Design of a modular and mixed neuromimetic ASIC

J. Tomas, Y. Bornat, S. Saighi, Timothée Levi, S. Renaud

Design of a modular and mixed neuromimetic ASIC

J; Tomas, Y. Bornat, S. Saïghi, T. Lévi, S. Renaud
IXL Laboratory, CNRS UMR 5818,
ENSEIRB – Université Bordeaux 1
Talence, France
tomas@ixl.fr

***Abstract*—This paper presents a new specific integrated circuit (ASIC) that emulates neurons electrical activity using a biophysical model (neuromimetic ASIC). Such ASICs form the computation core of a complete simulation system dedicated to the investigation of the dynamics of biomimetic neural networks. The circuits were designed using a modular approach. Simulations were realized by mixing the behavioral and the transistor-level description of modules. We present in this paper the ASICs specifications and architecture, together with simulation results.**

I. INTRODUCTION

Neuromorphic engineering is a field of engineering based on the design and fabrication of artificial neural systems, which architecture and design principles are based on those of biological nervous systems. Recent research in that field addresses the investigation of spiking neural networks, in a tentative understanding of the temporal coding of information by such networks [1], [2]. We engineered in our research group different systems to process real-time simulations of neural networks, using detailed and biologically-realistic models of neurons [3]. One innovation in those systems is the fact that artificial neurons are integrated on custom analog VLSI circuits (ASICs). These neuromimetic ASICs compute in real-time the electrical activity of neurons, represented by the voltage difference across their membrane (V_{mem}) [4].

After briefly presenting the simulation system architecture, we will describe in this paper a new generation of ASICs we designed using a modular approach. We will also display simulation results obtained using simulation tools merging behavioral and transistor-level descriptions.

II. CONTEXT AND SPECIFICATIONS

Neuron models exist at different precision levels. The computation mode of artificial neural networks can be classical software, but also digital and/or analog hardware. The choice results from a necessary compromise between precision and technical performance, such as computational speed or power consumption. We chose to design systems based on hardware artificial neurons and synapses, which improve the computation speed and allow the building of hybrid neural networks, where living cells communicate in real time with artificial neurons [5].

An active neuron presents action potential - "spikes" - at low frequencies (from 0.1 Hz to 300 Hz) on V_{mem}, that can be spontaneous or induced by synaptic connections from other neurons. The exact timing of spikes is then a crucial information on the neural network connectivity and activity. The synaptic interactions can be subject to short-term and long-term variations, and process cognitive mechanisms that rely on plasticity and learning. In our simulation systems neuronal activity is computed in real-time and in analog hardware. Synaptic interactions, that control the information transmission between the neurons, are digitally processed using hardware or software programmable medium. The systems are used by neuroscientists to investigate learning or plasticity algorithms that are presumably present in biological neural networks [6].

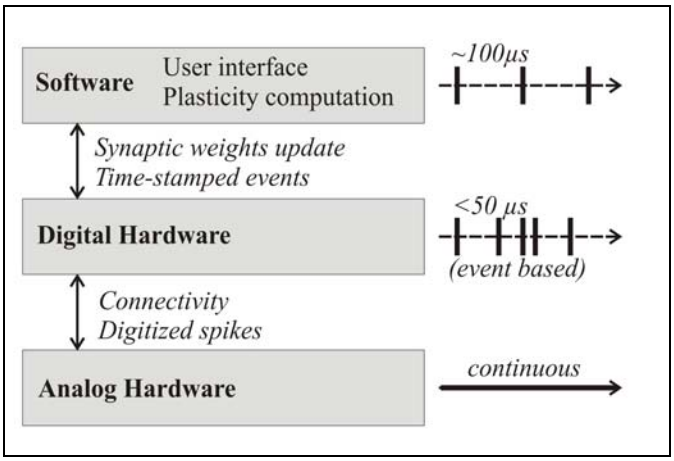

Fig. 1. : Architecture and data flow in the neural networks simulation system.

The whole simulation system is organized in 3 layers (figure 1). The analog hardware layer runs the continuous and real-time computation of the neurons activity. The analog integrated circuits (ASICs) are controlled by the digital hardware layer. This hardware is in charge of computing spike events information from the analog neurons, and of controlling the synaptic connectivity back to

the analog hardware. To optimize computational speed, the processing mode is event-based, with a specification for the maximum period of 100µs. Predefined stimulation patterns can also be applied to individual neurons. The next layer includes the software driver and interface, in charge of controlling the data bi-directional transfer to the software via a PCI bus. A computer running a real-time operating system hosts software functions to compute the connectivity dynamic functions in the neural network. The software also includes user interface functions to control off-line and on-line the simulation configuration.

A. The neurons model

As mentioned in the introduction, we chose to implement a biologically-realistic model of neurons, able to capture the main intrinsic and response properties of cortical neurons, and compatible with the required level of precision in the timing of spikes. The corresponding category of models is the conductance-based models (Hodgkin-Huxley type, [7]) In such a model, we compute the electrical activity of a single neuron by summing ionic and synaptic currents on a capacitance that represents the neuron membrane. The ionic currents are described by a set of non-linear and time and voltage-dependent equations, which parameters depend on the biophysical properties of the considered ionic specie. A set of parameters corresponds to a "model card", and describes a specific type of neurons. We chose to design configurable artificial neurons: the model parameters are not fixed, and can vary in a pre-defined range. This range is fixed in a way that the models can represent "prototypical" neurons and synaptic interactions that co-exist in the cerebral cortex.

For each neuron, pre-synaptic events are collected on synaptic inputs. The events are integrated with their respective synaptic weight to generate the multi-synapse digital control signal. Each of those signals generates a synaptic current added to the neuron ionic current. A synaptic conductance follows the same type of formalism than the ionic currents (voltage and time dependent, [8]), with parameters in a range that corresponds to the most classical types of synapses. The digital hardware layer is in charge of generating the synapses control signals. Using this multi-synapses scheme, the neural network can handle all-to-all connections, whatever the number of neurons in the network.

B. The architecture

The circuit is organized to provide to user a large variety of configurations for the simulated neural network. As the neural activity is generated by a sum of ionic and synaptic currents on a membrane capacitance, we decided to integrate a set of generic blocks, each able to compute a conductance-based model of ionic or synaptic current. Parameters of the model are stored on analog memory cells, which values are programmed during the configuration phase of the simulation. During that same phase, the user will also set the topology of the network, i.e. define the blocks connectivity. A set of connected blocks will form an artificial neuron, with their respective currents summed on an external capacitance.

The *Galway* chip we present here comprises (see figures 2 and 3): - a set of conductance modules, each able to generate an ionic or synaptic current following the conductance-based model; - spike-detection modules, to code on 1-bit the neuron membrane voltage; - a set of synaptic input modules, that activate synaptic conductance modules with a digitally-controlled weight; - an analog memory cells array, to store the conductances parameters; - a matrix of switches, to control the neurons topology (i.e. the arrangement of the conductance and synaptic modules that form the artificial neuron); - digital functions to control data transfer from and to external devices.

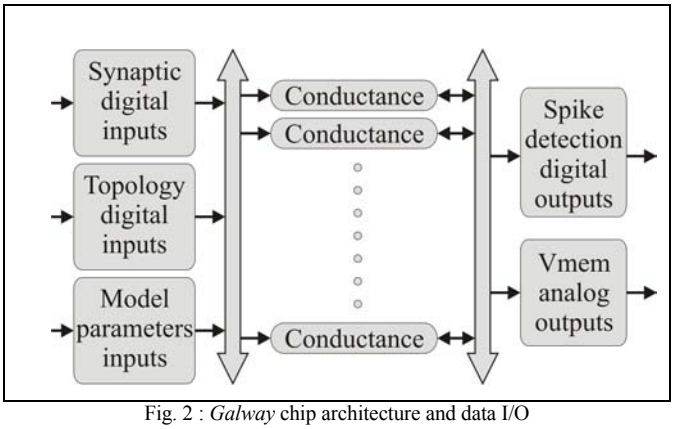

Fig. 2 : *Galway* chip architecture and data I/O

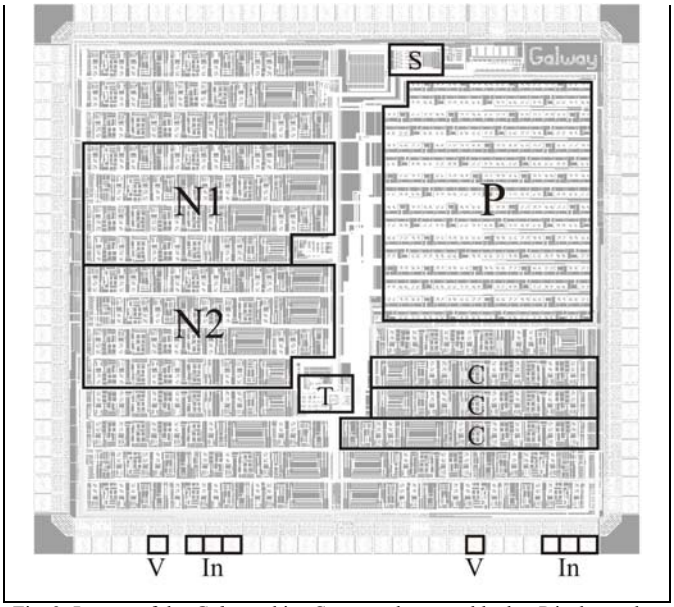

Fig. 3: Layout of the *Galway* chip. *C* are conductance blocks. *P* is the analog memory cells array, *T* the topology control block, *S* a spike detection block, *N1* and *N2* represent 2 "typical" artificial neurons, each formed by 4 ionic conductances and 3 synaptic conductances; the *V* pads convey *N1* and *N2* respective analog membrane potential, *In* receive the *N1* and *N2* respective 3 digital synaptic control inputs. The chip area is 10.5 mm^2, the number of transistors of the core is 47000 (93% for the analog part of the circuit).

The I/O pins are used to: - output the neuron analog membrane voltage (V_{mem}); - output the spike information (1-bit coding of V_{mem}); - input the synaptic weight digital control; - connect the external passive elements necessary for computing the kinetics in the model equations; - convey the control data; - input the analog values for the memory cells; - convey the digital and analog power supplies.

Successive generations of ASICs were designed, that integrate the conductance-based neuron models [9], [10]. These circuits were exploited to build a library of electronics function. Each mathematical function appearing in the neuron model corresponds to a generic analog module in the library. Conductance block that compute the ionic currents models are built using these generic modules. The model parameters are stored in the analog memory cells, and routed in the chip to the analog modules. The ionic and synaptic conductance blocks are interconnected to form a neuron following the topology control.

Each chip includes 21 ionic conductance blocks and 15 synaptic conductance blocks. 205 analog parameters are stored in the memory cells. A typical arrangement of the modules is the building of 5 artificial neurons comprising from 3 to 5 ionic conductances each receiving 3 synaptic inputs. Such a structure allows us to address standard configuration of neural networks. The analog signals (membrane voltage, ionic current) are exploited in specific experimental configurations, such as dynamic clamp experiments; in that case, dedicated on-chip conductances are used as artificial synapses, that communicate in real-time with in vitro biological neurons through an *Axoclamp* amplifier (*Axon Instruments*®) [5].

III. SIMULATIONS

As described in section II, the ASIC is a mixed signal circuit. Nevertheless, the digital part is limited to control blocks and to the state machine that manages the refreshing of the memory cells array. The analog part is the most important one and includes the ionic and synaptic conductances, the memory cells array and the spike detection blocks.

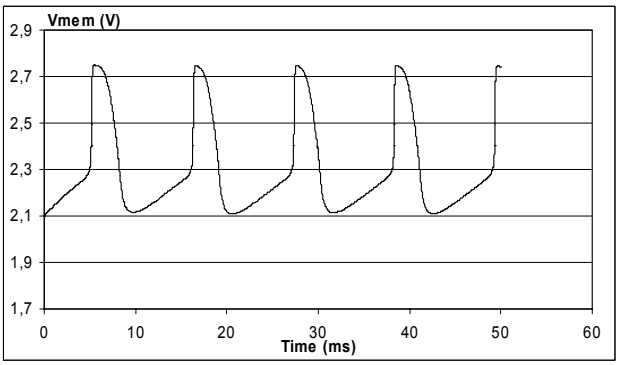

Fig. 4: electrical activity of a standard spiking neuron (VerilogA description)

Except for the state machine synthesized from a VHDL description, all the blocks have been validated using the analog simulator *Spectre* under *Cadence* environment. Due to the huge number of components that compose the chip, it is not possible to perform analog simulation of the final chip. Furthermore, the neural activity is a low frequency activity, and the neural activity has to be simulated for at least tens of ms as shown in figure 4. Let us precise that electrical potential are 5 times greater than biological potential and the biological reference potential (0 V) corresponds to 2.5 V in our simulation. The power supply of the analog part of *Galway* is 5 V.

We decided as a consequence to mix transistor level and behavioral description of the different blocks to validate their connectivity and ensure the functionality of the whole design.

The design of the chip - especially for the conductance blocks - respects a strict hierarchy that can be described as follows:

- a first (and highest) level includes the memory cells array, the neurons elements as a whole, and the digital block. Each block is described by its schematic view, layout view and behavioral view written in *VerilogA* language.
- a second level where, for instance, neurons are described using multi-synapses and different ionic current generators (ionic and synaptic conductances). Each sub-block also has a schematic, layout and behavioral description in *VerilogA* language.
- a final level where appear all the elementary analog functions, with no behavioral description. All these functions are gathered into a common resources library.

Each *VerilogA* description has been defined at the building of the library to validate the design of conductance blocks and, at a higher level, the standard types of neurons we intend to emulate. A neuron *VerilogA* view contains the conductance-based model equations adapted to the electrical representation. The *VerilogA* view of the memory cells array includes the power supplies connectivity and provides typical values for all the parameters needed by the ionic current generator.

As an example, a simulation for one neuron of type FS (Fast Spiking [11]) with null synapses inputs lasts 2'17" using *VerilogA* description and 14'56" using transistor level description. Both simulations give the same neural activity as depicted in figure 5.

To perform the validation of the whole circuit, we used the *Hierarchical Editor* of *Cadence* environment. For each simulation, only one block was described at the transistor level while the others were described using behavioral description.

We ran as many simulations as high-level blocks. Figure 6 illustrates one of these simulations, where only one

neuron of FS type is not simulated using *VerilogA* view. This simulation lasted 3h13' while an identical simulation using behavioral description for all blocks lasted 44'37".

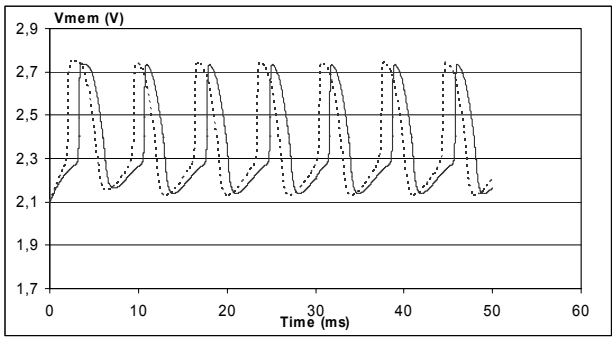

Fig. 5: electrical activity of a FS spiking neuron (dashed line corresponds to VerilogA description and continuous line to transistor level description)

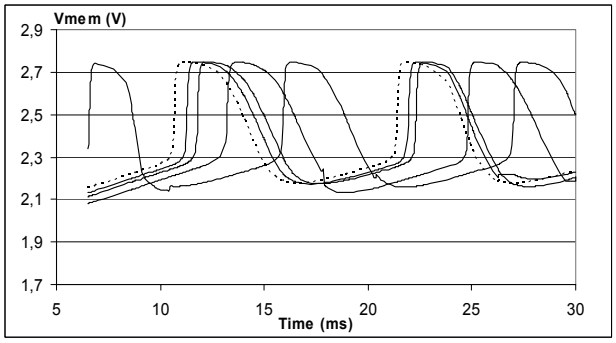

Fig. 6 : electrical activity of 5 neurons configured in the *Galway* chip. The dashed line corresponds to one of the neurons (FS type) simulated using its transistor-level description

IV. CONCLUSION

After describing its environment and functional specifications, we have detailed the different blocks of the Galway chip and their functionality. The circuit contains 36 ionic and synaptic conductance blocks. They can be organized in up to 5 neural elements that can represent the main types of neuron encountered in cortex. The neural network connectivity is performed by conductance-based multi-synapses, which weight is digitally-controlled. These ASICs constitute the analog computational core of a neural network simulator we developed for a pluri-disciplinary european project (*FACETS*, IST-15879). They are associated in the system with digital computation facilities (FPGA or software). It is then possible for example to emulate a neural network activity when its synaptic connections follow plasticity rules. The whole set-up has been dimensioned to simulate up to 512 neurons with all-to-all synaptic connections, dispatched on the ASICs. Such a network will be simulated in real-time.

When considering the whole simulation system, the largest design effort (in terms of time) is the design of the full-custom ASICs. To minimize this effort and improve the re-use, we have developed a library of analog elementary functions and a specific database of standard blocks (neurons). This database allows us to be able to build more rapidly the next generations of ASICs, with varying numbers of ionic and synaptic conductances. The library associates behavioral and schematic views of all blocks at all hierarchical level to be able to perform *Spectre* simulations of the whole chip in a reasonable CPU time. Simulation time is a key issue in our case, considering the low frequency activity of biological neurons.

REFERENCES

[1] R. Hahnloser, R. Sarpeshkar,. M. Mahowald, R. Douglas, S. Seung, "Digital selection and analog amplification co-exist in an electronic circuit inspired by neocortex" , *Nature*, vol.405, pp. 947-951, 2000.

[2] J. Schemmel, S. Hoffman, K. Meier, F. Schurman, "A mixed-mode analog neural network using current-steering synapse", *Analog Integrated Circuits and Signal Processing*, vol.38, pp.233-244, 2004.

[3] S. Renaud-Le Masson, A. Laflaquière, D. Dupeyron, T. Bal, G. Le Masson, "Analog circuits for modeling biological neural networks: design and applications", IEEE Transactions on Biomedical Engineering, vol. 46-6, pp. 638-645, 1999.

[4] V. Douence, A. Laflaquiere, S. Le Masson, T. Bal, G. Le Masson, "Analog electronic system for simulating biological neurons", *IWANN'99*, Alicante, Spain, June 1999.

[5] G. Le Masson, S. Renaud, D. Debay and T. Bal, "Feedback inhibition controls spike transfer in hybrid thalamic circuits", *Nature*, vol. 417, pp. 854-858, 2002.

[6] A. Destexhe, E. Marder, "Plasticity in single neuron and circuit computations", Nature Review Neuroscience, vol. 431, pp.789-795, 2004.

[7] C. Koch, I. Segev, Methods in neuronal modeling: from synapses to networks, MIT Press, Cambridge, 1989.

[8] A. Destexhe, Z.F. Mainen, T.J. Sejnowski, "An efficient method for computing synaptic conductances based on a kinetic model of receptors binding", *Neural Computation*, vol. 6, pp. 10-14, 1994.

[9] L. Alvado, J. Tomas, S. Saighi, S. Renaud, T. Bal, A. Destexhe, G. Le Masson, "Hardware computation of conductance-based neuron models", Neurocomputing, Vol. 58-60, pp. 109-115, 2004

[10] Y. Bornat, J. Tomas, S. Saighi, S. Renaud, "BiCMOS Analog Integrated Circuits for Embedded Spiking Neural Networks", XX Conference on Design of Circuits and Integrated Systems (DCIS), , Lisbon, Portugal, 2005.

[11] B. Connors, M. Gutnick, " Intrinsic firing patterns of diverse neocortical neurons", *Trends Neuroscience,* vol.13, pp. 99-104, 1990.