ADDING METADATA TO ORC TO SUPPORT REASONING ABOUT GRID PROGRAMS *

Marco Aldinucci
Dept. Computer Science – University of Pisa – Italy
aldinuc@di.unipi.it

Marco Danelutto
Dept. Computer Science – University of Pisa – Italy
marcod@di.unipi.it

Peter Kilpatrick
Dept. Computer Science – Queen's University Belfast – UK
p.kilpatrick@qub.ac.uk

Abstract Following earlier work demonstrating the utility of Orc as a means of specifying and reasoning about grid applications we propose the enhancement of such specifications with metadata that provide a means to extend an Orc specification with implementation oriented information. We argue that such specifications provide a useful refinement step in allowing reasoning about implementation related issues ahead of actual implementation or even prototyping. As examples, we demonstrate how such extended specifications can be used for investigating security related issues and for evaluating the cost of handling grid resource faults. The approach emphasises a semi-formal style of reasoning that makes maximum use of programmer domain knowledge and experience.

Keywords: Orc, grid, metadata, fault handling, security.

*This research is carried out under the FP6 Network of Excellence CoreGRID funded by the European Commission (Contract IST-2002-004265).

1. Introduction

Grid computing is intended to enable the development of both industrial and scientific applications on an unprecedented scale in terms of computing power and ubiquity. These applications are supposed to transparently handle dynamicity and heterogeneity of computing platforms [11] and they often exploit some flavour of component programming model. Component technology focuses (by its very nature) on the decoupled development of modules implementing single features [1, 4, 5], that should then be arranged and connected to realize the application. While several frameworks for developing grid-oriented components exist or are under design [7–8], the models to reason about their orchestration are still inadequate. Although a model for orchestration should necessarily subsume a notion of component/module behaviour, it can be specified along a spectrum of abstraction levels: from the full implementation itself to the fully logic/algebraic description. Currently, most of the effort is concentrated on the ends of the spectrum, which are far from the designer's viewpoint. For example, BPEL [6] is a recognized standard for orchestration of Web Services, but it is designed for machine processing and is therefore not suitable for supporting human "abstract reasoning" about orchestration. At the other extreme, π-calculus is a well-recognized formal tool for reasoning about distributed programs [12], but it comes with a heavyweight formal framework typically outside the interest and experience of system designers.

In earlier work we explored the use of Orc [10] as a means of specifying and reasoning about grid computations. Orc was developed as a notation for describing the orchestration of distributed systems, rather than the core computations themselves. Orc's primitive is the *site* which may be used to abstract basic computations. A site call returns a single value or remains silent. Site calls may be combined using three composition operators (plus recursion):
Sequential : $A > x > B(x)$. For each output, x, from A execute an instance of B taking x as parameter. If x is not used in B write $A \gg B$.
Parallel : $A \mid B$. The output is the interleaved outputs from each of A and B.
Asymmetric parallel : $A\ where\ x :\in B$. Execute A and B in parallel until A needs x. Take the first x delivered by B and terminate the remaining execution of B while A continues.

We believe that Orc lies in the middle ground of the spectrum of orchestration description: as described in previous work [3], Orc appears to be a suitable candidate to reason about certain non-functional properties (e.g. fault-tolerance) of the grid-oriented *muskel* system [2]. In this paper we present a further step along the same path. We enrich Orc with *metadata* to describe non-functional properties such as deployment information. This could be used, for example, to describe the mapping of application parts (e.g. components, modules) onto a grid platform. The approach is consistent with the current trend of keeping

decoupled the functional and non-functional aspects of an application. We believe that the use of metadata introduces a new dimension for reasoning about the orchestration of a distributed system by allowing a narrowing of the focus from the very general case. We expect this approach can be gracefully extended in order to allow reasoning – at design time – about several static invariants of the final implementation.

2. Orc metadata

A generic Orc program, as described in [10], is a set of Orc *definitions* followed by an Orc *goal expression*. The goal expression is the expression to be evaluated when executing the program. Assume $\mathcal{S} = \{s_1, \ldots, s_s\}$ is the set of *sites* used in the program, i.e. the set of all the sites *called* during the evaluation of the goal expression (the set does not include the pre-defined sites, such as *if* and *Rtimer*, as they are assumed to be available at any user defined site), and $\mathcal{E} = \{e_0, \ldots, e_e\}$ is the set including the goal expression (e_0) and all the "head" expressions appearing in the left hand sides of Orc definitions.

The set of *metadata* associated with an Orc program may be defined as the set: $\mathcal{M} = \{\mu_1, \ldots, \mu_n\}$ where $\mu_i = \langle t_j, md_k \rangle$ with $t_j \in \mathcal{S} \cup \mathcal{E}$ and $md_k = f(p_1, \ldots, p_{n_k})$. f is a generic "functor" (represented by an identifier) and p_i are generic "parameters" (variables, ground values, etc.). The metadata md_k are not further defined as, in general, metadata structure depends on the kind of metadata to be represented. In the following, examples of such metadata are presented.

As is usual, the semantics of Orc is not affected when *meta*data is taken into account. Rather, the introduction of metadata provides a means to restrict the set of actual implementations which satisfy an Orc specification and thereby eases the burden of reasoning about properties of the specification. For example, restrictions can be placed on the relative physical placement of Orc sites in such a way that conclusions can be drawn about their interaction which would not be possible in the general case.

Suppose one wishes to reason about Orc program site "placement", i.e. about information concerning the relative positioning of Orc sites with respect to a given set of *physical resources* potentially able to host one or more Orc sites. Let $\mathcal{R} = \{r_1, \ldots, r_r\}$ be the set of available physical resources. Then, given a program with $\mathcal{S} = \{siteA, siteB\}$ we can consider adding to the program metadata such as $\mathcal{M} = \{\langle siteA, loc(r_1)\rangle, \langle siteB, loc(r_2)\rangle\}$ modelling the situation where $siteA$ and $siteB$ are placed on distinct processing resources. Define also the auxiliary function $location(x) : \mathcal{S} \times \mathcal{E} \rightarrow \mathcal{R}$ as the function returning the location of a site/expression and consider a metadata set *ground* if it contains location tuples relative to *all* the sites in the program (that is, all sites have been allocated to a processor).

loc metadata can be used to support reasoning about the "communication costs" of Orc programs. For example, the cost of a communication with respect to the placement of the sites involved can be characterized by distinguishing cases:

$$k_{Comm} = \begin{cases} k_{nonloc} & \textit{if } location(s_1) \neq location(s_2) \\ k_{loc} & \textit{otherwise} \end{cases}$$

where s_1 and s_2 are the source and destination sites of the communication, respectively and, typically, $k_{nonloc} \gg k_{loc}$.

Consider now a second example of metadata. Suppose "secure" and "insecure" site locations are to be represented. Secure locations can be reached through trusted network segments and can therefore be communicated with while taking no particular care; insecure locations are not trusted, and can be reached only by passing through untrusted network segments, therefore requiring some kind of explicit data encryption to guarantee security. This representation can be achieved by simply adding to the metadata tuples such as $\langle s_i, trusted() \rangle$ or $\langle s_i, untrusted() \rangle$. Then a costing model for communications that takes into account that transmission of encrypted data may cost significantly more than transmission of plain data can be devised.

$$k_{SecComm} = \begin{cases} k_{InSecComm} & \textit{if } \langle s_1, untrusted() \rangle \in \mathcal{M} \\ & \quad \vee \langle s_2, untrusted() \rangle \in \mathcal{M} \\ k_{Comm} & \textit{otherwise} \end{cases}$$

2.1 Generating metadata

So far the metadata considered have been identified explicitly by the user. In some cases he/she may not wish, or indeed be able, to supply all of the metadata and so it may be appropriate to allow generation of metadata from partial metadata supplied by the user. For example, suppose the user provides only partial location metadata, e.g. metadata relative to the goal expression location and/or metadata relative to the location of the components of the topmost parallel command found in the Orc program execution. Metadata information available can be used to infer ground location metadata (i.e. location metadata for all $s \in \mathcal{S}$) as follows. Consider two cases: in the first (*completely distributed* strategy) it is assumed that each time a new site in the Orc program is encountered, the site is "allocated" on a location that is distinct from the locations already used. In the second case (*conservative* strategy) new sites are allocated in the same location as their parent (w.r.t. the syntactic structure of the Orc program), unless the user/programmer specifies something different in the provided metadata.

More formally, in the first case, we can state that when an Orc definition such as $E \triangleq f \mid g$, $E \triangleq f(x)$ *where* $x :\in g$, $E \triangleq f \gg g$ or $E \triangleq f > x > g$

is considered, both the metadata $\langle f, loc(freshLoc(\mathcal{M})\rangle$ and $\langle g$, $loc(freshLoc(\mathcal{M}))$ are added to $\mathcal{M}$. In the second case, the same Orc definitions will lead to insertion in the set $\mathcal{M}$ of the new metadata $\langle f, location(E)\rangle$ and $\langle g, location(E)\rangle$ (provided the user did not explicitly supply site metadata information for f and g).

Example To illustrate the use of metadata, consider the following description of a classical task farm (embarrassingly parallel computation):

$farm(pgm, nw) \triangleq tasksource \mid resultsink \mid workers(pgm, nw)$
$workers(pgm, nw) \triangleq \mid i : 1 \leq i \leq nw : worker_i(pgm)$
$worker(pgm) \triangleq tasksource > t > pgm > y > resultsink(y) \gg worker(pgm)$

A typical goal for this program will be of the form $farm(myPgm, 10)$. Suppose the user provides the metadata:

$\forall i \in [1, nw]\langle worker_i, loc(PE_i)\rangle \in \mathcal{M}$
$\langle farm(myPgm, 10), strategy(fullyDistributed)\rangle \in \mathcal{M}$

where $strategy(fullyDistributed)$ means the user explicitly requires that a "completely distributed implementation" be used. An attempt to infer metadata about the goal expression identifies $location(farm(myPgm, 10)) = \perp$ but, as the strategy requested by the user is *fullyDistributed* and as $farm(pgm, nw)$ is defined as a parallel command, the following metadata is added to $\mathcal{M}$:

$\langle tasksource, loc(freshLoc(\mathcal{M}))\rangle$
$\langle resultsink, loc(freshLoc(\mathcal{M}))\rangle$
$\langle workers(pgm, nw), loc(freshLoc(\mathcal{M}))\rangle$.

Next, expanding the *workers* term, gives the term

$\mid i : 1 \leq i \leq nw : worker_i(pgm)$

but in this case metadata relative to *worker*$_i$ has already been supplied by the user. At this point

$\mathcal{M} = \{ \langle tasksource, loc(freshLoc(\mathcal{M}))\rangle, \langle resultsink, loc(freshLoc(\mathcal{M}))\rangle,$
$\langle workers(pgm, nw), loc(freshLoc(\mathcal{M}))\rangle, \langle worker_1, loc(PE_1\rangle, \ldots,$
$\langle worker_{nw}, loc(PE_{nw}\rangle\}$

and therefore is *ground* w.r.t. the program.

Thus, in addition to the location metadata provided by the user it was possible to derive the fact that the locations of *tasksource* and *resultsink* are distinct and, in addition, are different from the locations of each *worker*$_i$. Suppose now that the user has also inserted the metadata item $\langle PE_2, untrusted()\rangle$ in addition to those already mentioned. That is, one of the placement locations is untrusted. This raises the issue of how it can be determined whether or not a communication must be performed in a secure way. This information may be inferred from the available metadata as follows. Let functions $source(C)$ denote a site "sending" data and $sink(C)$ denote a site "receiving" data in communication

C. Then C *must* be secured iff

$$source(C) = X \wedge sink(C) = Y \wedge \langle X, loc(LX\rangle \in \mathcal{M} \wedge \langle Y, loc(LY\rangle \in \mathcal{M}$$
$$\wedge (\langle LX, untrusted()\rangle \in \mathcal{M} \vee \langle LY, untrusted()\rangle \in \mathcal{M}).$$

Thus, for the farm example above, the metadata $\langle worker_2, PE_2\rangle$ and $\langle PE_2, untrusted()\rangle$ and the definition

$$worker_2(pgm) \triangleq tasksource > t > pgm > y > resultsink \gg worker_2(pgm)$$

together with the metadata $\langle tasksource, loc(TS)\rangle$, $\langle resultsink, loc(RS)\rangle$, $\langle TS, trusted()\rangle$, $\langle RS, trusted()\rangle$ lead to the conclusion that the communications represented in the Orc code by $tasksource > t > pgm.compute(t)$ and by $pgm.compute(t) > y > resultsink$ within *worker*$_2$ must be secured.

It is worth pointing out that the metadata considered here is typical of the information needed when running grid applications. For example, constraints such as the *loc* ones can be generated to force code (that is, sites) to be executed on processing elements having particular features, and information such as that modelled by *untrusted* metadata can be used to denote those cluster nodes that happen to be outside a given network administrative domain and may therefore be more easily subject to "man in the middle" attacks or to some other kind of security related leaks.

3. Metadata exploitation

In this section we consider two alternative versions of a tool and use their Orc specifications together with metadata to analyse their performance and security properties. `muskel` [9] is a skeleton-based parallel programming environment written in Java. `muskel` converts a user program to a data flow graph which is stored in a *taskpool*. Program input is handled as an input token to a fresh copy of the data flow graph placed in the taskpool. Fireable instructions (*tasks*) in the taskpool are computed by a set of *remote worker* processors that are recruited for the job. Each remote worker is under the supervision of a *control thread* that accesses the *taskpool*, sends a task to its worker and places the result in the *resultpool*.

The first version of `muskel` considered here includes a *manager* that is responsible for recruitment of remote workers, their allocation to control threads and the handling of remote worker failure. This represents the original (centralized) version of `muskel`, but the presence of such a manager was seen as a potential single point of failure. [3] describes how the original specification was analysed and modified to obtain a revised (decentralized) version in which this single point of failure was removed by making each control thread responsible for its own remote worker recruitment. Here, using metadata, we examine the efficiency implications of such a policy change. The Orc model of the decen-

tralized version is given in Figure 1; the Orc model of the centralized version can be found in [3].

$$
\begin{array}{l}
systemDistribManager(pgm, tasks, contract, G, t) \triangleq \\
\quad taskpool.add(tasks) \mid i : 1 \leq i \leq contract : ctrlthread_i(pgm, t, G) \\
ctrlthread_i(pgm, t, G) \triangleq discover(G, pgm) > rw > ctrlprocess(pgm, rw, t, G) \\
discover(G, pgm) \triangleq let(rw) \text{ where } rw :\in |_{g \in G}\ g.can\_execute(pgm) \\
ctrlprocess(pgm, rw, t, G) \triangleq taskpool.get > tk > \\
\quad (\ \text{if } valid \gg resultpool.add(r) \gg ctrlprocess(pgm, rw, t, G) \\
\quad \mid \text{if } \neg valid \gg taskpool.add(tk) \\
\qquad\qquad \mid discover(G, pgm) > w > \\
\qquad\qquad\quad ctrlprocess(pgm, w, t, G) \quad) \\
\quad \text{where } (valid, r) :\in \\
\quad\ (\ remoteworker(pgm, tk) > r > let(true, r) \\
\quad\ \mid Rtimer(t) \gg let(false, 0)\ )
\end{array}
$$

Figure 1. Decentralized manager `muskel` specification in Orc.

3.1 Comparison of communication costs

In comparing the two versions of `muskel`, as is typical in such studies, the focus will be on the "steady state" performance, that is, the typical activity of a control thread when it is processing tasks. There are two possibilities: the task is processed normally and the result placed in the *resultpool* or the remote worker fails and the control thread requires a new worker. In analysing the specifications a conservative placement strategy will be assumed; that is, the sub-parts of an entity are assumed to be co-located with their parent unless otherwise stated. Given the following metadata supplied by the developer:

$\forall rw_i \in G.\langle rw_i, loc(PE_i)\rangle \in \mathcal{M}$
$\langle system, loc(C)\rangle \in \mathcal{M}$
$\langle system(myPgm, tasks, 10, G, 50), strategy(conservative)\rangle \in \mathcal{M}$

the rules for propagation and the strategy adopted ensure that the following metadata are present for both versions:

$\langle rw_i, loc(PE_i)\rangle$, $\langle ctrlthread_i, loc(C)\rangle$, $\langle taskpool, loc(C)\rangle$, $\langle resultpool, loc(C)\rangle$, $\langle rworkerpool, loc(C)\rangle$.

In addition, for the decentralized version, $\langle cntrlprocess, loc(C)\rangle$ is present.

Normal processing For the centralized version, examination of the definition of *cntrlthread* shows that in the case of a normal calculation the following sequence of actions will occur:

$taskpool.get > tk > remw(pgm, tk) > r > let(true, r) \gg resultpool.add(r).$

Using the metadata, and reasoning in the same way as in the farm example, it can be seen that the communication of the task *tk* to the remote worker and the subsequent return of the result *r* to the control thread represent non-local communications; all other communications in this sequence are local.

Similar analysis of the decentralized version reveals an identical series of actions for normal processing and an identical pattern of communications. Naturally then, similar results from the two versions for normal processing would be expected, and indeed this is borne out by experiment - see section 4.

Fault processing Now consider the situation where a remote worker fails during the processing of a task. In both versions the *Rtimer* timeout occurs, the task being processed is returned to the *taskpool* and a new worker is recruited. In the centralized version the following sequence of events occurs:

$$taskpool.get \gg Rtimer(t) \gg let(false, 0) \gg taskpool.add(tk) \gg rworkerpool.get(remw)$$

while in the decentralized version the events are effectively:

$$taskpool.get \gg Rtimer(t) \gg let(false, 0) \gg taskpool.add(tk) \gg rw.can\_execute(pgm) > rw > let(g)$$

where *rw* is the first site in *G* to respond.

Analysis of these sequences together with the metadata reveals that the comparison reduces to the local communication to the *rworkerpool* in the centralized version versus the non-local call to the remote site *rw* in the decentralized version. This comparison would suggest that, in the case of fault handling, the centralized version would be faster than the decentralized version and, again, this is borne out by experiment.

3.2 Comparison of security costs

Consider now the issue of security. Suppose that one of the remote workers, say rw_2, is in a non-trusted location (that is $\langle PE_2, untrusted() \rangle \in \mathcal{M}$). The implications of this can be determined by analysing the specification together with the metadata. In this case, as $\langle rw_2, loc(PE_2) \rangle \in \mathcal{M}$ we can conclude that $cntrlthread_2$ will be affected (while it is operating with its initially allocated remote worker) to the extent that the communications to and from its remote worker must be secured. This prompts reworking of the specification to split the control threads into two parallel sets: those requiring secure communications and those operating exclusively in trusted environments. In this way the effect, and hence cost, of securing communications can be minimised. Experimental results in section 4 illustrate the cost of securing the communications with differing numbers of control threads.

4. Experimental results

We ran several experiments, on a distributed configuration of Linux machines, aimed at verifying that the results obtained from analysis of the Orc specifications of `muskel` together with metadata are consistent with practice.

We first verified that centralized and decentralized manager versions of `muskel` perform the same (up to a reasonable percentage difference) when no faults occur in the resources used for remote program execution. We ran the same `muskel` program with both the centralized and the decentralized `muskel` implementation, using up to 4 processing elements for the remote macro data flow interpreter instances: we obtained differences in completion time not exceeding 1.6% (1.05% average).

Then we considered remote resource failure. We measured the time spent in handling a single fault in several runs on both centralized and decentralized `muskel` versions. The distributed version takes longer to handle a single fault, as expected looking at the Orc models of the two implementations: 128.4 vs. 114.4 msecs, average. Finally, we attempted to verify the effectiveness of limiting secure mechanism usage to communications involving untrusted nodes, which may be identified by examination of the Orc specifications with associated metadata. Figure 2 shows the completion time of a `muskel` program whose remote worker sites are running on a variable mix of *trusted* and *untrusted* locations. The greater the number of remote interpreters exploited using secure mechanisms, the lower the performance values that are achieved. Therefore, restricting the classification of insecure nodes by analysis of metadata results in better efficiency on the target architecture.

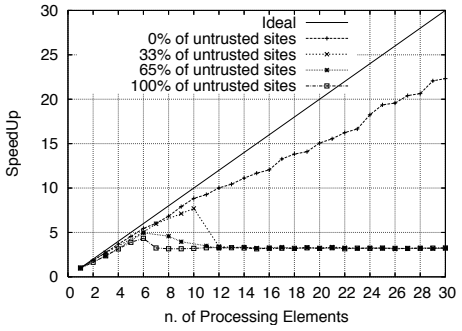

Figure 2. Comparison of runs involving different percentages of *untrusted* locations

5. Conclusions

We have shown how, by associating metadata with an Orc specification, we can reason about the specification and that this reasoning carries through to the actual grid code which implements the specification. In particular, we considered how user provided metadata can be associated with the Orc model of a real structured grid programming environment (`muskel`) and showed how this could be used to perform qualitative performance comparison between two different versions of the programming environment, as well as to determine how the overhead introduced by security techniques can be minimized. We compared

these theoretical results with actual experimental results and verified that they qualitatively match. Thus, the availability of an Orc model on which to "hang" the metadata allows metadata to be exploited *before* the actual implementation is available. We are currently working to formalize and automate the techniques discussed here. In particular, we are aiming to implement tools to support the metadata propagation and reasoning procedures adopted. It should be noted, however, that the whole approach, based on Orc, as described here and in [3] encourages the use of *semi*-formal reasoning to support program development (both program design and refinement). (Thus, for example, the equivalence of Orc specifications and the `muskel` implementations is not formally proven.) We believe this approach has the potential to reduce substantially experimentation by allowing the exploration of alternatives prior to costly implementation and *without* recourse to full-blown formal treatment.

References

[1] M. Aldinucci, S. Campa, M. Coppola, M. Danelutto, D. Laforenza, D. Puppin, L. Scarponi, M. Vanneschi, and C. Zoccolo. Components for high performance grid programming in grid.it. *Proc. of the Intl. Workshop on Component Models and Systems for Grid Applications*, CoreGRID series, pages 19–38, Saint-Malo, France, Jan. 2005. Springer.

[2] M. Aldinucci and M. Danelutto. Algorithmic skeletons meeting grids. *Parallel Computing*, 32(7):449–462, 2006. DOI:10.1016/j.parco.2006.04.001.

[3] M. Aldinucci, M. Danelutto, and P. Kilpatrick. Management in distributed systems: a semi-formal approach. TR-07-05, Univ. of Pisa, Dept. of Comp. Science, Feb. 2007.

[4] M. Alt, J. Dünnweber, J. Müller, and S. Gorlatch. HOCs: Higher-order components for grids. In *Component Models and Systems for Grid Applications*, CoreGRID series, pages 157–166. Springer, Jan. 2005.

[5] F. Baude, D. Caromel, and M. Morel. On hierarchical, parallel and distributed components for grid programming. *Proc. of the Intl. Workshop on Component Models and Systems for Grid Applications*, CoreGRID series, pages 97–108, Jan. 2005. Springer.

[6] Business Process Execution Language for Web Services version 1.1, 2007. `http://www-128.ibm.com/developerworks/library/specification/ws-bpel/`.

[7] The Common Component Architecture Forum, 2007. `http://www.cca-forum.org/`.

[8] CoreGRID NoE deliverable series, Institute on Programming Model. *Deliverable D.PM.04 – Basic Features of the Grid Component Model (assessed)*, Feb. 2007.

[9] M. Danelutto and P. Dazzi. Joint structured/non structured parallelism exploitation through data flow. *Proc. of ICCS: Intl. Conference on Computational Science, WS on Practical Aspects of High-level Parallel Programming*, LNCS, Reading, UK, May 2006. Springer.

[10] J. Misra and W. R. Cook. Computation orchestration: A basis for a wide-area computing. *Software and Systems Modeling*, 2006. DOI 10.1007/s10270-006-0012-1.

[11] Next Generation GRIDs Expert Group. *Future for European Grids: GRIDs and Service Oriented Knowledge Utilities. Vision and Research Directions 2010 and Beyond*, 2006. `http://cordis.europa.eu/ist/grids/ngg.htm`.

[12] H. Smith and P. Fingar. Workflow is just a pi process. *BPTrends*, pages 1–36, 2004.