Neural Networks 45 (2013) 4–26

Contents lists available at SciVerse ScienceDirect

Neural Networks

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

2013 Special Issue

Design of silicon brains in the nano-CMOS era: Spiking neurons, learning synapses and neural architecture optimization

Andrew S. Cassidy[a,1], Julius Georgiou[b], Andreas G. Andreou[a,b,*]

[a] Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA
[b] Department of Electrical and Computer Engineering, University of Cyprus, Nicosia 1678, Cyprus

ARTICLE INFO

Keywords:
Silicon brains
Neuromorphic engineering
Silicon neurons
Learning in silicon
FPGA neural arrays

ABSTRACT

We present a design framework for neuromorphic architectures in the nano-CMOS era. Our approach to the design of spiking neurons and STDP learning circuits relies on parallel computational structures where neurons are abstracted as digital arithmetic logic units and communication processors. Using this approach, we have developed arrays of silicon neurons that scale to millions of neurons in a single state-of-the-art Field Programmable Gate Array (FPGA). We demonstrate the validity of the design methodology through the implementation of cortical development in a circuit of spiking neurons, STDP synapses, and neural architecture optimization.

1. The computer and the brain

The brain is a massively parallel and efficient information processing system, with a radically different computational architecture from present day computers. Characteristics of neural computation include event based processing, fine-grained parallel computational units, robustness and redundancy, as well as adaptation and learning, all done under severe constraints of size, weight, and energy resources. This computational architecture excels at lower-level sensory information processing such as vision, and sensor–motor integration as well as cognitive tasks such as speech and language understanding.

Over the last half century computer scientists, architects and engineers have envisioned building computers that match the parallel processing capabilities of biological brains. Fifty years ago, the fathers of computer science Alan Turing (Turing, 1952) and John von-Neumann (Neumann, 1958) looked to the brain for inspiration in order to advance the science of computing.

Twenty-five years ago, the connectionist movement emerged as an alternative approach to artificial intelligence for solving the hard problems in perception and cognition. The central doctrine in the connectionist movement is that the cognitive abilities of the brain are a result of a highly interconnected network of simple processing units. These simple non-linear computational units abstract the function of neurons while synapses abstract the connections between neurons. The strength of the synaptic connections in networks of such units is determined through a learning algorithm. A two volume edited book-set by the “Parallel Distributed Research Group” (McClelland, Rumelhardt, & Group, 1987; Rumelhart, McClelland, & Group, 1987) defined the research agenda in the field of connectionist architectures and neural networks in the decades that followed. At about the same time, Carver Mead’s book “Analog VLSI and Neural Systems” (Mead, 1989) inspired a new generation of scientists and engineers to explore hardware implementation of neural models in state-of-the-art silicon integrated circuit technology. The book had a dual objective: (i) to create a new design discipline for collective computational systems using analog VLSI subthreshold CMOS integrated circuit technology and (ii) to promote a synthetic approach in the understanding of biology and the human brain. This was the birth of neuromorphic design as an engineering discipline.

1.1. *Neuromorphic engineering: the formative years*

“Neuromorphic” electronic systems, a term coined by Carver Mead in the late 1980s, describes systems that perform artificial computation based on the principles of neurobiological circuits. In the following two decades, inspired by Mead’s pioneering work (Mead, 1990) and colleagues at Caltech, a large number of CMOS neuromorphic chip designs have been reported in the literature.

These spanned a wide range of designs from analog VLSI models of neurons (Arthur & Boahen, 2010; Hsin, Saighi, Buhry, & Renaud, 2010; Saighi, Bornat, Tomas, Le Masson, & Renaud, 2010; Yu, Sejnowski, & Cauwenberghs, 2011) to silicon retina architectures

[*] Corresponding author at: Department of Electrical and Computer Engineering, Johns Hopkins University, Baltimore, MD 21218, USA.
E-mail addresses: andrewca@us.ibm.com (A.S. Cassidy), julio@ucy.ac.cy (J. Georgiou), andreou@jhu.edu (A.G. Andreou).
[1] Now with IBM Research, Almaden, USA.

0893-6080/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.neunet.2013.05.011

(Boahen & Andreou, 1992; Mahowald, 1992), and retinomorphic vision systems (Boahen, 1996), to attention circuits (Horiuchi & Koch, 1999), and biomorphic imagers (Culurciello, Etienne-Cummings, & Boahen, 2003) that abstract biology at a lower level. Other mixed-mode designs (Andreou, Meitzler, Strohben, & Boahen, 1995; Pardo, Dierickx, & Scheffer, 1998) and (Etienne-Cummings, Kalayjian, & Donghui, 2001) have also implemented silicon retinas and focal plane processing architectures that include processing beyond gain control and spatio-temporal filtering, including polarization sensing (Andreou & Kalayjian, 2002; Wolff & Andreou, 1995). Most of the above bio-inspired sensors have limited programmability as they employ analog computational circuits at the focal plane.

The shortcomings of non-programmable analog architectures motivated the exploration of analog vision chip architectures with programmable functionality (Serrano-Gotarredona, Andreou, & Linares-Barranco, 1999; Serrano-Gotarredona et al., 2009). Programmable architectures for associative memory (Boahen, Pouliquen, Andreou, & Jenkins, 1989; Pouliquen, Andreou, & Strohben, 1997), pattern classification (Genov & Cauwenberghs, 2001; Karakiewicz, Genov, & Cauwenberghs, 2007) and audition (Kumar, Himmelbauer, Cauwenberghs, & Andreou, 1998; Stanacevic & Cauwenberghs, 2005) have also been reported in the literature.

Programmable architectures have also been advanced by the adoption of a standard interface between chips known as Address Event Representation or (AER) in short. The time-multiplexed AER bus (Boahen, 2000; Lin & Boahen, 2009; Mahowald, 1992; Sivilotti, 1991) is a popular interconnect method for neuromorphic systems. Spike events from multiple channels are time-multiplexed onto a digital AER bus, transmitted, and decoded at the destination onto individual channels. Throughout this proposal, we use the terms spikes, events, and spike events interchangeably. AER has been used by many analog and digital spiking neural arrays, as well as to communicate events from off-chip neuromorphic sensors and even in 3D CMOS technology (Harrison, Özgün, Lin, Andreou, & Etienne-Cummings, 2010). The European Union project CAVIAR (http://www2.imse-cnm.csic.es/caviar/) demonstrated a board-level vision system architecture communicating using the AER protocol (Serrano-Gotarredona et al., 2009). Variants of AER to improve the efficiency of the protocol have also been proposed (Georgiou & Andreou, 2006, 2007). A probabilistic approach to AER has been exploited to perform computations in the address domain (Goldberg, Cauwenberghs, & Andreou, 2001b).

Learning in silicon has also been pursued intensively in the analog VLSI neuromorphic community. The early work by Diorio and colleagues (Diorio, Hasler, Minch, & Mead, 1996, 1997), the Field Programmable Analog Arrays (Sivilotti, 1991) and the research program of Hasler (Hall, Twigg, Gray, Hasler, & Anderson, 2005) paved the way to floating gate MOS transistors in configurable learning chips. Other designs employ dynamic circuits for implementing learning in analog VLSI with excellent results on small systems (Bartolozzi & Indiveri, 2007; Indiveri, Chicca, & Douglas, 2004; Mahowald, 1992). This work has continued with encouraging results for hardware models that abstract higher-level functions such as stimulus specific adaptation (Mill, Sheik, Indiveri, & Denham, 2011) and working memory using attractor dynamics (Giulioni et al., 2011).

Abstracting biology at a higher level, the Cellular Non-linear/Neural Networks (CNN) approach (Chua & Yang, 1988) offered another paradigm for an analog visual processor with programming capabilities. In CNN architectures, information processing is implemented through the evolution of a continuous-time non-linear dynamical network with nearest neighborhood connectivity. The CNN–UM (Universal Machine) is one of the earliest systems (Roska & Chua, 1993) that implemented CNN programmable functionality on a chip. Another example of CNN hardware implementation merges a CNN–UM type processor and an imager (Carmona et al., 1998; Dominguez-Castro et al., 1997). This system, while analog internally, has a digital interface with on-chip 7-bit A/D and D/A converters, improving the programmability and simplifying the interface to digital computers (Cembrano et al., 2004).

Programmable analog VLSI circuits and systems aimed at large-scale model simulation have also been under development in the last decade. The Neurogrid architecture in Kwabena's group (Arthur & Boahen, 2010; Choudhary et al., 2012; Silver, Boahen, Grillner, Kopell, & Olsen, 2007), the IFAT architecture (Goldberg, Cauwenberghs, & Andreou, 2001a; Vogelstein, Mallik, Culurciello, Cauwenberghs, & Etienne-Cummings, 2007), the PAX platform (Renaud et al., 2010) and the FACETS wafer-scale computational infrastructure (Bruederle et al., 2011) are notable projects in this direction.

1.2. Neuromorphic engineering: the nano-CMOS Era

In 1986, Mead's group at Caltech was employing bulk CMOS technology with λ between 2.5 micron and 0.7 micron (p. 59 of Mead, 1989). A quick review of our own publications and laboratory notebooks from that period, reveals that we were fabricating chips in 4 micron Silicon On Sapphire (SOS)–CMOS technology and in 3 micron *p*-well bulk CMOS. Alas! Twenty five years later, with foundry CMOS technologies at the 45 nm and 22 nm nodes, the neuromorphic engineering community at large has not been able to capitalize on the benefits of the ($\times 10\,000$) improvements in digital MOS transistor area density to engineer brain like structures and cognitive machines that match the effectiveness and energetic efficiency of the human brain. With the exception of the event-based, asynchronous vision sensors (Lichtsteiner, Posch, & Delbruck, 2008) and subsequent design (Posch, Matolin, & Wohlgenannt, 2011), the goals of endowing modern computer systems with industrial-strength robust bio-inspired sensoria or tackling the challenge of silicon cognition have been unrealized. And even though our lack of knowledge about the inner workings of brain function and behavior has contributed to this chasm and is limiting us today, matching the information processing capabilities of biological neural structures in state-of-the-art silicon technology has remained an elusive dream despite the stunning advances in microelectronics.

Even more elusive has been our quest to understand how to achieve the energy efficiency seen in biological brains. One would have thought that the research activities in the last two decades would have brought us closer to both a deeper understanding of brain function as well as to commercially-viable brain-inspired information technology at the scale. However, this is not the case. Many of the analog VLSI neuromorphic systems rely on analog devices and as such, scaling the density of these components (mostly MOS transistors and capacitors) did not follow Moore's law. Furthermore, the majority of neuromorphic hardware was based on traditional "analog" circuit models of neurons and synapses, a technology that does not offer flexibility in component models, nor in their level of description; an aspect which impedes rapid advances.

Mead advocated using analog transistor physics to perform neural computation, directly mimicking the currents in neuron ion channels (Mead, 1990), and speculated that an energy savings of approximately 10^4 could be gained over comparable traditional digital approaches. However the power dissipation of neuromorphic systems did not benefit from technology scaling either and our best circuits today hover between 10 and 100 nW per computational cell. Each cell has typically one or two single pole circuits with two or three current branches biased in the nano-ampere current level. Even though one could argue the power dissipation is manageable locally, the energy cost to send the