

Parallelizing Synchronous Data-Flow Graphs via Retiming*

Timothy W. O’Neil
Dept. of Comp. Science & Eng.
University of Notre Dame
Notre Dame, IN 46556

Edwin H.-M. Sha
Dept. of Comp. Science
Erik Jonsson School of Eng. & C.S.
Box 830688, MS EC 31
Univ. of Texas at Dallas
Richardson, TX 75083-0688

Sissades Tongsimma
High Perf. Comp. Lab.
Nat’l. Electronics & Comp. Tech. Center
11th Floor Bangkok Thai Tower Bldg.
108 Rangnam Road
Phyathai, Rachathewe, Bangkok 10400
THAILAND

Abstract

Many common iterative or recursive DSP applications can be represented by synchronous data-flow graphs (SDFGs). A great deal of research has been done attempting to optimize such applications through retiming. However, despite its proven effectiveness in transforming single-rate data-flow graphs to equivalent DFGs with smaller clock periods, the use of retiming for attempting to reduce the execution time of synchronous DFGs has never been explored. In this paper, we do just this. We develop the basic definitions and results necessary for expressing and studying SDFGs. We review the problems faced when attempting to retime a SDFG in order to minimize clock period, then present an algorithm for doing this. Finally, we demonstrate the effectiveness of our method on several examples.

1 Introduction

Since the most time-critical parts of DSP applications are loops, we must explore the parallelism embedded in the repetitive pattern of a loop. One of the most useful models for representing DSP applications has proven to be the *multirate* or *synchronous data-flow graph (SDFG)* first proposed by Lee [13]. The nodes of a SDFG represent functional elements, while edges between nodes represent connections between them. Each node consumes and produces a predetermined fixed number of *delays* (i.e., data tokens) on each invocation. Additionally, each edge may contain some initial number of delays. This model has proven popular with designers of signal processing programming environments [9, 11, 18, 23] with its use leading to numerous important results regarding the scheduling [7], hierarchization [21], vectorization [20] and multiprocessor allocation [8, 13] of DSP programs.

A great deal of research has been done attempting to optimize various aspects of an application’s execution by applying various graph transformation techniques to the application’s SDFG. One of the more effective of these techniques is *retiming* [15, 16], where delays are redistributed among the edges so that hardware is optimized while the application’s function remains unchanged. Retiming was initially applied to single-rate DFGs to optimize the application’s schedule of tasks so that the *clock period* of the graph (i.e., the total computation time of the longest zero-delay path) was decreased in order for the application to be more efficiently scheduled for execution on multiprocessors [3–5]. It was later extended to the more general SDFG model in order to extend vectorization capabilities [25] or minimize the total delay count of a SDFG [24]. However, the problem

*This work is partially supported by NSF grants MIP95-01006 and MIP97-04276, and by the A. J. Schmitt Foundation.

of using retiming to minimize the clock period of a multirate DFG has remained unexplored. In this paper, we will discuss this problem and propose a method for accomplishing this task.

To illustrate the benefits of the retiming transformation, consider the single-rate DFG in Figure 1(a). The numbers above the nodes represent computation times of individual tasks, while the short bar-lines cutting the edges are the delays. We can see that our clock period in this case is 4 due to the zero-delay edge from A to B , hereafter referred to as (A, B) . Retiming allows us to remove a delay from (C, A) and place it on (A, B) to create the retimed graph in Figure 1(b) with clock period 3. The function of the two graphs is the same, with the only complication being that we will have to provide the first value of A when we begin execution. The costs of doing this are miniscule when we consider that we will be saving a clock cycle each time we execute the loop.

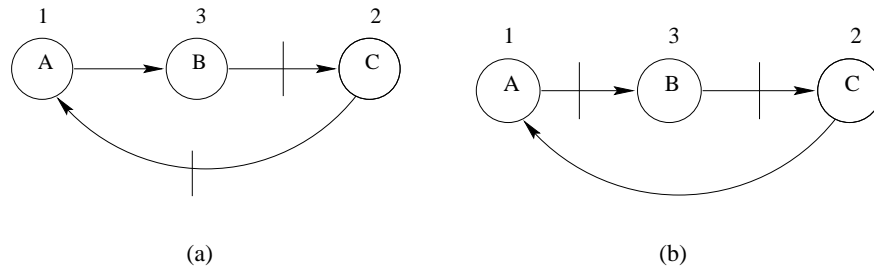


Figure 1: (a) A single-rate data-flow graph; (b) This DFG retimed to have clock period 3

An additional benefit is that scheduling alone usually yields a schedule requiring more resources than the schedule produced by retiming first. To illustrate this, consider the single-rate data-flow graph in Figure 2(a). It is clear that this graph has a clock period of 4, and we can derive the schedule in Figure 2(b) which has this clock period. Note that this schedule requires a minimum of 5 processing units to execute because of the work called for at any time-step greater than zero which is a multiple of 4.

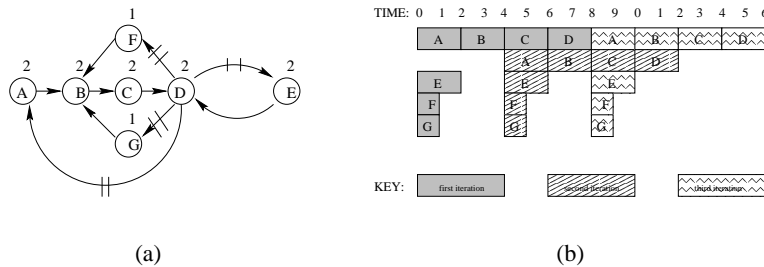


Figure 2: (a) A sample DFG; (b) Its schedule with cycle period 4

On the other hand, suppose that we retime our graph to become the DFG of Figure 3(a). This retimed graph permits us flexibility when scheduling node E , allowing us to compact our schedule and produce the one given in Figure 3(b) which requires only 3 processors, a 40% reduction in resources required for execution.

The benefits are clear, but reworking our retiming methods so that they may be applied to synchronous graphs is not easy. The difference between the single-rate and multi-rate models lies in the specification of production and consumption rates on each edge; in single-rate graphs all such rates are assumed to be the same, whereas different rates for different edges are typically specified when constructing SDFGs. Two pitfalls were

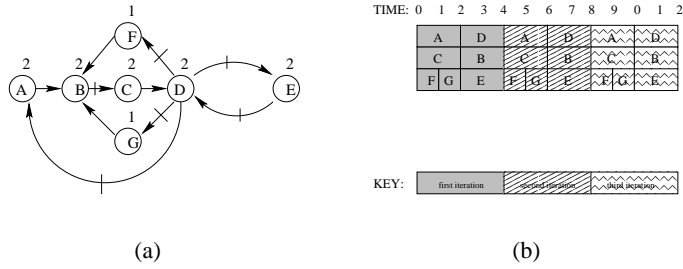


Figure 3: (a) Figure 2(a) retimed; (b) The retimed DFG’s schedule with cycle period 4

noted in [26]. First of all, a retiming may be derived for a single-rate DFG by solving a linear programming problem [16]. The introduction of rates on the edges potentially changes this to a more complicated integer linear programming problem. Second, the introduction of rates invalidates the traditional results regarding the delay counts of paths and cycles, depriving us of many useful results derived for the single-rate case. Specifically, in the single-rate case, we seek to remove zero-delay path with excessive total computation times. It isn’t clear what we want to avoid in the multi-rate case; a specific delay count on one path may or may not be adequate, depending on what rates have been specified.

Finally, the most popular method for retiming SDFGs has been to translate the SDFG to its single-rate equivalent, retime this new graph, then translate back [10, 24]. There are two problems with this idea. First, as we will demonstrate, it may be impossible to translate a retimed single-rate graph back to a retimed SDFG. Second, even if this method works, the costs in performing the necessary translations and dramatically increasing our problem size may be prohibitive. It is clearly preferable to work with the original SDFG as much as possible.

In this paper, we will develop the basic definitions and results necessary for specifying and manipulating a SDFG and its single-rate equivalent. We will review retiming and point out the problems which arise when it is applied to SDFGs. We will propose a polynomial-time algorithm which retimes a given SDFG to have a specified clock period. Finally, we will demonstrate the effectiveness of our algorithm by applying it to several examples.

In the next section, we will formalize the fundamental concepts related to the study of synchronous data-flow graphs. We then discuss retiming and the problems we face as we apply it to SDFGs. Next is our retiming algorithm, followed by detailed examples. Finally, we summarize our work and point to future directions for study.

2 Synchronous Data-Flow Graphs

The concept of a synchronous data-flow graph was developed and used extensively by Lee and Messerschmitt [12–14], but was not rigorously defined until the work of Zivojnovic *et al* [22, 24, 26]. In this section, we review their definitions and ideas in order to formalize these concepts.

2.1 Basic Definitions

A *synchronous data-flow graph (SDFG)* (sometimes called a *multirate* or *regular* data-flow graph) is a finite, directed, weighted graph $G = \langle V, E, d, t, p, c \rangle$ where:

1. V is the vertex set of nodes or *actors*, which transform input data streams into output streams;
2. $E \subseteq V \times V$ is the edge set, representing channels which carry data streams;
3. $d : E \rightarrow \mathbf{N} \cup \{0\}$ is a function with $d(e)$ the number of initial tokens (*delays*) on edge e ;
4. $t : V \rightarrow \mathbf{N}$ is a function with $t(v)$ the execution time of node v ;
5. $p : E \rightarrow \mathbf{N}$ is a function with $p(e)$ the number of data tokens produced at e 's source node to be carried by e ;
6. $c : E \rightarrow \mathbf{N}$ is a function with $c(e)$ the number of data tokens consumed from e by e 's sink node.

(In this definition \mathbf{N} is the set of natural numbers $\{1, 2, 3, \dots\}$.) If $p(e) = c(e) = 1$ for all $e \in E$, we say that G is a *homogeneous data-flow graph (HDFG)*. HDFGs are also sometimes referred to as *single-rate data-flow graphs* or simply *data-flow graphs*.

To illustrate, consider the SDFG given in Figure 4(a) below. The numbers above the nodes represent the execution times for the individual tasks, while the smaller numbers at either end of an edge denote tokens produced or consumed. As an example, $t(A) = 2$ while $t(B) = t(C) = 1$ in the figure. Furthermore, the numbers at either end of the edge connecting A and B indicate that node A produces one token on this edge when it executes, while node B consumes two tokens from this edge each time it fires.

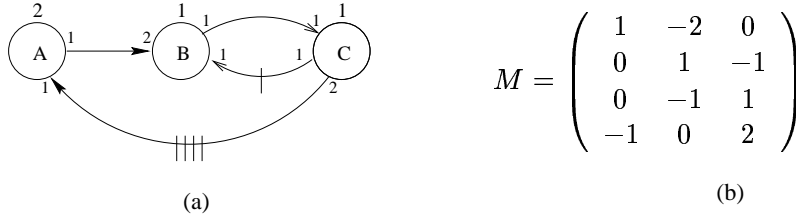


Figure 4: (a) A sample SDFG; (b) Its topology matrix M

It is sometimes useful to characterize an SDFG by its *topology matrix*, an $|E| \times |V|$ matrix similar to an incidence matrix. Each row corresponds to one edge in the graph, while each column corresponds to a node. A positive $(i, j)^{th}$ entry in the topology matrix indicates the number of tokens produced by the j^{th} node on the i^{th} edge, while a negative entry here gives the number of tokens consumed by node j from edge i . All other entries are zero. As an example, the topology matrix of Figure 4(a) is given in Figure 4(b).

In [13] it was demonstrated that a repeating sequential schedule can be constructed for a SDFG G if the rank of the graph's topology matrix is one less than the number of nodes in the SDFG. (The reverse is not necessarily true, as we will see shortly.) If this condition holds there is a positive integer vector q in the nullspace of the topology matrix called a *repetition vector* for G . The repetition vector for G with the smallest norm is called the *basic repetition vector (BRV)* for G [1]. For example, the BRV for the SDFG in Figure 4(a) is $q = [2 \ 1 \ 1]^T$. The elements of a BRV q indicate that q_j copies of node v_j must be executed during

every iteration of the static schedule. In our example we must schedule two copies of A and one copy each of B and C each time; see Figure 5(a). Finally, a SDFG is *consistent* if it has a BRV. An example of an inconsistent SDFG appears as Figure 5(b), with its rank 3 topology matrix in Figure 5(c). It is clear that, if we attempt to execute this circuit, each node will fire once before node A deadlocks the system waiting for its second token.

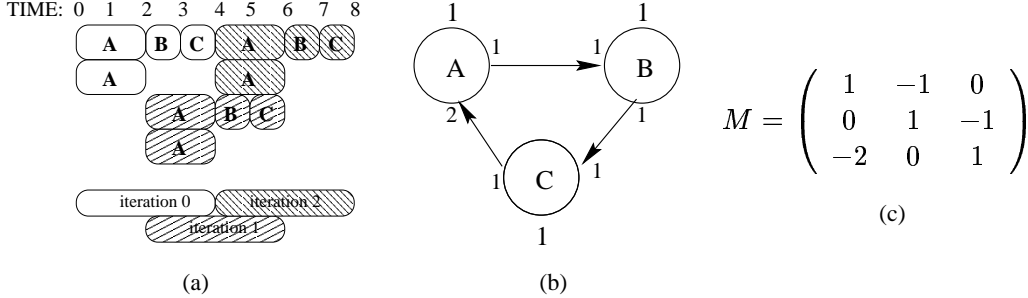


Figure 5: (a) The repeating schedule for Figure 4(a); (b) An inconsistent SDFG; (c) Its topology matrix

2.2 Constructing an Equivalent HDFG

In order to study an SDFG, it is sometimes useful to create its *equivalent homogeneous data-flow graph (EHG)*. As the name implies, an EHG performs the same function as the original SDFG, but is constructed so that each edge carries at most one token. Since each node is expecting to either produce or consume more data than this, an EHG compensates by inserting multiple edges between nodes.

An algorithm for creating a graph’s EHG appears as Algorithm 1 below. It is adapted from the method of [1] for constructing the EHG of *cyclostatic* DFGs, which not only permit multiple tokens to pass along edges but also specifies the pattern of their production or consumption. The algorithm first creates enough copies of each node to satisfy the specifications of the BRV. It then inserts edges. If nodes in a SDFG are connected by a zero-delay edge, then the first data token produced by the first copy of the source must be consumed by the first copy of the sink in the EHG. If there are delays on an edge, the data contained here is consumed first, so that the first new token produced is in fact needed by a later copy of the sink. The algorithm determines which copies of source and sink to map to one another based on how much data has been created and used. As an example of our algorithm in action, the EHG of Figure 4(a) appears in Figure 6.

There are two significant differences between our algorithm and that of [1]. First, the original algorithm was more concerned with making sure that the amount of data produced and consumed on an edge matched. This yields a simpler but more confusing graph. For purposes of clarity, we do not combine edges between nodes. If multiple tokens are to be sent between nodes in the EHG, each travels along its own edge. One benefit is that the delay counts between the original SDFG and the EHG match in our model. More significantly, the original algorithm also inserted control dependencies into the EHG, insuring that all copies of a node execute serially. Since we are concerned with maximizing parallelism, we concern ourselves only with the necessary data dependencies.

Finally, as derived in [1] and [6], we will say that a SDFG is *live* if its EHG has no zero-weight cycles. Otherwise the graph is *deadlocked*. An example of a consistent deadlocked graph appears as Figure 7(a), with its EHG in Figure 7(b). As we can see, the loop between nodes A and B_2 contains no delays, and so it is impossible to schedule them since each must precede the other. It should be clear that a SDFG must be both live and consistent in order for it to have a repeating static schedule.

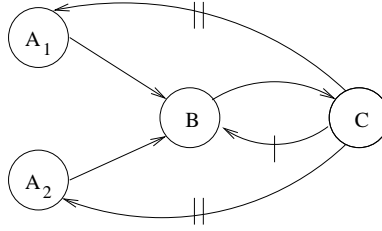


Figure 6: Figure 4(a)'s EHG

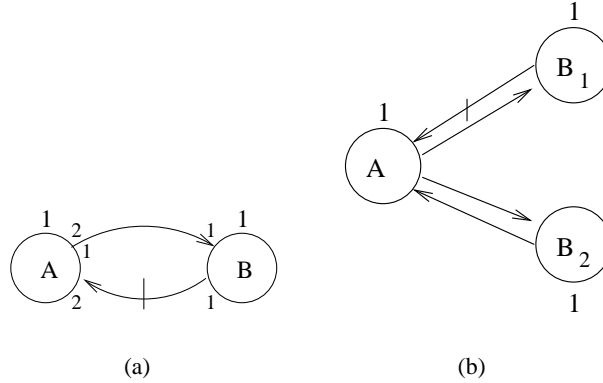


Figure 7: (a) A deadlocked SDFG; (b) Its EHG

3 Retiming

A great deal of research has been done attempting to optimize the schedule of an application's tasks after applying various graph transformation techniques to the application's HDFG. One of the more effective of these techniques is *retiming* [15, 16], where delays are redistributed among the edges so that the application's function remains the same while the execution time decreases. Despite its usefulness when applied to HDFGs, the application of retiming to SDFGs was explored only marginally prior to 1994 [10, 17] before being studied by Zivojnovic *et al* primarily as a way to minimize the delay count of a SDFG [24, 26]. In this section we intend to review the basics of retiming, explore some of the pitfalls which arise when studying retiming of SDFGs, demonstrate the effectiveness of retiming, and propose two algorithms for retiming SDFGs.

3.1 Basic Definitions

A *path* in either a SDFG or a HDFG is any sequence of nodes and edges. The *clock period* $cl(G)$ of a HDFG G is then defined to be the length of the longest zero-delay path [2]. This definition is problematic on two counts. First, it is not clear what the delay count of a path in a SDFG really is in light of the inconsistencies in the results from [24]. Second, suppose that we attempt to apply our definition directly to SDFGs, as demonstrated by Figure 7(a). We would conclude that the clock period equals 2, but in reality the graph must have an infinite clock period because of the problems scheduling nodes A and B_2 . Thus, we are forced to define the clock period of a SDFG to be equal to the clock period of its EHG. As an example, the clock period of the SDFG in Figure 4(a) is 4 by our definition.

Algorithm 1 Creating an EHG for a SDFG

Input: A SDFG $G = \langle V, E, d, t, p, c \rangle$ and its BRV q **Output:** An EHG $G' = \langle V', E', d', t' \rangle$

```
for all nodes  $v \in V$  do
  for  $i \leftarrow 1$  to  $q_v$  do
    add a copy of  $v$  as  $v_i$  to  $V'$  /* Create all needed copies of each node */
     $t'(v_i) \leftarrow t(v)$ 
  end for
end for
for all edges  $e = (u, v) \in E$  do
   $N \leftarrow d(e) \operatorname{div} c(e) + 1$ 
   $j \leftarrow (N - 1) \operatorname{mod} q_v + 1$  /* Find first sink node needing data */
   $no(v) \leftarrow N \cdot c(e) - d(e)$  /* How much new data this node needs */
  for  $i \leftarrow 1$  to  $q_u$  do
     $no(u) \leftarrow p(e)$  /* How much data can be given to this node */
    while  $no(u) \neq 0$  do
       $amt(u, v) \leftarrow \min\{no(u), no(v)\}$  /* While copy of source has data to give out loop */
      for  $k \leftarrow 1$  to  $amt(u, v)$  do
        add an edge  $e = (u_i, v_j)$  to  $E'$  with  $d'(e) \leftarrow (N - 1) \operatorname{div} q_v$ 
      end for
       $no(v) \leftarrow no(v) - amt(u, v)$ 
       $no(u) \leftarrow no(u) - amt(u, v)$ 
      if  $no(v) = 0$  then
         $N \leftarrow N + 1$  /* If copy of sink has enough data move to next copy */
         $j \leftarrow (N - 1) \operatorname{mod} q_v + 1$ 
         $no(v) \leftarrow c(e)$ 
      end if
    end while
  end for
end for
```

Similar problems arise when we attempt to minimize the clock period. We will say that an *iteration* of a SDFG is the execution of all nodes of its EHG once. The average computation time of one iteration is then called the *iteration period* of the SDFG and is equal to the iteration period of the EHG. (In Figure 4(a) the iteration period is also 4.) If a SDFG contains a loop, then the iteration period is bounded from below by the *iteration bound* [19], which is defined to be the maximum time-to-delay ratio of all cycles in the EHG. For example, the EHG in Figure 6 contains three loops: (A_1, B, C) and (A_2, B, C) each with total computation time of 4 and delay count 2; and (B, C) with computation time 2 and delay count 1. Thus the iteration period of the graph in Figure 4(a) is 2. This can be clearly seen from the schedule in Figure 5(a), where overlapped iterations create higher throughput. (The iteration period of an SDFG can be overestimated using the ideas from [22] without constructing the EHG, but our method yields a tighter bound, which is important as we attempt to minimize the iteration period of an SDFG next.)

A *retiming* $r : V \rightarrow \mathbb{N} \cup \{0\}$ is a function which specifies a transformation of a graph G . It labels each vertex with a number of delays to be removed from each incoming edge and placed on each outgoing edge, changing G into the retimed graph $G_r = \langle V, E, d_r, t, p, c \rangle$ where $d_r(e) = d(e) + p(e)r(u) - c(e)r(v)$ for each edge $e = (u, v)$ in E [24, 26]. As an example, a retiming with $r(A) = 2$ and $r(B) = r(C) = 0$ transforms

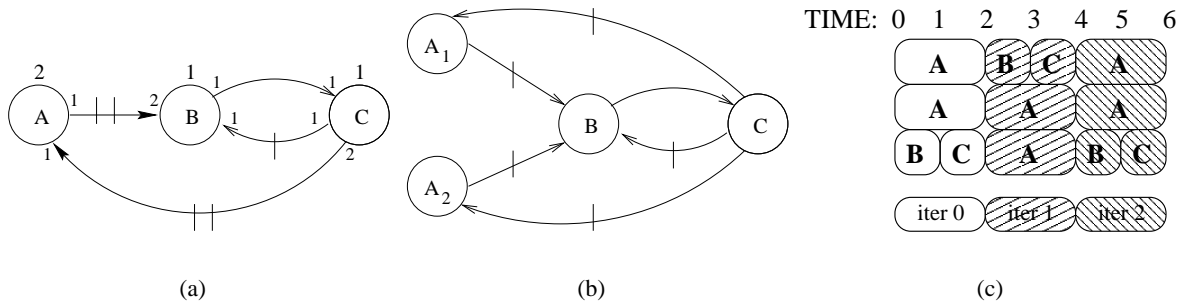


Figure 8: (a) Figure 4(a) retimed; (b) Its EHG; (c) Its repeating schedule

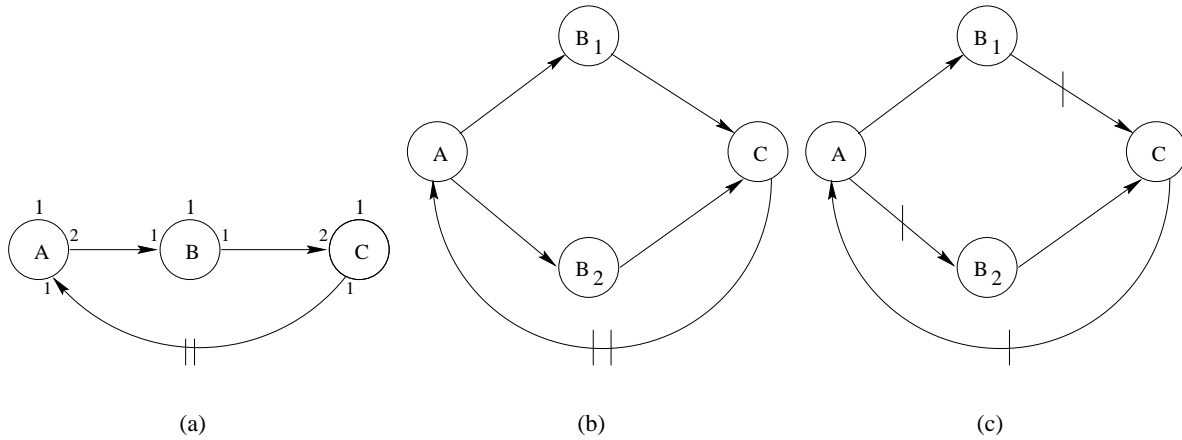


Figure 9: (a) A unit-time SDFG; (b) Its EHG; (c) Its retimed EHG

the SDFG of Figure 4(a) into that of Figure 8(a). Examining the EHG in Figure 8(b), we see that we have now achieved an optimal clock period of 2 which translates into the more compact schedule of Figure 8(c).

3.2 Problems Retiming EHG

On first glance, it appears that we should just be able to retime the EHG via traditional methods and then map back to the original SDFG, as was done by Lee originally [10]. Unfortunately, the initial translation from SDFG to EHG is too complex to permit this. As an example, consider the unit-time SDFG given in Figure 9(a), with its EHG appearing in Figure 9(b). A retiming with $r(A) = r(B_1) = 1$ and $r(B_2) = r(C) = 0$ transforms the EHG into the graph shown in Figure 9(c) with clock period 2. We now wish to try and match this with some retimed version of the original SDFG, but have a problem with the delay count of the edge between A and B. If the new delay count is 1, the EHG should have no delay on the edge (A, B_2) and one delay on (A, B_1) , exactly the opposite of what we actually have. On the other hand, if the retimed delay count is 2 or more, then both (A, B_1) and (A, B_2) should have non-zero delay counts, which also contradicts what we have. In any case, there can be no direct matching in this case. If we are to retime SDFGs, we must work directly on the original graph itself.

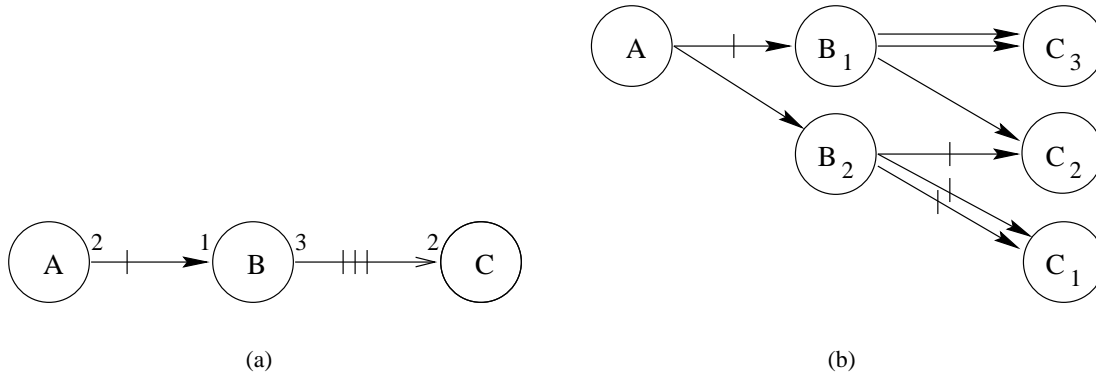


Figure 10: (a) A path in a SDFG; (b) Its homogeneous equivalent

4 Retiming a SDFG

Since we cannot retime a SDFG by working with its EHG, we must develop methods for retiming the SDFG directly. In this section we refine the methods of [16] to deal with this situation.

4.1 Initial Problems

Unfortunately, the retiming algorithms we will propose will either be pessimistic or expensive. The reason for this is that the original methods we are using as a basis were themselves built on one result from [16]:

Theorem 4.1 *Let G be a HDFG and c a potential clock period. $cl(G) \leq c$ if and only if every path in G with total computation time larger than c has delay count larger than 1.*

The problem now is that insufficient delays along a path in a SDFG do not necessarily translate into a zero-delay path in the EHG. As an example, consider the unit-time SDFG in Figure 10(a) below, with its EHG given in Figure 10(b). For $c = 2$, examining only the original SDFG would lead us to retime this path even though such an exercise is unnecessary. To avoid such false paths, we may need to construct intermediate EHGs for study, a very costly process.

In a similar vein, the nature of an EHG raises the question of what a path actually is. The traditional definition says that a path is a sequence of nodes and edges. Since we now have multiple edges between nodes, we must be very careful to consider all paths resulting from such multiple copies. To illustrate, the traditional definition would dictate that there is one path from A to C_3 in Figure 10(b). Because of the pair of edges between B_1 and C_3 , we will abuse our definition slightly and say that there are in fact two paths between A and C_3 in the EHG when we do our calculations below. While this makes sense, it is somewhat different from what has always been done and so must be noted.

Another additional cost that the problem of insufficient delays forces us to pay comes in the form of additional checks for legality. In the original algorithms from [16], only one delay at a time was moved, a stipulation which did not cause the proposed retiming to become illegal at any intermediate step (as proven in [16]). Because we are now pulling groups of delays through nodes, this situation no longer exists, and so we will have to check for legality at every stage of an algorithm.

The question now is to determine exactly how many delays to view as sufficient. Let $e : u \rightarrow v$ be an edge in a SDFG. Each copy of u in the EHG creates $p(u)$ tokens. By our construction each of these is to travel along a separate edge. Since there are q_u copies of node u in the EHG, there must then be a total of $q_u \cdot p(e)$ edges to carry all of the data, each of which we expect to require a delay when we retime the graph. Similarly, q_v copies of v are each receiving $c(e)$ tokens, and so there must be $q_v \cdot c(e)$ edges for these data. We will use either of these figures as the number of tokens required by an edge in the SDFG during retiming.

4.2 Retiming Algorithm

We will seek our retiming via *relaxation* on the edges of our graph. We do this by sorting our vertices and then sweeping along the sorted list. When we get to a point where the current path is too long, we insert enough delays so as to break the path up into sufficiently small pieces. We then verify that we are allowed to do this. If we can't then there is no retiming and we return with an error; otherwise we sweep further. Once our prospective retiming has been found, we test the retimed graph to make sure that the clock period is within our requirements. If it is we've found a way to retime the SDFG; otherwise there is no such retiming.

We begin our construction by considering Algorithm 2 below, the $O(|E|)$ -time algorithm from [16] for finding the length of the longest zero-delay path into each vertex of a HDFG. This procedure first sorts the vertices so that those occurring early in the list are connected to vertices later in the list by zero-delay edges. It then traces through the list assigning each vertex the length of its longest zero-delay path. If a vertex is not connected to a previous one, its path length must equal its own computation time; otherwise its path length equals its own time, plus the sum of the times of all the other vertices found along the path to this point. We require this algorithm not just for constructing our retiming, but also for verifying that our final retimed graph executes within the required time frame.

Algorithm 2 Find computation time of most expensive zero-delay path to all vertices

Input: A HDFG $G = \langle V, E, d, t \rangle$

Output: The $|V|$ -length vector τ

Topologically sort the vertices of G , with u preceding v if there is a zero-delay edge from u to v in G

for all v in order from the sorted list **do**

if v has no zero-delay incoming edge in G **then**

$\tau(v) \leftarrow t(v)$

else

$\tau(v) \leftarrow t(v) + \max\{\tau(u) \mid \exists e : u \rightarrow v \text{ in } G \text{ with } d(e) = 0\}$

end if

end for

return τ

With this in hand we can now proceed to our primary method, given as Algorithm 3 below. We begin by retiming our SDFG with the result to date and constructing its EHG. The EHG is then handed to Algorithm 2 to find the lengths of the maximum paths to all vertices. At this point the vertices in our SDFG fall into one of two groups. If all copies of a vertex in the EHG are isolated (i.e., connected to the rest of the graph only by edges containing delays), we don't wish to retime the node and remove it from consideration. Otherwise the node lies along some zero-delay path and we may have to retime it. In this case we assign it a longest path length equal to the longest path length of any of its copies in the EHG.

At this point we consider nodes for retiming. Since we want to push delays forward along our paths (rather than pulling them backward as was done in [16]), we retime those nodes which occur early in a path. This

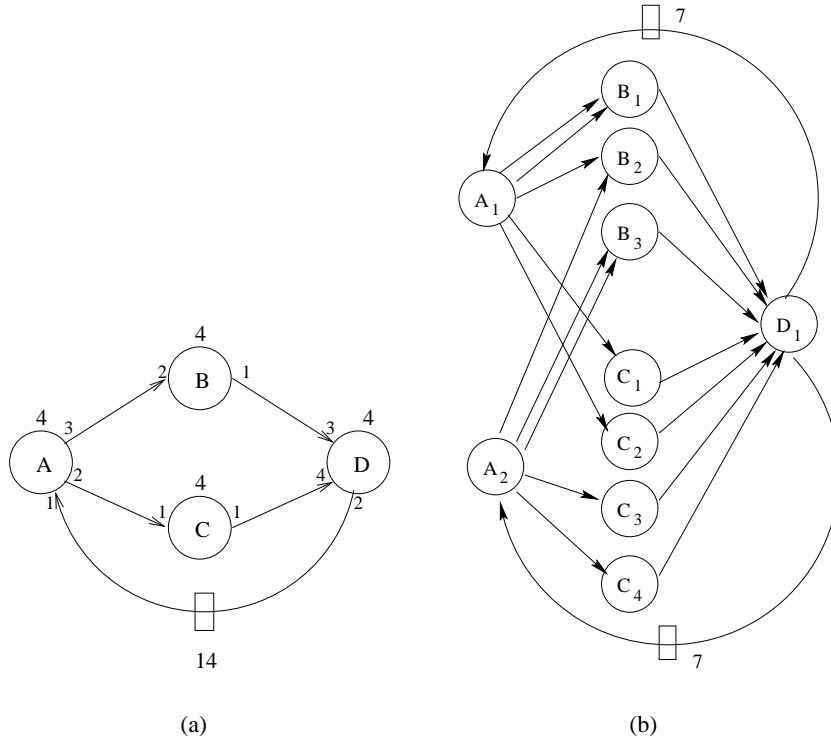


Figure 11: (a) An example SDFG; (b) Its EHG

process is complicated by the different rates of production and consumption on each node. For example, for each delay drawn into node A in Figure 11(a), three delays are pushed onto the edge from A to B and two delays onto the edge from A to C . Therefore, for each outgoing edge from such a node, we calculate the number of delays needed to retime all copies of the edge in the EHG, subtract the number of delays currently on the edge, adjust for the different rates of production and consumption, and retime by the maximum of these needs. Once all nodes are retimed, we test the prospective retiming for legality, i.e. we check that retiming by our function doesn't result in some edge containing a negative number of delays. If we pass this test, we look further along our path for other nodes in need of retiming.

Once we have checked all nodes at least once and have derived a legal retiming function, it is time to test our answer. We retime the SDFG, construct its EHG, and pass this to Algorithm 2 to find the maximum zero-delay paths to each node. Since the length of the largest zero-delay path in the EHG equals our clock period, this length is tested against our requested clock period. If the length is still too large we cannot retime this SDFG to execute in the time we wish and must return with an error. Otherwise we have found our retiming.

We now demonstrate our method by executing it on the SDFG of Figure 11(a) with $c = 4$. Sorting the vertices of Figure 11(b), computing longest path lengths and taking the maxima reveals that $\Upsilon(A) = 4$, $\Upsilon(B) = \Upsilon(C) = 8$ and $\Upsilon(D) = 12$ in this case. We thus only retime node A at this step. Since there are no delays on the edge (A, B) , our initial retiming has $r(A) = q_A = 2$ and $r(v) = 0$ for any other node v . Pulling 2 delays through A pushes 6 delays onto (A, B) and 4 onto (A, C) , as seen in the retimed graph in Figure 12(a), with its EHG appearing in Figure 12(b).

Looping back around, we see from Figure 12(b) that only node A is cut off; all other nodes lie along some zero-delay path. Thus $\Upsilon(A) = \infty$, $\Upsilon(B) = \Upsilon(C) = 4$ and $\Upsilon(D) = 8$ now, calling for us to retime nodes B

Algorithm 3 Retime a SDFG via Relaxation

Input: A SDFG $G = \langle V, E, d, t, p, c \rangle$, a potential clock period c

Output: A retiming r such that $cl(G_r) \leq c$ if one exists

```
for all  $v \in V$  do
   $r(v) \leftarrow 0$ 
end for
for  $i \leftarrow 1$  to  $|V|$  do
  Construct the retimed graph  $G_r$  given the current  $r$ 
  Construct the EHG  $H$  for  $G_r$ 
   $\tau \leftarrow \text{MaxPath}(H)$  /* Apply Algorithm 2 */
  for all  $v$  in  $V$  do
    if no copy of  $v$  in  $H$  is incident on a zero-delay edge in  $H$  then
       $\Upsilon(v) \leftarrow \infty$ 
    else
       $\Upsilon(v) \leftarrow \max\{\tau(v_i) \mid v_i \text{ is a copy of } v \text{ in } H\}$ 
    end if
  end for
  for all  $v$  with  $\Upsilon(v) \leq c$  do
     $r(v) \leftarrow r(v) + \max \left\{ \left\lceil \frac{q_v \cdot p(e) - d_r(e)}{p(e)} \right\rceil \mid e : v \rightarrow u \text{ in } G_r \text{ for some } u \right\}$ 
  end for
  for all  $e : u \rightarrow v$  in  $E$  do
    if  $d(e) + p(e)r(u) - c(e)r(v) < 0$  then
      return FALSE /* Retiming illegal; return with error */
    end if
  end for
end for
Construct the retimed graph  $G_r$  given the current  $r$  /* Determine clock period */
Construct the EHG  $H$  for  $G_r$ 
 $\Upsilon \leftarrow \text{MaxPath}(H)$  /* Apply Algorithm 2 again */
if  $\max\{\Upsilon(v) \mid v \in V\} > c$  then
  return FALSE /* No feasible retiming */
else
  return  $r$  /* Otherwise  $r$  is the retiming */
end if
```

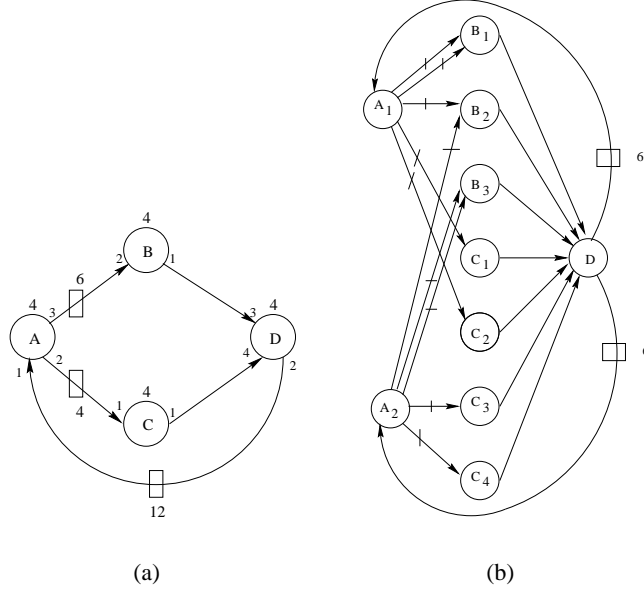


Figure 12: (a) The retimed SDFG after one pass of algorithm; (b) Its EHG

and C . Since neither of the edges (B, D) nor (C, D) currently contains delays, our retiming now has $r(A) = 2$, $r(B) = q_B = 3$, $r(C) = q_C = 4$ and $r(D) = 0$. The new retimed graph is given below as Figure 13(a), with its EHG in Figure 13(b). Note that, due to node B 's consumption rate of 2, a retiming of 3 applied to B requires us to pull $2 \times 3 = 6$ delays in so that the proper number of delays is pushed back out.

Studying the new EHG shows us that node D is now cut off, but node A requires further retiming. We have $\Upsilon(A) = 4$, $\Upsilon(B) = \Upsilon(C) = 8$ and $\Upsilon(D) = \infty$ now, so only node A must be retimed. Since both of the edges (A, B) and (A, C) are devoid of delays, we must add $q_A = 2$ onto the retiming for A , giving us the function $r(A) = 4$, $r(B) = 3$, $r(C) = 4$ and $r(D) = 0$. The application of this retiming to the original SDFG results in the graph of Figure 14(a) and we have found our answer.

However we have a final pass of the algorithm to perform. This time, all nodes in the SDFG have been isolated, so $\Upsilon(v) = \infty$ for all nodes v , insuring that no further retiming takes place. We now study the EHG of Figure 14(b), find that the maximum zero-delay path is an individual node with computation time 4 and conclude that we have found our retiming.

Let $G = \langle V, E, d, t, p, c \rangle$ be our SDFG with EHG $H = \langle V', E', d', t' \rangle$. Since Algorithms 1 and 2 each execute in $O(|E'|)$ time, Algorithm 3 only requires $O(|V|^2 + |V||E'|)$ time. However, while we suspect that its success is both a necessary and sufficient condition for a SDFG to be retimable to a given clock period, it is unknown whether or not this is the case. In our defense, the algorithm from [16] upon which this method is based was also never proven both necessary and sufficient, but has been extremely useful in practice. We suspect that the algorithm we've described here will prove just as valuable despite this logical gap.

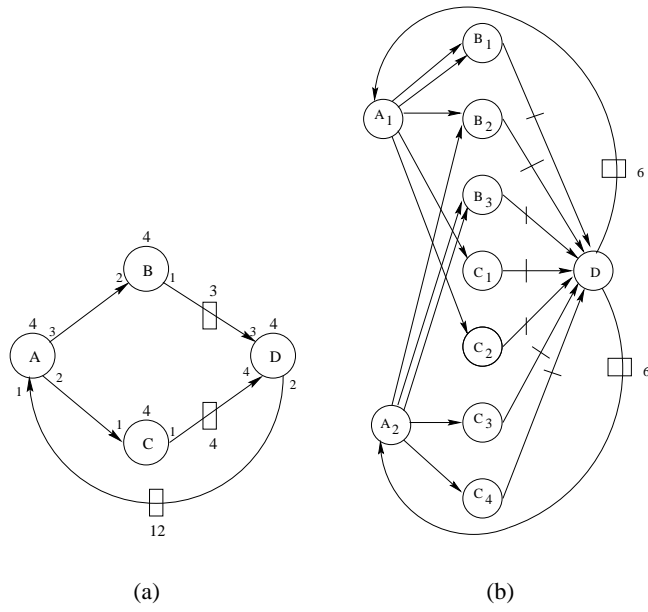


Figure 13: (a) The retimed SDFG after two passes of algorithm; (b) Its EHG

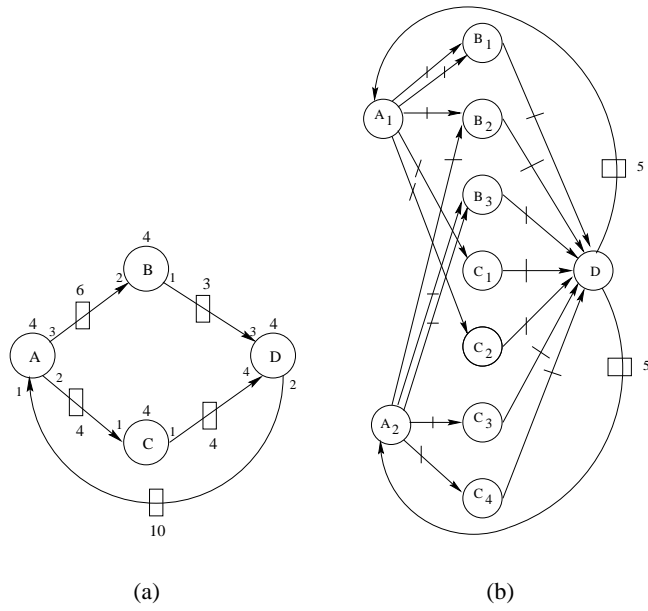


Figure 14: (a) Figure 11(a) retimed; (b) Its EHG

5 A Simple Example

To illustrate our method further, consider the SDFG in Figure 15, a variation on the example from [17] with a BRV of $q = [2 \ 1 \ 2]^T$. In our example, nodes A and B take 1 time unit to execute and C takes 2. We will attempt to retim it to have a clock period of 2. Our algorithm requires three passes to complete. At the outset, we compute the longest path lengths and find that $\Upsilon(A) = 1$, $\Upsilon(B) = 2$ and $\Upsilon(C) = 4$. Hence nodes A and B require retiming. Node A is source node for only one edge with production rate 1; thus $r(A) = q_A = 2$. On the other hand, two edges emanate from B : (B, A) with delay count 4 and production rate 2, and (B, C) contains no delays with production rate 4. Taking the maximum yields a value of $r(B) = q_B = 1$. Applying this retiming produces the graph of Figure 16(a).

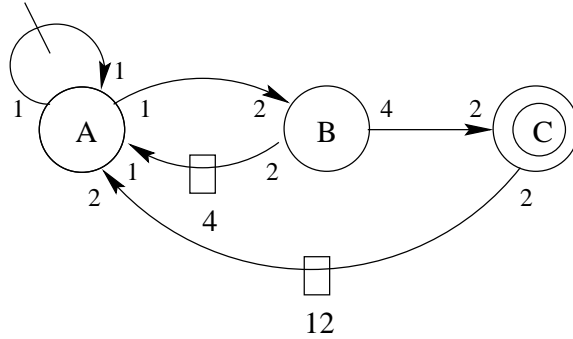


Figure 15: A sample SDFG

Looping back around, we find that node C has been cut off from the rest of the graph, but nodes A and B both need retimed again. Repeating the same line of reasoning given above leads to a value for $r(A)$ of $2 + 2 = 4$. This time, however, both edges originating at B contain delays. Applying our formula leads us to realize that $r(B)$ will not be incremented at this step. The function with $r(A) = 4$, $r(B) = 1$ and $r(C) = 0$ yields the retimed SDFG of Figure 16(b).

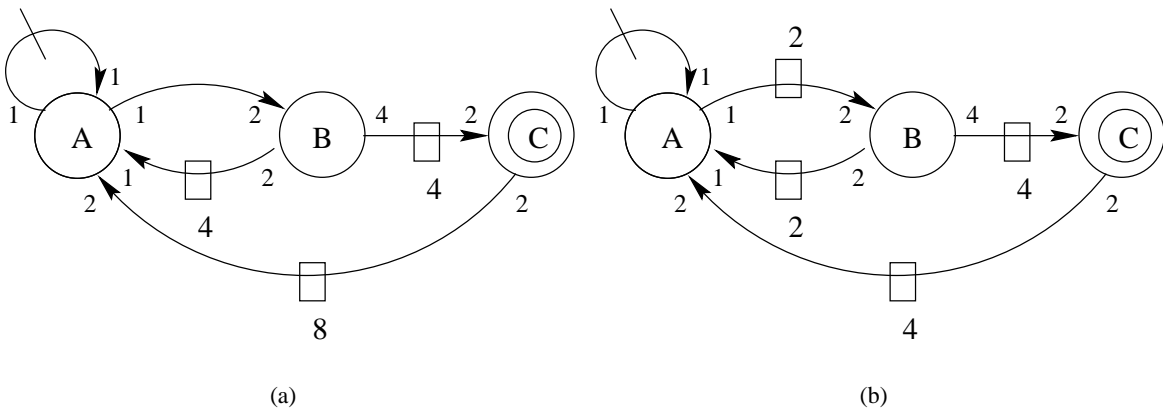


Figure 16: Figure 15 retimed: (a) after first pass of algorithm; (b) after second

We still have a pass of the algorithm to perform. This time the nodes are separated from each other by delays, and so $\Upsilon(v) = \infty$ for all nodes v . Hence no further retiming takes place, and we are finished. A check of the retimed graph reveals that all nodes are separated by delays, and since the maximum of their computation

times was 2, this is also the clock period of our retimed graph.

6 Example: A Simplified Spectrum Analyzer

Finally, let us apply our algorithm to a variation of the simplified spectrum analyzer from [24] which appears in Figure 17(a), with node descriptions in Figure 17(b). This graph has a BRV of $q = [16 \ 1 \ 1 \ 1 \ 4 \ 1]^T$, so in the interests of space we will not display the EHG at each step. Instead, we shall describe the pertinent information. It can be shown that the lower bound on the clock period for this SDFG is 3, and so we will attempt to retime it to be optimal.

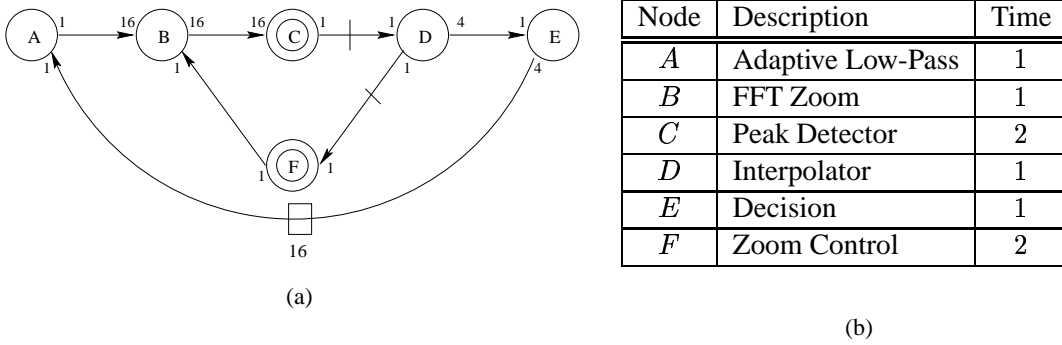


Figure 17: A simplified spectrum analyzer

At first, we see the zero-delay paths (F, B, C) and (A, B, C) which give us $\Upsilon(A) = \Upsilon(D) = 1$, $\Upsilon(B) = 3$, $\Upsilon(C) = 5$ and $\Upsilon(E) = \Upsilon(F) = 2$. Thus all nodes except C need retiming, and our formula gives an initial retiming of $r(A) = p((A, B)) = 16$, $r(B) = \frac{p((B, C))}{p((B, C))} = 1$, $r(C) = 0$, $r(D) = \max \left\{ \frac{p((D, E))}{p((D, E))}, p((D, F)) - 1 \right\} = 1$, $r(E) = \frac{q_E \cdot p((E, A)) - 16}{p((E, A))} = 0$ and $r(F) = p((F, B)) = 1$. (The value for $r(E)$ reveals a pattern to which we will refer again. If there are sufficient delays on an edge, the value of the retiming for the edge's source node will not change.) Applying this retiming to the graph in Figure 17(a) yields the graph of Figure 18(a). If our algorithm were rewritten so that this graph were checked, we would find that just this initial step has resulted in an optimal retimed graph.

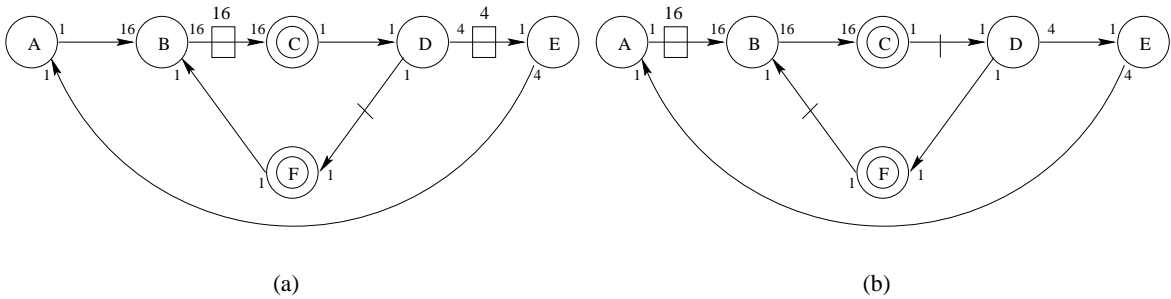


Figure 18: The analyzer: (a) after first pass; (b) after second pass

In the second pass, we see the zero-delay paths (B, C) , (D, F) and (D, E, A) , yielding longest path lengths

of $\Upsilon(A) = \Upsilon(C) = \Upsilon(F) = 2$, $\Upsilon(B) = \Upsilon(D) = 3$ and $\Upsilon(E) = 1$. Our algorithm requires us to retime all nodes at this step. The delay count on (A, B) has not changed, and so $r(A) = 16 + 16 = 32$ since the calculation performed above is simply repeated and added on; similarly $r(F) = 1 + 1 = 2$. For the remaining nodes, $r(B) = 1 + \frac{p((B,C))-16}{p((B,C))} = 1$, $r(C) = r(D) = 1$ and $r(E) = q_E = 4$. This function results in the retimed graph of Figure 18(b), which also has an optimal clock period.

The next pass of the algorithm finds longest path lengths of $\Upsilon(A) = \Upsilon(C) = \Upsilon(F) = 3$, $\Upsilon(B) = \Upsilon(D) = 1$ and $\Upsilon(E) = 2$. Again all nodes need retimed. The delays on (A, B) , (C, D) and (F, B) insure that the retimings for A , C and F don't change. Repeating above calculations causes the retimings of the other nodes to double in size, giving a function with $r(A) = 32$, $r(B) = r(D) = r(F) = 2$, $r(C) = 1$ and $r(E) = 8$. Applying this new function to our original graph yields Figure 19(a). We can see that the shortest zero-delay path to E has length 4, so this graph is not optimized.

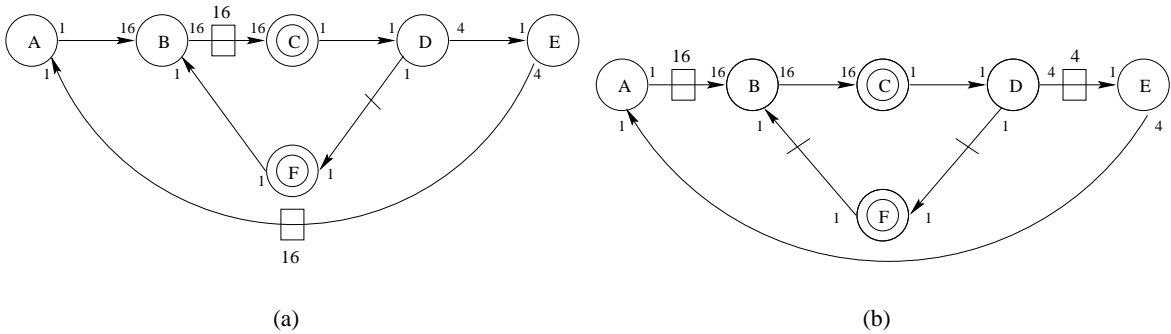


Figure 19: The analyzer: (a) after third pass; (b) after fourth pass

Looping back around, we find that all nodes but E need retimed based on lengths of shortest zero-delay paths. Furthermore, the delays on (B, C) insure that $r(B)$ remains fixed; the other values increase by one “unit” based on our previous calculations. Hence we now have $r(A) = 32 + 16 = 48$, $r(B) = 2$, $r(C) = 1 + 1 = 2$, $r(D) = 2 + 1 = 3$, $r(E) = 8$ and $r(F) = 3$. Applying this function produces the retimed graph in Figure 19(b), which is not optimal due to the zero-delay path (B, C, D) .

In the fifth pass, only nodes D and F do not require retiming, since D lies at the end of a zero-delay path of large enough size and delays have severed F from the rest of the retimed SDFG. We also do not retime A because of the delays on (A, B) , so our retiming function increases to $r(A) = 48$, $r(B) = r(C) = r(D) = r(F) = 3$ and $r(E) = 12$. Applying this retiming has no effect; the graph in Figure 17(a) is retimed right back to its original form.

Our sixth and final pass finds us back where we started. All but node C are retimed, and the delays on (E, A) insure that $r(E)$ isn't changed. Incrementing the others yields a final function with $r(A) = 64$, $r(B) = r(D) = r(F) = 4$, $r(C) = 3$ and $r(D) = 12$. Applying this function gives us Figure 18(a), the optimized graph we found at our second step, and we have finished.

7 Conclusion

In this paper, we have established a notation for expressing and studying synchronous data-flow graphs. We have presented the difficulties involved with retiming SDFGs, and then constructed a polynomial-time algorithm for retiming a synchronous graph so that it achieves a sufficiently small clock period. Finally, we have

demonstrated the effectiveness of our algorithm on several examples.

Regardless of how good our algorithm may be, it is still not proven to represent both a necessary and sufficient condition for retiming. This proof, or the construction of an alternate method which is necessary and sufficient, remain interesting open problems. Correcting the errors in [24] will definitely lead to greater understanding of our model and may open the door to removing this logical gap. It may also lead to a study of retiming applied to even more complicated models, such as cyclo-static or dynamic DFGs [1].

References

- [1] G. Bilsen, M. Engels, R. Lauwereins, and J. Peperstraete. Cyclo-static dataflow. *IEEE Transactions on Signal Processing*, 44:397–408, 1996.
- [2] L.-F. Chao. *Scheduling and Behavioral Transformations for Parallel Systems*. PhD thesis, Dept. of Computer Science, Princeton University, 1993.
- [3] L.-F. Chao and E. H.-M. Sha. Retiming and unfolding data-flow graphs. In *Proceedings of the International Conference on Parallel Processing*, pages II 33–40, 1992.
- [4] L.-F. Chao and E. H.-M. Sha. Static scheduling for synthesis of DSP algorithms on various models. *Journal of VLSI Signal Processing*, 10:207–223, 1995.
- [5] L.-F. Chao and E. H.-M. Sha. Scheduling data-flow graphs via retiming and unfolding. *IEEE Transactions on Parallel and Distributed Systems*, 8:1259–1267, 1997.
- [6] F. Commoner, A.W. Holt, S. Even, and A. Pnueli. Marked directed graphs. *Journal of Computer and System Sciences*, 5:511–523, 1971.
- [7] G. Gao, R. Govindarajan, and P. Panangaden. Well-behaved dataflow programs for DSP computation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 5, pages 561–564, 1992.
- [8] P.D. Hoang and J.M. Rabaey. Scheduling of DSP programs onto multiprocessors for maximum throughput. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 41:2225–2235, 1993.
- [9] R. Lauwereins, M. Engels, J.A. Peperstraete, E. Steegmans, and J. Van Ginderdeuren. GRAPE: A CASE tool for digital signal parallel processing. *IEEE ASSP Magazine*, 7:32–43, 1990.
- [10] E.A. Lee. *A Coupled Hardware and Software Architecture for Programmable Digital Signal Processors*. PhD thesis, Dept. of EECS, Univ. of California at Berkeley, 1986.
- [11] E.A. Lee, W.H. Ho, E. Goei, J. Bier, and S.S. Bhattacharya. Gabriel: A design environment for DSP. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 37:1751–1762, 1989.
- [12] E.A. Lee and D.G. Messerschmitt. Pipeline interleaved programmable DSP's: Synchronous data flow programming. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-35:1334–1345, 1987.
- [13] E.A. Lee and D.G. Messerschmitt. Static scheduling of synchronous data-flow programs for digital signal processing. *IEEE Transactions on Computers*, 36:24–35, 1987.

- [14] E.A. Lee and D.G. Messerschmitt. Synchronous data flow. *Proceedings of the IEEE*, 75:1235–1245, 1987.
- [15] C.E. Leiserson and J.B. Saxe. Optimizing synchronous systems. *Journal of VLSI and Computer Systems*, 1(1):41–67, 1983.
- [16] C.E. Leiserson and J.B. Saxe. Retiming synchronous circuitry. *Algorithmica*, 6:5–35, 1991.
- [17] K.K. Parhi. Algorithm transformation techniques for concurrent processors. *Proceedings of the IEEE*, 77:1879–1895, 1989.
- [18] J.L. Pino, S. Ha, E.A. Lee, and J.T. Buck. Software synthesis for DSP using Ptolemy. *Journal of VLSI Signal Processing*, 9:7–21, 1995.
- [19] M. Renfors and Y. Neuvo. The maximum sampling rate of digital filters under hardware speed. *Transactions on Circuits and Sampling*, CAS-28:196–202, 1981.
- [20] S. Ritz, M. Pankert, V. Zivojnovic, and H. Meyr. High-level software synthesis for the design of communication systems. *IEEE Journal on Selected Areas in Communications, Special Issue on Computer-Aided Modeling, Analysis and Design of Communication Links*, 11:348–358, 1993.
- [21] S. Ritz, M. Pankert, V. Zivojnovic, and H. Meyr. Optimum vectorization of scalable synchronous dataflow graphs. In *Proceedings of the International Conference on Application-Specific Array Processors*, pages 285–296, 1993.
- [22] R. Schoenen, V. Zivojnovic, and H. Meyr. An upper bound of the throughput of multirate multiprocessor schedules. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 655–658, 1997.
- [23] M. Veiga, J. Parera, and J. Santos. Programming DSP systems on multiprocessor architectures. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 2, pages 965–968, 1990.
- [24] V. Zivojnovic, S. Ritz, and H. Meyr. Optimizing DSP programs under the multirate retiming transformation. In *Proceedings of the 7th European Signal Processing Conference*, volume 3, pages 1597–1600, 1994.
- [25] V. Zivojnovic, S. Ritz, and H. Meyr. Retiming of DSP programs for optimum vectorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 465–468, 1994.
- [26] V. Zivojnovic and R. Schoenen. On retiming of multirate DSP algorithms. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 6, pages 3310–3313, 1996.