The CIAO Multi-paradigm Compiler and System: 
A Progress Report

M. Hermenegildo, F. Bueno, M. García de la Banda, and G. Puebla

Computer Science Department
Technical University of Madrid
{herme,bueno,maria,german}@dia.fi.upm.es

Extended Abstract

1 Introduction

In [HtCg93, HtCg94] we discussed several methodological aspects regarding the design and efficient parallel execution of logic programming systems and concurrent logic programming systems, and their generalization to constraint programming. We proposed a novel view of these systems, based on a particular definition of parallelism and argued that, under this view, a large number of the actual systems and models can be explained (and implemented) through the application, at different levels of granularity, of only a few basic principles, which include determinism, non-failure, independence (also referred to as stability), task granularity, etc. In identifying these fundamental principles, we also argued for a separation between those principles which have to do with the computation rule (i.e., to performing the least work possible) and those that are directly related to parallelism (i.e., to performing the such work in the smallest amount of time by splitting it among several processors). Finally, and basing our discussion on the convergence of concepts that this view brought, we sketched the design of the CIAO (Concurrent, Independence-based And/Or parallel) system, a platform for the implementation of several parallel constraint logic programming source languages and models based on a common, generic abstract machine and an intermediate kernel language.

The purpose of this paper is to report on recent progress in the implementation of the CIAO system itself, with special emphasis on the capabilities of the compiler and the techniques used for supporting such capabilities.

2 The CIAO System

CIAO is a multi-paradigm compiler, run-time, and program development system which is aimed at providing efficient implementations of a range of LP, CLP, and CC programming languages, on sequential and multiprocessor machines. It also supports distributed execution. The CIAO system is generic in the sense that the different source-level languages are supported by compilation via source to source transformations into a comparatively simple kernel language (which is also the native CIAO language). The analysis and transformation techniques used in this process are based on novel semantic modeling of CLP and CC program
behavior and on the exploitation of optimization principles, such as independence/stability, and techniques, such as specialization, abstract executability, etc.

The CIAO system can be used quite effectively for developing applications. However, one of its main objectives is to be useful as an experimentation and evaluation platform for the analysis and transformation techniques being developed, as well as for the underlying abstract machine.

More concretely, the design of the CIAO system is based on a series of ideas, which include:

- **Support for Multiple Models and Paradigms:** As mentioned above, the system supports a range of LP, CLP, and CC programming languages. It also supports several computation rules, including standard left-to-right SLD resolution and the determinate-first principle (as in the Andorra model [SCWY90, dMSC93]).

- **Support for Distributed Execution:** In the belief that distributed systems are one of the target applications of computational logic systems (not so much for performance but for the functionality of accessing remote resources such as knowledge bases) the CIAO system includes extensive distributed execution capabilities.

- **Implementation via Compilation into a Simple Kernel Language:** This is based on the belief that medium to high performance implementations of many LP systems can be obtained in this way, with the advantages that optimizations can be performed via source to source transformations and the low level machinery can be kept minimal. While performance may be limited somewhat in the approach, reasonable levels can be obtained, especially given the experimental nature of the intended system. Optimizations, which include parallelization, reduction of concurrency and synchronization, reordering of goals, code simplification, specialization, etc., are performed via source to source transformation. Most analysis phases are performed at the kernel language level, so that the same analyzer can be used for several models. For example, a single analyzer framework can handle Prolog programs with delay and concurrent (constraint) programs.

- **Explicit Control in the Kernel Language:** Explicit control in the kernel language makes it possible to perform many control-related optimizations at the source level. Such explicit control is performed via operators which include:

  - **Sequential, Parallel, and Concurrent Operators:** the presence of both sequentiality (“,”), concurrency (“&/1”) and parallelism (“&/2”, “&>/2”, “&</1”) operators allows performing optimizations such as parallelization (task creation based on dependencies), partitioning and schedule analysis (task coalescence based on dependencies), and granularity control (task coalescence based on task size considerations) as source to source transformations. The parallel operators support full backtracking. They assume independence among goals. Communication of bindings is not guaranteed until the join. No variable locking is performed. The concurrency operator allows concurrent programming in the style of CC languages. Variable communication (and locking) is performed. Backtracking is limited in the sense that no “active shared binding” can be undone via backtracking. An active shared binding is a binding to a variable that is shared among active processes.
- **Explicit And-Fairness Operator**: based on the observation that and-fairness in concurrent systems is still expensive to implement, a fair concurrency operator ("\&/1") is introduced which explicitly requests the (efficient) association of an operating system thread to a goal.

This also leaves open the possibility of implementing a fair source language that compiles efficiently into these operators (perhaps via an analysis which can determine the program points where fairness is really needed – to ensure, for example, termination).

- **Explicit Synchronization**: explicit synchronization is handled in the kernel language by means of "wait/1" and "ask/1" operators (the latter as in concurrent constraint programming, the former as in &-Prolog), augmented with some meta-tests on the variables (such as ground/1 or nonvar/1).

- **Explicit Placement Operator**: an explicit placement operator ("\@") allows control of task placement in distributed execution. These and other primitives for controlling distributed execution, and to implement the concept of active modules or active objects, are described in [CH95].

- **Generic Abstract Machine**: a comparatively simple abstract machine directly supports the kernel language. The design of the abstract machine is based on the belief that there is much in common at the abstract machine level among many of the LP, CLP, and CC models, and thus builds strongly on the parallelism and concurrency capabilities of the PWAM/&-Prolog abstract machine [Her86, HG91] and recent work on extending its capabilities and efficiency [PGT95a, PGH95, PGT+95b]. The abstract machine includes native support for attributed variables [Hol92, Hou90, Neu90] which are used extensively in the implementation of constraint solvers (as in other systems such as Eclipse [Eur93] and SICStus 3 [Swe95]) and in supporting communication among concurrent tasks [HCC95]. While the current abstract machine supports only ("dependent" and "independent") and-parallelism, it is expected that combination with or-parallelism will be possible by applying the techniques developed in [GC92, GHPC94, GSCYH91].

3 **The CIAO Compiler**

The CIAO compiler provides the required support for the different programming paradigms and their optimization. As mentioned before, it is strongly based on program analysis and transformation. The compilation process can be viewed as a translation process from the input language to (kernel) CIAO. The system is able to translate the input source, automatically extracting parallelism, compiling synchronization, and optimizing the final program. Optimizations include simplifying the code to avoid run-time tests and suspensions, and specializing predicates in order to generate much simpler and efficient code in the back end. Program analysis is instrumental in all the optimizations.

This compilation process is depicted in Figure 1, which illustrates the inputs and outputs, as well as the compilation options, which are selected via either menus, or program flags. The compilation process is structured into several steps. First, a module in a given input language is translated into the kernel language. Then, analysis is performed if required to support the rest of the compilation process. In some cases some degree of analysis may also be performed in the translation step to aid in the translation. After analysis, the program is optionally annotated for parallel execution, simplified and specialized.
The output can then be loaded for execution on the abstract machine. As an alternative, and using the transformational approach, most of the capability of the system (including distributed execution) is also supported (with sometimes somewhat lower efficiency) on any Prolog which supports delay declarations and attributed variables (e.g., SICStus Prolog Version 3 [Swe95]). In that sense, the CIAO compiler can also be viewed as a library package for Prolog systems with these capabilities.

The compiler steps and options are discussed in the following sections. Given the space limitations the aim is to offer a general description and provide references for publications or technical reports where the techniques used are described. An extended description of the capabilities of the compiler can be found in the User’s Manual [Bue95].

3.1 Source Languages Supported and Transformations Performed

3.1.1 Input Languages and Computation Rules The compiler can deal with several languages and computation rules simultaneously and perform several translations among them. Currently, in addition to (kernel) CIAO full syntax (backwards compatible with Prolog, plus the specific CIAO primitives), also concurrent logic programming syntax, including (flat) guards, is supported, as well as the basic Andorra model. Also, the constraint system over which the program is defined can be declared. Currently, the system supports
those of Prolog, CLP(R), and CLP(Q). The current set of choices is illustrated in Table 1.

The Compiler determines the language and computation rule under which to interpret a given program as follows. The default is the CIAO kernel language (which includes full Prolog), on the Herbrand domain. The mode of the system can be changed by typing at the top level the commands in the “Mode” column in Table 1. Programs read from then on will be interpreted in the new mode. Alternatively, the programs themselves can be annotated directly. This is done in the module declaration. One more argument is available in module declarations where the mode (again, one of those in the “Mode” column in Table 1) is specified.

Program transformations bridge the semantic gaps between the different programming paradigms supported. The methods used for supporting the (Basic) Andorra model are described in [BDGH95]. The methods used for supporting CC languages are an extension of those of [DGBH94, Deb93] and are described in [BH95c].

### 3.2 Analysis

The compiler includes a number of analysis modules, which can analyze programs over the Herbrand or some constraint domains. In addition, analysis of dynamically scheduled programs is provided in order to support the concurrent models. Note that, thanks to the transformational approach, only two frameworks are used (one for simple, left-to-right execution and another for the case when there are dynamically scheduled goals).

Local analysis of program clauses is quite straightforward but sometimes useful in some optimizations, as in program parallelization [BGH91]. Global analysis is performed in CIAO in the context of abstract interpretation [CC77, Deb92, CC92]. The underlying framework of analysis is that of PLAI [HWD92, MH90, MH92]. PLAI implements a generic, top-down driven\(^1\) abstract interpreter. It allows easily plugging into it several abstract domains. PLAI also incorporates incremental analysis [HMPS95] in order to deal with large programs and is capable of analyzing full languages (in particular, full standard Prolog [BCHP95, CRH94]). The CIAO analyzer incorporates the following domains, which are briefly explained below: \(SH, \ SH+FR, \ ASub, \ SH+ASub, \) and \(SH+FR+ASub,\) which are used in logic programming, and \(Def, \ Fr, \ Fd,\) which can be used either in logic or constraint logic programming, and \(LSign,\)

\(^1\)PLAI now also supports goal-independent analysis [CGBH94] and bottom-up analysis.
which is more specific to constraint logic programming.\(^2\) Analysis of dynamically scheduled languages can be carried out with the \(SH+FR\) and \(LSign\) domains.

### 3.2.1 Herbrand
For the analysis of (classical) logic programs (over the Herbrand domain) the CIAO compiler includes a number of traditional domains proposed in the literature for capturing properties such as variable groundness, freeness, sharing, and linearity information. This includes the set sharing \(SH\) [JL89, MH89], sharing and freeness \(SH+FR\) [MH91], and pair sharing \(ASub\) [Son86] domains. Combining the \(SH\) and \(SH+FR\) domains with \(ASub\) is also supported, resulting in the \(SH+ASub\) and \(SH+FR+ASub\) domains. The combination is done in such a way that the original domains and operations of the analyzer over them are re-used, instead of redefining the domains for the combination [CC79, CMB+93]. Two other domains, a modified version of Path sharing [KS95] and \(Aeqns\) (abstract equations) [MSJB95] are currently being incorporated to the system.

### 3.2.2 Constraint Programming
The definiteness abstraction \(Def\) [GH93, Gar94] derives definite interaction between constraints. More precisely, the analysis determines (1) which variables are definite, i.e. constrained to a unique value, with respect to the constraint store in which they occur, and (2) definite dependencies between program variables. These dependencies are used to perform accurate definiteness propagation and are also useful in their own right to perform several program optimizations.

The freeness abstraction \(Fr\) [DJBC93, DJ94, Dum94] derives possible interaction between constraints. This is different from freeness in the \(SH+FR\) domain. More precisely, the analysis determines (1) which variables act as degrees of freedom with respect to the satisfiability of the constraint store in which they occur, and (2) possible dependencies between program variables. These dependencies are used to perform accurate non-freeness propagation and are also useful in their own right to perform program optimizations.

A combination of the \(Def\) and \(Fr\) analyses derives both definiteness and freeness information at the same time. It is incorporated in the compiler as the analyzer \(Fd\). The information at each program point consists of: the set of ground or definite variables, the set of free variables, the set of definite dependencies, and the set of possible dependencies.

An additional domain supported is \(LSign\) [MS94]. This domain is aimed at inferring accurate information about possible interaction between linear arithmetic equalities and inequalities. The key idea is to abstract the actual coefficients and constants in constraints by their "sign". A preliminary implementation of this domain shows very promising accuracy, although at a cost in efficiency.

The information produced using these domains is instrumental in performing optimizations ranging from constraint/goal reordering to program parallelization.

### 3.2.3 Dynamically Scheduled Programs
CIAO also includes a version of the PLAI framework which is capable of accurately analyzing (constraint) programs with dynamic scheduling (e.g., including delay declarations [eA82, Car87]). Being able to analyze constraint languages with dynamic scheduling also allows analyzing CC languages with angelic nondeterminism.\(^3\) Initial studies showed that accurate analysis in such programs is possible [MGH94], although this technique involves relatively large cost in analysis time.

---

\(^2\)Some of these domains have been implemented by other users of the PLAI system, notably the K. U. Leuven, Monash University, and the U. of Melbourne.

\(^3\)This is a kind of nondeterminism which does not give rise to an arbitrary choice when applying a search rule.
The analysis integrated into the CIAO compiler uses a novel method which improves on the previous one by increasing the efficiency without significant loss of accuracy [GMS95]. The approach is based on approximating the delayed atoms by a closure operator. Experimental results show that this approach allows more efficient analysis with similar accuracy.

Another, direct method for analysis of CC programs has been developed and is currently being integrated into the compiler. As in the method mentioned above, this one is based on the observation that most implementations of the concurrent paradigm can be viewed as a computation which proceeds with a fixed, sequential scheduling rule but in which some goals suspend and their execution is postponed until some condition wakes them. Extending previous work of Debray [DGB94, Deb93], we show how, for certain properties, it is possible to extend existing analysis technology for the underlying fixed computation rule in order to deal with such programs [BH95b]. In particular, this idea has been applied using as starting point the original framework for the analysis of sequential programs. The resulting analysis can deal with programs where concurrency is governed by the Andorra model as well as standard CC models. The advantage with respect to the the method above is lower analysis time, in exchange for a certain loss of accuracy.

3.3 Parallelization

The information inferred during the analysis phase is used for independence detection, which is the core of the parallelization process [BGH94, GBH95]. The compile-time parallelization module is currently aimed at uncovering goal-level, restricted (i.e., fork and join), independent and-parallelism (IAP). Independence has the very desirable properties of correct and efficient execution w.r.t. standard sequential execution of Prolog or CLP. In the context of LP, parallelization is performed based on the well-understood concepts of strict and non-strict independence [HR95], using the information provided by the abstract domains. While the notions of independence used in LP are not directly applicable to CLP, specific definitions for CLP (and constraint programming with dynamic scheduling) have been recently proposed [GHM93, Gar94] and they have been incorporated in the CIAO compiler in order to parallelize CLP and CC programs [GHM95]. Additionally, the compiler has side-effect and granularity analyzers (not depicted in Figure 1) which infer information which can yield the sequentialization of goals (even when they are independent) based on efficiency or maintenance of observable behavior.

The actual automatic parallelization of the source program is performed in CIAO during compilation of the program by the so called annotation algorithms. The algorithms currently implemented are: mel, cdg, udg [Mut91, Bue94], and urlp [CH94]. To our knowledge, the CIAO system is the first one to perform automatic compile-time (And-)parallelization of CLP programs [GBH95].

3.4 Optimization

The CIAO compiler performs several forms of code optimization by means of source to source transformations. The information obtained during the analysis phase is not only useful in automatic program parallelization, but also in this program specialization and simplification phase.

The CIAO compiler can optimize programs to different degrees, as indicated by the user. It can just simplify the program, where simplification amounts to reducing literals and predicates which are known to always succeed, fail, or lead to error. This can speed up the
program at run-time, and also be useful to detect errors at compile-time. It can also specialize the program using the versions generated during analysis [PH95a]. This may involve generating different versions of a predicate for different abstract call patterns, thus increasing the program size whenever this allows more optimizations. In order to keep the size of the specialized program as reduced as possible, the number of versions of each predicate is minimized attaining the same results as with Winsborough’s algorithm [Win92].

As well as handling sequential code, the optimization module of the CIAO compiler contains what we believe is the first automatic optimizer for languages with dynamic scheduling [PH95b]. The potential benefits of the optimization of this type of programs were already shown in [MGH94], but they can now be obtained automatically. These kinds of optimizations include simplification and elimination of suspension conditions and elimination of concurrency primitives (sequentialization).

3.5 Output

The CIAO compiler produces several forms of output. It is possible to obtain the results of each of the intermediate compilation phases. This allows visualizing and affecting the transformation, analysis, parallelization, and optimization processes. Because of the source to source nature of the compiler, this output is always a (possibly annotated) kernel CIAO program.

The back end of the compiler takes the result of the previous program transformations and generates a number of final output formats. Normally, the result of the compiler is intended for the CIAO/\&-Prolog abstract machine. Output possibilities are then byte-code (".ql") files, stand-alone executables, and incore compilation (when the compiler is running inside the system rather than as a stand-alone application). As mentioned before, and as an alternative output, most of the capability of the system can also be handled by any Prolog which supports delay declarations and attributed variables. Alternatively, also AKL [JH91] can be used as a target, using the techniques described in [BH95a].

4 Future Work

We have briefly described the current status of the CIAO system. The current main objective of the system is to be an experimentation and evaluation vehicle for programming constructs and optimization and implementation techniques for the programming paradigms of LP, CLP, and CC, and their combinations. The current version of the system is available for experimentation (please contact the authors; further information can be obtained from http://www.dia.fi.upm.es).

We are continuing improving the system. Additionally, we are developing pilot applications with the system which should provide valuable feedback regarding its capabilities.

Much work remains to be done in several areas. While the CIAO system illustrates that analysis and optimization of concurrent programs is possible, much work remains in improving the efficiency and accuracy of the analysis and in improving the performance gains obtained with the resulting optimizations. As mentioned in Section 3.3, the automatic parallelization currently performed in the CIAO system is at the goal level. However, it is possible to parallelize at finer granularity levels, thus obtaining greater degrees of parallelism. The concept of local independence [MRB+94, BHMR94] can be used for this purpose. Although some promising progress has been made in this direction [HCC95], it remains as future work to implement a system fully capable of efficiently exploiting this very fine grained level of
Granularity control is a very important issue in both parallelization of sequential programs and sequentialization of concurrent ones. As mentioned in Section 3.3, the CIAO compiler already has some granularity control capabilities [DLH90, KS90, LHD94, DLHL94, LH95], but much work remains to be done in this very important area.

While our work in detection of parallelism in the CIAO compiler concentrates on compile-time detection of parallelism, run-time detection also needs to be explored. Significant progress has been made in this area by models and systems such as DDAS [She92], Andorra-I, and AKL.

Finally, there remains the issue of what is the ideal, future source language for LP/CLP/CC. CIAO sidesteps this issue by attempting to support several languages (including those that combine several paradigms). This allows concentrating on the implementation issues and developing basic techniques for analysis and optimization that, in the belief that the underlying principles are quite common to the approaches being explored, will hopefully be applicable to future languages. However, the issue of the next generation language is certainly important. Much promising work has been done in this direction in the design of the AKL [JH90] and OZ [Smo94] languages. In fact, interestingly, the kernel CIAO language also offers a (simplistic, but effective) solution to the problem, which is backwards compatible with Prolog and CLP.

References


[Smo94] G. Smolka. The Definition of Kernel Oz. DFKI Oz documentation series, German Research Center for Artificial Intelligence (DFKI), November 1994.

