High Speed VLSI Neural Network for High-Energy Physics

P. Masa, K. Hoen, H. Wallinga
MESA Research Institute, University of Twente,
P.O. Box 217, 7500 AE Enschede, The Netherlands,
E-mail: masa@ice.el.utwente.nl, Fax: 31 53 341903

Abstract

A CMOS neural network IC is discussed, which was designed for very high speed applications. The parallel architecture, analog computing and digital weight storage provides unprecedented computing speed combined with ease of use. The circuit classifies up to 70 dimensional vectors within 20 nanoseconds, performing 20 billion ($2*10^{10}$) multiply-and-add operations per second, and has as high as 28-42 Gbits/second equivalent input bandwidth with less than 1W dissipation. The synaptic weights can be directly downloaded from a host computer to the on on-chip SRAM. The full-custom, analog-digital chip implements a fully connected feedforward neural network with 70 inputs, 6 hidden layer neurons and one output neuron. A unique solution, a single chip neural network photon trigger for high-energy physics research is provided.

1. Introduction

Although neural networks (NNs) compute exceptionally parallel manner, this valuable characteristic has not been exploited as successfully as their learning capability. In case of fully parallel hardware, the processing time is independent of the amount of data to be processed by the network. Furthermore only a few computing steps have to be performed in serial manner, therefore computation time can be extremely short. This work concentrates on the benefits of unique parallel processing. One of the most challenging tasks of hardware realisation of neural nets is the inner product operation. Since it consumes too large chip area with digital circuitry, fully parallel digital architectures do not exist for large NNs. If high precision is not required, the compact and high speed analog approach has great advantage. With analog technique low cost, low power dissipation, single chip architectures of complex neural networks are possible. Although such systems are commercially available [4], offering as low as several microsecond processing time for as large as 128 dimensional input vector, it is almost impossible to find any solution for application domain demanding tens of nanoseconds processing delay for similarly large input vectors. The integrated circuit presented here is intended to provide the high computing performance needed for such applications.

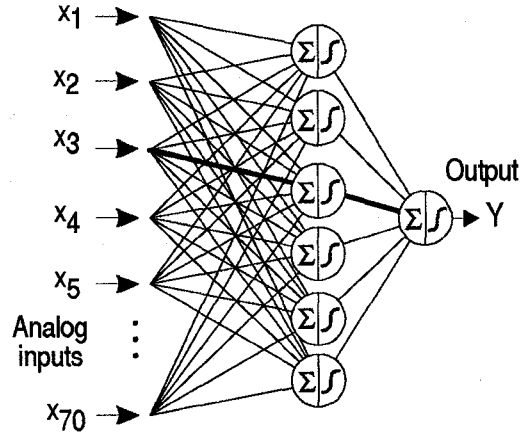

Figure 1. Implemented NN architecture

2. System description

The implemented NN architecture is shown on figure 1. It is a fully interconnected feedforward structure with 70 analog inputs, 6 hidden layer neurons and one output neuron. The neurons are inner product type, and have sigmoid-like activation function. The neural signal processing is fully analog, yielding high speed operation and compact circuitry for inner product operation. The synaptic weights are stored digitally on static RAM (SRAM) cells, to enable simple programming even from

0-8186-6710-9/94 $4.00

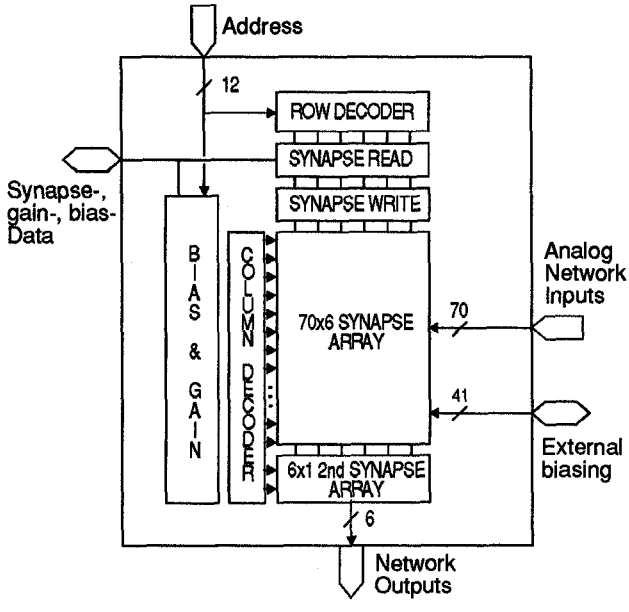

Figure 2. System block diagram

a personal computer. Digital weight storage also helps to eliminate weight decay and increases reproducibility. The SRAM cells are located nearby each synapse circuit to minimise wiring for communication. Downloading the approximately 3.5 Kbit synaptic weight and configuring information is relatively slow compared to normal operating speed of the NN circuit, and takes a few milliseconds. The chip block diagram is shown on figure 2.

The largest area is occupied by the 70x6 synapse array, including the 70x6 differential voltage to current converters as synapses and 70x6x5 SRAM cell for weight storage. Each 5 bit synaptic weight can be selected, red and written by the row-, column decoders and read, write circuitry. A programmable voltage source array is located nearby the synapse array which enables programmable biasing and control of the gain of neural activation functions.

3. Circuit operation

Figure 3 shows the analog circuitry along the signed signal path of figure 1. The processing delay of the NN pattern classifier is merely the delay introduced by this circuitry, since the rest of signal paths are parallel. The synapse circuit is a differential pair formed by T1 and T2, with a single ended voltage input and a reference voltage, which is equal for all the synapses in the NN. The outputs of synapses are differential currents, which are summed on the (differential) summing node of the corresponding neuron. Variable synaptic weight is achieved by programmable current source for the differential pair. The current source transistors T7, T8, T9, T10 are properly sized to deliver current with respect to the smallest, or unity current, according to ascending powers of 2. Any combination of these currents can be obtained by using the switch transistors T3, T4, T5, T6. The sign of the synapse can be varied by interchanging Vin and Vref, using an 8 transistor switch, which is not shown in figure 3.

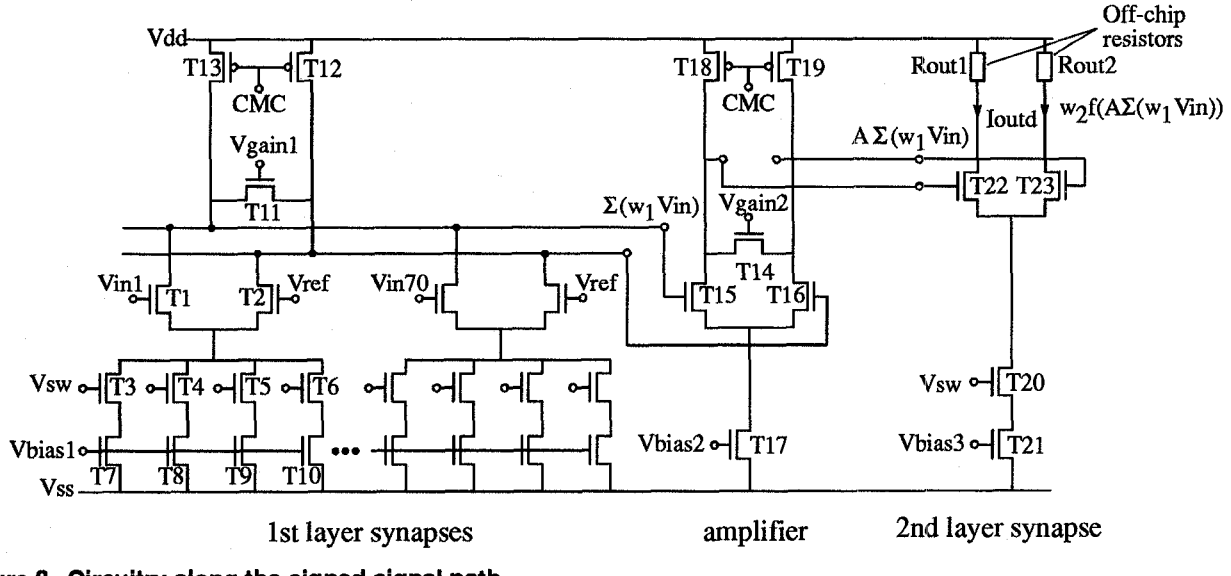

Figure 3. Circuitry along the signed signal path of figure 1

Simulated and measured synapse characteristic is shown in figure 4a and figure 4b. The sum of synaptic currents is transformed to voltage by the load transistors T12 and T13. T11 controls the differential load. The saturating, sigmoid-like activation function, shown in figure 5, is obtained by the saturating characteristic of the second layer synapse, rather than by a separate non-linear circuit.

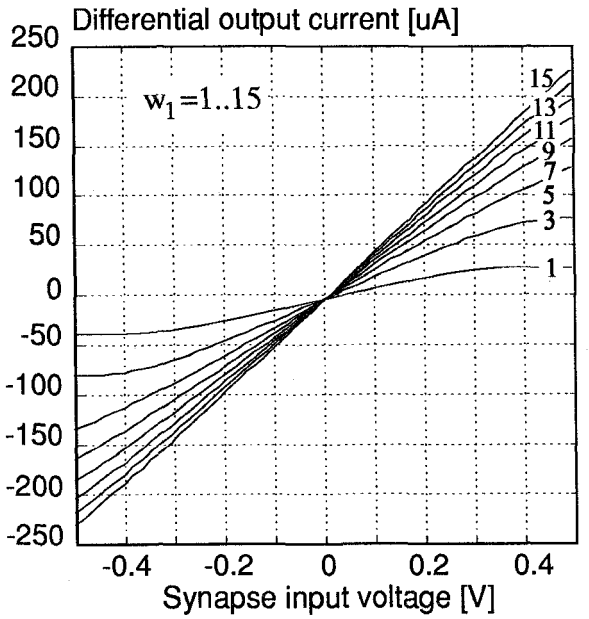

Figure 4a. Simulated synapse characteristic

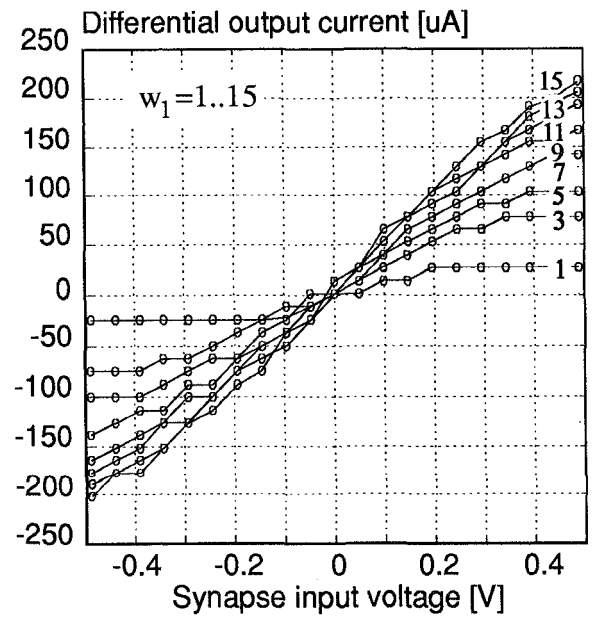

Figure 4b. Measured synapse characteristic

This method simplifies the circuitry and increases speed. The second layer synaptic weight is obtained by the number of parallel connected active synapse stages, formed by T20, T21, T22, T23. This stage is activated by the switch transistor T20. Every switch transistor of the circuit is wired to a separate SRAM cell. There are alto gether 3750 SRAM cells on chip.

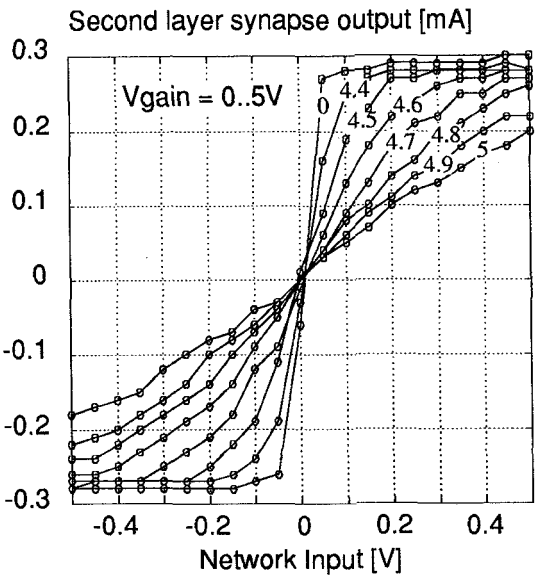

Figure 5. Activation function

The current summing node has intrinsically large parasitic capacitance, since all the synapse outputs and the common load are connected to this node. To increase the speed of the circuit, the node impedance has to be kept low. The consequence of low node impedance is a small voltage swing on the summing node. A voltage amplifier stage scales this voltage properly for the second layer synapse stage.

4. Application in high-energy physics research

The feasibility of a single chip NN photon trigger for the LHC[1] experiment in CERN[2] have been studied. We used a database containing "TGT calorimeter preshower" information[3], generated by photons and pions. The NN photon trigger should recognise data which was

1 LHC:Large Hadron Collider at CERN

2 CERN: Conseil européen pour la recherce nucléraire

3 "TGT calorimeter preshower" information, developed on the basis of CERN RD33 project.

generated by photons. Real-time data processing allows less than 25 nanosecond time period for the decision making. Within this time period a 32 dimensional analog input vector has to be evaluated. A similar high-energy physics (HEP) application is described for an experiment at DESY[4], in [1], [3], with higher dimensional analog, time discrete input vectors.

We trained a BackProp. network with the "labelled" database. The total set with more than 7000 samples, was divided into training and test set. The small difference between the performance on training- and test set indicate good generalisation. The percentage of incorrectly classified patterns is smaller than 4%, even for the test set. Synaptic weights obtained by the training procedure, can be downloaded to the NN chip.

Examining the decision making process of feedforward neural nets for pattern classification, reveals why and how this type of computation tolerates the non-ideal effects of analog hardware. Here only quantitative results are presented for the discussed application, one may refer to [1], [2], [3] for more detailed discussion.

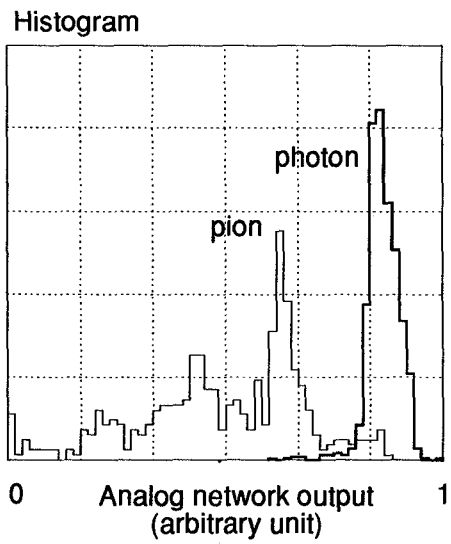

Figure 6. Histogram of analog classifier responses (test set, optimal)

Figure 6 shows, the histogram of analog classifier outputs over the test set. We can see, that although there is an overlap between the two classes, photon and pion data is clearly separated. Above or under a certain network output the data is classified "photon" or "pion" respectively. We call this value of network output "decision threshold".

[4]DESY: *Deutsches Elektronen Synchrotron (Hamburg)*

Figure 7 shows the classification efficiency as a function of the decision threshold. For example if we choose the decision threshold, where the percentage of misclassified photon patterns equals the percentage of misclassified pion patterns, the correctly classified data is 96% for both classes. When decreasing the decision threshold to correctly classify 99% of photon data, 15% of pion data is misclassified.

Figure 6 and figure 7 show performance in case of ideal hardware. Simulations have been made to examine the non-ideal effects, introduced by our analog NN hardware. Noise, synapse non-linearity, weight discretization and the effect of sigmoid-like shape of the activation function have been taken into account. Figure 8 and figure 9 show the results. The overlap between the two classes increases compared to the ideal case. In contrast to the 96% correctly classified data at the crossing of curves in figure 7, we get 93% with our hardware. The 75% increase of misclassified patterns is mainly due to the applied simple discretization technique. We expect even better performance with careful weight discretization.

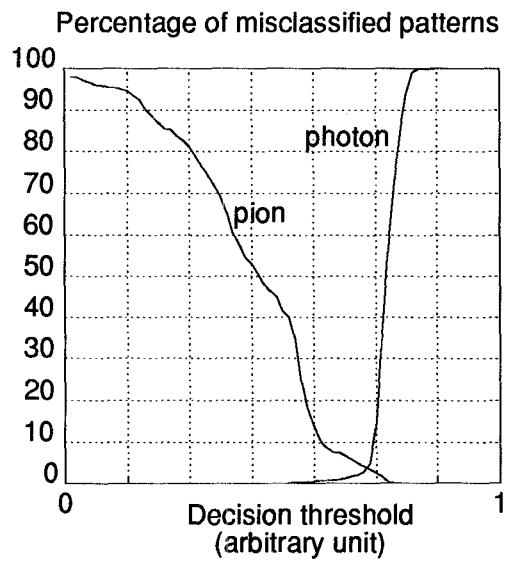

Figure 7. Classification error (test set, optimal)

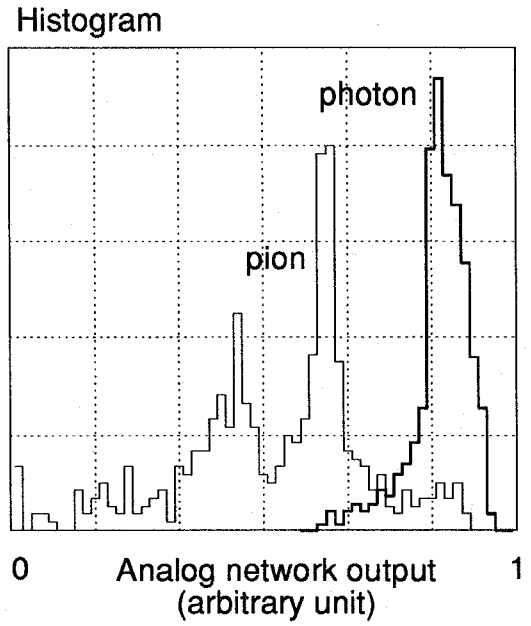

Figure 8. Histogram of analog classifier responses (test set, hardware)

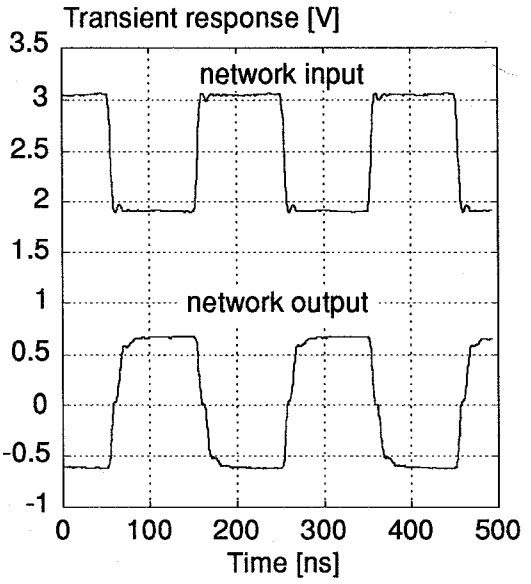

Figure 10. Transient response

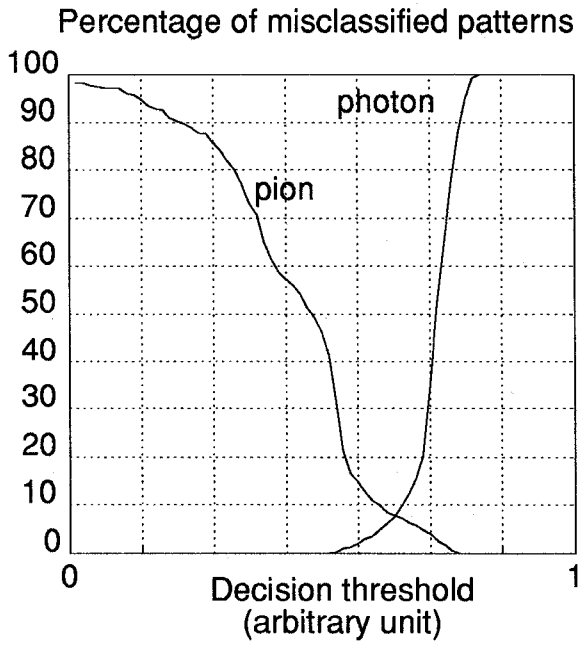

Figure 9. Classification error (test set, HW)

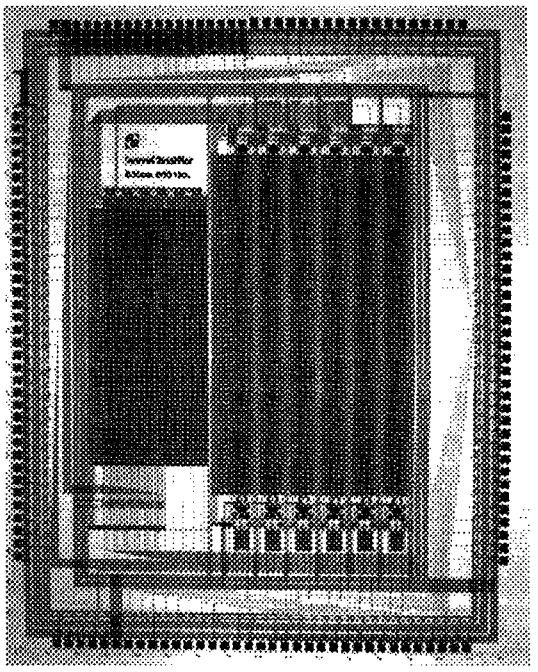

Figure 11. Chip Layout

The measured transient response is plotted in figure 10. Both the stimulus and the NN response are shown on the figure. Table 1. shows more details about the chip specifications. Equivalent input bandwidth was calculated by assuming 12 bit resolution (signal to noise ratio) and 50 MHz input rate for the time discrete analog input signals. Multiplying the maximal pattern rate by the number of synapses we get the computing performance in terms of multiplications and additions per second. The chip layout is shown in figure 11.

Table 1. Chip specifications

Total processing delay:	20 nanoseconds
Computation speed:	20 billion multiplications and additions per second
Equivalent input bandwidth	4 GBytes/second
Number/resolution of synapses	426, 5 bits (4 bits + sign)
Synapse size	$400x70\mu m^2$
No. of transistors:	40 000
PGA package:	144 pins
Chip size	10mmx9mm (1.5μm DLM CMOS, ES2, EUROCHIP)
On-chip static RAM	3750 bits
Power dissipation:	<1W

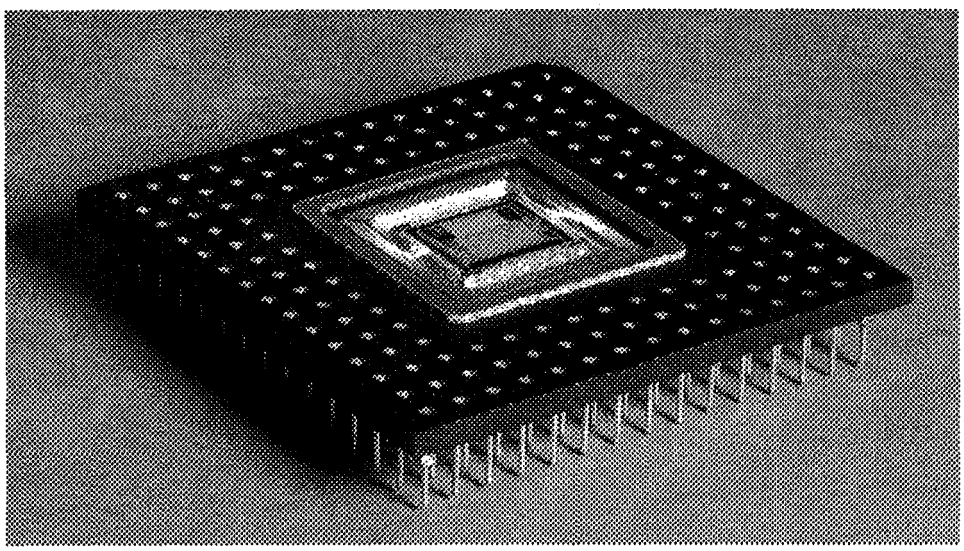

Figure 12. Chip photo

Table 2. Comparison of high speed integrated circuits for pattern classification

	intel® Ni1000 Recognition accelerator	intel® ETANN	this chip
Input Bandwidth [MBytes/s]	42	20	4000
Patterns/s	33K	300K	50000K
Multiplications/s	(not inner product neurons)	1.3G	20G

Table 2 compares the parameters of available high speed integrated circuits for pattern classification. The comparison is not comprehensive, because only those parameters are shown which were crucial for the presented high-energy physics application. Although the two Intel chips implement more neurons [4],[7], it was shown, that more hidden layer neurons doesn't result in better classification performance for the presented application and for the similar (HEP) application described in [1].

5. Conclusions

Although the chip does not take advantage of a state-of-the-art technology [5],[6] it provides unprecedented computing speed, due to the parallel architecture and the analog processing. Using the more advanced ES2 1.0um CMOS process, the computing performance could be further increased by an order of magnitude. The analog computing - digital storage concept combines very high speed with ease of use. The single chip pattern classifier performs 20 billion ($2*10^{10}$) multiplications per second, with less than 1W power dissipation and has 4 GBytes per second input bandwidth. Synaptic weights can be directly downloaded from a host computer without special interfacing, due to the SRAM synapse memory. With this unique performance classification up to 70 dimensional vectors within tens of nanoseconds becomes possible. The feasibility of a single chip neural network photon trigger for high-energy physics research have been confirmed.

Acknowledgements

This work in the program of the Foundation for Fundamental Research on Matter (FOM) have been supported by the Netherlands Technology Foundation (STW)

References

[1] P. Masa, K. Hoen, H. Wallinga, "High-Speed Analog Neural Processor", IEEE Micro, Special Issue on Analogue VLSI and Neural Networks, June 1994

[2] P. Masa, K. Hoen, H. Wallinga, "20 Million Patterns Per Second Analog CMOS Neural Network Pattern Classifier", Proc. European Conf. on Circuit Theory and Design, Davos, Switzerland, 1993

[3] P. Masa at. al., "20 Million Patterns Per Second VLSI Neural Network Pattern Classifier" Proc. International Conference on Artificial Neural Networks, Amsterdam, 1993

[4] Hernan A. at al., "Implementation and Performance of an Analog Nonvolatile Neural Network," Analog Integrated Circuits and Signal Processing 4, 97-113 1993

[5] Edward McLellan "The Alpha AXP Architecture and 21064 Processor", IEEE Micro, Vol. 13, No.3, June 1993 pp 36-47

[6] D. Alpert, D. Avnon, Architecture of the Pentium Microprocessor, IEEE Micro, Vol. 13, No.3, June 1993 pp 11-21

[7] J. Diamond, C. Park, S.C. The, U. Santoni, K. Buckmann, N. Holler, "Design and implementation of a Recognition Accelerator", Proceedings of the Canadian Conference on VLSI, 1993