A Multi-Chip Module Implementation of a Neural Network

Matthew G. Stout
Linton G. Salmon

Department of
Electrical and Computer Engineering
Brigham Young University
Provo, UT 84602

George L. Rudolph
Tony R. Martinez

Department of Computer Science
Brigham Young University
Provo, UT 84602

Abstract

The requirement for dense interconnect in artificial neural network systems has led researchers to seek high-density interconnect technologies. This paper reports an implementation using multi-chip modules (MCMs) as the interconnect medium. The specific system described is a self-organizing, parallel, and dynamic learning model which requires a dense interconnect technology for effective implementation; this requirement is fulfilled by exploiting MCM technology. The ideas presented in this paper regarding an MCM implementation of artificial neural networks are versatile and can be adapted to apply to other neural network and connectionist models.

1 Introduction

Artificial neural networks offer an exciting area of research because of their ability to solve difficult problems, typically those dealing with pattern recognition. This ability is due, in part, to their densely interconnected parallel architecture. However, often neural networks are simulated on sequential computers and lose much of their potential speed capability, since this inherent parallelism is lost [11]. Hardware implementations of neural networks offer a solution to this drawback but also present several challenges.

One particular challenge to the development of neural network hardware implementations is the typically high interconnect density that these implementations require. The required density can be appreciated by considering the equations that govern the dynamics of neural network models. Ramacher and Schürmann [11] have shown that the operation of most neural network models can be summarized by three general equations, the first of which is presented here without derivation. The output characteristics of a *neuron* (*processing element*) are described by

$$y_{i,p} = f_i(x),\ 1 \leq i \leq N,\ p \,\epsilon\, P, \tag{1}$$

where the argument of f_i is

$$x = \sum_{j=-1}^{N} W_{i,j} y_{j,p}. \tag{2}$$

In Equations 1 and 2, N is the number of neurons, i and j are arbitrary neurons, P is the index set for the patterns, p is an element of this set of patterns, f_i is the activation function of neuron i, $W_{i,j}$ is the connection strength or *weight* from neuron i to neuron j, and $y_{j,p}$ is the input to neuron j corresponding to pattern p. Specifically, $y_{-1,p}$ is the neuron's individual input, $W_{i,0}$ is the neuron's individual threshold, and

$$W_{i,-1} = y_{0,p} = 1.$$

The other general equations reported in [11] are not presented here. Briefly, they represent the changes in the values of the weights between neurons and the updated values of these weights.

These general equations can be used to estimate the general hardware requirements of neural networks. For instance, Equations 1 and 2 show that for the most general neural network consisting of N completely interconnected neurons (each neuron is connected to every other neuron), the outputs of the neurons $y_{i,p}$ are determined by the sum of products of the interconnecting weights $W_{i,j}$ and corresponding node inputs $y_{j,p}$. For this case, the number of connections is $O(N^2)$, which shows that present neural models could potentially require huge amounts of interconnections.

This observation has additional significance when practical implementations of neural networks are considered. A DARPA study [2] has reported characteristics and requirements of various neural network applications. These proposals include radar pulse identification, robot arm movement, isolated word recognition, low-level vision, and risk evaluation. The number of neurons ranges from 312 to 64,000, and the number of *synapses* ranges from 3,600 to approximately 4,000,000. In this sense, a synapse refers to a weighted connection, or in other words, the product $W_{i,j}y_{j,p}$ from Equation 2. Since the required interconnect density for an implementation is directly related to the number of synapses, it can be seen that practical hardware implementations of neural networks could potentially require many interconnections.

This requirement for dense interconnect in artificial neural network systems has led researchers to seek high-density interconnect technologies [1, 3, 14]. This paper reports an implementation of a connectionist model using a multi-chip module (MCM) as the interconnect medium. In this sense, a connectionist model refers to a system composed of fairly simple processing elements which are densely interconnected; artificial neural networks constitute a subset of connectionist systems. In general, the dynamics of all connectionist systems are *not* accurately described by the general equations presented earlier. However, like neural networks, typically the processing elements of general connectionist models require many interconnections. The authors believe that MCMs can be used as an effective interconnect medium for artificial neural networks as well as other general connectionist systems.

The integrated circuits (ICs) used in this MCM system are modeled after PASOCS (Priority Adaptive Self-Organizing Concurrent System), which is based on the connectionist architecture ASOCS (Adaptive Self-Organizing Concurrent System) [7, 8]. Although this connectionist model differs significantly in its mechanisms from the neural network models whose characteristics are described by the general equations mentioned earlier, the goal of both types of connectionist models is the same—both attempt to learn an arbitrary set of vector mappings. As mentioned previously, the models' similarities extend even further, since a practical system based on the ASOCS architecture would also typically require many interconnections.

The paper begins by briefly describing the main differences between MCM and other high-density packaging technologies, and then introduces the MCM process used for this implementation. Next, general overviews of the learning model and its hardware implementation are presented. Finally, several important issues pertaining to this general implementation approach are briefly explained.

2 High-density interconnection

In this section, the characteristics of multi-chip module packaging technology are briefly reviewed. Then, two other technologies that offer high-density interconnect are presented and compared to MCMs: wafer-scale integration and high-density printed circuit boards. Finally, the specific multi-chip module process used for this implementation is described.

2.1 Overview of multi-chip modules

The typical process used in the production of electronic systems consists of fabricating many identical ICs on a single wafer, separating and packaging the ICs, and then externally connecting the packaged ICs on some type of interconnect medium, such as a printed circuit board (PCB). In an MCM, unpackaged ICs are mounted on a substrate which contains multiple layers of interconnect. The ICs are mounted to the substrate and electrically connected (by wire-bonding or other methods) to the substrate [4, 5].

Multi-chip module packaging technology is an important technique which results in high chip density, small interchip propagation delay, low power consumption, and high interconnect density. A four-layer thin film MCM process may have a maximum interconnection density as high as 300–800 linear centimeters of interconnections per square centimeters of die area (cm/sq. cm) [4, page 97]. In addition, as will be explained in Section 3.3, an MCM can accommodate many different types of ICs, such as logic, memory, analog, etc. For these reasons, an MCM approach was taken for the system implementation described in this paper. As will be discussed later, other neural network models may also benefit from the characteristics of MCM packaging technology described above.

2.2 Overview of other technologies

Wafer-scale integration (WSI) and printed circuit boards (PCBs) are two other packaging approaches that can have relatively high interconnect densities. However, these methods suffer from other disadvantages compared to MCMs.

In WSI, ICs *and* their interconnect are fabricated on a wafer concurrently. The result is a very dense collection of ICs and interconnect which can be packaged as a single unit [5]. As with MCMs, this technique results in high chip density, small interchip propagation delay, low power consumption, and high interconnect density, typically somewhat higher than the interconnect density that can be obtained with MCMs. However, a WSI fabrication approach has several limitations: it suffers from relatively low yield, it demands a high entry-level cost to evaluate, and the types of ICs fabricated must be very similar. Due to this last constraint, WSI has proven to be primarily useful in applications where a large number of identical or similar ICs are required. Many connectionist systems, such as neural networks, fall into this category [1, 3, 5, 14]. However, as will be explained later, some connectionist models may benefit from the ability to use different types of ICs in a hardware implementation.

A printed circuit board is another approach that can be used to implement neural networks. Even though this approach offers high yield and is relatively simple and inexpensive compared to MCMs, it suffers from lower wiring densities and therefore may not be capable of providing the high interconnect density required for some connectionist system applications. One estimate of the interconnect density of a 20-layer PCB is about 140–260 cm/sq. cm [4, page 97], which is substantially lower than the density that can be obtained using MCMs. In addition, if packaged components are mounted on PCBs, much more space is required than a comparable system using MCMs and unpackaged die.

2.3 Specific MCM approach

A hardware implementation of the PASOCS model is being developed at Brigham Young University. The implementation consists of three identical die mounted on an MCM substrate; future implementations will use different types of ICs as will be explained later. The die are 2μm digital CMOS ICs fabricated through the MOSIS service and are 2.5mm $\times$ 2.7mm in dimension.

The entire MCM structure used in this implementation was fabricated in the class-10 clean room facility of the Integrated Microelectronics Laboratory (IML) at Brigham Young University. The MCM substrate was fabricated using two levels of the standard four-level metal IML process as described below. After substrate fabrication, the custom ICs are mounted to the MCM substrate and tested using a digital tester. The process used for this PASOCS implementation was optimized for ease of fabrication and maximum interconnect density. As a result, comparatively thin metal layers were used. Only two levels of metal were used in this three-die feasibility study, but future applications will require greater interconnect density and will use the additional metal layers available in the process. The entire four-level process with 40μm lines and pitch results in an interconnect density of about 500 cm/sq. cm.

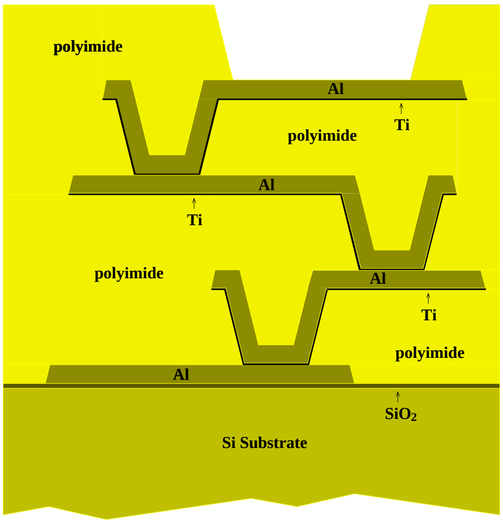

Figure 1: Cross-section of the IML MCM substrate. Although the four-level process is shown, only two levels of metal were required for this implementation.

A cross section of the IML MCM substrate process is shown in Figure 1. The thicknesses of the various layers are to scale, although they are not at the same scale as the lateral dimensions. The slope of the vias and metal layers is also approximately correct. Deviation from vertical is caused by the characteristics of the metal and via definition etches.

The first step of the process is the growth of a 1250Å layer of SiO_2 on the MCM substrate to isolate the interconnect structure from the conductive silicon substrate. Following SiO_2 growth, the first level metal layer of 2μm of aluminum is sputter deposited and patterned using a wet etch process. The first level metal layer is followed by spin coating of the substrate with an 8μm interlevel dielectric of Dupont 2611D polyimide. Vias are plasma etched in the polyimide to provide connection between the first and second level metals. Following via definition, a second level of 2μm TiAl is deposited and wet etched. In the full four-level metal process, the polyimide/metal steps are repeated two additional times, although in this application only two levels of metal were required. Following definition of the final metal layer, a top coat of polyimide is spin deposited for protection of the interconnect structure, and pad vias are etched to provide access for wirebonding.

3 The learning model

In this section a brief overview of the PASOCS model is given. Then, a description of the hardware implementation, including the individual VLSI ICs and the final MCM system, is presented. Finally, important characteristics of this implementation approach are examined which are applicable to other connectionist models.

3.1 Overview

The system implemented as an MCM is based on the connectionist architecture ASOCS. The primary goal of an ASOCS is similar to that of many decision-making connectionist systems—the system attempts to learn an arbitrary set of vector mappings. However, an ASOCS differs from many other connectionist systems in that it learns by the introduction of rules rather than a training set of input/output vectors. The system is able to learn by keeping itself consistent with the rules and by dynamically changing its topology as new rules are introduced. In this way, an ASOCS can change its structure to suit a particular problem.

The particular model implemented in this study is PASOCS, which is one of a class of ASOCS connectionist models. See [7, 8, 9] for background information on the ASOCS and PASOCS models. Briefly, a PASOCS is a network of self-organizing digital processing elements (or *nodes*) which accomplishes the following [9]:

- processes inputs in the form of boolean variables and outputs boolean results;
- accepts rules made up of a conjunction of boolean inputs which imply a boolean output;
- learns new rules over time;
- automatically resolves rule conflicts;
- combines specific rules into more general rules where appropriate; and
- maintains an associated priority with each rule.

In order to accomplish these goals, the nodes in the network self-modify using local information. The rules (inputs) to a PASOCS are stored in the nodes. Typically, a practical implementation of a PASOCS would require hundreds or thousands of nodes which are densely interconnected. This is one of the requirements that makes a multi-chip module implementation advantageous.

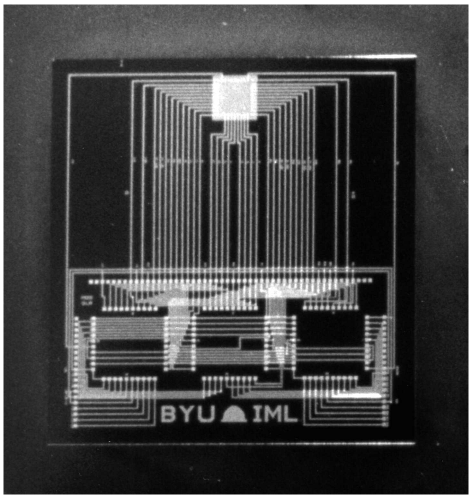

Figure 2: The multi-chip module substrate.

3.2 Hardware implementation and testing procedures

Each IC described in Section 3.1 contains the functional hardware for one node of a PASOCS. Details of these ICs can be found in [13]. For this study, four packaged ICs and four unpackaged ICs were fabricated. First, the individually packaged ICs were tested and compared to the simulation results obtained during the initial design. Then, three of the packaged ICs were connected on a PCB as a three-node PASOCS and tested. Next, as a prerequisite to mounting the ICs on the MCM substrate, the unpackaged ICs were tested separately. The three ICs are mounted on the MCM substrate using a die epoxy and tested by using a pad ring on the MCM which allows access to all of the I/O of the PASOCS. This pad ring can be seen at the top of Figure 2. The 40-pin die are mounted and wire-bonded to the three pad rings at the bottom of the figure.

The results of the tests have generally been positive. Most of the functions of this three-node PASOCS are performing according to original design specifications except for two functions associated with overall network minimization and rule relationships. It should be noted that these are problems with the specific implementation of the ICs and not with the functionality of the PASOCS model or with the original conceptual design of the ICs as shown in [13].

3.3 Important characteristics of an MCM approach

MCM technology has important characteristics that make it an attractive option for implementing connectionist systems. Some of these, which have already been mentioned in Section 2.1, include high chip density, small interchip propagation delay, low power consumption, and high interconnect density. Another important characteristic is the potentially high yield that can be obtained from this approach as opposed to WSI. Additionally, improvements in bare die testing techniques will help make higher MCM yield possible [6, 10], since this will assure that only KGD (known-good die) are mounted on the MCM. The die in this project were individually tested using an approach which is not appropriate for general manufacturing requirements.

Another important advantage of this approach is the ability to use different types of ICs in the design. The model described in this paper, for instance, includes both logic and memory circuits on the individual ICs. The memory on the ICs, however, is bulky and inefficient. The memory requirements would be met much more elegantly if smaller and denser memory cells (DRAMs, for instance) were used. The logic could be designed on CMOS ICs and DRAMs could be mounted next to these logic ICs on the substrate. In addition, other connectionist systems, including artificial neural network models whose general equations are discussed in Section 1, may also benefit from this ability. In some cases, it may be most efficient to fabricate the neurons and associated logic circuits in digital CMOS, while the adjustable weight ICs may be more easily or efficiently implemented as DRAMs or some type of analog device. The neural network could then be built with these different types of ICs using an MCM as the interconnect scheme.

Present research at Brigham Young University includes testing and simulation of the ideas presented as well as initial research into other connectionist systems that may benefit from an MCM implementation approach. Current work has shown that the general ideas presented are versatile and can be modified to reflect other models [12].

4 Conclusion

This paper described a multi-chip module implementation of a connectionist system. This system differs significantly from many other commercially available neural network systems and research projects, since MCMs are used as the interconnect medium.

The paper described the specific MCM process currently in use by the Integrated Microelectronics Laboratory at Brigham Young University. The specific connectionist system implemented as an MCM is also described. The general ideas pertaining to MCM implementation of connectionist systems also apply to other models, including artificial neural networks.

This research was supported in part by the Endowed Chair of Engineering, occupied by Dr. Linton G. Salmon, and by grants from the National Science Foundation and Novell, Inc.

References

[1] M. Campbell et al. 3-D wafer stack neurocomputing. In *Proceedings of the IEEE International Conference on Wafer Scale Integration*, pages 67–74, 1993.

[2] DARPA. Neural network study. AFCEA Int. Press, 1988.

[3] F. Distante et al. A general configurable architecture for WSI implementation for neural nets. In *Proceedings of the IEEE International Conference on Wafer Scale Integration*, pages 116–23, 1990.

[4] D. A. Doane and P. D. Franzon, editors. *Multichip Module Technologies and Alternatives: The Basics.* Van Nostrand Reinhold, 1993.

[5] D. Herrell and H. Hashemi. *Frontiers in Computing Systems Research.* Plenum Press, 1990.

[6] D. Keezer. Bare die testing and MCM probing techniques. In *Proceedings of the IEEE Multi-Chip Module Conference*, pages 20–3, 1992.

[7] T. R. Martinez and D. M. Campbell. A self-adjusting dynamic logic module. *Journal of Parallel and Distributed Computing*, 11(4):303–13, 1991.

[8] T. R. Martinez and D. M. Campbell. A self-organizing binary decision tree for incrementally defined rule based systems. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(5):1231–7, Sep/Oct 1991.

[9] T. R. Martinez, D. M. Campbell, and B. W. Hughes. Priority ASOCS. To appear in *Journal of Parallel and Distributed Computing*, 1993.

[10] R. Parker. Bare die test. In *Proceedings of the IEEE Multi-Chip Module Conference*, pages 24–7, 1992.

[11] U. Ramacher and B. Schürmann. Unified description of neural algorithms for time-independent pattern recognition. In U. Ramacher and U. Rückert, editors, *VLSI Design of Neural Networks*, pages 255–70. Kluwer Academic Publishers, 1991.

[12] G. Rudolph and T.R. Martinez. An efficient transformation for implementing multilayer feedforward neural networks. Brigham Young University TR#: BYU-CS-93-14.

[13] M. G. Stout. An approach to connectionist system hardware implementation using a multi-chip module interconnection and packaging scheme. Master's thesis, Brigham Young University, April 1994.

[14] M. Yasunaga. A self-learning neural network composed of 1152 digital neurons in wafer-scale LSIs. In *IEEE Joint Conference on Neural Networks*, volume 3, pages 1844–9, 1991.