Efficient systolic high speed architectures for delayed multipath two-dimensional FIR and IIR digital filtering

Prof. H.K. Kwan, MPhil, PhD, CEng, MIEEE
M.T. Tsim, Phil

Indexing terms: Filters and filtering, Digital filters, Recursive filters, Systolic architectures

Abstract: In this paper, two novel efficient block-level systolic architectures for the high speed realisation of delayed multipath two-dimensional FIR and IIR digital filters are presented. In practice, difficulty arises in the implementation of an IIR digital filter because of the inherent timing constraint in its recursive loop. With the new transformation method presented in the paper, an extra loop delay is allowed in the recursive part of the multipath structure of an IIR digital filter. Consequently, the problem of the inherent timing constraint in the recursive loop is solved. Two methods for the stabilisation of the transformation method are also introduced. The resultant block-level systolic structures remove the global communication requirements, which further increases the efficiency of the final realisation.

1 Introduction

Recently, two-dimensional (2-D) delayed multipath finite impulse response (FIR) and infinite impulse response (IIR) digital filtering architectures have been advanced, in References 1-3. The problem of inherent recursive loop delay, which affects the overall speed of a conventional 2-D IIR digital filter (as in the case of a 1-D IIR digital filter) still exists in our work, as presented in Reference 1. More recently, this problem has been solved by us, as reported in References 2-3, by the introduction of a new efficient transformation method. In this paper, we give a more detailed account of the work presented in References 2-3. Moreover, we present two methods for the stabilisation of the transformation method for IIR digital filters, as well as two novel block-level systolic architectures for the realisation of delayed multipath 2-D FIR and IIR digital filters.

In the following Section, the concept of latency in delayed multipath structure is first formulated for a 2-D FIR digital filter. Two block-level systolic architectures for the delayed multipath 2-D FIR digital filter realisation are then presented. A new transformation method, to solve the problem of extra loop delay in the recursive part of a 2-D FIR digital filter, is introduced, along with two stabilisation methods. On the basis of the block-level systolic structures developed, we then present two efficient block-level systolic architectures for 2-D IIR digital filters. Finally, a comparison of the hardware requirements and throughput performance is given.

2 Latency in delayed multipath 2-D FIR digital filters

Consider the transfer function of a 2-D FIR digital filter:

\[ H(z_1^{-1}, z_2^{-1}) = \sum_{m=0}^{M} \sum_{n=0}^{N} a_{mn} \cdot n \cdot z_1^{-m} \cdot z_2^{-n} \quad (1) \]

In general, eqn. 1 can be decomposed into a delayed multipath form in the \( z_1 \) dimension as

\[ H(z_1^{-1}, z_2^{-1}) = \sum_{i=0}^{N-1} z_1^{-i} \cdot A(z_1^{-i}, z_2^{-1}) \quad (2) \]

where

\[ A(z_1^{-i}, z_2^{-1}) = \sum_{L=0}^{L-1} \sum_{k=0}^{K-1} d_{kN} \cdot i \cdot n \cdot z_1^{-i} \cdot z_2^{-1} \quad (3) \]

and

\[ L = [(M + 1)/N] \quad (4) \]

The symbol \([x] \) denotes the smallest integer greater than or equal to \( x \). The structure of the direct form realisation of an FIR digital filter as given in eqn. 2 is shown in Fig. 1. Each of the FIR filter sections \( A(z_1^{-i}, z_2^{-1}) \) can be readily realised by an N-path structure as shown in Fig. 2. An N-path structure and its equivalent form is shown in Fig. 2. Because of the input and output switches of an N-path structure, there is an input/output processing delay of \( (N - 1) \) sample intervals in the implementation of the N-path structure for the transfer function \( A(z_1^{-i}, z_2^{-1}) \). The equivalent transfer function of a multipath structure with the input/output delay taken into account, is the original transfer function multiplied by a \( z_1^{-i} \) term, as shown in Fig. 2. In addition to the delay incurred in the N-path structure, we should also consider the delay required for the adders that add up the outputs of the FIR filter sections, as shown in Fig. 1. If we assume that the operations of the output adders require \( K \) sample intervals, the total extra delay is \( (N - 1 + K) \) sample intervals and can be easily obtained by multiplying both sides of eqn. 2 by a \( z_1^{-i}(N-1-K) \) term

\[ z_1^{-i}(N-1-K) \cdot A(z_1^{-i}, z_2^{-1}) \quad (5) \]
The transfer function, as given in eqn. 5, is now suitable for realisation by means of the delayed N-path structure. The resultant realisation is shown in Fig. 3. As we can see from eqn. 5, the resultant transfer function of the realisation is

$$H_n(z_1^{-1}, z_2^{-1}) = z_1^{-(N-1+k)}H(z_2^{-1})$$  (6)

Therefore, the equivalent transfer function of the structure shown in Fig. 3 is the same as the original transfer function, except that there is an overall extra input/output delay of $(N-1+K)$ sample intervals. This difference is not important and is often ignored in the realisation of FIR digital filters as it increases only the input/output latency. However, this is an important consideration in the realisation of an IIR digital filter as processing delay in the recursive loop is critical.

3 Block-level systolic structures for delayed multipath 2-D FIR digital filter realisation

The structure proposed in Fig. 3 requires global data communications in which the outputs from all filter sections, $A_j(z_1^{-N}, z_2^{-1})$, for $j = 0$ to $N-1$, are added together. Clearly, such nonlocal communication paths are undesirable in a large-scale multiprocessing implementation.

The delayed multipath structure can be further enhanced by systolic realisation of the structure, such that the global data-communications requirement can be eliminated. Basically, there are two forms of block-level systolic structure for the delayed multipath realisation of eqn. 5, which will be described in the following subsections.

3.1 Type I block-level systolic structure (input and output moving in same direction)

The first form of block-level systolic structure, type I, is shown in Fig. 4. With this systolic structure, both input and output move in the same direction and the input stream moves at half the speed of the output stream. The realisation comprises $N$ repetitive blocks, each of which consists of an $N$ path digital filter section, a delay element of two sample intervals, a delay element of one
sample interval and an adder, as shown in Fig. 5. With the designations of the input and output of the $k$th block as shown in Fig. 5, we have the following intrablock relationships:

$$W_k = T_k z_k^{-1}$$
$$U_k = \left[T_k z_k^{-(N-1)} \cdot A_k (z_k^{-1}, z_k^{-1}) + V_k \right] z_k^{-1}$$

With the systolic blocks connected in cascade, the outputs from the $k$th block are fed to the inputs of the $(k + 1)$th block. Hence, we have the following interblock relationships:

$$T_{k+1} = W_k$$
$$V_{k+1} = U_k$$

Also, we have the following boundary conditions for the zeroth and $(N - 1)$th block:

$$T_0 = X(z_1^{-1}, z_2^{-1})$$
$$V_0 = 0$$
$$U_{N-1} = Y(z_1^{-1}, z_2^{-1})$$

From eqns. 9, 7, and 11, we have

$$T_k = T_{k-1} z_1^{-2} = \cdots = T_0 z_1^{-2(N-1)} = X(z_1^{-1}, z_2^{-1}) z_1^{-2N}$$

If we apply eqns. 10, 11, and 14, eqn. 8 becomes

$$U_k = \left[X(z_1^{-1}, z_2^{-1}) z_1^{-2(N-1)} \cdot A_k (z_1^{-N}, z_2^{-1}) + U_k \right] z_1^{-1}$$

By expanding the recursive relation in eqn. 15, we obtain

$$U_k = X(z_1^{-1}, z_2^{-1}) \sum_{i=0}^{N-1} z_1^{-2i} z_2^{-i} \cdot A_k (z_1^{-N}, z_2^{-1}) z_1^{-1}$$

From eqn. 13, we get

$$Y(z_1^{-1}, z_2^{-1}) = U_{N-1}$$

$$= X(z_1^{-1}, z_2^{-1}) \sum_{i=0}^{N-1} z_1^{-2i} z_2^{-i} \cdot A_k (z_1^{-N}, z_2^{-1}) z_1^{-1}$$

$$= X(z_1^{-1}, z_2^{-1}) \sum_{i=0}^{N-1} z_1^{-2i} z_2^{-i}$$

$$= z_1^{-N} \cdot A_k (z_1^{-N}, z_2^{-1})$$

The transfer function of the systolic structure is therefore

$$T_k = T_{k-1} z_k^{-2}$$
$$U_k = U_{k-1} z_k^{-N} + T_k z_k^{-(N-1)} A_k (z_k^{-N}, z_k^{-1})$$

From the above derivation, the block-level systolic structure shown in Fig. 4 gives the realisation of the original FIR digital filter (eqn. 5). It is trivial, from eqn. 19, that an extra input/output delay of $(N - 1)$ sample intervals is incurred in this systolic structure. Hence, the total input/output delay that appears in this structure is $2(N - 1)$ (which takes into account the delay of $(N - 1)$ sample intervals incurred by the $N$-path structure). The design has the drawback of increased input/output delay (especially for large $N$, i.e. where $N > 2$), which is particularly unfavourable in the realisation of the recursive part of an IIR digital filter. A realisation example of a type I block-level systolic delayed three-path 2-D FIR digital filter is shown in Fig. 6.

3.2 Type II block-level systolic structure (input and output moving in opposite directions)

The second form of the block-level systolic structure, type II, is shown in Fig. 7. The realisation comprises $N$ repetitive blocks, each of which consists of an $N$-path digital filter section, two delay elements of one sample interval and an adder, as shown in Fig. 8. With this systolic structure, input and output move in opposite directions at the same speed. However, the systolic structure is a two-stage design, which requires the input and, hence, the output to be sampled once every two systolic clock cycles, which means that the input and output sequences are interleaved with a 'blank' data stream. To maintain a unity sampling rate at both input and output, we can transform the delay elements in each systolic block to delay elements of a half sample interval, as shown in Fig. 9. With the designations of the input and output of the $k$th block as shown in Fig. 9, we have the following intrablock relationships:

$$W_k = T_k z_k^{-1}$$
$$V_k = U_k z_k^{-N} + T_k z_k^{-(N-1)} A_k (z_k^{-N}, z_k^{-1})$$

With the systolic blocks connected in cascade, the outputs $W_k$ and $V_k$ from the $k$th block are fed to the input $T_{k+1}$ of the $(k + 1)$th block and $U_k$ of the $(k - 1)$th block.
From eqns. 22, 20 and 24, we have
\[
T_k = T_{k-1}z_1^{-1/2} = \ldots = T_0z_1^{-k/2} = X(z_1^{-1}, z_2^{-1})z_1^{-k/2}
\]  
(27)

If we apply eqns. 23 and 27, eqn. 21 becomes
\[
V_k = V_{k+1}z_1^{-1/2} + X(z_1^{-1}, z_2^{-1})z_1^{-k/2} \times z_1^{-(N-1)}A(z_1^{-N}, z_2^{-1})
\]  
(28)

By expanding the recursive relation in eqn. 28, we obtain
\[
V_k = V_k z_1^{-1} + \sum_{i=0}^{N-1} X(z_1^{-1}, z_2^{-1})z_1^{-(N-1)} \times z_1^{-(N-1)}A(z_1^{-N}, z_2^{-1})
\]  
(29)

From eqns. 23 and 26, \( V_0 = U_{N-1} = 0 \). Hence, eqn. 29 becomes
\[
V_k = X(z_1^{-1}, z_2^{-1}) \sum_{i=0}^{N-1} z_1^{-i} \times z_1^{-(N-1)}A(z_1^{-N}, z_2^{-1})
\]  
(30)

From eqns. 25 and 22, we get
\[
Y(z_1^{-1}, z_2^{-1}) = V_0 = X(z_1^{-1}, z_2^{-1}) \sum_{i=0}^{N-1} z_1^{-i} \times z_1^{-(N-1)}A(z_1^{-N}, z_2^{-1})
\]  
(31)

The transfer function of the block-level systolic structure is therefore
\[
Y(z_1^{-1}, z_2^{-1}) = \frac{X(z_1^{-1}, z_2^{-1})}{\sum_{i=0}^{N-1} z_1^{-i} \times z_1^{-(N-1)}A(z_1^{-N}, z_2^{-1})}
\]  
(32)

From the above derivation, the block-level systolic structure shown in Fig. 7 gives the realisation of the original FIR digital filter (eqn. 5). Also, we can observe from eqn. 32 that there is no extra input/output processing delay (i.e. \( k = 0 \)) incurred by this systolic structure, apart from the extra delay of \((N-1)\) sample intervals incurred by the N-path realisation of \(A(z_1^{-N}, z_2^{-1})\).

However, the systolic structure is actually a two-slow down design, which requires two clock cycles of the systolic array to process one input sample, and is only capable of generating one output every two clock cycles. It should also be noted that, because of the two cycles per unit sample requirement, the multipath switches in this structure will also switch once every two cycles. One way to relieve this drawback is to insert a buffer delay between the output of each filter section and the adder, as shown in Fig. 10. This serves to separate the block cycle of a filter section from that of the output adder. This has the overall effect of reducing the length of a cycle by an amount equal to the time required for one addition. However, such a modification also implies an extra input/output delay of one sample interval. The sampling period (though it has been effectively reduced) of the input, output and all multipath switches is still slowed down by a factor of two. For a two-slow down design like this, the processing time of a filter section can be fully used by having two independent convolution computations interleaved (instead of interleaving a 'blank' data stream) in the structure. This idea is directly applicable to HR digital filter realisation, as will be discussed in Section 5. A realisation example of a type II block-level
systolic (with preadder delay) delayed three-path 2-D FIR digital filter is shown in Fig. 11.

\[ B(z_1^{-1}, z_2^{-1}) = \sum_{n=0}^{M_2} z_1^n \cdot B_d(z_2^{-1}) \]  
(35)

and

\[
\begin{align*}
B_d(z_1^{-1}) & = \sum_{m=0}^{M_1} b(m, 0) \cdot z_1^{-m-1} \\
B_d(z_2^{-1}) & = \sum_{m=0}^{M_2} b(m, n) \cdot z_2^{-m-1}
\end{align*}
\]
(36)

for \( n = 1, 2, \ldots, M_2 \)

(37)

The structure of the direct form realisation of eqn. 33 is shown in Fig. 12. The realisation consists of two parts, namely, the 2-D FIR filter section \( A(z_1^{-1}, z_2^{-1}) \) of the nonrecursive part and the 2-D FIR filter section \( B(z_1^{-1}, z_2^{-1}) \) of the recursive part. As noted in eqn. 6, the realisation of a 2-D FIR filter section by means of the delayed N-path structure requires an extra input/output processing delay of \((N - 1 + K)\) sample intervals. This extra delay can be allowed in the processing of the nonrecursive part \( A(z_1^{-1}, z_2^{-1}) \), since this will not affect the overall filter characteristics. However, in the operation of the recursive loop, as shown in Fig. 12, the computation by the 2-D FIR filter section \( B(z_1^{-1}, z_2^{-1}) \) must be completed before the next input sample can be processed. In the present arrangement, there is no extra input/output processing delay allowed for the FIR filter section \( B(z_1^{-1}, z_2^{-1}) \). Consequently, an extra input/output delay must be added to realise the 2-D FIR filter section \( B(z_1^{-1}, z_2^{-1}) \) in the recursive loop of an IIR digital filter.

Fig. 12 Direct form 2-D IIR digital filter

So that the delayed N-path structure can be used to realise the FIR filter section \( B(z_1^{-1}, z_2^{-1}) \), the delay in the recursive loop of the IIR digital filter, must be increased to \((N + K)\) sample intervals (where \( K \) is the number of sampling intervals required from the computation in the output adder). This extra delay delay of \((N - 1 + K)\) sample intervals can be obtained by using a polynomial transformation technique. The transformation is to multiply both the numerator and denominator of \( H(z_1^{-1}, z_2^{-1}) \) by a polynomial \( Q(z_1^{-1}) \) of degree \((N - 1 + K)\), which is expressed as

\[
Q(z_1^{-1}) = 1 + q(1)z_1^{-1} + q(2)z_1^{-2} + \cdots + q(N - 1 + K)z_1^{-(N - 1 + K)}
\]
(38)

such that

\[
q(m) = -b(m, 0) - b(m - 1, 0)q(1) - \cdots - b(1, 0)q(m - 1) \quad \text{for } 1 \leq m \leq N - 1 + K
\]
(39)

Multiplication of both the numerator and denominator of the 2-D IIR filter transfer function by \( Q(z_1^{-1}) \) gives

\[
H(z_1^{-1}, z_2^{-1}) = \frac{Q(z_1^{-1})A(z_1^{-1}, z_2^{-1})}{Q(z_1^{-1})(1 + z_1^{-1}B(z_1^{-1}, z_2^{-1}))}
\]
(40)

That is

\[
\frac{A(z_1^{-1}, z_2^{-1})}{1 + z_1^{-1}B(z_1^{-1}, z_2^{-1})}
\]
(41)
where

\[
A(z_1^{-1}, z_2^{-1}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a(m, n) z_2^{-m} z_1^{-n}
\]

\[
B(z_1^{-1}, z_2^{-1}) = \sum_{m=-\infty}^{\infty} z_2^{-m} B_0^m(z_2^{-1})
\]

and

\[
B_0^m(z_2^{-1}) = \sum_{n=-\infty}^{\infty} b(m, n) z_2^{-m} z_1^{-n}
\]

for \(n = 1, 2, \ldots, M_2^e\).

Note that the recursive loop delay term attached to \(B(z_1^{-1}, z_2^{-1})\) in eqn. 41 has been increased to \(z_1^{-N-K}\) after the transformation.

In general, the roots of the transformation polynomial \(Q(z_1^{-1})\), as given by eqns. 38 and 39, may not lie within the unit circle. Therefore, the transformed transfer function of eqn. 41 may experience an instability problem. There are some methods to ensure that \(Q(z_1^{-1})\) is a stable transformation polynomial. However, these stabilisation methods inevitably increase the length of \(Q(z_1^{-1})\), which then decreases the efficiency of the final structure. If the resulting \(Q(z_1^{-1})\), as given by eqns. 38 and 39, is stable, we can use it as the transformation polynomial. Otherwise, we must use some methods to stabilise \(Q(z_1^{-1})\), as introduced in the following subsections:

4.1 Successive transformation by choosing stable roots

If the \(Q(z_1^{-1})\) obtained in eqns. 38 and 39 is unstable, we can construct \(Q(z_1^{-1})\) by the following method. In this method, one extra delay is obtained on each transformation. The total number of extra delays \(D\) is obtained by successively applying such a transformation \(D\) times. The \(i\)th transformation (for \(i = 1, 2, \ldots, D\)) is given by

\[
1 + z_1^{-i} B(z_1^{-1}) = [1 + z_1^{-i} B_i(z_1^{-1})]
\]

\[
\times [1 - b_{-i}(i, 0) z_1^{-i} S_i]
\]

where \(b_{-i}(i, 0)\) is the first nonzero coefficient of \(z_1^{-i}\) of the polynomial \([1 + z_1^{-i} B_i(z_1^{-1})]\), the initial polynomial \(B_0(z_1^{-1})\) is given by eqn. 36 and \(S_i\) is an integer to be chosen. The validity of the transformation given by eqn. 46 can easily be checked by expanding the right-hand side of the equation and noting the coefficients of \(z_1^{-k}\) (\(k = 1, 2, \ldots, i+1\)), which should be zero. To obtain a stable transformation, the integral scaling factor \(S_i\) should be chosen such that \(b_{-i}(i, 0) S_i\) is less than one and so is the modulus of the roots of the transformation polynomial. If \(D\) is the number of extra delays required, the overall transformation polynomial is given by

\[
Q(z_1^{-1}) = \prod_{i=1}^{D} [1 - b_{-i}(i, 0) z_1^{-i} S_i]
\]

and the length of the transformation polynomial is given by \((S_1 + 2S_2 + \cdots + DS_D)\).

4.2 Numerical approximation

Numerical techniques can also be used to find a stable transformation polynomial, which can be optimal. In this method, the unstable transformation polynomial \(Q(z_1^{-1})\) given by eqns. 38 and 39 is first computed. On the basis of \(Q(z_1^{-1})\), extra high-order terms are added one by one to the polynomial until the overall polynomial is within a stable margin. The extra term to be added can be numerically chosen such that the largest root (in terms of the modulus of the roots) of the polynomial is minimised. The minimisation process is done in a computer using numerical approximation. Addition of an extra term can cease when the overall polynomial is stabilised, i.e. all roots of the polynomial are within the unit circle. In general, the number of extra terms added is less than that of the previous method, independent of the original order of the denominator polynomial of eqn. 33.

5 Block-level systolic structures for delayed multipath 2-D IIR digital filter realisation

If we take into account the stabilisation procedure, which may be required in case the original transformed polynomial is unstable, the degree of the resultant stabilised transformed polynomial \(Q(z_1^{-1})\) will be \((N - 1 + K + E)\):

\[
Q(z_1^{-1}) = 1 + q(1) z_1^{-1} + q(2) z_1^{-2} + \cdots + q(N - 1 + K + E) z_1^{-N - 1 + K + E}
\]

\[
+ q(N - 1 + K + E) z_1^{-N - 1 + K + E}
\]

such that

\[
q(m) = -b(m, 0) - b(m, 1), 0 \leq q(1) \cdots
\]

\[
- (b(1), 0 \leq q(1) \cdots)
\]

for \(1 \leq m \leq N - 1 + K + E\) (49)

and \(q(N + K), \ldots, q(N + K + E)\) are the extra high-order terms required by the stabilisation procedure. The general form of the stabilised transformed transfer function is therefore given by

\[
H(z_1^{-1}, z_2^{-1}) = \frac{Q(z_1^{-1}) A(z_1^{-1}, z_2^{-1})}{Q(z_1^{-1}) \left[1 + z_1^{-1} B(z_1^{-1}, z_2^{-1})\right]}
\]

\[
= \frac{A(z_1^{-1}, z_2^{-1})}{1 + z_1^{-1} B(z_1^{-1}, z_2^{-1})}
\]

where

\[
A(z_1^{-1}, z_2^{-1}) = \sum_{m=-\infty}^{\infty} \sum_{n=-\infty}^{\infty} a(m, n) z_2^{-m} z_1^{-n}
\]

\[
B(z_1^{-1}, z_2^{-1}) = \sum_{m=-\infty}^{\infty} z_2^{-m} B_0^m(z_2^{-1})
\]

\[
B_0^m(z_2^{-1}) = \sum_{n=-\infty}^{\infty} b(m, n) z_2^{-m} z_1^{-n}
\]

\[
A(z_1^{-1}, z_2^{-1}) = \sum_{m=-\infty}^{\infty} b(m, n) z_2^{-m} z_1^{-n}
\]

\[
for n = 1, 2, \ldots, M_2^c
\]

The transformed FIR filter sections \(A(z_1^{-1}, z_2^{-1})\) and \(B(z_1^{-1}, z_2^{-1})\) given in eqns. 52-55 can then be decomposed in a similar way as to that for eqns. 1-4, such that the numerator and the denominator polynomials can be readily implemented by means of the N-path structure. Hence,

\[
H(z_1^{-1}, z_2^{-1}) = \frac{\sum_{i=0}^{N-1} z_1^{-i} A(z_1^{-1} N - z_2^{-1})}{1 + z_1^{-N-K} \sum_{i=0}^{N-1} z_1^{-i} B(z_1^{-1} N - z_2^{-1})}
\]
where

\[ A(z_1, z_2) = \sum_{j=0}^{M_1} \sum_{x=0}^{M_2} a^*(j) z_1^j z_2^x \]

\[ B_k(z_1, z_2) = \sum_{x=0}^{M_2} b_k^*(x) z_2^x \]

\[ B_{k,j}(z_1, z_2) = \sum_{x=0}^{M_2} b_{k,j}(x) z_2^x \]

\[ B_{k,j}(z_1, z_2) = \sum_{x=0}^{M_2} h_k(x) z_1^x z_2^x \]

for \( n = 1, 2, \ldots, M_2 \)

and

\[ L^* = [(M_1 + N + K + E)/N]^* \]

\[ L^* = [(M_1 + E)/N]^* \]

\[ L^* = [(M_1 + N + K + E)/N]^* \]

In principle, the transformation preserves all characteristics of the original 2-D IIR digital filter as it is achieved by multiplying both the numerator and denominator polynomials of the original filter transfer function by the same \( z_1^{-1} \). The transformed IIR filter transfer function, as shown in eqns. 56-63, is now suitable for realisation by means of the delayed \( N \)-path structure in both the recursive and nonrecursive parts. The resultant structure is as shown in Fig. 13.

Similarly to the case of a 2-D FIR digital filter, the delayed multipath realisation shown in Fig. 13 can be systolised at block-level. Both types of block-level systolic structures (Figs. 4 and 10) described in the realisation of FIR digital filters can be applied to the realisation of IIR digital filters:

5.1 Type I block-level systolic structure (with input and output moving in the same direction)

The FIR digital filter systolic structure as shown in Fig. 4 can be used for the FIR digital filter sections in both the recursive and nonrecursive parts of eqns. 56-63. The resulting IIR digital filter structure is shown in Fig. 14. As partly mentioned in Section 3.1, this realisation has the disadvantages of an extra increase in the input/output delay of \( (N + 1) \) sampling intervals (on top of the \( (N - 1) \) sampling delays produced by an \( N \)-path structure) in the nonrecursive part and an extra \( (N + 1) \) sampling delay in the recursive part. This implies that an extra delay of \( (N + 1) \) sampling intervals must be added to the recursive loop. Hence, the length of the transformed polynomial is longer and the computational efficiency of the final structure will be decreased, especially for a large \( N \) (where \( N \geq 2 \)). A realisation example of a type I block-level systolic delayed three-path 2-D IIR digital filter is shown in Fig. 15.

5.2 Interleaved type II systolic structure (with input and output moving in opposite directions)

Fig. 16 gives the resultant structure of the direct application of the type II block-level systolic structure shown in Fig. 10 to both the recursive and nonrecursive parts of a 2-D IIR digital filter. The structure has the advantage of having an extra increase of only one sampling delay in the nonrecursive part and an extra increase of two sam-

![Fig. 13 Delayed multipath 2-D IIR digital filter](image)
Fig. 14 Type I block-level systolic delayed multipath 11R digital filter

Fig. 15 Realisation example of the type I block-level systolic delayed three-path 2-D 11R digital filter
pling delays in the recursive part (on top of the \((N - 1)\) sampling delays produced by an \(N\)-path structure).

![Diagram](image)

**Fig. 16** Type II block-level systolic delayed multipath 2-D IIR digital filter

However, the design gives only one sample output every two cycles (i.e., it is a two-slow down design) and the efficiency is effectively halved.

To overcome this two-slow down drawback, we may interleave the filter sections \(A_0(z^{N/2})\) and \(B_0(z^{N/2})\) such that one cycle is used for a nonrecursive section while the other is used for a recursive section. With the interleaved mechanism, a two-way output switch is required to segregate the nonrecursive output and the recursive loop output. The input data stream is clocked twice into the systolic array, instead of interleaving a ‘blank’ data stream, as discussed in the Section 3.2. Duplicated input data streams are used by two sets of interleaved 2-D FIR digital filter sections, \(A_0(z^{N/2})\) and \(B_0(z^{N/2})\). The interleaved type II block-level systolic structure of a delayed multipath 2-D IIR digital filter is shown in Fig. 17. It should be noted that, because of the interleaved arrangement, the multipath switch must switch once every systolic cycle, although the input and output streams are still clocked in every two cycles. With this arrangement, the utilisation of the structure can be kept at 100% and, hence, no efficiency is lost.

As the input, the output and each of the subtransfer functions of either of the nonrecursive or recursive part are all operating at a sampling rate of two clock cycles per second (i.e. \(2T\)), the overall throughput rate can be increased to \(T\) clock cycles per second by simply increasing the sampling rate of the input, the output and each of the subtransfer functions of either the nonrecursive or recursive part to \(T\) clock cycles per second. In so doing, the sampling rate of the switch connecting alternately to the output and recursive loop of the filter must be increased to \(T/2\) clock cycles per second. Hence, both the type I and the type II systolic structures for an IIR digital filter will result in the same throughput rate. Furthermore, the interleaved version of the type II realisation

**Fig. 17** Interleaved type II block-level systolic delayed multipath 2-D IIR digital filter

*IEEE PROCEEDINGS, Vol. 137, Pt. G, No. 6, DECEMBER 1990*
structure is more efficient in terms of hardware requirement for $N > 2$ (as compared to the type I structure with an extra $(N + 1)$ sampling delay) and its overall structure is also more compact and regular (as compared to that of the type I realisation structure). A realisation example of the interleaved type II block-level systolic delayed three-path 2-D IIR digital filter is shown in Fig. 18.

![Fig. 18 Realisation example of the interleaved type II block-level systolic delayed three-path 2-D IIR digital filter](image)

6 Hardware requirement and throughput performance

As we can see from Figs. 4 and 10, the number of processors required to implement each of the two block-level systolic delayed N-path 2-D FIR digital filters is $N^2$. For the case of a direct form 2 2-D IIR digital filter, it is decomposed into two FIR digital filter sections. Hence, the number of processors required for each of the block-level systolic structures of a delayed N-path 2-D IIR digital filter is $2N^2$ (see Figs. 14 and 17).

As the computational complexity of each of the two block-level systolic delayed multipath realisations for a 2-D FIR digital filter is the same as the original filter transfer function, the maximum system throughput improvement can be regarded as equal to the number of processors, i.e. $N^2$.

For the case of a 2-D IIR digital filter, the maximum system throughput improvement is slightly less than the number of processors, since the computational complexity of the final realisation is slightly greater than that of the original 2-D IIR digital filter transfer function. The increase in computational complexity is brought about by the polynomial transformation required to provide an extra computational delay in the recursive loop. Consider the original 2-D IIR digital filter transfer function given in eqns. 33–37, the number of coefficients in $A(z_1^{-1}, z_2^{-1})^*$ is $(M_1 + 1)(M_1 + 1)$ and that for $B(z_1^{-1}, z_2^{-1})^*$ is $(M_1 + 1)(M_1 + 1) - 1$. Therefore, the computational complexity, measured in terms of number of coefficients, of the original transfer function is

$$(M_1 + 1)(M_1 + 1) + (M_1 + 1)(M_1 + 1) - 1$$

However, for the stabilised transformed transfer function given in eqns. 50–55, the number of coefficients in $A(z_1^{-1}, z_2^{-1})^*$ is

$$(M_1 + N + K + E)(M_1 + 1)$$

and that for $B(z_1^{-1}, z_2^{-1})^*$ is

$$(M_1 + E + (M_1 + N + K + E)M_1$$

Therefore, the computational complexity of the established transformed transfer function is

$$(M_1 + N + K + E)(M_1 + 1) + M_1 + E + (M_1 + N + K + E)M_1$$

which is larger. Hence, the maximum achievable throughput improvement of each of the block-level systolic structures for a delayed N-path direct form 2 2-D IIR digital filter as compared to that of the direct form structure is therefore

$$2N^2[(M_1 + 1)(M_1 + 1) + (M_1 + 1)(M_1 + 1) - 1]$$

Table 1 shows the numerical hardware requirement and maximum throughput improvement of a 2-D FIR digital filter and a 2-D IIR digital filter of filter order eight, i.e. $M_1 = M_2 = M_3 = M_4 = 8$. The realisation example of the 2-D FIR digital filter assumes the use of either the type I or type II block-level systolic delayed three-path structure (see Figs. 6 and 11), whereas the realisation example of the 2-D IIR digital filter assumes the use of an interleaved type II block-level systolic delayed three-path structure (see Fig. 18). Also, it is assumed that the transformation polynomial for the 2-D IIR digital filter requires no extra higher term for the purpose of stabilisation, i.e. $E = 0$.

<table>
<thead>
<tr>
<th>Structure</th>
<th>Maximum throughput improvement</th>
<th>Number of processors required</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conventional direct form</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Delayed three-path 2-D FIR digital filter</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>Delayed three-path direct form 2 2-D IIR digital filter</td>
<td>13.67</td>
<td>18</td>
</tr>
</tbody>
</table>

$M_1 = M_2 = M_3 = M_4 = 8; N = 3; E = 0$
7 Concluding remarks

Two efficient block-level systolic structures for the high speed realisation of delayed N-path 2-D FIR and IIR digital filters with $N^2$ and $2N^2$ processors, respectively, have been presented to eliminate global data communication requirement and to enhance throughput improvement. Of the two proposed systolic structures, the type I systolic structure is recommended for high-speed 2-D FIR digital filter implementation if an increase of latency of $(N-1)$ sampling intervals (see Fig. 4) is not a problem. On the other hand, the type II systolic structure (with an increase of only one sampling interval in latency, see Fig. 10) can be used with the same throughput as that of the type I structure by increasing the sampling rate of the input, the output and the subfilter switches by a factor of two. For an IIR digital filter, the type I systolic structure is recommended for high-speed 2-D IIR digital filter implementation for small $N$ ($N \leq 2$), with an extra loop delay of $(N+1)$ sampling intervals (see Fig. 14) in the recursive part and an increase of $(N+1)$ sampling intervals (see Fig. 14) in the latency of the overall filter (or the nonrecursive part). The interleaved type II systolic structure is recommended for high-speed 2-D IIR digital filter implementation for large $N$ ($N \geq 2$), with an increase of two sampling intervals (see Fig. 17) in the recursive part and an increase in latency of only one sampling interval (see Fig. 17) in the overall filter (or the nonrecursive part). In general, the choice of which type of block-level systolic realisation is appropriate depends on the speed (large $N$ or small $N$) and latency requirements of the application under consideration. In this paper, we have not discussed the systolic realisation along $z^{-1}$ of each of the subfilters of the type I and type II systolic structures of FIR and IIR digital filters, shown in Figs. 4, 10, 14 and 17. In general, each subfilter can also be systolised along $z^{-1}$ in a similar manner to a conventional convolution section, such as those shown in Figs. 4 and 10 (with each of the $N$-path sections replaced by a real multiplier).

8 References