Real-Time Neuromorphic System for Large-Scale Conductance-Based Spiking Neural Networks

Shuangming Yang, Jiang Wang, Bin Deng, Chen Liu, *Member, IEEE*,
Huiyan Li, Chris Fietkiewicz, and Kenneth A. Loparo, *Life Fellow, IEEE*

***Abstract*—The investigation of the human intelligence, cognitive systems and functional complexity of human brain is significantly facilitated by high-performance computational platforms. In this paper, we present a real-time digital neuromorphic system for the simulation of large-scale conductance-based spiking neural networks (LaCSNN), which has the advantages of both high biological realism and large network scale. Using this system, a detailed large-scale cortico-basal ganglia-thalamocortical loop is simulated using a scalable 3-D network-on-chip (NoC) topology with six Altera Stratix III field-programmable gate arrays simulate 1 million neurons. Novel router architecture is presented to deal with the communication of multiple data flows in the multinuclei neural network, which has not been solved in previous NoC studies. At the single neuron level, cost-efficient conductance-based neuron models are proposed, resulting in the average utilization of 95% less memory resources and 100% less DSP resources for multiplier-less realization, which is the foundation of the large-scale realization. An analysis of the modified models is conducted, including investigation of bifurcation behaviors and ionic dynamics, demonstrating the required range of dynamics with a more reduced resource cost. The proposed LaCSNN system is shown to outperform the alternative state-of-the-art approaches previously used to implement the large-scale spiking neural network, and enables a broad range of potential applications due to its real-time computational power.**

Manuscript received September 25, 2017; revised January 26, 2018 and March 22, 2018; accepted April 3, 2018. Date of publication April 19, 2018; date of current version April 17, 2019. This work was supported in part by the National Natural Science Foundation of China under Grant 61701336, in part by the Natural Science Foundation of Tianjin, China under Grant 17JCQNJC00800, in part by the Foundation of Tianjin University under Grant 2017XZC-0053, and in part by the Hong Kong Scholars Programs under Grant XJ2016006. This paper was recommended by Associate Editor P. P. Angelov. *(Corresponding author: Chen Liu.)*

S. Yang, J. Wang, and B. Deng are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yangshuangming@tju.edu.cn; jiangwang@tju.edu.cn; dengbin@tju.edu.cn).

C. Liu is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China, and also with the Department of Physics, Centre for Nonlinear Studies, Institute of Computational and Theoretical Studies, Hong Kong Baptist University, Hong Kong 999077 (e-mail: liuchen715@tju.edu.cn).

H. Li is with the School of Electrical and Information Engineering, Tianjin University of Technology and Educations, Tianjin 300222, China (e-mail: lhy2740@126.com).

C. Fietkiewicz and K. A. Loparo are with the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: cxf47@case.edu; kenneth.loparo@case.edu).

This paper has supplementary downloadable multimedia material available at http://ieeexplore.ieee.org, provided by the authors.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2823730

***Index Terms*—Brain model, field-programmable gate array (FPGA), large-scale spiking neural network (SNN), modified conductance-based neuron models, real-time system, scalable neuromorphic implementation.**

I. Introduction

In recent years, there have been significant attention given to simulating systems of neural networks in order to explore deeply the mechanisms for coding and transmission of neuronal information [1]–[4]. However, this has posed great challenges on both the scalability that supports large-scale neural networks with random connectivity and the capability to simulate complex dynamics of neuron models based on biophysiological phenomenon.

In a biological neuron, physical membrane channels control the flow of ions across the membrane by opening and closing in response to voltage changes due to intrinsic currents and externally applied signals [5]. The dynamics of ion channel behavior are highly nonlinear and stochastic, and they are often modeled using the Hodgkin–Huxley conductance-based neuron model [6], [7]. This model is regarded to be biologically realistic and describes essential dynamical characteristics of ion channels by nonlinear differential equations [8]. The challenge that arises is the implementation of the computationally intensive functions in an efficient realization of a large-scale brain network.

The last decade has seen great progress in understanding brain dynamics, but the processing mechanisms underlying the dynamics of large-scale networks remain poorly understood [9], [10]. One example is the cortico-basal ganglia-thalamocortical loop, which includes several interconnected subcortical nuclei [13]. Several pathological conditions, including Parkinson's and Huntington's diseases, have been related to the cortico-basal ganglia-thalamocortical network. The network functionality includes action selection, reinforcement learning, and dimensionality reduction in both motor and cognitive fields by the extensive interconnections or reciprocal projections to the brain stem [11]–[14]. Previous studies have also revealed that this loop plays a significant role in decision-making [15], [16].

Our approach toward the investigation of the collective behavior of the brain is to simulate large-scale neural networks. For the basal ganglia system, several studies have been proposed as shown in Fig. 1 [17]–[26]. Most models of the basal ganglia networks have used the leaky integrate-and-fire (LIF) neuron model. Previously we proposed a cost-efficient field-programmable gate array (FPGA) implementation of the basal ganglia network [25], however, it is not scalable. Izhikevich and Edelman [23] implemented a thalamocortical system to more deeply understand brain dynamics, but used a reduced representation of ionic conductance dynamics. Reduced spiking neuron models such as the LIF and Izhikevich models do not accommodate a full range of ionic conductance and therefore are not appropriate for studies of specific ionic current types. A graphics processing unit (GPU)-based system can provide for an efficient simulation of the basal ganglia network with conductance-based models, but its ability to scale is still limited [26]. Considering the computational speed, we distinguish real-time designs from the nonreal-time designs by solid and dotted boxes, respectively, in Fig. 1. Existing real-time designs either lack network biological accuracy or network scaling [21], [25], [26], while other works simulating the basal ganglia cannot obtain the real-time computation capacity [17], [20], [22], [24]. However, real-time interactions between an organism and a real environment are important in understanding neuronal dynamics [27]. Therefore, it is useful to build a real-time system of SNNs.

This paper focuses on the scalable real-time implementation of a large-scale cortico-basal ganglia-thalamocortical network, combining the advantages of high biological accuracy and large network scale (see Fig. 1). Our key motivation is to build a real-time system that can implement large-scale neural networks with cellular-level details. From a neuroscience view, this system should reproduce both the cellular and network levels of dynamical activities. From an engineering view, it poses the challenge of simulating specific ionic conductance in a large-scale network. To address this challenge, we built large-scale conductance-based spiking neural (LaCSNN), a reconfigurable, real-time system that is efficient, scalable, and flexible. At the cellular level, a set of piecewise linear approximation (PLA)-based biologically realistic neuron models are presented as the foundation, and digital multiplier-less and PLA techniques are presented for efficient realization. At the network level, a novel router is proposed for the routing of data flows with information from different nuclei, and an address event routing (AER) infrastructure is presented for the multichip 3-D scalable network-on-chip (NoC) topology. Thus, the first advantage of the LaCSNN system is the scalable implementation of the large-scale neural network with ionic channel dynamics, which bridges the gap between the cellular level and the network level of brain. The second advantage of the LaCSNN system is its real-time computational power which enables several applications, including brain–machine interfaces, neuro-robotic control, and robotic decision-making. The third advantage is its reconfigurability that makes the scalable LaCSNN system extensible for considering other nuclei that may interact with the cortico-basal ganglia-thalamocortical loop, the implementation of other brain regions, or even the realization of the full mammalian brain.

The remainder of this paper is organized as follows. Section II gives a general overview of the presented system and PLA models for the cortico-basal ganglia-thalamocortical loop. In Section III, a detailed hardware implementation of the LaCSNN system is presented, and a set of designs are proposed to address the challenges for a large-scale SNN. Experimental results are presented in Section IV, which includes dynamics invalidation of the presented PLA models, the hardware performance analysis and the precision analysis. Section V discusses the application of the LaCSNN system, makes a comparison with state-of-the-art techniques, and states the limitations and future work. Finally, this paper is concluded in Section VI.

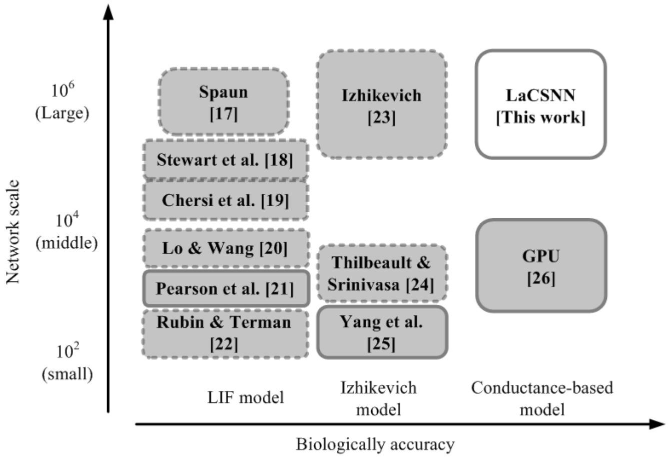
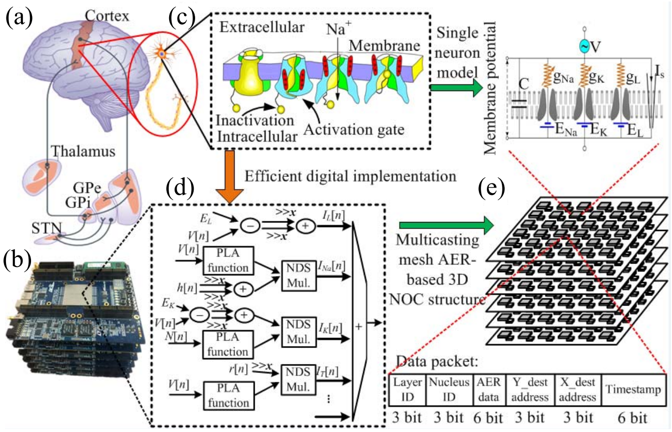


Fig. 1. Comparison between several recent representative implementations of the basal ganglia system and the proposed LaCSNN system. The comparison is divided into two aspects: network scale (or number of neurons and synapses) and biological accuracy (or neuron complexity). The dotted boxes represent the nonreal-time implementations, and the real-time works are highlighted by solid boxes.



Layer ID	Nucleus ID	AER data	Y_dest address	X_dest address	Timestamp
3 bit	3 bit	6 bit	3 bit	3 bit	6 bit

Fig. 2. LaCSNN system for the scalable implementation of the real-time large-scale biologically realistic cortico-basal ganglia-thalamocortical network. (a) Cortico-basal ganglia-thalamocortical loop in the human brain. (b) LaCSNN system implemented by six DE3 340 development boards. (c) Ionic channels across the neuronal membrane. (d) Digital multiplier-less and memory-less implementation of the ionic currents. (e) Multicasting mesh AER-based 3-D NoC structure for large scale SNN with multiple nuclei.

II. Design of the LaCSNN System and Description of the PLA-Based Models

A. General Framework of the LaCSNN System

The LaCSNN system uses high-end Altera Stratix III FPGAs to establish a real-time simulation platform of the

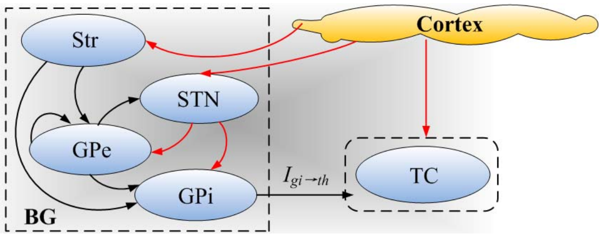


Fig. 3. Network structure of the cortico-basal ganglia-thalamocortical network.

large-scale neural network as shown in Fig. 2(a) and (b). In order to build an SNN, we start with the cortico-basal ganglia-thalamocortical network for its vital roles in the human brain and the validation of the computation performance of the LaCSNN system. As shown in Fig. 2(c) the implemented neural network is composed of biologically realistic conductance-based models to reproduce the biological dynamics of ion channels that are known to play critical roles in neuronal activities. The ionic channels are implemented digitally using multiplier-less and PLA methods as shown in Fig. 2(d), and the neuron model is then established based on the pipeline technique for enhanced operation frequency. As shown in Fig. 2(e), in order to realize the scalable of the *in silico* network, the 3-D multicasting mesh AER-based topology is mapped into the presented LaCSNN system, and a 24-bit data packet is used for both the intrachip and interchip data communication. Each data packet contains a 3-bit layer ID, a 3-bit nucleus ID, a 6-bit AER data, a 3-bit Y_dest address, a 3-bit X_dest address, and a 6-bit timestamp. The realization framework of the LaCSNN system can be applied in a wide range of event-driven, large-scale implementations of neural models, especially biologically realistic models, and even artificial neural networks for applications, such as pattern recognition and classification [36]–[39].

B. Model of the Cortico-Basal Ganglia-Thalamocortical Network

The major regions and connectivity of the cortico-basal ganglia-thalamocortical "motor" loop are well known, as shown in Fig. 3 [28], [29]. Input stimulation and output response data are required before implementing the input–output model. In this paper, the cortico-basal ganglia-thalamocortical loop model contains the following, as described in Rubin and Terman model [30]–[32]: subthalamic nucleus (STN), external segment of the globus pallidus (GPe), and internal segment of the GPe (GPi). Inhibitory synapses from the primary output nucleus of the BG system to the thalamus have strong effects on the dynamics of the thalamocortical (TC) cells. The thalamus receives synaptic inhibition from the GPi and the excitatory sensorimotor input current. GPi and GPe both receive excitatory input from the STN, and the GPe receives an applied current that represents input from the striatum. Additionally, there is interpallidal inhibition among the GPe neurons, the GPi receives inhibition from the GPe nucleus, and the STN nucleus receives inhibition from the GPe. In addition, the STN can receives a periodic stimulus that represents deep brain stimulation (DBS) used as a treatment for Parkinson's disease. Each nucleus is implemented using Hodgkin–Huxley type spiking neurons, and the membrane potentials of this type of neuron model are dominated by the ionic channels.

In this paper, we use the conductance-based neuron model and high accurate biological synapses, because it may contribute to a further investigation to reveal the potential mechanisms underlying the ionic dynamics and synaptic dynamics. The equations for each neuron model are described in the following form:

$$\begin{cases} C_m \frac{dV_{\mathrm{GPe}}}{dt} &= I_{\mathrm{app\_GPe}} - I_L - I_K - I_{Na} - I_T - I_{Ca} \\ &\quad - I_{\mathrm{GPe}\rightarrow\mathrm{GPe}} + I_{\mathrm{STN}\rightarrow\mathrm{GPe}} - I_{AHP} \\ C_m \frac{dV_{\mathrm{STN}}}{dt} &= I_{\mathrm{app\_STN}} - I_L - I_K - I_{Na} - I_T - I_{Ca} \\ &\quad - I_{\mathrm{GPe}\rightarrow\mathrm{STN}} - I_{AHP} \\ C_m \frac{dV_{\mathrm{GPi}}}{dt} &= I_{\mathrm{app\_GPi}} - I_L - I_K - I_{Na} - I_T - I_{Ca} \\ &\quad - I_{\mathrm{GPe}\rightarrow\mathrm{GPi}} + I_{\mathrm{STN}\rightarrow\mathrm{GPi}} - I_{AHP} \\ C_m \frac{dV_{\mathrm{TC}}}{dt} &= I_{SM} - I_L - I_K - I_{Na} - I_T \\ &\quad - I_{\mathrm{GPi}\rightarrow\mathrm{TC}} \end{cases} \tag{1}$$

where $C_m = 1\ \mathrm{pF}/\mu\mathrm{m}^2$ is the membrane capacitance, and $V_i (i \in \{\mathrm{STN, GPe, GPi, TC}\})$ represents the membrane potential of the STN, GPe, GPi, and TC neurons, respectively. The ionic currents $I_L, I_K, I_{Na}, I_T, I_{Ca}$, and I_{AHP} are the leak current, the potassium current, the sodium current, the low-threshold T-type calcium current, the high-threshold calcium current, and the after hyperpolarization potassium current, respectively. The detailed model equations and the corresponding parameter values are based on previous studies [30], [31]. Additionally, $I_{a\rightarrow b}$ $(a, b \in \{\mathrm{STN, GPe, GPi, TC}\})$ is the synaptic coupling from presynaptic cell a to postsynaptic cell b. The constant bias currents, represented by I_{app_i} $(i \in \{\mathrm{STN, GPe, GPi, TC}\})$, can be regarded as the net synaptic current projected into these neurons from other brain regions, including striatum and cortex. The cortex input I_{SM} from the sensorimotor region is modeled by a series of monophasic pulses with amplitude $3.5\ \mathrm{pA}/\mu\mathrm{m}^2$ and duration 5 ms. The pulse frequency follows a gamma distribution with an average rate of 14 Hz and a variance of 0.2, which mimics the irregular properties of incoming currents from the sensorimotor cortex to the TC cells [32].

Recent experimental studies have revealed that strong synaptic plasticity exists among the excitatory connections of the STN nucleus. Additionally, model-based computational experiments have also confirmed that both the pathway imbalances and motor impairments can be explained by the dysfunctional synaptic plasticity. Lourens *et al.* [33] have used a biologically realistic STN-GPe network model with synaptic plasticity. They demonstrated that synaptic plasticity can facilitate the training of the network to fire in a less synchronized form as a short-duration desynchronization modulation is applied for a sufficiently long time and with sufficiently high amplitude. The network model in this paper uses synaptic plasticity according to the following revised form of synaptic currents:

$$I_{i\rightarrow j} = g_{i\rightarrow j}(V_j - E_{i\rightarrow j}) \frac{1}{N} \sum_k S_{ij} s_i^k \tag{2}$$

where S_{ij} is the coefficient indicating the synaptic weight from the ith neuron to the jth neuron. The coefficient $g_{i\to j}$ is the maximum synaptic conductance, and $E_{i\to j}$ is the synaptic reversal potential. Each synaptic variable s_i^k is determined by a first order differential equation of the form

$$\frac{ds_i^k}{dt} = A_i\left(1 - s_i^k\right)H_\infty\left(V_i^k - \theta_i\right) - B_i s_i^k \tag{3}$$

where H_∞ is a smooth approximation of the Heaviside step function. The coefficients A_i and B_i control the synaptic time courses. The changes of the coupling coefficients S_{ij} are controlled by the following symmetric and smooth functions:

$$\begin{cases} S_{ij}(t_{n+1}) = S_{ij}(t_n) + \delta \cdot \Delta S_{ij} \\ \Delta S_{ij} = a_P e^{-|ISI_{ij}|/b_P} - a_D e^{-|ISI_{ij}|/b_D} \end{cases} \tag{4}$$

where $a_P = 0.038$, $a_D = 0.02$, $b_P = 10$ ms, $b_D = 25$ ms, and $\delta = 0.004$ are constants given by the previous work [34]. ISI_{ij} is the interspike interval of different cells in the coupled neuronal population.

C. Modified PLA-Based Neuron Models

Although previous studies have proposed conductance-based basal ganglia network models, they are unsuitable for real-time, large-scale implementation due to their considerable computational complexity [30]–[32]. A general hardware-based method to implement the conductance-based neuron model is based on the look-up table (LUT). However, for an LUT with a 10-bit address and 30-bit data, the resource cost is up to 30 720 bits, which means the LUT-based approach will require significant on-chip memory resources. Therefore, the number of LUTs has been reduced in this paper. In order to improve computational efficiency without an increase of the implementation cost for the network model, a modified network model is presented based on the PLA approach.

In this method, the nonlinear functions of the ionic current models are replaced by the PLA functions, which are described as follows:

$$f_{PLA}(V) = \begin{cases} k_1 V + b_1 \text{ when } V < \frac{b_2 - b_1}{k_1 - k_2} \\ k_2 V + b_2 \text{ when } \frac{b_2 - b_1}{k_1 - k_2} \le V < \frac{b_3 - b_2}{k_2 - k_3} \\ \dots \\ k_n V + b_n \text{ when } V \ge \frac{b_n - b_{n-1}}{k_{n-1} - k_n} \end{cases} \tag{5}$$

where k_i and b_i are the slope and intercept of the piecewise linear functions separately ($i = 1, 2, \ldots, n$), which can be implemented using shift and addition/subtraction operations. The intersection of the adjacent two lines is denoted by the coefficients. Three factors needs to be considered in this method, which are: 1) high fitting degree of the original function; 2) high fitting degree of the ionic currents; and 3) accurate neural dynamics of the modified model. A cost function for the evaluation of the approximating error is defined as

$$\mathrm{CF_{RE}} = \frac{1}{n}\sqrt{\sum_{i=1}^{n} \frac{(f_{\mathrm{ori}}(i) - f_{\mathrm{mod}}(i))^2}{f_{\mathrm{ori}}(i)^2}} \tag{6}$$

where n is the total sampling points. An exhaustive search algorithm is used in the proposed approach to get a set of coefficients in the modified functions. In this procedure, first, a few points on the nonlinear curves are chosen based on the number of segments to obtain intervals. In each interval, the cost function $\mathrm{CF_{RE}}$ is used to fit linear segment on the original nonlinear curve. In this method, the initial slopes and intercepts can be obtained and then the range of each slope and intercept can be defined. For instance, the range of k_1 in the first line segment of $f_1(V)$ was 0~10 k_1^0 by step size of 0.1 k_1^0. The range of b_1 in the first line segment of $f_1(V)$ was 0~10 b_1^0 by step size of 0.1 b_1^0. The parameters k_1^0 and b_1^0 are the initial slope and intercept obtained by the presented cost function. Then the cost function $\mathrm{CF_{RE}}$ is calculated between PLA and original nonlinear function. If the value of $\mathrm{CF_{RE}}$ is more than 10^{-5}, then the values of slopes and intercepts will be changed and the intersections of adjacent lines are consequently updated. Finally, the search procedure is terminated when the cost function $\mathrm{CF_{RE}}$ is less than 10^{-5}.

The value ranges of the segment points are limited so that the amount of computations can be reduced. If the modification of the nonlinear function cannot meet the standards aforementioned, the segment number of the piecewise linear function will be added until the three criteria can be guaranteed. The PLA functions are summarized in Table S1 in the supplementary material and the detailed parameter values of each PLA function are listed in Tables S2–S4 in the supplementary material.

III. Hardware Implementation of the LaCSNN System

A. Digital Implementation of the PLA Neuron Model

In the proposed design of PLA neuron models, including the modified GPe, STN, GPi, and TC neuron models, the Euler method of numerical integration is used for the digital implementation, which can obtain both high precision and low resource cost. The hardware topology for the GPe neuron model is presented in Fig. 4(a), which has seven pipelines for the seven variables in the model equations. The "V" pipeline implements the pipeline of variable V, and the number of stages required in the digital processing is V_{stage}. The pipelines of variables "h," "n," "r," "Ca," "s_{GPe}," and "$s_{\mathrm{GPe\to GPi}}$" are implemented by the corresponding pipeline modules in n_{stage}, h_{stage}, r_{stage}, Ca_{stage}, s_{GPestage}, and $s_{\mathrm{GPe\to GPistage}}$ stages, respectively. The latency numbers of the pipelines are V_{delay}, n_{delay}, h_{delay}, r_{delay}, Ca_{delay}, s_{GPedelay}, and $s_{\mathrm{GPe\to GPidelay}}$. Accordingly, the conditions must be satisfied in the following form:

$$\begin{cases} V_{\mathrm{delay}} = V_{\mathrm{stage}} \\ V_{\mathrm{delay}} = n_{\mathrm{delay}} = h_{\mathrm{delay}} = r_{\mathrm{delay}} = Ca_{\mathrm{delay}} = s_{\mathrm{GPedelay}} \\ \qquad = s_{\mathrm{GPe\to GPidelay}} \\ V_{\mathrm{stage}} = n_{\mathrm{stage}} = h_{\mathrm{stage}} = r_{\mathrm{stage}} = Ca_{\mathrm{stage}} \\ \qquad = s_{\mathrm{GPestage}} = s_{\mathrm{GPe\to GPistage}} \end{cases} \tag{7}$$

which guarantees the synchronization of the digital pipelines.

The detailed digital topology of the V pipeline module is depicted in Fig. 4(b), and the other pipelines have similar digital structures with the V pipeline. The "ADD" and

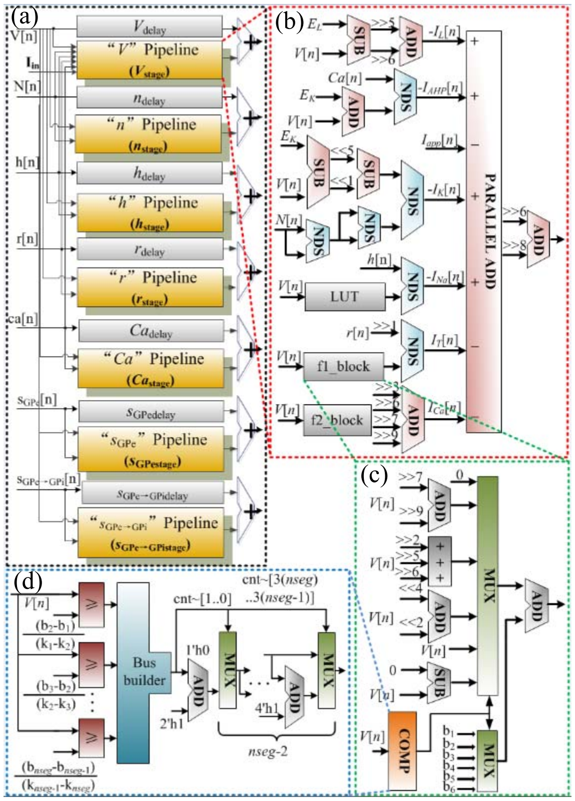


Fig. 4. Neural processor structure and data path of the GPe neuron model. (a) General overview of the pipelining structure. The V pipeline, n pipeline, h pipeline, r pipeline, Ca pipeline, s_{GPe} pipeline, and $s_{\text{GPe}\rightarrow\text{GPi}}$ pipeline represents the six pipelines for computing the seven variables in the GPe neuron model. (b) Digital architecture of the V pipeline. (c) Digital architecture of the "$f_1(V)$" function. (d) Detailed implementation of the "COMP" block in the digital architecture of $f_1(V)$ function.

"SUB" blocks realize the addition and subtraction operations, respectively. The neuronal dedicated shifting (NDS) block is a dedicated digital circuit for the multiplication operation without DSP resources and is referred to as an NDS multiplier. It should be noted that the nonlinear function in the sodium ionic channel in the GPe neuron model cannot be fitted by sixth-order piecewise approximation or less while maintaining its accurate dynamical characteristics. PLA methods with higher orders will induce undesirable amount of logic resources. Thus, the nonlinear function in the sodium ionic channel is realized by an LUT. The PLA functions $f_i(V)$ have been implemented in the corresponding "fi_block." The detailed hardware structure of the PLA function $f_1(V)$ is shown in Fig. 4(c), which has digital structures similar to other PLA functions. A dedicated circuit for the comparison operation in the digital circuit of PLA functions is also implemented by the COMP block. The digital topology of the COMP block is shown in Fig. 4(d). A bus builder is used to construct the output from inputs with a single bit. The number "*nseg*" is the segment number of the piecewise linear functions, which is based on the order number of the PLA functions. The implementation methods of other neurons, including STN, GPi, and TC neurons, are the same with the proposed design for the GPe neuron in Fig. 4. To reduce the computational cost of multiplier resources, the NDS multiplier is used in the network implementation, including both neurons and synapses.

B. Digital Architecture of the Proposed 3-D Multicasting Mesh AER Topology

The digital topology of the NoC architecture is critical and determines the simulation performance of the proposed system. In this paper, a layered approach is used to divide the 3-D NoC architecture into two parts, which includes the vertical crossbar and the horizontal network layers (i.e., 2-D NoC). This structure avoids the closure of the nodes on the edges and vertexes, and combines the NoC structure with the benefits of 3-D integration, has remarkable abilities to improve the system performance. Each horizontal layer of neural networks is implemented on an FPGA, and the vertical crossbar is implemented by the high speed Terasic connector (HSTC) interface that is equipped on a DE3 development board for high-speed interconnection and configurable I/O standards. The reduction of the hops between nodes can effectively enhance throughput and decrease the latency which are two basic and crucial factors of the system performance. In this paper, a 6 × 6 × 6 structure is presented, which focuses on the high-performance implementation of a spiking neural network (SNN).

As shown in Fig. 5(a), four HSTC connectors are used in the interchip data communication. Each HSTC connector can transmit 120-bit data at each clock cycle and is responsible for the data transmission from five nucleus processors. The bit width of data transmission from a nucleus processor to another layer is 24 bit. We used the multicasting mesh architecture on each layer because it provides the highest performance/cost ratio and offers a high bandwidth due to its high level of parallelism [35]. As shown in Fig. 5(b), each layer contains 36 nucleus processors. Each layer of the mesh-based network communicates bidirectionally using the data communication interface.

The digital structure of the nucleus processor is shown in Fig. 6, which includes GPe, GPi, STN, and TC processors. Each multiple-synaptic-information-processing (MSIP) router has six ports to communicate the data with the neighbor nucleus processors, which are up port, down port, north port, west port, east port, and south port. Although previous studies have presented the multicasting AER method for convolutional neural network [35], they were unable to solve the following three problems: 1) events with synaptic weighting; 2) implementation of physical synapses; and 3) multiple information routing. The data of the synaptic variables is required to be transmitted by the router across each nucleus processor for the computation of synaptic current. In GPe processor and STN processor, two kinds of synaptic variables are required as output. Taking GPe processor for example, the GPe processor contains two silicon synapse units, one GPe neuron unit, an MSIP router and a configuration unit. The silicon synapse units compute the synaptic currents $I_{\text{GPe}\rightarrow\text{GPe}}$ and $I_{\text{STN}\rightarrow\text{GPe}}$ based on (2), respectively. The synaptic variables s_{GPe}, s_{STN} from the router and the membrane potential V_{GPe} from the GPe neuron unit are used in the silicon synapse units. The GPe neuron unit computes and output the s_{GPe} and $s_{\text{GPe}\rightarrow\text{GPi}}$

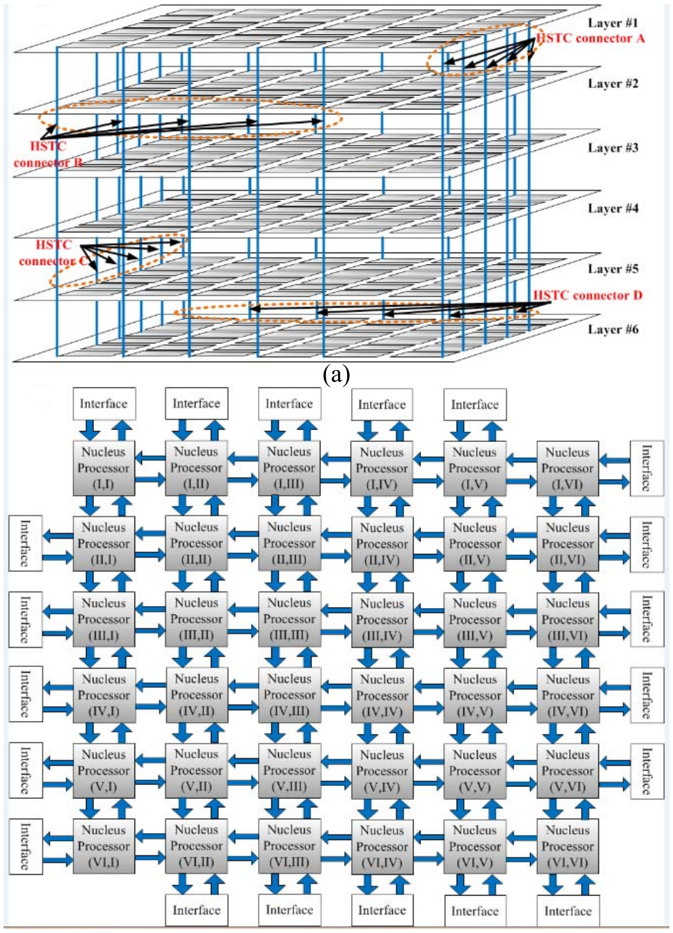

Layer #1
HSTC connector A
Layer #2
HSTC connector B
Layer #3
Layer #4
HSTC connector C
Layer #5
HSTC connector D
Layer #6
(a)
Interface
Nucleus Processor (I,I)
Nucleus Processor (I,II)
Nucleus Processor (I,III)
Nucleus Processor (I,IV)
Nucleus Processor (I,V)
Nucleus Processor (I,VI)
Nucleus Processor (II,I)
Nucleus Processor (II,II)
Nucleus Processor (II,III)
Nucleus Processor (II,IV)
Nucleus Processor (II,V)
Nucleus Processor (II,VI)
Nucleus Processor (III,I)
Nucleus Processor (III,II)
Nucleus Processor (III,III)
Nucleus Processor (III,IV)
Nucleus Processor (III,V)
Nucleus Processor (III,VI)
Nucleus Processor (IV,I)
Nucleus Processor (IV,II)
Nucleus Processor (IV,III)
Nucleus Processor (IV,IV)
Nucleus Processor (IV,V)
Nucleus Processor (IV,VI)
Nucleus Processor (V,I)
Nucleus Processor (V,II)
Nucleus Processor (V,III)
Nucleus Processor (V,IV)
Nucleus Processor (V,V)
Nucleus Processor (V,VI)
Nucleus Processor (VI,I)
Nucleus Processor (VI,II)
Nucleus Processor (VI,III)
Nucleus Processor (VI,IV)
Nucleus Processor (VI,V)
Nucleus Processor (VI,VI)
(b)

Fig. 5. 3-D network topology for the hardware architecture for the large scale cortico-basal ganglia-thalamocortical neural network. (a) Top-level design of the 3-D NoC implementation for the cortico-basal ganglia-thalamocortical network. (b) Multicasting mesh AER structure in each layer in the digital cortico-basal ganglia-thalamocortical system.

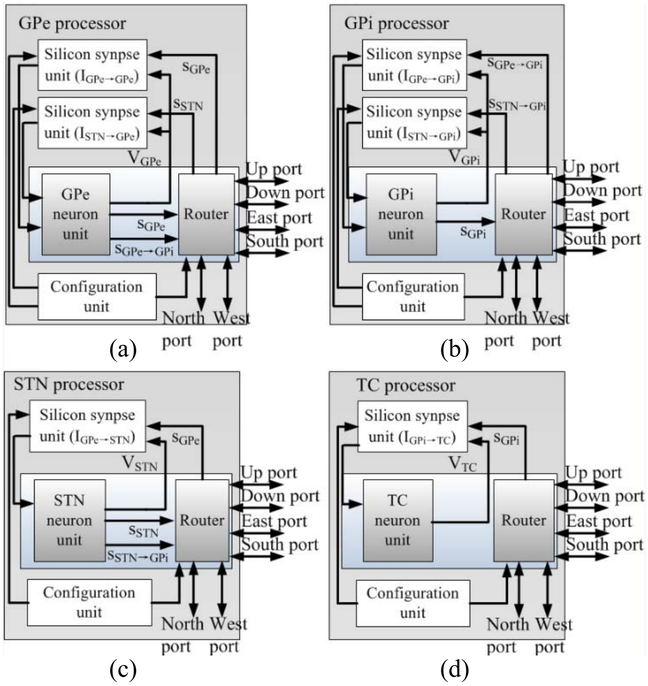


Fig. 6. Digital structure of the nucleus processors including (a) GPe processor, (b) GPi processor, (c) STN processor, and (d) TC processor.

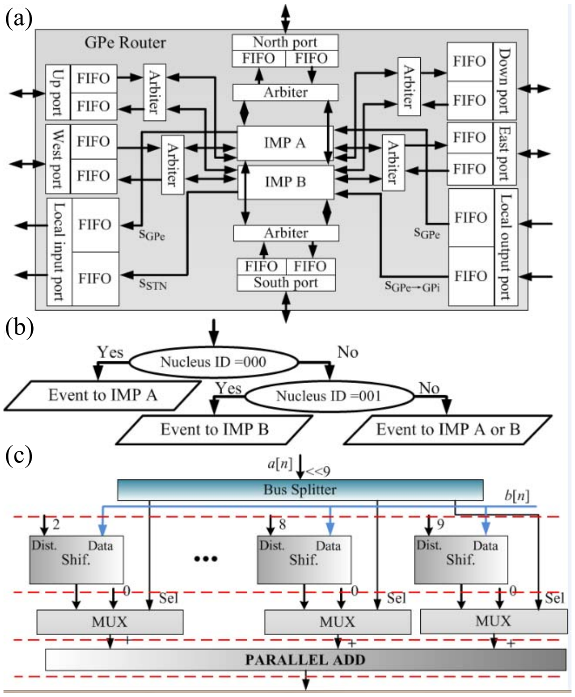


Fig. 7. Digital structure of the router in a nucleus processor. (a) Digital structure of the GPe router. (b) Arbitration algorithm in the GPe router for handling incoming events and determining the destination IMP of the incoming events. (c) NDS multiplier for digital multiplier-less implementation of large-scale SNNs.

signals into the router. The configuration unit is responsible for the configurable parameters setting for the MSIP router. The GPe neuron unit is time-multiplexing, using on-chip memory to achieve 4800 virtual neurons with one physical neuron unit. The digital implementations of other nucleus processors are similar with the GPe processor.

C. Digital Architectures of the MSIP Router and the NDS Multiplier

Unlike the router in the 2-D mesh-based implementation of convolutional neural network presented in a previous study [35], the router in the proposed system needs to deal with the problems of multiple information processing and interchip data communications. The router receives external events from the four neighbor nucleus processors and determines the data transmission direction according to the programmed routing tables in the configuration unit. Since multiple synapse information is demanded in the routing algorithm of the SNN with multiple nuclei, an improved multiple-information processor (IMP) in the MSIP router is presented, as depicted in Fig. 7(a). Although some neuromorphic chips have implemented synaptic weights by memristors, which is more efficient than digital router, the digital router is reconfigurable and more flexible than the conventional crossbar method. Thus, we use the router for synaptic transmission in the presented system.

The MSIP router is divided into four types according to the computation requirement of the silicon synapse unit, which are: GPe router, GPi router, STN router, and TC router. It should be noted that the AER data in the packet transmitted by the router is the synapse information rather than spike information, which is different from the conventional AER-based implementations of neural networks. We selectively show the detailed structure of the GPe router in Fig. 7(a) because it has the most complicated structure, as compared to the other three types of routers. Two IMPs are used in the GPe router for the routing of the incoming events of two different variables separately. A nucleus ID is used in the event packet to separate the variables s_{GPe}, "s_{STN}," $s_{\mathrm{GPe}\rightarrow\mathrm{GPi}}$, "$s_{\mathrm{STN}\rightarrow\mathrm{GPi}}$," and "$s_{\mathrm{GPi}}$" with the ID numbers "000," "001," "010," "011," and "100." An arbiter is used to determine the destination of the data flow to the IMP for routing scheduling. The arbitration algorithm is presented in Fig. 7(b). When an event arrives at a router, the arbiter analyzes the nucleus ID and decides the output, IMP A or IMP B, to which the event is to be forwarded. If the nucleus ID equals 000, the event is sent to IMP A, and if the nucleus ID equals 001, the event will be sent to IMP B. If the nucleus ID in an event is neither 000 nor 001, the event will be sent to IMP A or B randomly. Using the MSIP router, as shown in Fig. 7, we can handle the problem of multiple-information separation processing in an SNN with multiple nuclei. In the IMP unit, a routing algorithm is designed for the routing of multiple synaptic information flows, as shown in Fig. S1 in the supplementary material.

The method to implement the NDS multiplier is shown in Fig. 7(c). The NDS multiplier is used to implement the multiplication of two variables with powers of 2 using a logic shift, which can be multiplier-less with lower hardware cost. Two inputs of variable "$a[n]$" and "$b[n]$" are contained in the NDS multiplier, and the value of $a[n]$ is expected to be a positive value that is less than one. The bus splitter is used to split a bus into single-bit outputs, and the output ports are numbered from least significant bit to most significant bit. The "MUX" block implements the multiplexer operation, which is used to select the input data flow. In the first data line of the MUX block, the input value is zero. If the output of the bus splitter is 1, the second data line is selected and the variable $b[n]$ is shifted rightwards based on the bit number of the bus splitter. Then the single-bit outputs are used as the distances of the shift operations and inputted into the barrel shifter. The "shift" block represents the barrel shifter which can shift the input data by the amount set by the distance bus. The value of variable $b[n]$ is inputted into each data port of shift block, and each output of the shift block is then added together to obtain the final output value of the shift multiplication module. By using this method we are able to calculate the multiplication of the two neuronal variables with a significant reduction of hardware resource cost.

IV. Experimental Results

In this section, we present experimental results, including dynamical behaviors of PLA neuron models, hardware performance of single neuron and the neural network on the LaCSNN system that is realized by six Altera Stratix III 340 FPGAs. DE3 340 development boards are used, each of which has four HSTCs for the multiboard concatenation, one DDR2 memory channel for off-chip data storage and a USB 2.0 interface for the data transmission to a workstation. System tests are performed for the validation of capability to deal with the challenges in both the neuron complexity and the network scalability, and demonstrate the network dynamics for a large-scale SNN model. In general FPGA, chips are mounted on a development board, and add a further contribution that is unrelated with the emulation. As a result, the power consumption of the overall emulation platform is weakly dependent on the number of emulated neurons. In the presented LaCSNN system, the platform equipped with six Altera EP3SE340 FPGA chips dissipates 10.578 W. In terms of efficiency, LaCSNN's power density is 143.92 mW/cm^2, whereas that of a typical central processing unit (CPU) is 50–100 W/cm^2. The neuron model in (1) is easily laid out as a pipelined structure that can be clocked at 100 MHz. We can evaluate neurons with the maximum number of 1 million using 216 copies of the evaluation pipeline under the situation of fully usage of hardware resource with six FPGA chips. Besides, with the hardware infrastructure and the NoC topology presented in the previous section, we can simulate up to 60 million synapses with firing at an average rate of 50 Hz in real time. In fact, the number of layers of the proposed system can be increased by adding the FPGA develop boards, so that more number of neurons and synapses can be simulated.

A. Comparison of the Dynamics of the Original and Proposed Cortico-Basal Ganglia-Thalamocortical Network Model

To evaluate the neural dynamics of the PLA-based models, the investigation of ionic current provides insight into the level of similarity in dynamics in comparison with the original model. The dynamics of the ionic currents in the proposed models are shown in Fig. 8. The ionic current is modeled as a function of voltage and steady-state currents are calculated with slow variables set to their limiting values for fixed voltages. Results show that the ionic current of the modified model is consistent with the original model. The simulation is completed using MATLAB. The original model is based on the studies by Rubin and Terman [22] and Terman *et al.* [30]. The phase portrait of the relationship between the membrane potential and its deviation is also depicted in Fig. 8. It reveals that the proposed models can reproduce the biological dynamics of the desired network model accurately.

In order to further explore the neuronal dynamics of the proposed PLA-based models, a bifurcation analysis is presented in Fig. 9. It reveals that there exists a saddle-node on invariant circle bifurcation of these three neuron models, which is composed of two trajectories, called heteroclinic trajectories, connecting the saddle and the node. It reveals that the neuronal dynamics of the PLA-based models are consistent with the original models, which ensures the accurate reproduction of neuronal dynamical behaviors. The error evaluation is presented in Fig. S2 in the supplementary materials.

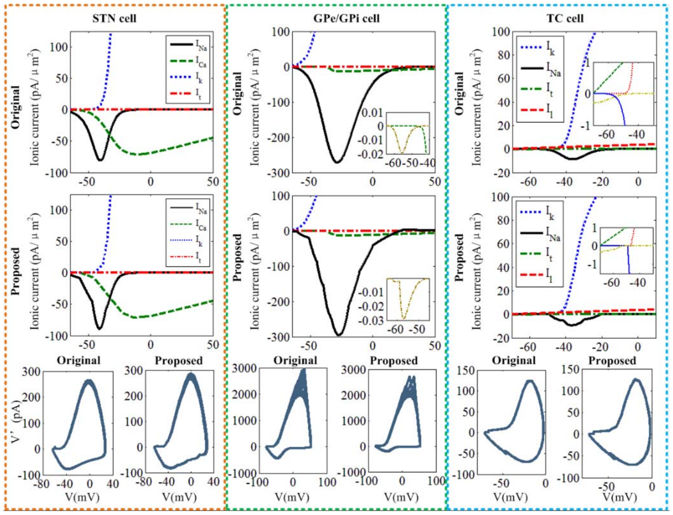


Fig. 8. Dynamical investigation of the proposed cortico-basal ganglia-thalamocortical network model.

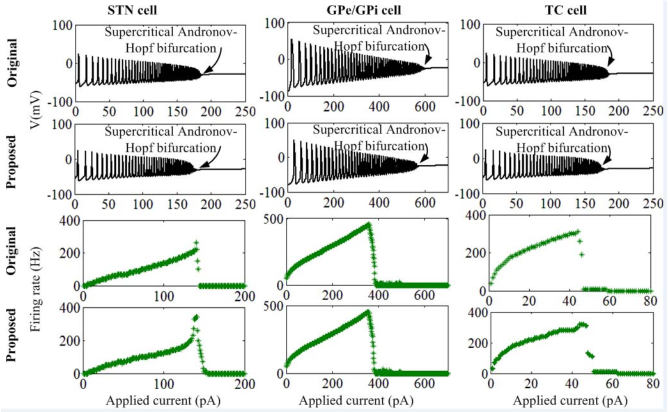


Fig. 9. Bifurcation analysis of the proposed PLA-based neuron models.

B. Network Dynamical Behaviors of the Proposed PLA-Based Models

The network dynamics have been explored under both the normal and the pathological states with different parameter values for I_{app_GPe} and $\text{I}_{\text{GPe}\to\text{GPe}}$, as described in previous works [22], [30]. The definitions of the normal and pathological states are based on the previous experimental results [40], [41]. Fig. 10 shows the accurate neuronal activities using the proposed PLA models in the cortico-basal ganglia-thalamocortical loop under the normal and pathological states. It shows the accurate firing rate and firing patterns according to biophysical studies. Under the normal state the TC cells respond to the excitatory sensorimotor currents faithfully.

C. Hardware Implementation Results of the Cortico-Basal Ganglia-Thalamocortical Network Model

The LaCSNN system uses six Altera Stratix III EP3SL340 FPGAs to establish a real-time simulation platform for the large-scale neural network. Fig. 11 shows the experimental results of the outputs of the LaCSNN system on an oscilloscope. The results are the spiking patterns of GPe, GPi, STN, and TC neurons randomly chosen in the cortico-basal ganglia-thalamocortical loop, respectively, under the normal state. Fig. 12 shows results in the form of a raster plot for 500 ms of simulation, where each dot represents a spike from a neuron at a given time. It shows that the behavior of the basal ganglia network can be reproduced with accurate dynamics. Besides, the electrophysiological patterns of activities reproduced by the hardware implementation of the cortico-basal ganglia-thalamocortical neural network on the LaCSNN system were compared with the software-based simulation, which is shown in Fig. S3 in the supplementary materials.

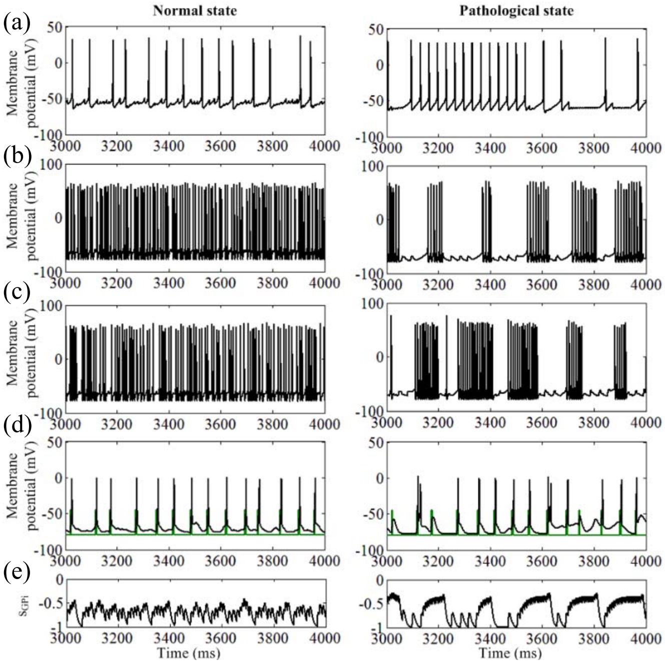


Fig. 10. Network responses and dynamical characteristics under both the normal and the Parkinsonian states. (a) Firing activities of the STN cells under both the normal and the pathological states. (b) Firing activities of the GPe cells under both the normal and the pathological states. (c) Firing activities of the GPi cells under both the normal and the pathological states. (d) Firing activities of the TC cells under both the normal and the pathological states. The green pulse trains represent the sensorimotor currents I_{SM} from the cortex.

D. Accuracy Evaluation and Hardware Performance Analysis

Path cost is the time taken by the packet to be transmitted from a source node to a destination node at a given time, which is measured by clock cycles. In the LaCSNN system, it only takes one clock cycle to convert the neural information into a packet and one clock cycle to convert the packet back to neural information. The cost function of the path cost for the transit through intermediate blocks can be described by

$$f_{\text{PCI}} = \left(f_{\text{hop}} - 2\right) \cdot \left(f_R + 2 \cdot f_{\text{FIFO}}\right) \tag{8}$$

where the cost function f_{hop} is the number of nodes a packet needs to pass during its transmission, which includes source node and destination node. The cost functions f_R and f_{FIFO} are the average of all the time spent by the packet inside the

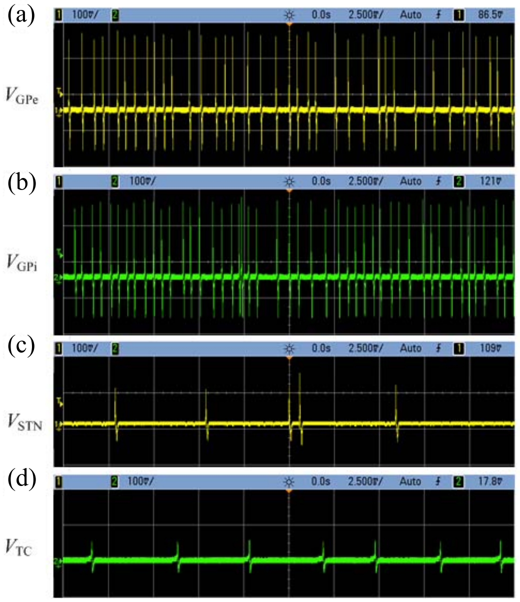


Fig. 11. Experimental results on oscilloscope device. The time division is 2.5 ms and the amplitude division is 100 mV. (a)–(d) Four graphs of spiking activities represent the dynamical behaviors of GPe, GPi, STN, and TC neurons in the large-scale network implemented on the LaCSNN system.

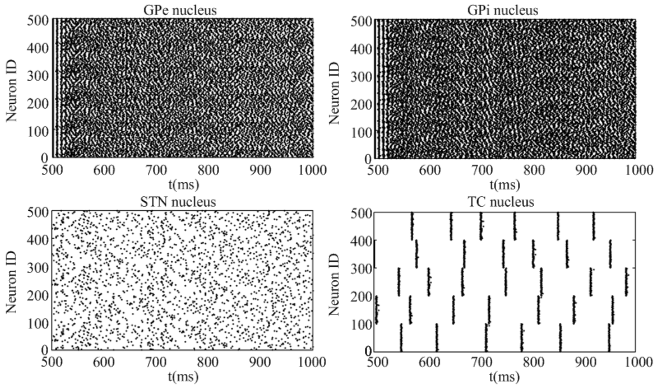


Fig. 12. Raster plot of the spiking patterns of 500 GPe, GPi, STN, and TC neurons chosen randomly in the LaCSNN system.

router and the time spent by the packet in the FIFO in the transmission path, respectively. The path cost in the source node or the destination node is defined as

$$f_{\mathrm{PCSD}} = 2f_R + 4f_{\mathrm{FIFO}} + f_{\mathrm{src}} + f_{\mathrm{dst}} \tag{9}$$

where the cost functions f_{src} and f_{dst} represent the time taken to convert the synapse information into a packet and to convert the packet back into the synapse information, respectively. In fact, the packet conversion takes one clock cycle to process the synapse information into a packet, so $f_{\mathrm{src}} = f_{\mathrm{dst}} = 1$. Thus, the path cost spent in the 3-D multicasting NoC can be described by

$$f_{PC} = \begin{cases} f_{\mathrm{hop}} \cdot (f_R + 2 \cdot f_{\mathrm{FIFO}}) + f_{\mathrm{src}} + f_{\mathrm{dst}}, & f_{\mathrm{hop}} \geq 2 \\ f_R + 2 \cdot f_{\mathrm{FIFO}} + f_{\mathrm{src}} + f_{\mathrm{dst}}, & f_{\mathrm{hop}} = 1 \end{cases} \tag{10}$$

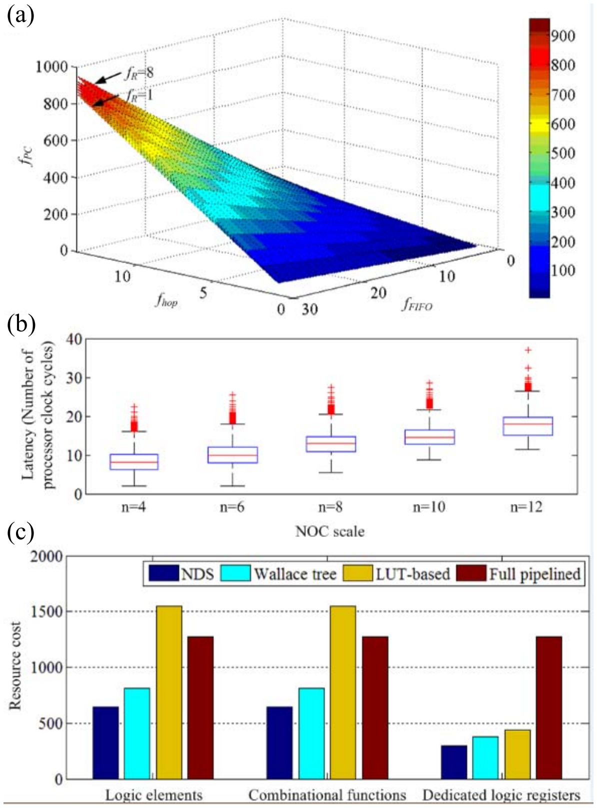


Fig. 13. Performance analysis of the proposed design. (a) Relationship between path cost, number of hops and transmission time in the FIFO. (b) Comparison between the proposed NDS multiplier and other designs of multipliers.

The relationship between the cost functions f_{PC}, f_{hop}, and f_{FIFO} is shown in Fig. 13(a). For a 3-D NoC, the maximum distance between any two nodes is dependent on the network width ($X_{\max}$), network length ($Y_{\max}$), and network height ($Z_{\max}$). Thus, it can be described by

$$\forall n > 2, \exists f_{\mathrm{hop}} \ni, 1 \leq f_{\mathrm{hop}} \leq ((X_{\max} + Y_{\max} + Z_{\max} - 3) - 1). \tag{11}$$

Measured average event latency as a function of network scale is shown in Fig. 13(b) with the synaptic event rate of 100 Hz. Since the average synaptic event rates of all the nuclei are less than 100 Hz, they are within the defined network performance characteristics. The network scale is quantified by the formation of $n \times n$ nodes. As the network scale enlarges, the median and ranges of latencies increases. Nevertheless, all packets are transmitted in under 100 processor clock cycles, which is the time it takes to update the state of the proposed network. No packet is lost at any of the measured input frequencies, which means that the network dynamics can be calculated accurately.

Multipliers are bulky and power-hungry devices that should be avoided in the hardware implementation [42], [43]. They are high-cost building blocks in terms of area, delay, and power consumption. Thus, logic elements are suggested to realize the multiplication. In this paper, we propose an NDS multiplier containing adder, shifter, bus splitter, and multiplexer. In order to obtain results with high precision, the proposed NDS multiplier is designed for variables with 10 bits for the integer part and 10 bits for the fractional part. A comparison is presented with several conventional de-facto efficient multipliers including Wallace tree multiplier, LUT-based multiplier, and full pipelined multiplier, as shown in Fig. 13(c), which was performed on an Altera Cyclone-IV EP4CE115 FPGA. It shows a lower cost of the proposed NDS multiplier compared to other multiplication designs.

V. Discussion

An important approach toward further comprehension of brain dynamics is to build a large-scale biologically realistic functional brain network with detailed neuronal activities. In this section, some essential issues are discussed in detail.

A. *Comparison With State-of-the-Art Hardware Implementations*

This paper presents a digital solution for large-scale implementation of a real-time SNN. In order to demonstrate the computational power and scalability of the LaCSNN system, we compare the hardware performance with three alternative approaches, including CPU, GPU, and multicore bus implementation. A cost function for the computational efficiency is defined as

$$Q = t_{\mathrm{exp}}/t_{\mathrm{bio}} \tag{12}$$

where t_{exp} and t_{bio} are the experimental computational time on the simulation system and the biological activity time, respectively. The CPU-based version ran on an Intel Core2 2.4 GHz CPU. The multicore bus implementation is based on FPGA, and GPU implementation is based on NVIDIA GTX 280 GPU. The computational efficiency of these two methods is estimated according to previous studies [26], [44]–[46]. In addition to the higher computational speed, the LaCSNN system with scalable 3-D NoC structure shows a higher scalability in comparison with the other hardware platforms. As shown in Fig. 14, as the network size increases, the computational efficiency of the LaCSNN system is consistently the highest. The GPU simulation setup is NVIDIA Geforce GTX 280 GPU, and multicore bus implementation uses the Altera EP3SE340 FPGA chip. The CPU-based simulation uses the Quad-core Intel Xeon CPU for the performance evaluation. The network type is the presented cortico-basal ganglia-thalamocortical neural network with the conductance-based neuron model.

There have been several studies aimed at large-scale neural network implementation, including Neurogird [47], SpiNNaker [4], Truenorth [2], BrainScaleS [48], and HiAER [49]. A comparison between the LaCSNN system and other systems, considering biological accuracy of the neuron model, run-time plasticity and the reconfigurability of the hardware system, is listed in Table I. The LIF and conductance-based neuron models are considered with low and high biological accuracy [55]. The Neurogrid project uses a quadratic IF neuron model for the somatic compartment and it uses four Hodgkin–Huxley conductance for the dendritic compartments. However, the drawback for Neurogrid is the lack of reconfigurability, which limits its application in other kinds of neural networks with other neuron models. The SpiNNaker project uses the Izhikevich model, which uses a limited number of nonspecific ionic conductances. Its implemented neuron model can be replaced by a complicated Hodgkin–Huxley type model, but doing so would increase computation speed and reduce network scaling. Truenorth has considerably lower biological plausibility and extendibility, as compared to the proposed work. Although SpiNNaker is also general to simulate arbitrary spiking neuron models, it is initially presented to simulate the Izhikevich model. Its lower computational resource is not preferable to support complex conductance-based neuron models. Besides, it uses ARM CPUs, which is less computational efficient than FPGAs with parallel computational capacity used in this paper. TrueNorth has chosen a fixed SNN model with leaky LIF neurons and limited programmable connectivity, and there is no on-chip learning. It is highly optimized for the chosen model and topology of the network. The BrainScaleS and HiAER projects make excellent contributions; however, neither of them cannot reproduce the dynamics of ionic conductance and have reconfigurability due to the analog approach. Regarding other similar projects, the NeuroDyn system is not for large-scale implementation [50], and the Blue Brain project is not a real-time system [1]. The proposed LaCSNN system is superior to the state-of-the-art

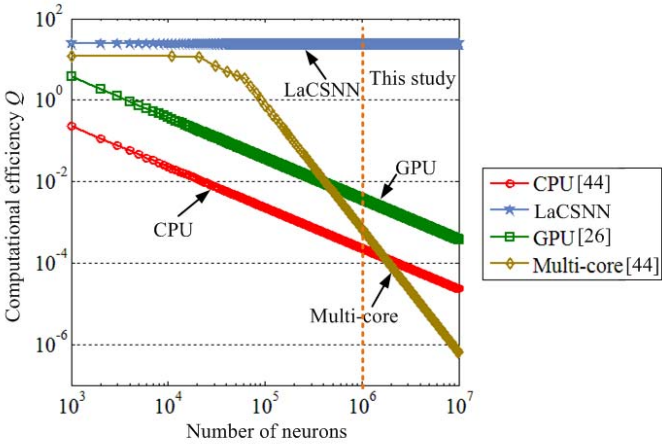


Fig. 14. Computational efficiency and scalability of four different digital approaches. The orange dotted line represents the neuron number considered in this paper. The computational efficiency of the LaCSNN system contains invariant because of its parallel feature as well as the scalable 3-D NoC topology.

TABLE I
Comparison With the State-of-the-Art Neuromorphic Engineering Projects

Project	Biological accuracy	Plasticity	Reconfigurability
Neurogird	Adaptive quadratic IF with Hodgkin-Huxley channels (High)	No	No
SpiNNaker	Izhikevich (Moderate)	Programmable	Yes
Truenorth	LIF (Low)	No	Yes
BrainScaleS	AdEXP (Moderate)	STDP	No
HiAER	IF (Low)	STDP	No
LaCSNN	Conductance-based (High)	Programmable	Yes

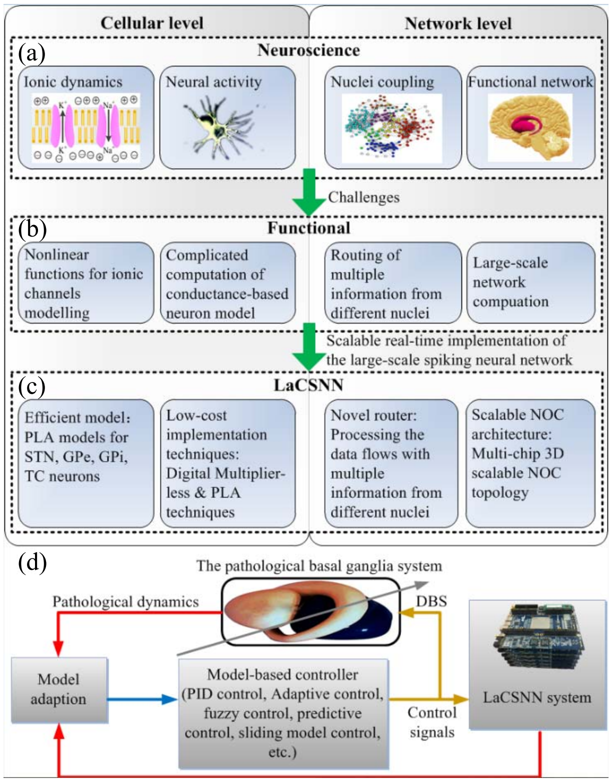

Cellular level
Network level
(a)
Neuroscience
Ionic dynamics
Neural activity
Nuclei coupling
Functional network
Challenges
(b)
Functional
Nonlinear functions for ionic channels modelling
Complicated computation of conductance-based neuron model
Routing of multiple information from different nuclei
Large-scale network compuation
Scalable real-time implementation of the large-scale spiking neural network
(c)
LaCSNN
Efficient model: PLA models for STN, GPe, GPi, TC neurons
Low-cost implementation techniques: Digital Multiplier-less & PLA techniques
Novel router: Processing the data flows with multiple information from different nuclei
Scalable NOC architecture: Multi-chip 3D scalable NOC topology
(d)
The pathological basal ganglia system
Pathological dynamics
DBS
Model adaption
Model-based controller (PID control, Adaptive control, fuzzy control, predictive control, sliding model control, etc.)
Control signals
LaCSNN system

Fig. 15. General overview of the LaCSNN system. (a) Neuroscience view. (b) Functional view. (c) LaCSNN view. (d) Simplified example of how the LaCSNN system fit in a model-based control paradigm of movement disorders.

neuromorphic projects, considering biological plausibility, extendibility as well as reconfigurability, as shown in Table I. All the projects in Table I simulate the large-scale neural network with more than one million neurons and in real time.

B. Applications and Future Works

As shown in Fig. 15, the LaCSNN system is motivated by neurophysiology and implemented with a set of novel techniques for an efficient, scalable, flexible, and reprogrammable design. Four challenges have been overcome through the realization of LaCSNN, which considers both the cellular and network dynamics with a real-time computational performance. Although the field of neuromorphic engineering has attracted the attention of neuroscientists and engineers, there are not yet applications that exploit this technology to its full potential. The presented LaCSNN system is a general and programmable computational engine for neuroscientists, cognitive scientists and brain researchers to investigate brain dynamics, functions and diseases with detailed neuron models. For a low-level case, it can be used as a sensory processing system or biosensors [51]. For a mid-level case, it can be applied to brain–machine interfaces based on the online information processing from the environment [52]. For a high-level case, it can be potentially applied in the decision making of a humanoid robot and visual processing [53].

Furthermore, the motivations for embedding a large-scale cortico-basal ganglia-thalamocortical model in hardware systems go beyond the obvious applications to intelligent agents and neuro-robotics. Previous studies have shown that the model-based control strategies have several clinical and practical applications [54]. The implemented LaCSNN system can be used for numerical calculations that are needed for simulating the model aspect of an observer. Fig. 15(d) shows an example of how the LaCSNN system could be used in a model-based control scheme. The model-based controller can use all the available closed-loop control strategies, including PID control, adaptive control, fuzzy control, predictive control, and sliding mode control. This is only one example, and a more extensive review is discussed in [54].

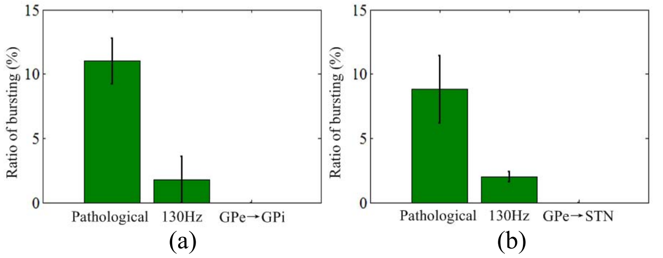


Fig. 16. Effects of the synaptic block on the oscillatory activity. (a) Ratio of bursting of the GPi activities under the condition of synaptic block into the GPi nucleus. (b) Ratio of bursting of the STN activities under the condition of synaptic block into the STN nucleus.

In addition, the presented advanced system can be applied in the exploration of the brain mechanisms and revelation of new findings in the pathological and medical fields. The increment of the bursting activities in the basal ganglia is closely related with the symptom of the movement disorders [56], [57]. Thus, we attempt to explore the possible mechanism underlying these bursting behaviors of the basal ganglia by using the presented LaCSNN system. Specifically, the effects of the synaptic currents from GPe to other nuclei (including GPe→GPi and GPe→STN) on the bursting activities are our main concerns. Different conditions are considered, including the pathological state, standard DBS with 130 Hz on STN nucleus and synaptic block under the pathological state. The DBS stimulation is described by

$$I_{\mathrm{DBS}} = I_d H(\sin(2\pi t/\rho_d)) \cdot \left[1 - H(\sin(2\pi(t+\delta_d))/\rho_d)\right] \tag{13}$$

where $H(\cdot)$ represents the Heaviside function. The amplitude $i_d = 300\ \mu\mathrm{A/cm}^2$ and the period $\rho_d = 1000/130 \approx 7.69$ ms. The parameter $\delta_d = 0.3$ ms. The synaptic block is modeled by setting the corresponding synaptic currents to zero. In Fig. 16, "GPe→GPi" and "GPe→STN" represent the block of the synaptic currents into the corresponding nucleus. As shown in Fig. 16(a), when the synaptic currents from GPe into GPi are blocked, the bursting activity of the GPi in the beta band disappears. As shown in Fig. 16(b), the bursting activity of the STN can also be effectively suppressed by blocking the synaptic input from the GPe. Thus, it can be deduced that the synaptic currents from GPe into other nuclei are vital for the bursting activity in the whole neural network, which may suggest the possibility of GPe acting as the target of DBS.

Further possible improvement of the LaCSNN system is divided into three aspects. First, there is no user interface for the direct observation of the network dynamics and configuration of the network parameters in real time. A graphical user interface can be developed to enable a user to change the model parameters and plot spike rasters from a selected neural layer as well as enter commands. Second, the interchip connection of the LaCSNN system can use FPGAs for the configuration. Other FPGAs can be used to regulate vertical transactions to enhance the system flexibility. However, this would increase the hardware cost of the overall system, which is a tradeoff with the system flexibility. Third, although not pursued in this paper, the LaCSNN system can be extended to any other brain region of interest. The high reconfigurability of LaCSNN the system enables the support of electrophysiology experiments.

VI. Conclusion

In this paper, we presented the LaCSNN system for a scalable, large-scale SNN and its efficient hardware design on reconfigurable FPGA, which aims at a high-performance realization of the SNNs. We started with the cortico-basal ganglia-thalamocortical system because it is necessary for human consciousness and is closely related to movement disorders. A real-time system is established for the implementation of SNNs, which contains multiple nuclei implemented by biologically realistic neuron models with ionic conductance dynamics. It provides a significant perspective for the efficient implementation of a detailed brain network with multiple nuclei, especially with biologically realistic neuron models. For the implementation of a single neuron model, a set of techniques are presented for the multiplier-less and PLA implementation, which overcomes the problems of resource shortage of embedded multipliers and on-chip memory. For the network level, a 3-D NoC structure is presented for the extendibility of a multichip system for the simulation of the SNNs. Additionally, a novel router is designed for information processing in the multiple-nuclei SNN, which is particularly useful in the simulation of a network with multiple nuclei. In comparison with other implementations, including CPU, multicore, and GPU-based system, it has advantages in both computational speed and scalability. As compared with state-of-the-art neuromorphic projects, it is superior in both the biological accuracy and the reconfigurability. The LaCSNN system is also applicable to other applications due to its high scalability, flexibility, and computational power.

References

[1] H. Markram, "The blue brain project," *Nat. Rev. Neurosci.*, vol. 7, no. 2, pp. 153–160, Feb. 2006.

[2] P. A. Merolla *et al.*, "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, Aug. 2014.

[3] H. De Garis, C. Shuo, B. Goertzel, and L. Ruiting, "A world survey of artificial brain projects, part I: Large-scale brain simulations," *Neurocomputing*, vol. 74, nos. 1–3, pp. 3–29, Dec. 2010.

[4] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The SpiNNaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652–665, May 2014.

[5] P. Dayan and L. F. Abbott, "Theoretical neuroscience: Computational and mathematical modeling of neural systems," *J. Cogn. Neurosci.*, vol. 15, no. 1, pp. 154–155, Jan. 2003.

[6] J. H. Goldwyn, N. S. Imennov, M. Famulare, and E. Sheabrown, "Stochastic differential equation models for ion channel noise in Hodgkin–Huxley neurons," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 83, no. 1, pp. 4190–4208, Apr. 2011.

[7] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952.

[8] E. M. Izhikevich, "Simple model of spiking neurons," *IEEE Trans. Neural Netw.*, vol. 14, no. 6, pp. 1569–1572, Nov. 2003.

[9] H. J. Kang *et al.*, "Spatio-temporal transcriptome of the human brain," *Nature*, vol. 478, no. 7370, pp. 483–489, Oct. 2011.

[10] R. A. Poldrack and M. J. Farah, "Progress and challenges in probing the human brain," *Nature*, vol. 526, no. 7573, pp. 371–379, Oct. 2015.

[11] J. W. Mink, "The basal ganglia: Focused selection and inhibition of competing motor programs," *Progr. Neurobiol.*, vol. 50, no. 4, pp. 381–425, Nov. 1996.

[12] I. Bar-Gad and H. Bergman, "Stepping out of the box: Information processing in the neural networks of the basal ganglia," *Current Opinion Neurobiol.*, vol. 11, no. 6, pp. 689–695, Dec. 2001.

[13] G. E. Alexander, M. R. DeLong, and P. L. Strick, "Parallel organization of functionally segregated circuits linking basal ganglia and cortex," *Annu. Rev. Neurosci.*, vol. 9, no. 1, pp. 357–381, 1986.

[14] J. Yelnik, "Functional anatomy of the basal ganglia," *Movement Disorders*, vol. 17, no. 3, pp. 15–21, Mar. 2012.

[15] E. Gleichgerrcht, A. Ibáñez, M. Roca, T. Torralva, and F. Manes, "Decision-making cognition in neurodegenerative diseases," *Nat. Rev. Neurol.*, vol. 6, no. 11, pp. 611–623, Nov. 2010.

[16] L. Ding and J. I. Gold, "The basal ganglia's contributions to perceptual decision making," *Neuron*, vol. 79, no. 4, pp. 640–649, Aug. 2013.

[17] C. Eliasmith, T. C. Stewart, and X. Choo, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, Nov. 2012.

[18] T. C. Stewart, T. Bekolay, and C. Eliasmith, "Learning to select actions with spiking neurons in the basal ganglia," *Front. Neurosci.*, vol. 6, no. 2, p. 2, Jan. 2012.

[19] F. Chersi, M. Mirolli, G. Pezzulo, and G. Baldassarre, "A spiking neuron model of the cortico-basal ganglia circuits for goal-directed and habitual action learning," *Neural Netw.*, vol. 41, no. 5, pp. 212–224, May 2013.

[20] C. C. Lo and X. J. Wang, "Cortico-basal ganglia circuit mechanism for a decision threshold in reaction time tasks," *Nat. Neurosci.*, vol. 9, no. 7, pp. 956–963. Jun. 2006.

[21] M. J. Pearson *et al.*, "Implementing spiking neural networks for real-time signal-processing and control applications: A model-validated FPGA approach," *IEEE Trans. Neural Netw.*, vol. 18, no. 5, pp. 1472–1487, Sep. 2007.

[22] J. E. Rubin and D. Terman, "High frequency stimulation of the subthalamic nucleus eliminates pathological thalamic rhythmicity in a computational model," *J. Comput. Neurosci.*, vol. 16, no. 3, pp. 211–235, May/Jun. 2004.

[23] E. M. Izhikevic and G. M. Edelman, "Large-scale model of mammalian thalamocortical systems," *Proc. Nat. Acad. Sci. USA*, vol. 105, no. 9, pp. 3593–3598, Mar. 2008.

[24] C. M. Thibeault and S. Narayan, "Using a hybrid neuron in physiologically inspired models of the basal ganglia," *Front. Comput. Neurosci.*, vol. 7, no. 7, pp. 88–105, Jul. 2013.

[25] S. Yang *et al.*, "Cost-efficient FPGA implementation of basal ganglia and their Parkinsonian analysis," *Neural Netw.*, vol. 71, pp. 62–75, Nov. 2015.

[26] J. Igarash, O. Shouno, T. Fukai, and H. Tsujino, "Real-time simulation of a spiking neural network model of the basal ganglia circuitry using general purpose computing on graphics processing units," *Neural. Netw.*, vol. 24, no. 9, pp. 950–960, Jun. 2011.

[27] A. M. Andrew, *Understanding Intelligence*, vol. 29. Cambridge, MA, USA: MIT Press, Apr. 2000, pp. 1333–1340.

[28] G. E. Alexander and M. D. Crutcher, "Functional architecture of basal ganglia circuits: Neural substrates of parallel processing," *Trends Neurosci.*, vol. 13, no. 7, pp. 266–271, Jul. 1990.

[29] A. Buot and J. Yelnik, "Functional anatomy of the basal ganglia: Limbic aspects," *Rev. Neurol.*, vol. 168, nos. 8–9, pp. 569–575, Jun. 2012.

[30] D. Terman, J. E. Rubin, A. C. Yew, and C. J. Wilson, "Activity patterns in a model for the subthalamopallidal network of the basal ganglia," *J. Neurosci.*, vol. 22, no. 7, pp. 2963–2976, Apr. 2002.

[31] Y. Guo, J. E. Rubin, C. C. McIntyre, J. L. Vitek, and D. Terman, "Thalamocortical relay fidelity varies across subthalamic nucleus deep brain stimulation protocols in a data-driven computational model," *J. Neurophysiol.*, vol. 99, no. 3, pp. 1477–1492, Mar. 2008.

[32] R. Q. So, A. R. Ken, and W. M. Grill, "Relative contributions of local cell and passing fiber activation and silencing to changes in thalamic fidelity during deep brain stimulation and lesioning: A computational modeling study," *J. Comput. Neurosci.*, vol. 32, no. 3, pp. 499–519, Oct. 2012.

[33] M. A. J. Lourens, J. A. Nirody, H. G. E. Meijer, T. Heida, and S. A. V. Gils, "The effect of spike time dependent plasticity on activity patterns in the basal ganglia," *BMC Neurosci.*, vol. 12, no. 1, pp. 351–362, Jul. 2011.

[34] M. A. J. Lourens, "Neural network dynamics in Parkinson's disease," Ph.D. dissertation, Univ. Dept. Elect. Eng., Math. Comput. Sci., Twente, Enschede, The Netherlands, Apr. 2013.

[35] C. Zamarreno-Ramos, A. Linares-Barranco, T. Serrano-Gotarredona, and B. Linares-Barranco, "Multicasting mesh AER: A scalable assembly approach for reconfigurable neuromorphic structured AER systems. Application to ConvNets," *IEEE Trans. Biomed. Circuits Syst.*, vol. 7, no. 1, pp. 82–102, Jun. 2013.

[36] N. Kasabov, K. Dhoble, N. Nuntalid, and G. Indiveri, "Dynamic evolving spiking neural networks for on-line spatio-and spectro-temporal pattern recognition," *Neural Netw.*, vol. 41, no. 5, pp. 188–201, May 2013.

[37] S. Ghosh-Dastidar and H. Adeli, "Spiking neural networks," *Int. J. Neural Syst.*, vol. 19, no. 4, pp. 295–308, Aug. 2009.

[38] S. Mitra, S. Fusi, and G. Indiveri, "Real-time classification of complex patterns using spike-based learning in neuromorphic VLSI," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 1, pp. 32–42, Feb. 2009.

[39] J. J. Wade, L. J. McDaid, J. A. Santos, and H. M. Sayers, "SWAT: A spiking neural network training algorithm for classification problems," *IEEE Trans. Neural Netw.*, vol. 21, no. 11, pp. 1817–1830, Nov. 2010.

[40] P. Brown *et al.*, "Dopamine dependency of oscillations between subthalamic nucleus and pallidum in Parkinson's disease," *J. Neurosci.*, vol. 21, no. 3, pp. 1033–1038, Feb. 2001.

[41] A. Raz, E. Vaadia, and H. Bergman, "Firing patterns and correlations of spontaneous discharge of pallidal neurons in the normal and the tremulous 1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine vervet model of parkinsonism," *J. Neurosci.*, vol. 20, no. 22, pp. 8559–8571, Nov. 2000.

[42] H. Soleimani, A. Ahmadi, and M. Bavandpour, "Biologically inspired spiking neurons: Piecewise linear models and digital implementation," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 59, no. 12, pp. 2991–3004, Dec. 2012.

[43] M. Storace and T. Poggi, "Digital architectures realizing piecewise-linear multivariate functions: Two FPGA implementations," *Int. J. Circuit Theory Appl.*, vol. 39, no. 1, pp. 1–15, Jan. 2011.

[44] J. Luo *et al.*, "Real-time simulation of passage-of-time encoding in cerebellum using a scalable FPGA-based system," *IEEE Trans. Biomed. Circuits Syst.*, vol. 10, no. 3, pp. 742–753, Jun. 2016.

[45] K. Cheung, S. R. Schultz, and W. Luk, "A large-scale spiking neural network accelerator for FPGA systems," in *Artificial Neural Networks and Machine Learning—ICANN*, vol. 1. Heidelberg, Germany: Springer, Sep. 2012, pp. 113–120.

[46] T. S. T. Mak, P. Sedcole, P. Y. K. Cheung, and W. Luk, "On-FPGA communication architectures and design factors," in *Proc. Int. Conf. Field Program. Logic Appl. (FPL)*, Aug. 2006, pp. 1–8.

[47] B. V. Benjamin *et al.*, "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[48] J. Schemmel *et al.*, "A wafer-scale neuromorphic hardware system for large-scale neural modeling," in *Proc. ISCAS*, Aug. 2010, pp. 1947–1950.

[49] J. Park, T. Yu, S. Joshi, C. Maier, and G. Cauwenberghs, "Hierarchical address event routing for reconfigurable large-scale neuromorphic systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 10, pp. 2408–2422, Oct. 2017.

[50] T. Yu and G. Cauwenberghs, "Biophysical synaptic dynamics in an analog VLSI network of Hodgkin–Huxley neurons," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Sep. 2009, pp. 3335–3338.

[51] S.-C. Liu and T. Delbruck, "Neuromorphic sensory systems," *Current Opinion Neurobiol.*, vol. 20, no. 3, pp. 288–295, Jun. 2010.

[52] F. Corradi and G. Indiveri, "A neuromorphic event-based neural recording system for smart brain-machine-interfaces," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 5, pp. 699–709, Oct. 2015.

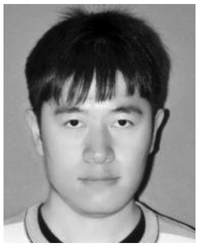
[53] C. Bartolozzi *et al.*, "Embedded neuromorphic vision for humanoid robots," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Aug. 2011, pp. 129–135.

[54] S. J. Schiff, *Neural Control Engineering: The Emerging Intersection Between Control Theory and Neuroscience*, vol. 65. Cambridge, MA, USA: MIT Press, 2012, pp. 337–356.

[55] C. Eliasmith and O. Trujillo, "The use and abuse of large-scale brain models," *Current Opinion Neurobiol.*, vol. 25, no. 25, pp. 1–6, Apr. 2014.

[56] P. Brown. "Oscillatory nature of human basal ganglia activity: Relationship to the pathophysiology of Parkinson's disease," *Movement Disorders*, vol. 18, no. 4, pp. 357–363, Dec. 2002.

[57] M. Weinberger *et al.*, "Oscillatory activity in the globus pallidus internus: Comparison between Parkinson's disease and dystonia," *Clin. Neurophysiol.*, vol. 123, no. 2, pp. 358–368, Feb. 2012.

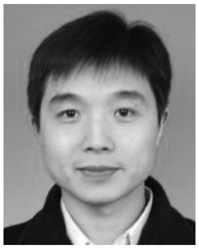


Shuangming Yang received the B.S. degree from the Hebei University of Technology, Tianjin, China, in 2013, and the M.S. degree from Tianjin University, Tianjin, in 2016, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering.

His current research interests include neuromorphic engineering, computational modeling of neural system, neural control engineering, robotic control, and machine learning.

Jiang Wang was born in China, in 1964. He received the master's degree in power and automation engineering and the Ph.D. degree from the University of Tianjin, Tianjin, China, in 1989 and 1996, respectively.

He is a Professor with the School of Electrical and Information Engineering, Tianjin University, Tianjin. His current research interests include nonlinear dynamical systems, neuroscience, and information processing and detecting.

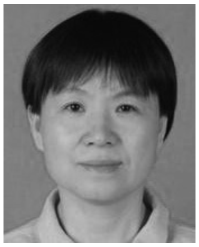


Bin Deng received the B.S., M.S., and Ph.D. degrees in electrical engineering from Tianjin University, Tianjin, China, in 2001, 2004, and 2007, respectively.

After his Post-Doctoral training with the School of Electrical and Information Engineering, Tianjin University, he served as a Research Assistant with the Department of Electrical Engineering, Hong Kong Polytechnic University, Hong Kong. He is currently a Professor with the School of Electrical Engineering and Automation, Tianjin University. His current research interests include the dynamic analysis of neuron model, the design of neuron model based on digital devices, and the nonlinear analysis of neuron electrical information.

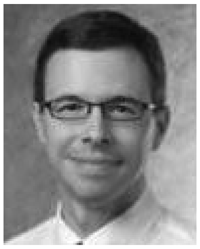
Chen Liu (M'17) was born in China, in 1988. She received the B.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2011 and 2016, respectively.

She was a Research Scholar with the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA, in 2015. She is currently a Lecturer with the School of Electrical and Information Engineering, Tianjin University and a Post-Doctoral Fellow with the Department of Physics, Hong Kong Baptist University, Hong Kong. Her current research interests include neural control engineering and computational modeling.



Huiyan Li received the Ph.D. degree from Tianjin University, Tianjin, China, in 2007.

She is currently a Professor with the School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin. Her current research interests include nonlinear systems and neural networks.



Chris Fietkiewicz received the B.S. degree in electrical engineering through a joint program between Messiah College, Mechanicsburg, PA, USA, and Temple University, Philadelphia, PA, USA, in 1991, and the Ph.D. degree in computer science from Case Western Reserve University, Cleveland, OH, USA, in 2010.

He was with the industry as an Electrical Engineer and a Software Engineer. He was a Post-Doctoral Associate with the State University of New York, New York, NY, USA. He uses techniques from computational neuroscience to better understand the nervous system. His current research interests include biological modeling and neuroinformatics.

Kenneth A. Loparo (F'99–LF'16) received the B.S. and M.S. degrees in mechanical engineering from Cleveland State University, Cleveland, OH, USA, and the Ph.D. degree in systems and control engineering from Case Western Reserve University, Cleveland.

He is the Nord Professor of engineering with faculty appointments with the Department of Electrical Engineering and Computer Science, Mechanical Engineering and Biomedical Engineering, Case Western Reserve University. His current research interests include stability and control of nonlinear systems with applications to large-scale electric power systems, nonlinear filtering with applications to monitoring, fault detection, diagnosis and reconfigurable control, etc.

Dr. Loparo is a fellow of the American Institute for Medical and Biological Engineering.