Tutorial

Toward Silicon-Based Cognitive Neuromorphic ICs—A Survey

Georgios Volanis, Angelos Antonopoulos, and Yiorgos Makris
The University of Texas at Dallas

Alkis A. Hatzopoulos
Aristotle University of Thessaloniki

Editor's notes:
This Tutorial describes the building blocks of neuromorphic VLSI systems and the way they are engineered. From learning mechanisms through the gap between reactive and cognitive systems, all major aspects are covered.
—Jörg Henkel, Karlsruhe, Institute of Technology

The principles of cognition still remain to be unraveled. Nevertheless, neuroscience has made great strides toward understanding the complex operations of the brain and the characteristics of neurobiological processing systems. In general, instead of Boolean logic, synchronous operation, and precise digital computations, neurobiological systems are hybrid analog/digital structures, event driven, distributed, fault tolerant, and massively parallel. They make extensive use of adaptation, self-organization, and learning, and outperform existing most powerful computers in everyday tasks, such as vision and motor control, yet remain energy efficient [1].

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.
Digital Object Identifier 10.1109/MDAT.2016.2545159
Date of publication: 22 March 2016; date of current version: 28 April 2016.

To mimic the operation of neural computing systems, neuromorphic engineering (NE) was introduced by Mead in the late 1980s [2] as the research field that uses electronic neural networks whose architectures and operations are based on those of biological nervous systems [3]. Mead's initial contribution to the evolution of NE can be distinguished in two parts. First, he observed that despite the advantages of digital computers in terms of precision and noise-free computation, the principles of the physics of neural computation are analog, rather than digital. For example, to compensate for the lack of parallelism, which is a key feature of the brain, computers have to run instructions faster, albeit at the cost of energy. Therefore, Mead was the first in the NE area to exploit the analog properties of transistors, rather than simply operating them as on–off switches [3]. His second observation regarding the common physical characteristics between protein channel in neurons and analog neuromorphic circuits resulted in the exploitation of the low power that transistors consume when operating below their threshold voltage. Specifically, he noticed that neuronal ion channels have a sigmoid

Copublished by the IEEE CEDA, IEEE CASS, IEEE SSCS, and TTTC

input–output characteristic similar to the one that output current and input voltage have in complementary metal–oxide–semiconductor (CMOS) transistors operating in subthreshold [4].

The ultimate goal of NE is to demonstrate cognitive systems using hardware neural processing architectures integrated with physical bodies, e.g., humanoid robots [3]. Therefore, real-time response is a fundamental requirement. However, simulating large networks of neurobiological systems with digital computers proves to be unrealistic in terms of time. For example, the Blue Gene rack, a 2048-processor computer, needs 1 h and 20 min to simulate 1 s of neural activity in a relatively complex scenario entailing eight million neurons and four billion synapses. Similar time limitations are also observed when graphic processing units (GPUs) and field-programmable gate arrays (FPGAs) are used for simulating complex networks. These observations resulted in the advent of analog VLSI implementations of silicon neurons (SiNs) which lie between biological neurons and digital computers, in terms of area and power, as shown in Figure 1 [4], and which can emulate brain computation mechanisms in real time. To date, neuromorphic silicon architectures have been used for implementing silicon retinas [5] and cochleas [6], SiNs and synapses, and networks of SiNs and synapses.

In this survey paper, we focus on presenting the general concepts of NE VLSI implementations and analyzing their building blocks. To this end, we first describe the basic structures of hardware models of spiking neurons, ranging from conductance-based and integrate-and-fire (I&F) structures to a mixture thereof. Next, we address plastic synapses and long-term weight storage mechanisms, as well as spike-based learning algorithms. Then, we discuss the asynchronous address–event-representation (AER) protocol, which is used to transmit spikes across chip boundaries and which contains both analog and digital components for local computation (on-chip) and long distance communication (off-chip), respectively. Combination of the above neuromorphic components results into single-chip or multichip spiking neural networks (SNNs), such as recurrent and winner take all (WTA) architectures. These neural networks have been proposed for performing a number of tasks, such as pattern recognition, working memory, and decision making. Finally, we discuss challenges on bridging the gap between reactive and cognitive systems, we present existing frontiers in NE, and we investigate potential solutions to emerging issues.

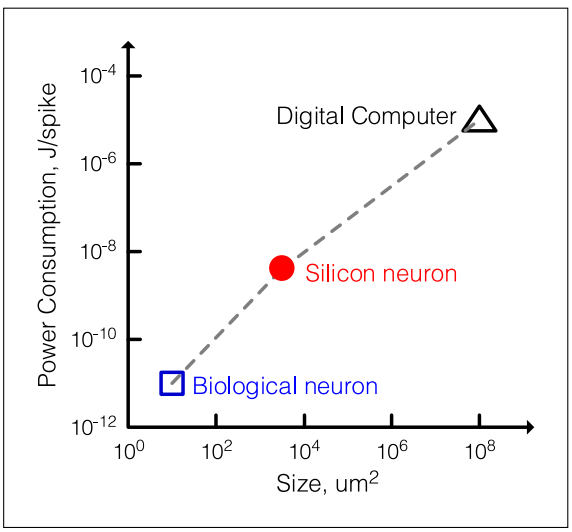

Figure 1. SiN compared to biological neuron and digital computer in terms of area and power [4].

Neuron circuits

The history of artificial neurons extends back to the 1940s with the first model proposed in 1943 by McCulloch and Pitts and its implementation appearing soon after. A first approach to neurorobotics was made in the 1950s by Walter implementing an electronic tube-based, neuron-like element to control a simple mobile robot. In 1958, Rosenblatt introduced the perceptron, an artificial neuron with a learning rule for classification tasks, which was also implemented in hardware.

Considering the neurobiological perspective, in 1952, Hodgkin and Huxley described the electrical activity of squid axons in a series of papers [7], eventually receiving the Nobel Prize in 1963. They showed that two types of channels are essential to generate an action potential (neuronal voltage response, whose typical form is shown in Figure 2 from the snail helisoma trivolvis) and they developed an electrical model to describe them. This model, which is shown in Figure 3, has become a typical circuit model simulating the physical neural activity. Despite its simplicity, this model has a key drawback for the circuit designer: the variable resistors (or conductances) g_K and g_{Na} are difficult to realize using simple circuit elements.

Inspired by the natural neural performance, extensive research efforts have been invested on spiking neurons and their hardware implementations. Based on the conductance model of Hodgkin and Huxley, Mahowald and Douglas proposed a SiN circuit with properties quite similar to those of real cortical neurons [8]. Other implementations of conductance-based models have also been proposed [9], [10]. The major drawback of these models is the silicon area they require, which makes them impractical for large neural network implementations. Simpler models are the Axon–Hillock circuit (Figure 4), proposed by Mead in the late 1980s [2]. In this circuit, an integrating capacitor is connected to two inverters, a feedback capacitor, and a reset transistor driven by the output inverter. The Axon–Hillock circuit is very compact and allows for large dense SiN arrays, but it suffers from a drawback of large power consumption due to the slow transition time (time constants in the order of milliseconds) imposed to the internal digital inverters, which are used as amplifiers. A further drawback is that it has a spiking threshold that only depends on CMOS process parameters (i.e., the switching threshold of the inverter) and does not model additional neural characteristics, such as spike-frequency adaptation properties or refractory period mechanisms [11]. A newer implementation is found in [12], consuming less power than previously proposed ones, but still lacking spike-frequency adaptation.

Several other variants have also been proposed for modeling many additional neural characteristics which are not included in the simple initial model. A recent circuit introduces a compact leaky I&F circuit which is optimized for power consumption and which implements spike-frequency adaptation, as well as tunable refractory period and voltage threshold modulation [11]. An even more recent neuron circuit integrated in the ROLLS neuromorphic processor chip [14] is derived from the adaptive exponential I&F circuit and can exhibit a wide range of neural behaviors, such as spike-frequency adaptation properties, refractory period mechanism, and adjustable spiking threshold mechanism.

Analog neuron implementations lying between the biology-inspired yet complex Hodgkin and Huxley model and the simplified Axon–Hillock circuit have also been developed. An extensive overview of such propositions can be found in [13].

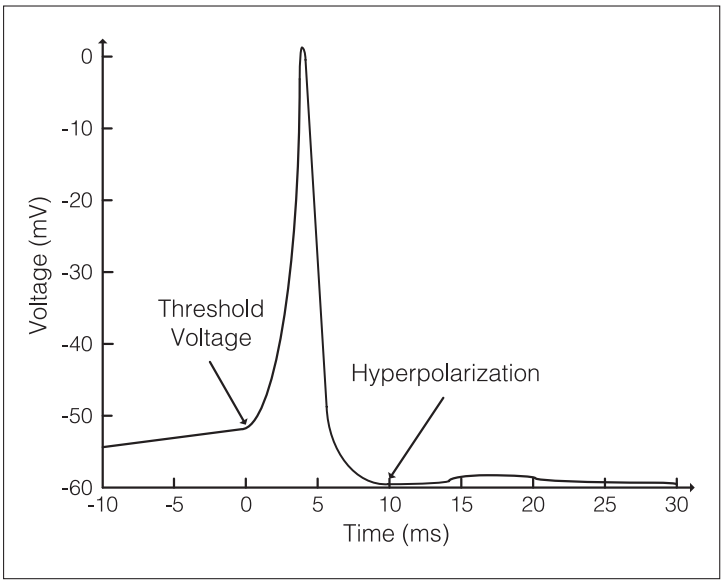

Figure 2. Typical action potential.

Synapses

The role of synapses in biological systems was first defined in 1897 by Charles Sherrington, who introduced the term synapse as the structure at the point of contact between two neurons which communicate. Projecting this definition onto VLSI circuits, synapses can implement a multiplication between the neuron input signal and its corresponding synaptic weight, in the case of classical neural networks [15]. In pulse-based neural networks, along with multiplication, synapses can also carry out linear or nonlinear integration of the input spikes with elaborate temporal dynamics and

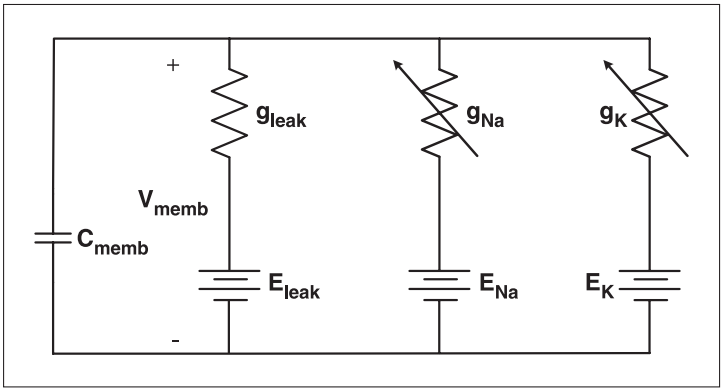

Figure 3. Original circuit model of neural electrical conductivity as devised by Hodgkin and Huxley [7].

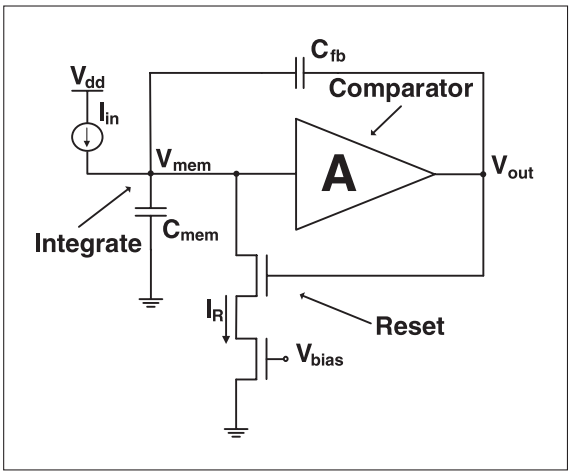

Figure 4. The Axon–Hillock circuit [13].

short-term plasticity (STP) and long-term plasticity (LTP) mechanisms. Synaptic plasticity, i.e., the ability of synapses to adapt their gain over time in response to increases or decreases in their activity, is extremely important in biological neural systems, as it is generally accepted that learning in the brain and formation of memories arise from synaptic modifications [16]. Indeed, the ability of biological synapses to exhibit STP and LTP is one of their fundamental features. In general, STP produces dynamic modulation of the synaptic strength by the timing of the input stimulation, whereas LTP produces long-term modifications on the synaptic gain based on the presynaptic and postsynaptic history [17]. A more detailed description on plasticity mechanisms is provided in the next section.

Silicon synapses which are able to reproduce the physics of real synapses have been demonstrated by several groups [14], [18]. In these implementations, the response of the synapse I_{syn}, i.e., the postsynaptic current (PSC), is a decaying exponential and is typically modeled as

$$I_{\text{syn}} = te^{-t/\tau_{\text{fall}}} \quad (1)$$

where τ_{fall} is typically in the order of 0.5–2 ms [19].

The main concern about neuromorphic synaptic models with learning capability is the storage of the synaptic weights, which are bounded and have limited precision. This constraint makes memory retrieval impossible, when new experiences, continuously producing new memories, saturate the storage capacity [20]. To address this problem, several synaptic storage implementations have been proposed, as follows.

1) Digital memory cells: A viable solution to the problem is digital memory cells. However, if the processing of the synapse is to be analog, additional analog-to-digital converter (ADC) and digital-to-analog-converter (DAC) are required to interface between the analog and digital worlds, thus, occupying more space on the chip die and increasing complexity and power consumption.
2) Capacitive storage: Capacitive storage provides a simple and analog alternative to digital memory cells. However, if the storage capacitance is connected to a transmission gate, the unavoidable leakage would call for mechanisms to mitigate the issue, e.g., using a bigger capacitor, at the cost of sacrificing area and power.
3) Floating-gate (FG) transistors: A completely analog, long-term, asynchronously accessible and nonvolatile storage can be realized using FG transistors. Employing a single FG device has enabled both PSC generation and long-term storage. Moreover, arraying single FG devices in a mesh architecture allows support of LTP learning approaches [15]. The major drawback of the FG device stems from the difficulties of precisely controlling the erase and programming process.
4) Bistable synapse: Using only two stable synaptic states can solve the problem of long-term storage, as has been shown in [20]. In this work, memory is preserved even in the absence of stimuli, or when the presynaptic activity is low, by using a bistable circuit that restores the synaptic state to either its high or its low rail, depending on whether the weight is above or below a certain threshold. Bistable synaptic dynamics make the synaptic weight more robust against spurious spikes, yet at the cost of plasticity sensitivity to temporal spike patterns, since multiple spike patterns may lead to the same binary synaptic weights [21].
5) Nanodevice technologies: Recent advances in nanotechnology have resulted in new devices which can carry out long-term multivalued weight storage and also allow for synaptic plasticity. Three such devices are the memristor [21], [22], the phase change memory [23], and the spin-transfer torque magnetic memory [24].

Learning algorithms

As previously mentioned, the two main plasticity mechanisms are STP and LTP. STP is effective for representing temporal signals and dynamics and comes in two different forms, depression and potentiation. The former occurs when postsynaptic potential falls rapidly, i.e., within 1 s or less, during repetitive stimulation, whereas the latter corresponds to a rapid growth in the postsynaptic potential, after repeated occurrences of a stimulus, also resulting in an increase of synaptic efficacy. Circuit implementations for these types of dynamics can be found in [3]. On the other hand, LTP occurs at excitatory synapses lasting minutes or more and plays a crucial role in learning. Particularly, in SNNs, spike-timing-dependent plasticity (STDP) constitutes the most intensively studied neural learning mechanism and can be considered as a spike-based formulation of the Hebbian learning rule [25]. In 1949, Hebb postulated the existence of synaptic strengthening, when a presynaptic neuron repeatedly takes part in firing a postsynaptic one. This process is commonly referred to as long-term potentiation in STDP, whereas long-term depression corresponds to the process in which anti-causal spike sequences with a postsynaptic spike preceding a presynaptic one lead to a decay in the synaptic weight. The change of synaptic connections $(\Delta w/w)$, where w is the synaptic weight, versus the relative timing of presynaptic and postsynaptic spikes is shown in Figure 5. STDP mechanisms, effective in learning how to classify spatio–temporal spike patterns, have been implemented in both analog and mixed-signal VLSI technologies, as described in [1].

To mitigate the constraints of bounded weight and limited precision of synapses, which was described in the previous section, Brader et al. [20] proposed a spike-based plasticity rule with bistable synaptics, which was implemented in silicon [1], [26]. This rule uses two stable states for every synapse, i.e., the depressing and potentiated states, using a stochastic mechanism for transitioning between them, based on spiking history. The protocol modifies only a random portion of all stimulated synapses with a small probability, significantly increasing storage capacity and lifetime memory of SNNs, which has been shown to increase inversely proportional with the probability of synaptic modification [27]. Unlike the STDP protocol, synaptic weight updates depend on: 1) the timing of the presynaptic spike; 2) the state of the postsynaptic neuron potential; and 3) a slow variable, related to the calcium concentration in biological neurons, proportional to the postsynaptic neuron mean firing rate. Brader's model also implements a "stop-learning" mechanism, preventing overfitting when the input pattern is highly correlated to the pattern stored in the synaptic weights. Overall, this bistable synaptic model is able to reproduce the STDP protocol and supports both unsupervised and supervised learning.

Information transmission

The number of connections between neurons in the brain is staggering. The human brain contains about 10^{11} neurons, while a cortical neuron typically makes 10^4 connections with other neurons. Thus, there is a total of 10^{15} point-to-point connections between neurons in the brain. Accordingly, the density of connections in large neural system implementations using analog integrated circuits comprises a major bottleneck. Although connections on an integrated circuit can be thinner than an axon, the brain uses its 3-D volume to route those connections, whereas integrated circuits and/or printed circuit boards have been limited, until recently, to a discrete number of 2-D layers.

To overcome this obstacle, the basic idea of AER, shown in Figure 6, is widely used. It practically trades in the advantage of speed of

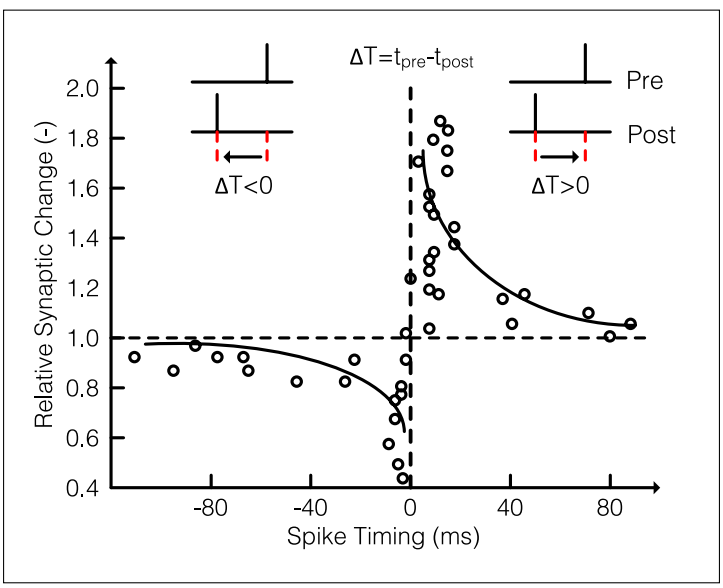

Figure 5. Change of synaptic connections as a function of timing between presynaptic and postsynaptic spikes.

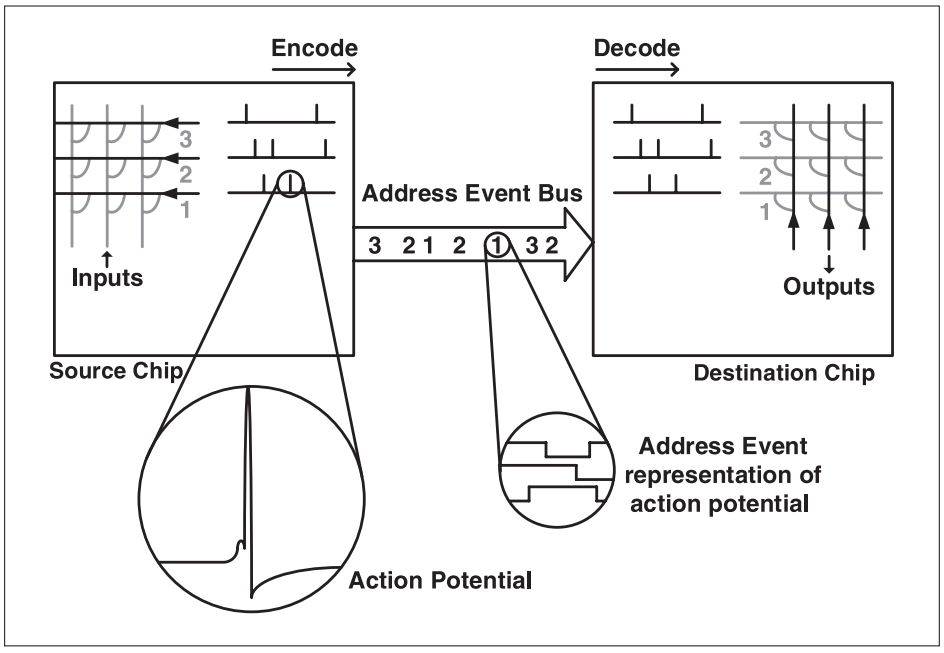

Figure 6. AER communication protocol [29].

electronics for its inferior interconnection density. Instead of an individual connection between each pair of neurons, two assemblies of neurons share one digital bus. An event (i.e., an action potential from a neuron) is encoded as a digital ID/address (i.e., a number that identifies the neuron producing the action potential) and is transmitted on this time-multiplexed, digital bus. On the receiver side, this address is, then, again converted into pulses, which are distributed to the receiving neurons that are connected to the sender [28].

The AER event-driven data representation and communication protocol was initially studied in Mead's lab by Mahowald and Sivilotti. It is an asynchronous handshaking protocol used to transmit signals between neuromorphic systems over a common communication bus, which is shared between chips and in which addresses are explicitly transmitted. Each transmission of an address is referred to as a spike. A unit in a system triggers a request for transmission when its internal state has crossed a threshold; its address is transmitted onto a common bus once this request is granted. Arbitration circuits on the periphery of the chip ensure that the addresses are sent off sequentially. The AER handshaking protocol ensures that the sender and the receiver write and read from the bus, respectively, only when they are allowed to. The activity level of each unit is represented by the frequency at which its address is transmitted.

The information being transmitted may be analog or digital, but must be communicated via spikes, thus raising the critical and exciting issue of signal encoding, which is currently a very active topic in neuroscience. Digital AER infrastructures allow construction of large multichip networks with nearly arbitrary connectivity and dynamic reconfiguration of the network topology for experimentation [3].

The AER protocol has been utilized in several neuromorphic systems. In cases of large neuromorphic chips, such as the ROLLS neuromorphic processor, SpiNNaker, HICANN, which are discussed in the following section, or in systems like CAVIAR [30], implementation variants of the AER technique can be found [17].

Neural networks

Combining SiNs, synapses, and learning mechanisms presented in the previous sections, networks of spiking neurons can be formed. In 1943, McCulloch and Pitts formulated the first neural network computing model. Based on this model, in 1958 Rosenblatt introduced a two-layer network, namely the perceptron, which was capable of solving linearly separable classification problems. Research in neural networks decelerated until the early 1980s when the backpropagation training algorithm was proposed, allowing for the construction of multilayer neural networks capable of solving more complicated tasks. These classical artificial neural networks (ANNs) can be viewed as an interconnection of processing elements where the strength of connections is controlled by synapses which act as multipliers of input signals and their local weight values. The sum of synaptic products is passed through a nonlinear activation function of a neuron. In contrast, SNNs differ from ANNs in two main points. First, SNNs incorporate the concept of time in neural simulation. Second, spike-based neurons and synapses emulate their biological counterparts.

Similarly to the classical neural networks, SNNs can be feedforward or recurrent. In a feedforward SNN, spike signals flow in only one direction, from input to output, one layer at a time, as depicted in Figure 7a. Addition of feedback loops allows the spike signals to flow in both directions and forms the recurrent neural network (RNN) shown in Figure 7b. Feedback adds some new properties in these networks such as associative memory [27] and context-dependent pattern classification, i.e., speech recognition [31], as compared to the feedforward neural networks which are mainly used for complex pattern classification. RNNs that perform a WTA computation are believed to play a central role in cortical processing. They can perform powerful computations, including nonlinear selection, signal restoration, and state-dependent processing [1]. In its simplest abstract form, a WTA network consists of a group of interacting neurons which compete with each other for activation. Neurons that receive the strongest input signal will suppress the activation of other neurons to win the competition. However, a variation that allows the activation of more than one neuron also exists, namely the soft WTA network.

Recently, spiking deep networks have also been proposed, aiming to overcome the large computational cost of the current state-of-the-art deep networks, such as convolutional and deep belief networks. Training these spiking deep networks is challenging because, instead of applying spike-based rules, a conversion from a conventional ANN, fully trained using backpropagation, into a spiking ANN is required. This conversion comes at a cost of performance losses [32].

There are numerous mixed-signal VLSI implementations of SNNs [9], [11]. These general-purpose computational networks consist of analog neurons that emulate the biophysics of real spiking neurons, synapses with STDP learning mechanisms, and the asynchronous AER communication protocol. The latter enables the configuration of a wide range of network topologies, including feedforward and recurrent networks. Additionally, due to their ability to simulate spike-based algorithms in real time, these VLSI networks may be interfaced to neuromorphic AER sensors, constructing VLSI sensory systems [30].

Six large-scale neuromorphic systems capable of simulating SNNs with a large number of neurons and synapses are presented below. Four of them are mixed-signal designs, in the sense that the circuits for neurons and synapses are analog, whereas the control of the analog circuit parameters, the network connectivity, and the multichip connectivity are based on asynchronous digital logic circuits. The remaining two are fully digital implementations. A more extensive review of these platforms can be found in [17].

The FACETS project [33] aims to develop a large-scale neuromorphic system, capable of implementing most of the neural systems modeled in computational neuroscience. To this end, a mixed-signal hardware neural network architecture consisting of analog neurons and a digital two-layer bus communication scheme was integrated on a wafer. The central element of this architecture is an analog neural network chip (HICANN), containing a total of 131072 synapses and up to 512 conductance-based adaptive exponential I&F neurons, which can be grouped together to form neurons with up to 14336 synapses. An important aspect of this type of neuron model is its operation at accelerated biological time. The acceleration factor ranges from 10^3 up to 10^5 as compared to the biological real time (BRT). Its synapses support both STP and LTP based on the STDP protocol mechanisms, whereas the synaptic weights are stored locally using 4-b SRAM cells. Following the AER paradigm, the digital communication scheme allows the integration of 384 HICANN chips on a single wafer. The main drawback of this wafer-scale system is the power consumption of 1 KW, which is mainly attributed to the acceleration scheme.

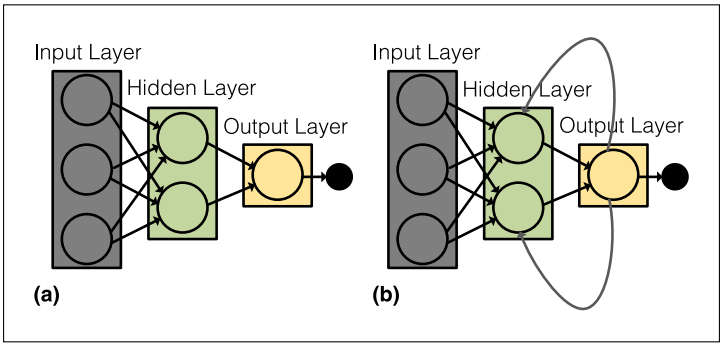

Figure 7. (a) Feedforward neural network. (b) Recurrent neural network.

The Neurogrid project [18] implements a mixed-signal neuromorphic hardware platform which serves as a brain simulation tool for neuroscientists. It uses analog computation to emulate ion-channel activity and a digital communication scheme to support synaptic connections. Its central processing element is called neurocore and contains 65536 quadratic I&F neurons which operate in BRT and use energy-efficient subthreshold operation. Neurogrid consists of 16 neurocores, yielding a total of 1048576 neurons and up to six billion synapses, which is much higher than in the FACETS project. Digital communication between neurocores is achieved by employing an external FPGA and a bank of SRAMs. Unlike the fully connected network of the HICANN chip, with separate circuits for every single synapse, all synapses of a neurocore neuron share only four synaptic population circuits, also feeding its neuron's neighbors. Although this results in a more compact implementation with considerably larger numbers of synapses per neuron, it also imposes a limitation as it precludes synaptic plasticity. It should be mentioned that the individual weights of the synapses are stored in a 1-b dedicated RAM. Finally, the overall system consumes only 5 W.

The HRL SyNAPSE project [34] developed a mixed-signal fully integrated neuromorphic chip which is scalable and can be configured to implement various large-scale neural network topologies. It uses analog neurons and, instead of the AER protocol, the synaptic time multiplexing (STM) paradigm is applied. STM is used both as a communication protocol and for configuring the network topology. The chip contains 576 nodes, each of which emulates a neuron with 128 synapses. As a result, the maximum number of neurons is 576 and the corresponding number of synaptic connections is 73728. Contrary to the two previous architectures, a more basic leaky I&F neuron model was designed. The neuron operates in BRT and has very low power consumption. Each node contains only a single synaptic circuit but, due to STM, this circuit can compute multiple (up to 128) logical synapses, provided that it is able to operate at much higher speeds than the neurons. The corresponding synaptic weights can be stored in memristor arrays. Furthermore, the synapses support STDP but, compared to the HICANN chip, only a limited number of different STDP types is supported. Similarly to neurogrid, the overall power consumption of the whole neuromorphic chip is as low as 130 mW.

The ROLLS neuromorphic processor [14] is also a mixed-signal VLSI architecture which can be used as an experimental platform for exploring the properties of computational neuroscience, as well as for developing brain-inspired large-scale neuromorphic systems. Once again, the synapse and neuron circuits are implemented in the analog domain and are combined with digital circuits, such as latches and asynchronous digital AER logic blocks. The ROLLS neuromorphic processor operates in real time, consumes approximately 4 mW and contains, in total, 256 neurons and 131072 synapses. An adaptive exponential I&F neuron model similar to the one used in the HICANN chip is also employed here. By default, each neuron is connected to a specific set of 512 synapses. Half of them are learning synapses, modeling the LTP mechanisms, and the other half are STP synapses with programmable synaptic weights. LTP is implemented based on the bistable spike-driven plasticity rule [20]. Additionally, a control circuit is used to enable allocation of multiple sets of synapses to the neurons by disconnecting and sacrificing the unused neurons. Moreover, there are 512 extra virtual synapses for modeling all types of synapses that have shared weights and time constants.

SpiNNaker [35] is a massively parallel, fully digital multiprocessor architecture for modeling and simulating large-scale SNNs. It can model neural networks up to a billion neurons and a trillion synapses with computations performed in BRT. The main processing building block is the SpiNNaker chip multiprocessor (CMP), which contains 18 ARM microprocessors and two routers. One of the routers handles the communication between the microprocessors and the peripherals, while the other handles the communication between microprocessors of different CMPs. Each of the microprocessors is capable of simulating up to 1000 spiking neurons and around 1000 synapses. However, out of the 18 microprocessors, only 16 are available for emulating neurons, as the rest are used for monitoring and fault tolerance purposes. The architecture allows for an integration of up to

65536 CMPs, with each CMP consuming 1 W. The fact that ARM microprocessors are used makes it possible to implement arbitrary neuron models and learning rules, a big advantage when compared to the analog neuromorphic chips previously analyzed. Furthermore, the synaptic weights are 16-b quantities stored in a tightly coupled memory (TCM) or SDRAM and the asynchronous AER protocol is used to model the massive interconnectivity of SNNs.

IBM TrueNorth [36] is also a fully digital neuromorphic chip. It is fabricated in Samsung's 28-nm CMOS process technology and is well-suited for many applications that use complex low-power neural networks in real time, such as classification and multiobject detection tasks. It is assembled from 4096 parallel and distributed neurosynaptic cores, interconnected in an on-chip mesh network that integrates one million programmable spiking neurons and 256 million configurable synapses. The digital neurons in the neurosynaptic cores are an implementation of the standard leaky I&F neuron model. Each core has allocated 100 kb on-chip SRAM memory in order to store synaptic and neuron parameters, as well as a router. The router passes spike packets between adjacent cores and delivers them to target cores, asynchronously. Although TrueNorth is a flexible architecture that leverages advances in packaging, 3-D integration, and novel devices, synaptic plasticity mechanisms are not supported.

Table 1 summarizes the most important features of the aforementioned six neuromorphic platforms. Measurements related to power are not from the same benchmark so they cannot be compared directly. Specifically, for the HICANN chip, there are no power data available in the literature. Also, the fact that, in some cases, the number of neurons and synapses varies is because the supported neural network topologies are not fixed, i.e., they can be reconfigured.

Discussion

In the low-power, mixed-signal, neuromorphic implementations surveyed herein, analog CMOS circuits are used for the design of SiNs and synapses. These silicon circuits typically operate in the subthreshold region and are, thus, susceptible to 1) noise and 2) mismatch, which become dominant when the device operating voltage falls well below the device threshold voltage. A common mismatch reduction technique is to increase device dimensions. However, in the case of large neuromorphic networks with many neurons and

TABLE 1 Main characteristics of six large neuromorphic platforms.

Architecture	HICANN	Neurogrid	HRL SyNAPSE	ROLLS	SpiNNaker CMP	IBM TrueNorth
Type	Analog+AER	Analog+AER	Analog+STM	Analog+AER	ARM CPU+AER	Digital
Neurons	1-512	1048576	576	1-256	1000 per CPU	10^6
Synapses per neuron	224-14336	$6*10^9$ total	128	512-131072	1000	256
Learning rules - Synapse storage	STP,STDP, 4-bit SRAM	None, 1-bit RAM	STDP, 8-level memristors	STP,STDP, 1-bit bistable synapses	Any, 16-bit TSM,SDRAM	None, SRAM
Speed (BRT)	10^3-10^5	1	1	1	1	1
Power (W)	n/a	5	130m	4m	1	63m
Size (mm^2)	50	168 (neurocore)	42.25	51.4	102	430
CMOS Process (nm)	180	180	90	180	130	28

synapses, this is impractical, since the area required for integrating these structures into a single chip would become prohibitive. To alleviate this problem, processes that are dedicated to low-power, subthreshold, system-on-chip (SoC) circuits, e.g., depleted silicon-on-insulator (SOI), thereby reducing the threshold voltage variation, have been available for implementing neuromorphic analog VLSI circuits [4]. In [27], the inherent CMOS device mismatch is addressed at the network and system level, where it is argued that plasticity mechanisms (both STP and LTP) are robust to device mismatch and do not require precisely matched transistors.

The NE community has made a remarkable progress in building technology platforms for simulating neurobiological models. However, additional steps are required toward the ultimate goal of implementing hardware neural processing systems which are: 1) autonomous; 2) able to interact with the environment in real time; and 3) able to express cognitive abilities. These steps are not hinging upon scaling or hardware restrictions but, rather, on a better understanding of computational principles used by the brain and computational models that neurons can support. Thus, the role of neuroscience and psychology in identifying these principles is crucial, before neuromorphic cognitive systems can become a reality. ■

■ References

[1] G. Indiveri, E. Chicca, and R. J. Douglas, "Artificial cognitive systems: From VLSI networks of spiking neurons to neuromorphic cognition," *Cognit. Comput.*, vol. 1, no. 2, pp. 119–127, 2009.

[2] C. Mead, *Analog VLSI and Neural Systems*. Boston, MA, USA: Addison-Wesley, 1989.

[3] G. Indiveri and T. K. Horiuchi, "Frontiers in neuromorphic engineering," *Front. Neurosci.*, vol. 5, no. 118, 2011, DOI: 10.3389/fnins.2011.00118.

[4] C.-S. Poon and K. Zhou, "Neuromorphic silicon neurons and large-scale neural networks: Challenges and opportunities," *Front. Neurosci.*, vol. 5, no. 108, 2011, DOI: 10.3389/fnins.2011.00108.

[5] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 × 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE J. Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, Feb. 2008.

[6] V. Chan, S. C. Liu, and A. van Schaik, "AER EAR: A matched silicon cochlea pair with address event representation interface," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 54, no. 1, pp. 48–59, Jan. 2007.

[7] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *J. Physiol.*, vol. 117, no. 4, pp. 500–544, Aug. 1952.

[8] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, no. 6354, pp. 515–518, Dec. 1991.

[9] R. J. Vogelstein, U. Mallik, J. T. Vogelstein, and G. Cauwenberghs, "Dynamically reconfigurable silicon array of spiking neurons with conductance-based synapses," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 253–265, Jan. 2007.

[10] L. Alvado et al., "Hardware computation of conductance-based neuron models," *Neurocomputing*, vol. 58–60, pp. 109–115, 2004.

[11] G. Indiveri, E. Chicca, and R. Douglas, "A VLSI array of low-power spiking neurons and bistable synapses with spike-timing dependent plasticity," *IEEE Trans. Neural Netw.*, vol. 17, no. 1, pp. 211–221, Jan. 2006.

[12] A. van Schaik, "Building blocks for electronic spiking neural networks," *Neural Netw.*, vol. 14, no. 6–7, pp. 617–628, 2001.

[13] G. Indiveri et al., "Neuromorphic silicon neuron circuits," *Front. Neurosci.*, vol. 5, no. 73, 2011, DOI: 10.3389/fnins.2011.00073.

[14] N. Qiao et al., "A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128K synapses," *Front. Neurosci.*, vol. 9, no. 141, 2015, DOI: 10.3389/fnins.2015.00141.

[15] D. Maliuk and Y. Makris, "An experimentation platform for on-chip integration of analog neural networks: A pathway to trusted and robust analog/RF ICs," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 8, pp. 1721–1734, Aug. 2015.

[16] R. M. Wang, T. J. Hamilton, J. Tapson, and A. van Schaik, "A neuromorphic implementation of multiple spike-timing synaptic plasticity rules for large-scale neural networks," *Front. Neurosci.*, vol. 9, no. 180, 2015, DOI: 10.3389/fnins.2015.00180.

[17] F. Walter, F. Röhrbein, and A. Knoll, "Neuromorphic implementations of neurobiological learning algorithms for spiking neural networks," *Neural Netw.*, vol. 72, pp. 152–167, 2015.

[18] B. V Benjamin et al., "Neurogrid: A mixed-analog-digital multichip system for large-scale neural simulations," *Proc. IEEE*, vol. 102, no. 5, pp. 699–716, May 2014.

[19] J. Hasler and H. B. Marr, "Finding a roadmap to achieve large neuromorphic hardware systems," *Front. Neurosci.*, vol. 7, no. 118, 2013, DOI: 10.3389/fnins.2013.00118.

[20] J. M. Brader, W. Senn, and S. Fusi, "Learning real-world stimuli in a neural network with spike-driven synaptic dynamics," *Neural Comput.*, vol. 19, no. 11, pp. 2881–2912, Nov. 2007.

[21] H. Mostafa et al., "Implementation of a spike-based perceptron learning rule using TiO(2-x) memristors," *Front. Neurosci.*, vol. 9, 2015, DOI: 10.3389/fnins.2015.00357.

[22] A. Thomas and E. Chicca, "Tunnel junction based memristors as artificial synapses," *Front. Neurosci.*, vol. 9, no. 241, 2015, DOI: 10.3389/fnins.2015.00241.

[23] S. Ambrogio et al., "Unsupervised learning by spike timing dependent plasticity in phase change memory (PCM) synapses," *Front. Neurosci.*, vol. 10, no. 56, 2016, DOI: 10.3389/fnins.2016.00056.

[24] X. Fong et al., "Spin-transfer torque devices for logic and memory: Prospects and perspectives," *IEEE Trans. Comput. Design Integr. Circuits Syst.*, vol. 35, no. 1, pp. 1–22, Jan. 2016.

[25] "Hebb, D. O. Organization of behavior. New York: Wiley, 1949, p. 335, 4.00," *J. Clin. Psychol.*, vol. 6, no. 3, p. 307, 1950.

[26] S. Mitra, S. Fusi, and G. Indiveri, "Real-time classification of complex patterns using spike-based learning in neuromorphic VLSI," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 1, pp. 32–42, Feb. 2009.

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

[28] P. Hafliger, "Adaptive WTA with an analog VLSI neuromorphic learning chip," *IEEE Trans. Neural Netw.*, vol. 18, no. 2, pp. 551–572, Mar. 2007.

[29] G. Indiveri, T. Horiuchi, E. Niebur, and R. J. Douglas, "A competitive network of spiking {VLSI} neurons," in *Proc. World Congr. Neuroinf.*, 2001, pp. 443–455.

[30] R. Serrano-Gotarredona et al., "CAVIAR: A 45k Neuron, 5M Synapse, 12G connects/s AER hardware sensory–processing–learning–actuating system for high-speed visual object recognition and tracking," *IEEE Trans. Neural Netw.*, vol. 20, no. 9, pp. 1417–1438, Sep. 2009.

[31] M. Rigotti, D. B. D. Rubin, S. E. Morrison, C. D. Salzman, and S. Fusi, "Attractor concretion as a mechanism for the formation of context representations," *Neuroimage*, vol. 52, no. 3, pp. 833–847, 2010.

[32] P. U. Diehl et al., "Fast-classifying, high-accuracy spiking deep networks through weight and threshold balancing," in *Proc. Int. Joint Conf. Neural Netw.*, 2015, DOI: 10.1109/IJCNN.2015.7280696.

[33] J. Schemmel, J. Fieres, and K. Meier, "Wafer-scale integration of analog neural networks," in *Proc. IEEE Int. Joint Conf. Neural Netw./IEEE World Congr. Comput. Intell.*, 2008, pp. 431–438.

[34] J. M. Cruz-Albrecht, T. Derosier, and N. Srinivasa, "A scalable neural chip with synaptic electronics using CMOS integrated memristors," *Nanotechnology*, vol. 24, no. 38, 2013, 384011.

[35] E. Painkras et al., "SpiNNaker: A 1-W 18-core system-on-chip for massively-parallel neural network simulation," *IEEE J. Solid-State Circuits*, vol. 48, no. 8, pp. 1943–1953, Aug. 2013.

[36] P. A. Merolla et al., "A million spiking-neuron integrated circuit with a scalable communication network and interface," *Science*, vol. 345, no. 6197, pp. 668–673, 2014.

Georgios Volanis is currently working toward a PhD in electrical engineering at The University of Texas at Dallas, Richardson, TX, USA. His research interests include hardware implementations of learning systems for robust and trustworthy analog/RF ICs. Volanis has an MSc in electronics from The University of Edinburgh, Edinburgh, U.K. He is a Student Member of the IEEE.

Angelos Antonopoulos is a Postdoctoral Research Associate at The University of Texas at Dallas, Richardson, TX, USA. His research interests include the design of robust and trusted integrated circuits and systems. Antonopoulos has a PhD in electronic engineering from Technical University of Crete, Chania, Greece. He is a Member of the IEEE.

Yiorgos Makris is a Professor of Electrical Engineering at The University of Texas at Dallas, Richardson, TX, USA. His research focuses on applications of machine learning in test, reliability, and security of ICs, with particular emphasis in the analog/RF domain. Makris has a PhD in computer engineering from the University of California San Diego, La Jolla, CA, USA. He is a Senior Member of the IEEE.

Alkis A. Hatzopoulos is with the Electronics Laboratory, Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki, Thessaloniki, Greece, where he is the Director. His research interests include design and fault diagnosis of integrated circuits and systems (analog, mixed-signal, RF, and 3-D). Hatzopoulos has a PhD from the same Department (1989). He is a Senior Member of the IEEE.

■ Direct questions and comments about this article to Georgios Volanis, Department of Electrical Engineering, The University of Texas at Dallas, Richardson, TX, 75080, USA; gxv130830@utdallas.edu.