On Lower Bounds for Scheduling Problems in High-Level Synthesis

M. Narasimhan† (mnarasim@ichips.intel.com)
J. Ramanujam§ * (jxr@ee.lsu.edu)

†Intel Corporation, Dupont, WA 98327
§Dept. of Electrical and Computer Engineering, Louisiana State University, Baton Rouge, LA 70803

Abstract

This paper presents new results on lower bounds for the scheduling problem in high-level synthesis. While several techniques exist for lower bound estimation, comparisons among the techniques have been experimental with few guarantees on the quality of the bounds. In this paper, we present new bounds and a theoretical comparison of these with existing bounds. For the resource-constrained scheduling problem, we present a new algorithm which generalizes the bounding techniques of Langevin and Cerny [6] and Rim and Jain [11]. This algorithm is shown to produce bounds that are provably tighter than other existing techniques. For the time constrained scheduling problem, we show how to generate the tightest possible bounds that can be derived by ignoring the precedence constraints by solving a linear programming formulation. These bounds are therefore guaranteed to be tighter than the bounds generated by the techniques of Fernandez-Bussell [2] or Sharma-Jain [12]. As a result, we show that the linear relaxation of the ILP formulation of the time constrained scheduling problem produces tighter bounds than the two techniques mentioned above.

1 Introduction

High-level synthesis (HLS) is the translation of a behavioral level specification into a register-transfer level description. A behavioral level description is typically converted to a data flow graph (DFG) that is used as an intermediate representation in the HLS system. A central task in HLS is scheduling, which is the mapping of operations of the data flow graph to appropriate control steps (c-steps), where a c-step is one cycle of the system clock. Often scheduling is the first task in an HLS system. The mapping of operations to c-steps must not violate precedence constraints between operations in a DFG. In addition to the precedence constraints in the DFG, a designer may face additional design criteria such as latency (time), area, clock cycle, etc.

The Resource-Constrained Scheduling (RCS) problem involves the minimization of the number of c-steps required to execute all the operations given that the number of instances of each resource available is fixed. The Time-Constrained Scheduling (TCS) problem involves minimizing the number of instances of each resource required when the number of time steps is given. Unfortunately both these problems are NP-complete; so the best known algorithms for solving these problems have exponential time complexity. Because of the intractable nature of these problems, heuristics are often used. While heuristics very often produce solutions which are near optimal, if not optimal, these methods do not produce any performance guarantee, and very often, it is desirable to be able to bound the quality of the solutions produced. An easy way of doing so is to compute a lower bound on these quantities for all schedules. So, for example, if the lower bound on the latency of all schedules in the RCS problem equals the latency of the schedule produced by a heuristic, then clearly, the schedule returned by the heuristic has optimal latency. Verifying the optimality of schedules in this manner depends crucially on the quality (or tightness) of the the bounds obtained. Such a technique is useful for improving the performance of *exact scheduling techniques* such as branch-and-bound and Integer Linear Programming (ILP).

In this paper we present techniques to calculate tight lower bounds on the latency of schedules for the RCS problem and on the number of instances of each resource class required for the TCS problem. There has already been a considerable amount of work done in the area of bounding techniques for the scheduling problem. Chaudhuri and Walker [1] and Hu et al. [4] have shown algorithms for computing lower bounds for the Time-Constrained Scheduling problem. Rim and Jain [11] and Langevin and Cerny [6] have shown algorithms for computing lower bounds for the Resource-Constrained Scheduling problem. Several authors [5, 9, 10, 12] have shown different algorithms that compute lower bounds for both these problems or for related problems. However, in most cases, it is difficult to guarantee that any one of these techniques will always produce a tighter bound than the others, and all comparisons between these algorithms so far have been experimental in nature. In this paper, in addition to deriving tighter bounds, we present a theoretical comparison of several techniques.

This paper is organized is follows. In Section 2, we discuss the problems considered in this paper along with the associated terminology and notation. Section 3 presents tighter lower bounds for the resource-constrained scheduling problem and shows that the bounds obtained by our technique are better than the bounds from other techniques. In Section 4, we show that the the linear relaxation of the ILP formulation of time-constrained scheduling yields tighter bounds than other techniques for this problem. Section 5 concludes with a summary and discussion.

2 Preliminaries

Given the problem of scheduling the tasks(operations) T with dependencies E between these tasks, we model it as a directed acyclic graph $G = (T, E)$, where T represents the vertices of G and E the (directed) edges of G. An operation that begins executing at some step c is said to be scheduled to time step c. Each operation t_i takes $\mathsf{delay}(t_i)$ time steps to complete executing. A dependence $\langle t_i, t_j \rangle \in E \Rightarrow$ operation t_j can begin executing only on or after time step $k + \mathsf{delay}(t_i)$. This is modeled by assigning a weight $\mathsf{delay}(t_i)$ to all such edges. We assume that there are R resources

*Corresponding author.

classes (adders, multipliers etc.) and that each operation can execute only on one resource belonging to a particular class[1]. We will let the function Resource : $T \to \{1, 2, \ldots, R\}$ be a mapping of operations to the corresponding resource classes.

A schedule of G is a function $\mathcal{S} : T \to \mathbb{Z}^+$, where each operation t_i is scheduled to begin execution at time step $\mathcal{S}(t_i)$. Note that in order for the schedule to be valid, we require that the schedule satisfy all the dependence constraints, and that no resource and time constraints (if any) are violated. The latency of a schedules is defined to be the maximum completion time of all operations under that schedule; therefore, $\mathsf{Latency}(\mathcal{S}) = \max_{t_i \in T} \{\mathcal{S}(t_i) + \mathsf{delay}(t_i)\}$.

We use $T_{pred}(t_i)$ (resp. $T_{succ}(t_i)$) to denote the set of direct and indirect predecessors (resp. successors) of task t_i. $G_{pred}(t_i)$ (resp. $G_{succ}(t_i)$) is the subgraph induced by $T_{pred}(t_i)$ (resp. $T_{succ}(t_i)$). Clearly, all elements in $T_{pred}(t_i)$ must have completed execution before t_i can be scheduled, and no operation in $T_{succ}(t_i)$ can begin execution before t_i completes executing.

Longest weighted directed paths in G are called *critical paths* of G. The length of a critical path of G is denoted by $\mathsf{CP}(G)$. A lower bound on the execution time of G is clearly $\mathsf{CP}(G) + 1$. For any operation (vertex) $t_i \in G$, we define

$$\mathsf{ASAP}(t_i) = \mathsf{CP}\big(G_{pred}(t_i)\big) + 1$$

and

$$\mathsf{ALAP}(t_i, UB) = UB - \mathsf{CP}\big(G_{succ}(t_i)\big)$$

We require that the value of UB be at least the latency of the optimal schedule. For the RCS problem, the value of UB is not specified, and so the value used could the latency of any schedule (possibly produced by a heuristic). $\mathsf{ASAP}(t_i)$ (As Soon As Possible) represents a lower bound on the time step that operation t_i can be scheduled at in any valid schedule. Similarly, $\mathsf{ASAP}(t_i, UB)$ (As Late As Possible) represents an upper bound on the time step at which operation t_i can be scheduled in any (valid) schedule that completes execution of all operations by step UB. The values of $\mathsf{ASAP}(t_i)$ and $\mathsf{ALAP}(t_i, UB)$ represent exact bounds only in the absence of (resource) constraints. Discarding constraints corresponds to increasing the size of the solution space, and minimizing over a larger solution space, gives a lower bound on the optimal solution in the original solution space. The scheduling problem as presented is NP-Complete. In order to find good bounds for this problem, we present related problems, which we can solve efficiently, obtained by replacing some of the constraints with other ones. In the next two sections, we present results that show that the solutions to these problems will be no greater than the optimal solution to the scheduling problems, and therefore valid bounds.

3 Bounds for the RCS Problem

In this section, we will obtain bounds on the time taken to schedule all the tasks in T, given constraints on the number of resources of each type available. Our strategy for obtaining bounds will be to discard all the precedence constraints, and then calculate the optimal value of the latency of the schedule; this will be a lower bound on the optimal latency of schedules that preserve the precedence constraints.

Before we describe how to compute this bound, we will describe a technique to compute the optimal solution to simpler problem, and then show how this procedure can be applied to solve the problem at hand.

```
Procedure SimpleBound(T, Release, D, M_1)
NumberResAvail[step] ← M_1  ∀step ∈ Z^+
Demerit ← 0
for i ← 1 to N do
   step ← Release(t_i)
   while (NumberResAvail[step] = 0) do
      incr step
   enddo
   S(t_i) ← step
   decr NumResAvail[step]
   Demerit ← max(Demerit, step + D(t_i))
endif
enddo
return Demerit
enddo
```

Figure 1: Algorithm to minimize Demerit

3.1 Scheduling Independent Operations with Release Times and Resource Constraints

In this section, we will consider the problem of scheduling independent tasks with release times. Let $T = \{t_1, t_2, \ldots, t_N\}$ be the set of operations to be scheduled. Each task $t_i \in T$ has a release time $\mathsf{Release}(t_i)$, and so the operation cannot begin executing before time step $\mathsf{Release}(t_i)$. We initially assume that all operations execute on the same type of resource, i.e., there is only one resource class, and that each resource of this class has unit delay. The number of instances of this resource available is M_1. Further, we assume that we are given another function $\mathsf{D} : T \to \mathbb{Z}^+$, and we seek a schedule $\mathcal{S} : T \to \mathbb{Z}^+$ to minimize the value of

$$\mathsf{Demerit}(\mathcal{S}) = \max_{t_i \in T} \big(\mathcal{S}(t_i) + \mathsf{D}(t_i)\big)$$

Without loss of generality, we assume that the operations in T have been ordered in decreasing order of $\mathsf{D}(t_i)$, i.e., $\mathsf{D}(t_1) \geq \mathsf{D}(t_2) \geq \mathsf{D}(t_3) \geq \cdots \geq \mathsf{D}(t_N)$

Theorem 3.1. *The algorithm shown in Figure 1 will always produce a schedule $\mathcal{S}$ such that the value of* $\mathsf{Demerit}(\mathcal{S})$ *is minimum.*

Proof. See [7] □

Lemma 3.2. *If* $\mathsf{Release}_1(t_i) \geq \mathsf{Release}_2(t_i)\ \forall t_i \in T$, *then* $\mathsf{SimpleBound}(T, \mathsf{Release}_1, \mathsf{D}, M_1) \geq \mathsf{SimpleBound}(T, \mathsf{Release}_2, \mathsf{D}, M_1)$.

Proof. See [7] □

Lemma 3.3. *If* $\mathsf{D}_1(t_i) \geq \mathsf{D}_2(t_i)\ \forall t_i \in T$, *then* $\mathsf{SimpleBound}(T, \mathsf{Release}, \mathsf{D}_1, M_1) \geq \mathsf{SimpleBound}(T, \mathsf{Release}, \mathsf{D}_2, M_1)$.

Proof. See [7] □

The bounding procedure SimpleBound has been defined only for sets of operations that are all mapped onto the same resource class. Now, suppose the operations in T are mapped to R different resource classes. In this case, we partition T into R subsets $\{T_1, T_2, \ldots, T_R\}$ such that $T = T_1 \cup T_2 \cup \cdots \cup T_R$

[1]The problem of determining the resource class on which each operation is scheduled is called the *module selection problem*. We assume that module selection has already been completed.

and each partition T_j consists of all tasks mapped to resource class j, i.e., $\mathsf{Resource}(t_i) = j \quad \forall t_i \in T_j$. Given any function f, we will denote by $f\big|_{T_j}$ the restriction of f to the set T_j. We define the bound as $\mathsf{SimpleBound}(T, \mathsf{Release}, \mathsf{D}) = \max_{1\le j\le R} \mathsf{SimpleBound}\big(T_j, \mathsf{Release}\big|_{T_j}, \mathsf{D}\big|_{T_j}, M_r\big)$

3.2 Application to the original problem

Since we are considering bounds that can be obtained by relaxing precedence constraints, it is useful to (for this section alone) extend our definition of schedules. We will refer to all functions $\mathcal{S} : T \rightarrow \mathbb{Z}^+$ that satisfy the resource constraints, i.e., $|\{t \in T\,|\mathsf{Resource}(t_i) = r, \mathcal{S}(t_i) = s\}|$ for each $r \in 1, 2, \ldots, R$ and $s \in \mathbb{Z}^+$ as schedules. We will refer to schedules that do not violate precedence constraints as valid schedules. We will also extend the definition of schedules to subsets of the set of tasks T. We will refer to $\mathcal{S}'$ as a valid schedule of $T' \subseteq T$ if it satisfies all precedences E' induced by T'. For any $T' \subseteq T$, we define $\mathsf{Opt}(T') = \min_{\mathcal{S} \text{ a valid schedule for } T'} (\mathsf{Latency}(\mathcal{S}))$. Let us assume that UB is an upper bound on the latency of the optimal valid schedule (i.e., the schedule does not violate any precedence constraints). Any valid schedule is an upper bound on the latency of an optimal valid schedule, and so this value can be easily obtained as the latency of a heuristic solution.

We will derive results for two functions $\mathcal{H}_1$ and $\mathcal{H}_2$. Suppose we are given functions $\mathcal{H}_1$ and $\mathcal{H}_2$ which satisfy the following properties.

- $UB - \mathsf{Opt}\big(G_{succ}(t_i)\big) \le \mathcal{H}_2(t_i) \le UB - \mathsf{delay}(t_i)$
- $0 \le \mathcal{H}_1(t_i) \le \mathsf{Opt}\big(G_{pred}(t_i)\big)$

Note that the ASAP and the ALAP functions as we defined before satisfy these conditions for $\mathcal{H}_1$ and $\mathcal{H}_2$ respectively. Also notice that $\mathcal{H}_1$ and $\mathcal{H}_2$ need not be schedules as they need not satisfy the resource constraints. $\mathcal{H}_1(t_i)$ represents an upper bound on $\mathsf{ASAP}_{ideal}(t_i) = \mathsf{Opt}(G_{pred}(t_i))$, while $\mathcal{H}_2(t_i)$ is a lower bound on $\mathsf{ALAP}_{ideal}(t_i) = \mathsf{Opt}(G_{succ}(t_i)$.

Let $\Lambda(\mathcal{S}) = \max\limits_{t_i\in T}\big(\mathcal{S}(t_i) + UB - \mathcal{H}_2(t_i)\big)$ and $\lambda(\mathcal{S}) = \max\limits_{t_i\in T} \mathcal{S}(t_i) + \mathsf{delay}(t_i)$.

Theorem 3.4. *For any valid schedule $\mathcal{S}$, $\lambda(\mathcal{S}) = \Lambda(\mathcal{S})$.*

Proof. See [7] □

As a result, it does not really matter whether we optimize for $\lambda(\mathcal{S})$ or for $\Lambda(\mathcal{S})$ when computing exact solutions. However, the choice of the optimizing function does make a difference when we consider a relaxation where all the precedences are discarded. It can be seen that the inequality $\Lambda(\mathcal{S}) \ge \lambda(\mathcal{S})$ holds irrespective of the validity (in terms of precedence constraints) of $\mathcal{S}$. However, the two need not be equal if any precedences are violated in $\mathcal{S}$ as the $UB - \mathcal{H}_2(t_i)$ term is effectively a penalty associated with precedence violations. Therefore, when we relax all precedence constraints, the optimal value of Λ is a tighter bound than the optimal value of λ. Solving for an optimal value of Λ can be computed by calling the Lower-Bound routine with $\mathcal{H}_1$ as the Release function and $UB - \mathcal{H}_2$ as the D function. The complete algorithm for calculating a lower bound on the latency of the optimal valid schedule given the functions $\mathcal{H}_1$ and $\mathcal{H}_2$ is shown in Figure 2. Table 1 shows the results of solving for the optimal value of both λ and Λ (indicated in the table as LP(λ) and LP(Λ) respectively) for various benchmarks, and compares them to bounds achieved by relaxing the integrality constraints in the ILP formulation. It can be seen that solving for optimal Λ produces strictly tighter bounds in most of the cases.

Table 1: Lower bounds on latency

Bench.	$*$	ALU	Opt.	LB(λ)	LB(Λ)	LB(LP)
EWF	1	1	28	27	28	23
EWF	2	1	28	27	28	23
EWF	1	2	21	20	21	19
EWF	2	2	18	17	18	18
FDCT	1	1	34	33	34	21
FDCT	2	2	18	17	18	13
FDCT	2	3	18	17	18	12
FDCT	3	3	14	12	13	10
ARF	1	1	34	32	34	29
ARF	1	2	34	32	34	29
ARF	2	1	18	16	18	16
ARF	2	2	18	16	18	16
HAL	1	1	13	12	13	12
HAL	2	1	8	6	7	8
HAL	3	2	6	6	6	6

Procedure $\mathsf{LowerBound}(T, \mathcal{H}_1, \mathcal{H}_2, UB)$
for $i \leftarrow 1$ **to** $|T|$ **do**
  $\mathsf{Release}(t_i) \leftarrow \mathcal{H}_1(t_i)$
  $\mathsf{D}(t_i) \leftarrow UB - \mathcal{H}_2(t_i)$
enddo
return $\mathsf{SimpleBound}(T, \mathsf{Release}, \mathsf{D})$

Figure 2: Lower Bounding Routine

3.3 Obtaining $\mathcal{H}_1$ and $\mathcal{H}_2$

As stated in Lemmas 3.1 and 3.2, the tightest bounds on the latency of optimal schedules of T can be obtained when the tightest values of the $\mathcal{H}_1$ and $\mathcal{H}_2$ functions that satisfy the requirements are used. The values of $\mathcal{H}_1(t_i)$ and $\mathcal{H}_2(t_i)$ are actually lower and upper bounds on the value of $\mathsf{ASAP}_{ideal}(\phi, t_i)$ and $\mathsf{ALAP}_{ideal}(\phi, t_i)$ respectively, where ϕ is the set of all schedules that satisfy all the precedence and all the resource constraints and $\mathsf{Latency}(\mathcal{S}) \le UB$.

Define $T(t_i, t_j) = T_{succ}(t_i) \cap T_{pred}(t_j)$ and a family of functions $\mathcal{H}_{2_j} : T_{pred}(t_j) \rightarrow \mathbb{Z}^+$ where $\mathcal{H}_{2_j}(t_i) = \mathsf{CP}\big(T(t_i, t_j)\big)^2$. Now, we use the fact that operation t_i cannot be scheduled until all the operations in $G_{pred}(t_i)$ have been completed. It can be shown that $\mathsf{ASAP}_{ideal}(\phi, t_i) \ge \mathsf{LowerBound}(T_{pred}(t_i), \mathsf{ASAP}, \mathcal{H}_{2_i}, UB) \ge ASAP(t_i)$. So we use the routine shown in Figure 3 to compute $\mathcal{H}_1$, which is a better bound for ASAP_{ideal} than is $ASAP$.

To compute the value of $\mathcal{H}_2$ we make use of the following Lemma.

Lemma 3.5. *A lower bound on the* ASAP *function for a graph G is also an upper bound on the* ALAP *function for the graph G^r obtained from G by reversing all the edge directions (dependencies) and vice-versa.*

[2] It can be shown that tighter results can be obtained by defining $\mathcal{H}_{2_j}$ as $\mathcal{H}_{2_j}(t_i) = \mathsf{LowerBound}(T(t_i, t_j), \mathsf{ASAP} - \mathsf{ASAP}(t_i), \mathsf{ALAP}(t_j, UB) - \mathsf{ALAP}(UB), UB)$.

```
H_1 ← ASAP_G

for i ← 1 to |T| do
  t_i ← i^th operation in topologically sorted order
  H_1(t_i) ← LowerBound (G_pred(t_i), H_1|_{G_pred(t_i)}, ALAP|_{G_pred(t_i)}, UB)
enddo
```

Figure 3: Code fragment to calculate $\mathcal{H}_1$ (and $\mathcal{H}_2$)

Proof. See [8] for a proof. □

Theorem 3.6. *Let $S \in \phi$ be a valid schedule for the DFG $G = (V, E)$. Let $G^r = (V, E^r)$ be another DFG, where $\langle x_i, x_j \rangle \in E^r \Leftrightarrow \langle x_j, x_i \rangle \in E$. Then, $S^r : V \to Z$ where $S^r(x_j) = Latency(S) - S(x_j) + 1$ is a valid schedule for G^r and $Latency(S) \geq Latency(S^r)$.*

PROOF: In order to prove the validity of S^r, we must prove the following

- All dependencies are satisfied;
- All resource constraints are satisfied.

Consider the dependency $\langle x_j, x_i \rangle \in E^r$. By construction, $\exists \langle x_i, x_j \rangle \in E$. Since S is a valid schedule, all dependencies must have been satisfied in S. Therefore, we have $S(x_j) > S(x_i)$. So $S^r(x_i) - S^r(x_j) = Latency(S) + 1 - S(x_i) - Latency(S) - 1 + S(x_j) = S(x_j) - S(x_i) > 0$. Therefore, the dependency $\langle x_j, x_i \rangle \in E^r$ is satisfied in S^r. To show that all resource constraints are satisfied in S^r, we notice that $S(x_j) = S(x_i) \Leftrightarrow S^r(x_j) = S^r(x_i)$. If we denote by $R(step, class, S)$ the set $\{x_j \in V | S(x_j) = step \bigcap type(x_j) = class\}$, then $R(step, class, S) = R(Latency(S) - step + 1, class, S^r)$. Therefore all resource constraints are satisfied in S^r.
Finally, $S^r(x) = Latency(S) + 1 - S(x) \leq Latency(S)$ because $S(x) \geq 1$.

Corollary 3.7. *If $S : V \to Z$ is an optimal schedule for $G = (V, E)$, then $S^r : V \to Z$ is an optimal schedule for $G^r = (V, E^r)$.*

A key point to notice is that the ASAP values of the operations of the DFG G become the ALAP values of the corresponding operations of the DFG G^r. This is because, the set of predecessors of an operation become the set of successors of the corresponding operation in the reversed DFG. So finding the ASAP values of G (resp. G^R) is the same as finding the ALAP values of G^R (resp. G).

Corollary 3.8. *If $S(x_j) \geq ASAP(x_j)\ \forall x_j \in V(G)$, S a valid schedule, then $S^r(x_j) \leq Latency(S) - ASAP(x_j) + 1$.*

So we can use the same routine to compute the values of $\mathcal{H}_2$.

Lemma 3.9. *For any task graph $G = (T, E)$ and any $\mathcal{H}_1 \geq \mathsf{ASAP}$, the value of $\mathsf{LowerBound}(T, \mathcal{H}_1, \mathcal{H}_2) \geq \mathsf{CP}(G)$.*

Proof. See [7] □

Lemma 3.10. *The code fragment in Figure 3 calculates values of $\mathcal{H}_1$ and $\mathcal{H}_2$ which satisfy the required properties, and which are tighter than the* ASAP *and the* ALAP *functions.*

Proof. From the code in Figure 3, it can be seen that $\mathcal{H}_1(t_i)$ is calculated only after $\mathcal{H}_1(t_j)$ has been calculated for all predecessors t_j of t_i. If t_i, has no predecessors, then the statement $\mathcal{H}_1(t_j) \geq \mathsf{ASAP}(t_j) \quad \forall t_j \in G_{pred}(t_i)$ holds vacuously. Therefore, the value calculated for $\mathcal{H}_1(t_i) \geq \mathsf{ASAP}(t_i)$ by Lemma 3.4. Now, suppose that for some t_i (with predecessors), $\mathcal{H}_1(t_j) \geq \mathsf{ASAP}(t_j) \quad \forall t_j \in G_{pred}(t_i)$ holds. Then again by Lemma 3.4, we have $\mathcal{H}_1(t_i) \geq \mathsf{ASAP}(t_i)$. Hence, by induction, $\mathcal{H}_1(t_i) \geq \mathsf{ASAP}(t_i) \quad \forall t_i \in T$.
The proof that $\mathcal{H}_2(t_i) \leq \mathsf{ALAP}(t_i) \quad \forall t_i \in T$ is similar to the one above. □

The bounds calculation is done in topological order so that any improvement in the value of either of these functions will be propagated to their successors or predecessors.

3.4 Discussion

It is easy to see that the bounds derived by Rim and Jain [11] are the same as the bounds obtained from $\mathsf{LowerBound}(T, \mathsf{ASAP}, UB - \mathsf{ALAP}, UB)$, and that the bounds computed by Langevin and Cerny [6] is the same as the bounds obtained from $\mathsf{LowerBound}(T, \mathcal{H}_1, UB - \mathsf{ALAP})$. Therefore, the bounds obtained from the procedure shown above are tighter than the bounds obtained by Langevin and Cerny [6] which in turn are tighter than those obtained by Rim and Jain [11].

The algorithm shown has O(N) time complexity, once $\mathcal{H}_1$ and $\mathcal{H}_2$ have been calculated; the functions $\mathcal{H}_1$ and $\mathcal{H}_2$ can be calculated in O(N^2) time. Typically in a branch-and bound scheme, the main bounds calculation routine is called several times, while the calculation of $\mathcal{H}_1$ and $\mathcal{H}_2$ needs to be done only once. Therefore, in such a situation, it is especially advantageous to use our algorithm.

We also compared the bounds that we get from our technique and those in [11] and [6] with bounds from LP relaxation of an ILP formulation of the RCS problem; our results demonstrate that LP relaxation for RCS produces significantly weaker bounds for many of the cases.

4 Bounds on Time-Constrained Scheduling (TCS)

In this section, we consider the time-constrained scheduling problem. As before, we will consider the bounds that can be obtained by relaxing all the precedence constraints. We will show how the optimal solution can be calculated in the absence of resource constraints using an LP formulation. Then, we show that these bounds are no worse than bounds due to [2] and [12] and therefore that the LP relaxation of an ILP formulation of the TCS problem produces tighter bounds than bounds due to [2, 12].

4.1 Scheduling Independent Operations with Release Times and Deadlines (SIRD)

In this section we consider the following problem: Given a set of independent tasks T, with each task $t_i \in T$ having a release step $\mathsf{Release}(t_i)$ and a deadline $\mathsf{Deadline}(t_i)$, find the minimum number of resources that is required for executing all the tasks within the ranges specified. As before, we initially start by restricting all the operations to execute on the same resource class.

This problem can be modeled and solved using Integer Linear Programming. In this section, we will show that the polyhedra corresponding to this problem is (almost) integral, and that the ceiling of the optimal solution to the linear relaxation of this problem is the optimal solution to this problem; therefore, it is sufficient to solve the linear relaxation problem to find optimal integral solutions to the original problem.

4.2 ILP formulation for this (SIRD) problem

We use a linear programming formulation similar to the one used by Chaudhuri and Walker [1] and Gebotys and Elmasry [3]; this is a restriction of their formulation.

Each task t_i has to be scheduled at some step between $\mathsf{Release}(t_i)$ and $\mathsf{Deadline}(t_i)$, and we have to decide which step to schedule the task in. This is modeled by a set of decision variables $\{x_{i,n_1}, x_{i,n_1+1}, \ldots, x_{i,n_2}\}$, where $n_1 = \mathsf{Release}(t_i)$ and $n_2 = \mathsf{Deadline}(t_i)$. Since the task should only be executed on one of these steps, we have

$$\sum_{\mathsf{Release}(t_i) \le j \le \mathsf{Deadline}(t_i)} x_{i,j} = 1 \qquad (1)$$

This is equivalent to the two constraints

$$\sum_{\mathsf{Release}(t_i) \le j \le \mathsf{Deadline}(t_i)} -x_{i,j} \le -1 \qquad (2)$$

and

$$\sum_{\mathsf{Release}(t_i) \le j \le \mathsf{Deadline}(t_i)} x_{i,j} \le 1 \qquad (3)$$

The number of resources required is the maximum of the number of resources used at any of the steps. If R is the number of resources required, then, we require that the number of resources used at each step j does not exceed R:

$$\forall \text{steps } j \quad \sum_{\{i \mid \mathsf{Release}(t_i) \le j \le \mathsf{Deadline}(t_i)\}} x_{i,j} \le R. \qquad (4)$$

Since the number of resources is a variable whose value has to be determined, we will express Inequality 4 as

$$\left(\sum_{\{i \mid \mathsf{Release}(t_i) \le j \le \mathsf{Deadline}(t_i)\}} x_{i,j} \right) - R \le 0 \qquad (5)$$

Of course, we require that all the decision variables be either 0 or 1; also, we require that the value of R be integral.

We can represent the conditions, i.e., Inequalities 2, 3 and 5, as a matrix inequality of the form $A'\vec{x} \le \vec{b}$ where A' is the coefficient matrix and $\vec{x}$ is the column vector corresponding to the decision variables and R. The first $|T|$ rows of A' represent the coefficients of Inequalities 2, and the next $|T|$ rows represent the coefficients of Inequalities 3. We will let $N = 2 \cdot |T|$. The remaining $Latency$ rows represent the coefficients of Inequalities 5 for the various steps. Let the number of variables (=the number of decision variables+1) be M. Therefore A' is a $(N + Latency) \times M$ matrix and $\vec{x}$ a $M \times 1$ matrix(vector). Without loss of generality, we will assume that the first $M-1$ elements of vector $\vec{x}$ represent the decision variables, and the last element represents R. Then

$$A' = \left(\begin{array}{c:c} A_1 & \vec{0} \\ A_2 & -\vec{1} \end{array} \right) = \left(\begin{array}{c:c} A & B \end{array} \right)$$

where A_1 is an $N \times (M-1)$ matrix and A_2 is a $Latency \times (M-1)$ matrix and A is the $(N + Latency) \times (M-1)$ matrix obtained by deleting the last (i.e., Mth) column from A. Let $\vec{b_1}$ be the column vector of $|T|$ 1's. Let $\vec{b_2}$ the column vector of $Latency$ 0's. We denote by $\vec{b}$ the column vector $\left(-\vec{b_1} \quad \vec{b_1} \quad \vec{b_2} \right)^T$ We denote by X the set of decision variables.

Lemma 4.1. *The matrix A is totally unimodular.*

Proof. See [7] □

Although we have shown that A is TU, A' need not be TU. Next, we show that the optimal solution to the SIRD problem can be obtained by taking the ceiling of the value returned by the LP relaxation of the ILP formulation.

4.3 Obtaining an optimal solution to the SIRD problem

The problem of finding the optimal (minimum) number of resources is equivalent to solving the following ILP model:

$$\text{(I)} \quad \text{Minimize } R \text{ subject to } \begin{cases} \left(\begin{array}{c:c} A & B \end{array} \right) \vec{x} \le \vec{b} \\ R, x \text{ integral } \forall x \in X \end{cases}$$

We denote this problem as (I). The linear relaxation of this problem (LR) is then

$$\text{(LR)} \quad \text{Minimize} \quad R \quad \text{Subject to} \quad \left(\begin{array}{c:c} A & B \end{array} \right) \vec{x} \le \vec{b}$$

Lemma 4.2. *Let r be the optimal solution to (LR). Then $\lceil r \rceil$ is the optimal solution to (I).*

Proof. See [7] □

Hence the values returned by the LP solution to this problems must be no less than the values returned by those bounding techniques. Consider the case when the precedence constraints are added. Then the optimal solution must certainly be larger, because the space has reduced.

Let P denote matrix such that $P\vec{x} \le \vec{q}$ represents the set of precedence constraints (see [3]). We define the following set of bounds:

$$\begin{array}{lllll}
(\mathbf{FB}) & LB_0 & = & \text{Bounds from [2, 9, 10, 12]} & \\
& LB_1 & = & \left\lceil \min_{A\vec{x} \le \vec{c}} R \right\rceil & \\
& LB_2 & = & \min_{A\vec{x} \le \vec{c}, \vec{x} \text{ integral}} R & \\
(\mathbf{LPR}) & LB_3 & = & \left\lceil \min_{A\vec{x} \le \vec{c}, P\vec{x} \le \vec{q}} R \right\rceil & \text{used in [3]} \\
(\mathbf{OPT}) & Sol & = & \min_{A\vec{x} \le \vec{c}, P\vec{x} \le \vec{q}, \vec{x} \text{ integral}} R &
\end{array}$$

Table 2: Comparisons of bounds for the TCS problem

Benchmark	#steps	F-B LB		LP LB		UB	
		$*$	$+$	$*$	$+$	$*$	$+$
FDCT	9	6	3	8	3	8	3
FDCT	13	3	2	4	2	4	2
ARF	13	3	2	4	2	4	2
DCT	8	8	4	8	5	8	5
DCT	9	7	3	8	4	8	4

It is clear that $LB_1 \leq LB_2 \leq Sol$ and $LB_1 \leq LB_3$. We have shown that $LB_1 = LB_2$. It can be shown that the $LB_0 \leq LB_2$. Hence, we must have $LB_0 \leq (LB_1 = LB_2) \leq LB_3 \leq Solution$. Hence we have shown that the ceiling of the solution to the linear relaxation of the ILP formulation is tighter than the bounds obtained using [2, 12].

4.4 Discussion

Chaudhuri and Walker [1] show that their technique generates tighter bounds than those produced by techniques in [2, 9, 10, 12]. They also show that the techniques in [2, 9, 10, 12] generate the same bounds; we will refer to these collectively as Fernandez-Bussell techniques. In this section we showed that LP relaxation is tighter than Fernandez-Bussell [2] and Sharma-Jain [12]; our approach was to cast the TCS bounding problem (SIRD) solved by [2, 12] as an LP instance and then show that the ceiling of the optimal solution to this LP instance is indeed optimal for the TCS bounding problem (SIRD). Both Walker and Chaudhuri [1] and Gebotys and Elmasry [3] have observed that LP relaxation found the exact bounds in all of their benchmark examples. Unlike the bounding techniques in [1, 2, 12] based on precedence relaxation, which solve the bounding problem per resource class, LP relaxation can also be used for lower bound estimates on area of resources. In addition, LP relaxation can generate useful information needed for branch node selection in a branch-and-bound technique, unlike other techniques. Thus, LP relaxation may be advantageous for both initial bound estimation as well as in a brach-and-bound technique. Experimental results comparing the Fernandez-Bussel (F-B LB) and LP relaxation (LP LB) bounds are shown in Table 2, along with the best achievable lower bound (UB). It is seen that in most of the cases, the bounds returned by LP relaxation are the same as those returned by the Fernandez-Bussell techniques. However, there are cases in which LP produces strictly tighter bounds.

In addition, given the tightness of the bounds produced by LP relaxation of the TCS problem—and the noticeable weakness of LP relaxation as a bounding technique for RCS—it may be better to use the TCS problem as the core problem in an exact design space exploration strategy using ILP scheduling in high-level synthesis.

5 Summary

This paper has presented new results on lower bounds for the scheduling problem in high-level synthesis. For the RCS problem, we presented a new way of characterizing lower-bounding, which results in a new bounding technique that is shown to be guaranteed to be no worse than the bounds using the Langevin-Cerny [6] approach, which is turn are no worse than Rim-Jain [11] bounds. For the TCS problem, we showed that by throwing away all precedence constraints, one could derive an LP formulation whose optimal solution is guaranteed to be integral. The solution to this LP problem is shown to be a lower bound for TCS and is guaranteed to be no worse than Fernandez-Bussell [2] or Sharma-Jain [12] bounds.

ACKNOWLEDGMENTS

The work of J. Ramanujam is supported in part by an NSF Young Investigator Award CCR–9457768 and an NSF Research Initiation Award CCR–9210422.

References

[1] S. Chaudhuri and R. Walker. Computing bounds on functional units before scheduling. *IEEE Transactions on Very Large Scale Integration*, 4(2):273–279, June 1996.

[2] E. B. Fernańdez and B. Bussell. Bounds on the number of processors and time for multiprocessor optimal schedule. *IEEE Transactions on Computers*, C-22:745–751, Aug 1973.

[3] C. Gebotys and M. Elmasry. Global optimization approach for architectural synthesis. *IEEE Trans. Computer-Aided Design*, 12(9):1266–1278, September 1993.

[4] Y. Hu and B.S. Carlson. A unified algorithm for the estimation and scheduling of data flow graphs. *Journal of Circuits, Systems and Computers*, 6(3):287–318, 1996.

[5] R. Jain, A. Parker, and N. Park. Predicting area-time tradeoffs for pipelined design. *IEEE Trans. Computer-Aided Design*, 11:955–965, August 1992.

[6] M. Langevin and E. Cerny. A recursive technique for computing lower-bound performance of schedules. *ACM Trans. Design Automation of Electronic Systems*, 1(4):443–456, Oct. 1996.

[7] M. Narasimhan. Exact scheduling techniques for high-level synthesis, 1998. M.S. thesis, Louisiana State University.

[8] M. Narasimhan and J. Ramanujam. A fast approach to computing exact solutions to the resource-constrained scheduling problem. *ACM Trans. Design Automation of Electronic Systems*, to appear (2001).

[9] S. Ohm, F. Kurdahi, and N. Dutt. Comprehensive lower bound estimation from behavioral descriptions. In *Proceedings of the IEEE/ACM International Conference on Computer-Aided Design*, San Jose, Calif., 1994. IEEE Computer Society.

[10] J. Rabaey and M. Potkonjak. Estimating implementation bounds for real time DSP application specific circuits. *IEEE Trans. Computer-Aided Design*, 13:669–683, June 1994.

[11] M. Rim and R. Jain. Lower bound performance estimation for the high-level synthesis scheduling problem. *IEEE Trans. Computer-Aided Design*, 13(4):451–458, April 1994.

[12] A. Sharma and R. Jain. Estimating architectural resources and performance for high-level synthesis applications. *IEEE Transactions on Very Large Scale Integration*, 1(2):175–190, June 1993.