# Hardware Implementation of Artificial Neural Network Using Field Programmable Gate Array

Esraa Zeki Mohammed and Haitham Kareem Ali

***Abstract*—In this paper a hardware implementation of an artificial neural network on Field Programmable Gate Arrays (FPGA) is presented. A digital system architecture is designed to realize a feed forward multilayer neural network. The designed architecture is described using Very High Speed Integrated Circuits Hardware Description Language (VHDL).**

**The parallel structure of a neural network makes it potentially fast for the computation of certain tasks. The same feature makes a neural network well suited for implementation in VLSI technology. Hardware realization of a Neural Network (NN), to a large extent depends on the efficient implementation of a single neuron. FPGA-based reconfigurable computing architectures are suitable for hardware implementation of neural networks. FPGA realization of ANNs with a large number of neurons is still a challenging task.**

***Index Terms*—Artificial neural network, hardware description language, field programmable gate arrays (FPGAs), sigmoid activation function.**

## I. Introduction

Artificial neural networks (ANN) have found widespread deployment in a broad spectrum of classification, perception, association and control applications [1].

The aspiration to build intelligent systems complemented with the advances in high speed computing has proved through simulation the capability of Artificial Neural Networks (ANN) to map, model and classify nonlinear systems. Real time applications are possible only if low cost high-peed neural computation is made realizable. Towards this goal numerous works on implementation of Neural Networks (NN) have been proposed [2].

Artificial neural networks (ANNs) have been mostly implemented in software. This has benefits, since the designer does not need to know the inner workings of neural network elements, but can concentrate on the application of the neural network. However, a disadvantage in real-time applications of software-based ANNs is slower execution compared with hardware-based ANNs.

Hardware-based ANNs have been implemented as both analogue and digital circuits. The analogue implementations exploit the nonlinear characteristics of CMOS (complementary metal-oxide semiconductor) devices, but they suffer from thermal drift, inexact computation results and lack of reprogrammability.

Digital hardware-based implementations of ANNs have been relatively scarce, representive examples of recent research can be found in. Recent advances in reprogrammable logic enable implementing large ANNs on a single field-programmable gate array (FPGA) device. The main reason for this is the miniaturisation of component manufacturing technology, where the data density of electronic components doubles every 18 months [3].

ANNs are biologically inspired and require parallel computations in their nature. Microprocessors and DSPs are not suitable for parallel designs. Designing fully parallel modules can be available by ASICs and VLSIs but it is expensive and time consuming to develop such chips. In addition the design results in an ANN suited only for one target application. FPGAs not only offer parallelism but also flexible designs, savings in cost and design cycle [4].

## II. Artificial Neuron

Artificial neural networks are inspired by the biological neural systems. The transmission of signals in biological neurons through synapses is a complex chemical process in which specific transmitter substances are released from the sending side of the synapse. The effect is to raise or lower the electrical potential inside the body of the receiving cell. If this potential reaches a threshold, the neuron fires. It is this characteristic of the biological neurons that the artificial neuron model proposed by McCulloch Pitts attempts to reproduce. Following neuron model shown in Fig. 1 is widely used in artificial neural networks with some variations.

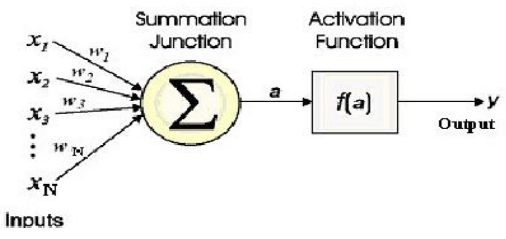


Fig. 1. Structural diagram of a neuron

The artificial neuron given in this figure has *N* inputs, denoted as *x*1, *x*2,…..*xN*,. Each line connecting these inputs to the neuron is assigned a weight, denoted as *w*1, *w*2,…..*wN* respectively. The activation, a, determines whether the neuron is to be fired or not. It is given by the formula:

$$a = (\sum_{j=1}^{N} w_j x_j) \quad (1)$$

Manuscript received November 18, 2012; revised February 27, 2013.

Esraa Zeki Mohammed is with the State Company for Internet Services, Ministry of Communcation, Kirkuk, Iraq (e-mail: Isra_mohammed2@yahoo.com).

Haitham Kareem Ali is with the Communication Engineering Department, Sulaimani Technical College, Sulaimani, Iraq (e-mail: haitham_elect@yahoo.com).

A negative value for a weight indicates an inhibitory connection while a positive value indicates excitatory connection.

The output, y of the neuron is given as:

$$y = f(a) \tag{2}$$

Originally the neuron output function f(a) in McCulloch Pitts model was proposed as threshold function, however linear, ramp, and sigmoid functions are also used in different situations. The vector notation

$$a = w^T x \tag{3}$$

can be used for expressing the activation of a neuron. Here, the *jth* element of the input vector x is *xj*, the *jth* element of the weight vector of *w* is *wj*. Both of these vectors are of size *N*. A Neuro-computing system is made up of a number of artificial neurons and a huge number of interconnections between them. Fig. 2 shows architecture of feedforward neural network [5].

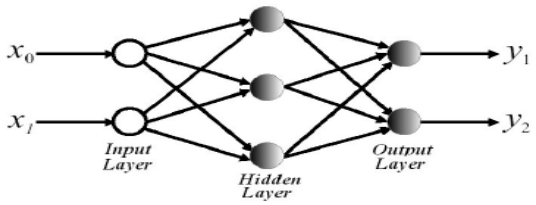


Fig. 2. Layered feedforward neural network

In layered neural networks, the neurons are organized in the form of layers. The neurons in a layer get inputs from the previous layer and feed their output to the next layer. These type of networks are called feedforward networks.

Output connections from a neuron to the same or previous layer neurons are not permitted. The input layer is made of special input neurons, transmitting only the applied external input to their outputs. The last layer is called the output layer, and the layers other than input & output layers are called the hidden layers. In a network, if there are input and output layers only, then it is called a single layer network. Networks with one or more hidden layers are called multilayer networks [6].

## III. Overview of VHDL

VHDL is a language meant for describing digital electronic systems. In its simplest form, the description of a component in VHDL consists of an interface specification and an architectural specification. The interface description begins with the ENTITY keyword and contains the input- output ports of the component. The name of the component comes after the ENTITY keyword and is followed by IS, which is also a VHDL keyword. The description of the internal implementation of an entity is called an architecture body of the entity. There may be a number of different architecture bodies of an interface to an entity corresponding to alternative implementations that perform the same function. The alternative implementations of the architecture body of the entity is termed as Behavioral Description or Structural Description.

After describing a digital system in VHDL, simulation of the VHDL code has to be carried out for two reasons. First, we need to verify whether the VHDL code correctly implements the intended design. Second, we need to verify that the design meets its specifications. The simulation is used to test the VHDL code by writing test bench models. A test bench is a model that is employed to exercise and verify the correctness of a hardware model and it can be described in the same language.

Some synthesis tools are capable of implementing the digital system described by the VHDL code using a PGA (Programmable gate array) or CPLD (Complex programmable logic devices). The PLDs are capable of implementing a sequential network but not a complete digital system. Programmable gate arrays and complex programmable logic devices are more flexible and more versatile and can be used to implement a complete digital system on a single chip. A typical PGA is an IC that contains an array of identical logic cells with programmable interconnections. The user can program the functions realized by each logic cell and the connections between the cells. Such PGAs are often called FPGAs since they are field programmable [7], [8].

## IV. The Proposed Design

The proposed design consists of neuron architecture design, activation function problem solving and artificial neural network design which consist of two layers.

### A. Neuron Architecture

The processing element of an ANN is the Neuron. A Neuron can be viewed as processing data in three steps; the weighting of its input values, the summation of them all and their filtering by sigmoid function. The summation can be calculated by a serial accumulation. For the weighted inputs to be calculated in parallel using conventional design techniques, a large number of multiplier units would be required. To avoid this, Multiplier/Accumulator architecture has been selected. It takes the input serially, multiplies them with the corresponding weight and accumulates their sum in a register. The processes are synchronized to clock signal. The number of clock cycles for a neuron to finish its work, equals to the number of connections from the previous layer. The accumulator has a load signal, so that the bias values are loaded to all neurons at start-up. Fig. 3 shows the proposed neuron design.

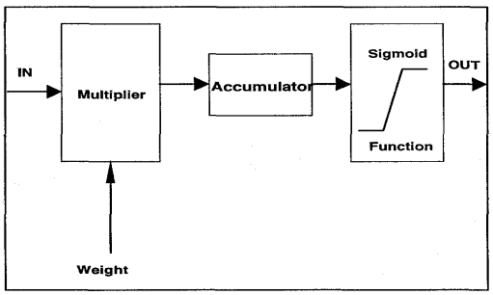


Fig. 3. Neuron architecture

### B. Activation Function

One of the most important parts of a neuron is its activation function. The nonlinearity of the activation function makes it possible to approximate any function. In the hardware implementation concept of neural networks, it is not so easy to realize sigmoid activation functions [9].

Special attention must be paid to an area-efficient implementation of every computational element when implementing large ANNs on digital hardware. This holds true for the nonlinear activation function used at the output of neurons.

A common activation function is the sigmoid function

$$y = \frac{1}{1+\varepsilon^{-x}} \tag{4}$$

Efficient implementation of the sigmoid function on an FPGA is a difficult challenge faced by designers. It is not suitable for direct implementation because it consists of an infinite exponential series. In most cases computationally simplified alternatives of sigmoid function are used.

Direct implementation for non-linear sigmoid transfer functions is very expensive. There are two practical approaches to approximate sigmoid functions with simple FPGA designs. Piece-wise linear approximation describes a combination of lines in the form of y=ax+b which is used to approximate the sigmoid function. Especially if the coefficients for the lines are chosen to be powers of two, the sigmoid functions can be realized by a series of shift and add operations. The second method is lookup tables, in which uniform samples taken from the centre of a sigmoid function can be stored in a table for look up. The regions outside the centre of the sigmoid function are still approximated in a piece-wise linear fashion.

This research presents an approximation approach to implement sigmoid function. A simple second order nonlinear function can be used as an approximation to a sigmoid function. This nonlinear function can be implemented directly using digital techniques. The following equation is a second order nonlinear function which has a tansig transition between the upper and lower saturation regions:

$$G_S(z) = \begin{cases} 1 \; for L \le z \\ z(\beta - \theta z) \; for 0 \le z \le L \\ z(\beta + \theta z) \; for -L \le z \le 0 \\ -1 \; for \le -L \end{cases}$$

Where β and θ represent the slope and the gain of the nonlinear function Gs(z) between the saturation regions -L and L.

The structural diagram of the approximated sigmoid function implemented using this process Gs(z) is shown in Fig. 4.

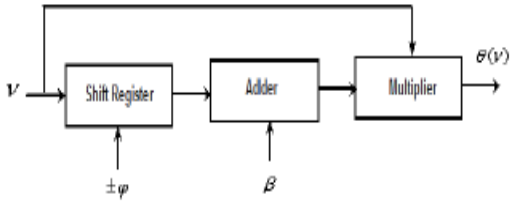


Fig. 4. Structural diagram for sigmoid function

The VHDL code for the approximated sigmoid activation function shown in code (1) below.

```
Code (1)
Library IEEE;
Use IEEE.std_logic_1164.all;
Use IEEE.std_logic_arith.all;
Use IEEE.std_logic_unsigned.all;

Entity Sigmoid is
Port(
        Z: in INTEGER range 0 to 511;
        S: out INTEGER range 0 to 255
        );
End Sigmoid;
Architecture Sigmoidarch of Sigmoid is
Signal TEMP:integer range  0 to 255;
Signal B: integer range o to 511;
Signal A: integer range 0 to 511;
Signal Z: integer range 0 to 65535;
Signal ZZ: integer range 0 to 262144;
Costant L: integer range 0 to 255:=255;
Constant M: integer range 0 to 1023:=512;
Begin
A<=Z;
B<=M-A;
ZZ<=B*A;
Z<=ZZ/265;
TEMp<=Z;
S<=TEMP when A<L else L;
End Sigmoidarch;
```

### C. Layer Architecture

In first Implentation of an ANN layer they take an input from their common input line, multiply it with the corresponding weight from their weight ROM and accumulate the product. If the previous layer has 3 neurons, present layer takes and processes these inputs in 3 clock cycles. After these 3 clock cycles, every neuron in the layer has its net values ready. Then the layer starts to transfer these values to its output one by one for the next layer to take them successively by enabling corresponding neuron's three-state output. The block diagram of a layer architecture including 3 neurons is shown in Fig. 5.

Since only one neuron's output have to be present at the layer's output at a time, instead of implementing an activation function for each neuron it is convenient to implement one activation function for each layer. In this layer structure pipelining is also possible. A new input pattern can enter the network while another is propagating through the layers.

A second Implementation of a fully parallel neural network is possible in FPGAs. A fully parallel network is fast but inflexible. Because, in a fully parallel network the number of multipliers per neuron must be equal to the number of connections to this neuron. Since all of the products must be summed, the number of full adders equals to the number of connections to the previous layer minus one. For example in a 3-1 network the output neuron must have 3 multipliers and 2 full adders. So different neuron architectures have to be designed for each layer. Because multipliers are the most

resource using elements in a neuron structure, a second drawback of a fully parallel network is gate resource usage. Fig. 6 show single layer artificial neural network with three input nodes and one output node.

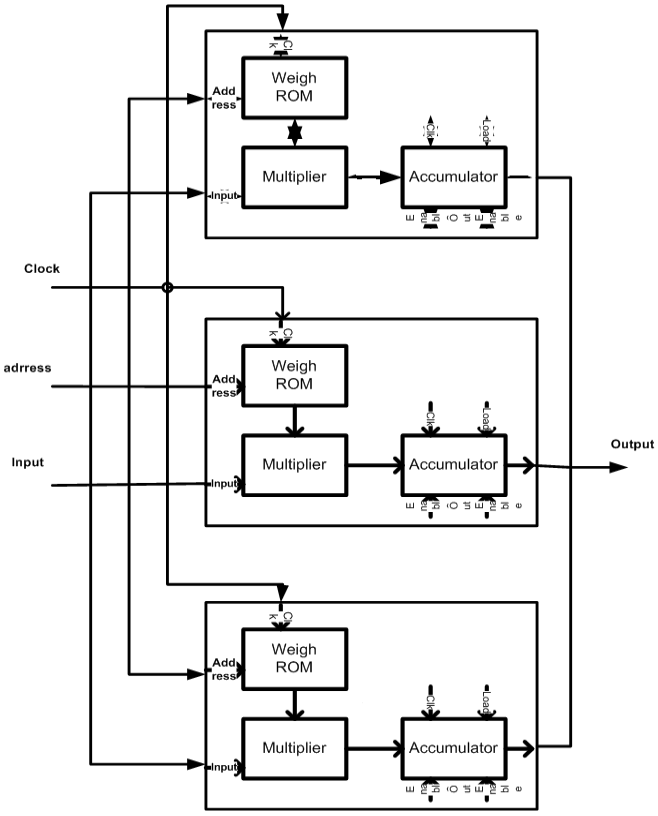


Fig. 5. Block diagram of a layer consist of three neuron

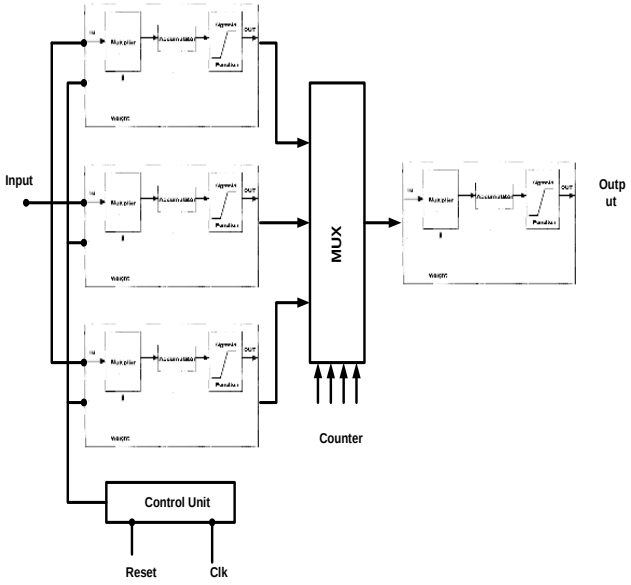


Fig. 6. Artificial neural network design

## V. Conclusions

This paper has presented the implementation of neural networks by FPGAs. The proposed network architecture is modular, being possible to easily increase or decrease the number of neurons as well as layers.

The motivation for this study stems form the fact that an FPGA coprocessor with limited logic density and capabilities can used in building Artificial Neural Network which is widely used in solving different problems.

Future work involves estimating the maximum size of ANNs in modern FPGAs. The main points are the size and parameterisability of multipliers and the number of interlayer interconnections. The first defines mainly the required area resources and the second defines the required routing.

## References

[1] B. Widrow, D. E. Rumelhart, and M. A. Lehr, "Neural networks: applications in industry, business and science," *Communications of the ACM*, vol. 37, no. 3, pp. 93-105, 1994.

[2] A. Muthuramalingam, S. Himavathi, and E. Srinivasan, "Neural network implementation using fpga: issues and application," *The International Journal of Information Technology*, vol. 4, no. 2, pp.86-92, 2008.

[3] M. T. Tommiska, "Efficient digital implementation of the sigmoid function for reprogrammable logic," *IEEE Proceedings, Computers and Digital Techniques*, vol. 150, no. 6, pp. 403- 411, 2003.

[4] A. Savran and S. Ünsal, "Hardware implementation of a feedforward neural network using FPGAs'," Ege University, Department of Electrical and Electronics Engineering, 2003.

[5] S. Haykin, *Neural Networks-a Comprehensive Foundations*, Second Edition, ISBN: 0132733501, 1998.

[6] C. S. Rai and A. P. Singh, "A review of implementation techniques for artificial neural networks," University School of Information Technology, GGS Indraprastha University, Delhi, 2006.

[7] S. P. J. V. Rani and P. Kanagasabapathy, "Design of Neural Network on FPGA," *International Conference on VLS*, USA, 2004.

[8] J. Hamblen and M. Furman, *Rapid prototyping of digital systems*, Kluwer Academic Publisher 2nd Edition, Boston, 2001.

[9] M. Avc ý and T. Yýldýrým, "Generation of tangent hyperbolic sigmoid function for microcontroller based digital implementation of neural networks," in *Proc. International XII. Turkish Symposium on Artificial Intelligence and Neural Networks*, 2003.

**Haitham Kareem** received the B.S. in electrical and electronic(AVIONICS) engineering from Al-Rashied College of Engineering & Science in Iraq in 1992, he received his M.S. degrees in communication and radar engineering from Al-Rashied college of Engineering & Science in Iraq in 1997 and Ph.D. in electronic engineering from University of Technology, Al-Rashied College of Engineering & Science in Iraq in 2006. From 1992 to 1997 he was working as an assistant lecturer in University of Technology-College of Engineering. From 1998 to 1999 he was working as an assistant lecturer at Academic Cararey for Technicians, Khartoum, Sudan. From 2000 to 2004 he was working as an assistant lecturer at the University of Technology, College of Engineering, Iraq. From 2004 to 2006 he was working as a lecture in the University of Diyala, College of Engineering, Iraq. From 2006 until now he has been working as a lecture in the Foundation of Technical Education, Sulaimany Technical College, Communication Engineering Department, Iraq.

**Esraa Zeki** received the B.S. in computer science from Mosul University in 2000-2001in Iraq. She obtained her M.S. degrees in computer science from Sulaimani University in 2008-2009, Iraq. From 2001 to 2002 she was working as external lecturer in Kirkuk technical institute, Software Department in Kirkuk technical Education, Iraq, from 2002 until now she has been working in State Company for internet services in Kirkuk, Iraq.