Large-Scale Neuromorphic Computing Systems

DOI:
10.1088/1741-2560/13/5/051001

Document Version
Final published version

TOPICAL REVIEW • OPEN ACCESS

Large-scale neuromorphic computing systems

To cite this article: Steve Furber 2016 *J. Neural Eng.* **13** 051001

OPEN ACCESS
IOP Publishing
Journal of Neural Engineering
J. Neural Eng. 13 (2016) 051001 (14pp)
doi:10.1088/1741-2560/13/5/051001

Topical Review

Large-scale neuromorphic computing systems

Steve Furber

School of Computer Science, The University of Manchester Oxford Road, Manchester M13 9PL UK

E-mail: steve.furber@manchester.ac.uk

Received 12 May 2016, revised 1 July 2016
Accepted for publication 28 July 2016
Published 16 August 2016

Abstract

Neuromorphic computing covers a diverse range of approaches to information processing all of which demonstrate some degree of neurobiological inspiration that differentiates them from mainstream conventional computing systems. The philosophy behind neuromorphic computing has its origins in the seminal work carried out by Carver Mead at Caltech in the late 1980s. This early work influenced others to carry developments forward, and advances in VLSI technology supported steady growth in the scale and capability of neuromorphic devices. Recently, a number of large-scale neuromorphic projects have emerged, taking the approach to unprecedented scales and capabilities. These large-scale projects are associated with major new funding initiatives for brain-related research, creating a sense that the time and circumstances are right for progress in our understanding of information processing in the brain. In this review we present a brief history of neuromorphic engineering then focus on some of the principal current large-scale projects, their main features, how their approaches are complementary and distinct, their advantages and drawbacks, and highlight the sorts of capabilities that each can deliver to neural modellers.

Keywords: neuromorphic systems, brain-inspired computing, large-scale neuromorphics

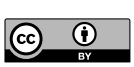
(Some figures may appear in colour only in the online journal)

1. Introduction

The brain is an enigma, and remains as one of the great frontiers of science. The 1.4 kg of fatty material that we each carry around in our heads defines our personalities, stores our memories, and controls all aspects of our behaviour. Yet, despite huge progress in neuroscience over the last century, the fundamental principles of information processing and storage in the brain are far from understood.

With the widespread introduction over the last half century of the stored-programme computer into all areas of human activity we have an alternative information processing paradigm against which the brain may be compared, yet despite superficial similarities in their roles as information processing and control systems, it is clear that the principles of operation of the brain and the computer are very different. While computers excel at speed and precision, the brain still wins in coping with novelty, complexity and ambiguity, and in practical tasks such as facial recognition and controlling bipedal locomotion.

This difference in capabilities has encouraged computer engineers from the earliest days of computing to wonder whether there might not be different principles at work in the brain that could profitably be applied in the design of machines. Where these principles are derived from our (partial) understanding of the structure and characteristics of neurons—the basic components from which the brain is constructed—and the complex networks of neurons in the brain, the resulting systems are categorized as neuromorphic computers. This terminology has come to be applied particularly where the medium in which the system is built is



based upon very large-scale integrated (VLSI) circuits—microchips.

Advances in microchip technology have followed Moore's Law [1], which predicted exponential growth in the number of transistors that could be manufactured on a single microchip. The exponential time constant is short—with a doubling every18 months to two years—and this has been sustained for half a century. The primary mechanism for delivering Moore's Law has been reducing transistor size, and as CMOS transistors are made smaller they become cheaper, faster and more energy-efficient, a win–win scenario that has led to the ubiquity of cheap, powerful and mobile computer technology. The only downside is that the costs of designing a microchip and building a facility to manufacture it have also risen exponentially, effectively restricting the growth rate to ensure that each technology generation generates enough revenue to pay for its investment before the next generation renders it obsolete.

These advances in microchip technology, funded from conventional computing applications, are also available to neuromorphic microchip designers. This has led to growth in the scale and capabilities of neuromorphic devices which, although still a long way short of the scales employed by biology, has led recently to the development of some very large scale neuromorphic systems, which are the focus of this review.

The main contributions of this paper are:

- A survey of the features of the brain that are modelled in neuromorphic systems (section 2);
- a brief history of the early development of neuromorphic technology (section 3) and more recent advances (section 4);
- descriptions of the main features of current large-scale neuromorphic systems (section 5), including the IBM TrueNorth chip (section 6), the Stanford Neurogrid (section 7), the Heidelberg BrainScaleS machine (section 8) and the Manchester Spiking Neural Network Architecture (SpiNNaker) machine (section 9);
- a discussion of the relative strengths and weaknesses of each of these large-scale systems (section 10).

2. The brain

The singular characteristic of neuromorphic computing is that it takes inspiration from what is known about the structure and operation of the brain. We are still a long way from having a full understanding of how the brain represents, stores and processes information, though a great deal is known about neurons and how they are organised into diverse structures and topologies to perform different functions in different regions of the brain.

Neurons have been studied in great detail since the pioneering work of Hodgkin and Huxley [2] on the squid giant axon yielded the first detailed mathematical model of the mechanisms that are responsible for the generation of the action potential, or 'spike', that characterizes the primary high-speed communication from one neuron to the next. The full biological details of the neuron are dauntingly complex [3], but widely accepted models abstract away most of the biological detail to yield a multiple-input single-output device where a range of different mathematical formulations can be used to describe the input-out transfer relationship.

Much of the 'action' in the brain is not in the neuron cell itself but in the synapse—the junction where the output signal from one neuron is coupled into the input to the next neuron. Synapses are also dauntingly complex in their full biological detail [3], but again their characteristics have been abstracted into various mathematical formulations that capture the function of the synapse with varying degrees of biological fidelity. These models attempt to capture synaptic plasticity, wherein the efficacy of the synapse is adjusted to allow the network to learn the statistics of inputs, and in some case structural plasticity, where neurons rewire to form new synapses to store more permanent memories.

The scale of the mammalian brain is immense, with each human brain comprising some 85 billion neurons connected through a quadrillion (10^{15}) synapses. However, some other species have much simpler brains. *C. elegans* has 302 neurons, drosophila has around 100 000; examples can be found of the neuron playing a useful role in survival in nervous systems at all scales up to human and beyond.

A feature of the biological brain is that while communication uses digital techniques for fast signalling over all but the shortest distances, all of the processing in neurons and synapses uses much more efficient analogue chemical techniques [3]. This mixed-signal approach contrasts with the all-digital designs of current general-purpose computers where digital logic delivers the noise-immunity and deterministic behaviour that is expected of the universal Turing machine. Biology sacrifices determinism for efficiency—an approach that may be of interest to future computer engineers designing systems, such as robot vision systems, where absolute accuracy is in any case unachievable, and energy-efficiency is a primary requirement.

The goal of neuromorphic computing is to observe the formidable complexity of the biological brain and to somehow extract from what is known about its structure and principles of operation some more abstract principles that can be applied in a practical engineered system. No neuromorphic system attempts to reproduce all of the biological detail, but all adhere to the idea that computation is highly distributed across small computing elements analogous in some way to neurons, connected into networks, with some degree of flexibility in the way connections are formed. That much is common; the details vary greatly.

3. Neuromorphic origins

During the 1980s Carver Mead led a number of developments in bio-inspired microelectronics, culminating in the publication of his book entitled 'Analog VLSI and Neural Systems'

by Addison Wesley in 1989 [4]. He founded companies such as Synaptics Inc. (established in 1986), who established a very successful business developing analogue circuits based on neural networks for laptop touch pads, and he advised Misha Mahowald in her prize-winning PhD thesis in which she described the development of a silicon retina [5, 6]. An international prize for neuromorphic engineering has recently been established in memory of Misha Mahowald. Among Mead's more recent PhD students is Kwabena Boahen, who led the development of the Stanford Neurogrid, about which more later.

Mead's approach to neuromorphic engineering was grounded in the analogy between the physics that determines the behaviour of transistors operating in the sub-threshold region (which digital designers simply describe as 'off'!) and the physics at work in biological neurons.

By the end of the 1980s there were a number of related activities going on in different parts of the world, and a workshop in May 1989 on 'Analog Integrated Neural Systems' was held in connection with the International Symposium on Circuits and Systems in Portland, Oregon, focusing on working chips in this area. The diversity of approaches represented in the published proceedings [7] is impressive, as is the list of authors of the papers, many of whom are still leading figures in neuromorphic engineering and/or related areas of research a quarter of a century later!

4. Ongoing neuromorphic development

One of the world's major centres for neuromorphic engineering is the Institute for Neuroinformatics (INI) which was established at the University of Zurich and ETH Zurich in 1995 by Rodney Douglas and Kevan Martin. Several notable contributors to neuromorphic development are based at INI, including Tobi Delbruck, Shih-Chii Liu and Giacomo Indiveri, and there is a strong sense of continuity from Mead's work at Caltech through to current work at INI.

Recent work at INI includes the development of neuromorphic vision sensors [8], silicon cochlea [9], and medium-scale neuromorphic processors such as the Reconfigurable On-Line Learning Spiking (ROLLS) and cxQuad chips. These chips use sub-threshold analogue circuits and have been used to demonstrate spiking deep neural networks. A circuit board with nine cxQuad chips and one ROLLS chip has been demonstrated [10]. The cxQuad chips are used to implement a hierarchical convolutional network; each chip incorporates 1024 neurons and 65 536 digital synapses. The ROLLS chip implements the classification layer for the deep network and incorporates 256 neurons and 128 k analogue synapses. The system demonstrates low latency and very low power consumption compared with a standard deep network running on a large digital cluster machine.

Research at UCSD led by Gert Cauwenberghs (also a Caltech alumnus) includes the development of a 65 536-neuron two-compartment integrate-and-fire transceiver modules that implement spike-driven continuous time analogue membrane dynamics interconnected through a Hierarchical Address Event Representation (HiAER) communications fabric [11, 12].

In Sylvie Renaud's lab in Bordeaux, France, analogue neuromorphic circuits have been interfaced to biological networks and digital neuron models to investigate thalamic mechanisms that gate sensory signals on their route to the cortex [13]. Network research has included developing a system based upon biologically-realistic analogue neuron circuits with connectivity and plasticity implemented in a digital system connected in a real-time closed loop [14]. Both of these exemplify the use of small-scale neuromorphic techniques to investigate biological phenomena, and demonstrate a very close match between the models and those phenomena.

Recently there has been growing interest in novel devices such as memristors and phase-change memory and their intrinsic similarities to biological synapses [15, 16]. Although these devices do not feature in current large-scale projects, they may well play a major role in future neuromorphic systems.

The examples cited above, and others not mentioned here, show the widespread and diverse interest in using neuromorphic techniques to model biological processes. This is a two-way process whereby engineers learn principles from neuroscience to incorporate into their models, and in return neuroscientists learn from the results those engineers get from their models. In some cases the goal is very close reproduction of detailed biological phenomena; in others the goal is to apply more abstract biological principles to the solution of engineering problems. All underline the point that there is still a great deal to learn about, and to learn from, biological nervous systems, and building models is a great way to learn!

5. Large-scale neuromorphic systems

A number of large-scale neuromorphic systems have emerged over recent years, taking advantage of the enormous transistor resource now available on a single microchip and, in one case, a full silicon wafer. The capabilities of the technology combine with scalable architectures to allow neuromorphic capabilities to extend to support scales of neural network incorporating many millions of neurons with many billions of synapses. These new capabilities enable modellers to contemplate building models of the complete brains of animals from insects up to smaller mammals, or substantial sub-areas of the human brain, and the same systems also offer platforms capable of supporting new scales of cognitive architecture.

Although these platforms now exist, in some cases at considerable scale, their full potential has yet to be realized. But already they display a diverse range of approaches motivated by a common conviction that novel brain-inspired approaches to computation have much to offer.

The large-scale systems described below offer a range of complementary approaches and divergent goals:

- The IBM TrueNorth chip is based upon distributed digital neural models aimed at real-time cognitive applications.

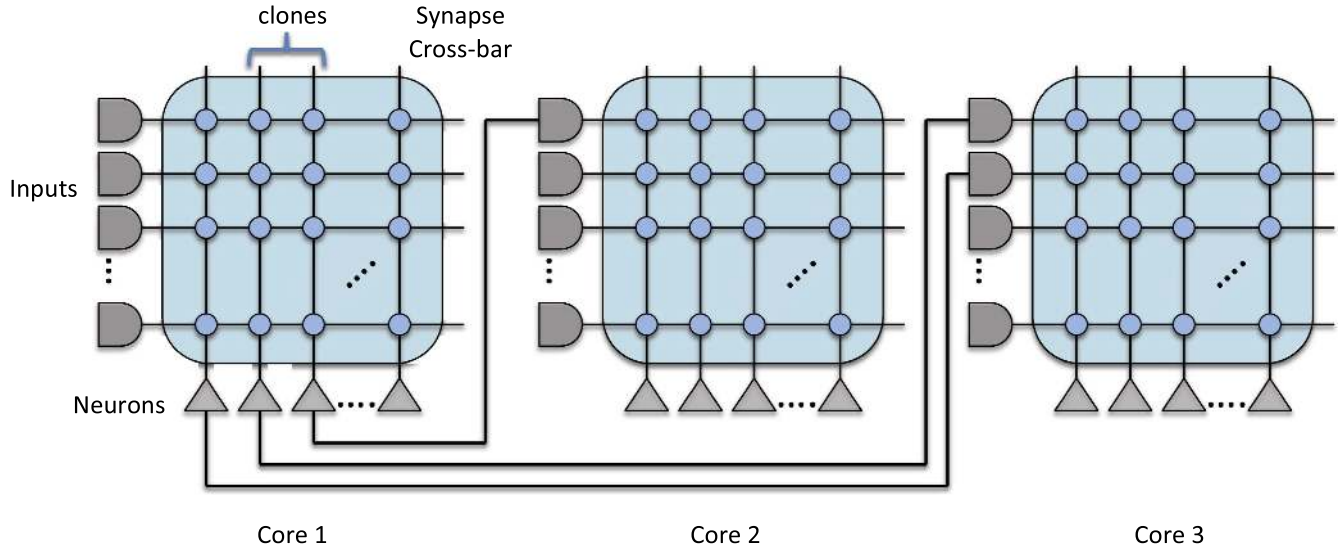


Figure 1. TrueNorth communications are based on point-to-point links conveying spikes from one neuron to one neurosynaptic core, where those spikes can connect into any or all of the 256 neurons in that core. Here the leftmost neuron in core 1 connects to core 3. To connect to cores 2 and 3, the 2nd and 3rd neurons in core 1 duplicate each other, and each makes one connection.

- The Stanford Neurogrid uses real-time sub-threshold analogue neural circuits.
- The Heidelberg BrainScaleS system uses wafer-scale above threshold analogue neural circuits running 10 000 times faster than biological real time aimed at understanding biological systems, and in particular, long-term learning.
- The Manchester SpiNNaker machine is a real-time digital many-core system that implements neural and synapse models in software running on small embedded processors, again primarily aimed at modelling biological nervous systems.

All of these approaches represent trade-offs—compromises between a set of desirable objectives. Energy-efficiency, integration density, flexibility and configurability, analogue versus digital algorithms, hardware versus software—all of these factors find different balances in the systems that are described below.

6. IBM Truenorth

The IBM TrueNorth chip [17, 18] is the outcome of a decade of work under the DARPA SYNAPSE programme aimed at delivering a very dense, energy-efficient platform capable of supporting a range of cognitive applications. The key component is a very large, 5.4 million transistor 28 nm CMOS chip that incorporates 4096 neurosynaptic cores where each core comprises 256 neurons each with 256 synaptic inputs [19]. The chip is all digital, and operates asynchronously apart from a 1 kHz clock that defines the basic time step. As a result, the hardware behaves deterministically exactly as predicted by a software model, which can therefore be used for application development and to implement learning algorithms.

6.1. TrueNorth design

The central design of a TrueNorth neurosynaptic core is a 256×256 cross-bar that selectively connects incoming neural spike events to outgoing neurons. The cross-bar inputs are coupled via buffers that can insert axonal delays. The cross-bar switches are binary, although each input is associated with one of four synapse 'types', and each neuron assigns an integer weight in the range -255 to $+255$ to each of the four types to give a synaptic weight to each connection—all active synapses associated with a particular input have the same type which is mapped independently to one of four weights by each neuron.

The outputs from the cross-bar couple into the digital neuron model, which implements a form of integrate-and-fire algorithm with 23 configurable parameters [19] that can be adjusted to yield a range of different behaviours, and digital pseudo-random sources are used to generate stochastic behaviours through modulating the synaptic connections, the neuron threshold and the neuron leakage.

6.2. TrueNorth communications

Neuron spike event outputs from each core follow individually-configurable point-to-point routes to the input to another core, which can be on the same or another TrueNorth chip. Where a neuron output is required to connect to two or more neurosynaptic cores, the neuron is simply replicated within the same core (see figure 1). The deterministic nature of the digital model ensures that all replicants will produce identical spike trains.

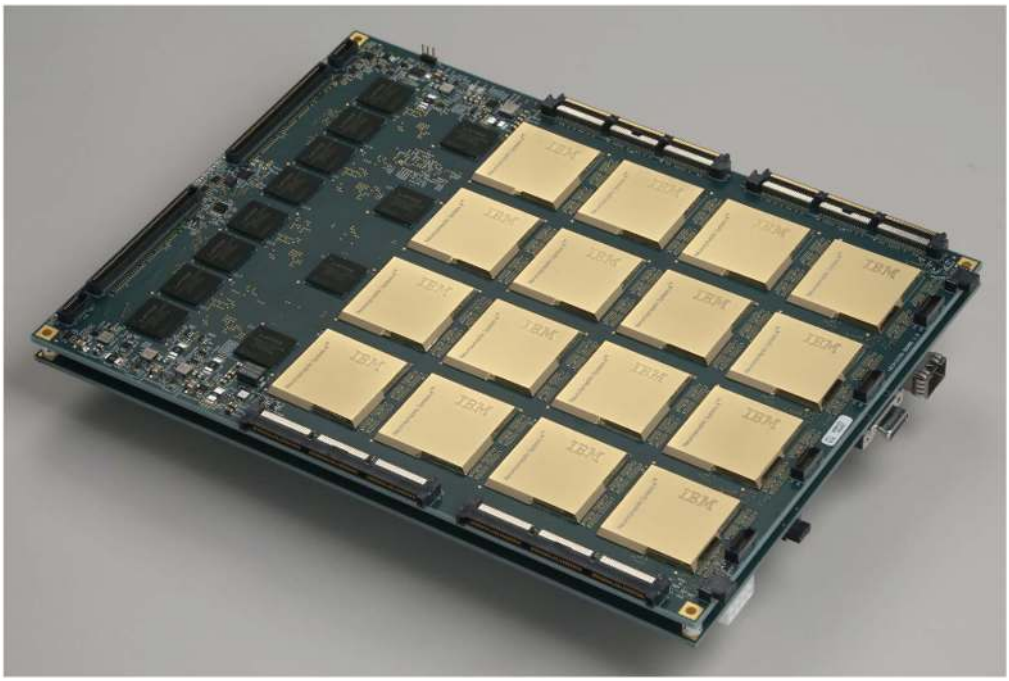


Figure 2. The NS16e circuit board incorporating 16 IBM TrueNorth chips. (Photo courtesy of IBM Corp and reproduced with permission.)

Inter-chip connections are multiplexed to reduce the number of electrical connections between chips, but the effect is to enable the network to extend seamlessly across systems of multiple TrueNorth chips.

6.3. TrueNorth systems

TrueNorth chips can be connected directly together to form larger systems, and a circuit board with 16 chips has been developed (see figure 2), incorporating a total of 16 million neurons and 4 billion synapses. Larger systems can be assembled by connecting multiple boards together.

6.4. TrueNorth support software

The TrueNorth hardware is supported by a software emulator which, exploiting the deterministic nature of the hardware, can be relied upon to predict the performance of the hardware exactly. The deterministic behaviour extends to the role of noise in inducing stochastic behaviour in the system, as the software model can predict the pseudo-random sequences generated from a given seed. The binary nature of the synapses makes on-line learning problematic, so the training takes place off line in the software environment.

The philosophy underpinning the TrueNorth support software is to raise the level of abstraction at which applications are conceived from the level of the individual neuron to the level of cognitive modules, where each module occupies one neurosynaptic core, and a library of such modules can be pre-generated and made available with tested and tried performance and behaviour. This is a very promising approach if it can be achieved since designing complex systems at the level of individual neurons incurs the risk of exposure to a very large parameter search space and can be a very open-ended endeavour.

6.5. TrueNorth applications

TrueNorth, more than any of the other large-scale neuromorphic platforms described here, is designed as an application delivery platform. It is intended to address problems across the range from vision (in particular, using event-based vision sensors such as those developed by INI) to audition and multi-sensory fusion. It offers very power-efficient real-time processing for high-dimensional, noisy, sensory data.

Applications such as real-time object recognition have been demonstrated running on TrueNorth at remarkably low power levels [17].

7. Neurogrid

The large-scale neuromorphic development that displays the strongest association with the heritage of Carver Mead at CalTech is the Stanford Neurogrid [20], which is perhaps not surprising since the leader of the Neurogrid project, Kwabena Boahen, was advised by Mead during his PhD at CalTech.

7.1. Neurogrid design

Neurogrid uses subthreshold analogue circuits to model neuron and synapse dynamics in biological real time, with digital spike communication. All inputs to a neuron go to one of four shared synapse circuits. The neuronal model uses shared leaky integrator dendritic structures whereby an input to one neuron affects neighbouring neurons through a

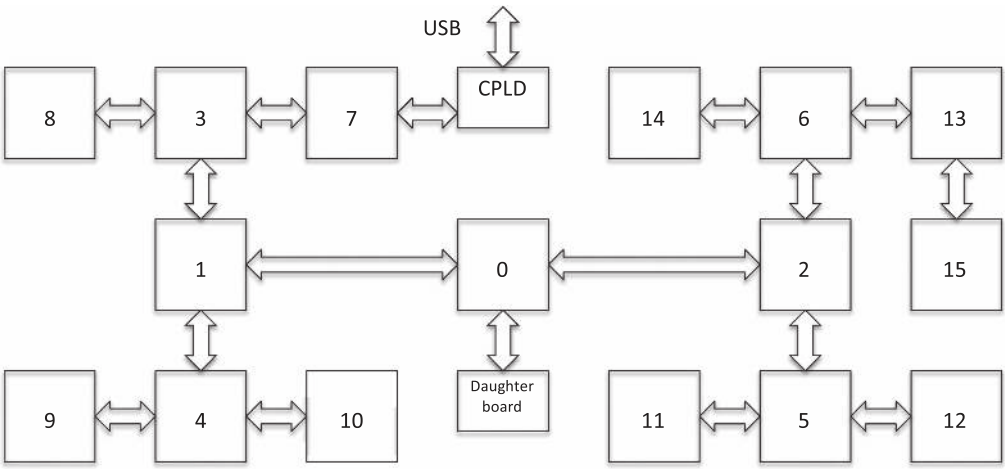


Figure 3. The Neurogrid tree hierarchy. The layout of the tree structure shown here can be seen reflected in the physical layout of the circuit board shown in figure 4. The organisation is a binary tree where each Neurocore chip (e.g. chip 1) connects to one parent (0) and two child chips (3 and 4). Multicast routing is achieved by passing a spike packet up to a chip which sits above all of the destination nodes, and then back down to those destination nodes, with duplication where required.

Figure 4. The Stanford Neurogrid system. 16 Neurocore chips are connected in a tree structure (see figure 3) on the circuit board. (Photo courtesy of Kwabena Boahen, Stanford University and reproduced with permission.)

resistive network. The neuron dynamics are defined by a quadratic integrate and fire model [20].

7.2. Neurogrid communications

Each Neurocore chip includes a router that is able to route spike packets between its local chip, its parent chip, and its two child chips. The routers support a multicast tree routing organisation (see figure 3), where spikes are passed up point-to-point to a node that sits above all of the intended destinations in the tree, and thence down to all of the destinations, being duplicated when required.

7.3. Neurogrid system

The Neurogrid system comprises a software suite for configuration and visualisation of neural activity together with a hardware platform to support real-time simulation of the neural network. The hardware is a circuit board incorporating 16 Neurocore chips plus some support circuitry (figure 4),

where each Neurocore chip supports 65 536 sub-threshold analogue neurons.

7.4. Neurogrid support software

The Neurogrid system is controlled by a host computer via a USB connection. The USB connects to a CPLD that translates USB packets into Neurogrid packets and vice versa.

The support software includes [20]: a user interface that allows the model to be specified using Python, real-time visualization of results, and run-time user interaction; a hardware abstraction layer that maps the model description onto the hardware; and hardware drivers that load the mapped problem onto the hardware using Neurogrid packets.

7.5. Neurogrid applications

The real-time operation of Neurogrid makes it suitable for robotic control, and it has been interfaced to a robotic arm with the ultimate goal of controlling a prosthetic limb. Further funding is aimed at exploiting the very low-power nature of the technology to develop a chip that can be implanted in the brain to control a prosthetic limb and to develop technology for drone control.

8. BrainScaleS

The BrainScaleS neuromorphic system has been developed at the University of Heidelberg over a series of projects funded by the European Union, including the FACETS projects and the BrainScaleS Project. Ongoing support for BrainScaleS comes from the EU ICT Flagship Human Brain Project.

8.1. BrainScaleS design

The key concepts employed in the design of BrainScaleS are:

- The use of above-threshold analogue circuits to implement *physical models* of neuronal processes. Physical models exploit the analogy between ionic circuits in biological neurons and electronic circuits. The above-threshold circuits used here contrast with the sub-threshold circuits favoured by Carver Mead and used in the Stanford Neurogrid, and yield much faster circuits, running at 10 000 times biological speeds.
- The use of wafer-scale integration to deliver large numbers of analogue neurons that can be interconnected very efficiently to accommodate the 10 000 times speed-up [21, 22].

Wafer-scale integration of high-speed analogue circuits is a very aggressive technological approach to large-scale neuromorphic computing, and has required several innovative technological solutions to be found. For example, although silicon wafers are used for all microchip manufacture, the high-precision manufacturing process can only be applied within a single reticle area of a few square centimetres at a time, and step-and-repeat is used to produce multiple copies of the same circuit across the wafer. These copies cannot be connected to each other during the standard manufacturing process, so the BrainScaleS wafer uses post-processing to add an additional metal layer (with coarser resolution than the metal lines used within a microchip) to achieve global connectivity across the wafer.

Within a BrainScaleS wafer each of the 48 reticles holds eight High-Count Analogue Neural Network (HiCANN) die, each of which implements 512 adaptive exponential integrate and fire (AdExp) [21] neurons and over 100 000 synapses.

8.2. BrainScaleS communications

The primary communication layer in the BrainScaleS system operates within a wafer. Hi-speed serial channels each convey the output of 64 neurons from HiCANN die to HiCANN die in continuous time, with only small timing errors introduced if two simultaneous output spikes contend for the same channel. These high-speed channels pass through cross-bar switches (which implement only sparse connectivity to reduce the capacitive load on the channels) to route the channels across the wafer.

In order to support simulations across multiple wafers a second layer of communication is supported using FPGAs to implement high-speed serial communications between wafers. The wafer post-processing is used both to make lateral connections between adjacent HiCANN die in separate reticles (connections within the same reticle use conventional wafer metal layers), and also to present signals from within each HiCANN die for external connection. These external connections are made through elastomeric strip connectors to the system board where FPGA communication PCBs time-stamp the spikes for onward distribution (see figure 5).

The overall wafer module structure integrates the wafer, power supplies, communications and analogue support circuits.

8.3. BrainScaleS systems

BrainScaleS hardware is available in small portable form, based around a single HiCANN die, and also in the form of the EU Human Brain Project 20-wafer platform which incorporates a substantial cluster server that acts as host, executing the network mapping functions and controlling the overall operation of the machine (see figure 6).

8.4. BrainScaleS support software

The development of the BrainScaleS system has gone hand-in-hand through a series of European projects with the development of PyNN [23], a python-based neural network description language. PyNN supports a *population-projection* based view of neural networks wherein a group of similar neurons are collectively viewed as a population, and this population projects to synaptic connections with other populations where a projection can be all-to-all, one-to-one, sparse with a given probability and so on.

PyNN not only specifies the network but can also define the network inputs and how the user wishes to visualise the

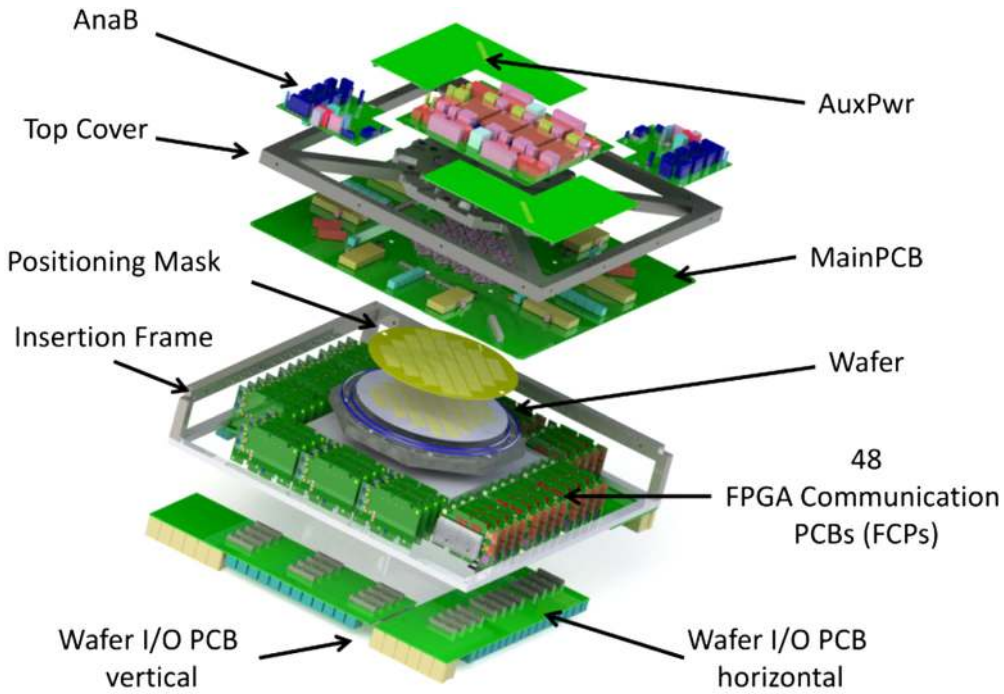


Figure 5. The BrainScaleS wafer module structure. (Diagram courtesy of Karlheinz Meier, Heidelberg University.)

Figure 6. The Brainscales 20-wafer machine. The wafer modules are mounted four per cabinet in five rack cabinets distributed either side of a cluster server that manages the system. (Photo courtesy of Karlheinz Meier, Heidelberg University and reproduced with permission.)

outputs, offering a sophisticated environment for specifying and managing neural network modelling.

8.5. BrainScaleS applications

The very high speed of the BrainScaleS system renders it highly suited to applications that take a very long time in biological terms. Examples of these are long-term learning tasks, such as modelling several years of childhood development, where the 10 000 times speed-up potentially turns years into hours. Different application domains would include very large-scale parameter searches, where high-speed batch-mode operation has direct benefit in terms of run-time.

9. SpiNNaker

The SpiNNaker project [24] has developed a massively-parallel digital computer whose communication infrastructure is motivated by the objective of modelling large-scale spiking neural networks with connectivity similar to the brain in biological real time. The current largest SpiNNaker machine (available as one of the EU Flagship Human Brain Project platforms) incorporates 500 000 processor cores, with a goal of doubling this to a million cores over the coming year.

In many respects SpiNNaker resembles a conventional supercomputer, with the following notable differences:

- The processors in SpiNNaker are small integer cores originally intended for mobile and embedded applications, rather than the high-end 'fat' cores preferred by supercomputer designers.
- The brain-inspired communications fabric in SpiNNaker [25] is optimised for sending large numbers of very small data packets (each typically conveying one neural spike) to many destinations following statically configured multicast paths, whereas supercomputers typically use large point-to-point packets with dynamic routing.

These differences mean that SpiNNaker should not be viewed as a general-purpose computer but rather as a specialised neurocomputer, although it is not, in fact, limited to modelling neural networks, and could potentially be suitable for a wider range of applications characterised by large numbers of relatively simple coupled processes with neuron-like communication properties. Examples might include cellular automata and finite-element problems. It is, perhaps, stretching the term to its limits to describe SpiNNaker as a neuromorphic system, but its inclusion here is justified because its primary purpose is to model neural networks.

9.1. SpiNNaker design

The design of SpiNNaker is motivated by two primary considerations:

- Scalability: brains, and especially the human brain, incorporate very large numbers of components, and modelling them is computationally very demanding. As a result, any system aspiring to approach the scale of the human brain must embody the principle of scalability.
- Energy-efficiency: because of the large scale of the system, its energy consumption risks becoming uneconomically large. Energy-efficient design is a holistic discipline, and SpiNNaker's design is influenced by this objective throughout.

These considerations lead to the fundamental design of SpiNNaker, which is based around a small plastic 300 bga (ball grid array) package which incorporates a custom processing chip [26] and a standard 128 Mbyte SDRAM memory chip. The processing chip, designed on a 130 nm CMOS technology, contains 18 ARM968 processor cores, each with 32 Kbytes of instruction memory and 64 Kbytes of data memory, a multicast packet router, and sundry support components. The principle at work here is to minimise the distances over which frequently accessed data must be moved: the code and most frequently-used data are within a millimetre or two of the core, and the less frequently-accessed data is on the SDRAM which is about 1 cm away from the core.

Energy-efficiency is achieved by delivering the full 18-core package with a maximum 1 W power dissipation when all cores are fully loaded, and managing the power down from this level when the compute load is lower.

Scalability is achieved by designing the package such that an (almost) arbitrarily large 2D surface can be tiled with these packages [27].

9.2. SpiNNaker communication

The SpiNNaker communication fabric is based on a 2D triangular mesh with each node formed from a processor layer and a memory layer (both in the same package) as described above. The router accepts packets from all of the 18 resident processor cores and the 6 incoming inter-chip links, and then uses an associative lookup table to decide how to copy the packet to any subset (or all) of its local processors and any subset (or all) of the outgoing inter-chip links. The result is that a single spike can propagate through an arbitrary tree to an arbitrary number of destinations within the machine (see figure 7).

The routing is based upon packet-switched Address Event Representation and relies on the fact that the connections from a particular neuron are static, or at most slowly changing. Each neuron can route through a unique tree, though in practice routing is based on populations of neurons rather than individual neurons, and the restricted size of each routing table makes this optimisation necessary on most cases.

9.3. SpiNNaker systems

SpiNNaker is delivered in two basic circuit board configurations (see figure 8):

- A 4-node (72-core) board, for training and small network development, which can be powered from a 5 V 1 A USB

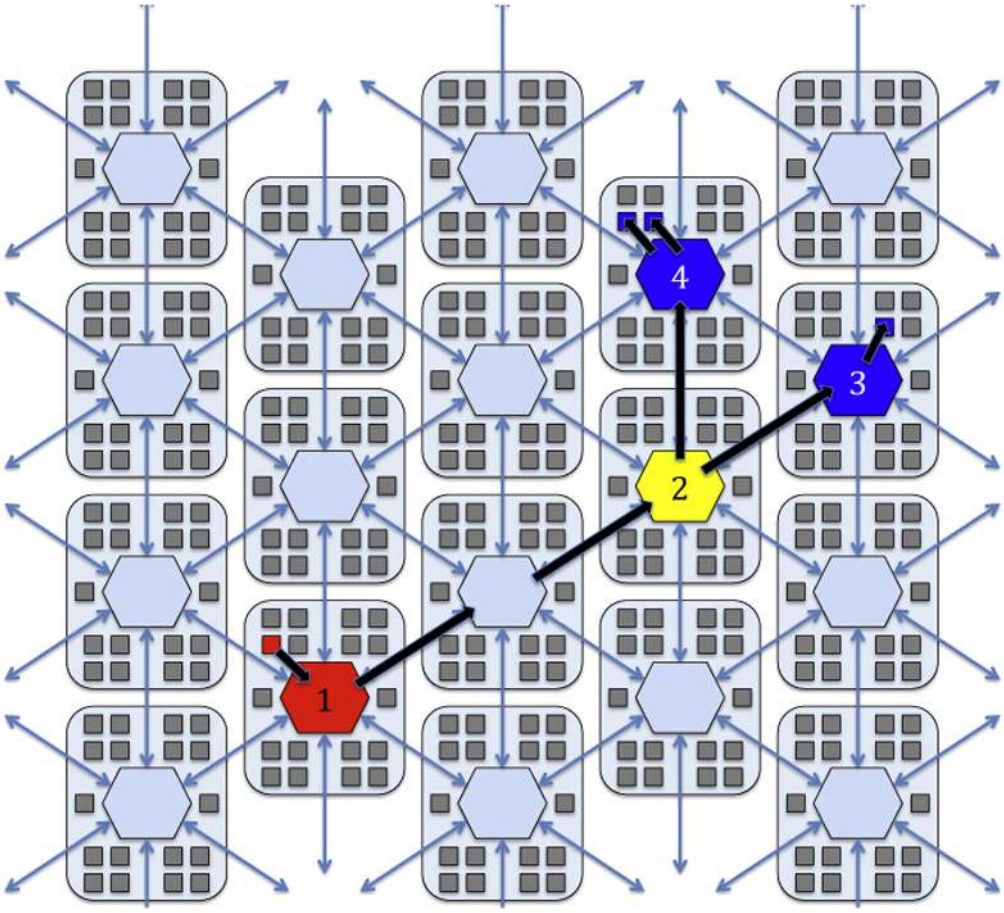


Figure 7. Multicast communications on SpiNNaker. The sending core (shown as a square) on chip 1 sends a spike packet to its local router (shown as a hexagon) which then passes it on towards the three destination cores on chips 3 and 4. The first router on the path can use default routing to pass the packet straight through. The router on chip 2 makes two copies, one to each destination router. The destination routers then send copies on to the receiver cores on their respective chips. The paths are configured in the router tables and can be arbitrary trees.

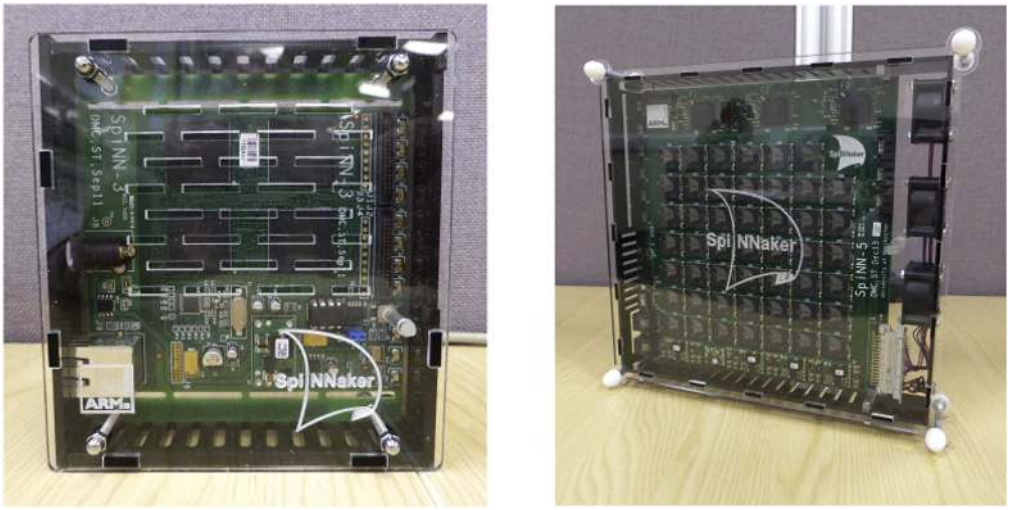


Figure 8. Small-scale SpiNNaker systems. The 4-node (72-core) system (left) can be powered from a USB port and is useful for familiarisation and small projects, for example in robotics. The 48-node (864-core) system (right) can be used for larger projects and is the basis of the larger machines. (Photos courtesy of the University of Manchester.)

Figure 9. The 500 000-core SpiNNaker machine. Five rack cabinets each hold 120 48-node SpiNNaker boards, with the high-speed serial wiring that connects the boards into a 2D toroidal surface visible across the fronts of the cabinets (whose doors have been removed for the photo). The sixth cabinet contains the server that manages the machine. (Photo courtesy of the University of Manchester.)

port and connected to a host machine (PC or laptop) through a wired Ethernet connection.

- A 48-node (864-core) board, which implements a hexagonal array of SpiNNaker nodes with FPGAs that support high-speed serial board-to-board interconnect, allowing multiple boards to be assembled into a single 2D toroidal mesh. This board is the basis of larger SpiNNaker systems.

The 48-node board uses a standard double extended Eurocard circuit board configuration that allows multi-board systems to be assembled in off-the-shelf card frames and racking. A single 19″ card frame can accommodate 24 boards with 20 000 cores requiring a 2 kW 12 V power supply and forced-air cooling. Five such card frames can be assembled into a 19″ rack cabinet giving 100 000 cores at 10 kW (peak), and multiple cabinets can be wired together to give a single SpiNNaker toroid incorporating up to a million cores in ten cabinets. The current large-scale system is half this size at 500 000 cores (see figure 9).

Each 48-node board has two 100 Mbit Ethernet connections, one to load application data onto the board and to recover results, the other to communicate with a Board Management Processor (BMP) that can monitor and configure the board's power and communication systems. Only one BMP Ethernet connection is required per 19″ card frame as all of the BMPs in a frame communicate through a backplane bus.

Large SpiNNaker systems require a substantial host machine (a server or cluster) to map applications onto the machine and offload results. Ideally the host should be able to handle concurrent transfers across the Ethernet to all (up to 1200) boards.

9.4. SpiNNaker support software

The host machine is responsible for managing all aspects of the SpiNNaker system, which can be viewed simply as a neural accelerator for the host machine.

Applications will normally be described in a high-level neural description language such as PyNN [23], Nengo [28], or similar. Software running on the host will map the topology of the high-level model onto SpiNNaker's routing and compute resources and map the functionality of the model onto real-time library functions to be run on the respective cores.

Some aspects of detailing the model can be implemented more effectively by postponing their expansion until after the model has been loaded onto SpiNNaker, because then SpiNNaker's inherent high levels of parallelism can be exploited to accelerate this expansion. This also helps overcome the bottleneck presented by the 100 Mbit Ethernet connection into each board.

9.5. SpiNNaker applications

Small-scale SpiNNaker systems have been used for a range of real-time applications such as robot control and vision processing, and also for non-real-time modelling of biological

Table 1. A comparison of the major features of the human brain and the four large-scale neuromorphic systems described in this paper.

Platform:	Human brain	Neurogrid	BrainScaleS	TrueNorth	SpiNNaker
Technology:	Biology	Analogue, sub-threshold	Analogue, over threshold	Digital, fixed	Digital, programmable
Microchip:		Neurocore	HiCANN		18 ARM cores
Feature size:	10 μm[a]	180 nm	180 nm	28 nm	130 nm
# transistors:		23 M	15 M	5.4 B	100 M
die size:		1.7 cm^2	0.5 cm^2	4.3 cm^2	1 cm^2
# neurons:		65 k	512	1 M	16 k
# synapses:		~100 M	100 k	256 M	16 M
power:		150 mW	1.3 W	72 mW	1 W
Board/unit:		PCB	20 cm wafer	PCB	PCB
# chips:		16	352	16	48
# neurons:		1 M	200 k	16 M	768 k
# synapses:		4 B	40 M	4B	768 M
power:		3 W	500 W	1 W	80 W
Reference system:	1.4 kg		20 wafers in 7 × 19″ racks		600 PCBs in 6 × 19″ racks
# neurons:	100 B		4 M		460 M
# synapses:	10^{15}		1 B		460 B
power:	20 W		10 kW		50 kW
Energy/connection:	10 fJ	100 pJ	100 pJ	25 pJ	10 nJ
Speed versus biology:	1×	1×	10 000×	1×	1×
Interconnect:	3D direct signalling	Tree-multicast	Hierarchical	2D mesh-unicast	2D mesh-multicast
Neuron model:	Diverse, fixed	Adaptive quadratic IF	Adaptive exponential IF	LIF	Programmable[b]
Synapse model:	Diverse	Shared dendrite	4-bit digital	Binary, 4 modulators	Programmable[c]
Run-time plasticity:	Yes!	No	STDP	No	Programmable[d]

[a] This is the approximate diameter of a cortical neuron soma. There are, of course, internal processes at much smaller (molecular) scales.
[b] SpiNNaker is optimised for point-neuron models, such as leaky integrate and fire (LIF), Izhikevich, and similar. It is less suited to models with high biological detail.
[c] As above, computational limits encourage the use of the simpler synapse models such as voltage step or conductance-based with single- or double-exponential post-synaptic potentials.
[d] Many synaptic learning rules, such as spike-time dependent plasticity (STDP), are computationally tricky (because of the event-driven software model) and expensive on SpiNNaker, and approximate implementations can improve efficiency significantly.

circuits. There are few examples of large-scale applications to date. Perhaps most notable among these is a real-time implementation of the 2.5 million neuron Spaun model [29], representing a 9000× speed-up compared with the 2–3 h the model takes to simulate a second of real time on a high-end desktop machine.

10. Discussion

The four large-scale neuromorphic systems described above represent a diverse range of approaches to modelling neural systems, and with somewhat diverse objectives. Each has its merits and its drawbacks, and there is no question of any one approach being superior to the others in every respect.

The diversity in approach makes direct comparison difficult, but table 1 summarizes some of the main points for comparison. Of particular note are:

- The energy per (synaptic) connection. This is the average energy required to pass one spike through one (static) synapse. Since synaptic processing dominates the system energy cost, this is the best guide to the overall energy-efficiency of the system. Here it is notable that the two analogue systems have very similar measures despite their significant differences in circuit implementation (above versus sub-threshold) and speed (10 000× versus 1× biological real time). Perhaps more surprising is that TrueNorth, using a digital approach, surpasses both analogue systems, perhaps because it uses a much more aggressive semiconductor technology, which its digital nature facilitates.
- Modelling speed. Three of the four systems run at biological speeds, but BrainScaleS runs 10 000× faster, thereby potentially compressing years of learning into hours.
- Model flexibility. Here the programmable nature of SpiNNaker offers significant benefits, albeit at a significant cost in terms of energy-efficiency. Novel neuron models, such as those incorporating stochasticity, can be incorporated into the library of run-time options relatively easily, as can novel learning algorithms.

- Physical modelling. In Carver Mead's seminal work on neuromorphic systems [4] he made much of the analogous behaviour of sub-threshold electronics and neural ion channel behaviour. Both analogue systems implement the neural equations in continuous time whereas the digital systems use some form of discrete time approximation, but only Neurogrid keeps close to Mead's original conceptual approach.

So each of the large-scale systems described here has its strengths: TrueNorth offers a platform for very highly-integrated and energy-efficient application delivery; SpiNNaker offers maximum flexibility for researching different neural models and plasticity rules; BrainScaleS brings high acceleration for long-term learning; and Neurogrid offers very good energy-efficiency with models that are closest to the physics at work in the biology.

All of these systems will evolve over time, and next-generation versions of most of them are already under development. But these first-generation large-scale systems have established a new base-line from which future systems will emerge, guided by extensive user experience with this first generation, so that progress is visible not only in terms of performance and efficiency but also in terms of flexibility and user-accessibility.

11. Conclusions

A new computational paradigm is emerging based on biologically-inspired principles, directed both towards enhancing our understanding of the clear front-runner in table 1—the human brain itself—and towards applying our current very partial knowledge to the development of cognitive systems, neural prosthetics, and similar applications. At this early stage in the development of such systems much is unknown, and the diversity represented by the four large-scale systems discussed here is to be welcomed—we are still some way from identifying an optimal approach that is as general-purpose in this domain as is the general-purpose programmable processor in the conventional computing domain.

Acknowledgments

The design and construction of the SpiNNaker machine was supported by EPSRC (the UK Engineering and Physical Sciences Research Council) under grants EP/D07908X/1 and EP/G015740/1, in collaboration with the universities of Southampton, Cambridge and Sheffield and with industry partners ARM Ltd, Silistix Ltd and Thales. Ongoing development of the software is supported by the EU ICT Flagship Human Brain Project (FP7-604102), in collaboration with many university and industry partners across the EU and beyond, and our own exploration of the capabilities of the machine is supported by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013)/ERC grant agreement 320689. SpiNNaker has been 15 years in conception and 10 years in construction, and many folk in Manchester and in our various collaborating groups around the world have contributed to get the project to its current state. We gratefully acknowledge all of these contributions.

References

[1] Moore G E 1965 Cramming more components onto integrated circuits *Electronics* **38** 114–7
[2] Hodgkin A and Huxley A F 1952 A quantitative description of membrane current and its application to conduction and excitation in nerve *J. Physiol.* **117** 500–44
[3] Sterling P and Laughlin S 2015 *Principles of Neural Design* (Cambridge, MA: MIT Press)
[4] Mead C 1989 *Analog VLSI and Neural Systems* (Reading, MA: Addison-Wesley)
[5] Mahowald M A 1992 VLSI analogs of neuronal visual processing: a synthesis of form and function *PhD Thesis* California Institute of Technology
[6] Mahowald M A and Mead C 1991 The silicon retina *Sci. Am.* **264** 76–82
[7] Mead C and Ismail M (ed) 1989 *Analog VLSI Implementation of Neural Systems* (Dordrecht: Kluwer)
[8] Brandli C, Muller L and Delbruck T 2014 Real-time, high-speed video decompression using a frame- and event-based DAVIS sensor *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS) (Melbourne, VIC, 1–5 June)* pp 686–9
[9] Yang M-H, Chien C-H, Delbruck T and Liu S-C 2016 A 0.5 V 55_uW 64X2-channel binaural silicon cochlea for event-driven stereo-audio sensing *Proc. IEEE Int. Solid-State Circuits Conf. (ISSCC)*
[10] Indiveri G, Corradi F and Qiao N 2015 Neuromorphic architectures for spiking deep neural networks *Proc. IEEE Int. Electron Devices Meeting (IEDM)* pp 4.2.1–4.2.4
[11] Yu T, Park J, Joshi S, Maier C and Cauwenberghs G 2012 65k-neuron integrate-and-fire array transceiver with address-event reconfigurable synaptic routing *Proc. IEEE Biomedical Circuits and Systems Conf. (BioCAS) (Hsinchu)* pp 21–4
[12] Park J, Yu T, Maier C, Joshi S and Cauwenberghs G 2012 Live demonstration: hierarchical address-event routing architecture for reconfigurable large scale neuromorphic systems *Proc. IEEE Int. Symp. on Circuits and Systems (ISCAS) (Seoul)* pp 707–11
[13] LeMasson G, Renaud-LeMasson S, Debay D and Bal T 2002 Feedback inhibition controls spike transfer in hybrid thalamic circuits *Nature* **417** 854–8
[14] Zou Q, Bornat Y, Saighi S, Tomas J, Renaud S and Destexhe A 2006 Analog-digital simulations of full conductance-based networks of spiking neurons with spike timing dependent plasticity *Netw., Comput. Neural Syst.* **17** 211–33
[15] Jo S H *et al* 2010 Nanoscale memristor device as synapse in neuromorphic systems *Nano Lett.* **10** 1297–301
[16] Chang Y-F *et al* 2016 Demonstration of synaptic behaviors and resistive switching characterizations by proton exchange reactions in silicon oxide *Sci. Rep.* **6** 21268
[17] Merolla P A *et al* 2014 A million spiking-neuron integrated circuit with a scalable communication network and interface *Science* **345** 668–73
[18] Hsu J 2014 IBM's new brain *IEEE Spectrum* pp 17–9

[19] Cassidy A S *et al* 2013 Cognitive computing building block: a versatile and efficient digital neuron model for neurosynaptic cores *Proc. IJCNN*

[20] Benjamin B V *et al* 2014 Neurogrid: a mixed-analog-digital multichip system for large-scale neural simuations *Proc. IEEE* **102** 699–716

[21] Schemmel J *et al* 2010 A wafer-scale neuromorphic hardware system for large-scale neural modeling *Proc. Int. Symp. Circuits System* pp 1947–50

[22] Scholze S *et al* 2012 A 32 GBit s^{-1} communication SoC for a waferscale neuromorphic system *Integr. VLSI J.* **45** 61–75

[23] Bruderle D *et al* 2009 Establishing a novel modeling tool: a python-based interface for a neuromorphic hardware system *Front. Neuroinform.* **3** 2009

[24] Furber S B *et al* 2014 The SpiNNaker project *Proc. IEEE* **102** 652–65

[25] Plana L A *et al* 2011 SpiNNaker: design and implementation of a GALS multi-core system-on-chip *ACM J. Emerg. Technol. Comput. Syst.* **7** 1–17

[26] Painkras E *et al* 2013 SpiNNaker: a 1 W 18-core system-on-chip for massively-parallel neural network simulation *IEEE J. Solid-State Circuits* **48** 1943–53

[27] Furber S B *et al* 2013 Overview of the SpiNNaker system architecture *IEEE Trans. Comput.* **62** 2454–67

[28] Stewart T C and Eliasmith C 2014 Large-scale synthesis of fuctional spiking neural circuits *Proc. IEEE* **102** 881–98

[29] Eliasmith C 2013 *To Build a Brain* (Oxford : Oxford University Press)