International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:1, 2016

Massively-Parallel Bit-Serial Neural Networks for Fast Epilepsy Diagnosis: A Feasibility Study

Si Mon Kueh, Tom J. Kazmierski

Abstract—There are about 1% of the world population suffering from the hidden disability known as epilepsy and major developing countries are not fully equipped to counter this problem. In order to reduce the inconvenience and danger of epilepsy, different methods have been researched by using a artificial neural network (ANN) classification to distinguish epileptic waveforms from normal brain waveforms. This paper outlines the aim of achieving massive ANN parallelization through a dedicated hardware using bit-serial processing. The design of this bit-serial Neural Processing Element (NPE) is presented which implements the functionality of a complete neuron using variable accuracy. The proposed design has been tested taking into consideration non-idealities of a hardware ANN. The NPE consists of a bit-serial multiplier which uses only 16 logic elements on an Altera Cyclone IV FPGA and a bit-serial ALU as well as a look-up table. Arrays of NPEs can be driven by a single controller which executes the neural processing algorithm. In conclusion, the proposed compact NPE design allows the construction of complex hardware ANNs that can be implemented in a portable equipment that suits the needs of a single epileptic patient in his or her daily activities to predict the occurrences of impending tonic conic seizures.

Keywords—Artificial Neural Networks, bit-serial neural processor, FPGA, Neural Processing Element.

I. Introduction

THE hidden disability of epilepsy is suffered by 1% of the world's population. Of these, 50%-75%, live in countries which are ill-equipped to manage this condition.

This paper outlines the motivation for undertaking Neural Processing Element(NPE) design which can be viewed from two different perspectives. From the author's point of view, this can be seen as a personal goal and a way to tackle his own disability. Additionally, recent developments in technology to detect or predict epilepsy will be explored. An NPE is a basic building block of a complex neural network analogous to a human brain neuron and, in theory it is possible to predict a seizure within the range of 10ms. However, there is the need to reduce the size of the logic circuit, while facing the requirements of hardware complexity, thus bit-serial architecture is chosen to be used as a design choice to be implemented. It would be imperative to explain the basics of bit-serial architecture and the advantages of incorporating this form of architecture in designing this custom neuron design. The bit-serial architecture functions by transferring data, bit by bit, along a wire during a single clock cycle, whilst the state of the art design uses bit-parallel word architecture which sends all input bits along a bus during a single clock cycle. Bit-parallel is faster. However, when the designer requires a lower power design it is far better to use bit-serial architecture as the hardware needed to develop the design are far cheaper in this decade compared to when it was first developed in the 1950's. Further minimisation occurs during the training phase due to the use of the leap-frog algorithm. Once the network is fully trained, with new weight coefficients and the latest network topology, it is configured onto the available existing hardware, described in detail by another paper [1]. It should be noted that in the current research, the training is done entirely offline to reduce complexity of the project. In order to complete this neuron, the basic plan is to have a control and a data path. The building blocks for both paths are listed in Table I. The decoder in this processor would also need to be split into an instruction decoder and a state machine.

S. Kueh and Tom J. Kazmierski are with the Department of Electronics and Computer Science, University of Southampton, Southampton, Hampshire SO17 3SX UK (https://secure.ecs.soton.ac.uk/people/smk1g10, https://secure.ecs.soton.ac.uk/people/tjk)

TABLE I
Basic Building Blocks for Bit-Serial Processor

Control Path	Data Path
PC	Decoder + State Machine
Program Memory	ALU
	Synchronous RAM/Register

The neuron implementation can be derived from two simple equations shown here:

$$u = \sum_{i=1}^{I} w_i x_i \tag{1}$$

In this equation, *x* is the real number and *w* is the probability of any bit is logic '1'. Proper normalization should be enforced due to the range limitation of the system.

$$y = \Phi(u) \tag{2}$$

It should be noted that this paper shows the custom design of a neuron and its testing of functionality. Fig. 4 shows a more fundamental planning of the hardware implementation of a single neuron which is completed and tested for its functionality. The main contribution of this paper is the proposed NPE design which is a purely bit-serial implementation of a complete neuron. The salient features of the NPE are that it is extremely small in size and with very low power consumption.

It is known that epilepsy can be characterized by recurrent seizure spike patterns within an EEG signal. Furthermore, an EEG signal can be broken down into four distinct waveforms: Delta; theta; alpha; beta waves. Those four distinct waves have their own range of frequencies. When an epileptic seizure occurs, delta (0-4Hz) and theta (4-8Hz) waves, which have the characteristics of lower frequency and higher magnitude,

can be observed and recorded using a form of sensory nodes attached to the scalp of the patient [2]. There are several waveform analysis methods available to suit this research once the EEG waveform has been obtained. Such methods include: Short Time Fourier Transform (STFT); Lyapunov Exponent; Wavelet Transform; Autoregressive modelling. In the next subsection, the focus of this research will become clear, as will the reason why the research team chose not to start from the point of waveform acquisition due to limited time and resources.

Compact bit vector (CBV) is used for executing core correlation matrix memory (CMM) operations within this type of processor [3]. The positive aspect of using such representation is an increase in system storage capacity, however, there is a compromise in processing performance. The architecture used is the Advanced Uncertain Reasoning Architecture (AURA)[3].

CMM [3] can be described as a form of binary neural network which is useful in approximate search and match operations that involve massive unstructured datasets. These operations would need to be conducted at high speed. Furthermore, this type of memory is also known as weightless neural network and can be used to implement associative memory structures discussed in later sections.CMM can be considered as two dimensional array M, whose elements can be set at 1, or 0. The two main function of CMM are training and recalling an object. To train an object into the CMM, the input I, and output pattern O, can be expressed. Here the y is the row index and x is the column index. The recall process requires a query input pattern I, which can be expressed in (3):

$$O_x = \sum_{p=1}^{x} (M_{py} AND I_y) \tag{3}$$

The efficient search and retrieval of information efficiently from large lookup tables has not always been a simple computing problem. By using hardware hashing it is possible to minimise the search time. The hardware hashing functionality is based on a combination of a few other functions which include bit folding, exclusive OR and a pseudo random number generator based on the cellular automata (CA) [4]. The literature [3] illustrates the design of a hardware hashing memory structure which gives three distinct advantages over RAM based designs.

1) The memory would be limited to the number of column IDs in a single query.
2) All the valid IDs and totals are stored in the current stack frame.
3) By resetting the stack pointer, the memory is cleared.

This architecture is heavily pipelined and requires a burst, or stream optimised SDRAM controller to efficiently access the selected row data. The details are explained in the literature mentioned above [3] and the main idea expressed is one of its functional units. This ITS can be combined to perform logical operations; and this is necessary for the computing problem expressed in [3].

Another work [5] has also some different form of multiplier. This multiplier is also efficient in certain applications. Such multipliers include the quasi-serial multiplier. This multiplier takes in two different input operands, one serial and one parallel. However, the output is in a serial fashion. This multiplier still requires $2 * N$ to perform a multiplication of 2 N-bit numbers. An approach was used to extend the design to a fully serial design, and currently the design requires $1.5 * N$ cycles to complete the same calculations. The pipelined design will only require N cycle to return a product, but this would, of course, involve much hardware cost.

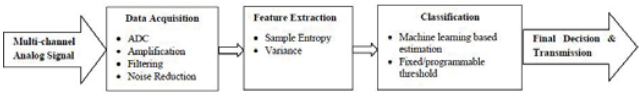

Fig. 1 Automated Seizure Detection based on EEG data flow [6]

A group of researchers used a combination of statistical function of variance as a means of feature extraction [6] and a general ANN classifier as depicted in Fig. 1 to achieve a high detection accuracy (99.18%). Their result is commendable as they achieved a acceptable hardware footprint (44%) for this type of application. However, the research team decided to focus only on building the simplest and smallest possible ANN classifier that would still provide the same functionality shown in Fig. 2. In order to build the simplest and smallest ANN classifier, our paper has given the proof of the basic building block of the NPE needed for this ANN.

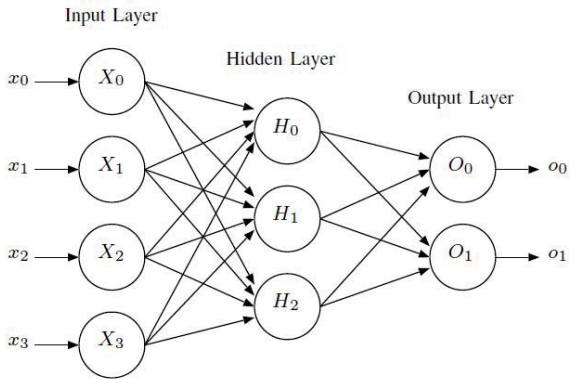

Fig. 2 Multiple Layered Perceptron Design [6]

Through different types of neural network, such as Back Propagation Neural Network (BPN), Probabilistic Neural Network (PNN), Elman Network (EN), Spiking Neural Network (SNN), the most common approaches are based on Multi-Layer Perceptrons (MLP) or Block based Neural Networks (BbNNs). Analysis of EEG waveforms using ANNs of massive complexity, require significant processing power and are difficult to implement in portable equipment suited to the needs of an individual patient.

SNN is different from other forms of ANN as each individual spiking neuron, propagates information by the timing of the neuron rather than the rate of the spikes. The supervised rule known as SpikeProp [7] is also used

for training purposes though the assumption that the internal state of the neuron increases linearly within a small enough region for neuronal firing. Hui Juan Fang et al. [8] proposed several methods to increase the learning rate adaptability. The methods are then tested using four different experiments which are: exclusive OR problems; Iris data-set classification problems; fault diagnosis in Tenesse Eastman (TE); decoding information from a Poisson spike train. It was also found that SNN are mainly used in brain modeling [9, 10]. SNN only need a single spiking neuron for pattern recognition making this a very efficient network topology. A form of hardware implementation was done by using the NVIDIA CUDA as these have the capability to simulate and implement this network.

By using the wireless sensor network (WSNs)[11] technology available, the automated seizure detection system will be capable of patient vital sign monitoring in hospital, and it will also enhance the performance of emergency responders in large disasters. Additionally, WSN improves the quality of life of the elderly in many situations and enables large field studies of human behaviour and chronic diseases. In order to accomplish the above goals, there will exist many challenges involved in choosing the best sensory input device, as well as incorporating our neural network with these wireless devices.

In 1990, some work was done on implementing an array of bit-serial processors and using it to demonstrate the mapping of neural net models. The work is useful for pattern recognition applications and the system was built with a board of 1024 processors using state of the art technology at that time [12]. In the 1990 study, Single Instruction Multiple Date(SIMD) type of architecture was considered appropriate, and the algorithm simulated was written using parallel language, Pascal. It was thought that Pascal would allow the construction of an implementable Bit-Array Processor(BAP). The study clearly stated that precision could be traded for speed by using BAP. However, it was decided to make a compromise between speed and precision.

II. Bit-Serial Neural Processing Element

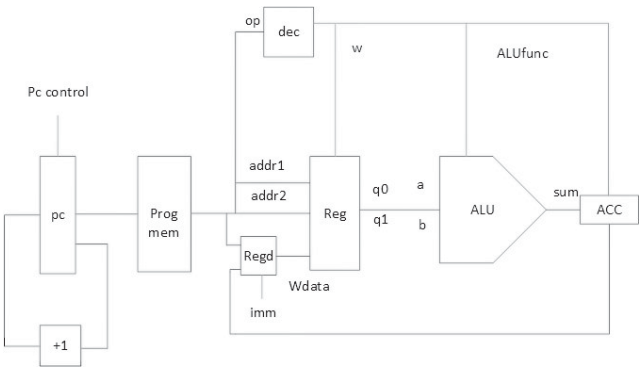

Fig. 3 PicoMIPs Block Diagram

The concept of an extremely small neural processing hardware is based on a bit-serial implementation of a low power consumption processor core [13], known as the picoMIPS. The general architecture of this processor is shown in Fig. 3. In hindsight, a bit-serial architecture seems to provide a better compromise between performance speed and hardware cost, which better serves the purpose of this custom NPE design.

Fig. 4 illustrates and a further explanation of this is given in the following text. The PC is a program counter connected to the program memory which consist of four simple instructions that cycle through multiplication and addition, until the all input data has been processed. The bit-serial Arithmetic Logic Unit (ALU) of the NPE consists of a full adder and an accumulator. The input data is multiplexed and parallel loaded into the register Mreg. Similarly, the register Qreg data is fed from the Wmem, the weight memory. Multiplication results are saved in Res1 which is a double length register. All these data processing functions are controlled by the controller, which consists of two main parts, the instruction decoder and a state machine. The NPE also contains a lookup table which consists of a fixed point representation of the hyperbolic sigmoid function used in this particular neuron, from which the NPE output data y is obtained.

The instruction format for this custom design has been designated to be a 17 bit instruction code. The break down of this format is included in Fig. 6.

The terms used in the block diagram on Fig. 5 are explained below. In the loading state, the data from Wmem and data from the selected input X data are saved into Qreg and Mreg respectively. The bit Q0 of the data in Qreg is checked. If it is logic '1', the adding state is enabled. In the adding state, the addition of single bits through a series of counter, N2 is performed. If it is logic '0', the shifting state is enabled. A regular shift operation is performed. Once the counter linked to N is down to 0, the operation is stopped. The full answer of the multiplication is saved into a register Res1 once a READY signal is activated during the stopped state. When the START ADD signal is activated, the Adding sum state will perform the addition of Res1 with the accumulator value which is saved each time. When N reaches 0 again, Stop sum state is called and a output value can be read from the lookup table which corresponds to the y data. If a START MUL signal is activated, the operation will restart from the loading state.

The bit serial multiplication logic (Fig. 8) was embedded in the NPE custom design in order to mimimize the overall hardware size, as such the individual components in Figs. 8 and 7 yielded circuits amounting to 11 and 5 Logic Elements respectively, on an Altera Cyclone IV FPGA. Thus, the total size of the NPE of 16 logic elements is indeed efficient and cost effective as, despite its extremely small size, it has the capability to execute the Shift-and-Add multiplication algorithm. The registers Mreg, Qreg and ACC are basic shift registers. Each of these shift registers have the parallel load functionality which is essential to this NPE design. Through the process of shifting, the calculations of arbitrary binary length can be performed serially within the single full adder. The final multiplication result is stored in register Res1.

Through thorough testing, it was found that a 4-bit accuracy of weights and synapse inputs are sufficient to obtain correct neural operation. It should be stressed, however, that the

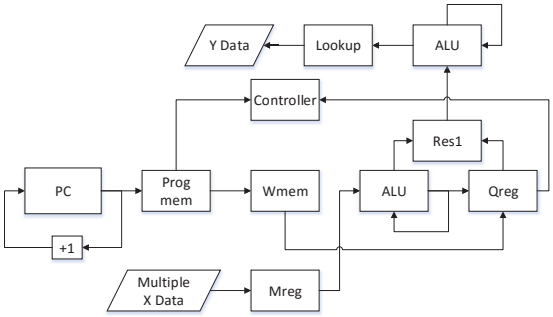

Fig. 4 Processor Block Diagram

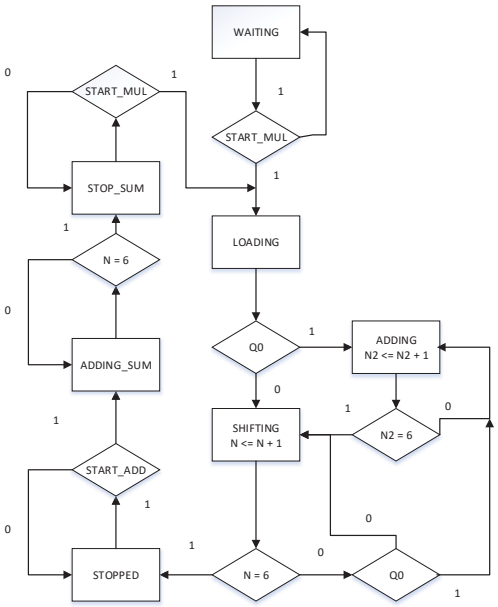

Fig. 5 Neural Operation State-Diagram

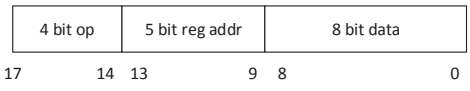

Fig. 6 Instruction Format for Bit-Serial Processor

proposed NPE design will work for an arbitrary number of bits with no changes to the basic design, other than the register length and the size of the look up table which implements the sigmoid function. Sample simulation waveforms are shown in Figs. 9 and 10.

III. Conclusion

In conclusion, experiments with a bit-serial neuron confirm that an extremely small logic system, can successfully implement a neuron suitable to be used in a massively parallel environment. The sample four-input system presented here, as a case study, uses a bit-serial multiplier with only 16 Cyclone IV FPGA logic elements, a lookup table (30 logic elements) and a controller (55 logic elements) which can drive multiple NPEs in a vector implementation. Further work will involve the testing of complex neural networks based on NPEs, and their application to epilepsy diagnostics. The proposed compact NPE design will allow for the construction of complex hardware ANNs that can be implemented in portable equipment to suit the needs of a individual epileptic patients in their daily activities, in order to predict impending tonic conic seizures.

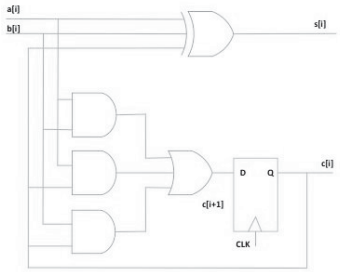

Fig. 7 Circuit for the Bit-Adder in this Bit-Serial Processor

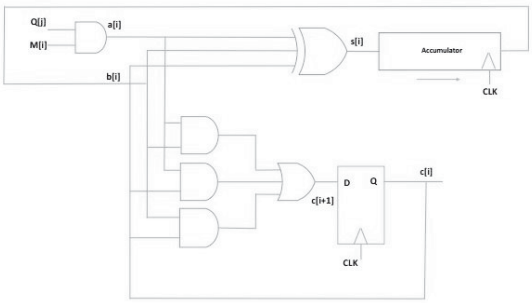

Fig. 8 Circuit for the Bit-Serial Multiplier in this Processor

Acknowledgment

Thanks are extended to colleagues in ESS at University at Southampton who have provided valuable insight into certain

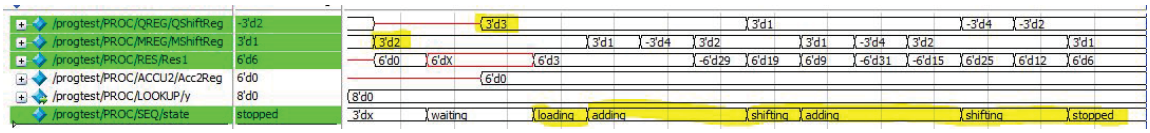

Fig. 9 Single Waveform of the Multiplication Process of the Nueron

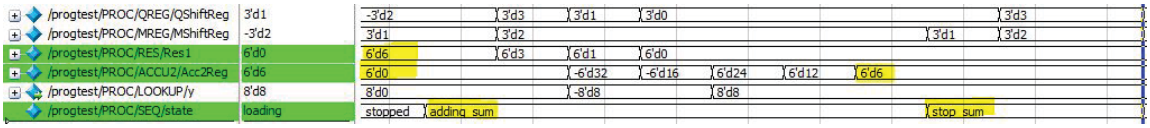

Fig. 10 Single Waveform of the Addition Sum and Lookup Process of the Nueron

design issues that we faced during the past year. Grateful thanks also go to the technicians who gave their support in providing the necessary tools for this research to be completed.

REFERENCES

[1] Y. Chen and W. du Plessis, "Neural network implementation on a fpga," in *Africon Conference in Africa, 2002. IEEE AFRICON. 6th*, vol. 1, Oct 2002, pp. 337–342 vol.1.

[2] Akin, M. and Arserim, M. A. and Kiymik, M.K. and Turkoglu, I., "A new approach for diagnosing epilepsy by using wavelet transform and neural networks," in *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE*, vol. 2, 2001, pp. 1596–1599 vol.2.

[3] M. Freeman and J. Austin, "Designing a binary neural network co-processor," in *Digital System Design, 2005. Proceedings. 8th Euromicro Conference on*, Aug 2005, pp. 223–226.

[4] D. Roy Chowdhury, I. Gupta, and P. Pal Chaudhuri, "A low-cost high-capacity associative memory design using cellular automata," *Computers, IEEE Transactions on*, vol. 44, no. 10, pp. 1260–1264, Oct 1995.

[5] A. Shafer, L. Parker, and E. Swartzlander, "The fully-serial pipelined multiplier," in *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, Nov 2011, pp. 1817–1822.

[6] M. Saleheen, H. Alemzadeh, A. Cheriyan, Z. Kalbarczyk, and R. Iyer, "An efficient embedded hardware for high accuracy detection of epileptic seizures," in *Biomedical Engineering and Informatics (BMEI), 2010 3rd International Conference on*, vol. 5, Oct 2010, pp. 1889–1896.

[7] T. Matsumoto, Y. Shin, H. Takase, H. Kawanaka, and S. Tsuruoka, "A learning method for extended spikeprop without redundant spikes #x2014; automatic adjustment of hidden units," in *Soft Computing and Intelligent Systems (SCIS), 2014 Joint 7th International Conference on and Advanced Intelligent Systems (ISIS), 15th International Symposium on*, Dec 2014, pp. 1465–1469.

[8] H. Fang, Y. Wang, and J. He, "Spiking neural networks for cortical neuronal spike train decoding," *Neural Computation*, vol. 22, no. 4, pp. 1060–1085, April 2010.

[9] Izhikevich, E.M., "Simple model of spiking neurons," *Neural Networks, IEEE Transactions on*, vol. 14, no. 6, pp. 1569–1572, Nov 2003.

[10] Cheng-Wen Ko, Hsiao-Wen Chung , "Automatic spike detection via an artificial neural network using raw EEG data: effects of data preparation and implications in the limitations of online recognition," *Clinical Neurophysiology - 1*, vol. 111, no. 3, pp. 477–481, March 2000.

[11] J. Ko, C. Lu, M. Srivastava, J. Stankovic, A. Terzis, and M. Welsh, "Wireless sensor networks for healthcare," *Proceedings of the IEEE*, vol. 98, no. 11, pp. 1947–1960, Nov 2010.

[12] B. Svensson and T. Nordstrom, "Execution of neural network algorithms on an array of bit-serial processors," in *Pattern Recognition, 1990. Proceedings., 10th International Conference on*, vol. ii, June 1990, pp. 501–505 vol.2.

[13] T. J. Kazmierski and C. Leech, "Synthesis of application specific processor architectures for ultra-low energy consumption," in *Small Systems Simulation Symposium*, February 2014. [Online]. Available: http://eprints.soton.ac.uk/366668/