Dense-Near/Sparse-Far Hybrid Reconfigurable Neural Network Chip

Robin Emery, Alex Yakovlev, Graeme Chester
Newcastle University
{r.a.emery,alex.yakovlev,graeme.chester}@ncl.ac.uk

***Abstract*—An adaptable device that will be a useful tool to further understanding of the operation of the mammalian nervous system, including the brain, is discussed. An example spiking neuron model that has been manufactured at 130nm is presented. In hardware platforms for neural networks that implement some degree of realism of interest to neuroscientists, connectivity between neurons can be a major limitation. Distal connectivity is an active area of Neuroscience, exhibiting a recent growth in results. Through the combination of new data and the application of the FPGA principle of providing many simple functional blocks connected by well designed routing resources (a hybrid of direct connectivity and AER), a useful degree of inter-neuron connectivity will be provided. Autonomous synaptic plasticity will also be investigated.**

I. Introduction

The reconfigurable nature of the FPGA lends itself to the study of naturalistic neural networks, but the architecture of modern FPGAs is not well suited to the task. Routing resources are not optimal and real-time learning is awkward. First-hand experience of work on a small FPGA neural network showed that so long as the area occupied by the neuron model was large (many CLBs), the routing resource available was significantly greater than that required by a neural network built using these large neurons [1]. As FPGA routing resources can occupy 70-90% of the chip area [2], it is desirable to optimise the connectivity to fit the less general case presented by a neural network.

Connectivity is currently of great interest to neuroscientists, but is a limiting factor of hardware neural network models. A greater understanding of the nature of inter-neuronal connectivity in the brain may reveal how information is routed to relevant centers of action and how it is subsequently processed. Observations of neural networks in the brain show that a single neuron may make between 100 and 3000 connections to other neurons [3], and that the majority of these connections are proximal to the originating neuron or form dense groups at several distal points, as shown in figure 1.

It is understood that the influence of a synapse becomes stronger or weaker over time [4], and it is believed that this behaviour is a major contributor to learning, along with the higher level addition and removal of connections. This synaptic plasticity is real-time and is local to the synapse and the post-synaptic neuron. Autonomous learning, that can be observed, would make a hardware neural model significantly more viable as an experimentation platform.

A new routing topology is proposed that is optimised for reconfigurable neural networks through the combination of a local mesh topology for spikes with a global AER-based (Address Event Representation [5]) topology. A series of protocols are defined for this topology, and a reconfigurable array will be designed by combining this topology with general circuits for all the elements of a neural network.

Following some background material, the structure of the system will be presented, followed by the results so far obtained.

II. Motivation

The investigation of the function of the brain can be approached by characterising it into anatomically and physiologically distinct levels [6], from molecules through synapses and neurons to topographical maps and whole systems. Hodgkin and Huxley [7] characterised the mechanism of a neural action potential (or spike) in terms of ion channels and a leakage current, and our understanding of the neuron has subsequently greatly increased. Neuroscientists are very interested in higher level structures - those of networks and topographical maps formed by those networks. They are interested in questions of how information is transformed, and how it is routed around various cognitive areas of the brain. Detailed imaging and reconstruction techniques, especially for distal portions of neural connectivity, are of great interest [8]. Artificial systems capable of modelling these observed network structures are likewise of great interest. Such systems must be able to model large networks, must be accurate enough, must be observable, and must produce results in a reasonable period of time.

Experimental data on local connectivity in networks in [9] indicates that a neuron may receive a few thousand synaptic inputs over various parts of the dendritic structure. Results presented in [10] indicate that a neuron may receive about 30 synaptic inputs directly from similar neurons and an order of magnitude more indirectly through inter-neurons. The paper also shows that while the probability of connection decreases as one moves away from the soma to about a tenth of what it was, the actual *number* of connections increases. This is due to the high density of neurons in the observed volume; however the volume was too small to determine the trend for more distal connectivity.

Work undertaken by Binzegger et al. [3] provides some results for distal connectivity. They find that axons of neocortical neurons form connections in multiple, separate clusters (see figure 1). Neurons formed between one and seven clusters, with almost all forming at least two and most forming about two or three. The cluster with the highest number of connections (primary cluster) typically formed near to the parent

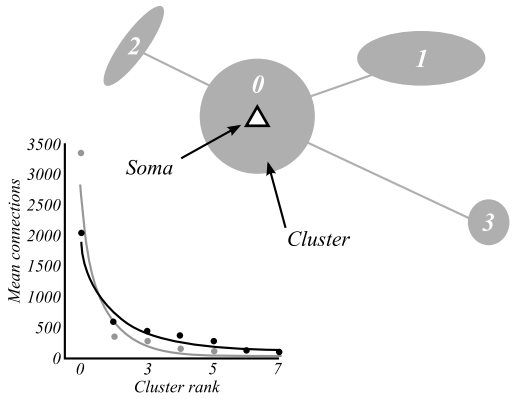

Figure 1. Neural connectivity in clusters (for a single neuron). The graph shows the relationship between the cluster rank and the number of connections formed (black: excitatory, grey: inhibitory). Adapted from [3].

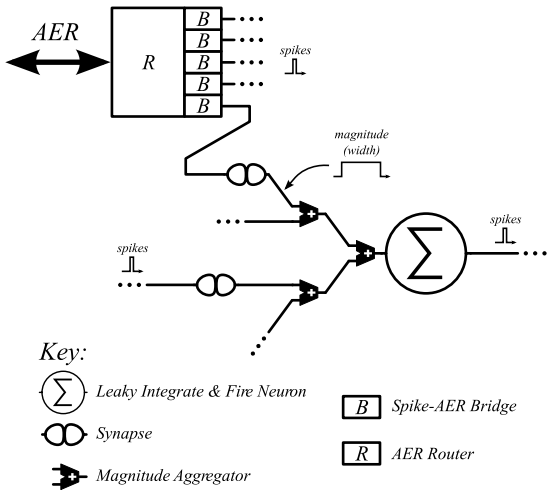

Figure 2. System Elements

soma, and the distance to the next cluster was proportional to the size of the parent cluster. They theorize that the "spoke-like" arrangement of clusters may be a means of routing information.

The operational mechanics of the neuron are well understood. There is currently much research into the mechanics of synaptic connectivity and how it relates to learning. Monitoring such long-term behaviours *in vivo* is difficult, although there are some interesting results from new 3D imaging technologies [8].

The available data supports an artificial network that has a hybrid of dense local connectivity and sparse connectivity over longer distances. Note that the depiction of clusters in figure 1 is for only a single neuron - in reality, there are many neurons with similar connectivity in close proximity.

Neuromorphic Engineering [11] concerns the application of VLSI to biology. Most commonly using analogue VLSI, parallels are drawn between the principals of biological computation and the behaviour of silicon. Mead observed that, in a similar way to the nervous system, analogue computational primitives emerge from fundamental laws of physics. Since the term was coined, "neuromorphic" has come to represent most systems that take a direct influence from natural biological computation.

Address Event Representation (AER) is a means of asynchronously multiplexing stereotypical "spikes" or "events" over a connection between two groups of neural processing elements. It was first used in Mahowald's stereoscopic vision system [12], and a draft standard has subsequently been produced [5]. In AER, an event is represented by at least a source or target address, so that the event can be received by the target. Events are digital - abstract - so any degradation can be restored without affecting the information conveyed. In those areas of the system where AER is used for communication, time can be seen as being the same throughout the system, with the same frame of reference for all nodes. Information is then conveyed in the time and number of events, providing a robustness to process variation between chips. When used in an adaptive system, a small number of lost events will not massively affect the operation of the system, providing an inbuilt robustness to transient faults.

Existing work on hardware for neural networks is mostly focused on application specific embedded hardware, rather than on accelerating general purpose computing. Artificial classifier networks tend to be the most common. Of neuromorphic devices, perhaps the Silicon Retina [12] is the most well known, but many other devices exist, mostly affected by connectivity or limited or cumbersome neurons. The most successful reconfigurable system of recent years is the Silicon Cortex (SCX) [13], a board-level infrastructure for multiple chip neuromorphic systems in which analogue VLSI chips are connected by digital hardware. However, a limited neuron population and a complex infrastructure makes this system unsuitable for large-scale networks.

Previous work that based a reconfigurable neural network on a Xilinx FPGA [1] determined that a hardware system must operate neurons asynchronously to be of interest to neuroscientists, and that a large area was required to implement a useful neuron due to the general nature of the FPGA logic. As a result of the large area required and the ample routing resources available, much of the interconnect was unused and was wasted.

In summary, a reconfigurable FPGA-like neural network device would be of interest to neuroscience, and observations of the structure of the brain support a hybrid of local and distal connectivity with a distinct transition from one to the other. Further, such a device would be of significantly more use if real-time learning is incorporated.

III. Reconfigurable Neural Array

In the light of the current direction of neuroscience research discussed in the previous section, a reconfigurable neural network architecture is presented. The architecture is not intended for classical artificial neural networks, but at networks inspired by a desire to understand large-scale observations of naturally occurring neural networks.

The basic structural elements of the system are depicted in figure 2. A simple neuron model, consisting of an integrator with a decay over time and a mechanism that emits a spike when a threshold level is reached, is connected to other

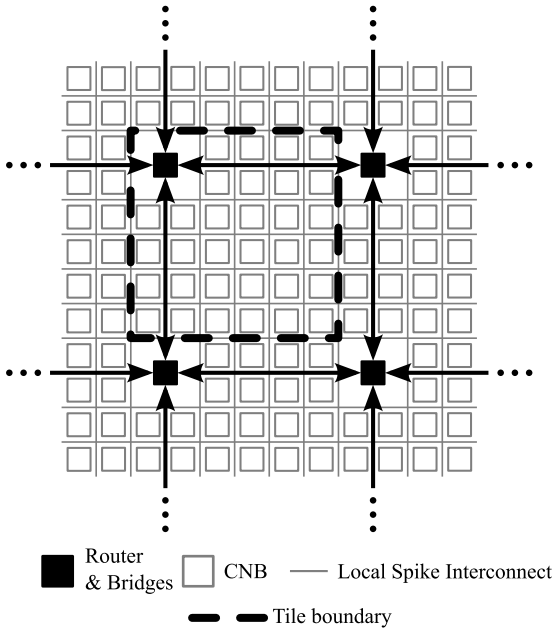

Figure 3. Hierarchical communication infrastructure. The distribution of routers and Configurable Neural Blocks over a tile has been chosen to aid illustration, and will be more numerous for the implemented design.

neurons via synaptic junctions. In response to the arrival of a spike at the input, a synapse will emit a value that can vary in magnitude according to a learning rule. The magnitude is represented by the width of a pulse, which when collected together through the dendritic tree of the neuron drive the integrator. If a spike needs to be conveyed over longer distances it is converted to an AER packet and put onto the AER network.

A neuron, several synapses and a tree of magnitude aggregators are collected to form a configurable block. These blocks are repeated many times to form a Cartesian grid, with spike links between blocks on a dedicated configurable routing resource, forming the local connectivity. For the distal connectivity, a large number of configurable blocks are grouped together along with an AER router and bridges to form a tile, as in figure 3. Tiles are connected together using fast serial AER links, packets being routed using a simple "*X,Y,bridge_id*" addressing scheme.

The tiles are not isolated in the traditional network-on-chip sense - the network is best thought of as "superimposed" over the shorter range network of neurons. Thus, the usage of the hybrid routing resources is flexible, with specific emphasis on the lack of a requirement to packetise a spike to cross a tile boundary.

As information is conveyed through the system in three forms - abstract "spikes", magnitudes in the form of pulse-widths, and AER packets - a series of protocols is required to correctly define the interaction of these communications. The two spike protocols are summarised in table I.

Magnitudes are communicated as long pulses that are used to drive an integrator, and do not travel very far. The pulses propagate through a binary tree towards the root, and if both incoming links at a node are active, the signals are simply OR'ed together. This method of aggregation can result in an

Layer	*Direct Connection*	*Indirect Connection*
Network (OSI 3)	*Not required* (point-to-point only, no routing)	AER. All nodes have unique address. Full-duplex mesh topology. Unicast
Data-link (OSI 2)	Spikes are buffered at the the receiver; new spikes are dropped if there is contention for the link. Source and destination are implicit, no access control required. No error detection service on physical layer.	Spikes can be ingress and egress queued if link is busy. Error detection service on the physical layer (malformed frames).
Physical (OSI 1)	Single wire per link; RZI; spike is falling edge. Simplex, fixed at configuration time. Pulse has set minimum width, glitches will be filtered.	High-speed asynchronous serial on-chip link; simplex. Simple error detection on frames; acknowledge signal.

Table I
SUMMARISED DIRECT AND INDIRECT SPIKE PROTOCOLS

inaccurate integration in the neuron, but has the advantage of requiring very little area.

Long-range connectivity is similar to existing work on networks using AER, except that a broadcast network is not used. The protocol stack does not make much provision for tolerance of error or saturation, in keeping with the biological influence. However, a greater level of error detection is required in areas that have been more greatly abstracted from biology: the loss or delay of many AER packets - representing many links - would affect many more elements than the loss of a few links between neurons in the brain.

The direct spike connectivity will consist of static wires of different lengths that form "hops" between configurable blocks. Connections to these wires will be governed by configurable switch blocks.

The focus of this system lies with the connectivity between neurons. A simple model neuron is enough; it needs only to exhibit the abstract leaky integrate & fire behaviour to be of service to this focus.

Real-time learning is implemented in the synapses, which modify their "weight" according to a spike-timing dependant plasticity rule. Within a small time window, if a pre-synaptic spike arrives at a synapse before a post-synaptic spike, the strength of the synapse is increased (it is probable that the incoming spike contributed to the post-synaptic spike); if the reverse is true, the strength is decreased. If the two spikes do not coincide within the window period, the strength is unchanged.

The reconfigurable neural network will be evaluated by producing a chip with enough neurons to implement an informative portion of a topographical map, such as an orientation map as seen in the visual cortex of the brain. The performance of the chip will be compared to the same network modelled using a popular software tool, such as NEURON [14].

The design presents several interesting avenues for further investigation. For example, the aggregation of pulse-widths over the dendritic tree of a neuron could be optimised by rescaling at various points to improve the range of the input magnitude and minimise information loss; the saturation point

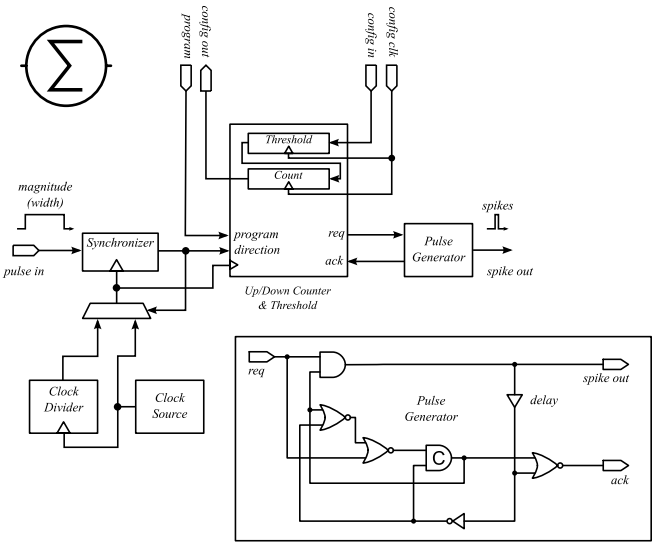

Figure 4. Leaky integrate & fire model

Area	$1145.6\mu m^2$ (90nm: $700\mu m^2$)
Gates	390
Density	$873\,p.\,mm^2$ (90nm: $1429\,p.\,mm^2$)
Spike Period	$4.5ns$
Generated clock frequency	$160MHz$
Max. Spike Rate (threshold=100)	2.35 million p. second

Table II
130NM LI&F NEURON STATISTICS

of the AER link could be raised by collecting spikes together as the rate increases; a global, distributed plasticity process could add and remove links in real time; direct links for spikes could be enhanced by using a limited multiplexing technique.

IV. RESULTS

A "leaky integrate and fire" model of a neuron has been designed using standard cells, simulated, and manufactured at 130nm using the Europractice mini@sic service. The schematic is shown in figure 4, and table II has some area and timing statistics. The neuron is self-contained - except for the configurable threshold register - and is asynchronous relative to the rest of the system, including other neurons. At the heart of the neuron is a 7-bit up/down counter, driven by a generated clock (as some measure of time is required). A clock divider provides a slower clock that is used for decay. The direction of the counter is controlled by the input pulse which also chooses between the original clock and the divided form, producing a quick increment when an input pulse is present and a slow decay when it is not.

The resolution of the counter, and the range of threshold values, is derived from real data on post-synaptic potentials and neural thresholds. For this model, a typical threshold value would be about 100, and each incoming pulse would add between 1 and 3 to the count value.

When the counter reaches the configured threshold value, the counter halts and requests that a spike be emitted by the pulse generator. The asynchronous pulse generator, synthesized using the Petrify tool [15], produces a single stereotypical spike in response to a request. Once the spike is emitted, the counter resets and the neuron starts the cycle again. The width of the pulse is determined by the delay element; the size of this element ($3.4ns$) was chosen to safely produce the shortest possible pulse while still being visible through the IO.

This is a simple neuron model, but it will enable a network to show interesting behaviour as the connectivity is altered and adapts of its own accord. The neuron occupies a small area such that enough neurons will fit on even a small mini@sic chip, when the anticipated size of the other system elements is accounted, such that a network built with these should be able to produce an interesting behaviour.

V. CONCLUSIONS AND FUTURE WORK

A reconfigurable, adaptive neural network system was presented. The system should possess enough real qualities to be of interest to neuroscientists, and provides several interesting avenues for further work. A neuron model and spike generator has been manufactured using 130nm standard cells; the result works as specified and in a reasonable area. A protocol for a hybrid of short- and long-range communication of the spikes produced by these circuits was also presented. A thorough analysis of spike traffic for the purposes of refining the design before manufacture are part of the future work, together with a tool for mapping networks to the system architecture.

REFERENCES

[1] R. Emery and S. Kolbeinson, "Examining the Limitations of an FPGA based Neural Network," Electrical, Electronic and Computer Engineering, Newcastle University, Tech. Rep., 2006.

[2] R. Huang and R. Vemuri, "Analysis and evaluation of a hybrid interconnect structure for FPGAs," in *Computer Aided Design, 2004. ICCAD-2004. IEEE/ACM International Conference on*, 2004, pp. 595–601.

[3] T. Binzegger, R. J. Douglas, and K. A. C. Martin, "Stereotypical Bouton Clustering of Individual Neurons in Cat Primary Visual Cortex," *J. Neurosci.*, vol. 27, no. 45, pp. 12 242–12 254, 2007.

[4] P. D. Roberts and C. C. Bell, "Spike timing dependent synaptic plasticity in biological systems," *Biological Cybernetics*, vol. 87, no. 5-6, pp. 392–403, 2002.

[5] Extended Address Event Representation Draft Standard v0.4. [Online]. Available: http://www.stanford.edu/group/brainsinsilicon/Downloads.htm

[6] P. S. Churchland and T. J. Sejnowski, *The Computational Brain*. Cambridge, Mass. : MIT Press, 1992.

[7] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *Journal of Physiology*, vol. 117, pp. 500–544, 1952.

[8] P. Hagmann, L. Cammoun, X. Gigandet, R. Meuli, C. J. Honey, V. J. Wedeen, and O. Sporns, "Mapping the structural core of human cerebral cortex," *PLoS Biology*, vol. 6, no. 7, 2008.

[9] C. Beaulieu, Z. Kisvarday, P. Somogyi, M. Cynader, and A. Cowey, "Quantitative distribution of gaba-immunopositive and-immunonegative neurons and synapses in the monkey striate cortex (area 17)," *Cerebral Cortex*, vol. 2, no. 4, pp. 295–309, 1992.

[10] C. Holmgren, T. Harkany, B. Svennenfors, and Y. Zilberter, "Pyramidal cell communication within local networks in layer 2/3 of rat neocortex," *J Physiol*, vol. 551, no. 1, pp. 139–153, 2003.

[11] C. Mead, *Analog VLSI and Neural Systems*. Addison-Wesley, 1989.

[12] M. Mahowald, *An analog VLSI system for stereoscopic vision*. Kluwer Academic Publishers, 1994.

[13] Z. Institute of Neuroinformatics, "Silicon Cortex (SCX) Project Page." [Online]. Available: http://www.ini.unizh.ch/ amw/scx/scx.html

[14] M. Hines and N. Carnevale, "NEURON: a Tool for Neuroscientists," *The Neuroscientist*, vol. 7, no. 123-135, 2001.

[15] J. Cortadella, M. Kishinevsky, A. Kondratyev, L. Lavango, and A. Yakovlev, "Petrify: a tool for manipulating concurrent specifications and synthesis of asynchronous controllers," in *XI Conference on Design of Integrated Circuits and Systems*, 1996.