Programming Finite-Domain Constraint Propagators in Action Rules[1]

Neng-Fa Zhou
Department of Computer and Information Science
Brooklyn College & Graduate Center
The City University of New York
New York, NY 11210-2889, USA
zhou@sci.brooklyn.cuny.edu

Abstract

In this paper, we propose a new language construct, called AR (*Action Rules)*, and describe how various propagators for finite-domain constraints can be implemented in it. An action rule specifies a pattern for agents, an action that the agents can carry out, and an event pattern for events that can activate the agents. AR combines the goal-oriented execution model of logic programming with the event-driven execution model. This hybrid execution model facilitates programming constraint propagators. A propagator for a constraint is an agent that maintains the consistency of the constraint and is activated by the updates of the domain variables in the constraint. AR has a much stronger descriptive power than *indexicals*, the language widely used in the current finite-domain constraint systems, and is flexible for implementing not only interval-consistency but also arc-consistency algorithms. As examples, we present the implementation of a weak arc-consistency algorithm for the `all_distinct` constraint and a hybrid algorithm for n-ary linear equality constraints. B-Prolog has been extended to accommodate action rules. Benchmarking shows that the performance of B-Prolog as a CLP(FD) system compares favorably with the fastest systems available now.

1 Introduction

CLP(FD), the constraint logic programming language over finite-domains, has been proved effective for solving a large number of real-life optimization problems [6, 12]. The key operation employed in CLP(FD) is called *constraint propagation* [13, 23], which uses constraints actively to prune search spaces as follows: whenever a variable changes, i.e., the variable has been instantiated or its domain has been updated, the domains of all the remaining variables are filtered to contain only those values that are consistent with this variable. There may exist different propagation rules for a constraint depending on the level of consistency to be achieved. Constraint propagation has been employed to solve not only constraints over finite-domains but also constraints over trees, lists, finite sets, floating-point numbers, and many other domains [12].

In early CLP(FD) systems, such as the CHIP system [7], constraints are interpreted rather than compiled. Constraints are first transformed into canonical-form

[1]TR-2002010, CUNY Computer Science, 2002.

terms and are then executed by an interpreter that performs, among other things, constraint propagation. The propagation procedure adopted is general enough for handling all types of constraints. Learning from the experience of compiling Prolog programs into the Warren Abstract Machine (WAM) [26], a former research group at ECRC extended the WAM for compiling CLP(FD) [1]. The CHIP compiler compiles constraints into low-level instructions such that different specialized propagation procedures are used for different types of constraints. This black-box approach has proved problematic because it is too complicated and lacks flexibility and extendibility. The extended WAM in the CHIP system [1] has over 100 instructions for compiling finite-domain constraints alone!

A language construct, called *indexicals*, has been quite popular as an intermediate language for compiling finite-domain constraints. The language was first proposed by Hentenryck et al [25] and then popularized by [3]. This language is also adopted by other systems [2, 22]. An indexical is a primitive constraint in the form of X *in* r, where X is a domain variable and r specifies the range for X. For each indexical, a propagation procedure specific to it is used. Indexicals are claimed to be a glass-box approach to compiling constraints in contrast with the black-box CHIP compiler. Nevertheless, as the delaying mechanism is embedded in range expressions, indexicals are not as open as claimed. Indexicals can be used to compile arithmetic constraints, but are too weak to be used to program many other kinds of propagators.

CHR (Constraint Handling Rules) [8] may currently be the most powerful implementation language for constraints. It can be used to program not only constraint propagators but also constraint reasoning rules. CHR has been implemented and integrated with Eclipse, Sicstus, and Oz. CHR resembles a production system. In CHR, the left-hand side of a rule specifies a pattern of constraints in the constraint store and the right-hand side specifies new constraints to replace those on the left-hand side or to be added into the store. The left-hand side of a rule may have multiple constraint patterns. This feature is helpful for reasoning about the constraint store. For example, $A > B$ & $B > C \rightarrow A > C + 1$ is a CHR rule that generates the constraint $A > C + 1$, which is helpful albeit redundant. The strong descriptive power, however, is not offered without cost. For CHR, a sophisticated matching algorithm is needed to match constraint patterns against the constraint store. Now, constraint solvers implemented in CHR are still an order of magnitude slower than constraint interpreters implemented in C [11].

This paper proposes a new language construct, called AR (*Action Rules*), that can be used to program event-handling in general and constraint propagation in particular. An action rule specifies a pattern for agents, an action that the agents can carry out, and an event pattern for events that can activate the agents. An agent is like a subgoal in Prolog but behaves in an event-driven manner. An agent can be suspended when certain conditions on it are satisfied and can be activated when certain events are posted. AR is an extension of delay constructs such as delay clauses [15] that allows for the descriptions of not only delay conditions but also activating events and actions [29]. The syntax and semantics of AR will be described in Section 3.

The focus of this paper is on how to implement various propagators for finite-domain constraints in AR. A propagator for a constraint is an agent that maintains the consistency of the constraint and is activated by the updates of the domain variables in the constraint. In Section 4, we present propagators for binary, non-binary, and the global constraint `all_distinct`. AR is more expressive than indexicals. Some of the propagators presented such as the one for maintaining arc-consistency for binary equality constraints and the one for maintaining weak arc-consistency for `all_distinct` cannot be implemented in indexicals as efficiently.

B-Prolog has been extended to accommodate AR and several constraint solvers including the ones over finite-domain, Boolean, trees, and sets have been developed in AR [30]. Section 5 compares the performance of the finite-domain solver of B-Prolog with GNU-Prolog (GP), a state-of-the-art implementation of CLP(FD) [5] and two other CLP(FD) systems. The benchmarking results show that B-Prolog is faster than GP for most of the benchmarks and on average as well.

Readers are assumed to be familiar with logic programming and constraint satisfaction, but no knowledge about the compilation is assumed. In Section 2, we define some preliminary terms and concepts about CLP(FD) and constraint propagation. Readers are referred to [14], [24] and [13] for the details. A brief description of the implementation of AR is given in Section 5, and a more detailed description can be found in [31].

2 Preliminaries

2.1 CLP(FD)

CLP(FD) [24] is an extension of Prolog that supports built-ins for specifying domain variables, constraints, and strategies for instantiating variables.

The domains of variables are declared as follows:

```
Vars in D
```

where `Vars` is a variable or a list of variables, and `D` is a list of ground terms or a range between two integers $l..u$. A domain variable is normally represented as a Prolog variable with attributes. A CLP(FD) system provides primitives for accessing and updating attribute values.

A CLP(FD) system provides *equality* (=), *disequality* ($\neq$), and *inequality* constraints. In addition, a CLP(FD) system also provides some other constraints such as global constraints. The global constraint `all_distinct(L)` ensures that the elements in the list `L` must be all pair-wisely different.

2.2 Constraint propagation

Constraint Propagation [13, 23] is a key operation employed in CLP(FD) systems for maintaining the consistency of constraints. The basic idea of constraint propagation is to activate the propagators of constraints whenever the domains of the variables in the constraints are updated. Propagating the updates to other variables may result in the shrinking of the domains of the variables or the instantiation of the variables.

There are different levels of consistency for constraints such as *node*, *interval*, *bounds*, *arc*, and *path* consistency [23, 14]. We define below three levels of consistency needed in this paper, namely node, interval and arc consistency, and define the propagators that maintain them.

A unary constraint *p(X)* where X has the domain D is said to be *node-consistent* if for any element x in D *p(x)* is satisfied.

$$\forall_{x \in D} p(x)$$

For example, for the equality constraint $X = Y + 1$, when X is instantiated to 3, Y must be instantiated to 2 to make the constraint node-consistent. As another example, for the disequality constraint $X \neq Y$, when X is instantiated to 3, 3 must be excluded from the domain of Y to make the constraint node-consistent. The propagation rule that maintains node-consistency is called *forward checking.* A propagator for a constraint that performs forward checking is activated whenever the constraint becomes unary.

Let C be an linear equality constraint $f(X_1, X_2, ..., X_n) = 0$ where X_i is defined over the domain D_i $(i = 1, ..., n)$. Suppose $X_i = g_i(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)$, the inverse function of f with respect to X_i, and the functions l and u are defined as follows:

$$l(g_i(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)) = min\{g_i(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) | x_1 \in D_1 \ldots x_n \in D_n\}$$

$$u(g_i(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)) = max\{g_i(x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) | x_1 \in D_1 \ldots x_n \in D_n\}$$

The constraint C is said to be *interval consistent w.r.t.* X_i if:

$$\forall_{x \in D_i}(l(g_i(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)) \leq x \leq u(g_i(X_1, ..., X_{i-1}, X_{i+1}, ..., X_n)))$$

To make the constraint interval consistent w.r.t. X_i, we have to exclude all the elements from D_i that are not in the range. The constraint C is said to be *interval consistent* if C is interval consistent w.r.t. all the variables. For example, the constraint $X = Y + 1$, where X and Y have the domain 1..5, is not interval-consistent. To make the constraint interval-consistent, we have to exclude 1 from the domain of X and 5 from the domain of Y. Propagators for maintaining interval consistency are activated whenever a bound of a variable is updated or whenever a variable is instantiated. The definition can be easily extended to inequality constraint $f(X_1, X_2, ..., X_n) \geq 0$.

Consider a binary constraint $p(X, Y)$ where X and Y are defined over the domains D_x and D_y, respectively. The constraint is said to be *arc-consistent w.r.t.* X if for any element in D_x there exists a supporting element in D_y such that the constraint is satisfied:

$$\forall_{x \in D_x} \exists_{y \in D_y} p(x, y)$$

Similarly, the constraint is arc-consistent w.r.t. Y if for any element in D_y there exists a supporting element in D_x such that the constraint is satisfied:

$$\forall_{y \in D_y} \exists_{x \in D_x} p(x, y)$$

The constraint is *arc-consistent* if it is arc-consistent w.r.t. both X and Y. For example, the equality constraint $X = Y + 1$ ($X \in \{2, 4, 5\}$, $Y \in \{1..4\}$) is not arc-consistent since there is no element in the domain of X that supports 2 in the domain of Y. To make the constraint arc-consistent, we must exclude 2 from the domain of Y. Propagators for maintaining arc-consistency are triggered whenever changes occur to the domain of a variable. It is very costly to maintain arc-consistency for arbitrary constraints. For this reason, some CLP(FD) systems maintain arc-consistency only for binary constraints and many others do not consider arc-consistency at all.

2.3 Domain variables

A *domain variable* is a suspension variable to which there are suspended propagators and some other information attached. A domain variable is represented in B-Prolog as a record that has the following fields:

`ref`	reference to the value
`type`	type of the domain
`min`	minimum element in the domain
`max`	maximum element in the domain
`size`	number of elements that remain in the domain
`ins_cs`	list of propagators to be executed when the variable is instantiated
`bound_cs`	list of propagators to be executed when a bound is updated
`dom_cs`	list of propagators to be executed when an inner element is excluded
`elms`	pointer to the bit vector representation of the elements

where `ref` refers to the variable itself if the variable is not instantiated, and `elms` is a pointer to a bit vector that represents the elements. When a domain is an interval without holes, no bit vector is necessary and `elms` is a null pointer.

The following built-in predicates and functions are available on domain variables.

- `dvar(X)`: Succeeds if `X` is a domain variable.
- `min(X)`, `max(X)`: Functions that return, respectively, the minimum and maximum elements of the domain of `X`.
- `size(X,Size)`: The size of the domain of `X` is `Size`.
- `exclude(X,E)`: Excludes the element `E` from the domain of `X`.
- `X in D`: The new domain for `X` is the intersection of its existing domain and `D` where `D` is a set of ground terms or a range `l..u` of integers. .

A failure occurs when the domain of a variable becomes empty. When the domain of a variable becomes a singleton, the variable will be instantiated to the element automatically.

An event is posted whenever the domain of a variable is updated. For a domain variable `X`, instantiating `X` posts the event `ins(X)`, updating the lower or upper bound of the domain posts the event `bound(X)`, and excluding an inner element `E` from the domain posts the event `dom(X,E)`. Notice that the event `bound(X)` is not posted when `X` is instantiated and the event `dom(X,E)` is not posted when either bound of the domain of `X` is updated. This implies that a propagator that maintains arc-consistency has to handle not only `dom(X,E)` events but also `bound(X)` and `ins(X)` events.

Each event on a domain variable activates its corresponding list of propagators. The event `ins(X)` activates the propagator list `ins_cs` of `X`, `bound(X)` activates the list `bound_cs`, and `dom(X,E)` activates the list `dom_cs`.

3 The AR Language

AR is designed for programming active agents. An agent is like a subgoal in Prolog but behaves in an event-driven manner. In this section, we describe the syntax and operational semantics of AR.

3.1 Syntax

An *action rule* takes the following form:

```
<Agent> <Condition> {<Event>} '=>' <Action>
```

where `Agent` is an atomic formula that represents a pattern for agents, `Condition` is a conjunction of conditions on the agents, `Event` is a disjunction of patterns for events that can activate the agents, and `Action` is a sequence of actions performed by the agents when they are activated. An action rule is called a *commitment rule* if `Event` together with the enclosing braces are missing. Conditions, event patterns, and actions are all atomic formulas where the delimiter ',' is used to separate the constituents.

An *agent definition* consists of a sequence of rules, and a *program* consists of a sequence of *agent definitions*. Agents correspond to subgoals in Prolog, and agent definitions correspond to predicates. In this paper, we use the term *agents* for subgoals that can be suspended and activated.

All conditions must be *in-line*[2] *tests*[3]. In the implementation of AR in B-Prolog, the following types of conditions are allowed:

- Type and mode checking: Predicates like `integer(X)`, `var(X)`, and `nonvar(X)`.
- Matching: A matching call takes the form of `X=Y` where one of the arguments must be a non-variable term and the other must be a variable that occurs before in the rule. The non-variable term serves as a pattern and the variable refers to a term to be matched against the pattern. This call succeeds if the pattern and the term become identical after a substitution is applied to the pattern. For instance, the condition `f(X)=Y` succeeds if `Y` is a structure whose functor is `f/1`.

[2]An in-line call is compiled into instructions that do not invoke any predicates.

[3]A test does not change the instantiation status of the variables in its arguments.

- Term inspection: Several built-in predicates including `arg/3`, `functor/3`, `==/2`, `\==/2`, and `n_vars_gt/2` can be used in the condition of a rule to inspect the arguments of an agent. The call `n_vars_gt(Term,N)` succeeds if the number of variables occurring in `Term` is greater than `N`.

- Arithmetic comparison: Checks the arithmetic equality (`=:=`), disequality (`=\=`), or inequality (`>`, `>=`, `<`, and `=<`) of two terms which must be ground at runtime.

A set of built-in events is provided.[4] As far as programming constraint propagators is concerned, an event pattern can be one of the following:

- `ins(X)`: Matches an event `ins(X)` which is posted when the domain variable `X` is instantiated.

- `bound(X)`: Matches an event `bound(X)` which is posted when the lower or upper bound of the domain of `X` is updated.

- `dom(X)` and `dom(X,E)`: Matches an event `dom(X,E)` which is posted when an inner element `E` is excluded from the domain of `X`.

A user program can create and post its own events and define agents to handle them. A user-defined event takes the form of `event(X,T)` where `X` is a suspension variable that connects the event with its handling agents, and `T` is a Prolog term that contains the information to be transmitted to the agents. If the event poster does not have any information to be transmitted to the agents, then the second argument `T` can be omitted. The built-in action `post(E)` posts the event `E`.

In an action rule, `dom(X,T)` or `event(X,T)` is not allowed to coexist with any other event patterns so when the action of the rule is executed `T` always refers to the second argument of the event.

3.2 Examples

The following defines an agent that echoes the messages sent to it by event posters.

```
echo_agent(X), {event(X,Message)} => write(Message).
```

The following query,

```
echo_agent(Ping), echo_agent(Pong),
post(event(Ping,ping)), post(event(Pong,pong))
```

creates two echo agents `echo_agent(Ping)` and `echo_agent(Pong)`, and activate them by posting two events. The event `event(Ping,ping)` activates the agent `echo_agent(Ping)`, and the event `event(Pong,pong)` activates `echo_agent(Pong)`.

The following defines the freeze predicate in Prolog-II [4].

[4]In the real implementation, built-in events are provided for programming constraint propagators, graphical user interfaces, and interactive agents.

```
freeze(X,G), var(X), {ins(X)} => true.
freeze(X,G) => call(G).
```

The primitive `freeze(X,G)` is logically equivalent to `call(G)` but the execution of `G` is delayed until `X` is instantiated to a non-variable term. The agent `freeze(X,G)` is suspended waiting for an event `ins(X)` when `X` is a variable. When an event `ins(X)` is posted, the condition `var(X)` is tested *again*. If it succeeds, then the action `true` is executed and the agent becomes suspended again. As long as `X` is a free variable, the agent `freeze(X,G)` will be suspended. Only when `X` becomes a non-variable term, can the second rule be applied.

Consider, as another example, how to implement the following indexical:

```
X in min(Y)+min(Z)..max(Y)+max(Z).
```

which ensures that the constraint `X = Y+Z` is interval-consistent w.r.t. `X`.

```
'V in V+V'(X,Y,Z),{ins(Y),bound(Y),ins(Z),bound(Z)} =>
      reduce_domain(X,Y,Z).

reduce_domain(X,Y,Z) =>
      L is min(Y)+min(Z), U is max(Y)+max(Z),
      X in L..U.
```

The propagator is activated whenever a bound of `Y` or `Z` is updated or either one is instantiated. The action `reduce_domain(X,Y,Z)` enforces that the domain of `X` be in the range `min(Y)+min(Z)..max(Y)+max(Z)`. The original indexical is equivalent to the following two subgoals:

```
'V in V+V'(X,Y,Z),reduce_domain(X,Y,Z)
```

where `reduce_domain(X,Y,Z)` enforces interval consistency w.r.t. `X` when the constraint is generated.

3.3 Operational Semantics

The operational semantics of AR can be formulated as a system of transition states. Each state is a pair $<A, S>$, where A is a *sequence* of actions and S is a *set* of suspended agents in the form (α, Γ) where α is an agent and Γ is a set of registered events α is waiting for. An initial state is $<A_0,\emptyset>$ where A_0 is a sequence of actions given by the user and an ending state is $<\emptyset,S>$ where the sequence of actions is empty. A failure occurs when no rule can be applied to the selected agent or a built-in such as unification fails. Since a failure is handled by backtracking as in Prolog, its treatment will not be further discussed.

Posting events

Let $<post(e).A, S>$ be the current state. After e is posted, all the agents waiting for e in S will be activated.

$$\frac{<post(e).A,S>, A'=\{\alpha|(\alpha,\Gamma)\in S, e\in\Gamma\}}{<A'+A,S>} \qquad \text{(Event posting)}$$

The activated agents are added to the front of the action sequence in some order. It is up to the implementers of the language to use a strategy to schedule activated agents. Whatever scheduling strategy is adopted, the user should not rely on the strategy to guarantee the correctness of programs. In the implementation in B-Prolog, the *first-created-first* strategy is used to order agents.

In practice, for the sake of efficiency, event posting is postponed until before the execution of the next non-inline subgoal. Therefore, at a point during execution, there may be multiple events posted that are all expected by an agent. If this is the case, then the agent has to be activated once for each of the events.

Application of rules

When an agent is created or activated, the system searches in its definition in textual order for a rule whose agent-pattern *matches* the agent and whose condition is satisfied. This kind of rule is said to be *applicable* to the agent. Formally, an action rule "H, C, E => B" or a commitment rule "H, C => B" is applicable to an agent α if there exists a unifier θ such that "$H\theta = \alpha \wedge C\theta$". Notice that since one-directional matching rather than full-unification is used to search for an applicable rule and in the condition no variable in α can be instantiated, the agent will remain the same after an applicable rule is found.

After an applicable rule is found, the agent will behave differently depending on the type of the rule. If the rule found is a commitment rule "H, C => B" in which no event pattern is specified, then the action B will be added to the action sequence and the agent is removed from the suspension set if it is included.

$$\frac{<\alpha.A,S>,(\alpha,\Gamma)\notin S}{<B\theta+A,S>} \qquad \frac{<\alpha.A,\{(\alpha,\Gamma)\}\cup S>}{<B\theta+A,S>} \qquad \text{(Application of a commitment rule)}$$

The agent will commit to the action and a failure of the action will lead to the failure of the agent.

A commitment rule is similar to a guarded clause in concurrent logic languages [21], but an agent can never be blocked while it is being matched against an agent pattern.

If the rule found is an action rule "$H, C, \{E\}$ => B" and E contains unregistered events on the agent, then the events will be registered and the agent will be suspended until it is *activated* by one of the registered events.

$$\frac{<\alpha.A,S>,(\alpha,\Gamma)\notin S}{<A,\{(\alpha,E)\}\cup S>} \qquad \frac{<\alpha.A,\{(\alpha,\Gamma)\}\cup S>,\exists_{e\in E}e\notin\Gamma}{<A,\{(\alpha,E\cup\Gamma)\}\cup S>} \qquad \text{(Suspension)}$$

If all the events in E have been registered on the agent, then the action will be added to the action sequence.

$$\frac{<(\alpha.A,\{(\alpha,\Gamma)\}\cup S>,\forall_{e\in E}e\in\Gamma}{<B\theta+A,\{(\alpha,\Gamma)\}\cup S>} \qquad \text{(Application of an activation rule)}$$

The agent does not vanish after the application, but instead turns to wait until it is activated again. So, aside from the difference in event-handling, the action rule "$H, C, \{E\}$ => B" is similar to the guarded clause "H:$-C \mid B, H$", which creates a clone of the agent after the action B is executed.

There is no primitive for killing agents explicitly. An agent never disappears as long as action rules are applied to it. An agent vanishes only when a commitment rule is applied to it. An agent flounders if it waits for an event that can never happen. It is the programmer's responsibility to prevent this situation from happening.

4 Programming Constraint Propagators in AR

The high descriptive power of AR opens new ways to implementing constraint propagators. In this section, we implement propagators that maintain node, interval, and arc consistency for binary constraints, a hybrid algorithm for non-binary constraints, and a weak arc-consistency algorithm for the global constraint `all_distinct`.

4.1 Binary constraints

We consider how to implement propagators for the binary constraint `A*X = B*Y+C`, where `X` and `Y` are domain variables, `A` and `B` are positive integers, and `C` is an integer of any kind. Similar propagators can be implemented for other types of binary constraints.

4.1.1 Forward checking

Recall that forward checking enforces node-consistency. The following shows a propagator that performs forward checking for the binary constraint.

```
'aX=bY+c'(A,X,B,Y,C) =>
       'aX=bY+c_forward'(A,X,B,Y,C).

'aX=bY+c_forward'(A,X,B,Y,C),var(X),var(Y),{ins(X),ins(Y)} => true.
'aX=bY+c_forward'(A,X,B,Y,C),var(X) =>
       T is B*Y+C, X is T//A, A*X=:=T.
'aX=bY+c_forward'(A,X,B,Y,C) =>
       T is A*X-C, Y is T//B, B*Y=:=T.
```

The operation `op1//op2`, which is equivalent to `truncate(op1/op2)`, gives the integer quotent of the division. When both `X` and `Y` are variables, the propagator is suspended. When either variable is instantiated, the propagator computes the value for the other variable.

4.1.2 Interval-consistency

The following propagator, which extends the forward-checking propagator, maintains interval-consistency for the constraint.

```
'aX=bY+c'(A,X,B,Y,C) =>
       'aX=bY+c_reduce_domain'(A,X,B,Y,C),
       'aX=bY+c_forward'(A,X,B,Y,C),
       'aX=bY+c_interval'(A,X,B,Y,C).
```

The subgoal `'aX=bY+c_reduce_domain'(A,X,B,Y,C)` preprocess the constraint to make it interval-consistent when the constraint is generated.

```
'aX=bY+c_reduce_domain'(A,X,B,Y,C) =>
      'aX in bY+c_reduce_domain'(A,X,B,Y,C),
      MC is -C,
      'aX in bY+c_reduce_domain'(B,Y,A,X,MC).

'aX in bY+c_reduce_domain'(A,X,B,Y,C) =>
      L is (B*min(Y)+C) /> A,
      U is (B*max(Y)+C) /< A,
      X in L..U.
```

The operation `op1 /> op2` returns the lowest integer that is greater than or equal to the quotient of `op1` by `op2` and the operation `op1 /< op2` returns the greatest integer that is less than or equal to the quotient. It can be proved easily that no value outside the range `L..U` satisfies the constraint.

The subgoal `'aX=bY+c_interval'(A,X,B,Y,C)` maintains interval-consistency for the constraint.

```
'aX=bY+c_interval'(A,X,B,Y,C) =>
      'aX in bY+c_interval'(A,X,B,Y,C),  % reduce X when Y changes
      MC is -C,
      'aX in bY+c_interval'(B,Y,A,X,MC). % reduce Y when X changes

'aX in bY+c_interval'(A,X,B,Y,C),var(X),var(Y),{bound(Y)} =>
      'aX in bY+c_reduce_domain'(A,X,B,Y,C).
'aX in bY+c_interval'(A,X,B,Y,C) => true.
```

Notice that the action `'aX=bY+c_reduce_domain'(A,X,B,Y,C)` is executed only when both variables are free. If either one turns to be instantiated, then the forward-checking rule will take care of that situation.

4.1.3 Arc-consistency

The following propagator, which extends the one shown above, maintains arc-consistency for the constraint.

```
'aX=bY+c'(A,X,B,Y,C) =>
      'aX=bY+c_reduce_domain'(A,X,B,Y,C),
      'aX=bY+c_forward'(A,X,B,Y,C),
      'aX=bY+c_interval'(A,X,B,Y,C),
      'aX=bY+c_arc'(A,X,B,Y,C).

 'aX=bY+c_arc'(A,X,B,Y,C) =>
      'aX in bY+c_arc'(A,X,B,Y,C),  % reduce X when Y changes
      MC is -C,
      'aX in bY+c_arc'(B,Y,A,X,MC). % reduce Y when X changes
```

```
'aX in bY+c_arc'(A,X,B,Y,C),var(X),var(Y),{dom(Y,Ey)} =>
     T is B*Ey+C,
     Ex is T//A,
     (A*Ex=:=T -> exclude(X,Ex);true).
'aX in bY+c_arc'(A,X,B,Y,C) => true.
```

Whenever an element `Ey` is excluded from the domain of `Y`, the propagator `'aX in bY+c_arc'(A,X,B,Y,C)` is activated. If both `X` and `Y` are variables, the propagator will exclude `Ex`, the counterpart of `Ey`, from the domain of `X`. Again, if either `X` or `Y` becomes an integer, the propagator does nothing. The forward checking rule will take care of that situation.

4.2 Non-binary Constraints

In indexical-based CLP(FD) systems, constraints are split into indexicals that contain no more than three variables. This algorithm has several advantages. Firstly, it generates linear-size code. Secondly, indexicals can be implemented in a low-level language to achieve better performance. Thirdly, information propagation can be restricted to only those constraints for which the domains have the possibility to be reduced [3]. For example, consider the two ternary constraints `T1 = X1+X2` and `T1+X3+X4 = 0`. If `T1 = X1+X2` is activated by an update of `X1`, as long as the shared variable `T1` does not change the other constraint needs not be activated. The disadvantages of this algorithm are that new domain variables have to be introduced and the granularity of constraints becomes smaller and thus context-switching becomes more costly. In B-Prolog, each domain variable takes at least 10 words, letting alone the space for the constraints and data structures for the elements. The space overhead cannot be neglected when the number of variables is large.

In comparison with indexicals, the high descriptive power of AR opens new ways to compiling non-binary constraints. We present two algorithms. One is called *unite*, which adopts one propagator for each constraint to maintain the interval-consistency. The other one, called *hybrid*, maintains interval-consistency when the constraint contains more than two variables and maintains arc-consistency when the constraint turns into binary.

4.2.1 Unite: use one propagator for each constraint

Let $A_1 * X_1 + \ldots + A_n * X_n + C = 0$ be an n-ary constraint where each A_i(i=1,...,n) is a non-zero integer and each X_i is a domain variable or an integer. The propagator for the constraint takes the following form:

```
'A1*X1+...+An*Xn+C=0'(C,A1,A2,...,An,X1,X2,..,Xn),
     {ins(X1),bound(X1),...,ins(Xn),bound(Xn)}
=>
     reduce the domains of X1,..,Xn.
```

In the action, attempts are made to reduce the lower and upper bounds of the domain of every variable.

To facilitate the generation of the code for reducing domains, the compiler splits the expression $A_1 * X_1 + \ldots + A_n * X_n + C$ into the following list of sub-expressions each of which contains at most three variables:

$T_0 = C,$
$T_1 = T_0 + A_1 * X_1,$
$T_2 = T_1 + A_2 * X_2,$
$\ldots$
$T_n = T_{n-1} + A_n * X_n$

The generated reducer first computes the lower and upper bounds of the temporary variables[5] by propagating information forward from T_0 to T_n. The lower and upper bounds of T_i are computed from those of T_{i-1} and $A_i * X_i$ $(i = 1, \ldots, n)$. After that, the reducer propagates information backward from T_n to T_1. For each tuple $T_i = T_{i-1} + A_i * X_i$, the new bounds of T_{i-1} and X_i are computed from those of T_i.

For example, the following shows the propagator generated for the constraint X1+X2+X3+C = 0.

```
'X1+X2+X3+C=0'(C,X1,X2,X3)
      {ins(X1),bound(X1),ins(X2),bound(X2),
       ins(X3),bound(X3)}
=>
      'X1+X2+X3+C=0_reducer'(C,X1,X2,X3).

'X1+X2+X3+C=0_reducer'(C,X1,X2,X3) =>
      Lt1 is C+min(X1), Ut1 is C+max(X1),         % T1 = C+X1
      Lt2 is Lt1+min(X2), Ut2 is Ut1+max(X2),     % T2 = T1+X2
      Lt3 is Lt2+min(X3), Ut3 is Ut2+max(X3),     % T3 = T2+X3
      Lt3 =< 0, Ut3 >= 0,                         % T3 = 0
      %
      NewLx3 is 0-Ut2, NewUx3 is 0-Lt2,           % T3 = T2+X3
      X3 in NewLx3..NewUx3,
      NewLt2 is 0-max(X3), NewUt2 is 0-min(X3),
      %
      NewLx2 is NewLt2-Ut1, NewUx2 is NewUt2-Lt1,% T2 = T1+X2
      X2 in NewLx2..NewUx2,
      NewLt1 is NewLt2-max(X2), NewUt1 is NewUt2-min(X2),
      %
      NewLx1 is NewLt1-C, NewUx1 is NewUt1-C,     % T1 = C+X1
      X1 in NewLx1..NewUx1,
      NewLt1-max(X1) =< C, NewUt1-min(X1) >= C.
```

The advantage of this algorithm is that only one propagator is used for each constraint whose code size is linear in the number of variables in the constraint.

[5]Temporary variables are plain variables not domain variables. Therefore, this compilation scheme is different from compiling constraints into indexicals.

The weakness of this algorithm is that the reducer is not fast. Whenever a variable is instantiated or a variable's bound is updated, the reducer tries to reduce the domains of all the variables including the seed variable that triggers the propagator.

4.2.2 Hybrid: combining interval and arc consistency algorithms

For a non-binary constraint, it is too expensive to maintain arc-consistency. One practical strategy is to maintain interval-consistency while there are multiple variables in the constraint and to maintain arc-consistency when the constraint turns into binary. The following shows the propagator for the linear non-binary constraint `A1*X1+...+An*Xn+C = 0`.

```
'A1*X1+...+An*Xn+C=0'(C,A1,A2,...,An,X1,X2,..,Xn),
      n_vars_gt([X1,...,Xn],2),
      {ins(X1),bound(X1),...,ins(Xn),bound(Xn)}
=>
      reduce domains of X1,..,Xn to achieve interval-consistency.
'A1*X1+...+An*Xn+C=0'(C,A1,A2,...,An,X1,X2,..,Xn) =>
      nary_to_binary([C,A1,X2,A2,X2,...,An,Xn],NewC,B1,B2,Y1,Y2),
      call_binary_constraint_propagator(NewC,B1,Y1,B2,Y2).
```

The propagator is activated whenever any variable is instantiated or its bound is updated. When `n_vars_gt([X1,...,Xn],2)` succeeds, i.e. when there are multiple variables in the constraint, the domains are reduced to make the constraint interval-consistent. When the constraint becomes binary, the condition `n_vars_gt` fails and the second rule will be tried. The subgoal `nary_to_binary` transforms the constraint into the binary constraint `B1*Y1+B2*Y2+NewC = 0`, and the next subgoal invokes an appropriate propagator for the binary constraint. In the real implementation, the two built-ins `n_vars_gt` and `nary_to_binary` do not take the constraint as an argument but instead access the constraint in the parent subgoal. In this way, no copy of the constraint needs to be made.

4.3 Propagators for `all_distinct(L)`

The constraint `all_distinct(L)` holds if the elements in `L` are pair-wisely different. One naive implementation method for this constraint is to generate binary disequality constraints between all pairs of variables in `L`. This implementation has two problems: First, the space required to store the constraints is quadratic in the number of variables in `L`; Second, splitting the constraint into small granularity ones may lose possible propagation opportunities [19, 18]. This subsection presents two propagators for the constraint. The propagation algorithms are not new. The goal of this subsection is to illustrate the expressive power of AR.

4.3.1 A linear-space propagator

To solve the space problem, we define `all_distinct(L)` in the following way:

```
all_distinct(L) => all_distinct(L,[]).
all_distinct([],Left) => true.
all_distinct([X|Right],Left) =>
      outof(X,Left,Right),
      all_distinct(Right,[X|Left]).

outof(X,Left,Right), var(X), {ins(X)} => true.
outof(X,Left,Right) => exclude_list(X,Left),exclude_list(X,Right).

exclude_list(X,[]).
exclude_list(X,[Y|Ys]):- exclude(Y,X),exclude_list(X,Ys).
```

For each variable `X`, let `Left` be the list of variables to the left of `X` and `Right` be the list of variables to the right of `X`. The subgoal `outof(X,Left,Right)` holds if `X` appears in neither `Left` nor `Right`. Instead of generating disequality constraints between `X` and all the variables in `Left` and `Right`, the subgoal `outof(X,Left,Right)` suspends until `X` is instantiated. After `X` becomes non-variable, `exclude_list(X,Left)` and `exclude_list(X,Right)` exclude `X` from the domains of the variables in `Left` and `Right`, respectively.

There is one propagator `outof(X,Left,Right)` for each element `X` in the list, which takes constant space. Therefore, `all_distinct(L)` takes linear space in the size of `L`. Notice that the two lists `Left` and `Right` are not merged into one bigger list; otherwise, the constraint would take quadratic space.

4.3.2 A weak arc-consistency algorithm

In terms of pruning ability, the linear-space propagator is the same as the naive one that splits `all_distinct(L)` into binary disequality constraints. In this subsection, we present a weak arc-consistency algorithm for `all_distinct(L)` and show how to implement it.

For each variable `X` in `L`, let `L-{X}` be the list of elements in `L` but `X`, n be the size of the domain of `X`, and m be the number of variables in `L-{X}` whose domains are subsets of that of `X`. If $m+1 > n$, then the constraint is unsatisfiable since it is impossible to assign n values to more than n variables such that each variable gets a different value. If $m+1 = n$, then for each value v in `X`'s domain, we can safely exclude v from the domains of all the variables whose domains are not subsets of that of `X`.

Consider the following query,

```
X in {1,2}, Y in {1,2}, Z in {1,2}, all_distinct([X,Y,Z]).
```

the weak arc-consistency algorithm detects the inconsistency of the constraint without labeling any variables. For the following query,

```
X in {1,2}, Y in {1,2}, Z in {1,2,3}, all_distinct([X,Y,Z]).
```

the algorithm assigns `3` to `Z` without labeling any variables.

The weak arc-consistency algorithm is not as powerful as the algorithm proposed by Regin [19] in terms of pruning ability but is much easier to implement.

To incorporate the weak arc-consistency checking algorithm into the linear-space propagator, we only need to redefine `outof(X,Left,Right)` as follows:

```
outof(X,Left,Right), var(X), {ins(X),bound(X),dom(X)} =>
      outof_reducer(X,Left,Right).
outof(X,Left,Right) => exclude_list(X,Left),exclude_list(X,Right).
```

where `outof_reducer(X,Left,Right)` first counts the variables in `Left` and `Right` whose domains are subsets of the domain of `X` and then decides what action to take depending on the count and the size of the domain of `X`.

The key operation is to decide whether a domain is a subset of another domain. In the worst case, the two domains have to be scanned. There are several facts that can be used to avoid scanning domain elements. A domain `D1` cannot be a subset of another domain `D2` if `D1` has a larger size or has a larger interval. Also if two domains are intervals without holes, then scanning the elements is unnecessary. Another fact that can be used in the detection is that if the event is `dom(X,E)` meaning that `E` has been excluded from `X`'s domain, then another domain `Y` cannot be a subset of `X` if `E` is included in `Y`. To take advantage for this fact, the propagator can be rewritten into the following:

```
all_distinct(L) => all_distinct(L,[]).

all_distinct([],Left) => true.
all_distinct([X|Right],Left) =>
      outof(X,Left,Right),
      outof_dom(X,Left,Right),
      all_distinct(Right,[X|Left]).

outof(X,Left,Right), var(X), {ins(X),bound(X)} =>
      outof_reducer(X,Left,Right).
outof(X,Left,Right) => exclude_list(X,Left),exclude_list(X,Right).

outof_dom(X,Left,Right),var(X), {dom(X,E)} =>
      outof_reducer(X,E,Left,Right).
outof_dom(X,Left,Right) => true.
```

The subgoal `outof_reducer(X,E,Left,Right)` takes `E` into account when detecting whether a domain is a subset of that of `X`.

5 Implementation and Performance Evaluation

B-Prolog has been extended to accommodate AR and the finite-domain constraint solver described in this paper has been developed in AR. In this section, we first briefly describe the implementation of AR and then evaluate the performance of the finite-domain constraint solver. The description of the implementation serves as the grounds for the explanations of the experimental results. A more detailed description can be found in [31].

5.1 Implementation

The abstract machine of B-Prolog, called ATOAM [28], is extended to support agents. The ATOAM is a variant of the Warren Abstract Machine (WAM) [26]. Unlike in the WAM where arguments are passed through argument registers, arguments in the ATOAM are passed through stack frames and only one frame is used for each subgoal. Each time when a predicate is invoked by a subgoal, a frame is placed on top of the control stack unless the frame currently at the top can be reused. Frames for different types of predicates have different structures. Agents are stored as frames on the control stack. The frame for an agent has the following slots in addition to those included in a normal frame: the status of the agent, the activating event, the program pointer to continue when the agent is activated, and the pointer to the previous agent's frame in the chain of agents. The frames on the control stack comprise three chains, namely the chain of *active subgoals* that are being executed, the chain of *choice points*, i.e., subgoals that have alternative clauses to be tried when execution backtracks to them, and the chain of *agents*.

Storing agents on the stack facilitates context-switching for agents [27] but complicates memory management. With frames of agents on the stack, the chronological order of frames is no longer preserved and therefore run-time checking is needed to determine whether the frame at the top can be reused and a garbage collector is needed to collect useless frames on the control stack [32].

Action rules are compiled into matching trees [28] such that shared tests among different rules do not need to be executed multiple times. This technique is useful for speeding-up constraint propagators.

The actions of constraint propagators are to reduce the domains of variables. This characteristic is exploited to improve the performance of constraint propagators. Some events that cannot lead to the shrinking of any domains are ignored. For example, if multiple events of `bound(X)` are posted at the same time, then only one of them needs to be handled, and if `bound(X)` and `ins(X)` are posted at the same time, then the `bound(X)` event is ignored. In this way, many redundant activations of rules that do not contribute to the reduction of any domains can be suppressed.

5.2 Performance Evaluation

We compared the performance of B-Prolog version 6.4 (BP)[6] with three other CLP(FD) systems: Eclipse 5.5 #46 (EP), GNU-Prolog version 1.2.16 (GP), and Sicstus 3.10 (SP). There are two solvers available in BP: one is called BP-AC which adopts the hybrid algorithm presented in this paper for equality constraints and the other called BP-IC which maintains only interval consistency for equality constraints. BP-AC is the default solver. The linear-space propagator is used for `all_distinct` in both solvers.

Table 1 shows the CPU times taken by the four solvers to run a set of benchmarks,[7] assuming the time taken by BP-IC be 1. Most of the benchmarks have been widely

[6]Available from www.probp.com.

[7]Available from probp.com/bench.tar.gz.

Table 1: Comparison of CPU times.

`Program`	**XP**					**Linux**	
	`BP-IC`	`BP-AC`	`EP`	`GP`	`SP`	`BP-IC`	`GP`
alpha	1	0.74	7.36	0.91	3.24	1	0.77
bridge	1	1.00	2.21	0.52	2.59	1	0.52
cars	1	1.00	4.22	1.00	2.60	1	1.11
color	1	1.00	5.50	0.88	2.26	1	1.03
eq10	1	1.16	4.85	3.83	4.50	1	2.95
eq20	1	1.00	4.12	2.01	2.95	1	1.75
magic3	1	1.32	8.35	1.95	4.66	1	1.17
magic4	1	1.20	8.80	1.80	6.20	1	1.50
olympic	1	0.90	9.02	1.60	3.80	1	1.61
queens1	1	1.00	3.82	0.38	4.28	1	0.34
sendmoney	1	1.00	8.01	4.48	9.66	1	2.00
sudoku81	1	1.00	11.00	4.00	12.00	1	1.21
zebra	1	1.07	8.10	2.53	8.38	1	1.39
<arithmetic mean>	1	1.05	6.57	1.99	5.16	1	1.37
<geometric mean>	1	1.04	6.02	1.55	4.48	1	1.21

used by other authors to compare CLP(FD) systems [2, 3, 24]. Three new programs were added by the author into the set: `color` is a program that colors a map with 110 regions; `olympic` is a puzzle taken from a Mathematics Olympic game for elementary students; and `sudoku81` is a program for solving a puzzle. The left-to-right labeling strategy is used to instantiate variables in all the benchmarks. The CPU times were measured on a 1.7GHz CPU running Windows XP. Each program was run at least 10 times and the average was taken. For some programs, execution was repeated up to 1000 times to obtain a stable average. Garbage collection was disabled. GP has a native code compiler for Linux. The comparison with GP was also conducted on Linux.

The BP solvers compare favorably with GP and are significantly faster than EP and SP. BP outperforms GP remarkably for programs that contain non-binary equality constraints, such as `eq10`, `eq20`, and `sendmoney`. This result reveals that the disadvantages of splitting n-ary constraints into indexicals outweigh the advantages. On the other hand, GP is three times as fast as BP for `queens(25)` and twice as fast for `bridge`. The high speed for `queens` may be attributed to an optimization technique adopted in GP that combines indexicals. If the propagators for the the disequality constraints were combined for the program in BP,[8] the speed would be doubled. For `bridge`, BP is slower than GP because in BP updates of lower and upper bounds are not treated as different events. Therefore, for the constraint $X \leq Y$, the propagator is activated even if the upper bound of X or the lower bound of Y is updated.

Comparing BP-AC and BP-IC reveals that the hybrid algorithm is effective for `alpha` and `olympic` only. The BP-AC solver is adopted as the default one since for

[8]For the three disequality constraints $X \neq Y$, $X \neq Y + N$, and $X \neq Y - N$, we can use one propagator to handle the `ins(X)` and `ins(Y)` events.

Table 2: Comparison of numbers of backtracks.

`Program`	`BP-AC`	`BP-IC`	`GP`
alpha	*4605*	8440	8440
bridge	0	0	0
crypta	52	52	52
cars	53	53	*34*
color	560	560	560
eq10	49	49	49
eq20	49	49	49
magic3	2	2	2
magic4	18	18	18
olympic	*36*	50	50
queens(25)	7255	7255	7255
sendmoney	2	2	2
sudoku81	0	0	0
zebra	2	2	2

some programs, such as the queens program given in [17],[9] BP-AC is exponentially faster than BP-IC. BP-AC is slightly slower than BP-IC for most of the programs. In general, this happens for programs for which the efforts to reduce domains do not pay off.

Table 2 gives the numbers of backtracks performed by the three solvers. `BP-AC` makes the same number of backtracks as `BP-IC` except for `alpha` and `olympic`, and `GP` makes fewer backtracks than `BP-AC` for `cars`. Basically, the three solvers explore the same search trees for most of the programs. Therefore, the comparison results shown in Table 1 reflect the real performance of the solvers.

GP and BP are quite different. In GP constraints are compiled into indexicals defined in C while in BP constraints are compiled into propagators defined in action rules. Although the GP Prolog engine may not have much impact on the performance of constraint programs, the BP Prolog engine does have a great impact since all the propagators are defined in action rules. One evidence for this observation is that the BP constraint solver becomes 20-30 percent faster after the main switch statement in the emulator is changed to a jump table. A further speed-up is expected if a native code compiler is employed.

There are other factors that affect the performance of a solver, such as domain representation, interaction with other solvers, and garbage collection. GP supports only finite domains of positive integers, while BP supports not only finite integer domains but also trees and finite domains of ground terms and sets [30]. In BP, integer domains are represented as described in Subsection 2.3. BP adopts a sound arithmetic that guarantees that no solution is lost. For example, when excluding

[9]This program is not included in the benchmark set since it requires support of negative integer domains and thus cannot run on GP.

an inner element from a large interval domain, the system generates a disequality constraint rather than brutally changes the interval into a bit vector.[10] BP has a garbage collector that collects garbage on the heap and the control stack, but GP does not support garbage collection yet. It is likely that garbage collection will suppress some optimization techniques.

6 Related Work

CLP(FD) systems have undergone an evolution process, from closed to open and from low level to high level. Several constructs have been proposed to facilitate the implementation of constraint propagators. Examples include attributed variables [10], indexicals [3], extended indexicals called projection constraints [22], delay clauses [15, 29], and constraint handling rules [8]. An action rule is an extension of a delay clause that allows for the descriptions of not only delay conditions on subgoals but also activating events and actions. This section compares AR with these constructs introduced into Constraint Logic Programming. Constructs introduced into other languages such as ILog [17] and Oz [20] are not compared.

AR is more powerful and flexible than indexicals. We have described in this paper several propagation algorithms in AR, some of which are impossible to encode in indexicals (e.g., the hybrid algorithm for n-nary constraints) and some of which cannot be implemented as efficiently (e.g., arc-consistency for binary constraints). Consider the following indexical taken from [2],

```
    X in dom(Y)+C
```

which maintains the arc-consistency for the constraint `X = Y+C` w.r.t. `X`. Whenever an element `y` is excluded from the domain of `Y`, the indexical will be activated. Because the indexical does no know what the excluded element is, it has to go through the domain elements of `Y` in the worst case to locate a possible no-good value in the domain of `X`. In contrast, in the propagator implemented in AR, the propagator knows exactly what element is excluded from the domain of `Y` and thus can compute the counterpart in the domain of `X` in constant time.

Compiling constraints into indexicals enables the use of more specialized propagators and restricts propagation to within a small range of constraints [3]. Nevertheless, this approach has to introduce new temporary variables and lower the granularity of propagators. Our experiment reveals that B-Prolog outperforms GNU-Prolog for almost all the benchmarks that contain non-binary constraints. This result reveals that the disadvantages of splitting constraints outweigh the advantages. Similar observations have been made independently in [2, 9, 29]. In [9], the same two-phase algorithm is used to reduce domains of linear constraints.

Attributed variables [10] are variables with attached attributes each of which has a list of handlers. Touching an attribute will trigger the corresponding list of handlers. In order to make context-switch swift for handlers, systems such as Eclipse treats handlers as demons rather than as normal subgoals. A demon

[10]Generating a disequality constraint is less efficient than changing an interval into a bit vector since the disequality constraint needs to be checked each time the variable is instantiated.

is different from a normal subgoal in that it does not disappear after execution but instead waits for another activation. In this sense, agents in our system are similar to demons. Nevertheless, an agent can be activated by different kinds of events and an agent may take different actions depending on the conditions. An agent can be defined by multiple action rules and the rules are compiled into a tree by the compiler such that shared tests are combined and conditions that are tested failure once need not be tested again. Sicstus [2] provides interfaces for implementing propagators and also some sort of delay construct similar to attributed variables that triggers propagators when events are posted.

AR is an extension of our early delay construct proposed in [29] that allows for the event `dom(X,E)` and user-defined events. The support of the event `dom(X,E)` is essential for implementing arc consistency algorithms and also propagators for set constraints [30]. Our delay construct is an extension of Meier's delay clause construct [15] that allows for not only delay conditions but also events and actions. In Meier's delay clause, events are implicitly extracted from delay conditions and a delayed subgoal never takes actions as long as the delay condition is satisfied. In retrospect, all these constructs were inspired by early work by Colmerauer and Naish [4, 16].

Other rule-based languages have been designed for implementing constraint propagators. CHR resembles a production system. In CHR, the left-hand side of a rule specifies a pattern of constraints in the constraint store and the right-hand side specifies new constraints to replace those on the left-hand side or to be added into the store. It should be possible to implement in CHR all the propagation algorithms described in this paper provided certain built-ins are added. If events are treated as constraints, then an action rule can be translated into a CHR rule. Treating events as constraints, however, can hardly achieve the same performance. Events are removed automatically after all the agents that are waiting for them are activated. In CHR, there must be rules to remove the events explicitly. The left-hand sides of CHR rules can have multiple constraint patterns. Therefore, it is impossible in general to translate a CHR rule into action rules straightforwardly. It is not clear whether or not it is possible to simulate CHR rules in action rules and how if the answer is yes. It would be an interesting direction to explore in the future.

7 Conclusion

There is a need for an implementation language for constraint propagators that is expressive enough and can be implemented efficiently. This paper has presented such a language called AR. The expressiveness of the language is illustrated though several examples that cannot be implemented in indexicals: the propagator for maintaining arc-consistency of binary equality constraints; a weak arc-consistency algorithm for the `all_distinct` constraint; and a hybrid algorithm for non-binary equality constraints that combines interval and arc-consistency ones. The efficiency is evaluated through benchmarking. For a set of widely used benchmarks, our solver implemented in B-Prolog competes favorably with that in GNU-Prolog, one of the fastest finite-domain constraint solvers available now.

The results are encouraging and promising since our solver is implemented in a high-level language and B-Prolog is an emulator-based system which provides more facilities than GNU-Prolog such as garbage collection and constraint solving over other domains. The high-performance of our solver stems from the following facts. Firstly, only one propagator is generated for each non-binary equality constraint that maintains the interval-consistency. Our solver performs especially well for the benchmarks that contain non-binary equality constraints. This reveals that compiling non-binary equality constraints into indexicals has more cons than pros. Secondly, the hybrid algorithm adopted in our solver is a good compromise between the need to achieve high-level consistency to cut search spaces and the need to reduce the cost. The cost of achieving arc-consistency for binary constraints is relatively small, but the effect can be very big for certain programs. Thirdly, our solver employs optimization techniques that reduce redundant activations of propagators.

Our solver can be improved further in the following aspects: (1) develop new optimization techniques for further avoiding redundant activations of propagators; (2) compile programs into native code to enhance the performance of action rules; and (3) implement consistency algorithms beyond interval and arc consistency such as path consistency and Regin's algorithm for `all_distinct`.

References

[1] Aggoun, A., and Beldiceanu, N. Overview of the CHIP compiler system. In *ICLP'91: Proceedings 8th International Conference on Logic Programming* (Paris, 1991), K. Furukawa, Ed., MIT Press, pp. 775–789.

[2] Carlsson, M., Ottosson, G., and Carlson, B. An open-ended finite domain constraint solver. *PLILP (LNCS) 1292* (1997), 191–205.

[3] Codognet, P., and Diaz, D. Compiling constraints in clp(FD). *Journal of Logic Programming 27*, 3 (1996), 185–226.

[4] Colmerauer, A. Equations and inequations on finite and infinite trees. In *Proceedings of the International Conference on Fifth Generation Computer Systems (FGCS-84)* (Tokyo, Japan, 1984), ICOT, pp. 85–99.

[5] Diaz, D., and Codognet, P. Design and implementation of the GNU Prolog system. *Journal of Functional and Logic Programming 2001(1)* (2001), 1–29.

[6] Dincbas, M., Simonis, H., and van Hentenryck, P. Solving large combinatorial problems in logic programming. *Journal of Logic Programming 8* (1990), 75–93. Special Issue: Logic Programming Applications.

[7] Dincbas, M., van Hentenryck, P., Simonis, H., Aggoun, A., Graf, T., and Berthier, F. The constraint logic programming language CHIP. In *Proceedings of the International Conference on Fifth Generation Computer*

Systems. Volume 2 (Berlin, FRG, 1988), I. for New Generation Computer Technology (ICOT), Ed., Springer Verlag, pp. 693–702.

[8] Frühwirth, T. Theory and practice of constraint handling rules, special issue on constraint logic programming. *Journal of Logic Programming 37* (1998), 95–138.

[9] Harvey, W., and Stuckey, P. J. Improving linear constraint propagation by changing constraint representation. *Constraints: An International Journal 8*, 2 (2003), 173–207.

[10] Holzbaur, C. Metastructures vs. attributed variables in the context of extensible unification. In *Proceedings of the Fourth International Symposium on Programming Language Implementation and Logic Programming* (Leuven, Belgium, 1992), M. Bruynooghe and M. Wirsing, Eds., LNCS 631, Springer-Verlag, pp. 260–268.

[11] Holzbaur, C., and Fruhwirth, T. Compiling constraint handling rules into Prolog with attributed variables. In *International Conference on Principles and Practice of Declarative Programming* (1999), ACM Press, pp. 117–133.

[12] Jaffar, J., and Maher, M. J. Constraint logic programming: A survey. *Journal of Logic Programming 19/20* (1994), 503–582.

[13] Kumar, V. Algorithms for constraint satisfaction problems: A survey. *AI Magazine 13* (1992), 32–44.

[14] Marriott, K., and Stuckey, P. *Programming with Constraints: An Introduction.* MIT Press, 1998.

[15] Meier, M. Better late than never. In *Implementations of Logic Programming Systems*, E. Tick and G. Succi, Eds. Kluwer Academic Publishers, 1994.

[16] Naish, L. *Negation and Control in Prolog.* PhD thesis, University of Melbourne, 1985.

[17] Puget, J., and Leconte, M. Beyond the glass box: Constraints as objects. In *Proc. International Logic Programming Symposium* (1995), MIT Press, pp. 513–527.

[18] Puget, J.-F. A fast algorithm for the bound consistency of alldiff constraints. In *Proceedings of the National Conference on Artificial Intelligence (AAAI-98)* (Menlo Park, 1998), AAAI Press, pp. 359–366.

[19] Regin, J. A filtering algorithm for constraints of difference in CSPs. In *Proceedings of the National Conference on Artificial Intelligence(AAAI-94)* (1994), AAAI Press, pp. 362–367.

[20] Schulte, C. *Programming constraint services: high-level programming of standard and new constraint services*, vol. 2302 of *Lecture Notes in Computer Science*. Springer-Verlag Inc., New York, NY, USA, 2002.

[21] Shapiro, E. The family of concurrent logic programming languages. *ACM Comput. Surveys 21* (1989), 412–510.

[22] Sidebottom, G., and Havens, W. Nicolog: A simple yet powerful cc(fd) language. *Journal of Automated Reasoning 17* (1996), 371–403.

[23] Tsang, E. *Foundations of Constraint Satisfaction.* Academic Press, 1993.

[24] van Hentenryck, P. *Constraint Staisfaction in Logic Programming.* MIT Press, 1989.

[25] van Hentenryck, P., Saraswat, V., and Deville, Y. Constraint Processing in `cc(FD)`. Tech. rep., Brown University, 1992.

[26] Warren, D. H. D. An abstract Prolog instruction set. Tech. rep., SRI International, 1983.

[27] Zhou, N. A novel implementation method of delay. In *Proc. Joint International Conference and Symposium on Logic Programming* (1996), MIT Press, pp. 97–111.

[28] Zhou, N. Parameter passing and control stack management in Prolog implementation revisited. *ACM Transactions on Programming Languages and Systems 18*, 6 (1996), 752–779.

[29] Zhou, N. A high-level intermediate language and the algorithms for compiling finite-domain constraints. In *Proc. Joint International Conference and Symposium on Logic Programming* (1998), MIT Press, pp. 70–84.

[30] Zhou, N. Implementing constraint solvers in B-Prolog. In *IFIP World Congress, Intelligent Information Processing.* Kluwer Academic Publishers, 2002, pp. 249–260.

[31] Zhou, N. A high-performance abstract machine for Prolog and its extensions. Technical Report TR-2003014, CUNY Computer Science (www.cs.gc.cuny.edu/tr/), 2003.

[32] Zhou, N.-F. Garbage collection in B-Prolog. In *First Workshop on Memory Management in Logic Programming Implementations* (2000), pp. 16–25.