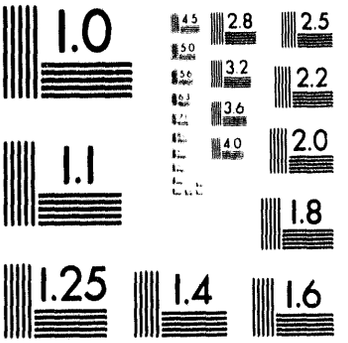
1.0
2.8
2.5
3.2
2.2
3.6
2.0
1.1
1.8
1.25
1.4
1.6

1 of 1

Conf-9206102--6

UCRL-JC-109597
PREPRINT

A Novel Technology for Fabricating Customizable VLSI Artificial Neural Network Chips

Chi Yung Fu, Benjamin Law, and George Chapline
Lawrence Livermore National Laboratory

Dick Swenson
Naval Air Warfare Center
China Lake, California

This paper was presented at the
IJCNN '92 Meeting, Baltimore, MD
June 7, 1992

February 5, 1992

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.

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.

A Novel Technology for Fabricating Customizable VLSI Artificial Neural Network Chips

Chi Yung Fu, Benjamin Law, and 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

Abstract

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. These neural networks are programmed using precision laser cutting and deposition. The fast turnaround of laser-based customization allows us to explore different neural network architectures and to rapidly program the synaptic weights. Our customizable chip allows us to expand an artificial network laterally and vertically. This flexibility permits us to build very large neural network systems.

Overview

The human cerebral cortex contains approximately 10^{11} neurons and 10^{14} synapses [1]. This extremely high degree of parallelism is perhaps the main factor that gives humans their powerful cognitive capabilities. It thus seems logical that any technology intended to mimic human capabilities and intelligence should have the ability to fabricate a very large numbers of neurons and an even larger numbers of synapses. Furthermore, one must be able to program the synaptic strengths, as these strengths contain the information that controls the function of the neural network.

Our approach is to address the synaptic problem with a relatively simple and yet highly reliable and scalable technology. We use a resistive material to form the synapse. This approach reduces the number of masks from a typical eighteen-mask EEPROM process down to only five. 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. This scheme allows us to build synaptic arrays based on single-valued resistors and thus further simplifies the processing issues, making possible the fabrication of a complex artificial neural network chip. Our resistor technology therefore offers the advantages of linearity over a wide range of operations, and simple fabrication processing which can be scaled to extremely large and complex neural network chips or systems.

A wide variety of neural network applications do not require synaptic weight variation once the application is fixed. This makes the fixed resistors weighting ideally suited for these types of applications. Well-known examples may be found in the areas of signal processing and artificial vision. Also, our proposed synaptic technology is not exactly a fixed-weight scheme. It is a programmable technology using direct laser cutting and laser-induced deposition [7,8] to define the synaptic strengths. In addition, these laser technologies allow us to connect or disconnect the paths between any two pairs of points and thus allow us to define almost any neural network architecture, for example, a Hopfield net with feedback paths or a multi-layer feedforward net.

MASTER

The combination of laser-programmable synaptic weights and laser-defined architecture allows us to define the concept of a customizable neural network. Such a customizable chip can be prefabricated, and thus enable us to customize different applications in a short time frame of minutes to hours, depending on the complexity of the neural circuits. This fast prototyping of new designs should greatly accelerate the development of neural network hardware and applications. Such a combination of design and technology would also allow rapid, cost-effective, low-volume prototype production. This approach will permit all functional defects to be addressed and corrected before beginning high-volume production of a neural network.

Artificial Synapses

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.

Measurements of the 315-resistor structure yielded a standard deviation of 0.79% (Fig. 1) for resistance measurements. This sampling covers the equivalent of an area that will contain 68,000 resistors for the resistor array (Fig. 2). The standard deviation for linearity of the resistors over a +/- 3 V range is 0.08% (Fig. 3).

Our success has several implications. First, our demonstration of large arrays of resistors in the range of ~200,000 Ω to 2 MΩ is a significant step toward artificial neural networks with sufficient complexity to mimic human capabilities. Second, the extremely tight tolerance in uniformity means that the laser programming of the chip can be done in one single pass. Third, the highly uniform synaptic value distribution thus obtained suggests that we can build an even larger chip and still retain excellent uniformity as we improve our processing techniques. Fourth, the high degree of linearity makes it easier for potential users to transfer their neural network designs from standard software simulations to hardware. Fifth, our wider range of linearity implies better dynamic range, which simplifies the neuron body circuit design.

Input Buffers and Output Neurons

All the active circuitry on the IC was designed in CMOS for 5 V operation and laid out for 2-micron, P-well, double-poly, double-metal fabrication using typical design rules. These circuits consist of input buffer amplifiers (Fig. 4) and the neuron body output amplifiers (Fig. 5). The neuron body amplifiers contain a binary weighted summer section, a sigmoid limiter, and an output buffer amplifier. The only difference between the output buffer amplifier section and the input buffer amplifier is the fixed unity gain of the input amplifier.

All amplifiers are fully differential about an user-selectable reference voltage (Vref). For minimum power dissipation and maximum linearity the best value for Vref is 1.5 V, with the sigmoid limits set at 0.5 V and 2.5 V. This gives the neuron amplifiers a dynamic range of +/- 1.0 V. Although the resistors are linear over a much greater range, the 2.0 V output limit was selected to maximize the amplifier linearity and minimize the resistor power dissipation. The sigmoid limit voltages are common to all output amplifiers on the chip and are set by the user through external pins. The sigmoid limits can be disabled at any amplifier during laser cutting if linear processing is desired on some of the neurons. The amplifier gain on each neuron body amplifier is selectable by laser cutting of the feedback resistors. 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 at the worst case differential of 2.0 V. Each resistor will dissipate between 0 to 2 μW. The power dissipation of each resistor is due to the voltage differential between its input and the average voltage of all the inputs on the summing line. The summing section at each neuron body amplifier is designed to have an 8-bit binary weighting function with a maximum error of one least significant bit (LSB). The average measured error is 0.6 LSB. The summer error must not be greater than the resistor

fabrication tolerances. The amplifier voltage output offset error was found to be an average of 7 mV and the worst error measured was only 16 mV. These measurements indicate that each neuron amplifier can easily be nulled to Vref with a single bias weight, although nulling does not appear to be necessary with such a small offset error.

Laser Programming

Our laser apparatus was designed and built with the goal of patterning features as small as 1 micrometer. The beam source for the system is a water-cooled argon laser, operated at a 514 nanometer wavelength. The laser beam is positioned within a scanning window by applying rf modulation on the two acoustic optics scanner crystals to separately control the horizontal and vertical scans. After the window is scanned, the stage then moves to a new position for the patterning of another window.

Figure 6 shows scanning electron micrographs of a resistor cell before and after a laser cut. No metal traces or debris can be seen with either an optical or a scanning electron microscope. X-ray microanalysis of the cut region reveals no traces of metals either left on the surface or originating from any possible reaction of the metal with the underlying oxide because of the extreme high reaction temperature induced by the laser. Electrical testing confirms that these resistors are open circuits after laser cutting. The present programming speed for a 50% cut of the 34,816 resistor array is 1.5 minutes. One major reason for the relatively slow speed is due to the slow settling time of the mechanical stage. Another way to improve the cutting speed is to increase the packing density of the resistors.

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 0.25 million resistors in 2 minutes, assuming again that 50% of the resistors will be cut. Based on this projected speed, the time needed to cut 1 million resistors will be 8 minutes, excluding the time needed for wafer loading, pump down, unloading, and alignment. We believe this programming speed is high enough to give a fast turnaround for prototyping a specified set of synaptic weights for a neural network chip. A further substantial improvement in programming speed can be achieved by using a multi-beam laser system.

The synaptic strengths are defined by the laser cutting and laser deposition for repair and rework. This implies that synaptic strength variation due to charge leakage or hot electron is not a factor. This inherent reliability is a major consideration for using artificial neural network technology in areas such as space exploration or smart missile.

A 96-Neuron Architecture

With all the components and programming technology successfully demonstrated, we are now integrating these components to form a neural network chip.

While a self-contained single-chip neural network with I/O interface built in may seem appealing, we believe that such chip may offer only a limited growth potential for addressing complex future neural networks. Since our goal is to build extremely large and complicated neural network systems, we decided to develop a modular neural network structure. This modular approach is somewhat similar to NAND or NOR gates in digital technology. Our customizable module thus becomes the basic building block for complex neural network systems. This chip contains an input buffer amplifier for each input line, a fully-connected synaptic array, and output neurons, representing a single-layer neural network. Our current design has 96 neurons and 12,288 synapses in one single chip (Fig. 7). Based on current design rules, the chip's dimensions are 1.25 by 1.25 cm. Only the passive resistors and metal interconnects are positioned in the center section of the integrated circuit. All active circuits lie on the boundary of the IC. Power distribution is thus confined to the periphery, minimizing bussing problems and voltage variations across the chip. In most architectures, the input amplifiers would only be used for the first layer. The output amplifiers would drive the synapse resistors in the next layer.

With this single customizable chip we can build many different architectures. With feedback paths formed by our fast-turnaround laser processing, we can readily implement a Hopfield net. The amplifier gain on each neuron body amplifier is individually selectable by laser cutting of the feedback resistors. This gives the capability of having multiple transfer functions on a single layer, and thus allows us to implement multi-layer networks on our single layer neural network chip, provided that the total number of neurons in the application is equal to or less than 96 and the number of synapses is no more than 12,288. This is possible because the feedback resistors are physically identical to the synapse resistors, and therefore have the same fabrication tolerances.

Our present projected power consumption for the resistor array is 0.2 W maximum and that for the active circuitry for input buffers and output neuron bodies is less than 0.5 W. This implies that no special provision is needed for heat dissipation. Our current design is expected to operate at a minimum of 1 MHz. As a result, the throughput of the system for a single chip will be at least 1.2×10^{10} interconnects/sec.

Conclusions

We have described a novel approach to the development of VLSI artificial neural network hardware. With all components and programming technology successfully demonstrated, we are integrating these components to form a customizable neural network chip. Our processing technology yields resistors with a highly linear behavior over a wide dynamic range. Laser processing provides an unique and extremely rapid means of defining the synaptic weights of a neural network as well as its architecture. The 8-bit binary weight scheme allows us to define synaptic strengths with sufficient resolution. The customizable chip combined with laser processing will allow fast-turnaround customization of neural network chips for many different applications. A fast fabrication process of this kind gives system designers a high degree of freedom and flexibility to prototype neural network hardware, and thus allows timely insertion of neural networks into different applications. Our approach may be extremely cost-effective for low-volume production. Finally, another major advantage of this unique approach is the ability to scale up to extremely large and complex neural network systems such as an artificial retina.

Acknowledgement

The authors wish to thank Dr. L. Petrich for his assistance and discussion. The 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. R. P. Lippmann, "An Introduction to Computing with Neural Nets," IEEE ASSP Magazine, April 1987.
3. M. Holler, S. Tam, H. Castro, and R. Benson, "An Electrically Trainable Artifiical Neural Network (ETANN) with 10240 "Floating Gate" Synapses", Proc. IEEE International Conference on Neural Network, Washington, D.C. July 1987.
4. L.D. Jackel, R.E. Howard, H. P. Graf, B. Straughn, and J.S. Denker, "Artificial Neural Networks for Computing," J. Vac. Sci. Tech. B 4 (1), Jan/Feb 1986, p 61-63.
5. J. L. Lamb, A. P. Thakoor, A. Moopenn, and S.K. Khanna, "Resistive Synaptic Interconnects for Electronic Neural Networks," J. Vac. Sci. Tech. A 5 (4), Jul/Aug 1987, p 1407-1410.
6. C. Mead, *Analog VLSI and Neural System*, Addison Wesley, 1989.

7. B. McWilliams, C. Y. Fu, F. Mitlitsky, and K. Weiner, "Laser Processing: Novel Techniques for Integrated Circuits Fabrication and Repair," (Invited Paper), Conference on Lasers and Electro-Optics, Anaheim, CA, April 1988.
8. C. Y. Fu, B. Law, N. Raley, V. Malba, R. Hsu, R. Hills, and L. Spencer, "A Laser Direct Write Double-Level-Metal Technology for Rapid Fabrication," IEEE Custom Integrated Circuits Conference, Boston, MA, May 1990.
9. R. Hsu, C. Y. Fu, and Ben Law, "The Magnetron-Enhanced Etching of Double-Level-Tungsten Interconnect," The Electrochemical Society Symposium, Washington, D.C., May 1991.
10. C. Y. Fu, R. Hsu, and V. Malba, "Magnetron-Enhanced Reactive-Ion-Etching of Al-1%Si-2%Cu Alloy," Materials Research Society Symposium, Anaheim, CA, April 1991.

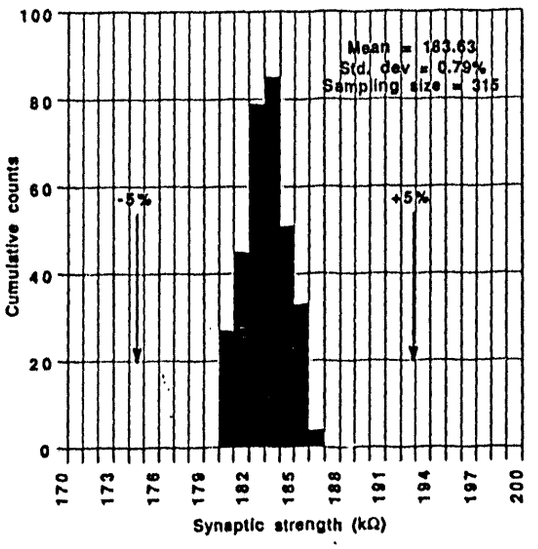

Fig.1. Histogram of resistance distribution

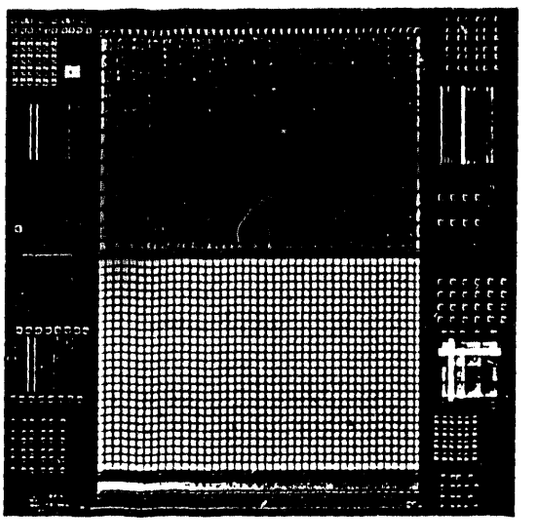

Fig.2. A VLSI synaptic array chip

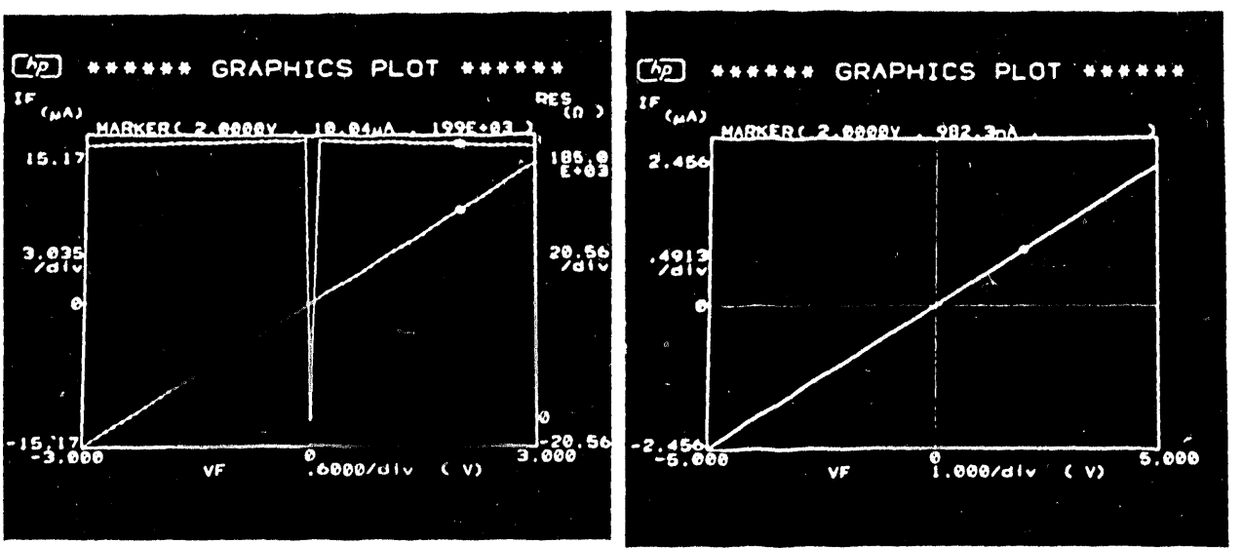

Fig.3. Electrical characteristic of the resistors (a) 200 KΩ and (b) 2 MΩ

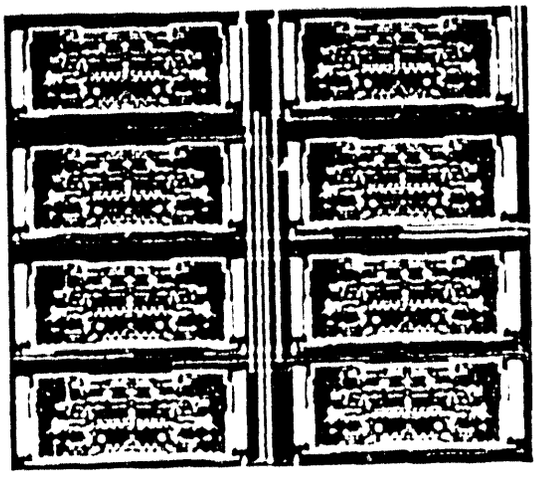

Fig.4. Eight input buffer amplifiers

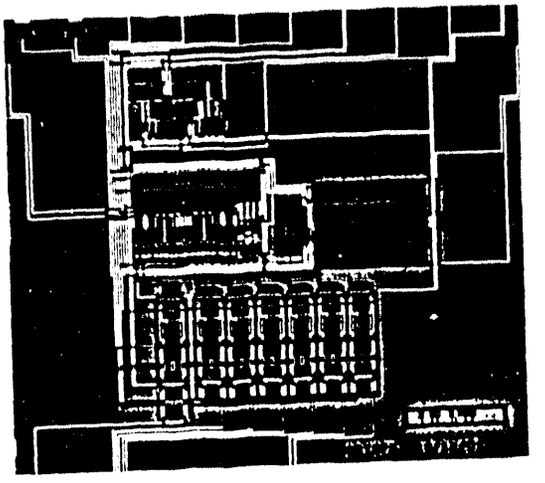

Fig.5. Output neuron body amplifier

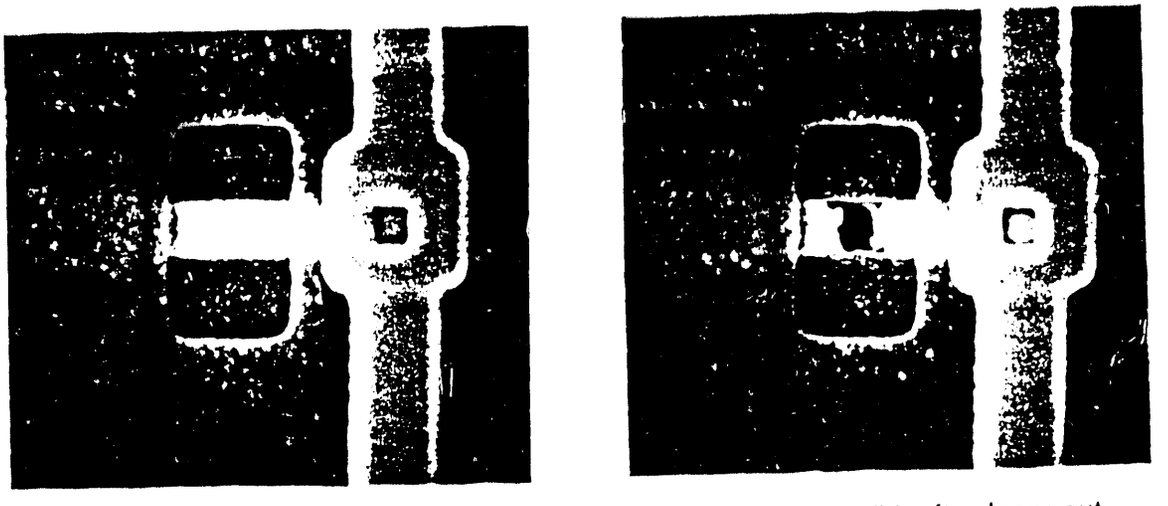

Fig.6. Laser programmable resistors (a) before laser cut, and (b) after laser cut.

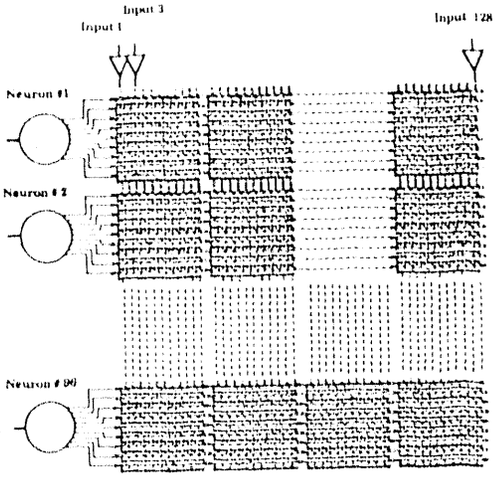

Fig.7. Block diagram for 96-neuron chip

DATE
FILMED
11/17/93

END