

Neuromorphic Electronic Systems for Reservoir Computing

Fatemeh Hadaeghi

Department of Computer Science and Electrical Engineering, Jacobs University
Bremen, 28759 Bremen, Germany

Institute of Computational Neuroscience, University Medical Center
Hamburg-Eppendorf (UKE), 20251 Hamburg, Germany
f.hadaeghi@uke.de

16 April 2019

Abstract. This chapter provides a comprehensive survey of the researches and motivations for hardware implementation of reservoir computing (RC) on neuromorphic electronic systems. Due to its computational efficiency and the fact that training amounts to a simple linear regression, both spiking and non-spiking implementations of reservoir computing on neuromorphic hardware have been developed. Here, a review of these experimental studies is provided to illustrate the progress in this area and to address the technical challenges which arise from this specific hardware implementation. Moreover, to deal with challenges of computation on such unconventional substrates, several lines of potential solutions are presented based on advances in other computational approaches in machine learning.

Keywords: Analog Microchips, FPGA, Memristors, Neuromorphic Architectures, Reservoir Computing

1 Introduction

The term “neuromorphic computing” refers to a variety of brain-inspired computers, architectures, devices, and models that are used in the endeavor to mimic biological neural networks [1]. In contrast to von Neumann architectures, biologically inspired neuromorphic computing systems are promising for being highly connected and parallel, incorporating learning and adaptation, collocating memory and processing and requiring low-power. By creating parallel arrays of connected synthetic neurons that are asynchronous, real-time, and data- or event-driven, neuromorphic devices offer an expedient substrate to model neuroscience theories as well as implementing computational paradigms to solve challenging machine learning problems.

The accustomed growth rates of digital computing performance levels (Moore’s Law) is showing signs of flattening out [2]. Furthermore, the exploding energy demand of digital computing devices and algorithms is approaching the limits of what is socially and environmentally tolerable [3]. Neuromorphic technologies suggest escape routes from both predicaments due to their potentials for

unlocked parallelism and minimal energy consumption of spiking dynamics. Moreover, neuromorphic systems have also received increased attention due to their scalability and small device footprint [4].

Significant keys to such advancements are remarkable progress in material science and nanotechnology, low-voltage analog CMOS design techniques and theoretical and computational neuroscience. At the device level, using the new materials and nanotechnologies for building extremely compact and low-power solid-state nanoscale devices, has paved the way towards on-chip synapses with characteristic properties observed in biological synapses. For instance, “memory resistor” or the memristor, a non-linear nanoscale electronic element with volatile and non-volatile modes, is ubiquitously used in neuromorphic circuits to store multiple bits of information and to emulate dynamic weights with intrinsic plasticity features (e.g., spike time dependent plasticity (STDP)) [5]. It has been argued that hybrid memristor-CMOS neuromorphic circuit may represent a proper building block for implementing biological-inspired probabilistic/stochastic/approximate computing paradigms that are robust to memristor device variability and fault-tolerant by design [6–8]. Similarly, conductive-bridging RAM (CBRAM) [9], which is a non-volatile memory technology, and atomic switches [10] – nanodevices implemented using metal-oxide based memristors or memristive materials – have also been fabricated to implement both short-term plasticity (STP) and long-term plasticity (LTP) in neuromorphic systems [11,12]. Both atomic switches and CBRAM have nano dimensions, are fast, and consume low energy [4]. Spike- time-dependent-depression and -potentiation observed in biological synapses have also been emulated by phase change memory (PCM) elements in hybrid neuromorphic architectures where CMOS “neuronal” circuits are integrated with nanoscale “synaptic” devices. PCM elements are commonly used to achieve high density and scalability; they are compatible with CMOS circuits and show a good endurance [13,14]. Programmable metallization cells (PMCs) [15] and oxide-resistive memory (OXRAM) [16] are another types of resistive memory technologies that have been demonstrated to present STDP-like characteristics. Beyond the CMOS technologies for neuromorphic computing, spintronic devices and optical (photonic) components have also been considered for neuromorphic implementation [4,17].

At the hardware level, a transition away from purely digital systems to mixed analog/digital implementation to purely analog, unlocked, spiking neuromorphic microchips has led to the emergence of more biological-like models of neurons and synapses together with a collection of more biologically plausible adaptation and learning mechanisms. Digital systems are usually synchronous or clock-based, rely on Boolean logic-based gates and discrete values for computation and tend to need more power. Analog systems, in contrast, tend to rely more on the continuous values and inherent physical characteristics of electronic devices for computation and more closely resemble the biological brain which takes the advantages of physical properties rather than Boolean logic for computation [18]. Analog systems, however, are significantly vulnerable to various types of thermal and electrical noise and artifacts. It is, therefore, argued that

only computational paradigms that are robust to noise and faults may be proper candidates for analog implementation. Hence, among the wide diversity of computational models, a leading role is emerging for mixed analog/digital or purely analog implementation of artificial neural networks (ANNs). Two main trends in this arena are the exploit of low-energy, spiking neural dynamics for “deep learning” solutions [19], and “reservoir computing” methods [20].

At the network and system level, new discoveries in neuroscience research are being gradually incorporated into hardware models and mathematical theory frameworks are being developed to guide the algorithms and hardware developments. A constructive paradigm shift has also occurred from strict structural replication of neuronal systems towards the hardware implementation of systemic/functional models of the biological brain [21]. Consequently, over the past decade, a wide variety of model types have been implemented in electronic multi-neuron computing platforms to solve pattern recognition and machine learning tasks [18, 22].

As mentioned above, among these computational frameworks, reservoir computing (RC) has been variously considered as a strategy to implement useful computations on such unconventional hardware platforms. An RC architecture comprises three major parts: the *input layer* feeds the input signal into a random, large, fixed recurrent neural network that constitutes the *reservoir*, from which the neurons in the *output layer* read out a desired output signal. In contrast to traditional (and “deep”) recurrent neural networks (RNNs) training methods, the input-to-reservoir and the recurrent reservoir-to-reservoir weights in an RC system are left unchanged after a random initialization, and only the reservoir-to-output weights are optimized during training. Within computational neuroscience, RC is best known as *liquid state machines* (LSMs) [23], whereas the approach is known as *echo state networks* (ESNs) [20] in machine learning. Reflecting the different objectives in these fields, LSM models are typically built around more or less detailed, spiking neuron models with biologically plausible parametrizations, while ESNs mostly use highly abstracted rate models for its neurons. Due to its computational efficiency, simplicity and lenient requirements, both spiking and non-spiking implementations of reservoir computing on neuromorphic hardware exist. However, proper solutions are still lacking to address a variety of technological and information processing problems. For instance, regarding the choice of hardware, variations of stochasticity due to device mismatch, temporal drift, and aging must be taken into account; in case of exploiting the spiking neurons, an appropriate encoding scheme needs to be developed to transform the input signal into spike trains; the speed of physical processes may require further adjustment to achieve on-line learning and real-time information processing; depending on the available technology, compatible local learning rules must be developed for on-chip learning.

Here, a review of recent experimental studies is provided to illustrate the progress in neuromorphic electronic systems for RC and to address the above-mentioned technical challenges which arise from such hardware implementations. Moreover, to deal with challenges of computation on such unconventional sub-

strate, several lines of potential solutions are presented based on advances in other computational approaches in machine learning. In the remaining part of this chapter, we present an overview of the current approaches to implement reservoir computing model on digital neuromorphic processors, purely analog neuromorphic microchips and mixed digital/analog neuromorphic systems. Since the neuromorphic computing attempts to implement more biological-like models of neurons and synapses, spiking implementations of RC will be highlighted.

2 RC on Digital Neuromorphic Processors

Field programmable gate arrays (FPGAs) and application specific integrated circuit (ASIC) chips – two categories of digital systems – have been very commonly utilized for neuromorphic implementations. To facilitate the application of computational frameworks in both embedded and stand-alone systems, FPGAs and ASICs offer considerable flexibility, reconfigurability, robustness, and fast prototyping [24].

As illustrated in Figure 1, an FPGA-based reservoir computer consists of neuronal units, a neuronal arithmetic section, input/output encoding/decoding components, an on-chip learning algorithm, memory control, numerous memory blocks such as RAMs and/or memristors wherein the neuronal and synaptic information is stored, and a Read/Write interface to access to these memory blocks. Realizing spiking reservoir computing on such substrates entails tackling a number of critical issues related to spike coding, memory organization, parallel processing, on-chip learning and tradeoffs between area, precision and power overheads. Moreover, in real-world applications such as speech recognition and biosignal processing, spiking dynamics might be inherently faster than real-time performance. In this section, an overview of experimental studies is provided to illustrate how these challenges have been so far addressed in the literature.

In the first digital implementation of spiking RC, [25] suggested storing the neuronal information in RAM, exploiting non-plastic exponential synapses and performing neuronal and synaptic operations serially. In this clock-based simulation, each time step consists of several operations such as adding weights to the membrane or synapse accumulators, adding the synapse accumulators to the membrane's, decaying these accumulators, threshold detection and membrane reset. On a real-time speech processing task, this study shows that in contrast to “serial processing, parallel arithmetic” [26] and “parallel processing, serial arithmetic” [27], serial implementations of both arithmetic operations that define the dynamics of the neuron model and processing operations associated with synaptic dynamics yield to slower operations and low hardware costs. Although this architecture can be considered as the first compact digital neuromorphic RC system to solve a speech recognition task, the advantage of distributed computing is not explored in this framework. Moreover, for larger networks and in more sophisticated tasks, a sequential processor – which calculates every neuronal unit separately to simulate a single time step of the network – does not

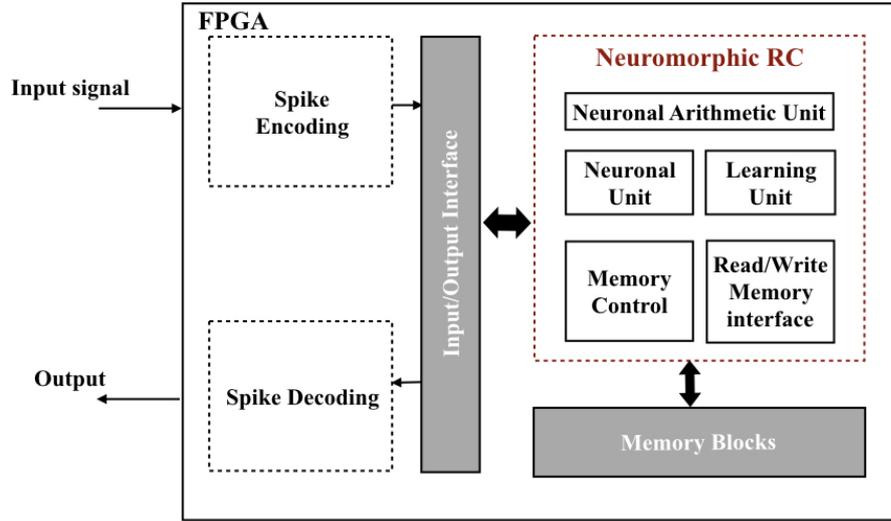


Fig. 1. A top-level schematic of an FPGA-based neuromorphic reservoir computer.

seem to be efficient. Stochastic arithmetic was, therefore, exploited to parallelize the calculations and obtain a considerable speedup [28].

More biologically plausible models of spiking neurons (e.g., Hodgkin–Huxley, Morris–Lecar and Izhikevich models [29]) are too sophisticated to be efficiently implemented on hardware and have many parameters which need to be tuned. On the other hand, the simple hardware-friendly spiking neuron models, such as leaky-integrate-and-fire (LIF) [29], often have a hard thresholding that makes supervised training difficult in spiking neural networks (SNNs). In spiking reservoir computing, however, the training of the readout mechanism amounts only to solving a linear regression problem, where the target output is a trainable linear combination of the neural signals within the reservoir. Linear regression is easily solved by standard linear algebra algorithms when arbitrary real-valued combination weights are admitted. However, for on-chip learning, the weights will be physically realized by states of electronic synapses, which currently can be reliably set only to a very small number of discrete values. It has been recently proved that computing optimal discrete readout weights in reservoir computing is NP-hard and approximate (or heuristic) methods must be exploited to obtain high-quality solutions in reasonable time for practical uses [30]. The spike-based learning algorithm proposed by [31] is an example of such approximate solutions for FPGA implementations. In contrast to offline learning methods, the proposed online learning rule avoids any intermediate data storage. Besides, through this abstract learning process, each synaptic weight is adjusted based on only the firing activities of the corresponding pre- and post-synaptic neurons, independent of the global communications across the neural network. In a speech recogni-

tion task, it has been shown that due to this locality, the overhead of hardware implementation (e.g., synaptic and membrane voltage precision) can be reduced without drastic effects on its performance [31]. This biological-inspired supervised learning algorithm was later exploited in developing an RC-based general-purpose neuromorphic architecture where training the task-specific output neurons is integrated into the reconfigurable FPGA platform [32]. The effects of catastrophic failures such as broken synapses, dead neurons, random errors as well as errors in arithmetic operations (e.g., comparison, adding and shifting) in similar architecture were addressed in [33] where the reservoir consists of excitatory and inhibitory LIF neurons that are randomly connected with non-plastic synapses governed by second-order response functions. The simulation results suggest that at least 8-bit resolution is needed for the efficacy of plastic synapses if the spike-based learning algorithm proposed by [31] is applied for training the readout weights. The recognition performance only degrades slightly when the precision of fixed synaptic weights and membrane voltage for reservoir neurons are reduced down to 6 bits.

To realize the on-chip learning on digital systems with extremely low-bit precision (binary values), [34] suggested an online pruning algorithm based on variances of firing activities to sparsify the readout connections. The cores to this reconfiguration scheme are STDP learning rule, firing activity monitoring, and variance estimation. The readout synapses projected from low-variance reservoir neurons are, then, powered off to save energy. To reduce the energy consumption of the hardware in FPGA, the reservoir computer developed by [24, 35] utilizes firing activity based power gating by turning off the neurons that seldom fire for a particular benchmark, and applies approximate arithmetic computing [36] to speed up the runtime in a speech recognition task.

Exploiting the spatial locality in a reservoir consisting of excitatory and inhibitory LIF neurons, a fully parallel design approach was also presented for real-time biomedical signal processing in [37, 38]. In this design, the memory blocks were replaced by distributed memory to circumvent the long access time due to wire delay. Being inspired by new discoveries in neuroscience, [39] developed a spiking temporal processing unit (STPU) to efficiently implement more biologically plausible synaptic response functions in digital architectures. STPU offers a local temporal memory buffer including an arbitrary number of memory cells to model delayed synaptic transmission between neurons. This allows multiple connections between neurons with different synaptic latency. Utilizing excitatory and inhibitory LIF neurons with different time scales and exploiting the second-order response function to model synaptic transmission, the effects of the input signal in a spoke digit recognition task will only slowly “wash out” over time, enabling the reservoir to provide sufficient dynamical short-term memory capacity. In order to create a similar short-term memory on a reconfigurable digital platform with extremely low bit resolutions, an STDP mechanism was proposed by [34] for on-chip reservoir tuning. Applying this local learning rule to the reservoir synapses together with a data-driven binary quantization algorithm creates sparse connectivities within the reservoir in a self-organized

fashion, leading to significant energy reduction. Stochastic activity-based STDP approach [40], structural plasticity-based mechanism [41], and correlation-based neuron gating rule [42] have also been introduced for efficient low-resolution tuning of the reservoir in hardware. Neural processes, however, require a large number of memory resources for storing various parameters, such as synaptic weights and internal neural states, and lead to a heavy clock distribution loading and a significant power dissipation. To tackle this problem, [42] suggested partitioning the memory elements inside each neuron that are activated at different phases of neural processing. This leads to an activity-based clock gating mechanism with a granularity of a partitioned memory group inside each neuron.

Another technique for power reduction is incorporating memristive crossbars into digital neuromorphic hardware to perform synaptic operations in on-chip/online learning and to store the synaptic weights in offline learning [43]. In general, a two terminal memristor device can be programmed by applying a large enough potential difference across its terminals, where the state change of the device is dependent on the magnitude, duration, and polarity of the potential difference. The device resistance states, representing synaptic weights, will vary between a high resistance state (HRS) or low resistance state (LRS), depending on the polarity of the voltage applied. At a system level, various sources of noise (e.g., random telegraph, thermal noise, and 1/f noise) arise from non-ideal behavior in memristive devices and distort the process of reading from the synapses and updating the synaptic weights. Given theoretical models for these stochastic noise processes, the effects of different manifestations of memristor read and write noise on the accuracy of neuromorphic RC in a classification task were investigated in [43]. A hardware feasible RC structure with memristor double crossbar array has also been proposed in [44] and was tested on a real-time time series prediction task. The two studies not only confirm that reservoir computing can properly cope with low-precision environments and noisy data, but they also experimentally demonstrate how RC benefits from memristor device variation to secure a more random heterogeneous weight distribution leading to more diverse response for similar inputs. Although memristive crossbars are area and energy efficient, the digital circuitry to control the read/write logic to the crossbars is extremely power hungry, thus preventing the use of memristors in large-scale memory systems.

In RC literature, it has also been shown that in time series prediction tasks, a cyclic reservoir –where neurons are distributed on a ring and connected with the same synaptic weights to produce the cyclic rotations of the input vector– is able to operate with an efficiency comparable to the standard RC models [45,46]. In order to minimize the required hardware resources, therefore, a single cyclic reservoir with stochastic spiking neurons was implemented for a real-time time series prediction task [47]. The architecture shows a high level of scalability and a prediction performance comparable to the software simulations. In non-spiking reservoir computing implementations on reconfigurable digital systems, the reservoir has sometimes been configured as a ring topology and the readout

weights are trained through gradient descent algorithm or ridge regression [48–50].

3 RC on Analog Neuromorphic Microchips

Digital platforms, tools and simulators offer robust and practical solutions to a wide range of engineering problems and provide convenient approaches to explore the quantitative behavior of neural networks. However, they do not seem to be ideal substrates for simulating the detailed large-scale models of neural systems where density, energy efficiency, and resilience are important. Besides, the observation that the brain operates on analog principles and relies on the inherent physical characteristics of the neuronal system for computation motivated the investigations in the field of neuromorphic engineering. Following the pioneering work conducted by Carver Mead [1] – therein biologically inspired electronic sensors were integrated with analog circuits and an address-event-based asynchronous, continuous time communications protocol was introduced – over the last decade, mixed analog/digital and purely analog very large scale integration (VLSI) circuits have been fabricated to emulate the electrophysiological behavior of biological neurons and synapses and to offer a medium in which neuronal networks can be presented directly in hardware rather than simply simulated on a general purpose computers [51]. However, the fact that such platforms provide only a qualitative approximation to the exact performance of digitally simulated neural systems may preclude them from being the ideal substrate for solving engineering problems and achieving machine learning tasks where detailed quantitative investigations are essential. In addition to the low resolution, realizing global asynchrony and dealing with noisy and unreliable components seem to be formidable technical hurdles. An architectural solution can be partly provided by the reservoir computing due to its bio-inspired principle, robustness to the imperfect substrate and the fact that only readout weights need to be trained [52]. Similar to other analog and digital computing chips, transistors are the building blocks of analog neuromorphic devices. It has been experimentally shown that in the “subthreshold” region of operation, the current-voltage characteristic curve of the transistor is exponential and analogous to the exponential dependence of active populations of voltage-gated ionic channels as a function of the potential across the membrane of a neuron. This similarity has paved the way towards the fabrication of compact analog circuits that implement electronic models of voltage-sensitive conductance-based neurons and conductance-based synapses as well as computational circuitry to perform logarithmic functions, amplification, thresholding, multiplication, inhibition, and winner-take-all selection [53, 54]. An example of biophysically realistic neural electronic circuits is depicted in Figure 2. This differential pair integrator (DPI) neuron circuit consists of four components: 1) the input DPI filter ($M_{L1} - M_{L3}$) including the integrating membrane capacitor C_{mem} , models the neuron’s leak conductance which produces exponential sub-threshold dynamics in response to constant input currents, 2) a spike event generating amplifier ($M_{A1} - M_{A6}$) together with a

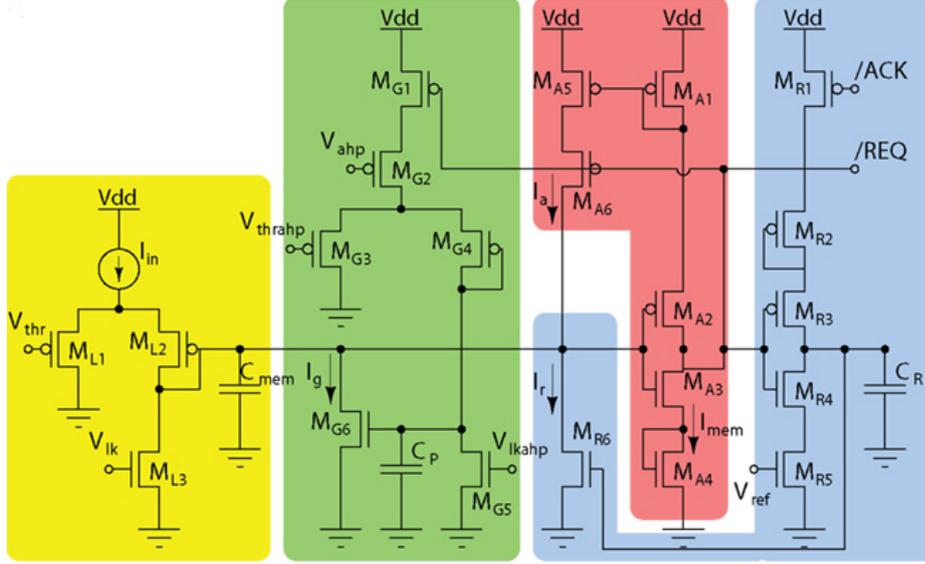


Fig. 2. A schematic of (DPI) neuron circuit (reproduced from [51]).

positive-feedback circuit represent both sodium channel activation and inactivation dynamics, 3) a spike reset circuit with address event representation (AER) handshaking signals and refractory period functionality ($M_{R1} - M_{R6}$) emulates the potassium conductance functionality and 4) a spike-frequency adaptation mechanism implemented by an additional DPI filter ($M_{G1} - M_{G6}$) produces an after hyper-polarizing current proportional to the neuron's mean firing rate. See [51] for an insightful review of different design methodologies used for silicon neuron fabrication.

In conjunction with circuitry that operates in subthreshold mode, exploiting memristive devices to model synaptic dynamics is a common approach in analog neuromorphic systems for power efficiency and performance boosting purposes [11, 18]. In connection with the fully analog implementation of RC, it has been shown that by creating a functional reservoir consisting of numerous neurons connected through atomic nano-switches, the functional characteristics required for implementing biologically inspired computational methodologies in a synthetic experimental system can be displayed [11]. A memristor crossbar structure has also been fabricated to connect the analog non-spiking reservoir neurons – which are distributed on a ring topology – to the output node [54, 55]. In this CMOS compatible crossbar, multiple weight states are achieved using multiple bi-stable memristors (with only two addressable resistance states). The structural simplicity of this architecture paves the way to independent control of each synaptic element. However, the discrete nature of memristive weights places a limit on the accuracy of reservoir computing; thus the number of weight states per synapse, that are required for satisfactory accuracy, must be deter-

mined beforehand. Besides, appropriate learning algorithms for on-chip training purpose have not yet been implemented for fully analog reservoir computing. Another significant design challenge associated with memristor devices is cycle-to-cycle variability in the resistance values even within a single memristor. With the same reservoir structure, therefore, [56] showed by connecting memristor devices in series or parallel, a staircase memristor model could be constructed which not only has a delayed-switching effect between several somewhat stable resistance levels, but also can provide more reliable state values if a specific resistance level is required. This model of synaptic delay is particularly relevant for time delay reservoir methodologies [46].

From the information processing point of view, the ultimate aim of neuromorphic systems is to carry out neural computation in an energy-efficient way. However, firstly, the quantities relevant to the computation have to be expressed in terms of the spikes that spiking neurons communicate with. One of the key components in both digital and analog neuromorphic reservoir computers is, therefore, a neural encoder which transforms the input signals into the spike trains. Although the nature of the neural code (or neural codes) is an unresolved topic of research in neuroscience, based on what is known from biology, a number of neural information encodings have been proposed [57]. Among them, hardware prototypes of rate encoding, temporal encoding, inter-spike interval encoding, and latency (or rank order) encoding schemes have been implemented, mostly on digital platforms [49, 56, 58, 59]. Digital configurations, however, requires large hardware overheads associated with analog-to-digital converters, operational amplifiers, digital buffers, and electronic synchronizers and will increase the cost of implementation. Particularly, in analog reservoir computers, fabricating a fully analog encoding spike generator is of crucial importance both to speed the process up and to optimize the required energy and hardware costs. Examples of such generators have been proposed, mainly, for analog implementation of delayed feedback reservoirs [59, 60].

4 RC on Mixed Digital/Analog Neuromorphic Systems

The majority of analog neuromorphic systems designed to solve machine learning/signal processing problems tend to rely on digital components, for instance, for pre-processing, encoding spike generation, storing the synaptic weights values and on-chip programmability/learning [61–63]. Inter-chip and chip-to-chip communications are also primarily digital in some analog neuromorphic platforms [61, 64, 65]. Mixed analog/digital architectures are, therefore, very common in neuromorphic systems. In the context of hardware reservoir computing and inspired by the nonlinear properties of dendrites in biological neurons, [66] proposed a readout learning mechanism which returns binary values for synaptic weights such that the “choice” of connectivity can be implemented in a mixed analog/digital platform with the address event representation (AER) protocols where the connection matrix is stored in digital memory. Relying on the same communication protocol and exploiting the reconfigurable online learning

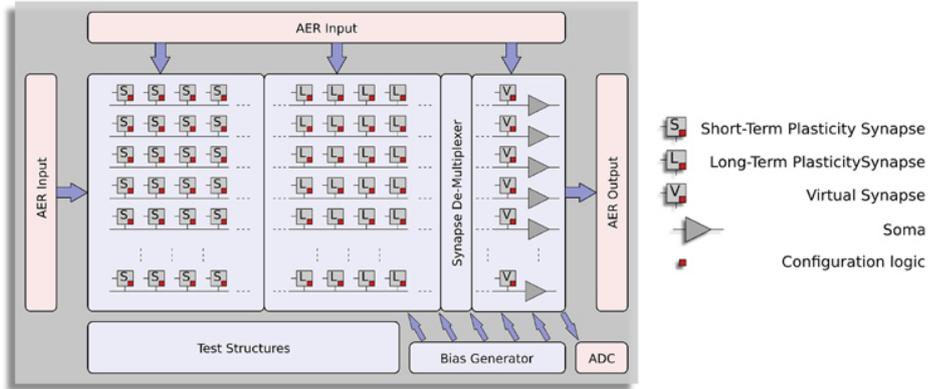


Fig. 3. Architecture of ROLLS neuromorphic processor (reproduced from [67]).

spiking neuromorphic processor (ROLLS chip) introduced in [67], [68] designed and fabricated a reservoir computer to detect spike-patterns in bio-potential and local field potential (LFP) recordings. The ROLLS neuromorphic processor (Figure 3) contains 256 adaptive exponential integrate-and-fire (spiking) neurons implemented with mixed signal analog/digital circuits. The neurons are connected to an array of 256×256 learning synapse circuits for modeling long-term plasticity mechanisms, an array of 256×256 programmable synapses with short-term plasticity circuits and 256×2 row of “virtual synapses” for modeling excitatory and inhibitory synapses that have shared synaptic weights and time constants. The reservoir in this model comprises 128 randomly connected spiking neurons that receive spike event inputs from a neural recording system and aim to enhance the temporal properties of input patterns. The readout layer is afterwards trained by applying a spike-based Hebbian-like learning rule previously proposed in [69]. The results of this study suggest that the device mismatch and the limited precision of the analog circuits result in a diversity of neuronal responses that might be beneficial in a population coding scheme within the reservoir computing framework.

Later, in another implementation of reservoir computing on mixed analog/digital substrates endowed with learning abilities, the neuronal circuits were chosen to be analog for power and area efficiency considerations and an on-chip learning mechanism based on recursive least squares (RLS) algorithm was implemented on an FPGA platform [49]. The proposed architecture was then applied to modeling a Multiple-Input Multiple-Output Orthogonal Frequency Division Multiplexing (MIMO-OFDM) system. A completely on-chip feasible design for RC has also been proposed in [70], where non-spiking reservoir neurons are distributed in a hybrid topology and memristive nano synapses were used in the output or regression layer. The proposed hybrid topology consists of a ring RC and a center neuron connected to all the reservoir neurons which transmit the

information about the state of the whole reservoir to each neuron. To fabricate non-spiking silicon neurons, current-mode differential amplifiers operating in their subthreshold regime have been used to emulate the hyperbolic tangent rate model behavior. For epileptic seizure detection and prosthetic finger control, it has been shown that a random distribution of weights in input-to-reservoir and reservoir synapses can be obtained by employing mismatches in transistors threshold voltages to design subthreshold bipolar synapses.

Recently, in order to create a functional spiking reservoir on the Dynap-se board [71], a “Reservoir Transfer Method” was proposed to transfer the functional ESN-type reservoir into a reservoir with analog, unlocked, spiking neurons [72]. The Dynap-se board is a multicore neuromorphic processor chip that employs hybrid analog/digital circuits for emulating synapse and neuron dynamics together with asynchronous digital circuits for managing the address-event traffic. It offers 4000 adaptive exponential integrate-and-fire (spiking) neurons implemented with mixed signal analog/digital circuits with $64/4k$ fan-in/out. From the computational point of view, implementing efficient algorithms on this neuromorphic system encounters general problems such as bit resolution, device mismatch, uncharacterized neural models, unavailable state variables, and physical system noise. Realizing the reservoir computing on this substrate, additionally, leads to an important problem associated with fine-tuning the reservoir to obtain the memory span required for a specific machine learning problem. The reservoir transfer method proposes a theoretical framework to learn ternary weights in a reservoir of spiking neurons by transferring features from a well-tuned echo state network simulated on software [72]. Empirical results from an ECG signal monitoring task showed that this reservoir with ternary weights is able to not only integrate information over a time span longer than the timescale of individual neurons but also function as an information processing medium with performance close to a standard, high precision, deterministic, non-spiking ESN [72].

5 Conclusion

Reservoir computing appears to be a particularly widespread and versatile approach to harness unconventional nonlinear physical phenomena into useful computations. Spike-based neuromorphic microchips, on the other hand, promise one or two orders of magnitude less energy consumption than traditional digital microchips. Implementation of reservoir computing methodologies on neuromorphic hardware, therefore, has been an attractive practice in neuromorphic system design. Here, a review of experimental studies was provided to illustrate the progress in this area and to address the computational bottlenecks which arise from specific hardware implementations. Moreover, to deal with challenges of computation on such unconventional substrates, several lines of potential solutions are presented based on advances in other computational approaches in machine learning.

Acknowledgments

This work was supported by European H2020 collaborative project NeuRAM3 [grant number 687299]. I would also like to thank Herbert Jaeger, who provided insight and expertise that greatly assisted this research.

References

1. Carver Mead and Mohammed Ismail. *Analog VLSI implementation of neural systems*, volume 80. Springer Science & Business Media, 2012.
2. Laszlo B Kish. End of moore’s law: thermal (noise) death of integration in micro and nano electronics. *Physics Letters A*, 305(3-4):144–149, 2002.
3. Ronald G Dreslinski, Michael Wieckowski, David Blaauw, Dennis Sylvester, and Trevor Mudge. Near-threshold computing: Reclaiming moore’s law through energy efficient integrated circuits. *Proceedings of the IEEE*, 98(2):253–266, 2010.
4. Catherine D Schuman, Thomas E Potok, Robert M Patton, J Douglas Birdwell, Mark E Dean, Garrett S Rose, and James S Plank. A survey of neuromorphic computing and neural networks in hardware. *arXiv preprint arXiv:1705.06963*, 2017.
5. J Joshua Yang, Dmitri B Strukov, and Duncan R Stewart. Memristive devices for computing. *Nature nanotechnology*, 8(1):13, 2013.
6. Giacomo Indiveri, Bernabé Linares-Barranco, Robert Legenstein, George Deligeorgis, and Themistoklis Prodromakis. Integration of nanoscale memristor synapses in neuromorphic computing architectures. *Nanotechnology*, 24(38):384010, 2013.
7. Gina C Adam, Brian D Hoskins, Mirko Prezioso, Farnood Merrih-Bayat, Bhaswar Chakrabarti, and Dmitri B Strukov. 3-d memristor crossbars for analog and neuromorphic computing applications. *IEEE Transactions on Electron Devices*, 64(1):312–318, 2017.
8. Sung Hyun Jo, Ting Chang, Idongesit Ebong, Bhavitavya B Bhadviya, Pinaki Mazumder, and Wei Lu. Nanoscale memristor device as synapse in neuromorphic systems. *Nano letters*, 10(4):1297–1301, 2010.
9. M Suri, O Bichler, D Querlioz, G Palma, E Vianello, D Vuillaume, C Gamrat, and B DeSalvo. Cbram devices as binary synapses for low-power stochastic neuromorphic systems: auditory (cochlea) and visual (retina) cognitive processing applications. In *2012 International Electron Devices Meeting*, pages 10–3. IEEE, 2012.
10. Masakazu Aono and Tsuyoshi Hasegawa. The atomic switch. *Proceedings of the IEEE*, 98(12):2228–2236, 2010.
11. Audrius V Avizienis, Henry O Sillin, Cristina Martin-Olmos, Hsien Hang Shieh, Masakazu Aono, Adam Z Stieg, and James K Gimzewski. Neuromorphic atomic switch networks. *PloS one*, 7(8):e42772, 2012.
12. Henry O Sillin, Renato Aguilera, Hsien-Hang Shieh, Audrius V Avizienis, Masakazu Aono, Adam Z Stieg, and James K Gimzewski. A theoretical and experimental study of neuromorphic atomic switch networks for reservoir computing. *Nanotechnology*, 24(38):384004, 2013.
13. Manan Suri, Olivier Bichler, Damien Querlioz, Boubacar Traoré, Olga Cueto, Luca Perniola, Veronique Sousa, Dominique Vuillaume, Christian Gamrat, and Barbara DeSalvo. Physical aspects of low power synapses based on phase change memory devices. *Journal of Applied Physics*, 112(5):054904, 2012.

14. Stefano Ambrogio, Nicola Ciocchini, Mario Laudato, Valerio Milo, Agostino Pirovano, Paolo Fantini, and Daniele Ielmini. Unsupervised learning by spike timing dependent plasticity in phase change memory (pcm) synapses. *Frontiers in neuroscience*, 10:56, 2016.
15. Shimeng Yu and H-S Philip Wong. Modeling the switching dynamics of programmable-metallization-cell (pmc) memory and its application as synapse device for a neuromorphic computation system. In *2010 International Electron Devices Meeting*, pages 22–1. IEEE, 2010.
16. Shimeng Yu, Yi Wu, Rakesh Jeyasingh, Duygu Kuzum, and H-S Philip Wong. An electronic synapse device based on metal oxide resistive switching memory for neuromorphic computation. *IEEE Transactions on Electron Devices*, 58(8):2729–2737, 2011.
17. Gouhei Tanaka, Toshiyuki Yamane, Jean Benoit Héroux, Ryosho Nakane, Naoki Kanazawa, Seiji Takeda, Hidetoshi Numata, Daiju Nakano, and Akira Hirose. Recent advances in physical reservoir computing: a review. *Neural Networks*, 2019.
18. Giacomo Indiveri and Shih-Chii Liu. Memory and information processing in neuromorphic systems. *Proceedings of the IEEE*, 103(8):1379–1397, 2015.
19. Paul A Merolla, John V Arthur, Rodrigo Alvarez-Icaza, Andrew S Cassidy, Jun Sawada, Filipp Akopyan, Bryan L Jackson, Nabil Imam, Chen Guo, Yutaka Nakamura, et al. A million spiking-neuron integrated circuit with a scalable communication network and interface. *Science*, 345(6197):668–673, 2014.
20. Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *science*, 304(5667):78–80, 2004.
21. Conrad D James, James B Aimone, Nadine E Miner, Craig M Vineyard, Fredrick H Rothganger, Kristofor D Carlson, Samuel A Mulder, Timothy J Draelos, Aleksandra Faust, Matthew J Marinella, et al. A historical survey of algorithms and hardware architectures for neural-inspired and neuromorphic computing applications. *Biologically Inspired Cognitive Architectures*, 19:49–64, 2017.
22. Janardan Misra and Indranil Saha. Artificial neural networks in hardware: A survey of two decades of progress. *Neurocomputing*, 74(1-3):239–255, 2010.
23. Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural computation*, 14(11):2531–2560, 2002.
24. Qian Wang, Youjie Li, Botang Shao, Siddhartha Dey, and Peng Li. Energy efficient parallel neuromorphic architectures with approximate arithmetic on fpga. *Neurocomputing*, 221:146–158, 2017.
25. Benjamin Schrauwen, Michiel D’Haene, David Verstraeten, and Jan Van Campenhout. Compact hardware for real-time speech recognition using a liquid state machine. In *2007 International Joint Conference on Neural Networks*, pages 1097–1102. IEEE, 2007.
26. Andres Upegui, Carlos Andrés Pena-Reyes, and Eduardo Sanchez. An fpga platform for on-line topology exploration of spiking neural networks. *Microprocessors and microsystems*, 29(5):211–223, 2005.
27. Benjamin Schrauwen and Jan Van Campenhout. Parallel hardware implementation of a broad class of spiking neurons using serial arithmetic. In *Proceedings of the 14th European Symposium on Artificial Neural Networks*, pages 623–628. d-side publications:, 2006.
28. David Verstraeten, Benjamin Schrauwen, and Dirk Stroobandt. Reservoir computing with stochastic bitstream neurons. In *Proceedings of the 16th annual Prorisc workshop*, pages 454–459, 2005.

29. Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.
30. Fatemeh Hadaeghi and Herbert Jaeger. Computing optimal discrete readout weights in reservoir computing is np-hard. *Neurocomputing*, 338:233–236, 2019.
31. Yong Zhang, Peng Li, Yingyezhe Jin, and Yoonsuck Choe. A digital liquid state machine with biologically inspired learning and its application to speech recognition. *IEEE transactions on neural networks and learning systems*, 26(11):2635–2649, 2015.
32. Qian Wang, Yingyezhe Jin, and Peng Li. General-purpose lsm learning processor architecture and theoretically guided design space exploration. In *2015 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pages 1–4. IEEE, 2015.
33. Yingyezhe Jin and Peng Li. Performance and robustness of bio-inspired digital liquid state machines: A case study of speech recognition. *Neurocomputing*, 226:145–160, 2017.
34. Yingyezhe Jin, Yu Liu, and Peng Li. Sso-lsm: A sparse and self-organizing architecture for liquid state machine based neural processors. In *2016 IEEE/ACM International Symposium on Nanoscale Architectures (NANOARCH)*, pages 55–60. IEEE, 2016.
35. Qian Wang, Youjie Li, and Peng Li. Liquid state machine based pattern recognition on fpga with firing-activity dependent power gating and approximate computing. In *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*, pages 361–364. IEEE, 2016.
36. Botang Shao and Peng Li. Array-based approximate arithmetic computing: A general model and applications to multiplier and squarer design. *IEEE Transactions on Circuits and Systems I: Regular Papers*, 62(4):1081–1090, 2015.
37. Anvesh Polepalli, Nicholas Soures, and Dhireesha Kudithipudi. Digital neuromorphic design of a liquid state machine for real-time processing. In *2016 IEEE International Conference on Rebooting Computing (ICRC)*, pages 1–8. IEEE, 2016.
38. Anvesh Polepalli, Nicholas Soures, and Dhireesha Kudithipudi. Reconfigurable digital design of a liquid state machine for spatio-temporal data. In *Proceedings of the 3rd ACM International Conference on Nanoscale Computing and Communication*, page 15. ACM, 2016.
39. Michael R Smith, Aaron J Hill, Kristofor D Carlson, Craig M Vineyard, Jonathon Donaldson, David R Follett, Pamela L Follett, John H Naegle, Conrad D James, and James B Aimone. A novel digital neuromorphic architecture efficiently facilitating complex synaptic response functions applied to liquid state machines. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2421–2428. IEEE, 2017.
40. Yingyezhe Jin and Peng Li. Ap-stdp: A novel self-organizing mechanism for efficient reservoir computing. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1158–1165. IEEE, 2016.
41. Subhrajit Roy and Arindam Basu. An online structural plasticity rule for generating better reservoirs. *Neural computation*, 28(11):2557–2584, 2016.
42. Yu Liu, Yingyezhe Jin, and Peng Li. Online adaptation and energy minimization for hardware recurrent spiking neural networks. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, 14(1):11, 2018.
43. Nicholas Soures, Lydia Hays, and Dhireesha Kudithipudi. Robustness of a memristor based liquid state machine. In *2017 international joint conference on neural networks (ijcnn)*, pages 2414–2420. IEEE, 2017.

44. Amr M Hassan, Hai Helen Li, and Yiran Chen. Hardware implementation of echo state networks using memristor double crossbar arrays. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 2171–2177. IEEE, 2017.
45. Ali Rodan and Peter Tino. Minimum complexity echo state network. *IEEE transactions on neural networks*, 22(1):131–144, 2011.
46. Lennert Appeltant, Miguel Cornelles Soriano, Guy Van der Sande, Jan Danckaert, Serge Massar, Joni Dambre, Benjamin Schrauwen, Claudio R Mirasso, and Ingo Fischer. Information processing using a single dynamical node as complex system. *Nature communications*, 2:468, 2011.
47. Miquel L Alomar, Vincent Canals, Antoni Morro, Antoni Oliver, and Josep L Rossello. Stochastic hardware implementation of liquid state machines. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 1128–1133. IEEE, 2016.
48. Piotr Antonik, Anteo Smerieri, François Duport, Marc Haelterman, and Serge Massar. Fpga implementation of reservoir computing with online learning. In *24th Belgian-Dutch Conference on Machine Learning*, 2015.
49. Yang Yi, Yongbo Liao, Bin Wang, Xin Fu, Fangyang Shen, Hongyan Hou, and Lingjia Liu. Fpga based spike-time dependent encoder and reservoir design in neuromorphic computing processors. *Microprocessors and Microsystems*, 46:175–183, 2016.
50. Miquel L Alomar, Vincent Canals, Nicolas Perez-Mora, Víctor Martínez-Moll, and Josep L Rosselló. Fpga-based stochastic echo state networks for time-series forecasting. *Computational intelligence and neuroscience*, 2016:15, 2016.
51. Giacomo Indiveri, Bernabé Linares-Barranco, Tara Julia Hamilton, André Van Schaik, Ralph Etienne-Cummings, Tobi Delbruck, Shih-Chii Liu, Piotr Dudek, Philipp Häfliger, Sylvie Renaud, et al. Neuromorphic silicon neuron circuits. *Frontiers in neuroscience*, 5:73, 2011.
52. Felix Schürmann, Karlheinz Meier, and Johannes Schemmel. Edge of chaos computation in mixed-mode vlsi-a hard liquid. In *Advances in neural information processing systems*, pages 1201–1208, 2005.
53. Shih-Chii Liu and Tobi Delbruck. Neuromorphic sensory systems. *Current opinion in neurobiology*, 20(3):288–295, 2010.
54. Colin Donahue, Cory Merkel, Qutaiba Saleh, Levs Dolgovs, Yu Kee Ooi, Dhireesha Kudithipudi, and Bryant Wysocki. Design and analysis of neuromemristive echo state networks with limited-precision synapses. In *2015 IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, pages 1–6. IEEE, 2015.
55. Cory Merkel, Qutaiba Saleh, Colin Donahue, and Dhireesha Kudithipudi. Memristive reservoir computing architecture for epileptic seizure detection. *Procedia Computer Science*, 41:249–254, 2014.
56. Xiao Yang, Wanlong Chen, and Frank Z Wang. Investigations of the staircase memristor model and applications of memristor-based local connections. *Analog Integrated Circuits and Signal Processing*, 87(2):263–273, 2016.
57. André Grüning and Sander M Bohte. Spiking neural networks: Principles and challenges. In *ESANN*, 2014.
58. Olivier Bichler, Damien Querlioz, Simon J Thorpe, Jean-Philippe Bourgoin, and Christian Gamrat. Extraction of temporally correlated features from dynamic vision sensors with spike-timing-dependent plasticity. *Neural Networks*, 32:339–348, 2012.

59. Chenyuan Zhao, Bryant T Wysocki, Clare D Thiem, Nathan R McDonald, Jialing Li, Lingjia Liu, and Yang Yi. Energy efficient spiking temporal encoder design for neuromorphic computing systems. *IEEE Transactions on Multi-Scale Computing Systems*, 2(4):265–276, 2016.
60. Jialing Li, Chenyuan Zhao, Kian Hamedani, and Yang Yi. Analog hardware implementation of spike-based delayed feedback reservoir computing system. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 3439–3446. IEEE, 2017.
61. Alan F Murray and Anthony VW Smith. Asynchronous vlsi neural networks using pulse-stream arithmetic. *IEEE Journal of Solid-State Circuits*, 23(3):688–697, 1988.
62. Mostafa Rahimi Azghadi, Saber Moradi, and Giacomo Indiveri. Programmable neuromorphic circuits for spike-based neural dynamics. In *2013 IEEE 11th International New Circuits and Systems Conference (NEWCAS)*, pages 1–4. IEEE, 2013.
63. Daniel Briiderle, Johannes Bill, Bernhard Kaplan, Jens Kremkow, Karlheinz Meier, Eric Müller, and Johannes Schemmel. Simulator-like exploration of cortical network architectures with a mixed-signal vlsi system. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pages 2784–8787. IEEE, 2010.
64. Syed Ahmed Aamir, Paul Müller, Andreas Hartel, Johannes Schemmel, and Karlheinz Meier. A highly tunable 65-nm cmos lif neuron for a large scale neuromorphic system. In *ESSCIRC Conference 2016: 42nd European Solid-State Circuits Conference*, pages 71–74. IEEE, 2016.
65. Elisabetta Chicca, Patrick Lichtsteiner, Tobias Delbruck, Giacomo Indiveri, and Rodney J Douglas. Modeling orientation selectivity using a neuromorphic multi-chip system. In *2006 IEEE International Symposium on Circuits and Systems*, pages 4–pp. IEEE, 2006.
66. Subhrajit Roy, Amitava Banerjee, and Arindam Basu. Liquid state machine with dendritically enhanced readout for low-power, neuromorphic vlsi implementations. *IEEE transactions on biomedical circuits and systems*, 8(5):681–695, 2014.
67. Ning Qiao, Hesham Mostafa, Federico Corradi, Marc Osswald, Fabio Stefanini, Dora Sumislawska, and Giacomo Indiveri. A reconfigurable on-line learning spiking neuromorphic processor comprising 256 neurons and 128k synapses. *Frontiers in neuroscience*, 9:141, 2015.
68. Federico Corradi and Giacomo Indiveri. A neuromorphic event-based neural recording system for smart brain-machine-interfaces. *IEEE transactions on biomedical circuits and systems*, 9(5):699–709, 2015.
69. Joseph M Brader, Walter Senn, and Stefano Fusi. Learning real-world stimuli in a neural network with spike-driven synaptic dynamics. *Neural computation*, 19(11):2881–2912, 2007.
70. Dhireesha Kudithipudi, Qutaiba Saleh, Cory Merkel, James Thesing, and Bryant Wysocki. Design and analysis of a neuromemristive reservoir computing architecture for biosignal processing. *Frontiers in neuroscience*, 9:502, 2016.
71. Saber Moradi, Ning Qiao, Fabio Stefanini, and Giacomo Indiveri. A scalable multi-core architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (dynaps). *IEEE transactions on biomedical circuits and systems*, 12(1):106–122, 2018.
72. Xu He, Tianlin Liu, Fatemeh Hadaeghi, and Herbert Jaeger. Reservoir transfer on analog neuromorphic hardware. In *9th International IEEE EMBS Neural Engineering Conference*, 2019.