Reduced Interconnects in Neural Networks Using a Time Multiplexed Architecture Based on Quantum Devices

Peter M. Kelly[1], Fergal Tuffy[1], Valeriu Beiu[2], and Liam J. McDaid[1]

[1] Intelligent Systems Research Centre, University of Ulster
Northland Road, Derry, N Ireland, BT48 0RE
[2] College of Information Technology, Center for Neural Inspired Nano Architectures
United Arab Emirates University, P.O. Box 17551, Al Ain, UAE
{pm.kelly,f.tuffy,lj.mcdaid}@ulster.ac.uk, vbeiu@uaeu.ac.ae

Abstract. The interconnection problem associated with large scale hardware-based neural networks is well known. A time multiplexed neural network architecture using silicon based quantum devices with MOS/CMOS devices is described and shows significant increased functional density compared to conventional devices.

Keywords: Quantum device, resonant tunneling device (RTD), time multiplexed architecture (TMA), neural network (NN), interconnects.

1 Introduction

As hardware based Artificial Neural Networks (ANNs) are scaled towards the very large neuron numbers associated with even the most rudimentary biological systems major interconnection problems arise. Although the semiconductor industry is rapidly reducing transistor sizes the same does not apply to interconnect, where delay and energy dissipation are significantly greater than that of transistors fabricated in the same process. Thus in fully interconnected highly parallel architectures such as ANNs there will undoubtedly be issues in terms of latency, energy dissipated and signal integrity particularly at high frequencies. These inherent problems act to limit the scalability of the architectures, especially when it is understood that interconnection lines scale exponentially with linear increases in neuron numbers. The simple NN shown in Fig. 1 illustrates the problem that interconnection creates. Each time a neuron is added there is a requirement for another vertical bus and extra neuronal connections. Thus the computational advantages of parallelism in hardware based ANNs are quickly lost as the circuits are scaled. Various schemes such as Pulsed Wave Interconnect (PWI), Address Event Decoding (AED), and Multiple Valued Logic (MVL), have attempted to reduce the interconnection overhead [1–3]. However these have deficiencies which prevent them from scaling towards biologically plausible architectures. On the other hand recent research conducted by the authors has shown that Time Multiplexed Architectures (TMAs) show significant potential for reducing interconnection resources in NN applications [4].

A. Schmid et al. (Eds.): Nano-Net 2009, LNICST 20, pp. 242–250, 2009.

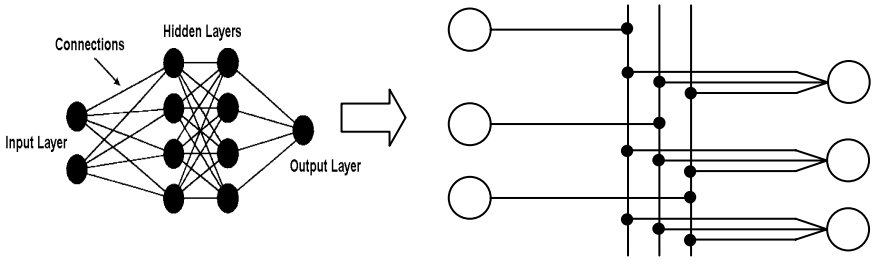

Fig. 1. Neural network interconnection implemented in two dimensions

This paper sets out to show how SiGe based Inter-band Tunneling Diodes (ITDs) can be combined with MOS/CMOS devices to create a high speed TMA design suited to ANN circuits. When the TMA is designed using the ITD/CMOS based circuits there is a very significant increase in functional density over the same architecture implemented in conventional circuit components.

2 TMA in Neural Networks

In previous work by the authors TMAs have been shown to significantly improve the ratio of interconnect to useful device area [4]. These previous designs used conventional CMOS/NMOS technology and significantly increased device count for the TMA control. These extra devices can detract from the impact in overall interconnect reduction as the devices are not directly associated with the actual processing carried out by the ANN.

2.1 Implementation Using D-Type Flip-Flops

Recent work by the authors has produced an architecture that makes use of D-type flip-flops in a daisy chain arrangement to sequentially switch the inputs of the first layer of an ANN to the neurons in the next layer. The diagram shown in Fig. 2 illustrates the approach. Although the advantages at the small scale are not obvious, it has been shown that for large numbers of neurons the circuit has a much higher functional density compared to conventional metal interconnections. In the architecture shown input signals are transmitted through *n*-channel enhancement mode MOSFETs acting as switches. For this arrangement spike signals are characterized as digital (0 – 1 – 0) pulses and routed between neuron layers using the TMA synchronized to a global clock. The TMA system can be best visualized with the aid of the two layer ANN fragment shown in Fig. 2 which has two input neurons, I_1 and I_2, and one output neuron, O_1. Firstly we shall consider the portion of the circuit to the left of the bus wire containing two D-latches (latch 1 and 2), which are configured in a daisy chain arrangement, and the two MOSFET transistors, M1 and M2. Initially one of the D-latches is preset to logic 1, before the clock signal C_K is applied. Thereafter C_K rotates a logic 1 between the two D-latches, alternatively switching M1 and M2, and in doing so I_1 and I_2 are sampled sequentially. The architectural arrangement to the left hand

side of the bus wire is also repeated on the right of the wire to allow the bus itself to be sampled. For the layout of Fig. 2, consider the case where the input neuron, I_1, generates a pulse, (0 – 1 – 0), lasting for a time T_P, and this pulse is connected to the drain terminal, D, of M1. Note that the Q output of each D-latch is used to control the gate terminal of the associated transistor. When the Q output of latch 1 is asserted, I_1 will be sampled and because it is in the firing state, a logic 1 is transferred to the bus wire; note that the gate of M2 will be held at logic 0 while M1 is on (sampling), to ensure that only one neuron can be sampled at any one time.

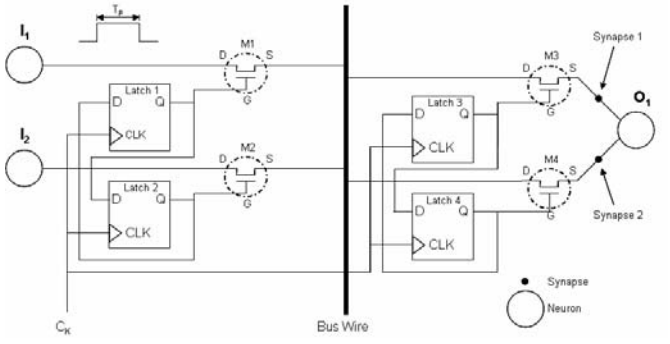

Fig. 2. TMA for a 2-input 1-output ANN

The sampling circuits on the left and right hand side of the bus line driven from C_K, M3 and M1, are on concurrently. This ensures that the pulse signal from I_1 is directed to the correct synapse on the output neuron O_1 (synapse 1). Neuron I_2 will be sampled immediately after I_1 whereby M2 and M4 will be turned on by the sampling circuits allowing the pulse from I_2 (if fired) to reach synapse 2. Clearly the sampling frequency is governed by the number of input neurons in the sampled layer and also the duration of their pulses. It can be shown that the minimum sampling frequency F_S (Hz) in a system of n-input neurons is given by:

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

So if we take a pulse duration of 1ms for all four neurons the sampling frequency calculated from eq. (1) would be set at 4KHz. The ANN layout, shown in Fig. 3, was arranged with four input neurons, I_0–I_3, and two output neurons, O_0–O_1. In a modification from the architecture in Fig. 2, the transistors at the input to each synapse have been replaced by D-latches D13–D20 because the "gating" of these high frequency pulses causes glitches at the input to the synapses. Because M1-M4 are not ideal and have an inherent rise and fall time, the transitions from logic 1 to logic 0, and vice versa, are not instantaneous. To take account of this, a two phase clocking system is used where one clock C_{K1} is used to drive the input sampling circuitry to the left of the bus wire, and a second clock C_{K2} is utilized to trigger the sampling circuit to the right of the bus wire; note that C_{K1} and C_{K2} are in anti-phase but operate at the same frequency.

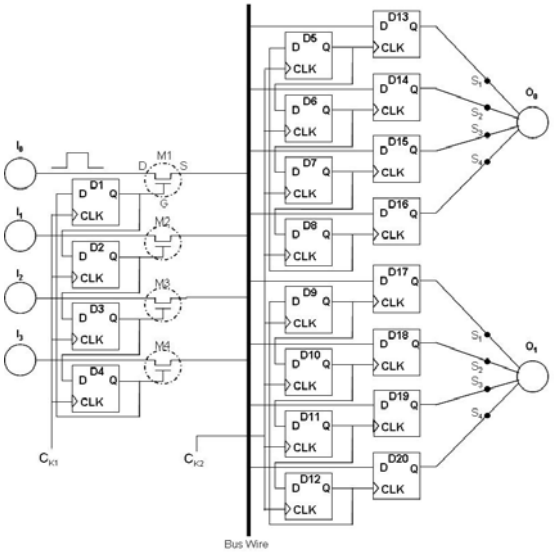

Fig. 3. A TMA implementation of a 4-input 2-output ANN

3 TMA Implementation Using ITD Based Latches

The authors have demonstrated that the method shown in Section 2 significantly reduces interconnect, increases the functional density and permits scaling beyond the limits of conventional interconnect [4]. A further improvement can be achieved by reducing the number of devices required for the TMA circuit design. In this research the authors have focused on the D-type latches as these have a significant device count. The latches proposed to replace the conventional circuits are designed using ITDs and NMOS transistors. It has been shown recently that it is possible to integrate these devices as monostable-bistable transition logic elements (MOBILES) [7]. These circuits have the advantage of very low complexity combined with high functional density due to the natural latching effect that is a characteristic of the ITDs.

3.1 ITD Based Latches

The voltage-current characteristic of a typical ITD is shown in Fig. 4. The curve is typical of resonant tunneling devices (RTDs) and displays negative differential resistance over part of its characteristic curve.

The main features of the characteristic curve can be subdivided into three main regions:

- PDR1 A positive differential resistance region as the bias voltage is increased from 0V to the peak voltage (V_P).
- NDR A negative differential resistance region as the bias voltage increases to the valley voltage (V_V).
- PDR2 A second positive differential resistance region as the bias voltage is increased beyond the valley voltage.

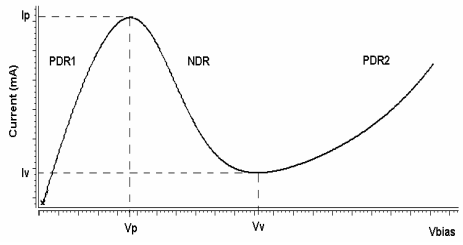

Fig. 4. ITD characteristic curve

Other features labeled in Fig. 4 include the peak current (Ip) and the valley current (Iv). The ability to alter the peak current by manufacturing devices with different areas is fundamental to the design of circuits that will ultimately create latching circuits. The valley current (Iv) is related to the off state of the device, therefore low valley currents are desirable for low power operation [5, 6]. When two of these devices are connected together in series (as shown in Fig. 5(b)) and driven by a clocked power supply it is possible by carefully selecting device dimensions to achieve a latched output [5]. Fig. 5(a) shows how the ITDs act as load and driver and create a latched output when the clocked power supply switches from zero to V_{CK}.

By including a transistor in the circuit as shown in Fig. 5(c) it is possible to control latching. In this instance if there is a logic '1' at the input of the transistor the circuit will latch, whilst a logic '0' will inhibit latching. The circuit of the latch shown here is based on the principle of operation of MOBILE [5, 6]. These latches are extremely compact having only three components. It may also be possible to construct the ITDs directly on the drain or source of the transistor which would further reduce the footprint of the latch.

The prospect of a latch with the footprint of a single transistor is obviously desirable. The impact of this on the TMA described in Section 2 would be a very large reduction in the number of devices required to implement the architecture.

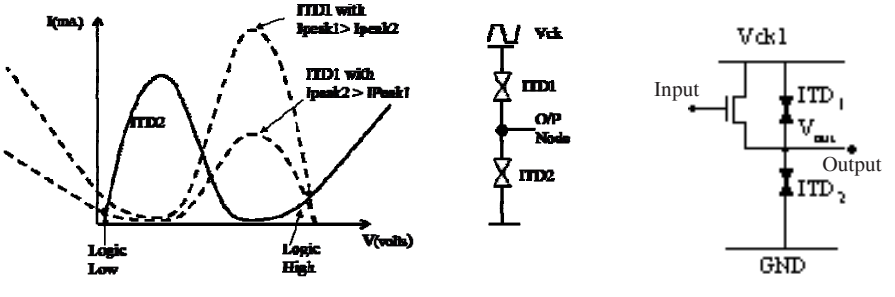

Fig. 5. (a-b) ITD characteristics and latching effect of two series devices. (c) Complete ITD based latch.

3.2 TMA Based on ITD Latches

Whilst the inherent latching effect of ITDs allows extremely low circuit complexity it can also create a problem because the power supply must be switched off before the latched state can be changed. Whilst this characteristic might be problematic for combinatorial logic circuits it is in fact a useful quality for the TMA proposed here. A fragment of the architecture is shown in Fig. 6. It is immediately obvious that the circuit complexity is significantly reduced when compared to conventional D-type latches. The number of clocks required to implement the daisy chain of latches is two.

In this arrangement the clocks Vck1, Vck2 are phase shifted to allow a logic '1' introduced at the input to be swept through the series connected latches. The first latch is connected to Vck1, if a logic '1' is introduced at the gate of its input transistor the output will become high when the power supply goes high. The output remains high until Vck1 falls to its low state. By overlapping the clocks so that Vck2 is rising whilst Vck1 is falling the second latch will switch to high at its output because there is a logic '1' at the gate of its input transistor for a sufficient period of time while Vck2 is in transition to high. During this process transistors mn5, mn6, mn7 and mn8 are switched on in sequence to sample the inputs VT1, VT2, VT3 and VT4. Two clocks were found to be sufficient for any number of inputs as only one latch output is high at any time so even though all the latches are clocked repetitively only the latch containing the '1' will turn on a sampling transistor.

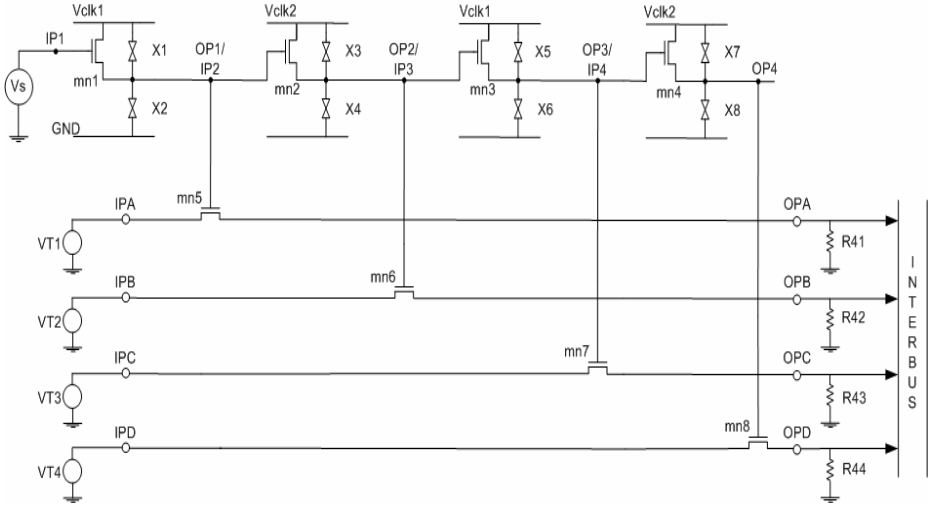

Fig. 6. Fragment of TMA using ITDs

Thus the arrangement at the input side of the single line bus is highly compact and simple in operation. The diagram in Fig. 6 presents a fragment of the TMA architecture. This shows a 4-input single-output working on a two-phase clock.

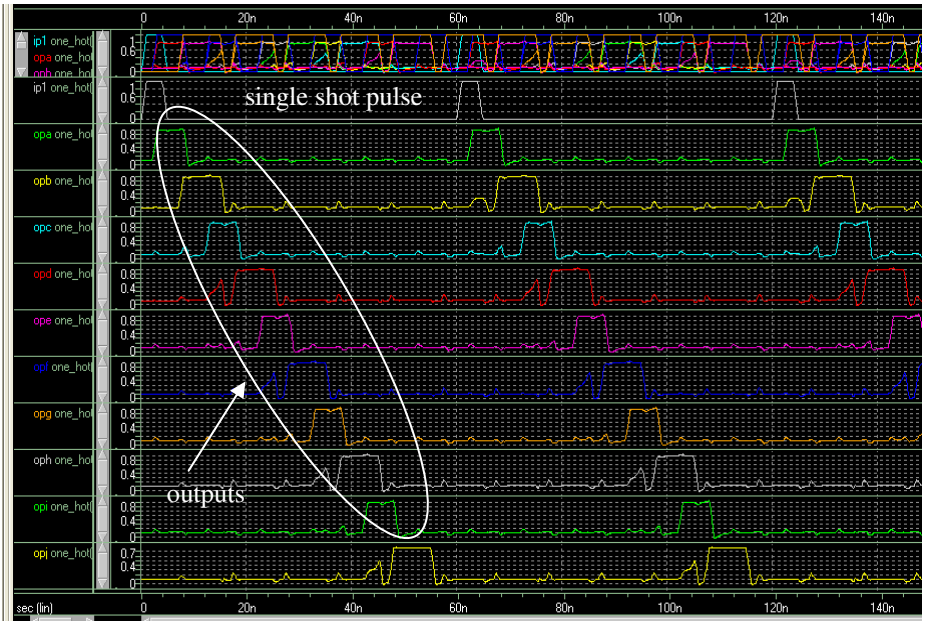

Fig. 7. Waveforms transmitted by ITD latches for 10-input neuron

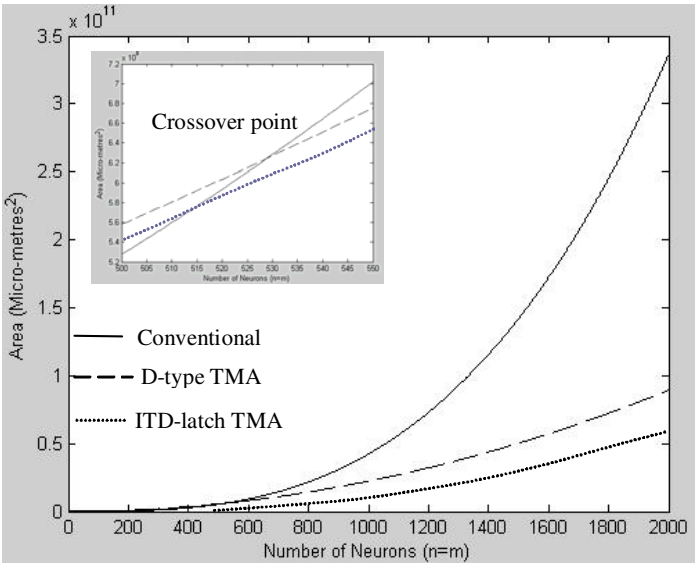

Fig. 8. Area consumed by metal and TMAs versus neuron density

Fig. 7 shows the signals for each of the lines of a ten input neuron. HSPICE models for the ITDs were developed from measurements derived from actual manufactured devices. The models reflect the performance of the devices accurately. It is clear from these simulations that the input data is successfully transmitted sequentially by the ITD latches. The signals demonstrated that a two phase clock produces a stable sequential switching of the series latches. Larger scale architectures were simulated, which showed that the design will scale whilst retaining the three phase clock at the input and output side of the single line bus. The reduction in device count compared to the authors' earlier design is significant with an order of magnitude fewer devices required to implement the architecture. The impact of this is to make more space available for neurons and synapses thus further increasing the functional density of the ANN. The graph shown in Fig. 8 compares functional density of the three approaches described in this paper.

It is clear from the curves that the conventional interconnection method is the worst case scenario with the smallest practical number of neurons. The TMA architecture implemented with conventional devices is better showing a significant increase in neuron numbers. Finally, the TMA architecture using ITD latches gives the best results with the maximum number of neurons.

4 Conclusion

The research reported in this paper shows how TMA can help to address the interconnection problem associated with ANNs. The results showed that TMA designed using D-type flip-flops increases the functional density and neuron count in an ANN when compared to conventional architecture. The introduction of ITD-based latches in this architecture further improves the functional density and neuron count. ITD based latches to create a TMA is realistic prospect for the future.

Acknowledgments. This work was supported partly by a British Council PMI2 Connect grant *Brain-inspired Interconnects for Nanoelecronics*, partly by an EPSRC project *Biologically Inspired Architecture for Spiking Neural Networks in Hardware*, and partly by the UAE National Research Foundation under the *Emirates Center for Nanoscience and Nanoengineering*.

This document is an output from the PMI2 Project funded by the UK Department for Innovations, Universities and Skills (DIUS) for the benefit of the United Arab Emirates Higher Education Sector and the UK Higher Education Sector. The views expressed are not necessarily those of DIUS, nor British Council.

References

1. Wang, P., Pei, G., Kan, E.C.-C.: Pulsed Wave Interconnect. IEEE Transactions on Very Large Scale Integration (VLSI) Systems 12(5), 453–463 (2004)
2. Chicca, E., Indiveri, G., Douglas, R.J.: An Event Based VLSI Network of Integrate and Fire Neurons. In: International Symposium on Circuits and Systems (ISCAS 2004), vol. 5, pp. 357–360. IEEE Press, New York (2004)

3. Smith, K.C.: Multiple-Valued Logic: A Tutorial and Appreciation. IEEE Computer 21(4), 17–27 (1988)
4. Tuffy, F., McDaid, L.J., Kwan, V.W., Alderman, J., McGinnity, T.M., Santos, J.A., Kelly, P.M., Sayers, H.: Inter-Neuron Communication Strategies for Spiking Neural Networks. Neurocomputing 71(Special Issue) (1-3), 30–44 (2007)
5. Chen, K.J., Maezawa, K., Yamamoto, M.: InP-Based High-Performance Monostable-Bistable Transition Logic Element (MOBILE): An Intelligent Logic Gate Featuring Weighted-Sum Threshold Operations. Japanese Journal of Applied Physics 35, 1172–1177 (1996)
6. Pacha, C., Glösekötter, P., Goser, K., Prost, W., Auer, U., Tegude, F.-J.: Resonant Tunneling Device Logic Circuits. Technical Report (MEL-ARI) ANSWERS and LOCOM (July 1999-July 2000)
7. Sudirgo, S., Pawlik, D.J., Kurinec, S.K., Thompson, P.E., Daulton, J.W., Park, S.Y., Yu, R., Berger, P.R., Rommel, S.L.: NMOS/SiGe Resonant Interband Tunneling Diode Static Random Access Memory. In: Device Research Conference, pp. 265–266 (2006)