Microelectronic Implementation of Neural Networks

U. Rückert

Technical University of Hamburg-Harburg,
Faculty of Electrical Engineering
POB 901052, 21071 Hamburg, Germany

Abstract

The paper reviews the major trends in microelectronic implementation of artificial neural networks. Despite the remarkable progress in silicon-based VLSI technology, there is no clear consensus at the moment on how to exploit these technological capabilities for massively parallel neural network hardware implementation. In the following some of the key implementational issues will be considered and some representative VLSI hardware realizations will be sketched.

Introduction

Even though there are still many open problems in the theory of ANNs, there is unanimous agreement that computational speed will ultimately become a stumbling block in the research and application of artificial neural networks (ANNs). At present, ANNs are still mainly simulated on conventional, sequential computers. Software simulators offer a high flexebility and serve well to test an ANN concept, but don't offer the performance required to run real-world applications in general. Therefore, the availability of specially designed neurocomputing hardware (neurocomputer) offers the most advantageous alternative for multifarious utilization of ANNs. Not only would the processing time drastically decrease for neurocomputing hardware, but also the smaller volume, the reduced power supply requirements, and the higher reliability would render microelectronic neurocomputers very attractive.

State-of-the-art VLSI (Very Large Scale Integration) and the emerging ULSI (Ultra Large Scale Intergration) technologies are able to integrate millions of microelectronic devices on a single chip. Clock rates are approaching 100MHz boosting the chip-computational power to 10^5-10^6 MIPS (million instructions per second). Hence, modern microelectronic technology offers a speed-up factor of several orders of magnitude compared with simulation on todays

sequential computers. Even more computational power may be obtained by emerging technologies like optoelectronics or molecular electronics.

Despite the impressive development of microelectronics during the last decades, there is no clear consensus on how to exploit these technological capabilities for massively parallel ANN algorithms. It is currently not possible to determine the best way to perform ANN calculations for any given application. This is one reason for the huge variety of approaches for ANN hardware implementation known in literature. In the following, an overview of the major trends in ANN hardware implementation with an emphasis on integrated circuits (ICs) will be given by grouping the different approaches into few categories and by discussing the key features of each of these categories. The first category contains neurocomputers based on standard ICs. Task or model dedicated neural ASICs (application specific integrated circuits) build the second group which are further subdivided into digital and analog circuits (Fig. 1).

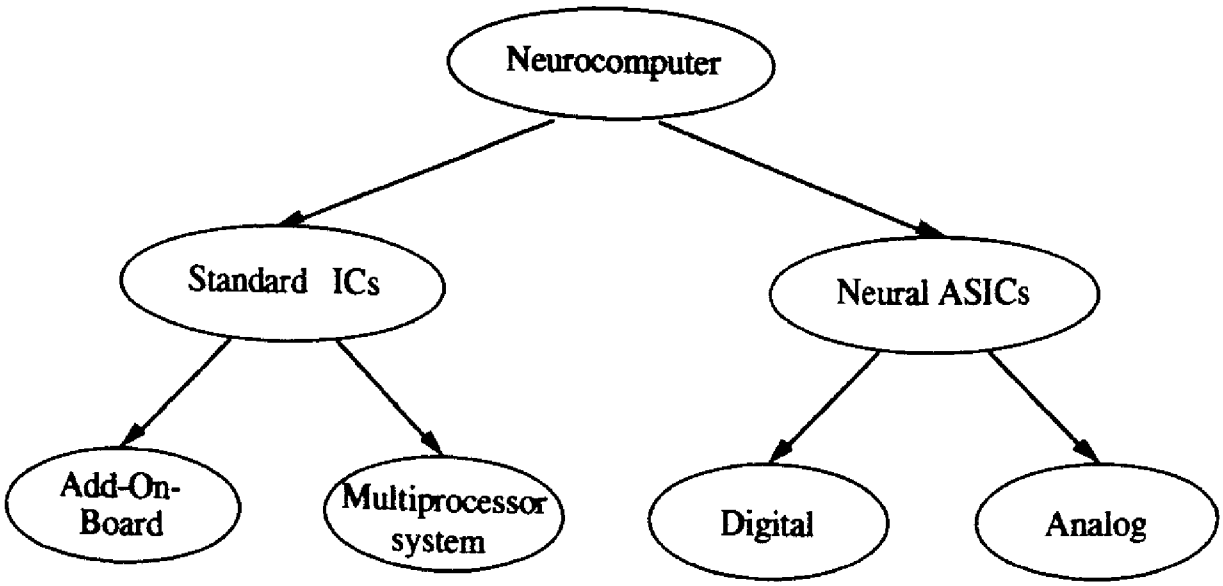

Figure 1 Neurocomputer categories

Neurocomputers based on standard ICs

Many ANN operations are based on a sum of products and are quite similar to operations (e.g. filtering) in digital signal processing (DSP). This similarity suggests that much of the DSP technology could be applied to accelerate ANN operations. Several institutions have developed board-level systems, so called add-on-boards or accelerator boards, built around a modern micro-

processor (proprietary, DSP, or RISC (reduced instruction set computer)) that excel at the floating-point operations required by many ANNs. Accelerator boards speed up ANN processing by at least an order of magnitude albeit at a significantly greater cost. Most of them are designed for IBM PCs and compatibles and simulate ANNs with supercomputer-like performance of up to 50MCPS (million connections per second) and 20MCUPS (million connection updates per second) [1]. MCPS, which often refers to the number of multiply-and-add operations that can be performed in a second, and MCUPS are common speed measures for the retrieval and learning phase of ANN simulators. Obviously, both terms have to used with great care. Firstly, because speed is more a matter of convenience than of functionality. The quoted MCPS may or may not include the time to fetch the variables that are to be multiplied, for example. Hence, more data are needed to appraise computational speed. Secondly, both terms are only useful in connexion with a neural network model and its application, because there exist no widely accepted benchmarks for ANNs yet. Nevertheless, the terms MCPS and MCUPS are both quite important measures and are very common in the ANN hardware community.

The next step towards higher performance are multiprocessor neurocomputers which are very similar to general-purpose parallel computers. The connections between processors can either be through a single high speed data path (bus-oriented) or via short point-to-point links between processors. For a bus-oriented system the number can vary between two and a few hundred complex processors, for example, whereas in the other case the system can have several thousand simple processors (e.g. up to 16384 for the Connection Machine [1]). The interconnection needs for ANNs pose a special challenge for parallel neurocomputers. As the nature of ANNs is to emphasize a high degree of connectivity, a massively parallel system can sustain dramatic decrease in throughput because of communication delays. Another difficulty is the inconsitency of the need for flexibility and the difficulty of efficiently programming parallel systems. In literature, performance figures of up to 500MCPS and 200MCUPS for multiprocessor neurocomputers can be found [1-3].

Since the components for such neurocomputers already exist they are being improved as fast as the state of the art of technology will allow it. This is obviously a strong point for neurocomputers based on standard ICs. The challenge in ANN implementation lies in developing appropriate architectures to effectively use these standard components.

Neural ASICs

An alternative way in implementing neural networks is the design of special-purpose integrated circuits. There has been much work in the design

of VLSI ICs implementing ANN algorithms. In IC design the options are digital or analog or a combination thereof.

Digital neural ASICs

Digital ASICs are the most flexible and mature ANN implementations. The obvious advantages of digital designs are high computational precision, high scalability, and the possibility to implement modifiable synapses with almost unlimited precision. Furthermore, powerful design tools are available for digital full- and semi-custom design. The disadvantages are the relatively large circuit size for low and medium sized ANNs and a semi-parallel implementation of the weighted sum of inputs.

Digital neural ASICs can be categorized into general-purpose approaches, for emulating different ANN models, and special-purpose approaches which are dedicated to a specific ANN model. However, the more general a neurocomputer is the slower it is. The synaptic weights can be stored on or off chip. The advantage of on-chip weight storage is the fast weight access. The disadvantage is that the number of necessary chips grows linear with the number of neurons.

The N6400 chip from Adaptive Solution Inc. [4] is an example for a general purpose digital neural ASIC with on chip storage of weights. Each N6400 contains 64 processing units and 256KBytes weight memory (1-16 bit weight resolution) on chip. A single chip (25 MHz) can perform 1.6GCPS (256MCUPS, backprop) for 8 or 16 bit weights and 12,8 GCPS for 1 bit weights, respectively. The CNAPS neurocomputer contains in the standard version four N6400 and has four-times the performance. Another interesting example of this category is the WSI (wafer scale integration) neural net from Hitachi [5]. It comprises 576 digital neurons and 36KBytes weight memory. It performs 1.25 GCPS and 118MCUPS.

The MA16 neural signal processor from Siemens is an example of a general-purpose digital neural ASIC with off chip weight storage. The MA16 chip has a systolic architecture build by four processing units each with four 16bitx16bit multipliers. The chip performs about 800MCPS at a clock rate of 50MHz. A prototype of the Synapse-1 neurocomputer has eight MA16 neural signal processors, two MC68040 CISC processors for control purposes, and a 128MByte DRAM bank. The prototype performs 4.2GCPS and 330MCUPS [6].

These examples of digital neural ASICs show the potential for further improvement of ANN simulation speed compared with general-purpose parallel computers. A speed of up to two orders of magnitude higher seems to be possible with current technologies.

Analog neural ASICs

Analog neural ASIC approaches can be divided into continuous-time and discrete-time analog electronics. Some additional options arise relating to the connectivity (local/global or low/full connetivity) within the ANN and the transistor's mode of operation (weak inversion or strong inversion) [7,8]. The advantages of analog designs are their compactness and their high speed due to the saving of device functions by functional integration. On the contrary, the design of analog circuits demands much more time and a good theoretical knowledge about transistor physics [8]. Analog integrated circuits are susceptible to noise and process-parameter variations resulting in a limited computational precision. Last but not least it is still difficult to integrate adaptive synapses (weights).

Nevertheless, the majority of microelectronic ANN implementations use analog computation at least to some extent. Analog processing derives it main advantage when physical processes can be used to perform required computational functions. For example, the weighted sum of input signals (activation function) which incurs the largest computational load in the recall phase can be efficiently implemented in analog circuit technique by means of current or charge summing [7-9]. Most of the proposed analog neural circuits make use of current summing as, for example, the associative memory chips from AT&T [9] or from the University of Dortmund [10]. With current state-of-the-art microelectronics simple neural associative memory chips with more than 1000 neurons and 1000 inputs each can be integrated on a single chip performing about 100GCPS. Such systems have been used for pattern classification and image segmentation, for example.

The associative memory chips mentioned above are programmable, but not trainable. Learning, or selforganization, does require incremental adjustment of the synapses (weights) in small steps. The design of multivalued weights must balance the cell size and the resolution of the weight. Whereas the implementation of digital memories are well-mastered techniques, storage in analog memories is still difficult. Proposals for analog synapses include charge-coupled devices (CCDs), MNOS (metal nitride oxide silicon) transistors combined with CCDs, and concepts based on special materials like bismuth sesquioxide or a-silicon [11]. Very promising concepts offer floating-gate transistors as used in EEPROMs [11] and analog storage cells with ferro-electric films, e.g. PZT [11,12]. For these proposals several problems have to be analyzed, for example, the amount of process-parameter variations across the chip, the defect yield, storage time (volatility), and compatibility to standard VLSI processing technology.

Analog and programmable weights are a significant feature of an effective and flexible VLSI implementation of ANNs. At the moment, there is one such analog neural ASIC commercially available, which uses two floating-gate

transistors as an adaptive weight [13]. It contains 64 neurons, 10240 synapses and performs about 2GCPS. Learning has to be done off chip.

A rather new analog design methodology is the so called *neuromorphic approach*, which can achieve significant improvements in computing hardware capability compared with conventional analog and digital techniques. This neuromorphic approach starts by identifying several structural levels in the nervous system, and then attempts to capture these organizing principles. At the lowest level, the computational primitives of the nervous system are identified and silicon analogies are designed by creatively harnessing the available physics of semiconductors. At the next level various ensembles of primitives can be organized to perform complex computational tasks, such as signal preprocessing. This novel methodology was inspired by the work of C. Mead [8,14] and has let to several interesting prototype chips [8], for example, a silicon retina, an electronic cochlea, or ear prothesis. Analog circuits, hard-wired for a specific function and with local connections, can deliver extremely high performance while dissipating very low power. The combination of sensors and computing circuitry on one chip makes them look promising for smart sensors. At the moment, they are still in research state where their potential is being explored.

Analog circuits can boost the performance beyond that of digital designs. But in general they are special-purpose implementations of a selected ANN model for a specific application. For small and medium ANNs analog neural ASICs are more compact than the digital counterparts, but not as flexible as digital systems. Analog systems hold undisputed claim to the interfaces between digital systems and the real, analog world. Fast, very high-accuracy analog-to-digital and digital-to-analog converters are often the critical items that make the application of intermediate digital processing possible. In the near future analog techniques will make possible dynamic neurons, which will lead to networks with new features [15].

Discussion

One of the most important differences between ANN research today and what was possible 30 years ago is the huge improvement in the technological capabilities. The progress in microelectronics provides a powerful basis to implement large neural networks in electronic or opto-electronic hardware. State-of-the-art VLSI and the emerging ULSI technologies are able to integrate thousands of neurons on a single chip with clock rates reaching 1 GHz. In the far future even more computational power for ANN may be obtained by using optoelectronic and molecular electronics. Optoelectronic devices and light wave guides integrated on silicon offer interesting aspects for ANN. First of all we get a flexible interconnectivity with high data rates and many data can be processed in parallel. Furthermore, new alternatives

for analog storage of weights (synapses) are offered by using PZT films, for example [16].

Much higher integration densities follow from nano and molecular electronics. For example, the crystal Zeolith with its regular pipeline structure, where conducting polymers and semiconductor molecules are embedded, offers a very high density of weights, about 10^9 mm^{-3}. The drawback is the very simple architecture with low interconnectivity, which is similar to cellular nets [16]. Molecular electronics offers a huge potential for ANN [17]. The tactile molecular processing unit with proteins has a relative low processing speed, but solves the interconnectivity in an elegant way: The information packets flow parallel in the cell liquid all searching their neural goals.

Despite of such features the discussion about the technological way to large neural systems, in the long term to so-called artificial brains, is open. At long term it might be the way of biological concepts with proteins or the physical way with nanostructured devices. For the next decade microelectronics will dominate the field of ANN implementation. A large number of design studies on ANN hardware implementation have been carried out in U.S., in Japan and in Europe. In Europe and Japan, digital implementation precedes analog and optoelectronic approaches. On the contrary, in U.S. the analog approach has been prevailing [2,3]. A number of neurocomputers - specialized machines able to efficiently implement neural networks - have been built, and a few are now commercially available [18].

The neurocomputer categories as shown in Fig. 1 and reviewed in this paper are summarized in respect to their speed performance in Fig. 2. The most commonly used category of neurocomputers are the desktop workstations. They offer a large amount of ANN simulators as well as graphics and support software, but they are quite slow compared with the other implementations. A straight forward way to enhance their performance is to couple them with Add-On-Boards offering the best price/performance ratio. Neurocomputers of this category can perform up to 100 MCPS at present. An order of magnitude faster are multiprocessor neurocomputers which are very similar to general-purpose parallel computers, provided software exists.

The next step towards higher performance are special purpose VLSI components. The digital approach is currently the best solution to implement ageneral purpose neurocomputer with high precision. Analog circuits find applications in front-end signal processing with low precision, in particular pattern recognition. Microelectronic implementations of neurocomputers promise to make cost-effective and physically small neural-based products possible. General-purpose computers and simulators, however, often fail to meet the size or cost constraints of designs that must be deployed in the real world.

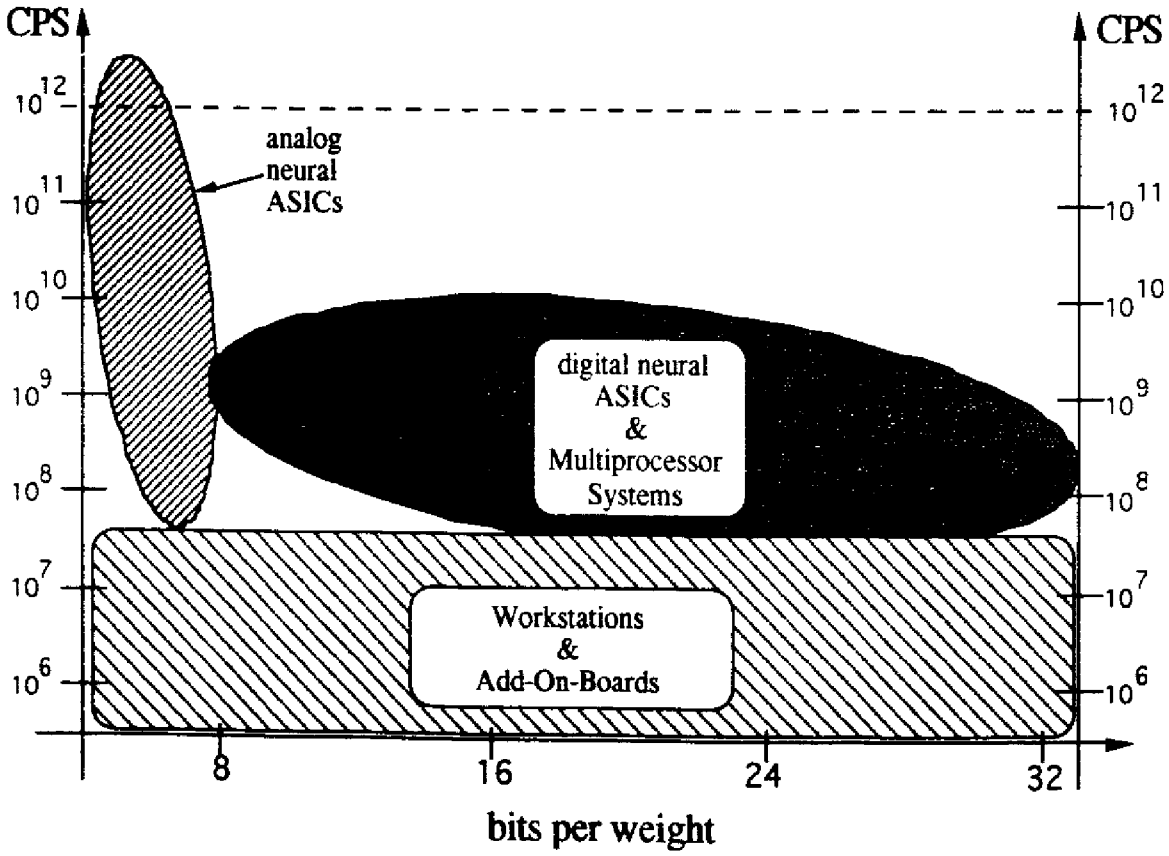

Figure 2 Speed of computation versus resolution in the computation

For the next decade the perspective of microelectronics are not narrowed by technical limitations, even if progress advances as fast as in the past. This implies for all neurocomputer categories an increase in performance by about two orders of magnitude. Consequently, neurocomputers are approaching the TeraCPS (10^6MCPS) performance. In other words, the output of an ANN with 100 million neurons each having about 10.000 inputs (synapses) can be computed within one second. As already mentioned, these performance figures have to be used with great caution. The simulation of ANN paradigms by means of a simple weighted summer and a threshold activation is much simpler than the emulation of biological oriented ANNs (pulse-coding, spatio-temporal parallelism). But they give an idea of the large potential of available technologies. Of course, comparing this number with the 10^{12} neurons of the human brain this potential may not be impressive. On the other hand, there are living organisms with smaller brains showing clever as well as fascinating behaviour. The forthcoming neurocomputer generation will offer new opportunities to alternative ANN design for discovering such neural organizing principles. It is in that spirit that progress in hardware implementations will hopefully contribute to a

better understanding of paradigms and biological systems as well as a number of useful applications.

Acknowledgement

I gratefully acknowledge the help of Prof. Karl Goser (University of Dortmund) for his advice, encouragement, and support of this work. I would like to thank the DFG (Deutsche Forschungsgemeinschaft) and the BMFT (german ministry of research and technology) for financial support.

References

[1] DAPRA, Neural Network Study, AFCEA Int. Press, Fairfax Verginia, 1988.

[2] Goser, K., Ramacher, U., Rückert, U.: Microelectronics for Neural Networks, Proceedings of the 1st Int. Workshop, University of Dortmund, 1990.

[3] Ramacher, U., Rückert, U., Nossek, J.A., Microelctronics for Neural Networks, Proceedings of the 2nd Int. Workshop, Kyrill&Method Verlag, München 1991.

[4] Hammerstrom, D.: A VLSI Architecture for High-Performance, Low-Cost, On-Chip Learning, Proc. IJCNN, Vol. 2, June 1990. pp. 537-544.

[5] Yasunaga, M., et.al.: Design, Fabrication and Evaluation of a 5-inch Wafer Scale Neural Network LSI composed of 576 Digital Neurons, Proc. IJCNN, Vol. 2, June 1990, pp. 527-535.

[6] Ramacher, U.: SYNAPSE-A Neurocomputer That Sysnthezis Neural Algorithms on a Parallel Systolic Engine, Journ. of Parallel and Distributed Computing 14, 1992, pp. 306-318.

[7] Ramacher, U., Rückert, U.: VLSI Design of Neural Networks, Kluwer Academics, Boston 1991.

[8] Mead, C.: Analog VLSI and Neural Systems, Addison-Wesley, 1989.

[9] Graf, H.P., Henderson, D.: A Reconfigurable CMOS Neural Network, IEEE Int. Solid State Circuits Conf. Dig. Tech. Papers, Feb. 1990, pp. 144-145, 285.

[10] Rückert, U.: An Associative Memory with Neural Architecture and its VLSI Implementation, Proc. of the HICSS '91, IEEE Computer Society Press, 1991, pp. 212-218.

[11] Goser, K., Hilleringmann, U., Rückert, U., Schumacher, K.: VLSI Technologies for Artificial Neural Networks, IEEE Micro, Vol. 9, No. 6, pp. 28-44.

[12] Goser, K., Hilleringmann, U., Rückert, U.: Applications and Implementations of Neural Networks in Microelectronics - Overview and Status, in: Monaco, V.A., Negrini, R. (ed.): Advanced Computer Technology, Reliable Systems and Applications, IEEE Comp. Society Press, Bologna 1991, pp. 531-536.

[13] Holler, M., Tam, S., Castro, H., Benson, R.: An Electrically Trainable Artificial Neural Network (ETANN) with 10240 'Floating Gate'Synapses, Proc. of the Int. Joint Conference on Neural Networks, Washington, June 1989.

[14] Mead, C.: Neuromorphic Electronic Systems, Proc.IEEE Vol.78, No.10, 1629-1636.

[15] Murray, A.F.: Pulse Arithmetic in VLSI Neural Networks, IEEE Micro, Vol. 9 No. 6, Dec. 1989, pp. 64-74.

[16] Goser, K: Challenge of ANN to microelectronics, Proc. of the ICANN 93, Amsterdam, to be published.

[17] Conrad, M.: Molecular Computing, Computer, Vol.25,No.11, 6-81.

[18] Rückert, U., Spaanenburg, L., Anlauf, J.: Hardwareimplementierung Neuronaler Netze, (in german) atp 7/93, to be published.