MULTI-OPERAND ASSOCIATIVE ARITHMETIC

Isaac Soherson and Sagi Ruhman
The Weizmann Institute of Science
Department of Applied Mathematics
Rehovot, Israel

ABSTRACT

Multi-operand associative techniques attain their full power in algorithms where the data may be recast into disjoint data sets, all acted upon concurrently, each by a different operand common to the set. But the multi-operand approach can also serve to enhance arithmetic operations significantly. The speed-up of associative multiplication by handling a number of multiplier bits at a time is described and analyzed, including an effective algorithm for a limited sum of products. The most complex process treated is convolution, which serves to illustrate the enhancement of an extended sum of products. Any number of vectors stored in memory can be convolved simultaneously by a common filter vector. Execution time is 45 milliseconds for 1024 element data and filter vectors, 2048 element results, and 16-bit precision.

INTRODUCTION

Parallel processing with associative memory has been the subject of considerable interest and much investigation for over twenty years [1]. However, fully parallel associative memory never developed as a standard component of high density. A number of efforts to develop a compact associative cell were reported [2,3], some resulting in an experimental chip, but most lacked some essential associative function, and none developed commercially. The only chip to reach the market was a 8x4 array (INTEL 3104). On the other hand, a very clever idea was conceived of emulating associative operation with conventional semiconductor memory [4], and was implemented in STEAMAN, the only associative processor to appear commercially. But emulated memory can access only a bit-slice at a time, operate on it off-chip, then return the resulting bit-slice. This leads to a direct loss in speed, due both to off-chip processing and inability to perform multi-bit compare and write operations. Furthermore, emulation limits the scale of integration to n bits, where n is the required word-length. In 1980 the authors proposed a partitioned associative architecture [5] in which many operands act concurrently, each on its own disjoint data set. This multi-operand approach requires fully parallel associative memory to take advantage of the added concurrency which is virtually lost in emulated associative memory. Hence a renewed effort was made to realize fully parallel associative memory as a high density integrated component. The results, which are summarized below, indicate that a 16K chip with 50 nanoseconds cycle time can be achieved using current MOS technology. The multi-operand approach has been successfully applied to such problems as tomographic back-projection [5,8] which may be recast into disjoint data sets (pixels 'belonging' to a given ray), each acted upon by a different operand (ray attenuation) common to the set. Our main topic here is the application of multi-operand techniques to enhance arithmetic processes such as summed multiplica-

BASIC ASSOCIATIVE MEMORY

Figure 1

VL$IA$ ASSOCIATIVE MEMORY

An associative memory is a device in which stored data words are identified according to their contents, hence it is sometimes called a content-addressable memory. Such a memory is to be distinguished from the more widely used coordinate addressable memories, such as Random Access Memory (RAM), in which stored data are accessed by their location or address. An associative
memory accepts as inputs a comparand (or operand) word and a mask word, searches all stored data locations simultaneously for a match between the unmasked bits of the comparand and the corresponding bits in all the stored words, and identifies matching data words by setting a marker or tag in a TAG register. Our basic model of an associative memory is shown in Figure 1 and its primitive operations given below.

For all \( j = 0, 1, \ldots, J - 1 \) and all \( k = 0, 1, \ldots, K - 1 \)

1. **SETAG**  \( t_j := 1 \)
2. **SHIFTAG**  \( t_j := t_{j-1} \)
3. **LOAD c**  \( c_k := 0, c_k := 1 , \text{ or } c_k := i_k \)
4. **LOAD m**  \( m_k := 0, m_k := 1 , \text{ or } m_k := i_k \)
5. **COMPARE**  \( t_j := t_j \sum_k m_k (a_{jk} - c_k) \)
6. **WRITE**  \( a_{jk} := t_j a_{jk} + t_j (a_{jk} + m_k a_{jk}) \)
7. **READ**  \( c_k := \sum_j t_j a_{jk} \)

The symbols \( + , - \) and \( \sum \) stand for OR, Exclusive-OR, and OR expansion respectively. \( \overline{m} \) denotes the complement of \( m \). Both WRITE and COMPARE operate just on masked ON bits \( (m_k=1) \) of words (rows) that are tagged \( (t_j = 1) \). Hence, the SETAG operation is required initially to mark all words eligible for processing. Up to four operations may be done concurrently during a given memory cycle: SETAG or SHIFTAG, LOAD c, LOAD m, and COMPARE, WRITE or READ. Clearly, any subset of the above may also be concurrent, and when registers \( c \) and \( m \) are loaded from the input bus simultaneously, they must receive the same data.

To implement these functions, a 12-transistor, 5-bus, static MOS associative memory cell was derived, and was laid out in a 6 micron silicon-gate technology, yielding a cell area of 79 x 123 microns. Allowing for peripheral registers and logic, a 32 x 32 array will fit on a 4x6 millimeter chip. To gauge the performance of such a chip, a 2x2 cell array with its bus drivers were simulated at circuit level using SPICE 2 [5]. The buses were represented by lumped capacitances corresponding to a 2x2 array. The results obtained were a maximum cycle time of 100 nanoseconds and a maximum dissipation of 50 milliwatts when operating continuously at the full rate. Scaling down the cell to a line width of 1.5 microns indicates a capacity of 16K bits per chip and a cycle time of 50 nanoseconds at the same power dissipation. It should be noted that only a major operation, COMPARE, WRITE, or READ, requires the full cycle time. A half-cycle will suffice for a minor operation, or any legal combination thereof.

**PARTITIONED ASSOCIATIVE ARCHITECTURE**

The distinctive associative COMPARE operation is limited to a single comparand. Iagyby [7] proposed simultaneous bit-by-bit comparison against a number of comparands, \( F \), by providing \( F \) flags per word and denoting agreement with any comparand by setting a ONE in the corresponding flag. When examining a single bit-slice of the comparands, they all fall into two groups, those having a ONE in that bit, and those having a ZERO, hence it suffices to deal with each group separately, regardless of the number of comparands involved. Referring to Figure 2, main memory \( A \) contains the data to be processed and an \( F \)-bit flag field. Although the comparands could be stored in a set of serial registers, they are shown residing in an auxiliary associative memory, \( A' \). The flag field is initially set to ONE throughout \( A \), and the data is processed bit serially. For any \( k \in \{0, 1, \ldots, K - 1\} \) we select the data cells in \( A \) having a ONE in bit position \( k \), and write ZERO in all flags corresponding to mismatching comparands — those having a ZERO in this bit position. We do the same for data cells with a ZERO, and comparands having a ONE in bit position \( k \). After repeating this process for each of the bits to be compared, the operation is complete.
<table>
<thead>
<tr>
<th>STEP</th>
<th>MEMORY A</th>
<th>MEMORY A'</th>
<th>CONTROL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$m_0 = d(N, N+1, \ldots, N+W-1)$; $m'_0 = d(0)$; $m_1 = d(0)$; $\text{CNT} = 0$</td>
<td>$\text{SETAG; WRITE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>1</td>
<td>$s_0 = d(W, k, 0)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>2</td>
<td>$s_0 = d(W, k, 0)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>3</td>
<td>$s_0 = d(W, k, 0)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>4</td>
<td>$s_0 = d(W, k, 0)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
</tbody>
</table>

\[ a_{k-1} \quad \text{MS} \quad \text{LS} \]
\[ a_{N+W-1} \quad a_N \quad a_0 \]

F-bit flag N-bit data

Main memory format

and every cell is labeled with the comparator it matches. Cells not matching any comparator will have an all-ZERO flag-field. Clearly, if the comparands are distinct, no cell can be labeled in more than one flag position.

This 'many-to-many' algorithm, in which 'many' comparands are concurrently matched against 'many' data words, is detailed in Table 1. In describing associative algorithms it is convenient to define two binary vector operators which can only appear in assignment statements and assume the length of the vector being assigned. The operator is akin to Kemeny's delta,

\[
d(1, i_1, \ldots, i_n) = d_j = 1 \quad \forall j = 1, \ldots, n
\]

\[
d_j = 0 \quad \forall j = 1, \ldots, n
\]

where the i's are integral constants, expressions, or symbols. The other is a shift operator,

\[ s(t', N, B) \Rightarrow \text{vector } t' \]

The algorithm executes in 4 cycles per bit and offers little advantage in emulated associative memory where writing is bit-serial. For our parallel memory, and a data field-length of N, an appreciable advantage may be gained when PXN. Unfortunately, efforts to extend the many-to-many idea to other useful algorithms were not successful.

<table>
<thead>
<tr>
<th>STEP</th>
<th>MEMORY A</th>
<th>MEMORY A'</th>
<th>CONTROL</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>$m_0 = d(N)$; $m'_0 = d(0)$; $m_1 = d(0)$; $\text{CNT} = 0$</td>
<td>$\text{SETAG; WRITE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>1</td>
<td>$s_0 = d(N)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>2</td>
<td>$s_0 = d(N)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>3</td>
<td>$s_0 = d(N)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
<tr>
<td>4</td>
<td>$s_0 = d(N)$; WRITE</td>
<td>$\text{SETAG; COMPARE}$</td>
<td>$\text{SETAG; COMPARE}$</td>
</tr>
</tbody>
</table>

\[ a_{k-1} \quad a_{k-2} \quad \text{MS} \quad \text{LS} \]
\[ a_{N+W-1} \quad a_N \quad a_0 \quad a_{N-1} \quad a_0 \]

F-bit flag Cond. Carry N-bit number mark

Main memory format

The authors introduced 'multi-operand' algorithms operating in the 'partitioned' architecture shown in Figure 2. Main memory A is dynamically partitioned into F disjoint subsets by an F-bit flag field, thus enabling F operands, stored in memory A', to act concurrently, each on its own distinct subset of data in A. To illustrate this idea, consider multi-operand addition. Memory A holds the F addends, and its tag output is routed to the flag-field of the mask register in memory A, permitting the flag-field to be masked with a bitwise slice of the addends (or their complement). Each one of the F data sets in memory A is distinguished by setting the flag-bit corresponding to its added. Elements in memory A not belonging to any one of the F data sets, and not to be involved in the operation, are distinguished by a ONE in the candidate mark. The algorithm is...
carried out sequentially by bit and is detailed in Table 2. The truth table for serial addition into an accumulator, shown in Table 3, lists four cases requiring action, and gives a valid order for their execution. Word-format in main memory assigns bit N to store the sequential carry, which is initially set to ZERO (step 0). Steps 1 to 5 operate on data sets belonging to operands whose current bit is ZERO. The current bit of operand memory A is obtained by executing a COMPARE to ONE (step 0 or step 10), and is applied as a mask to the flag field of main memory A (step 2). A COMPARE against ZERO in A (step 2 and step 4) will then select only cells whose flags correspond to operands with a ZERO in the current bit position. Similarly, steps 6 to 10 operate on data sets belonging to operands whose current bit is ONE. After repeating steps 1 to 10 for each operand, the process is complete and each operand has been added to its corresponding data set. It should be noted that this algorithm requires a multi-bit compare to achieve any speed-up, but also takes advantage of a multi-bit write. With the fully parallel associative memory assumed here, execution time is 9 cycles per bit.

Multi-operand subtraction in two's complement notation, with the "difference" replacing the "minuend", can be implemented in a similar manner and executes in the same time. Hence multi-operand multiplication and division follow directly by successive application of addition and subtraction. Consider multi-operand multiply, with main memory partitioned into F disjoint sets of multipliers, and F multiplicands stored in operand memory. The word format in main memory is as shown below.

<table>
<thead>
<tr>
<th>F</th>
<th>1</th>
<th>1</th>
<th>N</th>
<th>M\times N</th>
</tr>
</thead>
</table>

Part, flag, Candi. Carry, Multiplier, Product
mark

where each N-bit multiplier operates on the M-bit multiplier belonging to its set, and the full (M\times N) bit product is formed in a field adjacent to the multiplier. The algorithm consists of N successive multi-operand additions, one for each bit of the multiplier, starting with the least significant; addition of the appropriate multiplicant is conditional upon the candidate mark and current multiplier bit. The multiplier bit and the starting bit position for addition are incremented at each iteration. Execution time in memory cycles will be N(M\times 2.5)\times M\times N, where the added 2.5N cycles are due to one place carry propagation after each addition. At a modest cost in program complexity, word-length in main memory can be reduced by 1 bits if the multiplier may be overwritten as the product is formed. Multi-operand multiplication can be viewed as a vector-scalar product for a number of vectors and scalars simultaneously. It should be noted that the vition in which each multiple is placed may be of different length, and their elements may be arranged in any order and intermingled in any fashion.

In computations requiring a single vector-scalar multiplication, this simple successive addition method will run about twice as fast as the multi-vector case, because the addition is now conventional rather than multi-operand. Execution time is now N(M+2)/2\times M\times N. Nevertheless, it is possible to speed up single vector-scalar multiplication significantly through the multi-operand approach by using it to handle several bits of the multiplier at each iteration. Let b denote the number of bits handled at a time, then auxiliary memory is loaded with the binary codes 0 to 2^{b-1}, alongside the corresponding multiples of the common multiplier. The number of iterations performed is now N/b, and each iteration starts with a many-to-many comparison to partition main memory into 2^b disjoint sets according to the b-bit multiplier code being handled. This is followed by a multi-operand addition in which each multiple from auxiliary memory is added into the product field of the corresponding set of multipliers. The starting bit position for addition is incremented by b at each iteration. Execution time now becomes,

(N/b)(2^{b-2}) + 2N(M+2)/2 = N(M+1)/b + 2N/b \approx N(M+1)/b + 2N/b/2

where the first expression in parenthesis accounts for many-to-many comparison. Multiplication time is plotted in Figure 3 for M=50 and b ranging from 1 to 10. The value plotted for b=1 is that for conventional associative multiplication by a constant, N(M+1)/b. The word-length was chosen to be the least common multiple of 2, 3, 5, 6. As was to be expected, multi-operand enhancement of multiplication starts paying off at b=3.

The speed-up factor here approaches b/2(\log_2(F/2)), whereas an algorithm directly fitted to the multi-operand model would yield a factor of F/2.

An interesting variation of multi-operand multiplication enhancement arises in extending it to a limited sum of products. Such an expression occurs frequently in coordinate transformation of an image into a new representation. A simple example is two-dimensional coordinate rotation, where the new abscissa is given by,

x' = x \cos \theta - y \sin \theta

The angle of rotation is common to all elements in the image, and so are the trigonometric multipliers applied. Assume main memory contains a description of the image and its word format is as follows:

<table>
<thead>
<tr>
<th>2</th>
<th>x</th>
<th>y</th>
<th>Carry</th>
</tr>
</thead>
</table>

Part, flag, x, y, Carry x'

126
As before, we treat \( x \) and \( y \) as multipliers, \( \cos \theta \) and \( \sin \theta \) as multipliers. A straightforward approach to enhancement might be to multiply \( \cos \theta \) by field \( x \), then \( \sin \theta \) by field \( y \), in both cases summing into field \( x' \) and handling \( b \) bits of the multiplier at a time. A better idea, however, is to handle one or more corresponding bits of \( x \) and \( y \) simultaneously. Assuming two bits of each at a time, the 16 precomputed multiples stored in auxiliary memory \( A' \) are:

\[
0, \sin \theta, 2 \sin \theta, 3 \sin \theta, \cos \theta, \sin \theta + \cos \theta, \ldots, 3 \sin \theta + 3 \cos \theta.
\]

The resulting algorithm is simpler and appreciably faster, its execution time being given by,

\[
4.5N + K(9M + 3 \log_2 T(b - 1) + 3.5)b
\]

where \( T \) is the number of products summed, and \( b \) is now the number of bits per multiplier handled at a time. Appropriate cases are plotted in Figure 3 for comparison against multi-operand enhancement of a single product. Enhancement of an extended sum of products will be treated under convolution below. It should be noted that all our expressions for maximum execution time have assumed the product field to be initially cleared. As a consequence, some improvement in speed may be achieved by executing the first iteration as a copy (rather than add) operation. The speed advantage grows with decreasing multiplier precision \( N \) and increasing number of multiplier bits handled at a time (\( b \) or \( 2b \)).

**CONVOLUTION**

Multi-operand enhancement of compound associative arithmetic operations will be discussed with reference to convolution, a function much used in signal and image processing. Let \( \mathbf{p} = [p_k] \) be the data vector and \( \mathbf{h} = [h_k] \) the filter vector, both of length \( P \). The discrete convolution of \( p \) by \( h \) is defined by:

\[
(h * p)_k = \sum_j h_j p_{k-j}
\]

where \( k = 0, 1, \ldots, 2(P-1) \). Or in matrix form,

\[
\begin{bmatrix}
  h_0 & 0 & \ldots & \ldots & 0 \\
  h_1 & h_0 & \ldots & \ldots & 0 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  h_{P-1} & \ldots & \ldots & h_0 & 0 \\
  0 & h_{P-1} & \ldots & \ldots & h_1 \\
  \vdots & \vdots & \ddots & \vdots & \vdots \\
  0 & \ldots & \ldots & h_{P-1} & 0
\end{bmatrix}
\begin{bmatrix}
  p_0 \\
  p_1 \\
  \vdots \\
  p_{P-1}
\end{bmatrix}
\]

Note that every element of \( p \) is multiplied by every element of \( h \). The resulting products are added up to form the convolution vector whose \( k \)-th element is the sum of all \( h_j p_{k-j} \) such that \( 1+j=k \). Let us examine the \( h \)-matrix:

its non-zero elements occupy the central parallelogram consisting of full diagonals, each diagonal filled with a common \( h \), starting with \( h_0 \) in the first diagonal and continuing top to bottom in increasing order of \( h \) subscript to the last diagonal which is \( h_{P-1} \). This fact

**TABLE 4**

<table>
<thead>
<tr>
<th>PHASE</th>
<th>OPERATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Initialize convolution:</td>
</tr>
<tr>
<td>2</td>
<td>Load b-bit code sequence in memory A'.</td>
</tr>
<tr>
<td>3</td>
<td>Partition memory A into two parts of equal bit-length.</td>
</tr>
<tr>
<td>4</td>
<td>Perform multi-operand enhancement of a single product.</td>
</tr>
<tr>
<td>5</td>
<td>Propagate result through bit-position ( b ).</td>
</tr>
<tr>
<td>6</td>
<td>Increment ( b ); if ( b ) exceeds the number of elements, go to 3.</td>
</tr>
<tr>
<td>7</td>
<td>Increment ( N ); if ( N ) exceeds the number of elements, exit.</td>
</tr>
<tr>
<td>8</td>
<td>Shift ( p )-field and marker down. Go to 2.</td>
</tr>
</tbody>
</table>

**Main memory format**

- **Phase** operates on 'marked' elements only.

**Figure 3**

- **I. Single product**
- **II. Extended sum of products**

**Execution time** in 10 memory cycles

<table>
<thead>
<tr>
<th>Multiplier bits per iteration</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
</tr>
<tr>
<td>0</td>
</tr>
</tbody>
</table>

*Limited sum of products*
TABLE 5

Carry Propagation

<table>
<thead>
<tr>
<th>STEP</th>
<th>OPERATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BCT := 0</td>
</tr>
<tr>
<td>1</td>
<td>m := d(b[!1]+N+BCT,IMP);SETAG</td>
</tr>
<tr>
<td>2</td>
<td>c := d(TMP);COMPARE</td>
</tr>
<tr>
<td>3</td>
<td>c := d(b[!1]+N+BCT);WRITE</td>
</tr>
<tr>
<td>4</td>
<td>c := d(b[!1]+N+BCT,IMP);SETAG;COMPARE</td>
</tr>
<tr>
<td>5</td>
<td>c := d(TMP);WRITE</td>
</tr>
<tr>
<td>6</td>
<td>BCT := BCT+1; if b[!1]+N+BCT&lt;2N+4log₂p (\text{go to 1, else exit.})</td>
</tr>
</tbody>
</table>

suggests a simple shifting algorithm summarized in Table 4, which includes the word format. The p's are entered in order and their marker bits set; a gap of P-1 words (or greater) is provided below the p vector to develop the last terms of the hp vector. We start by multiplying the elements of p by h₀ and sum the product into the hp field and shift the data vector down together with the marker. This cycle is repeated for each successive element of h. At the end of this process, the resulting vector will be sitting in proper order in the hp field. Note that the marker was used to limit arithmetic to the elements of p, regardless of their location in a given cycle. Returning to Table 4, the extended sum of products, hereafter referred to as "summed multiplication", is described in phases 2 through 6 and is speeded up by handling b multiplier bits at a time. The sub-algorithms of phases 5 and 8 are detailed in Tables 5 and 6 respectively. We now write expressions for the execution time of significant phases of the algorithm, given in memory cycles per vector element.

Phase 3 : N/(b + 9N/2)
4 : 9N(M+b)/b
5 : 9N(N + 24log₂p+4)/b
8 : 5(N+1)

We pause here to focus on summed multiplication and compare its performance with that of the multiplication algorithms discussed earlier. The execution time of summed multiplication is given by that of phases 3, 4, and 5; adding and simplifying we obtain:

\[ N(N+1)/b + 9N(24log₂p+5)/b \]

where \(b\geq 1\), \(N\) is the precision of \(p\), and \(M\) the precision of \(h\). For direct comparison, summed multiplication time is again plotted in Figure 3 for a 60-bit wordlength and \(b\) ranging from 1 to 6. As before, the value plotted for \(b=1\) is that for associative summed multiplication without enhancement:

\[ 9N(24log₂p+4)/b \]

It is interesting to note that multi-operand enhancement of summed multiplication pays off even at \(b=2\). We deliberately chose one of the simplest compound arithmetic operations to illustrate this point.

Returning to convolution, from the phase execution times given above we obtain the following expression for the total convolution time in memory cycles:

\[ P[N(N+1)/b + 9N(24log₂p+5)/b]/b + 5(N+1) \]

TABLE 6

Shift Field

<table>
<thead>
<tr>
<th>STEP</th>
<th>OPERATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>BCT := 0</td>
</tr>
<tr>
<td>1</td>
<td>c := d(TMP);SETAG;WRITE</td>
</tr>
<tr>
<td>2</td>
<td>c := d(FB+BCT,IMP);COMPARE</td>
</tr>
<tr>
<td>3</td>
<td>c := d(FB+BCT);SHIFTAG;WRITE</td>
</tr>
<tr>
<td>4</td>
<td>c := 0,m := d(TMP);SETAG;COMPARE</td>
</tr>
<tr>
<td>5</td>
<td>m := d(FB+BCT);WRITE</td>
</tr>
<tr>
<td>6</td>
<td>BCT := BCT+1; if BCT&lt;8, go to 1, else exit</td>
</tr>
</tbody>
</table>

This may be approximated by,

\[ 9N(24log₂p+5)/b + 9N(M+b)/b \]

Examination of the main algorithm reveals the memory utilization factor during processing to be 50 per cent (neglecting shift). Hence there was no point in looking for a possible partitioning into disjoint data sets that might be processed concurrently, and multi-operand enhancement was applied to the arithmetic instead. As vector and filter have 1024 elements given to 16-bit precision, and the multiplier bits are handled 4 at a time, then convolution will execute in 60 milliseconds. Both the wordlength and execution time can be reduced considerably if the hp field is shortened to 28 bits and 8-bit computation truncated accordingly, still producing the full attainable precision. In that case, convolution requires 8 chips of main memory per data vector to simultaneously convolve any number of data vectors by a common filter in under 45 milliseconds. For comparison we list some current commercial machines \([9, 10, 11]\), giving their approximate convolution time for two 1024 element vectors yielding a 2048 element result.

<table>
<thead>
<tr>
<th>COMPANY</th>
<th>MODEL</th>
<th>CONVOLUTION TIME - sec.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Digital Equipment</td>
<td>VAX-11/780</td>
<td>1.5</td>
</tr>
<tr>
<td>Floating Point Sys.</td>
<td>AP-1200</td>
<td>0.4</td>
</tr>
<tr>
<td>Cray Research</td>
<td>CRAY-1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

The VAX is representative of general purpose sequential machines. The Floating Point AP-1200 is an array processor especially designed for problems like convolution. Architecturally, it employs a pipeline to achieve parallelism in vector processing. Even the CRAY-1, a supercomputer with a pipeline architecture, takes more than double the time to execute convolution. Moreover, an associative convoler can filter many data vectors in parallel at a cost of 8 memory chips per additional 1024 element vector. Thus, for a main memory capacity of 64Kx64 bits (256 chips) the processing rate will be one vector every 1.4 milliseconds.
REFERENCES