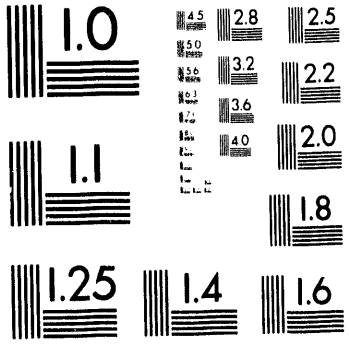
1.0
2.8
2.5
3.2
2.2
3.6
2.0
4.0
1.1
1.8
1.25
1.4
1.6

1 of 1

Conf-921188--3

UCRL-JC-110366
PREPRINT

Customizable VLSI Artificial Neural Network Chips Based on a Novel Technology

Chi Yung Fu
Benjamin Law
George Chapline
LLNL
and
Dick Swenson
Naval Air Warfare Center
China Lake, CA

This paper was prepared for submittal to the
1992 Government Microcircuit Applications Conference
Las Vegas, Nevada
November 9-12, 1992

September 14, 1993

Lawrence Livermore National Laboratory

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

MASTER

DISTRIBUTION OF THIS DOCUMENT IS UNLIMITED

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial products, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

CUSTOMIZABLE VLSI ARTIFICIAL NEURAL NETWORK CHIPS BASED ON A NOVEL TECHNOLOGY

Chi Yung Fu, Benjamin Law, ana George Chapline
Lawrence Livermore National Laboratory, L-271
University of California, Livermore, CA 94550

Dick Swenson
Naval Air Warfare Center, Code 3914
China Lake, CA 93555

Introduction

The human cerebral cortex contains approximately 10^{11} neurons and 10^{14} synapses[1]. It thus seems logical that any technology intended to mimic human capabilities should have the ability to fabricate a very large number of neurons and even larger numbers of synapses. This paper describes an implementation of hardware neural networks using highly linear thin-film resistor technology and an 8-bit binary weight circuit to produce customizable artificial neural network chips and systems.

Artificial Synapses

We use a resistive material to form the synapse. This approach reduces the number of masks from a typical 18-mask EEPROM process[2] down to only 5. Such a great reduction in complexity allows us to scale the technology up for the implementation of an immense number of resistors and still maintain a reasonable yield. By using an 8-bit binary weighted representation of a synapse, we are able to define 256 synaptic strengths with 8 identical resistors.

We have designed and successfully fabricated two prototype artificial synapse structures based on our resistor technology. The first structure is based on isolated resistors, allowing us to fully characterize each of the 315 resistors. The second one is a resistor array containing 34,816 resistors with a packing density of 2×10^5 resistors/cm^2. We can control resistance values from 200,000 Ω to 2 MΩ.

Measurements of the 315-resistor structure yielded a standard deviation of 0.79% for resistance measurements. The standard deviation for linearity of the resistors over a +/- 3 V range is 0.08% (Fig. 1).

Input Buffers and Output Neurons

All the neurons and input buffers on the IC were designed with 2-μm CMOS technology. The neuron body amplifiers contain a binary-weighted summer, a sigmoid limiter, and an output amplifier (Fig. 2). The summing section is designed to have an 8-bit binary weighting function. The average measured error is 0.6 LSB. The sigmoid limit voltages are set by the user through external pins. There are 128 possible gain selections, ranging from unity to 1024. Each leg of the differential amplifier is capable of driving 128 8-resistor synapses. The average amplifier voltage output offset error was 7 mV.

Laser Programming

The strength of each synapse is set by using a computer guided laser system to cut out the appropriate resistors of the 8 resistors for each synapse. Figure 3 shows a scanning electron micrograph of a resistor cell after a laser cut. The present programming speed for a 50% cut of the 34,816 resistor array is 1.5 minutes. The third generation of the resistor structure will have a fourfold increase in density. Combined with a faster mechanical stage, this would allow us to program 1 million resistors in 8 minutes, assuming again that 50% of the resistors will be cut and excluding the time needed for wafer loading, pump down, unloading, and alignment.

In addition, our laser technologies allow us to connect or disconnect the paths between any two pairs of points[3] and thus allow us to define different architectures. The combination of laser-programmable synaptic weights and laser-defined architecture defines a customizable neural network and thus enables us to customize different applications in a time frame of minutes to hours.

A 96-Neuron Architecture

Our customizable chip architecture includes an input buffer amplifier for each input line, a fully-connected synaptic array, and the output neurons,

representing a single-layer neural network (Fig. 4). It has 96 neurons and 12,288 synapses. This chip has the capability of having multiple transfer functions on a single layer, and thus allows us to implement multi-layer networks on our single layer chip. The chip is expected to operate at a throughput of 1.2 x 10^{10} interconnects/sec dissipating 0.7 W per chip. Furthermore, very large systems can be built by cascading these chips together.

Conclusions

We have demonstrated a novel approach to the development of VLSI artificial neural network hardware.

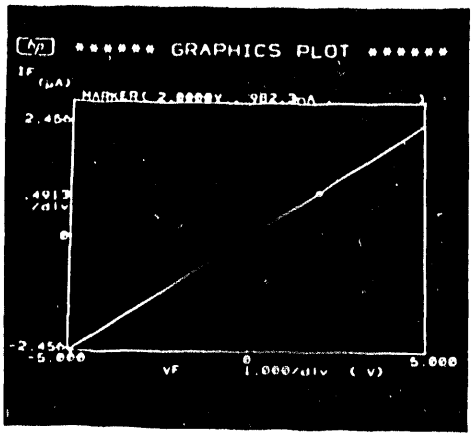

Fig. 1. Electrical characteristics of a 2 MΩ resistor. The standard deviation for linearity of the resistors over a +/- 3 V range is 0.08%.

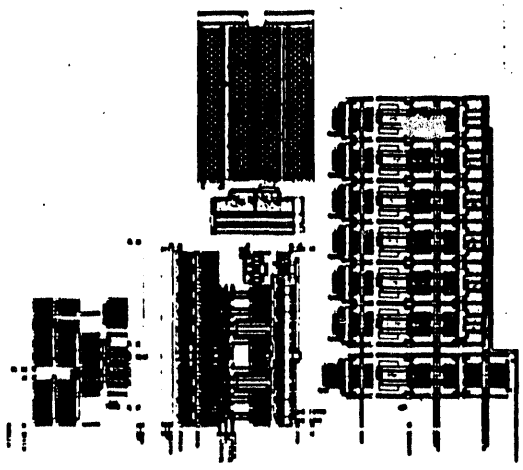

Fig. 2. An output neuron body amplifier. The neuron body amplifiers contain a binary-weighted summer, a sigmoid limiter, and an output amplifier.

This work was performed under the auspices of the DOE by Lawrence Livermore National Laboratory under contract No. W-7405-ENG-48.

References

1. "DARPA Neural Network Study," Armed Forces Communications and Electronics Association International Press, Nov. 1988.
2. M. Holler, et al., "An Electrically Trainable Artificial Neural Network (ETANN) with 10240 "Floating Gate" Synapses," Proc. IEEE International Conference on Neural Network, Washington, D.C. July 1987.
3. C. Y. Fu, et al., "A Laser Direct Write Double-Level-Metal Technology for Rapid Fabrication," IEEE Custom Integrated Circuits Conference, Boston, MA, May 1990.

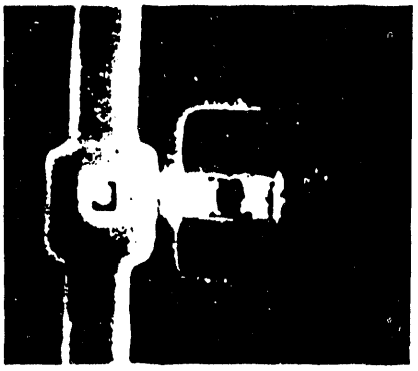

Fig. 3. SEM micrograph of a resistor cell after a laser cut. The present programming speed for a 50% cut of the 34,816 resistor array is 1.5 minutes.

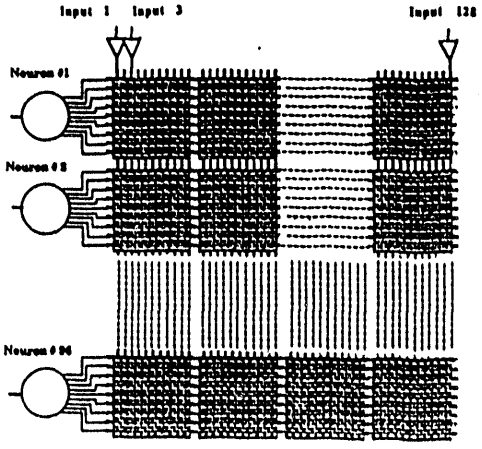

Fig. 4. Block diagram for a 96-neuron chip. The chip is expected to operate at a throughput of 1.2 x 10^{10} interconnects/sec dissipating 0.7 W per chip.

DATE
FILMED
11/30/93

END