An Analogue Neural Network using MCM Technology

Marwan Jabri, Philip Leong, Jim Burr [†], Barry Flower, Kam K. Lai,
Stephen Pickard, Edward Tinker and Richard Coggins

SEDAL, University of Sydney
[†] Starlab, Stanford Unversity

Abstract

This paper describes a large analogue neural network implemented using multichip module (MCM) technology. An array of 8 synapse chips, each containing 2500 synapses are connected to a neuron chip on an MCM to implement a (150,100,50) network containing more than 1 million transistors.

1 Introduction

The use of analogue VLSI technology to implement artificial neural networks has recently become a topic of active research interest. In order to produce even larger neural networks than possible on a single VLSI chip, several such chips can be interconnected. The interconnection of neural network chips can be achieved using normal printed circuit board (PCB) technology. However, the long wire lengths associated with PCBs introduces parasitics which degrade interchip signals, and pinout restrictions of the VLSI packaging limit the effectiveness of such an approach.

In order to address this problem, we use multichip module (MCM) technology [1] which enables multiple chips to be interconnected on the same substrate. MCMs have the following advantages over printed circuit board technology

- chips can be placed physically closer together, minimising the interchip wire length
- wiring density is improved
- chips with larger numbers of input/output pads can be accommodated
- the overall size of the design is reduced

This paper describes the scaling up of earlier designs [2, 3] to an MCM which contains more than 1 million transistors in a single package. The network size implemented is a (150,100,50) three layer perceptron.

Technology	MCM array of 8 7.2 × 7.2 mm synapse chips plus 1 7.2 × 7.2 mm neuron chip
Topology	(150,100,50) multilayer perceptron
Chip Technology	1.2 micron, 2 metal, 2 poly
Chip array	[3,2,1]
Modules	Built from two modules (8 × SYN & 1 × NEU)
Inputs	All are analogue
Output	All are analogue
Weight update time	one weight per bus cycle
Synapses	20,000 (120,000 bits)
Synapse design	6 bit MDAC
Neuron design	Switchcap with single ended output to cancel common mode

Table 1: MCM summary.

2 Architecture

The neural network implemented on the MCM is a (150,100,50) perceptron. Networks smaller than this can be obtained by setting weights to zero. The MCM consists of two separate VLSI chips, a synapse (SYN) chip and a neuron (NEU) chip, each being 7.2×7.2 mm in dimension. Table 1 presents a summary of this architecture.

The MCM is used to provide interchip connections between synapse and neuron chips. The first layer of the neural network is made from a 3 × 2 array of synapse chips and the second layer is a 2 × 1 array of synapse chips. Neurons are required in the middle and output layer of the design and routing for this is performed on the MCM.

0-8186-4260-2/93 $03.00

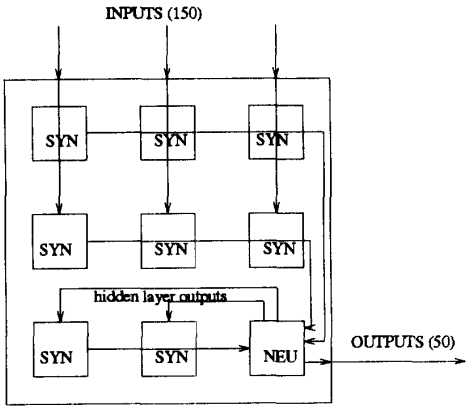

Figure 1: MCM Floorplan

2.1 Synapse Chip

The synapse (SYN) chips contain a 50×50 synapse array of 6 bit MDACs, current source and a set of row and column shift registers for weight addressing. These chips dominate the area of the MCM, and eight of these are used to form 20000 synapses.

Inputs to the synapse chips are voltages and outputs are currents, the design being the same as an earlier chip designed by the authors [2]. The synapse cell is a multiplying digital to analogue converter which can operate with bias currents of the order of 1 nA (see Figure 2). Weights can be written to a register in the synapse and the synapse multiplies the weight by the analogue voltage inputs to produce a current output. All synapses which connect to the same neuron are summed using Kirchoff's current law and neurons are used to convert this current into a voltage which is used as the inputs to the next layer of the neural network. The transfer function of the network can be described by the following equations

$$u_i = \sum_{j=1}^{N_l} w_{ij} \tanh(\frac{\kappa(a_j)}{2}) \quad (1)$$

$$a_i = \alpha u_i \quad (2)$$

where u_i is the summed output of the synapses, a_i is the neuron output, κ and α are constants, l denotes the lth layer ($0 \leq l \leq L-1$), L = total number of layers, N_l = number of neuron units at the lth level and i is the neuron number ($1 \leq i \leq N_l$).

Synapse chips are cascaded to build larger networks, inputs to the neural network running vertically and outputs horizontally so that a regular connection scheme can be realised.

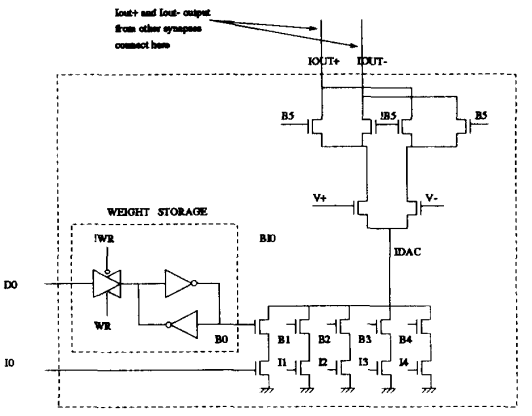

Figure 2: Synapse Circuitry.

Inputs and outputs from the synapse chip are differential and since there are 50 inputs and 50 outputs from this chip more than 200 pins are required. Using standard wirebond packaging, even with minimally spaced pads, a pinout of 132 can be achieved since the pads must be on the periphery of the chip.

Flip–chip technology, whereby contacts to the chip are made by small bumps on the MCM substrate, enables connections to be made to pads anywhere on the chip, and the required pinout is easily achieved.

Although flip–chip technology offers a solution to the high pinout problem, testing of the individual synapse chip dies remains a problem. This problem is addressed by using a customised probe card. This card has pins to which the die pads can be connected, and thus testing of the individual dies can be effected.

The synapse chip has been designed using 1.2 micron single poly, double metal technology. A chip plot is given in Figure 3. The design has been sent for fabrication. The synapse chip contains more than 140,000 transistors.

2.2 Neuron Chip

The NEU chip is a chip which simply contains an array of neurons which converts the accumulated sum of the synapse output currents to a voltage which can fanout to the next layer of synapses.

The synapse outputs are differential current sinks and the common mode value of the neuron outputs cannot be allowed to drop below the working range of the synapse inputs. We call this the "common mode

problem", and any neuron circuit must ensure that its outputs are within the synapse's input range.

The neuron design has not yet been finalised, but it is likely to be a circuit similar to the switched capacitor neuron described in [3] which offers the advantages of simplicity as well as having gain which is adjustable over a wide range.

3 Training

Although the MCM neural network could be applied to many different problems, we have concentrated on using NETtalk as an example [4] to test our architecture. NETtalk is a good problem for the MCM since it requires a moderate synapse resolution and has many inputs and outputs. The architecture used was (203,120,26) in size – larger than the MCM design. The MCM will implement NETtalk with a reduced alphabet, replacing characters with sequences of characters which make the same sound.

We have a simulator (MUME) which mathematically models the transfer function of the synapses using Equation 2. Although second order effects are not included in the model, we have found very close correspondence with the hardware in our previous chips.

MUME was used to simulate NETtalk using online backpropagation on a SUN Sparcstation 2 in approximately 50 hours of CPU time (130 iterations). Our previous chip had only 84 synapses, and we could train this without any gradient information [2]. However, the NETtalk problem is larger by three orders of magnitude, and so an analytic gradient of the mathematical model was used for training.

4 Conclusion

We have described the design of a large analogue neural network which uses MCM technology to provide interchip connections. MCM technology enables a higher synapse density to be achieved and benefits which include smaller area and reduced parasitics can be enjoyed. The MCM is under construction and its completion will provide a large analogue neural network in a single package which can be used for many applications which require such networks.

References

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

[2] P.H.W. Leong and M.A. Jabri. A low power analogue neural network classifier chip. In *Proceedings of the IEEE Custom Integrated Circuits Conference*, pages 4.5.1–4.5.4, San Diego, USA, May 1993.

[3] R.J. Coggins and M.A. Jabri. A comparison of three on chip neuron designs for a low power VLSI mlp. In *Proceedings of the Third Internation Conference on Microelectronics for Neural Networks*, pages 97–103, Edinburgh, UK, April 1993.

[4] Terry Sejnowski and Rosenberg C.R. NETtalk: A Parallel Network That Learns to Read Aloud. Technical report jhu/eecs-86/01, Johns Hopkins University, 1984.

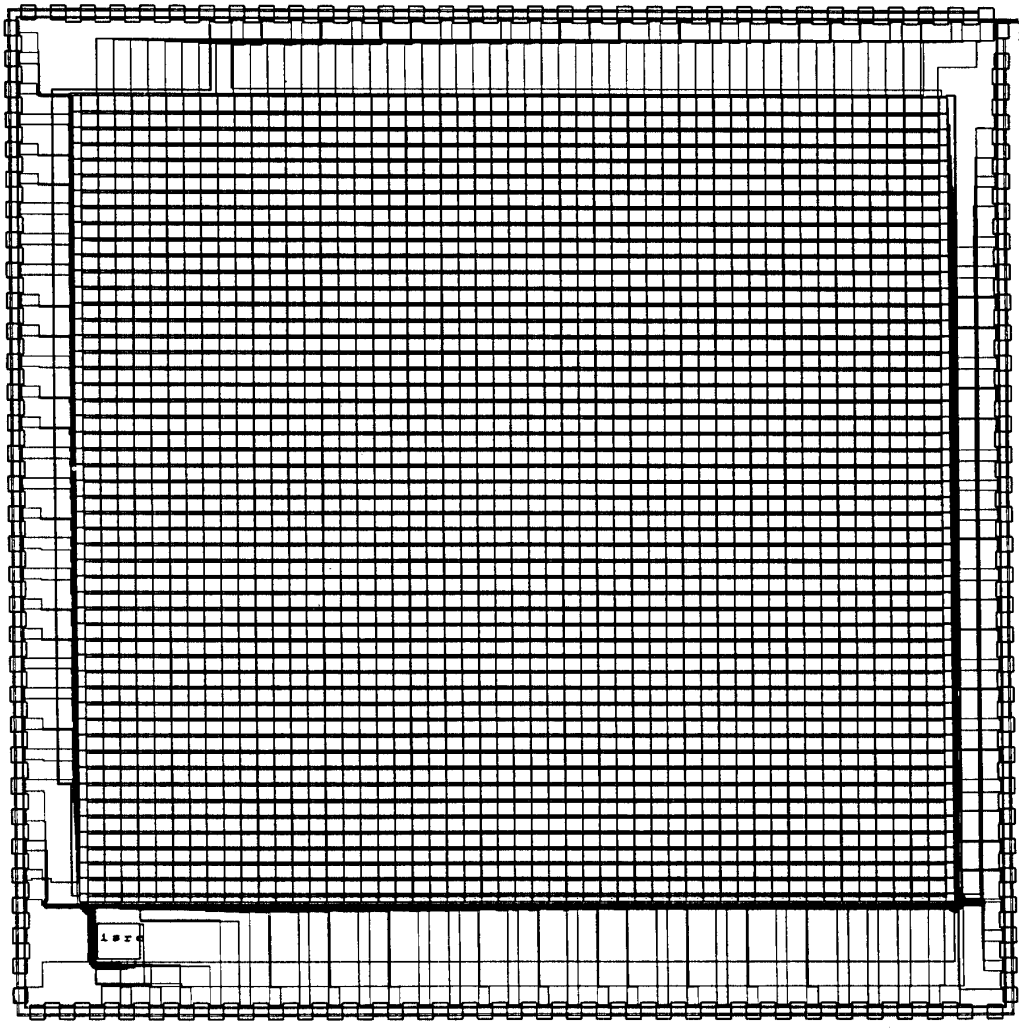

Figure 3: Synapse Chip Plot.