Helsinki, Finland, June 4-6, 1997, pp. 136 - 141

SOM Hardware-Accelerator

S. Rüping, M. Porrmann and U. Rückert

Heinz Nixdorf Institute, University of Paderborn,
System and Circuit Technology, Fuerstenallee 11,
33102 Paderborn, Germany
E-Mail: rueping@hni.uni-paderborn.de

Abstract. Many applications of Selforganizing Feature Maps (SOMs) need a high performance hardware system in order to be efficient. Because of the regular and modular structure of SOMs , a hardware realization is obvious. Based on the idea of a massively parallel system, several chips have been designed, manufactured and tested by the authors. In this paper a high performance system with the latest NBISOM_25 chips is presented. The NBISOM_25 integrated circuit contains 25 processing elements in a 5 by 5 array. Due to the scalability of the chips a VME-bus board was built with 16 ICs on it. The controller for the VME-bus and the SOM hardware are realized using FPGAs. The system runs SOM applications with up to 400 elements in parallel mode (20 by 20 map). Each weight vector can have up to 64 weights of 8 bit accuracy. The maximum performance of the board-system is 4.1 GCPS (recall) and 2.4 GCUPS (learning).

1 Introduction

Applications of Selforganizing Feature Maps (SOMs) [1] are controlling problems, data analysis and pattern matching. In several cases the performance of a workstation running a SOM software fits the requirements of the application. But there are still a lot of problems, that come with real-time or high performance requirements [e.g. 6, 7]. In these cases a custom hardware system is necessary to provide a solution. Especially embedded applications of selforganizing feature maps can not be realized with workstations due to their physical size.

The SOM hardware system presented by the authors is based on a massively parallel structure that provides a processing unit with on-chip memory for each neuron of the selforganizing feature map. There are two main fields where this hardware can be applied. The first is a single chip or MCM (Multi-Chip Module) embedded in an environment with the requirement of small physical size. The second is a large hybrid system with different components, where a board containing several SOM chips guarantees a high performance for applications with large data files. [8]

The principles of the former BISOM / NBISOM chips have been presented in [2, 3, 4, 5]. In this paper a complete VME-bus board containing 25 of the latest NBISOM_25 chips is presented. The board is part of a larger system that includes a VME-bus SparcStation as well as different processor-boards and custom hardware boards for neural systems.

Chapter 2 explains the internal structure of the NBISOM_25 chips. In chapter 3 the SOM board is described in detail. The NBISOM_25 chips are scalable and can be used to build different map sizes. On this board the maximum map size is 20 by 20. Performance numbers are given in chapter 4 and a summary can be found in chapter 5.

2 The NBISOM_25 Chip

As described in [5] the design of the NBISOM_25 chip is based on the idea to simplify the algorithm of selforganizing feature maps due to hardware aspects in order to minimize the necessary chip area and thus to maximize the number of processing elements per chip. First the Manhattan distance is used instead of the

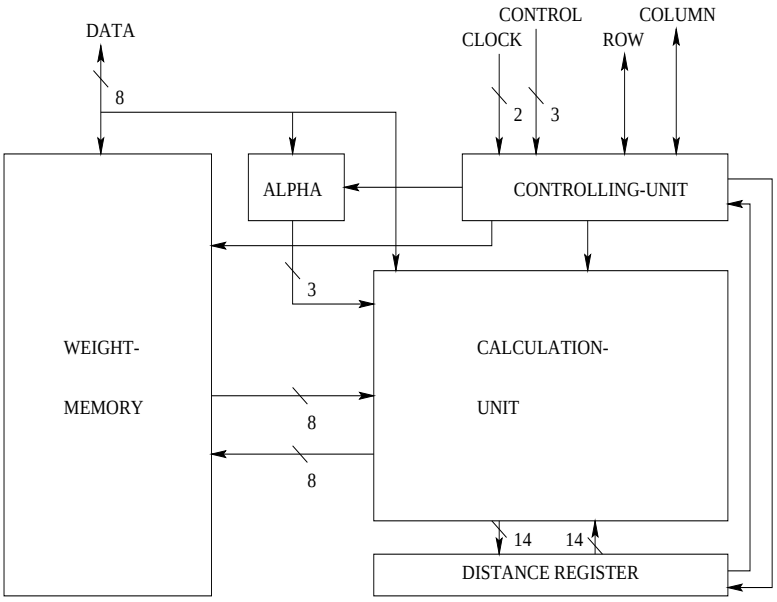

Figure 1: The main building blocks of a processor element

euclidean distance. Therefore no multipliers are necessary to calculate the distance. Secondly the values of the adaptation factor alpha are restricted to the set {1, 1/2, 1/4, 1/8, 1/16,...}. This makes it possible to use shifters instead of multipliers for the adaptation. It could be shown in [4] that the simplified algorithm is still working properly for most applications.

The NBISOM_25 chip contains 25 processing elements with on-chip memory on a chip area of 75 mm^2 (including pads). About 40 % of the core area is used for the SRAM memory blocks. Each element has 64 weights with 8 bit accuracy. The chip was manufactured using the EUROPRACTICE services (ES2 1.0 µm CMOS technology). It contains 6076 standard cells, 25 macro cells (SRAM) and 80 padcells. The package type is a pin grid array PGA 84.

The chip test has been finished successfully. During the test a proper function up to 16 MHz could be proofed. An advantage of the concept of many elements working in parallel on the chip is the graceful degradation of the system performance. Beside the proper working chips, the circuits with one or few manufacturing faults on it can also be used. If the processing element with a fault does not react at all, the chip can be run without changes of the control environment. If the element shows wrong reactions, the controller has to store the address and fade out the element by masking the position.

Each processor element is built with the blocks shown in figure 4. The weight memory block contains 64 bytes of SRAM and an special counter register for generating the necessary addresses. The alpha register is used to store the current value of the adaptation factor alpha. The distance register is necessary to store and accumulate the value of the distance (14 bit). It is also used to count down the distance during the best match search.

All necessary operations for the data processing are implemented in the calculation unit (14 bit accuracy). It can calculate the Manhattan distance, multiply by the adaptation factor alpha and decrement the distance register.

The controlling unit of the processor element handles the synchronization and the mode select of the other blocks. It is also the main interface to the outside world of the element. There are connections to the row and column line of the element, a control bus for the command bits, two clock lines (non overlapping, 16 MHz) and an 8 bit data bus.

A NBISOM_25 chip has 8 I/O pins for the data bus, 3 input pins for the control bus, 2 input pins for the clock, 5 I/O pins for the row lines (Open Drain) and 5 I/O pins for the column lines (Open Drain). All in all the number of necessary pins is 23 plus the pins for the power supply. This shows, that the number of 84

pins available in the used package PGA 84 is much to high. But this package was the choice with the lowest pin count, that fits to the necessary die size (7.67 mm x 9.69 mm). Figure 5 and 6 show the package and the chip area.

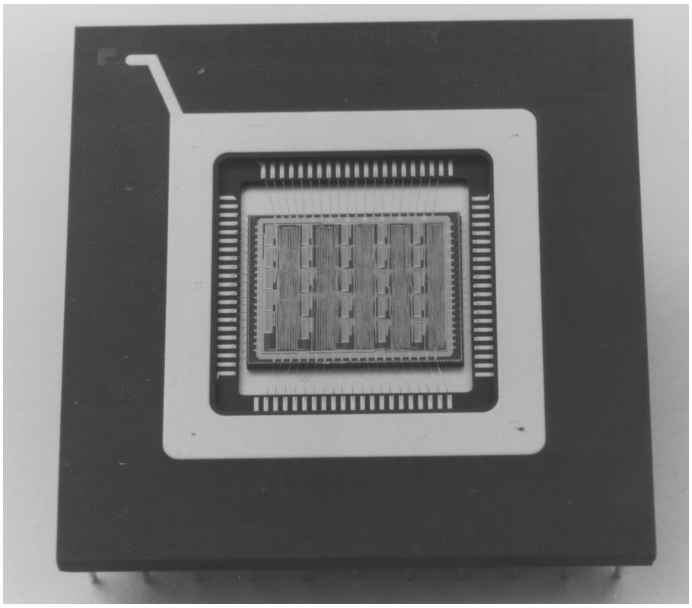

Figure 2: The NBISOM_25 chip

3 The NB25-VME Board

The most important component of the board is the NBISOM_25 chip, which is described in detail in chapter 4. Additionally there are controllers for the VME-bus and the NBISOM_25 chip array as well as a dual-port SRAM on it. The VME-bus controller is responsible for the communication tasks concerning the VME-bus port. The NBISOM controller handles the array of 16 NBISOM_25 chips. Each chip contains 25 processing elements that are arranged as a two-dimensional array with 5 rows and 5 columns. Each row and column has an input/output pin. These pins are connected on board-level, so that the controller has to handle 20 row and 20 column lines. The lines are used to address single or multiple elements as well as to find the so-called best match position. This is the position of the processing element, that stores the weight vector with the minimum distance to the current input vector.

Additionally the NBISOM controller generates the clock and the control-bits for the array. The data of the input vectors is stored in the dual-port SRAM, which means the controller has also to generate the address of the memory synchronized to the control sequence.

When an input vector or a sequence of vectors is to be learnt by the SOM, the data is transferred via the VME-bus into the dual-port SRAM. Then the NBISOM controller is started and the vectors are processed. The parameters for adaptation strength and width of the neighborhood function is stored in a parameter field in the dual-port SRAM. When the NBISOM controller has finished the job, it sends an interrupt to the VME-bus controller and the next data can be processed.

During recall phase the input vectors are transferred in the same way. After processing the best match position is stored in the dual-port SRAM and an interrupt is sent. The position can be read over the VME-bus.

Additionally there are commands for reading and writing data out of or into the array of processing elements. The addresses have to be stored in a special address field of the dual-port SRAM.

The layout of the NB25-VME board can be seen in figure 3. On the right side the array of 16 NBISOM_25

chips is shown. On the left side the field-programmable gate arrays XC3164A-1 and XC3195A-1 are placed on the board. The first handles the VME-bus, the second is responsible to control the NBISOM_25 array. Additionally the dual-port SRAM and an EPROM containing the bitstream for programming the FPGAs can be found.

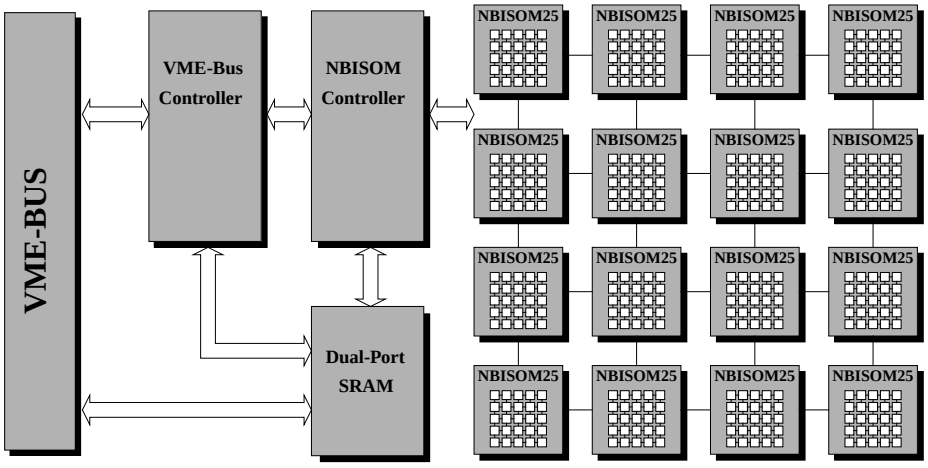

Figure 3: The NB25-VME board structure

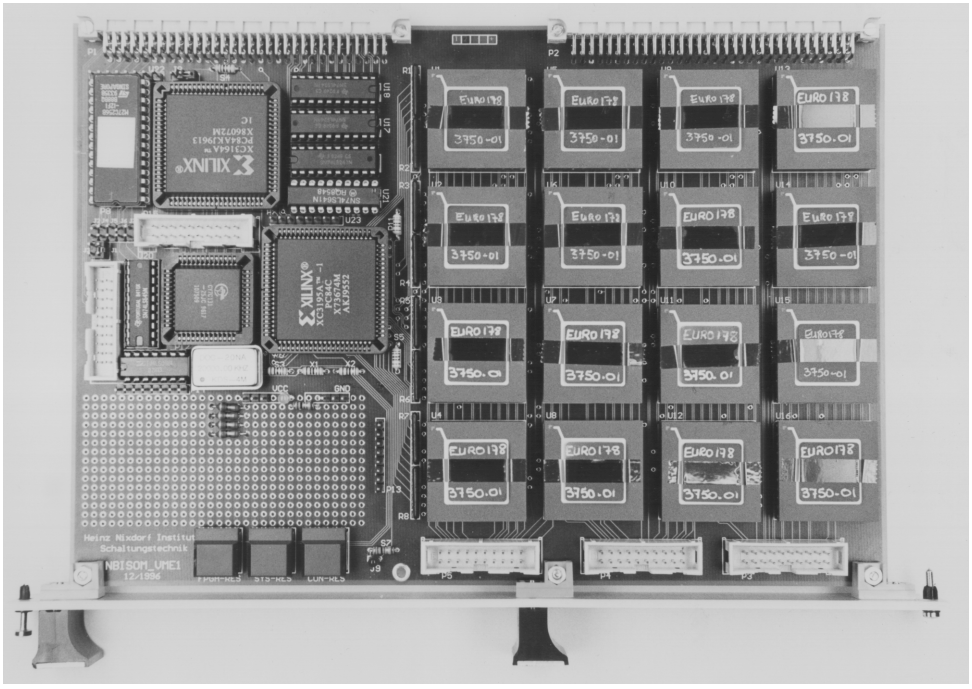

Figure 4: The VME-NB25 system board

4 Performance of the System

In order to estimate the performance of the NB25-VME board, the units MCPS (Million Connections Per Second) for recall phase and MCUPS (Million Connection Updates Per Second) for learning phase will be

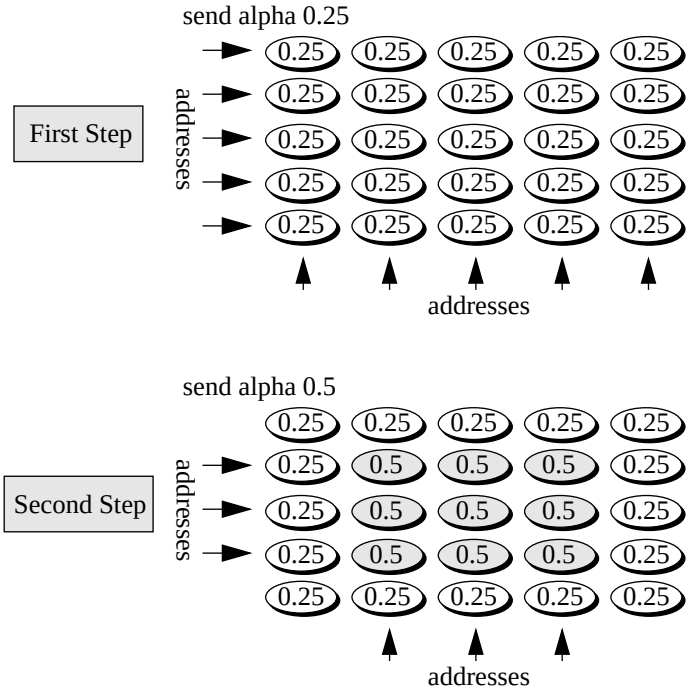

Figure 5: Distributing of the adaptation factor alpha

used. Each processing element can handle one input vector component per clock cycle. For an input vector with 64 components (8 bit accuracy), the calculation of the distance (14 bit accuracy) to the internal weight vector takes 64 clock cycles. After that the best match search is realized by decrementing the distance and detect the first element to be zero. Therefore the time that is necessary to find the best match depends on the amount of the minimum distance. Here we assume an average value of 36, which sums up to a total time for calculating the distance and finding the best match of 100 clock cycles.

During learning phase the weights of the best match element and of the surrounding elements have to be adapted. The strength of the adaptation depends on the learning step and the position of the element corresponding to the best match position. We have restricted the set of values that can be used for the multiplication during adaptation to {1, 1/2, 1/4, 1/8, 1/16, ...}. If we consider, that we have 8 bit accuracy of the weights, the maximum number of shift operations can be 8. This leads to 8 different adaptation factors.

Before an element can adapt it's weights, it must receive the adaptation factor. The row and column lines are used to address single elements as well as arrays of elements.

In the first step the smallest factor is send to an array of elements with the best match in the middle. During the following steps, a smaller array is addressed and higher values of the factor are send to the elements. Each send operation takes only one clock cycle, which means the distribution of all factors is finished after 8 clock cycles.

The elements can adapt their weights in parallel. In order to spare memory in the elements, the controller sends the input vector components once again during adaptation. Corresponding to the number of 64 components it takes 64 clock cycles to adapt all weights. This leads to a sum of 172 clock cycles for each input vector during learning phase including the best match search.

The board runs with 16 MHz. The are 400 elements with 64 weights working in parallel. Therefore the following performance numbers can be calculated. The time to send the data to the VME-board and to read the result from it is not included in these numbers. Nevertheless these operations can be done in parallel to the SOM calculations. (Pipelining)

	Mapsize	Components	Recall	Learning
Chip	5 x 5	64	256 MCPS	149 MCUPS
Board	20 x 20	64	4096 MCPS	2382 MCUPS

5 Summary

In this paper we present a high performance hardware for selforganizing feature maps. The hardware is a VME-bus board that is embedded in a hybrid VME-bus system with different boards for tasks like preprocessing, actuator/sensor interfacing, neural associative memories, interfacing to a workstation network and selforganizing feature maps.
The board for SOMs is based on a custom chip (NBISOM_25) that consists of 25 processing elements. There are 16 of these chips on the board which leads to a maximum map size of 400 elements (20 x 20). The weight vectors can have up to 64 components of 8 bit accuracy. All elements work in parallel and build a two-dimensional SIMD structure. The controller for the NBISOM_25 chip array and for the VME bus are realized using FPGAs. A dual-port SRAM is used to store the input/output data.
The chip was manufactured using the EUROPRACTICE services (ES2 1.0 µm CMOS). It runs with 16 MHz and has a die size of 75 mm^2. The processing elements have on-chip memory and do on-chip learning. In order to minimize the necessary chip area and maximize the number of processing elements per chip, the algorithm of selforganizing feature maps has been simplified due to hardware aspects.
With the board containing the chips we achieve high performance during recall and learning phase. For a 20 by 20 map the numbers are 4,1 GCPS (recall) and 2,4 GCUPS (learning).

Acknowledgments

This work has been partly supported by the Deutsche Forschungsgemeinschaft (German Research Council) DFG GR 948/14-2 and DFG RU 477/2-3.

References

1. Kohonen, T.: “Self-Organizing Maps“, Springer-Verlag, Berlin, 1995.
2. Rüping, S., Rückert, U., Goser, K.: "Hardware Design for Selforganizing Feature Maps with Binary Inputs", in J. Mira, J. Cabestany, A. Prieto (Eds.): New Trends in Neural Computation, Lecture Notes in Computer Science 686, Springer Verlag, Berlin (1993), pp. 488-493.
3. Rüping, S., Goser, K., Rückert, U.: "A Chip for Selforganizing Feature Maps", IEEE MICRO, Vol. 15, No. 3, June 1995,.pp. 57-59.
4. Rüping, S.: “VLSI-gerechte Umsetzung von selbstorganisierenden Karten und ihre Einbindung in ein Neuro-Fuzzy Analysesystem“, Fortschritt-Berichte VDI, Reihe 9: Elektronik, Düsseldorf: VDI Verlag, 1995.
5. Rüping, S., Rückert, U.: “A Scalable Processor Array for Self-Organizing Feature Maps“, Fifth International Conference on Microelectronics for Neural Networks and Fuzzy Systems, MicroNeuro’96, February 12-14, 1996, Lausanne, Switzerland, pp. 285 - 291.
6. Witkowski, U., Rüping, S., Rückert, U., Schütte, F., Beinecke, S., Grotstollen, H.: "System Identification Using Selforganizing Feature Maps", submitted to the Fifth International Conference on Artificial Neural Networks. Cambridge, UK, July 1997.
7. Schütte, F., Beinecke, S., Grotstollen, H., Witkowski, U., Rüping, S., Rückert, U.: "Structure- and Parameter Identification for a Two-Mass-System With Backlash and Friction Using a Self-Organizing Map", submitted to the European Conference on Power Electronics and Applications, Trondheim, Norway, September 1997.
8. Rückert, U., “A Hybride Knowledge Processing Architecture“, IEE-Proc. Publication No. 395 “Intelligent Systems Engineering“, Norwich, UK, 1994, pp.372 - 377.