ANALOG VLSI NEURAL NETWORK INTEGRATED CIRCUITS

F.J. Kub, K.K. Moon and E.A. Just
Code 6813
Naval Research Laboratory
Washington, DC 20375
(202) 767-2534
FAX (202) 767-0546

ABSTRACT

Two analog VLSI vector-matrix multiplier integrated circuit chips have been designed, fabricated and partially tested than can perform both vector-matrix and matrix-matrix multiplication operations at high speeds. The 32x32 vector-matrix multiplier chip and the 128x64 vector-matrix multiplier chip have been designed to perform 300 million and 3 billion multiplications per second, respectively.
An additional circuit that has been developed is a continuous-time adaptive learning circuits. The performance achieved thus far for this circuit is an adaptivity of 28dB at 300KHz and 11dB at 15MHz. This circuit has demonstrated greater than 2 orders-of-magnitude higher frequency of operation than any previous adaptive learning circuit.

INTRODUCTION

Analog VLSI vector-matrix multiplier circuits consist of a two dimensional array of multipliers with the matrix of analog weights stored at the multiplier sites, as shown in Fig. 1. The expression for the vector-matrix multiply operation,

$$V_{yi} = \sum_j W_{ij} V_{xj} \qquad (1)$$

where V_{xj} is an input-vector element, W_{ij} is a matrix of weight values, and V_{yi} is an output-vector element. The primary advantages of analog circuits for vector-matrix multiplier operations is that the large two-dimensional matrix of weights (on the order of n^2) are stored at the multiplier sites and do not have to be retrieved from memory as in the digital signal processing case. Only the input vector (on the order of n) has to be retrieved from memory, leading to significant increases in performance.

The vector-matrix multiplier operations is a general function utilized in the vast majority of neural network algorithms and also a large number of conventional signal processing operations. This paper describes fully multiplexed 32x32 and 128x64 programmable analog vector-matrix multiplier circuits. Four-quadrant analog multiplier circuits are used in both chip designs. The weights are X-Y addressed to the multiplier sites and are stored as the difference of analog voltages on two capacitors. (The Naval Research Laboratory holds the basic patent for capacitive weight storage for vector-matrix multiplier circuits [1].) Analog multiplexers are used for the analog input vector, the analog output vector, and the X-Y weight address. A fully differential design has been used throughout the signal path for cancellation of common-mode noise feedthroughs and reduction of offsets. The 32x32 and 128x64 vector-matrix multiplier circuit have been fabricated using a N-well CMOS foundry process. Possible applications of the vector-matrix multiplier circuits are implementing artificial neural network algorithms, implementing large banks of two or three dimensional convolution filters, and performing large area, high speed (1000 frames per second) two dimensional template matching in the Fourier domain.

An additional circuit that has been developed is a continuous-time adaptive learning circuits. The performance achieved thus far for this circuit is an adaptivity of 28dB at 300KHz and 11dB at 15MHz. This circuit has demonstrated over 2 orders-of-magnitude higher frequency of operation than any previous adaptive learning circuit. These circuits can be used as interference cancelers, linear predictors or equalizers. Possible commercial applications areas are smart house wiring over exiting power lines, removal of coherent noise in musical systems, or equalizers for modems or magnetic heads.

32X32 VECTOR-MATRIX MULTIPLIER CHIP

Chip Description

The chip block diagram is shown in Fig. 1. The circuit consists of a two dimensional array of analog multipliers, X-Y address weight decoders, input and output decoders, and current-to-voltage (I-V) converters at the output of each row. In this configuration, the multiplication of a matrix and a vector is performed by capacitively storing analog weights as differential voltages, $\Delta Vw_{ij} = Vw_{ij} - Vwr$, at each multiplier site, and applying a vector of analog inputs as differential voltages, $\Delta V_{xj} = V_{xj} - V_{xr}$, to the column busses. The output currents of the multipliers in a given row are summed on a bus and converted to voltages by the I-V converters to provide the output analog vector elements.

The circuit design approach for fully differential operation is shown in Fig. 1. Both the weight input and weight reference are sampled simultaneously so that feedthroughs from the switches are canceled by the common-mode differencing operation of the four-quadrant analog multipliers. The same procedure is used for the V_x inputs. This technique allows the use of differential D/A converters.

The PMOS Gilbert four-quadrant analog multiplier used in the 32x32 circuit is shown in Fig. 2. The weight values are capacitively stored at the gates of M1 and M2. The difference in current outputs is proportional to the product of two differential voltages, ΔVx and ΔVw. Row and column decoders are used to write the analog voltages to the capacitive weight storage nodes at each of the multiplier sites.

The I-V converters convert the currents on the row busses into voltages. Variable gain circuits are used to optimize the dynamic range. The gain stage is followed by a sample-and-hold circuit that is used to isolate the output from the input so that a new analog vector can be inputted to the vector-matrix multiplier circuit while reading the present output analog vector elements.

Circuit Characteristics

The double-capacitor storage arrangement, shown in Fig. 2, tends to cancel the effects of leakage currents at the capacitor storage sites, thereby significantly improving the weight retention. Measurements of the weight retention shown in Fig. 3 show a factor of 50 improvement for the double- capacitor configuration over the single-capacitor configuration. The bottom curve for the double-capacitor approach shows a 3mV change in the effective weight over a 10mS period at 90C. This change would provide better than 1 percent accuracy for the stored weights. It is necessary that the capacitively stored weights be refreshed. A refresh period of 10ms is reasonable for most applications. Measurements have shown that individual Gilbert multipliers have a total harmonic distortion less than 1.5 percent [2,3].

The cell size for the multiplier shown in Fig. 2 using 2μm design rules is 58μm x 60μm. Figure 4 shows the operation of the fully multiplexed 32x32 vector-matrix multiplier for the case of "1" and "0" weights loaded on alternating rows. An alternating sequence of "1s" and "0s" is observed at the output as expected. A low noise analog board is currently being implemented which will allow further characterization of the circuit's dynamic range.

128X64 VECTOR-MATRIX MULTIPLIER CHIP DESIGN

A schematic of the overall architecture of the 128x64 vector-matrix multiplier circuit is shown in Fig. 5. Four 1-to-32 analog multiplexers are used for inputting the 128 element analog input vector and two 1-to-32 analog mutiplexers are used for outputting the analog vector. The input decoders and weight decoders control transmission switches as shown in Fig. 1. Each of the input and weight decoders is a five-bit random-address decoder with an enable. The five-bit input, output, and weight decoders can address the eight 32x32 multiplier array blocks in parallel when all decoders are enabled. Alternately, if only one of the input, output, or weight decoders is enabled, one individual analog input, one individual analog output, or one individual analog weight can be addressed. The currents from 128 multipliers on each row are summed to produce the analog row output.

The row decoder is used to select 1 of 64 rows to write weights to. It is expected that the input and output multiplexer will operate at approximately 10MHz, thus requiring approximately 3μs to load the input vector. The projected performance for this circuit is approximately 3 billion connections per second.

A feature incorporated into the 128x64 vector-matrix multiplier is a power-saving mode of operation [4] which will likely reduce the power dissipation of large vector-matrix multiplier array by greater than an order-of-magnitude. It can be expected for large vector-matrix multiplier arrays that the power dissipation of each multiplier will be approximately 160μW. Thus, for arrays with 10^4 to 10^5 multipliers, the power dissipation can be in the range of 1.5W to 15W. The power saving mode in the 128x64 vector-matrix multiplier is achieved by turning on the MOSFET triode-mode I-V converter transistors only during the time that the sample-and-hold circuits are turned on. Simulations indicate that the summing bus turn-on time is less than 10ns and the turn-off time is about 30ns. These results indicate that the time period that the triode MOSFETs must be on can be < 100ns. Since the time to load the input vector will be typically 1 to 3μs, this power-saving mode will likely provide greater than an order-of-magnitude reduction in power dissipation.

A photomicrograph of the 128x64 programmable analog vector-matrix multiplier circuit fabricated using 2μm N-well CMOS foundry process is shown in Fig. 6. The chip size is 6.5mm x 6.5mm and the differential-pair multiplier cell size is 31μm x 60μm.

MATRIX-MATRIX MULTIPLY AND HIGH SPEED TEMPLATE MATCHING

The 32x32 and 128x64 circuits described above can also implement a matrix-matrix multiplication operations. The approach is to load one of the matrices into the two-dimensional array of multipliers and then to sequentially input the second matrix a row at a time to the input of the chip. The product of the matrix-matrix multiply operation appears a column at a time at the output of the chip. The multiplication rate performance is the same as that for the vector-matrix multiplier case as long as the first matrix is not reloaded. The performance degrades approximately a factor of two if the first matrix is reload for each matrix-matrix multiply operation. Two of the 128x64 circuits can be used to implement a 128x128 size matrix-matrix multiply operation. The performance rate for the 128x128 size matrix-matrix multiply operation is approximately 6 billion multiplications per second (assuming the matrix stored in the chip is reloaded infrequently).

Two dimensional template matching in the Fourier domain consists simply of a multiplication of a reference template matrix by an image template matrix. Fourier transforms of two dimensional images are readily performed by using conventional digital Fast Fourier Transform (FFT) circuits. Two one dimensional FFT transform operations plus a reformatting operation are necessary to implement a two dimensional Fourier transform. FFT circuits are now available commercially that will implement a 1024 point transform in approximately 100 microseconds. The envisioned operations of the template matcher is that a two dimensional image would be loaded in the vector-matrix multiplier circuit and reference templates would be applied at a high rate to determine the best match. For the 128x128 circuit, the time required to input a 128x128 reference template is approximately 400 microseconds. Thus, greater than 1000 template match comparisons per second are possible. Alternately, the reference template could be loaded in the circuit and two dimensional images applied at a high rate to determine the best match.

CONTINUOUS-TIME ADAPTIVE LEARNING CIRCUITS

A new approach for high frequency adaptive learning circuits using a continuous-time circuit, shown in Fig. 7, to implement the least mean square learning algorithm has been developed. Previous analog adaptive learning circuits have utilized either CCD sampled data circuits or switched-capacitor circuits. Previously, the highest performance achieved using analog circuits was for a four channel switched-capacitor circuit operating at 40KHz [5]. The advantages of the continuous-time approach are the potential for a large number of adaptive taps (>200) and the potential for a high frequency of operation (>50MHz).

The performance achieved thus far is an adaptivity of 28dB at 300KHz and adaptivity of 11dB at 15MHz (Fig. 8) in the linear predictive arrangement. Also, a notch filter with a 10KHz half-width and an isolation of 30dB (Fig. 9) was achieved for an interference canceler arrangement.

This adaptive filter circuit has been fabricated using 2μm CMOS foundry technology. The cell size for the learning circuitry at each tap is 43μm x 1150μm. Thus, an adaptive learning circuit with a large number of taps can be achieved in a typical integrated circuit chip size.

CONCLUSIONS

Fully multiplexed vector-matrix multiplier circuits using capacitive weight storage with a standard N-well CMOS technology have been described. Experimental results were presented for the operation of the Gilbert multiplier, and the 32x32 vector-matrix multiplier circuit. A method of reducing power dissipation by greater than an order-of magnitude was described. A fully multiplexed 128x64 analog vector-matrix multiplier circuit was implemented in an area of 6.5mm x 6.5mm using 2μm CMOS foundry design rules. A method to perform high frame rate (>1000 frames per second), large size (128x128) two-dimensional template matching was described.

An additional circuit type that has been developed is a continuous-time adaptive learning circuit. This circuit has thus far demonstrated high levels of adaptivity and greater than two orders-of-magnitude improvement in the frequency of operation over any previous analog learning circuit. Circuit designs to achieve higher levels of performance are being implemented.

ACKNOWLEDGMENT

The authors would also like to acknowledge the support of the Office of Naval Research and the Office of Naval Technology for this work.

REFERENCES

[1] Patent 4,931,674, F.J. Kub, I.A. Mack, and K.K. Moon, Programmable analog voltage multiplier circuit means, Issue Date: 5 June 1990.

[2] F.J. Kub, K.K. Moon, I.A. Mack and F.M. Long, "Programmable analog vector-matrix multiplier," IEEE Journal of Solid State Circuits, Vol. SC-25, pp. 207-214, February 1990.

[3] K.K. Moon, F.J. Kub and I.A. Mack, "Random address 32x32 programmable analog vector-matrix multiplier for artificial neural networks," Proceedings of the 1990 Custom Integrated Circuit Conference, May 13-16, Boston, Mass., pp. 26.7.1-26.7.4.

[4] F.J. Kub, K.K. Moon and J.A Modolo, "Analog Programmable Chips for Implementing ANNs using Capacitive Weight Storage," To be published in Proceedings of International Joint Conference Neural Networks-91-Seattle, July 8-12, 1991.

[5] J. Fichtel, J.H. Hoticka, and P. Sieber, "An analog adaptive filter in BICMOS Technology," Proc. of ISSCC Conf., San Francisco, 1990.

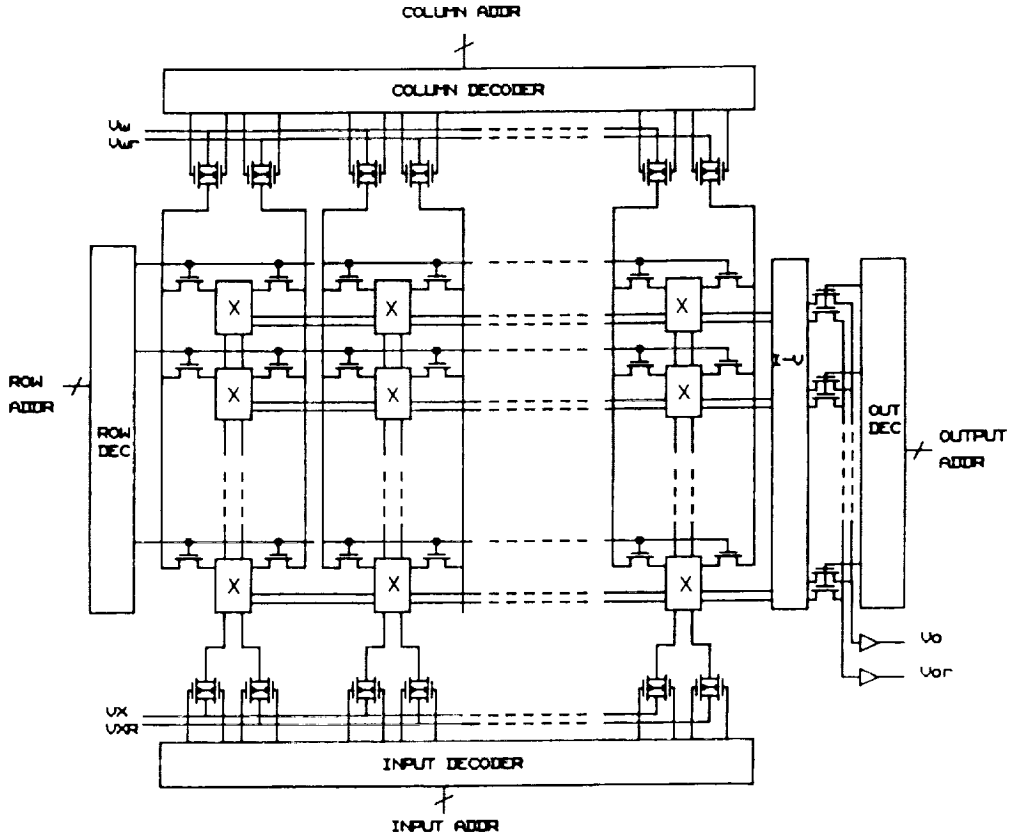


Fig. 1. Vector-matrix multiplier block diagram.

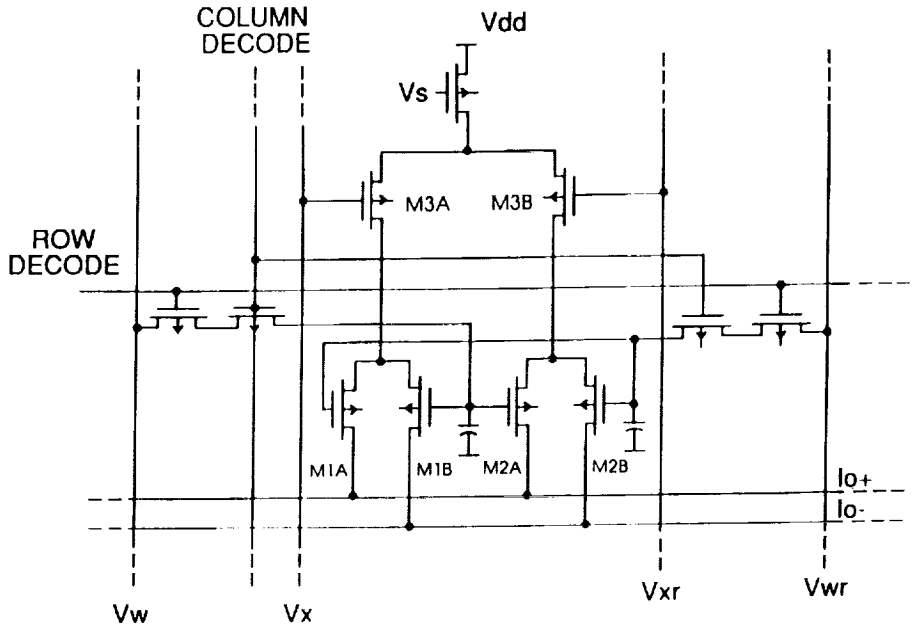


Fig. 2. Gilbert multiplier cell with capacitive weight storage.

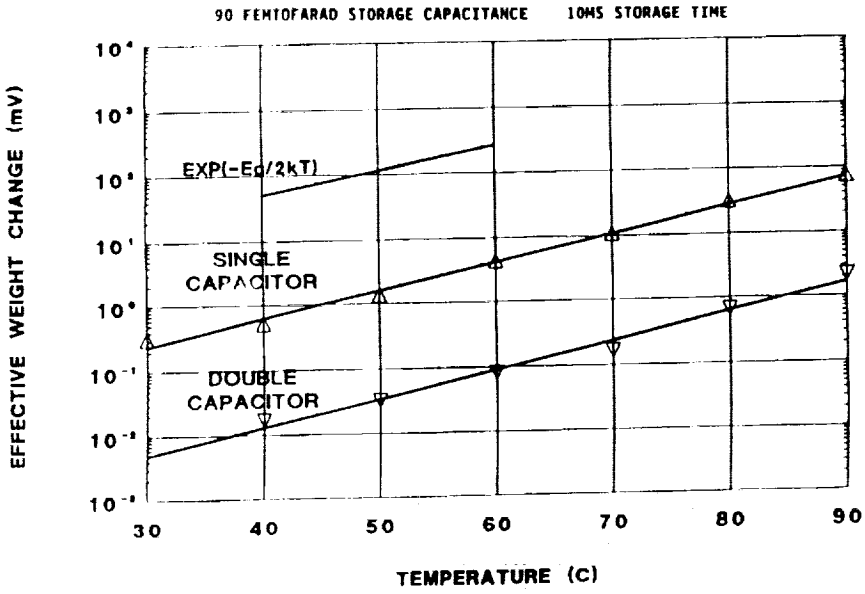


Fig. 3. Weight retention versus temperature.

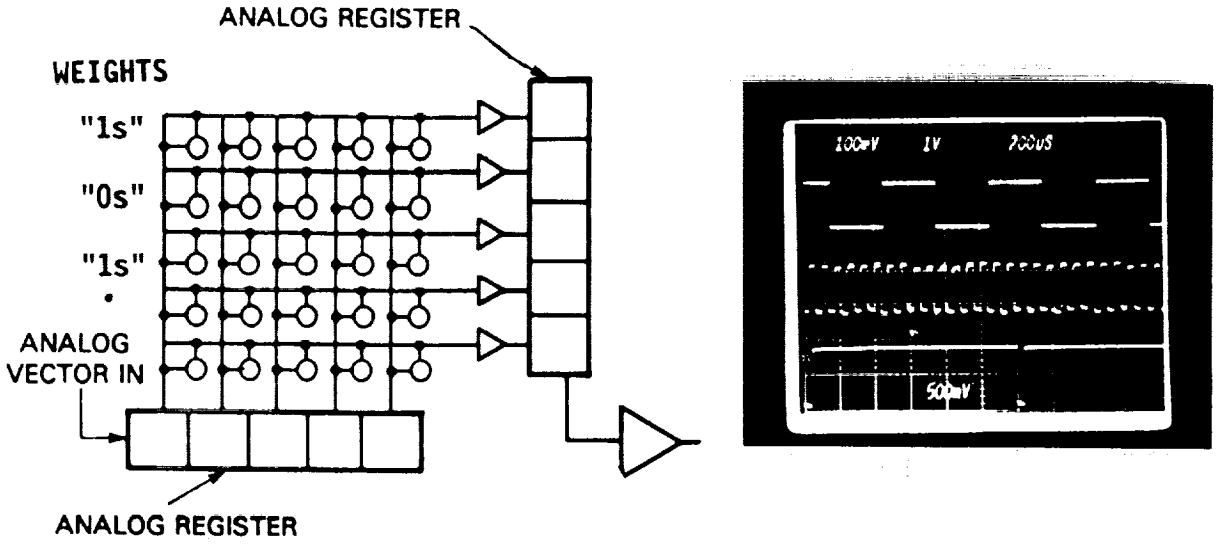


Fig. 4. Output of 32x32 vector-matrix multiplier with "1s" and "0s" loaded on alternating rows.

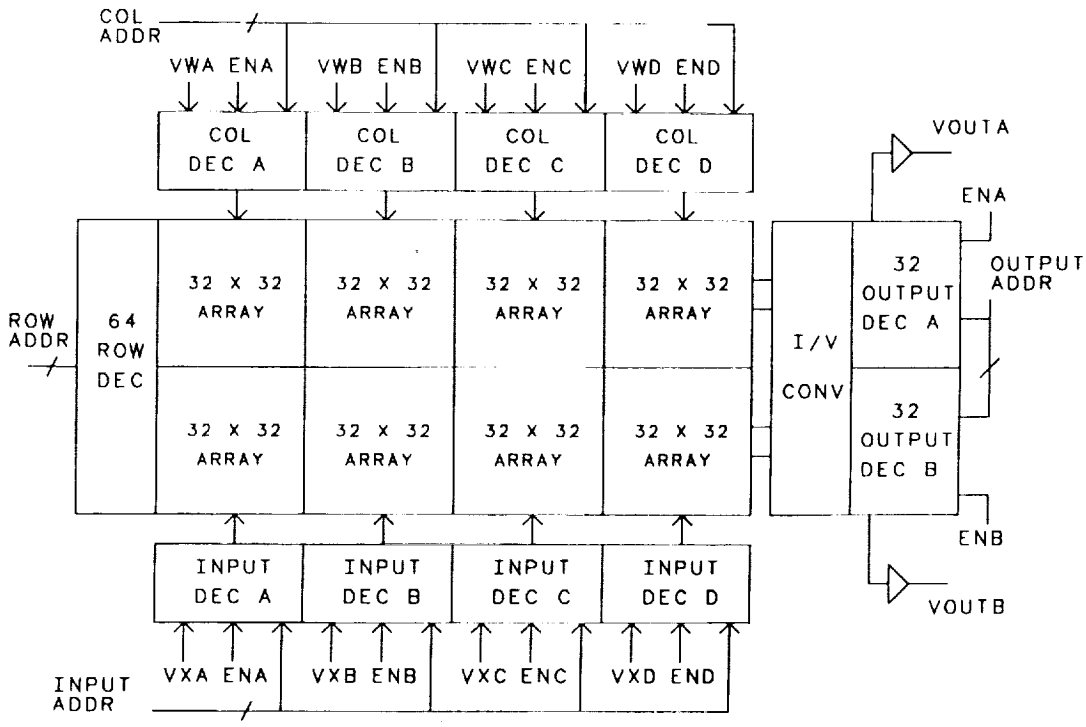


Fig. 5. Architecture of 128x64 vector-matrix multiplier.

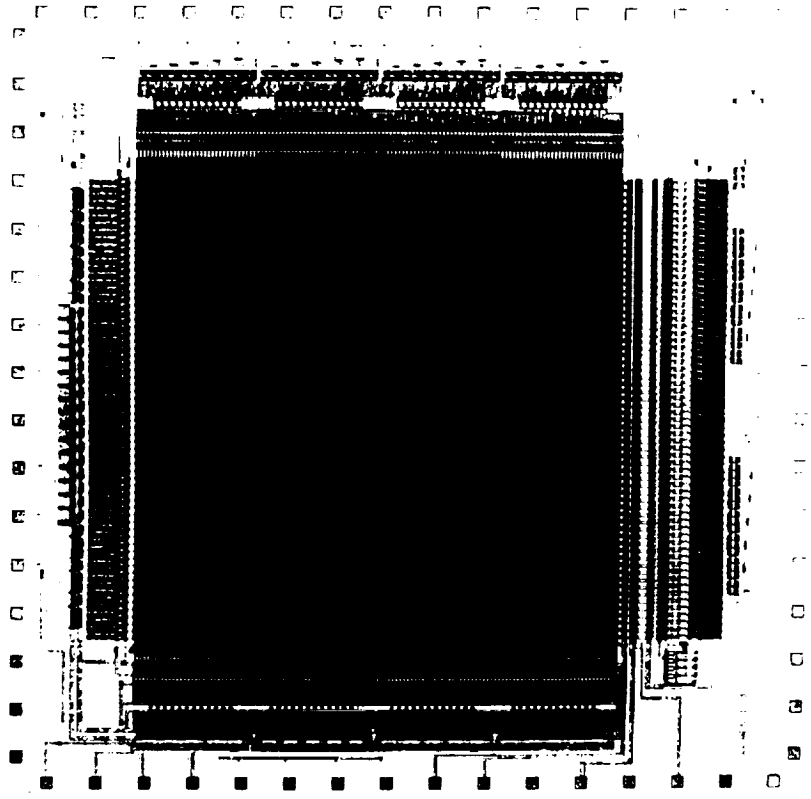


Fig. 6. Photomicrograph of 128x64 vector-matrix multiplier chip.

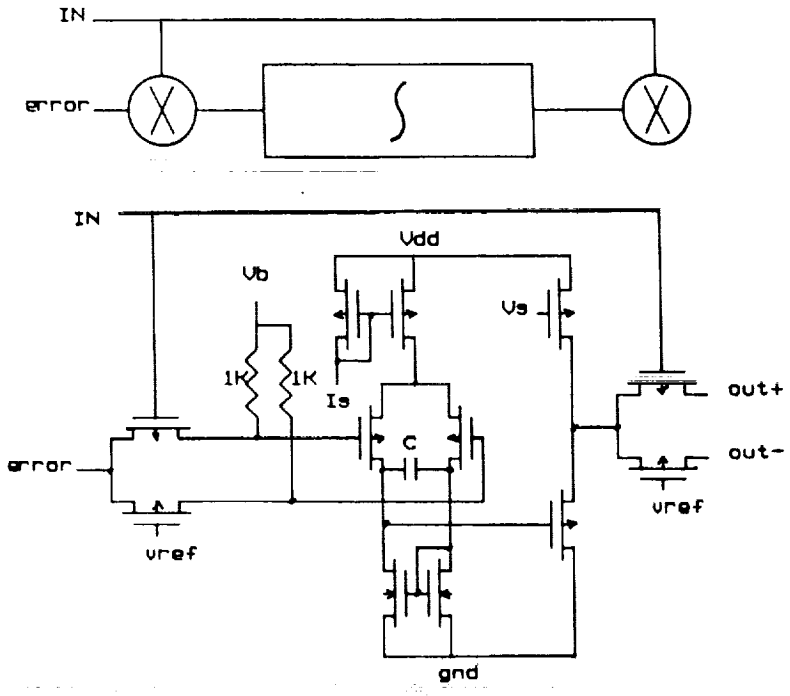
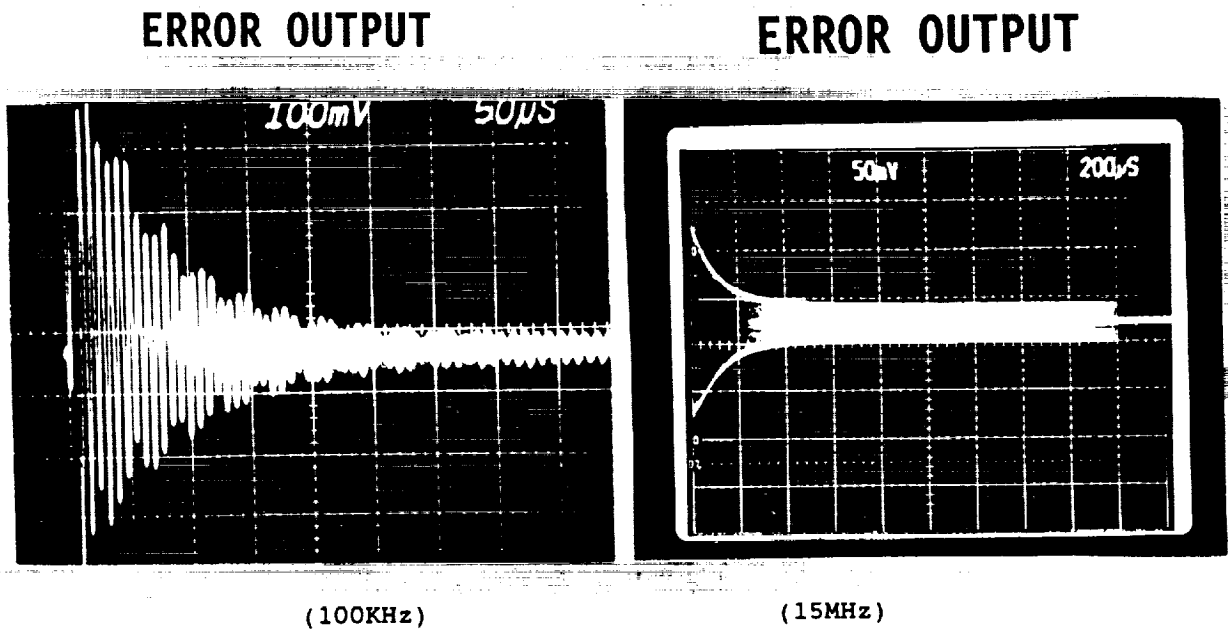
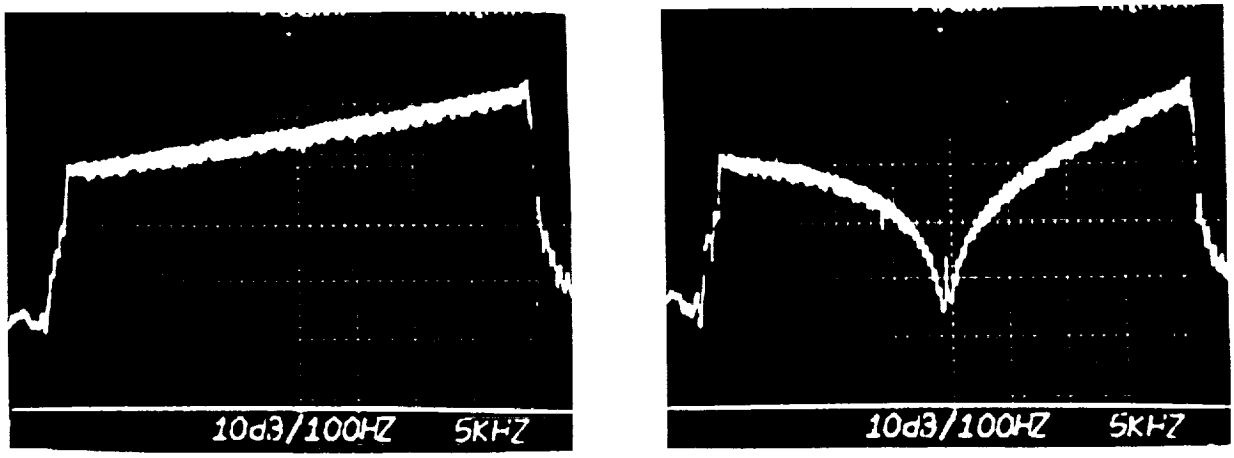


Fig. 7. Continuous-time least mean square weight learning circuit.



(100KHz) (15MHz)

Fig. 8. Experimental results showing error output for linear predictive arrangement.



SIGNAL IN NOTCH FILTER OUTPUT

Fig 9. Experimental results showing adaptive filter operated as a notch filter (80kHz).