Speed Enhancement with Soft Computing Hardware

Taher Daud, Adrian Stoica, Ricardo Zebulum, Tuan Duong, and Anil Thakoor
Jet Propulsion Laboratory, California Institute of Technology
Pasadena, California, USA

Abstract

During the past few years JPL has been actively involved in soft computing research encompassing theory, architecture, and electronic hardware applications. There are a host of soft computing applications that require orders of magnitude enhancement in speed compared to that obtained using simulations on digital machines. For on-board computing this is made possible by selecting suitable algorithms, designing compatible architectures and implementing them in parallel processing hardware. Compact low-power hardware designs include neural network multi-chip module performing high-speed object classification and recognition with 10^{12} multiply-sum operations per second (ops). Additionally, development on evolvable hardware implemented on suitable electronic hardware has shown exciting high-speed evolution of various digital and analog circuits. We will review our work on electronic neural networks and evolvable hardware to bring out speed advantage.

Key Words: Soft Computing, Neural Networks, Evolvable Hardware, Speed, Neuroprocessor.

1 INTRODUCTION

The Jet Propulsion Laboratory (JPL) is conducting research in various aspects of information processing using soft-computing algorithms and architectures. Our research focuses on orders of magnitude speed advantage and potential enabling technology promise. During neural networks revival stage in the eighties, a compact building block synapse-neuron design with hybrid (digital-analog) architecture was evolved and tested [1]. This design with later improvements became the center-piece of a varied electronic circuitry, providing unprecedented data-processing speed (250ns) and low power consumption. It provided a power-miser data processing circuit, which formed the basis for a multi-layer perceptron (MLP) architecture [2]. A more recent highly parallel neural network design was used to stack a multi-chip module (MCM) (with 64 ICs) in a 3D configuration (sugar cube size) for high-speed (10^{12} operations per second, ops) image processing [3].

Using these soft-computing paradigms, an architecture was evolved, which was very effective in providing a solution for sensor and data fusion in high-speed hardware. While these processors can be built in digital or analog hardware, the massive amount of interconnection lines of a parallel implementation and the power requirements encountered in certain space, military or commercial applications such as hand-held devices make the idea of a hybrid, digital-analog ASIC processor necessary. An example of such an application requiring low power and fast processing of multi-sensor data is that of object discrimination performed onboard a fast frame seeker.

In addition, the present focus for the new research activity is on evolvable hardware (EHW). An appreciable amount of theoretical, algorithmic, modeling, and

analytical work has been done. Further, a series of test chips termed "Field-Programmable Transistor Array (FPTA)" have been designed, fabricated, tested, and used successfully in various digital and analog circuit-evolution experiments [4]. Starting with evolution in simulation, integration has now progressed using FPTAs combined with field programmable gate array (FPGA) for its control that provide high-speed evolution of circuits with potential on-board fault tolerance and function morphing.

A description of these activities along with the motivation of high speed brought on by required applications is described. Selected results of the hardware implementation are presented. EHW work is described not only for the speed advantage but also its enabling promise for future long-life space missions.

2 OBJECT CLASSFICATION/ RECOGNITION

Present deployable object classification, recognition, and tracking applications are either bulky, power hungry, or not capable of providing reliable results in real time. Parallel computer systems capable of giga-operations/s, with a reasonable template size (32x32 or larger) and using <10W of power is still beyond today's computer/processor technology. This prevents the realization of a flexible, robust processor for such applications.

2.1 Test-bed

Our approach, based on 3D packaging of IC chips, is capable of processing large-size images (IR, UV, visible) by sequentially inputting consecutive 64x64 pixel chunks and performing high speed convolutions with 64 prestored 64x64 image templates [5]. The data processing architecture providing an image processing system in a small package performing 10^{12} ops of multiply-sum was designed and executed as a testbed [6]. It incorporates a frame grabber, a highly parallel 3D data convolver engine (performing 64x64 pixel operations with 64x64 templates in <250 ns) (Figure 1), an artificial neural network, and an output processing protocol. It can be coupled to a variety of sensors covering spectral range ultraviolet (UV) through infrared (IR).

2.2 Neural Networks

The ICs use a synapse design based on Multiplying Digital-to-Analog Converter (MDAC) technology using a hybrid approach reported in detail in Refs. [2]. Each circuit is digitally programmable, has an 8-bit resolution digital weight storage, and is a compact analog multiplier with a voltage-input and a current-output configuration [6]. As an MCM of 64 chips, it performs 64X64 complete inner products, each with a 4096 (i.e., 64x64) input array and can accomplish this in 250 nanoseconds (i.e., $\sim 10^{12}$ multiply and add operations in 1 second).

2.3 Functional Description Of Analog Processing

The 64 analog voltage inputs first get converted to currents by a set of V-I converters at the beginning of each row of the 64 x 64 synaptic array. These signals are then current mirrored into all 64 synapses along the row so that all the synapses in a given row receive an identical input signal [2].

At each synapse, a byte, which controls switches to scale current copies of the input, is stored in respective local static memories (SRAM) constituting as weights. By switching in different multiples of the input current and adding them together, the input current is effectively multiplied by the digital weight. Synapses on the same

column have their outputs, Iout, summed by attaching them all to the same wire. These 64 summed signals, one for each column of the array, are then sent directly out

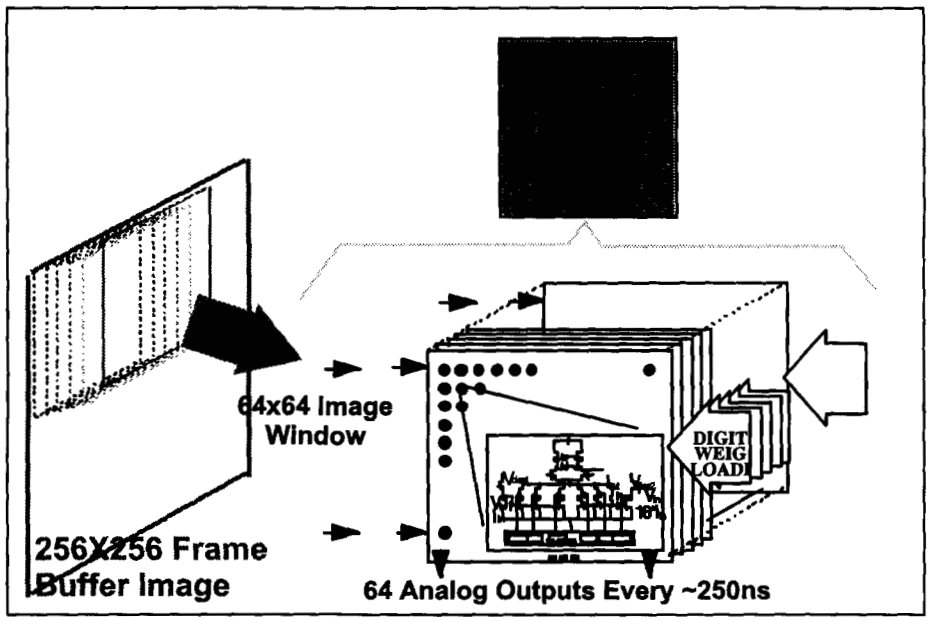

Figure 1. The 3D MCM network consists of 64 ICs each with a 64x64 synapse array based on 8-bit multiplying digital to analog conversion technology and incorporating an image input circuitry to interface with a larger size (e.g. 256x256-pixel) image frame. With parallel processing of inner products, its 64 outputs are fed to a neural net classifier. It can be realized (in-set) in a 10-gm, 3-cm^3 package, with power consumption of ~5 W. A speed of 10^{12} multiply-add operations/s is obtained.

every 250 nanoseconds as shown in Figure 1. In the 3D scheme, respective column outputs of all the 64 chips are added together through edge-wise metallization, so that a 64x64-pixel incoming image multiplies with 64, 64x64 templates on 64 chips to provide 64 analog inner-product current outputs. The process is repeated by sequentially inputting the adjoining 64x64 image windows, "row-by-row" and "column-by-column". Thus a 256x256-image convolution is completed in about 16 milliseconds.

2.4 Digital Weight Programming

Before the processing can begin, the synapse weights as 64x64templates are obtained separately. For this purpose, an approach based on eigenvectors is employed. Since each data point (image) is a 4096-element vector, finding a set of 4096 orthonormal eigenvectors is possible (64 of these can then reside on 3DANN at a time). Selecting the most significant 64 eigenvectors constructed from principal component analysis of target imagery reduces dimensionality of image sets, without losing much of the information relevant for classification. The selected 64 templates are loaded single row at a time. The data for a given row is clocked into a 64 long, 8-

bit wide shift register, one byte at a time. After 64 clock cycles, the data for an entire row of synapses is ready to be loaded into the local memory of each MDAC. A 6-bit row address is supplied and an active-low load signal is asserted, which dumps the data into the synapses on the row specified. The register loads from the bottom up so that the first data loaded corresponds to the first row. More details including its configuration as a recurrent neural network can be found in the literature [2,3].

Incorporating a multi-synapse circuit as an analog multiplier makes the network extremely powerful image-processing engine capable of carrying out in parallel 64 convolutions of the form:

$$\mathbf{c}_i(x,y) = \mathbf{f}(x,y) \otimes \mathbf{g}_i(x,y); \quad i = 1, 2, ..., 64; \qquad (1)$$

where $\mathbf{f}$ is the input image, $\mathbf{g}_i$ is the ith filter mask, and $\mathbf{c}_i$ is the ith output.

2.5 Processing Results

To estimate the speed advantage, an image-processing example was used. As a comparison, the convolution operations on a 256x256 image would only take 16 milliseconds using our 3D MCM. However, the same processing would take as much as 2.5 hours using SPARC-10, and a few minutes with parallel processor such as SHARC. Thus, the hardware provides orders of magnitude speed advantage.

Information about the object (its class, identity, or pose) is processed in a coarse-to-fine manner. For instance, after detecting an object in a frame, a rough estimate of image pose/scale is made, a result that can then be used to limit the variation that needs to be considered during object classification (i.e., plane, helicopter, and missile). Results (Figure 2) using the technique described here have achieved nearly

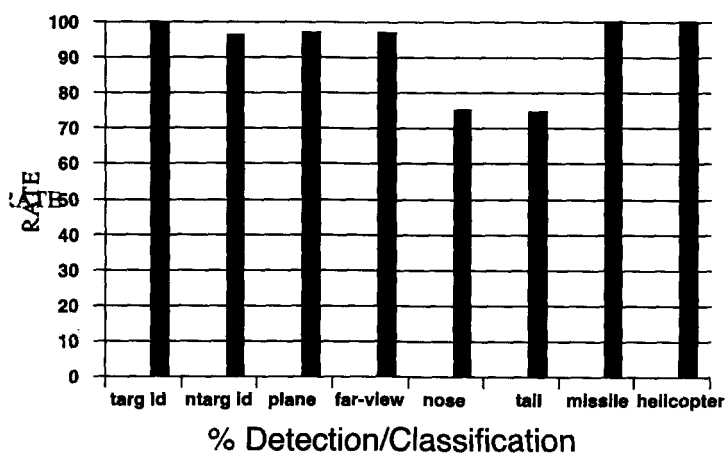

Figure 2. High detection/classification rates are achieved on selected data sets that include all possible orientations and scales of targets.

97% detection rates, 94% classification rates for determining the angle of the principal dimension of an object with respect to the image (±30°), and object classification rates approaching 95%. Results on object/non-object image

classification rates achieved with a helicopter/missile/plane data set were also very encouraging [6].

3. EVOLVABLE HARDWARE

An evolvable hardware system consists of two main components: the reconfigurable hardware (RH) and the reconfiguration mechanism (RM). In previously reported research, the evolutionary processor (EP) that acts as an RM was implemented on a variety of platforms including supercomputer, single PC, DSP, FPGA, and ASIC. The RH was approached as simulated model of unconstrained topology, real FPGA, FPAA model or actual chips, FPTA model or actual chip. For a survey on these different applications the reader is referred to [7,8].

Most simulation approaches have demonstrated that circuit evolution times can vary from a few seconds to hours and days. Therefore, real-world applications will require compact, low-power, autonomous, and stand-alone evolvable hardware. An effort in transitioning from PC-simulated or PC-controlled evolutions to embedded and ultimately to integrated system-on-a-chip evolvable systems is required, because in general, a circuit evolution using simulation models such as SPICE take many seconds to go through one iteration versus a circuit in hardware that takes only fraction of a second.

3.1 A Stand-Alone Board-Level Evolvable System (SABLES)

A board-level solution provides autonomous, fast (1,000 circuit evaluations per second), on-chip circuit reconfiguration. Its main components are a JPL-designed Field Programmable Transistor Array (FPTA) chip as a transistor-level reconfigurable hardware, and a Texas Instruments DSP implementing the evolutionary algorithm as the controller for reconfiguration.

An overview of the components of SABLE is provided below, including the FPTA2 chip and the DSP system. The evolution of a half-wave rectifier circuit is presented to illustrate how the system functions.

3.2 SABLES Components

As mentioned, SABLES integrates an FPTA and a DSP implementing the Evolutionary Platform (EP). The system is stand-alone and is connected to a PC only for the purpose of receiving specifications and communicating back the results of evolution for analysis.

The FPTA is an implementation of an evolution-oriented reconfigurable architecture (EORA) [7]. The lack of evolution-oriented devices, in particular for analog, has been an important stumbling block for researchers attempting evolution in intrinsic mode (with evaluation directly in hardware). Extrinsic evolution (using simulated models) is slow and scales badly when performed accurately e.g. in SPICE), and less accurate models may lead to solutions that behave differently in hardware than in software simulations. The FPTA has transistor level reconfigurability and supports any arrangement of programming bits without danger of damage to the chip (as is the case with some commercial devices).

Three generations of FPTA chips have been built and used in evolutionary experiments. The latest chip, FPTA-2, consists of an 8x8 array of reconfigurable cells. Each cell has a transistor array as well as a set of programmable resources, including programmable resistors and static capacitors. Figure 3 provides a block diagram of the chip architecture on the left and a schematic of the reconfigurable transistor array

cell on the right. The reconfigurable circuitry consists of 14 transistors connected through 44 switches and is able to implement different building blocks for analog processing, such as two- and three-stage operational amplifiers, logarithmic photo detectors, or Gaussian computational circuits. It includes three capacitors, Cm1, Cm2 and Cc, of 100fF, 100fF and 5pF value respectively.

The evolutionary algorithm was implemented in a DSP that directly controlled the FPTA, together forming a board-level evolvable system with fast internal communication ensured by a 32-bit bus operating at 7.5MHz. Details of the EP were presented in [9]. SABLES fits in a box 20.0x20.0x7.6 cm^3.

3.3 An Evolution on SABLES

The following experiments illustrate evolutions on SABLES. The objective of the first experiment is to synthesize a half-wave rectifier circuit. The testing of candidate circuits is made for an excitation input of 2kHz sine wave of amplitude 2V. A computed rectified waveform of this signal is considered as the target. The fitness function rewards those individuals exhibiting behavior closer to target (using a simple sum of differences between the response of a circuit and target) and penalizes those farther from it. After evaluation of 100 individuals, they are sorted according to fitness and a 9% portion (elite percentage) is set aside, the remaining individuals undergoing first crossover (70% rate), either among themselves or with an individual from elite, and then mutation (4% rate). In this experiment only two cells of the FPTA was allocated.

In the second experiment, evolution builds a circuit able to extract two sine waves (independent source signals) from linear combinations of these sources: the evolving circuit is excited with two different linear combinations of the source signals, and the two circuit outputs should respectively restore each source signal. In this signal separation experiment, the GA parameters were: 70% mutation rate; 20% crossover rate; replacement factor of 20%; population of 400; and 100 to 200 generations. Each execution took about 5 minutes in the SABLES system. More than 20 different GA executions were performed. The fitness was the sum of the absolute difference between the FFT of the output signals and the target values over a spectrum between 1kHz and 50kHz. The fitness equation is shown below.

Figure 4 displays snapshots of evolution in progress in the rectifier experiment, illustrating the response of the best individual in the population over a set of generations. Figure 4 a) shows the best individual of the initial population, while the subsequent ones show the best after 5, 50 and 82 generations. The solution, with a fitness value below a minimum for a satisfactory solution is shown on the right.

In the second experiment, evolution builds a circuit able to extract two sine waves (independent source signals) from linear combinations of these sources: the evolving circuit is excited with two different linear combinations of the source signals, and the two circuit outputs should respectively restore each source signal. More than 20 different GA executions were performed. The fitness was the sum of the absolute difference between the FFT of the output signals and the target values over a spectrum between 1kHz and 50kHz.

In the signal separation experiment we selected as source signals two sine waves of frequencies f_1 = 10kHz and f_2 = 20kHz. The set of experiments was performed using 10 cells of the FPTA-2. Figure 5 depicts the inputs and outputs of the best circuit achieved in this experiment. For further details the reader can refer to [10].

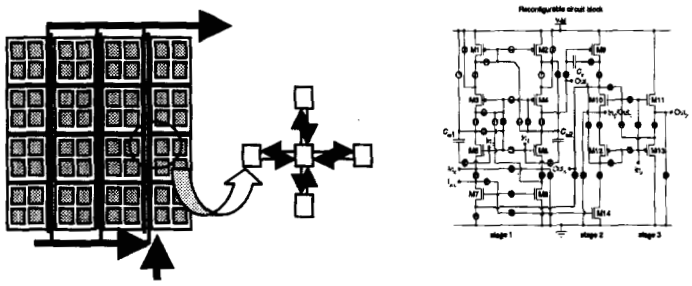

Figure 3. FPTA 2 architecture (left) and schematic of cell transistor array (right). The cell contains additional capacitors and programmable resistors (not shown).

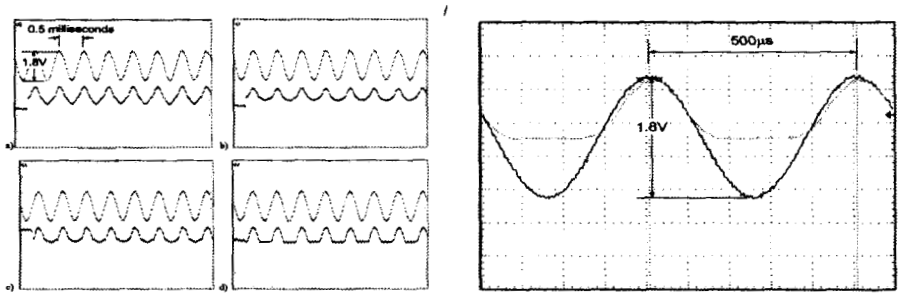

Figure 4. Evolution of a halfwave rectifier showing the response of the best individual of generation a) 1, b) 5, c) 50 and finally the solution at generation d) 82. The final solution is illustrated on the right.

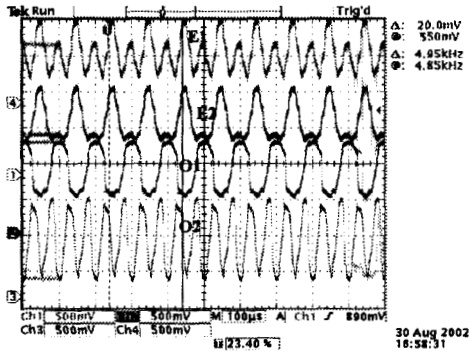

Figure 5. Result of a signal separation experiment. At the top inputs E1, E2, & at the bottom, outputs O1 (10kHz), O2 (20kHz) are shown.

3.4 Results

SABLES achieves about 1-2 orders of magnitude reduction in memory and nearly 4 orders of magnitude speed enhancement compared to systems evolving in simulations, and about 1 order of magnitude reduction in volume and 1 order of magnitude improvement in speed (through improved communication) compared to a PC-controlled system using the same FPTA chip. In the signal separation experiment, each trial takes about 5 to 10 minutes. This means a 1 to 2 orders of magnitude speed enhancement compared to systems evolving in simulations.

3.5 Conclusion

SABLES has shown its performance with an evolution time in seconds for a halfwave rectifier circuit. To date, this is the fastest, most flexible, and most compact stand-alone evolvable system for both analog and digital circuits. Similarly, a 1 to 2 orders of magnitude speed advantage was obtained with SABLES for the signal separation experiment.

ACKNOWLEDGEMENTS

The research described in this paper was performed at the Center for Advanced Avionics, Jet Propulsion Laboratory, California Institute of Technology and was sponsored by the National Aeronautics and Space Administration, the Missile Defense Agency, and the Defense Advanced Research Projects Agency.

REFERENCES

[1] A. Moopenn, T. Duong, and A.P. Thakoor, "Digital-analog hybrid synapse chips for electronic neural networks," Advances In: Neural Information Processing Systems –2 (NIPS-2); Ed: Touretzky 1990, Morgan Kaufmann, pp. 769-776.

[2] T. Daud, T. Duong, M. Tran, H. Langenbacher, and A. Thakoor, "High resolution synaptic weights and hardware-in-the-loop learning," Proc. SPIE Conf. On Nonlinear Image Processing VI: Ed: Dougherty, et al., vol. 2424, 1995, pp. 489-500.

[3] T. Duong, S. Kemeny, M. Tran, T. Daud, and A. Thakoor, "High speed low power analog ASICs for a 3-D neuroprocessor," Proc. SPIE Conf. On Nonlinear Image Processing VI; Ed: Dougherty, et al., vol. 2424, 1995, pp. 470-477.

[4] Stoica, D. Keymeulen, R. Tawel, C. Lazaro and Wei-te Li. "Evolutionary experiments with a fine-grained reconfigurable architecture for analog and digital CMOS circuits." Proc. of the First NASA/DoD Workshop on Evolvable Hardware; Ed: A. Stoica, et al., July 19-21, 1999, IEEE Computer Society Press, pp. 77-84.

[5] T. Duong, S. Kemeny, T. Daud, A. Thakoor, C. Saunders, and J. Carson, "Analog 3-D neuroprocessor for fast frame focal plane image processing," *SIMULATION*, vol. 65, no. 1, pp. 11-24, 1995.

[6] S. Udomkesmalee, A. Thakoor, C. Padgett, T. Daud, W-C. Fang, and S.C. Suddarth, "VIGILANTE: An advanced sensing/processing testbed for ATR applications," Proc. of the SPIE Conf. Automatic Target Recognition VII; Ed: F.A. Sadjadi, vol. 3069, 1997, pp. 82-93.

[7] A. Stoica, R. Zebulum, D. Keymeulen, R. Tawel, T. Daud, and A. Thakoor, Reconfigurable VLSI Architectures for Evolvable Hardware: from Experimental Field Programmable Transistor Arrays to Evolution-Oriented Chips. In IEEE Transactions on VLSI Systems, Special Issue on Reconfigurable and Adaptive VLSI Systems, vol. 9, No. 1,February 2001. (pp.227-232).

[8]A. Stoica, J. Lohn, J, D. Keymeulen (eds). Proceedings of the First NASA/DoDWorkshop on Evolvable Hardware. Los Alamitos, CA 90720, July 1999. IEEE Computer Society Press.

[9] Ferguson, M.I., Stoica A., Zebulum R., Keymeulen D. and Duong, V. "An Evolvable Hardware Platform based on DSP and FPTA". To be published in Proceedings of the Genetic and Evolutionary Computation Conference, July 9-13, 2002, New York, New York.

[10] Zebulum, R.S. , Stoica, A., Keymeulen, D., M.I Ferguson, Duong, V., Vorperian, V., "Automatic Evolution of Tunable Filters Using SABLES", submitted for the 2003 International Conference on Evolvable Systems (ICES) to be held in Norway, March, 2003.