2001

A compact multi-chip-module implementation of a multi-precision neural network classifier

Amine Bermak
Edith Cowan University

Dominique Martinez
LORIA - Campus Scientifique, France

10.1109/ISCAS.2001.921294

A COMPACT MULTI-CHIP-MODULE IMPLEMENTATION OF A MULTI-PRECISION NEURAL NETWORK CLASSIFIER

Amine Bermak

School of Engineering and Mathematics
Edith Cowan University
Perth, 6027 WA, Australia
a.bermak@ecu.edu.au

Dominique Martinez

LORIA - Campus Scientifique
Boite Postale BP239
54506 Vandoeuvre-Les-Nancy, France
dmartine@loria.fr

ABSTRACT

This paper describes a novel Multi-Chip Module (MCM) digital implementation of a reconfigurable multi-precision neural network classifier. The design is based on a scalable systolic architecture with a user defined topology and arithmetic precision of the neural network. Indeed, the MCM integrates 64/32/16 neurons with a corresponding accuracy of 4/8/16-bits. A prototype has been designed and successfully tested in CMOS $0.7\mu m$ technology.

1. INTRODUCTION

Most of the research on Artificial Neural Networks (ANN) has concentrated on theoretical studies and software simulations. However, a real need for hardware implementation has raised for real-world neural network applications mainly driven by an increasing demand for low power and portable detection systems. For this last category of applications, VLSI implementation appears as an inevitable solution which combines the sensor and the VLSI processor (ANN processor or other). Implementing the VLSI processor on a single chip is not always the best solution or the most economical one especially for laboratory prototypes. The low yield of large digital chips may result in a prohibitively expensive product [1]. An alternative solution is the use of advanced MCM and packaging technologies. MCMs are a significant advance in the field of packaging and interconnections due to the ability of MCMs to significantly increase electronic system performance and to reduce system size [2]. In our case the MCM is used in order to increase the computational power of the neural network classifier. This is mainly motivated by the need for a high number of neurons or/and a high accuracy for complex classification problems. The MCM is based on a scalable systolic architecture with a user defined topology and accuracy and a reconfigurable connectivity between the Processors Elements (PEs) of the systolic architecture and the VLSI chips of the MCM.

Section 2 of this paper explains the adopted neural network and describes the applications. Section 3 describes both the VLSI chip and the MCM design with a particular emphasis on the novel features of the hardware. Section 4 summarizes the performances of the MCM and compares them to those of well-known neural circuits reported in the literature. Finally, section 5 concludes the work described in this paper.

This work was done while the authors were at LAAS-CNRS Toulouse-France

2. THE NEURAL NETWORK STRUCTURE

The neural network studied here is a binary neural network called parity machine. Its main advantages are that it can be taught and tested using simple digital operations resulting in a simple and a compact VLSI implementation. A further advantage of the parity machine is related to the simplicity of the activation function which makes the neural network topology easily scalable and therefore suitable for MCM design. However, the parity machine does have drawbacks. It may not generalize as well as the general neural network [3], nor can it solve as many problem classes as a general neural network. More explanations on both the recall and the training of a parity machine could be found in [3]. It is basically composed of a first Threshold Logic Unit (TLU) and a decision output layer. The TLU is implemented through a weighted sum followed by a binary activation function:

$$V_j = \sum_{i=0}^{p} w_{ij} x_i \quad (1)$$

$$S_j = sign(\ V_j\ ) \quad (2)$$

where w_{ij} are the synaptic weights, x_i are the input activities S_j are the output states and p is the number of inputs.
The final output is carried out through a decision layer which, in the case of the parity machine, implements the parity of the TLU's outputs as described by Eq. (3):

$$s_f = \bigoplus_{j=0}^{q} S_j \quad (3)$$

where $\bigoplus$ is the xor function and q is the number of neurons. In the case of the committee machine, the decision layer implements the majority function of the TLUs outputs.
Several learning algorithms for the parity machine have been proposed in the literature such as the least action algorithm [4] and the Offset algorithm [3, 5].

The main goal behind the VLSI implementation of these networks is to build a low power and portable system for methane gas detection and driver non-vigilance detection. In both applications, the VLSI neural network is interfaced with a set of sensors providing real-time data. In the first application the goal is to detect the methane concentration in a gas mixture composed of CH_4, CO and H_2O. The neural network solution was selected in order to resolve the problems associated with the non-reproductibility and

0-7803-6685-9/01/$10.00©2001 IEEE

the non-selectivity of the gas sensors [6]. In the second application, the driver non-vigilance is detected by combining a set of heterogeneous sensors (such as speed sensor, vehicle position on the road sensor, eye blinks sensor, etc...) and a parity machine NN classifier in order to detect normal or abnormal driving behavior.

Extensive simulations and experiments were conducted in order to evaluate the hardware requirements for the two applications, particularly in terms of neural network structure and accuracy of the synaptic weights . It was found that 16 bits synaptic weight precision were needed for the non-vigilance detection while only 8 bits were required for the gas detection application. It was also found that a better generalization performance was obtained by combining a higher number of neurons; this obviously results in an increase of hardware requirements. Based on these considerations, we decided to implement a variable precision neural network classifier in which the accuracy is user-defined depending on the application. The parity machine is designed to achieve variable precision with 4, 8 or 16 bits of synaptic weights and arbitrary precision of the inputs. The number of neurons (and/or inputs) could be increased for lower accuracy and vice-versa. The number of neurons is also increased by interfacing more VLSI chips within the MCM.

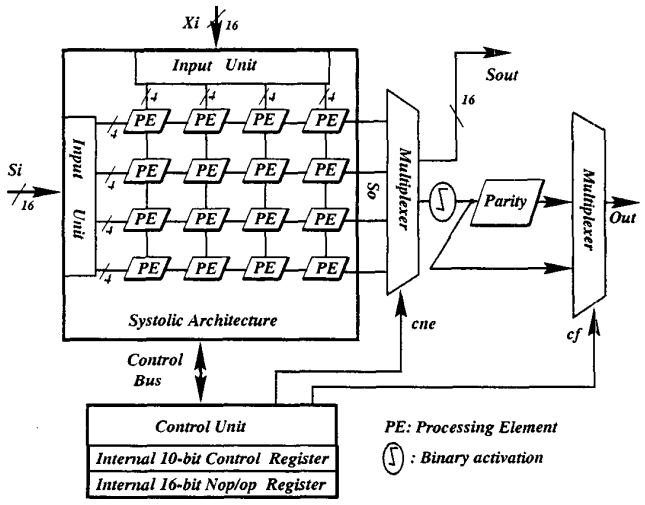


Figure 1: Internal architecture of a basic VLSI chip.

3. VLSI DESIGN

The implemented VLSI design is based on a compact MCM including four basic VLSI chips. The connections between the four basic VLSI chips is configured depending on the required neural network topology as well as the required weight and input precision. In this section we first explain the Basic VLSI chip architecture and then we explain the MCM implementation.

3.1. Basic chip design

The basic building block chip is based on 2D systolic array architecture. This array consists of 4 × 4 Processing Elements (PE) as shown on Figure 1. The array can be configured to perform a weighted sum, a parity or a comitte machine or even a multi-layer perceptron. Each processor PE includes a local reconfigurable memory to store the synaptic weights in the form of one 16-bit synaptic weight, two 8-bit synaptic weights or four 4-bit synaptic weights.

The inputs are fed serially from the least significant to the most significant bit with an arbitrary user defined precision. The 16-bit Xi bus is used to handle the inputs while the buses Si and Sout are systolic buses used to interface between basic VLSI chips when higher number of neurons and/or a higher accuracy are needed. After off-chip digital conversion of the analog sensor output, a processor PE_{ij} within the systolic array receives a digital sensor output X_i from the least significant bit to the most significant one. The processor computes the product of the locally stored synaptic weight W_{ij} by X_i. In order to perform the partial product within each processor, a novel multi-precision serial parallel multiplier [7] has been used. This last multiplier was well optimized in terms of silicon area which makes the VLSI chip cost effective. The output of the multiplier is then added to the partial sum received from the processor located on the left and transmitted to the processor PE_{i+1j} located on the right one clock cycle later. All results S_O are collected on the right side of the array. The weighted sum outputs S_O may be configured, by the control bit (Cne), as partial outputs Sout, in order to interface between the VLSI chips within the MCM. The outputs S_O may also be configured under internal control bit (cf) to perform the parity machine or the TLU. An internal 10-bit control register is used to store the control word used to configure the systolic array. In this register several parameters are stored and used to configure the array, such as the network structure (number of inputs and neurons), network type (TLU, parity machine, committee machine, matrix operation), the arithmetic coding (unsigned or signed two complement numbers), the weight accuracy (4, 8, 16 bits) and the input accuracy (arbitrary).

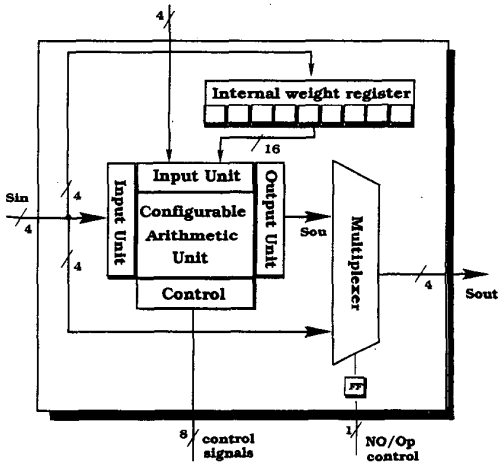


Figure 2: Processing Element (PE) building block diagram. for NO/Op = 1, Sout = Sou (Op) while for NO/Op = 0, sout = sin (NO). FF stands for Flip-Flop.

Figure 2 shows a block diagram of a basic Processing Element. The processor integrates a novel loading technique of the synaptic weight using the 4-bit systolic bus Sin. This same bus Sin is used to load the 16-bit internal weight register, to feed the partial sum inputs to the configurable arithmetic unit or to the 2I/1O

multiplexer. In this last situation, the processor is not operational (NO) and the output data Sout are passed to the adjacent processor located on the right without any processing. With this last NO programming technique, three interesting features are obtained:

1. Easy test procedure since each processor can be tested separately. The test of one processor within the systolic array is done by programming all other processors as NO.
2. High level of reliability is obtained within the circuit. If following a test, a processor is found faulty, then it will be programmed as NO and the systolic architecture will be still functional.
3. The loading of the synaptic weight is facilitated using this (NO) programming technique since the synaptic weights of all PEs are loaded using a single 16-bit bus (Sin). Moreover, the synaptic weights of the four interfaced VLSI chip within the MCM are all loaded using the same 16-bit bus.

Figure 3 shows the architecture of the reconfigurable arithmetic unit in which each row consists of a single 4-bit processor. Eight multiplexers within each PE are used to change the hardware connections between two adjacent rows of cells in order to obtain a synaptic weight precision of 4, 8 or 16-bit. The configuration with the lowest precision provides an increase in the number of inputs or neurons according to the following equation:

$$n \times p \times q = C \tag{4}$$

Where n, p and q are the weight precision, number of inputs and the number of neurons respectively. C is a constant term equal to $2^{12} = 4096$ for one single VLSI chip and $2^{14} = 16384$ for the MCM. We can note that the circuit limitation is a trade-off between three parameters (n,p and q). The MCM hardware resources are increased by a factor of four as compared to a single VLSI chip.

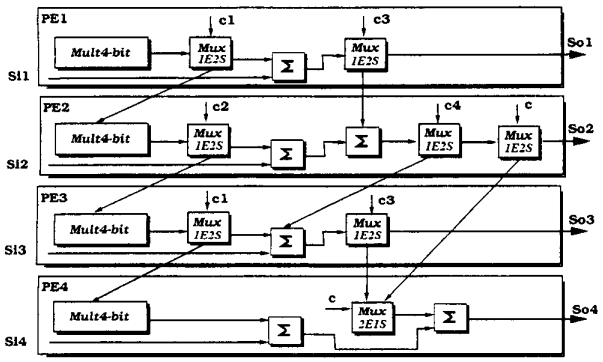


Figure 3: Internal schematic of the configurable arithmetic unit.

3.2. MCM design

The main objective behind the development of the MCM is to increase the neural network computational power needed for a better generalization performance, a higher number of neurons and a higher synaptic weight accuracy. The MCM offers the possibility of high level of integration needed for portable systems as it is the case for our application. Since a cost effective solution was an important criteria for our applications, we chose to implement an array of 4x4 systolic PEs and to increase the computational power by interfacing 4 basic chips within the prototype. The MCM-Laminated (MCM-L) has been adopted since it is a very cost effective solution compared to other MCM possibilities. Figure 4.A shows a microphotograph of the VLSI chip while Figure 4.B shows the MCM. This latter includes four basic VLSI chips and occupies an area of $2 \times 2cm^2$ and integrates 256 digital synapses.

4. PERFORMANCE

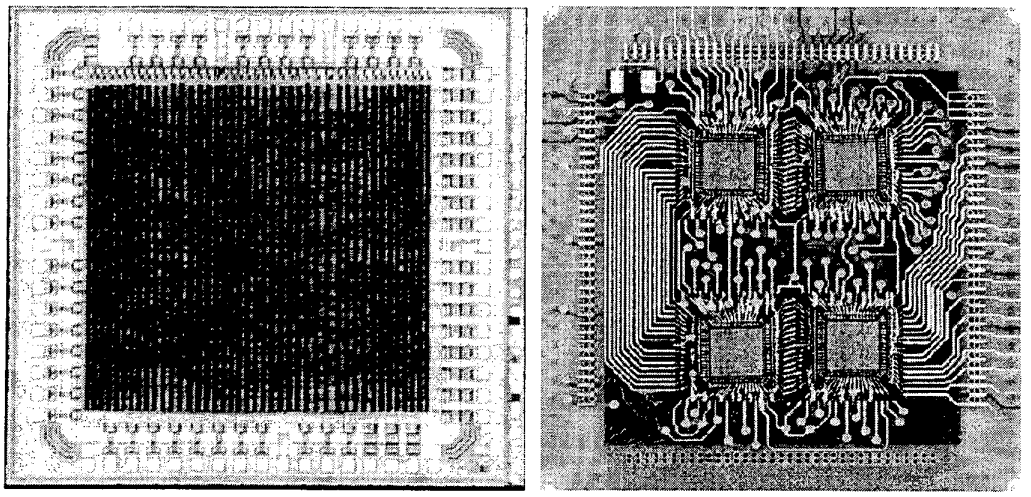
The basic VLSI circuit was fabricated in $0.7\mu m$ CMOS technology. Verilog digital simulator was used to check the logic correctness of the design and also to generate random test vectors for the test of the circuit and the MCM. The circuit and the MCM were successfully tested at a maximum frequency of 20 MHz for different topologies and types of the neural network, different precisions and signed and unsigned arithmetic. The MCM power dissipation is at most 92 mW/MHz. The loading time at the maximum frequency is about $4\mu s$ for all control words and synaptic weights. Table 1 shows the performance comparison between the proposed MCM neural classifier and most well-known circuits reported in the literature. The VLSI design proposed in this paper presents the advantage of being reconfigurable in terms of precision and neural network topology. The MCM is a powerful neural processor including 64 Processing Elements implementing 256 digital synapses. It exhibits a significant speed of 640MCPS. These last performances are comparable with those of the powerful neural processor N64000 from Innova, which implements more than 3.4M gates.

Name of IC	CMOS techno	Precision $b_{in} \times b_w$	NN	Speed CPS
our	$0.7\mu m$	**cfg**$\times(4/8/16)$	64PE	640M
MD1220[8]	NA	1×16	1PE	9M
SIOP2[9]	$0.8\mu m$	16×10	NA	350M
SAND[10]	$0.8\mu m$	16×16	NA	200M
LNeuro[11]	NA	$16 \times (4/8/16)$	16PE	26M
N64000[12]	NA	9×16	64PE	870M
MA16[13]	NA	16×16	16PE	400M

Table 1: *Performance comparison of neural circuits. In the table, NN stands for the Number of Neurons, cfg stands for configurable and NA stands for Not Available.*

5. CONCLUSION

A compact MCM digital neural network classifier has been fabricated and successfully tested. The VLSI circuit has been designed in order to be interfaced with a multi-sensor system for detection applications such as gas and non-vigilance detection. The circuit features a high level of flexibility and programability which makes it very suitable for a wide range of applications. The proposed architecture supports a variable precision of the synaptic weights (4/8/16) a reconfigurable network structure (number of inputs and neurons), a programable network type (TLU, parity machine, committee machine, matrix operation) and a signed or unsigned arithmetic. The design also includes other novel design features such as a systolic synaptic weight loading technique. The final MCM design occupies an area of $2 \times 2cm^2$ and implements 64 neurons with a processing speed of 640MCPS.



(A.) (B.)

Figure 4: *(A.) Microphotograph of the VLSI chip. (B.)* $2 \times 2cm^2$ *MCM including four VLSI chips.*

Acknowledgments

This work was done while the authors were at LAAS-CNRS Toulouse-France. The authors would like to thank Dr. C. Val from 3D-Plus Electronics and Dr. D. Esteve from LAAS-CNRS for their support and helpful discussions.

6. REFERENCES

[1] S. P. Larcombe, et al., "An Ultra-Miniature Camera and Processing System," *IEE Colloquium on Integrated Imaging Sensors and Processing*, 1994, London, UK.

[2] P. D. Franzon, "MultiChip Module Technologies and Alternatives," *New York: Van Nostrand Reinhold*, 1993, ch. 3, pp. 102-106.

[3] D. Martinez and D. Estève, "The Offset Algorithm: Building and Learning Method for Multilayer Neural Networks," *Europhys. Lett*, 18 (2), pp. 95-100, 1992.

[4] G. J. Mitchison and R. M. Durbin, "Bounds on the learning capacity of some multi-layer networks," *Biol. Cybern, 60*, pp. 345-356, 1989.

[5] M. B. Gordon, "A convergence theorem for incremental learning with real-valued inputs," *Proceedings of ICNN*, pp. 381-386, Washington, 1996.

[6] M. Dilhan, D. Martinez and N. Jaffrezic, "Chemical microsensors," *L'onde Electrique*, Vol. 74, no. 2 pp. 28-35, 1994.

[7] A. Bermak, D. Martinez, J.L Noullet, "High-density 16/8/4-bit configurable multiplier," *IEE Proc. Circuits Devices Syst*, Vol.144, N.5, pp.272-276, Oct. 1997.

[8] "*MD 1220 Data Sheet*," Micro Devices, 1990

[9] B. Friebe, S. Neusser and B. Hofflinger, "SIOP: Application-Specific Neural Hardware," *Proceedings of MicroNeuro 97*, pp.18-24, Dresden, Germany, 1997.

[10] W. Eppler, T. Fischer, H. Gemmeke, T. Becker and G. Kock, "High Speed NN Chip on PCI board," *Proceedings of MicroNeuro 97*, pp.18-24, Germany, 1997.

[11] N. Mauduit, M. Duranton, J. Gobert, and J. A. Sirat "Lneuro 1.0: A piece of Hardware LEGO for building neural network systems," *IEEE transactions on Neural Networks*, Vol.3, N.3, pp.414-422, May. 1992.

[12] D. Hammerstrom, "A VLSI architecture for high performance, low cost, on chip learning," *Proceedings of the IJCNN'90*, pp.537-544, USA, 1990.

[13] U. Ramacher, J. Beichter and N. Bruls "Architecture of a general-purpose neural signal processor," *Proceedings of the IJCNN*, Volume I, pp.443-446, July 1991.