BiCoSS: Toward Large-Scale Cognition Brain With Multigranular Neuromorphic Architecture

Shuangming Yang, *Member, IEEE*, Jiang Wang, *Member, IEEE*, Xinyu Hao, Huiyan Li, Xile Wei, *Member, IEEE*, Bin Deng, *Senior Member, IEEE*, and Kenneth A. Loparo, *Life Fellow, IEEE*

***Abstract*— The further exploration of the neural mechanisms underlying the biological activities of the human brain depends on the development of large-scale spiking neural networks (SNNs) with different categories at different levels, as well as the corresponding computing platforms. Neuromorphic engineering provides approaches to high-performance biologically plausible computational paradigms inspired by neural systems. In this article, we present a biological-inspired cognitive supercomputing system (BiCoSS) that integrates multiple granules (GRs) of SNNs to realize a hybrid compatible neuromorphic platform. A scalable hierarchical heterogeneous multicore architecture is presented, and a synergistic routing scheme for hybrid neural information is proposed. The BiCoSS system can accommodate different levels of GRs and biological plausibility of SNN models in an efficient and scalable manner. Over four million neurons can be realized on BiCoSS with a power efficiency of 2.8k larger than the GPU platform, and the average latency of BiCoSS is 3.62 and 2.49 times higher than conventional architectures of digital neuromorphic systems. For the verification, BiCoSS is used to replicate various biological cognitive activities, including motor learning, action selection, context-dependent learning, and movement disorders. Comprehensively considering the programmability, biological plausibility, learning capability, computational power, and scalability, BiCoSS is shown to outperform the alternative state-of-the-art works for large-scale SNN, while its real-time computational capability enables a wide range of potential applications.**

***Index Terms*— Brain-inspired computing, computational neuroscience, field-programmable gate array (FPGA), large-scale spiking neural network (SNN), neuromorphic.**

I. Introduction

THE further exploration and comprehension of neural processing mechanisms and biological dynamics in the human brain require building large-scale, anatomically realistic models of the functioning brain network to increase understanding of how neural systems process information to generate network behaviors for cognition tasks [1]–[3]. The most fundamental human cognitive behavior requires the large-scale collective spiking activities of millions of neurons [4]–[6]. An important problem in the field of neuroscience is to understand and explain the underlying mechanisms of high-level human cognition at the cellular level. The challenge is to simultaneously correlate the neural dynamics with brain cognition functions, which requires simulation systems with high computational power and a more biologically plausible computational architecture. In general, there are two critical problems in terms of this challenge.

1) The von Neumann architecture has its bottleneck and is intrinsically different from the computational mode of the human brain from the computational architecture point of view [7]–[9].
2) Current brain simulation ignores the relationship between the detailed dynamical behaviors of neurons and the cognitive functions of the human brain [10], resulting in a gap in comprehending the information processing mechanism of the human brain for cognition from the cellular point of view.

In recent years, high-performance neuromorphic systems have been developed as a powerful approach toward the study of neuroscience, which uses the non-von Neumann architecture inspired by the neural systems in the human brain, providing new research schemes and paradigms [11]–[13]. The study of neuroscience is, in turn, facilitates the establishment and improvement of the neuromorphic systems. Due to the rapid development of neuroscience, an ideal neuromorphic system should be characterized with the following four features: first, support the large-scale complicated neural network that can reproduce rich spatiotemporal dynamical activities; second, support a wide range of spike-based coding schemes and routing schemes for hybrid information flow; third, support computation for multibrain region with different cognitive tasks; and fourth, support multilevel multigranule computing in the field of neuroscience [14]–[16]. The multilevel multigranule computing refers to the following: at the level of neural morphology, it includes point neuron model and compartmental neuron model; at the level of biological plausibility, it includes leaky integrate-and-fire (LIF), Izhikevich, Hodgkin–Huxley (H-H) models, and so on; at the level

Manuscript received September 28, 2019; revised April 29, 2020 and September 10, 2020; accepted December 8, 2020. Date of publication January 11, 2021; date of current version July 7, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61701320, Grant 61871287, Grant 62071324, and Grant 62006170; in part by the Natural Science Foundation of Tianjin under Grant 18JCZDJC32000; and in part by the China Postdoctoral Science Foundation under Grant 2020M680885. *(Corresponding author: Bin Deng.)*

Shuangming Yang, Jiang Wang, Xinyu Hao, Xile Wei, and Bin Deng are with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yangshuangming@tju.edu.cn; dengbin@tju.edu.cn).

Huiyan Li is with the School of Automation and Electrical Engineering, Tianjin University of Technology and Education, Tianjin 300222, China.

Kenneth A. Loparo is with the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH 44106 USA (e-mail: kal4@case.edu).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2020.3045492.

Digital Object Identifier 10.1109/TNNLS.2020.3045492

of network topology, it includes feedforward and recurrent networks; at the level of the synapse, it includes electrical and chemical synapses; and at the level of learning, it includes various forms of learning rules. In the field of neuromorphic engineering, it requires a hybrid compatible system to support these features in a unified architecture.

No such hybrid neuromorphic architecture and system have existed yet. The existing works either support a limited kind of neuron model [17], [18] or cannot calculate the dendritic nonlinearity [2], [19]. Although SpiNNaker presents an excellent scheme for programmable neuromorphic computing, it uses von Neumann architecture on each core, which limits the further enhancement of its computing performance [20]. Since the field-programmable gate array (FPGA) is featured by the reconfigurable capability, parallel calculation, and distributed architecture, numerous studies on FPGA-based implementation of SNNs have been presented for different applications with different architectures [21]–[26]. These studies have presented essential frameworks for digital neuromorphic engineering. However, none of them can solve the aforementioned problem for large-scale neuromorphic computing. This study presents a multigranule hybrid compatible non-von Neumann paradigm using FPGA chips, which is based on scalable hierarchical heterogeneous multicore architecture, realizing the real-time computing of the biological activities for multiple cognitive tasks in multiple brain areas.

The remainder of the article is organized as follows. Section II presents the overview of the biological-inspired cognitive supercomputing system (BiCoSS). In Section III, the system architecture is described, and the method of neural information routing is introduced in Section IV. Section V presents the performance evaluation of BiCoSS in detail and the experimental results related to human cognition. Discussion in terms of the comparison with the state-of-the-art projects and other significant points is presented in Section VI. Finally, the article is concluded in Section VII.

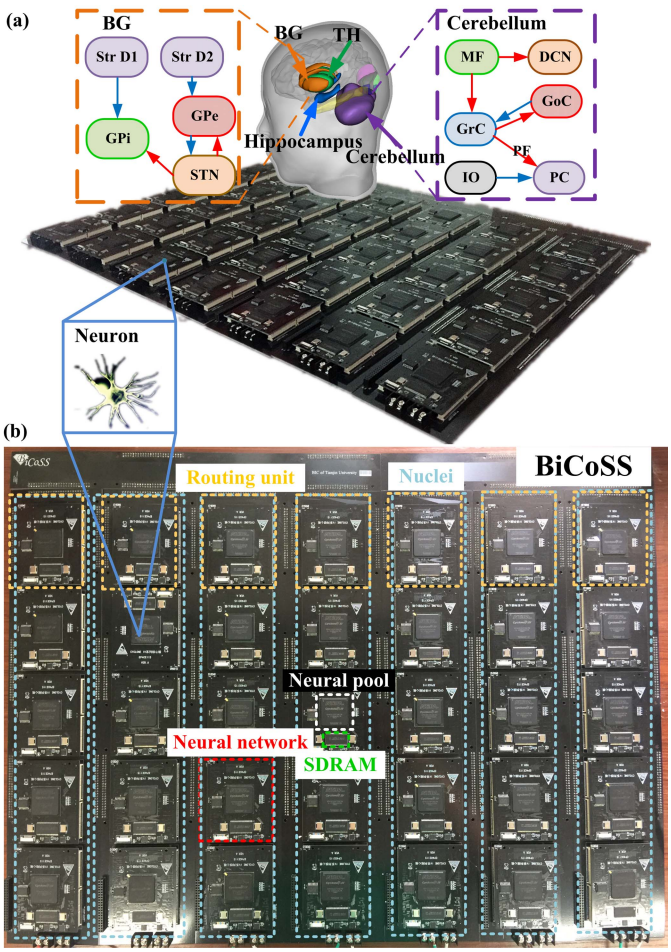


Fig. 1. Overview of BiCoSS. (a) BiCoSS for cognitive functions of the human brain. (b) BiCoSS system with 35 chips having a power consumption of 10.419 W and capable of real-time computing four million neurons.

II. Overview of BiCoSS

A. Neuromorphic Paradigm for Multigranule Neural Processing

Multiscale and multitype biological evidence of the human brain obtained by neuroscientists provides vital inspiration for the construction of more biologically meaningful computational models of the brain. BiCoSS is a specified platform that provides a unified description and solution for spiking-based models in computational neuroscience. The layout and physical view of the BiCoSS system are shown in Fig. 1. By determining multitype and multigranule neuron models for multiple cognitive tasks shown in Table III, the functions are realized on the unified BiCoSS architecture. By aligning data streams, the BiCoSS system can realize various kinds of cognition flexibly. Multiple nuclei are realized on BiCoSS based on multigranule modes, including basal ganglia (BG), cerebellum, hippocampus, and thalamus (TH). The BG consists of the subthalamic nucleus (STN), globus pallidus externus (GPe), globus pallidus internus (GPi), and striatum (Str) D1 and D2. The cerebellum contains mossy fiber (MF), deep cerebellar nucleus (DCN), granule cells (GrCs), Golgi cells (GoCs), Purkinje cells (PCs), and inferior olive (IO).

In addition, BiCoSS includes multiple spatial scales defined on three levels: the single neuron level, the network level, and the nucleus level. At the microscopic (neuron) level, deviations exist in types of neurons, synapses, and structure compatibility that depend on the complexity of tasks. At the mesoscopic (network) level, intrinsic connection and plasticity in each brain area are effectively integrated within networks, which realizes the characteristics of spiking neural networks (SNNs). At the macroscopic (nucleus) level, the cooperation among different brain regions is responsible for neural information processing, enabling highly intelligent cognitive processes. Therefore, the proposed BiCoSS system integrates information processing mechanisms at the microscopic, mesoscopic, and macroscopic levels to model a large-scale cognition brain to achieve brain-like intelligence.

B. Design Concept

BiCoSS uses scalable hierarchical heterogeneous multicore architecture with localized memory and significant parallelization. The critical designs for the hybrid compatible multigranule neuromorphic system are described as follows.

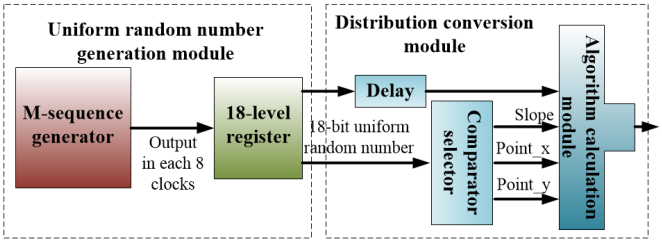


Fig. 2. Generator module of the stochastic neural heterogeneity.

1) Independency Reconfigurable Population Processor: The neural population can be reconfigured into different modes independently. Different types and granularities of models can be used for different types, levels, and scales of cognitive computing. Neuron models with different levels of biological plausibility, including LIF, Izhikevich, and H-H models, can be realized flexibly, aiming at different kinds of cognition tasks and different comprehensive levels of mechanisms underlying the human brain. FPGA is employed in this study aiming at reconfigurable parallel distributed computing of brain-inspired SNNs.

2) Stochastic Neural Heterogeneity: Neural heterogeneity exists in the mammalian brain, which can optimize the temporal coding for neural information processing. It can sparse sensory representation and enhance the response to a periodic external stimulus, which can play an essential role in enhancing the response of the neural system to weak signal detection. Therefore, BiCoSS presents an efficient algorithm and implementation method to enable neural heterogeneous in various kinds of large-scale SNNs, which is closer to the mechanisms in the biological brain. In order to cut down the hardware resource cost, a rapid generation strategy for the Gaussian white noise on FPGA is realized. There are two modules to generate the neural heterogeneity: the uniform random number generation module and the distribution conversion module. The detailed digital implementation method is depicted in Fig. 2. The uniform random number generation module contains the m-sequence generator and decorrelation processing. The m-sequence generators are employed to generate m-sequence by implementing a linear feedback shifting register (LFSR) sequence. The primitive polynomial of the LFSR is $x^{18} + x^7 + 1$. The distribution conversion module contains the comparison selector, delay registers, and arithmetic calculation module. All the multipliers are replaced by shifting and addition operations to cut down the hardware resource cost.

3) Routing Scheme for Hybrid Neural Information: There are two categories of hybrid neural information considered in BiCoSS design. First, different types of synapse models utilize various forms of neural spiking activities in terms of neural information routing. In general, there are two bit-width-level forms of neural information to calculate the synaptic dynamics, which are based on action potential (AP) and neural spikes. The AP-based neural information requires more than 8 bits in digital computation, and spike-based neural information only needs a 1-bit calculation that can be realized by a multiplexer based on logic elements. Therefore, BiCoSS presents a synergistic spike-action-potential-based routing scheme, with a predetecting and classified processing mechanism in the routing scheme to be compatible with various forms of neural information. Second, different nuclei will generate different specified information flow to realize cognitive activities in a collaborative manner, which is considered in BiCoSS design with a novel routing scheme on an address event routing (AER) infrastructure.

III. BiCoSS Architecture

BiCoSS is a custom system-on-a-programmable-chip (SoPC) with 35 Cyclone IV FPGA processor nodes, as shown in Fig. 1(b). For the BiCoSS system, five circuits are connected using a two-layer board with a PMC connector. The serial configuration device, i.e., the EPCS128N chip, is used to store the program after the power is OFF. For each neural network unit, the FPGA chip is connected with two synchronous dynamic random-access memories (SDRAMs) using a two-layer baseboard, and the BiCoSS circuit uses a six-layer baseboard. The seven neural network units communicate with each other using two-layer connection boards. There are extra pins for each circuit of the BiCoSS platform that can be used for digital-to-analog (DA) convertor, camera, robot, and system expansion. BiCoSS uses the Intel EP4CE115 FPGA chip that contains 115 vertically arranged logic elements and 4 Mb of embedded memory. There are two configuration modes: the joint test action group (JTAG) and active serial (AS) modes. Two associated phase-locked loops are used to produce clock signals for synchronization among the processor nodes. General-purpose input/output (GPIO) communication ports are implemented for input and output devices, which are used for real-time applications, including pattern recognition and decision-making.

A. Architecture Design

BiCoSS consists of seven subsystems, referred to as the nuclei units. A previous study has revealed that the scalability of linear, mesh, and tree for a network-on-chip (NoC) architecture is n, n^2, and $\exp(n)$, respectively [19]. Therefore, these nuclei units are interconnected with each other based on a tree structure that significantly enhances BiCoSS scalability. Neurons are used as the basic computing cores, and SNNs are used as the primary computing units. Each SNN unit contains 16 or 64 cores, whose number is based on the model complexity constrained by the maximum computing power. As shown in Fig. 3, each nuclei unit contains four SNN units and one routing unit, which avoids the frequent cross-level communication between computing units induced by the simple use of tree topology, reducing unnecessary communication latency and resource consumption. The four SNN units in each nuclei unit are implemented based on a mixed interconnection architecture and are connected with each other using the routing unit. The parallel neural processors in each SNN unit are connected based on the butterfly fat tree (BFT) architecture to enhance the communication efficiency and scalability of the multicore system. The proposed hierarchical heterogeneous multicore architecture presents a scalable and feasible solution for the neuromorphic system.

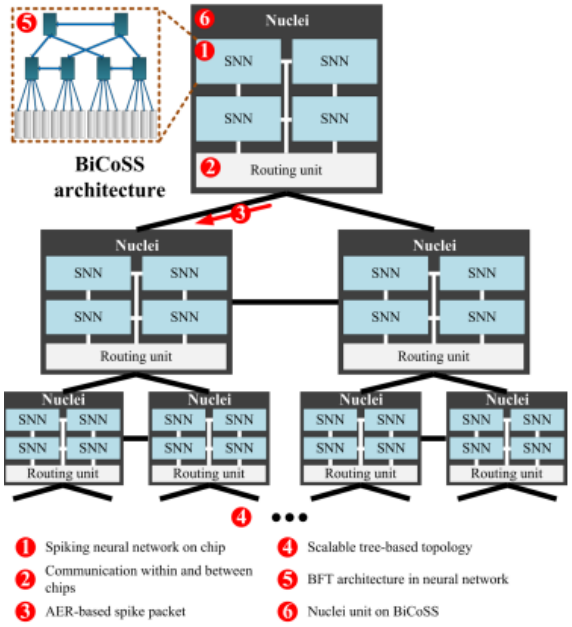


Fig. 3. System architecture of BiCoSS.

B. Network-on-Programmable-System (NoPS) Architecture on BiCoSS

In terms of the BiCoSS Network-on-Programmable-System (NoPS) realization, the BFT architecture is used for efficient communication within the SNN unit. Fig. 4(a) shows the operation of the neuromorphic BFT for the execution of an SNN model at algorithm time step Δt. Each computational core calculates its firing state that generates spike messages independently. It is connected to a special router, called a border router, which is responsible for regulating traffic across different BiCoSS nuclei units. The boarder router is the gateway for interlayer communications but in a different manner. The AER data packet contains five parts: AER data with 1 or 8 bits, chip address with 5 bits, layer address with 3 bits, node address with 4 bits, and timestamp with 6 bits.

Fig. 4(b) shows the overall NoPS architecture of a nuclei unit in the proposed BiCoSS digital neuromorphic topology. Four BFTs are connected with each other, with four root nodes each. In order to reduce network latency considering the hop counts, root nodes with the same number (i, ii, iii, iv, and v) are coupled together. From the neural routing perspective, BiCoSS is divided into four scopes of routing: system scope, nuclei scope, network scope, and population scope. The system scope consists of seven nuclei units, and the nucleus scope contains four SNN units. Both these two scopes use boarder routers. In each SNN unit, two network routers are used for network scope, which is based on BFT architecture. Four population routers are employed for population scope. The population scope contains four neural processors for neural information processing.

C. Architecture of Neural Processor

The neural processor generates spike events and receives AER spike packets from routers to update the spiking

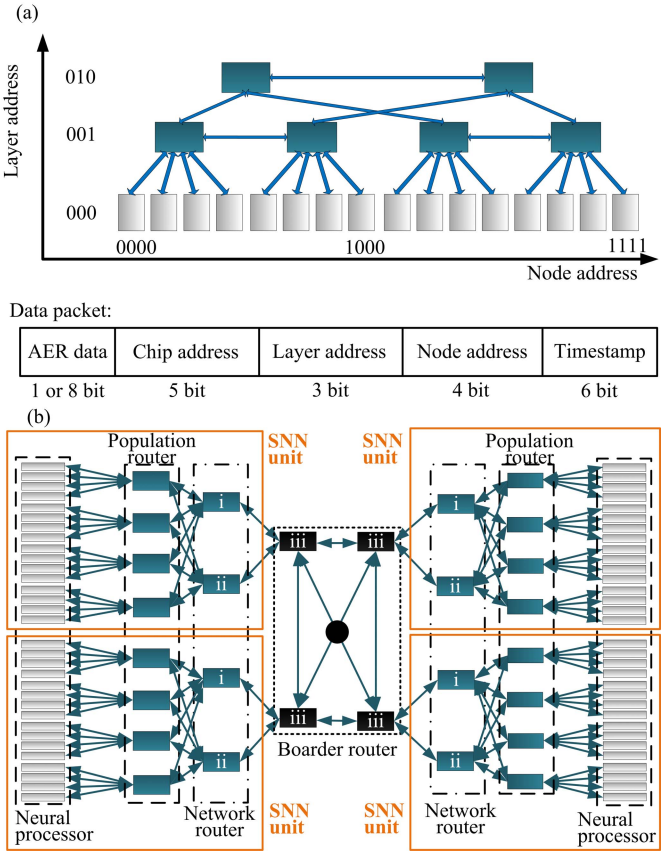


Fig. 4. NoC architecture on BiCoSS. (a) BFT topology for neural network unit of BiCoSS. (b) Logical diagram of four neural network units within a nuclei unit.

information. Time-multiplexing is used to implement the neuron model, and the architecture is shown in Fig. 5(a). According to the neuron model, pipelines are realized in the neuron unit, containing soma, dendrite, slow variable, and synaptic variable pipelines for the variables updating for the neuron model. The axon unit is used to store the variable values using dual-port block random access memories (RAMs) embedded on the FPGA. The spike-timing dependent plasticity (STDP) unit is implemented according to the STDP learning rule for updating the synaptic strength. The synapse computing unit (SCU) in the neural processor receives the AER spike packet from other routers and updated synaptic strength values from the STDP unit and outputs the synaptic current to the corresponding neural pipelines.

The schematic for the digital implementation of the LIF model in the cerebellar network is shown in Fig. 5(b), where all the multipliers are replaced with shifting, addition operations, and "shift MUL" module presented in our previous work [21]. Detailed implementation of the SCU is shown in Fig. 5(c), which includes parallel synaptic current processors. The multiplexer receives AER spike packets, with ports selected sequentially by a regular counter. AER spike packets are processed using decoders to extract the synaptic information data and timestamp. The timestamp data is used as the wri address, and a counter is used for control of the read address. The values of network connectivity C_{ij} are stored in

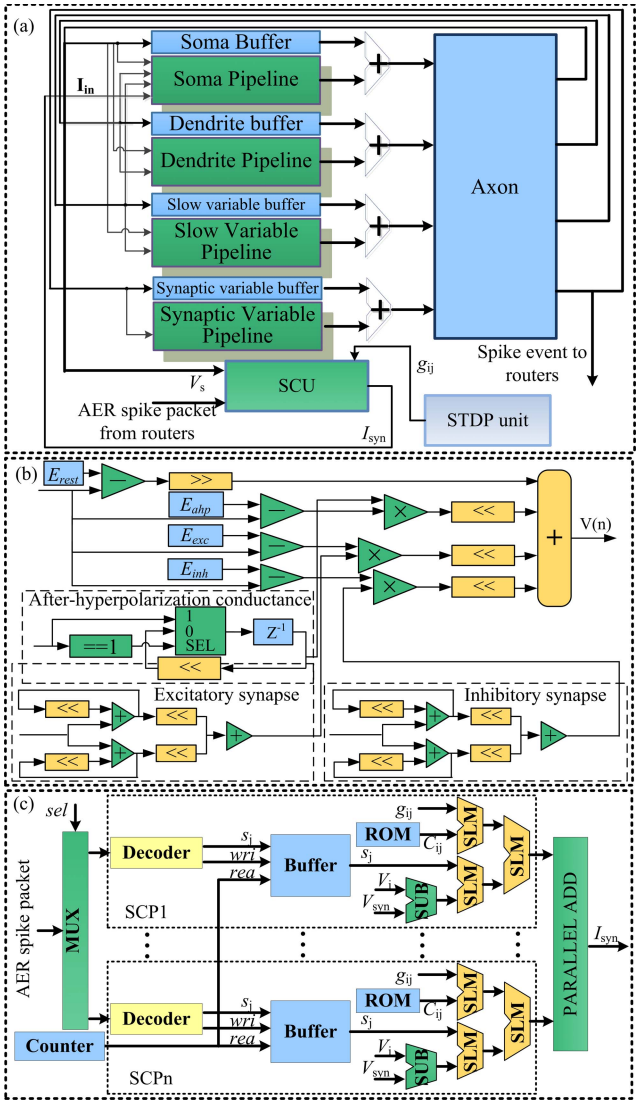


Fig. 5. Digital neuromorphic architecture of the neural processor. (a) Time-multiplexing pipeline architecture of the neural processor. (b) Digital implementation of the LIF neuron model. (c) Detailed digital implementation of the silicon synapse unit.

ROM, and a parallel adder is used to sum the output of the synaptic current. Then, in the next step, the synaptic current I_{syn} is output to the pipelines for neural state updating.

IV. Neural Information Routing on BiCoSS

A. Routing Algorithm of BiCoSS

The detailed routing algorithm in the router is shown in Fig. 6. Boarder and population routers, denoted by B and P, respectively, provide four levels of routing: system, nuclei, neural network, and population, with the exact topology shown in Figs. 3 and 4. Each population router is connected to four neuron units. P_s and P_d are population routers connected to the source and destination nodes, respectively, and *Bs* and *Bd* are corresponding boarder routers. The routing logic algorithm is implemented using the router in each neuron unit. A router contains only information that is related to the scope of the respective router. Different routing tables and their information

TABLE I
Different Routing Tables Based on Scopes

Population Routing Table	Table (Node Number, Link Number)
Network Routing Table	Table (Population Number, Link Number)
Boarder Routing Table	Table (Network Number, Link Number)
Nucleus Routing Table	Table (Boarder Number, Link Number)

Routing logic algorithm on BiCoSS

```
begin
  Nuclei, neural network, population for both source and destination;
  if Nucleuss = Nucleusd then
    if Boarders= Boarderd then
      if Networks = Networkd then
        if Populations = Populationd then
          Place flit into respective physical neuron;
        end
        else
          Reach Ns connected to Ps;
          Reach Pd, connected to Ns;
          Place the flit into respective physical neuron;
        end
      end
      else
        Reach Nd, connected to Ps;
        Reach Pd, connected to Nd;
        Place the flit into respective physical neuron;
      end
    end
    else
      Reach either of two Ns, connected to Ps;
      Reach respective Nd, connected to Ns;
      Reach Pd, connected to Nd;
      Place the flit into respective physical neuron;
    end
  end
  else
    Reach either of two Ns, connected to Ps;
    Reach Bs, connected to Ns;
    Reach NCs of the Nucleuss, connected to Bs;
    Reach NCd of the Nucleusd,, connected to NCs of the Nucleuss;
    Reach Bd, connected to NCd of the Nucleusd
    Reach any of the two Nd, connected to Bd;
    Reach Pd, connected to Nd;
    Place flit into respective physical neuron;
  end
end
```

Fig. 6. Pseudocode of the main routing methodology on BiCoSS.

content are summarized in Table I. Besides, a predetecting mechanism is used in BiCoSS routing to determine whether the event is spike-based or action-potential-based, which forms a synergistic routing scheme for hybrid neural information.

B. Digital Neuromorphic Implementation of BiCoSS Router

Six internal input ports are used in the first-level router to interface with the neighborhood routers or neural processor. As shown in Fig. 7(a), each input port is equipped with a virtual channel with two FIFOs. A counter is used to enable the sequential selection of the corresponding FIFO that is controlled by virtual channel arbiters in each virtual channel. The spike wrapper unit implements a packetization process at the beginning of spike information routing after the first-level router receives a spike event from a neural processor. The

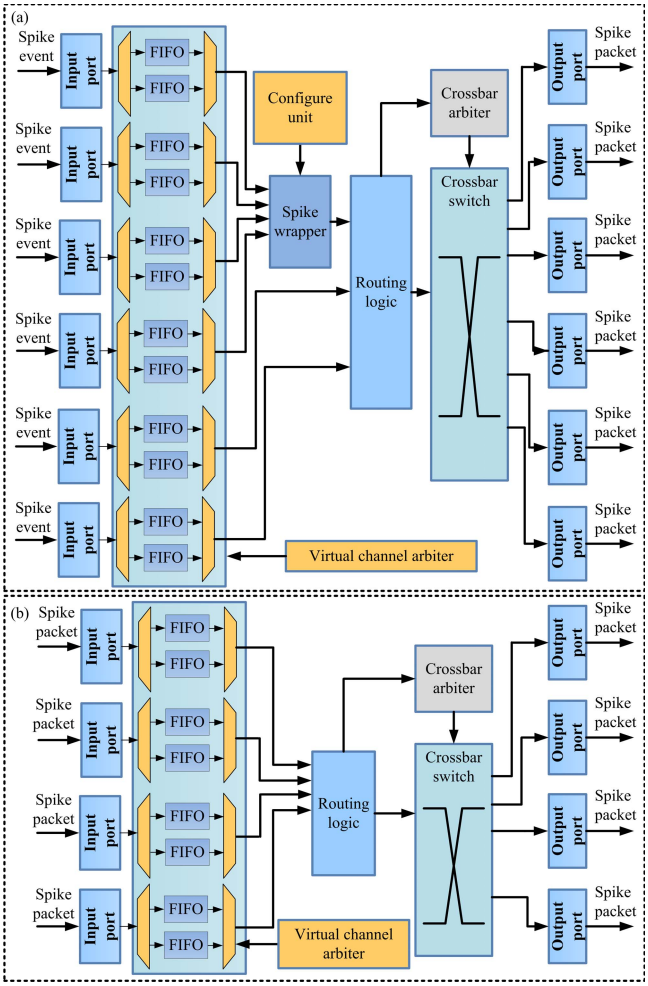


Fig. 7. Digital neuromorphic architecture of the BiCoSS router. (a) First-level router. (b) Second-level router.

spike wrapper unit is used to process spike events from a neural processor into an AER spike packet that contains 21 or 28 bits according to the 1-bit spike-based or 8-bit AP-based forms of the neural activity. The packet is composed of five parts: AER data, chip address, layer address, node address, and timestamp. The configure unit can be reconfigured at any time according to the neural connectivity specification that is required for the application to be processed, including the type of neural connectivity. The configure unit contains four types of registers: chip address register, layer address register, node address register, and timestamp register. The chip address register with 5-bit data and the layer address register with 3-bit data are used to select the correct chip and layer on each processor, respectively. The node address register and timestamp register store 6-bit data, respectively. Incoming spike events and the corresponding deliver-at timestamps are stored in memory until the deliver-at time is reached by the current global time. After the AER spike packet is made, a routing logic unit processes the packet according to the routing algorithm. Another important part of the first-level router is the crossbar switch that uses multiplexers controlled by signals from the crossbar arbiter according to the routing logic unit. The AER spike packets are then routed to the output ports with four spike events to the neuron units and two to the routers.

The detailed digital realization of the second-level router is shown in Fig. 7(b), which contains four internal input and output ports to interface with four other routers. The third-level router is connected with three other routers, so there are three input and output ports in each third-level router. The architecture of the third-level router is consistent with the second level. In this microcircuit architecture, the spike wrapper in the first-level router is removed because there is no spike event input from neuron units. The upstream and downstream data paths contain a 4×4 crossbar switch, providing the physical interconnection architecture to route a spike event packet to the output ports. If two spike events are transmitted to the same output port, one will be routed to the next output port to solve the packet communication conflict.

V. Performance Evaluation and Results

Large-scale neural modeling and simulation are characterized by high computational capability, communication, and power efficiency, as well as scalability and flexibility. It is beyond the ability of the most powerful computers based on the von Neumann architecture to even model a subsystem of the central nervous systems, let alone the entire mammalian brain. It would be possible to realize such large-scale neural networks using massively parallel systems. The large-scale SNN of the human brain requires a large number of processing cores, efficient computation and communications, high scalability and flexibility, and low energy consumption. The ambition of the BiCoSS system is to meet the requirements in a balanced situation in terms of hardware efficiency, scalability, flexibility, network scale, and biological plausibility.

A. Computation Efficiency and Accuracy

For the evaluation of computational efficiency, the hardware performance of BiCoSS is compared with three alternatives, which are CPU, GPU, and multicore bus systems. The cost function for the evaluation of computational efficiency is given by

$$Q = \tau_{\text{com}} / \tau_{\text{bio}} \tag{1}$$

where τ_{com} and τ_{bio} are the computational time by the computation system and the human brain biological system. The computational time of the biological system is determined based on the simulation results, and the time of the computational system is tested according to the time for the generation of the expected signals on the software/hardware platform. It is referred to as a small-world SNN, which contains four million LIF neuron models. In order to further evaluate the computational efficiency of BiCoSS, two alternative platforms, CPU and GPU, are used for the comparison, specifically the Intel Core 2 2.4-GHz CPU and NVIDIA GTX 280 GPU. As shown in Fig. 8(a), the computational efficiency of BiCoSS is highest along with the increment of network scale with the computational efficiency Q 5.4×10^5 times more than traditional CPU-based serial solutions. Profiting from the massively parallel computing characteristics and colocalized

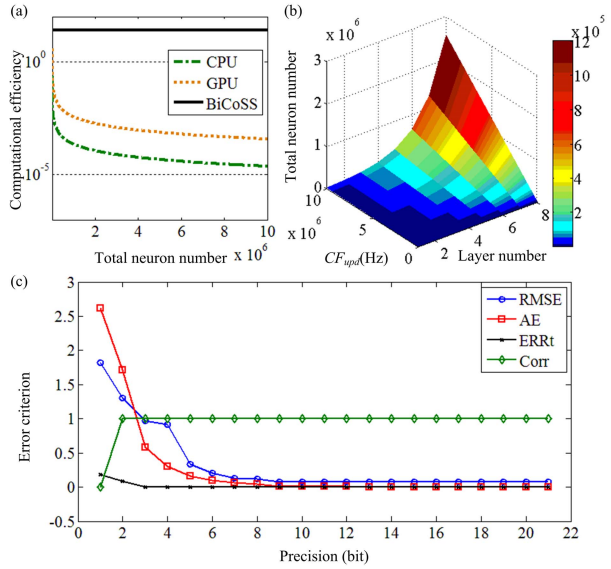


Fig. 8. Performance evaluation of BiCoSS. (a) Comparison of the computational efficiency with other alternative approaches. (b) Analysis of computational scale on BiCoSS. (c) Error evaluation of BiCoSS.

memory and computation features, BiCoSS has the potential to solve the von Neumann memory bottleneck problem. On-chip memory resource is used to realize the architecture with colocalized memory and computation, as shown in Fig. 5(a), which is a near-memory processing paradigm distinguished from the classical von Neumann architecture. Due to the hierarchical distributed multicore architecture with the non-von Neumann paradigm, the computational efficiency is enhanced significantly.

In order to explore the implementation of a large-scale neural network, the total available neuron number is investigated in terms of network updating efficiency $\mathrm{CF_{upd}}$ and layer number, which is shown in Fig. 8(b). The layer number determines the parallelism of the neuromorphic system. The updating efficiency $\mathrm{CF_{upd}}$ is determined by the multiplexed time and operating frequency, with BiCoSS parallelism determined by the layer number. Both the increment of layer number and enhancement of updating efficiency increase the total number of neurons that are available on BiCoSS.

In order to validate the proposed implementation, bit-level fixed-point simulation is compared with software-based simulation. For bit-level simulation, all hardware components were designed and realized based on the Very-High-Speed Integrated Circuit Hardware Description Language (VHDL) language. The root-mean-square error (RMSE), absolute error (AE), the error of spikes timing (ERRt), and correlation coefficient (Corr) are used as criteria for computational precision

$$\begin{cases} \mathrm{RSME} = \sqrt{\dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(x_{\mathrm{sof}}(i) - x_{\mathrm{har}}(i))^2} \\ \mathrm{AE} = \max|x_{\mathrm{sof}}(i) - x_{\mathrm{har}}(i)| \\ \mathrm{ERRt} = |(\Delta T_{\mathrm{har}} - \Delta T_{\mathrm{sof}})/\Delta T_{\mathrm{sof}}| \\ \mathrm{Corr} = \mathrm{cov}(x_{\mathrm{sof}}, x_{\mathrm{har}})/(\sigma(x_{\mathrm{sof}})\sigma(x_{\mathrm{har}})) \end{cases} \tag{2}$$

where $x_{\mathrm{sof}}(i)$ and $x_{\mathrm{har}}(i)$ are the values of software- and hardware-based computation results at the ith iteration. Variables ΔT_{har} and ΔT_{sof} represent the spiking time interval of the hardware and software computation. Corr is generally defined as the ratio of covariance to the product of the variances

$$\begin{cases} \mathrm{cov}(x_{\mathrm{sof}}, x_{\mathrm{har}}) = \displaystyle\sum_{i=1}^{n}(x_{\mathrm{sof}}(i) - \bar{x}_{\mathrm{sof}})(x_{\mathrm{har}}(i) - \bar{x}_{\mathrm{har}}) \\ \sigma(x) = \sqrt{\displaystyle\sum_{i=1}^{n}(x(i) - \bar{x})^2} \end{cases} \tag{3}$$

where $\bar{x}_{\mathrm{sof}}$ and $\bar{x}_{\mathrm{har}}$ represent the average values of $x_{\mathrm{sof}}(i)$ and $x_{\mathrm{har}}(i)$, respectively. The comparison results in Fig. 8(c) show that the proposed implementation obtains high computational precision. The main cause of any deviation is the fixed-point calculations of the FPGA implementation. The computational error can be further reduced by increasing the number of bits in the proposed system or adjusting the parameter values of the large-scale biologically meaningful network.

B. Communication and Power Efficiency

The BiCoSS average latency and average acceptance rate (AAR) in Fig. 9 show comparisons between BiCoSS and other NoC architectures, including mesh, torus, butterfly, and flattened butterfly. The average BiCoSS latency is considerably lower than mesh with an enhancement of 43%–89%. In comparison with the torus, BiCoSS has an 83%–88% improvement in average latency and 6%–9% in AAR. In addition, BiCoSS outperforms 31%–95% improvement in average latency and 1%–8% improvement in AAR in comparison with flattened butterfly.

Fig. 10 shows comparative results between mesh-based SpiNNaker and BFT-based BiCoSS with IP numbers of 16, 64, and 256. The BiCoSS topology is superior to SpiNNaker in terms of router number, network throughput, packet data latency, link utilization, router input, and the output buffer. However, the router node degree of BiCoSS is larger, which means that a more complicated routing algorithm may be required, inducing higher resource cost for the router.

The BiCoSS platform is equipped with 35 Intel EP4CE115 FPGA chips, which dissipates 10.419 W. In terms of power efficiency, the power density of the BiCoSS platform is 35.4 mW/cm^2, whereas that of a typical central processing unit (CPU) is 50–100 W/cm^2. The power consumption of the graphical processing unit (GPU) is 50–300 W, with a power density of 10–100 W/cm^2. The digital neuromorphic SpiNNaker board with 49 Advanced RISC Machine (ARM) chips consumes 62.5 W, with a power density of 360.1 mW/cm^2.

C. Cognitive Functions of the Autonomous Cognitive Brain on BiCoSS

BiCoSS can realize multigranule SNN models with different network topologies, including feed-forward and recurrent networks. The large-scale SNNs on BiCoSS can be capable to reproduce biological activities related to cognitive behaviors in the human brain. In order to further reveal mechanisms responsible for human cognition and analyze mechanisms from the neuron level, BiCoSS further

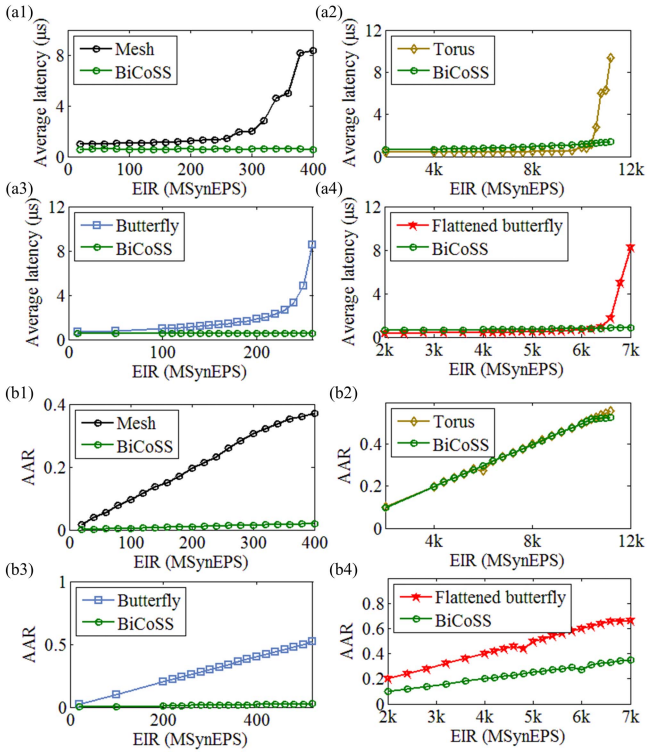


Fig. 9. Comparison of the on-chip performance between BiCoSS and other NoC architectures. (a) Comparison of the average latency between BiCoSS and other NoC architectures. (b) Comparison of the AAR between BiCoSS and other NoC architectures.

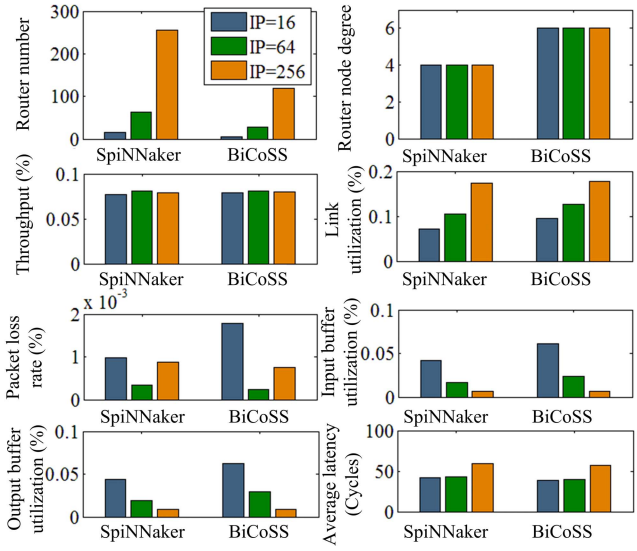


Fig. 10. Comparison of the performance between SpiNNaker and BiCoSS.

explores cognitive principles in decision-making, motor learning, context-dependent learning, and TH relay disability with movement disorders using large-scale SNNs with four million neurons.

1) Cerebellar Motor Learning: To explore the dynamic response of the granule (GR) layer in the cerebellum, BiCoSS is used for a cerebellar motor learning task based on the model of the previous study [31]. Two different frequencies

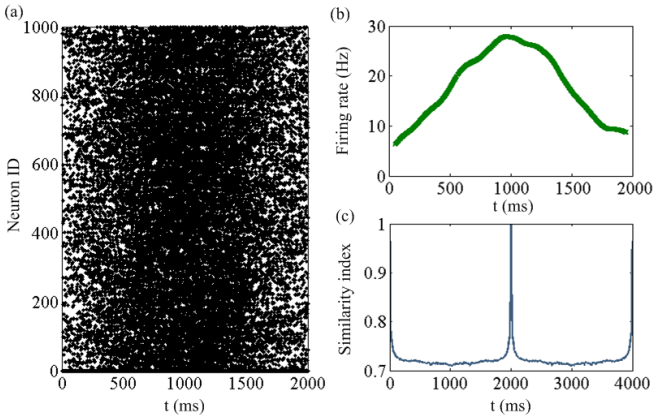


Fig. 11. Dynamics of cerebellar GR cells in response to sinusoidally oscillating MF signals at 0.5 Hz. (a) Spike patterns of 1000 out of 100k GR cells during a cycle of signal oscillation within MFs, where black dots represent spike firings. (b) Firing rate of GR cells during motor control. (c) Similarity index for the spike patterns in the cerebellar network.

of the input signal through MF are given, and the spiking behaviors of both the GrCs population and different cell groups are observed. First, we give 30-Hz sinusoidal Poisson spikes, and the dynamic responses are shown in Fig. 11. Fig. 11(a) shows the spike patterns of 1000 GrCs randomly chosen from the BiCoSS cerebellum. GrCs emit spikes sparsely and randomly with increasing firing activity, which fits the stimulation intensity. The corresponding average firing rate of the GrCs is depicted in Fig. 11(b). The result shows that the GrCs can transmit amplitude and time information of the input signals to PCs, which is consistent with previous findings [32]. To confirm that the population of active GrCs gradually changes with time, the similarity index between active GR-cell populations is calculated based on the following equation:

$$\begin{cases} S(\Delta t) = \dfrac{1}{T}\displaystyle\sum_{t=0}^{T} C(t, t+\Delta t) \\ C(t, t+\Delta t) = \dfrac{\sum_i z_i(t) z_i(t+\Delta t)}{\sqrt{\sum_i z_i^2(t)}\sqrt{\sum_i z_i^2(t+\Delta t)}} \\ z_i(t) = \dfrac{1}{\tau}\displaystyle\sum_{s=0}^{t} \exp(-(t-s)/\tau)\left(\dfrac{1}{N_c}\sum_{j=1}^{N_c}\delta_{ij}(s)\right) \end{cases} \tag{4}$$

where N_c is the number of GrCs, and $\delta_{i,j}(t) = 1$ when GrC j in the ith cluster spikes at time t and equals 0 otherwise. The parameter $\tau = 8.3$ ms represents the time constant of alpha-amino-3-hydroxy-5-methyl-4-isoxazole propionic acid receptors (AMPAR)-mediated excitatory post synaptic potential (EPSPs) at parallel fiber-PC synapses. The variable $z_i(t)$ represents the AMPAR-mediated EPSPs at a PC evoked by the ith GR-cell cluster at t, and $C(t, t + \Delta t)$ defines the autocorrelation of the activity pattern at time t and $t + \Delta t$. The temporal changes are not time-dependent, meaning that the one-to-one correspondence between active GR-cell populations can be generated from the beginning of a cycle of MF signal oscillations.

Fig. 12 shows the comparison of the interspike interval (ISI) distributions based on the software simulation and BiCoSS. There are four major types of neurons in the cerebellar network

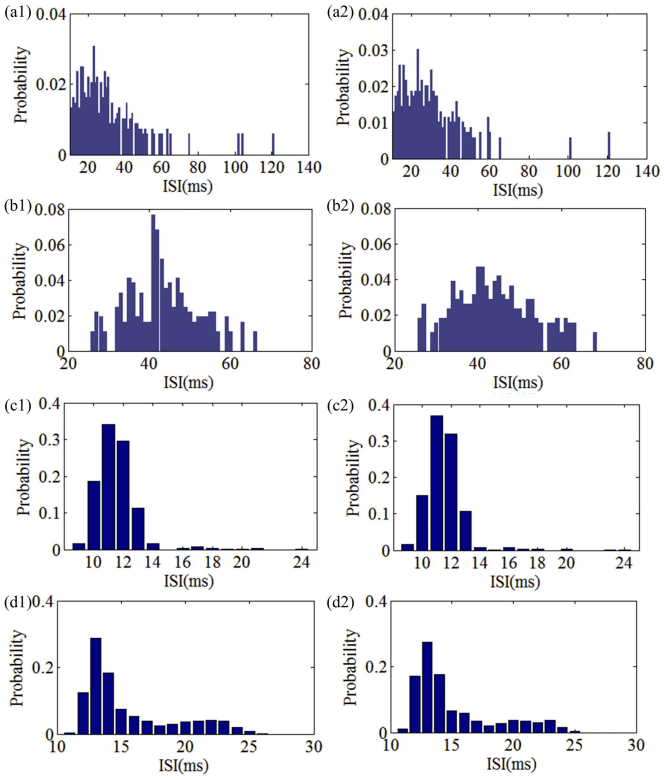


Fig. 12. Network dynamics comparison between software and hardware. (a1) ISI distributions of the GrC neurons on software. (a2) ISI distributions of the GrC neurons on BiCoSS. (b1) ISI distributions of the GoC neurons on software. (b2) ISI distributions of the GoC neurons on BiCoSS. (c1) ISI distributions of the PKJ neurons on software. (c2) ISI distributions of the PKJ neurons on BiCoSS. (d1) ISI distributions of the BS neurons on software. (d2) ISI distributions of the BS neurons on BiCoSS.

considered in the dynamics' comparison, which are GrC, GoC, Purkinje (PKJ), and BS. As shown in Fig. 12(a), the temporal positions of the peaks of the curves are 23 ms for the GrC neurons on both software and BiCoSS. The peaks of the ISI distributions are 41 and 40.5 ms for GoC neurons on software and BiCoSS, respectively, as shown in Fig. 12(b1) and (b2). As shown in Fig. 12(c) and (d), the peak values of the ISI distributions for PKJ neurons are both 11 ms, and both the peak values for BS neurons on the two platforms are 13 ms. It shows that good consistency is obtained between the two implementations based on the computation of BiCoSS.

2) Decision-Making of BG With Reinforcement Learning: In order to make decisions, the human brain evaluates all available options, compares options with the current aim, and selects the most rewarding one. The BG is considered a critical neural substrate for decision-making. Decision-making can be divided into three types: "explore," "exploit," or "take no action." In this study, we apply BiCoSS in a binary action selection task [27]. The model details of human decision-making are introduced in previous work [28].

Under high DA levels, the activity of the D1 stratum will be increased, and the direct pathway will dominate the indirect pathway. A stronger input is chosen because it can reach the threshold in a shorter time. The raster plots of STN, GPe, and GPi neurons randomly chosen from the large-scale network on BiCoSS are shown in Fig. 13(a)–(c), respectively. Therefore,

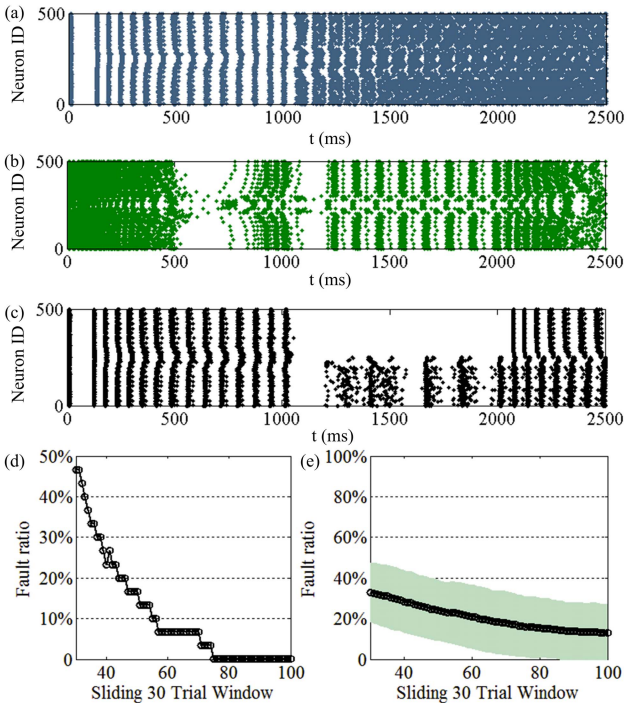


Fig. 13. Decision-making and learning of the context-dependent task. (a) Raster plot of STN neurons with "Go" strategy on BiCoSS. (b) Raster plot of GPe neurons with "Go" strategy on BiCoSS. (c) Raster plot of GPi neurons with "Go" strategy on BiCoSS. (d) Fault ratio of hippocampal learning for a single run on BiCoSS. (e) Fault ratio of hippocampal learning for 100 runs on BiCoSS.

the "Go" strategy of human action occurs at the higher DA level, and the experimental results are consistent with previous biophysical experiments [29], [30].

3) Learning of Context-Dependent Task in Hippocampus: We built a hippocampal SNN that can learn context-specific rules based on an efficient internal representation according to the model in [33]. The hippocampal SNN Fig. 13(d) shows the BiCoSS performance during the task learning, which represents the mean value of the fault ratio during learning. At the beginning of training, the network is unable to achieve the goal accurately; therefore, the error rate is rather high. As the number of trials increases, the output of the network is more accurate because, first, the network does not have the memory of how to finish the task, but, in the end, the network is able to get learn according to the STDP rule. Fig. 13(e) shows the performance of fault ratio during learning when we run the network 100 times. After several trials, the fault ratio of the network shows a downward trend. With the trial is ongoing, the fault ratio finally reaches zero, which realizes the learning ability of the hippocampus and substantiates the validity of the network.

4) Learning With and Beyond Pair-Based STDP Mechanism: Neuromorphic systems are good candidates for the artificial intelligence application due to their computational power and efficiency. SNNs have been investigated because of their energy-efficient advantage and closer mechanism to the human brain. SNN models are also the basic components of the neuromorphic systems, with STDP learning rules for online learning. In this study, the unsupervised learning capability of BiCoSS is further tested by the Mixed National

Institute of Standards and Technology (MNIST) data set and compared with other neuromorphic systems. The unsupervised learning algorithm presented by Diehl and Cook is employed with pair-based and triplet STDP learning rules, respectively, [48]. In fact, several neuromorphic studies have used the MNIST data set to test the learning performance previously. Wang *et al.* [49] presented an FPGA-based neuromorphic implementation method for the SNN model by Diehl and Cook, which achieved 89.1% accuracy with a total of 1591 neurons. Buhler *et al.* [50] presented an on-chip trainable SNN work with an accuracy of 88% on MNIST digits. In addition, neuromorphic hardware by Kim *et al.* [51] achieved 84% accuracy with 256 neurons. Due to the large-scale computational capability of BiCoSS, 94.6% and 94.8% accuracy are achieved with pair-based and triplet STDP, respectively. Diehl and Cook have pointed out that SNN models with a large number of neurons can obtain 95% accuracy using pair-based and triplet STDP learning rules. This is because of the hardware-based operating on the fixed-point arithmetic on BiCoSS.

Previously kinds of neuromorphic systems have been presented with different advantages and features [7], [41]. Another distinguished contribution of the presented study is the integration of various forms of spike-driven learning rules, which can bridge the gap between both brain-inspired intelligence and computational neuroscience.

The experimental results have shown that the STDP learning rule is not the only learning rule available to biological neural systems. BiCoSS provides an essential platform to implement multiple synaptic plasticity mechanisms to coexist in a single multicompartment neuron. This mechanism is closely related to the associative learning and memory retention of human cognition. The STDP learning rule is configured across the whole-basal dendritic tree, and long-term potentiation (dLTP) is configured about half of the basal dendritic tree. It allows for potentiation even when the network inputs cannot evoke APs and needs the functionally related synapses in nearby locations [52]. The neuromorphic network is the connections that are randomly distributed across distal and proximal compartments. The computational model is proposed in the previous study by Bono and Clopath [53].

Since the neurons from these different features are never activated together, proximal weights between different features are weakened. Item A is characterized by features I, II, and IV, while item B is characterized by features I, II, and III. Each learning procedure contains 300 ms, in which item B is activated in the first and last 100 ms, and item A will be activated in the middle 100 ms. As shown in Fig. 14(a), the proximal weight is cut down between the neurons with the same characteristics, and the distal weight is not significantly reduced between neurons with different features in comparison with proximal synapses. Suppose two kinds of items A and B to be learned, in which features I and II are the shared feature for these two items. Features III and IV are the distinguished features for the two items. Fig. 14(b) shows that the distal and proximal synaptic weights remain high, which causes that the neural connections are retained. As shown in Fig. 14(c), when item B is activated, the features II and III are activated, which induces that the proximal and distal connections are enhanced.

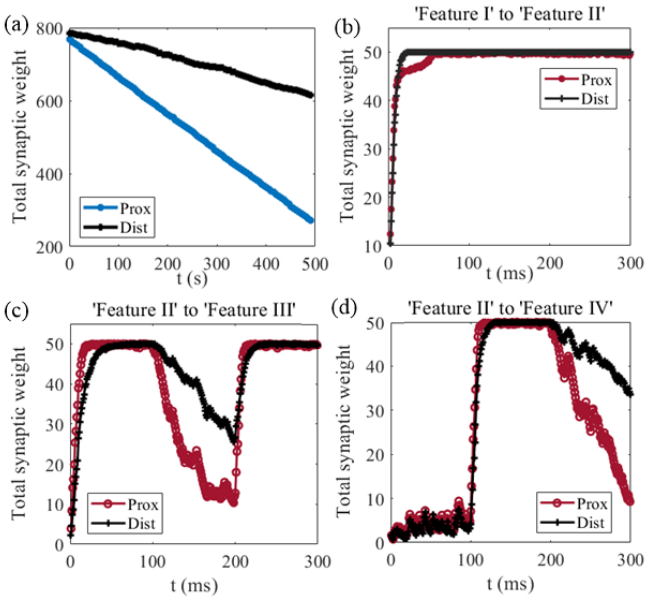


Fig. 14. Associative learning with hybrid learning mechanisms and dendritic processing on BiCoSS. (a) Evolution of total synaptic weight along with time; (b) Evolution of total synaptic weight with connection from "Feature I" to "Feature II" along with time; (c) Evolution of total synaptic weight with connection from "Feature II" to "Feature III" along with time; (d) Evolution of total synaptic weight with connection from "Feature III" to "Feature IV" along with time.

When item A is activated, feature III is not activated, which induces that the neural proximal connection with feature III is reduced and the distal connection is retained. When item A is activated in Fig. 14(d), the neurons representing features II and IV are activated, which means that the proximal and distal connections are enhanced for the neurons with these two features. The neurons representing feature IV do not respond, which induces the proximal connections are reduced and the distal connections are retained. Besides, we find that the distal weights between neurons of different features do not depress substantially compared with the proximal weights. It reveals how ongoing activity affects memory retention in neuromorphic networks.

5) Movement Disorders: The TH is the crucial gateway to the neocortex in the sense that no sensory signal, such as vision or taste, reaches the neocortex without going through a proper thalamic nucleus [34], [35]. The network structure is based on a previous study on movement disorders [36]. In the state of movement disorders, the relay capacity of the thalamocortical (TH) neurons is disturbed by activities of rebound bursting. An increasing value of the synaptic conductance g_{inc} is defined as follows:

$$\begin{cases} g_{\text{STN}\to\text{GPe}} = 0.075 + g_{\text{inc}} \\ g_{\text{GPe}\to\text{GPe}} = 0.025 + g_{\text{inc}} \\ g_{\text{GPe}\to\text{GPi}} = 0.015 + g_{\text{inc}} \\ g_{\text{STN}\to\text{GPi}} = 0.075 + g_{\text{inc}} \\ g_{\text{GPe}\to\text{STN}} = 0.01 + 5 * g_{\text{inc}} \\ g_{\text{GPi}\to\text{STN}} = 0.01 + 5 * g_{\text{inc}} \end{cases} \tag{5}$$

where the synaptic conductances $g_{\text{STN}\to\text{GPe}}$, $g_{\text{GPe}\to\text{GPe}}$, $g_{\text{GPe}\to\text{GPi}}$, $g_{\text{STN}\to\text{GPi}}$, $g_{\text{GPe}\to\text{STN}}$, and $g_{\text{GPi}\to\text{STN}}$ represent the

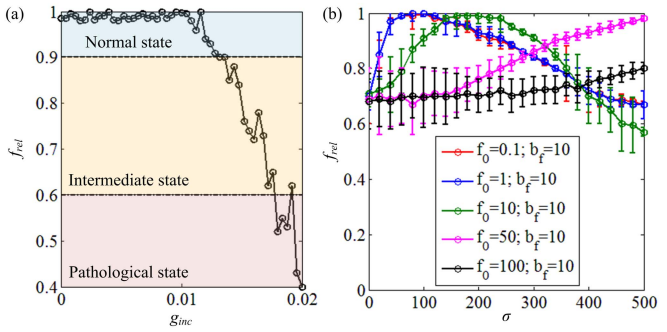


Fig. 15. Investigation of the relay ability of TH cell. (a) Impact of synaptic conductance. (b) Impact of noise.

coupling strengths for each type of synapses. In order to quantitatively evaluate the performance under different conditions, a relay reliability index f_{rel} is defined as follows:

$$f_{\text{rel}} = 1 - \frac{n_{\text{fau}}}{n_{\text{suc}} + n_{\text{fau}}} \tag{6}$$

where n_{fau} represents the fault spikes that have not been relayed, and n_{suc} represents the spikes that have relayed the neural information successfully. As shown in Fig. 15(a), in the normal state, the value approaches 1, and it is below 0.6 under the pathological state. As the coupling strength grows larger, the value of x_{rel} decreases, inducing the movement disorders with large synaptic coupling strength.

In addition, the effects of the noise stimulation on the movement disorders are explored on BiCoSS. Since frequency and intensity are the primary properties of the noise stimulation, the respective roles of the frequency and intensity parameters of the noise stimulation are investigated on BiCoSS, including the bandwidth b_f, the initial frequency f_0, and the noise intensity σ. As shown in Fig. 15(b), the values of f_{rel} always first increase and decrease with the increasing of σ. It reveals that there is an optimal noise intensity value for the maximum f_{rel} with a certain f_0. Besides, the optimal noise intensity increases with f_0 increasing, which reveals that higher frequency requires a larger intensity for the noise stimulation to obtain high reliability.

VI. Discussion

A critical path to achieving further comprehension of brain cognition is the development of large-scale, biologically realistic models of the brain based on a neuromorphic system. Currently, we are at the stage of realizing and further investigating large-scale brain models for different cognition tasks and pathological states. In this study, we have presented the BiCoSS cognitive brain and a neuromorphic platform developed for SNN models of the full brain. Its scientific core has been developed by concepts from computational and theoretical cognitive neuroscience, with the ambition of bridging the gap between low-level microscopic biological dynamics of a single neuron and high-level macroscopic cognitive behaviors. From a computational neuroscience perspective, BiCoSS lays the groundwork for paradigms in large-scale modeling of the human brain, by presenting a real-time powerful framework based on digital neuromorphic techniques. From a brain-like intelligence perspective, this work aims to investigate the dynamics involved in the emergence of biological intelligence, with the objective of comprehending how and why many different ion channels, neurons, synapses, networks, and nuclei in the human brain form various kinds of brain functions and intelligent cognitive behaviors. BiCoSS presents novel design philosophy and methodology for neuromorphic modeling and novel solutions for brain-inspired cognitive intelligence, which has both more powerful performance and more possibilities to generate cognition and learning behaviors. In this respect, it shows a strong contribution in the following.

1) A novel computational paradigm for the large-scale functioning brain, which provides a new perspective for the development of the large-scale cognitive brain model with the capability of real-time response to the real world.
2) A novel solution comprising various types of learning algorithms, which can produce more powerful learning capability. The learning capability on BiCoSS for the digit recognition of the MNIST data set is tested and is compared with other neuromorphic systems using MNIST as testing data set. In addition, hybrid learning mechanisms with dendritic processing are realized on BiCoSS, which results in the associative learning capability, as shown in Fig. 14. BiCoSS presents an opportunity to integrate hybrid learning rules on a unique neuromorphic system.
3) A novel neuromorphic method with multigranular neuron models, which can generate cognition behaviors across different brain regions. It can reproduce the network-level dynamical activities of different brain regions including cerebellum, BG, TH, and hippocampus, as shown in Figs. 11–13 and 15. The cognitive function refers to the cerebellar supervised learning, reinforcement learning of BG, and the context-dependent learning of the hippocampus.
4) A novel hierarchical heterogeneous multicore architecture is presented, which makes BiCoSS scalable and enhances its communication efficiency. The presented architecture is shown in Figs. 3 and 4, and the corresponding routing algorithm and router architecture are depicted in Figs. 6 and 7, respectively. Higher computational performances are achieved compared with the alternative computing platforms or neuromorphic systems, as shown in Figs. 8(a)–10. The computational power and scalability are superior to the state-of-the-art neuromorphic works, as shown in Table II. The powerful computational capability enables BiCoSS an expected candidate to build a neuromorphic brain with hybrid learning mechanisms and cognitive activities.

A. Comparison With State-of-the-Art Neuromorphic Approaches

Previous articles on neuromorphic systems have presented neuromorphic systems for large-scale simulation of the brain [2], [10], [17]–[20], [37]–[41]. A comparison between BiCoSS and other major neuromorphic systems, considering

TABLE II
REPRESENTATIVE OF THE STATE-OF-THE-ART NEUROMORPHIC SYSTEMS

Project	Method	Model	Learning	Scale	Scalability
BrainScaleS	Analog	AdEx IF	STDP	4M	N^2
Truenorth	Digital	LIF	None	1M	N^2
Neurogrid	Mixed	QIF	None	1M	2^N
SpiNNaker	Digital	Any type	STDP	1B	N^2
LaCSNN	Digital	H-H type	STDP	1M	N^2
ROLLS	Analog	AdEx IF	STDP	256	N^2
BlueHive	Digital	Izhikevich	None	256k	2^N
IFAT	Analog	LIF	None	65k	2^N
HiAER	Mixed	LIF	None	1M	2^N
SiElegans	Digital	LIF; H-H type	None	330	--
Loihi	Digital	LIF	STDP	131k	N^2
Dynap-SEL	Mixed	LIF	Hebbian-like	1k	N^2
BiCoSS	Digital	Any type	D-STDP; T-STDP; dLTP	4M	2^{N+2}

network scale, learning mechanisms, cognitive tasks, programmability, and computational speed, is provided in Table II. The scalability means an opportunity to scale up to the maximum number of computational nodes. As for the cognitive functions, only about half of these neuromorphic systems involve cognitive functions in the human brain. The types of hardware that are used determine programmability and simulation speed; analog circuits are inflexible and not programmable limiting their application in reconfigurable neuromorphic design. Digital systems require more power consumption; however, they are more flexible with higher computational precision.

In terms of biological plausibility, different types of neuron models can be realized, and both D-STDP and T-STDP are considered in the learning mechanism. As to network scale, four million large-scale SNNs can be emulated in real time, which can be further enhanced by cascading a number of BiCoSS boards. Previous neuromorphic systems use various kinds of synaptic connection topologies, including linear topology in Neurogrid [18] and mesh topology in BrainScaleS, Truenorth, and SpiNNaker [2], [17], [20]. By contrast, BiCoSS focuses on combining the advantages of high flexibility and high expandability by embedding random access addressing at all levels of scale in the BFT-based connection hierarchy. As shown in Table II, BiCoSS uses a tree-based topology that serves as a communication backbone to an event-driven SNN system. For a given number of hops N, the expandability of BiCoSS is $4e^N$, which is superior to other platforms. As shown in Fig. 7, for communication efficiency, the BFT topology used outperforms current state-of-the-art neuromorphic approaches, such as SpiNNaker. The BiCoSS system can also be scaled up using multiple BiCoSS boards without performance loss by building on its modular architecture and hierarchical communication scheme. Therefore, in comparison with current major projects, BiCoSS is a large-scale neuromorphic system with superior combined advantages considering network scale, learning mechanisms, cognitive tasks, programmability, biological plausibility, and expandability.

B. Toward Reverse Engineering the Cognitive Brain Using BiCoSS

This work employs the neuromorphic approach to bridge the gap between single neuron dynamics and high-level mammalian behaviors based on the bottom-up scheme. Neuromorphic engineering as a vital approach for reverse engineering of the cognitive brain is discussed in [42]. BiCoSS has successfully demonstrated that simple behaviors can be implemented in a real-time digital neuromorphic system composed of asynchronously communicating spiking neurons. The proposed neuromorphic approach is sufficiently general to be used on a wide range of SNNs or deep neural networks that have reconfigurable synaptic weights and reprogrammable connectivity and is vital for large-scale emulation of SNNs. Another critical application of BiCoSS is for brain–machine interface and bio-hybrid experiments. Recent studies have successfully realized the neuromorphic prosthesis by neuromorphic hardware [54]–[57]. The neuromorphic-based system opens avenues for the exploitation of neuroprosthetic devices in bioelectrical therapeutics for health care and brain–machine integration. Since BiCoSS has the capability to generate biologically plausible dynamics of different brain regions, it can be a satisfied candidate to be coupled with biological networks for the real-time brain–machine or neuron–machine interfaces. In the future study, we will further explore the capability of BiCoSS as a real-time neuroprosthetic *in silico* that can be coupled with the biological neural network.

While we believe that neuromorphic systems, such as BiCoSS, are moving us toward a better understanding of brain dynamics and cognitive functions, there remain many challenges ahead for reverse engineering of the human brain. Modeling overall brain cognition is not the aim of the current study, and the interconnections between each brain area and specified cognitive functions are not addressed. Future work involves integrating separated brain areas to generate multimodal intelligence. There are some studies and ideas that can be employed on BiCoSS for this ambition, including Spaun, which has successfully solved the problem of building a large-scale model with the capability of solving cognitive tasks in a comparable way to how humans do the same tasks [1]. The semantic pointer architecture proposed in Spaun is hypothesized as a model for the organization, function, and representational resources in the mammalian brain. In addition, Marbleston *et al.* [43] present an idea to integrate deep learning with neuroscience that proposes a large architecture containing several specialized systems. They use structured architectures, including detailed systems for attention, recursion, and memory storage. Cost functions and training are included, and they hypothesize the brain can optimize these cost functions, which can be helpful for the design and development of BiCoSS in the future. Besides, more biologically plausible spiking network models for the functioning nucleus are required, including biologically meaningful cortex and functional nucleus models [44]–[47]. By introducing more biological mechanisms in BiCoSS, the gap across spatial scales between neural dynamics and human behaviors can be bridged, and the underlying principles can eventually be unraveled.

VII. CONCLUSION

In this study, we propose a digital neuromorphic system BiCoSS, which is inspired by neurophysiology and neuroscience and realized by a reconfigurable neuromorphic

computational engine, with the goal of simulating microscopic neural dynamics on large-scale brain networks. BiCoSS builds brain models at the level of neurons, emphasizing the large-scale network characteristics and considering the relationship between the function of the brain region and neuronal dynamics within each region, which could be further related to behavior, physiology, and neurological diseases. BiCoSS uses FPGA chips to simulate four million spiking neurons in real time. It is also scalable to cascade several BiCoSS boards to realize larger SNNs with more memory resources. Compared with current major neuromorphic systems, BiCoSS has superior combined advantages considering network scale, learning mechanisms, cognitive tasks, programmability, biological plausibility, and expandability. The proposed BiCoSS system offers an affordable and scalable approach toward bridging the gap from the cellular dynamics level to explore cognition functions of the human brain and is meaningful for research and applications of neural cognition and brain-like computing.

APPENDIX

TABLE III
DESCRIPTION OF NEURAL NETWORK MODELS

Brain region	Model	Learning	Nuclei	Cognition mechanisms
Cerebellum	LIF	STDP	MF	Motor learning
			DCN	
			GrCs	
			GoCs	
			PCs	
			IO	
BG	Izhikevich	STDP	STN	Decision making
			GPe	
			GPi	
			Str	
Hippocampus	LIF	STDP	CA1	Context-dependent learning
CBT	H-H type	STDP	GPe	Movement disorders
			GPi	
			STN	
			TC	
			Sensorimotor	
Hippocampus	Izhikevich	No	CA3	Neuromodulation

ACKNOWLEDGMENT

The authors would like to thank the editor and the reviewers for their critical and constructive comments and suggestions.

REFERENCES

[1] C. Eliasmith *et al.*, "A large-scale model of the functioning brain," *Science*, vol. 338, no. 6111, pp. 1202–1205, Nov. 2012.

[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] E. M. Izhikevich 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.

[4] 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.

[5] M. Breakspear, "Dynamic models of large-scale brain activity," *Nature Neurosci.*, vol. 20, no. 3, pp. 340–352, Feb. 2017.

[6] M. Djurfeldt *et al.*, "Large-scale modeling-a tool for conquering the complexity of the brain," *Frontiers Neuroinform.*, vol. 2, p. 1, Apr. 2008.

[7] A. Neckar *et al.*, "Braindrop: A mixed-signal neuromorphic architecture with a dynamical systems-based programming model," *Proc. IEEE*, vol. 107, no. 1, pp. 144–164, Jan. 2019.

[8] S. Yang *et al.*, "Real-time neuromorphic system for large-scale conductance-based spiking neural networks," *IEEE Trans. Cybern.*, vol. 49, no. 7, pp. 2490–2503, Jul. 2019.

[9] E. Chicca, F. Stefanini, C. Bartolozzi, and G. Indiveri, "Neuromorphic electronic circuits for building autonomous cognitive systems," *Proc. IEEE*, vol. 102, no. 9, pp. 1367–1388, Sep. 2014.

[10] N. Qiao *et al.*, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Frontiers Neurosci.*, vol. 9, p. 141, Apr. 2015.

[11] C. S. Thakur *et al.*, "Large-scale neuromorphic spiking array processors: A quest to mimic the brain," *Frontiers Neurosci.*, vol. 12, p. 891, Dec. 2018.

[12] S. Yang *et al.*, "Scalable digital neuromorphic architecture for large-scale biophysically meaningful neural network with multi-compartment neurons," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 1, pp. 148–162, Jan. 2020.

[13] G. Indiveri and S.-C. Liu, "Memory and information processing in neuromorphic systems," *Proc. IEEE*, vol. 103, no. 8, pp. 1379–1397, Aug. 2015.

[14] A. V. Herz, T. Gollisch, C. K. Machens, and D. Jaeger, "Modeling single-neuron dynamics and computations: A balance of detail and abstraction," *Science*, vol. 314, no. 5796, pp. 80–85, Oct. 2006.

[15] E. Marder and A. L. Taylor, "Multiple models to capture the variability in biological neurons and networks," *Nature Neurosci.*, vol. 14, no. 2, pp. 133–138, Feb. 2011.

[16] C. Zednik, "Models and mechanisms in network neuroscience," *Phil. Psychol.*, vol. 32, no. 1, pp. 23–51, Jan. 2019.

[17] M. A. Petrovici *et al.*, "Characterization and compensation of network-level anomalies in mixed-signal neuromorphic modeling platforms," *PLoS ONE*, vol. 9, no. 10, Oct. 2014, Art. no. e108590.

[18] 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.

[19] 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.

[20] 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.

[21] J. V. Arthur *et al.*, "Building block of a programmable neuromorphic substrate: A digital neurosynaptic core," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jun. 2012, pp. 1–8.

[22] S. Yang *et al.*, "Cost-efficient FPGA implementation of a biologically plausible dopamine neural network and its application," *Neurocomputing*, vol. 314, pp. 394–408, Nov. 2018.

[23] R. Wang, G. Cohen, K. M. Stiefel, T. J. Hamilton, J. Tapson, and A. van Schaik, "An FPGA implementation of a polychronous spiking neural network with delay adaptation," *Frontiers Neurosci.*, vol. 7, p. 14, Feb. 2013.

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

[25] F. Khoyratee, F. Grassia, S. Saïghi, and T. Levi, "Optimized real-time biomimetic neural network on FPGA for bio-hybridization," *Frontiers Neurosci.*, vol. 13, p. 377, Apr. 2019.

[26] N. Qiao *et al.*, "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Frontiers Neurosci.*, vol. 9, p. 141, Apr. 2015.

[27] M. D. Humphries, R. D. Stewart, and K. N. Gurney, "A physiologically plausible model of action selection and oscillatory activity in the basal ganglia," *J. Neurosci.*, vol. 26, no. 50, pp. 12921–12942, Dec. 2006.
[28] A. Mandali, M. Rengaswamy, V. S. Chakravarthy, and A. A. Moustafa, "A spiking basal ganglia model of synchrony, exploration and decision making," *Frontiers Neurosci.*, vol. 9, p. 191, May 2015.
[29] M. Magnin, A. Morel, and D. Jeanmonod, "Single-unit analysis of the pallidum, thalamus and subthalamic nucleus in parkinsonian patients," *Neuroscience*, vol. 96, no. 3, pp. 549–564, 2000.
[30] A. Benazzouz, S. Breit, A. Koudsie, P. Pollak, P. Krack, and A.-L. Benabid, "Intraoperative microrecordings of the subthalamic nucleus in Parkinson's disease," *Movement Disorders*, vol. 17, no. S3, pp. S145–S149, Mar. 2002.
[31] T. Yamazaki and S. Tanaka, "A spiking network model for passage-of-time representation in the cerebellum," *Eur. J. Neurosci.*, vol. 26, no. 8, pp. 2279–2292, Oct. 2007.
[32] T. Yamazaki and S. Nagao, "A computational mechanism for unified gain and timing control in the cerebellum," *PLoS ONE*, vol. 7, no. 3, Mar. 2012, Art. no. e33319.
[33] F. Raudies and M. E. Hasselmo, "A model of hippocampal spiking responses to items during learning of a context-dependent task," *Frontiers Syst. Neurosci.*, vol. 8, p. 178, Sep. 2014.
[34] S. Béhuret, C. Deleuze, L. Gomez, Y. Frégnac, and T. Bal, "Cortically-controlled population stochastic facilitation as a plausible substrate for guiding sensory transfer across the thalamic gateway," *PLoS Comput. Biol.*, vol. 9, no. 12, Dec. 2013, Art. no. e1003401.
[35] O. Summ, A. R. Charbit, A. P. Andreou, and P. J. Goadsby, "Modulation of nocioceptive transmission with calcitonin gene-related peptide receptor antagonists in the thalamus," *Brain*, vol. 133, no. 9, pp. 2540–2548, Sep. 2010.
[36] 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.
[37] S. W. Moore *et al.*, "Bluehive–a field-programable custom computing machine for extreme-scale real-time neural network simulation," in *Proc. IEEE 20th Int. Symp. Field-Program. Custom Comput. Mach.*, Apr./May 2012, pp. 133–140.
[38] T. Yu, J. Park, S. Joshi, C. Maier, and G. Cauwenberghs, "65k-neuron integrate-and-fire array transceiver with address-event reconfigurable synaptic routing," in *Proc. IEEE Biomed. Circuits Syst. Conf. (BioCAS)*, Nov. 2012, pp. 21–24.
[39] P. Machado, J. Wade, and T. M. McGinnity, "Si elegans: FPGA hardware emulation of C. elegans nematode nervous system," in *Proc. 6th World Congr. Nature Biol. Inspired Comput. (NaBIC)*, Jul./Aug. 2014, pp. 65–71.
[40] M. Davies *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *IEEE Micro*, vol. 38, no. 1, pp. 82–99, Jan. 2018.
[41] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (DYNAPs)," *IEEE Trans. Biomed. Circuits Syst.*, vol. 12, no. 1, pp. 106–122, Feb. 2018.
[42] G. Cauwenberghs, "Reverse engineering the cognitive brain," *Proc. Nat. Acad. Sci. USA*, vol. 110, no. 39, pp. 15512–15513, Sep. 2013.
[43] A. H. Marblestone, G. Wayne, and K. P. Kording, "Toward an integration of deep learning and neuroscience," *Frontiers Comput. Neurosci.*, vol. 10, p. 94, Sep. 2016.
[44] M. R. Joglekar, J. F. Mejias, G. R. Yang, and X. J. Wang, "Inter-areal balanced amplification enhances signal propagation in a large-scale circuit model of the primate cortex," *Neuron*, vol. 98, no. 1, pp. 222–234, Apr. 2018.
[45] J. Bono and C. Clopath, "Synaptic plasticity onto inhibitory neurons as a mechanism for ocular dominance plasticity," *PLOS Comput. Biol.*, vol. 15, no. 3, Mar. 2019, Art. no. e1006834.
[46] J. P. Gallivan, C. S. Chapman, D. M. Wolpert, and J. R. Flanagan, "Decision-making in sensorimotor control," *Nature Rev. Neurosci.*, vol. 19, no. 9, pp. 519–534, Sep. 2018.
[47] J. P. Stroud *et al.*, "Motor primitives in space and time via targeted gain modulation in cortical networks," *Nature Neurosci.*, vol. 21, no. 12, pp. 1774–1783, 2018.
[48] P. U. Diehl and M. Cook, "Unsupervised learning of digit recognition using spike-timing-dependent plasticity," *Frontiers Comput. Neurosci.*, vol. 9, p. 99, Aug. 2015.
[49] Q. Wang, Y. Li, B. Shao, S. Dey, and P. Li, "Energy efficient parallel neuromorphic architectures with approximate arithmetic on FPGA," *Neurocomputing*, vol. 221, pp. 146–158, Jan. 2017.
[50] F. N. Buhler, P. Brown, J. Li, T. Chen, Z. Zhang, and M. P. Flynn, "A 3.43TOPS/W 48.9pJ/pixel 50.1nJ/classification 512 analog neuron sparse coding neural network with on-chip learning and classification in 40nm CMOS," in *Proc. Symp. VLSI Circuits*, Jun. 2017, pp. 30–31.
[51] J. K. Kim, P. Knag, T. Chen, and Z. Zhang, "A 640M pixel/s 3.65 mW sparse event-driven neuromorphic object recognition processor with on-chip learning," in *Proc. Symp. VLSI Circuits (VLSI Circuits)*, Jun. 2015, pp. 61–62.
[52] G. G. Turrigiano and S. B. Nelson, "Homeostatic plasticity in the developing nervous system," *Nature Rev. Neurosci.*, vol. 5, no. 2, pp. 97–107, Feb. 2004.
[53] J. Bono and C. Clopath, "Modeling somatic and dendritic spike mediated plasticity at the single neuron and network level," *Nature Commun.*, vol. 8, no. 1, pp. 1–17, Dec. 2017.
[54] Y. Mosbacher *et al.*, "Toward neuroprosthetic real-time communication from in silico to biological neuronal network via patterned optogenetic stimulation," *Sci. Rep.*, vol. 10, no. 1, pp. 1–16, Dec. 2020.
[55] S. Buccelli *et al.*, "A neuromorphic prosthesis to restore communication in neuronal networks," *iScience*, vol. 19, pp. 402–414, Sep. 2019.
[56] H. Keren, J. Partzsch, S. Marom, and C. G. Mayr, "A biohybrid setup for coupling biological and neuromorphic neural networks," *Frontiers Neurosci.*, vol. 13, p. 432, May 2019.
[57] A. Serb *et al.*, "Memristive synapses connect brain and silicon spiking neurons," *Sci. Rep.*, vol. 10, no. 1, p. 2590, Dec. 2020.

Shuangming Yang (Member, IEEE) received the B.S. degree from the Hebei University of Technology, Tianjin, China, in 2013, and the M.S. and Ph.D. degrees from Tianjin University, Tianjin, China, in 2016 and 2020, respectively.

He is currently a Lecturer with the School of Electrical and Information Engineering, Tianjin University. His research interests include neuromorphic engineering, computational neuroscience, brain-inspired computing, and machine learning.

Jiang Wang (Member, IEEE) was born in China, in 1964. He received the M.S. degree in power and automation engineering and the Ph.D. degree from the School of Management Engineering, Tianjin University, Tianjin, China, in 1989 and 1996, respectively.

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

Xinyu Hao received the B.S. degree from Tianjin University, Tianjin, China, in 2017, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering.

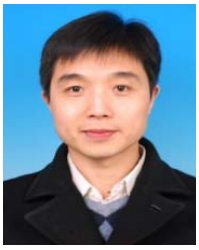
His research interests include field-programmable gate array (FPGA) design and brain-inspired computing.

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

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

Xile Wei (Member, IEEE) was born in China, in 1975. He received the B.S., M.S., and Ph.D. degrees from Tianjin University, Tianjin, China, in 1997, 2004, and 2007, respectively.

He is a Professor with the School of Electrical and Information Engineering, Tianjin University. His current research interests include analysis of bioelectromagnetics effects, neuromorphic modeling, neural control engineering, analysis of neural dynamics, and nonlinear control theory.



Bin Deng (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from Tianjin University, Tianjin, China, in 2007.

He is a Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include dynamic analysis of neuron model and nonlinear analysis of neuron electrical information.

Kenneth A. Loparo (Life Fellow, IEEE) is currently the Nord Professor of Engineering and the Chair of the Department of Electrical Engineering and Computer Science, Case Western Reserve University, Cleveland, OH, USA.

His current research interests include stability and control of nonlinear systems with applications to modeling, simulation, and variability analysis of biological systems.

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