Microprocessors and Microsystems 000 (2016) 1–11

Contents lists available at ScienceDirect

Microprocessors and Microsystems

journal homepage: www.elsevier.com/locate/micpro

The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells

Giordana Florimbi[a,*], Emanuele Torti[a], Stefano Masoli[b], Egidio D'Angelo[b], Giovanni Danese[a], Francesco Leporati[a]

[a] Department of Electrical, Computer and Biomedical Engineering, via Ferrata 5, I-27100 Pavia, Italy
[b] Dipartimento di Scienze del Sistema Nervoso e del Comportamento, via Forlanini 6, I-2700 Pavia, University of Pavia, Italy

ARTICLE INFO

Article history:
Received 21 January 2016
Revised 10 May 2016
Accepted 27 May 2016
Available online xxx

Keywords :
Application specific processors
Cerebellar neuron simulation
Neuromorphic elaboration
FPGA
GPU
European projects in digital systems design

ABSTRACT

Studying and understanding human brain is one of the main challenges of 21st century scientists.

The Human Brain Project was conceived for addressing this challenge in an innovative way, enabling collaborations between 112 partners spread in 24 European countries.

The project is funded by the European Commission and will last until 2023.

This paper describes the ongoing activity at one of the Italian units focused on innovative brain simulation through high performance computing technologies. Simulations concern realistic models of neurons belonging to the cerebellar cortex. Due to the level of biological realism, the computational complexity of this model is high, requiring suitable technologies. In this work, simulations have been conducted on high-end Graphical Processing Units (GPUs) and Field Programmable Gate Arrays (FPGAs). The first technology is used during model tuning and validation phases, while the latter allows to achieve real time elaboration, aiming at a possible development of embedded implantable systems. Simulations performance evaluations are discussed in the result section.

1. Introduction

The knowledge of neurophysiological principles that define the brain activities and allow to learn, understand, behave and remember are one of the most addressed challenges for neuroscientists and engineers. In particular the development of a model able to simulate accurately the brain behaviour is one of the targets of the next decade research.

In the central nervous system, brain circuits consist of neurons and synapses characterised by complex dynamic properties. Traditional computational approaches, like artificial neuronal networks, are built on connectivity rules featuring very simplified neurons. On the other hand, brain circuits work through complex sequences of not continuous signals while analysis of usual simulations introduce continuous signals.

Thus, a different point of view is required to understand brain processing and this is the reason why several research groups work on these issues trying to describe and simulate/emulate neurons behaviour following that approach called *neuromorphic computing.*

With this term it is meant both electronic networks/components able to simulate as much accurately as possible brain circuits but also architectures reproducing neuron physiology and in particular their low power, low size and very efficient computational capability [1–3]. While in the first case existing hw/sw technologies are employed to reproduce neurons behaviour, in the second case a complete rethink of traditional hw/sw implementations is required.

Those solutions that can be found in literature can be roughly assorted in two main groups featuring analog/digital design aiming at emulating neurons characteristics with a nanoscale design (neurochips) [4] or alternatively larger scale systems reproducing the neuron outputs (neurocomputers) [1].

Another issue is related to the comprehension and the description of neuronal properties and in particular how molecular and cellular mechanisms together with network connectivity can influence human capabilities like awareness, sensorimotor, cognition, emotion and so on [5,6]. The level of attention of the scientific community has been raised up significantly and several projects have been promulgated to explore models (Human Brain Project [7]), cellular imaging recording (Active Brain Mapping [8]) and connectivity (Human Connectome Project [9]). Other projects that similarly have been done or are still in progress are the American BRAIN project and the Blue Brain Project at EPFL [10].

In this paper we present the first results of the *realistic computational modeling* approach applied to the development of

[*] Corresponding author.
E-mail address: giordana.florimbi01@universitadipavia.it (G. Florimbi).

http://dx.doi.org/10.1016/j.micpro.2016.05.015
0141-9331/© 2016 Elsevier B.V. All rights reserved.

neurocomputers within the Human Brain Project Italian unit located at the University of Pavia. The work has been mainly done on cerebellar neurons who play a primary role in motor control, but will be extended to Golgi, Purkinje, basket and stellate neuronal cells. The technologies chosen for implementing computational units are GPUs for their implicit parallel architecture which resembles neurons organisation while providing fast tuning of models (brain computing). Then those models are implemented on FPGA for designing suitable hardware architectures that could turn out in final implantable neurochips providing further acceleration and close real time elaboration (neuromorphic computing).

First results show how GPUs significantly accelerate computation times with respect to modern multicore PCs and allow to perform effective simulations with significant number of neurons without using supercomputing sometimes expensive and not always available. On the other hand FPGAs demonstrate that only a specifically designed hardware chip is real time compliant but in this case significant simulations will be only possible through important availability of resources (i. e. cluster of processors).

The paper is organised as follows: Section 2 depicts challenges and platforms on which the Human Brain Project is based. Section 3 describes the physiological model of the implemented granular cells and relative synapses while in Section 4 GPU and FPGA implementations are described together with results in terms of processing times and quality of representation.

In particular, we characterized the computational complexity of the model in terms of execution time as function of the number of simulated neurons.

Moreover, we described the behaviour of a granule cell after different current injections.

Then, some concluding considerations together with the future developments are given in the final section.

2. The human brain project

The Human Brain Project (HBP) is the European Commission Future and Emerging Technologies Flagship that involves a consortium of 112 partners spread in 24 European countries. It aims at extending human brain knowledge, helping in identifying and diagnosing brain disorders, as well as designing new brain-inspired technologies [11].

The HBP has its origin in previous European projects on brain simulation and neuromorphic computing like FACETs (Fast Analog Computing with Emergent Transient States under the 6th Research Framework Program of the EU), Brain-i-Nets (Novel Brain-Inspired Learning Paradigms for Large-Scale Neuronal Networks under EU FP-7), BrainScaleS (Brain-inspired multiscale computation in neuromorphic hybrid systems under EU FP-7) and on the Swiss Blue Brain project.

Since the target of the Blue Brain Project was to explorate the feasibility of a large-scale supercomputer implementing biologically realistic simulations, the polytechnic of Lausanne was naturally established as the coordinator of HBP.

The HBP proposal was accepted in January 2013. It foresees a ramp-up phase, establishing collaborations among the network of researchers involved and establishing the structures for managing the project and the relationships with government entities.

This phase started on October, 2013, and will run until 30 September, 2016. The successive (operational) phase will begin in April 2016 and will last until September 2023.

The ramp-up phase has been subsidised with 54 M€ while the budget given for the overall project is 1,19 B€.

This project encourages a large-scale collaboration based on data sharing among multidisciplinary scientists and groups. It will be realized through the development of an integrated system of ICT-based research platforms which contain different types of data describing brain and diseases. Collected data concern different levels of organization, providing a *bottom-up* approach, giving information related to every single area of the brain (Fig. 1).

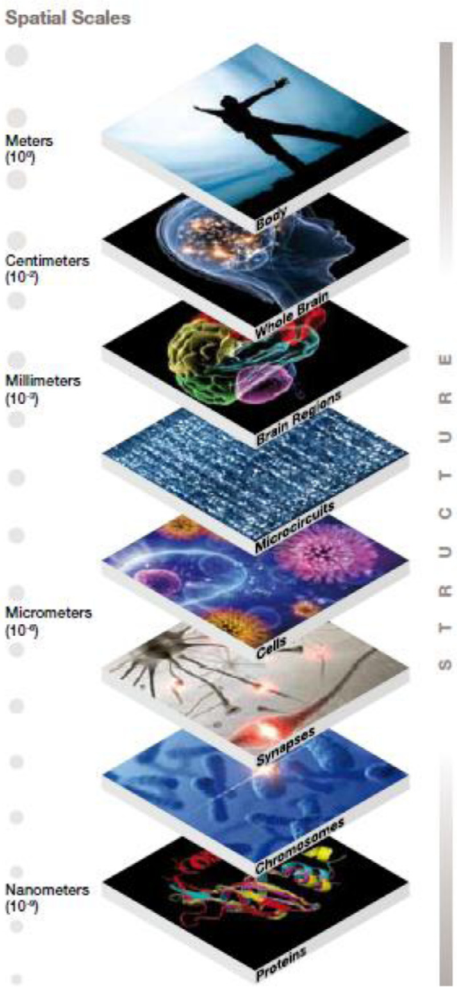

Fig. 1. Data concerning different levels of organization from proteins to the body [12].

In other words, collected data describe brain from the smallest elementary units (neurons and synapses) to the most complex ones (microcircuits and brain regions).

These platforms will be available to neuroscientists and neuroinformatics which can develop and simulate new mathematical models and integrate their experimental results [12]. What is important to identify with these simulations are the casual mechanisms which permit to understand, for example, the synaptic plasticity or the way the brain elaborates information from different sensorial inputs. These mechanisms are usually studied through in vivo experiments; however not always feasible since invasive and expensive [10]. For these reasons, researchers have to develop a simulator which allows to merge experimental data and to make in silico experiments which are impossible in the laboratory [11].

Moreover, in medicine, the simulator also allows researchers to analyse the *biological signatures* of psychiatric and neurological diseases. In particular, in silico experiments can help researchers to configure brain models to match biological signature of disease and to incorporate new information as their causes or effects. The simulator also allows to evaluate possible and/or innovative drug treatments and their effects at all levels of brain organization and in different animal species. Indeed, the simulator can represent a way to simplify the translation from animal drugs to human ones

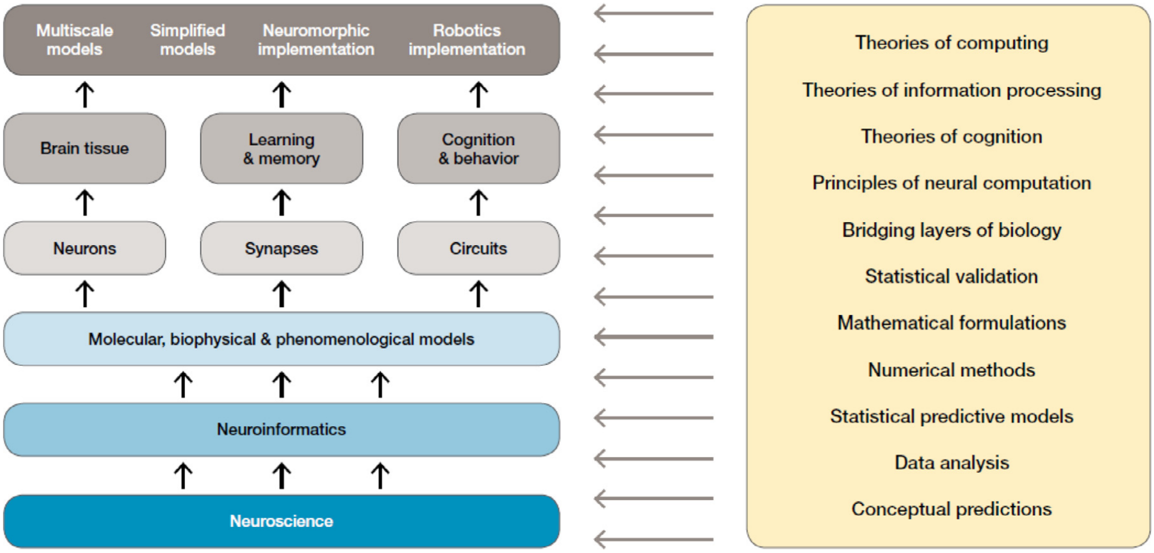

Fig. 2. Disciplines involved in the project [12].

for scientists and pharmaceutical companies [11]. A final turnout could be the introduction of a *personalised* neurology and psychiatry.

From the computer technology point of view, brain understanding will allow to develop innovative neuromorphic and neurorobotic systems based on neurons activity and brain elaboration capabilities. These systems will use the same basic principles of computation and the same cognitive architecture as the brain does. Moreover, these architectures will achieve high-energy efficiency and fault tolerance, together with learning and cognitive capabilities. Neurorobotic systems will be controlled by those devices, enabling a new category of closed-loop tools to analyse how different levels of brain organization affects behaviour and cognition [11].

Fig. 2 shows the different levels into which the project is structured foreseeing the modelisation of both elementary components and functions implemented through them. Fig. 2 shows also the implicit multi-disciplinary nature of the project that involves several contributes coming from Medicine, Mathematics, Computer Science, Biology, Electronics, Psychology.

The HBP structure foresees the development of six ICT platforms among the thirteen that globally constitute the entire project, specifically:

- Neuroinformatics (brain data mapped on atlases);
- Brain Simulation (focusing brain circuits but also functions);
- Medical Informatics (clinical data analysis vs. diseases);
- Neuromorphic Computing (hw implementation of brain functions);
- Neurorobotics;
- High Performance Computing (suitable computational power arrangements).

The platforms are also designed for use by scientists from outside the HBP Consortium who may have limited technical expertise. The HBP will develop technical support and training programmes for users. Access to the platforms will be through a competitive process similar to current methods for allocating time on major scientific instruments (e.g. telescopes).

The scientists involved into the project are formally undertaken to ensure ethical use of data and responsible research.

For sure the challenges that HBP faces are not trivial, starting from the required resources in terms of power supply (order of MW), power computation (ExaFlops) and memory size (multi scale simulation).

But also if we consider the difference from present solutions in terms of processors and algorithms we must conclude that the aim is to build a super machine. This will be made up of highly configurable custom computing units, able to learn. This will lead to biologically realistic algorithms that can turn out in an intelligent behaviour and used to carry out intelligent systems.

The role played by Italy into the project consists on the contributions of five italian units involved, i. e.:

- 3D mapping of the neuronal network, exploiting innovative optical microscopy technologies with resolution higher than traditional ones (LENS, European laboratory of non linear spectroscopy – University of Florence);
- neuromorphic hw system design through nanotechnologies (Polytechnic of Turin);
- collection of big set of data relative to Alzheimer or other neuro-degenerative diseases and analysis on the neuGRID platform so as to promote new approaches to brain pathologies studies (Alzheimer Italian Center Fatebenefratelli Hospital - Milan);
- supercomputing platform for analysis and classification of anatomy, physiology, genomic and other neuroscience related disciplines (CINECA supercomputing center – Bologna);
- new models for neural cell computer simulation (University of Pavia).

Specifically, our team is involved in the Brain Simulation Platform, i. e. scientists are developing innovative mathematical models for realistic neuronal activity description. They are focused on cerebellar neurons such as granules, Golgi, Purkinje, basket and stellate cells. The first step is to describe each cell individually; after that there is a phase of integration of those cells into a single network. This part of the work is based on the well-known Neuron simulator.

Another part of the group is working on exploiting those models on high performance computing technologies. The first step requires a careful analysis of mathematical models and their modification for better adapting to HPC platforms characteristics.

At the beginning granule and Golgi cells are considered in order to reproduce the *Granule layer*. Then Purkinje, basket and stellate cells (*Molecular layer*) are integrated to create an artificial *cerebellar cortex*. A further step is the extension of the model considering the *deep cerebellar nuclei* for obtaining a complete cerebellum and, then, the *inferior olivary nucleus*.

Please cite this article as: G. Florimbi et al., The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells, Microprocessors and Microsystems (2016), http://dx.doi.org/10.1016/j.micpro.2016.05.015

The proposed models are validated exploiting experimental data that were not used in the modelisation process. The implementation foresees at first a parallel GPU implementation for a quick tuning of the algorithms and then (when equations and models are well settled) a FPGA implementation that allows to identify the hw architecture most suitable for close real time elaboration (i. e. in silico neuroscience).

The development of neuronal mathematical models and their large-scale integration (carried on by the Pavia unit) represent a critical task in the entire project. Models will be integrated with experimental data collected by several HBP units so allowing the simulation of the brain functions at different levels of complexity (simple neuron cells description, circuits models, motor and cognitive functions but also the study of the pathologies of the entire human nervous system).

As it can be seen from the previous discussion, the chosen strategy features a “bottom-up” multi-scale modelisation process into which neural functions will be reproduced though realistic models of the single components of the nervous system (i. e. neurons and synapses). These “bricks” will be then linked to carry out those circuits that accomplish brain functions, exploiting available anatomical and functional data. This will turn out in a necessary integration and eventually development or extension of available competences in neurophysiology, cellular biophysics and neurophysiology and finally neuropathology. The initial foreseen activity (30 months) will allow the implementation of a suitable simulation platform applied to the study of the cerebellum and of the cerebellar cortex circuits.

The evolution of this approach will consist in emulate biological signatures of brain pathologies as new configurations of the model.

Finally the Brain Simulation Platform will bring to developing models and application specific processors necessary to reconstruct, first the complete mouse brain and then, the complete human brain.

3. Background and new Granule cell and Synapses models

In literature several models for describing granule cells and synapses are proposed, with different biological plausibility and computational efficiency. All these models can be written using ordinary differential equations (ODE).

The easiest model for spiking neurons is the leaky integrate-and-fire model (LIF), which is widely used due to its high computational efficiency [13]. The model provides a relation between the membrane potential, the synaptic current and the injected current. It is based on a threshold technique: when the membrane potential reaches the threshold value, a spike is generated, after that the potential is reset to the initial value. The main issue of this model is the fixed threshold, that does not consider spikes latency; therefore no biological plausibility is guaranteed and the LIF model is only suitable for proving analytical results.

Another well-known technique for granule cells simulation is the *spiking neuron model* by Izhikevich [14]. The model is described by the two following differential equations:

$$v' = 0.04v^2 + 5v + 140 - u + I \tag{1}$$

$$u' = a(bv - u) \tag{2}$$

where v is the membrane potential of the neuron, u is a variable considering the membrane recovery due to activation of K^+ ionic current and inactivation of Na^+ ionic current; finally, I is the total current in the soma, a describes the time scale of the recovery variable u and b describes the sensitivity of u to the sub-threshold fluctuations of the membrane potential v [14].

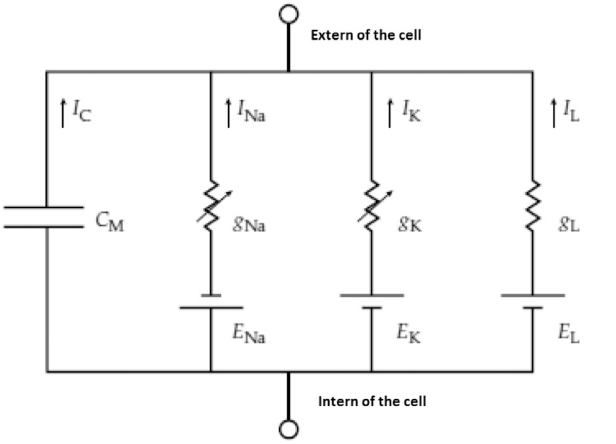

Fig. 3. The original Hodgkin-Huxley model.

Moreover, the after spike reset is modeled by the relation:

$$\text{if } v \geq 30\text{ mV then } \begin{cases} v = c \\ u = u + d \end{cases} \tag{3}$$

where 30 mV is the peak of the spike, c describes the after-spike reset value of the membrane potential v, and d describes after the spike-reset of u [14]. This model can describe different neurons, such as granules, Golgi and Purkinje cells. This model is computationally efficient since it only describes the membrane potential variations. This is a critical issue to achieve biological realism, since the model does not accurately describe all the physiological mechanisms of a neuron.

One of the most complex and accurate model that takes into account different biophysical contributions is the Hodgkin-Huxley (HH) model [15]. The original idea consists of four equations describing membrane potential, inactivation of Na^+ current and activation of Na^+ and K^+ currents. Fig. 3 shows the original proposal.

The membrane capacitance is modeled through a capacitor C_M while ionic channels have been modeled using a resistor. The ionic current I_{ion} is given by:

$$I_{ion} = g_{ion}(V_m - E_{ion}) \tag{4}$$

where g_{ion} is the conductance of the considered ionic channel, V_m is the membrane potential and E_{ion} is the Nernst potential, that does not enable the diffusion of the ion through the membrane.

It is important to notice that the conductance is not constant, since it depends from the membrane potential.

The total current in the circuit is given by the sum of the capacitance current I_C, the ionic currents I_{Na} and I_K calculated according to (4) and the leakage current I_L that takes into account contributions of channels that are always open.

This leads to a differential equation:

$$I = C_m \frac{dV_m}{dt} + \sum\nolimits_{ion} [g_{ion}(V_m - E_{ion})] \tag{5}$$

where I is the total current.

The conductance of each ionic channel is given by:

$$g_{ion} = \bar{g}_{ion}^{\max} x_i^z y_i^k \tag{6}$$

where $\bar{g}_{ion}^{\max}$ is the maximum conductance of the considered channel and x and y are state variables of the gating particles, which specify whether channels are active or not at a given instant of time. These variables assume values from 0 to 1; finally z and k represent the number of activation and inactivation particles of each channel. The probability of a particle of being in a permissive state is not constant, since it depends from two coefficients α_n and

Please cite this article as: G. Florimbi et al., The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells, Microprocessors and Microsystems (2016), http://dx.doi.org/10.1016/j.micpro.2016.05.015

β_n related to the velocity of transition. The relation is given by:

$$\frac{dn}{dt} = \alpha_n(1-n) - \beta_n n \tag{7}$$

where n is the probability of being in a permissive state. This equation can be simplified using the two relations:

$$n_\infty = \frac{\alpha_n}{\alpha_n + \beta_n} \tag{8}$$

$$\tau_n = \frac{1}{\alpha_n + \beta_n} \tag{9}$$

where n_∞ indicates the stationary part of the channel and τ_n is the activation time of the channel. These relations lead to rewrite Eq. (7) in the following manner:

$$\frac{dn}{dt} = \frac{n_\infty - n}{\tau_n} \tag{10}$$

that is solved by:

$$n(t) = n_\infty - (n_\infty - n_0)e^{-\frac{t}{\tau_n}} \tag{11}$$

where n_0 is the initial value of n.

The study reported in [15] has identified the number of gating particles of each channel.

It is possible to include all these relations in equation (5), obtaining the final HH model given by:

$$\begin{aligned} I = C_m\frac{dV_m}{dt} &+ \bar{g}_k s^4(V_m - E_k) \\ &+ \bar{g}_{Na} m^3 h(V_m - E_{Na}) + \bar{g}_L(V_m - E_L) \end{aligned} \tag{12}$$

where s, m and h are state variables; the first one is related to potassium channel while the other two are related to the sodium channel.

The HH model can be extended for achieving even better accuracy and realism. In [16] the HH model has been improved for taking into account some particular mechanisms related to ions. Authors of [16] introduced three currents related to sodium: a fast Na$^+$ current (I_{Na-f}), a persistent Na$^+$ current (I_{Na-p}) and a resurgent Na$^+$ current (I_{Na-r}). For what concerns potassium, five currents have been introduced: a current for rectified delayed channels (I_{K-V}), a current depending on intracellular calcium concentration (I_{K-Ca}), a current for inward rectified channels (I_{K-IR}), a current for type-A channels (I_{K-A}) and a current for slow kinetic channels (I_{K-slow}). Moreover, there is a current for describing calcium kinetic (I_{Ca}).

All these current contributions can be added for obtaining the total current of the soma.

This model can also be enriched by considering synapses, since a granular cell has both excitatory and inhibitory synapses, used for communication with other neurons.

The excitatory synapses are characterized by a neurotransmitter called *glutamate*; moreover there are two different kinds of receptors in the excitatory synapses:

- α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA)
- N-methyl-D-aspartate (NMDA).

These receptors have been modeled in [17] using a Markov chain approach for describing the transition between different states. The same strategy has been followed in [18] for describing the inhibitory synapse, whose receptor is called GABA, from the name of the used neurotransmitter. The three kinetic schemes of the aforementioned receptors are shown in Fig. 4.

For what concerns AMPA receptors, there are three possible channel states: open (O), closed (C) and desensitized (D).

The current contribution of these receptors is computed with the formula:

$$I_{AMPA} = g_{\max,AMPA}(V_m - V_{rev,AMPA})O(T) \tag{13}$$

where $g_{\max,\ AMPA}$ is the maximum conductance of the AMPA receptor (1200 pS), V_m is the membrane potential, $V_{rev,\ AMPA}$ is the ionic reversal potential and $O(T)$ is the probability of being in the open state, which depends from the concentration of neurotransmitter T.

NMDA receptors are more complex, since there are three closed states (C0, C1 and C2), an open state (O) and a desensitized state (D). Moreover this channel is heavily influenced by the Mg^{2+} ion, which is capable of blocking the channel.

In this case the total current is given by:

$$I_{NMDA} = g_{\max,NMDA}(V_m - V_{rev,NMDA})O(T)B \tag{14}$$

where $g_{\max,\ NMDA}$ is the maximum conductance of the AMPA receptor (18,800 pS), V_m is the membrane potential, $V_{rev,\ NMDA}$ is the ionic reversal potential and $O(T)$ is the probability of being in the open state, which depends from the concentration of neurotransmitter T; finally B is a term for taking into account the influence of Mg^{2+} ions.

The GABA (gamma-Aminobutyric acid) inhibitory receptors are made up of two parts, called α1-GABA and α6-GABA. These two parts can be modeled using the same Markov chain, which is made up of two open states (OA1 and OA2), three closed states (C, CA1 and CA2) and three desensitized states (DA1, DA2 and DA2f).

The current of each part of the GABA receptor is given by:

$$I_{GABA} = g_{\max,GABA}(V_m - V_{rev,GABA})(OA1(T) + OA2(T)) \tag{15}$$

where $g_{\max,\ GABA}$ is the maximum conductance (918 pS for α1-GABA and 132 pS for α6-GABA), V_m is the membrane potential, $V_{rev,\ GABA}$ is the ionic reversal potential and the sum $OA1(T) + OA2(T)$ represents the probability of being in an open state.

Now it is possible to improve Eq. (12) taking into account the new channels introduced by [15] and the synapses.

First of all, it is necessary to define the total applied current, which is given by:

$$I_{app} = I_{channels} + I_{synapses} \tag{16}$$

where $I_{channels}$ is the sum of all the currents of ionic channels and $I_{synapses}$ is the sum of all the synaptic currents.

Eq. (12) can be rewritten in the following form:

$$I = C_m\frac{dV_m}{dt} + \sum\nolimits_i g_i(V_m - E_i) + I_{synapses} = C_m\frac{dV_m}{dt} + I_{app} \tag{17}$$

where the index i is used for taking into account all the ionic channels, with the related conductance and the Nernst potentials.

It is important to stress that the obtained model has a grade of biological realism higher than the other models presented before. In fact, it can reproduce a greater number of physiological behaviour than I&F or Izhikevich's models. So it is described by several ODE but it has an enormous computational complexity.

4. Implementation

Research carried out at the University of Pavia involves the purpose of different HBP Platforms. Inside the Brain Simulation Platform we are developing a realistic simulator of the cerebellar cortex through suitable computing solutions.

The models of granule cell, excitatory and inhibitory synapses have been developed by the authors of [19] in Python, which is a suitable language for the NEURON simulator.

In particular, the authors simulated 4393 neurons and more than 40,000 synapses, described by very realistic models. The simulation of 3 s of activity on a Pentium dual core working at 2.6 GHz took about 20 h.

The execution times [19] are quite long and, for this reason, considering and thinking also to bigger simulations, it becomes necessary to exploit high performance computing (HPC) technologies for achieving better performance.

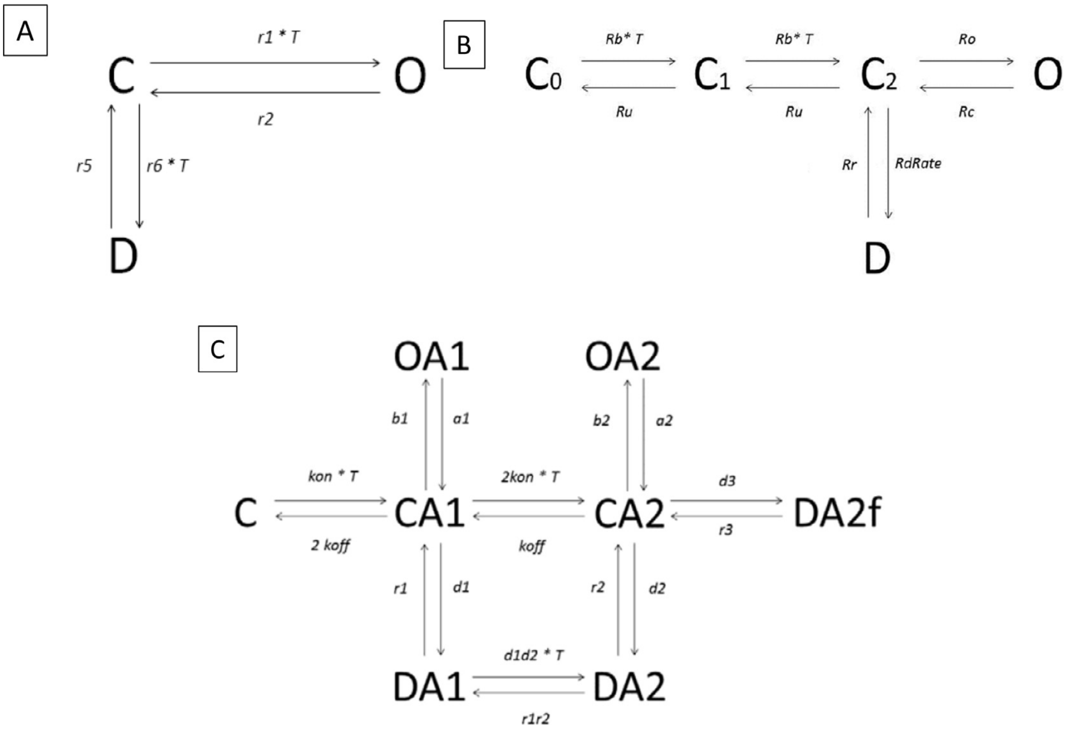

Fig. 4. Kinetic schemes of AMPA [A], NMDA [B] and GABA [C] receptors. All the parameters shown near the arrows represent kinetic constants of the reaction speed, while T is the amount of neurotransmitter.

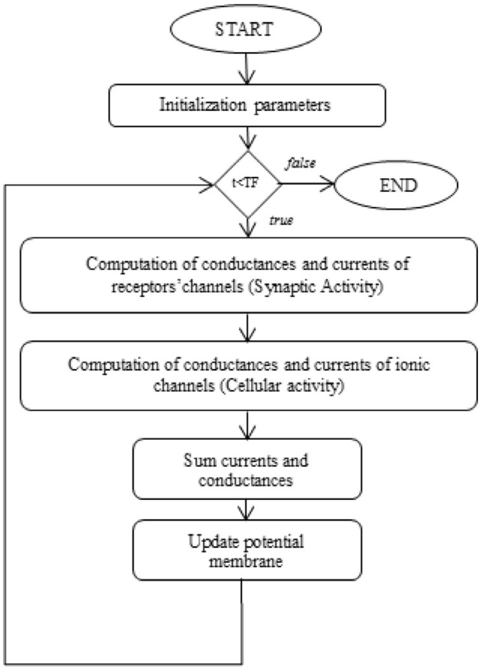
START
Initialization parameters
t<TF
false
END
true
Computation of conductances and currents of receptors'channels (Synaptic Activity)
Computation of conductances and currents of ionic channels (Cellular activity)
Sum currents and conductances
Update potential membrane

Fig. 5. Main flow of the algorithm.

The models have been written in C language starting from the NEURON *mod-files* [20]. This phase does not correspond to a simple algorithm translation from *mod-files* to C language, but requires the development of entire code parts, intrinsic in NEURON, which is a simulator specifically built for neuronal networks. The main flow of the algorithm is shown in Fig. 5.

The *Initialization parameters* phase sets the variables which describe the synapses and the granule's behaviour. Furthermore, in this phase, conductances are multiplied by a corrective factor which represents temperature effects.

The flow proceeds with a cycle which describes each granule's cellular activity. This iteration continues until $t < TF$ where t is simulation time and TF is the last instant of the cellular activity. The first phase of this cycle is made up of the following steps, that compute:

1. The glutamate amount released by the presynaptic terminal;
2. Kinetics states of the AMPA and NMDA channels;
3. Conductances and currents of these receptors;
4. The GABA amount released by the presynaptic terminal;
5. Kinetics states of the GABA channels;
6. Conductances and currents of this receptor.

Finally, all the excitatory and inhibitory synaptic currents are added up and stored.

The successive step is characterized by the computation of the conductances and currents of the nine ionic channels (according to Eq. (12) with modification introduced in [16]) and of the two leakage currents.

All the ionic and leakage currents and conductances are summed and stored in suitable variables.

The *Sum currents and conductances* step consists in the sum of the synaptic and ionic conductances and currents computed at the previous steps.

Finally, it is possible to update the potential membrane value according to the Eq. (17).

This algorithm allows to perform two different simulations: the first one, called in vitro, shows how the neuron reacts to a current injection in the soma, without taking into account synaptic activity. The variable which describes the injected current amplitude is added to the other currents. In this kind of simulation the synaptic current is of course forced to 0 pA (no synaptic activity).

On the other hand, an in vivo simulation should be performed, to understand how the neuron reacts to spikes generated by the presynaptic terminal while there is no current injection.

Please cite this article as: G. Florimbi et al., The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells, Microprocessors and Microsystems (2016), http://dx.doi.org/10.1016/j.micpro.2016.05.015

The cycle iterations describing cellular activity of each granule are independent, since this reflects the physiological cells behaviour. For this reason, it is possible to execute these iterations simultaneously using the multithreading *Application Program Interface* (API) OpenMP 2.0 where each thread calculates the cell activity of a granules group. The algorithm has been modified introducing suitable *#pragma* statements and specifying which variables are private or shared with other threads. This implementation is tested on an Intel i7 processor having four physical cores (eight logical cores), therefore each thread computes the cell activity of N_CEL/8 granules, where N_CEL is the number of simulated granules.

To exploit further acceleration, the algorithm has been developed in CUDA C exploiting GPU technology. In this paradigm each thread calculates the activity of each granule, executing three main phases:

1. Device memory allocation and variable transfer from host to device;
2. Kernel execution;
3. Results transfer from device to host and device memory deallocation.

This algorithm is characterized by a unique kernel which computes the entire cell activity cycle shown in Fig. 5.

The memory transfers from host to device and vice-versa are time consuming processes that have to be properly managed. All the transferred data are stored in 1D array consecutive locations. This strategy allows a considerable gain of time.

After the transfers from host to device, common data are stored in shared memory so threads of the same block can share them; otherwise, they are stored in private memory. Then, after kernel launch, the cell activity cycle is calculated for each granule and each iteration computes an updated value of membrane potential. All the values calculated for each granule are stored in a 1D array which is the only structure sent from device to host.

On the other hand, we are developing digital circuits that emulate cells behaviour. This activity is related to the Neuromorphic Computing Platform. One of the aim of this platform in fact is to explore the feasibility of application specific processors implementing the models conceived, evaluating execution times, real time constrains, satisfaction and power consuming.

For this reason, we have also exploited FPGA technology by designing a circuital model of the soma and a NMDA receptor. We chose this receptor since it is the most frequent one in excitatory synapses and its channel states number is halfway between AMPA's channel states number and GABA's one.

The circuital model of NMDA receptor has been designed using Altera Quartus Prime software. Data are represented with fixed point resolution by using 27 bits 4 of which are used for the integer part while the remaining 23 for the fractional one, since variables assume very small values. Equations in [20] implemented in Quartus Prime schematic entry describe NMDA behaviour reported in [16]. Implementation is made up of the following steps:

1. Computation of temporal variation of four kinetic states;
2. Each variation is added to the previous value of each state;
3. Computation of the fifth kinetic state;
4. Computation of receptor's channel conductance.

We also developed a circuital model of the soma. For a realistic representation, the soma model should be refined, requiring more bits for a suitable precision. In this case we adopted a fixed point representation that takes 64 bits 21 of which used for integer part while the remaining 43 for the fractional part. However, for achieving real-time performance, parallel elaboration of each channel is required. Since FPGA resources are limited, it is also necessary to optimize computations, in order to limit the number of operations performed for solving equations related to channels conductance. In particular, the gating particles kinetic can be described in mathematical terms (equations from (7) to (11)) and, after that, it is possible to simplify some steps. For example, to compute the new value of a gating particle, we need to solve Eq. (10). In the case of potassium channel, this value can be computed using the following equations:

1. $a_n = \frac{n_\infty}{\tau_n} = \frac{\alpha_n}{\alpha_n+\beta_n}(\alpha_n+\beta_n) = \alpha_n$
2. $b_n = \frac{1}{\tau_n} = \alpha_n + \beta_n$
3. $n = ne^{-b_n P} + \frac{a_n}{b_n}(1 - e^{-b_n P})$

where P is the discretization interval.

The new values of the gating particles are multiplied by suitable constants that represent the maximum conductance of each channel (see Eq. (6)).

This approach allows to minimize the number of logic elements needed to update the conductances. Moreover, it must be noticed that reducing the number of logic resources is also important for enabling parallel computation.

Fig. 6 shows the architecture of the circuit that implements the equations described for potassium channel. The computation of the exponential function is heavy in terms of resources, so it is important to find a way to minimize its use. Thus, we chose to employ only one exponential block in each channel. Moreover, we designed this block using Taylor series approximated to the fifth term, since this does not affect the correctness of the results. The blocks have been designed using pipeline philosophy, for increasing maximum working frequency. The latency for performing the computation needs 11 clock cycles.

The use of a single exponential block in each channel requires a suitable data routing strategy; in this work, we used a multiplexer with a custom control logic for selecting the right inputs for the calculation needed at a certain time.

Another required optimization consists in the multiplication by the discretization interval, which is equal to 0.025. The multiplication can be rewritten in the following form:

$$x * 0.25 = \frac{x}{10} * \frac{1}{4} \tag{18}$$

where the division by 10 can be approximated with the following algorithm, that uses only shifts and sums:

1. $Q = ((x \gg 1) + x) \gg 1$
2. $Q = ((x \gg 4) + Q)$
3. $Q = ((x \gg 8) + Q)$
4. $Q = ((x \gg 16) + Q) \gg 3$

the final multiplication for $\frac{1}{4}$ is implemented through a suitable shift operation.

The blocks for channels conductance computation work in parallel, and their output are added for obtaining the total conductance of the soma, as shown in Fig. 7. The blocks indicated as "CHANNEL" contain the solution of equations similar to (7)–(10) specialized for each channel. All the outputs of those blocks are added for obtaining the total conductance through a parallel adder. This value is used for updating the membrane potential. The calcium channel differs from the others, since the Nernst potential of this ion depends on intracellular calcium concentration, which is function of the calcium current.

The membrane potential value calculated at a given step is the input of the next iteration.

5. Results

The first part of this section describes results of the work related to Brain Simulation Platform.

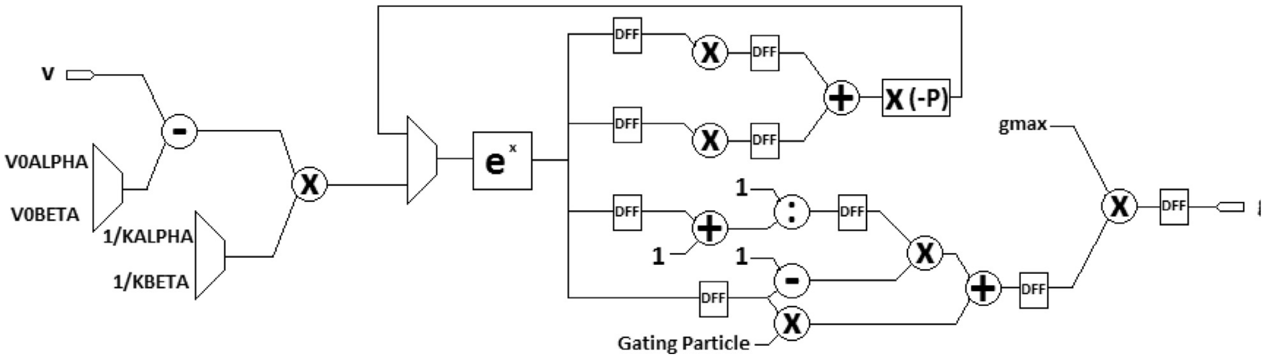

Fig. 6. Architecture of the potassium channel.

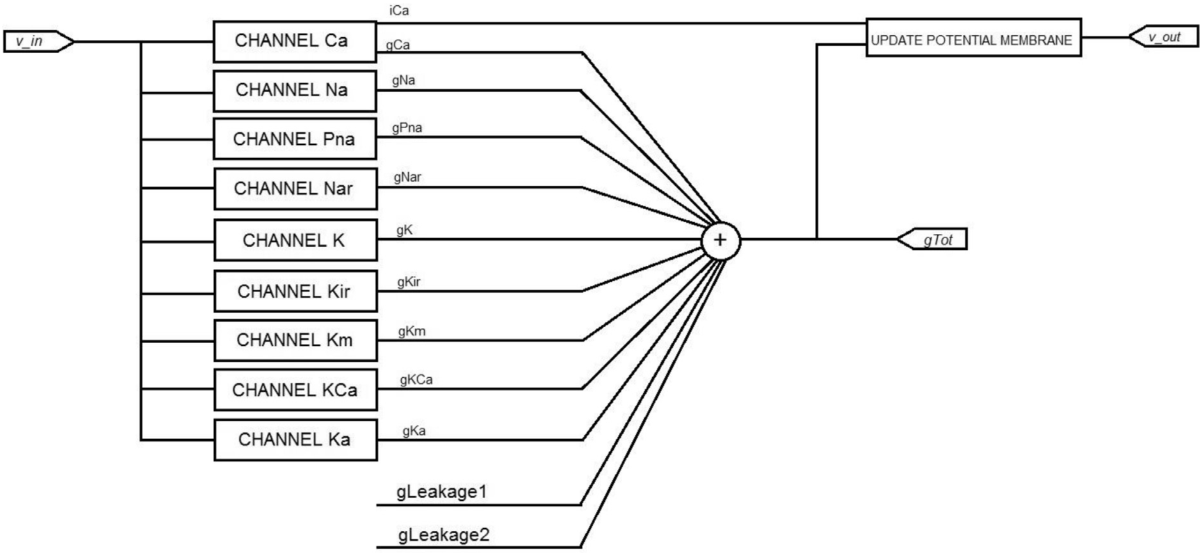

Fig. 7. Architecture for parallel computation of channels conductance.

Serial and OpenMP codes have been executed on an Intel i7 processor, working at 3.40 GHz, equipped with 8 GB RAM. CUDA version has been tested on an NVIDIA Tesla K40 active GPU (GPU 1) and on a NVIDIA GTX Titan Z (GPU 2). GPU 1 is based on Kepler architecture, working at 875 MHz, equipped with 2880 cores and 12 GB GDDR5 memory with a bandwidth of 288 GB/s. GPU 2 is a double GPU, each one is equipped with 2880 cores, working at 876 MHz with 6 GB GDDR5, with a bandwidth of 672 GB/s. GPU 2 is also based on Kepler architecture, but it is a more recent one so its performance is better than GPU 1.

The simulated cell activity lasts 3 s: in the first half there is a current injection (without synaptic activity) while in the remaining one there is only synaptic activity.

Simulations have been conducted with an increasing granule number, from 1 to 400,000, and an increasing synapse number, from 8 to 3200,000 (8 synapses for each granule).

Serial and OpenMP execution times are lower than CUDA ones for simulations of a few granules. If the simulated cells number grows up, GPU is the best solution as it can be seen in Table 1. It must be noted that GPU 2 has enough memory only for executing up to 100,000 granules and 800,000 synapses. As the number of simulated cells increases the speedup calculated between the serial code and the GPU version grows up until it becomes constant because sequential code parts prevail on the parallel ones. Higher GPU performances are due to massive parallelism of the algorithm and also to the optimization of variables transfer described before. The adopted strategy for data transfer is efficient as highlighted by the analysis conducted with NVIDIA Visual Profiler (see Fig. 8).

The graph in Fig. 8 is obtained as the mean of 18 executions. It shows that data transfer duration is negligible with respect to kernel one.

This concludes the description of the activity inherent to brain simulation. In the following we will discuss results obtained by the evaluation of FPGA technology for neuromorphic computing.

The circuital models developed with Quartus Prime has been deployed on an Altera Stratix V 5SGXEA7N2F45C2 FPGA. The architecture of NMDA receptor achieved a working frequency of about 140 MHz. Resource usage is less than 1% and total block memory used is less than 1%. Concerning the synaptic activity, the simulation of 3 s related to four independent NMDA receptors took 98.6 ms on an Intel i7 processor, while it took 12 ms on FPGA device.

For what concerns the architecture of the soma, we performed a simulation of 3 s with three different current injections: the first one is injected from 100 ms to 250 ms, the second from 250 ms to 1500 ms and the latter from 1500 ms to 3000 ms.

The first current is 10 pA and it is under threshold so it does not generate spikes. The second current is 16 pA, while the third one is 22 pA. These two currents generate spikes of different frequencies directly proportional to the amplitude of the injected current.

Results are shown in Fig. 9.

The design uses up all the logic elements available on the device. It also uses 78% of the DSP blocks and 130 pins. For comput-

Table 1
Execution times.

N° cells	Serial [s]	OpenMP [s]	Speedup serial – OpenMP	GPU 1[s]	GPU 2 [s]	Speedup serial – GPU 1	Speedup serial – GPU 2
1	0.34	0.36	0.93	10.11	9.44	0.03	0.04
5	1.71	1.00	1.71	10.15	10.19	0.17	0.17
10	3.34	2.06	1.62	10.17	11.37	0.33	0.29
50	16.70	8.97	1.86	18.75	15.97	0.89	1.05
100	33.54	16.54	2.03	18.77	16.08	1.79	2.08
500	167.11	83.95	1.99	36.85	16.21	4.53	10.31
1000	334.15	175.99	1.89	55.55	47.68	6.02	7.01
5000	1646.17	837.12	1.97	205.04	175.70	8.03	9.37
10,000	3243.43	1684.89	1.92	393.76	336.85	8.24	9.63
25,000	8183.94	4266.81	1.91	975.40	847.41	8.39	9.66
50,000	16225.42	7902.71	2.05	1949.47	1695.04	8.32	9.57
100,000	32620.71	16505.73	1.98	3892.24	3387.26	8.38	9.63
200,000	65282.04	30524.79	2.14	7778.74	–	8.39	–
400,000	129894.74	67268.86	1.93	15580.25	–	8.33	–

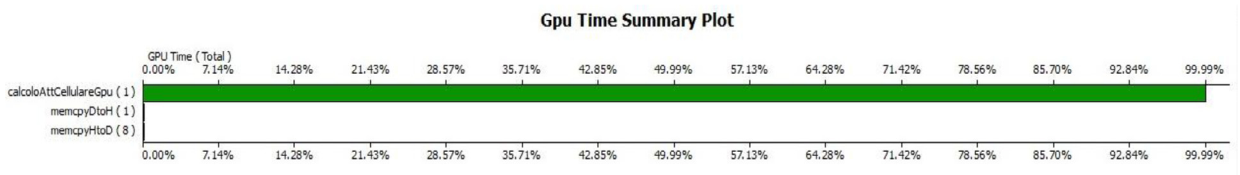

Fig. 8. Graph obtained with NVIDIA Visual Profiler. The first bar indicates the ratio of time taken by GPU for the computation, while the second and the third bars are not visible since they represent the ratio of time taken by memory transfers.

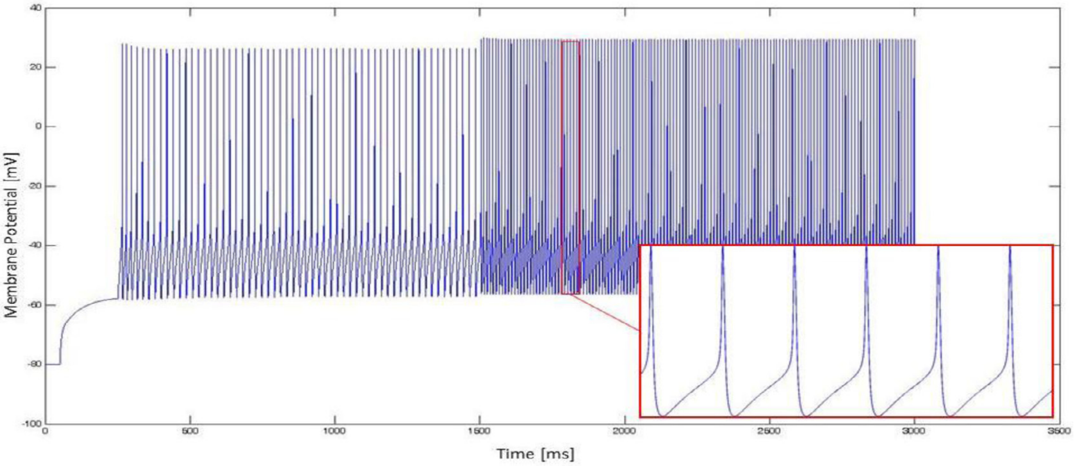

Fig. 9. Simulation results.

ing the membrane potential update for a single discretization interval, the proposed architecture takes 172 clock cycles. Assuming a working frequency of 50 MHz, the time needed for this computation is about 2.54 µs. So the architecture is real time compliant since one discretization interval is 25 µs. For the simulation described in Fig. 9 (3 s of cellular activity) the elaboration lasts for about 305 ms on the FPGA device, while it took 1.19 s on Intel i7 processor.

These results clearly indicate that, although we are at the beginning of the work, for algorithm tuning and set up GPU is one of the possible suitable technologies since it combines fast execution times with enough significant simulations, due to the relative high number of granular cells that can be included. When of course bigger network of cells will be simulated, supercomputers with cluster of computing units (maybe GPU) will be absolutely necessary. On the other hand, one of the target of the project is to design and carry out neurochips based on the accurate developed models to be locally implanted, if suitable biomaterials will be available. The work performed on FPGA, shows that this technology is not the solution but indicates the solution: due to their real time processing capabilities, custom ASIC processors will be the “bricks” by which to build these neurochips.

6. Conclusions

In this paper, we presented the current activity undergoing in Pavia’s University unit under the Brain Simulation, the Neuromorphic Computing and High Performance Computing platforms of the Human Brain Project.

The aim of the work, still in progress, is to provide scientists with a sufficient set of tools to identify the complex series of physiological events bringing from genomic analysis to cognition, perception, emotions, and decision-making; and to identify fundamentals related to learning and memory. This will permit to diagnose brain diseases at a preliminary step and to propose ad hoc treatments.

From the ICT point of view there is the potential to renew current computing theory and technologies. New devices will be able

Please cite this article as: G. Florimbi et al., The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells, Microprocessors and Microsystems (2016), http://dx.doi.org/10.1016/j.micpro.2016.05.015

to reproduce the low-cost, energy-efficient brain computing capability, and if possible to emulate “brain-like intelligence”.

Systems of this type could cooperate with traditional technologies, in a complementary role enabling new applications.

We also feel that the necessary research for HBP will be the catalyst of new approaches for high-performance computing in science and industry, eventually making elaboration performance available at the consumer level.

Concerning the research in progress (Brain Simulation) as described before, in literature there are neuronal network simulations with short execution time, even real time compliant in some cases [21–26]. Indeed, in those works, models are not biologically realistic but they feature a very low computational weight since they are typically based on Neural Networks or Leaky Integrate and Fire approaches. It must be noticed that the models used in those works are not suitable for the Human Brain Project which is focused on *realistic computational modelling*.

A human brain simulator requires neurons models as accurate as possible. Real time execution is certainly an important goal to achieve but it is not possible to exclude realistic models, if the goal is an accurate description of brain and its mechanisms. Nowadays, the most realistic goal is to develop a physiological accurate simulator which is able to reduce execution times. The work performed until now demonstrates how it is possible to simulate 3 s of cell activity of 400,000 granule cells with 3200,000 synapses in only four hours using GPU technology.

The next step of our work is the development of a simulator of cerebellar cortex including bigger number of different cells. In this case we will evaluate the use of multiple GPUs system.

Moreover, in this work we introduced and implemented two circuital models of NMDA receptor and soma based on FPGA technology. This technology shows execution times that are real time compliant. We performed simulations of in vitro activities that fully agree with the biological behavior of granule cells.

The activity in progress is a contribution connected to the activity of the Neuromorphic Computing and High Performance Computing platforms where two approaches are under analysis concerning the development of a highly efficient computing system. The first approach takes legacy from the European FACETS project that aimed at the development of a massively parallel neuronal system based on large scale integration of custom designed analog circuits [27]. The second one, instead, is based on a coarse grain parallel system featuring ARM processors, each of which hosts thousands of neurons and follows the idea introduced by the SpiNNaker group [28]. Our work could provide useful indications in both the cases.

The development of a cells network on FPGA is not possible on a single chip. So we will evaluate the connection of multiple boards. It must be noticed that the aim of this part of the work is to develop prototypes for future neurochips. In this case, it is not important to put a large number of cells on a single chip since the final device will be an ASIC. This will improve the area needed by the circuit together with elaboration times and power consumption.

The future work will be related to the continuous refinement of the implemented models together with the simulation of other types of neuronal cells (Purkinje, stellate, basket, Golgi).

Acknowledgment

The authors would like to thank NVIDIA Corporation for the donation of the NVIDIA Tesla K40 GPU used for this research.

References

[1] A. Calimera, B. Macii, M. Poncino, The Human Brain Project and neuromorphic computing, Funct. Neurol. 28 (3) (2013) 191–196.
[2] B.V. Benjamin, P. Gao, E. McQuinn, S. Choudhary, A.R. Chandrasekaran, J.M. Bussat, R. Alvarez-Icaza, J.V. Arthur, P.A. Merolla, K. Boahen, Neurogrid: a mixed-analog-digital multichip system for large-scale neural simulations, Proc. IEEE 102 (5) (2014) 699–716.
[3] A. Upegui, C.A. Peña-Reyes, E. Sanchez, An FPGA platform for on-line topology exploration of spiking neural networks, Microproc. Microsyst. 29 (5) (2005) 211–223.
[4] M.R. Azghadi, N. Iannella, S.F. Al-Sarawi, G. Indiveri, D. Abbott, Spike-based synaptic plasticity in silicon: design, implementation, application and challenges, Proc. IEEE 102 (5) (2014) 717–737.
[5] T.C. Stewart, C. Eliasmith, Large-scale synthesis of functional spiking neural circuits, Proc. IEEE 102 (5) (2014) 881–898.
[6] J. Tani, Self-organization and compositionality in cognitive brains: a neurorobotics study, Proc. IEEE 102 (4) (2014) 586–605.
[7] H. Markram, A countdown to a digital simulation of every last neuron in the Human Brain, Sci. Am. 306 (6) (2012).
[8] AP. Alivisatos, et al., *Neuroscience: the brain activity map*, Science 339 (2013) 1284–1285.
[9] JA. McNab, et al., The Human Connectome Project and beyond: initial applications of 300 mT/m gradients, Neuroimage 80 (2013) 234–245.
[10] H. Markram, Seven challenges for neuroscience, Func. Neurol. 28 (3) (2013) 145–151.
[11] www.humanproject.eu.
[12] www.humanbrainproject.eu/documents/10180/17648/TheHBPReport_LR.pdf.
[13] A.N. Burkitt, *A review of the integrate-and-fire neuron model: I. homogeneous synaptic input*, Biol. Cybern. 95 (1) (July 2006) 1–19.
[14] E.M. Izhikevich, Simple model of spiking neurons, IEEE Trans. Neural Netw. 14 (6) (2003) 1569–1572.
[15] A.L. Hodgkin, A.F. Huxley, A quantitative description of membrane current and its application to conduction and excitation in nerve, J. Physiol. 117 (4) (1952) 500–544.
[16] E. D'Angelo, T. Nieus, A. Maffei, S. Armano, P. Rossi, V. Taglietti, A. Fontana, G. Naldi, Theta-frequency bursting and resonance in cerebellar granule cells: experimental evidence and modeling of a slow k^+-dependent mechanism, J. Neurosci. 3 (21) (2001) 759–770.
[17] T.R. Nieus, E. Sola, J. Mapelli, E. Saftenku, P. Rossi, E. D'angelo, LTP regulates burst initiation and frequency at mossy fiber-granule cell synapses of rat cerebellum: experimental observations and theoretical predictions, J. Neurosci. 95 (2) (2006) 686–699.
[18] T.R. Nieus, L. Mapelli, E. D'Angelo, Regulation of output ppike patterns by phasic inhibition in cerebellar Granule Cells, Front. Cellular Neurosci. 12 (4) (2014) 1–18.
[19] S. Solinas, T. Nieus, E. D'Angelo, A realistic large-scale model of the cerebellum granular layer predicts circuit spatio-temporal filtering properties, Front. Cellular Neurosci. 12 (4) (2010) 1–17.
[20] http://senselab.med.yale.edu/ModelDB/default.cshtml.
[21] J. Luo, G. Coapes, P. Degenaar, T. Yamazaky, T. Mak, C. Thin, A real time Silicon Cerebellum Spiking Neural Model based on FPGA, in: International Symposium on Integrated Circuits (ISIC), Singapore, 2014, pp. 276–279.
[22] T. Yamazaki, J. Igarashi, Realtime cerebellum: a large scale network model of the cerebellum that runs in realtime using a graphics processing unit, J. Neural Netw. 47 (2013) 103–111.
[23] W.K. Li, M.J. Hausknecht, P. Stone, M.D. Mauk, Using a million cell simulation of the cerebellum: Network scaling and task generality, J. Neural Netw. 47 (2013) 95–102.
[24] F. Naveros, N.R. Luque, J.A. Garrido, R.R. Carrillo, M. Anguita, E. Ros, *A spiking neuronal simulator integrating event driven and time driven computation schemes using parallel cpu gpu", co-processing: a case study*, IEEE Trans. Neural Netw. Learn. Syst. (2014).
[25] A.A. Jalife, R.A. Vazquez, Implementation of configurable and multipurpose spiking neural networks on GPUs, in: IEEE World Congress on Computational Intelligence, WCCI, Brisbane, Australia, 2012, pp. 1–8.
[26] P. Pourhaj, D.H.Y. Teng, FPGA based pipelined architecture for action potential simulation in biological neural systems, in: 23rd Canadian Conference on Electrical and Computer Engineering, Calgary, Canada, 2010, pp. 1–4.
[27] www.facets-project.org.
[28] S. Furber, F. Galluppi, S. Temple, L. Plana, "The SpiNNaker Project, Proc. IEEE 102 (5) (2014) 652–665.

Giordana Florimbi was born in Teramo, Italy, in 1989. She received the Bachelor's degree in biomedical engineering from Università Politecnica delle Marche, Ancona, Italy, in 2012, and the Master's degree in bioengineering from the University of Pavia, Pavia, Italy in 2015. She is a Ph.D. student in bioengineering and bioinformatics at the University of Pavia. Her research interests include realistic simulations of neuronal activity on high performance technologies.

Emanuele Torti was born in Voghera, Italy, in 1987. He received the Bachelor's degree in electronic engineering and Master's degree in computer science engineering (cum laude) from the University of Pavia, Pavia, Italy, the Ph.D. degree in electronics and computer science engineering from the University of Pavia, in 2009, 2011, and 2014, respectively. He is a Postdoc Researcher with the Engineering Faculty, University of Pavia. His research interests include high performance architectures for real-time image processing and signal elaboration.

Stefano Masoli has a degree in cellular biology, a degree in neurobiology and a Ph.D. in biomedical sciences. He is currently a postdoctoral researcher at the University of Pavia. The scientific interest is the development of realistic computational models of cerebellar neurons, with the main focus on the reconstruction of the Purkinje cell, one of the most complex neuron of the Central Neurvous System. The techniques used in the creation of these models are the programming language Python and NEURON, which is the environment used to run the simulations. He is now working to translate his models in the format required by the Human Brain Project.

Egidio Ugo D'Angelo has a degree in medicine (cum laude) and a degree in neurology (cum laude). He is a full professor of physiology at the Department of Brain and Behavioral Sciences of the University of Pavia. He is supervisor of the unit of Neurophysiology at this department. He is the director of the Brain Connectivity Center (BCC) of the IRCCS C. Mondino of Pavia. His main scientific interests include the function of neurons, synapse and networks of the brain, with a special interest for cellular and synaptic mechanisms of synaptic plasticity.

Giovanni Danese received the Ph.D. degree in electronics and computer engineering from the University of Pavia, Pavia, Italy, in 1987. He is a Full Professor with the Computer Programming and Computer Architecture, Engineering Faculty, University of Pavia. His research interests include parallel computing, computerized instrumentation special-purpose computers, and signal and image processing.

Francesco Leporati received the Ph.D. degree in electronics and computer engineering from the University of Pavia, Pavia, Italy, in 1993. He is an Associate Professor with the Industrial Informatics and Embedded Systems and Digital Systems Design, Engineering Faculty, University of Pavia. His research interests include automotive applications, FPGA and applicationspecific-processors, embedded real-time systems, and computational physics. Dr. Leporati is a member of the Euromicro Society and Associate Editor of Microprocessors and Microsystems.

Please cite this article as: G. Florimbi et al., The Human Brain Project: Parallel technologies for biologically accurate simulation of Granule cells, Microprocessors and Microsystems (2016), http://dx.doi.org/10.1016/j.micpro.2016.05.015