Energy-Efficient Circuit Designs for Miniaturized Internet of Things and Wireless Neural Recording

by

Jongyup Lim

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical and Computer Engineering)
in the University of Michigan
2021

Doctoral Committee:

Professor Dennis M. Sylvester, Chair
Professor David T. Blaauw
Associate Professor Cynthia A. Chestek
Assistant Professor Hun-Seok Kim

Jongyup Lim

jongyup@umich.edu

ORCID iD: 0000-0003-0306-3966

DEDICATION

To my wife Suzie,

my parents, and family

with love and gratitude

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Internet of Things (IoT) have become omnipresent over various territories including healthcare, smart building, agriculture, and environmental and industrial monitoring. Today, IoT are getting miniaturized, but at the same time, they are becoming more intelligent along with the explosive growth of machine learning. Not only do IoT sense and collect data and communicate, but they also edge-compute and extract useful information within the small form factor. A main challenge of such miniaturized and intelligent IoT is to operate continuously for long lifetime within its low battery capacity. Energy efficiency of circuits and systems is key to addressing this challenge. This dissertation presents two different energy-efficient circuit designs: a 224pW 260ppm/°C gate-leakage-based timer for wireless sensor nodes (WSNs) for the IoT and an energy-efficient all analog machine learning accelerator with 1.2 μJ/inference of energy consumption for the CIFAR-10 and SVHN datasets.

Wireless neural interface is another area that demands miniaturized and energy-efficient circuits and systems for safe long-term monitoring of brain activity. Historically, implantable systems have used wires for data communication and power, increasing risks of tissue damage. Therefore, it has been a long-standing goal to distribute sub-mm-scale true floating and wireless implants throughout the brain and to record single-neuron-level activities. This dissertation presents a $0.19\times0.17mm^2$ 0.74μW wireless neural recording IC with near-infrared (NIR) power and data telemetry and a $0.19\times0.28mm^2$ 0.57μW light tolerant wireless neural recording IC.

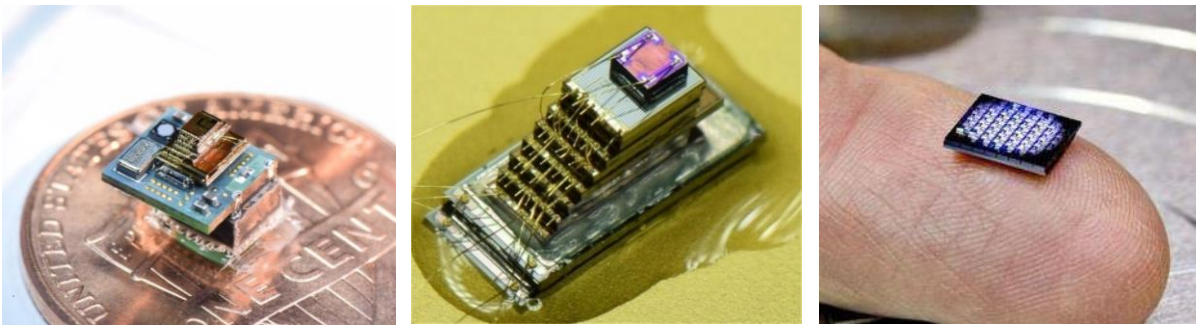
CHAPTER 1

Introduction

1.1 Miniaturized and Intelligent Internet of Things

Beginning from early mainframe computers in the 1950s, the computing platform has evolved to personal computers (PC) in the 1980s, smartphones in the 2000s, and now various kinds of Internet of Things (IoT). Development of complementary metal-oxide-semiconductor (CMOS) technology along with expansion of new hardware/software and increasing user demands for inexpensive and convenient access have driven the computing platform to be less expensive and physically smaller in size. Today, the IoT with smaller physical dimension relative to its predecessors have become ubiquitous over various applications that require seamless monitoring or sensing; e.g., healthcare, smart building, agriculture, and so on. Furthermore, recent research on millimeter-scale wireless sensor nodes [1]-[10] have enabled further shrinkage of IoT size, opening up new applications including medical implants, environmental monitoring, surveillance, and blockchain technology to the supply chain as shown in Fig. 1.1.

Main challenge of the miniaturized sensor nodes and IoT, however, is their limited energy budget from small battery capacity. Since a large number of the miniaturized IoT should be deployed all over different application areas wirelessly, it is infeasible to externally supply power



(a) (b) (c)

Figure 1.1 Miniaturized sensor nodes and IoTs (a) Stacked audio sensor node [5], (b) Stacked pressure sensor node [6], (c) IBM 1mm^3 computing platform [7].

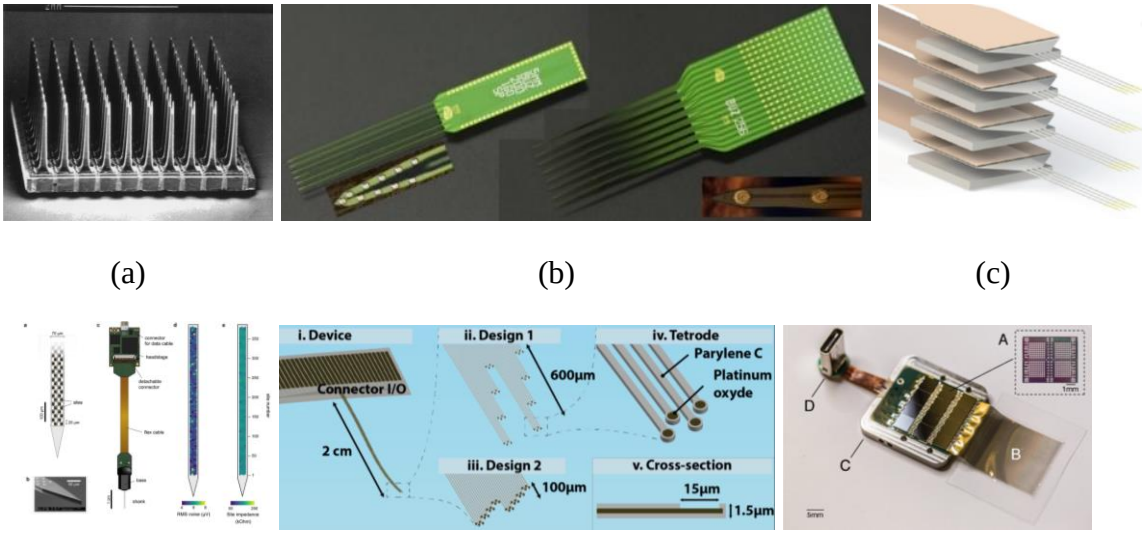
to individuals. Therefore, the IoT devices require battery as their power source. However, miniaturization of the system size has also forced the battery size to be shrunk limiting overall system energy budget. For instance, sub-mm^3 Li Thin-film battery has approximately $10^6\times$ lower energy capacity compared to the conventional alkaline AA battery with nearly 10 cm^3 of volume [11]. In other words, the miniaturized device using the sub-mm^3 Li Thin-film as a power source should consume only 10nW in average to achieve similar lifetime with the device using AA battery consuming 10mW in average. Therefore, energy-efficient and low power circuits and system design is key to addressing this challenge.

Another recent technological trend has been the explosive growth of deep learning [12] and its wide usage across numerous applications; e.g., self-driving cars, autonomous machines, medicine, entertainment, security, and so on. Along with the proliferation of IoT devices, demand of machine learning accelerators designed for edge computing has increased. Within small form factor and limited energy budget, the intelligent IoT need to not only sense and collect data and communicate, but also edge-compute to perform inference or even training. Therefore, the importance of energy-efficient circuit and system design has grown to meet the demand of increased computational capability of such IoT.

In chapter 2, an ultra-low power timer for wireless sensor nodes (WSNs) for the miniaturized IoT [13] is proposed. In chapter 3, an energy-efficient all analog machine learning accelerator for IoT edge computing [14] is presented.

1.2 True Wireless Neural Recording

Starting from 1970s, brain machine interfaces (BMI) [15] has been developed with an initial goal of restoring useful function of people who are paralyzed or disabled by neuromuscular disorders, such as spinal cord injury. Nowadays, active research on neural probe arrays have enabled high channel neural recording implants [16]-[21] (Fig. 1.2). However, implantable systems with these probe arrays use wires to connect the arrays to a bulky neural recording application-specific integrated circuit (ASIC). The use of wires and large form factor of system increase potential risks of tissue damage, infection, and cerebrospinal fluid leakage. Since the brain



(a) (b) (c)

(d) (e) (f)

Figure 1.2 Electrode arrays and neural recording system (a) Utah electrode array [16], (b) NeuroNexus Michigan Probes [17], (c) Caltech 3D electrodes[18], (d) imec Neuropixels [19], (e) Standford NeuroRoots [20], and (f) Neuralink prototype [21].

undergoes micro motion, even flexible wires can still create scraping and form scar tissue, making these systems unsuitable for long-term implantation.

Therefore, miniaturized, and true floating wireless neural recording motes that do minimal damage have been a long-standing goal. Recently, several miniaturized and wireless neural implants have been proposed including near-field RF, ultrasonic, and near-infrared (NIR) based powering and data telemetry [22]-[27]. However, RF-based motes [22]-[23] require 0.5W of transceiver power to operate exceeding the safety exposure limits by 10× [23], while ultrasonic-based motes remain relatively large in sizes ($0.8mm^3$) due to a bulky ultrasound transducer [24]-[25]. On the other hand, NIR-based approach using a photovoltaic (PV) cell and a light-emitting diode (LED) has shown promising wireless implementation shrinking down the mote sizes to 100s of microns [26]-[27]. Several challenges, however, still exist with NIR-based approach. One of the main challenges is its limited energy and area budget. NIR optical power density must be maintained under 190 $\mu W/mm^2$ due to safety limit of the brain [28], and the total energy that the PV cell can harvest is also limited by its small area (i.e. 1.5μW of electrical power from 150 $\mu W/mm^2$ NIR light with PV size of $190\times204\mu m^2$ [28]). Therefore, the energy efficient and highly compact neural recording circuit and system design is key to addressing this challenge.

In chapter 4 and 5, two generations of sub-μW and sub-mm wireless neural recording IC for motor prediction with NIR power and data telemetry are proposed [29]-[30].

Lastly, chapter 6 summarizes key contributions of the works presented from chapter 2 to chapter 5 and proposes future directions.

CHAPTER 2

A Gate-Leakage-based Frequency-Locked Timer for Ultra-Low Power Sensor Nodes with Second-Order Temperature Dependency Cancellation

2.1 Introduction

Wake-up timers are a critical component of wireless sensor nodes (WSNs) for the Internet of Things. Since they are on even when the sensor node is in sleep mode, they must consume extremely low power. In addition, they should ensure high timing accuracy for synchronization between devices and general timekeeping while remaining compact, leading to a highly constrained design space. An RC oscillator [31] or frequency-locked oscillator [32] based on temperature-compensated resistors achieves frequency stability across temperature of <50ppm/°C. However, these approaches consume ~100 nW or more, which far exceeds the power budget of state-of-the-art ultra-low power sensors. Extending these approaches to sub-nW requires extremely large resistors, unacceptably increasing the area and cost. Recently, a switch-resistor based timer achieved a high effective resistance without increasing resistor size and obtained a temperature coefficient (TC) of 13.8 ppm/°C [33]. However, the approach requires large capacitors, and power consumption remains relatively high at 4.7 nW. An alternative to resistor-based timers is gate-

leakage-based timers; several such timers have been proposed [34]-[35], providing sub-nW power consumption in compact silicon area. However, gate leakage exhibits significant first- and second-order temperature dependence, complicating temperature compensation, and it is also sensitive to the gate voltage. As a result, previous gate leakage timers have TCs in excess of several hundred ppm/°C and line sensitivities (LS) >150%/V. The gate leakage timer in [35] achieves 31 ppm/°C but requires 10-point calibration, and its 660 pW power consumption does not include the power of a required auxiliary temperature sensor. Further, its LS is unacceptably high at 420%/V.

This chapter proposes a 224-pW gate-leakage-based frequency locked timer with first- and second-order temperature dependency cancellation, yielding a TC of 260 ppm/°C across −5 to 95°C. Supply insensitive reference voltage generators and an on-chip low dropout (LDO) regulator decrease LS to 0.93%/V for 1.1−3.3 V, which marks a 150× improvement compared to previous gate-leakage-based timers.

2.2 Gate-Leakage-based Frequency-Locked Timer

The proposed design uses a frequency locked oscillator scheme [32]-[33] in which current I_1, set by the gate leakage of a standard-V_{th} NMOS N_1, is matched with current I_2, using modulation of the frequency of a switched capacitor C_2 (Fig. 2.1). The measured temperature dependence of N_1 gate leakage shows both first- and second-order components (Fig. 2.2). It is essential to cancel both components to achieve a good TC.

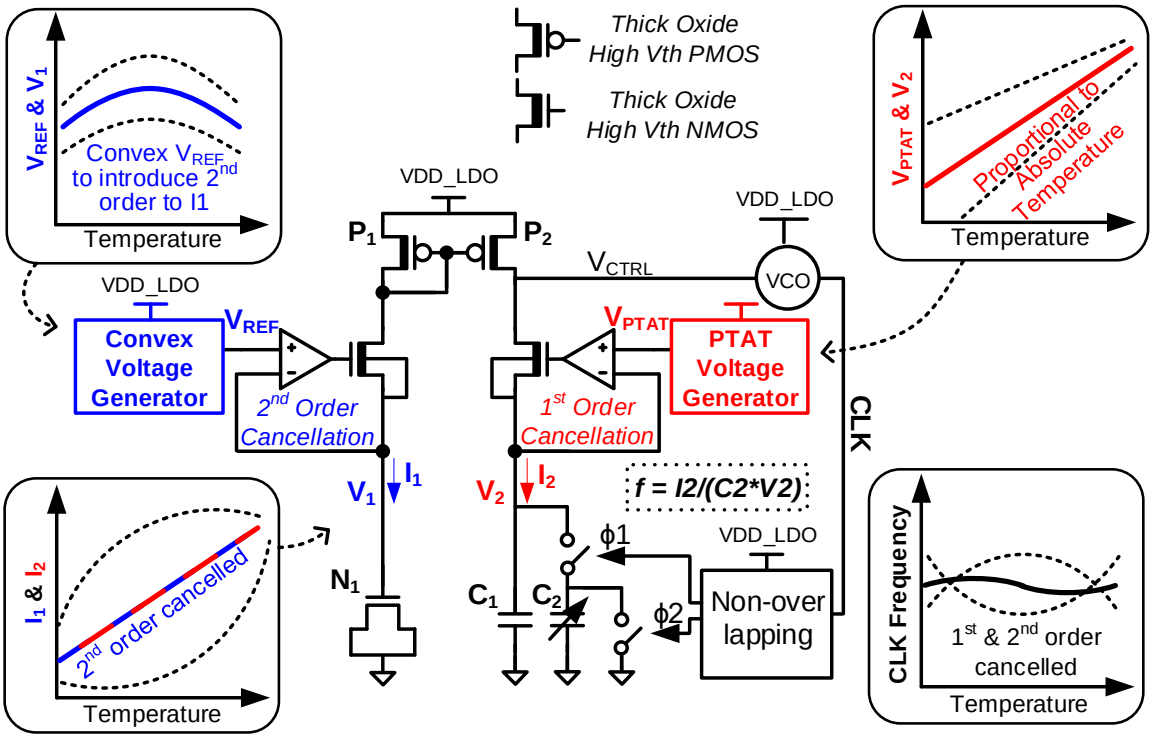


Figure 2.1 Proposed gate leakage based frequency locked timer.

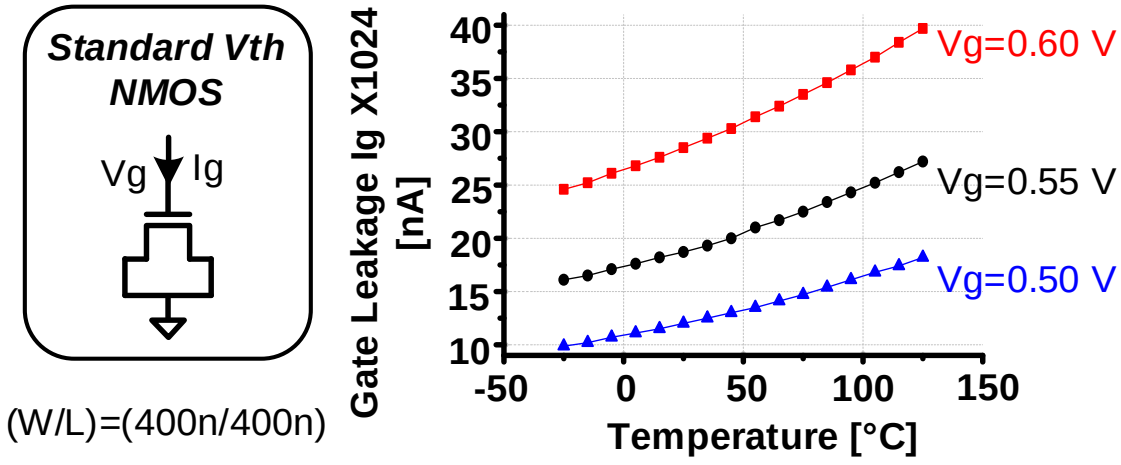


Figure 2.2 Measured gate leakage current across temperature in 55nm CMOS.

2.3 First- and Second-order Temperature Dependance Cancellation

In the proposed design, we use two tuning mechanisms. We cancel the first-order dependence by varying V_2 in a proportional to temperature (PTAT) fashion using a voltage reference with tunable temperature dependence (Fig. 2.1, right). This PTAT reference consists of two PMOS diode stacks, each with different threshold voltages and sizes to create a first-order dependence on temperature (Fig. 2.3b). Switches control the high-Vt PMOS size, which tunes the slope of V_{PTAT} from 0.5 to 0.68%/°C (simulation).

To cancel the second-order dependence, we use a 2T voltage reference, which has intrinsic convex temperature dependence [36] (Fig. 2.3a). However, the convexity of this reference is fixed and is not easily tuned to cancel the second-order dependence of gate leakage. Hence, we leverage the exponential dependence of gate leakage on voltage to provide this tuning mechanism, as follows: First, we remove first-order dependence by tuning the native NMOS and High-Vt PMOS sizes, resulting in $V_{2T} = V_{2T,0}+\alpha(T-T_0)^2$. We then amplify V_{2T}, and a mux structure selects the output voltage $V_{REF} = V_1 = k_{mux}(V_{2T,0}+ \alpha(T-T_0)^2)$ where k_{mux} varies with the mux selection. Note that this does not change the relative magnitude of the convexity of V_1. However, gate leakage I_1 is exponentially dependent on V_1, resulting in $I_1 \propto \exp(\beta k_{mux} V_{2T,0}) \times \exp(\beta k_{mux} \alpha(T- T_0)^2)$. Hence, by changing k_{mux} (i.e., the mux setting) we can modulate the relative magnitude of the convexity of I_1, which allows us to cancel the second-order temperature dependence of the gate leakage. Fig. 2.3c shows simulation results of this approach. After both first- and second-order temperature dependencies are canceled, only third- and higher-order terms remain. Finally, the center frequency (0th order) is adjusted by tuning C_2 in Fig. 2.1.

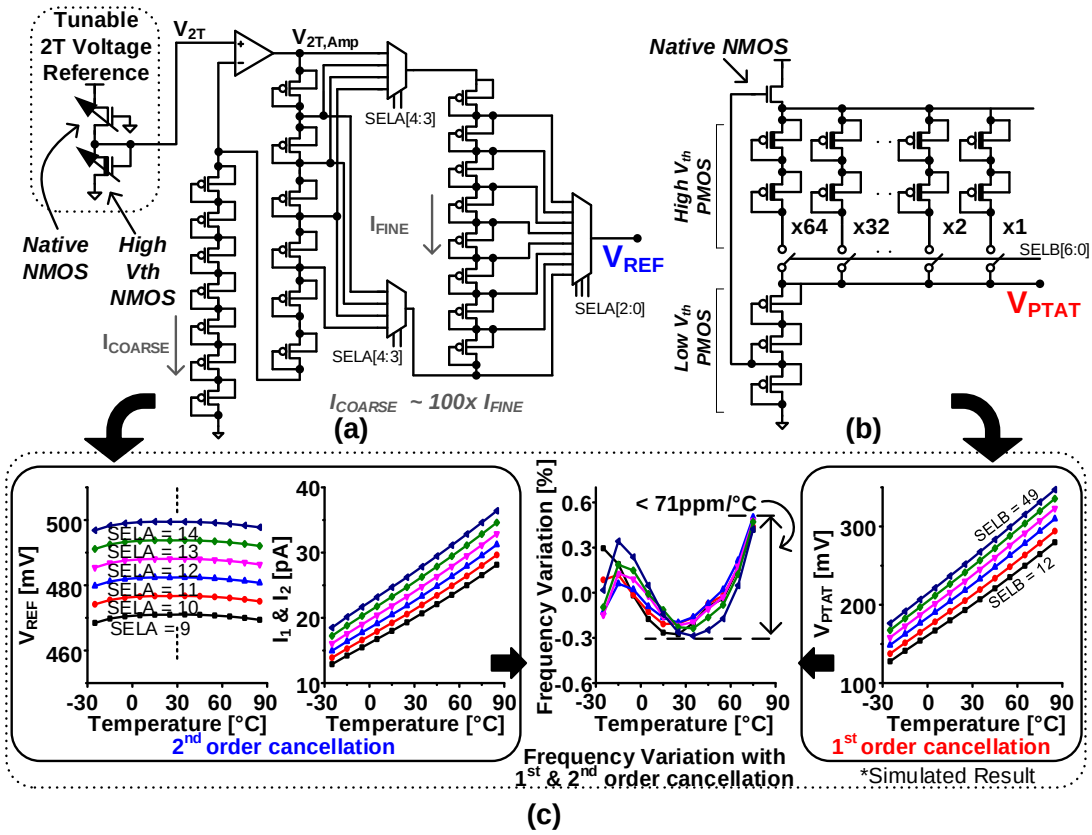


Figure 2.3 (a) Proposed convex voltage generator (b) PTAT voltage generator and (c) simulation result of first- and second-order cancellation.

2.4 VCO with Switched Capacitor Voltage Doubler

Two PMOS devices (P_1 and P_2, Fig. 2.1) implement the current mirror. The devices are high threshold thick-oxide PMOS transistors operating in subthreshold with VDS >5 kT/q, which significantly reduces mismatch between I_1 and I_2. The low power voltage controlled oscillator (VCO) in Fig. 2.1 provides the timer's output frequency and is composed of stacked high threshold inverters to minimize short circuit current (Fig. 2.4). The voltage range of V_{CTRL} across temperature, 0.67-0.9 V, is too narrow and situated at too high a voltage to compensate the VCO frequency across temperature. We double the voltage range and shift it lower using switch capacitor C_4, obtaining V_{CTRL}' with range 0.2-0.65 V. Capacitors C_3-C_5 also generate the dominant pole of the frequency lock scheme.

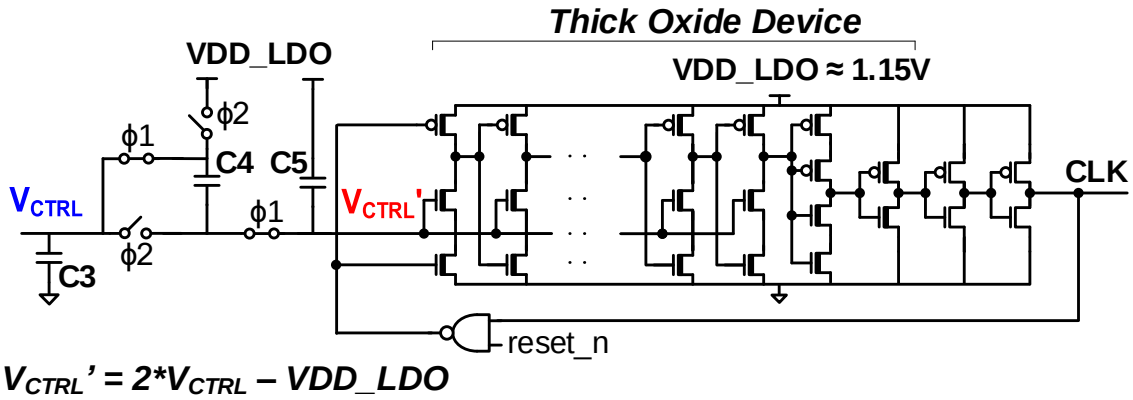


Figure 2.4 Diagram of stacked inverter VCO with switched capacitor voltage doubler.

2.5 Line Sensitivity Enhancement

Gate leakage has high voltage sensitivity, leading to strong frequency dependence on supply voltage in previous gate-leakage-based timers. The proposed design addresses this by placing native NMOS transistors in the convex voltage generator and PTAT voltage generator, enabling low line sensitivity (1.3%/V and 2.2%/V, respectively, simulation). An on-chip LDO further reduces supply voltage dependence while consuming only 18 pW (simulation).

2.6 Measurement Results

The proposed gate-leakage-based timer was implemented in 55nm CMOS (MIFS C55DDC) in 0.057 mm2. Fig. 2.5 show the measured frequency variation from −5 to 95°C for five typical corner dies. Fig. 2.5a gives results with no tuning, while Fig. 2.5b has 2-pt calibration to cancel first-order dependence. Fig. 2.5c uses the proposed second-order cancelation with 3-pt calibration, yielding measured TCs of 175−343 ppm/°C, which is 5× better than first-order cancelation only. The timer consumes 224 pW at 25°C with 90 Hz output frequency; power increases to 1.2 nW at

95°C (Fig. 2.8). Line sensitivity is 0.33−1.29%/V across 1.1−3.3 V supply voltage for the five dies (Fig. 2.6). Fig. 2.11 compares TC, LS, and energy per cycle to those of previous sub-nW timers and resistor-based timers.

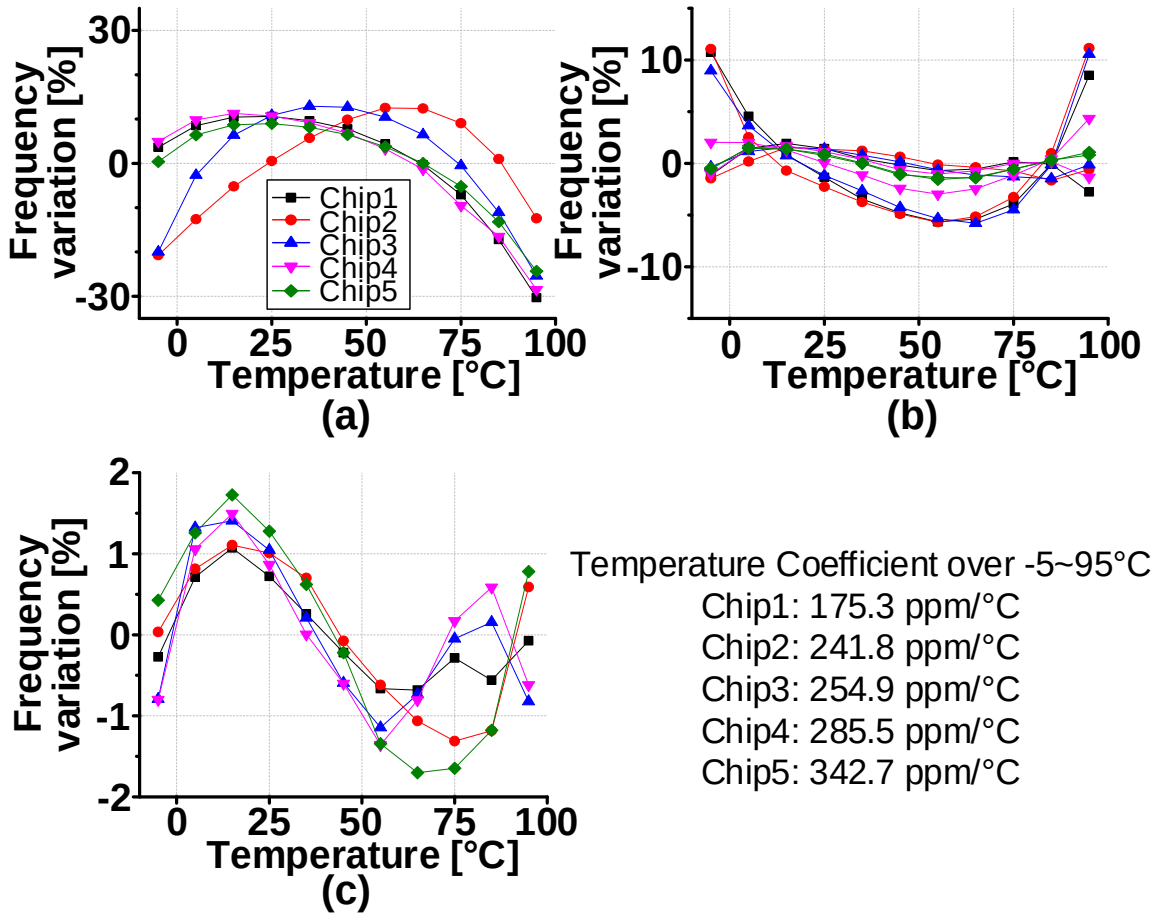


Figure 2.5 Measured frequency variation over temperature: (a) without calibration (b) with 2-pt and (c) with 3-pt calibration.

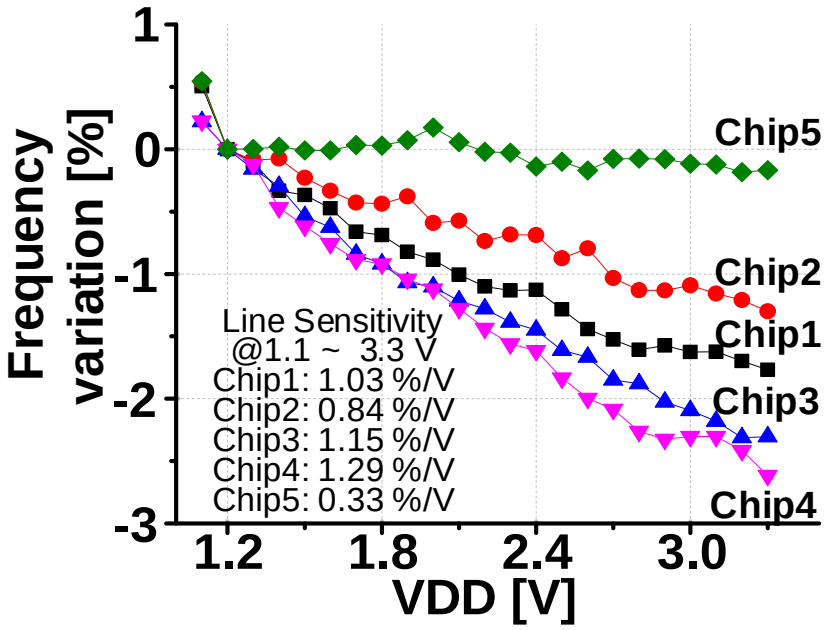


Figure 2.6 Measured line sensitivity of output clock frequency.

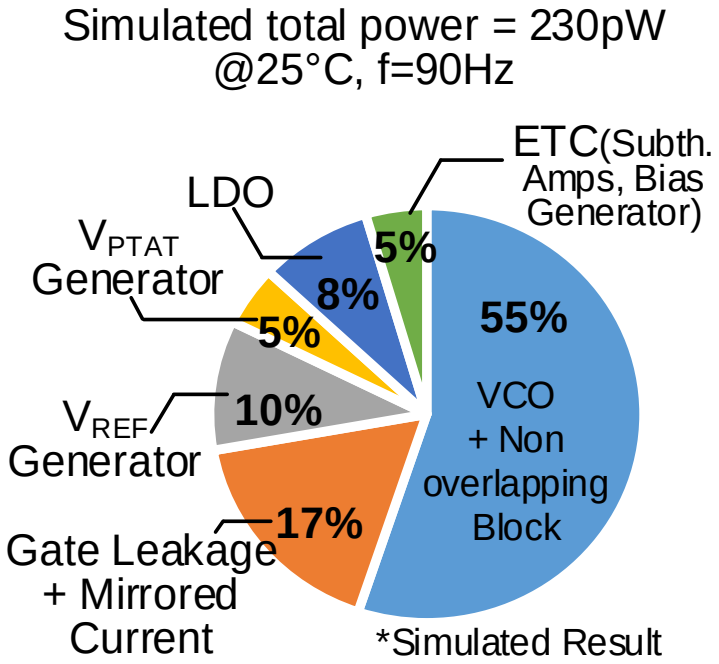


Figure 2.7 Simulated power breakdown of timer.

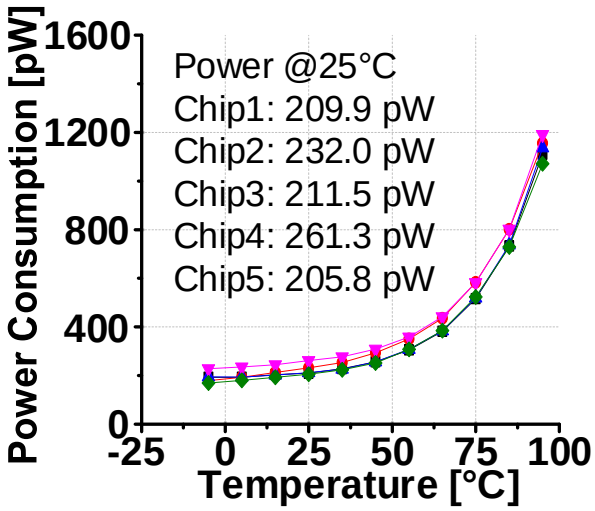


Figure 2.8 Measured power consumption.

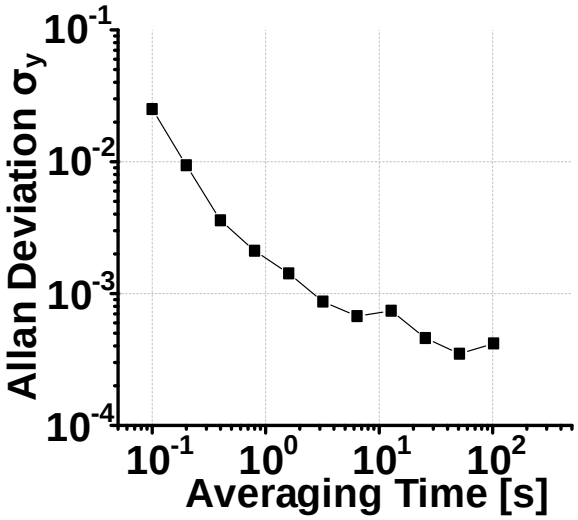


Figure 2.9 Measured Allan deviation.

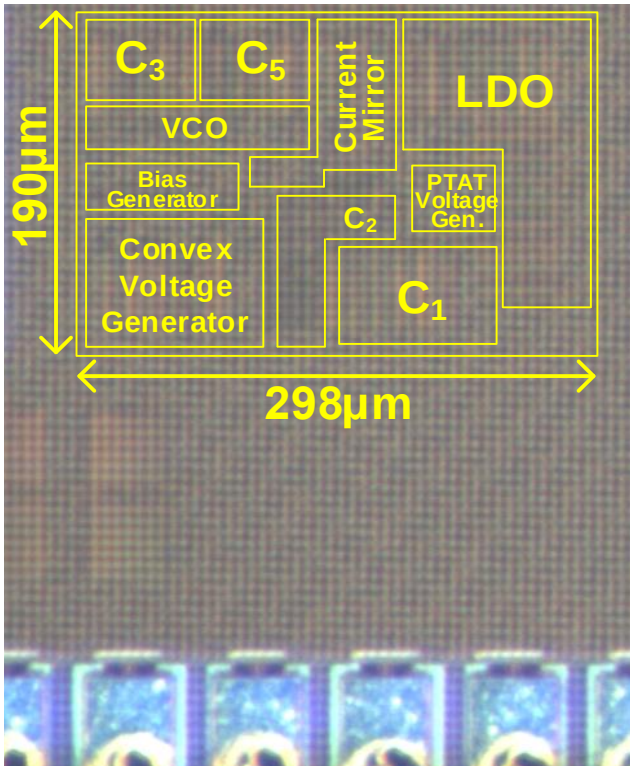


Figure 2.10 Die Photo.

2.7 Conclusion

The proposed timer is Pareto optimal in terms of TC and LS vs. power among the listed works, enabling a new ultra-low power timer design space. Energy per cycle of 2.49 pJ/cycle is comparable to the best reported among the listed works.

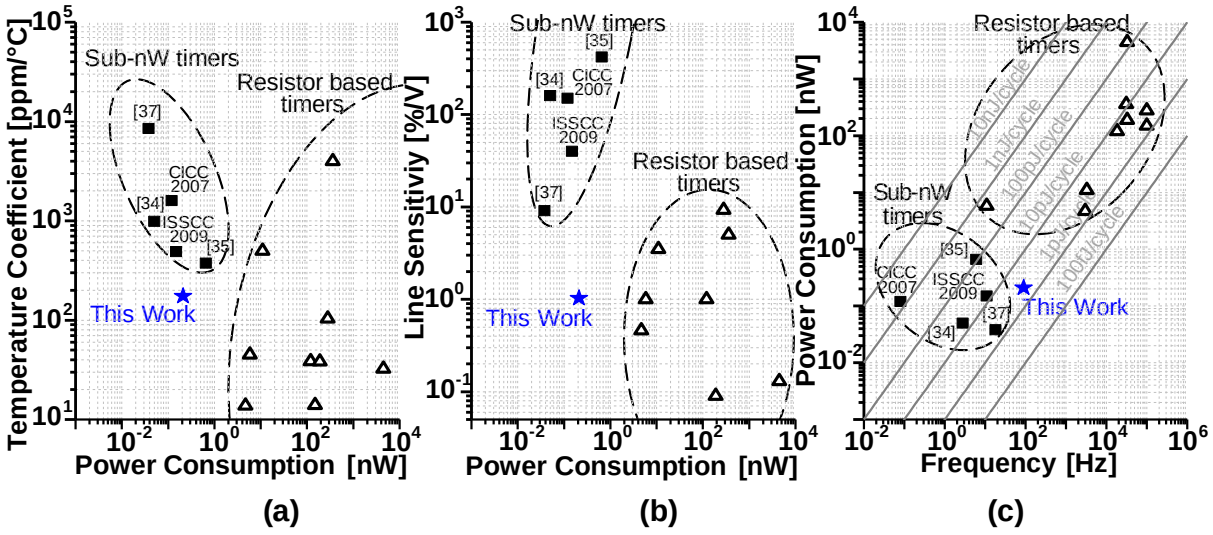


Figure 2.11 Comparison scatter plots with previous work (best-reported dies): (a) temperature coefficient (b) line sensitivity and (c) energy per cycle.

Table 2.1 Comparison table

	This Work	JSSC 2016 [37]	JSSC 2016 [34]	JSSC 2013 [35]	ISSCC 2009 (Y.-S. Lin)	CICC 2007 (Y.-S. Lin)
Process [nm]	**55**	180	65	130	130	130
Supply Voltage [V]	**1.2**	1.8 / 0.6	0.5	0.7 & 1.2	0.6	0.45
Power [pW]	**206~261 (avg 224[a])**	38 / 4.2	30 ~ 50[b] (avg 44[a])	660[c]	150	120
Frequency [Hz]	**90**	18	2.8	6	11	0.08
Temp. Coeff. [ppm/°C]	**175~343 (avg 260[a])**	8500 / 21000	990~2200[b] (avg 1260[a])	375 (31[d])	490	1600
Temp. Range [°C]	**-5 ~ 95**	-30 ~ 60	-40 ~ 60	-20 ~ 60	0 ~ 90	0 ~ 80
Line Sensitivity [%/V]	**0.33~1.29 (avg 0.93[a])**	9.1 / 240	160	420	40	150
Energy/cycle [pJ/cycle]	**2.49**	2.1 / 0.23	15.8	110	13.6	1500
Area [mm^2]	**0.057**	0.18	0.026	0.015	0.02	0.0005
# of reported dies	**5**	1	5	1	1	1

a. Average over multiple samples
b. Calculated from Fig15.
c. Power consumption without a temperature sensor
d. 10 point calibration with a temperature sensor

CHAPTER 3

An Energy-Efficient All-Analog ResNet Accelerator

3.1 Introduction

The development of machine learning hardware, along with deep learning algorithms, has allowed significant breakthroughs in a number of areas, including image classification, motion detection, and speech recognition. The proliferation of IoT devices has increased the demand of machine learning accelerators designed for edge computing and has reinforced the importance of energy efficiency of such accelerators. Most notably, the vast amount of energy consumed by frequent memory access [38]to load data (weights, features, and parameters) during inference must be reduced to meet the limited energy budget of edge computing. Recently, various in-memory and mixed-signal approaches [39]-[47] have attempted to address this issue and reduce energy consumption by replacing frequent memory read accesses and digital computations with in-memory and analog computations. In addition, recent studies have proposed modified training methods for mixed-signal-based accelerators with low bit precision [48] and in-situ methods for minimizing accuracy degradation due to process variation [49]. However, all of these approaches include digital-to-analog converters (DAC) and analog-to-digital converters (ADC) at the front and back of each hidden layer to store and broadcast features in digital representation [41]-[48] (Fig. 3.1a). Further, they implement the required nonlinear (NL) functions in the digital domain

[45]. The DACs/ADCs are energy bottlenecks, especially when high precision of weight or activation is required. The energy overhead gets even worse when implementing deep convolutional neural networks (CNNs) [50] that consist of many layers. Hence, prior in-memory or mixed-signal designs have been largely restricted to simple shallow networks. Other approaches implementing binarized CNN (BNN) [51] in a mixed-signal domain have been proposed using XNOR [52]-[53] for multiplication and charge-sharing techniques for addition [54]-[56]. BNN has the benefit of reducing computation complexity to a single bit, and as such, these mixed-signal accelerators reduce the DAC and ADC energy overhead since they only have a single bit precision for both weights and activations. However, the BNN works well only for moderately sized networks (e.g., AlexNet [57] and nine-layer networks with 328 KB [55]/ 295 KB [56] of weights) and has a critical limit on the scalability to support very large networks that are difficult to train with binary weights.

To address these challenges, this chapter introduces the first multi-layer (total, 18), all analog NN accelerator in 28-nm CMOS with 32.2 KB of weight storage, implementing not only convolution but also NL function, storage of value for subsequence use, and routing between layers all in the analog domain. Weights and activations are in 4-bit and 3-7-bit precision, respectively, thereby offering significantly better precision compared to BNNs. This work makes the following contributions:

- Energy-efficient all-analog structure without any DAC or ADC overhead between hidden layers (Fig. 3.1b): (1) activations are represented in the pulse-width domain with 3-7 bit precision; (2) convolution is performed using analog integration in the charge domain on the bit-lines to convert to voltage using a charge integrating amplifier to represent the activation output; (3) activation values are stored and broadcasted in the voltage domain by sample-

and-hold (S/H) buffers; (4) NL functions (such as sigmoid and recta-linear) are performed while transforming the voltage domain back to the pulse-width domain using a simple comparator and a purposefully shaped voltage ramp; and (5) resulting output pulses are then routed to subsequent layers for their activation or to the final inference outputs.

• Fully pipelined structures (18 layers) for a fast throughput rate of 325K image/sec with CIFAR-10 and SVHN datasets [58]-[59], which is more than 830× faster than conventional mixed-signal approaches [55]-[56].

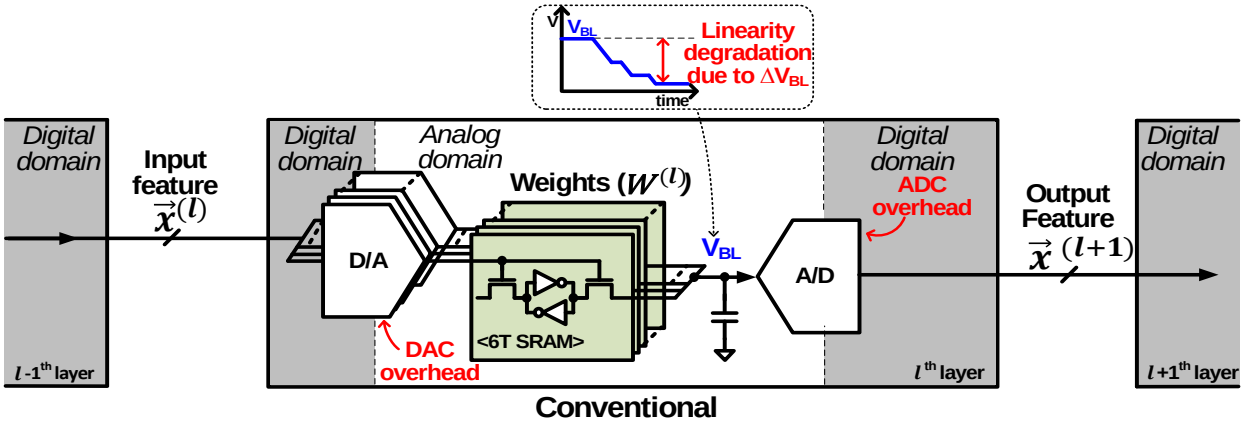
• Energy consumption per inference of 1.2 μJ over CIFAR10 and SVHN datasets, which is 3× lower than prior mixed-signal approaches [55]-[56].

• The first implementations of a deep ResNet [60]-[61] fully in the analog domain including convolution, batch normalization (BN), rectified linear units (ReLU), average pooling and fully-connect (FC) layer.

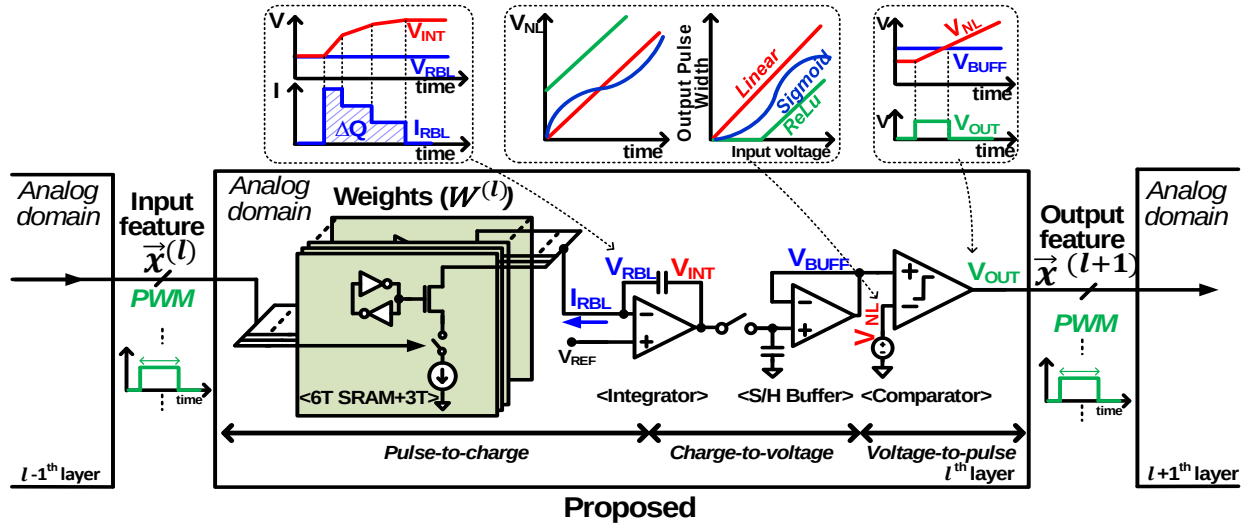
• Linearity enhancement of analog convolution by maintaining a constant read bit line (RBL) voltage.

• Evaluation of linearity, noise, effective bit precision, accuracy, and energy efficiency of the proposed accelerator using transistor-level SPICE simulation and Matlab.

The rest of the chapter describes the proposed AA-ResNet accelerator and implementation details and provides an evaluation and conclusion.



(a)



(b)

Figure 3.1 Comparison of (a) conventional approaches with in-memory/mixed-signal computing and (b) proposed all-analog approaches.

3.2 Main Concept of the AA-ResNet

In this section, we first describe the main concept of the AA-ResNet with analog operations of a single layer. Then we present the overall architecture of the AA-ResNet accelerator.

3.2.1 Convolution in Pulse-to-Charge Domain

For every layer, the input activations are represented in the pulse-width domain as shown in Eq (1):

$$x_{i,j,d}^{(l)} = \Delta t_{i,j,d}^{(l)} \tag{1}$$

where $x_{i,j,d}^{(l)}$ is the input activation value of the *i*-th row, *j*-th column, *d*-th depth of *l*-th layer, and $\Delta t_{i,j,d}^{(l)}$ is the corresponding pulse width.

In the proposed analog convolution, the weights of kernels, $w_{i,j,d,k}^{(l)}$, represented by Eq (2), are the product of the 4-bit sign-and-magnitude digital values $W_{i,j,d,k}^{(l)}$ stored in the 6T SRAM arrays and the weight control DC current $I_{LSB}^{(l)}$ that charges capacitors in the analog integrators (see Fig. 3.2). In Eq (2), *k* represents the kernel index.

$$w_{i,j,d,k}^{(l)} = W_{i,j,d,k}^{(l)} \cdot I_{LSB}^{(l)} \tag{2}$$

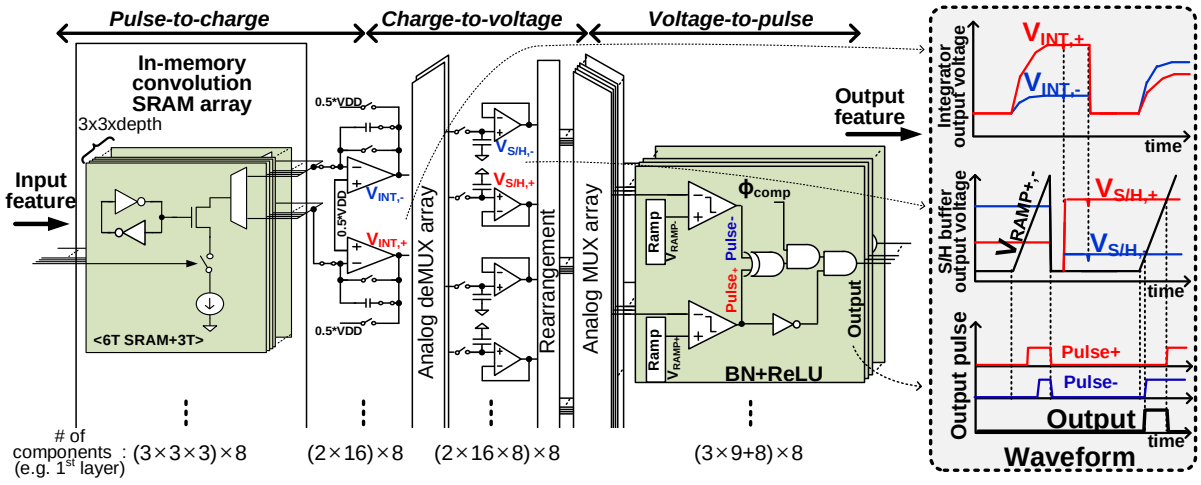


Figure 3.2 A single layer(L1) structure of proposed AA-ResNet accelerator.

The input activation value (i.e., pulse width) determines the on-time of the DC current $I_{LSB}^{(l)}$ in Eq (2). The accumulated charge is proportional to the DC current level $I_{LSB}^{(l)}$, the stored weight value, and the time period the current is turned on (i.e., the input pulse width). Therefore, the accumulated charge represents the multiplication of the input and weight. Multiple (e.g., 3×3×d as shown in Fig. 3.2) wires are shorted together at the input of an analog integrator, which merges all of the charge flowing through the tied wires. Thus, the total integrated charge is equivalent to the convolution output, as shown in Eq (3a) and (3b). Because the DC current has only a single polarity (pull down), a pair of the integrators integrate charge for the positive and negative convolution value separately; Eq (3a) and (3b).

$$Q_{i,j,k}^{+(l)} = \sum_d \sum_{j'=1,2,3} \sum_{i'=1,2,3} \left| W_{i',j',d,k}^{(l)} \right| \cdot I_{LSB}^{(l)} \cdot \Delta t_{i+i'-2,j+j'-2,d}^{(l)} \quad (3a)$$

$$sign\left(W_{i',j',d,k}^{(n)}\right) \geq 0$$

$$Q_{i,j,k}^{-(l)} = \sum_d \sum_{j'=1,2,3} \sum_{i'=1,2,3} \left| W_{i',j',d,k}^{(l)} \right| \cdot I_{LSB}^{(l)} \cdot \Delta t_{i+i'-2,j+j'-2,d}^{(l)} \quad (3b)$$

$$sign\left(W_{i',j',d,k}^{(n)}\right) < 0$$

$$y_{i,j,k}^{(l)} = Q_{i,j,k}^{+(l)} - Q_{i,j,k}^{-(l)} \quad (3c)$$

The final convolution result $y_{i,j,k}^{(l)}$ is obtained by the difference of charge accumulated in the capacitors of a pair of integrators as in Eq (3c). The positive and negative convolution results $Q_{i,j,k}^{+(l)}$ and $Q_{i,j,k}^{-(l)}$ are separately stored in the form of electric charge or voltage, while subtraction for $y_{i,j,k}^{(l)}$ is performed in the pulse domain after voltage-to-pulse conversion as explained in section 3.2.3.

In the proposed design, we also implement the addition of convolution results with the bypassed input activations. This feedforward shortcut is the key idea of ResNet [60]-[61] that improves accuracy of very deep networks through residual learning. The shortcut connection of ResNet, $y_{i,j,k}^{(l)} = \vec{w}^{(l)} \cdot \vec{x}^{(l)} + x_{i,j,k}^{(l-1)}$ is also calculated in the charge domain by tying the 6T SRAM arrays and current path for $x_{i,j,k}^{(l-1)}$ to the input of the integrators. The detailed circuit implementation of residual learning is discussed in section 3.3.1.

3.2.2 Sampling and Holding in Charge-to-Pulse Domain

The convolution results are stored in the analog (charge) domain and broadcasted at the proper timing to the NL function blocks and the next layers. The charge on the capacitors (Fig. 3.1) for the integrator pairs $Q_{i,j,k}^{+(l)}$ and $Q_{i,j,k}^{-(l)}$ can be directly converted into the voltage level $V_{INT\,i,j,k}^{+(l)}$ and $V_{INT\,i,j,k}^{-(l)}$, respectively, as in Eq (4a) and (4b), since the analog integrators hold the bottom plate of the integrator capacitors to a constant voltage of $\frac{1}{2} \cdot V_{DD}$. The voltages $V_{INT\,i,j,k}^{+(l)}$ *and* $V_{INT\,i,j,k}^{-(l)}$ are sampled when the charge integration is complete. The buffers hold the sampled voltage to be fed into the NL block (Fig. 3.2).

$$V_{INT\,i,j,k}^{+(l)} = \frac{1}{C} \cdot Q_{i,j,k}^{+(l)} + \frac{1}{2} \cdot V_{DD} \quad (4a)$$

$$V_{INT\,i,j,k}^{-(l)} = \frac{1}{C} \cdot Q_{i,j,k}^{-(l)} + \frac{1}{2} \cdot V_{DD} \quad (4b)$$

3.2.3 NL function in Voltage-to-Pulse Domain

In the proposed approach, NL function is performed in the analog domain, converting the convolution output from voltage to the pulse-width domain (Fig. 3.1). In the NL block, a ramp voltage, which monotonously rises in time, is compared with $V_{INT}{}_{i,j,k}^{+(l)}$ and $V_{INT}{}_{i,j,k}^{-(l)}$ using a comparator. The binary output of the comparator encodes the NL function value in the pulse-width domain (Fig. 3.3).

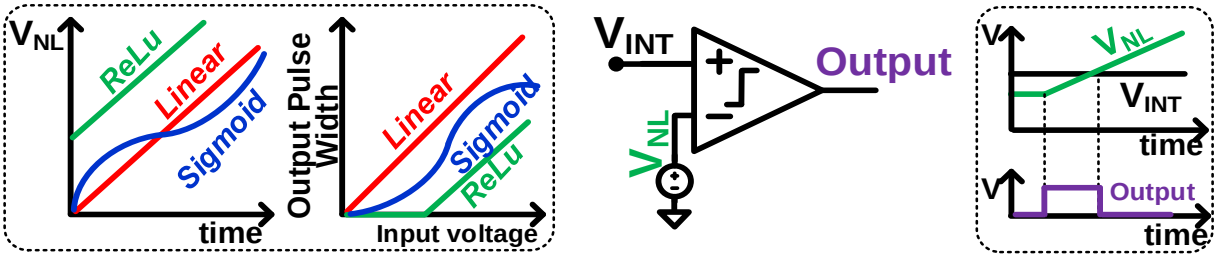


Figure 3.3 Time domain mapping of non-linear voltage.

Various non-linear functions such as sigmoid and ReLU can be realized by properly shaping the ramp voltage that monotonously rises with a non-constant slope. In the proposed design, ReLU with batch normalization (BN) is implemented using the ramping voltages $V_{RAMP,k}^{+}$ and $V_{RAMP,k}^{-}$ generated by the ramp voltage generator structure discussed in section 3.3.3.

$$V_{RAMP,k}^{+}\left(\Delta t_{i,j,k}^{+(l)}\right) = V_{INT}{}_{i,j,k}^{+(l)} \tag{5a}$$

$$V_{RAMP,k}^{-}\left(\Delta t_{i,j,k}^{-(l)}\right) = V_{INT}{}_{i,j,k}^{-(l)} \tag{5b}$$

We define $\Delta t_{i,j,k}^{+(l)}$ and $\Delta t_{i,j,k}^{-(l)}$ as the time between the start of the comparison and the point where $V_{RAMP,k}^{+}$ ($V_{RAMP,k}^{-}$) exceeds $V_{INT}{}_{i,j,k}^{+(l)}$ ($V_{INT}{}_{i,j,k}^{-(l)}$).

$$\Delta t_{i,j,k}^{(l+1)} = \begin{cases} \Delta t_{i,j,k}^{+(l)} - \Delta t_{i,j,k}^{-(l)}, & \Delta t_{i,j,k}^{+(l)} \geq \Delta t_{i,j,k}^{-(l)} \\ 0, & \Delta t_{i,j,k}^{+(l)} < \Delta t_{i,j,k}^{-(l)} \end{cases} \quad (6)$$

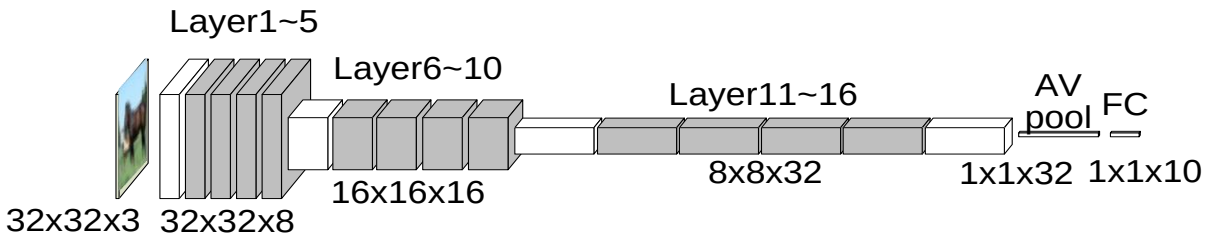
Using the logic gates in Fig. 3.1, the final output pulse width $\Delta t_{i,j,k}^{(l+1)}$ is obtained by the difference between $\Delta t_{i,j,k}^{+(l)}$ and $\Delta t_{i,j,k}^{-(l)}$ as in Eq (6), realizing ReLU. Note that until this point, the negative and positive convolution results are separately maintained. This final output pulse width $\Delta t_{i,j,k}^{(l+1)}$ is streamed into the next layer representing the input activation $x_{i,j,k}^{(l+1)}$ as in Eq (1).

3.2.4 Overall Structure

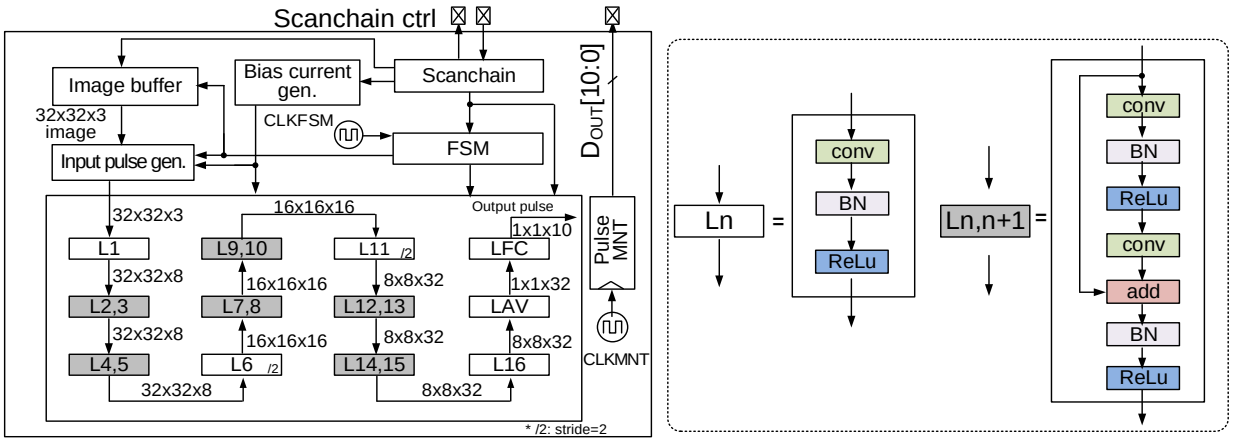
The proposed accelerator implements a modified ResNet [58]-[59] that consists of 16 convolution + BN + ReLU layers, an average pooling layer, and a fully connected (FC) layer as shown in Fig. 3.4a. The layers colored in grey in Fig. 3.4a have feedforward shortcut connections that are unique to ResNets.

The overall accelerator architecture is shown in Fig. 3.4b. The input image of 32×32×3(RGB) pixels is loaded to image buffers implemented with 6912 bytes of compiled SRAM. The input image buffers hold two images for ping-pong buffering. An input image is divided into 64 sub-images of 3× 18 × 3 pixels, and each of them is fetched into the input pulse generator in each 48-ns operation cycle. From the input pulse generator to the last FC layer, the entire ResNet datapath of 19 layers is fully pipelined to generate classification output at a very high throughput of one image per 64 cycles (64×48 ns) or 325,520 images/s, which is 830× faster

than reported in [55]-[56]. The final output pulses from ten channels of the FC layer are monitored by the TDC-based pulse monitor block to find the channel number with the longest pulse width. The channel index of the longest pulse width is the inferred image class number processed by the AA-ResNet. Trained weights and parameters are loaded to the accelerator via a scan-chain block (registers) before inference starts. The 48-ns operation cycle is composed of 24 clock cycles generated by an interal 500-MHz on-chip ring oscillator (included in the power estimation).



(a)



(b)

Figure 3.4 (a) A diagram of modified ResNet and (b) overall hardware architecture.

3.3 AA-ResNet Circuit Implementation

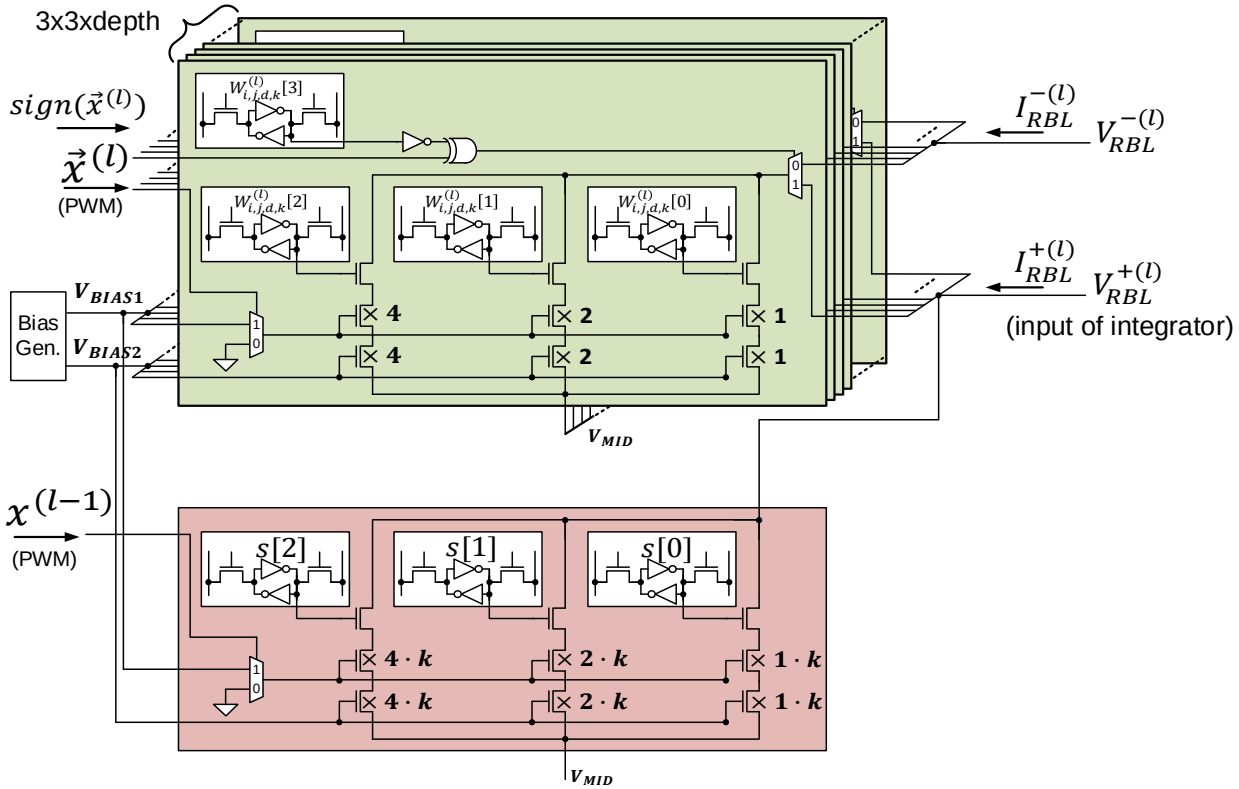
This section presents the implementation details of various components, including the in-memory convolution SRAM array cells, integrators, S/H buffers, and NL function blocks.

3.3.1 In-memory Convolution SRAM Array Cells

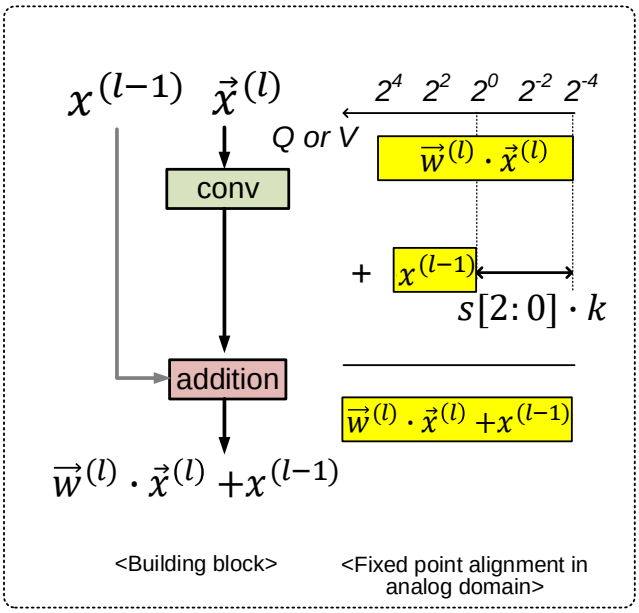
Fig. 3.5 shows in-memory convolution SRAM array cells with 3T-readout buffers that generate current proportional to the magnitude of the weight. Stacked NMOS transistors offer tolerance to V_{DS} variation, generating linear currents. The MSB of the weight (sign) selects one of the current-conducting paths.

The average pooling and FC layers are also implemented with in-memory SRAM array cells and integrators. In the average pooling layer, constant weights are stored in arrays.

The convolution layers with residual learning (i.e., L2,3 and L4,5 in Fig. 3.4) require an addition block, as shown at the bottom of Fig. 3.5. In the addition block, instead of weight, a scaling factor, $s[2:0]$, is stored in the SRAM cells. This is for aligning the fixed point of the convolution results and the input of the previous layer. Although activations are in the analog domain, we must consider their effective fixed-point representation (Fig. 3.5). The scaling factors vary over different training datasets and layers.



(a)



(b)

Figure 3.5 (a) Circuit structure of in-memory convolution SRAM array cells and (b) diagram of the shortcut connection.

3.3.2 Analog Integrators

The analog integrators are composed of a complementary folded cascode amplifier [62] with an auto-zeroing scheme [63] to cancel offset (Fig. 3.6). The amplifier holds the RBL voltage constant at 0.5 V = V_{DD}/2, and this further improves the linearity of the 3T-readout buffer (Figs. 3.2 & 3.5) by holding V_{DS} of the NMOS devices constant. Linearity does degrade as the output voltage approaches the upper headroom of the amplifier, which is ~0.8 V. Therefore, in the proposed design, we only use an integrator output range of 0.5–0.75 V.

To fully utilize the output voltage range of the integrator, $I_{LSB}^{(l)}$ in the in-memory convolution SRAM array block is determined based on the maximum value of the sum of the positive/negative parts of each convolution layer during off-line training. In addition, scaling of $I_{LSB}^{(l)}$ includes compensation of the final pulse output $\Delta t_{i,j,k}^{(l)}$ scaling from the previous layer for subtraction of two pulse widths, $\Delta t_{i,j,k}^{+(l-1)}$ and $\Delta t_{i,j,k}^{-(l-1)}$ (Eq 6). Note that the reduced pulse width can be recovered in the pulse-to-charge domain by scaling $I_{LSB}^{(l)}$; however, the dynamic range issue still remains in this step. The impact of pulse-width reduction and limited dynamic range is discussed in section 3.4.3.

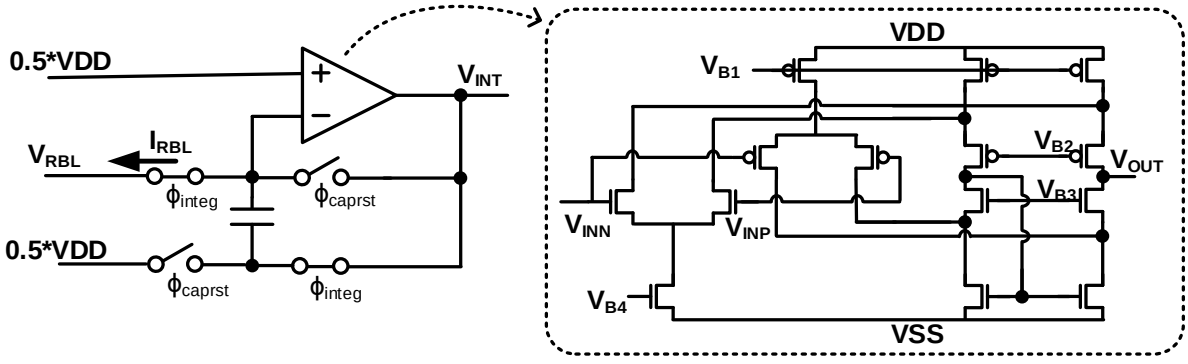
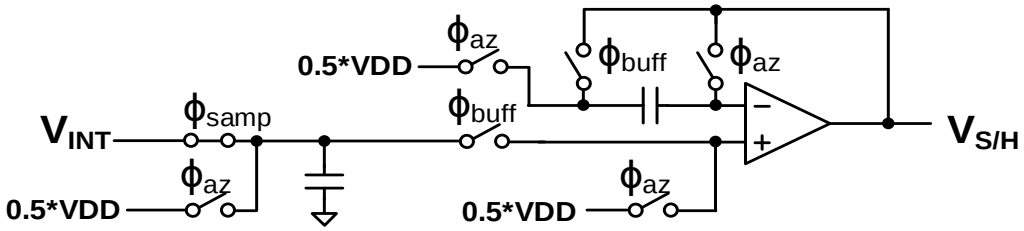


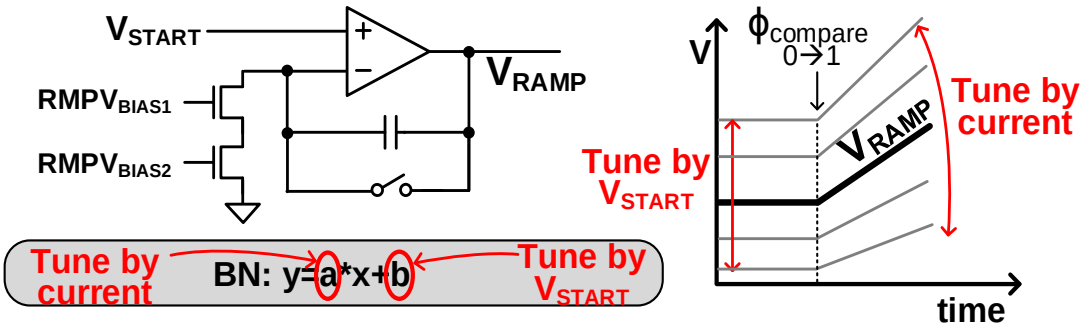
Figure 3.6 Circuit structure of analog integrator.

3.3.3 S/H buffers and Ramp Voltage Generators for BN and ReLU

The S/H buffer samples the integrator output voltage on a capacitor after the integrated voltage is settled. (Fig. 3.7a) The sampled voltage is buffered by an analog buffer for several operation cycles since the voltage (convolution result) can be reused for multiple convolution operations. An FSM controls both the analog deMUX arrays between the integrator pairs and S/H buffers and the analog MUX arrays between the S/H buffers and comparator pairs (Fig. 3.2).



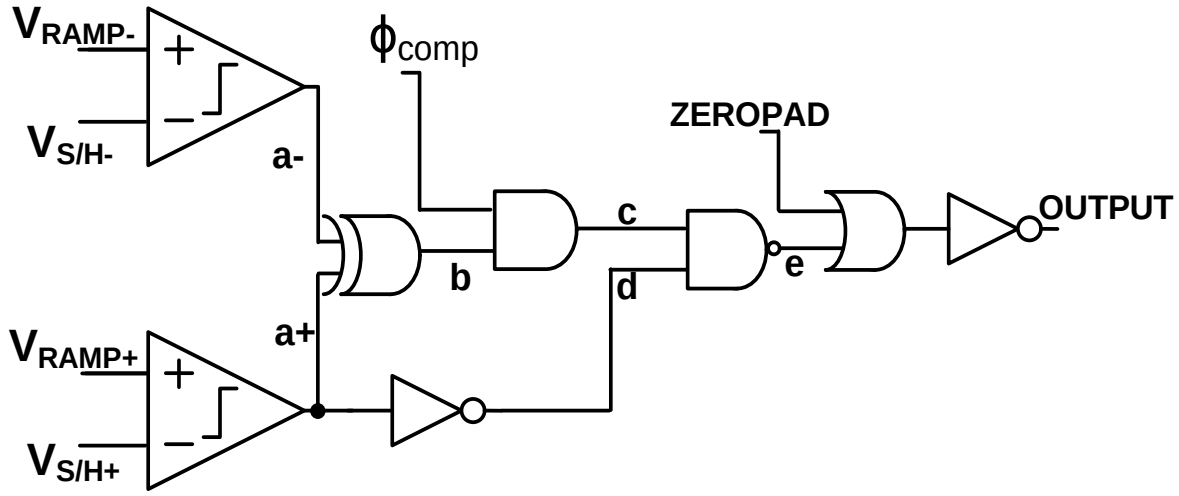
(a)



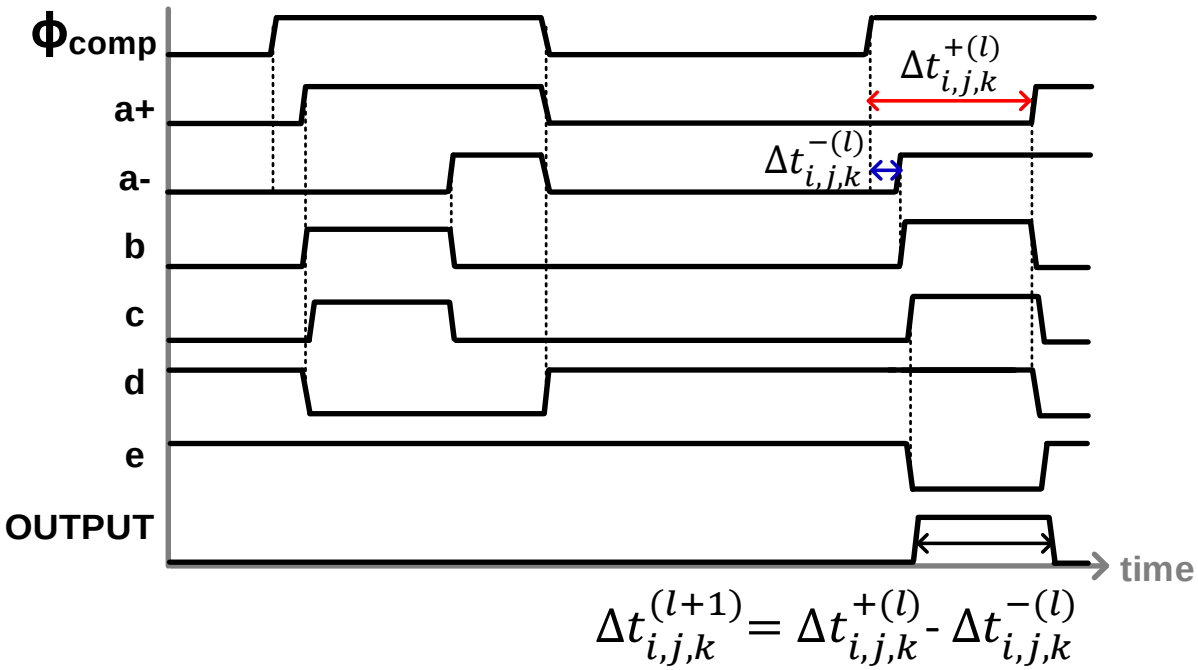
(b)

Figure 3.7 Circuit structure of (a) S/H buffer and (b) ramp voltage generator for BN and ReLU with tunable gain and offset.

The ramp voltage generator in Fig. 3.7b, used for non-linear voltage (BN+ReLU) generation, has a similar structure as the analog integrator except is employs a 2T-NMOS-based always-on current rather than a 3T-NMOS-based readout buffer. The slope of the ramp signal is determined by the DC current level, which corresponds to the gain of the BN function. In addition, the bias of BN can be tuned with $V_{START,}$ which is the starting voltage level of V_{RAMP}. A pair of ramp voltage generators share the DC current level (BN gain) but have separate bias voltages, V_{START}^{+} and $V_{START}^{-} = 1\text{V} - V_{START}^{+}$, allowing for identical BN to be applied separately to positive and negative convolution results. V_{START}^{+} and V_{START}^{-} are generated from a diode-stacked ladder. Continuous comparators evaluate V_{RAMP}^{+} / V_{RAMP}^{-} against buffered voltage pairs, generating a rising edge when V_{RAMP}^{+} and V_{RAMP}^{-} cross the buffered voltage pairs. Finally, ReLU is performed by passing these pulses through the logic shown in Fig. 3.8a. In the waveform in Fig. 3.8b, the final output pulse is generated when $\Delta t_{i,j,k}^{+(l)} \geq \Delta t_{i,j,k}^{-(l)}$ with the pulse width of $\Delta t_{i,j,k}^{(l+1)} = \Delta t_{i,j,k}^{+(l)} - \Delta t_{i,j,k}^{-(l)}$. If this inequality is not met, the output pulse width is zero (ReLU).



(a)



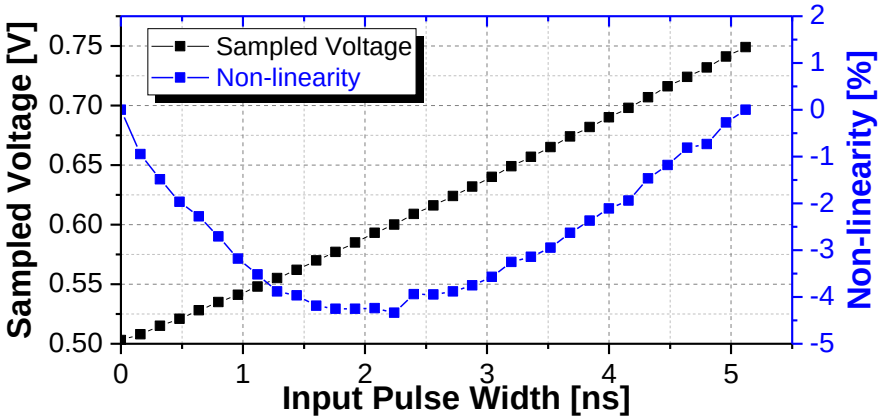
(b)

Figure 3.8 (a) Proposed ReLU structure and (b) output waveform.

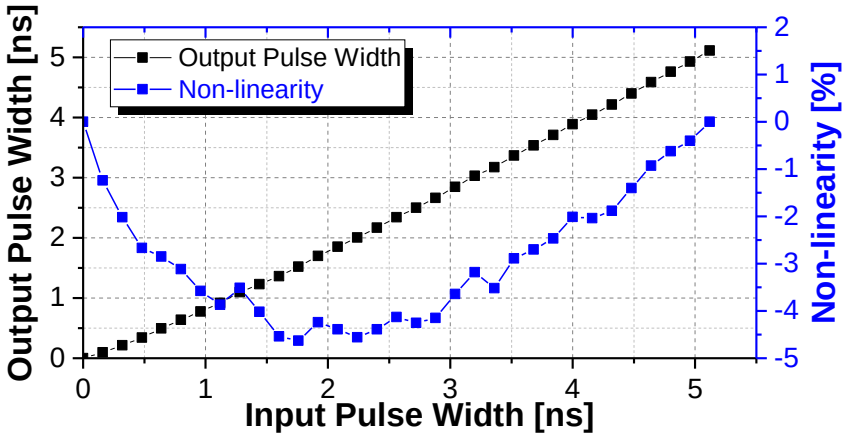
3.4 Performance Evaluation

3.4.1 Linearity of a Single Hidden Layer

A single layer including in-memory convolution SRAM arrays, integrators, S/H buffers, ramp generators, and the BN+ReLU block is simulated in transistor-level SPICE simulation. Nonlinearity is measured by sweeping the input pulse width of the parallel channels of a single convolution kernel from 0 to 5.12 ns with a 160 ps time step (Figs 3.9 & 3.10).



(a)



(b)

Figure 3.9 Linearity simulation result of (a) input pulse to voltage and (b) input pulse to output pulse.

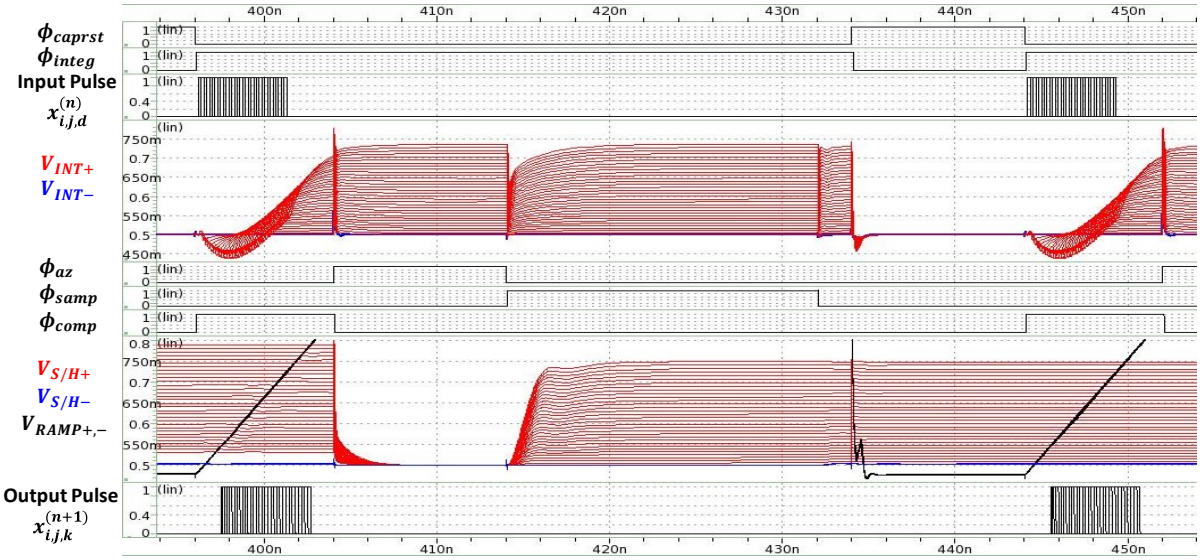


Figure 3.10 Transient simulation waveform with input pulse width sweep from 0 to 5.12 ns with 160 ps time step.

Thanks to the constant RBL voltage and stacked readout buffers, the nonlinearity incurred in the convolution SRAM arrays and analog integrators is limited to 4.5% in the worst case (Fig. 3.9a). This percentage is calculated by normalizing the difference between the output value and its ideal value to the full output range (250 mV). The nonlinearity in the voltage-to-output pulse domain is negligible compared to the nonlinearity from integrators, and hence the total nonlinearity in a single layer (input pulse to output pulse) is < 4.8%. This nonlinearity characteristic can be monitored outside of the chip and modeled in the offline training to further improve accuracy.

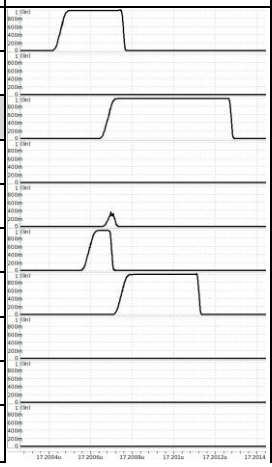
3.4.2 Multi-Layer Verification

Multi-level operation is verified by co-simulation using transistor-level SPICE simulation of the analog layers and VCS for the synthesized logics simultaneously. A sample image is loaded to the input image buffer, and the output pulse width of each layer is compared with output features

from Matlab to ensure correct functionality. In Fig. 3.11, the transistor-level SPICE simulation results of the average pooling and FC layer match well with the output feature obtained from Matlab.

After validating the individual components in SPICE simulations, we employ Verilog-A models of analog components to reduce the simulation time and fully verify the wiring and timing of the full system with all of the layers.

(a)

Class	SPICE sim. waveform	Pulse width [ps] (SPICE)	Normalized pulse width (SPICE)	Normalized output feature (Matlab)
0		302.81	0.5117	0.4943
9		0.00	0.0000	0.0000
8		**591.83**	**1.0000**	**1.0000**
7		0.00	0.0000	0.0000
6		0.00	0.0000	0.0000
5		109.10	0.1843	0.1624
4		373.00	0.6302	0.6347
3		0.00	0.0000	0.0000
2		0.00	0.0000	0.0000
1		0.00	0.0000	0.0000

(b)

Figure 3.11 (a) A sample SVHN image and (b) Transistor-level SPICE simulation result of average pooling + FC layers and comparison with Matlab results.

3.4.3 Analysis on Noise and Dynamic Range

The effective bit-precision of activations in the pulse-width domain can be calculated from the signal voltage range and noise level observed from the analog blocks (or from the pulse-width range and jitter). The noise statistics are estimated using transient noise simulations of a transistor-level SPICE netlist (Table 3.1).

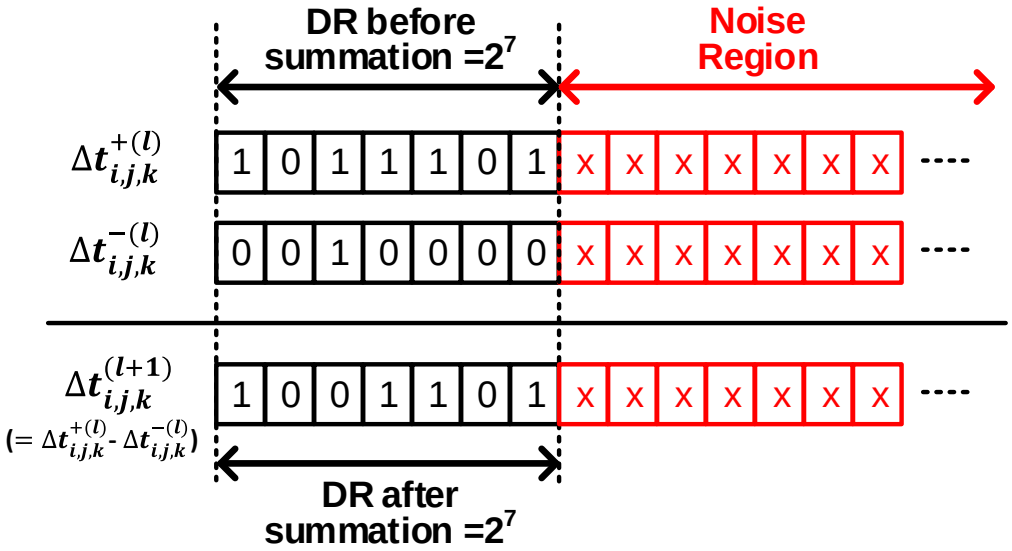
From Table 3.1, an effective bit precision is calculated to be $\log_2\left(\frac{250mV}{1.6815mV}\right) \approx 7.22$b, which ignores the dynamic range reduction after summation (subtraction) of the positive and negative convolution values.

Table 3.1 RMS noise from transient noise simulation.

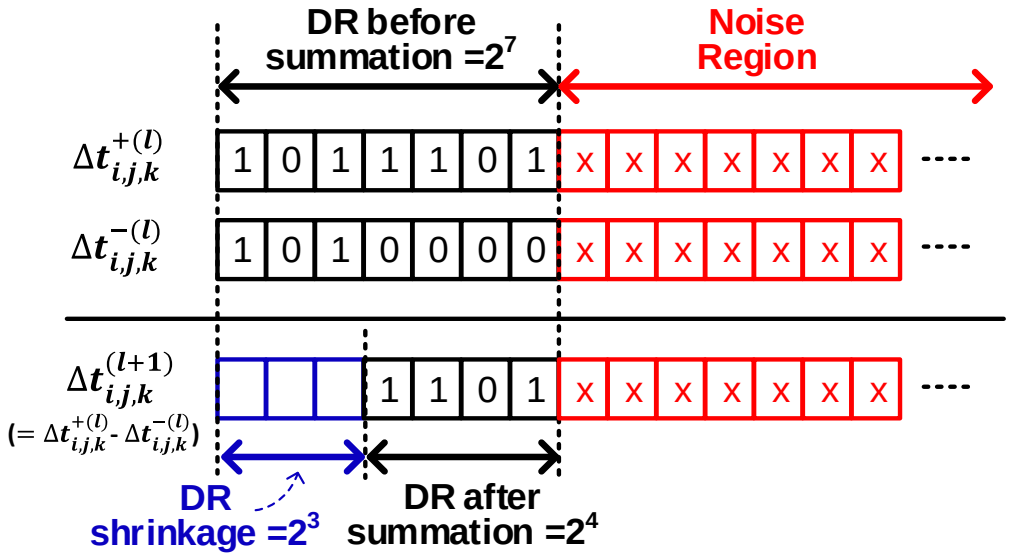
Noise source	RMS noise [mV] (Voltage domain)	RMS Jitter [ps] (Time domain)
SRAM array + Integrator	0.8838	18.1002
S/H buffer	0.7976	16.3345
Ramp generator	1.0787	22.0921
Comparator	0.4966	10.1700
Total	**1.6815**	**34.4373**

In Fig. 3.12, an output pulse width of a layer $\Delta t_{i,j,k}^{(l+1)} (= \Delta t_{i,j,k}^{+(l)} - \Delta t_{i,j,k}^{-(l)})$ is represented in binary format to visualize the effective bit precision of the analog values. For the case shown in Fig. 3.12a, there is no additional dynamic range loss except the loss from noise. On the other hand, the case shown in Fig. 3.12b incurs additional dynamic range loss after summation when $\Delta t_{i,j,k}^{+(l)} - \Delta t_{i,j,k}^{-(l)}$ is small. During training, the dynamic range reduction after summation of the positive and negative parts is modeled for each layer to minimize classification accuracy loss. Overall, the

dynamic range reduction varies from 2^0 to 2^4 among different layers, which corresponds to an effective bit loss of 0-4 bits. Including noise level and value reduction together, the effective bit precisions of activations vary from 3 to 7 bits among different layers.



(a)



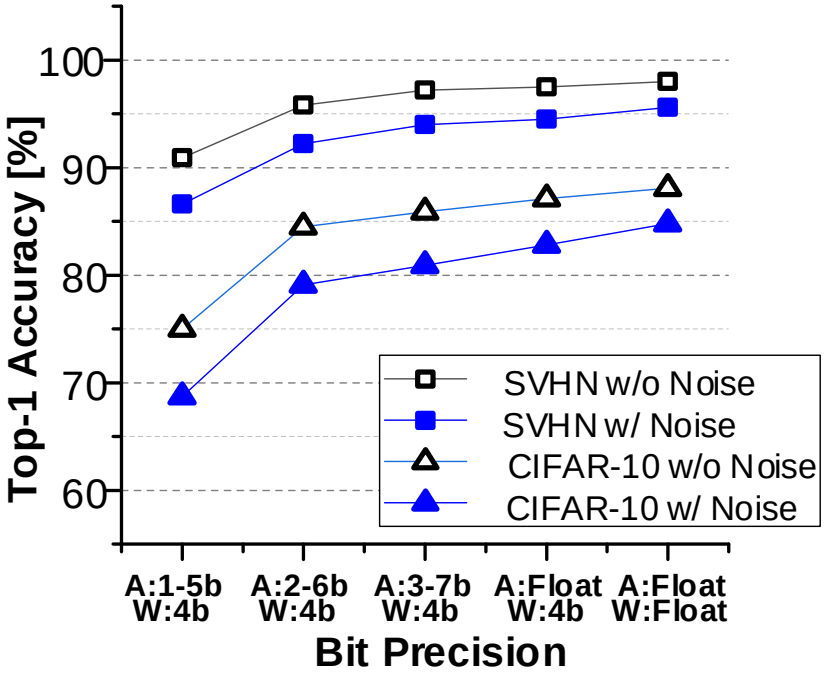
(b)

Figure 3.12 Diagram of effective bit precision activations (a) without dynamic shrinkage (b) with dynamic shrinkage = 2^3 after summation.

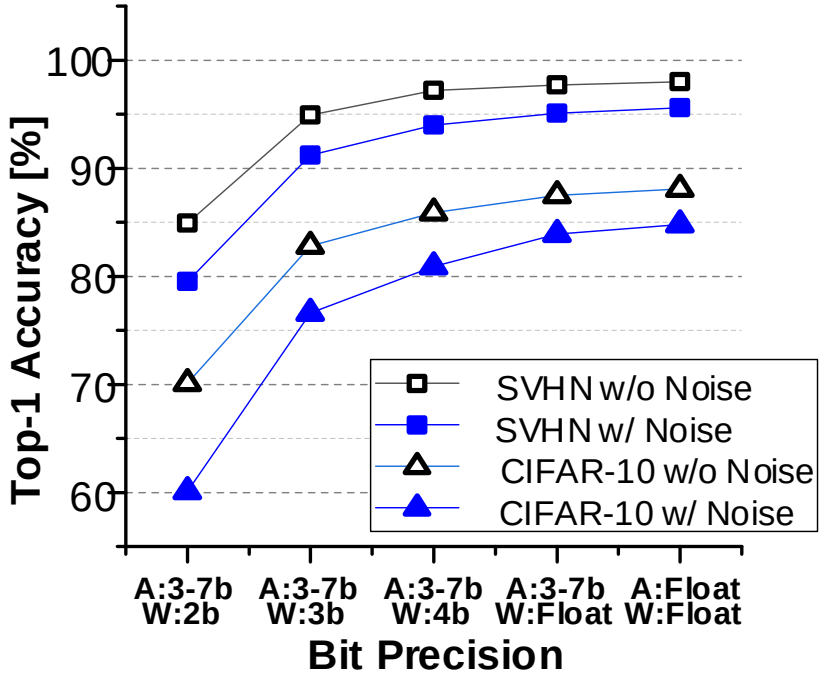
3.4.4 Accuracy Evaluation

To evaluate the classification accuracy of the proposed accelerator, a Matlab model is designed based on the circuit structure including separate accumulation of positive/negative convolution results. The fixed-point activations are truncated after convolution based on the dynamic range reduction of each layer. In addition, the Gaussian noise in Section IV.C is added in each layer during training and testing and compared with truncation-only cases. SVHN and CIFAR-10 data sets are used for training and inference.

From Fig. 3.13, with an effective activation of 3–7 bits and a 4b-weight, the proposed accelerator achieves 94.0% and 80.9% accuracy with the SVHN and CIFAR-10 datasets, respectively. Accuracy degradation occurs due to both finite bit precision and circuit noise (Fig. 3.13). Compared to results obtained with identical noise conditions but floating precision for both activations and weights, accuracy degradation is 1.6% and 3.9%, respectively. Compared to the case with identical bit precision without noise, the accuracy degradation is 3.2% and 5.0% for SVHN and CIFAR-10, respectively. Notice that using binary weights incurs severe accuracy loss for the evaluated ResNet.



(a)

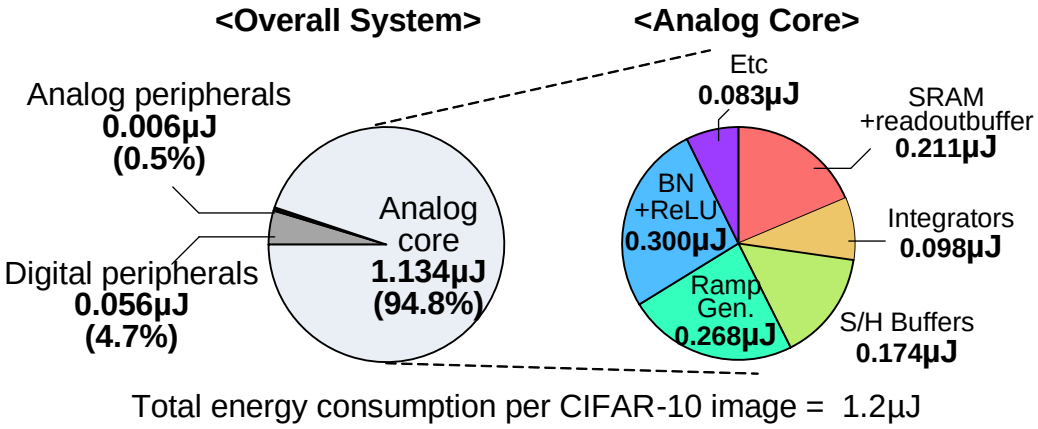


(b)

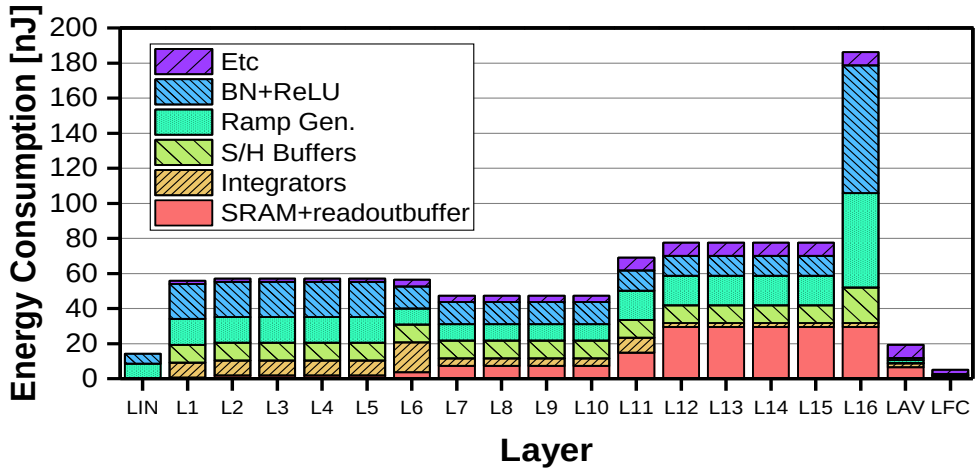
Figure 3.13 Top-1 accuracy (a) over different bit precisions of activation with 4b-weight and (b) over different bit precisions of weight with 3~7b activation.

3.4.5 Energy BreakdownAccuracy Evaluation

AA-ResNet energy consumption is measured by SPICE simulation for analog cores and Prime Time PX with synthesized digital logics on actual image input vectors. The analog cores consume 94.8% of the total energy (Fig. 3.14a), mostly by amplifier DC bias currents, while the remaining energy is consumed by digital peripherals (input image buffers with compiled SRAM, FSM controller, scanchain) and analog peripherals (ring oscillator and bias current generator).



(a)



(b)

Figure 3.14 (a) Energy breakdown of AA-ResNet accelerator and (b) energy distribution over different layers.

The energy consumption distribution among the different layers mainly depends on the layer dimension (Fig. 3.14b). The 16th layer dominates (186nJ), mainly by ramp generators and BN+ReLU, due to many parallel output channels that are connected to the average pooling layer. Overall the energy consumption is 1.2 µJ per inference, which is 3× smaller compared to state-of-the art [55]-[56], a reduction achieved by avoiding ADCs and DACs.

Table 3.2 Comparison with state-of-the-art.

		This work[a]	[55]	[56]	[44]	[45]
Technology		**28nm**	28nm	65nm	65nm	55nm
Area		**11.9**[b]	5.95	17.6	1.44	3.4
Supply [V]		**1**	0.8 / 0.6	0.68 / 0.94 / 1.2	0.675 ~ 0.925	0.4 ~ 1
Power [mW]		**389**	0.899 / 0.094	14.34[c]	1.36[c]	Max 0.69
Core Circuit Type		**Analog**	Mixed-signal	Mixed-signal	Mixed-signal	Mixed-signal
Algorithm		**ResNet**	Binary CNN	Binary CNN	SVM	Stochastic RL
Dataset		**CIFAR-10 / SVHN**	CIFAR-10	CIFAR-10 /SVHN /MNIST	MIT-CBCL face detection data	Online learning from ultrasonic sensors
On-chip Memory [kB]		**Image: 6.75 Weight :32.2**	328	295	16	N/A
Accuracy [%]		**80.9 / 94.0**	86	84/ 94/ 98.6	96	N/A
MAC precision	Weight	**4b**	1b	1b	8b	6b
	Input	**Pulse width (eff.3~7b)**	1b	1b	8b	6b
1b-scaled[d] MAC performance [TOPS]		**33.1~77.2**	0.478 / 0.072	9.438	0.272[c]	0.078[c]
1b-scaled[d] MAC Efficiency [TOPS/W]		**85.1~198.5**	532 / 772	658	200[c]	112.32[c]
# of 1b-scaled[d] MAC Ops per CIFAR-10 image		**237.47M**	818.52M[c]	1441.47M[c]	N/A	N/A
CIFAR-10 image throughput rate [image/sec]		**325,520**	237 / 36	N/A	N/A	N/A
Energy per CIFAR-10 image [µJ/image]		**1.2**	3.79 / 2.61	N/A	N/A	N/A

a. Simulation based results b. Layout based
c. Calculated based on other reported values
d. 1b-scaled: (weight precision) × (input precision) × original value

3.5 Conclusion

In this chapter, we proposed a multi-bit precision AA-ResNet accelerator design for performing all operations, including convolution, NL transform, BN, and multi-cycle value retention, in the analog domain to overcome DAC/ADC overhead present in conventional approaches. The proposed design achieved 1.2 µJ energy consumption and an inference rate of 325,520 images/s for the SVHN/CIFAR-10 data sets. We further analyzed the nonlinearity in convolution, effective bit precision of activations from noise and dynamic range shrinkage, and accuracy including the effects of noise and bit precision.

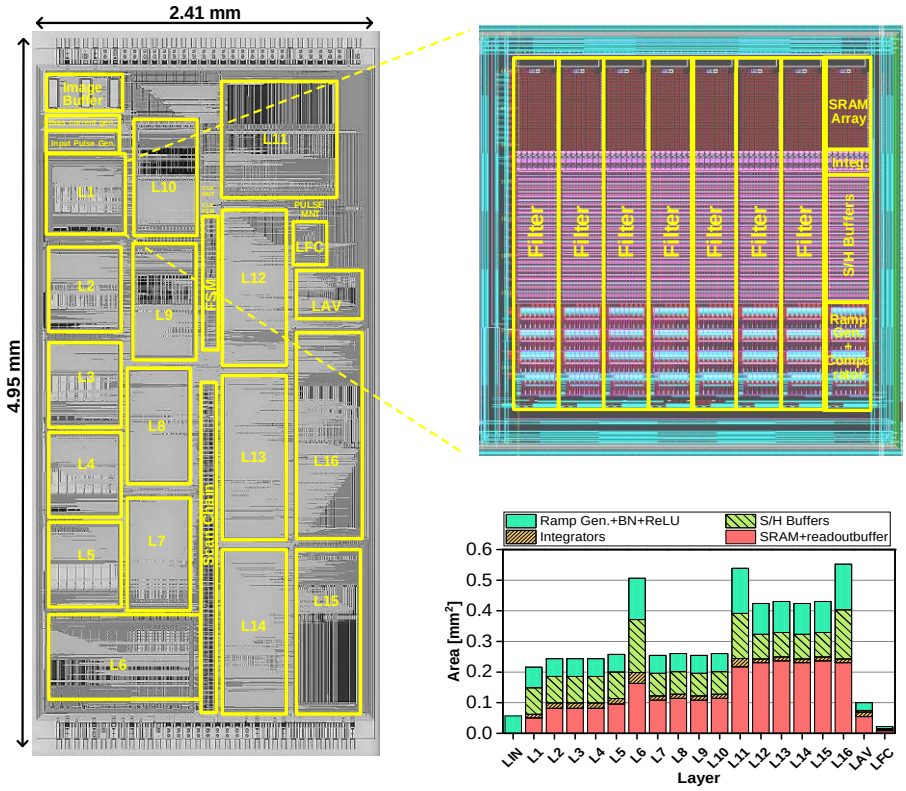


Figure 3.15 (a) Layout view of the proposed AA-ResNet accelerator, (b) layout view of the 1st layer, and (c) area distribution over different layers.

CHAPTER 4

A Miniaturized Wireless Neural Recording IC for Motor Prediction with Near-Infrared-Based Power and Data Telemetry

4.1 Introduction

Brain machine interfaces using neural recording systems [64]-[67] can enable motor prediction [68]-[69] for accurate arm and hand control in paralyzed or severely injured individuals. However, implantable systems have historically used wires for data communication and power, increasing risks of tissue damage, infection, and cerebrospinal fluid leakage, rendering these devices unsuitable for long-term implantation (Fig. 4.1). Recently, several wireless and miniaturized neural recording implants with various power and data transmission methods were proposed. References [22]-[23] propose an electrocorticography (ECoG) recording system with near-field RF power transfer and bilateral communication, but the 0.5W Tx exceeds maximum exposure limits by 10× [23]. Ultrasonic telemetry can safely send more power than RF; however, it requires mm-scale dimensions ($0.8mm^3$ in [24]) due to bulky ultrasound transducers. On the other hand, near infrared (NIR) light can provide power transfer and data downlink via a

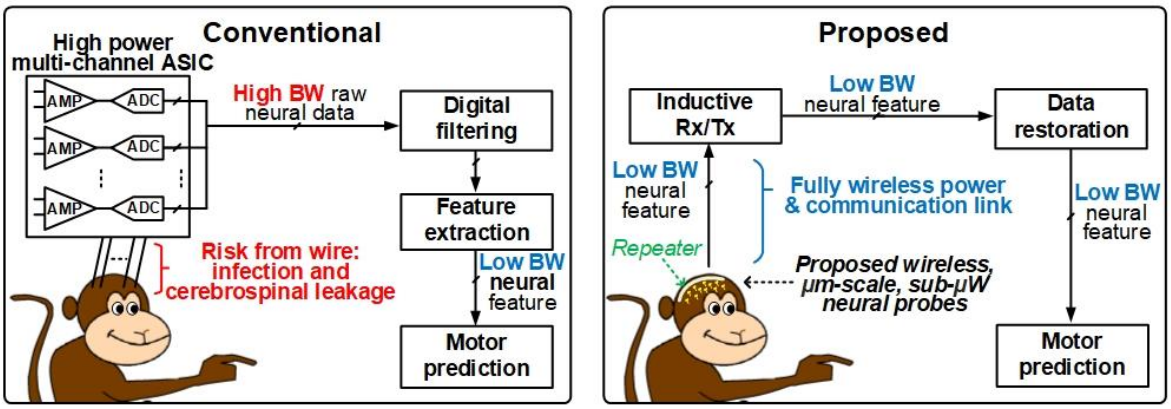


Figure 4.1 Conventional and proposed neural recording system.

photovoltaic cell (PV), and a data uplink via a light-emitting diode (LED). Dimensions can be scaled to 100s of microns [70], with [26] demonstrating a 0.0297mm^2 neural recording system using a 50mW/mm^2 light source (<1/6th of safety limit for the brain). However, this system is limited to a single channel, and since it only has a surface electrode, it can record only surface potentials (facedown, potentially blocking the light channel) or must itself be injected into brain tissue, creating significant tissue damage and danger of bleeding. In this chapter, we propose a 0.74μW, 0.19×0.17mm^2 IC designed for a wireless neural recording probe. It computes so-called spiking band power (SBP) [68], [71] on-chip to save 920× power while maintaining accurate finger position and velocity decoding.

4.2 System Overview and Top Circuit

A neural probe IC is designed for a larger neural recording system concept (Fig. 4.2) in which numerous micro-probes would be placed on the brain in the sub-dural space to record neural spikes using a carbon fiber electrode that penetrates several mm into brain tissue and has been

shown to incur minimal chronic scar formation [72]. The probes will be powered and globally programmed by 850nm NIR light emitted by a repeater placed in the epidural space. The LED in the probe will act as the data uplink; its light received by the repeater using a single-photon avalanche diode (SPAD). The repeater would service 100s of probes, which are distinguished by their on-chip ID and location. Given its larger size, the repeater can use an inductive link for wireless power and data communication with an external receiver.

The CMOS IC consists of an optical receiver followed by clock and data recovery, a random-number-generated-based chip ID [73], neural recording amplifier, SBP extractor, and LED driver (Fig. 4.3).

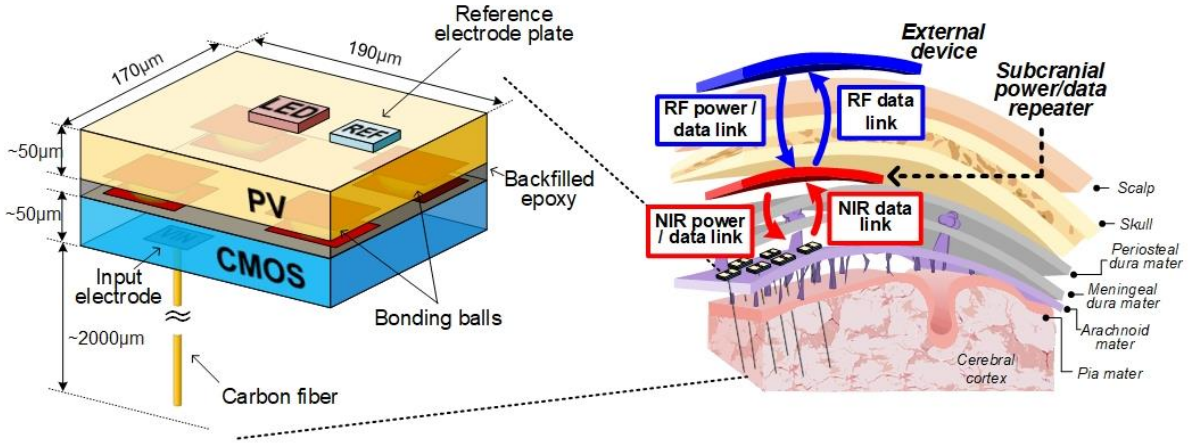


Figure 4.2 Concept diagram of proposed neural probe and two-step approach for recording and transmitting neural signals.

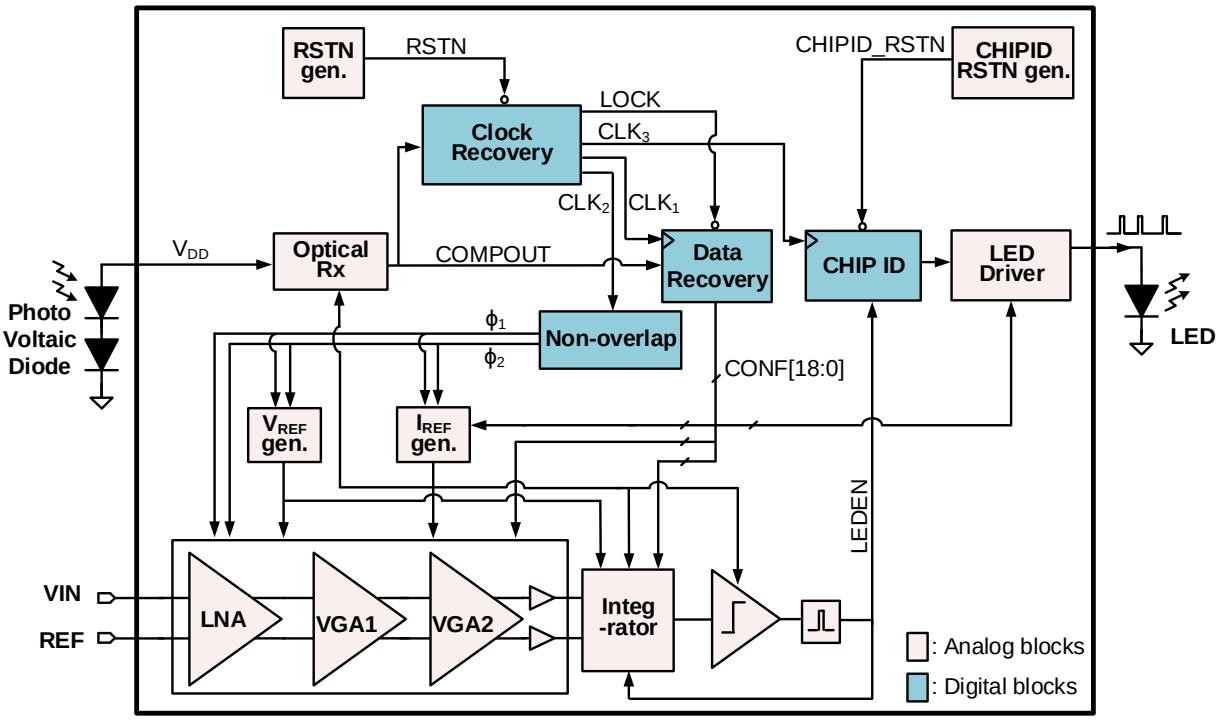


Figure 4.3 Top-level circuit diagram of the neural recorder.

4.3 Optical Receiver and Clock and Data Recovery

Fig. 4.4, 4.5 shows the schematic and measured signal diagram of the optical receiver (ORx). VDD is AC-coupled to a comparator input to convert modulated light from the repeater to a digital signal. The comparator has 80mV hysteresis to remove glitches due to unwanted VDD fluctuations. In the power-on reset phase, the clock recovery circuit locks the onchip recovery clock to the precise 8kHz modulated light from the repeater. This is critical since the clock is used to set the reference current, which must be precisely controlled for reliable amplification and signal filtering. The clock recovery circuit searches the digitally-controlled oscillator (DCO) thermometer-coded configurations to match the received modulation period with the DCO period. It then switches the system clock from the default to recovery clock using glitchfree multiplexers.

After clock locking, the repeater programs the system using pulse width modulated (PWM) light (downlink). An 8b hardwired passcode is implemented to prevent unwanted programming. The signal diagrams in Fig. 4.5 are measured from the proposed chip, wire-bonded with a custom dualjunction GaAs PV cell that generates 893nA I_{SC} and 1.67V V_{OC} under 120.5µW/mm^2 850nm light (Fig. 4.6).

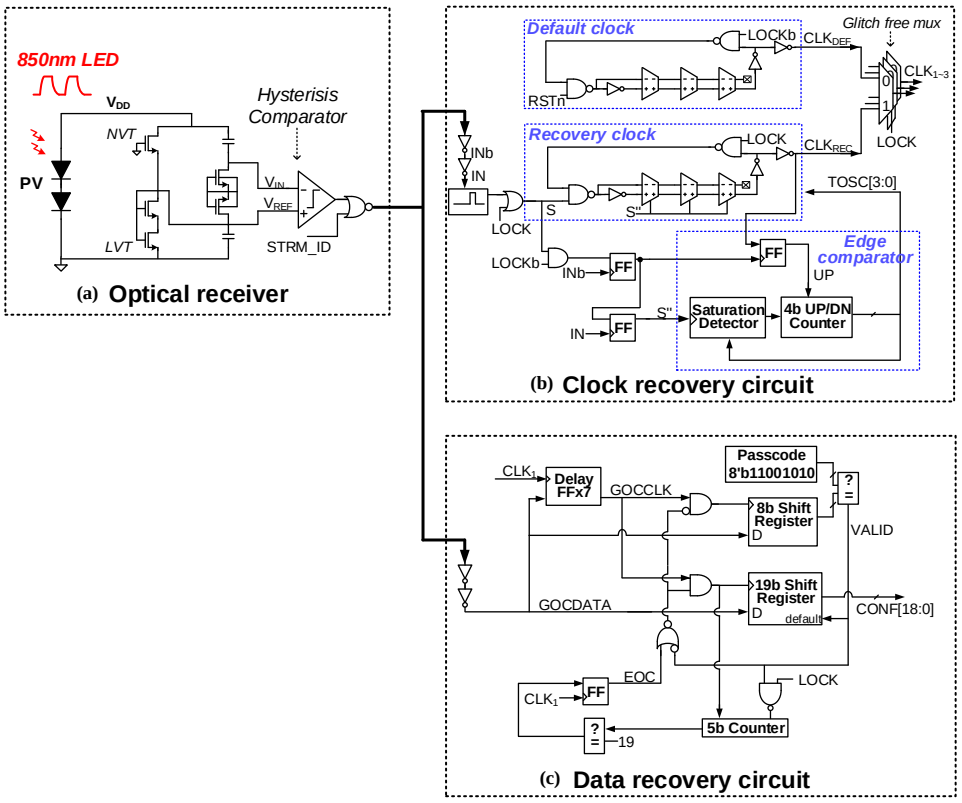


Figure 4.4 (a) Optical receiver, (b) clock recovery circuit, and (c) data recovery structure.

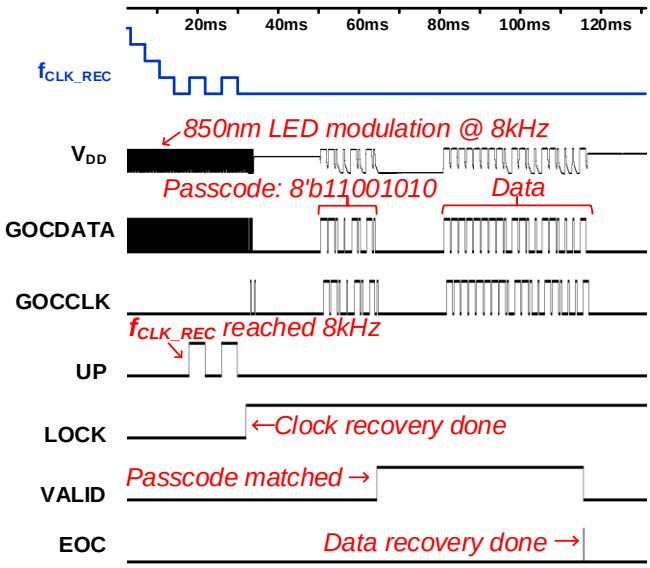


Figure 4.5 Measured signal diagram during clock and data recovery.

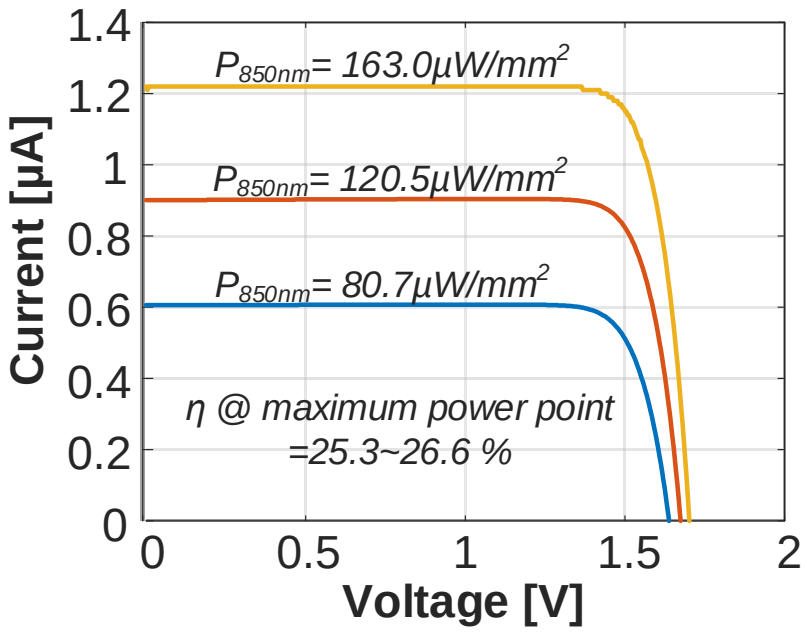


Figure 4.6 Measured performance of the PV.

4.4 Amplifier and Rectifier based Analog Integrator

The AFE is specifically designed to support SBP [68] based finger position / velocity decoding. SBP is the absolute average of signal amplitude in the 300-to-1000Hz band. When used as input to a trained linear decoding filter, SBP maintains finger position / velocity decoding accuracy relative to a standard 7.5kHz bandwidth neural recording while reducing the required communication bandwidth from probe to repeater to only 100s of Hz, thereby reducing uplink power. The AFE is composed of a three-stage bandpass differential amplifier chain with subsequent source follower and rectifier-based integrator to quantize the SBP (Fig. 4.7, 4.8). The LNA, with 60MΩ input impedance at 1kHz, is fully differential and achieves 30dB gain without bulky capacitors by implementing its gain using g_m ratio. VGA1 and VGA2 set the high-cut-off (f_H, 950Hz) and low-cut-off frequencies (f_L, 180Hz), respectively, and define the spiking band. f_H is set by VGA2 bias current, which is generated by a current reference implemented using a voltage reference and switched capacitor operating at f_{CLK}. f_L is defined by the VGA2 DC servo loop, whose feedback impedance is defined by $1/C_{SW}f_{CLK}$. Accuracy of f_H and f_L is ensured by locking f_{CLK} during clock recovery to the repeater. Peak gain is measured at 69dB while amplifying action potential (AP) spikes in 180−950Hz bandwidth for SBP-based motor prediction. (Fig. 4.9) Measured input-referred noise (IRN) is 4.8μVrms while consuming 510nW at 38°C.

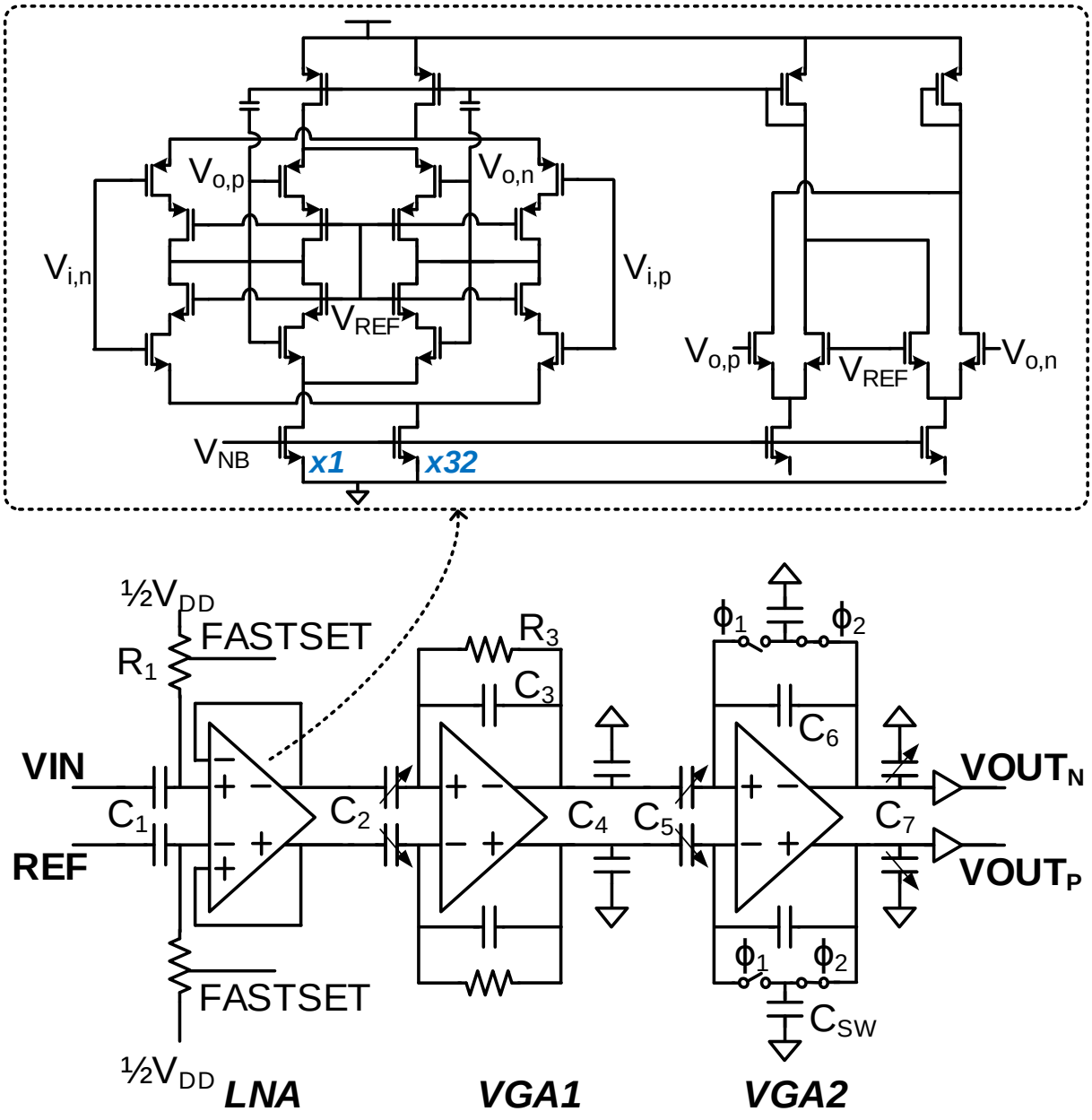


Figure 4.7 Amplifier structure.

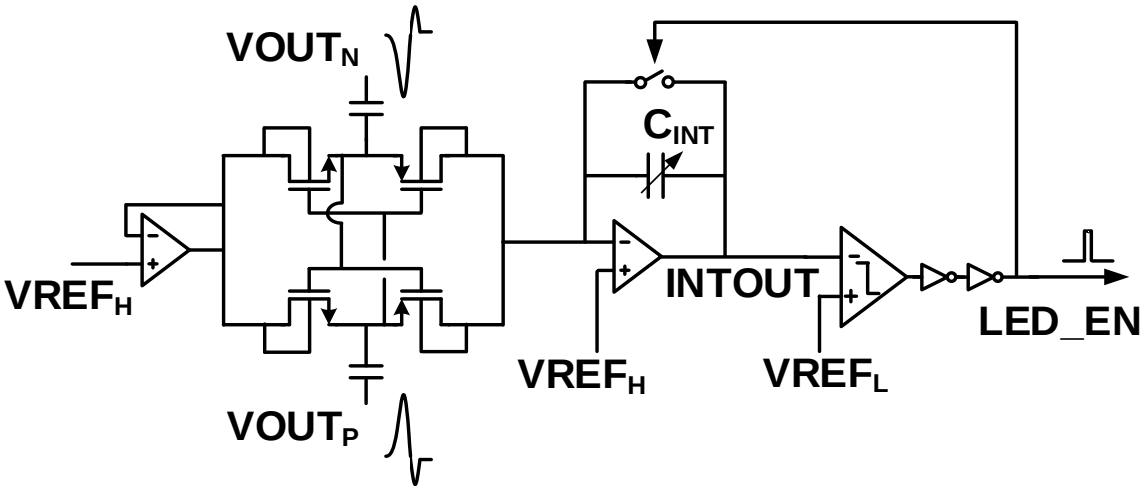


Figure 4.8 Rectifier based analog integrator structure.

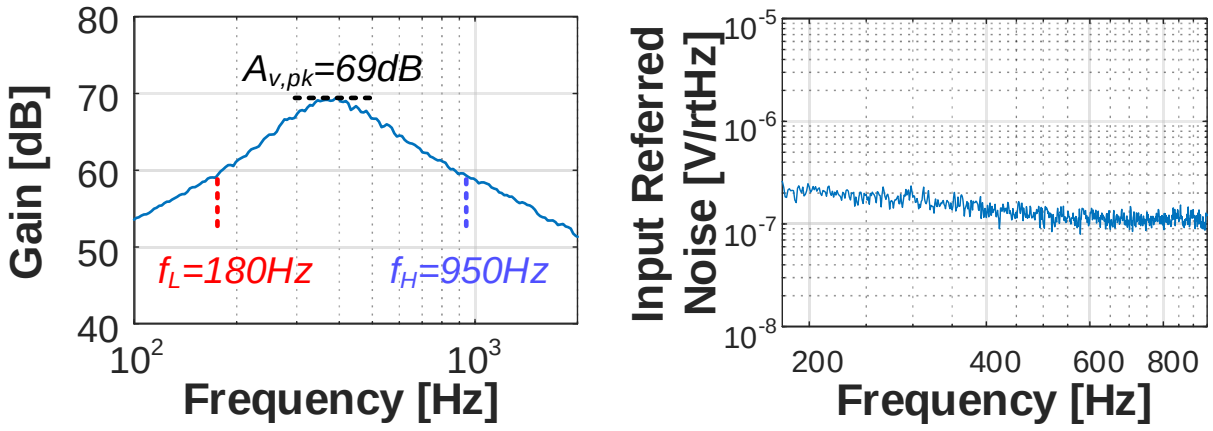


Figure 4.9 Measured AC gain and input referred noise of amplifier.

The 3-stage amplifier drives a rectifier (Fig. 4.8) whose output is initially precharged to $VREF_H$. The rectifier output decays at a rate proportional to its input amplitude. When it drops below VREFL, a pulse is generated on LED_EN. This triggers the LED driver to transmit a Manchester encoded (unique) chipID (Fig. 4.10) consuming 6.7pJ/bit (post layout simulation). Therefore, the LED firing rate or frequency is proportional to the SBP.

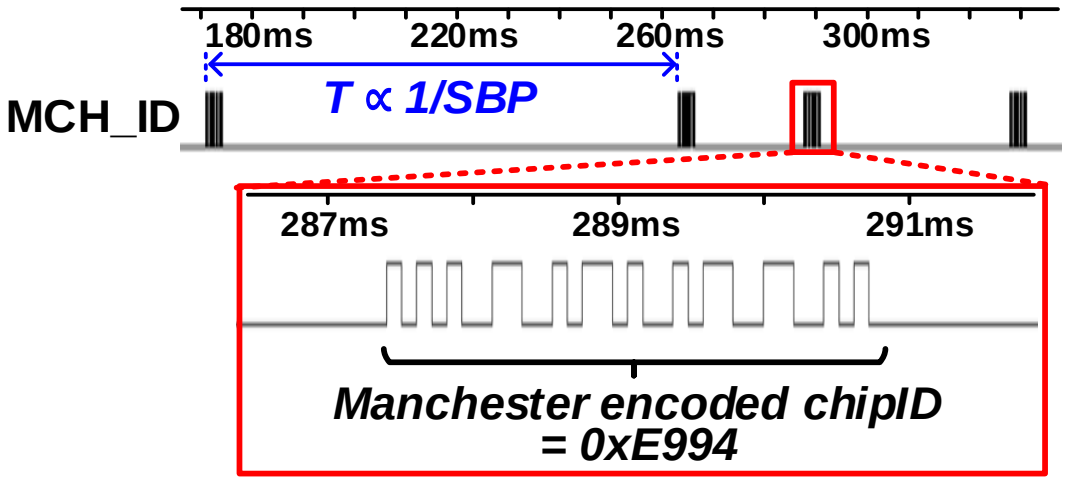


Figure 4.10 Measured Manchester encoded chipID.

4.5 Measurement Results

AFE functionality was also verified *in vivo* using a carbon fiber driven ~1.3mm into the motor cortex of an anesthetized Long Evans rat. A commercial recording system (24.414kSps, [2.2Hz, 7.5kHz] BW) is connected to the carbon fiber electrode in parallel to the IC for accuracy comparison (Fig. 4.11). All procedures complied with the Institutional Animal Care and Use Committee. VIN is the input of the proposed amplifier, measured by the high-power commercial recording system. VOUT($VOUT_P$-$VOUT_N$) is the amplifier measured output. Results show that the rectifier output (INTOUT) steps down at each motor cortex neuron spike and is restored to $VREF_H$ when it reaches $VREF_L$ (Fig. 4.12).

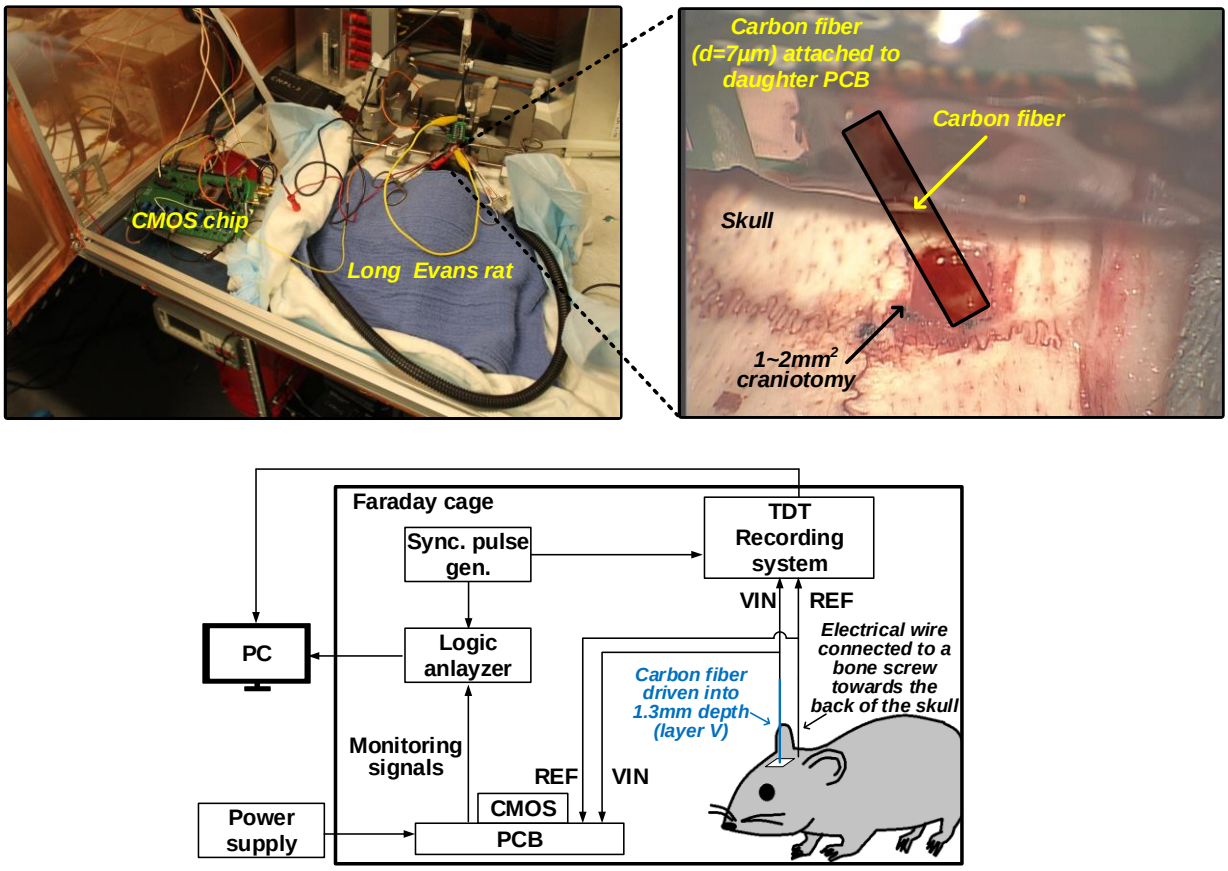


Figure 4.11 Photo of *in vivo* testing setup (top left). Carbon fiber mounted to PCB is inserted (top right) and a bone screw was placed at the most posterior portion of the skull. Recordings were taken with the IC in parallel with RA16AC headstage, RA16PA pre-amplifier, and RX7 Pentusa base station (Tucker-Davis Technologies, Alachua, FL, 2.2-7500Hz bandpass filtered) (bottom).

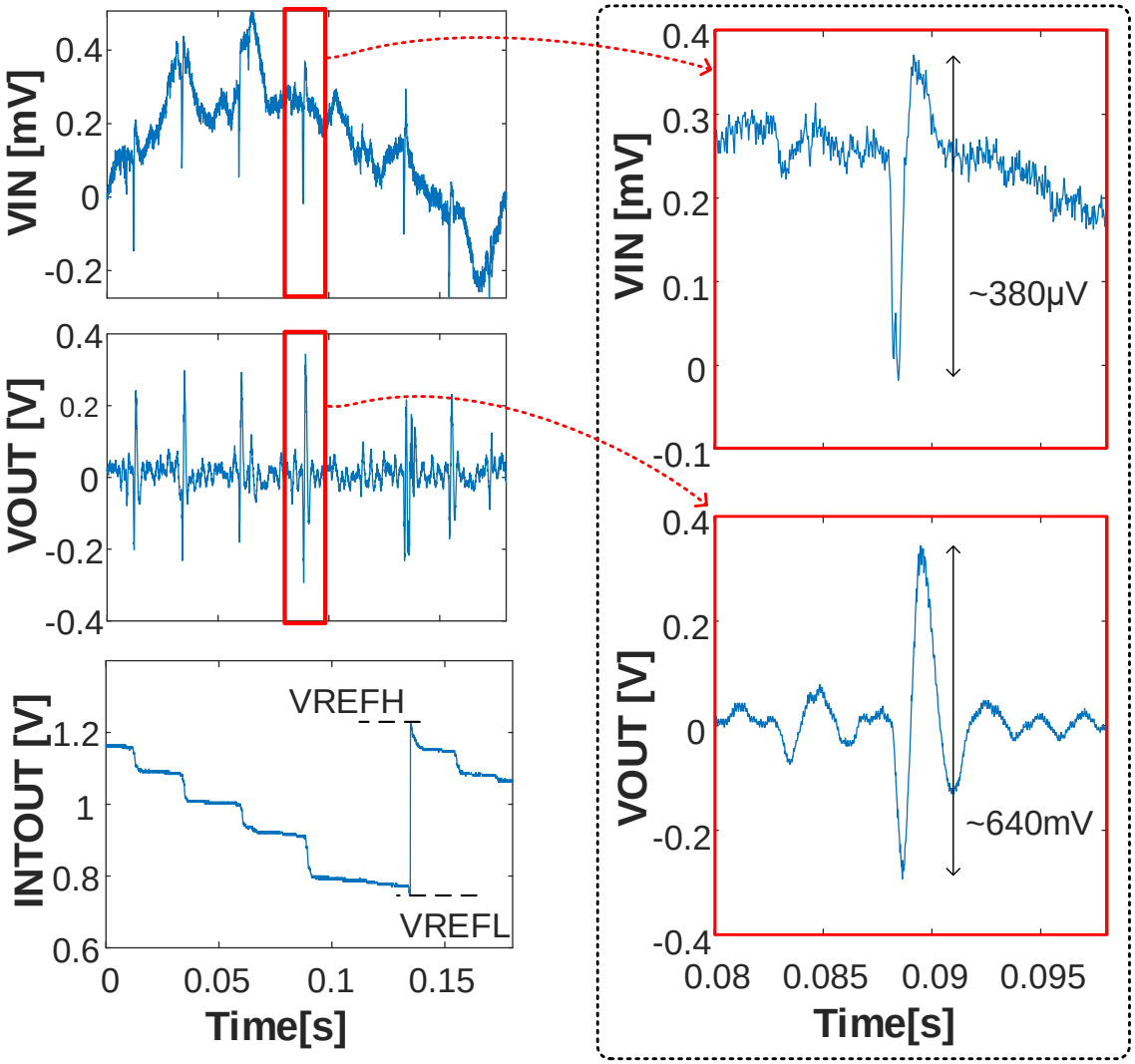


Figure 4.12 *In vivo* transient measurement results with rat motor cortex neural signal.

LED firing rate linearity across SBP is tested using synthesized AP spikes (240μVpk-to-pk, 1ms width) with varying rates from 0 to 100Hz (Fig. 4.13.). The measured LED firing rate is proportional to SBP with nonlinearity <2.9% and its sensitivity is programmable from 0.4 to 5.0 firings per μV. Overall functionality is verified using three different types of input signals; synthesized neural simulator, in vivo rat motor cortex, and pre-recorded monkey motor cortex (Fig.

4.14). Measured probe SBP is decoded from the measured time interval of LED_EN signal and compared with the result generated by a conventional high-power analog front-end and DSP SBP calculation [68]. The measured probe SBP accurately matches the conventional system results.

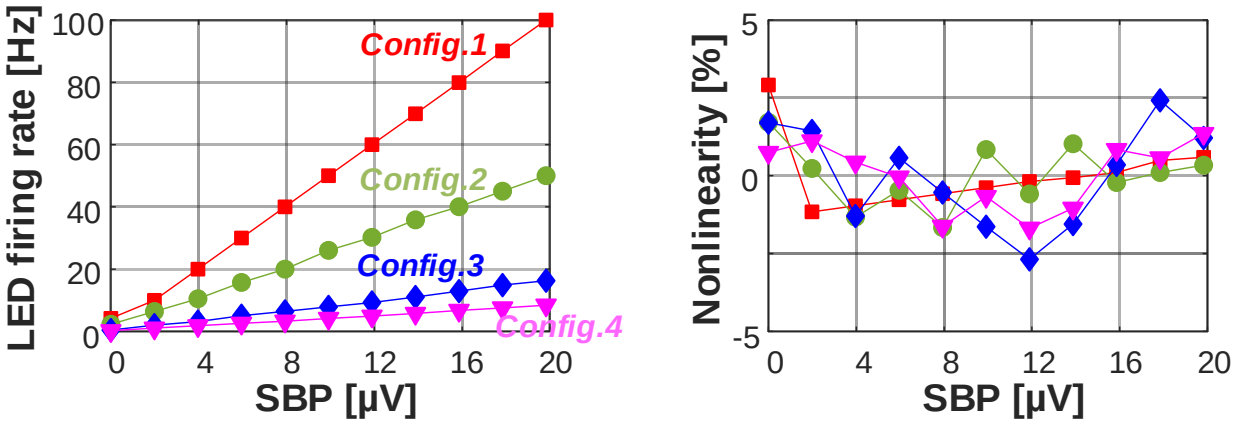


Figure 4.13 Measured linearity of LED firing rate.

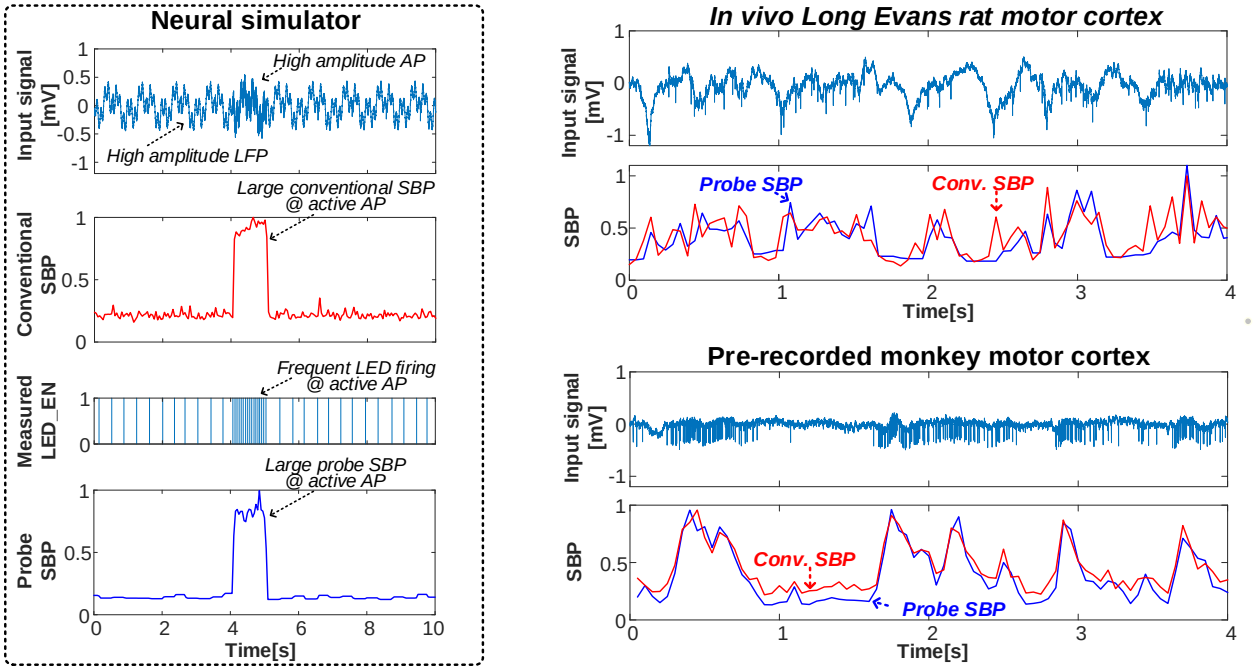


Figure 4.14 Measured transient waveform from three types of input neural signals.

Fig. 4.16 shows finger position / velocity decoding results using Kalman-Filter (KF) [69] with conventional and probe SBP from pre-recorded 20-channel neural signals of a male monkey. All procedures complied with the Institutional Animal Care and Use Committee. The system accurately predicts finger position / velocity with state-of-the-art correlation coefficient of 0.8587 / 0.5919 while a conventional high-power and wired system demonstrates 0.8886 / 0.6155 correlation coefficient.

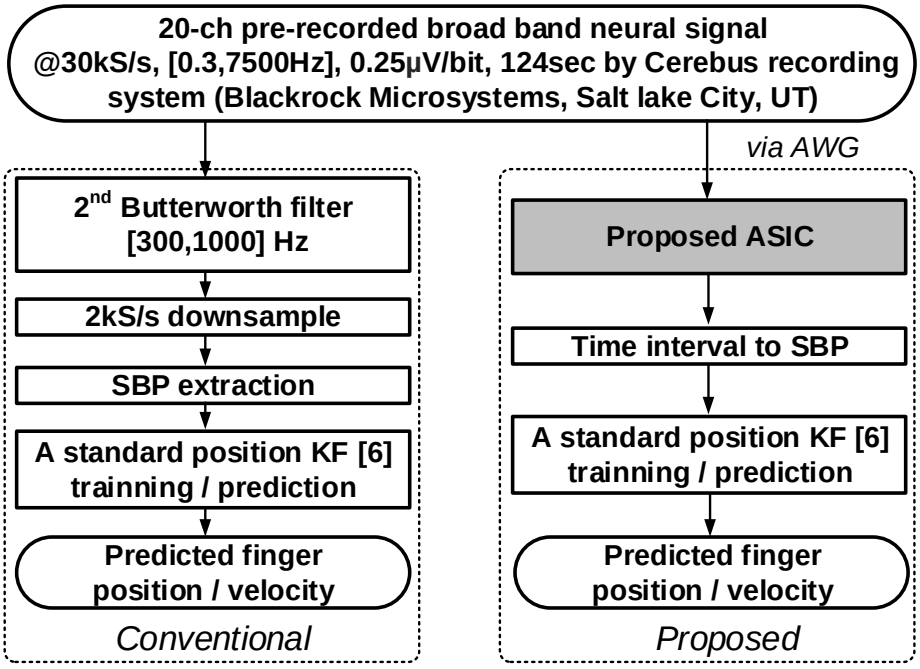


Figure 4.15 Flow chart of finger position and velocity decoding.

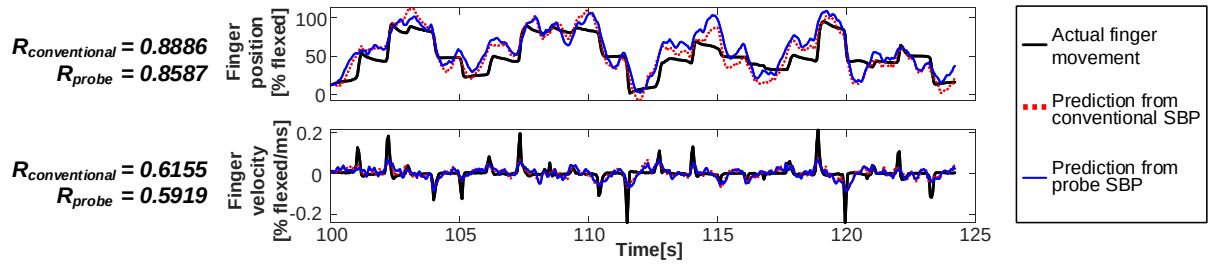


Figure 4.16 Finger position / velocity decoding result using KF with the probe and conventional SBP with pre-recorded 20-channel neural signals of a monkey.

4.6 Conclusion

The IC is fabricated in 180nm CMOS (Fig. 4.17). Table 4.1 compares to previously published wireless neural probe chip designs. It consumes 0.74μW with 3.76 amplifier NEF at 1.5V supply and 38°C, achieving best noise performance among comparable designs [22], [24], [26].

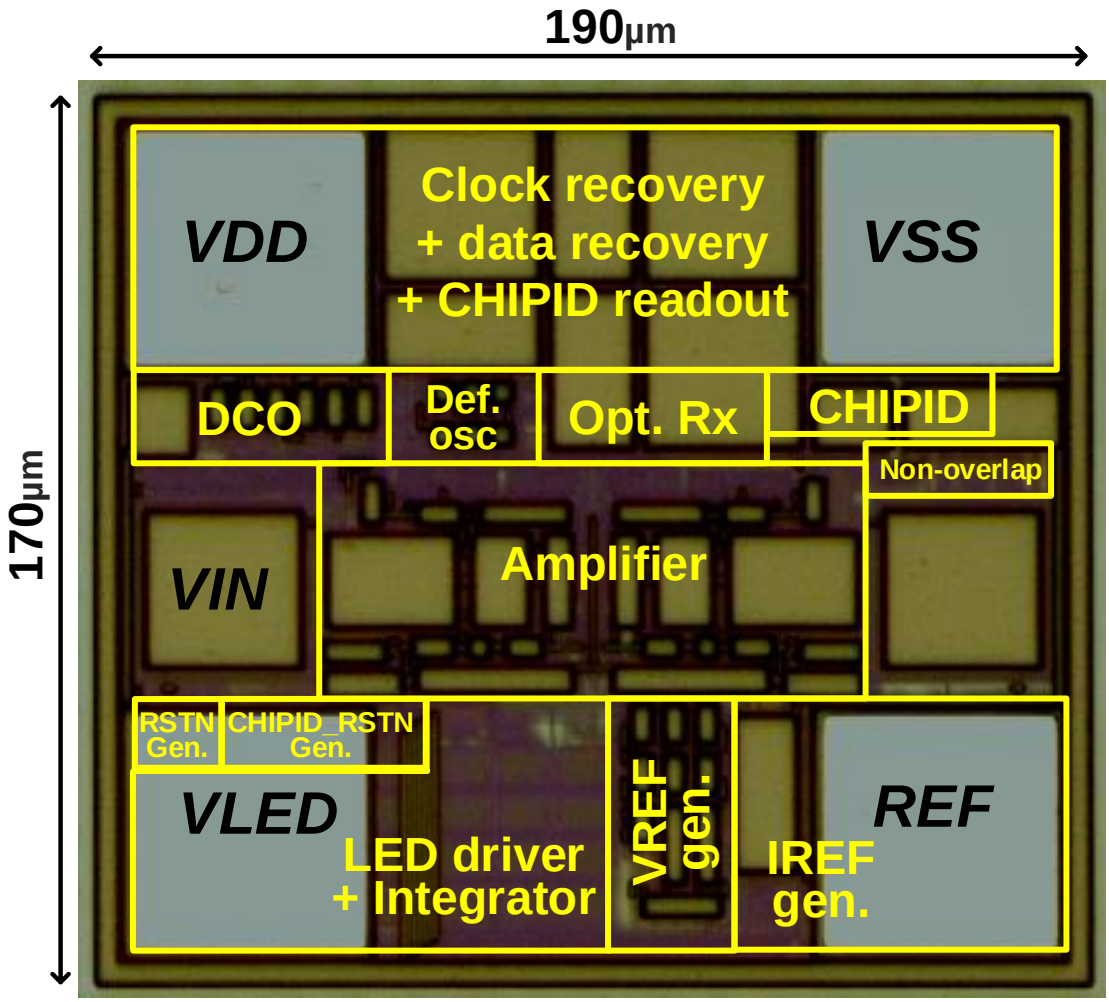


Figure 4.17 Die photo of the IC in 180nm CMOS.

Table 4.1 Comparison table.

		This work	NER 2019 [22]	ISSCC 2019 [24]	ISSCC 2018 [26]
Technology [nm]		180	65	65	180
Wireless power source		Optical	RF	Ultrasonic	Optical
Data transimmision method		Optical	RF	Ultrasonic	Optical
Data transimmision	Up-link	SIM (Symbol Interval Modulation)	BPSK-modulated RF backscatter	AM backscatter	PPM
	Down-link	PWM	ASK-PWM	No	No
On-chip feature extraction		SBP	No	No	No
Chip ID		16b	24b	No	No
Clock recovery		Yes	No	No	No
Supply [V]		1.5	0.6	1	0.9
Power [μW]	Total	0.74	40	28.8	< 1
	Amplifier	0.51	3.2	4	< 0.52
Area [mm^2] (W [mm] x L [mm])		0.0323 (0.19 x 0.17)	0.25 (0.5 x 0.5)	PZT: 0.5625 (0.75 x 0.75) IC :0.25 (0.5 x 0.5)	0.0297 (0.330 x 0.090)
Target neural signal		AP	ECoG (epicortical)	LFP, AP	LFP, AP
Gain[dB]		69	N/A	24	24
Bandwidth[Hz]		180 ~ 950*	500	5000	10000
Input referred noise [μVrms]		4.8†	2.2	5.3	42
NEF		3.76	8.7	5.87	12.3

* -10dB cutoff frequency
† Output noise divided by transfer function

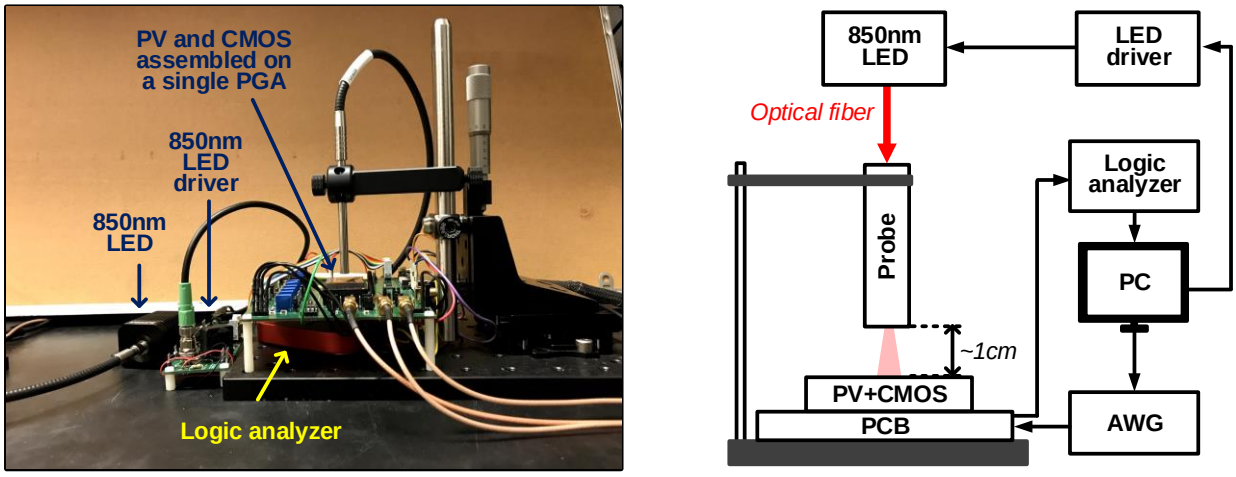


Figure 4.18 Optical setup with the IC wire-bonded with a custom dual-junction GaAs PV.

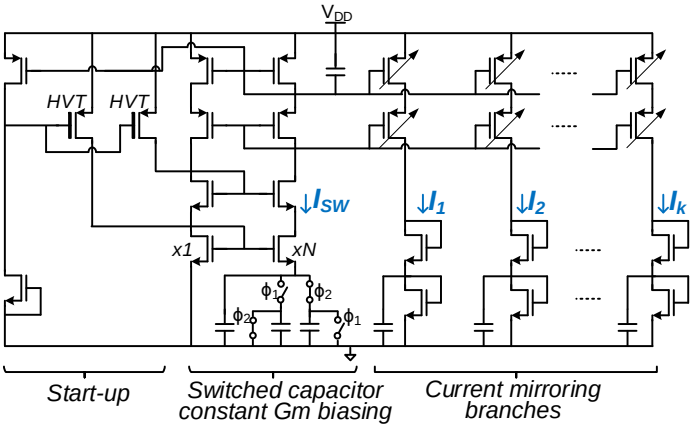


Figure 4.19 Structure of switched capacitor based current reference.

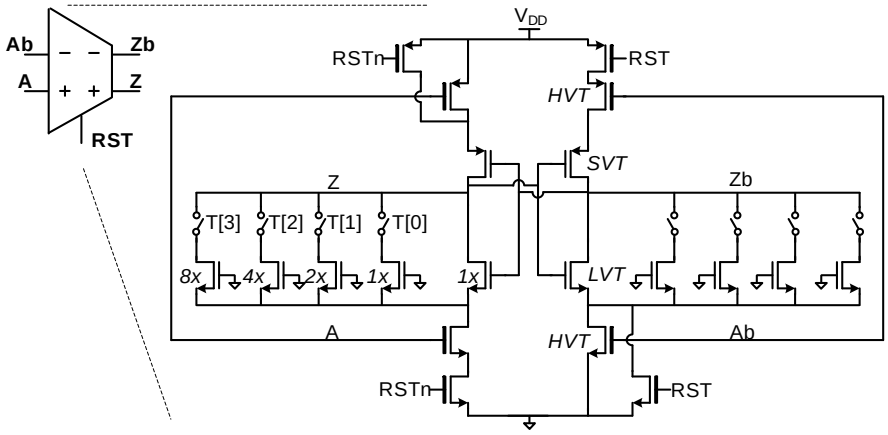


Figure 4.20 Structure of digitally controlled delay cell in clock recovery circuit.

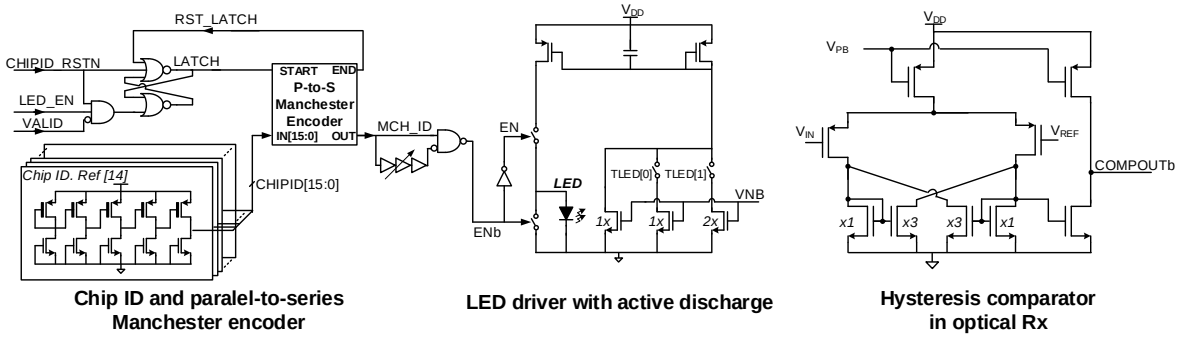


Figure 4.21 Structure of chipID Manchester encoder, LED driver and hysteresis comparator in optical receiver.

CHAPTER 5

A Light Tolerant Neural Recording IC for Near-Infrared-Powered Free Floating Motes

5.1 Introduction

Power transmission and communication are the key challenge for ultra-small (< 0.5mm) wireless neural recording motes and, among several approaches (RF, ultra-sound [22], [25]), NIR using an integrated PV and LED is unique in its ability to scale linearly to very small sizes (< 100μm) [27], [29]. Minimum size is critical to achieving dense recording arrays and minimum scarring and requires that radiated light power is maximized while chip power and currents are minimized. This leaves the circuits particularly susceptible to light-induced parasitic currents (Fig. 5.2). In conventional chips, light is blocked with an encapsulant. However, a partly transparent encapsulation that exposes the PV and LED while blocking light for sensitive circuits is infeasible at sub-mm scales leaving the solution to light tolerant circuit design. To our knowledge, this work is the first attempt to address this challenge.

The proposed IC achieves robust operation past the tissue limit NIR (150 μW/mm^2) while a baseline implementation fails at 8μW/mm^2. The chip maintains sub-μW power while incorporating advanced functionality, including on chip feature extraction and gain control. The

proposed work was tested with neural signals from a Long Evans rat and demonstrated high fidelity monkey finger motion decoding.

5.2 System Overview and Top Circuit

The envisioned system architecture is described in [29] and consists of a large number of free-floating motes on top of the brain that use NIR for power delivery, uplink and downlink to a repeater unit outside the dura (Fig. 5.1). Each mote consists of a custom GaAs chip with dual junction PV (I_{SC} > 1.1µA, V_{OC} = 1.6V at 150µW/mm^2 850nm light, measured) and LED

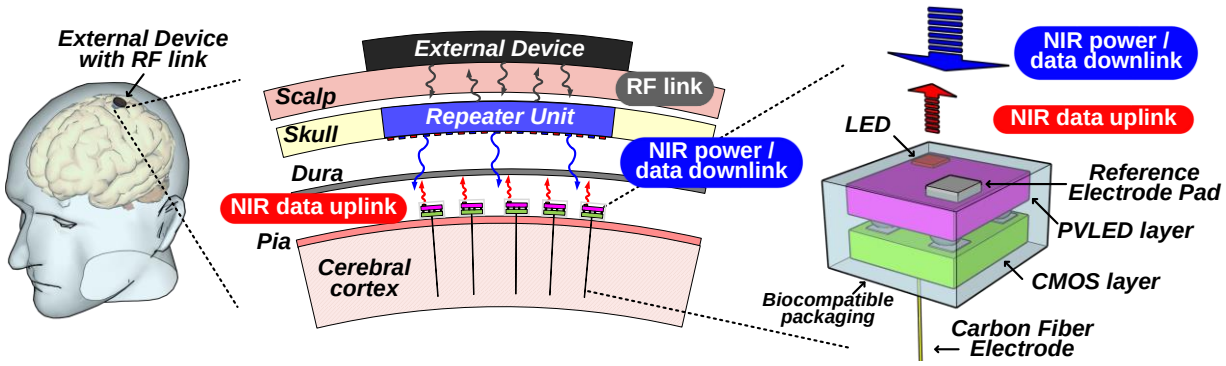


Figure 5.1 Conceptual illustration of NIR based wireless neural recording motes.

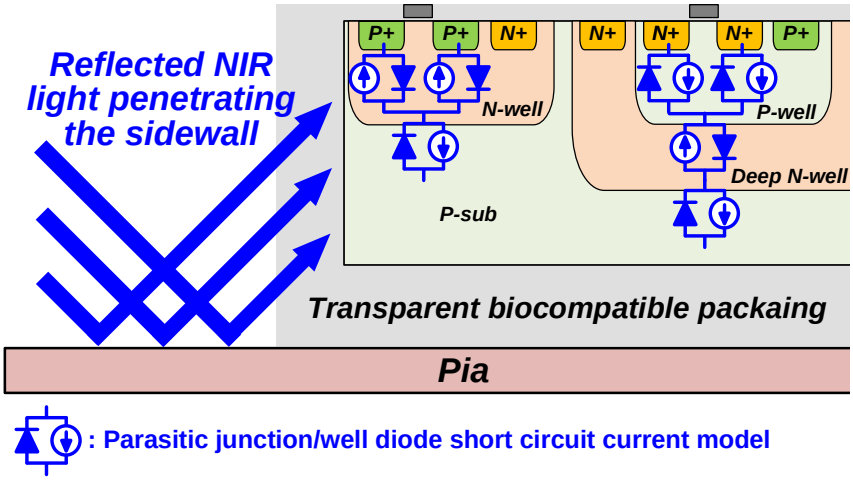


Figure 5.2 Cross section of the CMOS layer with parasitic diode short circuit currents.

sandwiched on top of the CMOS chip with an attached carbon fiber penetrating the brain to obtain neural signals.

The IC consists of a three-stage-amplifier for neural signal acquisition, signal processing that extracts a neural feature called spiking band power (SBP) [74], a pulse gap modulator (PGM) and LED driver for data uplink, and an optical receiver (ORx) for data downlink followed by clock and data recovery (Fig. 5.3).

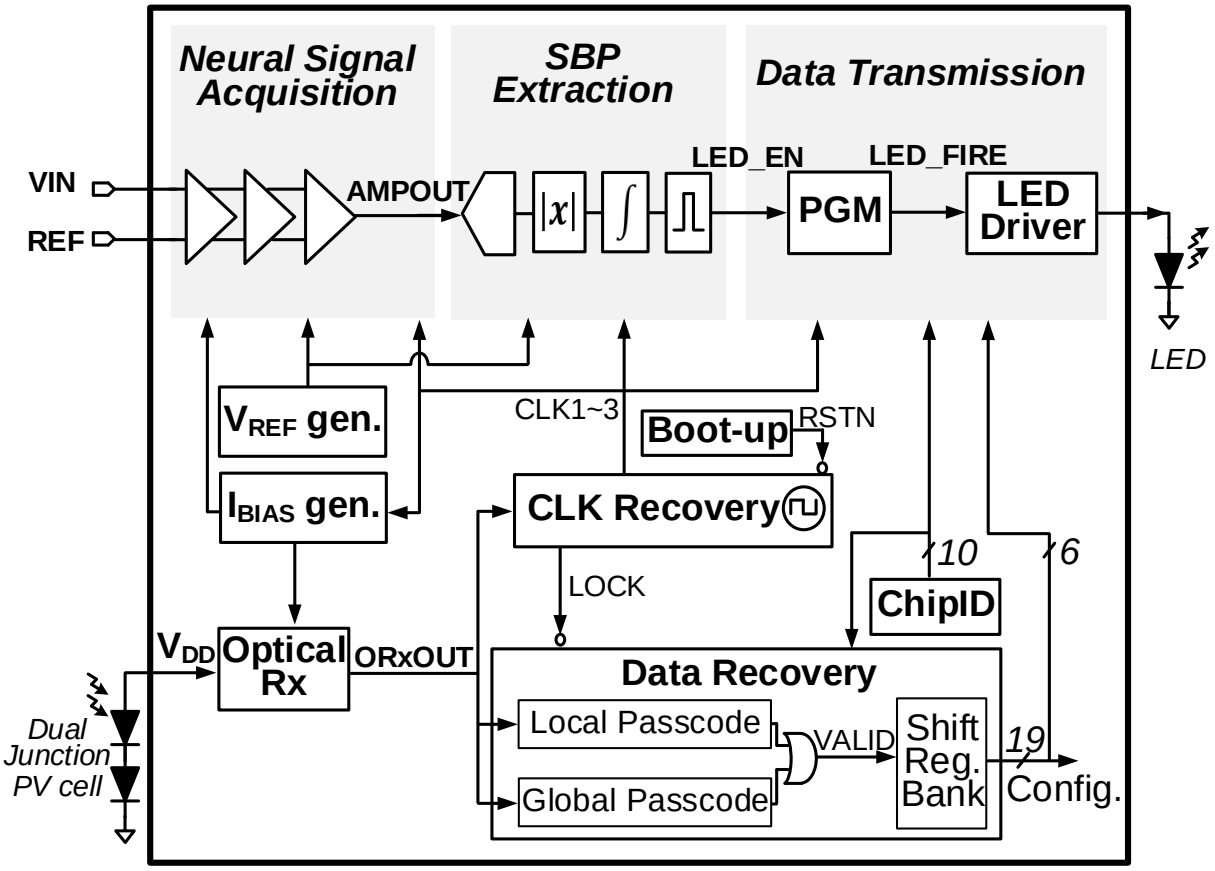
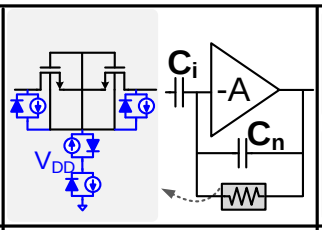
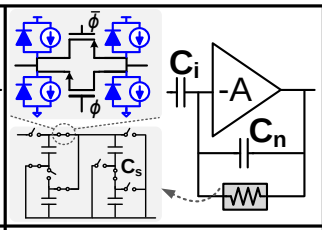
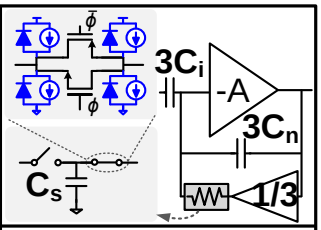


Figure 5.3 Top circuit diagram of the CMOS layer.

5.3 Light Tolerant Amplifier

A pseudo resistor (R_{PSD}) is frequently use for DC-feedback since its high, TΩ resistance [27], [29] can achieve the demandingly low high-pass corner and reduces resistor noise. However, its extremely low conductance, G_{FB}, also makes it susceptible to junction to substrate and deep n-well to p-well photo generated current (I_{SC_P}) (Fig. 5.4, left). This low conductance vs. I_{SC_P} results in a poor *light robustness ratio* ($R_{LR} = G_{FB} / I_{SC\_P}$) and the DC-bias level will drift at < 1μW/mm^2 (simulation). A series-to-parallel switched capacitor-based resistor [75] was proposed to address the process sensitivity of R_{PSD}. However, while it has higher conduction, its high number of switches results in a large total junction area and high I_{SC_P} and R_{LR} remains poor (Fig. 5.4, mid).

R_{FB} Types			
f_{HP}	$\left(\frac{1}{2\pi}\right)\left(\frac{1}{R_{PSD}C_n}\right)$	$\left(\frac{1}{20\pi}\right)\left(\frac{C_s}{C_n}\right)f_{clk}$	$\left(\frac{1}{18\pi}\right)\left(\frac{C_s}{C_n}\right)f_{clk}$
G_{FB} [pS]	0.16	9.1	117.3
I_{SC_P} [pA]	1375	63	18
$R_{LR}(=G_{FB}/I_{SC\_P})$ [1/V]	0.00012	0.14	6.52

* I_{SC_P} (at 150μW/mm^2 light power) proportional to junction area is modeled in simulation.
** Sim. parameters: R_{PSD} = 6.3TΩ, C_S = 11fF, f_{clk} = 8kHz, each junction-to-well area = 0.25μm^2, p-well-to-deep n-well area of R_{PSD}= 15μm^2

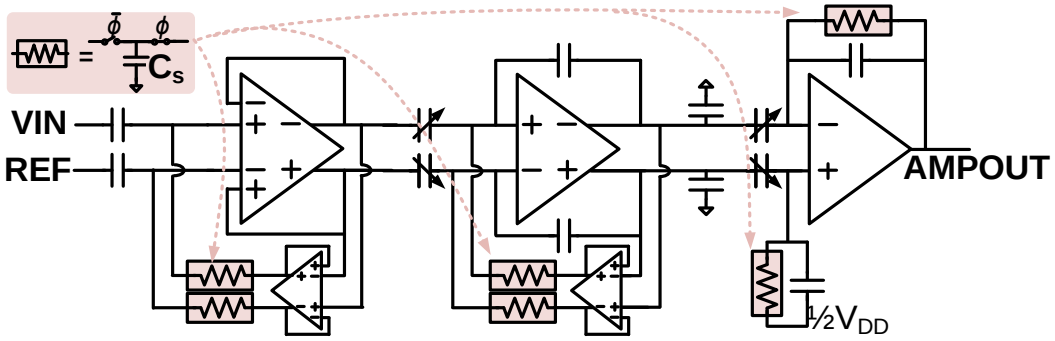


Figure 5.4 Simulated light robustness of three different feedback resistors (top) and proposed light tolerant amplifier (bottom).

Instead, this work adopts a hybrid approach combining a simple switched capacitor resistor with a 3× attenuator. It maintains a much larger G_{FB} while having a lower I_{SC_P} resulting a $5 \cdot 10^4$× improvement in R_{LR} and achieves light tolerance till 350µW/mm^2 in simulation (Fig. 5.4, right).

The amplifier achieves 68 dB peak gain, [380, 1060] Hz bandwidth, > 67dB of CMRR and PSRR and, IRN of 6.2µVRMS with 150µW/mm^2 of incident 850nm LED light at 38°C in

	Light OFF	Light ON(150µW/mm^2)
Peak Gain [dB]	67.2	68.0
CMRR [dB]	69.7	72.2
PSRR [dB]	67.5	69.2
BW [Hz]	[350,1080]	[380,1060]
ORN [mV_{RMS}]	12.7	13.5
IRN [μV_{RMS}]	6.2	6.2

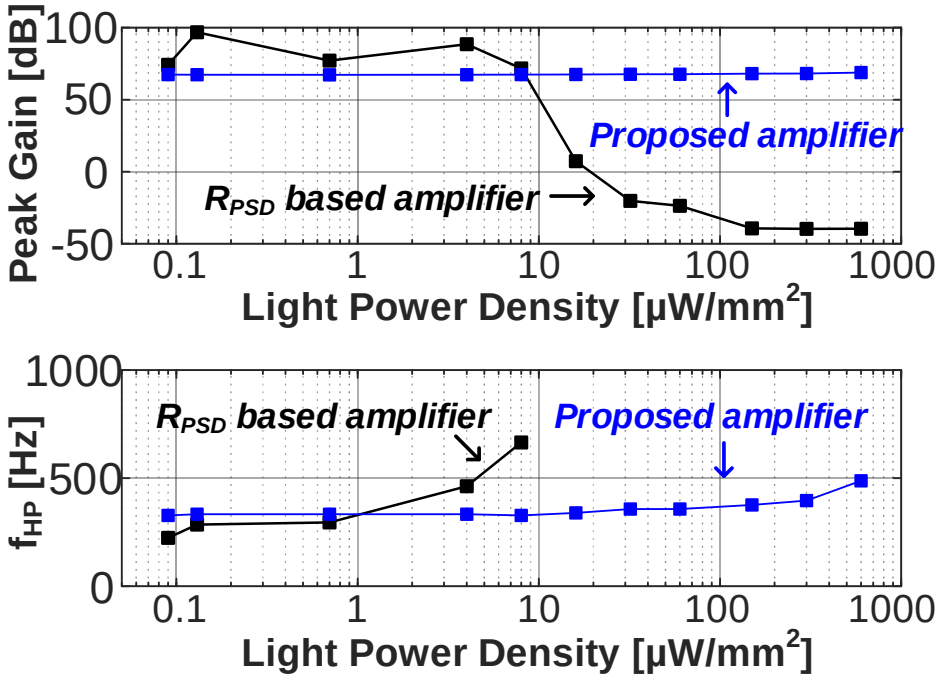


Figure 5.5 Measured amplifier performance with 850nm light (IRS4, CMVision).

measurement which is nearly unchanged from that measured without light (Fig. 5.5, table). The graph in Fig. 5 plots measured gain across light level for a baseline R_{PSD} and proposed structure showing that while the baseline structure fails at 8μW/mm^2, the proposed structure stays stable till 300μW/mm^2.

5.4 Flash ADC and Pulse-Counter-based SBP Computing Unit

Spiking band power (SBP) is a neural feature used for motor prediction and is defined as average of absolute signal amplitude in 300-1000Hz [74]. The analog SBP extraction in [29] is compact, but relies on tens of pA of on-current to charge an integration capacitor, which is susceptible to I_{SC_P}. We instead propose an area-efficient and light tolerant digital SBP extraction

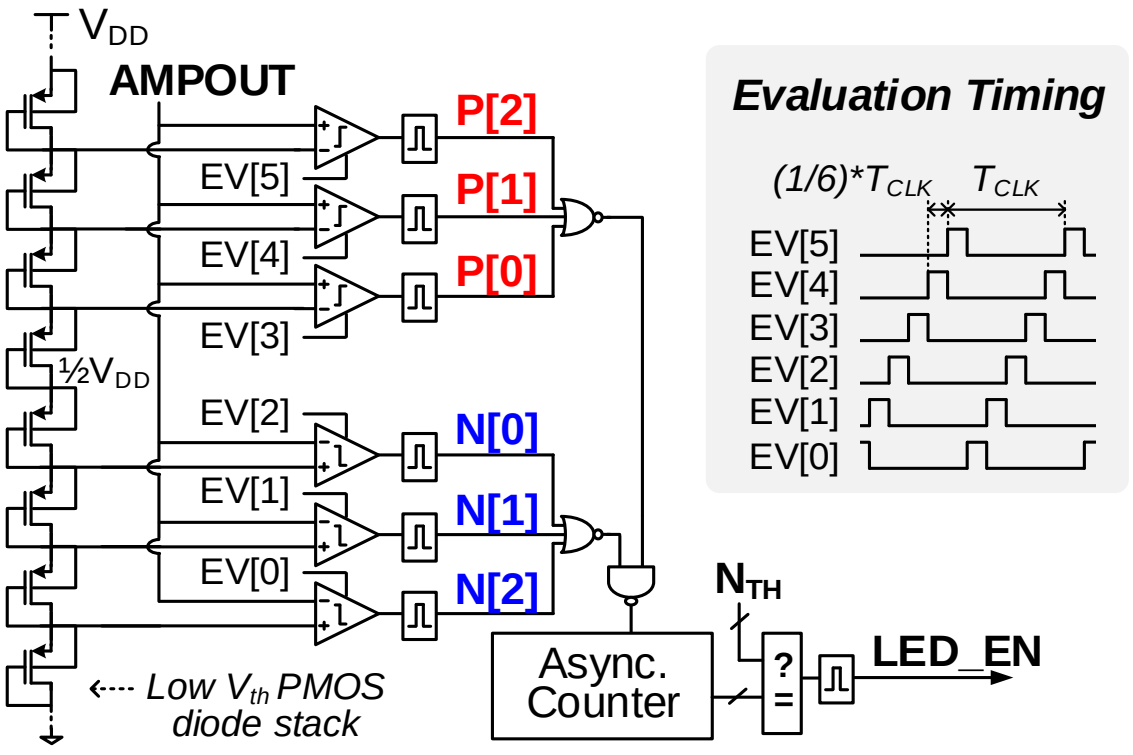


Figure 5.6 Flash ADC and pulse-counter-based SBP computing unit.

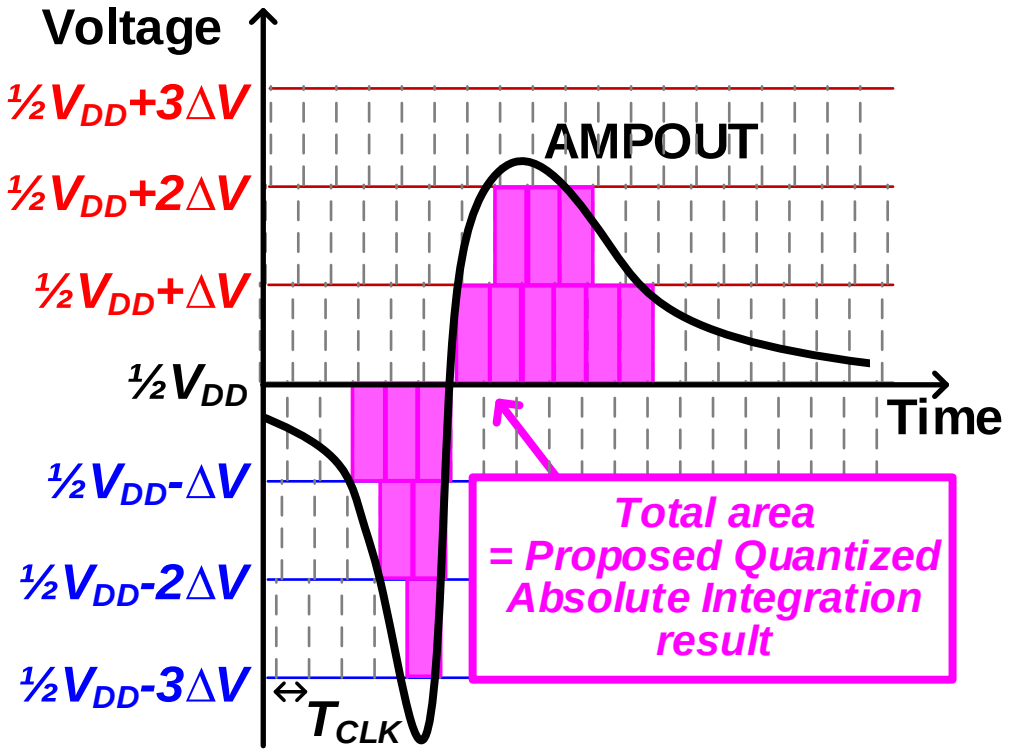


Figure 5.7 Quantization of absolute amplitude and width from the SBP computing unit.

unit using a flash ADC. It consists of a diode-stack-based V_{REF} generator (12nA, simulation), dynamic comparators with staggered clocks, followed by pulse generators. An asynchronous counter accumulates the total number of fired pulses which integrates the absolute amplitude over the pulse width (Fig. 5.6, 5.7). By comparing the counter to a threshold, SBP is symbol-interval-encoded (LED_EN, Fig. 5.6). LED_EN then fires the LED with a pulse-gap-modulated (PGM) encoding of the mote ID (Fig. 5.3). Each LED packet consists of a total 17 pulses where the pulse gap ($2*T_{CLK}/3*T_{CLK}$ for data 0/1) encodes the 10b unique chip ID (from PUF [73]) and 6b gain configuration (Fig. 5.10). The LED driver consumes 76nW (simulation) at 50Hz LED firing rate.

5.5 Optical Receiver and Remote Gain Control

The ORx allows for data downlink and remote gain control (RGC) (Fig. 5.8). Two matched 2T-VRs [36] provide DC-bias to the inputs of a hysteretic comparator, AC coupled to VDD and GND. Light modulation toggles the comparator which drives clock and data recovery. The 2T-VRs are size for 1.4nA (simulation) to ensure light robustness, eliminating the light sensitive R_{PSD} bias in [29].

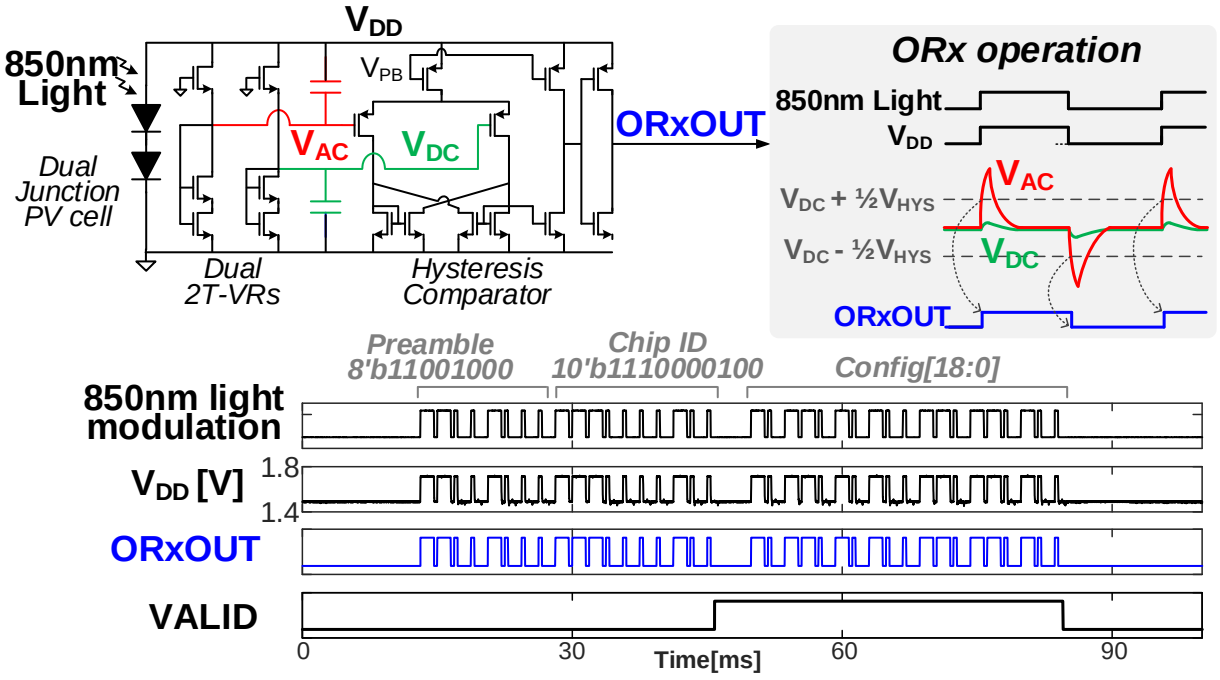


Figure 5.8 ORx structure and operation (top), and measured selective programming waveforms from wireless optical setup (bottom).

5.6 Measurement Results

The proposed IC was fabricated in 180nm CMOS (Fig. 5.11). In a fully wireless optical setup with an NIR laser for power transfer and downlink and SPAD detector for uplink reception (Fig. 5.10) the IC with custom PV/LED GaAs chip wirebonded side-by-side was fully functional. The LED_EN signal was successfully decoded from the measured SPAD output using the 16b match filter, shown in Fig. 5.10.

In vivo measurement using a carbon fiber inserted into the brain of an anesthetized Long Evans rat and wired to the CMOS chip verified the proposed SBP extraction. Compared to SBP measurement with a high-power commercial recording/signal-processing system, the proposed chip shows good accuracy for motor function decoding (Fig. 5.9). All procedures complied with the University of Michigan's Institutional Animal Care and Use Committee.

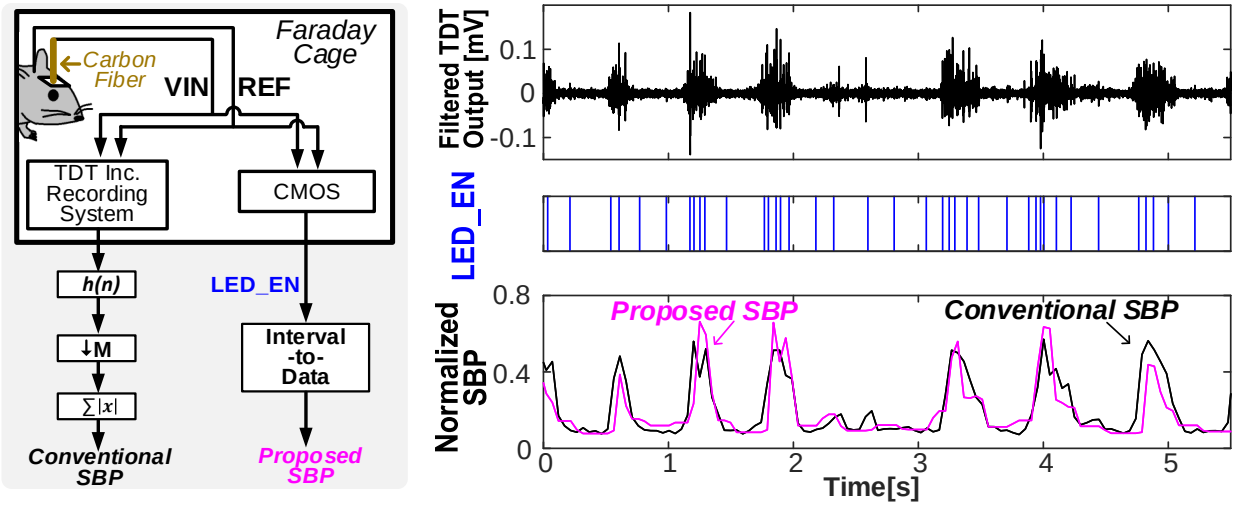


Figure 5.9 *In vivo* measurement setup with RA16PA pre-amp and RX7 Pentusa base station from TDT Inc. (left) and measured waveforms (right).

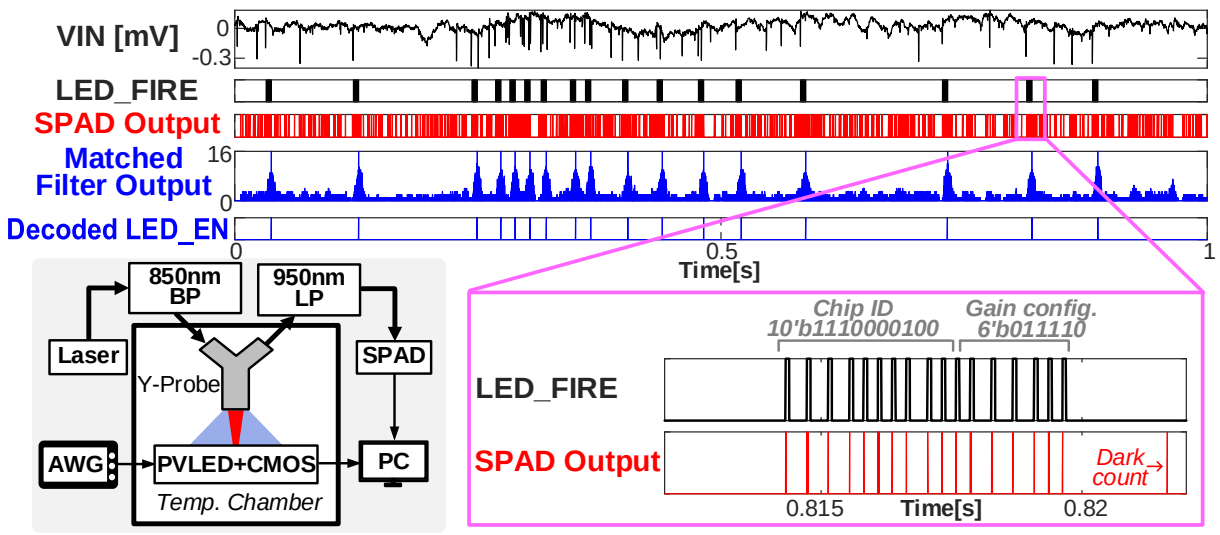


Figure 5.10 Measured matched filter decoding result (top), and wireless optical setup with NIR laser (QFLD850200S, Qphotonics) and SPAD (SPDOEMNIR, Aurea) (bottom, left).

Finger movement of a monkey was predicted using a 20-channel-prerecorded motor cortex signal and the resulting SBP from the IC with both fixed gain and off-chip RGC (based on average LED firing rate, Fig. 5.12a). A Kalman filter was used for training with the first 100s and predicting the next 24s of the movement. The proposed SBP successfully predicted the movement (Fig. 5.12b) with only slight accuracy degradation. With RGC, accuracy improves by several percent and LED firing rate remains below 50Hz across all the channels, allowing for increased channel utilization.

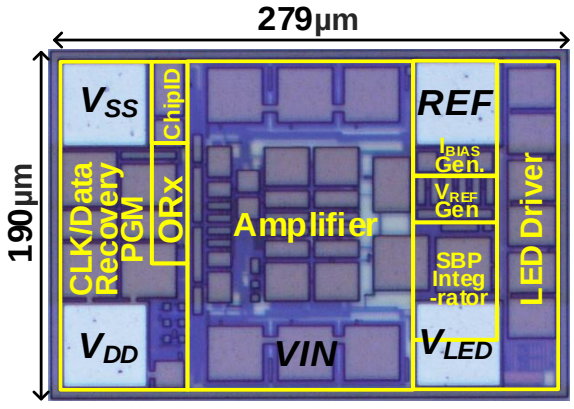
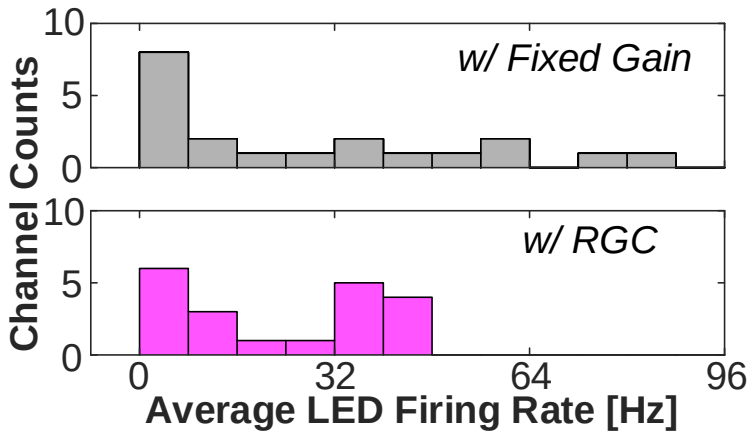
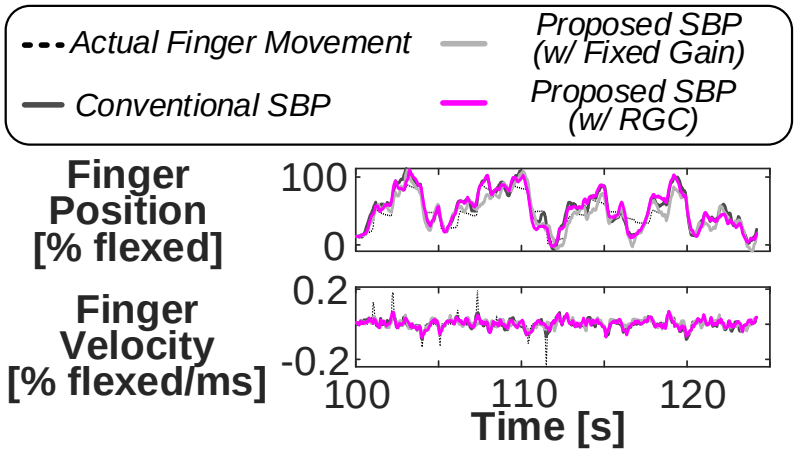


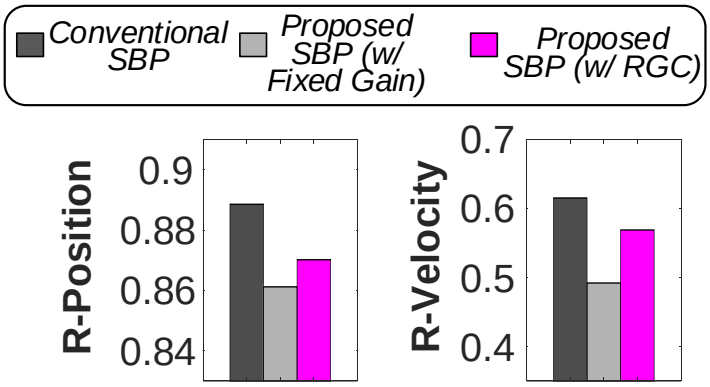
Figure 5.11 Die photo.



(a)



(b)



(c)

Figure 5.12 Finger movement decoding result (a) average LED firing rate histogram (b) predicted movement (c) correlation.

5.7 Conclusion

Table 5.1 compares the IC to state-of-the-art standalone wireless recorders. Only optical units scale below 0.5 mm and only the proposed optical mote can fully function under 300μW/mm^2 of light exposure. It also achieves the lowest power consumption of 0.57μW at 38°C with 4.1 NEF, pseudo-resistor-less amplifier, on-chip SBP extraction in digital domain, and individual mote downlink for RGC.

Table 5.1 Comparison table.

		This work	[29]	[27]	[22]	[25]
Technology [nm]		**180**	180	180	65	65
Wireless Method		**Optical**	Optical	Optical	RF	Ultrasonic
Area [mm^2] (W [mm] x L [mm])		**0.053 (0.19 x 0.28)**	0.032 (0.19 x 0.17)	0.014 (0.25 x 0.06)	0.250 (0.50 x 0.50)	0.250 (0.50 x 0.50)
Data Link	**Uplink**	**PGM-SIM**	Macnchester-SIM	PPM	RF	AM
	Downlink	**PWM**	PWM	No	ASK-PWM	No
	Mote-Level Gain Control	**Yes**	No	No	No	No
On-chip Feature Extraction		**SBP (digital)**	SBP (analog)	No	No	No
Chip ID		**Yes**	Yes	No	Yes	Yes
Clock Recovery		**Yes**	Yes	No	No	No
Use of Pseudo Resistor		**No**	Yes	Yes	No	No
Light Tolerant Design (max. light power density)		**Yes (300μW/mm^2)**	No	No	N/A**	N/A**
Supply [V]		**1.55**	1.5	0.9	0.6	1
Power	**Total [μW]**	**0.57***	0.74*	1	40	28.8
	Amplifier [μW]	**0.36***	0.51*	0.5	3.2	4
Target neural signal		**AP**	AP	LFP, AP	ECoG	LFP, AP
Gain [dB]		**67.2***	69.0*	30.0	N/A	24.0
Bandwidth [Hz]		**[350, 1080]***	[180, 950]*	10000	500	5000
NEF		**4.10***	3.76*	4.31	8.70	5.87

* Measured at 38°C ** Not Applicable

CHAPTER 6

Conclusion

6.1 Key Contributions

This dissertation proposes circuit designs for two different fields that highly demands energy-efficient systems: IoT and BMI. First, miniaturized and intelligent IoT these days need to operate for long lifetime to continuously sense, monitor, collect data, edge-compute, and communicate with other devices within their limited battery capacity. Therefore, energy efficiency of circuits and systems is key to addressing this challenge. Second, previous wired-base neural recorders have been inevitable from potential risks of tissue damage making these wire-based neural recorders unsuitable for long-term implantation. Therefore, sub-mm-scale and energy efficient wireless implants for a single neuron level activity recording has been a long-standing goal in BMI. The dissertation proposes different energy efficient circuit designs for IoT application in chapter 2 and 3, and two generations of sub-μW and sub-mm wireless neural recording IC in chapter 4 and 5.

In chapter 2, a gate-leakage-based frequency locked wake-up timer with first- and second-order temperature dependency cancellation is introduced. Wake-up timers are a critical component of WSNs for the IoT, and two key requirements are low power consumption and high timing accuracy. The proposed timer achieves a TC of 260ppm/°C across −5 to 95°C while burning only 224pW of power. In addition, the reported LS is 0.93%/V across 1.1−3.3 V of supply voltage,

achieving 150× improvement compared to previous gate-leakage-based wake-up timers. Overall, the proposed timer is Pareto optimal in terms of TC and LS vs. power among other resistor-less timers.

Chapter 3 proposes an energy-efficient AA-ResNet accelerator for edge-computing application. The proposed multi-bit precision accelerator performs all operations, including convolution, NL transform, BN, and multi-cycle value retention, in the analog domain to overcome DAC/ADC overhead present in conventional approaches. The proposed design achieves inference rate of 325,520 images/s for the SVHN/CIFAR-10 data sets in simulation while it consumes only 1.2 μJ energy per image. At the end of chapter, analysis on the nonlinearity in convolution, effective bit precision of activations from noise and dynamic range shrinkage, and accuracy including the effects of noise and bit precision is presented.

In chapter 4, a $0.19\times0.17mm^2$ IC designed for a wireless neural recording probe with NIR power and bidirectional data telemetry is proposed. It only consumes a 0.74μW of power with 3.76 amplifier NEF at 1.5V supply and 38°C, achieving best noise performance among other published standalone neural recorders. The proposed neural recording IC computes SBP on-chip in analog domain for accurate finger position and velocity decoding. Using the pre-recorded neural signal of a monkey, the IC predicts finger position / velocity with high correlation coefficient of 0.8587 / 0.5919.

Finally, chapter 5 introduces the second generation of the wireless neural recording IC with much enhanced light tolerance compared to the first generation in chapter 4. Since it is difficult to encapsulate the PV and LED partly transparent while blocking light for sensitive circuits at sub-mm scales, it is critical to design the system insensitive to the light exposure. By replacing all sub-blocks sensitive to photo generated current by novel light-tolerant structures, the proposed IC

maintains robust operation at 300μW/mm^2 of NIR light exposure on the bare die, while the first generation fails at 8μW/mm^2. The proposed IC includes individual mote downlink for mote-level gain control and consumes 0.57μW at 38°C, which is the lowest power consumption among state-of-the-arts standalone neural recorders.

6.2 Future Directions

There are various opportunities to further improve the works introduced in this dissertation. Although the wake-up timer proposed in chapter 2 is Pareto optimal in terms of TC and LS vs. power compared to other resistor-less timers, it still has higher TC relative to the nW-level resistor-based timers. This is because the gate leakage is intrinsically nonlinear across temperature compared to the resistance of poly resistors of the resistor-based timers, enabling much complex temperature dependency cancellation scheme. However, if WSNs, where the proposed gate-leakage based timer is implemented in, include temperature sensor, then an additional temperature compensation can be applied to further reduce TC below 100ppm/°C similar to the approach of [35]. Simplifying the current three-point calibration or implementation of auto-calibration would be another future opportunity for lowering the testing and production cost.

AA-ResNet accelerator in chapter 3 has a potential of improving energy efficiency of deep learning hardware. At the same time, there are various future directions to successfully develop the concept into the practical application. One is further research on improving flexibility and configurability of the analog accelerator similar to the conventional digital accelerator. Another direction could be a study on efficient simulation and layout methodology for such complex and heavy analog accelerator design. In addition, research on new in-situ methods or training methods

with the actual fabricated prototype of such analog based deep learning hardware would be helpful to minimize the accuracy degradation due to nonlinearity and PVT variations of the analog cores.

Lastly, NIR based wireless neural recording ICs in chapter 4 and 5 achieve overall great performance with a custom PV/LED GaAs chip wirebonded. In addition, the functionality of the ICs is verified with *in vivo* measurement at motor cortex of alive rats with a carbon fiber inserted. The next main step would be the full integration and verification of the actual floating neural probe composed of the proposed IC, a custom PV/LED GaAs chip, a carbon fiber electrode, and a hermetic biocompatible packaging described in Fig.4.2 and Fig 5.1. Along with the verification of a fully assembled mote, a design of the repeater unit that wireless powers and exchanges data with the floating probes implanted in brain and that communicates with the external device would be another important step toward the full wireless recording system described in section 4.2 and 5.2. Furthermore, theoretical and empirical study on decoding algorithms and channel utilization of the optical data up-link would potentially optimize the proposed wireless recording system with minimum accuracy degradation.

BIBLIOGRAPHY

[1] Y. Lee, S. Bang, I. Lee, Y. Kim, G. Kim, M. H. Ghaed, P. Pannuto, P. Dutta, D. Sylvester and D. Blaauw, "A Modular 1 mm^3 Die-Stacked Sensing Platform With Low Power I 2 C Inter-Die Communication and Multi-Modal Energy Harvesting," *in IEEE Journal of Solid-State Circuits*, vol. 48, no. 1, pp. 229-243, Jan. 2013

[2] D. Jeon, Y. Chen, Y. Lee, Y. Kim, Z. Foo, G. Kruger, H. Oral, O. Berenfeld, Z. Zhang, D. Blaauw and D. Sylvester, "An implantable 64nW ECG-monitoring mixed-signal SoC for arrhythmia diagnosis," *2014 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2014, pp. 416-417

[3] G. Kim, Y. Lee, Z. Foo, P. Pannuto, Y. Kuo, B, Kempke, M. H. Ghaed, S. Bang, I. Lee, Y. Kim, S. Jeong, P. Dutta, D. Sylvester and D. Blaauw, "A millimeter-scale wireless imaging system with continuous motion detection and energy harvesting," *2014 IEEE Symposium on VLSI Circuits*, Honolulu, HI, USA, 2014, pp. 1-2

[4] T. Jang, M. Choi, Y. Shi, I. Lee, D. Sylvester and D. Blaauw, "Millimeter-scale computing platform for next generation of Internet of Things," *2016 IEEE International Conference on RFID (RFID)*, Orlando, FL, USA, 2016, pp. 1-4

[5] M. Cho, S. Oh, S. Jeong, Y. Zhang, I. Lee, Y. Kim, L. Chuo, D. Kim, Q. Dong, Y. Chen, M. Lim, M. Daneman, D. Blaauw, D. Sylvester and H.-S. Kim, "A 6×5×4mm^3 general purpose audio sensor node with a 4.7μW audio processing IC," *2017 Symposium on VLSI Circuits*, Kyoto, Japan, 2017, pp. C312-C313

[6] S. Oh, Y. Shi, G. Kim, Y. Kim, T. Kang, S. Jeong, D. Sylvester and D. Blaauw, "A 2.5nJ duty-cycled bridge-to-digital converter integrated in a 13mm^3 pressure-sensing system," *2018 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, USA, 2018, pp. 328-330

[7] A. Krishna, "Changing the Way the World Works: IBM Research's "5 in 5"," *IBM Research Blog*, 19 March 2018 (https://www.ibm.com/blogs/research/2018/03/ibm-research-5-in-5-2018)

[8] T. Kang, I. Lee, S. Oh, T. Jang, Y. Kim, H. Ahn, G. Kim, S. Shin, S. Jeong, D. Sylvester and D. Blaauw, "A 1.7×4.1×2 mm^3 Fully Integrated pH Sensor for Implantable Applications using Differential Sensing and Drift-Compensation," *2019 Symposium on VLSI Circuits*, Kyoto, Japan, 2019, pp. C310-C311

[9] L. Chuo, Z. Feng, Y. Kim, N. Chiotellis, M. Yasuda, S. Miyoshi, M. Kawaminami, A. Grbic, D. Wentzloff, D. Blaauw and H.-S. Kim, "Millimeter-Scale Node-to-Node Radio Using a Carrier Frequency-Interlocking IF Receiver for a Fully Integrated 4×4×4 mm^3 Wireless Sensor Node," *in IEEE Journal of Solid-State Circuits*, vol. 55, no. 5, pp. 1128-1138, May 2020

[10] S. Jeong, Y. Kim, G. Kim and D. Blaauw, "A Pressure Sensing System with ±0.75 mmHg (3σ) Inaccuracy for Battery-Powered Low Power IoT Applications," *2020 IEEE Symposium on VLSI Circuits*, Honolulu, HI, USA, 2020, pp. 1-2.

[11] "Rechargeable Solid Stage Energy Storage: 12 μAh, 3.8 V, EnerChip CBC005." Datasheet, Cymbet Corp., Elk River, MN, USA, 2009.

[12] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, Art. no. 7553, 2015.

[13] J. Lim, T. Jang, M. Saligane, M. Yasuda, S. Miyoshi, M. Kawaminami, D. Blaauw and D. Sylvester, "A 224 pW 260 ppm/°C Gate-Leakage-based Timer for Ultra-Low Power Sensor Nodes with Second-Order Temperature Dependency Cancellation," *in 2018 Symposium on VLSI Circuits (VLSI Circuits)*, Honolulu, HI, 2018, pp. 117-118.

[14] J. Lim, M. Choi, B. Liu, T. Kang, Z. Li, Z. Wang, Y. Zhang, K. Yang, D. Blaauw, H.-S. Kim, and D. Sylvester, "AA-ResNet: Energy Efficient All-Analog ResNet Accelerator," in 2020 IEEE 63rd International Midwest Symposium on Circuits and Systems (MWSCAS), Springfield, MA, USA, 2020, pp. 603-606 (invited).

[15] J. J. Vidal, "Toward direct brain-computer communication," *Annual Review of Biophysics and Bioengineering*, 2 (1): 157–80, 1973.

[16] R.A. Norman, E. M. Maynard, P. J. Rousche and D. J. Warren, "A Neural Interface for a Cortical Vision Prosthesis," *Vision Research*, 1999

[17] G. Buzsáki, E. Stark, A. Berényi, D. Khodagholy , D.R. Kipke, E. Yoon and K.D. Wise, " Tools for probing local circuits: high-density silicon probes combined with optogenetics," *Neuron*, 86(1):92-105, April 2015

[18] G. Rios, E. V. Lubenov, D. Chi, M. L. Roukes and A. G. Siapas, "Nanofabricated Neural Probes for Dense 3-D Recordings of Brain Activity," *Nano Letters*, 16 (11), 6857-6862, 2016.

[19] J. J. Jun, N. A. Steinmetz, J. H. Siegle, D. J. Denman, M. Bauza, B. Barbarits, A. K. Lee, C. A. Anastassiou, A. Andrei, Ç. Aydın, M. Barbic, T. J. Blanche, V. Bonin, J. Couto, B. Dutta, S. L. Gratiy, D. A. Gutnisky, M. Häusser, B. Karsh, P. Ledochowitsch, C. M. Lopez, C. Mitelut, S. Musa, M. Okun, M. Pachitariu, J. Putzeys, P. D. Rich, C. Rossant, W. Sun, K. Svoboda, M. Carandini, K. D. Harris, C. Koch, J. O'Keefe and T. D. Harris, "Fully integrated silicon probes for high-density recording of neural activity," *Nature* 551, 232–236, 2017.

[20] M. D. Ferro, C. M. Proctor, A. Gonzalez, E. Zhao, A. Slezia, J. Pas, G. Dijk, M. J. Donahue, A. Williamson, G. G. Malliaras, L. Giocomo and N. A. Melosh, "NeuroRoots, a bio-inspired, seamless Brain Machine Interface device for long-term recording," *bioRxiv* 460949, Nov 2018.

[21] E. Musk and Neuralink, "An integrated brain-machine interface platform with thousands of channels," *bioRxiv* 703801, Jul 2019.

[22] J. Lee, E. Mok, J. Huang, L. Cui, A.-H. Lee, V. Leung, P. Mercier, S. Shellhammer, L. Larson, P. Asbeck, R. Rao, Y.-K. Song, A. Nurmikko, and F. Laiwalla, "An Implantable Wireless Network of Distributed Microscale Sensors for Neural Applications," *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, San Francisco, CA, USA, 2019, pp. 871-874.

[23] J. Lee, F. Laiwalla, J. Jeong, C. Kilfoyle, L. Larson, A. Nurmikko, S, Li, S. Yu, and V. W. Leung, "Wireless Power and Data Link for Ensembles of Sub-mm scale Implantable Sensors near 1GHz," *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, Cleveland, OH, 2018, pp. 1-4.

[24] M. M. Ghanbari, D. K. Piech, K. Shen, S. F. Alamouti, C. Yalcin, B. C. Johnson, J. M. Carmena, M. M. Maharbiz, and R. Muller, "A 0.8mm^3 Ultrasonic Implantable Wireless Neural Recording System With Linear AM Backscattering," *2019 IEEE International Solid-State Circuits Conference - (ISSCC)*, San Francisco, CA, USA, 2019, pp. 284-286.

[25] M. M. Ghanbari et al., "A Sub-mm^3 Ultrasonic Free-Floating Implant for Multi-Mote Neural Recording," *in IEEE Journal of Solid-State Circuits*, vol. 54, no. 11, pp. 3017-3030, Nov. 2019.

[26] S. Lee, A. J. Cortese, P. Trexel, E. R. Agger, P. L. McEuen and A. C. Molnar, "A 330μm×90μm opto-electronically integrated wireless system-on-chip for recording of neural activities," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, San Francisco, CA, 2018, pp. 292-294.

[27] S. Lee, A. J. Cortese, A. P. Gandhi, E. R. Agger, P. L. McEuen and A. C. Molnar, "A 250 μm × 57 μm Microscale Opto-electronically Transduced Electrodes (MOTEs) for Neural Recording," *in IEEE Transactions on Biomedical Circuits and Systems*, vol. 12, no. 6, pp. 1256-1266, Dec. 2018.

[28] E. Moon, M. Barrow, J. Lim, J. Lee, S. Nason, J. Costello, H. S. Kim, C. Chestek, T. Jang, D. Blaauw and J. Phillips, "Bridging the "Last Millimeter" Gap of Brain-Machine Interfaces via Near-Infrared Wireless Power Transfer and Data Communications," ACS Photonics Apr. 2021.

[29] J. Lim, E. Moon, M. Barrow, S.R. Nason, P.R. Ratel, P. G. Patil, S. Oh, I. Lee, H. Kim, D. Sylvester, D. Blaauw, C.A. Chestek, J. Philips, and T. Jang, "A 0.19×0.17mm^2 Wireless Neural Recording IC for Motor Prediction with Near-Inrared-Based Power and Data Telemetry," *in 2020 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2020, pp. 416-417.

[30] J. Lim, J. Lee, E. Moon, M. Barrow, G. Atzeni, J. Letner, J. Costello, S.R. Nason, P.R. Patel, P.G. Patil, H. Kim, C.A. Chestek, J. Phillips, D. Blaauw, D. Sylvester, and T. Jang, "A Light Tolerant Neural Recording IC for Near-Infrared-Powered Free Floating Motes," *in 2021 Symposium on VLSI Circuits (VLSI Circuits)*, to be published.

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

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

[33] T. Jang, M. Choi, S. Jeong, S. Bang, D. Sylvester and D. Blaauw, "A 4.7nW 13.8ppm/°C self-biased wakeup timer using a switched-resistor scheme," *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2016, pp. 102-103.

[34] H. Wang and P. P. Mercier, "A Reference-Free Capacitive-Discharging Oscillator Architecture Consuming 44.4 pW/75.6 nW at 2.8 Hz/6.4 kHz," *in IEEE Journal of Solid-State Circuits*, vol. 51, no. 6, pp. 1423-1435, June 2016.

[35] Y. Lee, B. Giridhar, Z. Foo, D. Sylvester and D. B. Blaauw, "A Sub-nW Multi-stage Temperature Compensated Timer for Ultra-Low-Power Sensor Nodes," *in IEEE Journal of Solid-State Circuits*, vol. 48, no. 10, pp. 2511-2521, Oct. 2013.

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

[37] P. M. Nadeau, A. Paidimarri and A. P. Chandrakasan, "Ultra Low-Energy Relaxation Oscillator With 230 fJ/cycle Efficiency," *in IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 789-799, April 2016.

[38] M. Horowitz, "1.1 Computing's energy problem (and what we can do about it)," *in 2014 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2014, pp. 10–14.

[39] B. E. Boser, E. Sackinger, J. Bromley, Y. L. Cun, and L. D. Jackel, "An analog neural network processor with programmable topology," *IEEE Journal of Solid-State Circuits*, vol. 26, no. 12, pp. 2017–2025, Dec. 1991.

[40] L. Fick, D. Blaauw, D. Sylvester, S. Skrzyniarz, M. Parikh, and D. Fick,"Analog in-memory subthreshold deep neural network accelerator," *in 2017 IEEE Custom Integrated Circuits Conference (CICC)*, 2017, pp. 1–4.

[41] J. Zhang, Z. Wang, and N. Verma, "In-Memory Computation of a Machine-Learning Classifier in a Standard 6T SRAM Array," *IEEE Journal of Solid-State Circuits*, vol. 52, no. 4, pp. 915–924, Apr. 2017.

[42] J. Oh, G. Kim, B.-G. Nam, and H.-J. Yoo, "A 57 mW 12.5 μJ/Epoch Embedded Mixed-Mode Neuro-Fuzzy Processor for Mobile Real-Time Object Recognition," *IEEE Journal of Solid-State Circuits*, vol. 48, no. 11, pp. 2894–2907, Nov. 2013.

[43] J.-Y. Kim, M. Kim, S. Lee, J. Oh, K. Kim, and H.-J. Yoo, “A 201.4 GOPS 496 mW Real-Time Multi-Object Recognition Processor With Bio-Inspired Neural Perception Engine,” *IEEE Journal of Solid-State Circuits*, vol. 45, no. 1, pp. 32–45, Jan. 2010.

[44] S. K. Gonugondla, M. Kang, and N. Shanbhag, “A 42pJ/decision 3.12TOPS/W robust in-memory machine learning classifier with on-chip training,” *in 2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, 2018, pp. 490–492.

[45] A. Amaravati, S. B. Nasir, J. Ting, I. Yoon, and A. Raychowdhury, “A 55-nm, 1.0–0.4V, 1.25-pJ/MAC Time-Domain Mixed-Signal Neuromorphic Accelerator With Stochastic Synapses for Reinforcement Learning in Autonomous Mobile Robots,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 1, pp. 75–87, Jan. 2019.

[46] S. Gopal, P. Agarwal, J. Baylon, L. Renaud, S. N. Ali, P. P. Pande, and D. Heo, “A Spatial Multi-Bit Sub-1-V Time-Domain Matrix Multiplier Interface for Approximate Computing in 65-nm CMOS,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 506–518, Sep. 2018.

[47] M. Kang, S. Lim, S. Gonugondla, and N. R. Shanbhag, “An In-Memory VLSI Architecture for Convolutional Neural Networks,” *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 8, no. 3, pp. 494–505, Sep. 2018.

[48] R. S. Amant, A. Yazdanbakhsh, J. Park, B. Thwaites, H. Esmaeilzadeh, A. Hassibi, L. Ceze, and D. Burger, “General-purpose code acceleration with limited-precision analog computation,” *in 2014 ACM/IEEE 41st International Symposium on Computer Architecture (ISCA)*, 2014, pp. 505–516.

[49] K. Jia, Z. Liu, Q. Wei, F. Qiao, X. Liu, Y. Yang, H. Fan, and H. Yang, “Calibrating Process Variation at System Level with In-Situ Low-Precision Transfer Learning for Analog Neural Network Processors,” *in 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 2018, pp. 1–6.

[50] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, ‘‘Gradient-based learning applied to document recognition,’’ *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[51] M. Courbariaux, I. Hubara, D. Soudry, R. El-Yaniv, and Y. Bengio, “Binarized Neural Networks: Training Deep Neural Networks with Weights and Activations Constrained to +1 or -1,” *arXiv:*1602.02830 [cs], Feb. 2016.

[52] S. Jeloka, N. B. Akesh, D. Sylvester, and D. Blaauw, “A 28 nm Configurable Memory (TCAM/BCAM/SRAM) Using Push-Rule 6T Bit Cell Enabling Logic-in-Memory,” *in IEEE Journal of Solid-State Circuits*, vol. 51, no. 4, pp. 1009–1021, Apr. 2016.

[53] Z. Jiang, S. Yin, M. Seok, and J. Seo., “XNOR-SRAM: In-Memory Computing SRAM Macro for Binary/Ternary Deep Neural Networks,” *in 2018 IEEE Symposium on VLSI Technology*, 2018, pp. 173–174.

[54] R. Liu, X. Peng, X. Sun, W-S. Khwa, X. Si, J.-J. Chen, J.-F. Li, M.-F. Chang, and S. Yu, “Parallelizing SRAM Arrays with Customized Bit-Cell for Binary Neural Networks,” *in 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC)*, 2018, pp. 1–6.

[55] D. Bankman, L. Yang, B. Moons, M. Verhelst, and B. Murmann, “An Always-On 3.8 μJ/86% CIFAR-10 Mixed-Signal Binary CNN Processor With All Memory on Chip in 28-nm CMOS,” *IEEE Journal of Solid-State Circuits*, pp. 1–15, 2018.

[56] H. Valavi, P. J. Ramadge, E. Nestler, and N. Verma, “A 64-Tile 2.4-Mb In-Memory-Computing CNN Accelerator Employing Charge-Domain Compute,” *IEEE Journal of Solid-State Circuits*, vol. 54, no. 6, pp. 1789–1799, Jun. 2019.

[57] M. Rastegari, V. Ordonez, J. Redmon, and A. Farhadi, “XNOR-Net: ImageNet Classification Using Binary Convolutional Neural networks,” *arXiv*:1603.05279, Mar. 2016.

[58] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, “Reading digits in natural images with unsupervised feature learning”, *in NIPS*, page 5, 2011.

[59] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images”, *Citeseer*, 2009

[60] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *arXiv*:1512.03385 [cs], Dec. 2015.

[61] K. He, X. Zhang, S. Ren, and J. Sun, “Identity Mappings in Deep Residual Networks,” *arXiv*:1603.05027 [cs], Mar. 2016.

[62] R. E. Vallee and E. I. El-Masry, “A very high-frequency CMOS complementary folded cascode amplifier,” *IEEE Journal of Solid-State Circuits*, vol. 29, no. 2, pp. 130–133, Feb. 1994.

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

[64] C. M. Lopez, S. Mitra, J. Putzeys, B. Raducanu, M. Ballini, A. Andrei, S. Severi, M. Welkenhuysen, C. V. Hoof, S. Musa, and R. F. Yazicioglu, "A 966-electrode neural probe with 384 configurable channels in 0.13µm SOI CMOS," *2016 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2016, pp. 392-393.

[65] B. C. Johnson, S. Gambini, I. Izyumin, A. Moin, A. Zhou, G. Alexandrov, S. R. Santacruz, J. M. Rabaey, J. M. Carmena, and R. Muller, "An implantable 700µW 64-channel neuromodulation IC for simultaneous recording and stimulation with rapid artifact recovery," *2017 Symposium on VLSI Circuits*, Kyoto, 2017, pp. C48-C49.

[66] H. Chandrakumar and D. Markovic, "A 2.8µW 80mVpp-linear-input-range 1.6GΩ-input impedance bio-signal chopper amplifier tolerant to common-mode interference up to 650mVpp," *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2017, pp. 448-449.

[67] C. Kim, S. Joshi, H. Courellis, J. Wang, C. Miller and G. Cauwenberghs, "A 92dB dynamic range sub-µVrms-noise 0.8µW/ch neural-recording ADC array with predictive digital autoranging," *2018 IEEE International Solid - State Circuits Conference - (ISSCC)*, San Francisco, CA, 2018, pp. 470-472.

[68] Z. T. Irwin, D. E. Thompson, K. E. Schroeder, D. M. Tat, A. Hassani, A. J. Bullard, S. L. Woo, M. G. Urbanchek, A. J. Sachs, P. S. Cederna, W. C. Stacey, P. G. Patil, and C. A. Chestek, "Enabling Low-Power, Multi-Modal Neural Interfaces Through a Common, Low-Bandwidth Feature Space," *in IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 5, pp. 521-531, May 2016.

[69] Z. T. Irwin, K. E. Schroeder, P. P. Vu, A. J. Bullard, D. M. Tat, C. S. Nu, A. Vaskov, S. R. Nason, D. E. Thompson, J. N. Bentley, P. G. Patil, and C. A. Chestek, "Neural control of finger movement via intracortical brainmachine interface," *Journal of Neural Engineering*, vol. 14, no. 6, p. 066004, Dec. 2017.

[70] X. Wu, I. Lee, Q. Dong, K. Yang, D. Lim, J. Wang, Y. Peng, Y. Zhang, M. Saligane, M. Yasuda, K. Kumeno, F. Ohno, S. Miyoshi, M. Kawaminami, D. Sylvester, and D. Blaauw, "A 0.04mm^{3}16nW Wireless and Batteryless Sensor System with Integrated Cortex-M0+

Processor and Optical Communication for Cellular Temperature Measurement," *2018 IEEE Symposium on VLSI Circuits*, Honolulu, HI, 2018, pp. 191-192.

[71] E. Stark, and M. Abeles, "Predicting Movement from Multiunit Activity," *Journal of Neuroscience*, vol. 27, no. 31, pp. 8387-8394, Aug. 2007.

[72] P. R. Patel, H. Zhang, M. T. Robbins, J. B. Nofar, S. P. Marshall, M. J. Kobylarek, T. D. Kozai, N. A. Kotov, and C. A. Chestek, "Chronic In Vivo Stability Assessment of Carbon Fiber Microelectrode Arrays," *Journal of Neural Engineering*, vol. 13, no. 6, p. 066002, Dec. 2016.

[73] K. Yang, Q. Dong, D. Blaauw and D. Sylvester, "A 553F^2 2-transistor amplifier-based Physically Unclonable Function (PUF) with 1.67% native instability," *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, San Francisco, CA, 2017, pp. 146-147.

[74] S.R. Nason, A.K. Vaskov, M.S. Willsey, E. J. Welle, H. An, P.P. Vu, A.J. Bullard, C.S. Nu, J.C. Kao, K.V. Shenoy, T. Jang, H.-S. Kim, D. Blaauw, P.G. Patil, and C.A. Chestek, "A low-power band of neuronal spiking activity dominated by local single units improves the performance of brain–machine interfaces," *in Nature Biomedical Engineering*, vol. 4, pp. 973–983, July 2020.

[75] N. Verma, A. Shoeb, J. Bohorquez, J. Dawson, J. Guttag and A. P. Chandrakasan, "A Micro-Power EEG Acquisition SoC With Integrated Feature Extraction Processor for a Chronic Seizure Detection System," *in IEEE Journal of Solid-State Circuits*, vol. 45, no. 4, pp. 804-816, April 2010.