VLSI -Implementation of artificial neural network for Sensor-array-system

Frank Stüpmann, Gundolf Geske, Ansgar Wego; Silicann Technologies GmbH, Joachim-Jungius-Str. 9, 18059 Rostock, stuepmann@silicann.com

Abstract

Neural networks are used because of their learning ability in cases in which neither rules nor mathematical models exist. If neural networks are simulated on sequential standard processors an essential advantage, the parallelism, remains unutilized. Therefore, much effort is invested on the implementation of artificial neural networks in integrated hardware. Processing speeds less than 5 µs are achievable with current circuit techniques due to the analog approach.
A special neural network is developed with three layers consisting of 10 neurons in the input layer, 6 neurons in the hidden layer and 10 neurons in the output layer. One application in sensor field is colour value processing.

Functionality

The functionality of a neural network implemented can be described as consecutive (as many as network layers) vector-matrix dot multiplications. Assuming two network layers, the process is as follows. The signals applied to the chip inputs form the input vector (in). This vector is dot multiplied by the first matrix of weight values (w1). The result is a vector that is again dot multiplied by the second matrix of weight values (w2) delivering the output vector (out). The following equation gives an example using a 3-4-2 network topology, i.e. 3 inputs, 4 hidden neurons and 2 outputs.

$$\begin{bmatrix} out_1 \\ out_2 \end{bmatrix} = \left[\begin{bmatrix} in_1 \\ in_2 \\ in_3 \end{bmatrix}^{\mathrm{T}} \bullet \begin{bmatrix} w1_{1,1} & w1_{2,1} & w1_{3,1} \\ w1_{1,2} & w1_{2,2} & w1_{3,2} \\ w1_{1,3} & w1_{2,3} & w1_{3,3} \\ w1_{1,4} & w1_{2,4} & w1_{3,4} \end{bmatrix}^{\mathrm{T}} \right] \bullet \ldots$$
$$\ldots \begin{bmatrix} w2_{1,1} & w2_{1,2} & w2_{1,3} & w2_{1,4} \\ w2_{2,1} & w2_{2,2} & w2_{2,3} & w2_{2,4} \end{bmatrix}^{\mathrm{T}}$$

Fig.1: equation gives an example using a 3-4-2 network topology

The procedure, the proper weight values are determined is called the adaptive process (training). The VLSI integrated artificial neural network chip performing any function trainable with the given network topology. The training is done by a host computer and special training software. The neural network part of the device is fully parallel and analog. It contains a 10-6-10 network topology (i.e. 10 inputs, 6 hidden neurons and 10 outputs). 120 regular synapses are employed. Additionally there are 6 bias synapses in the hidden layer and 10 bias synapses in the output layer. These bias synapses are supplied with a constant signal via an additional input bias neuron. In total, the network holds 136 synapses.
The digital part of the chip consists of decoders to address the synaptic weights. Since the weight storage is realized capacitively, a cyclic refresh of the weight values must be implemented externally [1].

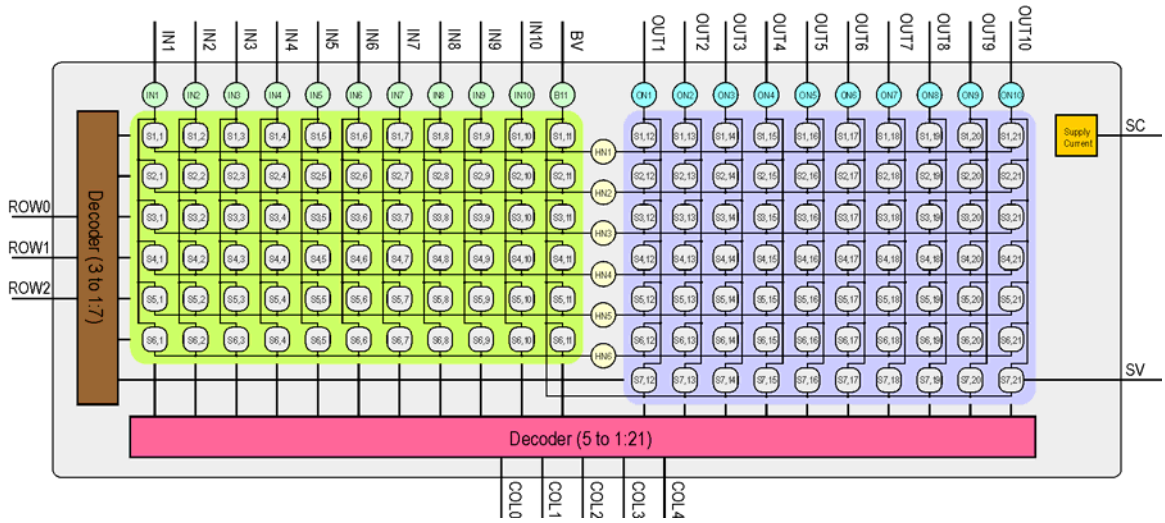

Fig. 1: Schematic structure of the block circuit diagram of the neural network chip

Fig.2 shows the block circuit diagram of the neural network chip. The synapses are arranged in rows and columns. To write a weight value into a synapse weight cell, the proper address must be applied to the row and column terminals. This selects the synapse and the value applied to the synaptic value terminal is loaded into the weight cell.

Refresh

Since the synaptic values are stored capacitively, they must be refreshed at regular intervals [2]. One refresh per 10000 ms is sufficient to garantee stable values. About 7 µs are required to address and load a single synapse. In total, a complete refresh cycle for all synapses takes less than a millisecond (shown in Fig. 3). The quality of the signal processing could be slightly affected during this short transaction.

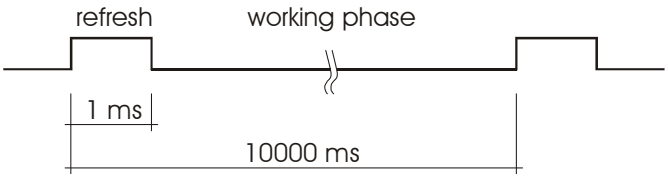

Fig. 3: Refresh Scheme

Response time

The time needed for the classification of one pattern is less than 5 µs (pin to pin). The response time and its variation for a small number of chips is depicted in Fig. 4. Due to the analog approach the chip is predestined for processing of analog sensor signals because no conversions have to be done. The range of values for inputs and outputs is continuous. Also, processing is continuous in the time domain since the neural net is not clocked. To verify the neural network chip's ability to function, several benchmark problems, such as breast cancer classification problems [3, 4, 5] known to neural network researchers, are tested with good results.

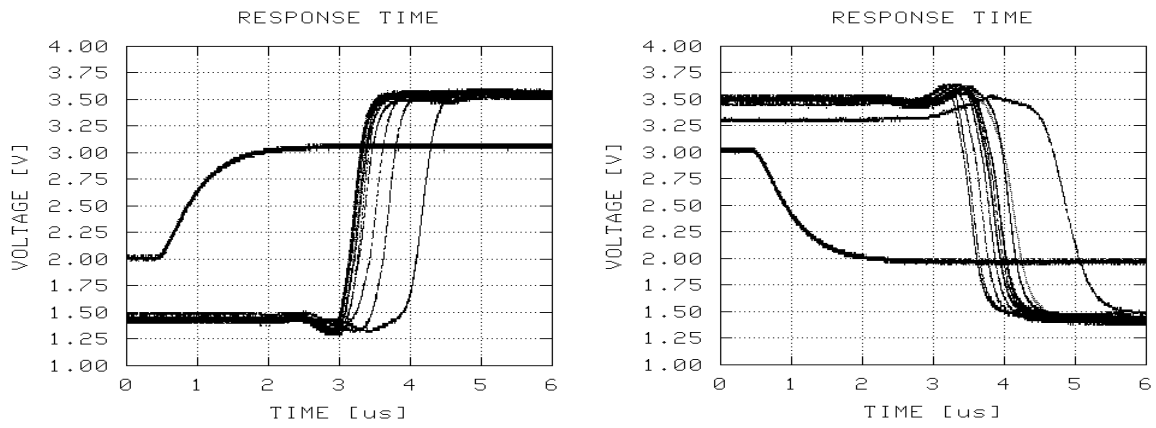

Fig. 4: Measurements of response time needed for the classification of one pattern

Results

The implementation of the synaptic weight storage was solved in a capacitive way by which a good area usage of the silicon can be obtained. One disadvantage of capacitively storing voltages is the drift. This was tried to hold as small as possible by suitable circuit technique. So averaged voltage drifts under 10 µV/s were measured over a larger number of storage elements. By the area efficient implementation of the storage cells we are able to implement a 10:6:10 neural network topology in a chip with a die size < 13 mm ². Therefore a later scaling of the network topology, based on the used circuit technique, is straightforwardly possible. As can be seen from
Table 1, the current consumption lies at 9.5 mA. But it has to be said that the half of the current consumption is caused by the quite robustly dimensioned output drivers. Despite the low power dissipation a high processing performance of 27.2 MCPS (mega-connections by second) can be achieved what corresponds to 27.2 million multiplication addition operations per second [6].

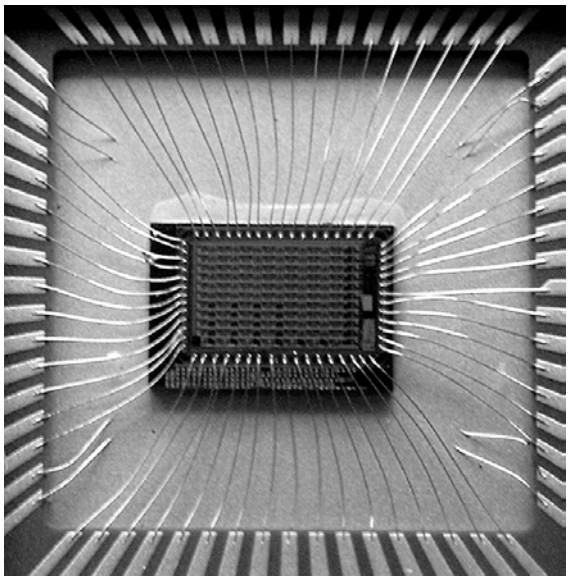

Fig. 5: Neural Network chip with an topology of 10 inputs, 10 outputs and 6 hidden neurons. The die size is less than 13 mm².

area	(2,880 x 4,500) mm ⇒ 12,96 mm²
supply current	9.5 mA
supply voltage	5V
power dissipation	47,5 mW
input voltage range	1,5 ..3,5 V
output current	± 500µA
processing speed	approx. 5µs
processing performance	27.2 MCPS
net topology	10:6:10

Table 1: Specification of the current neural network implementation.

Conclusions

Because of its enormous efficiency, the analog approach suits an extremely large spectrum of applications. Some fields are fast classification tasks, control systems, sensor signal processing or sensor signal adaptation

With such a current topology a classification of colours as well as a adaptation of the sensor circuit are possible. By the integration of DA and AD-converters the training of this analog network can be performed in a digital environment. The training of the neuro chip is realized in an "in the loop" strategy so that non idealities of the analog circuit technique are compensated.

References

[1] G. Geske, F. Stüpmann, S. Rode: "Artificial Neural Network in analog VLSI-Technology", Baltic Electronic Conference, BEC, October 6.-9. 2002, Tallinn, Estonia

[2] P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software, 1:23–34, 1992.

[3] O. L. Mangasarian, R. Setiono, and W. H. Wolberg. Pattern recognition via linear programming: Theory and application to medical diagnosis. SIAM Publications, Philadelphia, ”Large scale numerical optimization”, pages 22–30, 1990.

[4] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. SIAM News, 23(5):1 & 18, September 1990.

[5] F. Stüpmann, S. Rode, G. Geske: “SINGLE CHIP VLSI REALIZATION OF A NEURAL NET FOR FAST DECISION MAKING FUNCTIONS”, ICONIP 2002, 9th International Conference on Neural Information Processing (ICONIP'02), November 18 - 22, 2002, Singapore

[6] G. Geske, A. Wego, F. Stüpmann: „Recognition and Transformation of Colors with an analog Neural Network Chip for Applications with High Real Time Requirements”, 6th International Conference for Optical Technologies, Optical Sensors and Measuring Techniques, Mai, 25 -26, 2004, Nürnberg