Extended Abstracts of the 1994 International Conference on Solid State Devices and Materials, Yokohama, 1994, pp. 388-390

Parallel Implementations of Neural Networks Using the L-Neuro 2.0 Architecture

Christophe DEJEAN, Françoise CAILLAUD

Laboratoires d'Electronique Philips,
22, avenue Descartes
94453 Limeil-Brévannes Cedex (France)
Email : dejean@lep-philips.fr

Real-time and embedded applications of neural networks are often limited by the time required for both the learning and the recall phases and also by the preprocessing operations, especially when neural networks are used for image processing. In this paper, we present a VLSI chip successor of L-Neuro 1.0[1)]. The basic architecture is derived from the one of L-Neuro 1.0, with improvements due to a better knowledge and analysis of the problems of this kind of structure.

1. INTRODUCTION

From the studies of the systems using L-Neuro 1.0, we derived the improved chip L-Neuro 2.0 which is a general purpose parallel neuro-chip.It is able to perform several neuronal like applications, and it is not dedicated to a particular one. It is also able to perform some preprocessing functions, mainly for image treatment (convolution, filtering...). It is best suited to regular algorithms, with no (or few) conditional branches and tests (but with loops), so the control unit is rather simple and the internal pipeline of the chip can be fully used. The on-chip controller is powerful enough for stand-alone operations on the chip. The bus interface is versatile in order to be compatible with several microprocessor bus interfaces (multiplexed or not). The cascability is supported at several levels :

- Multi-chip cascadability.
- Word size extension.
- Possibility to time-multiplex neurons on the same hardware.
- The instruction set will be extensible in order to support one architecture but with several embodiments. It is also simple, easy to be automatically generated.

Lots of general operations can be performed on the chip (for example, dot product, the "Hebb learning step", minimum and maximum value extraction in a vector, norm calculation, steps for FIR calculations, vector operations...).

The L-Neuro 2.0 will be constituted of three main units :

- Operative part constituted by four sub-parts.
- Control part.
- Communication part.

which will be described in the following paragraphs.

2. THE OPERATIVE UNITS

There are four different operative units within a L-Neuro 2.0 :

- the Vector Processor is composed of N identical Elementary Vector Processors (EVPs). It is organized in a column of N 16-bit processors acting in an SIMD fashion. They operate on vector data. EVP is a small processor with a 16-bit register file of 128-words and with an appropriate memory mapping, it is

possible to have access to a matrix and to its transposed. It contains a 16-bit multiplier and a 32-bit ALU which is able to perform all common operations on 32 bit operands, and min-max function.

- The Vector to Scalar unit, which has to "compact" a vector into a scalar. This unit was a tree of adders in L-Neuro 1.0, but is extended here to extract the maximum (or the minimum) value in a vector, and its position.

- A Scalar Unit (SU), working on 32-bit words. It performs scalar operations and more complex operations like square root extraction and division, and may act as a look-up table controller. The unit has a saturation logic under control of overflow and underflow condition.

- A Scalar to Vector Unit which generates vectors for the EVPs from scalars issued from the Scalar Unit.

All these elements are shown in the figure 1.

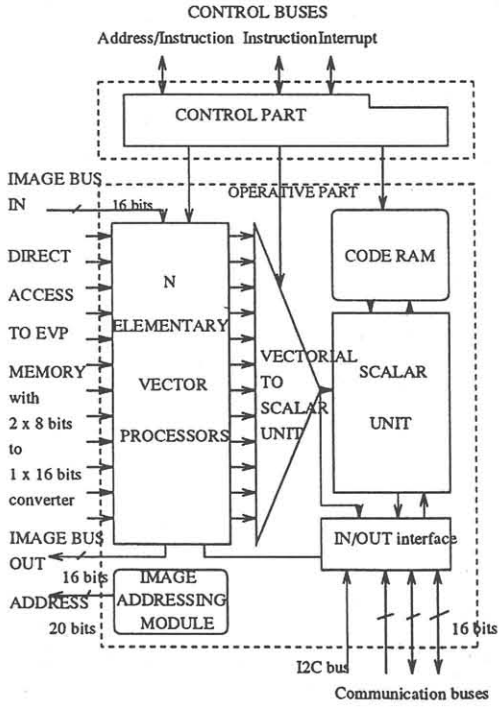

Fig. 1 : overall architecture of the L-Neuro 2.0 chip

3. THE CONTROL UNITS

L-Neuro 2.0 is a WISC (Writable Instruction Set Computer) controlled device which has a VLIW (Very Long Instruction Word) structure to simultaneously control in MIMD (Multiple Instruction Multiple Data) fashion the four operative parts: the EVPs, the Vector to Scalar Unit, the Scalar Unit and the Scalar to Vectorial Unit. A two level code allows to expand the 32 bit instruction into the final VLIW instruction.

4. THE COMMUNICATION PART

The communication part is divided into four parts:

- A 32-bit wide interface bus (MB) to the host microprocessor and/or to the dedicated controller. In autonomous mode, this bus can be split into two buses : one address bus and one data bus, 16-bit wide each. This allows to connect the chip to a standard RAM or ROM to store the program.

- Communication buses (CB) with other L-Neuro circuits, divided into one mono-directional input bus (ICB) and two bidirectional output port (OCB). These two buses can be two 16-bit parts of a general purpose 32-bit communication bus.

- Image buses (IB) : one input image bus and one output image bus that allow to manipulate sub-windows in an image.

- N parallel buses: one 8 bit bus for each EVP, and which allows to load data directly in the memory of EVP.

The interface bus is compatible with common buses.

5. CONCLUSION

Neural networks, besides their utility for applications, are also a good testbed for fine grain parallel architectures. To validate dedicated hardware, we have followed an incremental approach that allows to test and benchmark various pieces of the architecture. From these experiences, we have defined the architecture of the L-Neuro 2.0 chip which has an operating frequency of 40 MHZ in 0.6 μm CMOS technology and which can reach 1.2 Gops. The L-Neuro 2.0 chip will be sold with its development system which will consist in a modular system (several boards), and can be used for the application development phase and for the final application phase. The development system, the boards and the software, will be available by end of 1994.

6. REFERENCES

1) N. Mauduit, M. Duranton, J. Gobert, J.A. Sirat, Lneuro 1.0 : A Piece of Hardware LEGO for Building Neural Network Systems,IEEE Trans. on Neural Networks, Vol. 3, n.3, pp. 414-422, May 1992.