Low Power Analog Neurosynapse Chips for a 3-D "Sugarcube" Neuroprocessor

Tuan Duong, Sabrina Kemeny, Mua Tran, Taher Daud, and Anil Thakoor

Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology,
4800 Oak Grove Dr., Pasadena, CA 91109-8099

David Ludwig, Chris Saunders, and John Carson

Irvine Sensors Corporation
3001 Redhill, Costa Mesa, CA 92626

Abstract --- **Object discrimination and pattern recognition are computationally intensive and for many defense and commercial applications, speed is of the essence. A novel 3-dimensional VLSI architecture in which neural network integrated circuits (ICs) are stacked together and mated to an image sensor may be used to solve** such problems. ***New compact, high speed, low power, analog neuron and synapse circuits, suitable for such 3-D z-plain stacking are reported.*** **The neural circuits have been designed for incorporation into a reconfigurable multilayer perception consisting of 64 inputs, up to 64 hidden units, and up to** *6* **outputs, which can be utilized to solve a variety of pattern recognition problems. The circuits, fabricated in a 1.2 µm CMOS process have achieved 12S ns propagation through a synapse neuron pair, resulting in 4 MHz operation through the envisioned 3 layer feed forward network. Power dissipation at these speeds is expected to be under 30 mW per chip.**

I. Introduction

Pattern recognition is computationally intensive and for many defense and commercial applications, speed is of the essence. Neural networks, implemented in software, have been reported for such applications, but are slow [1]. VLSI realizations of neural networks have been utilized to greatly reduce processing time and are useful in a variety of applications [2,3]. However the size of the VLSI networks are often limited by available silicon area (constrained by increasing cost and decreasing reliability as die size increases). Silicon area can be increased through the use of wafer-scale integration, multichip modules or die stacking. Die stacking is particularly attractive since it is extremely compact. A cube, constructed from many (e.g. 64) thinned die, occupies approximately the same footprint as a standard die. In addition to the tremendous processing power afforded by such a dense IC cube, hybridization of a 3-D IC stack to an image sensor array enables spatially parallel signal processing to be performed on image data at extremely high data rates. In the current collaborative effort between JPL and ISC, an architecture has been conceptualized which combines the spatially parallel 3-D imager cube with neural network processing for the first time, promising tremendous speed and problem size improvements over conventional 2-D VLSI techniques.

A particularly challenging application that requires the tremendous processing capability afforded by such a 3-D neural image processing cube is missile defense which requires spatial-temporal recognition of both point and resolved targets at extremely high speed (milliseconds). A reconfigurable neural network architecture, trained properly, (loaded with appropriate weights), may discriminate targets from clutter or classify targets once resolved. By mating a 64 x 64 image sensor to a stack of 64 neural net ICs, each with different weights, a variety of image processing tasks could be performed in parallel at extremely high speeds and in an extremely small package (≈1 inch cube). JPL and ISC are currently pursuing such a VLSI neural image cube for a missile defense application (a fast frame seeker operating at 1000 frames per second). In order to minimize heat dissipation in the CUM, power most be limited to about 2 Watts for the entire IC stack. This requirement has led to the design of extremely low power analog circuits for implementation of the VLSI neural network ICs. Use of analog circuitry (as opposed to digital) enables very compact, low power neural network realizations [4,5]. In addition, for the 3-D stack mated to an imager array, the spatially parallel input to the neural networks is in analog form. Digital neural processing would require at least one (high speed) or up to 64 (moderate speed) analog to digital converters on each IC, impractical for the low power requirement of the proposed stack. This paper focuses on the analog neural network portion of the 3-D architecture. Test results of the neuron and synapse circuits are presented.