# Design and VLSI Implementation of a Unified Synapse-Neuron Architecture

*H. Djahanshahi, M. Ahmadi, G.A. Jullien and W.C. Miller*
Department of Electrical Engineering, University of Windsor,
Windsor, Ontario N9B 3P4, Canada
djahans@engn.uwindsor.ca

## Abstract

*We describe the design and VLSI implementation of a unified synapse-neuron architecture for multi-layer neural networks. A new hybrid building block proposed for this purpose is formed by integrating a partial S-shape neural nonlinearity within a Multiplying DAC synapse. MDAC synapse contains modifications to simplify sign-bit circuit. Small analog circuits generate a distributed S-shape neural function by combining quadratic characteristics of four MOS transistors. The proposed modular neural network architecture features design simplicity and scalability, area efficiency, reduced interconnection problem, improved robustness and digital programmability. Based on the proposed scheme, we have considerably increased the synaptic density in the improved version of a programmable optically-coupled neural network.*

## 1. Introduction

VLSI neural networks are becoming more popular because of providing real-time solutions for many real world problems. However, neural network VLSI designers face many challenges, for instance, in implementing massively interconnected networks, producing fully parallel input-outputs and developing modular and scalable architectures that can easily be adopted for different applications. Area and power efficiency, speed, storage and calculation accuracy are some other major issues. Compromise solutions now lead to hybrid circuits: mixed analog and digital, or even optoelectronics. We have been investigating alternative hybrid architectures for flexible and area-efficient, yet non-multiplexed implementation of optically-coupled neural networks designed for real-time applications [1], [2], [3].

Conventional electronic neural networks consist of two types of building block: synapse and neuron. A (linear) synapse performs multiplication while a neuron provides summation and sigmoid (S-shape) transfer characteristics.

If the output of synapse is a current, summation is done for free by connecting the outputs of synapses together (i.e. KCL). Moreover, a nonlinear resistive load can perform S-shape I-to-V operation and complete the neuron's task. A neuron of this type can be *distributed* among all associated synaptic inputs generating a *unified synapse-neuron* building block, also known as *distributed neuron-synapse*. Reference [4] provides a discussion of the subject and the details of an all-analog implementation. As an alternative, here we present a hybrid digital-analog distributed architecture and its basic building block. In section 2 this architecture is described and in section 3 its sub-blocks and circuit designs are introduced. The overall characteristics of the unified synapse-neuron is presented in section 4 and an application of the proposed architecture is briefly mentioned in section 5. Section 6 is conclusion.

## 2. Hybrid Distributed Architecture

We present a digital-analog building block for modular implementation of hybrid neural networks. The proposed architecture and building block are then used in an optically-coupled neural network designed for a process control and pattern classification application. Our design combines a Multiplying DAC synapse and a modified distributed neuron, to generate a unified synapse-neuron building block. This modular design integrates a partial S-shape neural nonlinearity in each MDAC synapse cell. As a result, it becomes the only block required, besides digital weight memory, to build a complete multi-layer neural network.

Parallel output connection of $n$ such building blocks, for instance, generates a neuron with $n$ digitally-programmable input synapses. Figure 1 shows 3 such neurons each having 4 inputs, i.e. a {4, 3} neural network built with 12 hybrid building blocks.

An {m, n, p} feed-forward network can be implemented simply by interconnecting regular ( $m \times n$ ) and ( $n \times p$ ) arrays of such building blocks. Figure 2 shows a typical {4, 3, 2} neural network built with $4 \times 3 + 3 \times 2 = 18$ hybrid blocks. Such a regular structure is very attractive



0-8186-7502-0/96 $5.00 

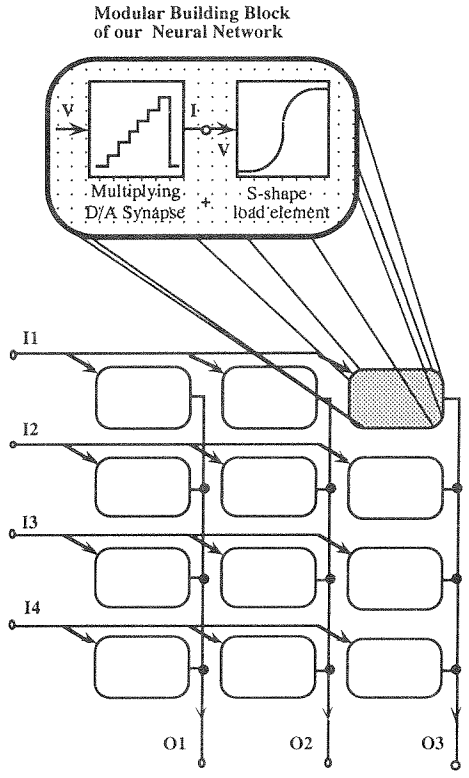


**Fig. 1:** A two-layer {4, 3} feed-forward Neural Network, I1...4: Input layer, O1...3: Output layer

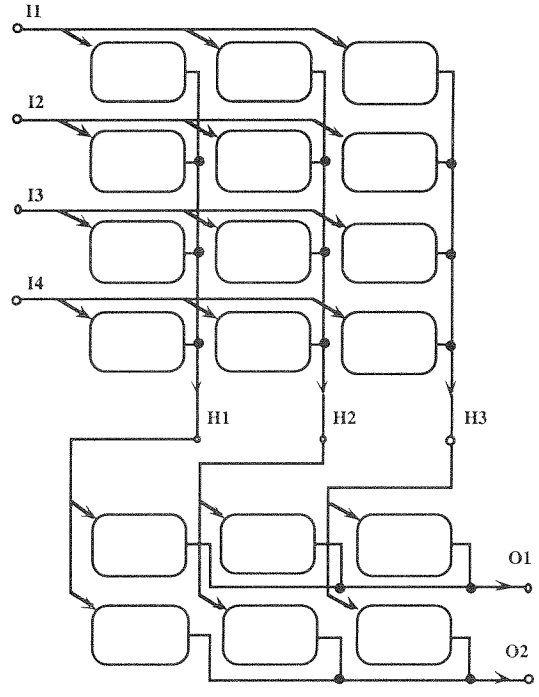


**Fig. 2:** A three-layer {4, 3, 2} Neural Network, I1...4: Input layer, H1...3: Hidden layer, O1,2 : Output layer

for VLSI implementation. Additional modules of the same type can be used for neural threshold (bias) adjustment. Threshold blocks need a constant non-zero analog input voltage and their threshold value is a digital input to MDAC from weight memory.

## 3. Components of Our Building Block

Two subcircuits of our design, namely MDAC and S-shape distributed neuron, both contain new modification and have been separately optimized through extensive simulations. The overall performance of the combined cell is then characterized in post-layout simulations. Cadence 4.3 software tools have been used from schematic .capture to layout and extraction level. Simulations are performed using *Spectre* ™ in *Analog Artist* ™ environment.

### 3.1. Multiplying DAC

Figure 3 shows the schematic and layout of our MDAC synapse circuit. MDAC receives a 5-bit sign-magnitude input from weight memory. The circuit consists of a set of binary-weighted current mirrors and a sign-bit circuit at the top. A layout technique known as " ΔW correction" is used in binary-weighted current mirrors [8].

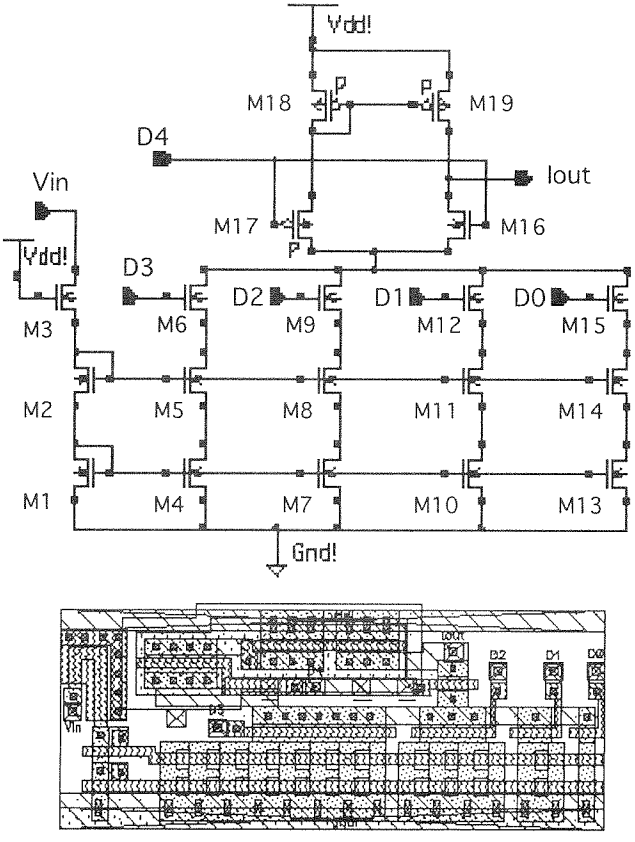


**Fig. 3:** Schematic and layout of Multiplying DAC Synapse

We have modified the sign-bit circuit compared to the circuit described in [5] such that it only needs D4 input (instead of both D4 and D4'), saving an inverter or an extra interconnection line per each synapse. Each saving related to synaptic circuit counts because of the great number of synaptic cells and interconnections that eventually occupy a considerable amount of chip area. The sign circuit consists of only 4 transistors, namely M16, M17, M18 and M19. They produce a bi-directional output current. When D4 (sign-bit) is HIGH, M17, M18 and M19 are OFF and only M16 is ON that *sinks* the binary-weighted current from output terminal. When D4 is LOW, M16 is OFF and the other three devices are ON that *source* the binary-weighted current to $I_{out}$ terminal.

Table 1 shows the device widths and lengths (W/L) in MDAC synapse circuit. Figure 4 shows MDAC output current simulation as binary weight increases successively from -15 to 15. MDAC is operating, within ±1% linearity margin, as a weight-dependent current source with an output current of -100μA to +100μA.

| | | | |
|---|---|---|---|
| M1-M2: | 3.2/2.0 | M3: | 2.4/24. |
| M4-M5-M6: | 25.6/2.0 | M7-M8-M9: | 12.8/2.0 |
| M10-M11-M12: | 6.4/2.0 | M13-M14-M15: | 3.2/2.0 |
| M16: | 6.0/2.0 | M17: | 14.0/2.0 |
| M18-M19: | 12.0/2.0 | | |

**Table 1:** MDAC device sizes (W/L) in μm

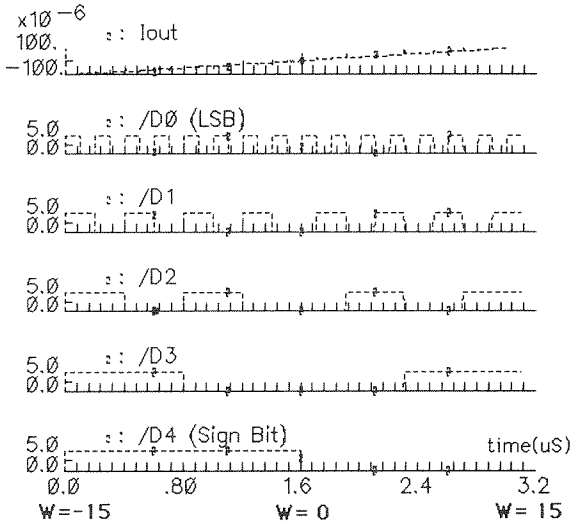


**Fig. 4:** MDAC output current vs. successive weights at $Vin_{max}$

## 3.2. Modified S-shape Distributed Neuron

An I-to-V neuron based on a nonlinear load is presented in [4] where both lumped and distributed implementations are discussed. Figure 5-a shows the schematic circuit of this neuron. Nonlinear characteristic of this neuron circuit is a combination of two quadratic curves (from M1 and M2) and a linear transition part corresponding to R. To realize the resistor's task, some designs rely on existing parasitic/leakage impedances while the others may use a MOS device.

Here, a modified circuit is presented with 4 MOS devices (and no R) that approximates a S-shape neural function with 4 quadratic characteristics. Circuit diagram and layout of the new circuit is shown in figure 5-b. With a careful selection of the bias voltages ($V_{Bias1}$ and $V_{Bias2}$) and device geometries, we have been able to use only 2 common bias voltages for 4 MOS transistors.

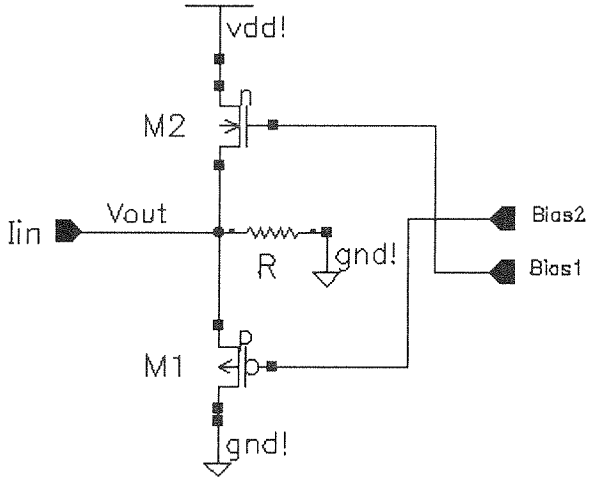


**Fig. 5 a:** Original I-to-V Neuron based on Nonlinear Load

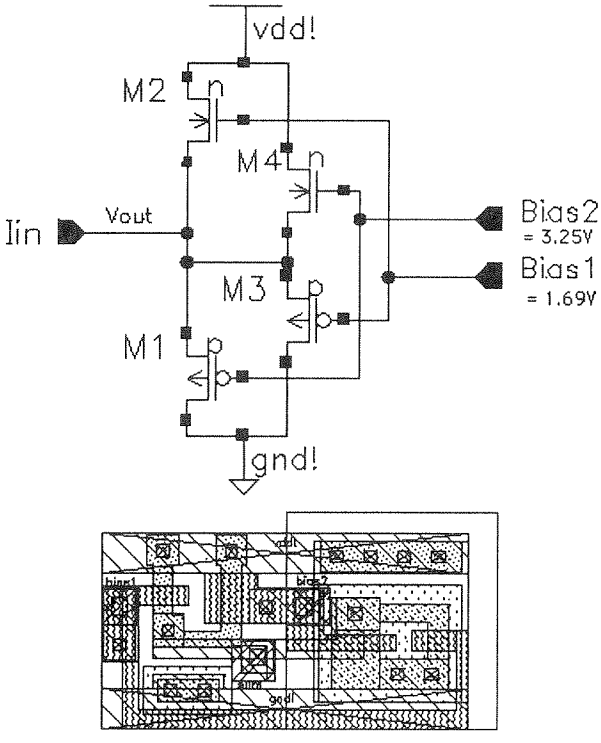


**Fig. 5 b:** Modified I-to-V Neuron (S-shape distributed element)

In fact, two additional devices, M3 and M4, are replacing R with a lightly S-shaped characteristic in the region where M1 and M2 are both OFF. Cell layout is 36.4μm x 19.4μm in 1.2μ CMOS4S process. The overall layout overhead of this design is not noticeable as each transistor now requires a smaller width and besides, R has been removed.

Simulation results comparing V-I characteristics of 4-MOS nonlinear neuron with 2-MOS version is shown in figure 6. In the absence of R, the circuit of figure 5-a shows a stepwise transition while the new circuit (figure 5-b) performs a smoother transition. This circuit can be designed to have a differentiable characteristic by slightly overlapping the conduction regions of M3 and M4. In the present example M3 and M4 are set just at the conduction threshold at Vout=2.5V. Note that sub-threshold conduction exists for both devices at this point, so the slope of transfer characteristic is still limited. Table 2 specifies the conduction regions of each of the 4 devices. $V_{TN}$ and $V_{TP}$ are the threshold voltages of NMOS and PMOS devices respectively. Device widths are adjusted to reach 0 and 5V at ±100μA input current.

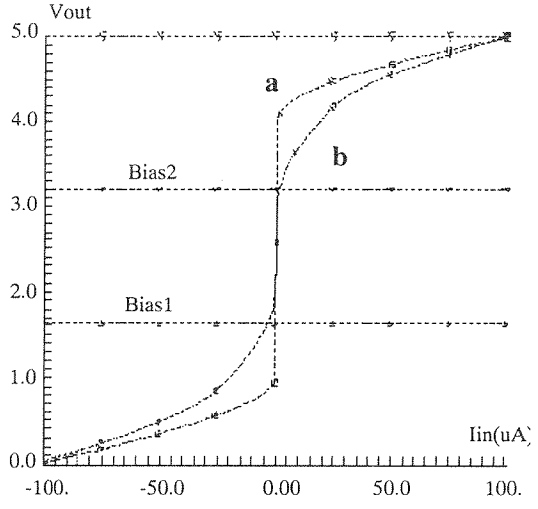


**Fig. 6:** Nonlinear characteristics of circuits in Fig. 5-a and 5-b

| Region | Vout | Devices conducting |
|---|---|---|
| (I) | $Vout < V_{Bias1} - V_{TN}$ | M4, M2 |
| (II) | $V_{Bias1} - V_{TN} < Vout < V_{Bias2} - V_{TN}$ | M4 |
| (III) | $V_{Bias1} + \|V_{TP}\| < Vout < V_{Bias2} + \|V_{TP}\|$ | M3 |
| (IV) | $Vout > V_{Bias2} + \|V_{TP}\|$ | M3, M1 |

**Table 2:** Device conduction regions

Now we investigate other properties of the modified distributed neuron:

**3.2.1. Automatic Scaling of Neuron.** In principle, we have to re-design a neuron for different number of input synapses, N, in order to scale the sigmoid function properly. If each synapse generates an output current between $-I_0$ and $+I_0$, then N synapses can maximally produce a current from $-N.I_0$ to $N.I_0$ . A sigmoid nonlinearity suitable for one synapse or two, would look like a hard-limiting function for moderate to large number of input synapses (e.g. N≥5) because of large saturating areas involved. A scaling scheme proportional to $N$ [4] or $\sqrt{N}$ (based on statistical analysis [6]) should be considered.

A distributed neuron design, e.g. the present design, provides a modular solution to scaling problem. $N$ building blocks (partial nonlinearities) in parallel, automatically generate an stretched S-shape scaled by a factor of $N$. Figure 7 depicts the auto-scaling property of our neuron sub-block for N=5 compared to N=1. When partial nonlinearities are distributed among MDAC synapses, each synapse brings its own share and

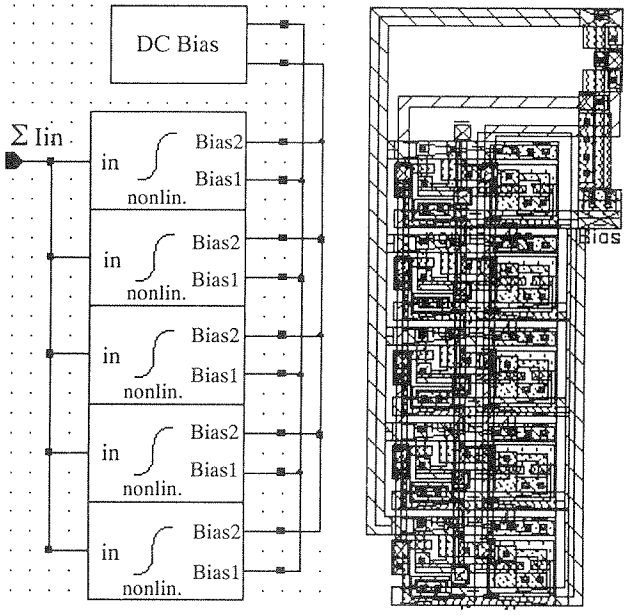


**Fig. 7 a:** Configuration of nonlinear neural elements (schematic and layout) to examine scaling property

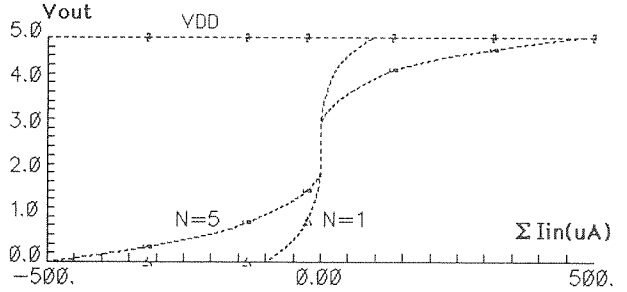


**Fig. 7 b:** Auto-scaling property of nonlinear blocks for N=1 and N=5 (post-layout simulation)

incrementally adjusts the overall neural S-shape characteristics. This greatly simplifies the design of a network for different number of neurons and synapses.

**3.2.2. Increased Robustness.** Since the neuron is distributed among $N$ sub-blocks, there is an increased robustness and fault-tolerance in this architecture. For example, a VLSI defect would affect only $1/N$th of a neuron instead of the whole.

In essence, the present circuit with S-shape characteristics provides a modular and scalable I-to-V neural function which is differentiable and well-defined in a narrow range despite some process variations. This makes the neuron transfer function suitable for in-loop training using popular gradient-based algorithms. While it is used as a distributed neuron element in our hybrid building block, the present design can equally be used in all-analog implementations. Moreover, a lumped neuron may be built with the same circuitry but with different device geometries.

## 4. Characteristics of Unified Synapse-Neuron

We have integrated our modified (4-MOS) S-shape nonlinearity within each MDAC synapse cell described earlier. The result is a unified synapse-neuron (USN) which is the basic building block of our neural network architecture. The layout of the unified synapse-neuron cell is 126μm × 42μm based on standard 1.2μm CMOS4S technology.

To store digital synaptic weights on chip, we have used compact Read/Write registers. This will make the network programmable after different training sessions. A 5-bit custom-designed memory register operating with double-phased clocks is used in conjunction with each USN building block. Figure 8 shows the post-layout simulation of a unified synapse-neuron output voltage versus successive synaptic weights at Vin=4.5 V.

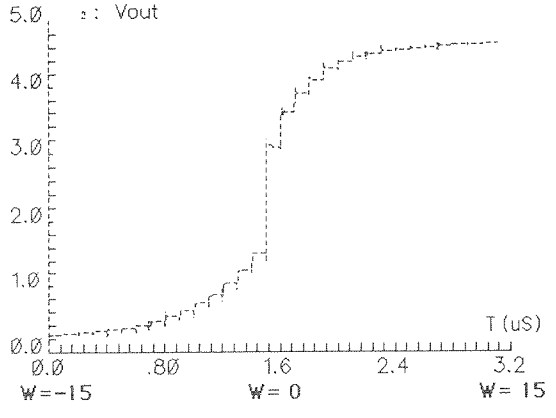


**Fig. 8** Overall characteristics of unified synapse-neuron (discrete S-shape for Vin=4.5 V)

Unified synapse-neuron block inherits the properties mentioned earlier for distributed neuron, e.g. automatic scaling. Moreover, there is a new property to be mentioned here which is obtained from parallel configuration of MDAC cells:

### 4.1. Automatic Fan-out Increase

In many practical applications considerable number of neurons exist in each layer of a multi-layer network. Each neuron must be able to drive all of the outgoing synapses to the next layer. Total load includes interconnection lines and input impedances of all synapses. Circuit techniques, especially the use of high-drive buffer amplifiers, have been proposed in this context to resolve the fan-out problem [7].

A neuron in our architecture is a "more-fan-in more-fan-out" entity: a neuron with a higher number of input synapses inherently consists of a higher number of parallel transistors from MDAC output and S-shape circuit. This is equivalent to output transistors with proportionally increased width and hence higher drive. Therefore, with more input synapses a neuron becomes potentially capable of driving more synaptic outputs to the next stage, if required.

## 5. Optically-Coupled Neural Network Application

We have been involved in the design and test of CMOS photosensor arrays and optically-coupled neural networks [1], [2], [3]. Our specific application requires the design and VLSI implementation of a smart photosensor that can be used in process control to determine the position or classify the surface geometry of an object whose image is captured on chip using LASERs or beam-steering methods. The sensor architecture is a mixed analog-digital CMOS VLSI realization of a multi-layer artificial neural network with an integrated photosensitive array. Light-sensitive elements are based on parasitic BJT phototransistors of CMOS technology [3], [9]. An integrated array of such photoreceptors act as the input nodes to a programmable neural network classifier.

The latest design of optically-coupled neural network (sensor) is based on the proposed unified synapse-neuron building block approach that results in a highly modular and scalable VLSI architecture. A chip containing an array of photosensitive elements and a fully-connected programmable neural network classifier with input, hidden and output neurons has been designed for fabrication in 1.2μ CMOS4S technology (details will be reported elsewhere). Weights will be programmed on chip after an off-board training session. In actual recall operation an object would be "imaged" onto the photosensitive array and the states of the neurons in the output layer would then define a control vector for on-line control based on a non-contact measurement.

Using the modular design approach mentioned in this paper, we have effectively doubled the number of synapses per die area in comparison with the previous version of our programmable neural-based sensor reported in [1]. Also interconnection problem has been greatly reduced. Therefore, we have been able to increase the dimensions (resolution) of on-chip photoreceptor array as well as the size of neural network itself. The total density of synapses in our new chip is increased by more than 100% compared to the previous design.

## 6. Conclusion

A hybrid analog-digital distributed architecture has been developed and proposed for VLSI implementation of neural networks. Basic building block of this architecture combines a linear Multiplying DAC synapse and a modified distributed S-shape neuron, hence providing a unified synapse-neuron hybrid block. This module is the only block required, together with digital weight memory, to build a complete hybrid multi-layer neural network. A properly interconnected $m \times n$ array of this building block implements a two-layer {m, n} neural network. Additional arrays can simply be added for multi-layer implementations. Regularity of this architecture well suits it for VLSI realization.

A test chip containing our basic subcircuits, cells and test networks has been fabricated in 1.2μ CMOS[1] technology. Preliminary test results have been satisfactory. To show the potential improvement, we have used the unified synapse-neuron cells in a larger design targeted for a real application. This is an optically-coupled neural network that can be used in process control and pattern classification applications. Using a modular and scalable approach, we have designed a sensor chip for fabrication in 1.2μ CMOS which contains a photosensitive input array integrated with a multi-layer neural network. The synaptic density in the new design has been increased by more than 100% compared to [1] which is designed and fabricated based on conventional architectures.

Modularity is the main feature of the present hybrid design. A neural network built with the proposed synapse-neuron architecture: a) can easily be designed or re-designed for different applications requiring various number of layers and/or number of neurons per layers, b) has a highly regular VLSI architecture hence the interconnection problem and inter-cell area is reduced, c) is more robust and fault-tolerant compared to conventional architectures and d) is digitally programmable.

[1] All fabrications are in CMOS4S, a 1.2um double-metal N-well CMOS process, through Northern Telecom fabrication facilities.


### Acknowledgements

This project would not have been possible without the financial support of NSERC and Ortho-McNeil Inc. We would also like to thank Canadian Microelectronics Corporation (CMC) and Northern Telecom for providing us with software tools and fabrication services.


### References


[1] K.W. Lei, G.A. Jullien and W.C. Miller, "A Programmable Intelligent Optical Sensor Realization", *Proc. of the 37th Midwest Symp. on Circuits and Systems* , August 1994, pp. 465-468.

[2] B. Lam, W.C. Miller and G.A. Jullien, "An Intelligent Optical Sensor", *Proc. of International Conf. on Applications of Photonic Technology (ICAPT)*, Toronto, June 1994.

[3] G. Liang and W.C. Miller, "A Novel Photo BJT Array for Intelligent Imaging", *Proc. of the 36th Midwest Symp. on Circuits and Systems* , August 1993, Detroit, pp. 1056-1059.

[4] S. Stayanarayana, Y. Tsividis and H.P. Graf, "A Reconfigurable VLSI Neural Network", *IEEE Journal of Solid-State Circuits* , Vol. 27, No. 1, January 1992, pp. 67 -81.

[5] A. Moopen, T. Duong and A.P. Thakoor, "Digital-Analog Hybrid Synapse Chips for Electronic Neural Networks", in *Advances in Neural Information Processing Systems* , Vol.2, Ed. San Mateo, 1990, pp. 769-776.

[6] H.C.A.M. Withagen, "Reducing the Effect of Quantization by Weight Scaling", *Proc. IEEE International Conf. on Neural Networks* , June 28- July 2 1994, pp. 2128-2130.

[7] N. Yazdi, M. Ahmadi, G.A. Jullien and M. Shridhar, "A Large-Swing High-Drive CMOS Buffer Amplifier for a Wide Load Range", *Journal of Circuits, Systems, and Computers,* Vol. 2, No. 4, 1992, pp. 323-333.

[8] P.E. Allen and D.R. Holberg, *CMOS Analog Circuit Design.* New York: Holt Rinehart and Winston, 1987.

[9] C. A. Mead, *Analog VLSI and Neural Systems.* MA: Addison-Wesley, 1989.