Technical Report #UMIACS-TR-2000-09, Institute for Advanced Computer Studies, University of Maryland at College Park, March, 2000.

Contention-conscious Transaction Ordering in Embedded Multiprocessors Systems[1]

Mukul Khandelia, and Shuvra S. Bhattacharyya

Department of Electrical and Computer Engineering, and
Institute for Advanced Computer Studies
University of Maryland, College Park

Abstract

This paper explores the problem of efficiently ordering interprocessor communication operations in statically-scheduled multiprocessors for iterative dataflow graphs. In most digital signal processing applications, the throughput of the system is significantly affected by communication costs. By explicitly modeling these costs within an effective graph-theoretic analysis framework, we show that ordered transaction schedules can significantly outperform self-timed schedules even when synchronization costs are low. However, we also show that when communication latencies are non-negligible, finding an optimal transaction order given a static schedule is an NP-complete problem, and that this intractability holds both under iterative *and* non-iterative execution. We develop new heuristics for finding efficient transaction orders, and perform an experimental comparison to gauge the performance of these heuristics.

1. Background

This paper explores the problem of efficiently ordering interprocessor communication (IPC) operations in statically-scheduled multiprocessors for iterative dataflow specifications. An iterative dataflow specification consists of a dataflow representation of the body of a loop that is

[1] This research was sponsored by the US National Science Foundation (CAREER, MIP9734275); and the Army Research Laboratory under Contract No. DAAL01-98-K-0075, and the MICRA program.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE **MAR 2000**	2. REPORT TYPE	3. DATES COVERED **00-00-2000 to 00-00-2000**

4. TITLE AND SUBTITLE
Contention-conscious Transaction Ordering in Embedded Multiprocessors Systems

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Maryland,Department of Electrical and Computer Engineering,Institute for Advanced Computer Studies,College Park,MD,20742

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES **34**	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**			

to be iterated indefinitely. Dataflow programming in this form is used widely in the design and implementation of digital signal processing (DSP) systems.

In this paper, we assume that we are given a dataflow specification of an application, and an associated multiprocessor schedule (e.g., derived from scheduling techniques such as those presented in [6, 9, 18, 22]). Our objective is to reduce the overall IPC cost of the multiprocessor implementation, and the associated performance degradation, since IPC operations result in significant execution time and power consumption penalties, and are difficult to optimize thoroughly during the scheduling stage. IPC is assumed to take place through shared memory, which could be global memory between all processors, or could be distributed between pairs of processors (e.g., hardware first-in-first-out queues or dual ported memory). Such simple communication mechanisms, as opposed to cross bars and elaborate interconnection networks, are common in embedded systems, due to their simplicity and low cost.

1.1 Scheduling dataflow graphs

Our study of multiprocessor implementation strategies in this paper is in the context of *homogeneous synchronous dataflow* (HSDF) specifications. In HSDF, an application is represented as a directed graph in which vertices (*actors*) represent computational tasks of arbitrary complexity; edges (*arcs*) specify data dependencies; and the number of data values (*tokens*) produced and consumed by each actor is fixed. An actor executes (*fires*) when it has enough tokens on its input arcs, and during execution, it produces tokens on its output arcs. HSDF imposes the restriction that on each invocation, each actor consumes exactly one token from each input arc, and produces one token on each output arc. HSDF and closely-related models are used extensively for multiprocessor implementation of embedded signal processing systems (e.g., see [6, 10, 11, 12]). We refer to an HSDF representation of an application as an *application graph*.

For multiprocessor implementation of dataflow graphs, actors in the graph need to be scheduled. Scheduling can be divided into three steps [13] — assigning actors to processors (*processor assignment*), ordering the actors assigned to each processor (*actor ordering*), and determining when each actor should commence execution. All of these tasks can either be performed at run-time or at compile time to give us different scheduling strategies. To reduce run-time overhead and improve predictability, it is often desirable in embedded applications to carry out as many of these steps possible at compile time [13].

Typically, there is limited information available at compile time since the execution times of the actors are often *estimated values*. These may be different from the actual execution times due to actors that display run-time variation in their execution times because of conditionals or data-dependent loops within them, for example. However, in a number of important embedded domains, such as DSP, it is widely accepted that execution time estimates are reasonably accurate, and that good compile-time decisions can be based on them. In this paper, we focus on scheduling methods that extensively make use of execution time estimates, and perform the first two steps — processor assignment and actor ordering — at compile time.

In relation to the scheduling taxonomy of Lee and Ha [13], there are three general strategies with which we are primarily concerned in this paper. In the *fully-static* (*FS*) strategy, all three scheduling steps are carried out at compile time, including the determination of an exact firing time for each actor. In the *self-timed* (*ST*) strategy, on the other hand, processor assignment and actor ordering are performed at compile time, but run-time synchronization is used to determine actor firing times: an ST schedule executes by firing each actor invocation A as soon as it can be determined via synchronization that the actor invocations on which A is dependent have all completed execution.

The FS and ST methods represent two extremes in the class of scheduling algorithms considered in this paper. The ST method is the least constrained scheme since the only constraints are the IPC dependencies, and it is tolerant of variations in execution times, while the FS strategy only works when tight worst case execution times are available, and forces system performance to conform to the available worst case bounds. When we ignore IPC costs, the ST schedule consequently gives us a lower bound on the average iteration period of the schedule since it executes in an ASAP (as soon as possible) manner.

The *ordered transaction* (*OT*) method [11, 23] falls in-between these two strategies. It is similar to the ST method but also adds the constraint that a linear ordering of the communication actors is determined at compile time, and enforced at run-time. The linear ordering imposed is called the *transaction order* of the associated multiprocessor implementation.

The FS and OT strategies have significantly lower overall IPC cost since all of the sequencing decisions associated with communication are made at compile time. The ST method, on the other hand, requires more IPC cost since it requires synchronization checks to guarantee the fidelity of each communication operation — that is, to guarantee that buffer underflow and overflow are consistently avoided. Significant compile-time analysis can be performed to streamline this synchronization functionality [3, 4].

The metric of interest to us in this paper is the *average iteration period* T. Intuitively, in an iterative execution of a dataflow graph, the iteration period is the number of cycles that it takes for each of the actors in the graph to execute exactly once — i.e., to complete a single graph iteration. Note that it is not necessary in a self-timed schedule for the iteration period to be the same from one graph iteration to the next, even when actor execution times are fixed [24]. The inverse of the average iteration period T gives us the *throughput* T^{-1}, which is the average number of graph iterations carried out per unit time.

1.2 Terminology and notation

We denote the set of positive integers by Z^+, the set of natural numbers $\{0, 1, 2, \ldots\}$ by $\aleph$, and the number of elements in a finite set S by $|S|$.

With each actor $v \in V$ in an HSDF specification (V, E), we associate an integer $exec(v)$, which denotes the execution time estimate of v, and an integer $proc(v)$, which denotes the processor that v is assigned to in the assignment step. Each edge $(v_i, v_j) \in E$ has a non-negative integer *delay* associated with it, which is denoted by $delay(v_i, v_j)$. These delays represent initial tokens, and specify dependencies between iterations of actors in iterative execution. For example, if the tokens produced by an actor v_i on its kth invocation are consumed by actor v_j on its $(k+2)$th invocation, the edge between v_i and v_j would have a delay of 2.

Every edge (v_i, v_j) induces the precedence constraint

$$start(v_j, k) \geq start(v_i, k - delay(v_i, v_j)) + t(v_i), \tag{1}$$

where $start(x, k) \in Z^+$ denotes the starting time of the kth invocation of an actor x. Here, $start(v_i)$ is set to 0 for $k \leq 0$ as initial conditions.

A *path* in a directed graph (V, E) is a finite sequence $(e_1, e_2, \ldots, e_n)$, where each e_i is in E, and $snk(e_i) = src(e_{i+1})$, for $i = 1, 2, \ldots, (n-1)$. We say that the path $(e_1, e_2, \ldots, e_n)$ is *directed from* $src(e_1)$ *to* $snk(e_n)$. A path that is directed from some vertex to itself is called a *cycle*. Given a path $p = (e_1, e_2, \ldots, e_n)$, the *path delay* of p, denoted $Delay(p)$, is given by

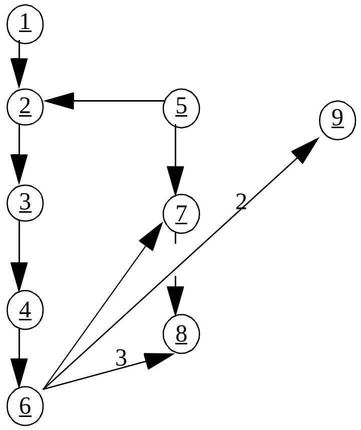
$$Delay(p) = \sum_{i=1}^{n} delay(e_i). \tag{2}$$

Each cycle c in a dataflow graph must satisfy $Delay(c) > 0$ to avoid deadlock.

The evolution of a self-timed implementation can be modeled by Sriram's *IPC graph* model [24]. Given an application graph and an associated self-timed schedule, the IPC graph, denoted G_{ipc}, is constructed by instantiating a vertex for each application graph actor, connecting an edge from each actor to the actor that succeeds it on the same processor, and adding an edge that has unit delay from the last actor on each processor to the first actor on the same processor. Also, for each application graph edge (x, y) that connects actors that execute on different processors, an *inter-processor edge* is instantiated in G_{ipc} from x to y. A sample application graph and a self-timed schedule are illustrated in Figure 1, and the corresponding IPC graph is illustrated in Figure 2.

IPC costs (estimated transmission latencies through the multiprocessor network) can be incorporated into the IPC graph model by explicitly including *communication* (*send* and *receive*) *actors*, and setting the execution times of these actors to equal the associated IPC costs.

The IPC graph is an instance of Reiter's *computation graph* model [20], also known as the *timed marked graph* model in Petri net theory [19], and from the theory of such graphs, it is well known that in the ideal case of unlimited bus bandwidth, the average iteration period for the ASAP execution of an IPC graph is given by the *maximum cycle mean* (*MCM*) of G_{ipc}, which is



Self-Timed Schedule

Proc 1: (1, 2, 3, 4, 6)

Proc 2: (5, 7, 8)

Proc 3: (9)

Figure 1. An example of an application graph and an associated self-timed schedule. The numbers on the edges $(6, 8)$ and $(6, 9)$ denote nonzero delays.

defined by

$$MCM(G_{ipc}) = \max_{\text{cycle C in } G_{ipc}} \left\{ \frac{\sum_{v \in C} exec(v)}{Delay(C)} \right\} \quad . \tag{3}$$

The quotient in (3) is referred to as the *cycle mean* of the associated cycle C.

A similar data structure that is useful in analyzing OT implementations is Sriram's *ordered transaction graph* model [24]. Given an ordering $O = \{o_1, o_2, \ldots o_p\}$ for the communication actors in an IPC graph $G_{ipc} = (V_{ipc}, E_{ipc})$, the corresponding ordered transaction graph $\Gamma(G_{ipc}, O)$ is defined as the directed graph, $G_{OT} = (V_{OT}, E_{OT})$ where $V_{OT} = V_{ipc}$, $E_{OT} = E_{ipc} \cup E_O$,

$$E_O = \{(o_p, o_1), (o_1, o_2), (o_2, o_3), \ldots, (o_{p-1}, o_p)\}, \tag{4}$$

$delay(o_i, o_{i+1}) = 0$ for $1 \le i < p$, and $delay(o_p, o_1) = 1$. Thus, an IPC graph can be modified by adding edges obtained from the ordering O to create the ordered transaction graph.

2. Previous work

In [23, 24], Sriram and Lee discuss some of the advantages and disadvantages of the OT strategy compared to the ST strategy — in particular, lower synchronization and arbitration costs for the IPC mechanism at the expense of some run-time flexibility. They also develop a method to compute an optimum transaction order when a fully-static schedule is given beforehand. In this approach, a set of inequalities is constructed using the timing information of the given FS schedule, and represented as a graph. The Bellman-Ford shortest path algorithm is applied to this graph to obtain new starting times of the actors, thereby modifying the original FS schedule. A transaction order is then obtained by sorting the starting times of the communication actors. We shall term this method of finding the transaction orders, which is an efficient polynomial-time algorithm, the *Bellman Ford Based* (*BFB*) method. Under an assumption that the cost (latency) of IPC is zero, Sriram shows that the transaction order determined by the BFB technique is always optimum.

However, in this paper, we show that when IPC costs are not negligible, as is frequently and increasingly the case in practice, the problem of determining an optimal transaction order is NP-hard. Thus, under nonzero IPC costs, we must resort to heuristics for efficient solutions. Furthermore, the polynomial-time BFB algorithm is no longer optimal, and alternative techniques that account for IPC costs are preferable.

Numerous approaches have been proposed for incorporating IPC costs into the assignment and ordering steps of scheduling (e.g., [2, 22]). The techniques that we propose in this paper are complementary to these approaches in that they provide a means for mapping the resulting schedules into efficient OT implementations, which eliminate the performance and power consumption overhead associated with run-time synchronization and contention resolution.

3. Comparison of self-timed and ordered transaction strategies

Given an application graph, an associated multiprocessor schedule, and an FS implementation, an OT implementation, and an ST implementation for the schedule, suppose T_{FS}, T_{OT}, and T_{ST}, respectively, denote the average iteration periods of the corresponding schedules. In general, when IPC costs are negligible, $T_{\mathrm{FS}} \geq T_{\mathrm{OT}} \geq T_{\mathrm{ST}}$ [24]. This is because the ST method has the least constraints. The ST schedule only has assignment and ordering constraints, while the OT schedule has transaction ordering constraints in addition to those constraints, and the FS schedule has exact timing constraints that subsume the constraints in the ST and OT schedules. ST schedules overlap in a natural manner, and eventually settle into a periodic pattern of iterations. This pattern can be exponential in size, and therefore, the ST schedule has the advantage that in successive iterations, the transaction order may be different, while this flexibility is not available for the OT and FS schedules.

In practical cases, however, the IPC cost is non-zero. Depending on the bandwidth of the bus, IPC costs may be quite significant. The throughput of the ST schedule can be computed easily when IPC costs are ignored by calculating the MCM of the corresponding dataflow graph (i.e., via (3)). However, when IPC costs are taken into account, this can no longer be done since the notion of bus contention comes into the picture. Not only do the communication actors in the dataflow graph have to wait for sufficient tokens on the input arcs to fire, they also have to wait for the bus to be available — i.e., no other communication actor should be accessing the bus at the same instant of time. Therefore, the throughput of the self-timed schedule is typically derived using simulation techniques, which are time-consuming. On the other hand, the throughput of the OT schedule can still be obtained by calculating the MCM of the transaction order graph since there will be no bus contention when a linear order is imposed on the communication actors [23].

The relation $T_{\mathrm{FS}} \geq T_{\mathrm{OT}} \geq T_{ST}$ is also no longer valid in the presence of non-zero IPC costs. To see why this is true, assume that two communication actors become *enabled* (have sufficient input tokens to fire) at more or less the same time. Then the ST method will schedule the communication actor that becomes enabled earlier. Doing this may result in a lower throughput since, for example, the processor that contains the communication actor that is scheduled later might be more heavily loaded. The FS and the OT methods avoid such pitfalls by analyzing the schedules at compile time, and producing an exact firing time assignment, or a transaction order that takes the entire schedule into consideration. Intuitively, the ST method follows a more greedy, ASAP approach in choosing which communication actor to schedule next, and this can result in inefficient execution patterns.

Example 1: To illustrate how an ST schedule might perform worse than an OT schedule, consider the IPC graph of Figure 2. Dashed edges represent inter-processor data dependencies. Numbers beside actors show their execution times, numbers beside edges indicate nonzero delays, xs_y denotes the *yth* send actor of computation actor *x*, and xr_y denotes the *yth* receive actor of *x*. Figure 3 shows the periodic pattern that the ST schedule eventually settles down into. Although Processor 1 is most heavily loaded, we see that there are instances when the processor is idling waiting for the bus to become free. In contrast, when the transaction order

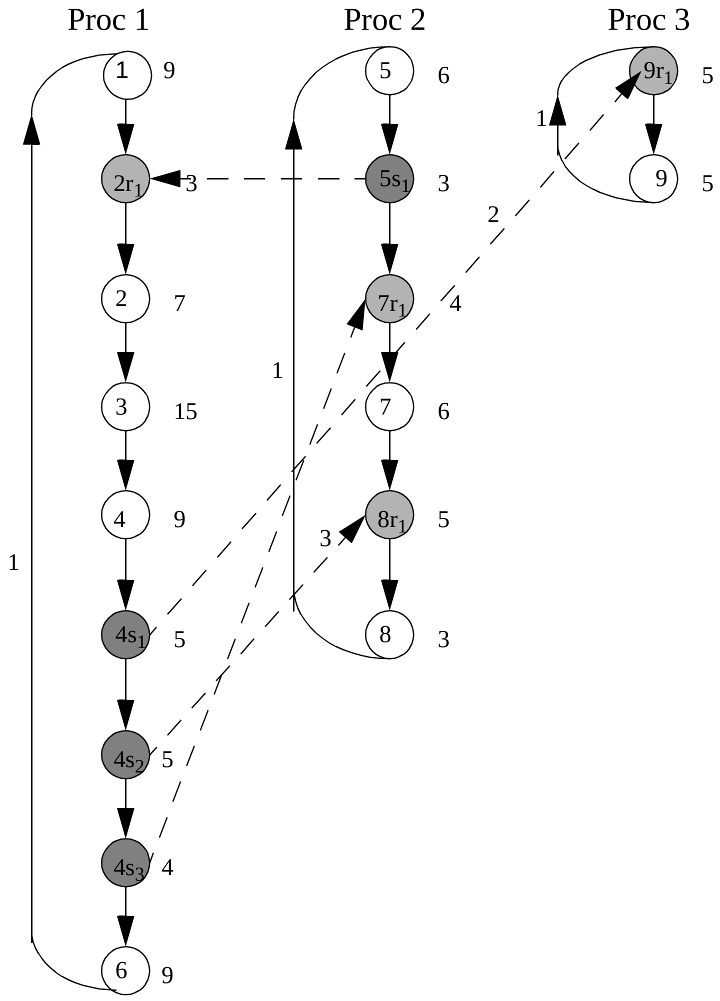


Figure 2. IPC graph constructed from application graph of Figure 1.

$(5s_1, 2r_1, 7r_1, 4s_1, 4s_2, 4s_3, 8r_1, 9r_1)$ is enforced (Figure 4), an 11% lower average iteration period results. This is because the transaction order is computed in a fashion that enables the heavily loaded Processor 1 to access the bus whenever it requires it. Such an ability to prioritize strategically-selected transactions is especially important in heterogeneous multiprocessors, which often have imbalanced loads due to large variations in processing capabilities of the computing resources.

The ST approach has the further disadvantage that in the presence of execution time uncertainties, there is no known method for computing a tight worst-case iteration period, *even using simulation techniques*. In particular, the period of the ST schedule obtained by using worst

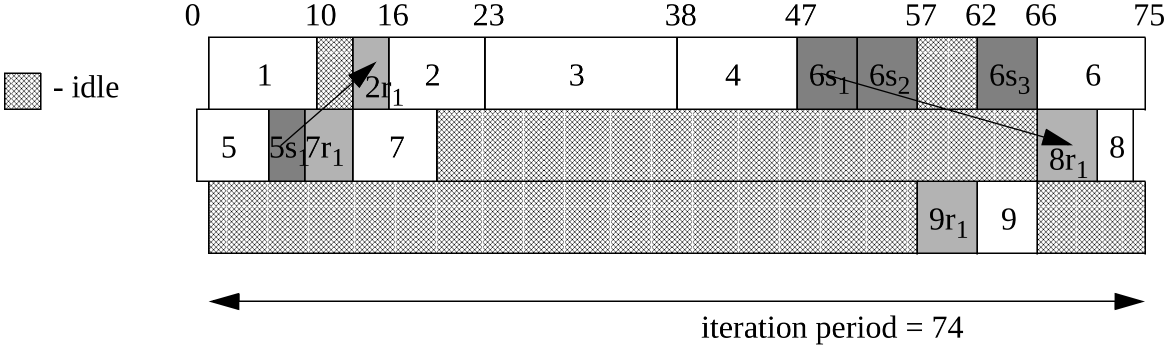


Figure 3. Gantt Chart for ST schedule in Example 1.

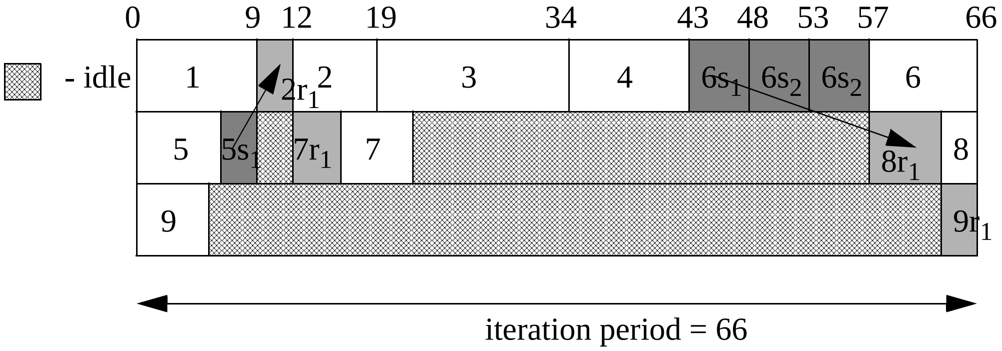


Figure 4. Gantt Chart for OT schedule in Example 1.

case execution time estimates of the actors does not necessarily give us the worst case iteration period of a schedule. This can prove to be a big disadvantage in real-time systems where worst-case bounds are needed beforehand.

Example 2: Consider the IPC graph of Figure 5, and suppose that Actor 1 has a worst-case execution time of 21, and a best case execution time of 19. Figure 6 shows the ST schedule that results when actor 1 has an execution time of 21. An iteration period of 50 is obtained. However, when the same schedule is simulated for an execution time of 19, we obtain an iteration period of 59 as shown in Figure 7 .

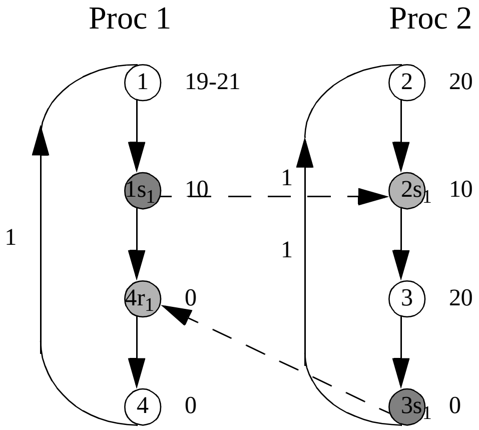


Figure 5. IPC graph for Example 2.

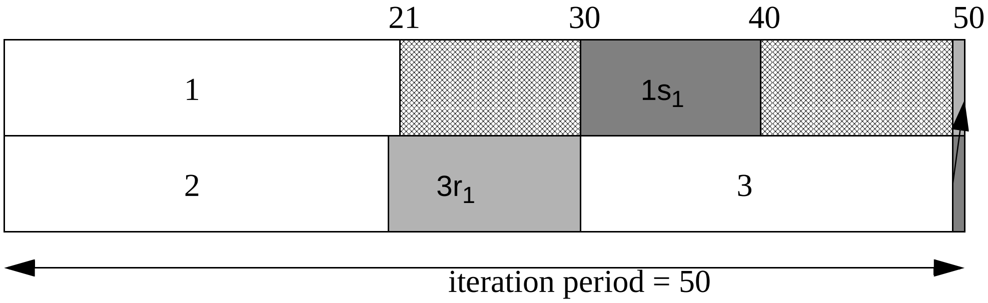


Figure 6. Gantt Chart for ST schedule when *exec(1)=21.*

In contrast, the iteration period obtained by computing the MCM of the ordered transaction graph with worst-case actor execution times is the worst-case iteration period. This is because the MCM is an accurate measure of performance for ordered transaction implementations [23,24], and the MCM can only increase or remain the same when the execution time of an actor is increased.

4. Finding optimal transaction orders

In the *transaction ordering problem*, our objective is to determine a transaction order O for a given IPC graph such that the MCM of the resulting ordered transaction graph is minimized (so that throughput is maximized). As mentioned in Section 2, it has been shown that this problem is tractable when IPC costs are ignored. In this section, we show that when IPC costs are considered, the transaction ordering problem becomes NP-complete.

We show this by first showing that determining an optimal transaction order for non-iterative implementations, which is a more restricted (easier) problem, is NP-complete. To convert an iterative IPC graph to a non-iterative one, it suffices to remove all edges in the graph that have delays of one or more. This results in an acyclic graph since any cycle in the original graph must

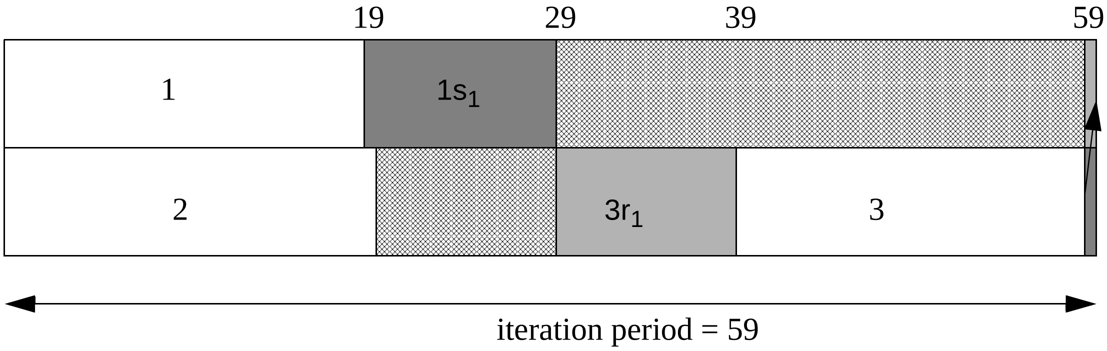


Figure 7. Gantt Chart for ST schedule with *exec(1)=19.*

have a delay of one or more for the graph not to be deadlocked.

Definition 1: Given an IPC graph $G_{ipc} = (V, E)$, the associated *non-iterative inter-processor communication* (*NIPC*) *graph* is defined as $G_{nipc} = (V, E_{nipc})$ where $E_{nipc} = \{e|(e \in E \textbf{ and } (delay(e) = 0))\}$.

Definition 2: Given an NIPC graph $G_{nipc} = (V, E_{nipc})$, and an ordering O, the corresponding *non-iterative ordered transaction (NOT) graph* $G_{NOT} = \Pi(G_{nipc}, O)$ is defined as $G_{NOT} = (V_{NOT}, E_{NOT})$, where $V_{NOT} = V$, $E_{NOT} = (E_{nipc} \cup E_O) - \{(o_p, o_1)\}$, and E_O is as defined in (4).

By definition, the total execution time (*makespan*) of a NOT graph G_{NOT} is finite, and this execution time can be determined in polynomial time — as the length of the longest cumulative-execution-time path in G_{NOT} — since G_{NOT} is acyclic and the execution times of all actors are nonnegative. However, given an IPC graph, finding a transaction order that minimizes the makespan of the associated NOT graph is intractable.

Definition 3: The *non-iterative transaction ordering* problem is defined as follows. Given an NIPC graph $G_{nipc} = (V, E_{nipc})$, and a positive integer k, does there exist a transaction order $O = \{o_1, o_2, \ldots o_n\}$ such that $G_{NOT} = \Pi(G_{nipc}, O)$ has a makespan that is less than or equal to k?

To show that non-iterative transaction ordering is NP hard, we derive a reduction from the *sequencing with release times and deadlines* (*SRTD*) problem, which is known to be NP-complete [8]. The SRTD problem is defined as follows.

Definition 4: (The *SRTD* problem). Given an instance set T of tasks, and for each task $t \in T$, a length (duration) $l(t) \in \aleph$, a release time $r(t) \in \aleph$, and a deadline $d(t) \in \aleph$, is there a single-processor schedule for T that satisfies the release time constraints and meets all the deadlines? That is, is there a one-to-one function (called a *valid SRTD schedule*) $\sigma : T \to \aleph$, with $(\sigma(t) > \sigma(t')) \Rightarrow (\sigma(t) \geq \sigma(t') + l(t'))$, and for all $t \in T$, $\sigma(t) \geq r(t)$, and $\sigma(t) + l(t) \leq d(t)$?

Theorem 1: The non-iterative transaction ordering problem is NP-complete.

Proof: This problem is clearly in NP since we can verify in polynomial time whether the longest path length (in terms of cumulative execution time) of the graph is less than or equal to a given positive integer.

Now suppose that we are given an instance of the SRTD problem (T, r, l, d) with $T = \{t_1, t_2 \ldots t_p\}$. We construct an NIPC graph G_{nipc} from this instance by carrying out the following steps. Here, all edges instantiated are delayless unless otherwise specified, and k is equal to the maximum deadline of the tasks in the given instance of the STRD problem.

For each $t_i \in T$,

i) instantiate a send actor u_i when i is odd, or a receive actor u_i when i is even with $exec(u_i) = l(t_i)$ and $proc(u_i) = i$.

ii) instantiate a computation actor m_i with $exec(m_i) = r(t_i)$ and $proc(m_i) = i$.

iii) instantiate a computation actor n_i with $exec(n_i) = k - d(t_i)$ and $proc(n_i) = i$.

iv) instantiate an edge (m_i, u_i) and another edge (u_i, n_i).

Each send actor u_i is connected to the receive actor u_{i+1} by an interprocessor edge (u_i, u_{i+1}) with a delay of unity. Since each of the interprocessor edges has a delay of unity, these edges are not present in G_{nipc}. Without loss of generality, we assume that there are an even number of tasks, so that the number of send and receive actors is the same (if the number of tasks is not even to begin with, we can instantiate an appropriately-defined dummy actor to generate an equivalent "even-task" instance). Observe from our construction that from the p tasks in the given instance of the SRTD problem, we construct a graph G_{nipc} that involves p processors, p communication actors, $2p$ computation actors, and $2p$ edges.

Claim: If there exists a transaction order O for $G_{NOT} = \Pi(G_{nipc}, O)$ that will have a makespan that is less than or equal to k, then there exists a valid SRTD schedule for the given instance of the SRTD problem.

The reasoning behind our construction and the above claim is that we make the communication actors of the ordered transaction graph correspond exactly to the tasks of the STRD problem. We do this by making the execution time of the computation actor before each corresponding communication actor equal to the release time of the associated task and, thus, guarantee that the communication actors cannot begin execution before their respective release times. Also since computation actors will begin execution from time 0 as each is on a different processor, the release times correspond to when they complete execution. Similarly, the execution time of the computation actors that follow the communication actors are chosen to be $k - d(t_i)$ so that the corresponding communication actors will have to complete their execution before $d(t_i)$ for the makespan to be less than or equal to k. This is true because the computation actor can begin execution immediately after the communication actor has finished. Therefore, the valid SRTD schedule corresponds exactly to the shared bus schedule in the derived instance of the non-iterative transaction ordering problem. If we can find a transaction order that has a makespan less than or equal to k, we have a bus schedule that schedules the communication actors in the same manner as an appropriate single-processor schedule for the corresponding SRTD tasks. Conversely, if a transaction order cannot be found that satisfies the given makespan constraint, it is easily seen that there is no valid SRTD schedule for the given instance of the SRTD problem. *Q.E.D.*

Note that in Theorem 1, we have simplified the problem greatly by assuming the inter-processor edges to have unit delays. This removes the inter-dependencies that are imposed by these edges, but even with this simplification, the problem remains NP-complete.

Example 3: Suppose that we are given an instance of the SRTD problem with task set $T = \{t_1, t_2, t_3, t_4\}$; and respective release times $r(T) = \{0, 4, 5, 6\}$, lengths $l(T) = \{5, 2, 3, 1\}$, and deadlines $d(T) = \{5, 8, 11, 8\}$. To construct an instance of the non-iterative transaction ordering problem with $k = 11$, we create 4 processors, each with 3 vertices. The execution times are determined from above — e.g., $u_1 = 5$, $m_1 = 0$, $n_1 = 6$. The resulting NOT graph is illustrated in Figure 8. Dash-dot edges indicate OT edges. Removing the dash-dot edges that represent the transaction order edges gives us the NIPC graph constructed from above. This figure shows a transaction order (u_1, u_2, u_4, u_3) where the schedule length of 11 is satisfied. This means that there exists a valid SRTD schedule for the given SRTD problem instance. The start times of the tasks can be obtained by finding the longest path lengths between the source nodes and the corresponding communication actors. Setting the starting times of the tasks (t_1, t_2, t_3, t_4) to equal $(0, 5, 8, 7)$, respectively, we obtain a valid SRTD schedule for the SRTD problem instance.

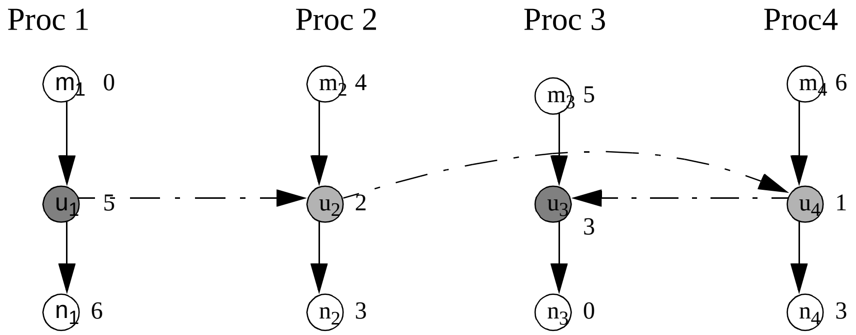


Figure 8. NOT Graph constructed in Example 3.

As demonstrated by the Theorem 2 below, we can extend the proof of Theorem 1 to show that the transaction ordering problem is NP-complete in the *iterative* context as well as the non-iterative case.

Definition 5: The *iterative transaction ordering* problem (also called the *transaction ordering problem*) is defined as follows. Given an IPC graph G_{ipc} and a positive integer k, does there exist a transaction order O such that $G_{OT} = \Gamma(G_{ipc}, O)$ satisfies $MCM(G_{OT}) \le k$?

Theorem 2: The iterative transaction ordering problem is NP-complete.

Proof: The MCM can be found in polynomial-time, therefore, the problem is in NP.

To establish NP-hardness, we again derive a reduction from the SRTD problem, and we modify the graph construction from the proof of Theorem 1 so that the MCM equals the makespan.

Now suppose we are given an instance of the SRTD problem (T, r, l, d) with $T = \{t_1, t_2 \ldots t_p\}$. We construct an IPC graph G_{ipc} from this instance by carrying out the following steps. All edges instantiated are delayless unless otherwise specified, and k is equal to the maximum deadline of the tasks in the given instance of the STRD problem.

For each $t_i \in T$,

i) instantiate a send actor u_i when i is odd, or a receive actor u_i when i is even with $exec(u_i) = l(t_i)$ and $proc(u_i) = i$.

ii) instantiate a computation actor m_i with $exec(m_i) = r(t_i)$ and $proc(m_i) = i$.

iii) instantiate a computation actor n_i with $exec(n_i) = k - d(t_i)$ and $proc(n_i) = i$.

iv) instantiate an edge (m_i, u_i) and another edge (u_i, n_i).

v) instantiate a send actor s_i with $exec(s_i) = 0$ and $proc(s_i) = i$.

vi) instantiate a receive actor r_i with $exec(r_i) = 0$ and $proc(r_i) = i$.

vii) instantiate a computation actor d_i with $exec(d_i) = 0$ and $proc(d_i) = i$.

viii) instantiate an edge (n_i, s_i), an edge (s_i, r_i), and another edge (r_i, d_i).

ix) instantiate another receive actor q_i with $exec(q_i) = 0$ and $proc(q_i) = p + 1$ (recall that $p = |T|$).

x) instantiate another send actor w_i with $exec(w_i) = 0$ and $proc(w_i) = p + 1$.

xi) instantiate an (interprocessor) edge (s_i, q_i) and another edge (w_i, r_i).

After completing all the above, join all q_is with edges in a linear chain, instantiate a computation actor h with $exec(h) = 0$ and $proc(h) = p + 1$, instantiate edges (q_p, h) and (h, w_1) and again join all w_is with edges in a linear chain. Finally for each of the $p + 1$ processors, add an edge with a delay of unity from the last actor on the processor to the first actor.

We again assume without loss of generality that there is an even number of tasks in T. Each send actor u_i is connected to the receive actor u_{i+1} with an interprocessor edge of unit delay. Note that in the OT graph $\Gamma(G_{ipc}, O)$, these interprocessor edges become *redundant* (in the sense of synchronization redundancy, as discussed in [3]) because of the ordered transaction edges added due to O: since the ordered transaction edges are connected by a cycle of delay unity, the constraints imposed by $\{(u_i, u_{i+1})\}$ are automatically met by the ordered transaction edges.

This graph effectively represents a blocked schedule for an iterative graph when the execution times of the actors that have been instantiated after step v) have execution times that are much less than the execution times of the other actors, and the MCM of the constructed graph represents the longest path or the schedule length of the graph. Note that each of the longest paths in the non-iterative graph will correspond to a cycle in the iterative case, where the cycle mean of the cycle is equal to the longest path (since the denominator of the associated quotient in (3) is unity).

Similarly, as in the non-iterative case, it is possible to find a one-processor schedule of the STRD instance that satisfies the constraints if we can determine a transaction order whose enforcement will guarantee that the MCM of the corresponding OT graph is less than or equal to k. This is true because the communication actors that have non-zero IPC cost in the bus schedule of the OT problem correspond to the tasks in the valid schedule of the STRD problem.

Hence, we can conclude that the (iterative) transaction ordering problem is NP-complete. *Q.E.D.*

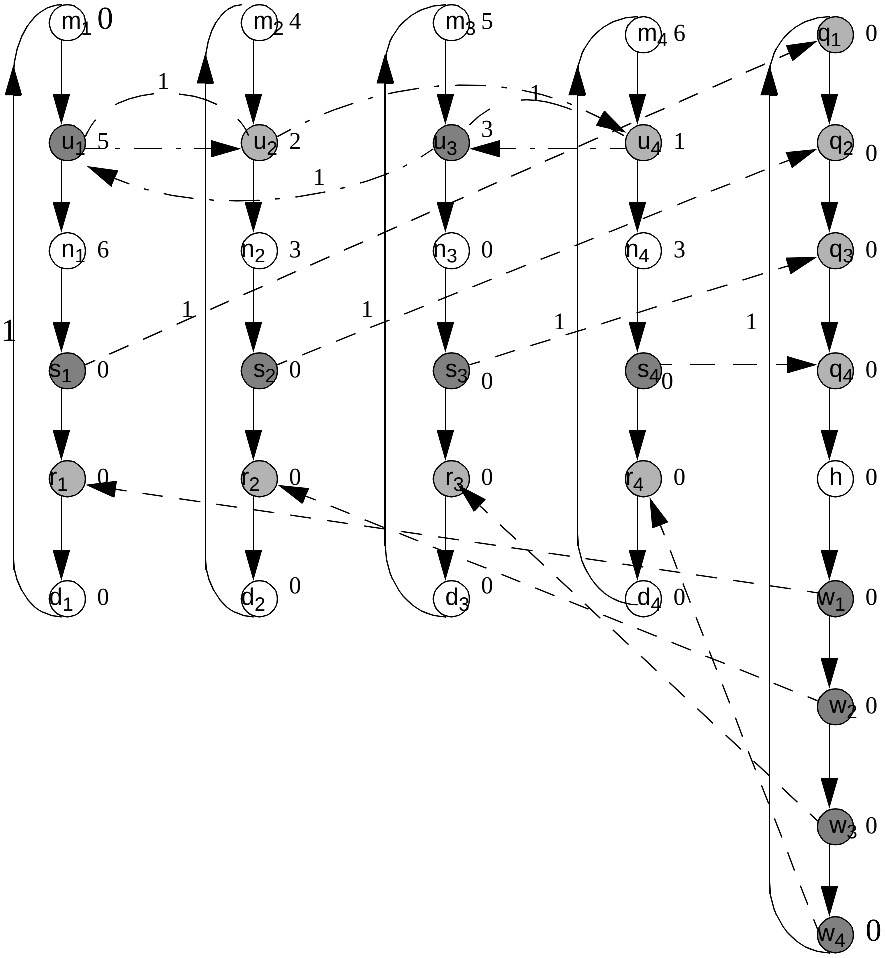


Figure 9. Constructed OT graph in Example 4.

Example 4: Consider again the SRTD instance of Example 4. Figure 9 shows the corresponding ordered transaction graph that results when the ordering (u_1, u_2, u_4, u_3) is imposed. Removing the OT edges $(u_1, u_2), (u_2, u_4), (u_4, u_3), (u_3, u_1)$ gives the constructed IPC graph. Note that the edges (u_1, u_2) and (u_3, u_4) introduced during construction are redundant in the OT graph due to the paths $((u_1, u_2))$ and $((u_3, u_1), (u_1, u_2), (u_2, u_4))$, respectively, that are imposed by the linear order and have delays of one or less.

5. The transaction partial order heuristic

The BFB technique does not take bus contention into consideration while scheduling the transaction order. Instead, it tries to find a transaction order that will be close to or equal to the associated self-timed schedule. However, we have demonstrated that in the presence of non-zero IPC, the OT method can, in fact, perform significantly better than the ST method, and thus, more direct consideration of OT execution is clearly worthwhile when scheduling transactions. For this purpose, we propose in this section a heuristic, called the *transaction partial order (TPO) algorithm*, that simultaneously takes IPC costs and the serialization effects of transaction ordering into account when determining the transaction order. Note that OT edges added to the IPC graph can only increase the MCM of the IPC graph, or leave the MCM unchanged. The MCM of the original IPC graph therefore represents a lower bound on the achievable average iteration period. By adding OT edges, we are effectively removing bus contention by making sure that no two communication actors submit conflicting bus requests, and this generally increases the MCM of the IPC graph. The TPO heuristic finds a transaction order on the basis that an OT edge that increases the MCM of the IPC graph by a comparatively smaller amount should be given preference. Therefore, to determine which communication actor should be scheduled first, we insert OT edges between communication actors that are contending for the bus (during the transaction ordering process), and calculate the corresponding MCM of the IPC graph. Actors whose corresponding MCMs are more favorable under such an evaluation are scheduled earlier in the transaction order.

More specifically, a partial order of the communication (send and receive) actors is first computed from the IPC graph G_{ipc}: the *transaction partial order (TPO) graph* G_{TPO} is computed by first deleting all edges in G_{ipc} that have nonzero delays, and then deleting all of the computation actors.

Example 5: The transaction partial order graph computed from the IPC graph of Figure 2 is illustrated in Figure 10. Notice that all the dependencies imposed by the IPC graph are retained in G_{TPO}, but only for the communication actors.

The heuristic proceeds by considering — one by one — each vertex of G_{TPO} that has no input edges (vertices in the TPO graph that have no input edges are called *ready* vertices) as a *candidate* to be scheduled next in the transaction order. Interprocessor edges are drawn from each candidate vertex to all other ready vertices in G_{ipc}, and the corresponding MCM is measured. The candidate whose corresponding MCM is the least when evaluated in this fashion is chosen as the

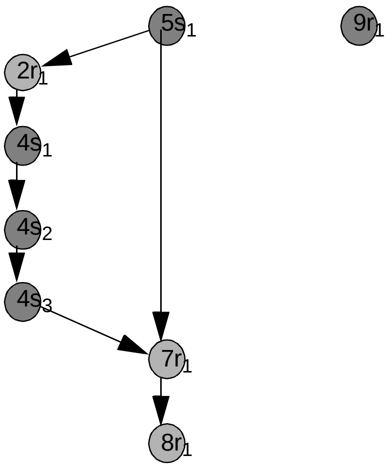


Figure 10. TPO Graph in Example 5.

next vertex in the ordered transaction, and deleted from G_{TPO}. This process is repeated until all communication actors have been scheduled into a linear ordering. A pseudocode specification of the TPO heuristic is given in Figures 11-13.

The algorithm makes sense intuitively since the dependencies imposed by the edges drawn from the candidate vertices will remain when the transaction ordering O is enforced. These edges represent constraints in addition to the interprocessor edges that are already present in G_{ipc} and thus, they can only increase the MCM or leave the MCM unchanged. Since we are interested in minimizing the MCM, we choose candidate vertices that increase the MCM by the least possible amounts. Thus, the algorithm follows a greedy strategy in choosing vertices, but it explicitly takes communication serialization and IPC costs into account

.

```
Function Choose-Communication-Actor
Input an IPC graph $G = (V, E)$, a TPO graph $G_{TPO}$ and a list of actors ReadyList
Output a communication actor $v$

For $x \in ReadyList$
    For $y \in ReadyList$
        If $x \neq y$
            e = G.addedge$(x, y)$
            temp.addedge(e)
        end if
    end for
    criteria$[x]$ = $MCM(G)$
end for
For $e \in temp$
        G.delete$(e)$
end for
$v = min\{criteria(x)\}$
return $v$
```

Figure 11. Function to choose the next communication actor in the transaction order.

Example 6: When we apply the TPO heuristic to the IPC graph of Figure 2, the schedule we obtain is illustrated by the Gantt chart of Figure 4. The corresponding OT graph is illustrated in Figure 14.

The OT edges corresponding to the actors that have already been scheduled are added as the heuristic proceeds since they represent the schedule of the bus, and hence, make the heuristic more accurate for the later stages of the transaction order. The maximum number of nodes in the ready list at any given instant is P (where P is the number of processors). The complexity of the algorithm is thus $O(P|V|^2|E_{OT}|)$ since the complexity of computing the MCM of a graph (V, E) is $O(|V||E|)$.

The edge of the transaction order that connects the last communication actor in the ordering with the first one has a delay of unity (to represent the transition to the next graph iteration). We can improve the performance of the TPO algorithm by introducing this edge at the beginning because it will give a more accurate estimate of the MCM in choosing vertices later as the heuristic proceeds. Under this modification, the heuristic proceeds as before, except that the “last” (unit-delay) transaction ordering edge is drawn at the beginning. Since G_{TPO} has a maximum of P

```
Function Initialize
Input an IPC graph $G$

compute $G_{TPO}$ from $G$
For $v \in G$
    mark[v] = FALSE
    If indegree(v) ==0
        ReadyList.append(v)
    end if
end for
```

Figure 12. Function to initialize data structures called by the TPO function.

communication actors that can be scheduled last in the transaction order, the modified heuristic is repeated for each of these candidate communication actors that can be scheduled in the end, and the best solution that results is selected. This increases the complexity of the algorithm by a factor of P to $O(P^2|V|^2|E_{OT}|)$.

```
Function TPO-heuristic
Input an IPC graph $G = (V, E)$
Output a linear list of communication actors LinearList

Initialize($(G)$
complete = FALSE
first = TRUE
while not (complete)
    $v$ = choose-communication-actor($(G, G_{TPO}, ReadyList)$
    mark[v] = TRUE
    LinearList.append(v)
    If not (first)
        G.addedge$(w, v)$
        $w = v$
    end if
    first = FALSE
    For $u \in \{(v, u) \in E\}$
        flag = TRUE
        For $s \in \{(s, u) \in E\}$
            If mark[s] = FALSE
                flag = FALSE
            end if
        end for
        If flag == TRUE
            ReadyList.append(u)
        end if
    end for
    If (ReadyList.empty) == TRUE)
        complete = TRUE
    end if
end while
return LinearList
end Function
```

Figure 13. Pseudocode for TPO heuristic.

6. Genetic algorithm for transaction scheduling

Since the transaction ordering problem is intractable, we are unable to efficiently find optimal transaction orders on a consistent basis. We have implemented a branch and bound strategy to explore the search space comprehensively, but this technique requires excessive amounts of time for graphs that have significant numbers of IPC edges. To develop an alternative to this branch and bound approach, and the TPO heuristic, we have implemented a genetic algorithm (GA) to search for the best transaction order. The GA exploits the increased tolerance for compile

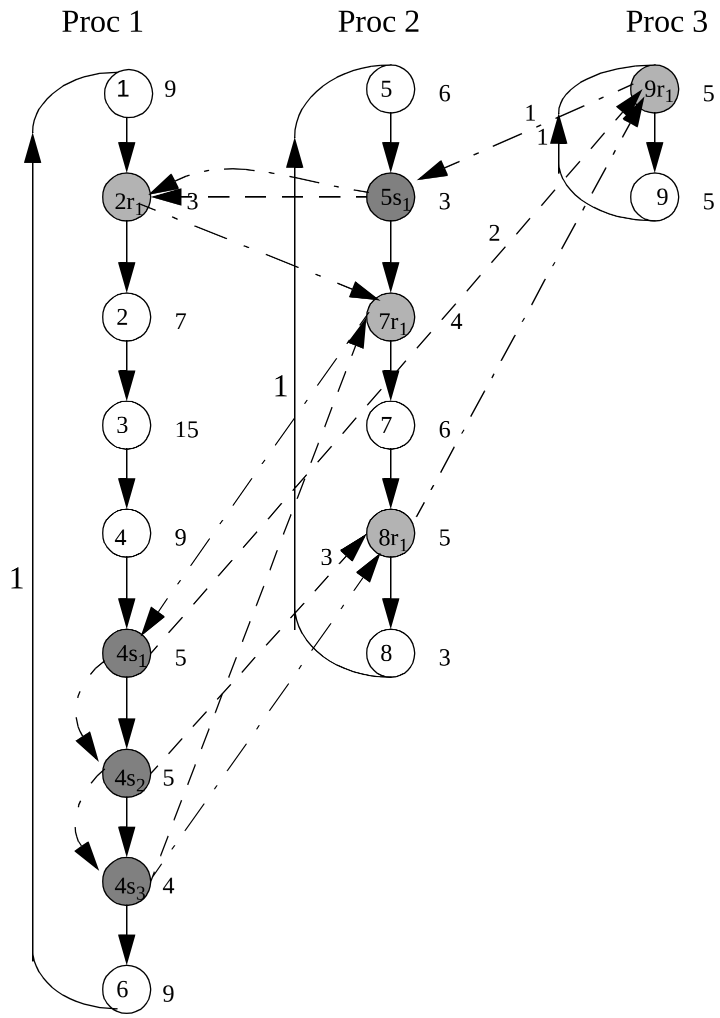


Figure 14. OT graph obtained by applying TPO heuristic in Example 6.

time that is available for many embedded applications [14], and can leverage the TPO heuristic by incorporating its solution in the “initial population.”

In our GA formulation, candidate transaction orders are encoded using the matrix-based sequence-encoding method described in [7]. Using this method, the partial order of the communication actors is converted into a precedence matrix and randomly completed to yield a random transaction order that is valid. Mutation is carried out by swapping rows and columns, and recombination is performed using the *intersection operator* explained in [7]. The intersection operator takes subsequences that are common among the parents by taking the boolean “and” of the two parent matrices to form the “offspring,” and the undefined part is randomly completed.

A pseudocode sketch of the GA is shown in Figure 15. For details on the underlying GA concepts (e.g., tournament selection), we refer the reader to [1]. The mutation step takes $O(|V|^2)$ time multiplied by the number of swaps carried out since each time we have to check whether the swap was valid by comparing it with the partial boolean matrix M_{TPO} corresponding to the trans-

```
Function TransOrderingGA
Input an IPC graph $G = (V, E)$
Output a linear list of communication actors LinearList

compute $G_{TPO}$ from $G$
convert $G_{TPO}$ to boolean matrix $M_{TPO}$
generate initial population $M$ by randomly completing $M_{TPO}$
For j = 1 to NoIterations
    For i = 1 to PopulationSize
        $P_i$ = mutate*$(M_i)$
        $R_i$ = recombine$(P_{i-1}, P_i)$
        $R_i$.FitnessValue = evaluate$(R_i, G)$
    end for
    $M$ = tournament_selection$(R, M)$
end for
```

Figure 15. Pseudocode for our GA approach to transaction ordering.

action partial order graph G_{TPO}. The recombination step takes $O(|V|^2)$ time, and the evaluation step takes $O(|V||E_{OT}|)$ time. The overall complexity of each iteration is also influenced by the population size and the overhead involved in generating random numbers.

7. Dynamic reordering

Once we obtain a transaction order (e.g., using the TPO heuristic or the GA approach defined in Section 6), it is possible to swap the position of consecutive communication actors in the transaction order as long as the new positions do not violate the dependencies imposed by the transaction partial order. This method has the advantage that it cannot degrade the transaction order since we can discard any solution that is worse. The concept is similar to *dynamic variable reordering* used in OBDD's (Ordered Binary Decision Diagrams) [17]. We have implemented an adaptation to ordered transaction scheduling, called *dynamic transaction reordering* (*DTR*), of the *Sifting Algorithm* introduced by Rudell [21], and have observed that from DTR, we consistently obtain improvements in the iteration period, regardless of the method used to find the transaction order.

8. Results

Experiments were carried out to compare the ST method and the OT method, and to measure the performance of the TPO, GA, and DTR heuristics in finding transaction orders. These heuristics were implemented in C/C++ using the *LEDA* [16] framework for fundamental graph-theoretic data structures and algorithms. The benchmarks are standard DSP applications that have been scheduled using the classic *HLFET* algorithm [8] with straightforward extensions to incorporate IPC costs.

The IPC graphs are fairly complicated, ranging from between 50-150 nodes, and the numbers of processors involved range from 2 to 8. The examples *fft1, fft2,* and *fft3* result from three representative schedules for Fast Fourier Transforms based on examples given in [15]; *karp10* is a music synthesis application based on the Karplus Strong algorithm in 10 voices; and *qmf4* is a 4 channel multi-resolution QMF filter bank for signal compression.

In the simulation of the ST schedule, we ignore the overhead of synchronization so as to give us a worst-case comparison with the OT schedule. In practice, of course, synchronization has nonzero cost, and thus, depending on the actual synchronization overhead in the target architecture, the benefit of the OT schedules examined will be even more that what the results here demonstrate. Thus, our analysis in this section gives a lower bound on the improvement we can expect using the OT implementation strategy in conjunction with our proposed transaction ordering techniques.

Table 1 compares the performance (iteration period) of the ST and the OT schedules. Here, the average iteration period (T_{OT}) of the OT schedule is obtained by taking the best performance using the algorithms proposed in Sections 5-7, and T_{ST} denotes the average iteration period of the corresponding ST schedule. In each of the cases, we see that the OT strategy can outperform the ST strategy, and that this holds even though we are ignoring synchronization costs, which gives us a very optimistic view of the performance under ST execution.

Table 1. Comparison of ST and OT schedules.

Application	T_{ST}	T_{OT}
fft1	263	245
fft2	312	300
fft3	263	245
karp10	312	308
qmf4	147	140

Table 2 gives us a comparison between the different heuristics in finding transaction orders. Each entry is the iteration period when the transaction order found by the heuristic is enforced. Column 2 shows the iteration period when a randomly-generated transaction order is enforced. From the table, we can conclude that all the heuristics work fairly well compared to the random transaction order. The TPO heuristic for which the results are shown is the modified version that inserts the unit-delay edge beforehand. This consistently gives us a slight improvement. Generally, the TPO heuristic works better than the BFB technique — especially for *fft1* and *fft3* — and the heuristic that combines the TPO heuristic and DTR performs best (even better than the GA which, takes significantly more time to execute). The GA was implemented with a population size of 100 and the number of iterations was set to 1000. The GA for the experiments that we tried generally stabilized before the 1000 iteration limit was reached.

Table 2. Comparison of algorithms.

Application	T_{random}	T_{BFB}	T_{TPO}	T_{GA}	$T_{\text{TPO+DTR}}$
fft1	392	280	245	255	245
fft2	395	340	320	300	300
fft3	390	300	255	255	245
karp10	482	312	309	308	309
qmf4	196	148	145	140	145

When we use the transaction ordering obtained by the TPO heuristic combined with DTR in the initial population of the GA, we achieve the best results since we simultaneously obtain the benefits of all three approaches. The results are shown in Table 3.

Table 3. Results when the GA is applied to the TPO heuristic in conjunction with DTR.

Application	$T_{\text{TPO+DTR+GA}}$
fft1	245
fft2	295
fft3	245
karp10	305
qmf4	140

9. Conclusions

We have demonstrated that in the presence of accurate estimates for actor execution times, the ordered transaction method — which is superior to the self-timed method in its predictability, and its total elimination of synchronization overhead — can significantly outperform self-timed implementation, even though ordered transaction implementation offers less run-time flexibility due to a fixed ordering of communication operations. We have also shown that in the presence of non-zero IPC costs, finding an optimal transaction order is an NP-complete problem, and we have developed a variety of heuristic techniques to find efficient transaction orders. These techniques include a low-complexity, deterministic heuristic for rapid design space exploration, and a genetic algorithm for exploiting extra compile time when generating final implementations. Useful directions for further work include integrating transaction ordering considerations into the scheduling process, and the exploration of hybrid scheduling strategies that can combine ordered transaction, self-timed, and fully-static strategies in the same implementation based on subsystem characteristics.

10. References

[1] T. Back, U. Hammel, and H-P Schwefel, “Evolutionary computation: Comments on the history

and current state," *IEEE Transactions on Evolutionary Computation*, 1(1):3-17, 1997.

[2] S. Banerjee, T. Hamada, P. M. Chau and R. D. Fellman, "Macro Pipelining Based Scheduling on High Performance Heterogeneous Multiprocessor Systems," *IEEE Transactions on Signal Processing*, Vol. 43, No. 6, pp 1468-1484, June, 1995.

[3] S. S. Bhattacharyya, S. Sriram and E. A. Lee, "Minimizing Synchronization Overhead in Statically Scheduled Multiprocessor Systems," *Proceedings of the International Conference on Application Specific Array Processors*, July, 1995.

[4] S. S. Bhattacharyya, S. Sriram and E. A. Lee. "Latency-constrained Resynchronization for Multiprocessor DSP Implementation," *Proceedings of the International Conference on Application Specific Systems, Architectures and Processors*, August, 1996.

[5] S. S. Bhattacharyya, P. K. Murthy, and E. A. Lee, *Software Synthesis from Dataflow Graphs*, Kluwer Academic Publishers, 1996.

[6] J. T. Buck, S. Ha, E. A. Lee, and D. G. Messerschmitt, "Ptolemy: A framework for simulating and prototyping heterogeneous systems," *International Journal of Computer Simulation,* Vol. 4, pp. 155-182, Jan. 1994.

[7] B. R. Fox and M. B. McMahon, "Genetic Operators for Sequencing Problems", G. Rawlins, *Foundations of Genetic Algorithms*, Morgan Kaufman Publishers Inc., 1991.

[8] M. R. Garey and D. S. Johnson, *Computers and Intractability: A Guide to the Theory of NP-Completeness,* W.H. Freeman and Company, New York, 1999.

[9] T. C. Hu, "Parallel sequencing and assembly line problems," *Operations Research*, Vol. 9, No. 6, pp. 841-848, Nov. 1961.

[10] S. Y. Kung and P. S. Lewis and S. C. Lo, "Performance Analysis and Optimization of VLSI Dataflow Array," *Journal of Parallel and Distributed Computing*, pp. 592-618, 1987.

[11] E. A. Lee and J. C. Bier, "Architectures for Statically Scheduled Dataflow," *Journal of Parallel and Distributed Computing*, Vol. 10, pp. 333-348, Dec. 1990.

[12] E. A. Lee and D. G. Messerschmitt, "Static Scheduling of Synchronous Dataflow Programs for Digital Signal Processing," *IEEE Trans. on Computers*, Vol. C-36, No. 2, Feb. 1982.

[13] E. A. Lee and S. Ha, "Scheduling strategies for Multiprocessor real-time DSP," *Proceedings of the Globecom Conference*, Dallas, Texas, pp. 1279-1283, Nov. 1989.

[14] P. Marwedel and G. Goossens, *Code Generation for Embedded Processors*, Kluwer Academic Publishers, 1995.

[15] C. L. McCreary, A. A. Kahn, J. J. Thompson and M. E. McArdle, "A Comparison of Heuristics for Scheduling DAGS on Multiprocessors," *International Parallel Processing Symposium*, 1994.

[16] K. Mehlhorn and S. Naher and M. Seel and C. Uhrig, *The LEDA User Manual*, Max-Planck--Institut for Informatik, Saarbrucken, Germany, Version 3.7.

[17] G. De. Micheli, *Synthesis and Optimization of Digital Circuits*, McGraw-Hill, 1994.

[18] K. Parhi and D. G. Messerschmitt, "Static Rate Optimal Scheduling of Iterative Dataflow Programs via Optimum Folding," *IEEE Transactions on Computers*, Vol. 40, No. 2, pp. 178-194, Feb. 1991.

[19] J. L. Peterson, *Petri Net Theory and Modelling of Systems*, Prentice-Hall Inc., Englewoods Cliffs, New Jersey, 1981.

[20] R. Reiter, “Scheduling Parallel Computations,” *Journal of the Association for Computing Machinery*, Vol. 15, No. 4, pp. 590-599, Oct. 1968.

[21] R. Rudell, *“Dynamic Variable Ordering for Ordered Binary Decision Diagrams,”* IEEE Transactions on Computer Aided Design, 1993.

[22] G. C. Sih, “Multiprocessor Scheduling to account for Interprocessor Communication,” Ph.D. Thesis, Memorandum No. UCB/ERL M91/29, Electronics Research Laboratory, University of California at Berkeley, April, 1991.

[23] S. Sriram and E. A. Lee, “Determining the Order of Processor Transactions in Statically Scheduled Multiprocessors,” *Journal of VLSI Signal Processing*, pp. 207-220, 1997.

[24] S. Sriram, “Minimizing Communication and Synchronization Overhead in Multiprocessors for Digital Signal Processing,” Ph.D. Thesis, Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, 1995.