An 8-Channel Neural Spike Processing IC with Unsupervised Closed-Loop Control Based On Spiking Probability Estimation

Tong Wu and Zhi Yang*

***Abstract*— This paper presents a neural spike processing IC for simultaneous spike detection, alignment, and transmission on 8 recording channels with unsupervised closed-loop control. In this work, spikes are detected according to online estimated spiking probability maps, which reliably predict the possibility of spike occurrence. The closed-loop control has been made possible by estimating firing rates based on alignment results and turning on/off channels individually and automatically. The 8-channel neural spike processing IC, implemented in a 0.13 μm CMOS process, has a varied power dissipation from 36 μW to 54.4 μW per channel at a voltage supply of 1.2 V. The chip also achieves a 380$\times$ data rate reduction for the testing *in vivo* data, allowing easy integration with wireless data transmission modules.**

I. Introduction

For neural recording experiments, the idea of neural assemblies has always been closely associated with the occurrence of spike patterns in convergently-divergently connected networks, where a sufficient number of neurons are a prerequisite to establish casual connectivity. Current recording system allows simultaneous data acquisition from more than 200 channels [1], generating over 100 Mb data per second to be processed. Therefore it is important to compress the data and extract useful features for information decoding. In addition, given an efficient circuit implementation that can compress neural data, it allows data bandwidth reduction thus transceivers with affordable power budget can be integrated.

Many works have been reported on developing efficient hardware for data compression and information decoding [2], [3], [4], where the absolute value (*Abs*) and the nonlinear energy operator (*NEO*) are popular spike detectors due to their computational simplicities. An unresolved challenge of these methods is how to set the detection threshold to ensure consistent performance in the presence of many recording imperfections. To address this difficulty, we have optimized and implemented a probability-based detector in our system, which allows us to directly specify the desired precision of detection instead of thresholding on different neural codes and having no idea of real-time detection performance [5].

Low power consumption is another primary requirement for neural recording and signal processing devices. Empirical studies with *in vivo* experiments have shown that around 70% of channels contain little spikes, implying a significant margin for power reduction if "spikeless" channels can be correctly identified and shut down for a while. In neural recording experiments with high channel counts (>64), manual identifying and closing channels require extensive human interventions thus infeasible. In this work, an unsupervised closed-loop control has been on-chip implemented, which can dynamically save power dissipations by turning off digital signal processing channels selectively for a user-specified period of time.

The rest of the paper is organized as follows. Section II describes the system design. Section III presents the chip prototyping and testing. Section IV concludes the paper.

Tong Wu and Zhi Yang* are with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore. e-mail: {elewut, eleyangz}@nus.edu.sg. *Asterisk indicates corresponding author*

II. System Architecture

The system accepts 8-channel time-multiplexed raw data digitized in 16-bit at a sampling rate of 20 kHz, and outputs three serialized data streams: the band-pass filtered neural data, spiking probability maps, and aligned spike packages. Block diagram of the proposed system is shown in Fig. 1.

A. Pre-processing of Raw Data

First, raw neural data are band-pass filtered to remove low-frequency components as well as artifacts. The filter can be programmed externally to select the higher corner frequency from 5, 6, 7, and 8 kHz, while the lower corner frequency is fixed at 300 Hz. Throughout the programming range, the filter has achieved >70 dB stop-band attenuations and <0.03 dB pass-band ripples.

Band-limited neural data are then Hilbert transformed, which is realized by cascading a 16-point fast Fourier transform (FFT) and an inverse-FFT (IFFT) with an intermediate rotation of the FFT outputs. We chose to implement Hilbert transform in pipelined FFT-IFFT structure instead of time-domain convolution to facilitate multichannel hardware sharing. Simulation results confirmed that a 16-point Hilbert transform ensures a >97% precision in terms of the histogram in the next step. Finally, Hilbert transformed neural data are normalized to their estimated variances. This is to represent neural data in a more compact form to facilitate data distribution approximation.

B. EC-PC Parameter Regression

We have previously reported an unsupervised and adaptive spike detection algorithm in [5], which proves that in neural data probability distributions, *in vivo* noise forms an exponential component (EC) and extracellular spikes form a polynomial component (PC). By estimating both EC and PC, the algorithm can quantitatively predict the occurrence of spikes based on a spiking probability map.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

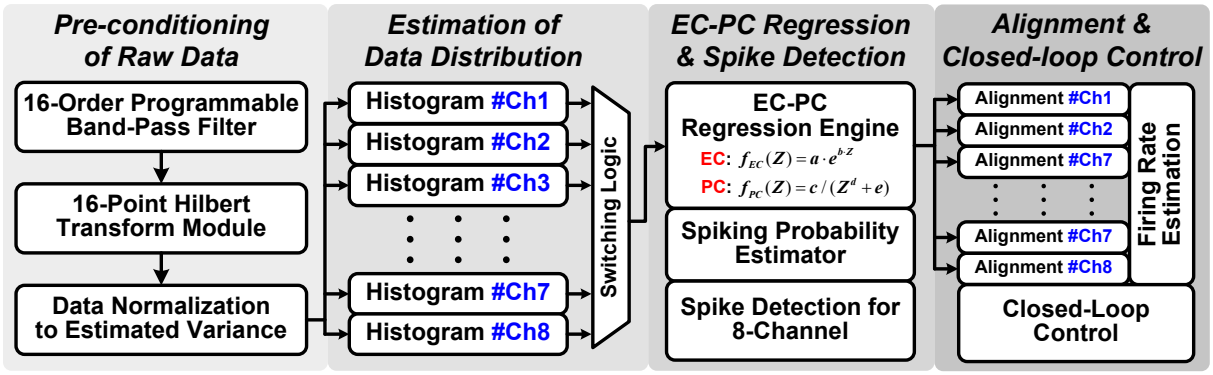

Fig. 1. Block diagram of the proposed neural spike processing system.

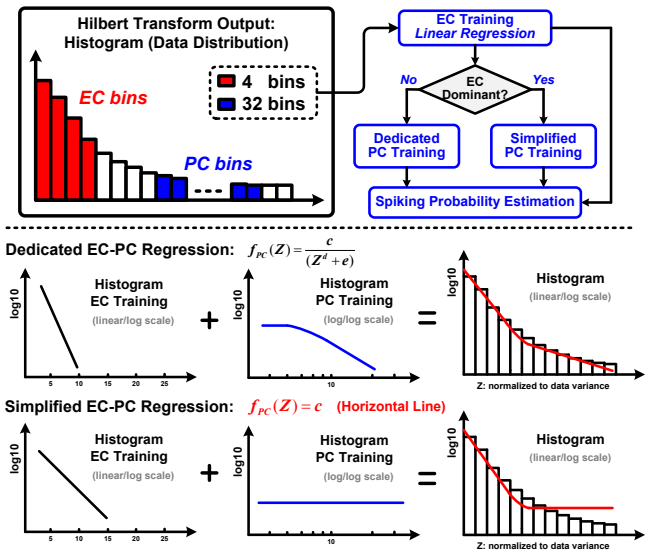

Fig. 2. Flowchart of EC-PC parameter regression engine and illustration of the regression algorithm.

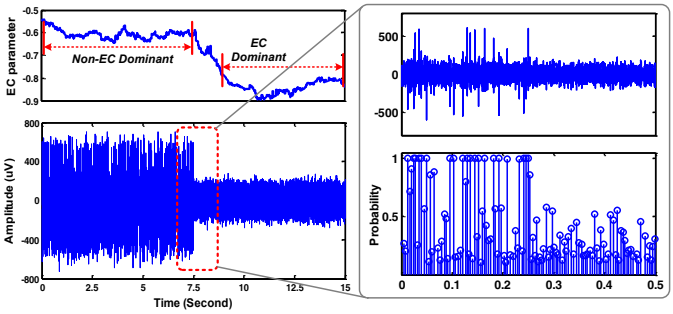

Fig. 3. Top left: trained EC parameters. Bottom left: synthesized neural data with an abrupt firing rate reduction. Right: zoom in of neural data and corresponding spiking probability map.

In this work, the distribution of each channel is approximated by a histogram, consisting of 4×14-bit bins for EC and 32×10-bit bins for PC, as shown in Fig. 2. The contents of histograms are updated adaptively as normalized neural data arrive. An EC-PC regression engine is switched to interface with 8-channel histograms sequentially for parameter estimation. Ideally, the parameters of EC and PC are trained by two regressions in the linear-log scale and log-log scale, respectively, where the log-log scale training consumes excessive circuit power due to the logarithmic arithmetics on the 32 PC bins. In this implementation, we have improved the regression scheme by introducing a simplified model which trains PC as a horizontal line in the linear-log scale, as shown in Fig. 2. Simulation results showed that the simplified model can save 46% circuit area and 25% power consumption compared with the dedicated one.

A finite state machine (FSM) has been implemented to automatically switch the regression engine between the dedicated and the simplified PC models based on the observation that inactive neurons fire only 1-10 spikes per second [6], resulting in small PC and making EC dominant. In Fig. 3, the trained EC parameter served as a reliable indicator of the spiking activity by tracking its variations. As the last step in the FSM, a spiking probability map is generated from the trained EC and PC, which represents the chance of the spike presence at a given time slot. For example, a 50% probability score indicates a 50% possibility of a data point being part of a spike, as illustrated in the zoom-in region of Fig. 3.

C. Spike Detection and Alignment

We propose to combine the EC-PC idea with existing detectors, e.g., *Abs*, where spiking probability maps can be directly thresholded to identify spikes. The operations of the spike detection and alignment in our implementation are illustrated in Fig. 4, and summarized as follows.

1) Pre-load: A buffer with a capacity of 20 data samples buffers the neural data stream. Once full, the buffer will be updated adaptively until a probability score exceeds the user-specified threshold.

2) Post-load: The *Pre-load* buffer stops working, and another buffer is initiated to hold the next 50 data samples. During the *Post-load* phase, the peak with the largest absolute value in the 50 samples is determined.

3) Comp-load (optional): The peak obtained in *Post-load* is supposed to be the centre of a 40-point spike segment. If samples on the right of the peak in the *Post-load* buffer are less than 20, the *Post-load* buffer will left-shift to fill in new data and compensate for the centre point deviation.

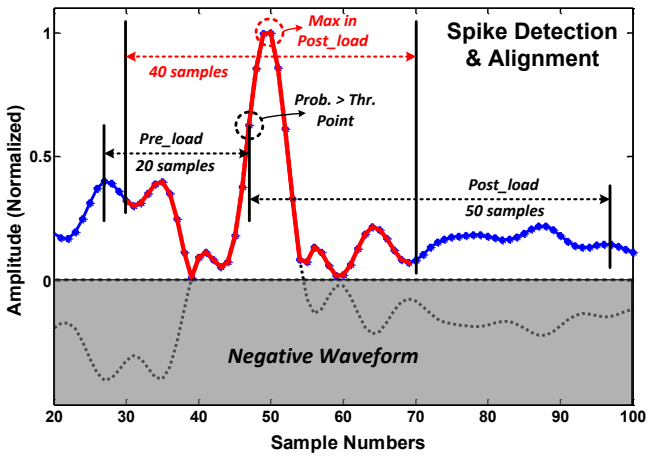

Fig. 4. Illustration of the spike detection and alignment scheme.

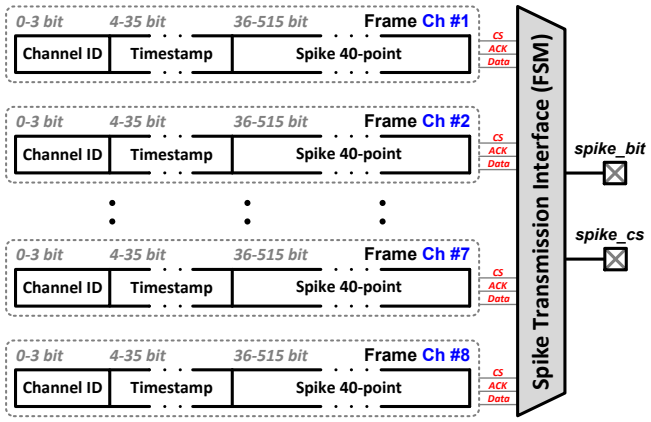

Fig. 5. Format of spike packages and configuration of 8-channel spike transmission module. Each spike segment consist of channel ID, timestamp, and 40-point spike waveform, requiring 516 bits in total.

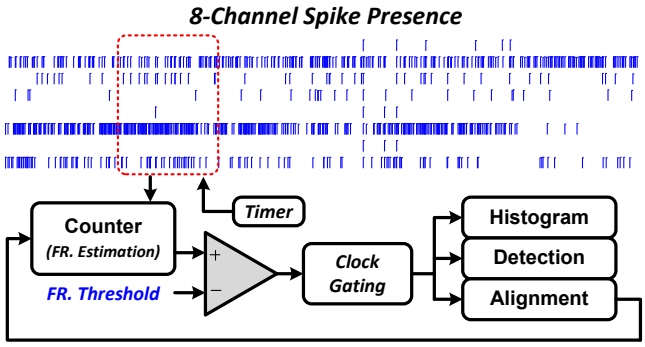

Fig. 6. Structure of single channel closed-loop control.

Whole System	
Technology	CMOS .13 µm Process
Area	0.76 x 5.28 mm
Reference Clock	5.12 MHz
Gate Count	269 K
Core Voltage	1.2 V
Num. of Channels	8
Power Density	108.45 µW/mm^2
Pre-processing (per channel)	
Power	36 µW
Area	0.19 mm^2
Detection (per channel)	
Power	11 µW
Area	0.20 mm^2
Buffer Size	376 bit
Alignment (per channel)	
Power	7.4 µW
Area	0.10 mm^2
Buffer Size	996 bit

Fig. 7. Die photo of designed chip and performance summary.

Due to the compulsory filling of the 70-point spike buffers in total during alignment, the system has a built-in refractory time of 3.5 ms. The *Post-load* buffer can hold 2.5 ms neural data, which to a large extent avoids misalignment to local maximums instead of real peaks. After alignment, spikes from 8 channels are packaged in frames along with channel IDs and timestamps, and transmitted off-chip through a pair of bit streams, as shown in Fig. 5. We assume that the 8-channel system can record activities of 20 neurons with an average firing rate of 20 Hz, corresponding to a net data rate of 20×20×516 = 2.06 Mbps. The spike transmission interface has been implemented as a FSM with intermediate buffers and achieved a highest throughput of 5 Mbps, satisfying the bandwidth requirement.

D. Firing Rate Estimation and Closed-Loop Control

In this work, the closed-loop control structures enable automatic and selective activation/deactivation of signal processing channels to save power consumption. As shown in Fig. 6, we have used firing rate as the criterion for closing channels, which is estimated by counting the number of spikes appearing in a time window. Both the firing rate threshold and the length of the time window are programmable. When channels are labeled as "spikeless", their histograms, detection, and alignment blocks are shut down for a user-specified period of time through clock gating. The pre-processing blocks (band-pass filtering and Hilbert transform) are not influenced by the closed-loop control due to their time-multiplexed hardware sharing configuration.

III. Prototyping and Measurements

A. Chip Implementation

The proposed system has been fabricated in a CMOS 0.13 μm process and occupies 0.76 × 5.28 mm^2 as shown in Fig. 7, where measured circuit performances are also given. The power consumption of single channel varies from 36 μW when being deactivated and only the pre-processing is working, to a peak 54.4 μW when the functionalities are fully activated, allowing a dynamic and adaptive power saving. The peak power density is 108.45 μW/mm^2, well below the 277 μW/mm^2 required for neural implants [7].

B. Benchtop Testing

We have conducted benchtop testings of the proposed chip on a customized prototyping board with neural data recorded

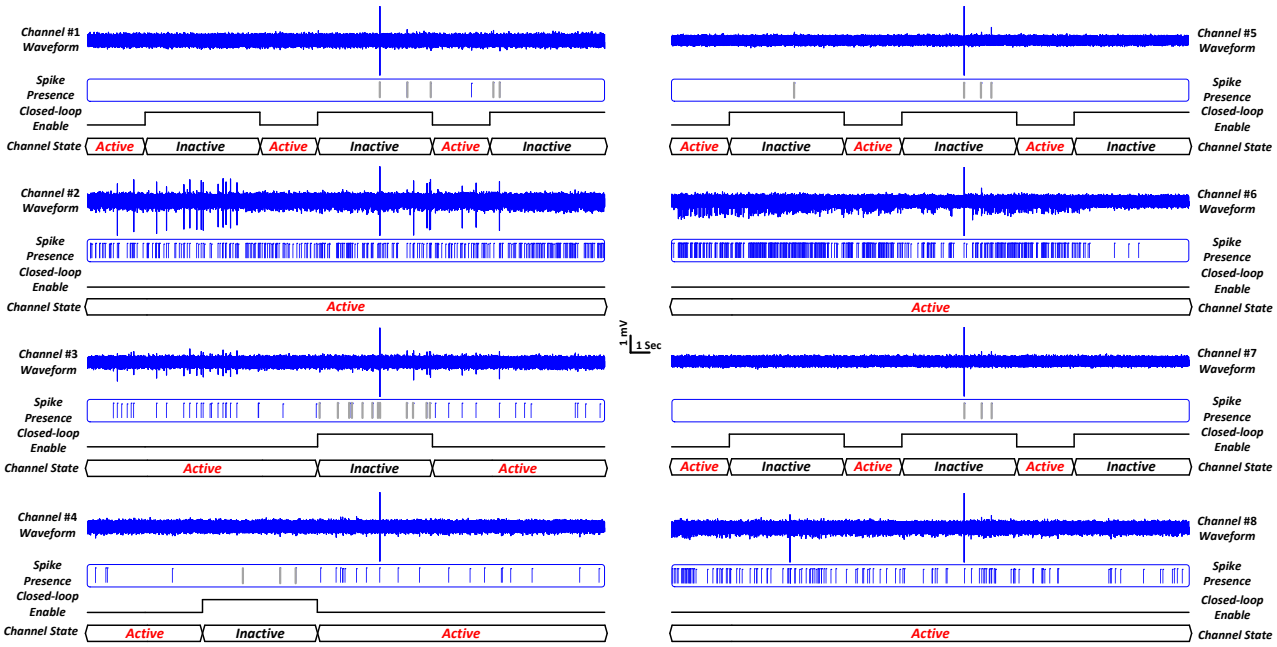

Fig. 8. 8-channel testing results with enabled closed loop control. In each channel, the waveform of neural data (60 sec), the estimated spike presence, the closed-loop enable signal, and the channel state are displayed from top to bottom. Spikes missed due to closed-loop control are in gray.

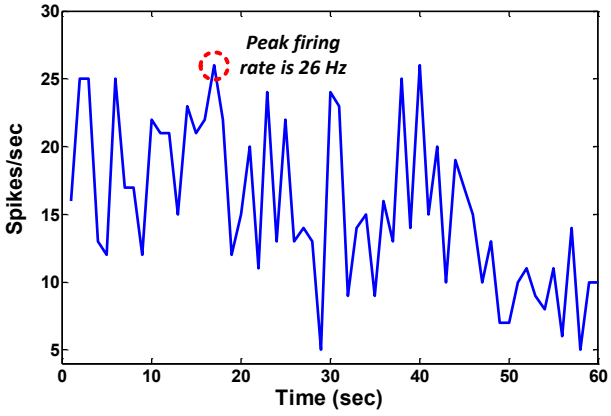

Fig. 9. Estimated firing rate for the simulated dataset in Fig. 8. The peak is 26 Hz.

from *in vivo* experiments, as shown in Fig. 8. During the testing, the closed-loop control has been enabled with a 6.7-sec firing rate estimation window and a 13.3-sec resting period, which are indicated as *Active* and *Inactive* in the figure. The firing rate threshold to turn off channels is set at 1 Hz. The result has clearly demonstrated the chip's ability of performing multichannel closed-loop control with single-channel resolution.

The chip has achieved a 380× data rate reduction for the given dataset, from 16-b×40 kHz×8-ch = 5.12 Mbps to 26 spike/s×516-b/spike = 13.416 kbps, as illustrated in Fig. 9. With the closed-loop control enabled, the data rate reduction can be further improved at acceptable information loss.

IV. Conclusion

We have reported a 8-channel neural spike processing IC to perform simultaneous spike detection, alignment, and transmission, featuring an online and unsupervised closed-loop operation. By adapting to neuronal firing activities and reallocating computational resources efficiently, the proposed system can find wide applications in robust neural recording experiments where power consumption is a major concern. The provided substantial bandwidth reduction is also favorable for wireless biomedical signal processing applications.

Acknowledgement

This work is supported by Singapore A*STAR and MOE grants R-263-000-699-305, R-263-000-A32-305, R-263-000-A29-133, and R-263-000-619-133.

References

[1] J. N. Y. Aziz, K. Abdelhalim, R. Shulyzki, R. Genov, B. L. Bardakjian, M. Derchansky, D. Serletis, and P. L. Carlen, "256-Channel Neural Recording and Delta Compression Microsystem With 3D Electrodes," *IEEE J. Solid-State Circuits*, vol. 44, no. 3, pp. 995–1005, Mar 2009.

[2] T. Chen, K. Chen, Z. Yang, K. Cockerham, and W. Liu, "A Biomedical Multiprocessor SoC for Closed-Loop Neuroprosthetic Applications," *Dig. Tech. Papers IEEE Int. Solid-State Circuits Conf.*, pp. 434–435, 2009.

[3] T.-T. Liu and J. M. Rabaey, "A 0.25 V 460 nW Asynchronous Neural Signal Processor With Inherent Leakage Suppression," *IEEE J. Solid-State Circuits*, vol. 48, no. 4, pp. 897–906, Apr 2013.

[4] V. Karkare, S. Gibson, and D. Marković, "A 75-μW, 16-Channel Neural Spike-Sorting Processor With Unsupervised Clustering," *IEEE J. Solid-State Circuits*, vol. 48, no. 9, pp. 2230–2238, Sep 2013.

[5] Z. Yang, W. Liu, M. R. Keshtkaran, Y. Zhou, J. Xu, V. Pikov, C. Guan, and Y. Lian, "A New EC-PC Threshold Estimation Method for in-*vivo* Neural Spike Detection," *J. Neural Eng.*, vol. 9, no. 4, 2012.

[6] R. R. Harrison, P. T. Watkins, R. J. Kier, R. O. Lovejoy, D. J. Black, B. Greger, and F. Solzbacher, "A Low-Power Integrated Circuit for a Wireless 100-Electrode Neural Recording System," *IEEE J. Solid-State Circuits*, vol. 42, pp. 123–133, 2007.

[7] S. Kim, P. Tathireddy, R. A. Normann, and F. Solzbacher, "Thermal Impact of an Active 3-D Microelectrode Array Implanted in the Brain," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 15, no. 4, pp. 493–501, 2007.