A SPACE INTEGRATING ACOUSTO-OPTIC MATRIX–MATRIX MULTIPLIER

Kelvin WAGNER and Demetri PSALTIS
Department of Electrical Engineering, California Institute of Technology, Pasadena, CA 91125, USA

Received 9 August 1984

An optical architecture is described for performing pipelined matrix–matrix multiplications. The architecture is implemented using multiple transducer acousto-optic devices and a wideband photodetector array. A variant of engagement formatting allows multiple inner products to be simultaneously computed by 1-D spatial integration, and through proper pipelining the full product matrix is produced at the output of the detector array. The output matrix in this architecture is in a format that is directly compatible with the input, a feature that can facilitate the implementation of iterative matrix algorithms. Digital multiplication by analog convolution can be incorporated for improved accuracy by using a frequency multiplexed representation of the binary data.

1. Introduction

The optical implementation of matrix operations has received considerable attention recently. Architectures and algorithms have been designed that have increased the speed, accuracy and flexibility of optical matrix processors, extending the potential applicability of such systems to a broader range of problems. Specific advances that have been accomplished in recent years include the initial demonstration of vector–matrix multiplication [1], the introduction of time integrating systolic [2], engagement [3,4], and outer product optical processors [5], a frequency multiplexed processor [6], improvements in accuracy with residue arithmetic [7], the utilization of the method of digital multiplication by analog convolution (DMAC) [8,9] in the above array processors [10,11], and a combination of systolic processing and the DMAC algorithm in a two dimensional implementation utilizing crossed multichannel Brag cells for matrix vector multiplications with digital accuracy [12].

Perhaps the most significant application of numerical optical processors is in $O(N_3)$ problems, i.e. matrix algebra problems that require a minimum of N^3 multiplications and additions, where N is the size of the matrix involved. The solution of a set of linear equations, matrix inversion, and singular value decomposition are examples of such problems [13]. Optical techniques can be applied to such problems by selecting an algorithm that can implement the required operation with successive matrix–matrix multiplications, such as Gauss eliminations, Givens transformations or Housholder reflections [13]. For these algorithms N optical matrix multiplications are required, and since each matrix–matrix multiplication requires N^3 multiplies and adds, optical systems usually solve an N^3 problem with N^4 operations. However, the speed and parallelism of optics can make the optical implementation advantageous, despite this inefficiency. The product matrix that is produced at each iteration during the execution of such an algorithm is used as one of the input matrices for the next iteration. It is therefore important that the format of the output product matrix is directly compatible with the input in order to avoid reformatting and minimize the iteration time. The architecture described in this paper was selected principally because the output can be amplified and applied directly to the input. A space integrating implementation using a parellel output wideband photodetector array is chosen for accuracy and speed considerations. Several candidate data flow optical architectures satisfying these requirements are possible. In this paper we present one such data flow matrix processor which uses crossed multichannel acousto-

0 030-4018/84/$03.00

optic devices (AOD). The operation of the system is based on time and space alignment of vectors which allows the formation of multiple inner product summations via spatial integration in a pipelined fashion. The matrix format is similar to an engagement array, but the data flow is transposed so that the local multiplications needed for each inner product operation form in parallel in space, rather than sequentially in time. Each inner product is summed by a 1-D space integrating condensing cylinder onto an output detector. Many such inner product accumulators are multiplexed in the orthogonal dimension onto separate detectors of a linear array.

2. Optical processor architecture

The proposed optical architecture for matrix–matrix multiplication is shown in fig. 1, along with the appropriate data flow. The principal components of the system are a pulsed laser, two orthogonal multichannel Bragg cells (one with N channels, and the other with $2N - 1$ channels), a linear array of N wideband photodetectors, and lenses. The optical processor is a 2-D array of N^2 analog multipliers configured as an array of N space integrating inner product processors. With the appropriate engagement format of the matrices an array of inner products is formed on the detector array during each processor cycle. As data flows through the Bragg cells the output appears in the same engagement format as the input, which allows direct feedback for iterative operations without latency.

Global system synchronization and sample definition at the output are provided at each time interval T, by the strobing action of a repetitively pulsed laser diode. The pulsed light is collimated and incident on the first multichannel acousto-optic device, AOD1, at the Bragg angle in the x dimension. The elements of an $N \times N$ matrix A are applied to the N transducers of AOD1 in an engagement representation. They propagate continuously along the x direction at a velocity equal to one inter transducer spacing of AOD2 each T s. Rows are represented in individual channels as sequential acoustic pulses separated by T s. The nth row of the matrix A is applied to the nth transducer of AOD1 with a delay of nT s. In this manner the matrix is folded back in time into a sliding parallelogram format we call time engagement. The optical field emerging from AOD1 is spatially filtered in the Fourier plane to remove the undiffracted component and the diffracted field is imaged onto AOD2 at the Bragg angle in y. For clarity the image reversal of the imaging system is ignored. Matrix B propagates in AOD2 in the y direction one channel separation of AOD1

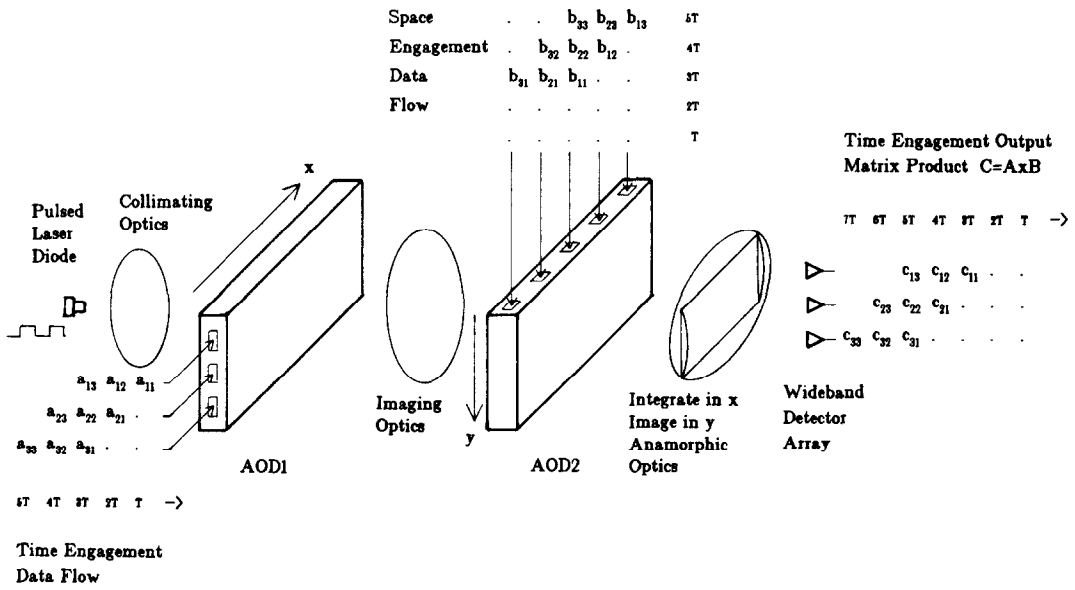

Fig. 1. Schematic representation of the space integrating acousto-optic matrix–matrix multiplier, with the associated engagement data flow. (For simplicity, the details of spatial filtering and the effect of image reversal are neglected in this figure.)

each T s, orthogonally to, and synchronously with, the motion of A. B is delayed by $(N-1)T$ s with respect to A to allow the first row of A to fully enter AOD1. At time $(N-1+k)T$, the kth column of the matrix B is applied to transducers k through $k+N$ of AOD2. In this representation, called space engagement, the matrix is folded over in space to $2N-1$ parallel channels which require N time cycles to be completed. The doubly diffracted light is imaged in y and space integrated in x by the anamorphic lens system following AOD2. The light collected on each photodetector during each cycle is the sum of the product of the elements that are aligned within the corresponding horizontal channels of the two AODs. As we will see in the following paragraph, at the output of the detector array we obtain the product matrix $C = A \times B$ in a time engagement format.

At the Nth time increment the first row of matrix A in AOD1 and the first column of matrix B in AOD2 are spatially aligned in the top channel of the system. The N local products $a_{1i}b_{i1}$ are calculated in parallel by imaging AOD1 through AOD2, and the sum $c_{11} = \Sigma_{i=1}^{N} a_{1i}b_{i1}$ is produced by spatially integrating all these products onto the top detector. One time increment T later the second row of A has fully entered the second channel of AOD1, and simultaneously the first row of A has moved one column away from the transducer. In the orthogonal dimension, the first column of B has moved down in AOD2 one channel to engage the second row of A and produce $c_{21} = \Sigma_{i=1}^{N} a_{2i}b_{i1}$ via space integration onto the second detector. Concurrently the second column of B has been applied to transducers 2 through $N+1$, in order to align with the first row of A and produce $c_{12} = \Sigma_{i=1}^{N} a_{1i}b_{i2}$ on the first detector. In a similar manner successive inner products are aligned, locally multiplied and globally accumulated to the N output detectors. In a total of $2N-1$ time increments T, the output product matrix is produced in a time engagement format identical to the format of the matrix A. Therefore it can be fed directly back to the N transducers of AOD1 with no reformatting or latency. This allows for fully efficient pipelining of iterative matrix algorithms, since after the matrix A is initially loaded into AOD1 there are no more waiting periods required to load new matrices. For instance, when the first element of the output matrix c_{11} is produced the first row of A has been fully entered into OAD1 so we can begin entering the first row of the output matrix in the top channel of AOD1. No interference with subsequent calculations will occur because of the zeroes included in the space engagement formatting of matrix B.

If the Bragg cells are operated in the linear amplitude diffraction regime, and coherent detection is used, then it is possible to make the outputs of the photodetectors appear at the center frequency of AOD1, simplifying direct feedback. The coherent implementation allows complex valued matrices to be represented by the magnitude and phase of the acoustic pulses. If the Bragg cells are operated in the linear intensity diffraction regime (incoherent implementation), then only real positive matrices can be represented, but simpler non-interferometric detection can be employed. In this case the output matrix would not be on a carrier, therefore mixers would be required to upconvert the output before amplifying and applying to AOD1.

3. Frequency multiplexed DMAC

An increase in the accuracy of this processor can be incorporated by the use of the digital multiplication by analog convolution algorithm (DMAC) [8–12], at the expense of additional complexity. It is well known that the multiplication of two time domain waveforms results in the convolution of their Fourier spectra. This can be utilized to implement the DMAC algorithm by simply multiplying the frequency multiplexed binary representations of two numbers.

The product $z = x * y$ of two M bit integers x and y can be expressed as follows:

$$z = \sum_{k=0}^{2(M-1)} z_k 2^k = \left(\sum_{i=0}^{M-1} x_i 2^i\right)\left(\sum_{j=0}^{M-1} y_j 2^j\right)$$

$$= \sum_{k=0}^{2(M-1)} 2^k \left(\sum_{i=0}^{M-1} x_i y_{k-i}\right), \qquad (1)$$

where x_i, y_i are the bits in the binary representation of the integers x and y and the coefficients $z_k = \Sigma_{k=0}^{M-1} x_i y_{k-i}$ can achieve M discrete levels. A time domain representation of a frequency multiplexed binary word is given by $f(t) = \Sigma_{n=0}^{M-1} x_n \exp(jnwt)$. The product of two such waveforms is

$$h(t) = f(t)\, g(t)$$

$$= \left(\sum_{n=0}^{M-1} x_n \exp(jnwt) \right) \left(\sum_{m=0}^{M-1} y_m \exp(jmwt) \right)$$

$$= \sum_{k=0}^{2(M-1)} \left(\sum_{n=0}^{M-1} x_n y_{k-n} \right) \exp(jkwt)$$

$$= \sum_{k=0}^{2(M-1)} z_k \exp(jkwt). \qquad (2)$$

Thus the weights of the frequency multiplexed product waveform correspond to an M level digitally weighted representation of the product of the two binary words. This pseudo binary representation can be channelized into $2M - 1$ frequency channels, each centered at kw. The amplitude of each spectral component can be quantized to one of M levels by an A/D converter, and true binary representation can be obtained with a digital shift and add register.

Digital multiplication by frequency convolution can be incorporated in the matrix multiplier of fig. 1 in order to improve the accuracy over that attainable with analog data representation. This will increase the required time bandwidth product of the AODs by a factor of at least the number of bits. The duration of the optical and acoustic pulses must be at least $2\pi/w$ s, to permit channelization at the detector output. The frequency multiplexed binary weighted data must be encoded in phase within each acoustic pulse so that all the frequency components add constructively. When interferometric detection is used, the RF output from each detector will be the coherent sum of N frequency multiplexed multilevel binary weighted signals, occupying up to twice the original bandwidth. This format is compatible with direct feedback to AOD1 without redigitization in each cycle. Further iterations would increase the required dynamic range of each frequency component and increase the number of nonzero frequencies. By examining the Fourier plane of AOD1 we can determine globally the number of frequencies occupied for matrix A. If the available number of frequency bins is exceeded by p excess Fourier components, then we can perform a global pseudo floating point rescaling of the product matrix by increasing the local oscillator frequency used for heterodyne detection by pw. The lower order p bits are then discarded by highpass filtering the detector outputs, or with a Fourier plane aperture. Redigitization is required if a detectors dynamic range is exceeded or when the iterative matrix operation has converged, and the resultant matrix must be output in binary form.

The major difficulty with incorporating DMAC in the matrix–matrix multiplier is the complexity of the frequency multiplexed encoding and decoding, and the large number of A/D converters required to convert back to true binary form. Each of the N detectors must be channelized into $2M$ frequency bands, and each of these must be digitized every T s to an accuracy of NM. This could require as many as $2NM$ A/Ds, which is probably impractical. Alternatively, N faster A/Ds could be multiplexed between $2M$ frequency channels each by performing a conversion each $T/2M$ s. To decrease the number of A/Ds even further it may be possible to digitize only one detector output at a time, and recycle the matrix for further conversions by multiplying by the identity matrix I.

4. Discussion

Acousto-optic devices are attractive transducers for data flow optical processors because of their sliding window nature, wide bandwidth (>1GHz), large number of resolvable spots (TB>1000) and wide dynamic range (>60dB). When dealing with multichannel Bragg cells, however, the constraints imposed by acoustic diffraction and electrical crosstalk limit all aspects of device performance. Today a practical limit on the number of channels is on the order of 100 or less. Eventually larger arrays may be realized through the use of anisotropic self collimation and effective RF isolation techniques. The properties of the multichannel acousto-optic devices determine the processing power of the matrix multiplier and to a lesser degree the accuracy obtainable. In order to perform an $N \times N$ analog matrix multiplication this architecture would require an N channel AOD with TB = $2(2N - 1)$ (AOD1), and a $2N - 1$ channel AOD with TB = $2N$ (AOD2), where a factor of 2 was included for intra pulse dead time. If frequency multiplexed binary encoding is used with M bits, then AOD1 requires a TB > $M(2N - 1)$, and AOD2 requires a TB > NM.

The detector array is composed of N parallel wideband photo-detectors whose outputs are bandpass filtered, combined with a steering matrix, amplified, and fed back to AOD1. Single photodetectors can have a wide dynamic range (>50dB), but large monolithically

fabricated arrays are limited by crosstalk. To obtain the full dynamic range capabilities of the detectors, a powerful optical source must be employed, and optical losses minimized.

To obtain an estimate of the processing power of the space integrating optical matrix multiplier described in this letter, let us consider a target system for multiplying 32 X 32 matrices. AOD1 requires 32 channels and a TB > 128, OAD2 requires 63 channels and a TB > 64. If we assume a bandwidth of 128 MHz, then the full matrix product could be obtained in 1 μs, yielding an analog processing rate of 3.2×10^{10} multiplications per second. At these rates the matrix could be inverted in as little as 32 μs with a direct algorithm, or a fraction of a millisecond with an iterative algorithm. If we desire digital accuracy of 8 bits input and 16 bits output, then the system parameters become much more stringent. AOD1 requires a TB > 1024, and AOD2 requires a TB > 512, which could be accomplished with a bandwidth of 100 MHz, and a matrix multiplication time of approximately 10 μs. This yields a processing rate of 3×10^9 DMAC multiplications per second. However, to obtain this additional accuracy we require an array of 32 frequency channelizers with 16 frequency bins each, and 32 8-bit A/D operating at 100 MHz, temporally multiplexed between the frequency channels. When the complexity of the electronic peripherals to the optical processor reaches this high level, it is important to keep in perspective what would be required for a fully digital implementation. For instance, for the same numbers as above a digital array of 32 8-bit multipliers/accumulators each operating at 100 MHz can have the same processing power (3.2×10^9 operations/s). In other words, comparable electronic hardware is needed for both the digital and the DMAC optical implementations. The advantage of the optical implementation is in the reduced communication rate (320×10^6 samples/s versus 3.2×10^9 samples/s in this example) between the high speed array processor and the system buffer memory. This advantage derives from the 2-D parallelism of optics which results in 1024 parallel multiplications being performed each 100 ns interval.

5. Conclusion

A highly parallel, pipelined, space integrating, acousto-optic processor for iterative matrix–matrix multiplication has been described. The architecture avoids the serial readout bottleneck and dynamic range limitations of CCD arrays used in time integrating architectures. The wideband nature of the input and output transducers can result in an analog processing rate exceeding 30 GOPs (Billion multiplies per second), and in excess of 3 GOPs for the DMAC implementation. Additional accuracy can be incorporated by the use of the DMAC algorithm and frequency multiplexing, but the improved accuracy is accompanied by an increase in complexity and expense. The analog optical processor can be implemented with currently available devices and simple electronic support circuitry. It provides extremely high processing power with reasonably good accuracy (equivalent to 8–10 bits) due to the high dynamic range that is achievable with non integrating photodiode arrays.

Acknowledgements

The research reported here is supported by the Air Force Office of Scientific Research and the Rome Air Development Center. K. Wagner is the recipient of an Army Research Office graduate fellowship. Illuminating discussions with T. Weverka and E. Miles are gratefully acknowledged.

References

[1] J.W. Goodman, A.R. Dias and L.M. Woody, Optics Lett. 2 (1978) 1.
[2] H.J. Caulfield, W.T. Rhodes, M.J. Foster and S. Horvitz, Optics Comm. 40 (1982) 86.
[3] P.S. Guilfoyle, Proc. SPIE 352 (1982) 2.
[4] R.P. Bocker, Appl. Optics 22 (1983) 804.
[5] R.A. Athale and W.C. Collins, Appl. Optics 21 (1982) 2089.
[6] D. Casasent, J. Jackson and C. Neuman, Appl. Optics 22 (1983) 115.
[7] J. Jackson and D. Casasent, Appl. Optics 22 (1983) 2817.
[8] H.J. Whitehouse and J.M. Speiser, in: Aspects of signal processing, pt 2, ed. G. Tacconi (Proc. NATO Advanced Study Institute) Boston (1976) pp. 669–702.
[9] D. Psaltis, D. Casasent, D. Neff and M. Carlotto, Proc. SPIE 232 (1980) 151.
[10] R.P. Bocker, S.R. Clayton and K. Bromley, Appl. Optics 22 (1983) 2019.
[11] R.A. Athale, W.C. Collins and P.D. Stilwell, Appl. Optics 22 (1983) 368.
[12] P.S. Guilfoyle, Opt. Eng. 23 (1984) 20.
[13] G.H. Golub and C.F. VanLoan, Matrix computations (John Hopkins University Press, Baltimore, 1983).