AN ULTRA-LOW-POWER IMPLANTABLE CHIP FOR SAFE NEUROSTIMULATION

Jiahe Chen

A THESIS

in

Electrical and Systems Engineering

Presented to the Faculties of the University of Pennsylvania in Partial
Fulfillment of the Requirements for the Degree of Master of Science in Engineering

2019

_____
Prof. Jan Van der Spiegel
Supervisor of Thesis

_____
Prof. Firooz Aflatouni
Supervisor of Thesis

_____
Prof. Victor Preciado
Graduate Group Chair

Acknowledgement

I would like to express my gratitude to my two research advisors: Prof. Jan Van der Spiegel and Prof. Firooz Aflatouni. They brought me to the field of integrated circuits and taught me how to conduct extraordinary research. I would like to thank Prof. Andrew G. Richardson from Penn Med, who helped me understand the neurological background of this project and offered many fantastic suggestions. Thanks to their strong support, I accomplished this project and gained a deep understanding in brain-machine interface. I believe that I will keep benefiting from this research experience for my future academic career.

I would also like to thank my colleagues in Electronic Photonics Microsystems Lab and Center for Sensor Technologies. I would like to thank Mr. Han Hao, Mr. Zhe Xuan, Mr. Vasant Iyer, Mr. Pouria Sanjari, Mr. Mohamad Hossein Idjadi, and many others. Without their help, I cannot finish this thesis.

I am especially grateful to my parents, for their unconditional love and tremendous support for my study at Penn. Thank you for bearing my inability to be with you for the past two years. I wish I could have spent more time with you.

Last but not least, I would like to thank my friends in Philadelphia, especially my friends from the University City Chinese Christian Church. This friendship is invaluable for my life.

Contents

List of Figures

List of Tables

Chapter 1 **Introduction**

1.1 Background and Motivation

Tremendous advances in circuit design techniques and silicon technology make silicon-based integrated circuits (IC) one of the most powerful platforms for realizing extremely complex systems on a small scale. These advances also drive future innovations in IC to take place at the system level and become more application-oriented. Benefiting from these innovations, electrical engineering has been able to create new modalities for interacting with biological systems, what is having an impact on biological sciences like no discipline has had before. [1] Neuroscience and neuroengineering pose particularly interesting and challenging research problems, such as how to use electronics to record and decode weak and noisy neural signals. A Brain-Machine Interface (BMI) which creates an artificial direct pathway between the brain and external devices was therefore proposed to tackle such kind of problems. [2] With rapid development since 1970s, BMI gains broad applications in fundamental neuroscience research, neural disease treatment and other high-impact fields. [3] Recently emerging areas of BMI research include untethered closed-loop bidirectional BMI system, application of BMI to early detection and prevention of neurological disorders, and implantable high-density distributed BMI systems. [2] [4] [5]

Inspired by previous work mentioned before, this thesis focuses on one critical component of the BMI system, the neural stimulator. Implantable electrical neural stimulator interfaces with neurons through electrodes and delivers low electrical energy to artificially excite neurons. One popular type of electrical stimulator is the current-regulated stimulator, due to its safety and well-controlled stimulation energy. However, one major drawback of such type of stimulator is its poor energy efficiency. For implantable devices, any waste of energy degrades the usability. Over the past decade, numerous circuit and system innovations have been proposed to improve the power efficiency of the current-regulated stimulator. [6] [7] [8] However, most of these techniques focus on improving the circuit efficiency of the stimulator. In this work, we try to solve this problem from the perspective of the biological system. We propose a closed-loop neurostimulation system that can automatically find the minimum stimulation current that is needed to evoke an action potential from neurons. We can then use the found minimum stimulation current to stimulate neurons with the least power consumption and safety risk later. Details about the proposed system are described in the next section.

1.2 Introduction to the Proposed System

The basic architecture and operational scheme of the proposed closed-loop neurostimulation system are shown in Figure 1.1. The operation of the system contains two phases: the stimulation phase executed by a digitally controlled stimulator, and the detection phase executed by an analog front-end (AFE) and a spike detector. Peripheral circuitry, including a stimulation timing generation module (STGM) and a counter, coordinates these two phases. The AFE and the stimulator interphase with the tissue through a same electrode. Therefore, a closed-loop control system is formed.

The operational scheme is to perform stimulation and detection sequentially and repeatedly, and gradually increase the stimulation current until an action potential is observed. As shown in Figure 1.1, the system first initializes the stimulation current (I_{Stim}) at the minimum value and then starts the first stimulation. After the stimulation phase, the system turns the stimulator off and starts the detection phase. The neural signal from the tissue gets amplified by the AFE and is then analyzed by the spike detector. If the system does not detect any action potential during the detection phase, I_{Stim} will increase by one least significant bit (LSB) and the system will start another stimulation and repeat subsequent tasks. If the system detects an action potential, it will recognize the current value of I_{Stim} as the minimum needed stimulation current and then stop the searching process.

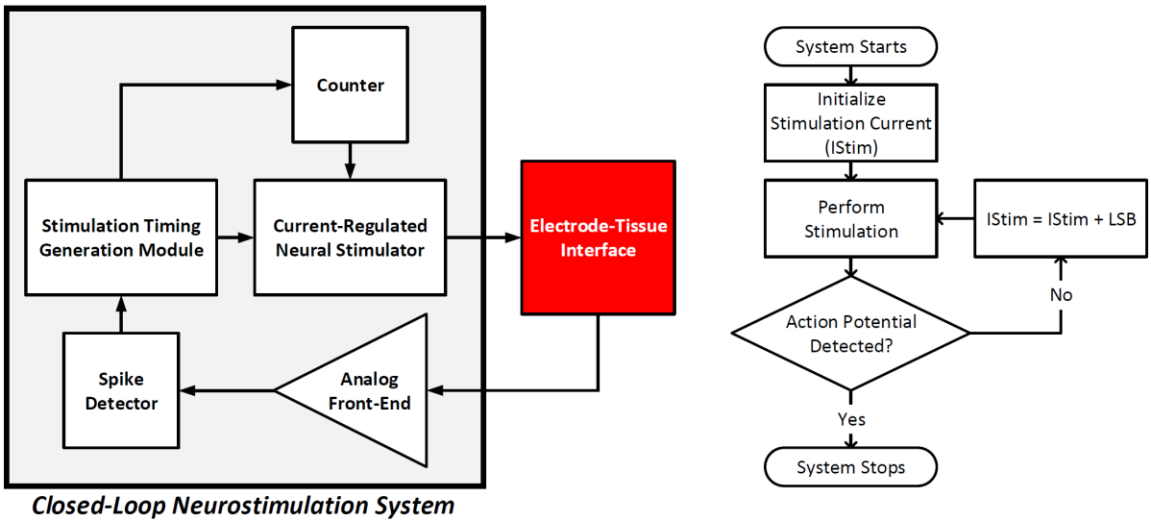

Figure 1.1: Left: Basic architecture of the proposed system. Right: Basic operation flow of the proposed system.

For the timing of the operation, one stimulation-detection process takes 2ms. The 4-bit digitally controlled stimulator provides 15 different current amplitudes. Therefore, one complete operation (assuming an action potential is detected) takes maximum 30ms.

In terms of the circuit implementation, the system does not pose extreme requirements for the AFE and the spike detector since no complex signal processing like spike classification is involved. Also, since the stimulation phase and the detection phase

happen sequentially, the stimulation artifact is not severe. This further alleviates the requirement for the AFE. However, the system does have rigorous requirements for the neural stimulator. Since several stimulations will be performed in a short duration, the stimulator needs to have excellent charge-balancing performance. Since we want to know the accurate value of delivered current, the stimulator also needs to have accurate control over the stimulation current amplitude. All these implementation details will be presented in later chapters.

The system is designed in GlobalFoundries 180nm silicon-on-insulator (SOI) process. Four types of transistors are used. Their circuit symbols are shown in Figure 2.1. For this work, generally T-body transistors are used for analog circuits and floating-body transistors are used for digital circuits.

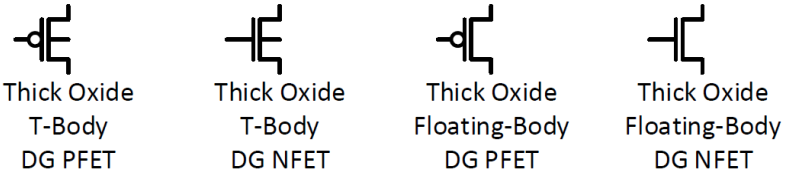

Figure 1.2: Transistors used in the design.

1.3 Outline of the Thesis

The thesis is organized as follows: Chapter 2 to Chapter 5 present the design and implementation of each block of the system. The sequence follows the signal flow in the block diagram: from the AFE to the stimulator and finally the peripheral circuitry. Chapter 6 describes the integration and implementation of the system. Chapter 7 is an independent chapter. A low-power linear-phase delay filter, which is part of a neural spike classification chip, is presented in this chapter.

Chapter 2 **Analog Front-End**

2.1 Introduction to Neural Signal Recording

The analog front-end (AFE) in a BMI system is an essential block for converting neural signals to more processable electrical signals on the chip. The first stage of the AFE is normally a neural amplifier which extracts and conditions microvolt to millivolt scale signals aroused from weak neural activities. [9] Based on Friis formulas for noise, this stage makes dominant contribution to the overall noise of the AFE and thereby should have particularly low noise level. [10] The later stages are optional and depend on system functions. For our system, the AFE has two stages and the second stage is a programmable gain amplifier (PGA), which can be used to further amplify the neural signal and adjust the signal amplitude for later signal processing. The conceptual circuit diagram of the designed AFE is shown in Figure 2.1.

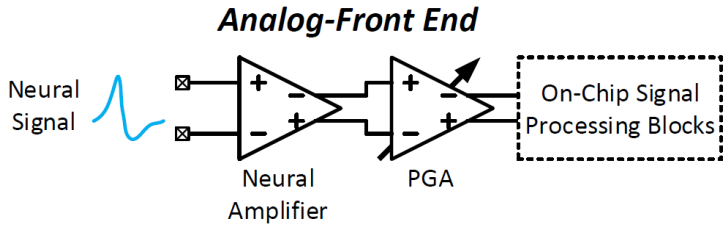

Figure 2.1: Conceptual circuit diagram of the AFE.

This chapter describes the design of a general-purpose analog front-end for recording action potentials from neurons. The chapter is organized as follows: Section 2.2 presents the overall architecture of the AFE. Section 2.3 describes the circuit implementation. Section 2.4 shows the performance of the AFE from Cadence simulations.

2.2 AFE Architecture

The overall circuitry of the designed AFE is shown in Figure 2.2. The neural amplifier stage is based on a widely used topology proposed by Harrison et al. and Olsson III et al. [11] [12] The topology utilizes capacitive feedback loops to set up the closed-loop gain and large MOS pseudo-resistors to set up the DC feedback. The PGA stage is based on the same topology but with variable capacitors for gain tuning. All four variable capacitor are controlled by digital signals $B_{0\sim4}$.

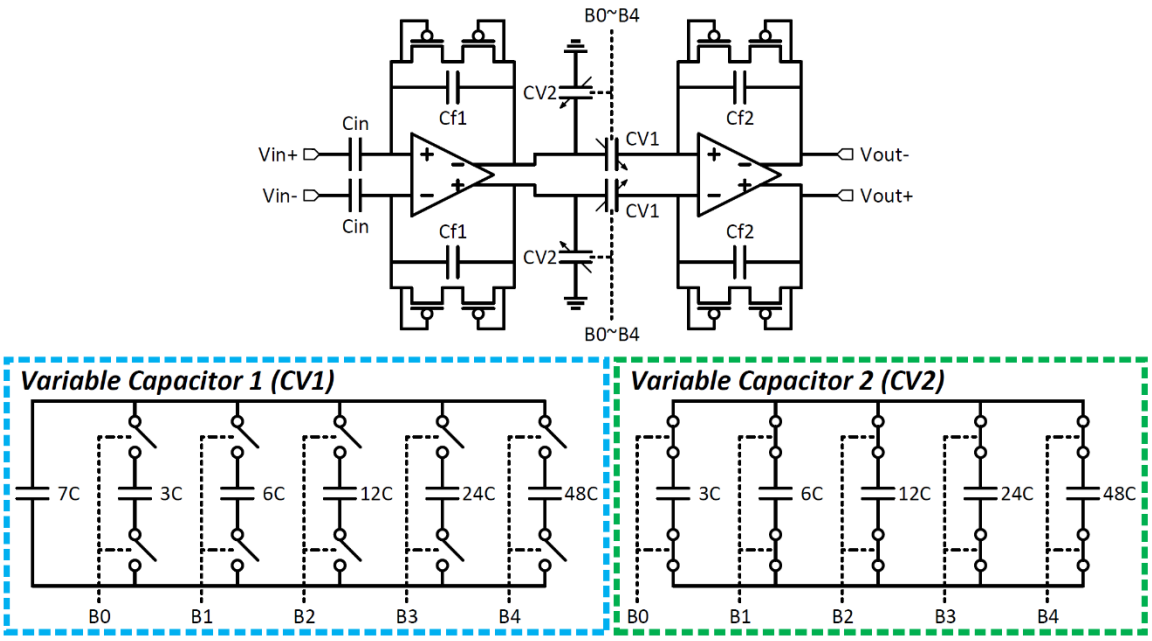

Figure 2.2: Architecture of the AFE.

To analyze the circuit, consider the simplified circuit diagram of each stage shown in Figure 2.3. Since in the recording setup, V_{in+} of the AFE is connected to the working electrode and V_{in-} is connected to the reference electrode, we can assume that the positive input of the single-ended OTA shown in the graph is connected to the signal ground.

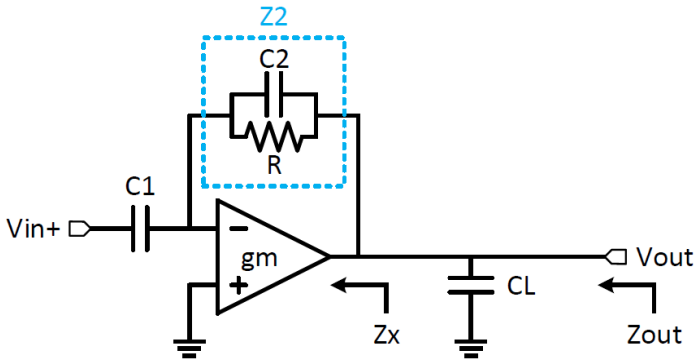

Figure 2.3: Simplified circuit diagram of each stage.

Suppose that the OTA has a voltage gain of A_v, the closed-loop gain is then given by:

$$G = -\frac{A_v Z_2 s C_1}{(A_v + 1) + Z_2 s C_1} = \frac{-\frac{A_v C_1}{C_1 + C_2 + A_v C_2}}{1 + \frac{1 + A_v}{sR(C_1 + C_2 + A_v C_2)}} = -\frac{G_M}{1 + \frac{\omega_L}{s}},$$

where G_M is the mid-band gain and ω_L is the lower cutoff frequency or the high-pass corner frequency. Assume that the gain is large, and the following criterion is met, we have:

$$A_v C_2 \gg C_1 + C_2 \Longrightarrow \begin{cases} G_M \approx -\dfrac{C_1}{C_2} \\ \omega_L \approx \dfrac{1}{RC_2} \end{cases}.$$

We can see that the closed-loop gain is set by the ratio of C_1 to C_2 and the lower cutoff frequency depends on R. Therefore, the resistance of the MOS pseudo-resistor is normally very large so that ω_L is in Hz range. Another design consideration is that the DC gain A_v of the OTA must be sufficiently larger than the desired closed-loop gain. Otherwise the closed-loop gain of the circuit will have a considerable error. For this design, the closed-loop gain of the neural amplifier is 40dB. To have a gain error less than 10%, the DC gain of the OTA should be at least 60dB.

To find the upper cutoff frequency or the low-pass corner frequency, we assume that the OTA has a transconductance g_m and an extremely high output impedance. The output impedance of the circuit is then:

$$Z_{out} = \frac{1}{sC_L} || Z_x,$$

$$Z_x = \frac{1 + sC_1 Z_2}{g_m + sC_1}.$$

Suppose that at higher frequency, the impedance of C_2 dominates over R. Also assume that C_1 is much larger than C_2. Then we have:

$$Z_x \approx \frac{G_M}{sC_1 + g_m} \Longrightarrow Z_{out} = \frac{1}{\dfrac{sC_1 + g_m}{G_M} + sC_L} \Longrightarrow A_v = g_m Z_{out} = \frac{g_m}{\dfrac{sC_1 + g_m}{G_M} + sC_L}.$$

Since A_v is normally very large, the closed-loop gain can be approximated as follows:

$$G \approx \frac{-1}{\dfrac{1}{sZ_2C_1} + \dfrac{1}{A_v}} = \frac{-G_M}{1 + \dfrac{G_M}{g_m}\left(\dfrac{sC_1 + g_m}{G_M} + sC_L\right)}.$$

Suppose that G_M is large and C_1 and C_L are comparable. Then we have:

$$\frac{sC_1 + g_m}{G_M} + sC_L \approx sC_L \Longrightarrow G \approx \frac{-G_M}{1 + \dfrac{sG_M C_L}{g_m}} = \frac{-G_M}{1 + \dfrac{s}{\omega_H}},$$

where ω_H is the upper cutoff frequency. Since ω_L is normally very small compared with ω_H, the bandwidth of the neural amplifier is therefore given by:

$$BW \approx \omega_H = \frac{g_m}{G_M C_L} = \frac{g_m C_2}{C_1 C_L}.$$

Normally G_M and g_m are set by other design requirements. Therefore, the bandwidth mainly depends on the loading capacitor C_L.

For this design the ratio of C_1 to C_2 is 100 to realize a closed-loop gain of 40dB. For the neural amplifier, we made C_{f1} and C_{in} equal to 100fF and 10pF respectively. For the PGA, C_{f2} is 50fF and C_{V1} changes from 0.35pF to 5pF with a tuning step of 0.15pF. The capacitive feedback of the neural amplifier has larger capacitance to reduce the effect of parasitic capacitors on the closed-loop gain. The capacitance tuning of C_{V1} enables the gain of the AFE to be changed from 0.7kV/V to 10kV/V with a tuning step of 0.3kV/V. This wide linear change enables accurate control over the signal amplitude. Conventionally, for the PGA, C_{f2} is designed as the variable capacitor to realize the tunable gain. [13] With such design, the gain of the AFE can hardly be linearly controlled. However, for our design, the change of C_{V1} means the change of loading capacitors of the neural amplifier since the input impedance of the PGA is approximately the impedance of C_{V1}. Therefore, change of C_{V1} will ultimately affect the bandwidth of the AFE. To compensate this change, a second variable capacitor C_{V2} is added to the circuitry to make the loading capacitors of the neural amplifier and hence the bandwidth of the AFE constant during gain tuning. The transconductance of the OTA of the neural amplifier is 16.71μS. With the loading capacitance of 5pF, the upper cutoff frequency is around 5.3kHz.

The MOS pseudo-resistor is realized by two 0.5μm/0.5μm diode-connected floating-body PMOS transistors. Simulation shows that such structure has the highest linearity of the resistance and is reciprocal due to its symmetry. The typical average resistance is 976GΩ. Therefore, the lower cutoff frequency of the AFE is around 3.3Hz. The bandwidth of the AFE is then 3.3Hz ~ 5.3kHz. Notice that this is only the theoretical value based on the circuit model. The implemented AFE normally has slightly smaller bandwidth due to parasitic capacitors.

2.3 Implementation of the AFE

This section describes details about the implementation of the AFE architecture, including the design and noise analysis of the OTA of the neural amplifier and the design of the OTA of the PGA. Specifications of the neural amplifier and the PGA will also be shown.

2.3.1 OTA of the Neural Amplifier

The OTA of the neural amplifier is the core of the AFE since it determines the overall performance of the block. There are several tradeoffs for the design: noise, power, gain, bandwidth, stability, output range and immunity to PVT variations. Among these tradeoffs, the noise performance is of the upmost importance. The input-referred noise of the neural amplifier is given by the equation below: [11]

$$\overline{v_{ni,NA}^2} = \left(\frac{C_{in} + C_{f1} + C_{in,OTA}}{C_{in}}\right)^2 \overline{v_{ni,OTA}^2},$$

where $C_{in,OTA}$ is the input capacitance of the OTA and $\overline{v_{ni,OTA}^2}$ is the input-referred noise of the OTA. We can see that with small $C_{in,OTA}$ and large G_M, the input-referred noise of the neural amplifier is approximately equal to the input-referred noise of its OTA. For a neural amplifier, the minimum required input-referred rms noise voltage is 50μV. [14] The typical extracellular neural background noise over the bandwidth of the action potential is 5μV ~ 10μV. [8] The goal of this design is to have an input-referred noise less than 5μV.

After considering all those tradeoffs, we choose to implement the OTA by a fully differential telescopic amplifier for the following reasons:

1) the Telescopic structure only has two current branches. This enables higher current through the input transistors without consuming too much power. Higher current provides larger g_m and hence larger gain and lower thermal noise. The overall noise performance of this structure is also better than the current mirror OTA or the two-stage amplifier.

2) The fully differential structure rejects differential noises from electrodes (such as background neural signal). Simulation also shows that for single-ended structure, the common-mode voltage of the output has a high chance of drifting away from the applied V_{CM} when the output swing is high. This is possibly due to the asymmetry of the structure and the large resistance of MOS pseudo-resistors.

3) It is easier to perform frequency compensation of this structure since it does not have the mirror pole. At low current, the mirror pole could be very close to the dominant pole, making it hard to compensate the circuit. If designed well, the frequency compensation of this structure can be done by simply placing loading capacitors at outputs to shift its dominant pole to higher frequency. This relates the phase margin of the OTA to the bandwidth of the neural amplifier. For our design, the loading capacitance is 5pF and is sufficient for compensating the OTA as well.

4) The telescopic structure provides higher gain which is essential for minimizing the error of closed-loop gain.

The proposed structure also has following tolerable disadvantages:

1) The telescopic structure consumes considerable voltage headroom. For the 180nm silicon-on-insulator (SOI) process we used, when the supply voltage is lower than 3V, although the design is still possible, the designed OTA does not have much immunity to PVT variations. At some process corners, the OTA simply fails to work.

2) The fully differential structure requires common-mode feedback (CMFB) circuit, which consumes power and may cause stability issues. The CMFB also decreases the bandwidth and increases the noise.

3) The telescopic structure has limited output range. For this design, the issue is alleviated since the maximum amplitude of the action potential is only around 500μV. [15]

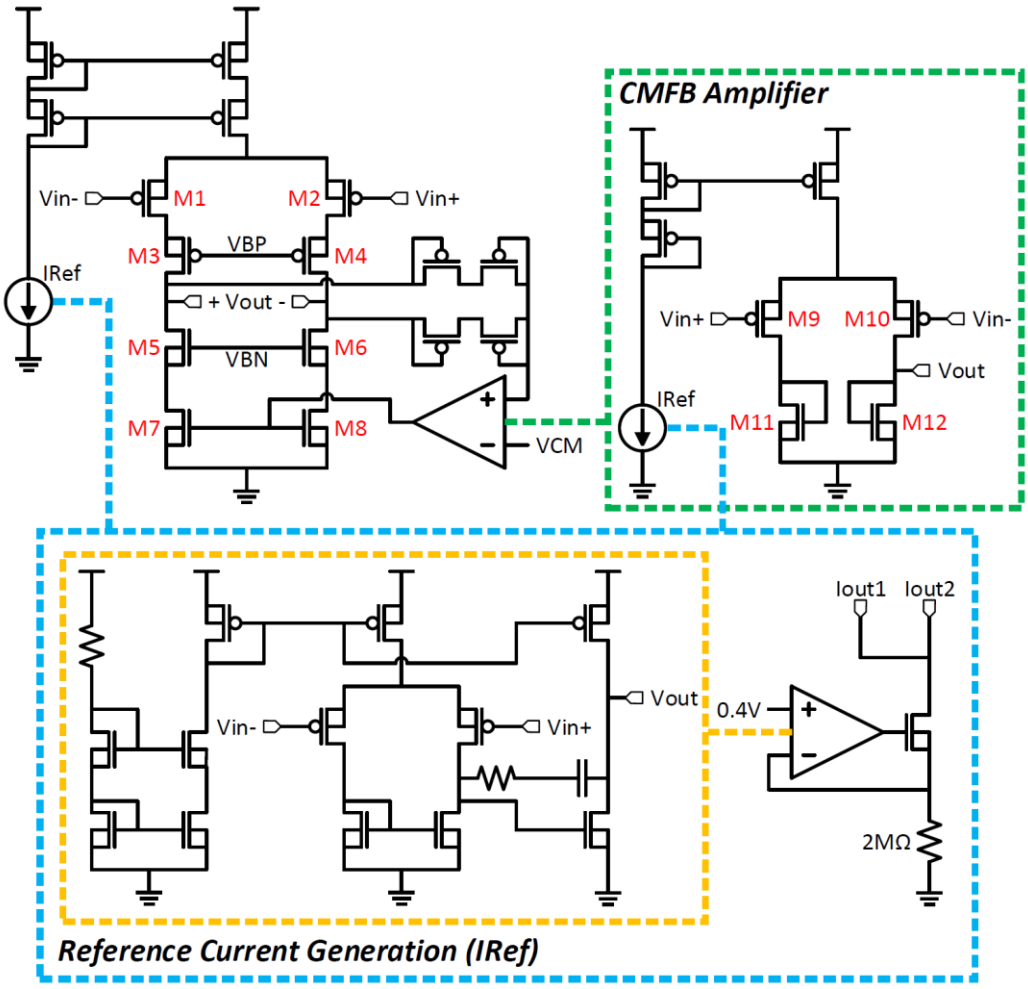

Figure 2.4: OTA of the neural amplifier.

Transistor	W/L (μm)
$M_{1,2}$	51/3
$M_{3,4}$	30/8
$M_{5,6}$	1/20
$M_{7,8}$	5 × 0.5/60
$M_{9,10}$	1/2
$M_{11,12}$	0.5/60

Table 2.1: Transistor sizing of the telescopic amplifier.

The final design of the OTA is shown in Figure 2.4. The sizing of some important transistors is shown in Table 2.1. The DC gain of the OTA is given by:

$$A_v = g_{m1}(r_n||r_p) \approx g_{m1}[(g_{m3}r_{o3}r_{o1})||(g_{m5}r_{o5}r_{o7})],$$

where r_n is the output impedance looking into the drain of $M_{5,6}$ and r_p is the output impedance looking into the drain of $M_{3,4}$. To avoid introducing low-frequency nondominant poles, r_n and r_p should have close values. To have a higher gain, we need to maximize $g_{m1,2}$, which can be done by operating $M_{1,2}$ in subthreshold region and increasing their aspect ratio. [11] We can also increase the size of $M_{3,4}$ and $M_{5,6}$ to boost the gain based on the following equation:

$$\begin{cases} g_m r_o = \dfrac{\sqrt{2\mu C_{OX}\dfrac{W}{L}I_D}}{\lambda I_D} \\ \lambda \propto \dfrac{1}{L} \end{cases} \Longrightarrow g_m r_o \propto \sqrt{\frac{WL}{I_D}}.$$

The input-referred noise of the OTA is given by: [11]

$$\overline{v_{ni,OTA}^2} = 2\left[\frac{4K_BT}{g_{m1}}\left(\frac{2}{3}\right)\left(1+\frac{g_{m7}}{g_{m1}}\right)\right] + 2\left[\frac{K_P}{C_{OX}W_1L_1} + \frac{K_N}{C_{OX}W_7L_7}\left(\frac{g_{m7}}{g_{m1}}\right)^2\right]\frac{1}{f},$$

where K_B is Boltzmann's constant, T is the absolute temperature, and K_N and K_P are the flicker noise coefficients of NMOS and PMOS transistors. The first part of the equation is the thermal noise and the second part is the flicker noise. We can see that larger g_{m1} provides lower overall noise and larger gain simultaneously. Therefore, the aspect ratio of $M_{1,2}$ should be as large as possible. However, the size of $M_{1,2}$ should not be too large since large input capacitance of the OTA increases the overall noise level of the neural amplifier. We can also see that to reduce the noise, g_{m7} should be small and the size of $M_{7,8}$ should be large. This means that the aspect ratio of $M_{7,8}$ should be as small as possible. However, the aspect ratio of $M_{7,8}$ should not be too small since large $r_{o7,8}$ may introduce low-frequency nondominant poles.

The CMFB circuit is formed by two MOS pseudo-resistors for sensing the common-mode voltage of the output and a CMFB amplifier for comparing the sensed signal to the applied V_{CM}. The CMFB circuit has two poles, which are located at the input and output of the CMFB amplifier respectively. To lower the effect of these two poles. The resistance of MOS pseudo-resistors should not be too high. The input capacitance and output impedance of the CMFB amplifier should be as low as possible. For this design, each MOS pseudo-resistor is realized by two 20μm/0.5μm floating-body PMOS transistors. The resistance is around 127GΩ.

A stable and accurate reference current is of critical importance for the noise performance of the OTA. Any drastic change of the reference current will affect the g_m and hence the noise of the circuit significantly. Therefore, a reference current generation circuit formed by regulated cascode stage is used. Its output current only depends on the source resistor

and the bias voltage applied. This type of circuit has strong immunity to PVT variations. Detailed simulation can be found in Chapter 4, Section 4.4.3. For this design, the reference current I_{Ref} is 100nA. The current of the telescopic amplifier is 2μA and of the CMFB amplifier is 400nA.

Specifications of the neural amplifier are shown in Table 2.2. The overall performance of the neural amplifier is evaluated by the noise efficiency factor (NEF) and the power efficiency factor (PEF) given below:

$$NEF = V_{ni,rms}\sqrt{\frac{2I_{tot}}{\pi \times \frac{K_B T}{q} \times 4K_B T \times BW}},$$

$$PEF = NEF^2 V_{DD},$$

where $V_{ni,rms}$ is the rms input-referred noise of the neural amplifier, I_{tot} is the total bias current, q is the elementary charge and V_{DD} is the supply voltage. From Table 2.2, we can see that the integrated noise of the neural amplifier is slightly higher than 5μV. This is due to that fact that the lower cutoff frequency of the neural amplifier is very close to the flicker noise corner frequency. For the overall AFE block, the integrated noise will be lower since its lower cutoff frequency is farther away from the flicker noise corner frequency. This will be shown in Section 2.4. The equation of the noise density is given below: [2]

$$Noise\ Density = \frac{V_{ni,rms}}{\sqrt{BW \times \frac{\pi}{2}}}.$$

Power	8.82μW
Supply Voltage	3V
Bias Current (Including CMFB)	2.6μA
Open-Loop Gain	83dB
Closed-Loop Gain	40dB
Gain Error	0.71%
Bandwidth (5pF Loading Capacitor)	1.44Hz ~ 4.78kHz
Input Range	$7.2mV_{pp}$
Integrated Noise over the Bandwidth	5.18μV
Noise Density	$59.78nV/Hz^{1/2}$
NEF	4.66
PEF	32.52

Table 2.2: Specifications of the neural amplifier.

2.3.2 OTA of the PGA

For the OTA of the PGA, the noise requirement is much less rigorous given sufficiently large gain of the neural amplifier. Some important design specifications include: power, output range, and gain error. For this design, the OTA is implemented by a fully differential two-stage amplifier for the following reasons:

1) This structure provides large output swing since the output stage has adequate voltage headroom.

2) It is easier to compensate this structure, especially at low current. However, the designer should make sure that the dominant pole occurs at the output of the first stage. Otherwise, Miller compensation will not be effective.

3) Since it is a cascaded structure, it is easier to achieve high gain without consuming too much power. More specifically, the gain required for each stage is not very high and hence the g_m requirement for each transistor is alleviated.

The circuit of the OTA is shown in Figure 2.5. Sizing of some important transistors is shown in Table 2.3. Input transistors $M_{1,2}$ are operating in subthreshold region to maximize the gain. Miller compensation is applied, and the phase margin is around 60°. The CMFB circuit is applied to the second stage only to avoid introducing too many poles in the feedback loop. The common-mode voltage of the output is sensed by two MOS pseudo-resistors realized by 20µm/0.5µm floating-body PMOS transistors. To set the common-mode voltage close to the applied V_{CM}, the CMFB amplifier needs to have a moderate gain. Therefore, a one-stage current mirror amplifier is used. Specifications of the PGA are shown in Table 2.4.

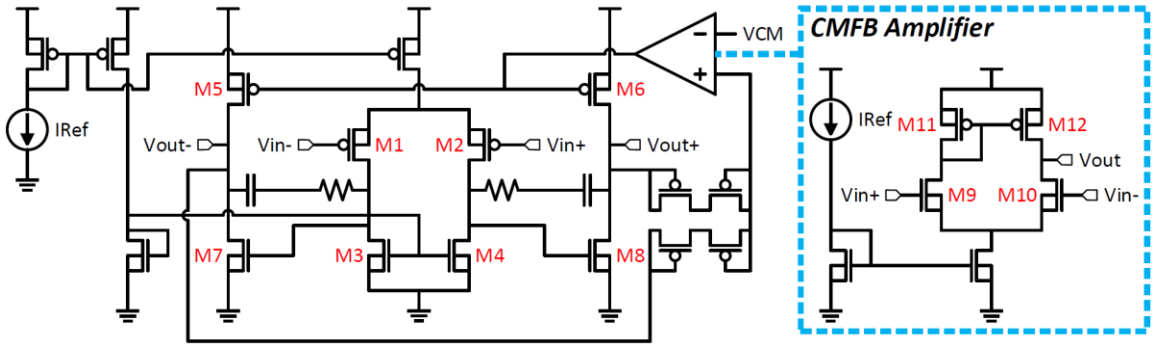

Figure 2.5: OTA of the PGA.

Transistor	W/L (μm)
$M_{1,2}$	20/1
$M_{3,4}$	1/2
$M_{5,6}$	1/10
$M_{7,8}$	1/2
$M_{9,10}$	0.5/0.5
$M_{11,12}$	1/10

Table 2.3: Transistor sizing of the two-stage amplifier.

Power	5.79μW
Supply Voltage	3V
Bias Current (Including CMFB)	1.4μA
Open-Loop Gain	71dB
Closed-Loop Gain Range	7V/V ~ 100V/V (17dB ~ 40dB)
Gain Tuning Step	3V/V
Maximum Gain Error	2.77%
Bandwidth (0.1pF Loading Capacitor)	2.70Hz ~ 133.59kHz
Single Output Range	$2.5V_{pp}$

Table 2.4: Specifications of the PGA.

2.4 Simulation of the AFE

Several tests are performed in Cadence simulations to evaluate the frequency response, noise, gain tuning function, gain error and signal distortion level of the AFE. All results are from **pre-layout** simulations.

The amplitude frequency responses of the AFE at all 32 gain setups are shown in Figure 2.6. Plot (a) shows the gain of the AFE in dB, from which we can see that the lower and upper cutoff frequency of the AFE remain the same at different gain setups. This means that the loading capacitance compensation technique described in Section 2.2 is effective. The bandwidth is 3.25Hz ~ 4.65kHz and covers most of the frequency spectrum of the action potential (approximately 100Hz ~ 7kHz). [15] Plot (b) shows the gain in linear scale, from which we can see the uniform gain tuning step. By comparing the mid-band gain of each response with the ideal gain, we can calculate the gain error at different setups, which is shown in Plot (c). We can see that the gain error increases as the closed-loop gain increases. The maximum gain error is less than 4%. With such small gain error, accurate control over the signal amplitude can be realized.

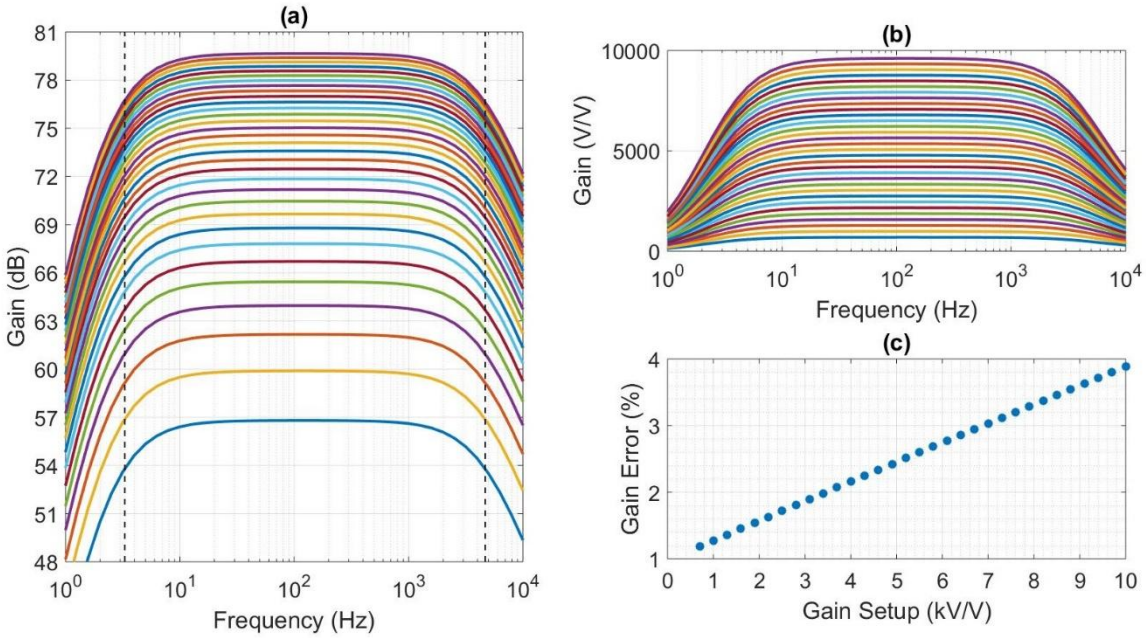

Figure 2.6: Amplitude frequency responses of the AFE at different gain setups.

The gain tuning function is also tested in real time. The result is shown in Figure 2.7. The input signal is a 1kHz sinusoidal wave with 50μV amplitude. The 5-bit digital control signal $B_{0\sim4}$ changes by 1 LSB every 10ms. We can see that the amplitude of the output signal changes from 35mV to 480mV approximately. Notice that the common-mode voltage of the output does not always stay at 0V since sometimes the gain switching does not occur at the zero-crossing point. However, if the simulation lasts longer, the common-mode voltage will ultimately settle down at 0V.

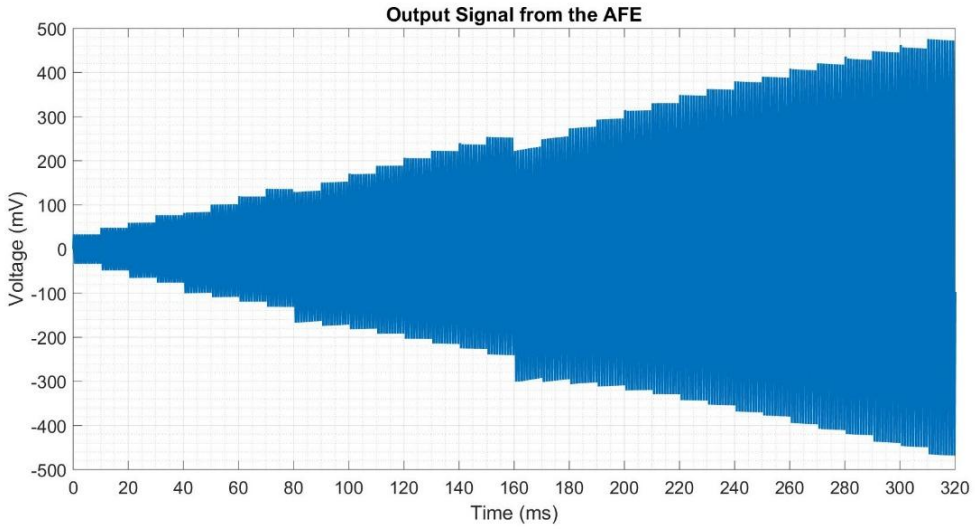

Figure 2.7: Real-time gain tuning of the AFE.

To test the signal distortion level of the AFE, a 100ms-long real neural signal recorded from the hippocampus region is used as the input signal. [16] The signal amplitude has been adjusted to a smaller value to test the AFE with the maximum gain. The output signal from the AFE is compared with an ideally amplified neural signal. The ideally amplified neural signal is created in MATLAB by following two steps: 1) The real neural signal (with adjusted amplitude and sampled at 1MHz) is filtered by a second-order Butterworth bandpass filter with edge frequencies equal to the lower and upper cutoff frequency of the AFE. 2) The filtered signal is then amplified by 10000 to generate the ideally amplified neural signal. Figure 2.8 shows a 60ms-long section of the simulation result which contains three neural spikes. We can see that the output signal from the AFE maintains the shape of the original signal well and is almost identical to the ideally amplified neural signal. This shows that the AFE has low signal distortion level and can keep features of the original signal for later signal processing.

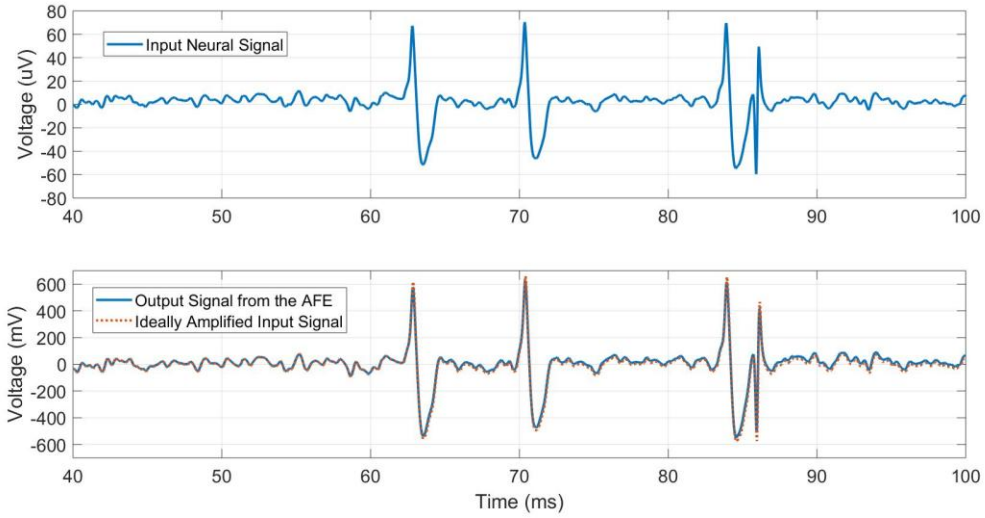

Figure 2.8: Amplification of a real neural signal by the AFE.

Specifications of the AFE are summarized in Table 2.5. We can see that the integrated noise of the AFE is less than 5μV. The overall gain can be tuned linearly with wide range. This enables accurate control over the signal amplitude. Due to the fully differential structure, the AFE has very high common-mode rejection ratio (CMRR). The AFE also has sufficiently large bandwidth and input range for recording action potentials. In conclusion, this general-purpose AFE meets the design goal and is suitable to be integrated into our closed-loop neurostimulation system.

Process	180nm Silicon-on-Insulator
Total Power	14.61μW
Supply Voltage	3V
Bias Current (Excluding I_{Ref} Generation)	4μA
Closed-Loop Gain Range	0.7kV/V ~ 10kV/V (57dB ~ 80dB)
Gain Tuning Step	0.3kV/V
Maximum Gain Error	3.89%
Bandwidth (0.1pF Loading Capacitor)	3.25Hz ~ 4.65kHz
Input Range	$7.2mV_{pp}$
Single Output Range	$2.5V_{pp}$
CMRR	>134dB
Integrated Noise over the Bandwidth	4.81μV
Noise Density	$56.28nV/Hz^{1/2}$

Table 2.5: Specifications of the AFE.

Chapter 3 **Spike Detector**

3.1 Introduction to Action Potential Detection

Real-time action potential detection plays an important role in on-chip neural signal processing. For our system, the event-triggered stimulation heavily relies on the fidelity of the spike detector. This chapter describes the design of a spike detector based on the absolute threshold detection (ATD) method. [17] Both algorithm design and circuit implementation will be presented.

3.2 Algorithm Design

The effectiveness of the spike detection method relies on the signal-to-noise ratio (SNR) and the robustness of the algorithm. Since the SNR is primarily set by the analog front-end, algorithm design becomes essential to the performance optimization. As shown in the conceptual diagram representing the working principle of ATD method, the neural signal is constantly compared to the predefined threshold voltage within the detection window. The comparison result is then used to detect the action potentials. Although the scheme is simple, for real-time action potential detection in systems with limited computation resources, ATD method is as effective as other power-hungry and computation-intensive methods. [17] [18]

One disadvantage of ATD method is its vulnerability to noises and artifacts. The implemented algorithm is thereby designed to improve the error tolerance of the spike detector. As shown in Figure 3.1, for each detection cycle, there are maximally twenty-five comparisons. The spike detector only recognizes a spike when at least four comparisons show that the neural signal crosses the threshold voltage. Otherwise, the spike detector clears comparison results and goes to the next detection cycle. This process helps to distinguish the action potential from fake spikes caused by the stimulator or improper setup of threshold voltage. A detailed example will be shown in Section 3.4.

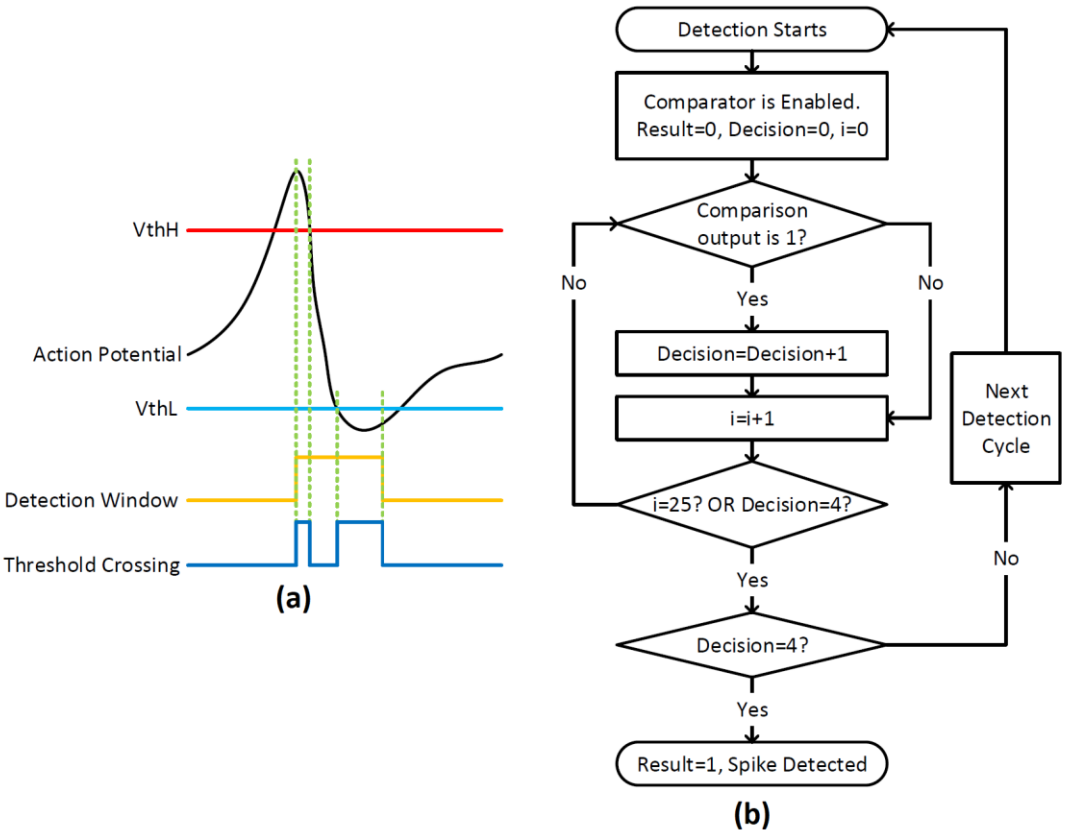

Figure 3.1: (a) The working principle of ATD method. (b) The flow chart of the implemented spike detection algorithm.

3.3 Circuit Implementation

The circuit of the spike detector is shown in Figure 3.2. There are three digital inputs to the detector: reset signal (RST), enabling signal (Detection) and input clock (CLK). The input clock is set to 100kHz. There are three analog inputs: neural signal (Vin), high threshold voltage (V_{thH}) and low threshold voltage (V_{thL}). For normal operation, V_{thH} and V_{thL} are set to 1.9V and 0.9V respectively. The spike detector has two outputs: “Result” and “Decision”. “Result” indicates the occurrence of action potential and “Decision” indicates the occurrence of threshold crossing.

The comparison of neural signal with V_{thL} and V_{thH} is accomplished by two latch comparators. The outputs from two comparators are fed into an OR gate to generate the output “Decision”. When the detection cycle starts (“Detection” turns to 1), the input clock is converted to 25kHz and then fed into the level shifter which drives the enabling port (EN) of two comparators. Since one detection cycle is 1ms, the comparator network will perform maximally twenty-five comparisons within one detection cycle.

The output “Decision” is used to clock a 4-bit serial-to-parallel shift register formed by four asynchronous resettable positive-edge-triggered D flip flops. The shift register’s input is tied to 1 (1.5V) and four outputs are connected to an AND gate. After four clock cycles (four threshold crossings), all four outputs are 1 and the final output “Result” turns to 1.

If after one detection cycle, “Result” does not turn to 1, the shift register will be reset at the end of detection. This makes sure that previous detection history will not be carried over to next detection cycle. If “Result” turns to 1, it will remain at 1 unless the spike detector is reset to initial state (RST turns to 1).

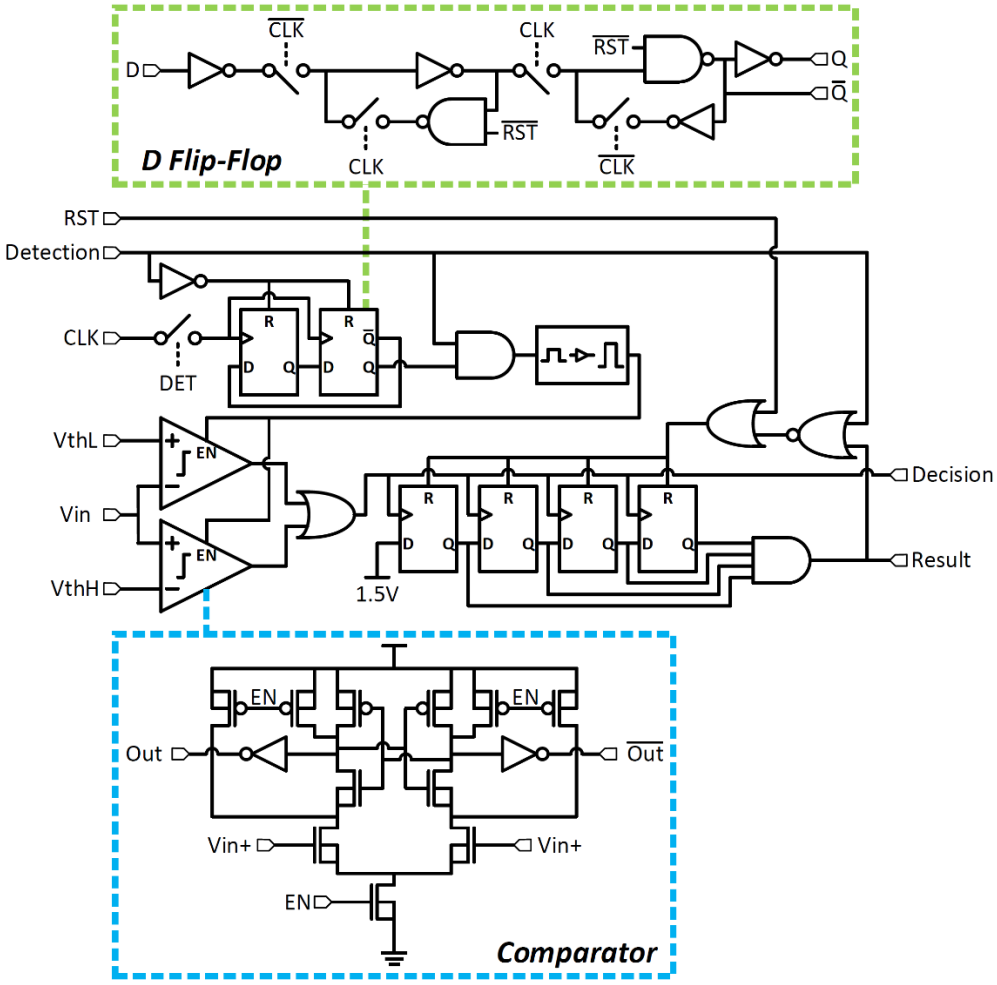

Figure 3.2: Circuit of the spike detector.

3.4 Cadence Simulation

The spike detector is tested by an amplified neural signal recorded from the hippocampus region. [3] Under normal operation, two threshold voltages V_{thL} and V_{thH} are set to 1.9V and 0.9V. The **pre-layout** simulation result is shown in Figure 3.3. We can see that within the detection window (1ms ~ 2ms), the neural signal has a wide region above V_{thH} and the spike detector successfully captures this region. After four comparisons which all indicate that the neural signal has crossed the threshold voltage, "Result" turns to and stays at 1, meaning that a spike has been detected.

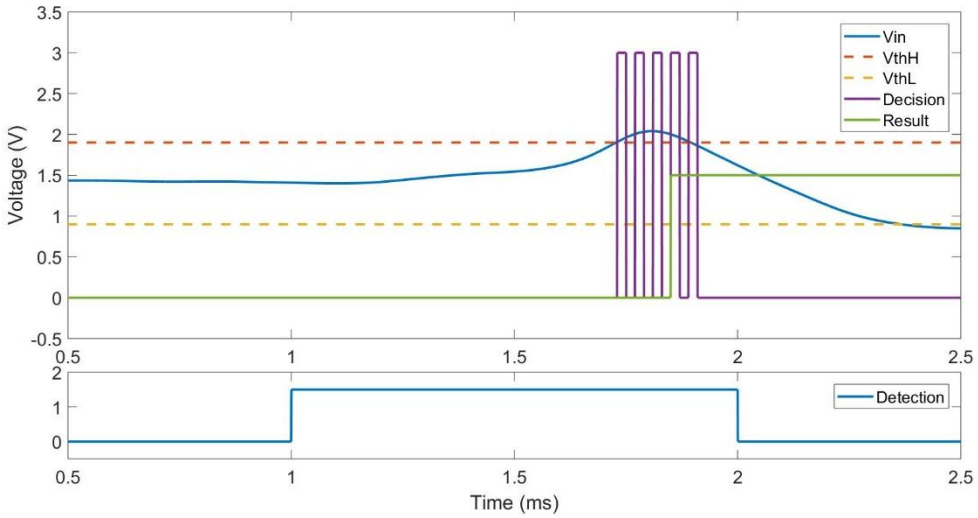

Figure 3.3: Cadence simulation of the spike detector under normal operation.

To test the error tolerance of the spike detector, the high threshold voltage V_{thH} is set to 1.5V which is much lower than normal value. Under such setup, noises or signal artifacts have a higher chance of crossing the threshold voltage. The **pre-layout** simulation result is shown in Figure 3.4. The top plot shows the complete 32ms simulation. We can see that the spike detector successfully differentiates the action potential signal from noises. From the middle plot, we can see that although the signal between 19ms and 19.5ms crosses the threshold voltage, the number of comparisons indicating the threshold crossing is smaller than three. Therefore, the spike detector does not recognize this signal as the action potential.

In conclusion, the designed spike detector presents strong error tolerance. The dynamic power of the spike detector is 239nW.

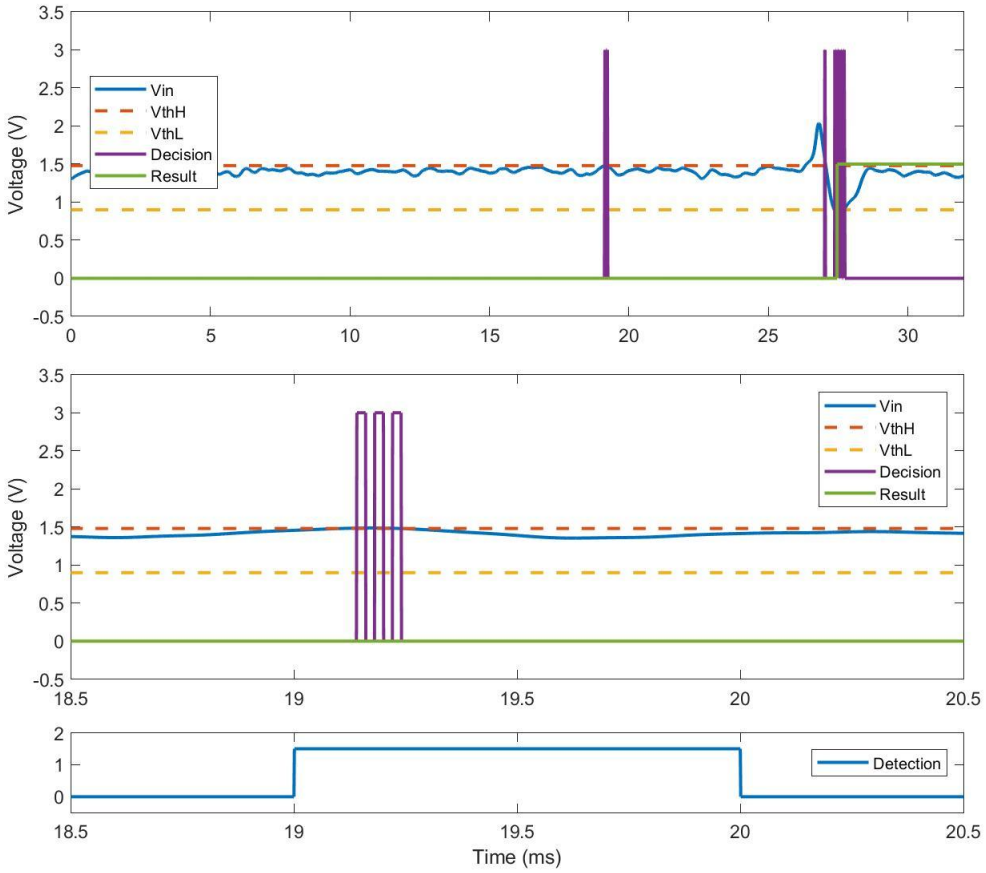

Figure 3.4: Cadence simulation of the spike detector when V_{thH} is 1.5V.

Chapter 4 **Neural Stimulator**

4.1 Introduction to Neurostimulation

Neurostimulation or neuromodulation is an important technique for today's neuroscience research. It uses controlled stimuli (in electrical, optical or chemical form) to artificially excite neurons. For implantable biomedical devices and systems, electrical stimulation is the prevalently used method due to its feasibility of being implemented on a CMOS platform with low power consumption. [19] For this work, we will focus on the electrical neural stimulator.

In this chapter, we will first review the concept of charge-balanced neurostimulation and the electrical circuit model of electrode-tissue interface. We will then investigate different types of stimulator design and discuss their pros and cons. Mathematical modeling of the electrical behavior of the electrode-tissue interface during stimulation will also be shown. Finally, the design of the stimulator used to form the closed-loop neurostimulation system will be presented.

4.2 Charge-Balanced Neurostimulation

There are three primary tradeoffs for neural stimulator design: safety, efficiency and performance. [2] For implantable devices, safety undoubtedly carries the highest weight among these tradeoffs. One major safety issue for electrical stimulation is the residual charge injected to the tissue caused by non-zero net current flow during stimulation or current leakage from the device. The accumulated residual charge raises the electrode potential and ultimately causes tissue damage and electrode corrosion. [9] For our closed-loop neurostimulation system, one complete operation contains multiple stimulations performed in short time. This imposes even more rigorous safety requirements for the design. To address this issue, charge-balanced neurostimulation is needed. Several important techniques have been reported to achieve or improve the charge-balancing behavior of the stimulator:

- **Blocking Capacitor:** A blocking capacitor can be placed at the output of the stimulator to limit charge accumulation on electrodes and block any DC path. The blocking capacitor requires a large capacitance to reduce the voltage drop across the capacitor and is thereby usually an off-chip capacitor. The voltage drop can be approximated by the equation below: [20]

$$\Delta V = I_{Stim}\frac{\Delta t}{C_B},$$

where Δt is the current pulse width and I_{Stim} is the amplitude of current pulse. Considering a 100μA 200μs current pulse going through a 1μF blocking capacitor, the voltage drop is then approximately 20mV which is acceptable in most design cases. For this work, all blocking capacitors used are 1μF.

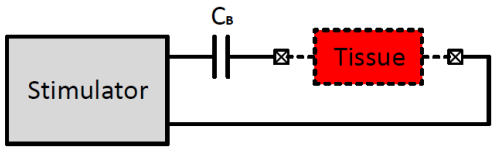

Figure 4.1: Blocking capacitor (C_B in the graph).

- **Passive Discharge:** Passive discharge can be used to clear residual charge by simply shorting the working electrode (the one connected to the blocking capacitor) to the reference electrode (the one connected to the common-mode voltage V_{CM}). Although the realization is simple, passive discharge does not provide controlled discharging current and discharging time. [2] Therefore, to remove significant amounts of residual charge, passive discharge may not be the most reliable solution.

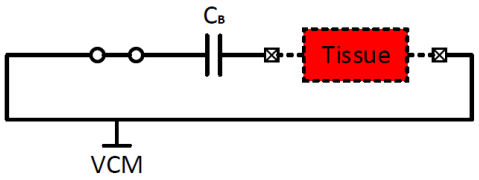

Figure 4.2: Passive discharge (assuming blocking capacitor is used).

- **Biphasic Stimulation:** Biphasic stimulation achieves charge balance by sending two current pulses sequentially. The cathodic pulse stimulates neuron cells and the anodic pulse actively discharges the electrode. As shown in Figure 4.3, the duration of cathodic phase, interphase delay and duration of anodic phase are given by: t_c, t_p and t_a respectively. The current amplitude of cathodic pulse and anodic pulse are I_c and I_a respectively.

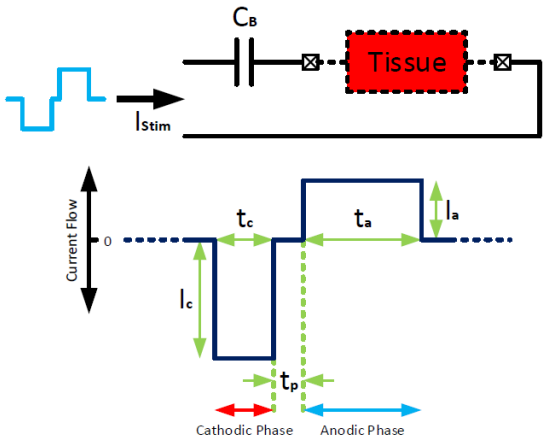

Figure 4.3: Biphasic stimulation (assuming blocking capacitor is used).

Biphasic stimulation can be realized by monopolar or bipolar stimulation waveform. In a monopolar stimulation, both cathodic phase and anodic phase are generated from a single electrode. In a bipolar stimulation, cathodic phase and anodic phase are generated by a pair of electrodes. [2]

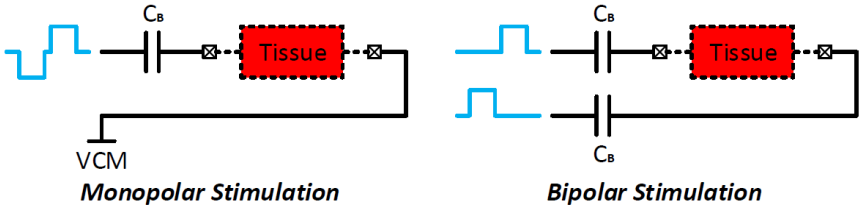

Figure 4.4: Monopolar and bipolar stimulation.

For charge-balanced stimulation, we want zero net charge flow to the tissue. Namely, we want:

$$Q_{total} = \int_{t_0}^{t_0+T} I(t)dt = 0,$$

where T is the duration/period of stimulation and t_0 is the initial time. The time-varying current flow depends on the driving ability of the stimulator and the impedance of electrode-tissue interface. A simplified electrical circuit model of electrode-tissue interface is shown in Figure 4.5. [2] [21] As shown in the graph: R_F is the Faradaic resistance; C_F is the double layer capacitor; R_S is the solution spreading resistance.

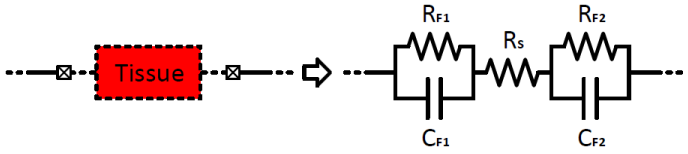

Figure 4.5: Electrical circuit model of electrode-tissue interface.

Values of R_F, C_F and R_S are based on the measurement of a low-cost 75µm/50µm-diameter tungsten electrode commonly used in neuroscience research. The impedance is measured while the electrode is in 0.9g/100mil Sodium Chloride solution. [2] For two perfectly identical electrodes in contact with the tissue, the circuit model is symmetrical ($R_{F1} = R_{F2}$ and $C_{F1} = C_{F2}$). However, in reality, these electrodes will always have certain asymmetry. To have a comprehensive study of the effect of electrodes on the charge-balancing behavior, we consider these cases: 1) symmetrical 75µm/50µm electrodes. 2) asymmetrical electrodes formed by a 75µm and a 50µm electrode. 3) asymmetrical electrodes formed by 75µm electrodes with 20% variation on both sides. Circuit parameters of them are shown in Table 4.1.

Parameter	**Symmetrical (75µm)**	**Symmetrical (50µm)**	**Asymmetrical (75µm+50µm)**	**Asymmetrical (20% Var.)**
C_{F1}	55nF	18nF	55nF	66nF
R_{F1}	7MΩ	19MΩ	7MΩ	8.4MΩ
R_S	12kΩ	20kΩ	12kΩ	12kΩ
C_{F2}	55nF	18nF	18nF	44nF
R_{F2}	7MΩ	19MΩ	19MΩ	5.6MΩ

Table 4.1: Circuit parameters of the electrode-tissue interface model.

4.3 Overview of Electrical Neural Stimulator Design

There are generally three types of electrical stimulators in terms of stimuli generation methods: voltage-regulated stimulator, current-regulated stimulator and charge-regulated stimulator. [2] In this section, we will focus on the first two and evaluate their charge-balancing performance by both circuit simulation and mathematical modeling.

4.3.1 Voltage-Regulated Bipolar Stimulator

This type of stimulator enables flexible stimulation waveform and has very high energy efficiency. However, it has poor control over the charge injected to the tissue and thereby has the worst charge-balancing behavior. The simulated stimulator circuit is shown in Figure 4.6. [13] The biphasic waveform is realized by the cross-coupled switch and the voltage stimulation waveform is generated by a class-AB amplifier.

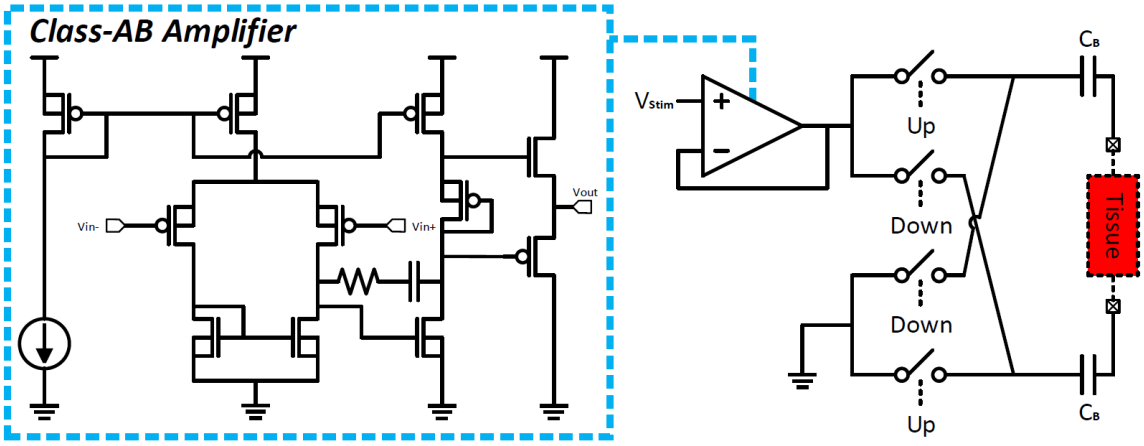

Figure 4.6: Voltage-Regulated bipolar stimulator.

In Cadence simulation, a biphasic voltage pulse with 200μs pulse width (PW) and 1V amplitude (V_{DC}) is injected to the 75μm electrode pair at t_0 = 10μs. The interphase delay (t_p) is set to 10μs.

For mathematical modeling, we can derive the current in Laplacian domain and use inverse Laplace transform to find the current in time domain. The current in Laplacian domain is given by:

$$I(t) = \mathcal{L}^{-1}\{I(s)\} = \mathcal{L}^{-1}\left\{\frac{V(s)}{Z(s)}\right\}.$$

Assume that the voltage waveform has a positive and negative pulse with same pulse width and amplitude but no interface delay. Parameters of Z(s) are from the symmetrical model described in section 4.2. Then we have:

$$V(s) = \frac{V_{DC}}{s}\left[\left(e^{-t_0 s} - e^{-(t_0+PW)s}\right) - \left(e^{-(t_0+PW)s} - e^{-(t_0+2PW)s}\right)\right],$$

$$Z(s) = R_{F1} || \frac{1}{sC_{F1}} + R_S + R_{F2} || \frac{1}{sC_{F2}}.$$

The calculation is done in Wolfram Mathematica. Figure 4.7 shows the **pre-layout** simulation result. We can see that the net charge injected to the tissue is clearly not zero. This can be proven by the mathematical modeling, which generates approximately an identical current waveform. This comparison also proves that the mathematical modeling can accurately describe the electrical behavior of the electrode-tissue interface given small artifacts generated by circuits.

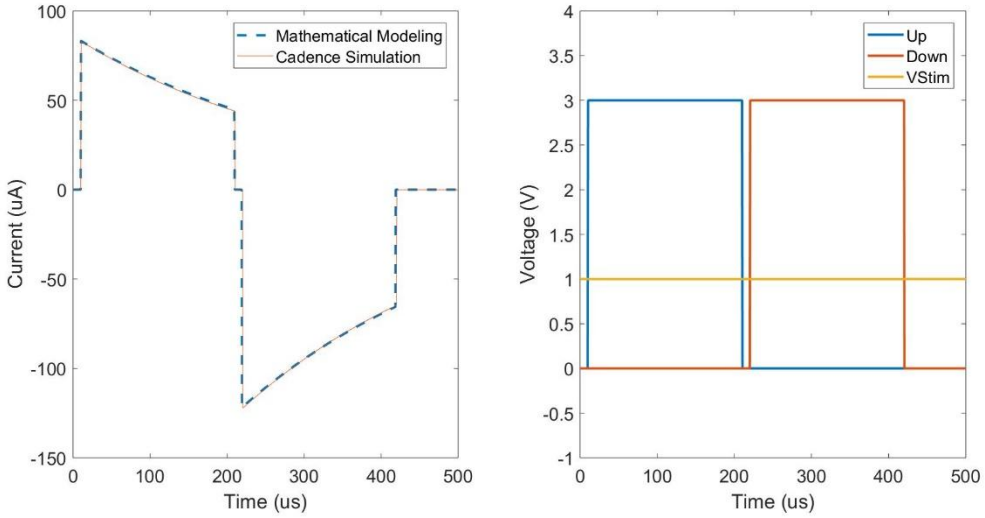

Figure 4.7: Cadence simulation and mathematical modeling of bipolar voltage-regulated stimulator. Left: current waveform. Right: stimulation waveform.

Although it is possible to compensate the net charge flow by engineering the voltage waveform, this type of stimulator is not suitable for our system due to its overall poor charge-balancing performance.

4.3.2 Current-Regulated Bipolar Stimulator

This type of stimulator can be realized by a single current source/sink or multiple current sources. Here we discuss the simplest form shown in Figure 4.8. [22] The variable current sink has an output impedance given by:

$$R_{out} \approx g_{mn} r_{on} r_{op} A_v,$$

where A_v is the DC gain of the auxiliary amplifier. For the simulated circuit, the auxiliary amplifier is a two-stage amplifier with DC gain of 60dB. This huge output impedance enables the current sink to drive the electrode-tissue interface given sufficient supply voltage. The negative feedback also regulates the source voltage of PMOS transistors. For sufficiently large gain, the source voltage is close to the reference voltage V_{REF} (shown in equation below). The binary weighted PMOS transistors form a 4-Bit current DAC. With regulated V_{SD}, accurate current control can be achieved.

$$V_S = \frac{A_v}{\frac{1}{r_{op} g_{mn}} + 1 + A_v} V_{REF} \approx V_{REF}.$$

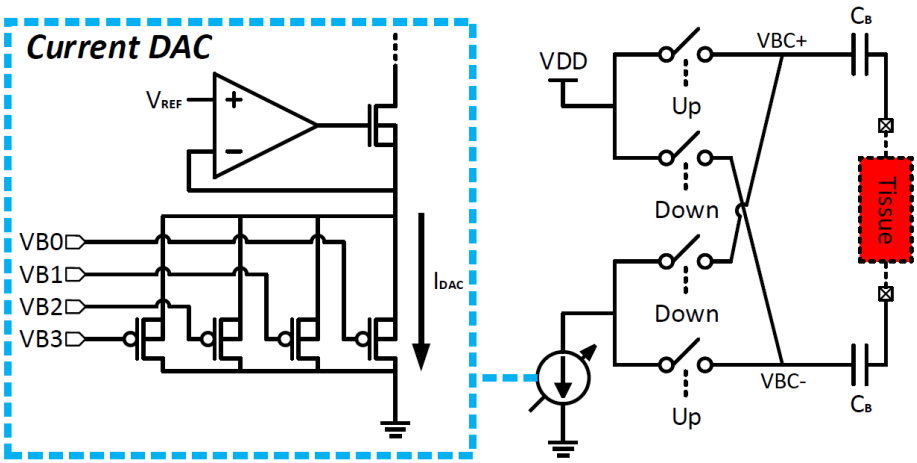

Figure 4.8: Current-Regulated bipolar stimulator.

Two types of loads are used to test the driving ability of the stimulator: symmetrical load formed by two 75µm electrodes and asymmetrical load formed by one 75µm electrode and one 50µm electrode. A biphasic current pulse with 200µs pulse width (PW) and 74µA amplitude (I_{DC}) is injected to the electrode pair at t_0 = 10µs. The interphase delay (t_p) is set to 10µs. The **pre-layout** simulation result is shown in Figure 4.9. We can see that for symmetrical load, the stimulator shows excellent charge-balancing performance. The voltage across the blocking capacitor is approximately zero at the end of stimulation. However, for asymmetrical load, we can see that the positive current pulse is "attenuated", and the net charge flow is non-zero.

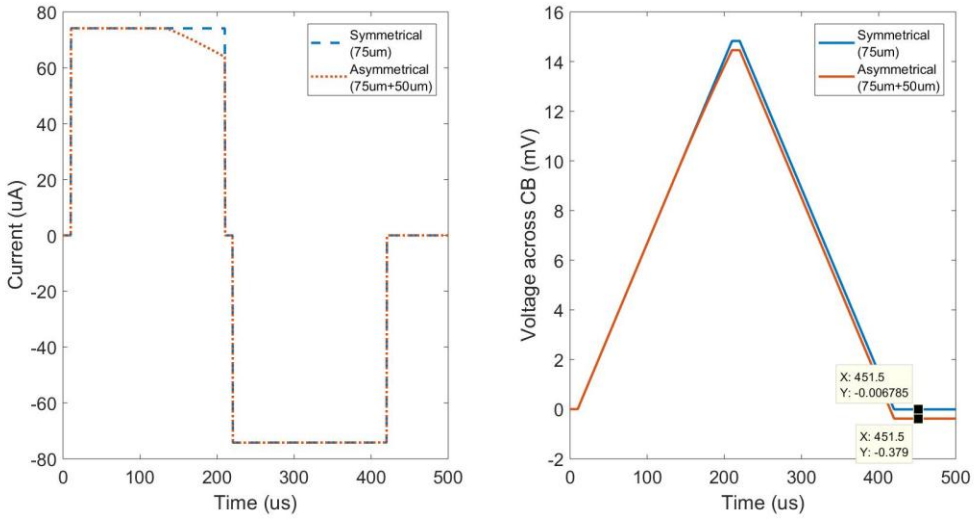

Figure 4.9: Cadence simulation of current-regulated bipolar stimulator. Left: Stimulation current. Right: Voltage across the blocking capacitor.

The failure of charge-balanced stimulation is mainly caused by the poor driving ability of the stimulator when small capacitor is present (introduced by the 50µm electrode). The

small capacitor causes more rapid drop of the NMOS transistor's drain voltage during current injection, which ultimately pulls the NMOS into triode region. When the NMOS falls out of saturation region, the output impedance of the stimulator decreases significantly, and the stimulator loses accurate control over the output current. This failure can be further revealed by observing the voltage on the blocking capacitors.

Another issue is the high voltage caused by the switching process. During the switching between two phases, blocking capacitors and capacitors in the load act like a charge pump. As shown in Figure 4.10, we can see that VBC+ jumps at the onset of anodic phase. This voltage jump could be very high and could ultimately damage transistors.

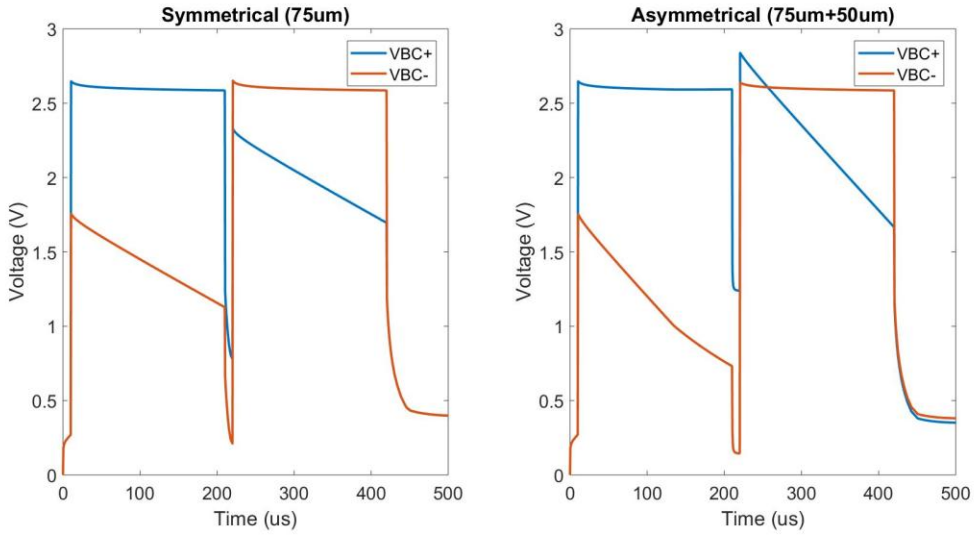

Figure 4.10: Voltage on blocking capacitors during stimulation.

Although such asymmetry may not appear practically, this design provides very limited output current range if charge-balanced stimulation is desired.

4.3.3 Current-Regulated Monopolar Stimulator

Monopolar stimulator is suitable for electrode arrays since multiple electrodes can share one reference electrode. If well designed, it also has larger driving ability. However, this type of stimulator usually needs high supply voltage and the circuitry is complex. [23] [24] [22] [25] The conceptual diagram of the stimulator is shown in Figure 4.11. Detailed circuit implementation is covered in Section 4.4. As shown in graph, the stimulator has both current sink and source which are usually controlled by current mirrors. A biphasic current pulse is generated from a single electrode. The reference electrode is connected to the common-mode voltage V_{CM} which is usually half of supply voltage. Also notice that the stimulator only needs one blocking capacitor.

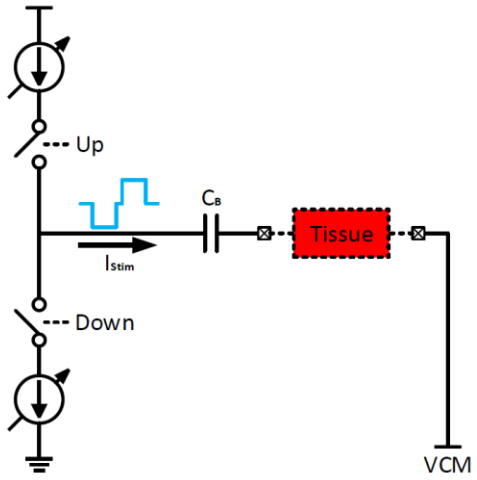

Figure 4.11: Current-Regulated monopolar stimulator.

4.4 Design of a Charge-Balanced Neural Stimulator

This section describes the design of a charge-balanced current-regulated monopolar stimulator. The conceptual circuit diagram of the design is shown in Figure 4.12. The stimulator sends biphasic current pulse to minimize net charge introduced to the tissue and then performs passive discharge by closing the switch at the output to clear residual charge. Current source and sink are controlled by two low-voltage current mirrors in series. The controlled current from a 4-bit current DAC is injected in between two current mirrors. The proposed structure minimizes the current mismatch introduced by current copying. The current copying ratio is 20.

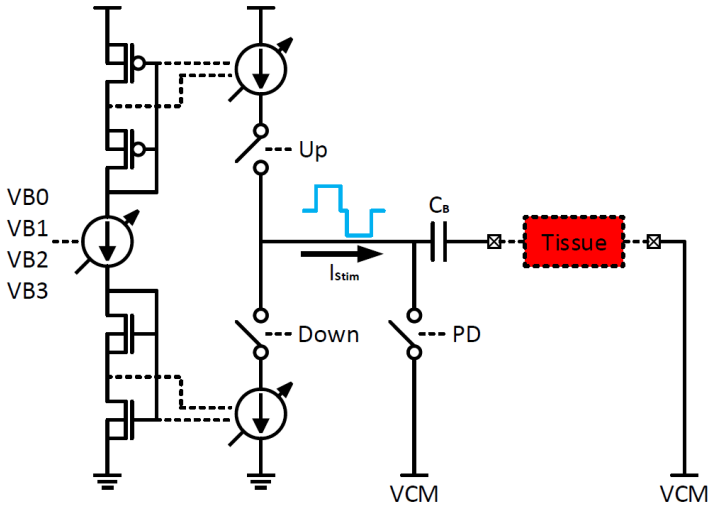

Figure 4.12: Conceptual circuit diagram of the final design.

4.4.1 Study of Passive Discharge

Before going into the circuit implementation, it is necessary to study the electrical behavior of the electrode-tissue interface during passive discharge. Assume that the current waveform has a positive and negative pulse with same pulse width (PW) but different amplitude (I_c and I_a respectively). The current is injected at t = 0 and there is no interphase delay. The current waveform in Laplacian domain is then given by:

$$I_{Drive}(s) = \frac{1}{s}[I_c(1 - e^{-PWs}) - I_a(e^{-PWs} - e^{-2PWs})].$$

The driving voltage for generating such current through the model impedance Z(s) is:

$$V_{Drive}(t) = \mathcal{L}^{-1}\{V_{Drive}(s)\} = \mathcal{L}^{-1}\{I_{Drive}(s)Z(s)\}.$$

We know that during passive discharge, two electrodes are shorted, and the output voltage is equivalently zero. If the passive discharge takes place immediately after stimulation, the total output voltage is:

$$V_{out}(t) = V_{Drive}(t) \times [\theta(t) - \theta(t - 2PW)],$$

where θ(t) is Heaviside step function. Total output current is then given by:

$$I_{out}(t) = \mathcal{L}^{-1}\{I_{out}(s)\} = \mathcal{L}^{-1}\left\{\frac{\mathcal{L}\{V_{out}(t)\}}{Z(s)}\right\}.$$

Suppose the pulse width is 200μs and the load is 75μm electrodes with 20% variation on both sides. Assume that cathodic current and anodic current have following relationship:

$$I_c = I_a(1 + \varepsilon),$$

where ε is the current mismatch in percentage. Assume that the discharging time is t_{PD}. The ratio of cleared residual charge to total residual charge introduced by current mismatch is then given by:

$$R_{PD} = \frac{Q_{cleared}}{Q_{net}} = \frac{\left|\int_{2PW}^{2PW+t_{PD}} I_{out}(t)dt\right|}{(I_c - I_a)PW}.$$

The ratio R_{PD} is evaluated in three cases: 1) Vary t_{PD} at I_a = 75.23μA and ε = 10%. 2) Vary ε at I_a = 75.23μA and t_{PD} = 570μs. 3) Vary I_a at ε = 10% and t_{PD} = 570μs. Plots of them are shown in Figure 4.13. The output current at I_a = 75.23μA, ε = 10% and t_{PD} = 570μs is also shown. All calculations are done in Mathematica. From the graph, we can see that R_{PD} only depends on ε and t_{PD}. For 10% mismatch, 570μs-long passive discharge can clear 83.71% residual charge which is enough for our design. As discharging time increases beyond 500μs, efficiency of passive discharge decreases significantly. For the final design, discharging time is set to 570μs.

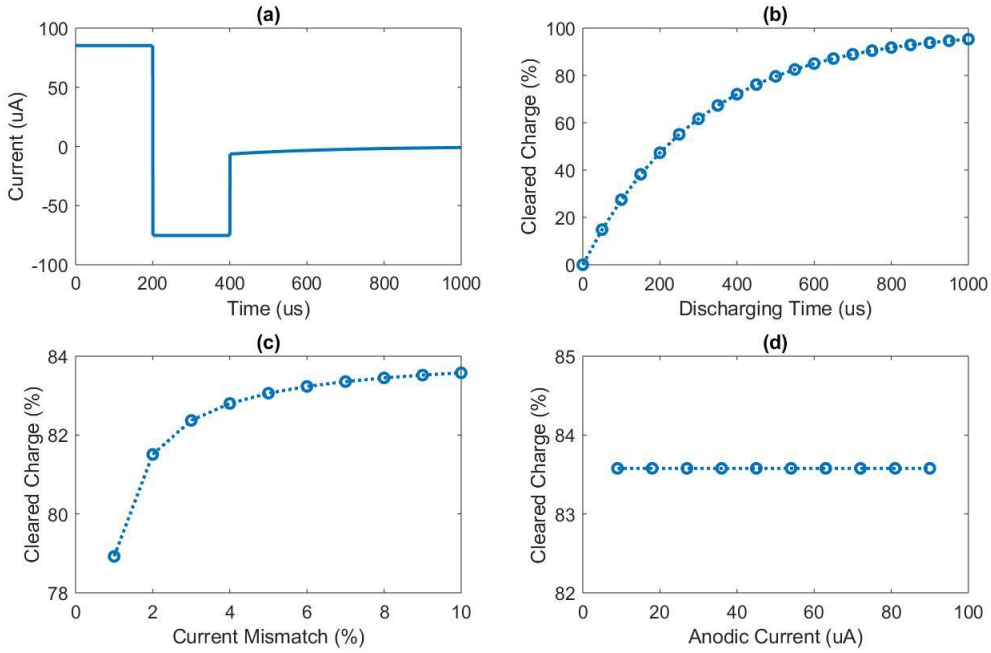

Figure 4.13: Mathematical modeling of passive discharge. (a) Output current waveform. (b) R_{PD} vs t_{PD}. (c) R_{PD} vs ε. (d) R_{PD} vs I_a.

4.4.2 Circuit Implementation

The circuit is shown in Figure 4.14. The output current amplitude of the current DAC is controlled by four binary weighted PMOS transistors $D_{0\sim3}$. Two auxiliary amplifiers are used to boost the output impedance at two ends and regulate $D_{0\sim3}$'s drain voltage to V_{Bias} and source voltage to 3.5V. Digital control is realized by four single-pole, double throw (SPDT) switches connected to the gates of $D_{0\sim3}$. Each SPDT switch is implemented by two 1μm/1μm floating-body transistors due to their high breakdown voltage. When the digital input is 1, the gate is connected to V_{Bias} (PMOS is on) and when the digital input is 0, the gate is connected to 3.5V (PMOS is off).

We can see that the resolution (LSB) of the current DAC is thereby set by V_{Bias}. For our system, the desired value is 300nA. To obtain such resolution, V_{Bias} is generated by a current mirror formed by transistor D_B (which has same size as D_0) and a 300nA current reference. The current reference is a regulated cascode stage whose current output only depends on the source resistor. This type of current reference has strong immunity to process variation and guarantees stable resolution of current DAC. With 300nA resolution, the output range of current DAC is 0.3μA ~ 4.5μA. The current DAC is driving the PMOS/NMOS low-voltage current mirror which controls output current source/sink respectively. With current copying ration of 20, the stimulator's output range is 6μA ~ 90μA with 6μA resolution.

One major issue about the large-ratio current copying is its significant noise current introduced by the flicker noise of the reference transistor (N_2 or P_2), which will translate to the noise voltage on the load and may affect the neural signal recording. However, for our system, since the stimulation and the recording do not happen simultaneously, and the stimulator is completely off during the detection phase, the noise issue is not severe. To further alleviate this issue, we can insert a resistor between gates of N_2 and N_4 and a same one between gates of P_2 and P_4 to reduce noises. [26]

The output stage contains a current source and a current sink, which are both realized by regulated cascode stage. The auxiliary amplifier also regulates the drain voltage of N_4/P_4 to the drain voltage of N_2/P_2. Therefore, the channel length modulation effect of N_4/P_4 is eliminated and accurate current copying can be achieved. Transistors N_5 and P_5 are used to lower $|V_{DS}|$ of $N_{3,4}$ and $P_{3,4}$ during stimulation to protect them from high-voltage breakdown.

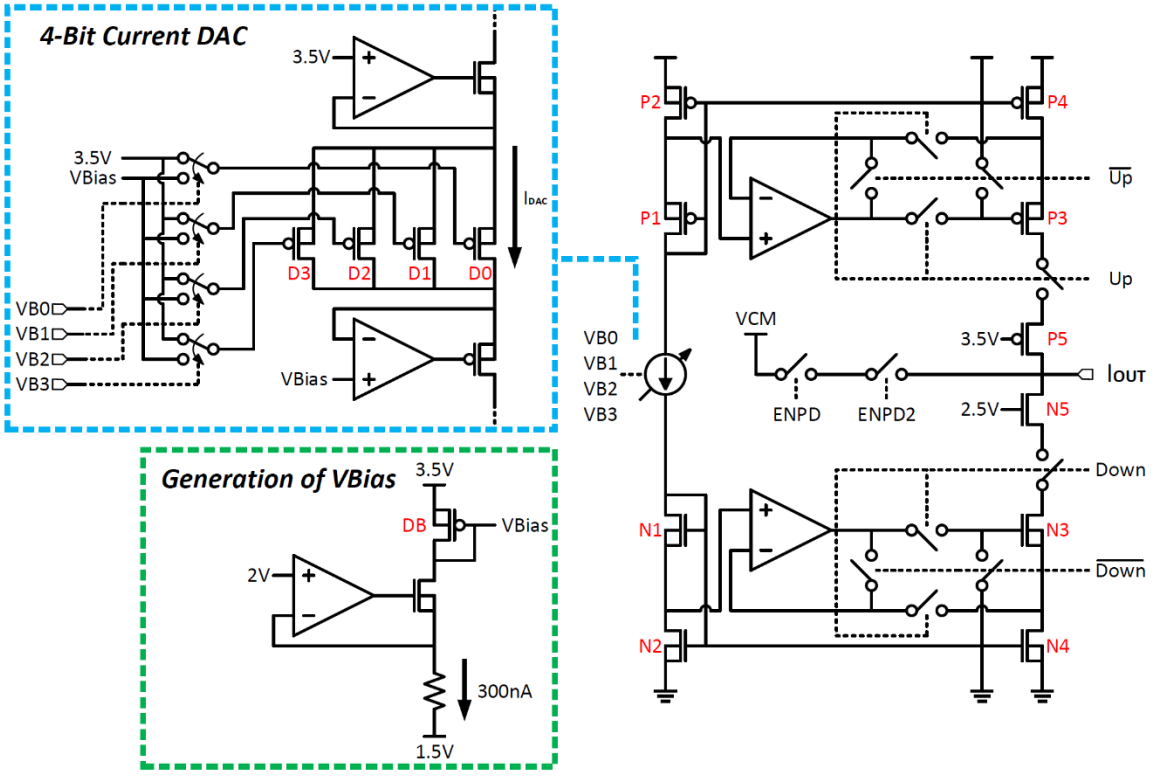

Figure 4.14: Circuit of charge-balanced current-regulated monopolar stimulator.

The phase switching is controlled by “Up” and “Down” which corresponds to cathodic phase and anodic phase respectively. Each controlled switch is implemented by a 1μm/1μm floating-body transistor. Switching procedure of the current sink is shown in Figure 4.15 (current source has a similar process). In state (a), the current sink is connected to the output and the regulated cascode stage is formed by auxiliary amplifier and $N_{3,4}$. In state (b), current sink is disconnected from the output and the gate of N_3 is tied to ground to fully turn off the current sink. The output of the axillary amplifier is now connected to its negative input to form a voltage buffer. The voltage buffer

maintains the voltage at the negative input in state (a) so that when the current sink goes back to state (a) to form the regulated cascode stage, the drain voltage of N_4 can quickly settle down. This procedure helps to eliminate the sharp current spike during phase switching.

Auxiliary amplifiers at the output stage play an important role in the stimulator's performance. Taking the current sink as an example, as I_{DAC} changes from 0.3μA to 4.5μA, drain voltage of N_2 can change widely. This means that the auxiliary amplifier needs to have large input common mode range (ICMR). Since the drain voltage of N_2 could be very low at minimum current output, auxiliary amplifier with PMOS input is used to regulate N_3. For current source, NMOS-input auxiliary amplifier is used. In the final design, both types are realized by two-stage amplifier. The PMOS-input amplifier has minimum gain of 60dB over V_{CM} of 0.1V ~ 2.5V. The NMOS-input amplifier has minimum gain of 52dB over V_{CM} of 0.3V ~ 2.9V.

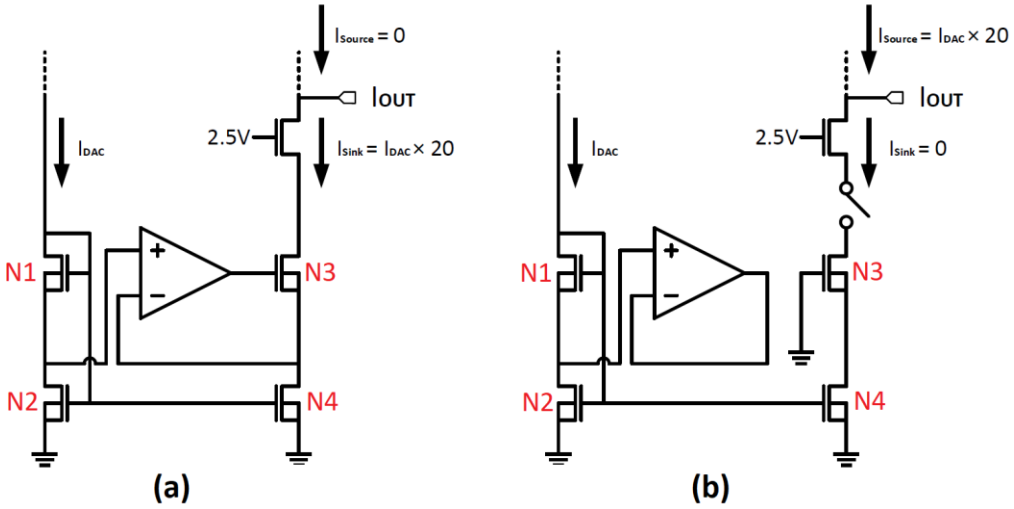

Figure 4.15: Switching of current sink. (a) current sink is on. (b) current sink is off.

The passive discharge takes place after the stimulation phase. During the passive discharge phase, the stimulator's output is shorted to V_{CM}. Outside the passive discharge phase, switches are open and ideally no current flows through them. There are three tradeoffs for the design: the discharging speed, the leakage current and the circuit's ability to endure high voltage. Cadence simulation shows that when the stimulator is driving the asymmetrical 75μm model with 90μA biphasic current pulse, the electrode voltage (V_{out} in the graph) can vary from 2V to 5V approximately. This wide voltage variation poses following constraints for the design: (a) The gate voltage cannot be too high during the stimulation phase. Otherwise the leakage current through the circuit will significantly degrade the charge-balancing performance. (b) At least two transistors need to be stacked to form the circuit. A single transistor can hardly handle such wide voltage change since V_{GS} will possibly exceed the breakdown voltage. (c) The aspect ratio of the transistor cannot be too large since the breakdown voltage of the transistor decreases as the length decreases or the width increases. The aspect ratio cannot be too small either. Otherwise the discharging will be very slow and not enough residual charge can be

cleared for the given time. For the reliability, $|V_{DS}|$ is kept below 2.75V and $|V_{GS}|$ is kept below 3.5V for all floating-body transistors.

The final design of the passive discharge circuit is shown in Figure 4.16. The circuit is formed by two switches controlled by "ENPD" and "ENPD2". These two digital control signals have the same waveform but different amplitude. "ENPD" has an amplitude of 1.5V/4V and "ENPD2" has an amplitude of 3V/5.5V. Each switch is implemented by a floating-body NMOS transistor. As shown in the graph, during the passive discharge phase, "ENPD" is 4V and "ENPD2" is 5.5V. Both S_1 and S_2 are on. V_{GS1} is 1V and V_{GS2}, for the worst case, is 2.5V. Outside the passive discharge phase, "ENPD" is 1.5V and "ENPD2" is 3V. When V_{out} is 2V, S_2 is on. However, V_{GS1}, for the worst case, is -0.5V. This ensures that S_1 is always off. When V_{out} is 5V, S_1 and S_2 are both off. S_2 acts like a protection transistor for S_1. Since V_{G2} is 3V, V_{S2} or V_{D1} is always below 3V. This means that V_{GS1} for the worst case is -1.5V. In conclusion, V_{DS} and V_{GS} of S_1 and S_2 do not exceed breakdown voltage in both two cases. Notice that S_1 has smaller width compared with S_2 since we want to minimize the leakage current when S_1 is off and S_2 is on.

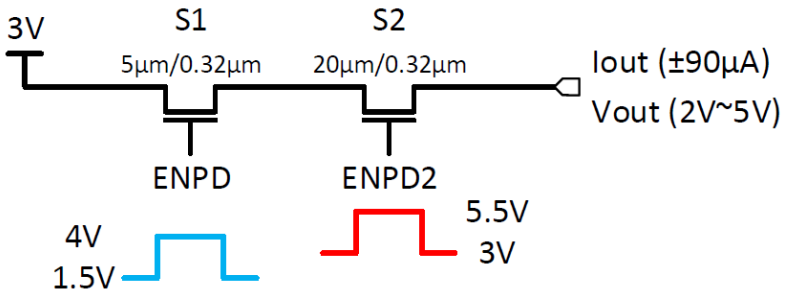

Figure 4.16: Implementation of passive discharge.

4.4.3 Cadence Simulation

In Cadence simulation, the stimulator is set to send 200µs-pulse-width biphasic pulse with 10µs interphase delay. Passive discharge takes place 10µs after stimulation and lasts for 570µs. The stimulator then stays idle till next stimulation. One stimulation cycle takes 2ms. The load is 75µm electrodes with 20% variation on both sides.

Three tests are performed to evaluate the static nonlinearity, passive discharge function and charge-balancing performance of the stimulator. The **pre-layout** simulation result is shown in Figure 4.17. Plot (a) shows all 15 output current values. The output current amplitude is taken as the average amplitude of cathodic pulse and anodic pulse. Calculation shows that the maximum differential nonlinearity (DNL) and the maximum integral nonlinearity (INL) are 0.0011 LSB and 0.0010 LSB respectively. Such low DNL/INL enables accurate current control and makes the stimulator a good candidate for generating programmable current waveform.

In the test of passive discharge, the output current amplitude is set to 75.23µA to compare its performance to the modeling in section 4.4.1. As shown in plot (b), the discharging

curve from Cadence simulation and mathematical modeling are very close to each other. This again proves that the modeling method in section 4.4.1 is very accurate. The circuit's discharging current is slightly less than the value predicted by modeling due to the nonideality of MOS switch. Despite that, the circuit is still able to clear 80.54% residual charge which is only slightly lower than the predicted value (83.71%).

In the test of successive stimulations, the stimulator operates for 32ms with output current amplitude drops from 90μA to 0μA. As shown in plot (c)-1, the stimulator faithfully changes output current and performs passive discharge every 2ms. Plot (c)-2 shows how the electrode voltage changes during stimulation. We can see that at the end of 15 stimulations, the electrode voltage goes back to V_{CM} (3V), which means very small net charge injection.

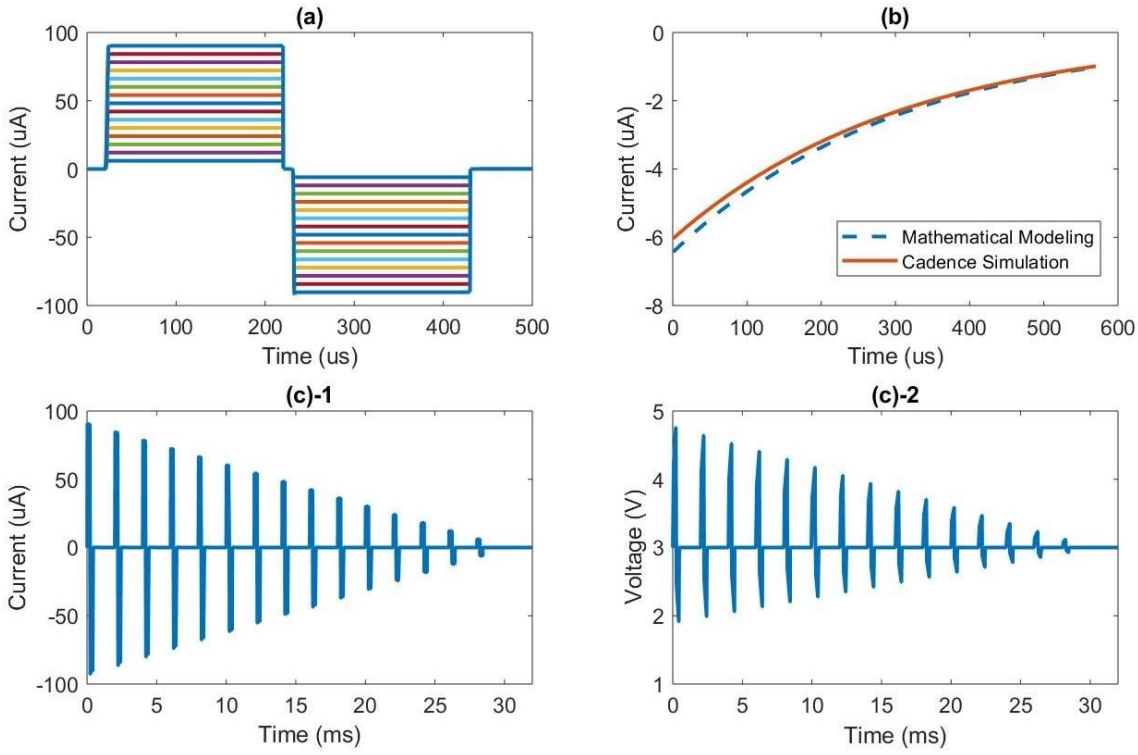

Figure 4.17: Cadence simulation of the final design. (a) All output current waveforms. (b) Test of passive discharge function. (c)-1 Output current during successive stimulation. (c)-2 Electrode voltage during successive stimulation.

To have a more comprehensive evaluation of the charge-balancing performance, the stimulator is also tested under maximum current output at different process corners. Current mismatch is calculated by the equation below. Testing result based on **pre-layout** simulation is shown in Table 4.2.

$$\varepsilon = \frac{I_{net}}{I_{out}} = \frac{\frac{1}{t_{Stim}}\left|\int_{t_0}^{t_0+t_{Stim}} I_{out}(t)dt\right|}{\frac{1}{2}(I_a + I_c)}.$$

Corner	tt	ff	ss	fs	sf	fff	ssf
I_c (μA)	90.29	84.72	96.12	91.12	90.27	88.67	97.08
I_a (μA)	90.29	84.72	96.12	91.13	90.28	88.68	97.08
I_{net} (nA)	94.65	90.54	113.10	257.00	116.20	96.42	126.50
ε (%)	0.10	0.11	0.12	0.28	0.13	0.11	0.13

Table 4.2: Current mismatch at different process corners.

We can see that for typical process, the stimulator has best charge-balancing performance and the current mismatch is only 0.1%. We can also see that the maximum change of output current due to process variation is only 7%. This shows that the design has very robust performance and strong immunity to process variation. Specifications of the final design are summarized in Table 4.3.

Stimulator Type	Monopolar Current-Regulated
Stimulation Current	6μA ~ 90μA
Stimulation Frequency	500Hz
Typical Current Mismatch	0.10%
Current DAC Resolution	4 Bit
Maximum DNL	0.0011 LSB
Maximum INL	0.0010 LSB
Idle Power	3.626uW
Power Supply	6V

Table 4.3: Specifications of the final stimulator design.

Chapter 5 **Peripheral Circuitry**

5.1 Introduction to the Peripheral Circuitry of the System

The peripheral circuitry plays an important role in the timing and mutual control of three major function blocks of the system: the analog front-end, the spike detector and the neural stimulator. It also modulates the current states of the system to more readable output. The peripheral circuitry is primarily formed by a stimulation timing generation module (STGM) and a 4-bit counter. The STGM controls three important phases of the system: the stimulation, the passive discharge and the detection phase. The 4-bit counter counts the number of stimulations performed and increases the stimulation current accordingly. Another important circuit is the level shifter. Since some circuit blocks (such as the stimulator) are operating at high voltage (>1.5V) and most digital logics are operating at low voltage to save power, level shifters are used for converting the output from digital logics to high-voltage control signal. Detailed implementation of the peripheral circuitry will be shown in following sections.

5.2 Design and Implementation of the Level Shifter

The level shifter converts a low-voltage digital signal to a high-voltage digital signal as shown in Figure 5.1. For our system, most outputs from digital logics are 1.5V/0V. The maximum driving voltage required is 6V/3.5V. For the 180nm silicon-on-insulator (SOI) process, to ensure that the transistor does not break down, V_{DS} needs to be kept below 2.75V. To adapt to this voltage limit and fulfill the driving requirement, two types of level shifters are designed. Specification of each type is shown in Table 5.1. The low-voltage type (LVLS) increases the output range but does not change the lower bound of power supply. The high-voltage type (HVLS) not only increases the output range but also levels up the lower bound of power supply.

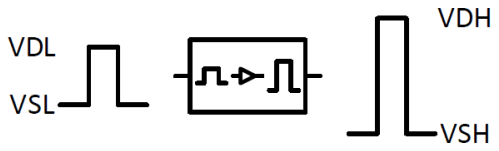

Figure 5.1: The input and output of a level shifter.

	VSL (V)	VDL (V)	VSH (V)	VDH (V)
LVLS	0	1.5	1.5	4
HVLS	0	1.5	0	2.5

Table 5.1: Specification of the level shifter.

The circuit of LVLS and HVLS are shown in Figure 5.2. [27] HV-INV and LV-INV are high-voltage inverter powered by V_{SH}/V_{DH} and low-voltage inverter powered by V_{SL}/V_{DL} respectively. For HVLS, four transistors are stacked on each side to alleviate V_{DS} on a single transistor. For LVLS, since the output voltage is not that high, protection transistors are not used.

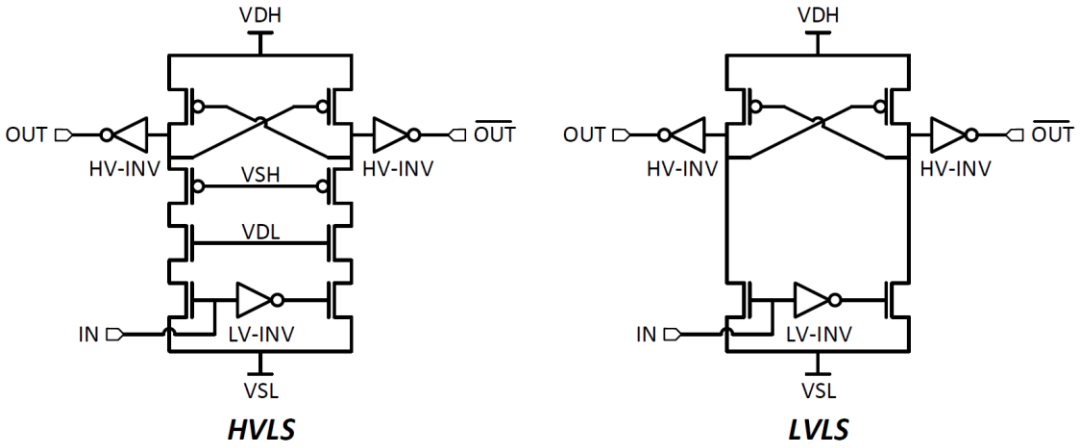

Table 5.2: Circuit of the level shifter.

5.3 Design and Implementation of the STGM

The conceptual circuit diagram of the STGM is shown in Figure 5.2. There are two inputs to the STGM: a global reset signal (RST) and a 100kHz clock signal (CLK). The input clock signal firstly goes through two 5-stage Johnson ring counters to generate a 1kHz clock signal with different delays. The 100kHz input clock and the 1kHz generated clock are then modulated by a clock modulation block (CMB). The CMB is primarily formed by combinational logics and is periodically reset by a 500Hz clock signal generated from the 1kHz clock. The CMB generates four low-voltage digital signals: “PD”, “Detection”, “Cathodic” and “Anodic”. These low-voltage digital signals are then converted to high-voltage control signals by six level shifters. The final outputs are: “ENPD” and “ENPD2” for controlling passive discharge, “Detection” for controlling the detection of action potential, “Up” and “$\overline{\text{Up}}$” for controlling the current source of the stimulator and “Down” and “$\overline{\text{Down}}$” for controlling the current sink of the stimulator.

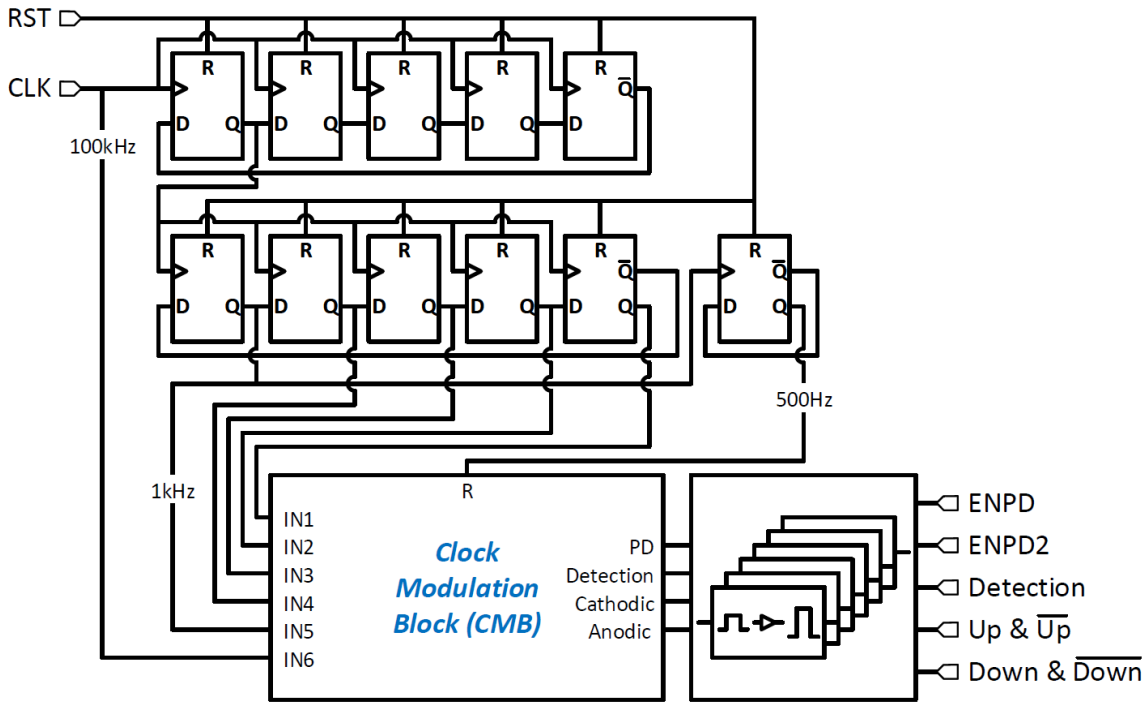

Figure 5.2: Circuit of the stimulation timing generation module (STGM).

Since all outputs have period of 2ms, a 2.1ms transient simulation in Cadence is run to test the STGM. As shown in Figure 5.3, the STGM is firstly reset for 1µs. After reset, there is a 20µs initial state. During the initial state, all output states are 0 except "Detection" which has a 10µs pulse. This initial pulse is intentionally set to ensure that the stimulation current starts from the lowest amplitude. The detailed mechanism is explained in the next section.

From the plot of the output control signals, we can find the operation sequence of the system. After the initial state, the system enters the stimulation phase which comprises a 200µs cathodic phase, a 10µs interphase delay and a 200µs anodic phase sequentially. After the stimulation phase, there is a 10µs interphase delay and then the system enters the 570µs-long passive discharge phase. After the passive discharge phase, the system immediately enters the 1ms detection phase. After the detection phase, the system goes back to the stimulation phase and repeats subsequent tasks.

The designed STGM shows robust performance at different process corners. The dynamic power consumption is 69nW.

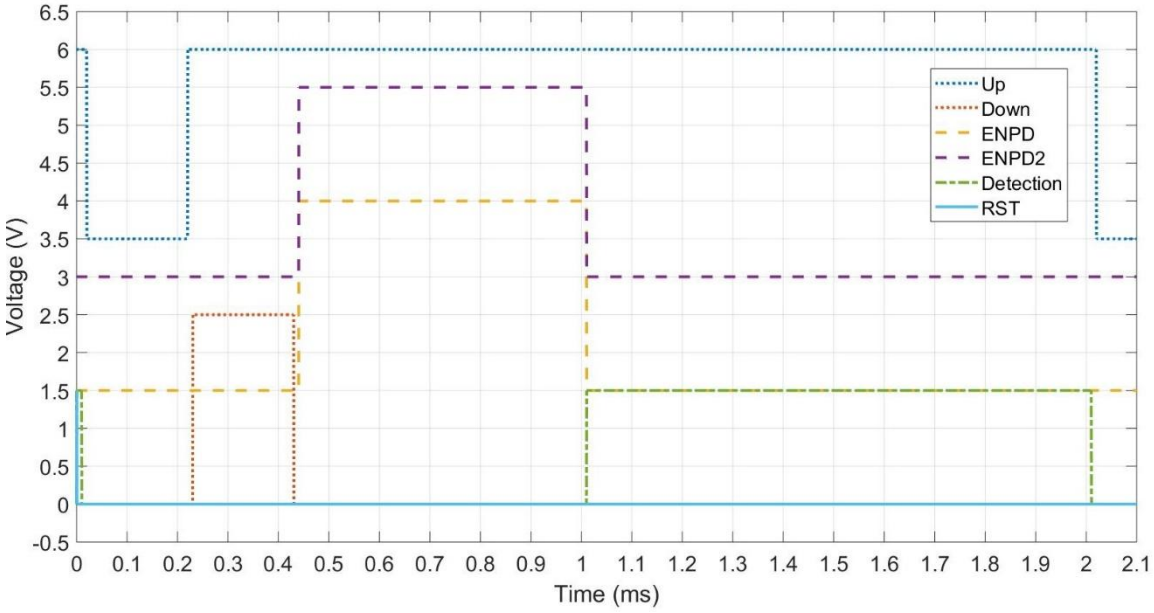

Figure 5.3: Outputs from the STGM.

5.4 Design and Implementation of the 4-Bit Counter

The counter has two inputs: a global reset signal (RST) and an input clock signal (CLK). The clock signal is counted by a 4-bit ripple counter. Each output of the ripple counter is then connected to a digital buffer. The buffered outputs from the ripple counter are converted to a voltage signal (V_{Bout}) by an R-2R DAC and converted to four high-voltage digital control signals ($B_{0\sim3}$) by four level shifters. Notice that when CLK is 1, all outputs from level shifters are forced to be 0.

When the counter is deployed in the system, its input CLK is connected to the output "Detection" from the STGM. Its four outputs $B_{0\sim3}$ are connected to the digital control ports of the neural stimulator $V_{B0\sim B3}$ respectively. Therefore, the counter counts the number of stimulations performed and encodes this number to a voltage output. From this voltage output, we can derive the current pulse amplitude of last stimulation. Besides counting, the counter also increases the amplitude of stimulation current by one LSB after each "Detection" pulse. Section 5.3 mentions that "Detection" has a 10μs initial pulse during the initial state of the STGM. This initial pulse is intentionally set to make sure that $B_{0\sim3}$ from the counter starts from 0001 or equivalently the stimulator starts from the minimum stimulation current.

Figure 5.5 shows a 32ms transient simulation in Cadence. We can see that both V_{Bout} and $B_{0\sim3}$ accurately represent the number of counted pulses. Notice that $B_{0\sim3}$ are set to 0000 when CLK is 1 to ensure that the stimulator is turned off during detection. The dynamic power consumption of the counter is 251nW.

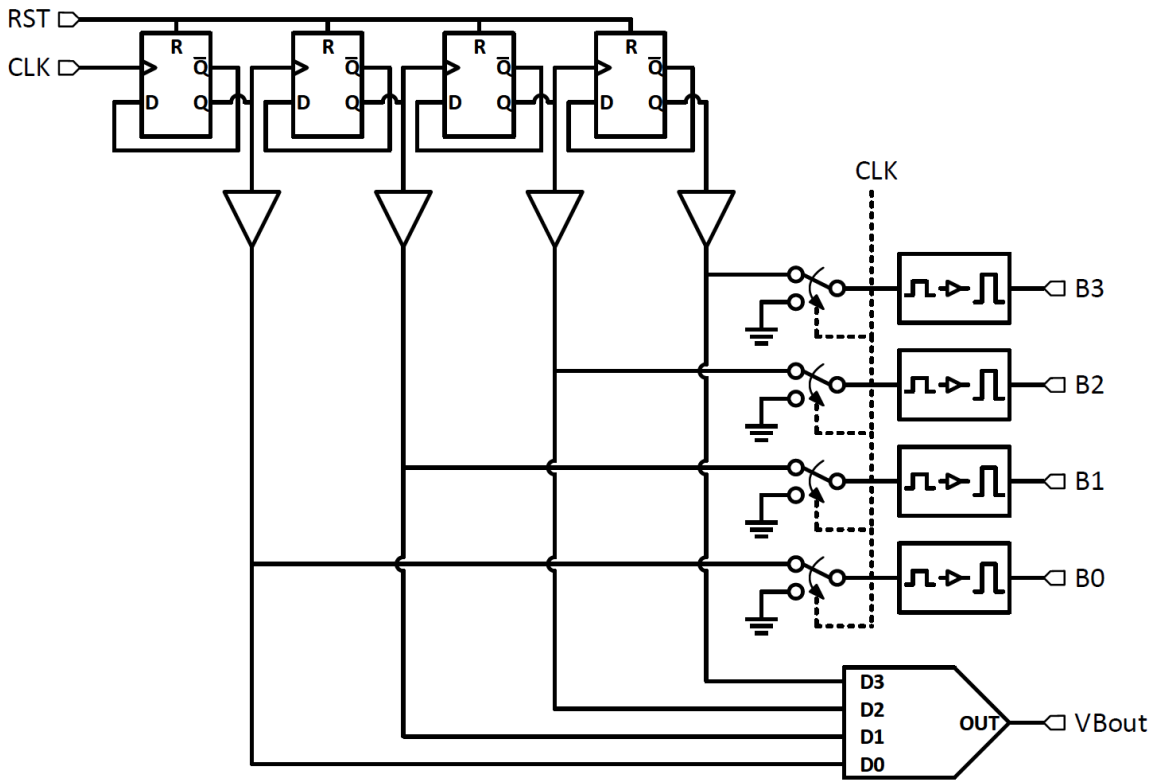

Figure 5.4: 4-Bit Counter.

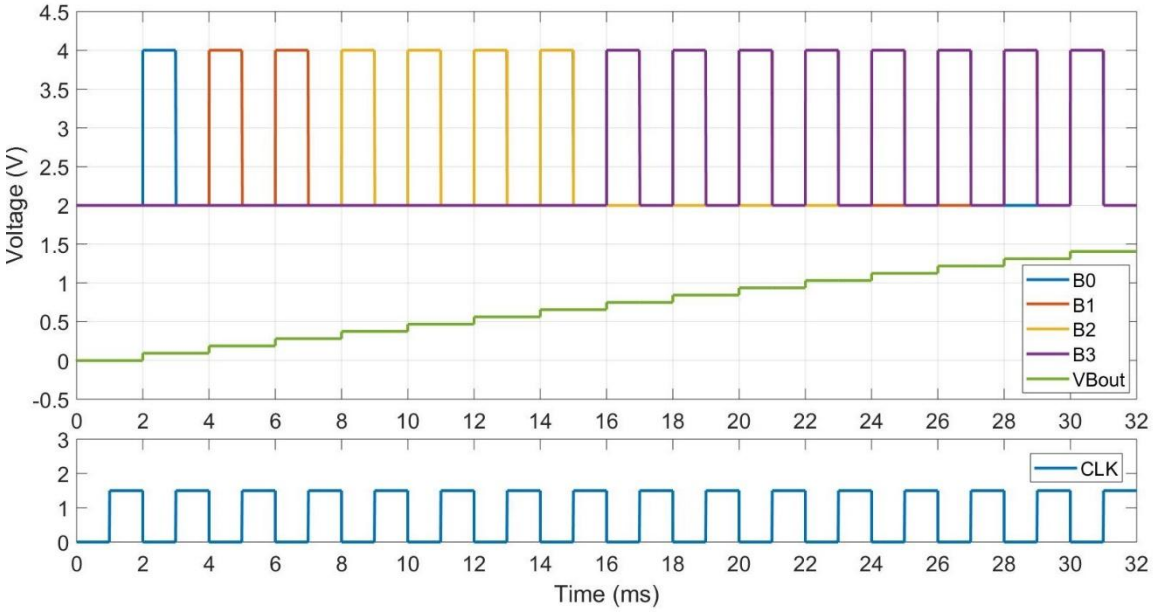

Figure 5.5: Outputs from the 4-bit counter.

Chapter 6 System Integration

6.1 Introduction to the Integrated System

This chapter describes the integration of the closed-loop neurostimulation (CLN) system by using circuit blocks described in previous chapters. The system aims to find the minimum stimulation current that is needed to evoke an action potential from neurons. The operational scheme is to perform stimulation and detection sequentially and repeatedly, and gradually increase the stimulation current until the action potential is observed. Details about the system can found in Chapter 1. The final design was implemented in 180nm silicon-on-insulator (SOI) process and was submitted to the manufacturer in November 2018.

The chapter is organized as follows: Section 6.2 describes the architecture and mechanism of the system. Section 6.3 shows the implementation of the system. Section 6.4 presents the simulation of a complete operation of the system.

6.2 System Architecture

Figure 6.1 shows the conceptual circuit diagram of the implemented system. The complete CLN system has five major blocks: the analog front-end (AFE), the spike detector, the stimulation timing generation module (STGM), the counter and the neural stimulator. Due to the limited time for the design, the AFE has not been integrated into the implemented system. However, since the core function of the system (detection and stimulation) only relies on the other four blocks, the system can still be tested in the simulation. For the chip measurement, an off-chip low-noise amplifier will be used to complement the implemented chip to form the closed-loop system.

Here we will look at the function of the complete system. As shown in the Figure 6.1, a closed loop is formed by the stimulator, the AFE, the spike detector and the STGM. The STGM controls the stimulator to perform charge-balanced current-mode neurostimulation to the tissue. The intensity of the stimulation current is controlled by the counter and will increase by one LSB (around 4.36μA) for the next stimulation. The working electrode is also connected to the AFE, whose output is connected to the spike detector. After each stimulation, the system turns the stimulator off and starts the action potential detection. The neural signal from the tissue gets amplified by the AFE and is then analyzed by the spike detector. If the spike detector does not detect any spike, the STGM will continue normal function. The system will start another stimulation with higher intensity and repeat subsequent tasks. If the spike detector detects the spike, the system will recognize the intensity of the last stimulation as the minimum stimulation current needed to evoke

an action potential. There is then no need to continue the stimulation-detection process. The spike detector will thereby terminate the STGM by keeping its reset signal high. This essentially stops the operation of the system.

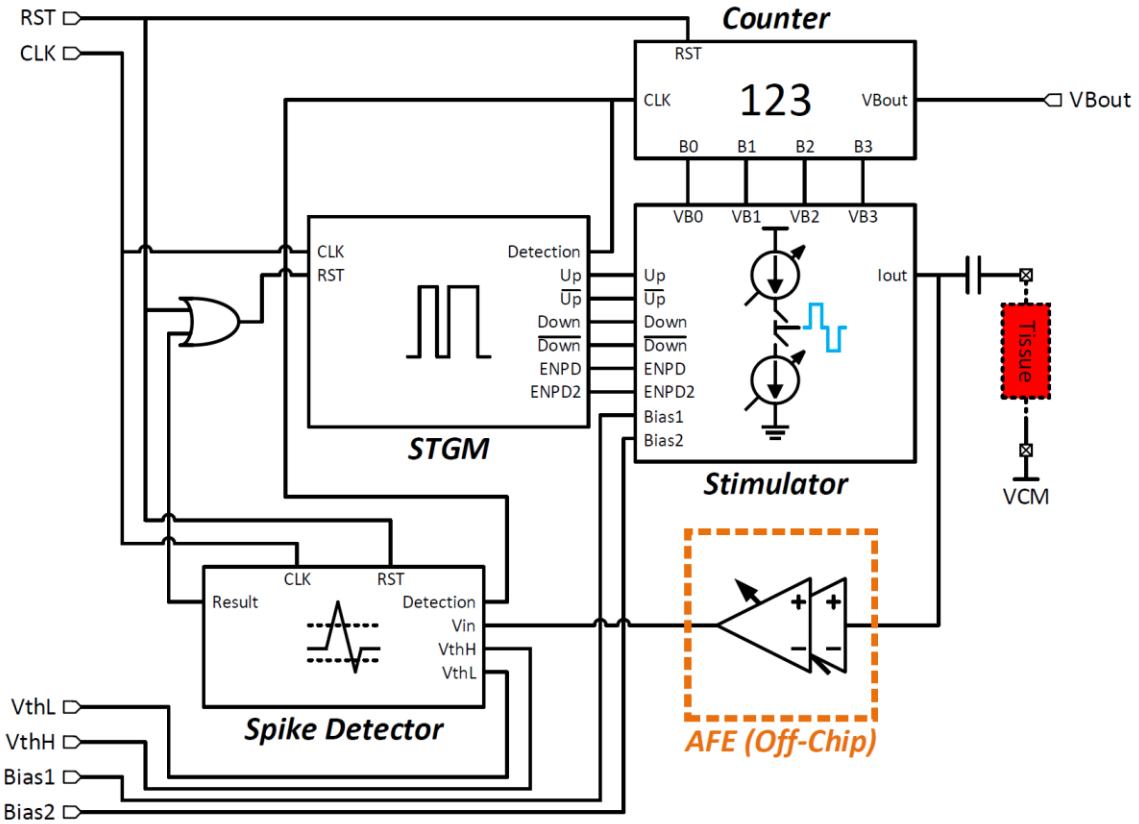

Figure 6.1: Conceptual circuit diagram of the implemented CLN system.

The system has six inputs: a global reset signal (RST), a 100kHz input clock signal (CLK), two bias voltage signals for setting up the ATD-based spike detector (V_{thL} and V_{thH}) and two bias voltage signals for setting up the LSB of the current-mode neural stimulator (Bias1 and Bias2). There is only one signal output from the system: V_{Bout} from the 4-bit counter. V_{Bout} reports the number of stimulations performed, from which we can calculate the intensity of the last stimulation. Details can be seen in Chapter 5.

Notice that there is one slight difference between the stimulator in the implemented system and the stimulator described in Chapter 4, Section 4.4.2. The former one needs external biasing to set up the LSB of its internal current DAC. Its circuit is shown in Figure 6.2. The latter one has on-chip biasing generation circuit for setting up the LSB. It was designed after chip fabrication and is thereby not part of the system. Other than the biasing, two stimulators have a fully identical structure. Details about the circuit structure can be found in Chapter 4, Section 4.4.2.

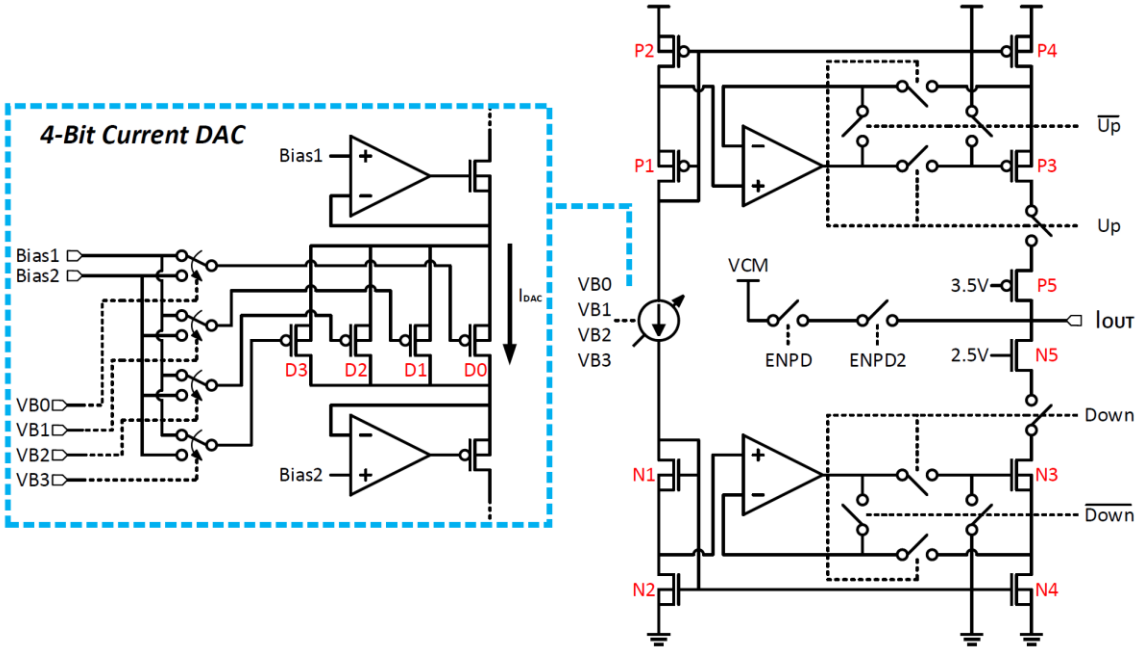

Figure 6.2: Circuit of the neural stimulator in the implemented system.

6.3 Implementation of the System

The layout of the chip is shown in Figure 6.3. Specifications of the chip are summarized in Table 6.1. Detailed specifications of each block can be found in previous chapters.

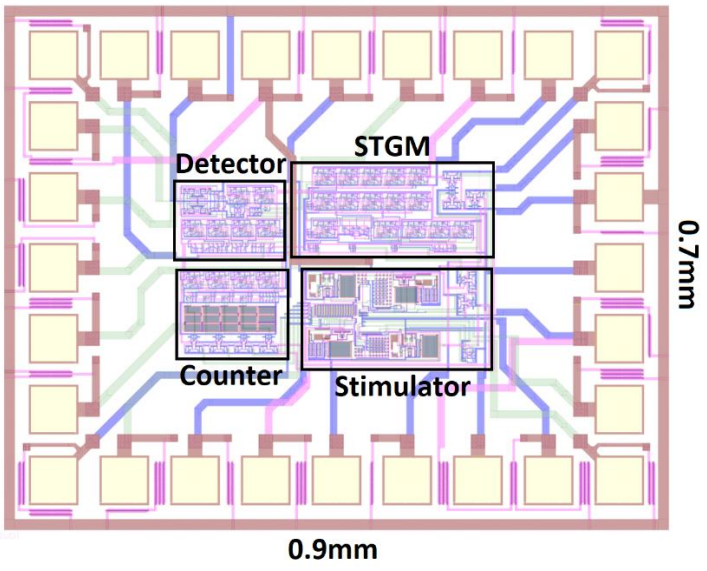

Figure 6.3: Layout of the implemented system.

Chip Dimension	0.9mm × 0.7mm
Idle Power Consumption	4μW
Power Supply	0V ~ 6V
Stimulator Type	Monopolar Current-Regulated
Stimulation Current	4.36μA ~ 65.40μA (Adjustable)
Stimulation Current Resolution	4.36μA (Adjustable)
Stimulation Frequency	500Hz
Typical Stimulation Current Mismatch	0.10%

Table 6.1: Specifications of the chip.

6.3.1 Overview of Applied Layout Techniques

This section describes some important layout techniques applied to the implementation of the system. Most layout issues can be avoided by complying with the layout rules from the PDK. However, some design-specific layout issues still need to be addressed by the designer.

- **Electromigration:** Due to the low-power and low-current feature of the system, the electromigration issue is not severe. However, for reliability, all power supply traces have larger width than the minimum specified one and are implemented in metal layers with better current handling ability.

- **Antenna Effect:** For our design, the antenna effect is most significant in the digital circuitry where many gates could be connected to one long metal trace. This issue is addressed by shunt wiring. More specifically, a tie-down diode realized by a diode-connected transistor is connected to the gate of each protected transistor to avoid charge accumulation on the metal trace connected to the gate.

- **Gate Shadowing Effect:** The gate shadowing effect normally affects the analog circuit sensitive to symmetry. [26] One layout example can be seen in Figure 6.4. For the differential pair, four transistors are vertically placed to ensure that drain and source region see same surrounding environment. This improves the symmetry of the circuit. Same approach is applied to the current DAC and the current mirror shown in the graph.

- **Well Proximity Effect:** For current copying and multiplication, multiple transistors with same size are connected in parallel to form a single transistor with equivalent aspect ratio. For the current mirror in Figure 6.4, the current copying ratio is 20. Any deviation from this ratio will cause current mismatch of the stimulator. To minimize the transistor mismatch, 20 transistors with the same size are used. To ensure that each transistor sees same surrounding environment, dummy transistors are also placed around the current mirror. Similar approach is applied to the current DAC.

- **ESD:** Double-gate ESD diodes are used for ESD protection. For high-voltage power supply, stacked diodes are used.

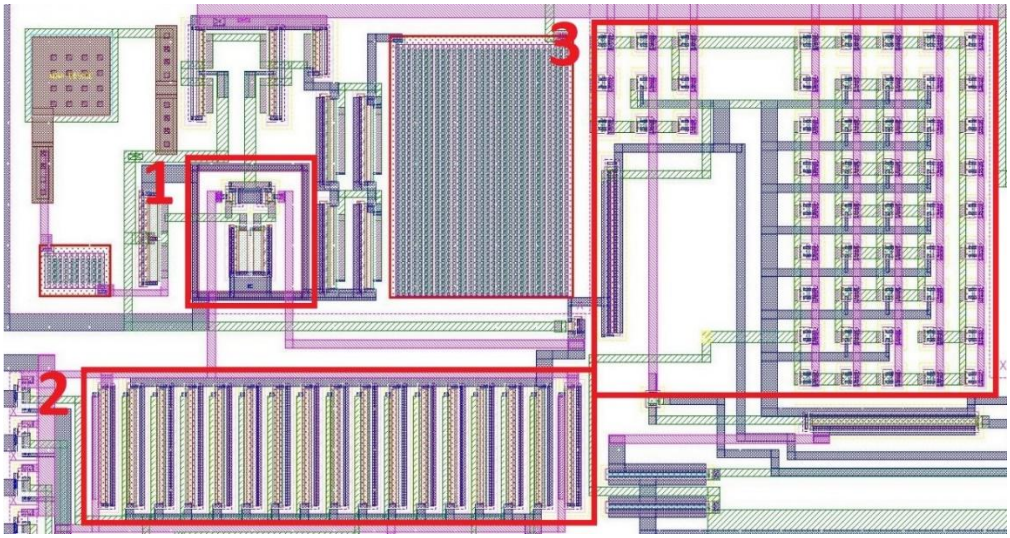

Figure 6.4: Partial layout of the stimulator: 1) Differential pair. 2) Current DAC. 3) Current mirror.

6.4 Simulation of the System

For the designed system, Cadence simulation can only capture its open-loop behavior. Some important side effects in closed-loop BMI systems (such as the stimulation artifact) can hardly be accurately modeled and need to be measured in vitro or in vivo. However, the open-loop simulation still provides valuable information about basic functions of the system and is essential for the evaluation of the performance of the system.

In terms of stimulation artifact, since the system does not perform stimulation and detection simultaneously, the issue is already alleviated. Besides that, Chapter 4 shows that the 570μs-long passive discharge can quickly clear any residual charge on the electrode so that when detection begins, the electrode voltage already settles down to V_{CM}. Therefore, we assume that the stimulation artifact has only a small impact on the system.

To test the system, we break the loop by disconnecting the electrode from the AFE and replace the AFE with an ideal neural signal source. The signal is recorded from the hippocampus region and has already been amplified. [16] The electrode-tissue interface model used is the asymmetrical model with 20% variation. Detailed information can be found in Chapter 4, Section 4.2. A blocking capacitor of 1μF is used.

The **post-layout** simulation results are shown below. Figure 6.5 shows one complete operation of the system. We can see that the intensity of the stimulation current increases by one LSB (4.36μA) every 2ms. V_{Bout} which represents number of stimulations performed also increases by one LSB (93.75mV) after each stimulation. After the fourteenth stimulation, a neural spike occurs at around 26ms. This triggers the spike detector which sends a flag signal ("Result") to the STGM. The system then stops the stimulation-detection process. The output V_{Bout} stays at 1.406V indicating that 14 stimulations have been performed and the intensity of final stimulation is 14 LSB or 61.04μA.

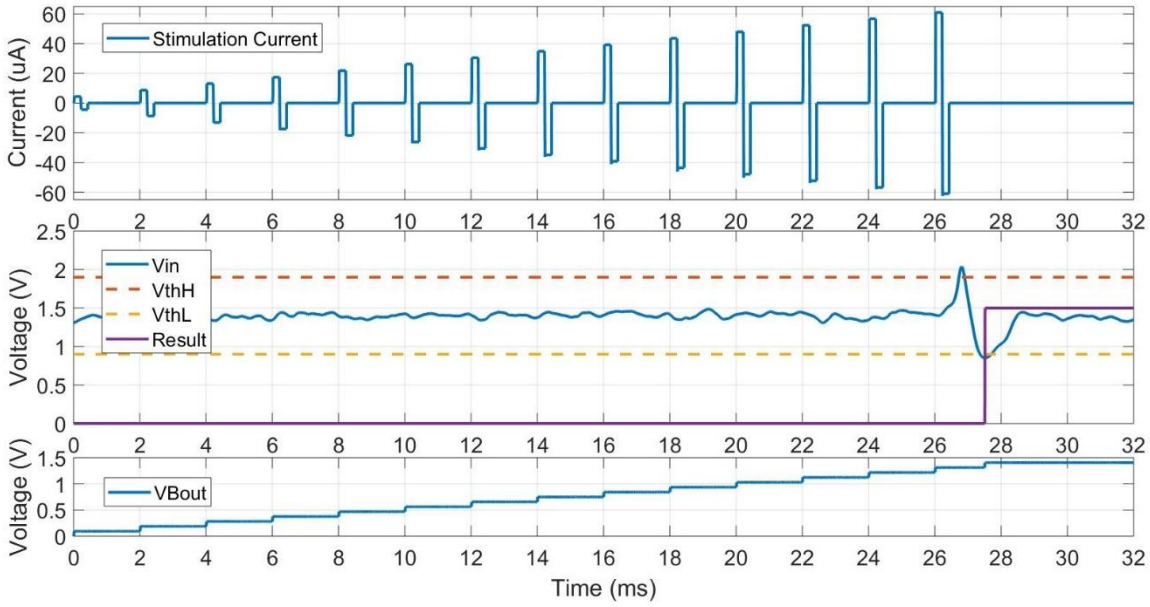

Figure 6.5: Transient simulation of one complete operation of the system.

Figure 6.6 shows the timing of the three phases. We can see that the detection phase has an initial pulse to set up the first stimulation. After that, these three actions: stimulation, passive discharge and detection happen sequentially and repeat themselves with period of 2ms until a spike is detected.

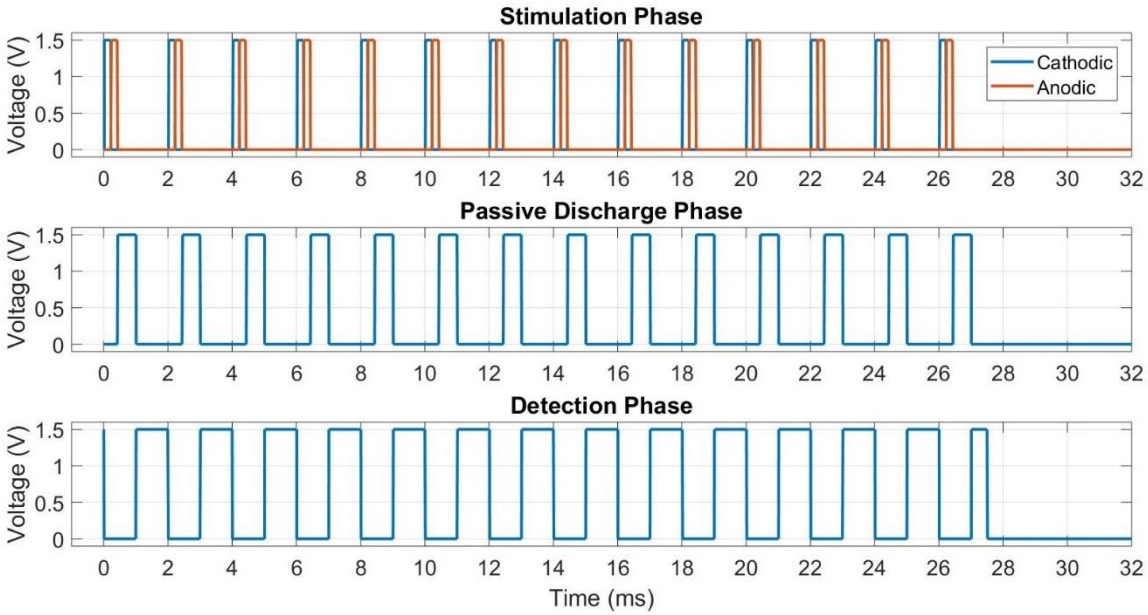

Figure 6.6: Timing of the three phases of the system.

Figure 6.7 shows the stimulation current and the electrode voltage during operation. We can see that at the end of operation, the electrode voltage settles down to V_{CM}. The intensity of the final stimulation is 61.00μA which is very close to the expected value (61.04μA). The small discrepancy is caused by the deviation from the designed current copying ratio, which is possibly due to the transistor mismatch in the layout.

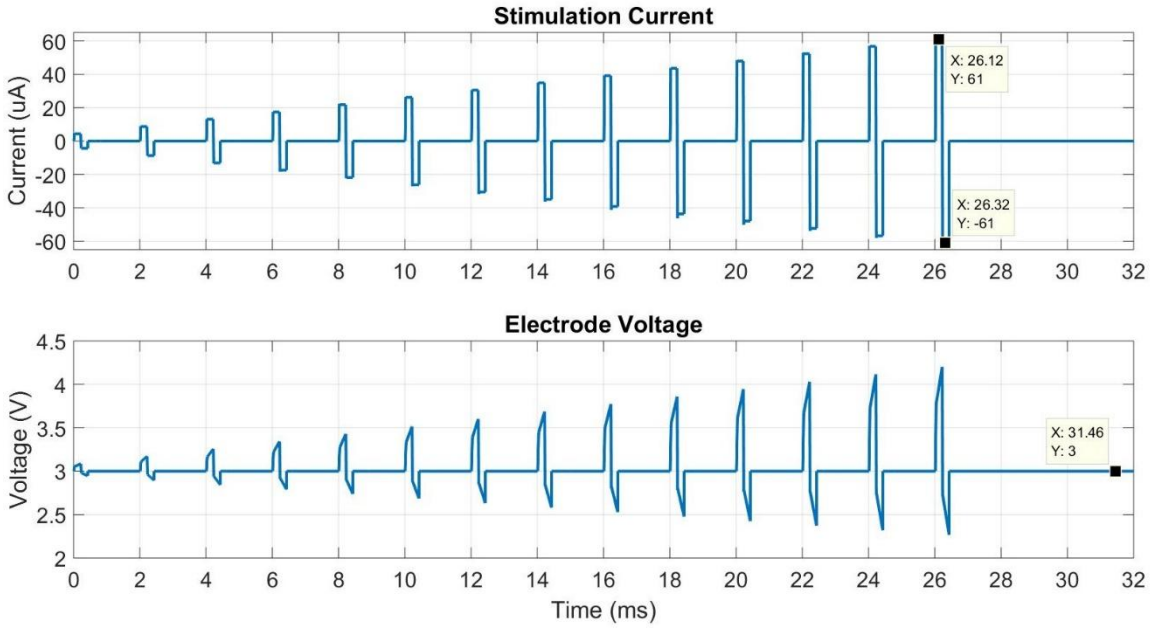

Figure 6.7: Stimulation current and electrode voltage during operation.

In conclusion, the integrated CLN system accomplishes well the designed operation flow. The implemented stimulator indicates good charge-balancing performance and high accuracy in current control mode. The whole system consumes only 4μW in idle state and occupies a chip area less than $1mm^2$.

Chapter 7 **Linear-Phase Delay Filter**

7.1 Design Background

This chapter describes the design of a low-power continuous-time linear-phase delay filter, which is part of the analog signal processing block of an on-chip neural spike classification chip. The designed filter aims to provide wideband large group delay and small signal distortion to avoid introducing errors for later signal processing.

7.2 Design of a Linear-Phase All-Pass Transfer Function

A typical all-pass transfer function is given by:

$$H(s) = H_0 \frac{P(-s)}{P(s)},$$

where P(s) is a polynomial function and H_0 is the gain of the filter. The group delay of the transfer function is:

$$\text{grd}[H(s)] = -\frac{d\phi_{H(s)}}{d\omega}.$$

We can easily find that:

$$\text{grd}[H(s)] = 2 \times \frac{d\phi_{P(s)}}{d\omega}.$$

Therefore, we only need to find the derivative of the phase of polynomial P(s) and then the total group delay is simply two times of that value. For a large group delay, generally a high-order polynomial function is required. Among the proposed polynomials, an equal-ripple function provides the largest delay bandwidth for a given order by uniformly introducing small ripples of delay time (<3%) over delay bandwidth. [28]

To ensure normal functioning of the spike classifier, the required delay is up to 600μs. Based on this requirement, a ninth-order equal-ripple polynomial function is chosen to construct the filter. The normalized-frequency function of the chosen polynomial is given below: [28]

$$P_{ER}(s) = (s + 0.5729)(s^2 + 0.8158s + 8.5927)(s^2 + 1.0355s + 5.2528)(s^2 + 1.1095s + 2.5840)(s^2 + 1.1380s + 0.9006).$$

The delay and the constant-delay bandwidth of the normalized-frequency function are given below:

$$\begin{cases} D_0 = 3.655s \\ \omega_0 = 2.93Hz \end{cases}.$$

The bandwidth-delay product of scaled-frequency function remains the same. Namely, we have:

$$D_s\omega_s = D_0\omega_0.$$

The frequency scaling factor is given by:

$$k = \frac{\omega_s}{\omega_0} = \frac{D_0}{D_s}.$$

We can see that the scaled transfer function can have larger delay over smaller bandwidth (more signal distortion) or smaller delay over larger bandwidth (less signal distortion). For this application, we implemented two delay cases: 400μs and 600μs. The corresponding delay bandwidth and frequency scaling factor is given below:

$$f_s = \begin{cases} 5.68kHz & for \quad D_s = 600\mu s \\ 8.52kHz & for \quad D_s = 400\mu s \end{cases},$$

$$k = \begin{cases} 12183 & for \quad D_s = 600\mu s \\ 18275 & for \quad D_s = 400\mu s \end{cases}.$$

Therefore, even the smallest delay bandwidth can still cover the frequency spectrum of typical action potential signal. Notice that this is only for the ideal transfer function. The implemented transfer function normally has smaller bandwidth due to the circuit nonlinearity.

To scale the frequency, consider the normalized-frequency second-order all-pass function shown below as an example:

$$H_{2nd}(\omega) = \frac{-\omega^2 + ja_n\omega + b_n}{-\omega^2 - ja_n\omega + b_n},$$

where a_n and b_n are normalized polynomial coefficients. The corresponding scaled-frequency function is then given by:

$$H_{2nd}(\omega_s) = \frac{\frac{-\omega_s^2}{k^2} + \frac{ja_n\omega_s}{k} + b_n}{\frac{-\omega_s^2}{k^2} - \frac{ja_n\omega_s}{k} + b_n} = \frac{-\omega_s^2 + ja_nk\omega_s + b_nk^2}{-\omega_s^2 - ja_nk\omega_s + b_nk^2} = \frac{-\omega^2 + ja_s\omega + b_s}{-\omega^2 - ja_s\omega + b_s}.$$

Therefore, the scaled polynomial coefficients are given by:

$$\begin{cases} a_s = a_n k \\ b_s = b_n k^2 \end{cases}.$$

Based on this property, we can configure the circuit to change the polynomial coefficients. This equivalently changes the delay time of the filter. Detailed implementation is explained in Section 7.3.

7.3 Circuit Implementation

This section describes the circuit implementation of the ninth-order equal-ripple all-pass filter (ERAPF). The transfer function of the filter is given by:

$$H_{ERAPF}(s) = H_0 \frac{P_{ER}(-s)}{P_{ER}(s)} = H_0 \times \frac{(-s + a_1)}{(s + a_1)} \times \frac{(s^2 - a_2 s + b_2)}{(s^2 + a_2 s + b_2)} \times \ldots \times \frac{(s^2 - a_5 s + b_5)}{(s^2 + a_5 s + b_5)}$$

$$= H_0 \times H_1(s) \times H_2(s) \times H_3(s) \times H_4(s) \times H_5(s),$$

where H_0 is the DC gain of the filter and $P_{ER}(s)$ is the scaled-frequency ninth-order equal-ripple polynomial function. We can see that $H_1(s)$ is a first-order all-pass filter and $H_{2\sim5}(s)$ are second-order all-pass filters. The strategy is thereby to implement them separately and cascade all five blocks to form the whole transfer function.

7.3.1 Implementation of a First-Order All-Pass Filter

The circuit of the first-order all-pass filter is shown in Figure 7.1.

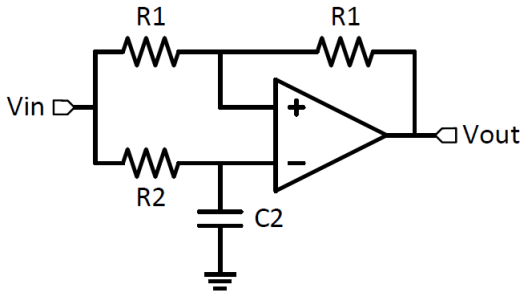

Figure 7.1: Circuit of a first-order all-pass filter.

For sufficiently large gain provided by the amplifier, the transfer function of the circuit is given by:

$$H_1(s) = \frac{\frac{1}{R_2C_2} - s}{\frac{1}{R_2C_2} + s} = \frac{(a_1 - s)}{(a_1 + s)}.$$

7.3.2 Implementation of a Second-Order All-Pass Filter

The second-order all-pass filter is implemented by Delyiannis circuit shown in Figure 7.2. [29]

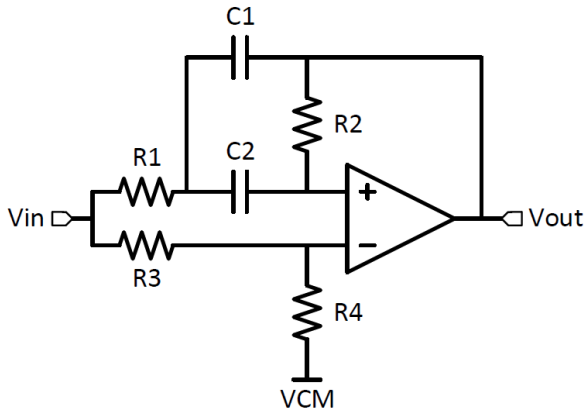

Figure 7.2: Circuit of a second-order all-pass filter.

The transfer function of the circuit is given by:

$$H_i(s) = \frac{A_v}{1 + A_v} \times \frac{R_4}{R_3 + R_4} \times \left[\frac{1 + \left(C_1R_1 + C_2R_1 - C_2R_3\frac{R_2}{R_4}\right)s + C_1C_2R_1R_2s^2}{1 + \left(C_1R_1 + C_2R_1 + \frac{C_2R_2}{1 + A_v}\right)s + C_1C_2R_1R_2s^2}\right],$$

where A_v is the gain of the amplifier. For very large gain, this transfer function is:

$$H_i(s) \approx \frac{R_4}{R_3 + R_4} \times \left[\frac{1 + \left(C_1R_1 + C_2R_1 - C_2R_3\frac{R_2}{R_4}\right)s + C_1C_2R_1R_2s^2}{1 + (C_1R_1 + C_2R_1)s + C_1C_2R_1R_2s^2}\right].$$

If we take C_1=C_2=C, R_1=R_3 and R_2=4×R_4, then we have:

$$H_i(s) = \frac{R_4}{R_3 + R_4} \times \left[\frac{1 - 2R_1Cs + R_1R_2C^2s^2}{1 + 2R_1Cs + R_1R_2C^2s^2}\right] = A_i \times \frac{(s^2 - a_is + b_i)}{(s^2 + a_is + b_i)},$$

which is an ideal second-order all-pass function with gain given below:

$$A_i = \frac{R_4}{R_3 + R_4}.$$

We can see that this transfer function has a gain of less than 1V/V. After cascading 4 second-order all-pass filters, the gain of the whole circuit is around 0.2V/V.

It should be noticed that if the amplifier's gain is not extremely large, the approximation of the transfer function is not accurate. If we consider the finite gain effect and choose C_1=C_2=C, R_1=R_3 and R_2=4×R_4, we have:

$$H_i(s) = \frac{A_v}{1 + A_v} \times \frac{R_4}{R_3 + R_4} \times \left[\frac{1 - 2CR_1 s + R_1 R_2 C^2 s^2}{1 + \left(2CR_1 + \frac{CR_2}{1 + A_v}\right) s + R_1 R_2 C^2 s^2} \right].$$

In fact, the term $CR_2/(1+A_v)$ may be comparable with $2CR_1$ when the gain is not sufficiently large. This additional term leads to a notch response of the filter and the notch frequency normally falls into the bandwidth of interest. One way to fix this problem without changing R_1 and R_2 is to increase R_3. We can write down the transfer function in following form:

$$H_i(s) = A_i \times \frac{(s^2 - a_{i1} s + b_{i1})}{(s^2 + a_{i2} s + b_{i2})}.$$

According to the transfer function derived, b_{i1} and b_{i2} are always the same. To obtain an all-pass function, we need to equate a_{i1} and a_{i2}. Namely, we want:

$$-\left(C_1 R_1 + C_2 R_1 - C_2 R_3 \frac{R_2}{R_4}\right) = C_1 R_1 + C_2 R_1 + \frac{C_2 R_2}{1 + A_v}.$$

This yields the equation of R_3 in terms of the finite gain:

$$R_3 = R_1 + \frac{R_2}{4(1 + A_v)}.$$

In circuit simulation, we can sweep value of R_3 to make the magnitude response flat. When a_{i2} is larger than a_{i1}, the filter tends to have a notch response. When a_{i2} is smaller than a_{i1}, the filter's amplitude response tends to peak.

7.3.3 Implementation of the ERAPF with Alterable Delay

Assume that all capacitors used have same capacitance C. From previous circuit analysis, we know that the scaled coefficient of first-order all-pass filter is:

$$a_s = \frac{1}{R_2 C} = a_n k.$$

The scaled coefficients of second-order all-pass filter are:

$$\begin{cases} a_s = \dfrac{2}{R_2 C} = a_n k \\ b_s = \dfrac{1}{R_1 R_2 C^2} = b_n k^2 \end{cases}.$$

Therefore, if R_1 and R_2 are fixed, we have:

$$\frac{1}{k} \propto C.$$

Notice that the delay of the filter is:

$$Delay = 2D_s = \frac{2D_0}{k} \Longrightarrow Delay \propto C.$$

This means that we can linearly change the delay by simply changing the capacitance C. The implemented ERAPF with alterable delay is shown in Figure 7.3.

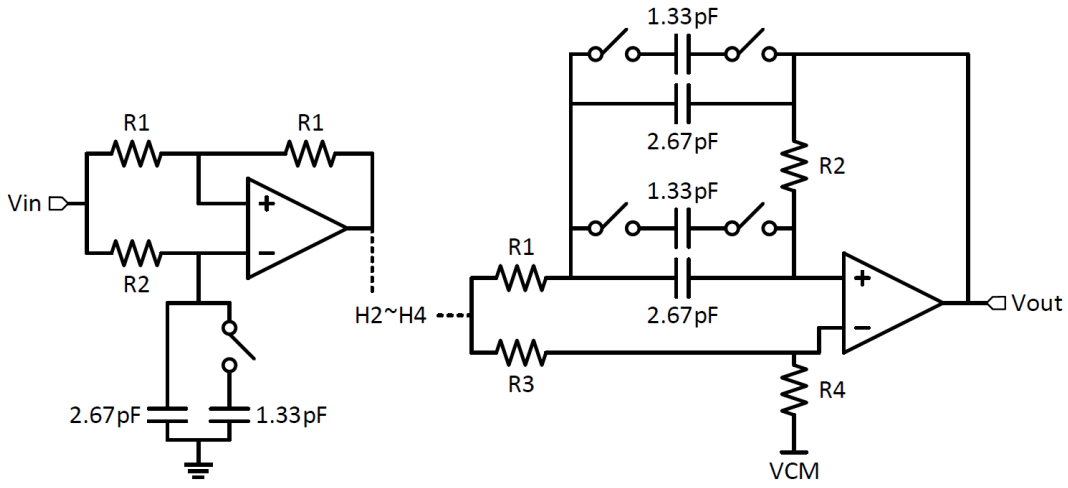

Figure 7.3: Circuit of ninth-order ERAPF with alterable delay.

As shown in the circuit diagram, when all switches are closed, the capacitance C is 4pF which corresponds to the 600μs-delay case. When all switches are open, C drops to around two thirds of 4pF which corresponds to the 400μs-delay case. Based on the capacitance C and considering the finite gain effect, we can calculate $R_{1\sim4}$ of each block. These resistance values are given in Table 7.1. The amplifier used in each block is a two-stage amplifier with Miller compensation.

Transfer Function	R1 (MΩ)	R2 (MΩ)	R3 (MΩ)	R4 (MΩ)
H1	20.8174	35.7563	N/A	N/A
H2	0.9750	50.3254	1.2149	12.6299
H3	2.0381	39.2726	2.1401	9.9338
H4	4.4016	36.9244	4.5730	9.2726
H5	13.0195	35.9899	13.2399	9.0522

Table 7.1: $R_{1\sim4}$ of implemented filter.

7.4 Cadence Simulation

The delay and amplitude response of the filter from the **pre-layout** AC simulation are shown in Figure 7.4. We can see that the amplitude response is very flat for both delay cases. This means that the filter has small amplitude distortion. The delay bandwidth of 400μs-delay case and 600μs-delay case are 5kHz and 3.16kHz approximately. These values are smaller than the values predicted by the mathematical model of the filter. One explanation is that we did not consider the loading effect when we cascaded all five blocks. Since the input impedance of each block is not infinite, it can change the transfer function of last block. This change of transfer function shrinks the delay bandwidth but has small impact on the delay time. For neural signal, most of the frequency components are concentrated at the low-frequency region. Therefore, the signal distortion, as shown in the transient simulation, is still acceptable.

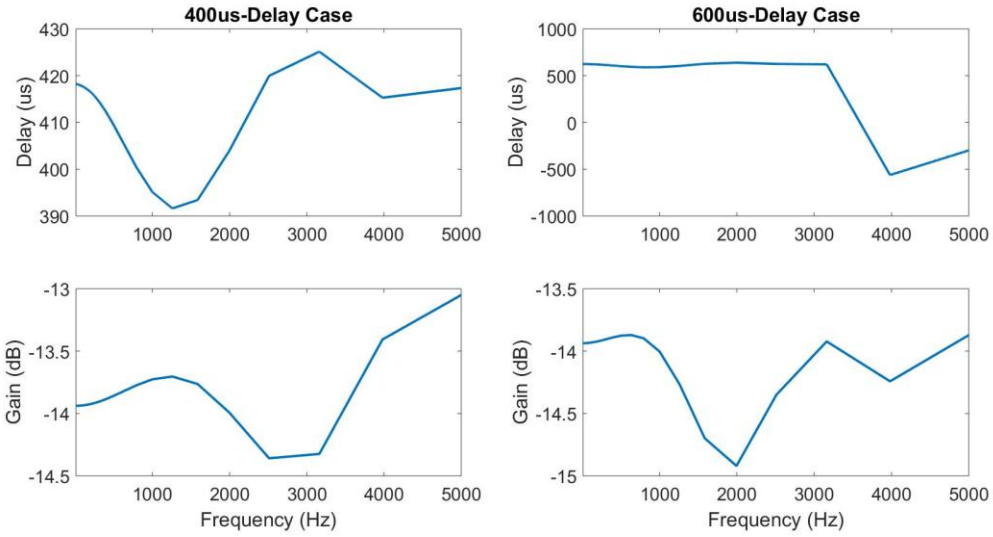

Figure 7.4: Delay and amplitude response of the filter.

The filter's delay function is tested by an amplified 200ms-long real neural signal recorded from hippocampus region. [16] The neural signal has peak-to-peak voltage up to 700mV and common-mode voltage of 700mV. The output of the filter is firstly amplified to compensate the gain of the filter and then compared with the ideally delayed neural signal to calculate the error introduced by the filter. The error equation is given below:

$$Error(t) = \frac{\left|V_{sig}(t-\tau) - \frac{1}{H_0} \times V_{out}(t)\right|}{\left|V_{sig}(t-\tau)\right|} \times 100\%,$$

where τ is the delay time and H_0 is the gain of the filter. Figure 7.5 shows a 15ms section of the 200ms **pre-layout** transient simulation. We can see that for both delay cases, the delayed signal maintains the shape of original signal well. The 400µs-delay case has lower error due to its larger delay bandwidth.

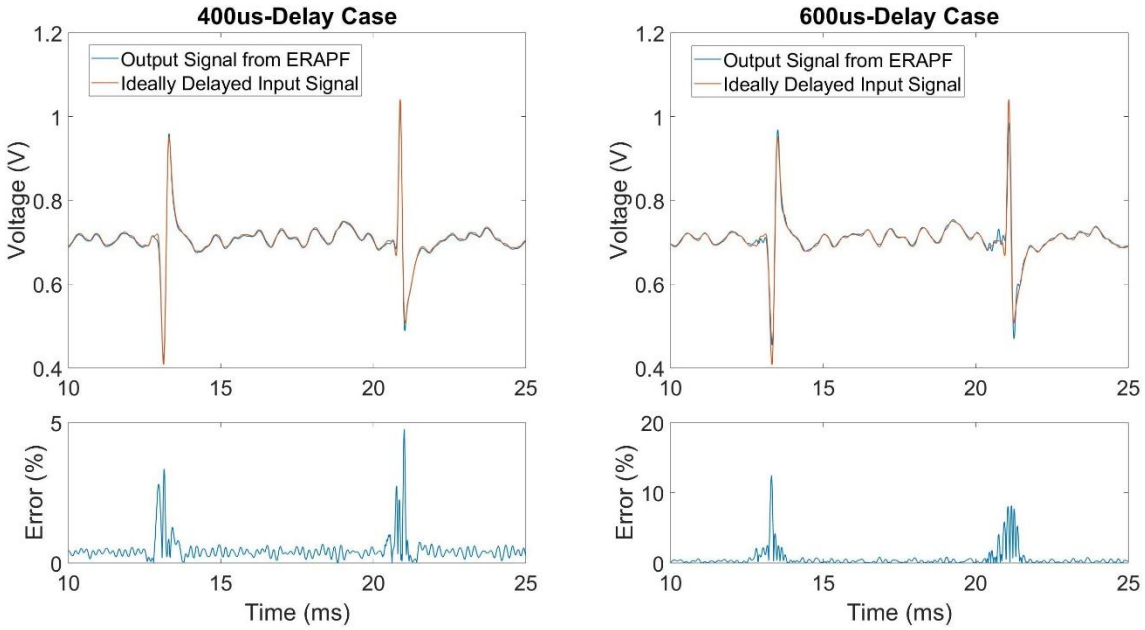

Figure 7.5: Transient simulation of the filter.

The filter has also been tested at different process corners. The **pre-layout** simulation shows that although the delay time fluctuates, the filter can reliably delay the signal with acceptable distortion level. Delay time at each corner is shown in Table 7.2.

	400µs-Delay Case		**600µs-Delay Case**	
Corner	**Delay (µs)**	**Bandwidth (Hz)**	**Delay (µs)**	**Bandwidth (Hz)**
tt	411	5000	613	3160
ff	466	5000	694	3160
fff	468	5000	698	3160
fs	409	5000	612	3160
sf	413	5000	615	3160
ss	372	5000	561	4000
ssf	367	5000	554	4000

Table 7.2: Simulation of the filter at different process corners.

The performance of the filter is assessed by the FOM equation proposed by Gosselin et al., with a small modification. [30] The equation is shown below:

$$FOM = \frac{Power \times V_{DD}}{n \times f \times DR},$$

where n is the number of poles plus zeros, f is the cutoff frequency of the filter and DR is the dynamic range in V/V. The equation of the DR is given below:

$$DR = \frac{Input\ Range}{RMS\ Input\ Referred\ Noise}.$$

For this work, the input range is defined as the range of input for which the total harmonic distortion (THD) of the output signal is below -40dB. Specifications of the implemented filter are summarized in Table 7.3.

	400μs-Delay Case	**600μs-Delay Case**
Power	1.533μW	1.533μW
Power Supply	1.5V	1.5V
Order of the Filter	18	18
Cutoff Frequency	5000Hz	3160Hz
Input Range (This Work)	$800mV_{pp}$	$800mV_{pp}$
Input Range (Gosselin et al.)	N/A	$40mV_{pp}$
Input Referred Noise	$0.5mV_{pp}$	$0.6mV_{pp}$
Dynamic Range	566V/V	471V/V
Delay	411μs	613μs
FOM (This Work)	4.5×10^{-14}	8.6×10^{-14}
FOM (Gosselin et al.)	N/A	18×10^{-14}

Table 7.3: Specifications of the filter.

We can see that the 600μs-delay case, compared with the state-of-the-art design, provides the same delay time but with much larger input range. The FOM has also been reduced to less than one half of the reported value, meaning a better overall performance. Besides the normal delay function, the filter also provides digitally controlled alterable delay time, which could be very useful in low-power analog signal processing applications. The filter also shows strong immunity to process variation.

Chapter 8 **Conclusion and Future Works**

8.1 Conclusion

In this work, a closed-loop neurostimulation system that automatically finds the minimum stimulation current to evoke an action potential from neurons has been proposed. The design and implementation of the system has also been presented. The motivation behind this work is to improve the energy efficiency of the neural stimulator from the perspective of the biological system. Besides the closed-loop neuromodulation system, the thesis has included the implementation of a low-power linear-phase delay filter as an additional but separate work. The filter is used to delay the neural signal and is part of a neural spike classification chip. All designs are implemented in GlobalFoundries 180nm silicon-on-insulator (SOI) process.

The main work is summarized as follows: Chapter 2 presents a general-purpose analog front-end (AFE). The AFE has an input-referred noise less than 5μV over a bandwidth of 3.25Hz ~ 4.65kHz and consumes 14.61μW. The low-power, low-noise and wideband feature makes the AFE suitable for recording action potentials from neurons. The AFE also provides wide and accurate gain tuning with a linear tuning step, which enables accurate control over the signal amplitude.

Chapter 3 presents a 239nW neural spike detector based on the absolute threshold detection (ATD) method. A robust spike detection algorithm is proposed to improve the error tolerance of the stimulator. Testing with a real neural signal confirms that the spike detector can reliably differentiate action potentials from various noise sources.

Chapter 4 gives a comprehensive study of the charge-balancing performance of different types of neural stimulators and the modeling of passive discharge. Based on this study, a charge-balanced monopolar current-regulated stimulator with built-in passive discharge function is designed and implemented. The stimulator has a typical current mismatch less than 0.1% and a strong immunity to PVT variations. The stimulation current can be accurately tuned from 6μA to 90μA with a tuning step of 6μA. In idle state, the stimulator consumes 3.626μW.

Chapter 5 describes the peripheral circuitry of the system, including high-voltage level shifters, a 69nW stimulation timing generation module (STGM) and a 251nW 4-bit counter.

Chapter 6 gives the integration of the overall system. Layout and post-layout simulations of the system are included. The integrated system accomplishes well the designed operation flow and demonstrates excellent charge-balancing performance and accurate control over the stimulation current. The whole system consumes 4μW in idle state and occupies a chip area less than $1mm^2$.

Chapter 7 presents a low-power continuous-time linear-phase delay filter, which provides wideband large group delay, small signal distortion and large dynamic range. The figure of merit (FOM) of the filter is less than one half of the FOM of the state-of-the-art design. Besides the normal delay function, the filter provides digitally controlled alterable delay time. The filter also shows a strong immunity to PVT variations.

8.2 Future Works

For the closed-loop neurostimulation system, following future works still need to be addressed:

- **Integration of the AFE:** Although we have a complete design for the AFE, it has not been integrated into the system yet. The integration of the AFE will help us study the performance of a complete closed-loop system.

- **Study of the Stimulation Artifact:** Although the neural signal recording and the stimulation happen sequentially in the proposed system, the stimulation artifact may still affect the performance of the system. For example, if the system is duplicated for large-scale stimulation and each channel does not perform stimulation or detection simultaneously, surrounding channels may introduce stimulation artifacts to channels that are currently recording. Therefore, it is necessary to study possible issues and methods to prevent stimulation artifacts.

- **Discrimination of Spontaneous Action Potentials:** Spontaneous action potentials deliver erroneous information to the system and make the system output wrong results. One way to solve this problem is to make the system run multiple complete operations and take the average value of all outputs.

- **AFE with Self-Calibration:** The designed AFE has accurate control over the signal amplitude. We could leverage this feature to construct an AFE that can automatically tune its gain to adapt to the threshold voltage set by the spike detector. This reduces the number of calibration steps for the system and makes the system more suitable for constructing large-scale stimulation systems.

For the linear-phase delay filter, the following future work still need to be addressed:

- **Study of the Loading Effect:** Compared with the mathematical model, the implemented filter has smaller delay bandwidth. One explanation is the loading effect of each stage of the filter. If we can use some circuit techniques to compensate or reduce this loading effect, the performance of the filter can be further improved.

Reference

[1] "Biological & Biomedical | Electrical Engineering - Princeton University." [Online]. Available: http://www.ee.princeton.edu/research/biomedical. [Accessed: 07-May-2019].

[2] X. Liu and J. V. der Spiegel, *Brain-Machine Interface: Closed-loop Bidirectional System Design*. Springer International Publishing, 2018.

[3] J. J. Vidal, "Toward Direct Brain-Computer Communication," *Annu. Rev. Biophys. Bioeng.*, vol. 2, no. 1, pp. 157–180, 1973.

[4] M. Shoaran *et al.*, "A 16-channel 1.1mm2 implantable seizure control SoC with sub-μW/channel consumption and closed-loop stimulation in 0.18μm CMOS," *2016 IEEE Symp. VLSI Circuits VLSI-Circuits*, pp. 1–2, 2016.

[5] V. W. Leung *et al.*, "A CMOS Distributed Sensor System for High-Density Wireless Neural Implants for Brain-Machine Interfaces," in *ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC)*, 2018, pp. 230–233.

[6] M. Azin, D. J. Guggenmos, S. Barbay, R. J. Nudo, and P. Mohseni, "A Battery-Powered Activity-Dependent Intracortical Microstimulation IC for Brain-Machine-Brain Interface," *IEEE J. Solid-State Circuits*, vol. 46, no. 4, pp. 731–745, Apr. 2011.

[7] X. Liu, M. Zhang, A. G. Richardson, T. H. Lucas, and J. V. der Spiegel, "Design of a Closed-Loop, Bidirectional Brain Machine Interface System With Energy Efficient Neural Feature Extraction and PID Control," *IEEE Trans. Biomed. Circuits Syst.*, vol. 11, no. 4, pp. 729–742, Aug. 2017.

[8] K. S. Guillory and R. A. Normann, "A 100-channel system for real time detection and storage of extracellular spike waveforms," *J. Neurosci. Methods*, vol. 91, no. 1–2, pp. 21–29, Sep. 1999.

[9] P. Cong, *Circuit Design Considerations for Implantable Devices*. River Publishers, 2018.

[10] H. T. Friis, "Noise Figures of Radio Receivers," *Proc. IRE*, vol. 32, no. 7, pp. 419–422, Jul. 1944.

[11] R. R. Harrison and C. Charles, "A low-power low-noise CMOS amplifier for neural recording applications," *IEEE J. Solid-State Circuits*, vol. 38, no. 6, pp. 958–965, Jun. 2003.

[12] R. H. Olsson, M. N. Gulari, and K. D. Wise, "Silicon neural recording arrays with on-chip electronics for in-vivo data acquisition," in *2nd Annual International IEEE-EMBS Special Topic Conference on Microtechnologies in Medicine and Biology. Proceedings (Cat. No.02EX578)*, 2002, pp. 237–240.

[13] F. Shahrokhi, K. Abdelhalim, D. Serletis, P. L. Carlen, and R. Genov, “The 128-Channel Fully Differential Digital Integrated Neural Recording and Stimulation Interface,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 3, pp. 149–161, Jun. 2010.

[14] “European Committee for Electrotechnical Standardization, Medical electrical equipment - Part 2–47: particular requirements for the safety, including essential performance, of ambulatory electrocardiographic systems.” [Online]. Available: https://webstore.iec.ch/publication/2666. [Accessed: 06-May-2019].

[15] K. Najafi and K. D. Wise, “An implantable multielectrode array with on-chip signal processing,” *IEEE J. Solid-State Circuits*, vol. 21, no. 6, pp. 1035–1044, Dec. 1986.

[16] “hc-1 — CRCNS.org.” [Online]. Available: https://crcns.org/data-sets/hc/hc-1. [Accessed: 25-Apr-2019].

[17] I. Obeid and P. D. Wolf, “Evaluation of spike-detection algorithms for a brain-machine interface application,” *IEEE Trans. Biomed. Eng.*, vol. 51, no. 6, pp. 905–911, Jun. 2004.

[18] G. Gagnon-Turcotte, C.-D. Camaro, and B. Gosselin, “Comparison of low-power biopotential processors for on-the-fly spike detection,” in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, 2015, pp. 802–805.

[19] K. Iniewski, *VLSI Circuits for Biomedical Applications*. Artech House, 2008.

[20] M. Bugbee, N. de N. Donaldson, A. Lickel, N. J. M. Rijkhoff, and J. Taylor, “An implant for chronic selective stimulation of nerves,” *Med. Eng. Phys.*, vol. 23, no. 1, pp. 29–36, Jan. 2001.

[21] D. R. Merrill, M. Bikson, and J. G. R. Jefferys, “Electrical stimulation of excitable tissue: design of efficacious and safe protocols,” *J. Neurosci. Methods*, vol. 141, no. 2, pp. 171–198, Feb. 2005.

[22] X. Liu, A. Demosthenous, and N. Donaldson, “An Integrated Implantable Stimulator That is Fail-Safe Without Off-Chip Blocking-Capacitors,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 2, no. 3, pp. 231–244, Sep. 2008.

[23] M. Ortmanns, A. Rocke, M. Gehrke, and H. Tiedtke, “A 232-Channel Epiretinal Stimulator ASIC,” *IEEE J. Solid-State Circuits*, vol. 42, no. 12, pp. 2946–2959, Dec. 2007.

[24] J. Sit and R. Sarpeshkar, “A Low-Power Blocking-Capacitor-Free Charge-Balanced Electrode-Stimulator Chip With Less Than 6 nA DC Error for 1-mA Full-Scale Stimulation,” *IEEE Trans. Biomed. Circuits Syst.*, vol. 1, no. 3, pp. 172–183, Sep. 2007.

[25] N. Butz, A. Taschwer, S. Nessler, Y. Manoli, and M. Kuhl, “A 22 V Compliant 56$\m u$W Twin-Track Active Charge Balancing Enabling 100% Charge Compensation Even in Monophasic and 36% Amplitude Correction in Biphasic Neural Stimulators,” *IEEE J. Solid-State Circuits*, vol. 53, no. 8, pp. 2298–2310, Aug. 2018.

[26] B. Razavi, *Design of Analog CMOS Integrated Circuits*, 1st ed. New York, NY, USA: McGraw-Hill, Inc., 2001.

[27] B. Choi, "Enhancement of current driving capability in data driver ICs for plasma display panels," *IEEE Trans. Consum. Electron.*, vol. 55, no. 3, pp. 992–997, Aug. 2009.

[28] *Design of Analog Filters 2nd Edition*, Second Edition, New to this Edition: Oxford, New York: Oxford University Press, 2009.

[29] T. Deliyannis, "High-Q factor circuit with reduced sensitivity," *Electron. Lett.*, vol. 4, no. 26, pp. 577–579, Dec. 1968.

[30] B. Gosselin, M. Sawan, and E. Kerherve, "Linear-Phase Delay Filters for Ultra-Low-Power Signal Processing in Neural Recording Implants," *IEEE Trans. Biomed. Circuits Syst.*, vol. 4, no. 3, pp. 171–180, Jun. 2010.