2017 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2017, Cancun, Mexico, December 4-7, 2017

Parallelization of a spiking neural network model of layered cortical sheet consisting of multiple cortical regions

Jun Igarashi†

†RIKEN Advanced Center for Computing and Communication
2-1Hirosawa, Wako-shi, Saitama-ken, Japan
Email: jigarashi@riken.jp

Abstract–A parallel computing of spiking neural networks of the cortex at whole-brain scale is a grand challenging in the next decade. In a whole-brain scale simulation, load imbalance and increasing communication of spikes reduce computational efficiency. To overcome the problems, we investigated tile partitioning parallelization of a spiking neural network model of the cortex with layer structure using supercomputer K. We added one communication feature to reduce communication frequency using signal transmission delay of long-range connections. The parallelization showed reduction of communication frequency and elapsed time, and ideal scaling performance. The tile partitioning parallelization may work for a simulation of the cortex at whole-brain scale.

1. Introduction

The brain performs information processing by communication among neurons with nonlinear dynamical property. A large-scale simulation of a spiking neural network using nonlinear neuron model is one of ways to investigate contribution of nonlinearity of neurons. Thanks to exponential growth of recent computational performance according to Moore's low, the number of neurons in large-scale simulation has reached to more than 1 billions [1], and still growing year by year toward simulation at whole-brain scale including 100 billions of neurons to investigate the brain in which neurons work as a whole for its function.

However, there are problems in realizing the whole-brain simulation, for instance, increase in the communication overhead of spike time information, insufficient memory for representing whole neurons and synapses, and computing resources to calculate neurons and synapses.

In this study, we investigate parallel computing of a spiking neural network model of the cortex that is one of major parts in the brain, focusing on spatial partitioning of the model and efficient communication of spike time information.

2. Scenario of parallelization of cortical sheet

2.1 Overview of physiological property of the cortex

First, we summarize physiological features of the cortex to be considered for parallelization of a model of the cortex in the next section.

The cortex accounts for large part of the volume (~80% for human) and the numbers of neurons (~20% for human) in mammalian brains [2]. The cortex is located in the outer part of the brain, and the neurons in the cortex are located within the depth from 1-2 mm from the brain surface, which forms sheet-shape structure over a hemisphere, which is called "cortical sheet". The cortical sheet consists of 6 layers that are hundred microns of thickness. Numbers of neurons, neuron types and connectivity differ across layers. The cortical sheet is divided into multiple cortical regions which are specialized for different information processing, sensing, motor control, decision, and so forth. Patterns of afferent and efferent connections also differ depending on cortical regions. Cortical neurons receive/send thousands of synaptic connections per a neuron. Connection probability is high between neighboring neurons in the same cortical region, and low between neurons in different cortical regions [3].

2.2. Tile partitioning of cortical sheet

In physical simulations, spatial partitioning methods are used for parallelization. Here, we consider tile partitioning, which is one of the spatial partitioning method that divides target plane into tiles. The tiles are calculated by computational nodes of parallel computer in parallel with communicating information among them. In this section, we consider the advantage of the tile partitioning of cortical sheet, and propose communication technique combined with the tile partitioning.

2.2.1. Load balancing and neighboring connectivity

In the cortices, the cell density, and the numbers of synaptic connections per mm^2 are in a similar range. Tile partitioning cortical sheet makes cortical tiles with similar amount of the above-mentioned neural elements, which may work for load balancing in parallel computing. Tile partitioning may also contribute to reduction of communication among computational nodes by putting connected neighboring neurons in the same cortical region together into the same computational node.

2.2.2. Communication of spike time information and postsynaptic currents

In calculation of postsynaptic currents (PSCs) in parallel computing, communication of spike time information between computational nodes must finish during signal transmission delay (STD). In the condition, it is possible that spike time information is kept in computational nodes with presynaptic neurons during STD, and send it to the computational nodes with postsynaptic neurons at once before calculation of PSCs, which leads to reduction of communication frequency. Representative neural simulators, NEST [4] and NEURON [5] have the feature to decrease in frequency of communication using minimum STD in all connections, where a typical minimum STD is 1ms.

In tile partitioning of cortical sheet, frequency of communication of spike time information can be reduced more by communicating between only tiles with connected neuron pairs. STDs between connected neurons in one pair of tiles range in similar extent because the STDs are almost determined by the distance between the tiles. Then, spike time information can be kept for longer STD in a pair of distant tiles, and reduce frequency of communication (Fig. 1, A to C) compared with neighboring tiles (Fig. 1, A to B). To investigate whether the reduction of communication frequency is effective for parallel computing of a cortical sheet model, we implemented the communication function of variable communication frequency depending on minimum STD of connections within each pair of tiles.

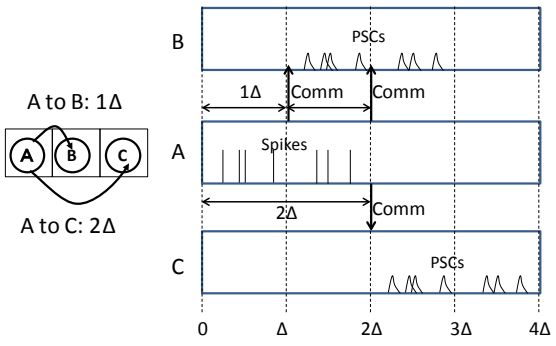

Figure 1 Variable communication frequency of spike time information depending on STDs. Left: Three tiles and connected neural population A, B and C. STD of A to B (1Δ) is a half of that of A to C (2Δ). Right: Spikes of A and the following PSCs in B and C. "Comm" denotes onset of communication. Note that the "Comm" appears twice in neighboring connection A to B, and once in long-range connection A to C.

3. Model description and parallelization

3.1. A spiking neural network model of cortical sheet with layer structure

Based on anatomical and electorophysiological data of mouse primary motor cortex, density of neurons and thickness of layers [6], neuron types [7], connections between neurons [8-11], we developed a model of cortical sheet that consists of 4 neighboring cortical regions of a regular square shape, assuming primary motor cortex (M1), secondary motor cortex (M2), somatosensory cortex (S1), and secondary somatosensory cortex (S2) (Fig. 2 right). The all cortical regions have the same structure in layer, neuron types, and connection probability of primary motor cortex.

Different neuron types were set for different layers, elongated neurogliaform cells (ENGCs) and single bouquet cells (SBCs) in layer 1 (L1)[11], corticocortical cells (CCs), fast spiking neuron (FSs), low threshold spiking neurons (LTSs) in layer 2/3 (L2/3), corticostariatal neurons (CS), CC, FS, LTS in layer 5A (L5A), pyramidal-tract neurons (PT), CC, FS, LTS, in layer 5B (L5B), corticothalamic cells (CT), FS, LTS, in layer 6 (L6). The cell density of neurons (count/mm^2) were as follows: L1 SBC:1259, L1 ENGC:540, L2/3 CC:14659, L2/3 FS:2290, L2/3 LTS:1374, L5A CS:1702, L5A CC:1702, L5A CT:1702, L5A FS:774, L5A LTS:516, L5B PT:3036, L5B CS:3036, L5B CC:3036, L5B FS:1822, L5B LTS:1215, L6 CT:14102, L6 FS:1763, and L6 LTS:1763.

We used integrate-and-fire neuron model for all types of neurons.

$$\tau_m \frac{dv}{dt} = -v + I_{syn} + I_{bias} \qquad (1)$$

$$if \quad v > v\_threshold, \quad then \quad v \leftarrow v\_reset$$

v, I_{syn}, I_{bias}, and τ_{m} are membrane potential, synaptic current, bias current, and membrane time constants, respectively. Membrane time constants were set to 10 ms for FS and 20ms for the other neurons. Amount of excitatory constant bias current was randomly set so that mean firing rate of neurons was about 10 Hz. We used conductance-based synapse modeled by an alpha-function kernel, where the time constants were 2 ms for excitatory synapses and 5ms for inhibitory synapses.

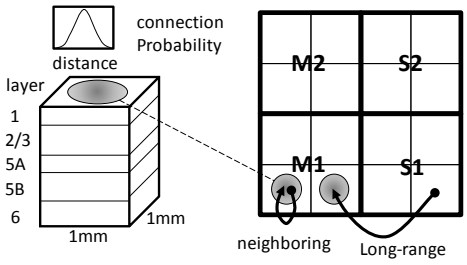

Figure 2 Network architecture of a spiking neural network model of cortical sheet. Right, cortical sheet consisting of multiple cortical regions assuming M1, M2, S1, and S2. S1 projects long-range connection to M1. Left, layered structure of one tile.

We set intra-regional connections within each cortical region using Gaussian probability functions of Euclidean distance between neurons that were estimated from the experimental data [8-11]. Average numbers of connections was about 6000 per neuron, and total number of connections was about three hundred million per 1mm^2 tile of the cortical sheet.

We set inter-regional connections from layer 2/3 CC and layer 5A CC in S1 to layer 2/3 CC, FS and LTS in M1 [12]. The inter-regional connection was also set using Gaussian probability functions of Euclidean distance between neurons in topological manner assuming topological connections between somatotopic maps of M1 and S1 [13].

STD was determined by STD = D *R_{acv} + Δ_{sd}, where D, $R_{acv,}$ and Δ_{sd} represent Euclidean distance between neurons, reciprocal of axonal conduction velocity (0.001 ms/ μ m), and synaptic delay (1.5 ms), respectively. Forward Euler method used for numerical calculation of neurons and synapse models. Calculation step was 0.1 ms.

We implemented the model of the cortical sheet using C programming language and MPI communication library. One mm^2 of tile was assigned to one computational node. The total sizes of the cortical sheet used in this study were, 16, 64, 1024, 4096 mm^2, which were regular square shapes (Fig. 2, right). The cortical sheet was equally divided into 4 cortical regions with a regular square shape, assuming M1, M2, S1, and S2 (Fig. 2, right).

Communication interval of spike time information was set according to minimum STD for each pair of connected tiles. To perform asynchronous communication, we set communication interval to a half of minimum STD consisting the phase to keep spike time information and the phase to send them asynchronously.

3.2. Calculation environment

For compiling the C program of the cortical sheet, Fujitsu C compiler, mpifccpx was used. Communication between computational nodes was performed using asynchronous communication, Isend, and Irecv function. Single precision floating-point number was used for representing state variables of neuron and synapse models. Calculation time was measured using “gettimeofday” function of C programming language.

For calculation of the cortical sheet, we used the K computer which is located at Riken Advanced Institute for Computational Science (AICS) in Kobe, Japan [14]. The K computer has 88,128 CPUs and 1.4 peta byte of DRAM memory. One computational node includes the one CPU and 16 giga Byte memory. The CPU consists of 8 cores and runs at 2 GHz. Computational nodes are connected by 6-dimensional mesh/torus interconnect (Tofu), whose peak interconnect link bandwidth is 5GB/s. The operating system of the K computer is customized Linux.

4. Results

4.1. Variable communication frequency using different STDs of pairs of tiles

We tested the variable communication frequency for different STDs of connections in tile partitioning parallelization, which is described in 2.2, using 16 mm^2 of cortical sheet model. We counted the numbers of calls of MPI communication of spike time information through long-range connection from S1 to M1 (Fig. 3).

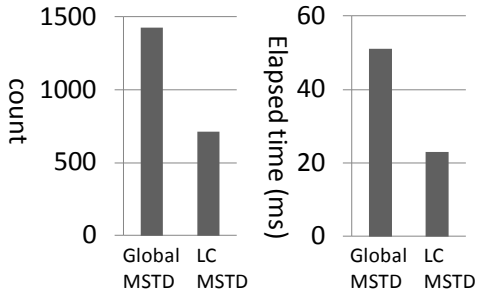

Figure 3 Reduction of communication frequency of spike time information in long-range connection. Left, counts of call of mpi communication function for global minimum STD (MSTD) and long-range connection’s minimum STD (LC MSTD). Right, elapsed time.

When minimum STD in whole network was used without variable communication frequency, the numbers of the call was 1429. On the other hand, when minimum STD between the S1 and M1 tiles was used for variable communication frequency, the number of the call was 715 (Fig. 3 left). The elapsed times for the communication were 51 ms for minimum STD in whole network, and 31 ms for minimum STD between S1 and M1 tiles (Fig. 3 right). This result suggests that the variable communication frequency was effective for reduction of communication frequency and actual elapsed time.

4.2 Scaling performance of tile partitioning of cortical sheet

Next, we tested whether the tile partitioning is effective for larger scale of simulation of cortical sheet. We measured calculation times with changing the size of the cortical sheets from 64 (8x8) to 4096 (64x64) mm^2 and the number of computational nodes from 64 to 4096 with fixed assignment of 1mm^2 tile to one computational node, which is a way of investigating parallel computing performance, weak scaling performance.

Fig. 4 shows the calculation times of the simulations of 1 second of biological time for different sizes of cortical sheet. The computational times were on the almost same level even with increase in the size of cortical sheet and computational nodes. This result demonstrated that the tile partitioning works for scaling the size of cortical sheet in parallel computing system.

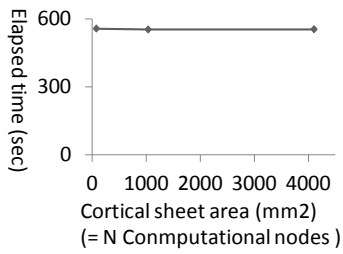

Figure 4 weak scaling performance of cortical sheet

5. Discussion

We investigated effectiveness of tile partitioning of the model of cortical sheet with additional communication method using STD.

Although we tested only one inter-regional connection, there are various inter-regional connections with different STD [3]. We should be able to reduce communication frequency more for inter-regional connection with longer STDs. In the next step, we will test variable communication frequency for multiple of inter-regional connections with different STDs and extend the cortical sheet to the size of whole-brain scale.

Acknowledgments

This work is supported by FLAGSHIP 2020 Project, Post-K Exploratory challenge 4.

References

[1] S. Kunkel et al., "Spiking network simulation code for petascale computers," *Front. Neuroinfor.*, vol.8:78, 2014.

[2] S. Herculano-Houzel, "The human brain in numbers: a linearly scaled-up primate brain," *Front. Hum. Neurosci*., vol.3:31, 2009.

[3] S. W. Oh et al., "A mesoscale connectome of the mouse brain," *Nature*, vol.508, pp.207-214, 2014.

[4] M.-O. Gewaltig and M. Diesmann, "NEST (Neural Simulation Tool)," *Scholarpedia* , 2(4):1430, 2007.

[5] N. T. Carnevale and M. L. Hines, "The NEURON Book," Cambridge, UK: Cambridge University Press, 2006.

[6] D. L. Lev, and E. L. White, "Organization of pyramidal cell apical dendrites and composition of dendritic clusters in the mouse: emphasis on primary motor cortex," *Eur. J. Neurosci*., vol.9, pp.280-290, 1997.

[7] M. J. Oswald et al., "Diversity of layer 5 projection neurons in the mouse motor cortex," *Front. Cell. Neurosci*. vol.7:174, 2013.

[8] N. Weiler et al., "Top-down laminar organization of the excitatory network in motor cortex," *Nat. Neurosci*. Vol. 3, pp.360-366, 2008.

[9] D. Kätzel et al., "The columnar and laminar organization of inhibitory connections to neocortical excitatory cells," *Nat. Neurosci*. vol.1, pp.100-107, 2011.

[10] A. J. Apicella et al., "Laminarly orthogonal excitation of fast-spiking and low-threshold-spiking interneurons in mouse motor cortex," *J. Neurosci.* vol.32, pp.7021-7033, 2012.

[11] X. Jiang X, et al., "The organization of two new cortical interneuronal circuits," Nat Neurosci., vol.16, pp.210-218, 2013.

[12] B. M. Hooks, et al., "Organization of cortical and thalamic input to pyramidal neurons in mouse motor cortex," vol.33, pp.748-760., 2013.

[13] M. Takada et al., "Organization of Two Cortico–Basal Ganglia Loop Circuits That Arise from Distinct Sectors of the Monkey Dorsal Premotor Cortex, " *INTECH*, Book "Basal Ganglia - An Integrative View," chapter 5, 2013.

[14] Top500 K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect, https://www.top500.org/system/177232