A Time Multiplexing Architecture for Inter-Neuron Communications

Fergal Tuffy[1], Liam McDaid[1], Martin McGinnity[1], Jose Santos[1], Peter Kelly[1], Vunfu Wong Kwan[2], John Alderman[2]

[1] University of Ulster, Intelligent Systems Engineering Laboratory, School of Computing and Intelligent Systems, Faculty of Engineering, Magee Campus, Northland Road, Derry, N. Ireland, BT48 OLY
{f.tuffy, lj.mcdaid, tm.mcginnity, ja.santos, pm.kelly}@ulster.ac.uk

[2] Tyndall National Institute, Lee Maltings, Prospect Row, Cork, Rep. of Ireland
{vunfu, john.alderman}@tyndall.ie

Abstract. This paper presents a hardware implementation of a Time Multiplexing Architecture (TMA) that can interconnect arrays of neurons in an Artificial Neural Network (ANN) using a single metal wire. The approach exploits the relative slow operational speed of the biological system by using fast digital hardware to sequentially sample neurons in a layer and transmit the associated spikes to neurons in other layers. The motivation for this work is to develop minimal area inter-neuron communication hardware. An estimate of the density of on-chip neurons afforded by this approach is presented. The paper verifies the operation of the TMA and investigates pulse transmission errors as a function of the sampling rate. Simulations using the Xilinx System Generator (XSG) package demonstrate that the effect of these errors on the performance of an SNN, pre-trained to solve the XOR problem, is negligible if the sampling frequency is sufficiently high.

1 Introduction

Biological research has accumulated an enormous amount of detailed knowledge about the structure and functionality of the brain. It is widely accepted that the basic processing units in the brain are neurons which are interconnected in a complex pattern, communicate through pulses and use the timing of the pulses to transmit information and perform computations [1-3]. Significant research has focused on "biological equivalent" neural network models that can be implemented in hardware and used to inspire new techniques for real time computations [4-6]. However, the standard network topologies employed to model the biological networks are proving difficult to implement in hardware, even for moderately complex networks. Existing inter-neuron connection schemes are achieved through metallization and thus as the size of the neuron array increases there is a rapid increase in the ratio of metal to device area which eventually self-limits the network size [7-8]. Given that the density of the interconnect pathways in the human brain is of the order of 10^{14} [9], it is inconceivable

that existing interconnect technologies will even remotely approach this order of magnitude and thus new approaches need to be explored.

This paper presents a novel Time Multiplexing Architecture (TMA) as a possible solution to the interconnect problem for Spiking Neural Networks (SNNs) in hardware. A single bus wire is used to transmit the signals between neuron layers, where timing is guaranteed by using a clocking system that is synchronized to a "global" clock. This implementation removes the requirement of dedicated metal lines for every synaptic pathway and therefore a significant saving in the silicon surface area is achieved. Section 2 of this paper discusses the TMA while section 3 highlights results that verify the approach. Errors in spike timing "across the TMA" due to the sampling frequency are investigated in section 4 where a simple SNN is initially trained, using a supervised approach, to solve the XOR problem. Using the Xilinx System Generator (XSG) package the output firing times that results from the TMA architecture are compared with those obtained when conventional metal interconnect is used, and from the subsequent analysis it is clear that the sampling frequency must be at least twice the minimal sampling frequency: note the minimal sampling frequency is set by the duration of the spike and the number of neurons in the sampled layer. Section 5 presents a quantitative analysis underpinning the scalability of the TMA and section 6 makes concluding remarks.

2 Time Multiplexing Architecture (TMA)

This section presents a novel inter-neuron communication architecture where biologically compatible neuron spikes are represented as digital pulses and connectivity between neuron layers is achieved using the TMA. Figure 1 shows a two layer neural network fragment containing two input neurons, I_0 and I_1, and one output neuron, O_0. The sampling circuit to the left of the bus wire contains two D-type latches in a daisy chain configuration where one of the latches is preset to logic 1, prior to the application of the clock, C_K. Effectively the clock input, C_K, rotates a logic 1 between the two latches, switching on transistors *M1* and *M2* sequentially: *M1*, *M2*, *M3* and *M4* are *n*-channel enhancement mode MOSFETs. This sampling to the left of the bus wire is repeated on the right of the bus wire. Consider the case where the input neuron, I_0, fires a spike, {0, 1, 0}, of duration T_P which forms the input to the drain, *D*, of *M1*. The gate terminal, *G*, of *M1* is controlled by the *Q* output of a D-type latch and when *Q* is asserted, I_0 is sampled and a logic 1 is placed on the bus wire: note that the gate of *M2* will be held at logic 0 while *M1* is on (sampling). Because both sampling circuits are driven from the same clock input, C_K, the bus line is now sampled by *M3* ensuring that the pulse from I_0 is directed to the correct synapse, Synapse 1. I_1 will be sampled directly after I_0 whereby *M2* and *M4* will be turned on by the sampling circuits allowing the pulse from I_1 (if I_1 has fired) to reach Synapse 2. Clearly the sampling frequency is a function of the number of neurons in the sampled layer and the duration of the spike pulse. In a layer of *n* neurons which are sampled sequentially, it can be shown that the minimum sampling frequency F_S (Hz) is given by,

$$Fs = \frac{n}{T_P} \tag{1}$$

The authors are aware that pulse transmission errors can exist between the time a neuron in one layer fires and the time required for this pulse to be observed at the synaptic inputs associated with the neurons in the subsequent layer. These are caused by the sampling circuitry operating in a synchronous mode while all the neurons that are sampled will fire in an asynchronous mode. Pulse transmission errors and their effect on a pre-trained SNN are investigated in section 5.

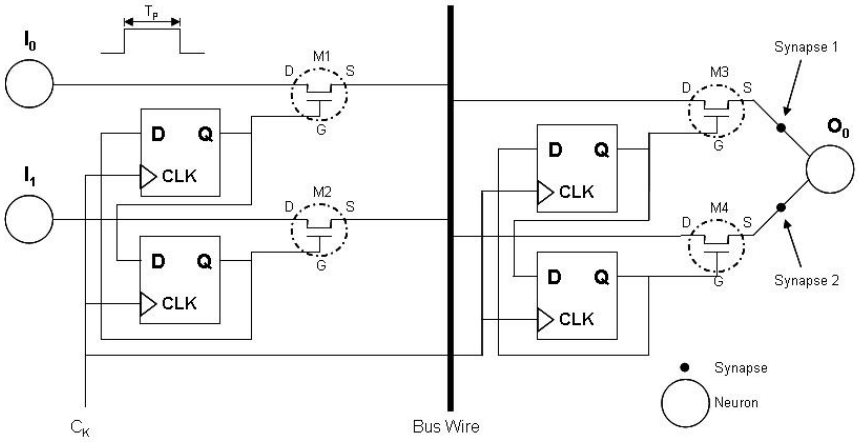

Fig. 1. TMA for a 2-input 1-output SNN

3 Simulation Results

The proposed TMA was simulated using the Mentor Graphics mixed signal simulation package, System Vision Professional. Figure 2 represents the layout used in the simulation where the SNN has four input neurons, I_0-I_3, and two output neurons, O_0, O_1 (note that this architecture is modified from that shown in figure 1 in that the MOSFET transistors at the input to each synapse are replaced by D-latches, *D13-D20*). It will be shown later that in order to reduce pulse transmission errors it is necessary to sample at a rate that is in excess of the minimum sampling frequency defined by equation (1). However, "gating" these high frequency pulses using MOSFETs causes glitches at the input to the synapses. This problem is avoided by the additional layer of D-latches, *D13-D20*.

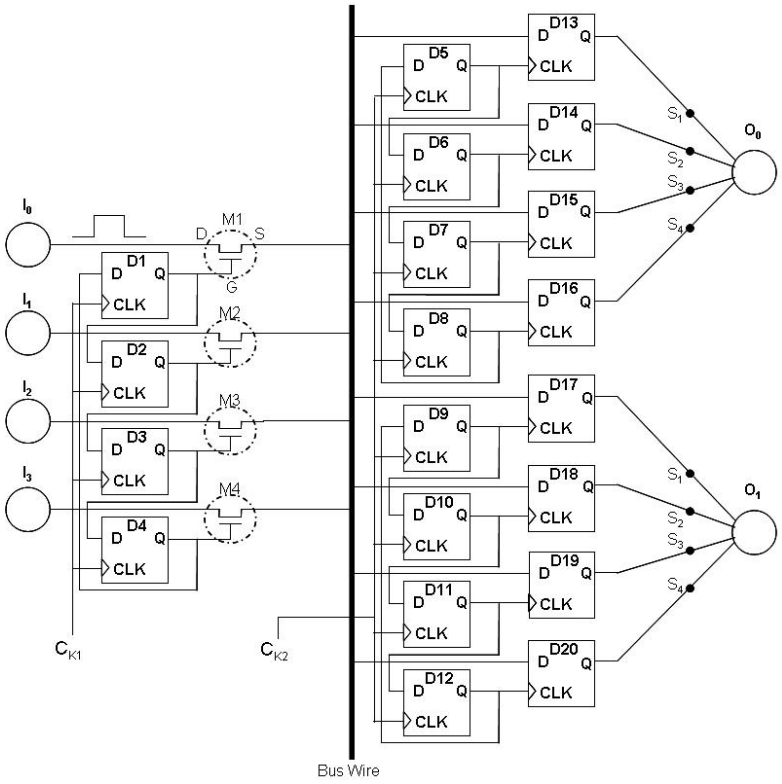

Fig. 2. TMA system layout for 4-input, 2-output NN

In the simulations, as shown in figures 2 and 3, the pulse length for all neurons, T_P, was set to 1ms and since there are four input neurons, the sampling frequency was calculated from equation (1) to be 4KHz. Because M_1-M_4 are not ideal the transitions from logic 1 to logic 0, and vice versa, are not instantaneous. Therefore, to avoid any overlap between the turn on transient of one transistor and the turn off of another a two phase clock system is used where one clock C_{K1} operates on the sampling circuit to the left of the bus wire and another clock C_{K2} operates on the sampling circuit to the right: note that C_{K1} and C_{K2} are in anti-phase but operate at the same frequency (8 KHz), as shown in figure 3(a). Figure 3 (b) shows random firing of neurons $I_0 - I_3$, and their arrival times at the appropriate synapses. It can be seen that there exists a time error ∂t_{Io} between I_0 firing and the arrival time of the pulse at the appropriate synapses. Note that from figure 3(b) similar errors exist for all pulses and therefore while TMA provides inter-neuron communication, transmission errors exist. The following section analyses these errors to determine their effect on the dynamics of a pre-trained SNN.

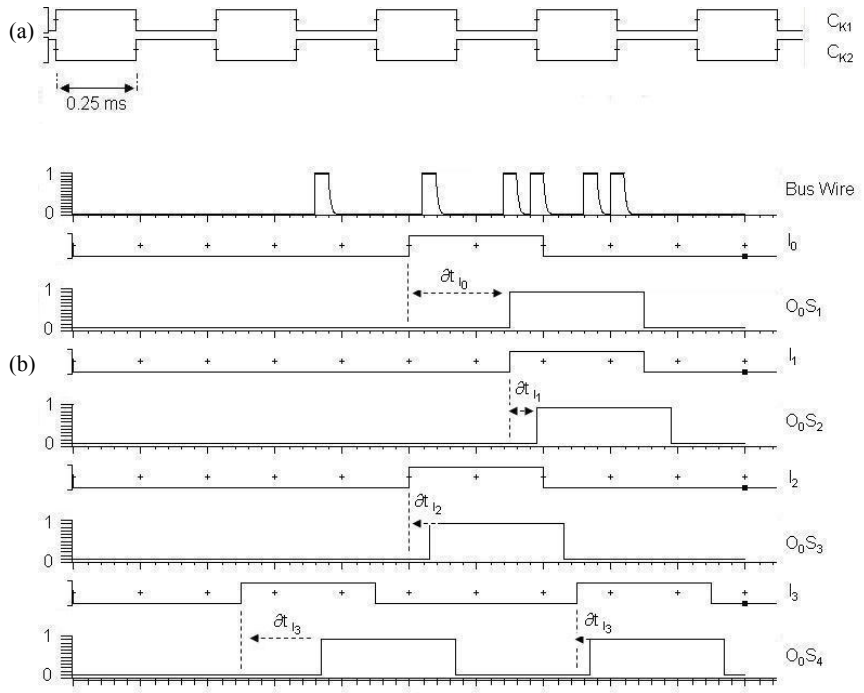

Fig. 3. (a) Timing diagram where the clock signal to the output sampling D-latches, *D5-D12*, is delayed by 0.25ms, quarter of the sampling pulse period. (b) bus wire signals caused by random firing neurons I_0 - I_4, and O_0S_1 - O_0S_4, show the time of arrival of pulses at the appropriate synapses

4 Xilinx System Generator implementation

In order to investigate pulse transmission errors both conventional interconnect and the TMA were used to interconnect neuron layers in a SNN topology that has been pre trained to solve the benchmark XOR problem [10]. Both topologies were simulated using the XSG toolset from Xilinx [11], as illustrated in figure 4. The SNN was trained off-line by an Evolutionary Strategy (ES) [10].

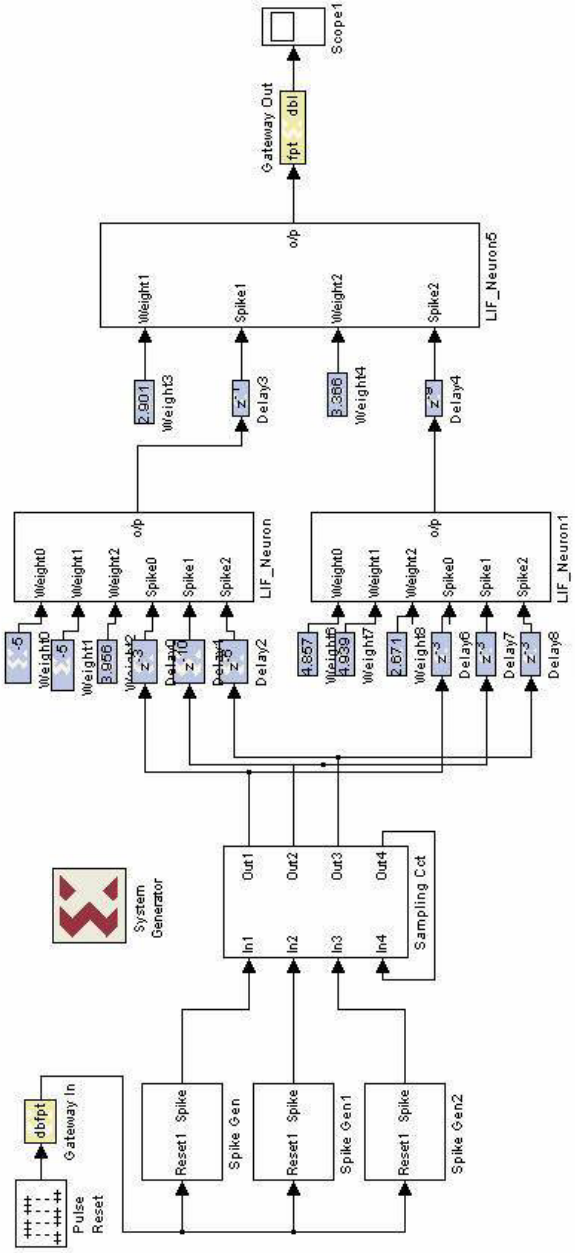

Fig. 4. SNN for XOR problem containing TMA in XSG simulator

Table 1 shows three input neurons where neuron 1 is a biasing neuron [10], which fires at 0ms, and neurons 2 and 3 provide the conventional 2 inputs for the XOR truth table. Note that column four is the post-trained firing times of the output neuron where the simulation used conventional metal interconnect. Columns 5 and 6 are the firing times of the same output neuron where the simulation used the TMA and clearly transmission errors are appreciable if the minimum sampling rate is used (column 4), this is the worst case firing times. However, if the sampling frequency is increased to $2*F_S$, then satisfactory agreement between column 6 and column 4 is obtained. Therefore, for effective pulse transmission without significant error the sampling frequency must be maintained such that

$$F_S \geq 2 \frac{n}{T_P} \tag{2}$$

Table 1. XOR dataset simulation results with and without TMA. Table includes 3 neuron inputs where neuron 1 is a biasing neuron and neurons 2 and 3 provide the conventional 2 inputs for the XOR truth table. The trained firing times (without TMA) is compared with the actual firing times for a sampling frequency of F_S and $2*F_S$.

Neuron 1 Firing Time	Neuron 2 Firing Time	Neuron 3 Firing Time	Firing time without TMA (ms)	Firing Times with TMA F_S (ms)	Firing Times with TMA $2F_S$ (ms)
0	0	0	14	15	15
0	0	6	20	15	21
0	6	0	20	15	21
0	6	6	14	22	15

5 TMA Scalability

To demonstrate the scalability of the TMA consider a network where we have n input neurons and m output neurons. It should be noted that the number of input neurons n, afforded by the TMA technique, is a function of the maximum possible operating frequency of the global clock while the theoretical limit on the scale of an $n*m$ network is determined by the physical size of the sampling circuits. To estimate n, consider a 1ms spike and assume a realistic sampling frequency F_S of 1GHz [12]. Equation (1) is then used to predict the number of neurons that can be accommodated on the input layer which equates to approximately one million. Even if we sample at $2*F_S$ to minimise pulse transmission errors, then equation (2) predicts an upper limit for n of half a million. This is an improvement over what is currently achieveable [13]. However, it is clear that the scale of a SNN implemented using the proposed TMA is unlikely to be severely limited by the frequency of the global clock, rather scaleability will be limited by the real estate occupied by circuitry, and the following is an estimate of this limit.

Consider again the case where we have n input neurons and the number of output

neurons, m, is allowed to increase. If we assume a fully connected feedforward network then the number of associated synapses increases according to the product nm. To calculate the limit on the network size, an estimate of the area consumed by the associated sampling circuits is required. Given that the sampling circuit is dominated by n D-type latches in the transmitting layer and $2*n*m$ D-type latches in the receiving layer, then we can write that the total area, A_T, occupied by the sampling circuitry is given by

$$A_T = \left(n + 2nm\right) A_D \approx 2nmA_D \tag{3}$$

for large m where A_D is the area of a D-type latch. It has been reported for a 0.18-μm process technology that a D-type latch can be designed to occupy a silicon area of approximately $4\mu m^2$ [14] and if the area occupied by sampling circuitry is restricted to 10% of the total chip area (assumed to be $1cm^2$), then a simple calculation (taking n = m) predicts that the TMA approach permits over three thousand neurons to be fabricated in each layer using a planar submicron process. For a fully connected feedforward NN this equates to 9 million synapses. While this is a significant improvement from what is reported elsewhere [13], it will be further enhanced as technology improvements continue [15]. Furthermore, given that the interconnect density will be substantially reduced by the proposed TMA then the real estate given over to the sampling circuitry is expected to be in excess of the 10% estimate. Hence, the above estimate is viewed as conservative and it is expected that the proposed TMA approach will advance the synaptic density even further.

6 Conclusion

This paper has proposed a novel time sampling architecture for the hardware implementations of SNNs. This work has shown that the optimal sampling frequency depends on the number of neurons in the sampled layer and the duration of the "digital spikes" they emit. However, with on-chip clock frequencies typically in the GHz range, the limitations placed on this approach by the sampling frequency are negligible. The TMA has been verified using the Mentor System Vision Pro software package and issues such as pulse transmission errors have been investigated using the XSG platform. It has been shown that these errors can be minimized by ensuring that the sampling frequency is maintain to at least twice the minimum sampling frequency ($2*F_S$). The authors wish to note that this paper has demonstrated the potential of the TMA for inter-neuron communication where the target implemented for this approach is a mixed signal Application Specific Integrated Circuit (ASIC) layout, given the asynchronous firing nature of neurons. Moreover, the authors are confident that if this approach is optimized in terms of minimal area circuitry and timing issues are addressed for large implementations, then this approach has the potential to implement well over a million inter-neuron pathways using a very simple and compact sampling architecture. Future work shall involve a comparative analysis with alternative interconnect strategies such as Address Event Decoding (AED).

7 Acknowledgment

This work was part supported by the European Union under the Irish Government National Development Plan, as part of a Higher Education Authority North-South program for collaborative research project – Interwave.

The authors would also like to thank Simon Johnston and Brendan Glackin, at the Intelligent Systems Engineering Laboratory (ISEL), for assistance in the XSG and Xilinx FPGA simulations and numerous fruitful discussions.

References

1. Roche, B., McGinnity, T.M., Maguire, L.P., McDaid, L.J.: Signalling Techniques and their Effect on Neural Network Implementation Sizes", Information Sciences 132, pages 67-82, NH Elsevier, 2001
2. Murray, F., and Woodburn, R.: The Prospects for Analogue Neural VLSI, International Journal of Neural Systems, Vol. 8, No. 5 & 6, pages 559-579, Oct/Dec. 1997
3. Liu, S.C., Kramer, J., Indiveri, G., Delbruck, T., Burg, T., and Douglas, R.: Orientation-selective VLSI Spiking Neurons, Neural Networks, Special Issue on Spiking Neurons in Neuroscience and Technology , Vol. 14, Issues 6-7, pages 629-643, July 2001
4. Diorio, C., Hsu, D., and Figueroa, M.: Adaptive CMOS: from biological inspiration to systems-on-a-chip, Proceedings of the IEEE, Vol. 90, Issue 3, pages 345 – 357, March 2002
5. Goldberg, D.H., Cauwenberghs, G., Andreou, A. G.: Probabilistic Synaptic Weighting in a Reconfigurable Network of VLSI Integrate-and-Fire Neurons, Neural Networks, Vol. 14, no. 6–7, pages 781–793, Sept 2001
6. Maass, W.: Computation with Spiking Neurons: the Handbook of Brain Theory and Neural Networks, MIT Press, 1998.
7. Noory, B., Groza, V.: A Reconfigurable Approach to Hardware Implementation of Neural Networks, IEEE CCECE 2003. Canadian Conference on Electrical and Computer Engineering, pages 1861 - 1864 Vol. 3, 4-7 May 2003
8. Chun, L., Shi, B., Chen, L.: Hardware Implementation of an Expandable On-chip Learning Neural Network with 8-Neuron and 64-Synapse, TENCON '02. Proceedings 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering, Vol. 3, pages 1451 – 1454, 28-31 Oct. 2002
9. Miki, T., Editor: Brainware: Bio-Inspired Architectures and its Hardware Implementation, World Scientific Publishing Co. Ltd, 2001.
10. Johnston, S.P, Prasad, G., Maguire, L. P., McGinnity, T. M.: Comparative Investigation into Classical and Spiking Neuron Implementations on FPGAs, 15th International Conference on Artificial Neural Networks, ICANN 2005, Part 1: pages 269-274, 11-15 Sept. 2005
11. http://www.xilinx.com/ise/optional_prod/system_generator.htm
12. Tu, S.-W., Jou, J.-Y., Chang, Y.-W.: RLC Coupling-Aware Simulation for On-Chip Buses and their Encoding for Delay Reduction, 2005 ISCAS IEEE International Symposium on Circuits and Systems, 23-26 May 2005 Page(s):4134 - 4137 Vol. 4
13. Chicca, E., Badoni, D., Dante, V., D'Andreagiovanni, M., Salina, G., Carota, L., Fusi, S. and Del Giudice, P.: A VLSI Recurrent Network of Integrate and Fire Neurons Connected by Plastic Synapses with Long Term Memory", IEEE Trans. on Neural Networks, Vol.14, No.5, Sept. 2003

14. Yamaoka, M., Osada, K., Ishibashi, K.: 0.4-V Logic-Library-Friendly SRAM Array Using Rectangular-Diffusion Cell and Delta-Boosted-Array Voltage Scheme, IEEE Journal of Solid-State Circuits, Volume 39, Issue 6, June 2004 Page(s):934 – 940
15. Naeemi, A., Meindl, J.D.: Monolayer Metallic Nanotube Interconnects: Promising Candidates or Short Local Interconnects, IEEE Electron Device Letters, Volume 26, Issue 8, Aug. 2005 Page(s):544 - 546