Synthesis of Efficient Constraint Satisfaction Programs

Stephen J. Westfold and Douglas R. Smith
Kestrel Institute
3260 Hillview Avenue
Palo Alto, California 94304

Introduction

Software synthesis is the process of transforming a formal problem specification into software that is efficient and correct by construction. We have used KIDS (Kestrel Interactive Development System) over the last ten years to synthesize very efficient programs for a variety of scheduling problems (Smith, Parra, & Westfold 1996; Smith & Westfold 1995; M.H.Burstein & Smith 1996). The efficiency of these schedulers system is based on the synthesis of specialized constraint management code for achieving arc-consistency. Previous systems for performing scheduling in AI (e.g. (Fox & Smith 1984; Fox, Sadeh, & Baykan 1989; Smith, Fox, & Ow 1986; Smith 1989)) and Operations Research (Applegate & Cook 1991; Luenberger 1989) use constraint representations and operations that are geared for a broad class of problems, such as constraint satisfaction problems or linear programs. In contrast, synthesis techniques can derive specialized representations for constraints and related data, and also derive efficient specialized code for constraint checking and constraint propagation. Synthesis technology allows us to specialize the code by exploiting both problem-independent knowledge (theory of linear programming, finite domains, etc.) as well as problem-specific information obtained from the problem specification.

In this paper we focus on the underlying ideas that lead to the efficiency of our synthesized scheduling programs and explore their generality. Our constraint satisfaction algorithms fall into the class of *global search* algorithms. We have explored this class in more depth elsewhere (Smith 1987; Smith & Westfold 1995). Here we focus on a formulation that uses points in a semilattice to represent solution spaces. This semilattice framework naturally allows for heterogeneous variables (variables of different types), multiple constraints on each variable, indexed variables, conditional constraints and a dynamic set of variables.

The basic idea of global search is to represent and manipulate sets of candidate solutions or *solution spaces*. The principal operations are to *extract* candidate solutions from a solution space and to *split* a space into subspaces. Derived operations include (1) *filters* which are used to eliminate spaces containing no feasible or optimal solutions, and (2) *cutting constraints* that are used to eliminate non-feasible elements from a space of candidate solutions. Global search algorithms work as follows: starting from an initial space

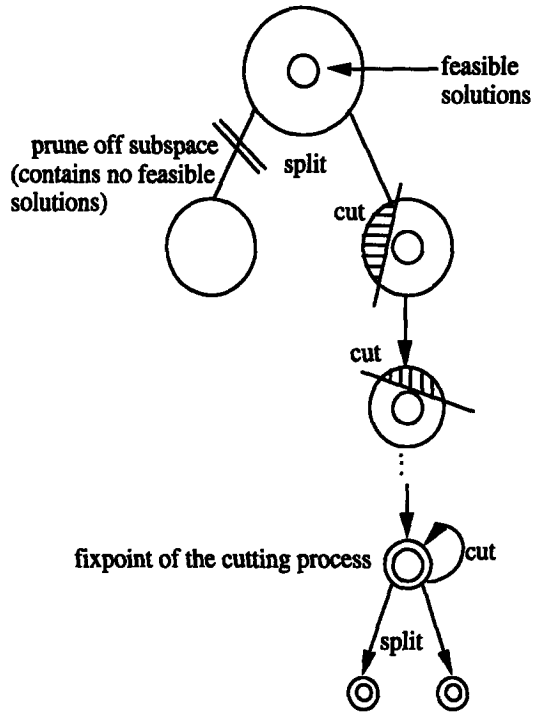

Figure 1: Pruning and Constraint Propagation

that contains all solutions to the given problem instance, the algorithm repeatedly splits spaces into subspaces— *refinements* of the space— eliminates spaces via filters, and contracts spaces by propagating cutting constraints until no spaces remain to be split. The process is often described as a tree (or DAG) search in which a node represents a solution space and an arc represents the split relationship between space and subspace. The filters and constraint propagators serve to prune off branches of the tree that cannot lead to solutions. See Figure 1 which illustrates the working of pruning and constraint propagation on solution spaces.

The key to the efficiency of global search algorithms is the reduction of the size of the search space by an effective representation of solution spaces that allows constraint propagation and pruning at the level of solution spaces rather

than individual concrete solutions.

There is a certain amount of freedom in the formulation of solution spaces. We have found that it is desirable to choose a formulation with a (meet) semilattice structure[1]. A solution space is represented by a point in the semilattice which is the greatest lower bound of the solution space. This structure gives you the property that a solution space for one constraint can be combined with the solution space of another constraint (using the *meet* operation—$\sqcup$) to produce the solution space for the conjunction of the two constraints.

Rehof and Mogensen have shown that satisfiability of sets of *definite* inequality constraints involving monotone functions in a finite meet semilattice can be decided by a linear-time algorithm (Rehof & Mogenson 1999) using a fixpoint iteration. Their notion of *definite inequality* generalizes the logical notion of Horn clauses from the two-point boolean lattice of the truth values to arbitrary finite semilattices. One indication of the power of this class is that arc consistency implies global consistency. We refer to the Rehof and Mogensen algorithm as *algorithm RM*. In CSP terms, a definite inequality problem has the desireable property that arc-consistency implies globally consistency. The RM algorithm is a (hyper)arc-consistency algorithm.

Our work can be viewed as adding disjunctive constraints to their framework[2]. The disjuncts are handled primarily by a search that incrementally generates conjunctive problems that are solved efficiently using essentially the algorithm RM.

Framework

We use the language of first-order predicate calculus for specification. A *binary relation* on a set S is a subset of the product $S \times S$. A *function* f from a set A to a set B, written $f: A \rightarrow B$, is a subset of $A \times B$ such that for each $a \epsilon A$ there is exactly one $b \epsilon B$ with $(a, b) \epsilon f$. In this case we write $f(a) = b$. A definitions of f is presented as follows:

definition $f(x) = e$

where e is an expression that may involve the variable x. If the domain of f is itself a product then f applied to the tuple (a, b) is written $f(a, b)$.

Definition. A binary relation $\sqsubseteq$ defined on a set S is a *partial order* in the set S if the following conditions hold:

$\forall(x)\ x \sqsubseteq x$ **(reflexivity)**
$\forall(x, y)\ (x \sqsubseteq y \wedge y \sqsubseteq x \Rightarrow x = y)$ **(antisymmetry)**
$\forall(x, y, z)\ (x \sqsubseteq y \wedge y \sqsubseteq z \Rightarrow x \sqsubseteq z)$ **(transitivity)**

A function f is *monotone* with respect to the partial order if $\forall(x, y)\ (x \sqsubseteq y \Rightarrow f(x) \sqsubseteq f(y))$.

Definition. Let $\sqsubseteq$ be a partial order on S and T a subset of S. An element p in S is a *least upper bound* of T if $\forall(x)\ (x \epsilon T \Rightarrow x \sqsubseteq p)$ and $\forall(y)\ (y \epsilon S \wedge \forall(x)\ (x \epsilon T \Rightarrow x \sqsubseteq y) \Rightarrow p \sqsubseteq y)$. Similarly an element p in S is a *greatest lower bound* of T if $\forall(x)\ (x \epsilon T \Rightarrow p \sqsubseteq x)$ and $\forall(y)\ (y \epsilon S \wedge \forall(x)\ (x \epsilon T \Rightarrow y \sqsubseteq x) \Rightarrow y \sqsubseteq p)$.

Definition. A set S with partial order $\sqsubseteq$ is a *meet semilattice* iff every pair a, b in S has a least upper bound. $a \sqcup b$ (or $\sqcup(a, b)$) is defined to be the least upper bound of $\{a, b\}$. Similary, it is a *join semilattice* iff every pair a, b in S has a greatest lower bound, and $a \sqcap b$ is defined to be this greatest lower bound. It is a *lattice* if it is both a meet semilattice and a join semilattice.

We refer to meet and join semilattices using the structures $(S, \sqsubseteq, \sqcup)$ and $(S, \sqsubseteq, \sqcap)$ and full lattices using the structure $(S, \sqsubseteq, \sqcup, \sqcap)$. Example lattices are $(Integer, \leq, max, min)$, $(Set, \subseteq, \cup, \cap)$, and the boolean lattice $(\{true, false\}, \Rightarrow, \vee, \wedge)$.

Two semilattices $(S_1, \sqsubseteq_1, \sqcup_1)$ and $(S_2, \sqsubseteq_2, \sqcup_2)$ can be combined to make a *product semilattice* $(S_1 \times S_2, \sqsubseteq_p, \sqcup_p)$ where

definition $x \sqsubseteq_p y = (x.1 \sqsubseteq_1 y.1 \wedge x.2 \sqsubseteq_2 y.2)$
definition $x \sqcup_p y = (x.1 \sqcup_1 y.1, x.2 \sqcup_2 y.2)$

A semilattice $(S, \sqsubseteq, \sqcup)$ or $(S, \sqsubseteq, \sqcap)$ may have *top* ($\top$) and/or *bottom* ($\bot$) elements defined as follows:

$\forall(x)\ (x \epsilon S \Rightarrow x \sqsubseteq \top)$
$\forall(x)\ (x \epsilon S \Rightarrow \bot \sqsubseteq x)$

The computational significance of the meet semilattice structure is that it allows two constraints on a variable X of the form $c_1 \sqsubseteq X$ and $c_2 \sqsubseteq X$, where c_1 and c_2 are constants in a lattice, to be replaced by the equivalent single constraint $c_3 \sqsubseteq X$ where $c_3 = c_1 \sqcup c_2$.

We follow the definitions of Rehof and Mogensen (Rehof & Mogenson 1999) in introducing the concept of *definite inequalities*.

Definition. An inequality is called *definite* if it has the form $\tau \sqsubseteq A$, where A is an atom (a constant or a variable) and τ is a term whose functions are all monotone.

Rehof and Mogensen showed that satisfiability of a set (conjunction) of definite inequalities can be decided in linear time for a meet semilattice domain. For the two-point boolean lattice this is exactly the satisfiability of propositional Horn clauses (HornSAT) problem (since Horn clauses have the form $P_1 \wedge \ldots \wedge P_n \Rightarrow Q$, which is the inequality $P_1 \sqcap \ldots \sqcap P_n \sqsubseteq Q$ in the boolean lattice).

The problem we are considering in this paper is solving sets (conjunctions) of disjunctions of definite inequalities over meet semilattices. This problem is NP-complete with propositional satisfiability (SAT) as an instance.

Although the problem formulation allows any number of variables, only one is necessary, as we can replace n variables by a single variable whose value is the n-tuple of the values of the n variables. The domain of the single variable is the product semilattice of the n semilattices of the domains of the variables. This allows heterogenous problems, where the variable domains are different, to be handled in our framework. Similarly, dynamic problems, where the number of variables constrained can increase during problem solving, can be handled via semilattice structure on dynamic maps and sequences.

For example, consider the heterogeneous problem with two semilattices $(S_1, \sqsubseteq_1, \sqcup_1)$ and $(S_2, \sqsubseteq_2, \sqcup_2)$, and func-

[1] This is not a limiting restriction as it is always possible to convert an arbitrary domain into a lattice by *lifting*: adding a bottom element.

[2] Our work was independent of that of Rehof and Mogensen and focused on the disjunctive aspects of the problem and the synthesis of constraint-solving codes.

tions $f_1 : S_1 \times S_2 \rightarrow S_1$ and $f_2 : S_1 \times S_2 \rightarrow S_2$, with the following two constraints on the variables $A_1 : S_1$ and $A_2 : S_2$.

$$f_1(A_1, A_2) \sqsubseteq_1 A_1$$
$$f_2(A_1, A_2) \sqsubseteq_2 A_2$$

These are equivalent to

$$f_1(A.1, A.2) \sqsubseteq_1 A.1$$
$$f_2(A.1, A.2) \sqsubseteq_2 A.2$$

where $A = (A_1, A_2)$.

The transformed constraint set is not yet in definite form because the right-hand sides are not A but they can be transformed using the following equivalence for a product semilattice $(S_1 \times S_2, \sqsubseteq_p, \sqcup_p)$:

$$\tau \sqsubseteq_i A.i \Leftrightarrow \text{tuple-shad}(A, i, \tau) \sqsubseteq_p A$$

where tuple-shad(tp, i, x) is the tuple tp with the i^{th} component replaced by x.

The transformed constraint set in definite form is

$$\text{tuple-shad}(A, 1, f_1(A.1, A.2)) \sqsubseteq_p A$$
$$\text{tuple-shad}(A, 2, f_2(A.1, A.2)) \sqsubseteq_p A$$

In the examples below we use the component forms of inequalities as they are simpler.

A naïve method for solving our problem of a conjunction of disjunctions of definite inequalities is to distribute conjunction over disjunction to get a disjunction of conjunctions. Algorithm RM can be used to solve each conjunction of definite inequalities independently and the derived solution sets can then be unioned together to create the total solution. However, this is very inefficient in general.

We now describe a global search algorithm for this problem that exploits the incremental nature of algorithm RM: the solution to a definite constraint set S can be used instead of $\perp$ as the starting point for the fixpoint iteration to find the solution to S with an extra definite constraint. Each node in the global search tree consists of a solution to a conjunctive problem plus the set of remaining disjunctions. The top node consists of the solution to the empty problem, $\perp$ of the semilattice, and the initial problem as a set of disjunctions. Children nodes are incrementally refined from their parent by choosing one remaining disjunction and creating a child node for each disjunct by adding the disjunct to the conjunctive problem solved by the parent and iterating to a fixpoint using algorithm RM.

Constraint propagation is used to eagerly reduce the size of remaining disjunct sets, pruning the space when a disjunct set becomes empty and incorporating the remaining disjunct of a unary disjunction into the current conjunctive problem. We add a $\top$ to the semilattice and all its components to represent the space becoming empty which means failure in the search tree. Whenever propagation leads to any component becoming $\top$, then the whole space becomes $\top$ and the branch can be pruned.

We do not discuss the choice of which disjunct set to expand and what order to explore disjuncts, as conventional considerations such as disjunct set size are applicable. In the large scheduling problems we have focused on, there are a large number of very large disjunct sets, so it is desirable to represent them procedurally rather than explicitly.

In the rest of the paper we illustrate the process of deriving a global search algorithm by applying it to an example. The steps are:

1. Specify problem to be solved.
2. Derive reformulation into disjunctions of definite inequalities.
3. Generate code for splitting solution spaces.
4. Generate constraint propagation code.
5. Generate code for extracting solutions for the original problem.

All steps after the problem specification are performed automatically by the KIDS system.

Simple Scheduling Example

We describe a simple transportation scheduling problem that includes essential features found in our real scheduling problems.[3]

The input is a set of *MVRs*, where an MVR is a *MoVement Requirement*—a description of cargo that has to be moved. In this simple version an MVR includes information about when the cargo is available to be moved (*release-time*) and by when it must arrive (*due-time*).

We have a single transportation resource to be scheduled (e.g. a plane), so a schedule is a single sequence of *trips*. Each trip has a start time and a set of MVRs it has been assigned to fulfill: *sched*(i).trip-start is the start time of the i^{th} trip in schedule *sched*, and *sched*(i).trip-MVRs is the manifest of the i^{th} trip in *sched*—the MVRs assigned to that trip.

Figure 2 gives an example of a scheduling problem and a solution schedule. The *release-times* and *due-times* of the five MVRs (labelled for convenient reference *a*, *b*, *c*, *d* and *e*) to be scheduled are plotted above the time axis (the vertical axis has no meaning) and a solution schedule consisting of three trips is plotted beneath it. Trips 1 and 2 start at the earliest possible time (given their manifests) whereas trip 3 starts somewhat later than the earliest possible time. There are many similar solution schedules with slight shifts in the start times.

The specification of valid schedules is

definition **valid-schedules**$(MVRs) =$
$\{sched \mid$ all-MVRs-scheduled$(MVRs, sched)$
$\wedge$ MVRs-ready$(sched)$
$\wedge$ MVRs-due$(sched)$
$\wedge$ trip-separations$(sched)\}$

Every MVR has to be scheduled on some trip[4].

[3]The problem is simplified to the extent that it is not NP-complete. Adding capacity bounds would be sufficient to make finding a feasible solution NP-complete.

[4]This constraint does not preclude an MVR from appearing on more than one trip, but the derived algorithm does not try to make a constraint true if it already is true, so it does not put an MVR on another trip if it is already on one.

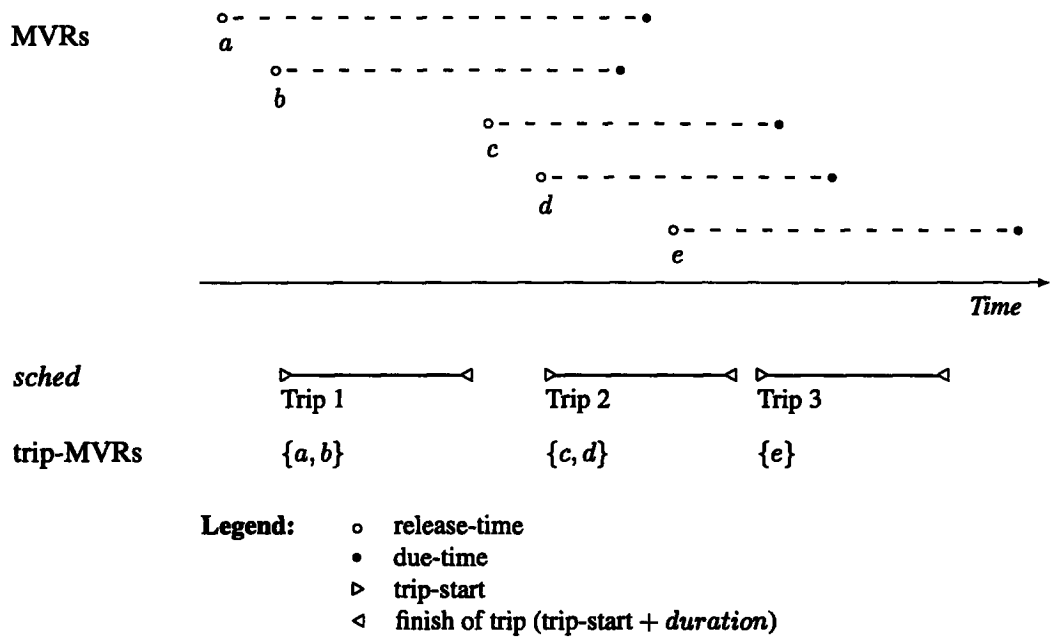

Figure 2: Scheduling Problem and Solution

definition **all-MVRs-scheduled**$(MVRs, sched) =$
$\forall(m : MVR)$
$(m \in MVRs$
$\Rightarrow \exists(i)(i \in \{1..size(sched)\} \wedge m \in sched(i).\text{trip-MVRs}))$

A trip cannot start until after the release dates of all the MVRs assigned to it.

definition **MVRs-ready**$(sched) =$
$\forall(i : Int, m : MVR)$
$(i \in \{1..size(sched)\} \wedge m \in sched(i).\text{trip-MVRs}$
$\Rightarrow \text{release-time}(m) \leq sched(i).\text{trip-start})$

A trip must complete before all the due dates of the MVRs assigned to it.

definition **MVRs-due**$(sched) =$
$\forall(i : Int, m : MVR)$
$(i \in \{1..size(sched)\} \wedge m \in sched(i).\text{trip-MVRs}$
$\Rightarrow \text{due-time}(m) - duration \geq sched(i).\text{trip-start})$

Typically the duration is a function of the properties of a trip, but this has little effect on the derivations in this paper so we make it a constant for simplicity.

A trip cannot begin until the previous trip has ended.

definition **trip-separations**$(sched) =$
$\forall(i : Int)$
$(i \in \{1..size(sched)-1\}$
$\Rightarrow sched(i).\text{trip-start} + duration$
$\leq sched(i+1).\text{trip-start})$

Again, typically there is a gap between the end of one trip and the beginning of the next, but our derivations only depend on the lack of overlap among trips.

Reformulation of Constraints

The scheduling constraints are not disjunctive definite inequalities as specified. The most fundamental problem is that they give both upper and lower bounds on the trip-start component of trips. In other words, the full lattice structure of time is being used. We address this problem by considering the lattice as two semilattices, one for increasing time and the other for decreasing time. We introduce components for greatest lower bound (earliest-trip-start) and least upper bound (latest-trip-start). These are related to trip-start:

$$x.\text{earliest-trip-start} \leq x.\text{trip-start} \leq x.\text{latest-trip-start}$$

From these bounds, we can infer versions of the constraints that have the form of definite inequalities, giving us a reformulated problem.

Figure 3 shows the reformulated version of the problem given in Figure 2 with its solution. The Figure 2 solution can be extracted by taking the earliest start and finish times of the first two trips and a slightly later time for the third trip. The earliest-trip-start of trips 1 and 2 is the same as the release-time of MVRs *b* and *d* respectively, because of the **MVRs-ready** constraint. The earliest-trip-start of trip 3 is the same as the earliest finish time of trip 2 which is later than the release-time of MVR *e* because of the **trip-separations** constraint. Similarly, the latest trip finish time of trips 2 and 3 is the same as the due-time of MVRs c and e respectively, because of the **MVRs-due** constraint, and the latest trip finish time of trip 1 is the same as the latest-trip-start of trip 2 because of the **trip-separations** constraint. The latest-trip-start of a trip is its latest trip finish less *duration*.

The reformulated problem has only one solution for a given assignment of MVRs to trip-MVRS. All solutions to the original problem can be extracted from the solution to

the reformulated problem. Not all values selected from the ranges are compatible, for example choosing the latest-trip-start for trip 1 and the earliest-trip-start for trip 2 is incompatible with the **trip-separations** constraint. However, there is at least one solution for any trip-start chosen from the range of earliest-trip-start and latest-trip-start.

Here is a simplified treatment of the inference for the definition of **MVRs-ready**. Ignoring the antecedent and incorporating the bounds on trip-start, we essentially have

$\forall(i : Int, m : MVR)$
  $(sched(i)$.earliest-trip-start $\leq sched(i)$.trip-start
    $\Rightarrow$ release-time$(m) \leq sched(i)$.trip-start)

then using the general law

$\forall(z)(a \leq z \Rightarrow b \leq z) \;=\; b \leq a$

we obtain the equivalent expression

$\forall(i : Int, m : MVR)$
  release-time$(m) \leq sched(i)$.earliest-trip-start

Using a generalized version of this key step, **MVRs-ready** and **MVRs-due** are easily reformulated to the equivalent definite inequalities

definition **MVRs-ready'**$(sched) =$
 $\forall(i : Int, m : MVR)$
  $(i \in \{1..size(sched)\} \wedge m \in sched(i)$.trip-MVRs
    $\Rightarrow$ release-time$(m) \leq sched(i)$.earliest-trip-start)

definition **MVRs-due'**$(sched) =$
 $\forall(i : Int, m : MVR)$
  $(i \in \{1..size(sched)\} \wedge m \in sched(i)$.trip-MVRs
    $\Rightarrow$ due-time$(m) -$ *duration*
      $\geq sched(i)$.latest-trip-start)

The constraint **trip-separations** is reformulated in a similar manner except that we obtain two constraints because it can be used to get lower bounds on $sched(i{+}1)$.trip-start and upper bounds on $sched(i)$.trip-start. Again ignoring the antecedent and incorporating the bounds, we essentially have

$\forall(i : Int, m : MVR)\ \forall(sched : schedule)$
  $(sched(i)$.earliest-trip-start $\leq sched(i)$.trip-start
    $\leq sched(i)$.latest-trip-start
   $\Rightarrow sched(i)$.trip-start + *duration*
    $\leq sched(i{+}1)$.trip-start)

then we can use monotonicity of $\leq$ on either occurrence of trip-start. Here we apply monotonicity to the first occurrence:

$\forall(i : Int, m : MVR)\ \forall(sched : schedule)$
 $(sched(i)$.earliest-trip-start $\leq sched(i)$.trip-start
  $\Rightarrow sched(i)$.trip-start + *duration*
    $\leq sched(i{+}1)$.trip-start)

$\Longrightarrow$ (monotonicity)

$\forall(i : Int, m : MVR)\ \forall(sched : schedule)$
 $(sched(i)$.earliest-trip-start $\leq sched(i)$.trip-start
  $\Rightarrow sched(i)$.earliest-trip-start + *duration*
    $\leq sched(i{+}1)$.trip-start)

$\Longleftrightarrow$ (using the law $\forall(z)(z \leq a \Rightarrow z \leq b) \;=\; a \leq b$)

$\forall(i : Int, m : MVR)$
  $sched(i)$.earliest-trip-start + *duration*
   $\leq sched(i{+}1)$.earliest-trip-start

The final result is shown below, after we do some normalization, introducing a variable t for the τ expression of the definite inequality. This reduces the cases that need to be considered when deriving specialized constraints (as described in the next section).

definition **trip-separations-ets**$(sched) =$
 $\forall(i : Int, t : Time)$
  $(i \in \{1..size(sched){-}1\}$
    $\wedge\ t = sched(i)$.earliest-trip-start + *duration*
   $\Rightarrow t \leq sched(i{+}1)$.earliest-trip-start)

An analogous derivation stemming from application of monotonicity to the second occurrence of trip-start results in

definition **trip-separations-lts**$(sched) =$
 $\forall(i : Int, t : Time)$
  $(i \in \{1..size(sched){-}1\}$
    $\wedge\ t = sched(i{+}1)$.latest-trip-start $-$ *duration*
   $\Rightarrow t \geq sched(i)$.latest-trip-start)

Generating Splitting Code

The **all-MVRs-scheduled** constraint provides the disjunction that leads to the global search splitting.

$\forall(m : MVR)$
 $(m \in MVRs$
  $\Rightarrow \exists(i)(i \in \{1..size(sched)\} \wedge m \in sched(i)$.trip-MVRs$))$

This is not in the form of a set of disjunctions of definite inequalities. However, it is equivalent to the form

$$\bigwedge_{m \in MVRs} \ \bigvee_{i \in \{1..size(sched)\}} \{m\} \subseteq sched(i).\text{trip-MVRs}$$

which is an indexed conjunction of disjunctions of definite inequalities in the meet semilattice of sets with union as meet for the trip-MVRs component of a trip.

Thus for every MVR there is a disjunction consisting of the MVR being on any one of the trips of the schedule. However, the number of trips is not known in advance. At each point in the elaboration of the search tree, the MVR could be added to an existing trip or a new trip could be created and the MVR added to it. Thus there are two kinds of incremental refinement to the solution space: adding an MVR to a trip and adding a new empty trip (to which we then add an MVR).

These two basic refinements can be expressed:

$sched'(i)$.trip-MVRs $= sched(i)$.trip-MVRs $\cup \{m\}$

$sched' = append(sched, trip_\perp)$ (**append-new-trip**)

where *sched* is the solution space of schedules before the refinement, *sched'* is the solution space of schedules after the refinement, and $trip_\perp$ is the empty trip, the bottom of the semilattice of trips[5].

[5]For completeness it is necessary to also consider the new trip

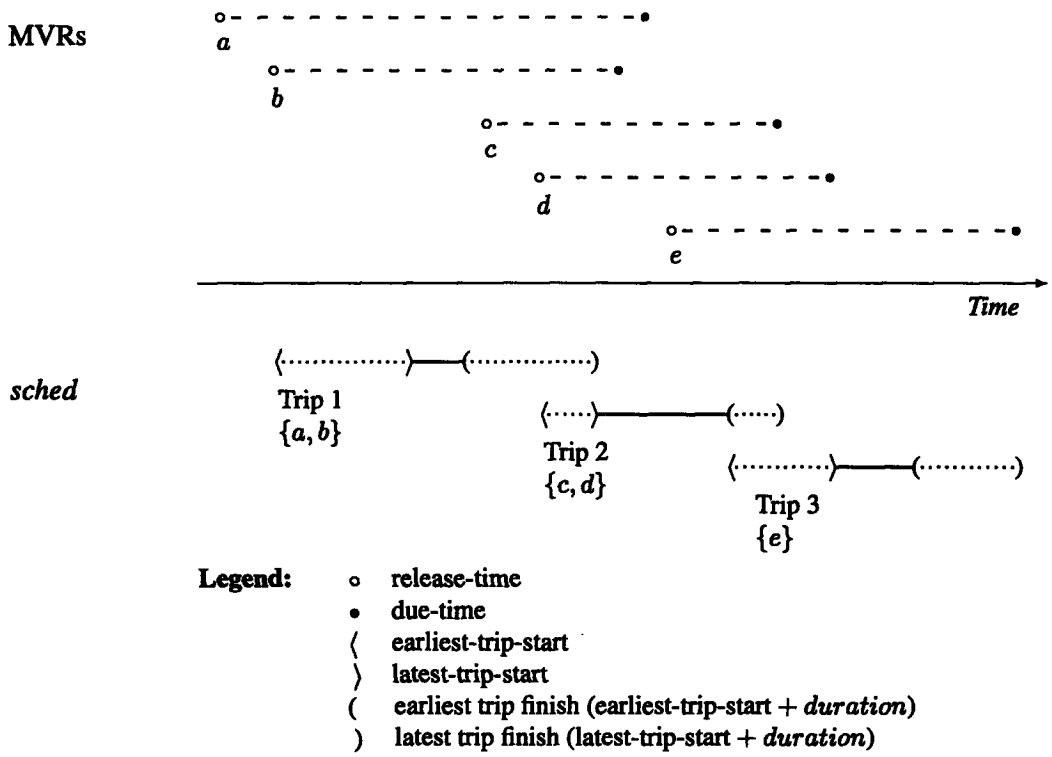

Figure 3: Reformulated Scheduling Problem Solution

To use these refinements it is convenient to have them in the form $A' = f(A)$. The second is already in this form; the first can be converted to

sched' **(update-trip-MVRs)**
= seq-shad(*sched*, *i*,
tuple-shad(*sched*(*i*), trip-MVRs,
sched(*i*).trip-MVRs $\cup$ {*m*}))

where seq-shad(S, i, x) is equal to sequence S except that at index i its value is x.

Specialized Constraint Propagation Code

The constraints are still not in the form of definite inequalities. Apart from **all-MVRs-scheduled** they have the form

$$\forall(x)\ (p(A, x) \Rightarrow x \sqsubseteq A).$$

However this is equivalent to either of the forms

$$reduce(\sqcup, \{x \mid p(A, x)\}) \sqsubseteq A$$

$$\bigwedge_{x \mid p(A,x)} x \sqsubseteq A$$

which are definite inequalities provided that p is monotonic, which is the case for our examples.

being inserted into the sequence, but to reduce the search space we just consider the append case. As we schedule MVRs in order of due date, this incompleteness has not proved to be a problem in practice.

Our derivation of propagation code works with the quantified implication form, the propagation being forward inference. After every refinement one could re-evaluate all the bounds from scratch but for efficiency it is desirable to specialize the constraint to the particular refinement. We exploit the properties that the constraint was true in the previous solution space and that we have just made an incremental refinement of this space.

We generate a procedure for each different kind of incremental refinement to the solution space (such as **append-new-trip** and **update-trip-MVRs**) and include in each procedure constraint checking and propagation code that is specialized to the particular incremental refinement. For example, when appending a new trip to the end of the sequence of trips, the earliest start time of the new trip may be dependent on the earliest start time of the previous trip, whereas the latest start time of the new trip may affect the latest start time of the previous trip. Our concern in this section is how to derive such dependencies automatically.

First we describe an abstract version of the derivations. In the next subsection we give a detailed derivation of specializing a particular constraint, and then we give brief derivations of specialization of the other constraints.

Consider an incremental refinement to A by some function g:

$$A' = g(A) \qquad (so\ A \sqsubseteq A')$$

and a definite constraint C on A given by

$$C(A) = \forall(x)\,(p(A,x) \Rightarrow x \sqsubseteq A)$$

We assume that the constraint C is true for the initial space A and we want it to be true in the refined space A'. $C(A)$ is true if for all x, $p(A,x)$ is false or $x \sqsubseteq A$ is true. In the latter case we have simply that $x \sqsubseteq A'$ by transitivity, so we can use A as an initial approximation of A'. To find additional values of x that belong in A' we assume the former case: we simplify $C(A')$ under the assumption that $p(A,x)$ is false. We call this simplified $C(A')$ the *residual* constraint for C given the refinement.

For the common case where p is a conjunction, this does not work well in practice because the negation is a disjunction which is difficult to use in simplification. Instead we treat each conjunct separately and combine the results as follows: if $p(A,x) = (p_1(A,x) \wedge p_2(A,x))$ and the residuals for $p_1(A,x)$ and $p_2(A,x)$ are $h_1(A,x)$ and $h_2(A,x)$ respectively then the full residual constraint is

$$\forall(x)\,(h_1(A,x) \wedge p_2(A',x) \Rightarrow x \sqsubseteq A')$$
$$\wedge\, \forall(x)\,(p_1(A',x) \wedge h_2(A,x) \Rightarrow x \sqsubseteq A')$$

If a constraint is unaffected by the refinement then its residual is *false*.

Specializing MVRs-ready'

We now use this derivation scheme to derive the residual constraint for **MVRs-ready'**:

$\forall(i : Int, m : MVR)$
$(i \in \{1..size(sched)\} \wedge m \in sched(i).\text{trip-MVRs}$
$\Rightarrow \text{release-time}(m) \leq sched(i).\text{earliest-trip-start})$

given the **update-trip-MVRs** refinement

$sched'$ **(update-trip-MVRs')**
$= \text{seq-shad}(sched, i_c,$
$\text{tuple-shad}(sched(i_c), \text{trip-MVRs},$
$sched(i_c).\text{trip-MVRs} \cup \{m_c\}))$

The variables are subscripted with a c to avoid name conflict in the derivation and to emphasize that in this context they have already been bound to particular values.

The residual for the conjunct $i \in \{1..size(sched)\}$ is *false* because $size(sched')$ simplifies to $size(sched)$ as seq-shad does not affect the size of the sequence.

Now we focus on the second conjunct,

$m \in sched'(i).\text{trip-MVRs}$

simplifying it under the assumption:

$m \in sched(i).\text{trip-MVRs} = false$ **(negation-assumption)**

We require the following rules:

(seq-shad-application)
$\text{seq-shad}(S, i, x)(j) = \textbf{if } i = j \textbf{ then } x \textbf{ else } S(j)$

$\text{tuple-shad}(x, f, y).f = y$ **(tuple-shad-deref)**

$(\textbf{if } x \textbf{ then } y \textbf{ else } false) = (x \wedge y)$ **(if-then-else-false)**

The simplification goes:

$m \in sched'(i).\text{trip-MVRs}$
{substitute **update-trip-MVRs'**}
$= m \in \text{seq-shad}(sched, i_c,$
$\text{tuple-shad}(sched(i_c), \text{trip-MVRs},$
$sched(i_c).\text{trip-MVRs} \cup \{m_c\}))$
$(i).\text{trip-MVRs}$
{**seq-shad-application**}
$= m \in (\textbf{if } i = i_c$
$\textbf{then } \text{tuple-shad}(sched(i_c), \text{trip-MVRs},$
$sched(i_c).\text{trip-MVRs} \cup \{m_c\})$
$\textbf{else } sched(i)).\text{trip-MVRs})$
{move **if** to top-level then use **tuple-shad-deref**}
$= \textbf{if } i = i_c \textbf{ then } m \in sched(i_c).\text{trip-MVRs} \cup \{m_c\}$
$\textbf{else } m \in sched(i).\text{trip-MVRs}$
{**negation-assumption** and **if-then-else-false**}
$= (i = i_c \wedge m \in sched(i_c).\text{trip-MVRs} \cup \{m_c\})$
{distribute $\in$ over $\cup$}
$= (i = i_c \wedge (m \in sched(i_c).\text{trip-MVRs} \vee m \in \{m_c\}))$
{**negation-assumption** and simplification}
$= (i = i_c \wedge m = m_c)$

The full residual constraint becomes, after simplification:

$\text{release-time}(m_c) \leq sched'(i_c).\text{earliest-trip-start}$

The proceduralization of this residual is

$\textbf{if } \neg\text{release-time}(m_c) \leq sched'(i_c).\text{earliest-trip-start}$
$\textbf{then } sched(i_c).\text{earliest-trip-start} \leftarrow \text{release-time}(m_c)$

Thus we have a third incremental refinement given by the equation

$sched'$ **(update-earliest-start)**
$= \text{seq-shad}(sched, i_c,$
$\text{tuple-shad}(sched(i_c), \text{earliest-trip-start}, t_c))$

where the refinement has been abstracted by introducing the variable t_c for $\text{release-time}(m_c)$.

Specializing trip-separations-ets

Now we consider specializing the **trip-separations-ets** constraint.

$\forall(i : Int, t : Time)$
$(i \in \{1..size(sched)-1\}$
$\wedge\, t = sched(i).\text{earliest-trip-start} + duration$
$\Rightarrow t \leq sched(i+1).\text{earliest-trip-start})$

It is only affected by **append-new-trip** and **update-earliest-start**.

Consider **append-new-trip**. The residual for the first conjunct, $i \in \{1..size(sched)-1\}$ is $i = size(sched)$. The full residual constraint becomes, after simplification:

$sched(size(sched)).\text{earliest-trip-start} + duration$
$\leq sched'(size(sched)+1).\text{earliest-trip-start}$

The residual for the second conjunct

$t = sched(i).\text{earliest-trip-start} + duration$

is

$i = size(sched) + 1$
$\wedge\, t = trip_{\perp}.\text{earliest-trip-start} + duration$

but the full residual constraint simplifies to *false* because the first conjunct becomes $i \in \{1..size(sched)\}$ which is inconsistent with $i = size(sched) + 1$.

The residual from the first conjunct has the same form as **update-earliest-start** so we can reuse the same refinement procedure, although passing different arguments.

Now we consider the effect of **update-earliest-start** on **trip-separations-ets**.

The residual for the first conjunct is *false* because the size of the schedule is unaffected. The residual for the second conjunct is $i = i_c \wedge t = t_c + duration$. The full residual constraint becomes, after simplification:

$i_c < size(sched)$
$\Rightarrow t_c + duration \leq sched'(i_c{+}1)$.earliest-trip-start

Again, except for the conditional, this has the same form as **update-earliest-start** so the same procedure can be used. The complete refinement procedure for **update-earliest-start** is as follows:

function update-earliest-start(i_c, t_c, *sched*) =
 if $t_c \leq sched(i_c)$.earliest-trip-start
  then *sched* *% Old bound is tighter*
 else if $\neg t_c \leq sched(i_c)$.latest-trip-start
  then $\top$ *% Fail because earliest is after latest*
 else *% Perform update and propagate*
 let (*sched'* =
  seq-shad(*sched*, i_c,
   tuple-shad($sched(i_c)$, earliest-trip-start, t_c)))
 if $i_c < size(sched)$
  then update-earliest-start($i_c + 1$, $t_c + duration$, *sched'*)
  else *sched'*

The specialization of **trip-separations-lts** is similar to that of **trip-separations-ets**.

Summary of Generated Propagation Code

The basic search routine generates new subspaces by calls to **append-new-trip** and **update-trip-MVRs**. The former has a call to **update-earliest-start** which initializes the earliest-start of the new trip based on the earliest start of the previous trip. **update-trip-MVRs** has calls to **update-earliest-start** and **update-latest-start** based respectively on the release-time and due-time of the MVR being added. **update-earliest-start** is a linear recursion that propagates a change to the earliest-start of a trip to earliest-starts of subsequent trips until one does not need to be updated because the separation is already adequate. **update-latest-start** is similar, but propagates from the latest-start of a trip to latest-starts of previous trips until one does not need to be updated.

This propagation code is very efficient, performing very few unnecessary tests. It is also space-efficient as the constraints are represented procedurally, and the solution spaces are represented intensionally by bounds. We have used a depth-first propagation control-structure because the dependency structure for this problem is a tree. In general, a breadth-first control structure is necessary to avoid unnecessary work when the constraint interactions are more complicated.

Extracting a Solution

For this problem the extraction process is straightforward. The only issue is extracting valid trip-starts given earliest- and latest-trip-starts. One cannot arbitrarily choose values from the intervals to get a valid schedule. For example, choosing the latest-trip-start from one trip and the earliest-trip-start from the next trip will likely violate the **trip-separation** constraint. Always choosing the earliest-trip-starts or always choosing the latest-trip-starts gives a valid schedule. A more flexible strategy is to choose a trip-start between the earliest- and latest-trip-starts for some trip, then call the procedures **update-earliest-start** and **update-latest-start** with the chosen value, so that the consequences of the choice are propagated. Then this choose-and-propagate process can be repeated for other trips until a trip-start has been chosen for each trip.

In general, it is possible that there is no valid solution to the original problem within a non-empty valid subspace (one that satisfies the transformed problem). If a single solution is required then one must enumerate valid subspaces until one is found that has at least one extractible solutions. If all solutions are required then all valid subspaces must be enumerated. If a minimal cost solution is required then a similar enumeration is necessary except we can add the constraint that the cost of a solution has to be less than the cost of the current best solution.

Related Work

Mackworth (Mackworth 1992) gives a characterization of constraint-satisfaction problems in relation to various logical representation and reasoning systems such as Horn First-Order Predicate Calculus and Constraint Logic Programs. Our system does not fit within his framework. Our work, along with that of Rehof and Mogensen, suggests the addition of an extra dimension to the framework in which the boolean lattice is generalized to an arbitrary lattice.

Our problem reformulation of replacing trip-start by upper and lower bounds on trip-start is similar to work that uses interval techniques (Van Hentenryck, McAllester, & Kapur 1995; Benhamou, McAllester, & Van Hentenryck 1994; Hyvnen 1992). Our framework can be seen as a generalization of these approaches to work with arbitrary kinds of bounds.

Our model of constraint propagation generalizes the concepts of cutting planes in the Operations Research literature (Nemhauser & Wolsey 1988) and the forms of propagation studied in the constraint satisfaction literature (e.g. (Van Hentenryck 1989)). Our use of fixed-point iteration for constraint propagation is similar to Paige's work on fixed-point iteration in RAPTS (Cai & Paige 1989). The main differences are (1) RAPTS expects the user to supply the monotone function as part of the initial specification whereas we derive it from a more abstract statement of the problem; (2) RAPTS instantiates a straightforward iteration scheme and then performs optimizations. Such an approach would be too inefficient for scheduling applications, so we use dependence analysis to generate code that is specific to the constraint system at hand. RAPTS uses finite differencing in order to make the iteration incremental. We have incorporated this into our framework and used it in more complex scheduling problems.

Discussion and Further Work

The main focus of our work has been on the synthesis of high-performance constraint-solving codes – some of the schedulers that we have synthesized using KIDS run several orders of magnitude faster than manually written schedulers

for the same problem. We believe that the speed is due to the specialized representation of constraints and the ability to optimize the propagation codes at design time. We have used KIDS to synthesize a variety of scheduling applications including ITAS (a theater airlift scheduler) (M.H.Burstein & Smith 1996) and the CAMPS Mission Planner (Emerson & Burstein 1999) which plans strategic airlift missions for the Air Mobility Command at Scott AFB. The speed of the generated scheduling algorithm has allowed us to tackle very complex constraint systems. For example, the CAMPS Mission Planner involves the routine scheduling of thousands of airlift missions, and the simultaneous handling of many classes of resource: aircraft and their configurations, crews and their duty days, fuel, parking capacity at ports, working and throughput capacity at ports, runway events, and others. Aircraft, crews and ports each have many capacity and usage constraints that must be modeled. Every time a scheduling decision is made, it is propagated through the constraint network to decide if it entails any inconsistency.

We have described our work as a generalization of various frameworks. These frameworks have been developed in more depth than ours so there are many opportunities to see how ideas developed in these frameworks can be carried over. Our focus has been mainly on scheduling algorithms. We have looked briefly at problems such as integer linear programming, but not sufficiently to make a good estimate of their potential compared to other methods. Working from such a model is unlikely to give the best results as these models capture information in a very limited lattice structure. Frequently problems must be reformulated to get them into the form required by these general methods. Working from the original problem we may be able to capture more of its structure in lattices and so get a smaller search space. On the other hand, it may be possible to recover lattice structure information from analysis of an integer linear programming problem specification.

Our framework is not directly applicable to incremental rescheduling if constraints are deleted as well as added. A possible way to handle the removal of a constraint is to follow dependencies of the deleted constraints to see which values may have depended on them. These values can be relaxed so that the space is large enough to find a solution when the new constraints are added.

References

Applegate, D., and Cook, W. 1991. A computational study of the job-shop scheduling problem. *ORSA Journal on Computing* 3(2):149–156.

Benhamou, F.; McAllester, D.; and Van Hentenryck, P. 1994. Clp(intervals) revisited. Technical Report CS-94-18, Department of Computer Science, Brown University.

Cai, J., and Paige, R. 1989. Program derivation by fixed point computation. *Science of Computer Programming* 11:197–261.

Emerson, T., and Burstein, M. 1999. Development of a constraint-based airlift scheduler by program synthesis from formal specifications. In *Proceedings of the Fourteenth Automated Software Engineering Conference*. IEEE Computer Society Press.

Fox, M. S., and Smith, S. F. 1984. ISIS – a knowledge-based system for factory scheduling. *Expert Systems* 1(1):25–49.

Fox, M. S.; Sadeh, N.; and Baykan, C. 1989. Constrained heuristic search. In *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, 309–315.

Hyvnen, E. 1992. Constraint reasoning based on interval arithmetic. the tolerance propagation approach. *Artificial Intelligence* 58:71–112. Also in: Freuder, E., Mackworth, A., (Eds.), Constraint-Based Reasoning, MIT Press, Cambridge, USA, 1994.

Luenberger, D. G. 1989. *Linear and Nonlinear Programming*. Reading, MA: Addison-Wesley Publishing Company, Inc.

Mackworth, A. K. 1992. The logic of constraint satisfaction. *Artificial Intelligence* 58:3–20.

M.H.Burstein, and Smith, D. 1996. ITAS: A portable interactive transportation scheduling tool using a search engine generated from formal specifications. In *Proceedings of the Third International Conference on Artificial Intelligence Planning System (AIPS-96)*.

Nemhauser, G. L., and Wolsey, L. A. 1988. *Integer and Combinatorial Optimization*. New York: John Wiley & Sons, Inc.

Rehof, J., and Mogenson, T. 1999. Tractable constraints in finite semilattices. *Science of Computer Programming* 35:191–221.

Smith, D. R., and Westfold, S. J. 1995. Synthesis of constraint algorithms. In Saraswat, V., and Van Hentenryck, P., eds., *Principles and Practice of Constraint Programming*. Cambridge, MA: The MIT Press.

Smith, S. F.; Fox, M. S.; and Ow, P. S. 1986. Constructing and maintaining detailed production plans: Investigations into the development of knowledge-based factory scheduling systems. *AI Magazine* 7(4):45–61.

Smith, D. R.; Parra, E. A.; and Westfold, S. J. 1996. Synthesis of planning and scheduling software. In Tate, A., ed., *Advanced Planning Technology*, 226–234. AAAI Press, Menlo Park.

Smith, D. R. 1987. Structure and design of global search algorithms. Technical Report KES.U.87.12, Kestrel Institute.

Smith, S. F. 1989. The OPIS framework for modeling manufacturing systems. Technical Report CMU-RI-TR-89-30, The Robotics Institute, Carnegie-Mellon University.

Van Hentenryck, P.; McAllester, D.; and Kapur, D. 1995. Solving polynomial systems using a branch and prune approach. *SIAM Journal of Numerical Analysis*.

Van Hentenryck, P. 1989. *Constraint Satisfaction in Logic Programming*. Cambridge, MA: Massachusetts Institute of Technology.