CerebelluMorphic: Large-scale Neuromorphic Model and Architecture for Supervised Motor Learning

Shuangming Yang, *Member, IEEE*, Jiang Wang, Nan Zhang, Zhicai Hu, Bin Deng, *Senior Member, IEEE*, Yanwei Pang, *Senior Member, IEEE*, Mostafa Rahimi Azghadi, *Senior Member, IEEE*

Abstract—The cerebellum plays a vital role in motor learning and control with supervised learning capability, while neuromorphic engineering devises diverse approaches to high-performance computation inspired by biological neural systems. This paper presents a large-scale cerebellar network model for supervised learning, as well as a cerebellum-inspired neuromorphic architecture to map the cerebellar anatomical structure into the large-scale model. Our multi-nucleus model and its underpinning architecture contain approximately 3.5 million neurons, upscaling state-of-the-art neuromorphic designs by over 34 times. Besides, the proposed model and architecture incorporate 3411k granule cells, introducing 284 times increase compared to a previous study including only 12k cells. This large scaling induces more biologically plausible cerebellar divergence/convergence ratios, which results in better mimicking biology. In order to verify the functionality of our proposed model and demonstrate its strong bio-mimicry, a reconfigurable neuromorphic system is used, on which our developed architecture is realized to replicate cerebellar dynamics during optokinetic response. In addition, our neuromorphic architecture is used to analyse the dynamical synchronization within the Purkinje cells, revealing the effects of firing rates of mossy fibres on the resonance dynamics of Purkinje cells. Our experiments show that real-time operation can be realized, with system throughput of up to 4.70 times larger than previous works with high synaptic event rate. These results suggest that the proposed work provide both a theoretical basis and a neuromorphic engineering perspective for the brain-inspired computing and the further exploration of cerebellar learning.

***Index Terms*—cerebellum model, supervised learning, motor learning, neuromorphic engineering, spiking neural network (SNN).**

I. Introduction

A deep understanding of the structural and dynamic complexity of the human brain is highly dependent on the development of large-scale, anatomically detailed models of the brain network, which can reveal the mechanisms of how neuronal and synaptic processes interact to generate the collective behaviors of the brain [1]. Large-scale network simulation is essential because even a simple human behavior involves several million neurons [2]. In addition, for the sake of further exploration of the neural information processing mechanism underlying collective behaviors of the brain, it is important to realize a large-scale biologically inspired model of the mammalian brain [3].

The cerebellum, a critical part of the human brain, is responsible for motor control, sensorimotor coordination and adaptive learning. It is connected with the most vital parts of the central nervous system, such as the brain-stem, basal ganglia, spinal cord, limbic system, cerebral cortex and thalamus [4]-[6]. Learning implicit memory tasks is another cerebellum function with strong plastic modifications [7]-[8]. Cerebellum also participates in the regulation of somatic balance, muscle tone and coordination of voluntary movements. From the engineering perspective, the cerebellum is considered an adaptive control system [9]-[10] and is critical for computations involving daily manipulation tasks. It implements a feedforward, nonlinear regulator through learning the intrinsic dynamics of a robotic arm. It is also responsible for coordinating emotional and visceral functions, making sensory predictions, and elaborating certain aspects of cognition [11]-[13]. The cerebellum contains several critical components, including granule cells (GrCs), Golgi cells (GoCs), deep cerebellar nucleus (DCN) cells, Purkinie (PKJ) cells, inferior olive (IO) cells, climbing fibers (CFs), mossy fibers (MFs) and parallel fibers (PFs). These components constitute a schematic of the neural circuit involved in OKR adaption presented by a previous experimental study [31]. GrCs and GoCs are responsible for processing signals from MFs to provide a sparse code and receiving excitatory input, respectively. PKJ cells are responsible for recognizing the activity pattern of GrCs. DCN and IO cells are responsible for receiving signals and outputting neural signals, respectively. CFs originate from IO neurons. CFs, MFs and PFs are responsible for transmitting the signals between each neural cluster. PFs are the neurons connecting the cerebellum to the outside. PKJ cells are a class of GABAergic neurons located in the cerebellar cortex. The overall input-output function of the cerebellar network model is adaptive based on spike timing dependent plasticity (STDP) mechanisms at different sites [14]. STDP is a Hebbian-based learning rule that adjusts the synaptic weight of neuron connections using the timing information of the presynaptic and postsynaptic spikes. The STDP learning mechanism is widely used in building the biologically plausible SNN models for various brain regions,

This work was supported in part by the National Natural Science Foundation of China (Grant No. 61701320，61871287) and Natural Science Foundation of Tianjin (Grant No. 18JCZDJC32000).

Shuangming Yang, Jiang Wang, Nan Zhang, Zhicai Hu, Bin Deng and Yanwei Pang are with School of Electrical and Information Engineering, Tianjin University, Tianjin, 300072 China. (e-mail: yangshuangming @tju.edu.cn; corresponding author: dengbin@tju.edu.cn). M. Rahimi Azghadi is with the College of Science and Engineering, James Cook University, Townsville, QLD 4814, Australia.

and contains two major mechanisms known as long-term potentiation (LTP) and long-term depression (LTD). Although several sites in the cerebellar circuitry have self-learning processes, one of the most vital learning mechanisms is LTD at the parallel fiber-Purkinje cell (PF-PKJ) synapses that is closely related to cerebellar motor learning [15]-[16]. Another type of learning is LTP, which does not require CF activation and can compensate for the effect of LTD [20]. The LTP will induce a weight increase when it receives the firing spike from the GrCs, and the LTD is responsible for the teaching signals from IO cells. According to previous studies, the LTD and LTP will not occur at the same time [21]. In the learning process, the IO output can be regarded as an error-related signal that induces plasticity [17]-[19]. Due to strong plasticity characteristics, the cerebellum can be regarded as a learning machine.

In this study, a non Von Neumann computing architecture is presented to emulate the large-scale cerebellar network model to provide more biologically plausible characteristics of the cerebellum based on large-scale conductance-based spiking neural network (LaCSNN) system, which is digital neuromorphic architecture designed for simulating large-scale spiking neural networks [22]. The LaCSNN includes six Altera EP3SE340 FPGA chips that can communicate with each other using a multicast router. The computational efficiency and scalability of LaCSNN are significantly higher than central processing unit (CPU), graphics processing unit (GPU) and multi-core systems [22]. Due to its powerful computational capability, it can bridge the gap between the cellular level and the network level of a large-scale brain, and is suitable to simulate conductance-based network models, hence, we use it for our large-scale cerebellar network emulations.

Some progress has been made on simulating large-scale cerebellum network on other hardware platforms. Yamazaki et al. [23] presented real-time cerebellum network simulations using graphics processing unit (GPU), but its scalability is still limited and it is constrained by memory and bandwidth issues [24]. Another work simulated the cerebellum network based on the custom EDLUT platform [25], but the number of neurons on the platform is only 2100, which cannot be considered as sufficiently large considering the approximately 10^{11} granule cells in the cerebellum. Luo et al. used a FPGA chip to simulate the passage-of-time encoding in a large-scale cerebellar network [26], but it lacks learning mechanisms. Besides, it only contains GrCs and GoCs, which cannot reproduce the cognition functions of the cerebellum. Solinas et al. presented a realistic large-scale model of the cerebellum granular layer, but it does not have self- learning mechanisms, and cannot simulate in real-time, limiting its application [27]. Significantly advancing the previous efforts in simulating large-scale cerebellum networks, this paper focuses on the scalable modeling and implementation of large-scale spiking neural network with self-learning mechanism that can simulate the relevant dynamical behaviors in real-time.

The remainder of the paper is organized as follows. Section II describes our self-learning cerebellar spiking neural network model. In Section III, the detailed hardware implementation of the cerebellar network using the LaCSNN system is presented, and a set of designs are proposed to address the challenges in implementing the large-scale spiking neural network. Experimental results are presented in Section IV, including exploration of the system dynamics, hardware performance and precision analysis. Section V discusses the application of the presented cerebellar network, with comparisons to state-of-the-art techniques. This Section also discusses the limitations of the implemented cerebellar network and suggests future research directions. Finally, the paper is concluded in Section VI.

II. The Cerebellar Spiking Network Model

A. Cerebellar network architecture and motor control

The proposed cerebellum architecture is shown schematically in Fig.1(a), which is based on the Marr-Albus-Ito theory of cerebellar function [28]-[30]. MFs are modeled to simulate individual Poisson spikes and provide excitatory signals to GrCs and VN cells. GrCs, GoCs and MFs transmit signals to each other through a structure called the cerebellar glomerulus, that are complex synaptic nests and closely packed collections of synaptic endings formed by the enlarged ends of the MFs, the dendrites of the GrCs and the axons or proximal dendrites of the GoCs. The PKJ cells receive synaptic current from GrCs groups through the PFs and from IO cells through CFs. The GrC neurons transmit the synaptic information to PKJ neurons through parallel fibers, where synaptic plasticity exists. Teaching signals are generated by the IO neuron and transmitted to PKJ neurons to change the synaptic weight of PF-PC. The output of the cerebellar network is given by VN cells that connect to all the PKJ cells. It is worth noting that CFs originated from IO cells.

There are two known pathways for updating synaptic strength during cerebellar learning. In the first pathway, MFs increase or decrease output responses using direct excitatory connections with the VN cells. In the second pathway, the excitatory synapses from GrCs to PKJ cells are modified based on the CF inputs, inducing synapses that activate before the weight of CF inputs decrease. The PKJ cells then reduce their activity when the same input occurs at the next time. The decrease in the activities of PKJ cells causes the weights of the excitatory synapses from MFs to VN cells to increase, resulting in more responsive VN cells to the same inputs of MFs. These modifications of synapses at the two sites are the basis of feedforward prediction and are considered to be the foundation of the capability of cerebellum to coordinate and fine-tune motor responses.

Fig. 1(b) shows a schematic of the cerebellar neural circuit involved in horizontal optokinetic response (OKR) eye movements, which is fully described in [31]. Information of visual motion is transmitted from the retina via the pretectum and nucleus reticularis tegmenti pontis (NRTP) by MFs to the DCN and to the zone of flocculus that manages the horizontal movement of the eyeball. In addition, it is also transmitted to the flocculus via CFs by the IO, receiving inputs from the pretectum. The vestibular nucleus (VN) is inhibited by the floccullus, driving extraocular muscle motor cells. In this study,

the cerebellar cortex, i.e. flocculus, and VN, IO, and MFs, are modeled to explore the intrinsic dynamics of the flocculus and VN. The abbreviation of AN, ON, LR, and MR represent abducens nucleus, oculomotor nucleus, lateral rectus and medial rectus, respectively.

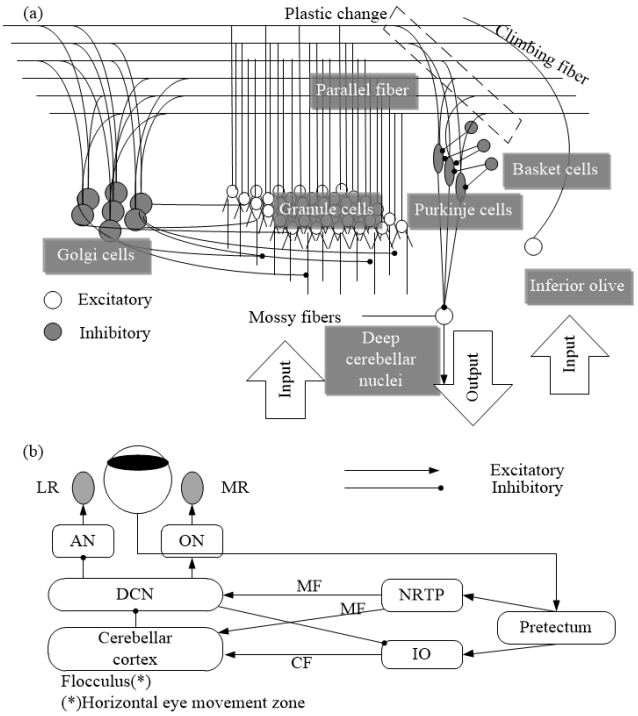


Fig. 1. Cerebellar structure and its motor learning. (a) The structure of the cerebellum model. (b) Schematics of the neural circuitry for optokinetic response (OKR) adaption in rabbits.

B. Network model

A previous study [32] suggests that our cerebellar neurons including GrCs, GoCs, VN, and IO cells, can be modeled as conductance-based leaky integrate-and-fire, dynamics of which is given by:

$$C\frac{dV}{dt} = -g_{leak}(V(t)-E_{leak}) - g_{AMPA}(t)(V(t)-E_{exc}) - g_{NMDA}(t)(V(t)-E_{exc}) - g_{inh}(t)(V(t)-E_{inh}) - g_{ahp}(t-t_0)(V(t)-E_{ahp}) + I_{app}, \quad (1)$$

where C is the capacitance and V is the membrane potential at each simulation step, t. The parameter E_{leak} represents the leakage potential, and E_{ahp} represents the post-hyperpolarization potential. The parameter g_x represents the resting conductance while E_x represents the resting potential where x can be either of *leak*, α-amino-3-hydroxy-5-methy1-4-isoxazolepropionic acid (*AMPA*), N-methy1-D-aspartate (*NMDA*), inhibitory (*inh*), after-hyperpolarization (*ahp*). The term I_{app} is spontaneous current that only exists in few types of cells. For synaptic currents, synapse conductance has an exponentially different time course, which is proportional to the probability of postsynaptic channel opening (P). The parameter t_0 is the presynaptic spike time. For each type x, the current is calculated from conductance g_x and reversal potential E_x, where subscript $x \in$ {leak, AMPA, NMDA, inh, ahp}.

Table 1 Summary of model parameters in the cerebellum network

Parameters	PKJ	GrC	GoC	BS	VN	IO
θ (mV)	-55.0	-35.0	-52.0	-55.0	-38.8	-50.0
C (pF)	107.0	3.1	28.0	107.0	122.3	10.0
g_{leak} (nS)	2.32	0.43	2.3	2.32	1.64	0.67
E_{leak}(mV)	-68.0	-58.0	-55.0	-68.0	-56.0	-60.0
$\hat{g}_{AMPA}$ (nS)	0.7	0.18	45.5	0.7	50.0	1.0
$\hat{g}_{NMDA}$ (nS)	--	0.025	30.0	--	25.8	--
E_{exc} (mV)	0	0	0	0	0	0
$\hat{g}_{inh}$ (nS)	1.0	0.028	--	--	30.0	0.18
E_{inh} (mV)	-75.0	-82.0	--	--	-88.0	-75.0
$\hat{g}_{ahp}$(nS)	0.1	1.0	20.0	0.1	50.0	1.0
E_{ahp} (mV)	-70.0	-82.0	-72.7	-70.0	-70.0	-75.0
τ_{ahp} (ms)	5.0	5.0	5.0	5.0	2.5	10.0
I_{spont} (nA)	0.25	--	--	--	0.7	--

Table 2 Parameter values of the exponential functions

Cell types	Exponential functions
PKJ	$\alpha_{AMPA}(t)=e^{-t/8.3}$, $\alpha_{inh}(t)=e^{-t/10.0}$
GrC	$\alpha_{AMPA}(t)=e^{-t/1.2}$, $\alpha_{NMDA}(t)=e^{-t/52.0}$, $\alpha_{inh}(t)=0.43e^{-t/7.0}+0.57e^{-t/59.0}$
GoC	$\alpha_{AMPA}(t)=e^{-t/1.5}$, $\alpha_{NMDA}(t)=0.33e^{-t/31.0}+0.67e^{-t/170.0}$
BS	$\alpha_{AMPA}(t)=e^{-t/8.3}$
VN	$\alpha_{AMPA}(t)=e^{-t/9.9}$, $\alpha_{NMDA}(t)=e^{-t/30.6}$, $\alpha_{inh}(t)=e^{-t/42.3}$
IO	$\alpha_{AMPA}(t)=e^{-t/10.0}$, $\alpha_{inh}(t)=e^{-t/10.0}$

The computation of conductance is defined by the convolution of the exponential function $\alpha_j(t)$ and the spike event $\delta_j(t)$ of presynaptic neuron j at time t as follows:

$$g_x(t) = \bar{g}_x \sum_j w_j \int_{-\infty}^{t} \alpha(t-s)\delta_j(s)ds, \quad (2)$$

where $\bar{g}_x$ represents the maximum conductance and w_j stands for the efficacy of signal transmission considered as the synaptic weight from the presynaptic neuron j. The neuron fires when the membrane potential reaches and exceeds the threshold θ, which is described by the after-hyperpolarization value and determines the refractory period. The after-hyperpolarization conductance is

$$g_{ahp}(t-\hat{t}) = \exp\left(-(t-\hat{t})/\tau_{ahp}\right), \quad (3)$$

where τ_{ahp} represents time constant of after-hyperpolarization, and $\hat{t}$ represents the last spiking time of the neuron. The parameter values used for our experiments are reported in Table 1 and Table 2. These parameters were taken from previous known physiological experimental papers that are listed in Sections S1 and S2 in Supplementary material.

C. Learning mechanism of the cerebellar network

The synaptic conductance of the proposed cerebellar model is changed based on the spike-timing-dependent plasticity (STDP) learning rule. Unlike the previous studies such as [34], which used only one learning mechanisms to govern the cerebellar synaptic conductance, this work uses three different learning mechanisms to implement a more biologically faithful model. The first learning mechanism used is STDP that modifies synaptic connections between PF and PC and is vital for motor

learning. This STDP mechanism includes LTP and LTD. According to [35]-[36], LTP is the default response in the STDP of PF-PC connections and the switching between LTD and LTP is determined by the local calcium concentration. The synaptic weight between GrC_j to PKJ_i at time t is represented by $w_{PKJi\rightarrow PFj}(t)$, which is

$$\begin{aligned} w_{PF_j\rightarrow PKJ_i}(t+1) = {} & 0.0005\left(w_{init} - w_{PF_j\rightarrow PKJ_i}(t)\right)PF_j(t) \\ & - 0.005 w_{PF_j\rightarrow PKJ_i}(t)\sum_{\Delta t=0}^{50} CF(t)\,PF_j(t-\Delta t), \\ & + w_{PF_j\rightarrow PKJ_i}(t) \end{aligned} \quad (4)$$

where $PF_j(t)$ and $CF(t)$ equal to PF_j or CF spikes at time t, otherwise equal to 0. The first term is LTP by PF stimulation only [36]-[37]. The second term is LTD by conjunctive activation of a CF and a PF, which is activated 0-50 ms earlier than the activation by CF. The constant w_{init}=1 denotes the initial synaptic weight. The parameter values are based on experimental findings of previous works that are listed in Section S1 in Supplementary material.

The second learning mechanism used alters the MF-VN synaptic connections, which were proposed in previous studies to explore the effect of multiple plasticity sites on cerebellar learning [38] as

$$\Delta W_{MFi\rightarrow VNj}(t) = \begin{cases} -10^{-5}\cdot\int_{-\infty}^{+\infty} K(t-x)\,\delta_{MFi}(t-x)\,dx \\ \qquad \text{if } MF_i \text{ active, } t = t_{PKJspikej} \\ 10^{-5} \qquad \text{if } MF_i \text{ active, } t \neq t_{PKJspikej} \\ 0 \qquad \text{otherwise} \end{cases}, \quad (5)$$

with the ancillary relationship

$$\delta_{MFi}(s) = \begin{cases} 1 & \text{if } MF_i \text{ is active at time } s \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

and the kernel function

$$K(z) = e^{-\frac{|z|}{\tau}}\left(\cos\left(\frac{z}{\tau}\right)\right)^2, \quad (7)$$

where $t_{PKJspikej}$ is the time when the corresponding jth PKJ cell spikes, and $K(z)$ is the integral kernel function. The parameter values are based on experimental findings of previous works that are listed in Section S2 in Supplementary material. The constant τ is employed to normalize the arguments in the learning rule. The standard spike-timing-plasticity method between PKJ and VN nuclei defines the third learning mechanism [39]. The inhibitory synapses from the two PCs to the corresponding VN strengthen when a PKJ cell spikes after VN spiking within an LTP time window (20 ms). Otherwise, the LTD synaptic weight changes when the opposite chronological ordering of events occurs within an LTD time window (60 ms).

D. Address event representation (AER) communication

In neural systems of the mammalian brain, action potentials, i.e., spikes, are transmitted along axons carrying long-distance neural information that is projected onto a large number of other cells distributed over different spatial domains. It is not trivial to implement this mechanism of distributed communications with spike events to realize efficient and scalable computation of large-scale neural networks on neuromorphic systems. LaCSNN uses the AER principle as an efficient point-to-point communication method among neural populations, where the addresses of neurons are communicated asynchronously whenever they fire. The AER communication principle routes address events directly using a synapse routing table in memory to make the synaptic connections in a dynamically reconfigurable manner, mapping pre-synaptic source addresses to post-synaptic target addresses. The virtual routing of AER-based synaptic connections among networks provides the flexibility to connect any pair of neurons. From a system perspective, AER-based synaptic connections allow multi-chip integration of neural cognitive systems based on spike events for various types of cognition tasks including object recognition and network learning.

E. Evaluation criterion

To determine the role of the network structure in the generation of cerebellar oscillations and PKJ fidelity, it is vital to repeat the simulations using regenerated networks based on the basic architecture described in Fig. 1. In order to explore how the activity patterns of GrC clusters evolves over time, the population average activity of GrC cluster i at time t is computed as

$$z_i(t) = \frac{1}{\tau_{PKJ}}\sum_{s=0}^{t}\exp\left(-(t-s)/\tau_{PKJ}\right)\left(\frac{1}{N_c}\sum_{j=1}^{N_c}\delta_{ij}(s)\right), \quad (8)$$

where N_c=100 is the number of GrCs in a cluster. The variable $\delta_{i,j}(t)$=1 when GrC_j in cluster i fires at time t, otherwise $\delta_{i,j}(t)$=0. Parameter τ_{PKJ}=8.3ms is the time constant of AMPAR-mediated EPSPs at the PF-Purkinje cell synapses, and $z_i(t)$ represents the AMPAR-mediated EPSPs at a PKJ cell induced by the ith GrC cluster at time t. The autocorrelation of the activity pattern at time t and t+Δt is defined as

$$C(t+\Delta t) = \frac{\sum_i z_i(t)\,z_i(t+\Delta t)}{\sqrt{\sum_i z_i^2(t)}\sqrt{\sum_i z_i^2(t+\Delta t)}}, \quad (9)$$

which represents the normalized inner product of population vectors of GrC clusters at times t and t+Δt. Because $z_i(t)$ only has positive values, the value of the correlation is between 0 and 1. When the population vectors at time t and t+Δt are identical, the correlation $C(t+\Delta t)$=1. The correlation equals to 0 for orthogonal vectors with no overlap in active populations.

The similarity index $S(\Delta t)$ is defined as

$$S(\Delta t) = \frac{1}{T}\sum_{t=0}^{T} C(t, t+\Delta t), \quad (10)$$

where T represents the inverse of the oscillation frequency of MF inputs. The reproducibility index $R(t)$ indicates how two activity patterns are differentiated in time as given by

$$R(t) = \frac{2}{KT}\sum_t\left(C^{(1)}(t) + C^{(3)}(t) + \ldots + C^{(K-1)}(t)\right) \quad (11)$$

where 10 pairs of successive cycles, i.e., K=20, is used to calculate the reproducibility.

III. Neuromorphic Cerebellar Architecture

The proposed digital neuromorphic cerebellar network is biologically meaningful and uses conductance-based leaky integrate-and-fire (LIF) neuron models to regenerate biologically plausible dynamic behaviors. Inspired by the computational architecture of the human brain, it is a multi-core digital neuromorphic system that is efficient, scalable and flexible. At the cellular level, the firing activities of cells for spiking information coding and processing is considered. At the network level, the large-scale cerebellar network can be realized using the neuromorphic architecture platform to comprehend the underlying mechanisms of cerebellar motor learning. At the synapse level, a key component is modeling the plasticity in different sites of cerebellum, which enables different nuclei to learn over time through changes in synaptic sensitivity and through modification of synaptic weights. In the following subsections, the architecture of the proposed CerebelluMorphic system is described in more details.

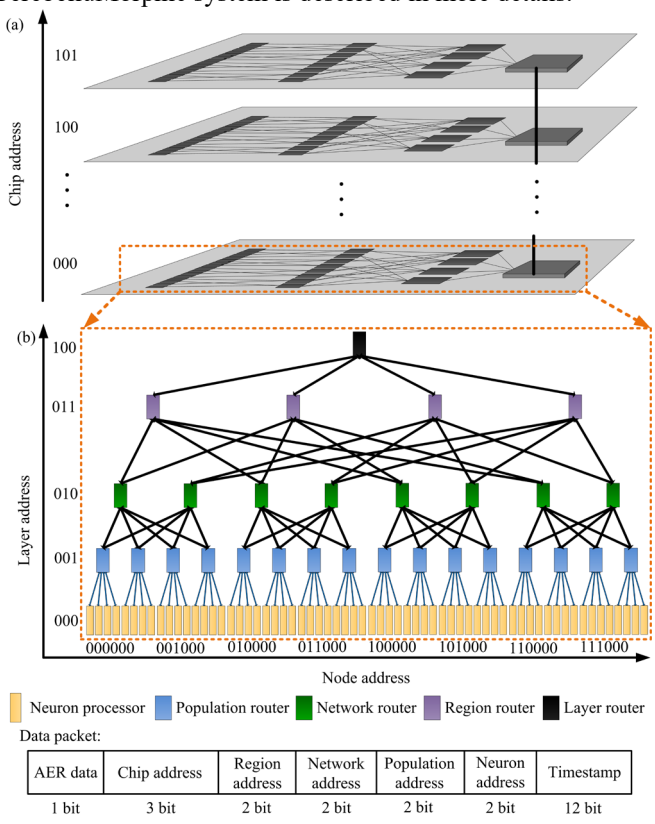


Fig. 2. The topology of the presented 3D BFT for large-scale neuromorphic realization. (a) The scalable connection structure for the proposed 3D NoC system. (b) The digital architecture on each chip.

A. Network-on-chip (NoC) architecture

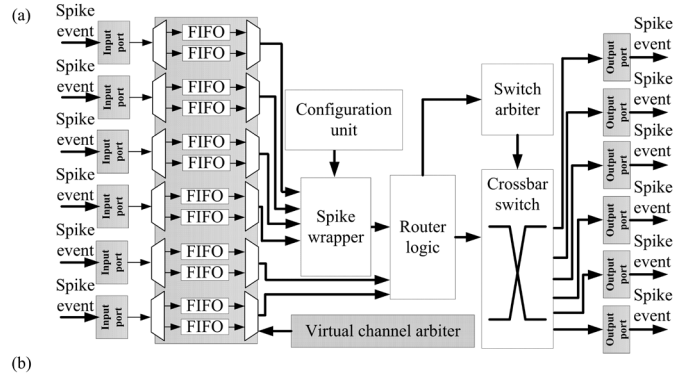
As shown in Fig. 2(a), the large-scale spiking cerebellar network is implemented using butterfly fat tree (BFT) topology. The scalable structure can be divided into two parts: the horizontal and the vertical BFT layers. The horizontal BFT layers are realized using FPGAs, and the vertical layers are implemented using the high-speed Terasic connector (HSTC). Network data is transmitted as 24-bit data packets, containing 1-bit AER data, 3-bit chip address, 2-bit region address, 2-bit network address, 2-bit population address, 2-bit region address and 12-bit timestamp. With a 12-bit timestamp, a maximum number of 4096 time-steps is coded. Therefore, each time-stamp is roughly 224 ns. LaCSNN runs with the clock frequency of 50 MHz. Each digital neuron completes its one-step computation within 10 clock periods (200 ns). Therefore, the utilized 12-bit timestamp is enough for the simulation of the time-multiplexed neurons, and will not limit the simulation maximum time. More spikes and longer time is not considered in this study, but will be investigated in our future studies. The synaptic connectivity is represented based on the AER communication protocol in a dynamically reconfigurable manner using routing address events from the synaptic routing tables, mapping the presynaptic source address to the postsynaptic destination address. The floor plan of a BFT topology in 3D NoC architecture is shown in Fig. 2(b), and the number of neuron units is determined by the available hardware resources; requiring three layers of BFT architecture for a network with 64 neuron units.

B. Routing of the neuromorphic cerebellar architecture

There are six internal input ports in the first-level router to interface with the second-layer neighboring routers or cerebellar neuron units (CNUs). Routers realize the routing and data flow control functions and are the key components of our digital neuromorphic system architecture. A new router for the digital neuromorphic cerebellum is shown in Fig. 3(a). Six bidirectional ports are contained in the proposed router, connecting two parent router nodes and four child router nodes respectively. A packetization process is initiated using the spike wrapper unit when the first-level router receives a spike event from a CNU. The spike wrapper unit is used to process a single spike event into a valid AER spike packet. It uses the information in the configuration processor for the data process. The configuration process can be reconfigured at any time according to the neural connectivity. The configuration processor contains four kinds of registers: chip address register, layer address register, node address register and timestamp register. Incoming spike events and the corresponding deliver-at time are stored in the on-chip memory after the deliver-at time stamps are reached. The routing logic unit processes the AER packet according to the routing algorithm as shown in Fig. 3(b). The crossbar switch in the first-level router is realized by multiplexers and is controlled by signals from the crossbar arbiter. The AER spike packets are then routed to the output ports with four AER spike events to the CNUs and two to the higher-level routers.

The detailed routing algorithm is shown in Fig. 3(b). In this algorithm, layers, regions, networks and populations are the source addresses of each layer, region, network and population router, while layerd, regiond, networkd and populationd represent routers with the corresponding destination addresses. For each router at the population, network and region layer, the destination addresses of the population routers and its bottom layer routers are compared with the corresponding addresses. If they are equivalent, then the AER data is routed to the corresponding node in the downstream layer, i.e. the neuron processor. If they are not equivalent, then the AER data is

transmitted up to the corresponding network router. The same procedure is realized for the network and region routers. For the layer routers, the current router address is compared with the destination layer address. The AER data is transmitted to the downstream corresponding region router if they are the same; otherwise it will be transmitted to other layers according to the destination layer address.



(b)

Routing logic algorithm in BFT architecture

begin
Neural network, population, region, layer for source and destination;
if *Layers* = *Layerd* **then:**
if *Regions*= *Regiond* **then:**
if *Networks* = *Networkd* **then:**
if *Populations* = *Populationd* **then:**
Place spike event into respective physical neuron;
else:
Reach *Nd* connected to *Ps*;
Reach *Pd*, connected to *Nd*;
Place the spike event into respective physical neuron;
end
else:
Reach *Nd*, connected to *Ps*;
Reach *Pd*, connected to *Nd*;
Place the spike event into respective physical neuron;
end
else:
Reach *Ns*, connected to *Ps*;
Reach respective *Rs*, connected to *Ns*;
end
else:
Reach either of two *Ns*, connected to *Ps*;
Reach *Rs*, connected to *Ns*;
Reach *Ls*, connected to *Rs*;
if the *address* of *Ld* > the *address* of *Ls* **then:**
while s<d**:**
Reach *Ls+1*,, connected to *Ls*;
s=s+1;
end
if the *address* of *Ld* < the *address* of *Ls* **then:**
while s>d**:**
Reach *Ls-1*,, connected to *Ls*;
s=s-1;
end
Reach *Rd,* connected to *Ld*;
Reach any of the two *Nd*, connected to *Rd*;
Reach *Pd*, connected to *Nd*;
Place spike event into respective physical neuron;
end
end

Fig. 3. Digital neuromorphic architecture of the routing unit. (a) Detailed digital architecture of the proposed router. It contains FIFO blocks, virtual channel arbiter, spike wrapper, router logic, configuration unit switch arbiter and crossbar switch. It is utilized for efficient routing of spike events from various neuron units. (b) Pseudo code of the routing logic for BFT architecture. Addresses of each layer, region, network and population router are considered in the proposed routing algorithm.

C. Digital implementation of the cerebellar neuron

The digital architectures of the nucleus processors are shown in Fig. 4, which includes GrC, GoC, PKJ, BC, IO and VN processors. Each router has six ports to communicate the AER event with the destination node of the BFT, using two up ports and four down ports. The AER spike event is transmitted by the router in each nucleus processor for the calculation of synaptic currents. In the PKJ processor, three kinds of spike events are required in the silicon synapse units: δ_{GrC}, δ_{BC} and δ_{IO}. The silicon synapse units calculate the synaptic currents $I_{\mathrm{GrC}\rightarrow\mathrm{PKJ}}$, $I_{\mathrm{BC}\rightarrow\mathrm{PKJ}}$, $I_{\mathrm{IO}\rightarrow\mathrm{PKJ}}$, respectively. The PKJ neuron unit calculates the spike event δ_{PKJ} and outputs it to the router. The configuration unit is responsible for the configuration of the router and silicon synapse units. The PKJ neuron unit uses time multiplexing technique and on-chip memory to achieve 9000 virtual neurons with one physical unit. The digital realizations of other nucleus processors use the same method as the PKJ processor with only one silicon synapse for the network computation.

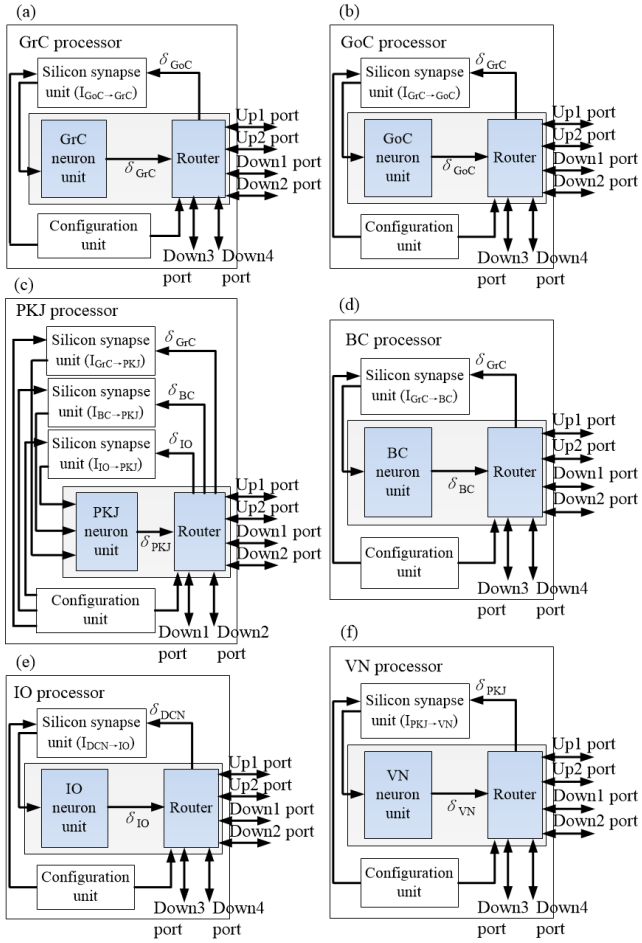


Fig. 4. Digital neuromorphic architecture of the nucleus processors. Each nucleus processor contains one or several silicon synapse units, a neuron unit, a router and a configuration unit. The router has six directions of ports, including two up ports and four down ports. The nucleus processors include (a) GrC processor, (b) GoC processor, (c) PKJ processor, (d) BC processor, (e) IO processor and (f) VN processor.

In order to realize different types of neurons in the cerebellar network, Euler method of numerical integration is used in the digital implementation to reduce the required computational resources compared to the Runge-Kutta method. The digital implementation for GrC neuron model shown in Fig. 5(a) includes one pipeline and one RAM module used for time multiplexing. Several RAM modules are used to store the variable values on FPGA. The RAM modules require $SI_{\mathrm{V}}*V_{\mathrm{b}}$ bits of on-chip memory, where SI_{V} is the data size of variable V and V_{b} is the bit width for each data. The latency number of the pipeline is V_{delay}, which has the relationship $V_{\mathrm{delay}}=V_{\mathrm{stage}}$ for pipeline synchronization. The detailed digital architecture of the V pipeline is shown in Fig. 5(b). The ADD and SUB blocks implement the addition and subtraction operations respectively. Detailed digital architecture of G_{ahp} module is shown in Fig. 5(c). The shift logic multiplier block, named SLM multiplier, is a dedicated digital circuit presented in this study for multiplication calculations without embedded multiplier resource on the FPGA as shown in Fig. 5(d). The SLM block is

used for multiplication between two variables in both neuron unit and silicon synapse unit.

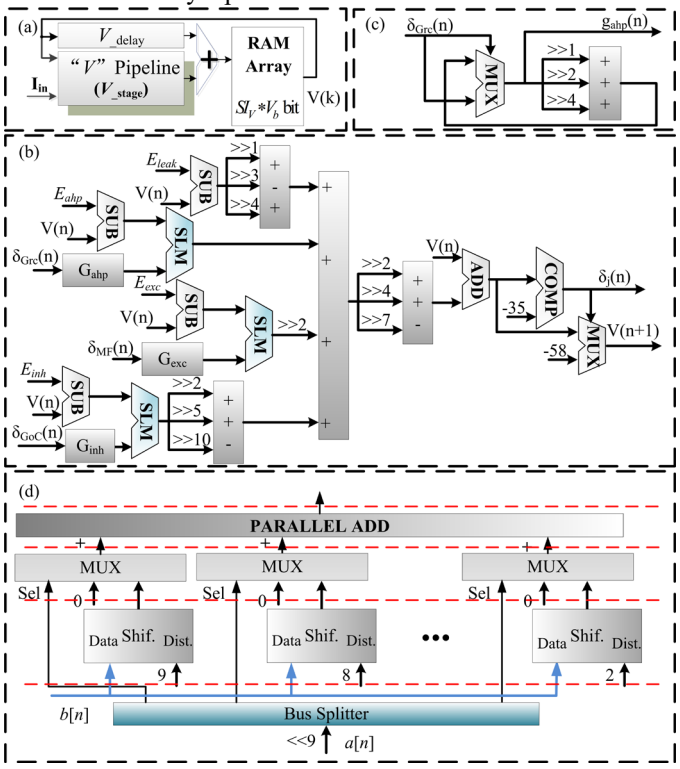


Fig. 5. Digital neuromorphic architecture of the neuron unit of the cerebellar GrC neuron, which is implemented based on a multiplier-less scheme. (a) The detailed "V" pipeline structure, which contains RAM array to store the transient variable values. (b) General overview of the neuron unit for GrC neuron, which uses SLM modules and shifter modules to replace multipliers. (c) The detailed digital architecture of "G_{ahp}" module, which uses shifter modules to replace multipliers. (d) SLM module for digital multiplier-less realization. It uses bus splitter, multiplier, barrel shifter and a parallel adder to realize the multiplication between the neural variables.

D. Digital architecture of the cerebellar synaptic plasticity

The detailed digital architecture of the synapse unit is shown in Fig. 6(a). There are 100 parallel groups of synaptic current processors (SCPs) in the silicon synapse unit. The AER spike packet is input to a multiplexer with 100 outputs, with its ports selected by a regular counter for sequential selection. The AER spike packets are processed by decoders to obtain the event data and its corresponding timestamp. The timestamp is used as the write address of the buffer, and the read address is controlled by a counter. In each SCP, the connectivity C_{ij} is determined by the configuration unit and all the multiplication operations use the SLM block. Cerebellar plasticity for the large-scale neuromorphic SNN is shown in Fig. 6(b)-(e). The ACC block represents the accumulator with two data ports and a synchronous clear port. The MUX block represents the multiplexer that selects the data path according to the control signal. In Fig. 6(d), the variable counter number (CN) represents the current number that is counted by a digital counter sequentially corresponding to the last firing activity in the fixed time window. The LUT block represents a look-up table to look up the prestored values when needed. The incoming information is the spiking activity of the jth neuron "$V_j[n]$". The variable NCI stands for the network connectivity information. The ABS block outputs the absolute value of the incoming input. In Fig. 6(e), if the peak value of the spike is detected, the corresponding counter number is sent to the output register. The incoming information is the spiking activity of the jth neuron "$V_j[n]$". The sclr signal represents synchronous clear signals to reset the counter at each period. The value of CN is obtained from the output register at the end of each time window and is then computed according to the STDP learning rule.

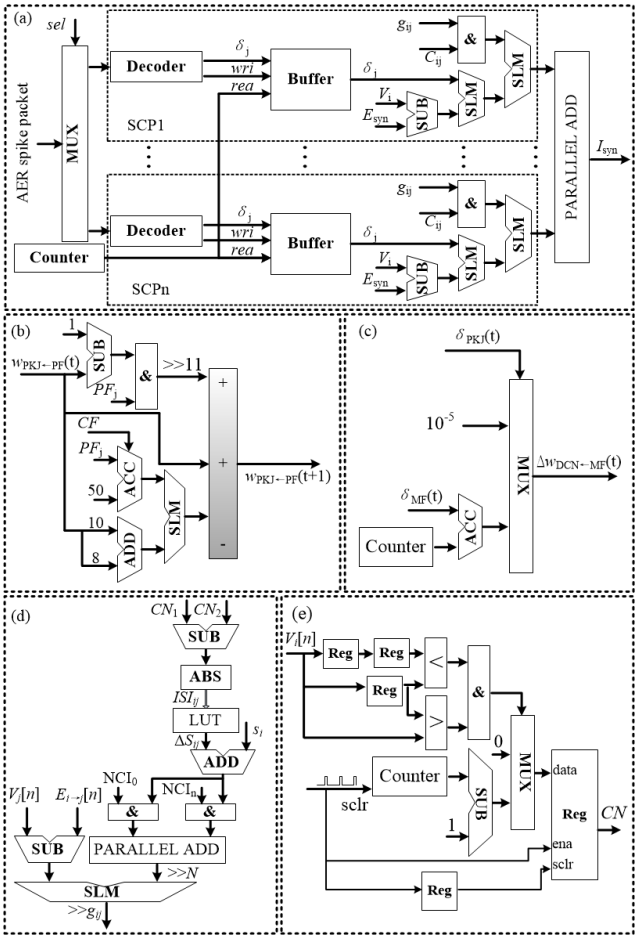


Fig. 6. Digital neuromorphic implementation of the cerebellar synaptic plasticity. (a) Digital architecture of the silicon synapse unit, which uses the parallel computational architecture. (b) Synaptic weight computation from PF to PKJ. (c) Synaptic weight change from VN to MF. (d) STDP learning computation from PKJ to VN. (e) Computation of the value of *CN*.

IV. Experimental Results

In this study, we use the LaCSNN neuromorphic system [22] to develop a detailed computational model of the large-scale cerebellar network with high biological plausibility and conduct dynamical analysis experiments of the proposed spiking network. The core component of LaCSNN is Intel Stratix III 340 FPGA. Its hardware resource contains 338000 logic elements, 16272 kbits of memory, and 576 18 × 18-bit multipliers blocks. To that end, the performance of the neuromorphic cerebellar network is analyzed. In addition, dynamical analysis of the synchronization properties within the PKJ cells is conducted. The cerebellar mechanisms have typically been studied independently using OKR eye movements in the Pavlovian delay eyeblink [40]. In OKR adaption, MFs and CFs convey retinal slip information, which

oscillates periodically in time. From the start of a cycle of the oscillation, different populations of GrCs become active one by one, sequentially. In this way, the cerebellum can learn the complete waveform instructed by the CFs. In this section, the learning ability of both PKJ and VN cells is investigated, and the dynamic response of the GrC neurons is also explored under the OKR adaptation condition.

A. Performance evaluation of the digital neuromorphic cerebellum

The proposed CerebelluMorphic system uses six Intel Stratix III EP3SL340 FPGAs to realize the large-scale neuromorphic cerebellar network with approximately 3.5 million neurons and 218.3 million synapses shown in Fig. 2(a). It contains 3411k GrC neurons, 1024 GoC neurons, 32 PKJ neurons, 128 BS neurons, 4 IO neurons and 8 VN neurons. Compared with previous studies [52], the proposed neuromorphic cerebellum model contains nearly 284.25 times more GrC neurons. As a result, the cerebellum divergence/convergence ratios can more closely approximate those ratios observed in biological cerebellums [14, 53].

In order to demonstrate the real-time computational capability of CerebelluMorphic system, the outputs of the spiking activities are sampled by oscilloscope, which is shown in Fig. 7. The input discrete spikes from the neuromorphic MFs are shown in Fig. 7(a), which is modeled by Poisson spikes. Fig. 7(b) shows the output discrete spikes from the GrC neurons randomly chosen on the CerebelluMorphic system. Raster plot of the output discrete spikes of neurons are shown in Fig. 8.

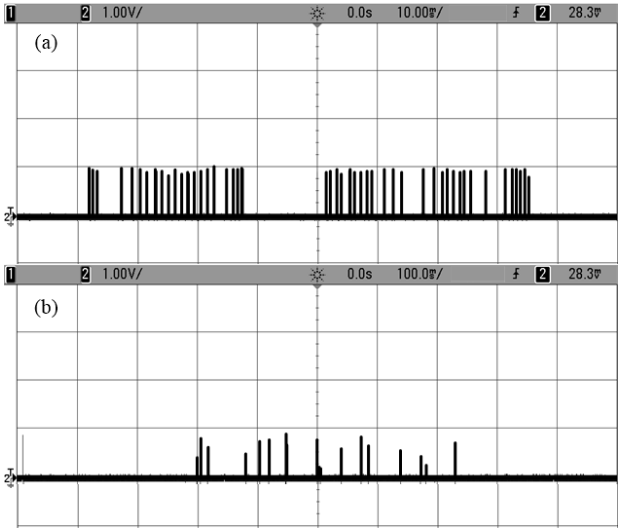


Fig. 7. The real-time spiking activities of the proposed CerebelluMorphic system on the oscilloscope. (a) The input discrete spikes from MFs. (b) The output discrete spikes from the proposed CerebelluMorphic network.

Bit-level fixed-point evaluation is proposed to investigate the computational precision. The evaluation criteria include root mean square error (RMSE), mean absolute error (MAE), correlation coefficient (CORR) and error of spike timing (ERRTT), that are computed as follows

$$\begin{cases} \mathrm{RSME} = \sqrt{\dfrac{1}{N}\sum_{i=1}^{N}(x_{sof}(i)-x_{har}(i))^2} \\ \mathrm{MAE} = \max\left|x_{sof}(i)-x_{har}(i)\right| \\ \mathrm{CORR} = \dfrac{\mathrm{cov}(x_{sof},x_{har})}{\sigma(x_{sof})\sigma(x_{har})} \\ \mathrm{ERRTT} = \left|\dfrac{\Delta T_{har}-\Delta T_{sof}}{\Delta T_{sof}}\right| \end{cases} \tag{12}$$

where $x_{sof}(i)$ and $x_{har}(i)$ represent the software and hardware computational results at the i^{th} iteration. Variables ΔT_{har} and ΔT_{sof} are the spiking time intervals of the hardware and software results. CORR is defined as the ratio of the covariance to variance product of the two data sets where

$$\begin{cases} \mathrm{cov}(x_{sof},x_{har}) = \sum_{i=1}^{n}(x_{sof}(i)-\bar{x}_{sof})(x_{har}(i)-\bar{x}_{har}) \\ \sigma(x) = \sqrt{\sum_{i=1}^{n}(x(i)-\bar{x})^2} \end{cases} \tag{13}$$

and $\bar{x}_{sof}$ and $\bar{x}_{har}$ represent the average values of $x_{sof}(i)$ and $x_{har}(i)$ respectively. As expected, the error evaluation results in Fig. 9 demonstrate that by increasing the bit width in the proposed digital neuromorphic cerebellar network the computational precision can be further enhanced.

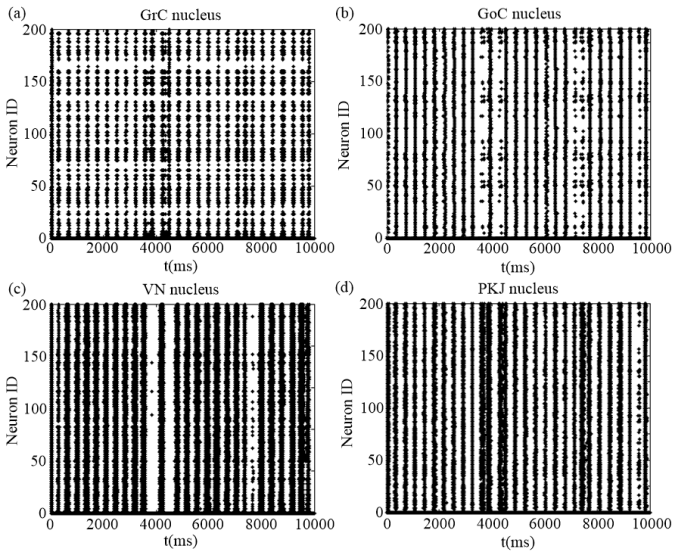


Fig. 8. Raster plot of the output discrete spikes of 200 neurons chosen randomly in the proposed neuromorphic cerebellar network. (a) GrC nucleus. (b) GoC nucleus. (c) VN nucleus. (d) PKJ nucleus.

In order to evaluate the system performance of the proposed CerebelluMorphic, a comparison of the throughput is performed between the proposed system and three state-of-the-art cerebellar digital neuromorphic systems [22, 54, 55]. In these comparisons, the CerebelluMorphic is configured to route neural events across the proposed architecture at different levels of synaptic events, from 20M synaptic outputs per second

(SynOPS) to 180MSynOPS. We use some typical destination distribution patterns, which include normalized traffic pattern, hotspot traffic pattern, and tornado traffic pattern, to evaluate the performance of the proposed system and compare it with state-of-the-art works. Fig. 10(a) shows the comparison of the system throughput with the normalized traffic pattern, and Fig. 10(b)-(c) demonstrate the comparison of the system throughput with 10% and 30% hotspot traffic respectively. The comparison of the system throughput with the tornado traffic pattern is shown in Fig. 10(d), where the node #*i* sends information to node ((*i*+*k*/2-1) mod *k*), where *k* represents network diameter. The throughput of the proposed system is significantly larger than the other two systems with the increment of the injected synaptic event rate. At 160MSynOPS event rate, the throughput of CerebelluMorphic is 4.09 and 1.18 times larger than the other systems respectively under the normalized traffic pattern, 4.70 and 1.26 times larger under 10% hotspot traffic pattern, 3.27 and 1.55 larger under 30% hotspot traffic pattern, 1.78 and 1.33 larger under tornado traffic pattern. This significant improvement is due to the BFT-based architecture of the proposed CerebelluMorphic digital neuromorphic system. The achieved throughput increase suggests that the CerebelluMorphic system can process larger information load within a certain period of time compared to the other two neuromorphic systems.

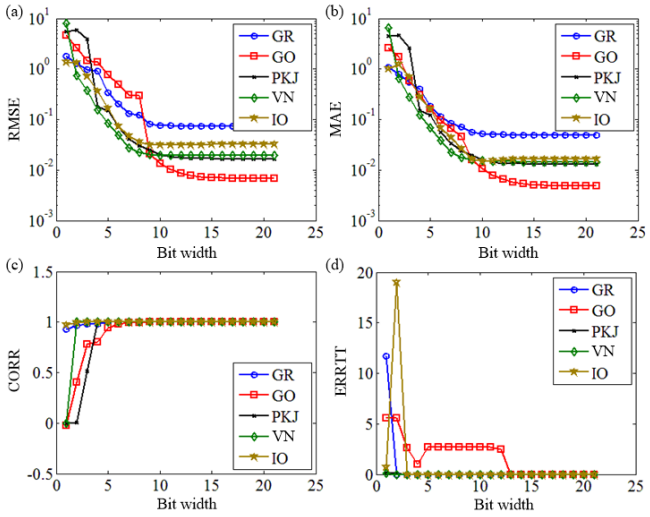


Fig. 9. Precision analysis of the digital neuromorphic computation for each nuclei in the neuromorphic cerebellar model. Different types of nuclei are considered, including GR, GO, PKJ, VN, and IO. Here, the analysis of the impact of the bit width is shown on (a) RMSE (b) MAE (c) CORR and (d) ERRTT.

B. Dynamical analysis of the neuromorphic PKJ cells

Cerebellar PKJ cells possess complex intrinsic biological behaviors, and can integrate numerous synaptic inputs. It is the sole output of the cerebellar cortex, thus understanding the dynamics of the PKJ cells is essential for the comprehension of cerebellar functions. Synchronization and resonance dynamics are essential mechanisms for neural information encoding and transmission. The coordination between neuron activities is featured with neural correlation in the population coding, and synchronization is a vital manifestation of the correlation between neurons. Resonance dynamics describe the firing output response to the input signal, which have been investigated in biological neural systems for years. In order to show the application of our CerebelluMorphic system, here we use it to explore the synchronization and resonance dynamics of the neuromorphic PKJ cells to study the impact of the synaptic weights on the PKJ population. The dynamical analysis of the PKJ cells were performed with active plasticity.

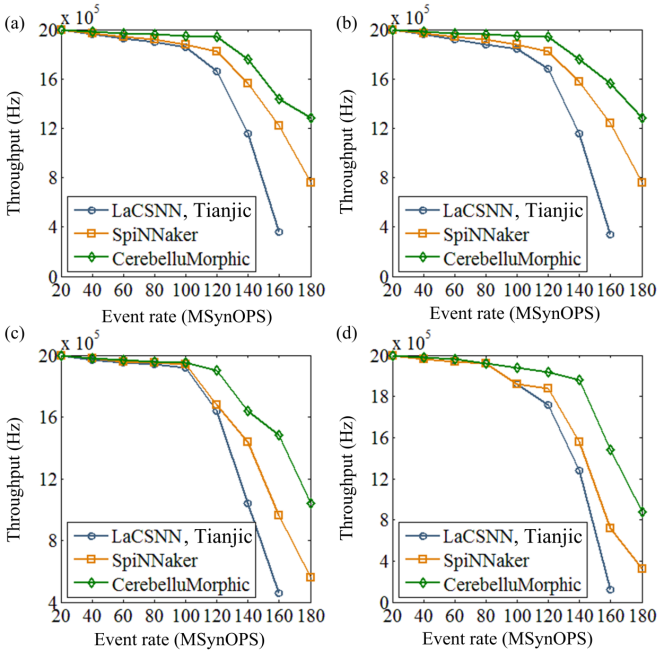


Fig. 10. System evaluation and comparison of the CerebelluMorphic throughput in different conditions. Three architectures of neuromorphic systems are considered, including LaCSNN/Tianjic, SpiNNaker and CerebelluMorphic (this study). Here, (a) Throughput under the normalized traffic pattern, (b) 10% and, (c) 30% hotspot traffic pattern, and (d) the tornado traffic pattern are shown against event rate.

In order to investigate the synchronization of the PKJ population, a network synchronization criteria is defined as

$$x_s = \frac{\left(\left\langle E(t)^2\right\rangle_t - \left\langle E(t)\right\rangle_t^2\right)}{\frac{1}{N}\sum_{i=1}^{N}\left(\left\langle E_i(t)^2\right\rangle - \left\langle E_i(t)\right\rangle_t^2\right)} \tag{14}$$

where $E(t)$ represents the average spike events of PKJ neurons, and N represents the total neuron number within the PKJ nucleus. $E(t)$ is defined as

$$E(n) = (1/N)\sum_{i=1}^{N} E_i(n) \tag{15}$$

As shown in Fig. 11(a) and (b), the increment of the synaptic weight $w_{BS\to PKJ}$ from BS to PKJ cells can increase the synchronization level of the PKJ population, while the weight increment of $w_{GrC\to PKJ}$ from GrC to PKJ will decrease the PKJ synchronization dynamics. Furthermore, with the oscillation frequency of the MFs increasing, the synchronization will be enhanced, and larger synaptic weights from GrC to PKJ will remove this effect. With low weights from BS to PKJ and high

weights from GrC to PKJ, the network synchronization of PKJ cells is not directly affected by the MF firing rate.

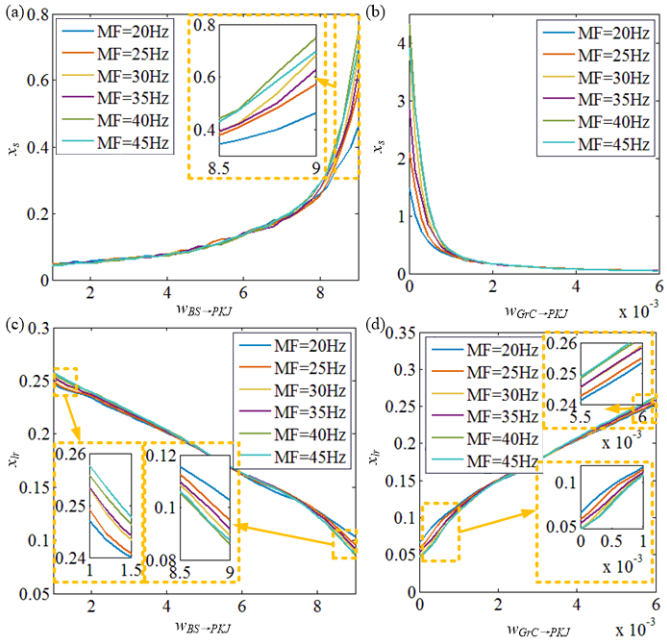


Fig. 11. Dynamical analysis of PKJ population. Different firing rates of MF is considered, which means the neuromorphic model receives different levels of stimulation. Here, different levels of MF are used to investigate the change in x_s of PKJ neurons with the increment of (a) $w_{BS\to PKJ}$ and (b) $w_{GrC\to PKJ}$. In addition, different levels of MF are used to investigate the change in x_{lr} of PKJ neurons with the increment of (c) $w_{BS\to PKJ}$ and (d) $w_{GrC\to PKJ}$

In order to describe the dynamical behaviors of the neural system quantitatively, a linear response criterion x_{lr} is defined as

$$x_{lr} = \sqrt{\left(x_{lr}^{\sin}\right)^2 + \left(x_{lr}^{\cos}\right)^2} \tag{16}$$

where $x_{lr}^{\sin}$ and $x_{lr}^{\cos}$ are

$$\begin{cases} x_{lr}^{\sin} = \dfrac{2}{Tt}\sum_{n=1}^{Tt} E(n)\sin(\omega n) \\ x_{lr}^{\cos} = \dfrac{2}{Tt}\sum_{n=1}^{Tt} E(n)\cos(\omega n) \end{cases} \tag{17}$$

where $\omega=2\pi/t$ is the angle frequency of the oscillation.

The linear response criterion x_{lr} reflects the resonance dynamics of a neural population and are depicted in Fig. 11(c) and Fig. 11(d). The resonance situation of the PKJ population is enhanced with synaptic weight decrement from BS to PKJ neurons, and with the increment from GrC to PKJ neurons. Interestingly, different levels of $w_{BS\to PKJ}$ and $w_{GrC\to PKJ}$ will influence the resonance dynamics conversely along with their increment. At low level of $w_{BS\to PKJ}$, the increment of MF firing rate will improve the resonance within the PKJ population, while it decreases the resonance at a high level of $w_{BS\to PKJ}$ that is larger. The enhancement of MF firing can depress the resonance of PKJ neurons at low $w_{GrC\to PKJ}$, and improves it when $w_{GrC\to PKJ}$ is larger.

C. *Neuromorphic learning of the PKJ and VN cells*

In order to explore the learning-induced change of PKJ cells and the corresponding VN responses, the firing rates of both PKJ cells and VN cells are investigated at the 1st, 50th, 100th, 150th, 200th, 250th and 300th cycles of MF input oscillation in Fig. 12(a). With the cycle number increasing, the maximal firing rate has a moderate change from 88.5 to 75.62 Hz, and the minimal firing rate decreases significantly from 83.59 to 34.7 Hz. Therefore, the firing modulation of the PKJ cell is caused by the decrease of the minimal firing rate, which is consistent with the firing rate change of PKJ neurons in OKR adaption [40]. The firing activities of the VN cells are shown in Fig. 12(b). Due to the modulation of the inhibition affects from the PKJ cells, the firing dynamics of the VN neurons are modulated in phase with the MF oscillation. The maximum firing frequency is increased from 60.37 to 109.6 Hz, and the minimum firing rate of the VN neuron changes from 23.17 to 42.43 Hz.

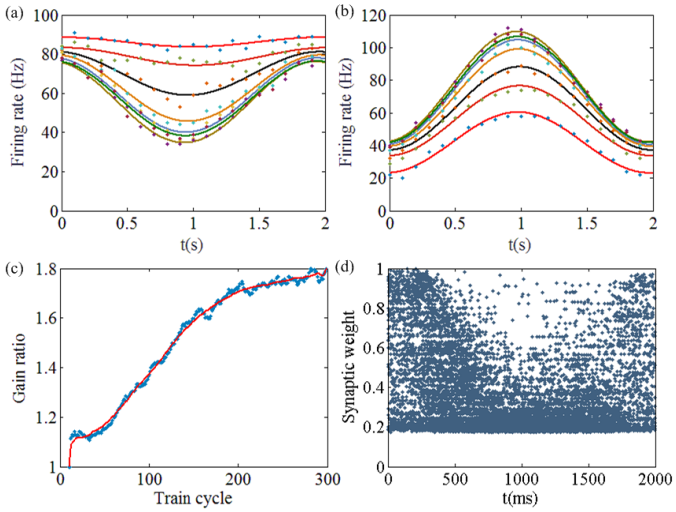


Fig. 12. Neuromorphic learning of OKR adaption experiment. Colored curves from top to bottom in turn represent the learning-induced firing rate change of (a) the PKJ cells and (b) the VN cells at the 1st, 50th, 100th, 150th, 200th, 250th and 300th cycles of MF signal oscillation. (c) Gain change ratio according to the MF train cycle number is shown. Here, the blue dots are discrete data of the gain ratio, while the red solid line represents a fitted curve. (d) Distribution of synaptic weights between PKJ cells and active GrC cells after 300 cycles of MF oscillation.

In order to explore the VN modulation by the MF signals, the gain ratio is defined as the modulation of the VN spiking activities at each cycle divided by that at the first cycle. Due to the inhibition of the IO information, the neuromorphic learning is changed with MF oscillation, and the gain ratio reaches 1.791 as shown in Fig. 12(c). The distribution of synaptic weights of active GrC neurons is shown in Fig. 12(d). The synaptic weights distribute uniformly from 0.175 to 1.0 at the beginning and the end of a modulation cycle, and most of the synaptic weights locate between 0.175 and 0.5 at the middle of a MF cycle with the largest MF and CF inputs. Therefore, the firing modulation of PKJ cells is induced by both the spatial distribution of the synaptic weights between PF and PKJ cells, and feedforward inhibition by the BC neurons.

D. Analysis of dynamic response in neuromorphic GrC layer

In order to explore the mechanism underlying the response of GrCs to temporary oscillations, we use the proposed model to simulate an OKR adaption experiment using retinal slip signal inputs. As shown in Fig. 13(a), in the OKR adaption experiment, MFs and CFs transmit retinal slip information that oscillates in real-time. It is based on a schematic of the neural circuit involved in OKR adaption presented by a previous experimental study [31]. From the start of a cycle of the sinusoidal oscillations, various groups of GrCs are activated one by one sequentially. In this study, Poisson spikes that oscillate sinusoidally at 0.5 Hz are input to the MFs according to the previous biophysiological study [40]. LTD shapes the spatial distribution of PF-PKJ cell synapses in a sinusoidal form, inducing the response of the PKJ cells to gradually increase sinusoidal modulation, which is consistent with the previous experimental study [40]. As shown in Fig. 13(b), at the beginning and end of a cycle of signal oscillation of the MFs, the firing rate of MF is low, so that the GrCs spike uniformly in a random manner. With an increment of the MF firing rate, the GrCs are activated with firing rates that are basically proportional to the firing activities of the MFs. This reveals that the GrCs can transmit MF amplitude information to the PKJ cells effectively. Based on this mechanism, the PKJ cells can learn both the scalar information including timing and gain signals, and the complete waveform processed by the CFs, so that gain and timing control could be unified.

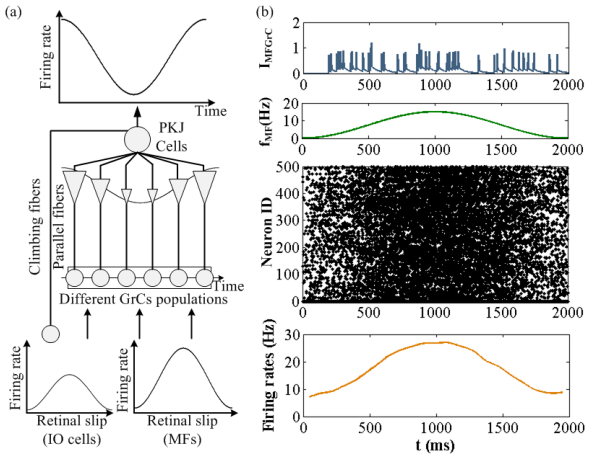


Fig. 13. Dynamics of the GrCs in response to sinusoidally oscillating MF input information at 0.5Hz. (a) The simulation results of OKR adaptability experiment of retinal slip signal input using the model proposed in this paper. (b) Spike patterns and firing rates of 500 granule cells during a cycle of MF signal oscillation. Black dots indicate spike discharges.

Because of the random recurrent connections between GrC and GoC cells, individual GrC neurons show a variety of temporal spiking activities as shown in Fig. 14(a), which shows the population of active GrC neurons gradually changes in time. It shows the firing rates of different GrC neurons are temporally fluctuating in response to MF signals, leading to the dynamics of passage-of-time in cerebellum. To evaluate this property, the similarity index $S(\Delta t)$ between active GrC neurons is calculated. The experimental result shown in Fig. 14(b) illustrates that the similarity decreases monotonically with Δt increasing. It reveals that the temporal change in the GrC layer is nonrecurrent, suggesting the one-to-one correspondence between an active GrC population and a time step under MF signal oscillation.

Although the passage-of-time dynamics exist in the GrC layer, the generation of temporally fluctuating spikes is reproducible under MF signal oscillation. The reproducibility index $R(t)$ between two spike patterns for all GrC neurons for two consecutive cycles is shown in Fig. 14(c). The value of $R(t)$ increases towards 0.9 at the beginning of a cycle, and then decreases slowly towards 0.8, revealing that the spike activities of the GrC neurons are highly reproducible across MF oscillation cycles. The GrC layer is responsible for the generation of the same sequence of active GrC neurons during different trials and cycles for the reliable transmission of MF signals. Because the dynamical activities of a recurrent network are based on external inputs and the initial state, the GrC layer should reset its internal state at the beginning of each cycle. Therefore, slowly increasing MF signals may be enough to reset the internal state.

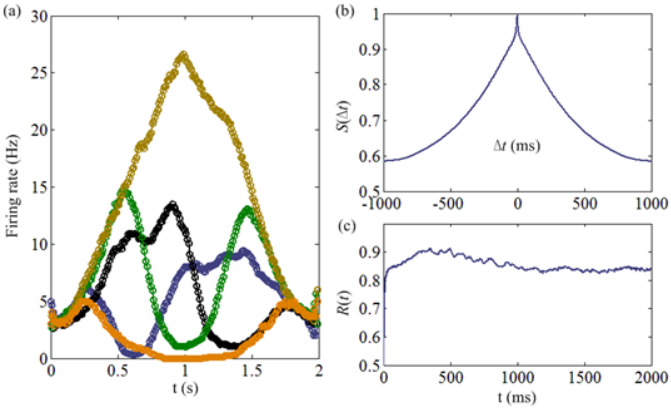


Fig. 14. Dynamical activities of GrC neurons in response to sinusoidally oscillating MF inputs. (a) The averaged firing rate of five representative GrC neurons. (b) The similarity index $S(\Delta t)$ for the spike patterns in GrC layer. (c) The reproducibility index $R(t)$ for the spike patterns for all GrC neurons across two successive cycles of MF oscillation.

E. Robustness analysis of the proposed model

In order to analyze the robustness of the proposed large-scale cerebellar model, the delayed eye blink classical conditioning paradigm is employed, which is divided into two sessions of 100 trials. As shown in Fig. 15(a), each session is composed of an acquisition phase with the presentation of conditioned stimuli-unconditioned stimuli pairs during 80 trials, and an extinction phase with the presentation of only conditioned stimuli for 20 trials. The inter-spike interval (ISI) is set to 300 ms, 400 ms, and 500 ms, respectively. The conditioned stimuli lasted 50 ms, which equals to ISI plus the duration of unconditioned with 100 ms. Between these two consecutive trials, a pause of 100 ms is set to make the network silent. Conditioned stimuli is input from the MFs, and the unconditioned stimuli is input from the IO neurons. As shown in Fig. 15(b)-(d), the first 100 trials are the acquisition phase, and the second 100 trials represent the extinction phase. The criteria %CR means the probability of conditioned response

(CR) from VN neurons during stimuli. It is shown that, the %CR of the proposed network is robust even when the ISR increases. In addition, in the second 100 trials, the proposed model can learn to generate CR more rapidly with its intrinsic plasticity mechanisms. It is also able to characterize learning in a physiological number of acquisition trials, even though some disturbed phenomenon emerged.

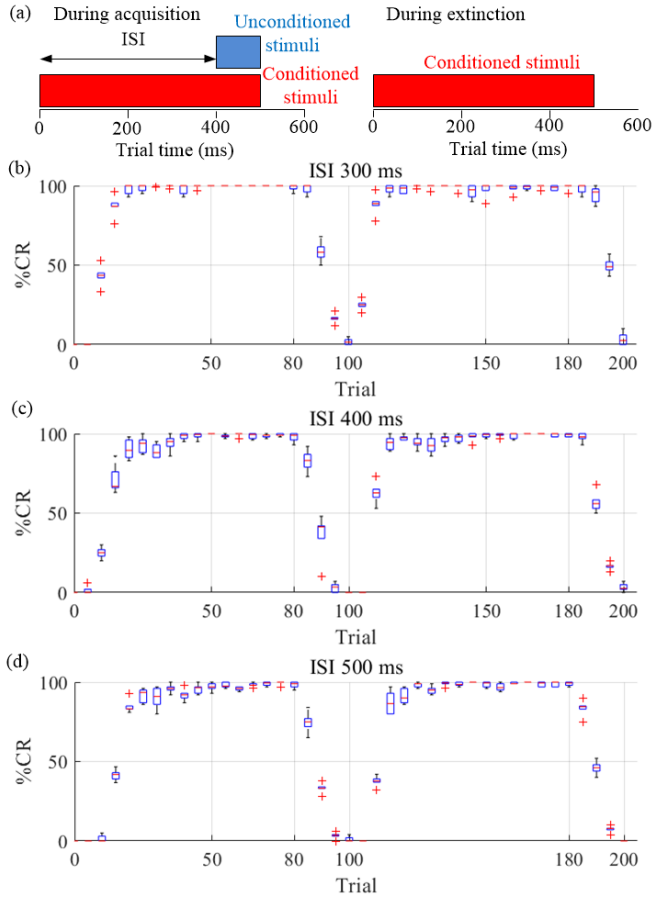


Fig. 15. Model robustness test. The behavioral CR outcomes with different eye blink classical conditioning protocols. (a) Eye blink classical conditioning protocol. (b)-(d) %CR with different ISI values from 300-500 ms.

V. Discussions

The cerebellum has inspired a number of theoretical models focusing on the combinational properties of the SNN due to its regular structure [9], [28]-[30]. Large-scale cerebellar network can generate biological dynamical behaviors including motor learning and memory consolidation. The cerebellum, as a vital region of mammalian brains, operates efficiently, in parallel, performing highly complex motor learning. Emulating the dynamics of cerebellum in a large-scale hardware and utilizing it for replicating brain functionalities and structure can be beneficial to better understanding the brain. It can also be useful in addressing real-world engineering challenges and building smart machines.

This study focuses on building a novel neuromorphic model with high biological relevance, which is inspired by the supervised learning in cerebellum. We developed a model that successfully maps the cerebellar physiological anatomical structure to a digital computing architecture. Real-time cerebellum model can provide a useful means to study a very slow neural process, as well as interact with external world for robotic control or brain-machine interface. As shown in Table 3, compared to previous models with only PF-PC synapses [23, 56, 59], we included more synaptic reversible plasticity sites, based on recent experimental observations at the cellular level [56-58]. The improvement of the proposed model enhances the neurophysiological plausibility and computational learning abilities of the cerebellar circuit. A previous study shows that the differential parts of multiple synaptic sites can reproduce more complex dynamical characteristics of supervised learning than when only a single synaptic plasticity site is considered [25].

Table 3. Comparison with previous real-time spiking cerebellum models

Study	Methodology	#Nucleus types	#Plasticity sites	#Neurons
Carrillo et al., 2008 [25]	EDLUT	8	2	2.1k
Luque et al., 2011 [56]	EDLUT	7	1	2k
Yamazaki & Igarashi, 2013 [23]	GPU Simulation	6	1	100k
Luo et al., 2016 [26]	Frame-based 2D mesh architecture on FPGA	1	None	100k
Antonietti et al., 2016 [39]	EDLUT	7	3	6.5k
Xu et al., 2017 [59]	Frame-based 2D mesh architecture on FPGA	6	1	10k
Hausknecht et al., 2017 [60]	GPU Simulation	8	2	1M
Naveros et al., 2018 [61]	EDLUT	5	2	2.7k
This study	Event-driven neuromorphic 3D BFT architecture	6	3	3.5M

Furthermore, in contrast to previous models [34-36], a novel neuromorphic methodology to model SNNs with a biologically meaningful mechanism is presented. Table 3 demonstrates that EDLUT is used in some previous studies [25, 39, 56, 61], which show high performance in real-time cerebellar motor control. GPU simulation is another powerful approach for real-time cerebellum modeling, which can scale up the network size to include 1M neurons [23, 60]. Neuromorphic computing is based on a non Von Neumann architecture inspired by the computational capabilities of the brain [62]. The brain has evolved to process neural sensory information in a highly parallel and asynchronous fashion, which is the computing principle of neuromorphic hardware. Luo et al. designs a frame-based neuromorphic 2D mesh architecture on FPGA, which can realize 100k large-scale SNN with GrC neurons without synaptic plasticity [26]. Xu et al. further used this neuromorphic architecture to simulate more neuron types while reducing the network scale. This study presents a novel neuromorphic methodology with event-driven 3D BFT neuromorphic

architecture, which can realize 3.5M neurons with 6 nucleus types and 3 plasticity sites. Thanks to the proposed architecture, biologically plausible mechanisms can be reproduced in real time, which is critical in the supervised motor learning function in biological cerebellum.

For our system design, we used a biologically constrained, bottom-up modeling approach for each cerebellar nucleus based on the Marr-Albus-Ito theory of cerebellar function [28]-[30]. By using this approach, the functional properties of the SNNs emerge from the properties of constitutional elements and the synaptic connections [42]-[43]. In order to verify the emergence of the cerebellar motor learning using the developed bottom-up model, several vital network dynamics related to the cerebellar motor learning, such as dynamic response of GrC cells during the OKR experiment, were investigated.

Another significant feature of our model is its high convergence ratio. Our neuromorphic architecture enables biologically plausible cerebellar divergence/convergence ratios. Electrophysiological studies consistently indicate that the cerebellum comprises a network of cells with known sites and learning rules for synaptic plasticity, numerical ratios, convergence and divergence ratios, and geometry of projections [7]. Previous studies reveal that the convergence ratio in the cerebellum is vital for cerebellar cognition [41]. Our proposed cerebellar neuromorphic model achieves a convergence ratio closer to that of the human brain, which means that more biologically realistic dynamics can be achieved using our hardware.

Apart from the above-mentioned benefits, our developed cerebellum hardware has limitations in a stand-alone setting. In order to address its limitations and expand its capabilities, additional brain regions can be combined with the proposed neuromorphic cerebellum to obtain a more complete and powerful model. These include the basal ganglia that are capable of reinforcement learning [43] and the cerebral cortex that is supposed to be responsible for unsupervised learning [44]. By combining these brain regions, deeper and more comprehensive knowledge of brain cognitive functions can be gained and brain structure could be much further explored.

In summary, the large-scale neuromorphic cerebellar model developed in this work and based on a bottom-up modeling approach that explicitly takes the brain-inspired computing architecture into account, (1) provides a theoretical and computational basis toward elucidating motor learning mechanisms with multiple plasticity rules in cerebellar learning with supervised learning capabilities beyond the Marr-Albus-Ito theory, (2) presents a novel perspective on reversely engineered large-scale cognition of the brain, and (3) develops a novel engineering framework for real-time motor learning.

Our LaCSNN hardware can be considered a spike-based HPC platform, which can be used to implement a general spike-based computational intelligence aided design framework. This framework can be utilized to implement various computational intelligence models such as our proposed cerebellar structure and its motor learning, to perform learning and computation in a biologically-plausible fashion. One possible future research direction is to find other performance-critical computational intelligence models, which cannot be efficiently run on conventional computers, and investigate their implementations on our neuromorphic platform.

Other future works may involve further exploration of the motor learning mechanisms of the cerebellum during motor control tasks such as those performed in [47-50]. However, generalization of our model to implement various cerebellar motor learning mechanisms is not straightforward and requires significant changes in our architecture. Hence, we leave exploring other motor learning mechanisms of the cerebellar cognition to the future.

VI. Conclusions

In this study, we presented a large-scale neuromorphic spiking neural network model of the cerebellum that incorporates LTP and LTD mechanisms at the PF-PC synapses, synaptic plasticity at the MF-VN synapses and learning mechanisms located between PKJ and VN cells. A novel neuromorphic system CerebelluMorphic and its architecture is proposed in detail. The MF-VN synapses can update the synaptic weight based on correlations between presynaptic MF activities and postsynaptic VN activities. The digital neuromorphic model consists of approximately 3.5 million neurons, which is 34.6 times more than the state-of-the-art digital neuromorphic design [26]. Our model successfully reproduces experimental results for specifically vital properties in cerebellar motor learning, including motor control with supervised learning ability, dynamic response to OKR adaption, passage-of-time properties and gain control. These properties explain how cerebellar cortex processes the neural information with motor learning cognition.

Appendix

Table A1. Absolute values of the synaptic weights are shown in nS. However, the inhibitory and excitatory connections are negative and positive, respectively.

Pre-	Postsynaptic neuron						
	MF	GrC	GoC	PKJ	BC	VN	IO
MF		4.0				0.002	
GrC			0.00004	0.003	0.003		
GoC		10.0					
PKJ						0.008	
BC				5.3			5.0
VN							
IO				1.0			

Acknowledgment

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

References

[1] E. M. Izhikevich and G. M. Edelman, "Large-scale model of mammalian thalamocortical systems," *Proc. Natl. Acad. Sci. U S A,* vol. 105, no. 9, pp. 3593-8, Mar. 2008.

[2] M. Berzish, C. Eliasmith, and B. Tripp, "Real-Time FPGA Simulation of Surrogate Models of Large Spiking Networks," *Artificial Neural Networks and Machine Learning - Icann 2016, Pt I,* vol. 9886, pp. 349-356, 2016.

[3] E. M. Izhikevich and G. M. Edelman, "Large-scale model of mammalian thalamocortical systems," *Proc. Natl. Acad. Sci. U S A,* vol. 105, no. 9, pp. 3593-8, Mar. 2008.

[4] N. L. Cerminara and R. Apps, "Behavioural significance of cerebellar modules,"*Cerebellum,* vol. 10, no. 3, pp. 484-94, Sep. 2011.

[5] R. Apps and R. Hawkes, "Cerebellar cortical organization: a one-map hypothesis,"*Nat. Rev. Neurosci.,* vol. 10, no. 9, pp. 670-81, Sep. 2009.

[6] M. Glickstein, F. Sultan, and J. Voogd, "Functional localization in the cerebellum," *Cortex*, vol. 47, no. 1, 2011.

[7] M. Ito, "Cerebellar circuitry as a neuronal machine," *Progress in Neurobiology,* vol. 78, no. 3-5, pp. 272-303, Feb-Apr 2006.

[8] M. Ito, "Control of mental activities by internal models in the cerebellum," *Nat. Rev. Neurosci.,* vol. 9, no. 4, pp. 304-13, Apr. 2008.

[9] P. Dean, J. Porrill, C. F. Ekerot, and H. Jorntell, "The cerebellar microcircuit as an adaptive filter: experimental and computational evidence," *Nat. Rev. Neurosci.,* vol. 11, no. 1, pp. 30-43, Jan. 2010.

[10] N. Schweighofer, K. Doya, and F. Lay, "Unsupervised learning of granule cell sparse codes enhances cerebellar adaptive control," *Neuroscience,* vol. 103, no. 1, pp. 35-50, 2001.

[11] J. M. Bower, "Is the cerebellum sensory for motor's sake, or motor for sensory's sake: the view from the whiskers of a rat?" In *Progress in brain research*, vol. 114, pp. 463-496, 1997.

[12] J. H. Gao *et al.*, "Cerebellum implicated in sensory acquisition and discrimination rather than motor control" *Science*, vol. 272, no. 5261, pp. 545-547, 1996.

[13] R. B. Ivry and J. V. Baldo, "Is the cerebellum involved in learning and cognition?" *Current opinion in neurobiology*, vol. 2, no. 2, pp. 212-216, 1992.

[14] N. R. Luque, J. A. Garrido, F. Naveros, R. R. Carrillo, E. D'Angelo, and E. Ros, "Distributed cerebellar motor learning: a spike-timing-dependent plasticity model," *Front. Comp. Neurosci.*, vol. 10, 2016.

[15] M. Ito and M. Kano, "Long-lasting depression of parallel fiber-Purkinje cell transmission induced by conjunctive stimulation of parallel fibers and climbing fibers in the cerebellar cortex,"*Neurosci. Lett.,* vol. 33, no. 3, pp. 253-8, Dec. 1982.

[16] G. Q. Bi and M. M. Poo, "Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type," *J. Neurosci.,* vol. 18, no. 24, pp. 10464-72, Dec. 1998.

[17] D. Jaeger, "No parallel fiber volleys in the cerebellar cortex: Evidence from cross-correlation analysis between Purkinje cells in a computer model and in recordings from anesthetized rats," *J. Comp. Neurosci.,* vol. 14, no. 3, pp. 311-327, 2003.

[18] A. Roth and M. Hausser, "Compartmental models of rat cerebellar Purkinje cells based on simultaneous somatic and dendritic patch-clamp recordings," *J. Physiol.,* vol. 535, no. Pt 2, pp. 445-72, Sep. 2001.

[19] S. Solinas, R. Maex, and E. De Schutter, "Synchronization of Purkinje cell pairs along the parallel fiber axis: a model," *Neurocomputing*, vol. 52, pp. 97-102, 2003.

[20] P. A. Salin, R. C. Malenka, & R. A. Nicoll, "Cyclic AMP mediates a presynaptic form of LTP at cerebellar parallel fiber synapses," *Neuron*, vol. 16, no. 4, pp. 797-803, 1996.

[21] W. Gerstner, and W. M. Kistler, "Spiking neuron models: Single neurons, populations, plasticity," *Cambridge university press*, 2002.

[22] S. Yang, J. Wang, B. Deng, *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.

[23] T. Yamazaki and J. Igarashi, "Realtime cerebellum: a large-scale spiking network model of the cerebellum that runs in realtime using a graphics processing unit,", *Neural Netw.,* vol. 47, pp. 103-11, Nov. 2013.

[24] M. C. Wang, B. Y. Yan, J. Z. Hu, and P. Li, "Simulation of Large Neuronal Networks with Biophysically Accurate Models on Graphics Processors,", *2011 International Joint Conference on Neural Networks (Ijcnn),* pp. 3184-3193, 2011.

[25] R. R. Carrillo, E. Ros, C. Boucheny, and O. J. Coenen, "A real-time spiking cerebellum model for learning robot control," *Biosystems,* vol. 94, no. 1-2, pp. 18-27, Oct.-Nov. 2008.

[26] J. Luo, G. Coapes, T. Mak, T. Yamazaki, C. Tin, and P. Degenaar, "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. 742-53, Jun. 2016.

[27] S. Solinas, T. Nieus, and E. D'Angelo, "A realistic large-scale model of the cerebellum granular layer predicts circuit spatio-temporal filtering properties," *Front. Cell Neurosci.,* vol. 4, p. 12, 2010.

[28] D. Marr, and W. T. Thach, "A theory of cerebellar cortex," In *From the Retina to the Neocortex*. Birkhäuser Boston, pp. 11-50, 1991.

[29] J. S. Albus, "A theory of cerebellar function," *Mathematical Biosciences*, vol. 10, no. 1-2, pp. 25-61, 1971.

[30] M. Ito, "Long-term depression," *Annual review of neuroscience*, vol. 12, no. 1, pp. 85-102, 1989.

[31] F. Shutoh, M. Ohki, H. Kitazawa, S. Itohara, and S. Nagao, "Memory trace of motor learning shifts transsynaptically from cerebellar cortex to nuclei for consolidation,"*Neuroscience,* vol. 139, no. 2, pp. 767-77, May 2006.

[32] W. Gerstner, "A framework for spiking neuron models: the spike response model," *In: Moss F, Gielen S (eds). The handbook of biological physics. Elsevier, Amsterdam*, chap 12 pp. 469 – 516, 2001.

[33] 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. 2279-92, Oct. 2007.

[34] T. Yamazaki and S. Nagao, "A computational mechanism for unified gain and timing control in the cerebellum,"*PLoS One,* vol. 7, no. 3, p. e33319, 2012.

[35] A. R. Gallimore, T. Kim, K. Tanaka-Yamamoto, and E. De Schutter, "Switching On Depression and Potentiation in the Cerebellum,"*Cell Rep.,* vol. 22, no. 3, pp. 722-733, Jan. 2018.

[36] V. Lev-Ram, S. B. Mehta, D. Kleinfeld, and R. Y. Tsien, "Reversing cerebellar long-term depression,"*Proc. Natl. Acad. Sci. U S A,* vol. 100, no. 26, pp. 15989-93, Dec. 2003.

[37] M. Coesmans, J. T. Weber, C. I. De Zeeuw, and C. Hansel, "Bidirectional parallel fiber plasticity in the cerebellum under climbing fiber control,"*Neuron,* vol. 44, no. 4, pp. 691-700, Nov. 2004.

[38] A. Antonietti, C. Casellato, J. A. Garrido, E. D'Angelo, and A. Pedrocchi, "Spiking Cerebellar Model with Multiple Plasticity Sites Reproduces Eye Blinking Classical Conditioning," *2015 7th International Ieee/Embs Conference on Neural Engineering (Ner),* pp. 296-299, 2015.

[39] A. Antonietti *et al.*, "Spiking Neural Network With Distributed Plasticity Reproduces Cerebellar Learning in Eye Blink Conditioning Paradigms," *IEEE Trans. Biomed. Eng.,* vol. 63, no. 1, pp. 210-9, Jan. 2016.

[40] S. Nagao, "Behavior of floccular Purkinje cells correlated with adaptation of horizontal optokinetic eye movement response in pigmented rabbits," *Experimental Brain Research*, vol. 73, no. 3, pp. 489-497, 1988.

[41] W. K. Li, M. J. Hausknecht, P. Stone, and M. D. Mauk, "Using a million cell simulation of the cerebellum: network scaling and task generality," *Neural Netw.,* vol. 47, pp. 95-102, Nov. 2013.

[42] S. Druckmann, Y. Banitt, A. Gidon, F. Schurmann, H. Markram, and I. Segev, "A novel multiple objective optimization framework for constraining conductance-based neuron models by experimental data," *Front. Neurosci.,* vol. 1, no. 1, pp. 7-18, Nov. 2007.

[43] K. Doya, "Complementary roles of basal ganglia and cerebellum in learning and motor control,"*Curr. Opin. Neurobiol.,* vol. 10, no. 6, pp. 732-9, Dec. 2000.

[44] K. Doya, "What are the computations of the cerebellum, the basal ganglia and the cerebral cortex?," *Neural Netw.,* vol. 12, no. 7-8, pp. 961-974, 1999.

[45] A. Giovannucci *et al.*, "Cerebellar granule cells acquire a widespread predictive feedback signal during motor learning," *Nat. Neurosci.,* vol. 20, no. 5, pp. 727-734, May 2017.

[46] Y. Yang and S. G. Lisberger, "Purkinje-cell plasticity and cerebellar motor learning are graded by complex-spike duration," *Nature,* vol. 510, no. 7506, pp. 529-32, Jun. 2014.

[47] A. L. Person and I. M. Raman, "Purkinje neuron synchrony elicits time-locked spiking in the cerebellar nuclei," *Nature,* vol. 481, no. 7382, pp. 502-5, Dec. 2011.

[48] M. Jelitai, P. Puggioni, T. Ishikawa, A. Rinaldi, and I. Duguid, "Dendritic excitation-inhibition balance shapes cerebellar output during motor behaviour," *Nat. Commun.,* vol. 7, p. 13722, Dec. 2016.

[49] D. J. Herzfeld, Y. Kojima, R. Soetedjo, and R. Shadmehr, "Encoding of error and learning to correct that error by the Purkinje cells of the cerebellum," *Nat. Neurosci.,* vol. 21, no. 5, pp. 736-743, May 2018.

[50] T. D. Nguyen-Vu *et al.*, "Cerebellar Purkinje cell activity drives motor learning,"*Nat Neurosci,* vol. 16, no. 12, pp. 1734-6, Dec 2013.

[51] J. Luo, G. Coapes, T. Mak, *et al.*, "Real-time simulation of passage-of-time encoding in cerebellum using a scalable FPGA-based system," *IEEE Trans. Biom. Circ. Syst.*, vol. 10, no. 3, pp. 742-753, 2015.

[52] J. F. Medina and M. D. Mauk, "Computer simulation of cerebellar information processing," *Nat. Neurosci.*, vol. 3, no. 11, pp. 1205, 2000.

[53] J. Voogd and M. Glickstein, "The anatomy of the cerebellum," *Trends Neurosci.*, vol. 21, no. 9, pp. 370–375, 1998.

[54] S. B. Furber, F. Galluppi, S. Temple, *et al.,* "The spinnaker project," *Proc. IEEE*, vol. 102, no. 5, pp. 652-665, 2014.

[55] J. Pei, L. Deng, S. Song, *et al.* "Towards artificial general intelligence with hybrid Tianjic chip architecture," *Nature*, vol. 572, no. 7767, pp. 106, 2019.

[56] N. R. Luque, J. A. Garrido, R. R. Carrillo, S. Tolu, and E. Ros, "Adaptive cerebellar spiking model embedded in the control loop: Context switching and robustness against noise," *Int. J. Neural Syst.*, vol. 21, no. 5, pp. 385–401, Oct. 2011.

[57] J. F. Medina, K. S. Garcia, W. L. Nores, N. M. Taylor, and M. D. Mauk, "Timing mechanisms in the cerebellum: Testing predictions of a large-scale computer simulation," *J. Neurosci.*, vol. 20, no. 14, pp. 5516–5525, 2000.

[58] A. Antonietti, C. Casellato, E. D. Angelo and A. Pedrocchi, "Model-Driven Analysis of Eyeblink Classical Conditioning Reveals the Underlying Structure of Cerebellar Plasticity and Neuronal Activity," *IEEE Trans. Neural Netw. Learn. Syst*., vol. 28, no. 11, pp. 2748-2762, Nov. 2017.

[59] T. Xu, N. Xiao, X. Zhai, *et al.,* "Real-time cerebellar neuroprosthetic system based on a spiking neural network model of motor learning," *J. Neural Eng.*, vol. 15, no. 1, pp. 016021, 2018.

[60] M. Hausknecht, W. K. Li, M. Mauk, *et al.,* "Machine learning capabilities of a simulated cerebellum," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 28, no. 3, pp. 510-522, 2016.

[61] F. Naveros, N. R. Luque, E. Ros, *et al.* "VOR adaptation on a humanoid iCub robot using a spiking cerebellar model," *IEEE Trans. Cybern.*, 2019.

[62] S. Yang, B. Deng, J. Wang, *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, 2019.