voltages slew rapidly initially due to the large charge pump current and gradually settle to the final value with smaller ramps due to the reduced charge pump current. We can also observe that the virtual ground voltages converge toward each other due to the negative feedback. Finally, we see the asynchronous charge pump control logic deactivating the charge pumps sequentially. Fig. 3.6 shows the ZCBA output spectra from a near-DC input and a near-Nyquist input. The ZCBA exhibits a signal-to-noise-and-distortion ratio (SNDR) and spurious-free-dynamic range (SFDR) of 59.8 dB and 72.1 dB at DC, respectively, and 57.4 dB and 72.1 dB at Nyquist, respectively.

Table I compares our ZCBA to state-of-the-art non-traditional amplifiers. Note that the reported data from [5], [8], [9] and [11] are from ZCBAs

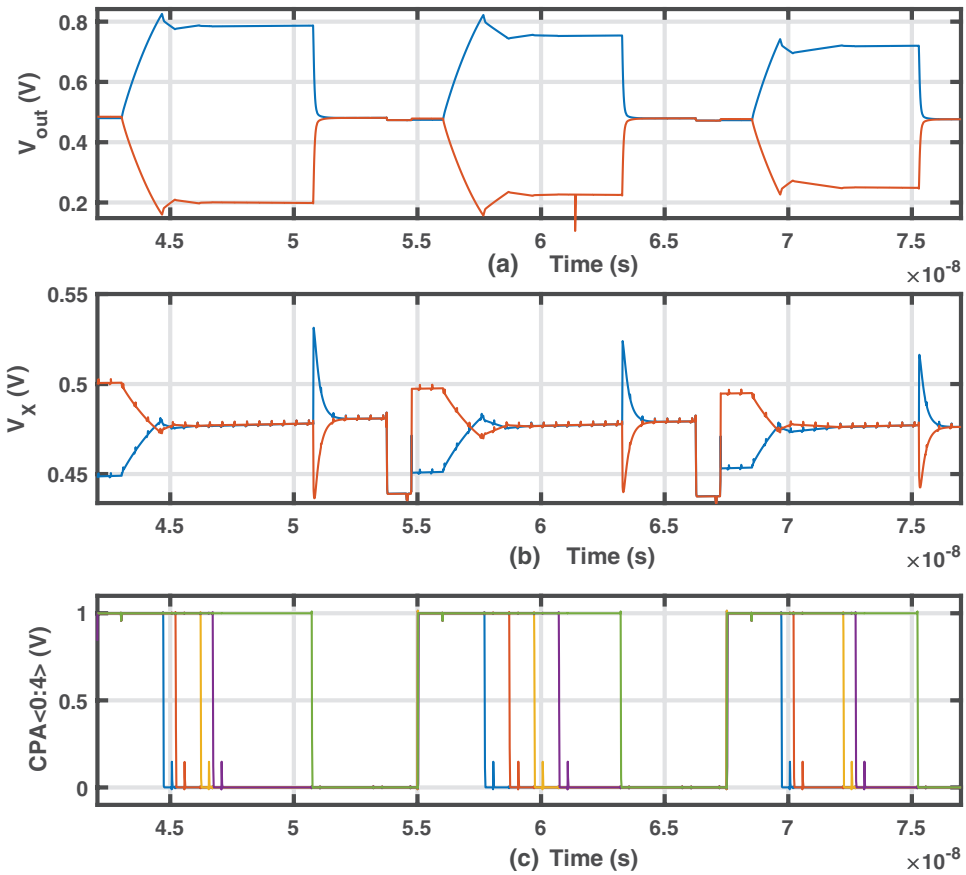

Figure 3.5: Time-domain waveforms at the (a) differential output, (b) virtual ground nodes V_X, and (c) charge pump control logic output.

embedded in ADCs. This work achieves a competitive sampling rate, SNDR, and power, and presents several unique circuit techniques including the FIA, self-timed circuit, background offset calibration, and multi-stage charge pump designs.

3.4 Conclusion

We present the design of a five-stage zero-crossing-based amplifier in a 28-nm CMOS technology. The proposed ZCBA utilizes five stages of scaled charge pumps to break the trade-off between speed and settling accuracy. Our proposed two-stage FIA incorporates background offset calibration using the zero-crossing detection comparator. In addition, a self-timed loop increases the available time for amplification to increase the gain. The scaled five-stage charge pump reduces the overshoot while offering a fast slew rate. The design achieves an SNDR of 57.4 dB at a sampling rate of 40 Mhz and consumes only 1.45 mW in simulations.

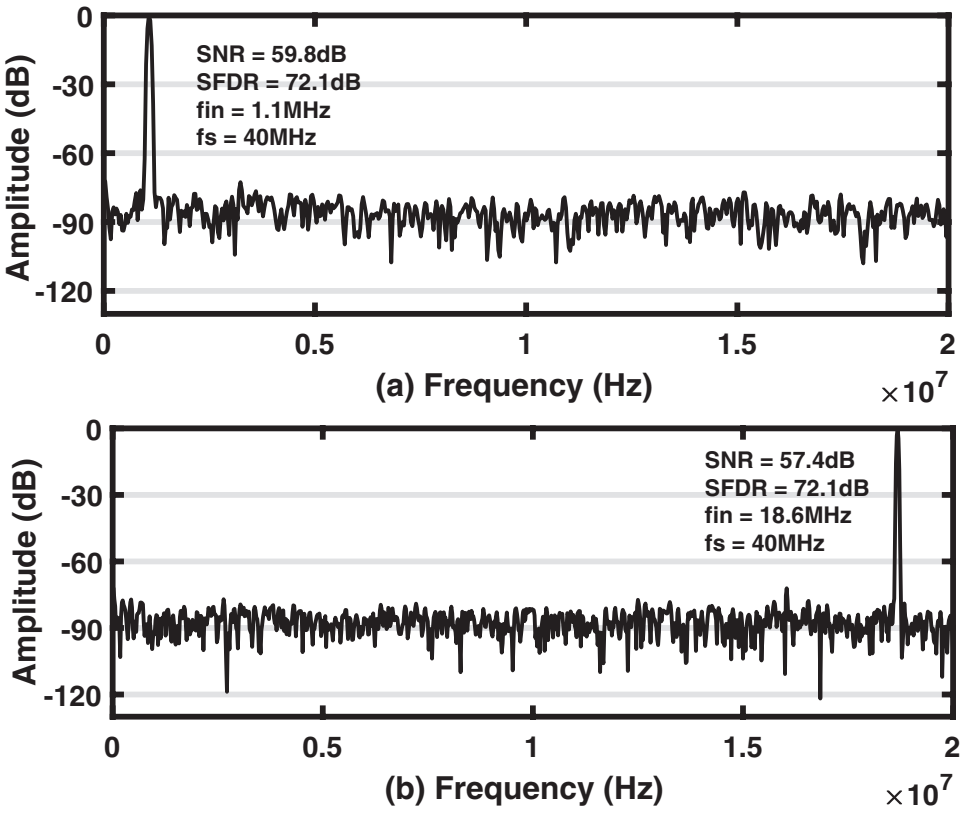

Figure 3.6: Simulated output spectra of the proposed ZCBA with (a) near-DC and (b) near Nyquist inputs.

Table 3.1: Performance Summary and Comparison.

	This Work	[1]	[6]	[9]	[11]
Architecture	**ZCBA**	MASH ADC	Pipeline ADC	Pipieline ADC	Pipeline ADC
Process (nm)	**28**	65	65	90	130
Supply (V)	**1**	1.2	1.2	1.2	1.2
F_s (MHz)	**56**	40	26	50	200
SNDR (dB) *	**59**	70.4	54.3	62	53
SFDR (dB) *	**62**	90	70.4	68	63
ENOB *	**9.25**	11.4	8.7	10	8.5
Power (mW)	**1.65**	3.73	1.78	3.8	38

*Reported values are for [1], [6]. [9] and [11] are for the ADC not the ZCBA only

4 Towards Low-Power Machine Learning Architectures Inspired by Brain Neuromodulatory Signalling

This chapter is composed from a paper entitled "Towards Low-Power Machine Learning Architectures Inspired by Brain Neuromodulatory Signalling" which is published in the journal "MDPI Journal of Low-Power Electronics and Applications" [34]. I hereby confirm that the use of this article is compliant with all publishing agreements. The authors on this work are myself, Hao Yu, and Kyle Rogers as lead authors, Nancy Fulda, Shiuh-hua Wood Chiang, Jordan Yorgason and Karl F. Warnick. I designed and simulated the artificial neuron circuit presented in this chapter. All learning algorithms, training tasks and biological foundations were developed by the other authors.

4.1 Introduction

Analog CMOS hardware has the potential to reduce energy consumption of deep neural networks by orders of magnitude, but the *in situ* training of networks implemented on such hardware is challenging. Once the chip has been programmed with the correct weight values for a task, typically no further learning occurs. We introduce a biologically-inspired knowledge transfer approach for neural networks that offers potential for *in situ* learning on the physical chip. In our method, the weight matrices of a spiking neural network [35]–[39] are initialized with values learned via offline (i.e., off-chip) methods, and the system is exposed to an analogous—but distinct—learning task. The bias inputs of the chip's spiking neurons are manipulated such that the network's outputs adapt to the new learning task.

This approach has applications for autonomous, power-constrained devices that must adapt to unanticipated circumstances, including vision and navigation in unmanned aerial vehicles (UAVs) deployed into unpredictable environments; fine-grained haptic controls for robotic manipulators; dynamically adaptive prosthetic devices; and bio-cybernetic interfaces. In these real-world domains, the system must deploy with initial knowledge relevant to its target environment, then adapt to near-optimal behavior given minimal training examples, a feat beyond the capability of current learning algorithms or hardware platforms. Neuromodulatory tuning offers

a path toward implementing such abilities on physical CMOS chips. The key contributions of our work are as follows:

1. We introduce a novel transfer learning variant, called *neuromodulatory tuning*, that is able to match the performance of traditional fine-tuning approaches with orders of magnitude fewer weight updates. This lends itself naturally to easier, lower power implementation on physical chips, especially because the proposed CMOS implementation of our the fine-tuning method does not involve writing to memory hardware.

2. We provide a biologically-inspired motivation for this tuning method based on recent findings in neuroscience, and discuss additional insights gleaned from modulatory neurotransmitter behaviors in biological brains that may prove valuable for neuromorphic computing hardware.

3. We outline the mechanisms by which neuromodulatory tuning can feasibly be implemented on CMOS hardware. We present an analog spiking neuron with neuromodulatory tuning capabilities. Post-layout simulations demonstrate energy/spike rates as low as 1.08 pJ.

4.2 Background

The current study lies at the intersection of three prodigious research fields: Transfer learning (Section 4.2.1), spiking neural networks (Section 4.2.2), and neuromorphic computing (Section 4.2.3). We outline key principles of each below. Our method also draws heavily on recent discoveries in neuroscience, documented alongside the motivating principles of this research in Section 4.3.1.

4.2.1 Transfer Learning

Transfer learning allows a network trained for one task to learn a new, similar task with less computational complexity than fully retraining the network. The field includes a broad range of techniques ranging from weighting, importance sampling, and domain adaptation in unsupervised contexts [40]–[45], to fine-tuning and multi-task learning in supervised settings [46]–[52]. Recent work in few-shot, one-shot, and zero-shot learning also contributes to this line of research [53]–[56].

Our approach can be combined with many of these methods, but is most closely related to feature learning from unsupervised data [47], whereby trained parameters from a related task are used to jump-start the learning process. Our method is distinct in that the activation sensitivity of individual neurons, rather than the strengths of their synaptic connections, are modified. In some sense, this can be viewed as a degenerate form of neural programming interface [57], in that activation patterns are modulated during each forward pass of the network; however, our method adjusts firing sensitivities via supplemental bias inputs rather than by overwriting output signals directly. Our work also has tangential relations to activation function learning [58],

although we adjust firing sensitivity only, rather than changing the shape of the activation curve.

Parallel to our work, [59] presented BitFit, which shows bias tuning is an effective sparse fine-tuning method that is competitive with traditional fine-tuning on Transformer-based Masked Language Models. Our work augments and expands upon the insights from this work in two key ways: We apply a bias tuning methodology much like [59] to a convolutional neural network in the domain of computer vision, where we discover that it is not able to match the performance of a traditional fine-tuning method, and we present a novel approach to bias tuning (neuromodulatory tuning) based on multiplicative rather than summative layer modifications, and demonstrate that this method is able to match traditional fine-tuning approaches.

4.2.2 Spiking Neural Networks

Spiking neural networks (SNNs) [35], [37], [38], [60]–[62] are artificial neural networks that attempt to mimic temporal and synaptic behaviors of biological brains. Rather than using continuous activation functions, spiking neurons utilize a series of binary pulses, called a spike train [63], to propagate information forward in a brain-like manner. SNNs are particularly well-suited to implementation on analog/mixed-signal hardware, which naturally supports the high parallel sparse activation pathways common in such networks [64].

Despite these potential advantages and their strong parallels with biological brain behavior, SNNs have not gained as much recent prominence as traditional (digital) feed-forward networks, in part because of the difficulty of propagating gradient information backwards through a spike train [65]. One means to compensate for this is by training a traditional (non-spiking) network using back-propagation and then applying a transfer function to convert the learned weights into their SNN equivalents [66]. We leverage this idea in our work, but instead of applying a transfer function, we copy the non-spiking weights directly, then use neuromodulatory tuning to adapt them to a new learning task.

Recent works detailing the conversion of traditional feed-forward networks to SNNs use algorithms which modify weights, biases and activation thresholds of the network to create a SNN from a feed-forward network [67], [68]. The difference between our work and others is that we do not train the network to match the behavior with existing feed-forward network. Instead, we seek to train network for different tasks. Therefore, we do not perform layer-wise comparison which is resource consuming. Moreover, our work tunes a single parameter per neuron which is far more implementable on physical chips compared to other more computationally expensive methods.

4.2.3 Neuromorphic hardware

Neuromorphic hardware uses dedicated processing units to implement neuronal connections and firing behavior directly on a physical chip, rather than simulating them mathematically. Analog neuromorphic hardware has been shown to be more power efficient than traditional digital computation

hardware, and doesn't suffer from the same bottleneck as Von Neuman computing [69]–[76]. Some designs take advantage of sub-threshold operation for ultra-low power neurons [77], [78]. Further power reductions have been achieved through sparse temporal coding [64].

The temporal nature of spiking neural networks naturally lends itself to on-chip, biologically plausible learning methods. Spike-time-dependent plasticity (STDP) uses analog hardware to directly implement learning rules on chip. Several works have shown impressive learning accuracies using this method [63], [69], [79]–[81]. However, direct hardware implementations for learning rules consume large amounts of space and power, limiting its potential learning capacity. Our work bridges this gap by offering the possibility of on-chip learning with similar performance but reduced space and component requirements.

4.3 Neuromodulatory Tuning

Neuromodulatory tuning is a novel fine-tuning method based on recent discoveries in neuroscience. Neuronal transmission in biological brains is highly complex in timing and can occur either via rapidly terminating signals that influence only immediately connected cells (synaptic transmission), or via chemical signals that spread further away to simultaneously influence larger groups of neurons (volumetric transmission) [82], [83]. Our work is motivated by and takes inspiration from this non-synaptic transmission method. Specifically, we observe that, rather than adjusting connection strengths between neurons directly, modulatory neurotransmitters impact system behavior by affecting the activation threshold of each neuron. Thus, a single trainable parameter, implemented in our case as a supplementary input, can be used in lieu of the large suite of trainable parameters typically employed during a fine-tuning process.

4.3.1 Biological Foundations

Modulatory neurotransmitters in biological brains use metabotropic g-protein coupled receptors as opposed to strictly ion conducting receptors propagate signals, and can include neurotransmitters such as the cathecholamines dopamine and norepinephrine [84]–[88]. Interestingly, glutamate is also used by neurons as a modulatory metabotropic signal, though it is largely discussed in the context of ion channel activity [89].

Artificial neural networks principally use neuronal ion channel activity, as represented by classical synapses, to represent synaptic strength. In contrast, metabotropic neuromodulators activate g-protein coupled receptors in neurons, whose downstream effectors can be stimulatory or inhibitory (depending on predefined cellular components) and work through a series of effectors that can amplify signals from traditional synaptic inputs, resulting in multiplicative tuning of the neuron's inputs. This is considered a tuning process since these neurotransmitters often do not directly change the membrane potential, but instead change the activation threshold by modulating the channels receiving inputs. Our neuromodulatory tuning method simulates this increase or decrease in sensitivity by including ad-

ditional inputs to the incoming signal, as shown in Section 4.3.2. In other words, neuromodulatory tuning increases a model's sensitivity to specific pre-learned features, rather than changing the functions represented by those features. To our knowledge, this is the first application of volumetric, as opposed to strictly synaptic, mesolimbic attention modalities within an analog CMOS system.

4.3.2 Neuromodulatory Tuning on Analog Hardware

One particularly advantageous aspect of neuromodulatory tuning (NT) is its suitability for implementation on analog neuromorphic hardware. The behavior of fine-tuned bias connections, implemented in digital simulations as additional bias neurons, can also be implemented in analog hardware as a current source with a variable supply voltage. This approach has the following advantages:

- Minimal additional chip area required
- Lower power consumption than digital hardware
- No need to re-load weights to the on-chip memory

To probe this possibility, we use Cadence Virtuoso to explore the feasibility of a NT approach on simulated analog hardware. Our hardware is designed and simulated at the transistor level in TSMC 28-nm CMOS. The analog neuron implements the leaky integrate-and-fire model [90]. Six binary-scaled current sources make up the synapse. A current is driven onto a 50-fF capacitor to produce an integrated membrane voltage that is quantized by a dynamically clocked latched comparator. An adjustable delay line generates a 100-ns spike when the membrane voltage reaches the activation threshold and resets the membrane voltage by connecting the capacitor to ground via a pull-down transistor. A schematic diagram of our proposed neuron is shown in Figure 4.1.

Synapse Design

Each synapse operates at a supply voltage between 0.5–1 V. A higher supply increases the current in the synapse. The neuron core operates at a constant supply of 1 V. Adjusting the supply voltage of individual synapses or groups of synapses effectively changes the weights of the synapse connections. This change in behavior is analogous to the bias neurons in the software implementation and to what is observed biologically [87], [88]. To make the synapse current dependent on the supply voltage V_{DD}, we use a current mirror with a resistive load. The drain-to-source current through an N-type MOSFET is given by eq. (4.1). In a current mirror, V_g is related to V_{DD} by equation (4.2). Substituting (4.1) into (4.2) and solving for I results in (4.3).

$$I_{ds,nfet} = \frac{1}{2}\beta(V_{gs} - V_{th})^2 \tag{4.1}$$

$$V_g = V_{DD} - IR_s \tag{4.2}$$

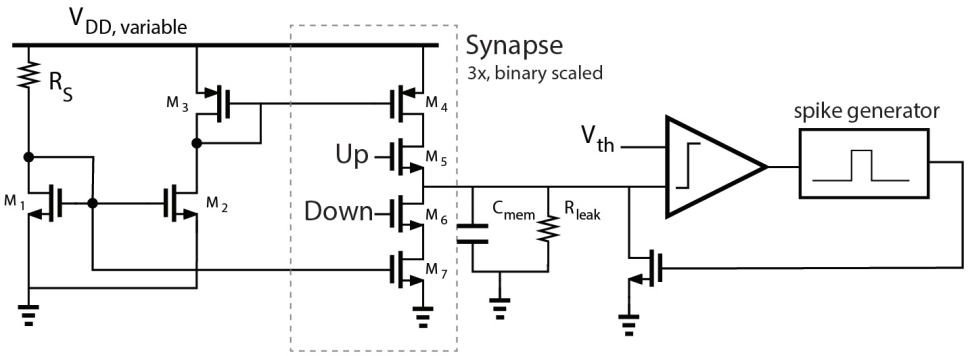

Figure 4.1: Schematic diagram of the proposed leaky integrate-and-fire neuron with NT ($V_{DD,variable}$) capabilities. The Up and Down signals are generated from the input spike and weight signals.

$$I = \frac{\sqrt{(4\beta R_s(V_{DD} - V_{th}) - 1) + 2\beta R_s(V_{DD} - V_{th}) + 1}}{2\beta R_s^2} \tag{4.3}$$

Eq. (4.3) shows that the synapse current I is a function of the supply voltage V_{DD}, which we tune to adjust the weights. Figure 4.2 shows the neuron behavior when we vary V_{DD} from 550 mV to 750 mV. The higher supply results in a larger current, producing more spikes.

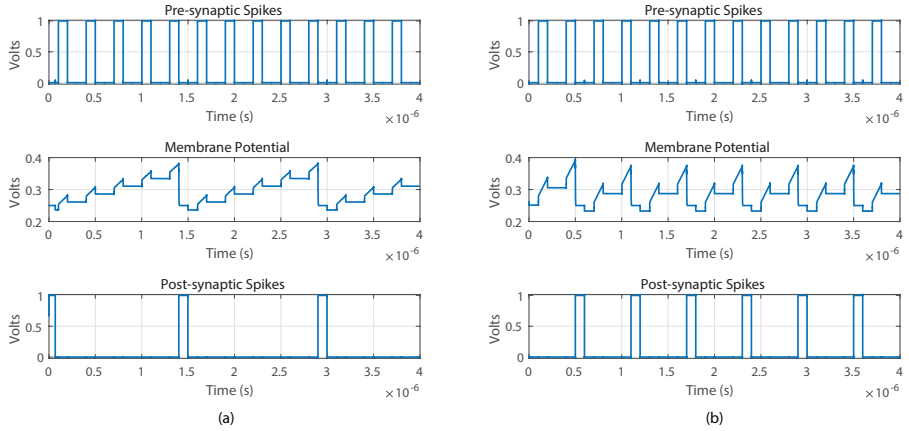

Figure 4.2: Neuron outputs with the same input spike pattern and synaptic weights, but with varied bias weights implemented as (a) V_{DD} = 550 mV and (b) V_{DD} = 750 mV.

The effect of a bias neuron with a weight of W_b on a synapse with weights W_s can be approximated as $I(W_b + W_s)$. The behavior of the analog implementation can be written as kIW where k represents the change in the synapse current due to adjusting V_{DD}. If $IW_b = kW$ then the behavior of the two implementations is identical.

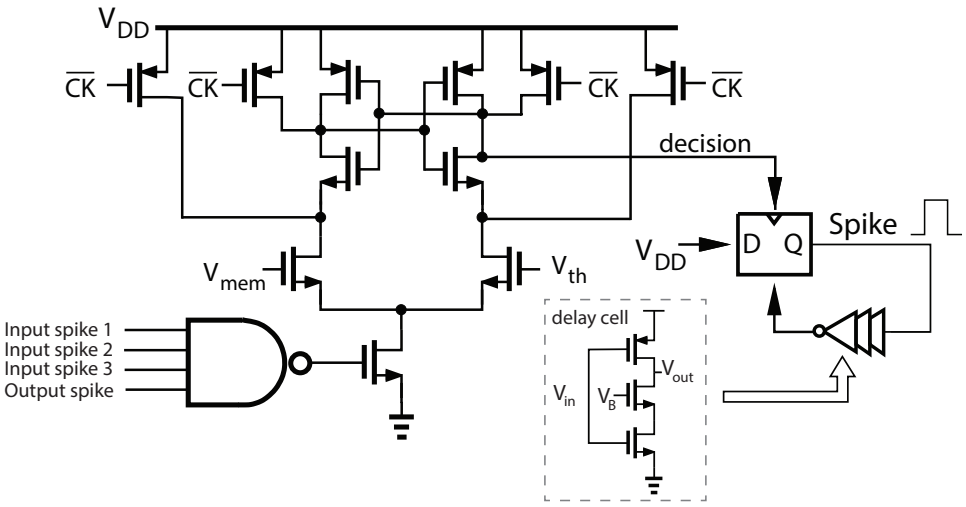

Figure 4.3: Schematic of the threshold comparator with dynamic clocking, and tunable spike generator circuit.

Neuron Core Design

A schematic of the neuron core is shown in Figure 4.3. The threshold comparator is implemented with the StrongARM topology. We choose a clocked topology to reduce static power, especially when compared to inverter based threshold detectors. Instead of a fixed-period clock, we only clock the comparator after an input spike or after an output spike. We use a 4-input NOR gate to generate the comparator clock. This ensures that power consumption is minimized in a network trained for minimal spiking activity. The membrane capacitance is always reset to V_{rest} = 250 mV and the comparator has a fixed threshold of V_{th} = 350 mV. We choose V_{rest} to give V_{mem} at least 100 mV of swing without driving the synapse current sources into the triode region, even when the synapse power supply is 0.5 V. Once the membrane potential crosses the preset threshold, the spike generation circuit is triggered. The spike is generated using a self-reset DQ fip-flop with current-starved inverter-based delay cells between Q and reset. The delay cells utilize parasitic capacitance to increase delay so as to decrease the number of stages needed for a certain spike width.

The membrane capacitor is a custom 50-fF finger capacitor which occupies only 27 μm^2. Because the membrane capacitance is only 50 fF, the neuron needs an extremely large resistor for a sufficiently low leakage current. Instead of using a polysilicon resistor which would occupy large area, we implement a CMOS pseudo resistor using a PMOS transistor which occupies only 0.7 μm x 0.5 μm and achieves approximately 400 MΩ (Figure 4.4). The pseudo-resistor is implemented as two PMOS transistors connected in a transdiode configuration. The simplest of pseudo-resistors have an asymmetric resistance-voltage characteristic, making them unusable for this neuron because the membrane potential can go both above and below V_{rest}, and must have the same up and down leakage current. To solve this, we use

two psuedo-resistors in parallel with opposite connections polarities. This halves the effective resistance, but creates a symmetric resistance-voltage characteristic.

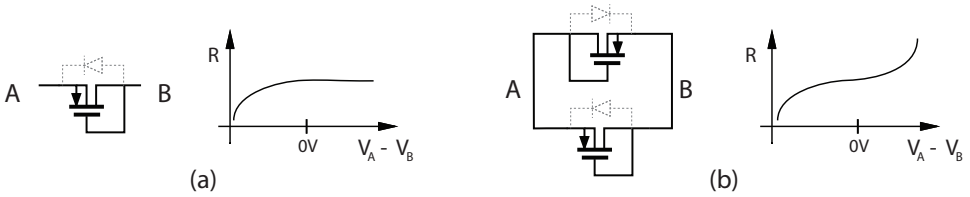

Figure 4.4: Schematic of (a) a one-directional pseudo-resistor and its asymmetric resistance characteristic and (b) the proposed pseudo-resistor showing symmetric resistance characteristics.

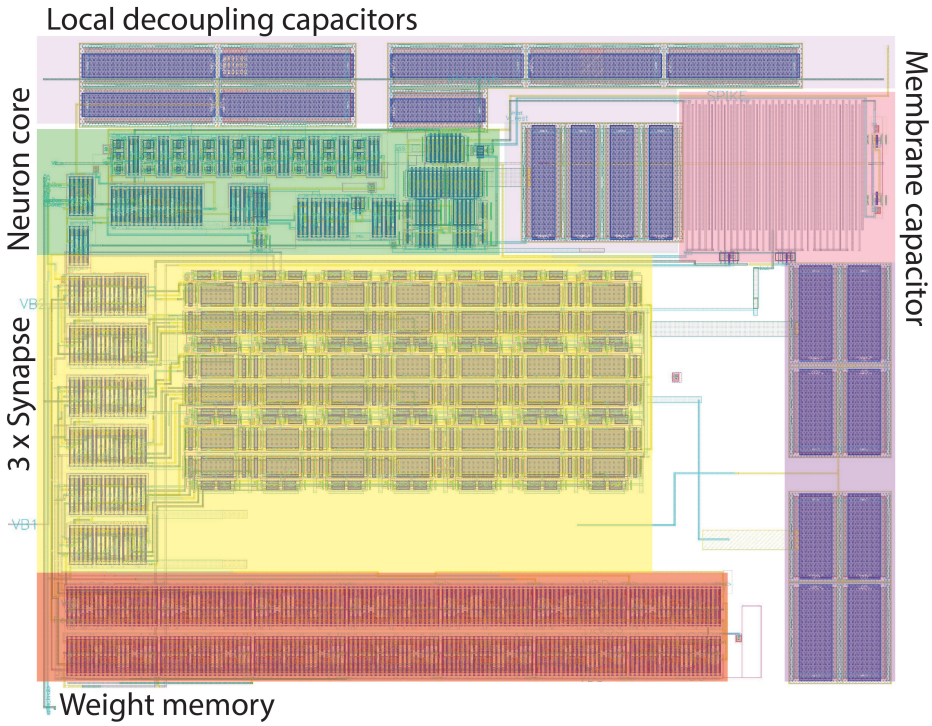

Figure 4.5: Neuron layout and annotations showing the regions of the neuron.

4.4 Results

Our long-term objective is to enable low-power analog learning behaviors *in situ* on physical analog chips. This requires both a viable mechanism for potential *in situ* learning that does not require large amounts of surface area for gradient calculations and a validated circuit design that can realistically

implement that mechanism. We present neuromodulatory tuning as a possible mechanism for this objective, and here provide results showing its performance in simulated (digital) spiking neural networks (Section 4.4.1) and a full chip design for its eventual implementation on physical CMOS hardware (Section 4.4.2).

4.4.1 Neuromodulatory Tuning on Spiking Neural Networks

To validate the performance of neuromodulatory tuning in spiking neural networks, we apply neuromodulatory tuning (NT) and traditional fine-tuning (TFT) to the SNN-VGG classification layers using the STL-10, Food-11, and BCCD datasets for comparison. We fix the batch size at 64 for all training, since our experiment with batch sizes (shown in Table 4.1) reveals that batch size does not impact the model performance dramatically. Both the Food-11 and BCCD datasets are singularly distinct from the ImageNet data [91] which was used to train VGG-19. VGG-19 therefore lacks output classes corresponding to labels from the Food-11 and BCCD datasets. To create the necessary output layer size, we added one extra fully connected layer at the end of each model. This extra layer functions as the output layer for corresponding classes in Food-11 and BCCD. Different from Food-11 and BCCD, STL-10 is a subset of ImageNet. Since VGG-19 is trained on ImageNet, VGG-19 contains classes that are contained within in STL-10 labels. Therefore, we do not add extra layers for the SNN STL-10 experiments. All SNN models were trained on an AMD Ryzen Threadripper 1920X 12-Core Processor. Results are shown in Table 4.2 and 4.3.

As expected, performance is poor when no tuning is applied. This is partially because SNN architectures, comprised of leaky integrate-and-fire neurons, differ drastically from traditional deep networks in both signal accumulation and signal propagation, resulting in almost 0% accuracy on all three transfer tasks. Tuning improves this accuracy, achieving up to 88% accuracy with TFT and 50% with NT on some tasks with certain learning rates. According to our results shown in Table 4.2, NT underperforms on the STL-10 dataset comparing to TFT, has equal performance to TFT on BCCD, and outperforms TFT on Food-11, which suggests that neuromodulatory tuning can positively impact learning behaviors on brain-like architectures.

Our performance comparison of the algorithms is influenced by differences between the three datasets. STL-10 is the subset of the dataset used to train VGG-19, so tasks in STL-10 is more native to the network. In contrast, Food11 and BCCD are foreign to the VGG-19 network, so those tasks will require VGG-19 to make adjustments in larger magnitudes or completely re-learn the task. Given that neuromodulatory tuning outperforms TFT on Food11, a foreign dataset, and that TFT requires changes of larger magnitudes, NT is superior for these cases. There are accuracies below random guessing, this might be caused by the low learning rate for NT and the absence of feed-forward to spiking network conversion algorithm for TFT.

Comparing two different types of NT, NT_1 performs better than NT_2 on STL-10 dataset, and has equal performance with NT_2 on Food-11 and BCCD dataset.

According to Table 4.3, TFT requires over 120 million parameters adjustment to achieve such performance, so the adjustments are impossible to implement on the physical chips. In contrast, NT method only requires 9000-20000 adjustments, which is implementable on physical chips.

Table 4.1: Validation accuracy on the Food-11 dataset on SNN after 10 epochs, mean of 10 training runs using bath sizes (bs) = {16, 32, 64, 128}.

	acc (bs=16)	acc (bs=32)	acc (bs=64)	acc (bs=128)
NT_1 (lr = 0.1)	0.4568	0.4605	0.4570	0.4647
TFT (lr = 0.1)	0.1304	0.1243	0.1145	0.0770

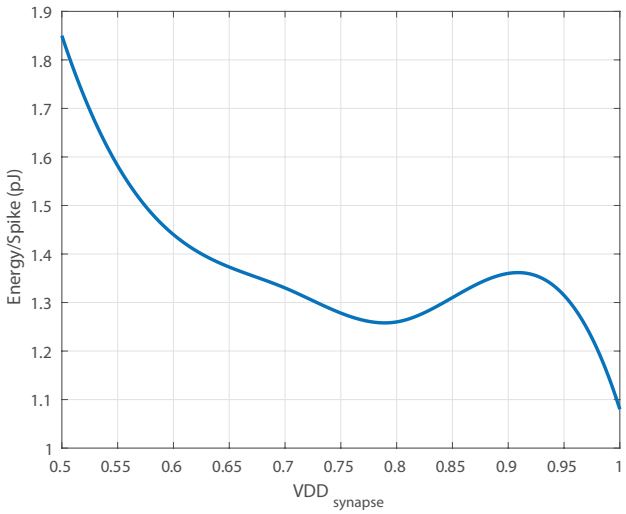

Figure 4.6: The energy/spike decreases as V_{DD} increases. This is because a higher V_{DD} yields a higher synapse current and therefore more output spikes for the same number of input spikes.

4.4.2 Analog Neuromorphic Hardware Simulation

The goal of this work is to develop a low-power CMOS chip architecture that implements neuromodulatory tuning. In addition to presenting the neuromodulatory tuning algorithm and exploring its performance, we also present a complete neuron design to implements this algorithm on analog CMOS hardware.

Figure 4.5 shows the layout of the proposed neuron implementing NT fine tuning. The entire neuron, synapse and weight storage occupies only $598um^2$, with the neuron core (including membrane capacitor) occupying only $132nm^2$. We have validated the simulation results from Section 4.4.1 using post-layout simulations in Cadence Virtuoso to model an XOR task using spiking neurons. Two neurons were chosen to be the inputs to the XOR "gate" and another designated as the output. A train of 10 spikes to an input neuron constituted a "1". No input spikes constituted a "0". The

Table 4.2: Validation accuracy on STL-10, Food-11, and the BCCD dataset in a spiking neural network (SNN) architecture. Models were trained for 50 epochs for STL-10, Food11, and the BCCD dataset, respectively. Average of five training runs. Best per-task performance of neuromodulatory tuning (NT_2) and traditional fine-tuning (TFT), respectively, is underlined. NT_2 refers to the modify existing bias implementation of NT and NT_1 refers to the additional bias implementation described in Section 4.3.

		lr 0.0001	lr 0.001	lr 0.01	lr 0.1
STL-10	no tuning	0.0007	0.0007	0.0007	0.0007
	TFT	0.8888	0.8014	0.2582	0.1274
	NT_2	0.0000	0.0000	0.3052	0.3062
	NT_1	0.0000	0.0009	0.5428	0.5731
	additive bias	0.0010	0.0008	0.0006	0.0025
Food-11	no tuning	0.0341	0.0341	0.0341	0.0341
	TFT	0.0147	0.0729	0.1017	0.1168
	NT_2	0.0063	0.3645	0.4537	0.4615
	NT_1	0.0020	0.3678	0.4564	0.4665
	additive bias	0.0840	0.1864	0.1404	0.1414
BCCD	no tuning	0.0005	0.0005	0.0005	0.0005
	TFT	0.1003	0.2508	0.2507	0.2508
	NT_2	0.2501	0.2509	0.1371	0.0680
	NT_1	0.2508	0.2509	0.2041	0.0591
	additive bias	0.1848	0.2137	0.2144	0.2505

Table 4.3: Validation accuracy and parameter on STL-10, Food-11, and the BCCD dataset in a spiking neural network (SNN) architecture. Models were trained for 50 epochs for STL-10, Food11, and the BCCD dataset, respectively. Accuracy from the learning rate with best average accuracy of five training runs. NT_2 refers to the modify existing bias implementation of NT and NT_1 refers to the additional bias implementation described in Section 4.3.

		best accuracy	Parameter Amount
STL-10	TFT	0.8888	123,642,856
	NT_2	0.3062	9,192
	NT_1	0.5731	9,192
	additive bias	0.0025	9,192
Food-11	TFT	0.0356	123,653,867
	NT_2	0.4615	20,203
	NT_1	0.4665	20,203
	additive bias	0.1864	20,203
BCCD	TFT	0.2508	123,646,860
	NT_2	0.2509	13,196
	NT_1	0.2509	13,196
	additive bias	0.2505	13,196

Table 4.4: Comparison of our proposed neuron implementing neuromodulatory tuning with the state of the art in standalone neurons. *Total area includes neuron core, synapse, and weight storage.

	This Work	Joubert et al., 2012	Cruz-Albrecht et al., 2012	Rangan et al., 2010	Jayawan 2008
Process (nm)	28	65	90	90	350
Area μm^2	598 (Total *) 132 (Core)	538	442	897	2800
Max f_{spike} (Hz)	3.3M	1.9M	100	7k	1M
Energy/spike (pJ)	1.08	41	0.4	1	9

spikes propagated through the network according to the trained weights. The output was "0" if less than three spikes were observed at the output, otherwise the output was a "1". The analog simulation showed 2 spikes at the output for a 0, and 4 for a 1.

The proposed neuron achieves performance competitive with the state-of-the-art in standalone neuron circuits (see Table 4.4). The total power for the neuron core varies with spike rate. Figure 4.6 shows the energy/spike vs spike rate, and figure 4.7 shows the distribution of power for two spike rates. The best case energy consumption is 1.08pJ/spike. The energy used to charge C_{mem} from V_{rest} to V_{th} can be calculated as $E_{charge} = \frac{1}{2}C_{mem}(V_{th} - V_{rest})^2$. For the values in this design, $E_{charge} = 0.25$fJ, which is completely negligible compared to the total energy consumption.

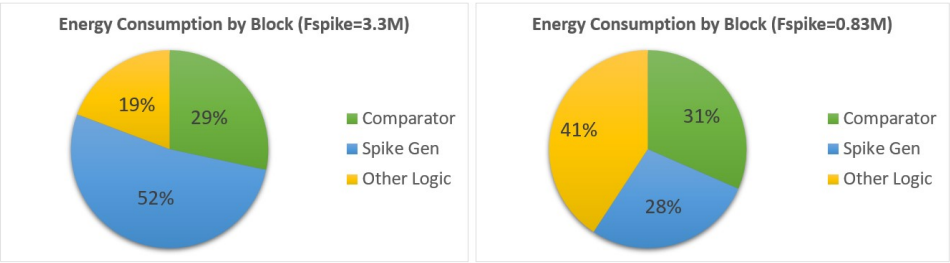

Figure 4.7: The distribution of power within the neuron core.

4.5 Conclusions

Low-power analog machine learning has the potential to revolutionize multiple disciplines, but only if novel and physically-implementable learning algorithms are developed that enable *in situ* behavior modification on physical analog hardware. This chapter presents a novel task transfer algorithm, termed neuromodulatory tuning, for machine learning based on biologically-inspired principles. On image recognition tasks, neuromodulatory tuning performs on test cases as well as traditional fine-tuning methods while requiring four orders of magnitude fewer active training parameters (although the total number of weights is comparable between methods). We verify this result using both deep forward networks and spiking neural network architectures. We also present a circuit design for a neuron that immplements

neuromodulatory tuning, a potential layout for the use of such neurons on an analog chip, and a post-layout verification of its capabilities.

Neuromodulatory tuning has the advantage of being well-suited for implementation on neuromorphic hardware, enabling circuit implementations that support life-long learning for applications that require energy-efficient adaptation to constantly changing conditions, such as robotics, unmanned air vehicle guidance, and prosthetic limb controllers. Future research in this area should focus on probing the performance of NT in domains beyond image recognition; exploring the possibility of paired bias links in which multiple neurons connect to a single power domain region; and designing improved SNN update algorithms with stronger convergence properties.

5 A Phase-Domain Spiking Neuron with Switched Capacitor Synapse.

This chapter is composed from a paper entitled "A Phase-Domain Spiking Neuron with Switched Capacitor Synapses" which will be submitted to the journal "IEEE Transactions on Circuits and System II: Express Briefs." I hereby confirm that the use of this article is compliant with all publishing agreements. The authors on this work are myself as lead author, Shea Smith, Yu Hao, Ryan Watson, Nancy Fulda, Jordan Yorgason, Karl Warnick, Yen-Chen Kuan and Shiuh-hua Wood Chiang. With support and input from the other authors, I architected and designed the time-domain neuron presented in this chapter.

5.1 Introduction

Analog spiking neurons have emerged as a competitive alternatives to power-hungry digital neurons. The development of spiking neural networks (SNN) have further motivated work in analog neurons, which are well suited to handle the time-domain components of SNN signaling [92]–[94].

A widely used, biologically inspired analog neuron model is called the leaky integrate-and-fire (LIF) neuron. The two dominant CMOS implementations of an LIF neuron are the op-amp based voltage-mode neuron, and the current-mode neuron.

A current-mode neuron is shown in Fig. 5.1(a) [92], [94]–[100]. Input spikes activate a current source I_{syn} and charge is integrated on C_{mem} until V_{mem} reaches the comparator threshold V_{th} at which point the neuron generates an output spike and resets C_{mem} to its resting potential. A resistor R_L or small conductor is placed in parallel with the capacitor to slowly leak charge. The comparator is either implemented as an inverter, which burns high short-circuit current as C_{mem} approaches V_{mem}, or a dynamic comparator requiring an external clock. A current-mode neuron's area is dominated by capacitors.[94] reports that 64% of the neuron area is consumed by C_{mem}.

Voltage-mode neurons use an op-amp as shown in Fig. 5.1(b) [93], [101]. A resistor R_L is placed in parallel with the feedback capacitor C_{mem} to achieve a leaky integrator. When the integrator output reaches a comparator threshold V_{th}, an output spike is generated and the integrator is reset. Op-amps consume large area and high power, high gain is difficult to achieve

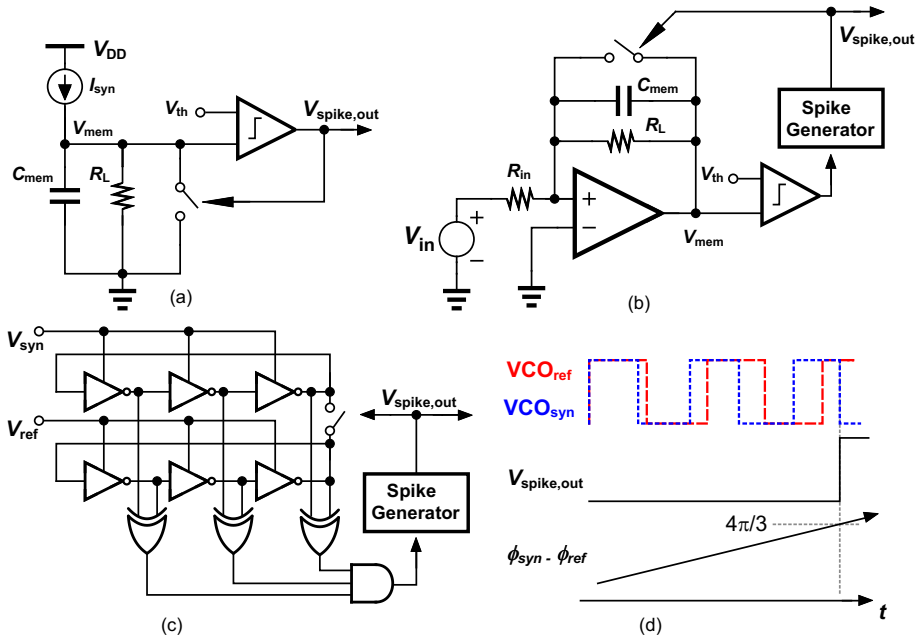

Figure 5.1: Block diagram of three implementations of an LIF neuron. (a) Current mode, (b) voltage mode, (c) phase domain. (d) Integration of phase between two VCOs.

under low supply. large area capacitors and resistors further limit this topology's scalability.

Some works have proposed to use a VCO in a time-domain LIF neuron. [102] proposes to use a VCO integrator in an analog low pass filter in place of C_{mem}. This design reduces area by removing a capacitor, but also uses many resistors which also occupy large area. Further, it uses time-domain circuity to do voltage domain filtering, which requires voltage-to-time conversion and time-to-voltage conversion. [103] proposes a time-domain neuron design using a current-controlled oscillator (ICO) and time-domain comparator. Their proposed circuits are not well described and limited verification is provided. Moreover, the comparator architecture is complex and consumes unecessary power.

We present the design and analysis of a VCO-based time-domain neuron in 28nm process which overcomes many of the challenges posed by existing neuron designs. Our proposed neuron design fully utilizes time domain computing and implements a simple, power-efficient phase-domain comparator. We propose a VCO-based time-domain spiking neuron with an XOR-based phase domain comparator with a fixed $4\pi/3$ radians phase threshold. We further propose a novel 5-bit switched-capacitor based synapse which utilizes the fast switching speed of small transistors and the unused area on the higher metal layers above the neuron and which bypasses the challenges associated with designing sub-nanoamp current sources. We further propose a phase-domain leak circuit inspired by the phase-locked-loop which replaces the leaky conductor in LIF neuron model. The neuron, synapse, and weight memory occupy a combined area of only 21x27μm. The

neuron achieves a maximum spiking frequency of 5.8MHz consuming only 134fJ/spike under a 0.35V supply.

5.2 Theory and Analysis of a Phase-Domain LIF Neuron

Time- and phase-domain circuits have emerged as potential replacements for analog circuits that face design difficulties in scaled technologies. Time based circuits already find application in analog-to-digital converters (ADC) and amplifiers [26]. In this section we analyze the behavior of the phase-domain circuit.

A block diagram of the phase-domain neuron is shown in Fig. 5.2. It works by comparing the phase between two VCOs. Bias voltages $V_{B,VCO}$ and V_{syn} controls the center frequency of the two VCO's VCO_{syn} and VCO_{ref}. Except during an input spike event, $V_{B,VCO} = V_{syn}$. During a synaptic input, f_{syn} temporarily changes meaning f_{syn} - $f_{ref} \neq 0$. A phase difference between VCO_{syn} and VCO_{ref} $\Delta\phi$ begins to accumulate. After the input spike is over, f_{syn} returns to the same value as f_{ref} and phase stops accumulating. In this way, we integrate phase the same way that current-mode neurons integrate charge on a capacitor. A phase comparator monitors the two VCOs. When The VCO's are $4\pi/3$ radians out of phase, the comparator triggers a spike generator. To 'leak' phase from our integrator, we use a phase-frequency detector (PFD) to generate leak pulses. A PFD provides a digital pulse whose outputs provide information about whether VCO_{syn} has accumulated positive or negative phase. Further, a PFD's output pulse width is proportional to the phase difference $\Delta\phi$ between its two inputs. Similar to a charge-pump phase-locked-loop, we use the PDF output to drive the phase of the VCO_{syn} towards VCO_{ref}. For this prototype chip, synaptic weights are stored in a shift-register.

This behavior closely matches the traditional LIF neuron, which models a biological neuron as a capacitor in parallel with a leaky conductor. The behavior of a LIF neuron is described by eq. 5.1.

$$C_{mem}\frac{dV}{dt} = I_{syn} - \frac{V_{mem}}{R_L} \tag{5.1}$$

Rearranging terms and combining I_{syn} and V_{mem}/R_L into I_{total} yields eq. 5.2

$$V_{mem} = \frac{1}{C_{mem}}\int_0^t I_{total}(t)dt \tag{5.2}$$

From signal processing theory we know that the integral of frequency is phase (eq. 5.3).

$$\phi(t) = \int_0^t f(t)dt \tag{5.3}$$

We expand this analysis to two VCOs: a VCO with constant frequency and a VCO with time-varying frequency. We call the VCO with a constant frequency f_{ref}, VCO_{ref} and the VCO with frequency $f_{syn}(t)$ that changes with synaptic input VCO_{syn}. Phase accumulates as the integral of the difference in frequency between VCO_{ref} and VCO_{syn} as in eq. 5.4.

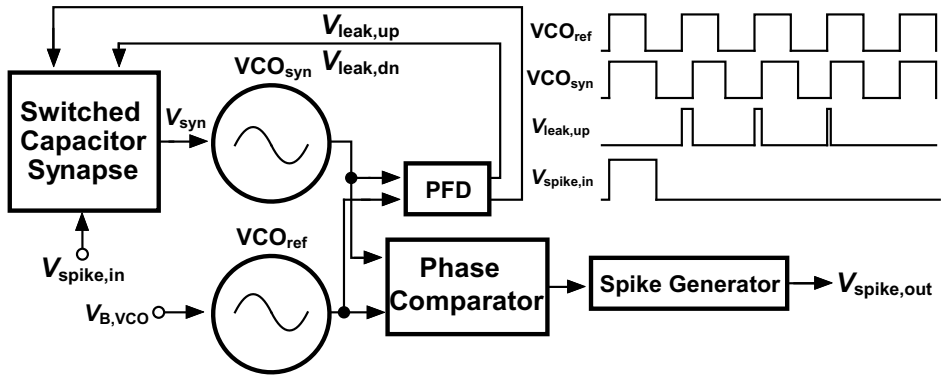

Figure 5.2: (a) Simplified block diagram of the proposed phase-domain LIF neuron. (b) Time-domain waveforms showing the operation of the proposed neuron

$$\Delta\phi(t) = \int_0^t f_{ref} - f_s(t)dt \tag{5.4}$$

$f_s(t)$ is determined by the product of the VCO gain k_{vco} and its bias voltage V_{syn}. V_{syn} is generated by the synapse and is changed by input spikes and PFD pulses. Let $V_{syn}(t)$ be the time varying bias voltage for VCO_{syn}. We rewrite eq. 5.4

$$\Delta\phi(t) = \int_0^t k_{vco}V_{B,VCO} - k_{vco}V_{syn}(t)dt \tag{5.5}$$

Again rearranging terms, we see that the eq. 5.6 describing a phase-domain neuron parallels eq. 5.2, where $1/C$ and k_{vco} dictate how much affect I_{syn} and V_{syn} have on V_{mem} and $\Delta\phi$ respectively.

$$\Delta\phi(t) = k_{vco}\int_0^t V_{B,VCO} - V_{syn}(t)dt \tag{5.6}$$

5.3 Design

5.3.1 VCO-Based Neuron Design

Both VCOs in the proposed neuron are implemented as 3-stage ring oscillators as shown in Fig. 5.3. The delay element is a current starved inverter with an analog control bias V_b to set the frequency. When V_{RST} is asserted, the VCO is reset to a known phase. When V_{init} is asserted, the VCO enters sleep mode and does not oscillate. This is used to conserve power when the neuron is not in use. The VCO's center frequency was chosen to be 15MHz, as a tradeoff between power and inference speed [103], and was also limited by the subthreshold speed of the phase comparator logic. The two VCO layouts are symmetrical and surrounded by shielding traces to equalize any parasitic loading and minimize any parasitic coupling to or from the VCO. The VCO outputs are routed with equal length traces to buffers before being routed to the phase comparator.

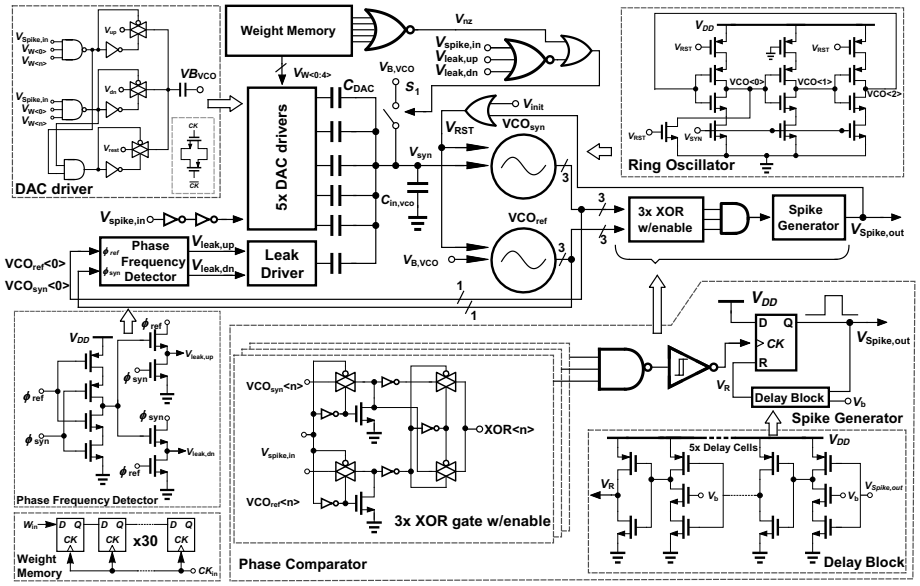

Figure 5.3: Full block diagram of the fabricated time-domain neuron.

Many voltage domain neurons use inverter based comparators. While simple and small, inverter-based comparators have thresholds that vary drastically over process, voltage and temperature, and are consume high short-circuit current as the membrane potential approaches the threshold. Clocked comparators have been used to solve this problem. Clocked comparators consume no static power but consume more area. Further, they require a network-wide clock distribution circuit, or local clock generator circuit which add area, power and complexity. This phase-domain comparator consumes no static power, has very low short-circuit current, requires no clock, and occupies small area.

The phase comparator is implemented as an array of three XOR gates, one XOR gate for each pair of VCO stages. This topology is also used in state-of-the-art time-to-digital converters [104]. A schematic of one of the XOR gates and the control logic is shown in Fig. 5.3. We only enable the comparator during an input spike, which reduces power consumption. All three XOR gates will assert high only when the VCOs are $4\pi/3$ rad out of phase. A NAND gate senses when all three XOR gates output high. A schmidt trigger buffer filters out any short pulses when the phase difference is approaching $4\pi/3$ rad.

The proposed spiking neuron has a simple and highly tuneable spike generator which consists of a clocked flip-flop with the input tied to VDD as shown in Fig. 5.3. The flip-flop is clocked by the output of the phase comparator. On a rising comparator edge, the flip flop output is pulled to VDD. A programmable delay line connects the flip-flop output and reset. The spike width is set by the flip-flop's internal propagation delays plus the delay of the delay line. The delay line is controlled by an analog bias V_b.

We also propose a novel time-domain leak circuit. We use a phase-frequency detector (PDF) to mimic the behavior of an RC decay in the phase

domain. The output of a PFD is a pulse in time proportional to the difference in phase of two VCOs. The PFD outputs connect to an auxiliary path in the synapse. A pulse from the positive output of the PFD causes a positive phase shift in VCO_{syn}, and a negative pulse causes a negative shift. The change in phase is proportional to length of the PFD pulse. This system will therefore 'leak' phase faster when the phase difference is larger.

This behavior is comparable to the RC-discharge-based leak in a voltage domain neuron which is modeled as $V = V_o e^{-t/RC}$. The leaked phase at time t $\phi_L(t)$ can be written as:

$$\phi_L(t) = -\beta\phi(t)K_{pfd}(t) \tag{5.7}$$

where β is the *synapse gain*, and $K_{pdf}(t)$ is the PFD gain. Because $K_{pdf}(t)$ is linearly proportional to $\phi(t)$, we can rewrite $K_{pdf}(t)$ as $\alpha\phi(t)$. The behavior of the leak circuit can now be written as

$$\phi_L(t) = -\alpha\beta\phi(t)^2 \tag{5.8}$$

A second order polynomial has a similar profile as a exponential function and effectively mimics the behavior of RC decay.

Fig. 5.4 shows the functionality of the phase-domain neuron. When $V_{spike,in}$ is high, f_{syn} drops causing a negative phase shift between ϕ_{ref} and ϕ_{syn}. Observe that each PFD output pulse $V_{leak,up}$ causes VCO_{syn} and VCO_{ref} to move closer in phase. As this happens, the $V_{leak,up}$ pulse widths decrease.

The leak circuit's closed loop transfer function $H(s) = \phi_{out}/\phi_{in}$ is written as

$$H(s) = \frac{\frac{C_A \Delta V K_{VCO}}{2\pi(C_A+C_B)}}{s + \frac{C_A \Delta V K_{VCO}}{2\pi(C_A+C_B)}} \tag{5.9}$$

where ΔV is the difference between $V_{B,VCO}$ and $V_{up/dn}$. This is a one pole system and is therefore stable.

5.3.2 Switched Capacitor Synapse

Existing current and voltage synapse designs are not suitable for a VCO-based neuron. The synapse must change the VCO bias only for the duration of an input spike. To meet this requirement, we propose a novel switched capacitor synapse. The operating principle is similar to a capacitive digital-to-analog converter (DAC). We construct a 4-bit binary-scaled capacitive DAC using high density custom-built finger capacitors. The voltage on the top plate of the DAC is V_{syn}. The bottom plate can be switched between three signals, V_{up}, V_{down} or V_{mid} as shown in Fig. 5.5. Except during an input spike, The synapse holds V_{syn} at $V_{B,VCO}$ as in Fig. 5.5 (a). When an input spike is detected, switches $\phi 2$ goes low to effectively sample V_{syn} on C_B. $\phi 1$,up or $\phi 1$,dn goes high if the synaptic weight W is positive or negative, respectively, which connects V_{bot} to either V_{up} or V_{down} as shown in Fig. 5.5 (b) and (c). This causes V_{syn} to change based on the values of V_{up} or V_{down} and the capacitor divider $C_B/C_B + C_A$ set by the number of switched DAC fingers.

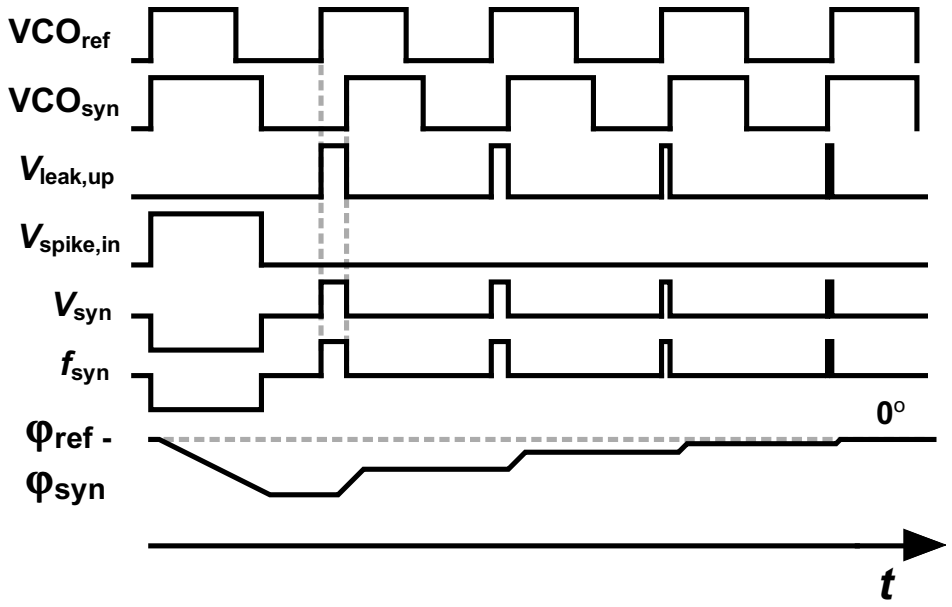

Figure 5.4: Operation of the proposed time-domain leak circuit.

A fundamental limitation of current source synapses is that they are inherently unidirectional. That means that a bi-directional current-source based synapse occupies approximately double the area of a unidirectional synapse. The proposed switch capacitor DAC is bidirectional because the DAC drivers can drive V_{bot} to a voltage higher or lower than V_{mid}. This allows us to use a 4-bit capacitor DAC for a 5-bit synaptic weight. In the DAC driver logic, We designate bit 1 $V_W < 0 >$ for weight polarity and $V_W < 1:4 >$ for weight magnitude. Logic in the DAC drivers determine how many fingers to drive to which voltages. A Nor gate detects a zero-value weight and holds V_{syn} at $V_{B,VCO}$ through S_1.

To further increase neuron density, we layout the DAC on a high metal layer on top of the DAC drivers and weight memory. We place a metal shielding layer between the DAC and any routing to equalize parasitic capacitance to neighboring metal layers. One finger has a capacitance of approximately 0.5fF. The DAC uses a layout technique from [105] which makes the DAC more robust against gradient mismatch. The layout of the proposed time-domain neuron is shown in Fig. 5.6.

5.3.3 Measurement Results

The proposed neuron and synapse were designed and fabricated in a 28nm cmos process. Fig. 5.9 shows the die photograph, with a core of 12 neurons in the center measuring 65x125μm. The chip consumes 15.61μW under a 0.35V supply, meaning each neuron consumes approximately 1.3μW. The power consumption varies based on the neuron's spiking frequency. At the maximum spiking frequency, the VCO, phase comparator, PFD, spike generator and synapse account for 20%, 16%, 17%, 5% and 29% respectively. At a spiking frequency of 1.7MHz, the VCO, phase comparator, PFD, spike

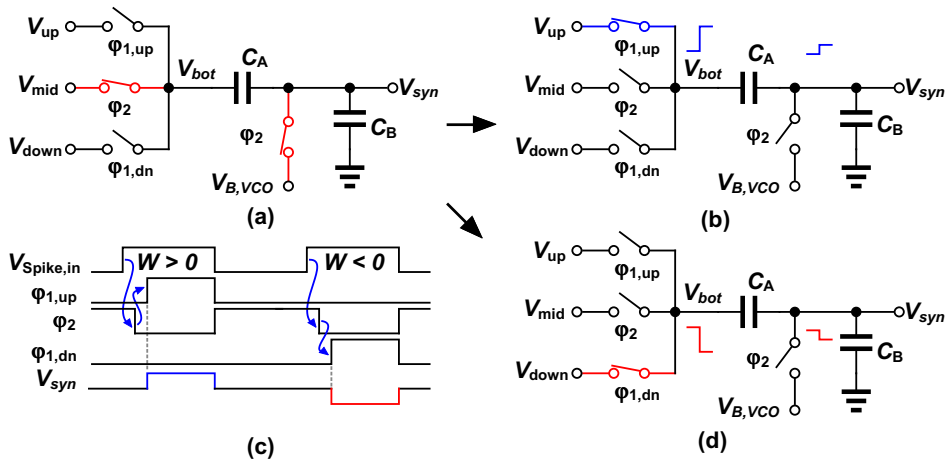

Figure 5.5: Operation of the switched capacitor synapse.

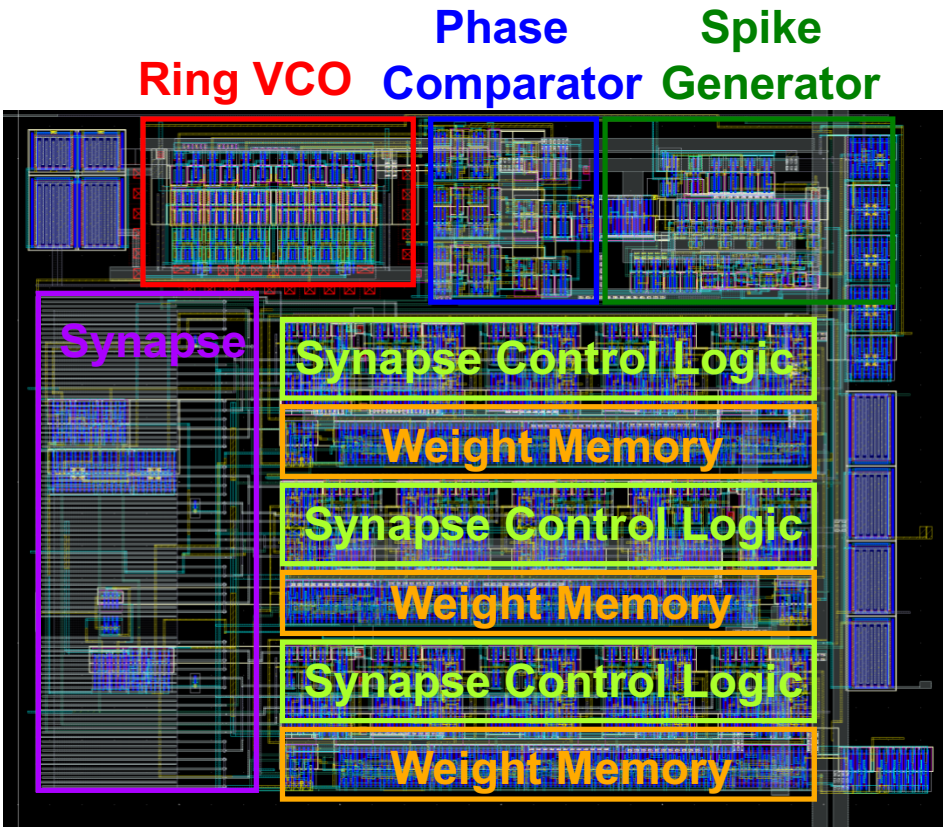

Figure 5.6: Layout of the proposed time-domain neuron.

	This Work	Sahoo ISCAS17	Moriya IJCNN22	Kuang TCASII21	Joubert IJCNN12	Rubino TCASI21
Process (nm)	28	65	65	65	65	22 (FDSOI)
Supply Voltage (V)	0.35	0.9	0.3	0.89	*	0.8
Topology	VCO	VCO	Current	Digital	Current	Current
Silicon	Y	N	N	Y	N	N
Neuron Area (um^2)	190	*	1050	*	120	900
Max Spiking Frequency (MHz)	5.8	1.5	2	24	1.9	30
Energy/Spike (fJ)	134	*	40.1	2064	2000	14000

*Not Reported

generator and synapse account for 27%, 15%, 20%, 4% and 25% respectively. With an measured average maximum spiking frequency of 5.5MHz, and 58% of power consumption from the neuron core, we achieve 134fJ/spike. We achieve 287fJ/spike at an output spike frequency of 2.7MHz and 1.07pJ/spike At an output spike frequency of 730kHz,.

Fig. 5.7 the output spike waveforms of two neurons, one each at high frequency and low frequency. Fig. 5.8 shows the measured spiking frequency as we vary several neuron parameters: input spike width, V_{up}, and synaptic weight. Fig. 5.8 (a) shows that as the width of an input spike increases, so does spiking frequency. This is expected, because the longer the spike width, the longer VCO_{syn} and VCO_{ref} are at different frequencies. In Fig. 5.8 (b) we sweep V_{up}. A larger V_{up} causes more phase change per input spike, which is what we observe in the measurements. In Fig. 5.8 (c) we sweep the input's synaptic weight. A larger weight results in a higher spiking frequency. The observed non-linearity in this data are due to the nonlinear VCO tuning curve. 5.8 (d) plots the power consumption of the neuron at a spiking frequency of 1.7MHz.

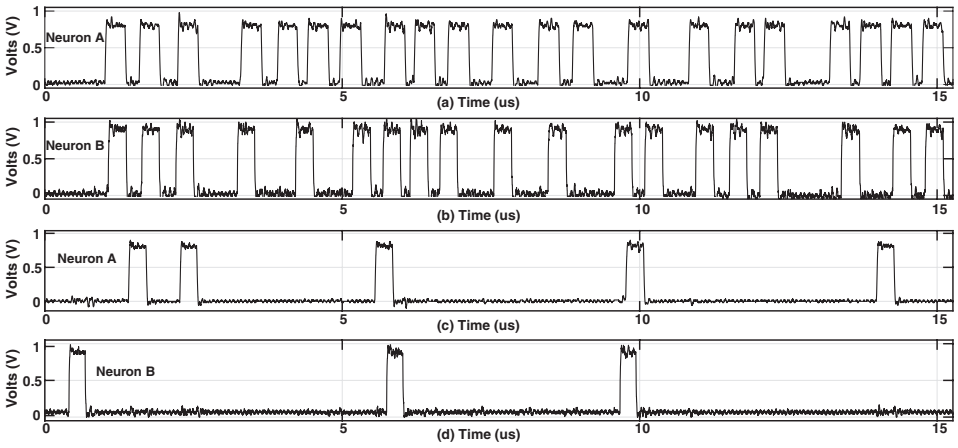

Figure 5.7: Measured waveforms showing two different spiking frequencies. (a) and (c) show neuron A's output at a high and low spiking frequency respectively. (b) and (d) shows neuron B's output at a high and low spiking frequency.

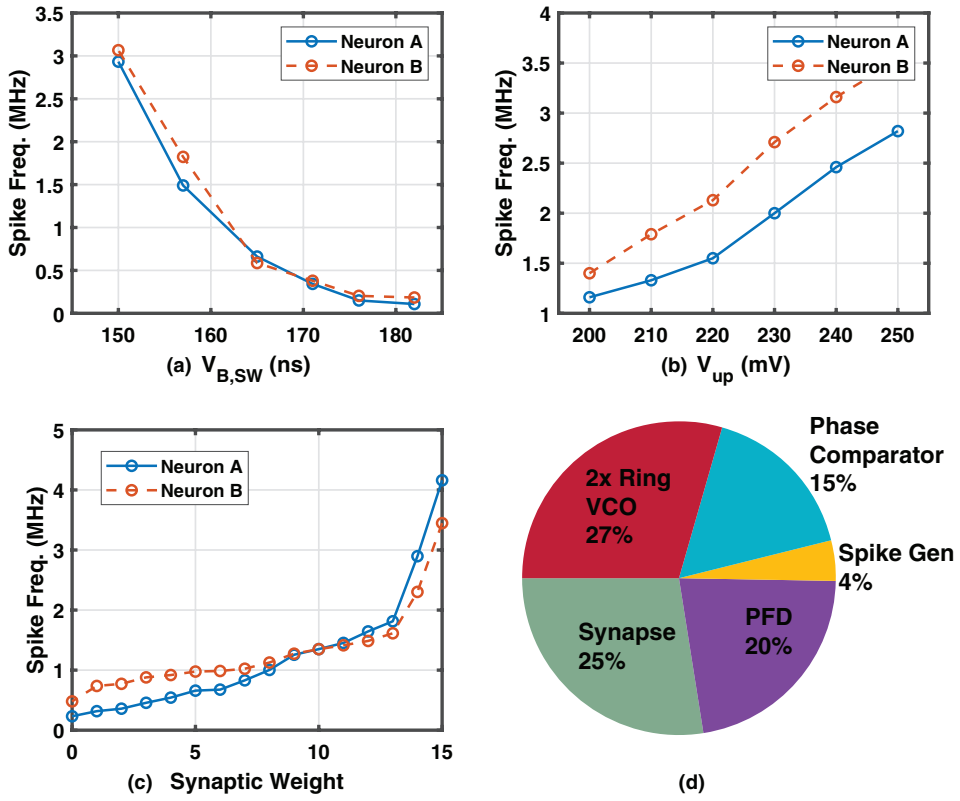

Figure 5.8: Measured spiking frequency versus (a) $V_{B,SW}$, (b) V_{up} and (c) synaptic weight. (d) Shows the power distribution at a spiking frequency of 1.7MHz.

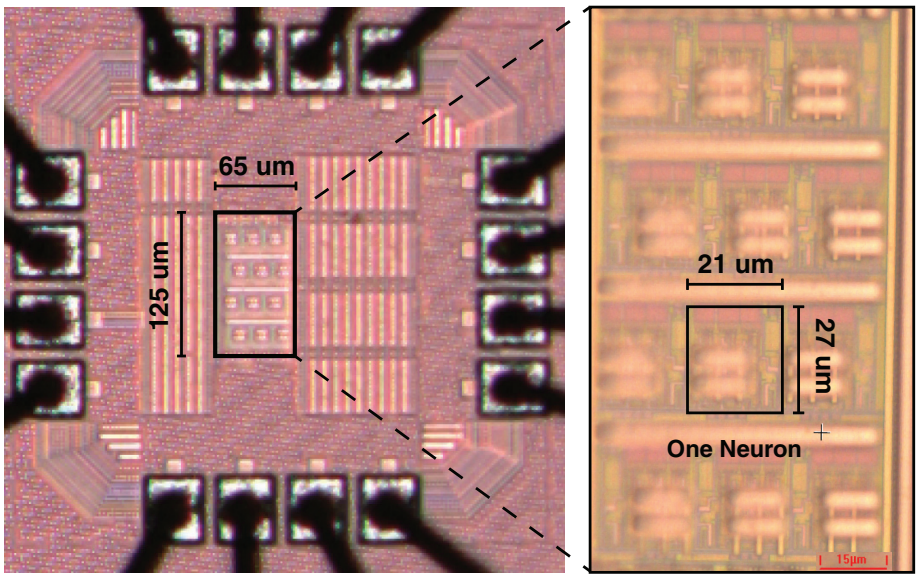

Figure 5.9: Die photo of fabricated chip.

5.3.4 Conclusion

We present a phase-domain spiking neuron circuit that achieves 134fJ/spike. The neuron is built from two voltage-controlled oscillators and an XOR-based phase comparator. A PFD is used to mimic the behavior of time-domain neuronal leak behavior. We proposed a novel bidirectional switched capacitor synapse which is more suited to a phase-domain neuron than existing synapse designs. Table 5.3.3 compares the performance of this neuron to the state-of-the-art.

This neuron design is more area efficient as compared to the voltage domain spiking neuron (VDSN) presented in Chapter 4. Note that both neurons are designed in the same 28nm CMOS process. The phase-domain neuron core occupies $110um^2$ versus $132um^2$ in the VDSN. The 5-bit synapse in the phase-domain neuron occupies $90um^2$ verses $133um^2$ for the 4-bit synapse in the VDSN. Further, because the phase-domain synapse consists mainly of a capacitor build on metal 5, it can be built on top of weight memory and synapse control logic. Phase-domain neurons have the potential to replace existing artificial neuron topolgies in scaled CMOS processes.

6 Conclusion

6.1 Thesis Contributions

The contributions of this thesis are:

- A novel optimization algorithm for zero-crossing-based amplifier design using a combination of MATLAB modeling and transistor level simulations
- The design and simulation of a zero-crossing-based amplifier to verify the aforementioned algorithm which achieves competitive performance and validates the design methodology
- The design of a high-speed two-stage background-calibration floating-inverter amplifier which occupies a significantly smaller area as compared to other works
- The design, layout, and simulation of a voltage-domain leaky integrate-and-fire neuron circuit which implements a novel fine-tuning algorithm
- The design, fabrication and measurement of a novel phase-domain artificial neuron. I show that phase-domain neurons are a scaling friendly alternative to voltage-domain neurons.
- The design of a bidirectional switched-capacitor synapse for VCO-based neurons. This synapse achieves lower area per bit as compared to current-base synapses.

6.2 Summary

This work presented the design of three circuits which break from traditional digital or analog architectures. Detailed analysis of each circuit was provided. Accompanying measurement and simulation results were also presented.

Chapter 2 discussed a ZCBA behavioral model. The model is used in a MATLAB script to optimize the charge pump design for a ZCBA. We showed that there exists a set of circuit parameters that minimize the amplifier's settling time within a set of predetermined design constraints. The design and simulation of a ZCBA based on the optimization algorithms were briefly discussed. Circuit simulation results showed that the MATLAB model and transistor-level design are in agreement.

Chapter 3 discussed in detail the design of a ZCBA in a 28nm CMOS process. The design of a novel two-stage background-calibrated floating-inverter amplifier was presented. A self-timed loop was used to relax the bandwidth requirement for the floating-inverter amplifier. The proposed ZCBA achieved an SNR of 57.4dB at a sampling rate of 40Mhz and consumes only 1.45mW under a 1V supply.

Chapter 4 presented a voltage-domain artificial neuron circuit compatible with neuromodulatory tuning. The behavior of spiking neurons and spiking neural networks was discussed. An analysis of a novel synapse design was presented. Simulation results showed the functionality of the neuron. The proposed neuron achieves a maximum spiking frequency of 3.3MHz and consumes only 1.08pJ/spike.

Chapter 5 detailed the design, fabrication and measurement of a novel phase-domain spiking neuron. A high-density bidirectional synapse design was presented. Both simulation and measurement results were shown. Simulation results showed the functionality of a phase-domain leak circuit which mimics an RC decay in the phase-domain. Measurement results showed how spiking frequency changes verses several input parameters. The phase-domain neuron consumes only 134fJ/spike under a 0.35V supply and occupies only 21μmx27μm.

6.3 Future Work

To continue research in the field of ZCBA, the following suggestions are provided:

1. Integrate the proposed ZCBA in an ADC. This would provide better comparison with existing ZCBA and provide further understanding of ZCBAs in system architectures.

2. Experiment with a reduced supply voltage. ZCBAs are well suited for low supply operation, and performance may not suffer dramatically. The behavioral model could be expanded to include subthreshold effects.

To continue research in phase-domain neuron design, the following suggestions are provided:

1. Design the ring oscillators to be robust to process variation. Much of the performance degradation in the proposed neuron is due to fabrication mismatch in the oscillators. A calibration circuit could be designed, similar to a PLL, to mitigate oscillator frequency mismatch.

2. Several system level parameters have not been explored. These include optimum spike width, input encoding scheme and VCO tuning range. Each of these parameters have the potential to significantly affect network performance.

Bibliography

[1] K. Yamamoto and A. Carusone, "A 1-1-1-1 MASH Delta-Sigma Modulator With Dynamic Comparator-Based OTAs," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 8, pp. 1866–1883, Aug. 2012, Conference Name: IEEE Journal of Solid-State Circuits, ISSN: 1558-173X. DOI: 10.1109/JSSC.2012.2196732.

[2] L. Brooks and H.-S. Lee, "A 12b, 50 MS/s, Fully Differential Zero-Crossing Based Pipelined ADC," *IEEE Journal of Solid-State Circuits*, vol. 44, no. 12, pp. 3329–3343, Dec. 2009, Conference Name: IEEE Journal of Solid-State Circuits, ISSN: 1558-173X. DOI: 10.1109/JSSC.2009.2032639.

[3] M. Chu, B. Kim, and B.-G. Lee, "A 10-bit 200-MS/s Zero-Crossing-Based Pipeline ADC in 0.13- \mu m CMOS Technology," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 23, no. 11, pp. 2671–2675, Nov. 2015, Conference Name: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2014.2371453.

[4] Y.-H. Kim and S. Cho, "A 1-GS/s 9-bit Zero-Crossing-Based Pipeline ADC Using a Resistor as a Current Source," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 7, pp. 2570–2579, Jul. 2016, Conference Name: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2015.2508564.

[5] J.-E. Park, Y.-H. Hwang, and D.-K. Jeong, "A 0.4-to-1 V Voltage Scalable \Delta \Sigma ADC With Two-Step Hybrid Integrator for IoT Sensor Applications in 65-nm LP CMOS," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 64, no. 12, pp. 1417–1421, Dec. 2017, Conference Name: IEEE Transactions on Circuits and Systems II: Express Briefs, ISSN: 1558-3791. DOI: 10.1109/TCSII.2017.2753841.

[6] S.-K. Shin, J. C. Rudell, D. C. Daly, *et al.*, "A 12 bit 200 MS/s Zero-Crossing-Based Pipelined ADC With Early Sub-ADC Decision and Output Residue Background Calibration," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 6, pp. 1366–1382, Jun. 2014, Conference Name: IEEE Journal of Solid-State Circuits, ISSN: 1558-173X. DOI: 10.1109/JSSC.2014.2322853.

[7] B. Li, J.-P. Na, W. Wang, J. Liu, Q. Yang, and P.-I. Mak, "A 13-bit 8-kS/s – Readout IC Using ZCB Integrators With an Embedded Resistive Sensor Achieving 1.05-pJ/Conversion Step and a 65-dB PSRR," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 27, no. 4, pp. 843–853, Apr. 2019, Conference Name: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2019.2895361.

[8] I.-H. Wang, H.-Y. Lee, and S.-I. Liu, "An 8-bit 20-MS/s ZCBC Time-Domain Analog-to-Digital Data Converter," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 56, no. 7, pp. 545–549, Jul. 2009, Conference Name: IEEE Transactions on Circuits and Systems II: Express Briefs, ISSN: 1558-3791. DOI: 10.1109/TCSII.2009.2022208.

[9] S.-K. Shin, Y.-S. You, S.-H. Lee, *et al.*, "A fully-differential zero-crossing-based 1.2V 10b 26MS/s pipelined ADC in 65nm CMOS," in *2008 IEEE Symposium on VLSI Circuits*, ISSN: 2158-5636, Jun. 2008, pp. 218–219. DOI: 10.1109/VLSIC.2008.4586013.

[10] J. K. Fiorenza, T. Sepke, P. Holloway, C. G. Sodini, and H.-S. Lee, "Comparator-Based Switched-Capacitor Circuits for Scaled CMOS Technologies," *IEEE Journal of Solid-State Circuits*, vol. 41, no. 12, pp. 2658–2668, Dec. 2006, Conference Name: IEEE Journal of Solid-State Circuits, ISSN: 1558-173X. DOI: 10.1109/JSSC.2006.884330.

[11] Y.-H. Kim and S. Cho, "A 1-GS/s 9-bit Zero-Crossing-Based Pipeline ADC Using a Resistor as a Current Source," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 24, no. 7, pp. 2570–2579, Jul. 2016, Conference Name: IEEE Transactions on Very Large Scale Integration (VLSI) Systems, ISSN: 1557-9999. DOI: 10.1109/TVLSI.2015.2508564.

[12] A. Joubert, B. Belhadj, O. Temam, and R. Héliot, "Hardware spiking neurons design: Analog or digital?" In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–5. DOI: 10.1109/IJCNN.2012.6252600.

[13] J. M. Cruz-Albrecht, M. W. Yung, and N. Srinivasa, "Energy-efficient neuron, synapse and stdp integrated circuits," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 3, pp. 246–256, 2012. DOI: 10.1109/TBCAS.2011.2174152.

[14] S. Moriya, H. Yamamoto, S. Sato, Y. Yuminaka, Y. Horio, and J. Madrenas, "A Fully Analog CMOS Implementation of a Two-variable Spiking Neuron in the Subthreshold Region and its Network Operation," in *2022 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-4407, Jul. 2022, pp. 1–7. DOI: 10.1109/IJCNN55064.2022.9891920.

[15] H. M. Lehmann, J. Hille, C. Grassmann, and V. Issakov, "Leaky Integrate-and-Fire Neuron with a Refractory Period Mechanism for Invariant Spikes," in *2022 17th Conference on Ph.D Research in*

Microelectronics and Electronics (PRIME), Jun. 2022, pp. 365–368. DOI: 10.1109/PRIME55000.2022.9816777.

[16] A. Rubino, C. Livanelioglu, N. Qiao, M. Payvand, and G. Indiveri, "Ultra-Low-Power FDSOI Neural Circuits for Extreme-Edge Neuromorphic Intelligence," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 45–56, Jan. 2021, Conference Name: IEEE Transactions on Circuits and Systems I: Regular Papers, ISSN: 1558-0806. DOI: 10.1109/TCSI.2020.3035575.

[17] Z. Yang, Z. Han, Y. Huang, and T. T. Ye, "55nm CMOS Analog Circuit Implementation of LIF and STDP Functions for Low-Power SNNs," in *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Jul. 2021, pp. 1–6. DOI: 10.1109/ISLPED52811.2021.9502497.

[18] S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, and K. Meier, "A highly tunable 65-nm CMOS LIF neuron for a large scale neuromorphic system," in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, Sep. 2016, pp. 71–74. DOI: 10.1109/ESSCIRC.2016.7598245.

[19] K.-J. Moon, D.-R. Oh, M. Choi, and S.-T. Ryu, "A 28-nm cmos 12-bit 250-ms/s voltage-current-time domain 3-stage pipelined adc," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 2843–2847, 2020. DOI: 10.1109/TCSII.2020.2990910.

[20] L. Qiu, C. Yang, K. Wang, and Y. Zheng, "A high-speed 2-bit/cycle sar adc with time-domain quantization," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 26, no. 10, pp. 2175–2179, 2018. DOI: 10.1109/TVLSI.2018.2837030.

[21] S. Zhu, B. Wu, Y. Cai, and Y. Chiu, "A 2-gs/s 8-bit non-interleaved time-domain flash adc based on remainder number system in 65-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 1172–1183, 2018. DOI: 10.1109/JSSC.2017.2774280.

[22] M. Zhang, Y. Zhu, C.-H. Chan, and R. P. Martins, "An 8-bit 10-gs/s 16× interpolation-based time-domain adc with <1.5-ps uncalibrated quantization steps," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 12, pp. 3225–3235, 2020. DOI: 10.1109/JSSC.2020.3012776.

[23] Y. Zhong, S. Li, X. Tang, *et al.*, "A second-order purely vco-based ct $\Delta\Sigma$ adc using a modified dpll structure in 40-nm cmos," *IEEE Journal of Solid-State Circuits*, vol. 55, no. 2, pp. 356–368, 2020. DOI: 10.1109/JSSC.2019.2948008.

[24] L. Liu, J. Jin, X. Liu, and J. Zhou, "A multi-modulus fractional divider with tdc free calibration scheme for mitigation of tx-vco pulling," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 67, no. 12, pp. 2848–2852, 2020. DOI: 10.1109/TCSII.2020.2983785.

[25] P. S. Locatelli, D. M. Colombo, and K. El-Sankary, "Time-domain multiply–accumulate unit," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 31, no. 6, pp. 762–775, 2023. DOI: 10.1109/TVLSI.2023.3266743.

[26] Y. Song, S. Smith, B. Karlinsey, A. R. Hawkins, and S.-H. W. Chiang, "The digital-assisted charge amplifier: A digital-based approach to charge amplification," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 8, pp. 3114–3123, 2022. DOI: 10.1109/TCSI.2022.3169056.

[27] J. Matsuno, D. Kurose, T. Sugimoto, H. Ishii, M. Furuta, and T. Itakura, "A power-scalable zero-crossing-based amplifier using inverter-based zero-crossing detector with cmfb," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2016, pp. 482–485. DOI: 10.1109/ISCAS.2016.7527282.

[28] Z. Li, W. He, F. Ye, and J. Ren, "A Low-Power Low-Noise Dynamic Comparator With Latch-Embedding Floating Amplifier," in *2020 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*, Dec. 2020, pp. 39–42. DOI: 10.1109/APCCAS50809.2020.9301705.

[29] M. S. Akter, K. A. A. Makinwa, and K. Bult, "A capacitively degenerated 100-db linear 20–150 ms/s dynamic amplifier," *IEEE Journal of Solid-State Circuits*, vol. 53, no. 4, pp. 1115–1126, 2018. DOI: 10.1109/JSSC.2017.2778277.

[30] S. Kalani, T. Haque, R. Gupta, and P. R. Kinget, "Using vco-ota tias to break the gain, linearity and power consumption trade-offs in passive mixer based direct-conversion receivers," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5. DOI: 10.1109/ISCAS.2018.8351469.

[31] S. Kalani, A. Bertolini, A. Richelli, and P. R. Kinget, "A 0.2v 492nw vco-based ota with 60khz ugb and 207 vrms noise," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2017, pp. 1–4. DOI: 10.1109/ISCAS.2017.8050503.

[32] X. Tang, X. Yang, W. Zhao, *et al.*, "9.5 a 13.5b-enob second-order noise-shaping sar with pvt-robust closed-loop dynamic amplifier," in *2020 IEEE International Solid- State Circuits Conference - (ISSCC)*, 2020, pp. 162–164. DOI: 10.1109/ISSCC19947.2020.9063058.

[33] M. Miyahara, Y. Asada, D. Paik, and A. Matsuzawa, "A low-noise self-calibrating dynamic comparator for high-speed ADCs," in *2008 IEEE Asian Solid-State Circuits Conference*, Nov. 2008, pp. 269–272. DOI: 10.1109/ASSCC.2008.4708780.

[34] T. Barton, H. Yu, K. Rogers, *et al.*, "Towards low-power machine learning architectures inspired by brain neuromodulatory signalling," *Journal of Low Power Electronics and Applications*, vol. 12, no. 4, 2022, ISSN: 2079-9268. DOI: 10.3390/jlpea12040059. [Online]. Available: https://www.mdpi.com/2079-9268/12/4/59.

[35] M. Pfeiffer and T. Pfeil, "Deep learning with spiking neurons: Opportunities and challenges," *Frontiers in neuroscience*, p. 774, 2018.

[36] A. R. Voelker, D. Rasmussen, and C. Eliasmith, "A spike in performance: Training hybrid-spiking neural networks with quantized activation functions," *arXiv preprint arXiv:2002.03553*, 2020.

[37] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," *International journal of neural systems*, vol. 19, no. 04, pp. 295–308, 2009.

[38] A. Tavanaei, M. Ghodrati, S. R. Kheradpisheh, T. Masquelier, and A. Maida, "Deep learning in spiking neural networks," *Neural networks*, vol. 111, pp. 47–63, 2019.

[39] F. Ponulak and A. Kasinski, "Introduction to spiking neural networks: Information processing, learning and applications.," *Acta neurobiologiae experimentalis*, vol. 71, no. 4, pp. 409–433, 2011.

[40] H. Daumé III, "Frustratingly easy domain adaptation," in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic: Association for Computational Linguistics, Jun. 2007, pp. 256–263. [Online]. Available: https://aclanthology.org/P07-1033.

[41] B. Sun, J. Feng, and K. Saenko, "Return of frustratingly easy domain adaptation," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, ser. AAAI'16, Phoenix, Arizona: AAAI Press, 2016, pp. 2058–2065.

[42] J. Blitzer, R. McDonald, and F. Pereira, "Domain adaptation with structural correspondence learning," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia: Association for Computational Linguistics, Jul. 2006, pp. 120–128. [Online]. Available: https://aclanthology.org/W06-1615.

[43] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto, "Few-shot adversarial domain adaptation," in *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, *et al.*, Eds., vol. 30, Curran Associates, Inc., 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/file/21c5bba1dd6aed9ab48c2b34c1a0adde-Paper.pdf.

[44] H. Liu, J. Wang, and M. Long, "Cycle self-training for domain adaptation," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 22 968–22 981. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/c1fea270c48e8079d8ddf7d06d26ab52-Paper.pdf.

[45] P. Stojanov, Z. Li, M. Gong, R. Cai, J. Carbonell, and K. Zhang, "Domain adaptation with invariant representation learning: What transformations to learn?" In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 24 791–24 803.

[Online]. Available: https://proceedings.neurips.cc/paper/2021/file/cfc5d9422f0c8f8ad796711102dbe32b-Paper.pdf.

[46] N. D. Lawrence and J. C. Platt, "Learning to learn with the informative vector machine," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 65.

[47] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng, "Self-taught learning: Transfer learning from unlabeled data," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 759–766.

[48] A. Argyriou, T. Evgeniou, and M. Pontil, "Multi-task feature learning," *Advances in neural information processing systems*, vol. 19, 2006.

[49] S.-I. Lee, V. Chatalbashev, D. Vickrey, and D. Koller, "Learning a meta-level prior for feature relevance from multiple related tasks," in *Proceedings of the 24th international conference on Machine learning*, 2007, pp. 489–496.

[50] D. Li and H. Zhang, "Improved regularization and robustness for fine-tuning in neural networks," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 27 249–27 262. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/e4a93f0332b2519177ed55741ea4e5e7-Paper.pdf.

[51] X. Dong, A. T. Luu, M. Lin, S. Yan, and H. Zhang, "How should pre-trained language models be fine-tuned towards adversarial robustness?" In *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 4356–4369. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/22b1f2e0983160db6f7bb9f62f4dbb39-Paper.pdf.

[52] Y. Zhang, B. Hooi, D. Hu, J. Liang, and J. Feng, "Unleashing the power of contrastive self-supervised visual models via contrast-regularized fine-tuning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34, Curran Associates, Inc., 2021, pp. 29 848–29 860. [Online]. Available: https://proceedings.neurips.cc/paper/2021/file/fa14d4fe2f19414de3ebd9f63d5c0169-Paper.pdf.

[53] O. Vinyals, C. Blundell, T. Lillicrap, D. Wierstra, *et al.*, "Matching networks for one shot learning," *Advances in neural information processing systems*, vol. 29, 2016.

[54] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," *Advances in neural information processing systems*, vol. 30, 2017.

[55] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 1877–1901. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[56] Z. Yue, H. Zhang, Q. Sun, and X.-S. Hua, "Interventional few-shot learning," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33, Curran Associates, Inc., 2020, pp. 2734–2746. [Online]. Available: https://proceedings.neurips.cc/paper/2020/file/1cc8a8ea51cd0adddf5dab504a285915-Paper.pdf.

[57] Z. Brown, N. Robinson, D. Wingate, and N. Fulda, "Towards neural programming interfaces," *Advances in Neural Information Processing Systems*, vol. 33, pp. 17 416–17 428, 2020.

[58] F. Agostinelli, M. Hoffman, P. Sadowski, and P. Baldi, "Learning activation functions to improve deep neural networks," *arXiv preprint arXiv:1412.6830*, 2014.

[59] E. B. Zaken, S. Ravfogel, and Y. Goldberg, *Bitfit: Simple parameter-efficient fine-tuning for transformer-based masked language-models*, 2021. DOI: 10.48550/ARXIV.2106.10199. [Online]. Available: https://arxiv.org/abs/2106.10199.

[60] J. Kendall, R. Pantone, K. Manickavasagam, Y. Bengio, and B. Scellier, "Training end-to-end analog neural networks with equilibrium propagation," *arXiv preprint arXiv:2006.01981*, 2020.

[61] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, 2014. DOI: 10.1109/JPROC.2014.2304638.

[62] K. Voutsas and J. Adamy, "A biologically inspired spiking neural network for sound source lateralization," *IEEE Transactions on Neural Networks*, vol. 18, no. 6, pp. 1785–1799, 2007. DOI: 10.1109/TNN.2007.899623.

[63] Z. Yang, Z. Han, Y. Huang, and T. T. Ye, "55nm cmos analog circuit implementation of lif and stdp functions for low-power snns," in *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, 2021, pp. 1–6. DOI: 10.1109/ISLPED52811.2021.9502497.

[64] B. Rueckauer and S.-C. Liu, "Conversion of analog to spiking neural networks using sparse temporal coding," in *2018 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2018, pp. 1–5. DOI: 10.1109/ISCAS.2018.8351295.

[65] W. Zhang and P. Li, "Spike-train level backpropagation for training deep recurrent spiking neural networks," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32, Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper/2019/file/f42a37d114a480b6b57b60ea9a14a9d2-Paper.pdf.

[66] J. R. Miquel, S. Tolu, F. E. T. Schx00F6;ller, and R. Galeazzi, "Retinanet object detector based on analog-to-spiking neural network conversion," in *2021 8th International Conference on Soft Computing Machine Intelligence (ISCMI)*, 2021, pp. 201–205. DOI: 10.1109/ISCMI53840.2021.9654818.

[67] J. Ding, Z. Yu, Y. Tian, and T. Huang, "Optimal ANN-SNN conversion for fast and accurate inference in deep spiking neural networks," *CoRR*, vol. abs/2105.11654, 2021. arXiv: 2105.11654. [Online]. Available: https://arxiv.org/abs/2105.11654.

[68] Y. Li, S. Deng, X. Dong, and S. Gu, *Converting artificial neural networks to spiking neural networks via parameter calibration*, 2022. DOI: 10.48550/ARXIV.2205.10121. [Online]. Available: https://arxiv.org/abs/2205.10121.

[69] G. Indiveri, E. Chicca, and R. Douglas, "A vlsi array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 211–221, 2006. DOI: 10.1109/TNN.2005.860850.

[70] B. Han, G. Srinivasan, and K. Roy, "Rmp-snn: Residual membrane potential neuron for enabling deeper high-accuracy and low-latency spiking neural network," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 555–13 564. DOI: 10.1109/CVPR42600.2020.01357.

[71] J. Li, C. Zhao, K. Hamedani, and Y. Yi, "Analog hardware implementation of spike-based delayed feedback reservoir computing system," in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 3439–3446. DOI: 10.1109/IJCNN.2017.7966288.

[72] S. Nitundil, G. Susi, and F. Maestú, "Design of an analog multi-neuronal spike-sequence detector (mnsd) based on a 180nm cmos leaky integrate amp; fire with latency neuron," in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2021, pp. 1–6. DOI: 10.1109/ICAECT49130.2021.9392419.

[73] Q. Sun, F. Schwartz, J. Michel, Y. Herve, and R. Dal Molin, "Implementation study of an analog spiking neural network for assisting cardiac delay prediction in a cardiac resynchronization therapy device," *IEEE Transactions on Neural Networks*, vol. 22, no. 6, pp. 858–869, 2011. DOI: 10.1109/TNN.2011.2125986.

[74] H. Mostafa, "Supervised learning based on temporal coding in spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 7, pp. 3227–3235, 2018. DOI: 10.1109/TNNLS.2017.2726060.

[75] H.-Y. Hsieh and K.-T. Tang, "Vlsi implementation of a bio-inspired olfactory spiking neural network," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 7, pp. 1065–1073, 2012. DOI: 10.1109/TNNLS.2012.2195329.

[76] M.-H. Kim, S. Hwang, S. Bang, *et al.*, "A more hardware-oriented spiking neural network based on leading memory technology and its application with reinforcement learning," *IEEE Transactions on Electron Devices*, vol. 68, no. 9, pp. 4411–4417, 2021. DOI: 10.1109/TED.2021.3099769.

[77] V. Cincon, E. I. Vatajelu, L. Anghel, and P. Galy, "From 1.8v to 0.19v voltage bias on analog spiking neuron in 28nm utbb fd-soi technology," in *2020 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, 2020, pp. 1–4. DOI: 10.1109/EUROSOI-ULIS49407.2020.9365302.

[78] F. Danneville, I. Sourikopoulos, S. Hedayat, C. Loyez, V. Hoël, and A. Cappy, "Ultra low power analog design and technology for artificial neurons," in *2017 IEEE Bipolar/BiCMOS Circuits and Technology Meeting (BCTM)*, 2017, pp. 1–8. DOI: 10.1109/BCTM.2017.8112899.

[79] I. Satyaraj and B. J. Kailath, "A simple pstdp circuit for analog implementation of spiking neural networks," in *2020 IEEE 4th Conference on Information Communication Technology (CICT)*, 2020, pp. 1–4. DOI: 10.1109/CICT51604.2020.9312100.

[80] D. Kim, X. She, N. M. Rahman, V. C. K. Chekuri, and S. Mukhopadhyay, "Processing-in-memory-based on-chip learning with spike-time-dependent plasticity in 65-nm cmos," *IEEE Solid-State Circuits Letters*, vol. 3, pp. 278–281, 2020. DOI: 10.1109/LSSC.2020.3013448.

[81] M. R. Azghadi, S. Al-Sarawi, N. Iannella, and D. Abbott, "Efficient design of triplet based spike-timing dependent plasticity," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*, 2012, pp. 1–7. DOI: 10.1109/IJCNN.2012.6252820.

[82] J. Clements, "Transmitter timecourse in the synaptic cleft: Its role in central synaptic function," *Trends in Neurosciences*, vol. 19, no. 5, pp. 163–171, 1996, ISSN: 0166-2236. DOI: https://doi.org/10.1016/S0166-2236(96)10024-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0166223696100242.

[83] L. F. Agnati, M. Zoli, I. Strömberg, and K. Fuxe, "Intercellular communication in the brain: Wiring versus volume transmission," English, *Neuroscience*, vol. 69, no. 3, pp. 711–726, 1995, Cited By :451. [Online]. Available: www.scopus.com.

[84] J. T. Yorgason, D. M. Zeppenfeld, and J. T. Williams, "Cholinergic interneurons underlie spontaneous dopamine release in nucleus accumbens," *Journal of Neuroscience*, vol. 37, no. 8, pp. 2086–2096, 2017, ISSN: 0270-6474. DOI: 10.1523/JNEUROSCI.3064-16.2017. eprint: https://www.jneurosci.org/content/37/8/2086.full.pdf. [Online]. Available: https://www.jneurosci.org/content/37/8/2086.

[85] J. .-. Beaulieu and R. R. Gainetdinov, "The physiology, signaling, and pharmacology of dopamine receptors," English, *Pharmacological reviews*, vol. 63, no. 1, pp. 182–217, 2011, Cited By :1628. [Online]. Available: www.scopus.com.

[86] R. A. Depue and P. F. Collins, "Neurobiology of the structure of personality: Dopamine, facilitation of incentive motivation, and extraversion," English, *Behavioral and Brain Sciences*, vol. 22, no. 3, pp. 491–517, 1999, Cited By :1391. [Online]. Available: www.scopus.com.

[87] M. J. Frank, "Dynamic dopamine modulation in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated parkinsonism," English, *Journal of cognitive neuroscience*, vol. 17, no. 1, pp. 51–72, 2005, Cited By :657. [Online]. Available: www.scopus.com.

[88] J. C. Stoof and J. W. Kebabian, "Two dopamine receptors: Biochemistry, physiology and pharmacology," English, *Life Sciences*, vol. 35, no. 23, pp. 2281–2296, 1984, Cited By :851. [Online]. Available: www.scopus.com.

[89] A. Reiner and J. Levitz, "Glutamatergic signaling in the central nervous system: Ionotropic and metabotropic receptors in concert," *Neuron*, vol. 98, no. 6, pp. 1080–1098, 2018, Cited by: 198; All Open Access, Bronze Open Access, Green Open Access. DOI: 10.1016/j.neuron.2018.05.018. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85047957321&doi=10.1016%2fj.neuron.2018.05.018&partnerID=40&md5=bacb458a82b8c6d7a5a666d6c94b844e.

[90] S. A. A. et al., "An accelerated lif neuronal network array for a large-scale mixed-signal neuromorphic architecture," *I,EEE Transactions on Circuits and Systems I: Regular Papers*, 2018.

[91] *Imagenet dataset*, https://www.image-net.org/.

[92] A. Joubert, B. Belhadj, O. Temam, and R. Héliot, "Hardware spiking neurons design: Analog or digital?" In *The 2012 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-4407, Jun. 2012, pp. 1–5. DOI: 10.1109/IJCNN.2012.6252600.

[93] J. M. Cruz-Albrecht, M. W. Yung, and N. Srinivasa, "Energy-Efficient Neuron, Synapse and STDP Integrated Circuits," *IEEE Transactions on Biomedical Circuits and Systems*, vol. 6, no. 3, pp. 246–256, Jun. 2012, Conference Name: IEEE Transactions on Biomedical Circuits and Systems, ISSN: 1940-9990. DOI: 10.1109/TBCAS.2011.2174152.

[94] S. Moriya, H. Yamamoto, S. Sato, Y. Yuminaka, Y. Horio, and J. Madrenas, "A Fully Analog CMOS Implementation of a Two-variable Spiking Neuron in the Subthreshold Region and its Network Operation," in *2022 International Joint Conference on Neural Networks (IJCNN)*, ISSN: 2161-4407, Jul. 2022, pp. 1–7. DOI: 10.1109/IJCNN55064.2022.9891920.

[95] V. Cincon, E. I. Vatajelu, L. Anghel, and P. Galy, "From 1.8V to 0.19V voltage bias on analog spiking neuron in 28nm UTBB FD-SOI technology," in *2020 Joint International EUROSOI Workshop and International Conference on Ultimate Integration on Silicon (EUROSOI-ULIS)*, ISSN: 2472-9132, Sep. 2020, pp. 1–4. DOI: 10.1109/EUROSOI-ULIS49407.2020.9365302.

[96] B. Joo, J.-W. Han, and B.-S. Kong, "Energy- and Area-Efficient CMOS Synapse and Neuron for Spiking Neural Networks With STDP Learning," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 69, no. 9, pp. 3632–3642, Sep. 2022, Conference Name: IEEE Transactions on Circuits and Systems I: Regular Papers, ISSN: 1558-0806. DOI: 10.1109/TCSI.2022.3178989.

[97] H. M. Lehmann, J. Hille, C. Grassmann, and V. Issakov, "Leaky Integrate-and-Fire Neuron with a Refractory Period Mechanism for Invariant Spikes," in *2022 17th Conference on Ph.D Research in Microelectronics and Electronics (PRIME)*, Jun. 2022, pp. 365–368. DOI: 10.1109/PRIME55000.2022.9816777.

[98] A. Rubino, C. Livanelioglu, N. Qiao, M. Payvand, and G. Indiveri, "Ultra-Low-Power FDSOI Neural Circuits for Extreme-Edge Neuromorphic Intelligence," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 68, no. 1, pp. 45–56, Jan. 2021, Conference Name: IEEE Transactions on Circuits and Systems I: Regular Papers, ISSN: 1558-0806. DOI: 10.1109/TCSI.2020.3035575.

[99] Z. Yang, Z. Han, Y. Huang, and T. T. Ye, "55nm CMOS Analog Circuit Implementation of LIF and STDP Functions for Low-Power SNNs," in *2021 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED)*, Jul. 2021, pp. 1–6. DOI: 10.1109/ISLPED52811.2021.9502497.

[100] S. A. Aamir, P. Müller, A. Hartel, J. Schemmel, and K. Meier, "A highly tunable 65-nm CMOS LIF neuron for a large scale neuromorphic system," in *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, Sep. 2016, pp. 71–74. DOI: 10.1109/ESSCIRC.2016.7598245.

[101] X. Wu, V. Saxena, K. Zhu, and S. Balagopal, "A CMOS Spiking Neuron for Brain-Inspired Neural Networks With Resistive Synapses and In Situ Learning," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 62, no. 11, pp. 1088–1092, Nov. 2015, Conference Name: IEEE Transactions on Circuits and Systems II: Express Briefs, ISSN: 1558-3791. DOI: 10.1109/TCSII.2015.2456372.

[102] B. Datta Sahoo, "Ring oscillator based sub-1V leaky integrate-and-fire neuron circuit," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*, ISSN: 2379-447X, May 2017, pp. 1–4. DOI: 10.1109/ISCAS.2017.8050980.

[103] J. Song, J. Shin, H. Kim, and W.-S. Choi, *Energy-Efficient High-Accuracy Spiking Neural Network Inference Using Time-Domain Neurons*, arXiv:2202.02015 [cs, eess], Apr. 2022. [Online]. Available: http://arxiv.org/abs/2202.02015 (visited on 02/21/2023).

[104] A. Mukherjee, M. Gandara, X. Yang, *et al.*, "A 74.5-dB Dynamic Range 10-MHz BW CT- ADC With Distributed-Input VCO and Embedded Capacitive- Network in 40-nm CMOS," *IEEE Journal of Solid-State Circuits*, vol. 56, no. 2, pp. 476–487, Feb. 2021, Conference Name: IEEE Journal of Solid-State Circuits, ISSN: 1558-173X. DOI: 10.1109/JSSC.2020.3012623.

[105] E. Swindlehurst, H. Jensen, A. Petrie, *et al.*, "An 8-bit 10-GHz 21-mW Time-Interleaved SAR ADC With Grouped DAC Capacitors and Dual-Path Bootstrapped Switch," *IEEE Solid-State Circuits Letters*, vol. 2, no. 9, pp. 83–86, Sep. 2019, Conference Name: IEEE Solid-State Circuits Letters, ISSN: 2573-9603. DOI: 10.1109/LSSC.2019.2931440.

Appendices

A Appendix A: Proof that no analytical minimum exists for ZCBA

Chaper 2 claims that there exists no closed form solution for Eq. 2.1. This appendix rigorously analyzes why this is the case.

The ZCBA settling time T_{settle} is given as

$$T_{settle} = T_{comp} \sum_{n=1}^{m} \lceil \frac{V_{x,n}}{\Delta V_n} \rceil \tag{A.1}$$

which includes a non-differentiable ceiling function, preventing us from using its derivative to find the minimum T_{settle}. Therefore, we approximate it as

$$T_{settle} = m\frac{T_{comp}}{2} + T_{comp} \sum_{n=1}^{m} \frac{V_{x,n}}{\Delta V_n} \tag{A.2}$$

which eliminates the ceiling function by assuming each stage overshoots by $T_{comp}/2$ on average over a set of random inputs. It follows that the setting time for stages 2 through $m-1$ will be the same for every value of V_{in} on average. We can therefore separate the settling time of the first stage and the rest of the stages:

$$T_{settle} = m\frac{T_{comp}}{2} + \frac{V_{in}}{V_{msb}}T_{comp} + T_{comp} \sum_{n=2}^{m} \frac{V_{x,n}}{\Delta V_n} \tag{A.3}$$

where $V_{msb} = I_{CP}T_{comp}/C_L$. We then expand $V_{x,n}$ and ΔV_n as

$$V_{x,n} = \frac{I_{cp,n-1}\beta T_{comp}}{2C_L} \tag{A.4}$$

and

$$\Delta V_n = \frac{I_{cp,n}\beta T_{comp}}{C_L} \tag{A.5}$$

where β is the ZCBA's feedback factor. We subsequently define $I_{cp,n}$ and $I_{cp,n-1}$ as

$$I_{cp,n} = \frac{I_{cp,total}}{k^{n-1}} \tag{A.6}$$

and

$$I_{cp,n-1} = \frac{I_{cp,total}}{k^{n-2}} \tag{A.7}$$

where k is the radix. Combining equations (A.4) - (A.7), we obtain

$$\sum_{n=2}^{m} \frac{V_{x,n}}{\Delta V_n} = \sum_{n=2}^{m} \frac{k}{2}. \tag{A.8}$$

Evaluating the sum and substituting the result into equation (A.3) yields:

$$T_{settle} = m\frac{T_{comp}}{2} + \frac{V_{in}}{V_{MSB}}T_{comp} + \frac{T_{comp}(m-1)k}{2}. \tag{A.9}$$

From Section II of the chapter 2, we have

$$I_{CP,total} = \sum_{n=0}^{m-1} I_{LSB}k^n. \tag{A.10}$$

Re-writing (A.10) as

$$k = \sqrt[m]{1 - \frac{(1-k)I_{CP}}{I_{LSB}}}, \tag{A.11}$$

we note that k does not have a closed-form solution. It follows that (A.9) which contains k does not have a closed-form solution. We thus conclude that no minimum of T_{settle} can be found through analytical methods, and numerical methods are necessary as described in our paper.