A High Throughput 3-D Inner Product Processor*

Tuan Duong and Taher Daud
Center for Space Microelectronics Technology
Jet Propulsion Laboratory, California Institute of Technology

A particularly challenging image processing application is the real time scene acquisition and object discrimination. It requires spatio-temporal recognition of point and resolved objects at high speeds with parallel processing algorithms. Neural network paradigms provide fine grain **parallelism** and, when implemented in hardware, of few orders of magnitude sped up. However, neural networks implemented on a VLSI chip are planer architectures capable of efficient processing of lineal vector signals rather than 2-D images. Therefore, for processing of images, a 3-D stack of neural-net ICs receiving planar inputs and consuming minimal power are required.

Using analog-digital hybrid techniques, innovative circuit design for convolution operation was developed and chips were fabricated in VLSI for 3-D packaging of stacks, each with 64 chips. Each chip has a 64x64 array of multiply-accumulate (inner-product) processors with low power (<8 microwatts) and high sped (250 nanoseconds) 8-bit storage per cell. Thus, a 64-chip module would consume <2 watts of power, and operating in parallel, would still perform a full convolution operation with 64 patterns at 4 MHz speed (Figure 1).

For such a package, the size of a sugarcube, the challenge of high throughput image input at >128 Gbits/s from any connected sensor or memory block in real-time was met by designing an innovative circuit for the column loading input chip (CLIC) operating at 32 MHz with 64 input channels. This chip would be electrically mated with the 3D stack via 4096 (64x64) indium bumps as shown in Figure 2, and connected to, say, an image grabber memory unit for scanning and inputting contiguous, 64x64 windows of the image, one every 250 nano seconds, to the 3D stack of inner-product processors. On the output side Of the stack, the respective outputs of the 64 chips would be connected together to obtain one output per convolution pattern (total 64 outputs). When this architecture is connected as a system with a point operation processor at the output end of the inner-product processor with suitably connected input and output memory units with a p6 controller, it has a potential of performing object discrimination function at incredibly high speed required of space-age technology for NASA and BMDO applications.

Details of the circuits with chip architectures will be described with need to develop ultra-low-power electronics. Further, use of the architecture in a system for high-speed processing will be illustrated.

---

*Sponsored in part by BMDO and NASA

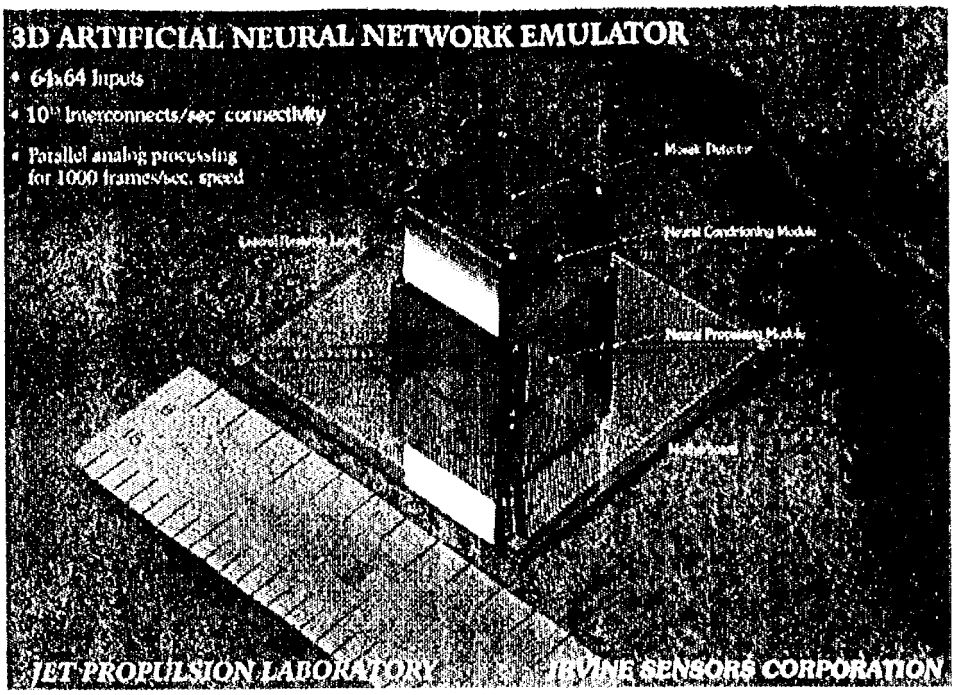

Figure 1. A 3D artificial Neural Network (3DANN) emulator with a 64x64 infrared detector array attached to it for parallel processing

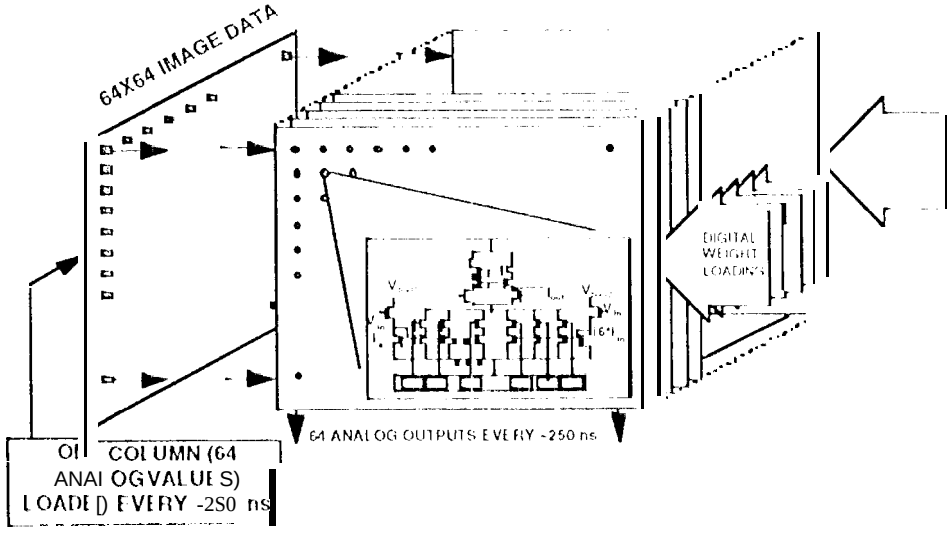

figure 2. 1 he column loading input chip (CLIC) to be mated to a 3D stack of neural processing module (NPM) with 64 chips to perform inner product / convolution operations for a fully parallel high speed object discriminati on/recognition function.