A Graph-Based Semantics Workbench for Concurrent Asynchronous Programs

Claudio Corrodi[1,2(✉)], Alexander Heußner[3], and Christopher M. Poskitt[1,4]

[1] Department of Computer Science, ETH Zürich, Zürich, Switzerland
[2] Software Composition Group, University of Bern, Bern, Switzerland
corrodi@inf.unibe.ch
[3] Software Technologies Research Group, University of Bamberg, Bamberg, Germany
[4] Singapore University of Technology and Design, Singapore, Singapore

Abstract. A number of novel programming languages and libraries have been proposed that offer simpler-to-use models of concurrency than threads. It is challenging, however, to devise execution models that successfully realise their abstractions without forfeiting performance or introducing unintended behaviours. This is exemplified by SCOOP—a concurrent object-oriented message-passing language—which has seen multiple semantics proposed and implemented over its evolution. We propose a "semantics workbench" with fully and semi-automatic tools for SCOOP, that can be used to analyse and compare programs with respect to different execution models. We demonstrate its use in checking the consistency of semantics by applying it to a set of representative programs, and highlighting a deadlock-related discrepancy between the principal execution models of the language. Our workbench is based on a modular and parameterisable graph transformation semantics implemented in the GROOVE tool. We discuss how graph transformations are leveraged to atomically model intricate language abstractions, and how the visual yet algebraic nature of the model can be used to ascertain soundness.

1 Introduction

To harness the power of multi-core and distributed architectures, software engineers must program with concurrency, asynchronicity, and parallelism in mind. Classical thread-based approaches to concurrent programming, however, are difficult to master and error prone. To address this, a number of programming APIs, libraries, and languages have been proposed that provide safer and simpler-to-use models of concurrency, such as the block-dispatch model of Grand Central Dispatch [14], or the message-passing-based model of SCOOP [40]. The concurrent programming abstractions that these languages provide rely on the existence of effective execution models for realising them; *effective* in the sense that they do

C. Corrodi and C.M. Poskitt—Research done whilst employed by the Chair of Software Engineering, ETH Zürich.

P. Stevens and A. Wąsowski (Eds.): FASE 2016, LNCS 9633, pp. 31–48, 2016.
DOI: 10.1007/978-3-662-49665-7_3

so without forfeiting performance or introducing unintended behaviours. Devising execution models that successfully reconcile these requirements, however, is challenging: a model that is too restrictive can deny desirable concurrency and lead to unnecessary bottlenecks; a model that is too permissive can lead to surprising and unexpected executions.

This challenge is exemplified by SCOOP [40], a message-passing paradigm for concurrent object-oriented programming that aims to preserve the well-understood modes of reasoning enjoyed by sequential programs, such as pre- and postcondition reasoning over blocks of code. Although the high-level language mechanisms for achieving this were described informally as early as the '90s [24,25], it took several years to understand how to effectively implement them: execution models [6,26,40], prototypes [28,37], and contrasting versions of a production-level implementation [11] gradually appeared over the last decade, and can be seen as representing multiple, partially conflicting semantics for realising SCOOP. They are also unlikely to be the last, as new language features continue to be proposed, prototyped, and integrated, e.g. [27]. Despite the possible ramifications to behavioural and safety properties of existing programs, little work has been done to support formal and automatic *comparisons* of the program executions permitted by these different semantics. While general, tool-supported formalisations exist—in *Maude*'s conditional rewriting logic [26], for example, and in a custom-built CSP model checker [6]—these are tied to particular execution models, do not operate on program source code, and are geared towards "testing" the semantics rather than general verification tasks. Furthermore, owing to the need to handle waiting queues, locks, asynchronous remote calls, and several other intricate features of the SCOOP execution models, these formalisations quickly become complex, making it challenging to ascertain their soundness with language designers who lack a formal methods background.

The Challenge. There is a need for languages like SCOOP to have tools that not only support the prototyping of new semantics (and semantic extensions), but that also facilitate formal, automatic, and practical analyses for comparing the executions permitted by these semantics, and highlighting where behavioural and safety-related discrepancies arise. The underlying formalism for modelling the semantics should not be ad hoc; rather, it should support re-use, a modular design, and be easily extensible for language evolutions and changes. Furthermore, such tools should be usable in practice: the modelling formalism must be accessible to and understandable by software engineers, and the analyses must support several idiomatic uses of the language mechanisms.

Our Contributions. We propose a "semantics workbench" equipped with fully and semi-automatic tools for SCOOP, that can be used to analyse and compare programs with respect to different execution models for the purpose of checking their consistency. We demonstrate its use by formalising the two principal execution models of SCOOP, analysing a representative set of programs with respect to both, and highlighting some behavioural and deadlock-related discrepancies that the workbench uncovers automatically. Our workbench is based on a

modular and parameterisable graph transformation system (GTS) semantics, built upon our preliminary modelling ideas in [18], and implemented in the general-purpose GTS analysis tool GROOVE [16]. We leverage this powerful formalism to atomically model complex programmer-level abstractions, and show how its inherently visual yet algebraic nature can be used to ascertain soundness. For language designers, this paper presents a transferable approach to checking the consistency of competing semantics for realising concurrency abstractions. For the graph transformation community, it presents our experiences of applying a state-of-the-art GTS tool to a non-trivial and practical problem in programming language design. For the broader verification community, it highlights a need for semantics-parameterised verification, and shows how GTS-based formalisms and tools can be used to derive an effective and modular solution. For software engineers, it provides a powerful workbench for crystallising their mental models of SCOOP, thus helping them to write better quality code and (where need be) port it across different SCOOP implementations.

Plan of the Paper. After introducing the SCOOP concurrency paradigm and its two most established execution models (Sect. 2), we introduce our formal modelling framework based on GTS, and show how to formalise different, parameterisable SCOOP semantics (Sect. 3). Implementing our ideas in a small toolchain (Sect. 4) allows us to check the consistency of semantics across a set of representative SCOOP programs (Sect. 5), and highlight both a behavioural and deadlock-related discrepancy. To conclude, we summarise some related work (Sect. 6), our contributions, and some future research directions (Sect. 7).

2 SCOOP and its Execution Models

SCOOP [40] is a message-passing paradigm for concurrent object-oriented programming that aims to preserve the well-understood modes of reasoning enjoyed by sequential programs; in particular, pre- and postcondition reasoning over blocks of code. This section introduces the programmer-level language mechanisms and reasoning guarantees of SCOOP, as well as its two most established execution models. These will be described in the context of SCOOP's production-level implementation for Eiffel [11], but the ideas generalise to any object-oriented language (as explored, e.g. for Java [37]).

Language Mechanisms. In SCOOP, every object is associated with a *handler* (also called a *processor*), a concurrent thread of execution with the exclusive right to call methods on the objects it handles. In this context, object references may point to objects with the same handler (*non-separate* objects) or to objects with distinct handlers (*separate* objects). Method calls on non-separate objects are executed immediately by the shared handler. To make a call on a separate object, however, a *request* must be sent to the handler of that object to process it: if the method is a *command* (i.e. it does not return a result) then it is executed asynchronously, leading to concurrency; if it is a *query* (i.e. a result is returned

and must be waited for) then it is executed synchronously. Note that handlers cannot synchronise via shared memory: only by exchanging requests.

The possibility for objects to have different handlers is captured in the type system by the keyword `separate`. To request method calls on objects of `separate` type, programmers simply make the calls within *separate blocks*. These can be explicit (we will use the syntax `separate x,y, ... do ... end`); but they also exist implicitly for methods with separate objects as parameters.

Reasoning Guarantees. SCOOP provides certain guarantees about the order in which calls in separate blocks are executed to help programmers avoid concurrency errors. In particular, method calls on separate objects will be logged as requests by their handlers in the order that they are given in the program text; furthermore, there will be no intervening requests logged from other handlers. These guarantees exclude object-level data races by construction, and allow programmers to apply pre- and postcondition reasoning within separate blocks independently of the rest of the program. Consider the following example (adapted from [40]), in which two distinct handlers are respectively executing blocks that set the "colours" of two `separate` objects:

```
separate x,y
do
    x.set_colour (Green)
    y.set_colour (Green)
end
```

```
separate x,y
do
    x.set_colour (Indigo)
    a_colour = x.get_colour
    y.set_colour (a_colour)
end
```

The guarantees ensure that whilst a handler is inside its `separate x,y` block, the other handler cannot log intervening calls on `x` or `y`. Consequently, if the colours are later queried in another `separate x,y` block, both of them will be Green or both of them will be Indigo; interleavings permitting any other combination to be observed are entirely excluded. This additional control over the order in which requests are processed represents a twist on classical message-passing models, such as the actor model [1], and programming languages like Erlang [2] that implement them.

Execution Models. The abstractions of SCOOP require an execution model that can realise them without forfeiting performance or introducing unintended behaviours. Two contrasting models have been supported by different versions of the implementation: initially, a model we call Request Queues (RQ) [26], and a model that has now replaced it which we will call Queues of Queues (QoQ) [40].

The RQ execution model associates each handler with a single FIFO queue for storing incoming requests. To ensure the reasoning guarantees, each queue is protected by a lock, which another handler must acquire to be able to log a request on the queue. Realising a `separate x,y,...` block then boils down to acquiring locks on the request queues attached to the handlers of `x,y,...` and exclusively holding them for the duration of the block. This coarse-grained solution successfully prevents intervening requests from being logged, but leads to performance bottlenecks in several situations (e.g. multiple handlers vying for the lock of a highly contested request queue).

In contrast, the QoQ execution model associates each handler with a FIFO queue that itself contains (possibly several) FIFO subqueues for storing incoming requests. These subqueues represent "private areas" for handlers to log requests without interference from other handlers. Realising a `separate x,y,...` block no longer requires vying for locks; instead, the handlers of `x,y,...` simply generate private subqueues on which requests can be logged without interruption for the duration of the block. If another handler also wants to log requests, then a new private subqueue is generated for it and its requests can be logged at the same time. The QoQ model removes the performance bottlenecks caused by the locks of RQ, while still ensuring the SCOOP reasoning guarantees by completely processing subqueues in the order that they were generated.

Figure 1 visualises three handlers (h_1, h_2, h_3) logging requests (green blocks) on another handler (h_0) under the two execution models. Note that the RQ and QoQ implementations (i.e. compilers and runtimes) include additional optimisations, and strictly speaking, can themselves be viewed as competing semantics.

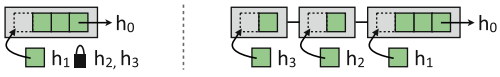

Fig. 1. Logging requests under the RQ (left) and QoQ (right) execution models

Semantic Discrepancies. Discrepancies between the execution models can arise in practice. In the mental model of programmers, with RQ, separate blocks had become synonymous with acquiring and holding locks—which are not implied by the basic reasoning guarantees or the QoQ model. This discrepancy comes to light with the classical dining philosophers program (as provided in the official SCOOP documentation [11]), which will form a running example for this paper. Under RQ, Listing 1 ("eager" philosophers) solves the problem by relying on the implicit parallel acquisition of locks on the forks' handlers; no two adjacent philosophers can be in their separate blocks (representing "eating") at the same time. Under RQ, Listing 2 ("lazy" philosophers) can lead to circular deadlocks, as philosophers acquire the locks in turn. With QoQ however—where there is no implicit locking—neither version represents a solution, and neither can cause a deadlock; yet the basic guarantees about the order of logged requests remain satisfied. We will return to this example in later sections, and show how such discrepancies can be detected by our workbench.

3 A Graph-Based Semantic Model for the SCOOP Family

There are several established and contrasting semantics of SCOOP [6,18,26,29,40], including a comprehensive reference semantics for RQ in *Maude*'s conditional rewriting logic [26], and a semantics for the core of QoQ in the form of

```
separate left_fork, right_fork
do
  left_fork.use
  right_fork.use
end
```

Listing 1. Eager philosophers

```
separate left_fork
do
  separate right_fork
  do
    left_fork.use
    right_fork.use
  end
end
```

Listing 2. Lazy philosophers

simple structural operational rules [40]. These formalisations, however, cannot easily be used for semantic comparisons, due to their varying levels of detail, coverage, extensibility, and tool support. Hence we present in this section "yet another" semantic model, called SCOOP-GTS, based on our preliminary modelling ideas for RQ in [18], using the formalism of graph transformation systems (GTS).

Our reasons to introduce SCOOP-GTS are manifold: (a) we need a common modelling ground that can be parameterised by models of RQ and QoQ; (b) known models based on algebra, process calculi, automata, or Petri nets do not straightforwardly cover SCOOP's asynchronous concurrent nature, or would hide these features in intricate encodings; (c) existing approaches are often proposed from a theoretician's point of view and are not easily readable by software engineers, whereas graphs and diagrammatic notations (e.g. UML) might already be used in their everyday work. Choosing graph transformations as our base formalism is well-justified, as they satisfy the above requirements, and reconcile the goal to have a theoretically rigorous formalisation with the goal to be accessible to software engineers, e.g. for expert interviews with the language implementers (see [31] for a detailed discussion of the pros and cons of GTS in this setting). The "non-linear" context of graph rewriting rules proves to be a powerful mechanism for defining semantics and their interfaces for parameterisation.

We formalised SCOOP-GTS using the state-of-the-art GTS tool GROOVE [17]. Due to limited space, we provide all the files necessary to browse our GTS model as supplementary material [36], including input graphs generated from the example programs of Sect. 5 that can be simulated, analysed, and verified.

SCOOP-Graphs. Each global configuration of a SCOOP program, i.e. snapshot of the global state, is represented by a directed, typed attributed graph consisting of (i) *handler nodes* representing SCOOP's handlers, i.e. basic execution units; (ii) a representation of each handler's *local memory* (i.e. "heap" of non-separate objects) and its known neighbourhood, consisting of references to separate objects that can be addressed by queries and commands; (iii) a representation of each handler's *stack*, via stack frames that model recursive calls to non-separate objects; (iv) *requests* for modelling separate calls, which are stored in (v) subgraphs representing each handler's *input work queue*; (vi) a global *control flow graph* (CFG) presenting the program's execution blocks (consisting of states and actions/transitions in-between); (vii) relations to model inter-handler and handler-memory relations (e.g. locking, waiting, etc.) and to assign each

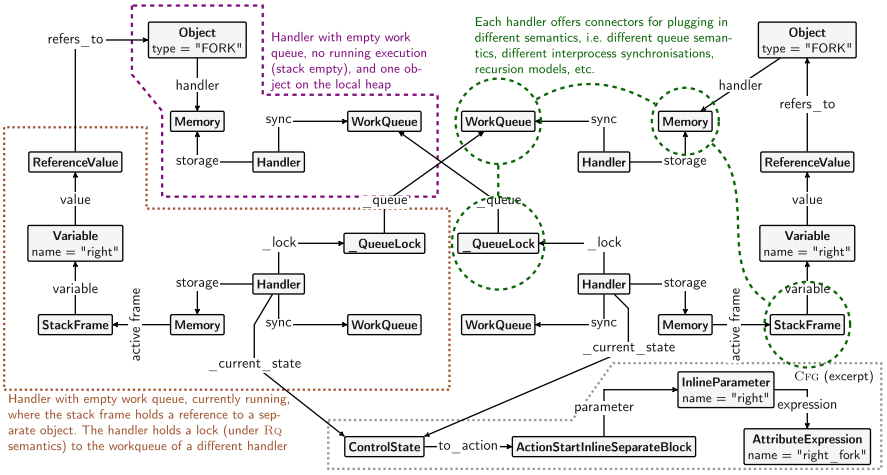

Fig. 2. Reachable deadlock under RQ for the lazy philosophers program (Listing 2) simplified from GROOVE output with additional highlighting and information in colour

handler to its current state in the CFG; and (viii) additional bookkeeping nodes, e.g. containing information on detected deadlocks, and nodes to model the interfaces/contexts for semantic parameterisation. An example SCOOP-Graph can be seen in Fig. 2, depicting a configuration with two concurrently running and two idle handlers.

GTS-Based Operational Semantics. The operational semantics of SCOOP-GTS is given by graph-rewriting rules that are regimented by *control programs*. An example rule, concisely written using nesting as supported by GROOVE, can be seen in Fig. 3. Note that nested rules (including ∀- and ∃-quantification) allow us to express complex, atomic rule matchings in a relatively straightforward and brief way (compared to rules in classical operational semantics, e.g. in [40] for multiple handler reservations). A simplified, example control program can be seen in Listing 3. Control programs allow us to make an execution model's scheduler explicit (and thus open to parameterisation) and help us to implement "garbage collection" for the model (e.g. removing bookkeeping edges no longer needed). Furthermore, they provide a fine-grained way to control the atomicity of SCOOP operations, by wrapping sequences of rule applications into so-called *recipes*.

Semantic Modularity of SCOOP-GTS. We support semantic parameterisation for SCOOP-GTS by providing fixed module interfaces in the graph via special "plug-in nodes/edges" (e.g. `WorkQueue`, `Memory`, `StackFrame` in Fig. 2), and changing only the set of GTS rules that operate on the subgraphs that they guard. We have modelled both RQ and QOQ with distinct sets of rules that operate on the subgraphs guarded by `WorkQueue`: we call the model parameterised by RQ and QOQ respectively SCOOP-GTS(RQ) and SCOOP-GTS(QOQ).

```
initialize_model;                                  // call gts rule for initialisation
while (progress & no_error) {
  for each handler p:                              // choose handlers under some scheduling strategy
    alap handler_local_execution_step(p)+;         // each handler executes local actions as long as possible
  try synchronisation_step;                        // then try (one) possible global synchronisation step
}
recipe handler_local_execution_step (p){
  try separate_object_creation(p)+;                // try local actions that are possibly applicable
  else try assignment_to_variable(p)+;
  else try ... ;                                   // sequentially try all other possible actions
  try clean_up_model+;                             // do some"garbage collection" to keep the model small
}
recipe synchronisation_step(){
  reserve_handlers | dequeue_task | ...;           // non-deterministically try to synchronise
}
...                                                // remaining recipes (core functionality)
// ---------- plug in ----------------------------------------------------------------------------------
recipe separate_object_creation(p){                // provide different implementations for RQ and QoQ
  ...                                              // and parameterise the control program
}
...                                                // remaining recipes that are plugged in
```

Listing 3. Simplified control program (in Groove syntax)

As well as parameterising the queue semantics, it is possible to model different recursion schemes, memory models, and global interprocess synchronisations.

This semantic modularity also permits us to directly apply abstractions to Scoop-Gts, e.g. changing the queue's semantics to a bag's counting abstraction, or flattening recursion. This could prove useful for providing advanced verification approaches in the workbench.

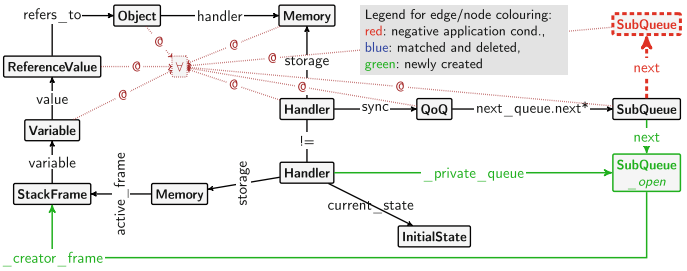

Fig. 3. Simplified QoQ rule for entering a `separate x,y,...` block, which uses ∀-quantification to atomically match arbitrarily many handlers. The rule assumes that the handlers' queues already contain some other private subqueues open

Soundness/Faithfulness. The relation of Scoop-Gts to the most prominent execution models and runtimes is depicted in Fig. 4. Due to the varying levels of detail in the formalisations of the RQ and QoQ execution models (and lack of formalisations of their implementations/runtimes), there is no universal way to prove Scoop-Gts's faithfulness to them. We also remark that Scoop-Gts currently does not support some programming mechanisms of the Eiffel language (e.g. exceptions, agents), but could be straightforwardly extended to cover them.

We were able to conduct expert interviews with the researchers proposing the execution models and the programmers implementing the Scoop compiler and runtimes, which helped to improve our confidence that Scoop-Gts faithfully

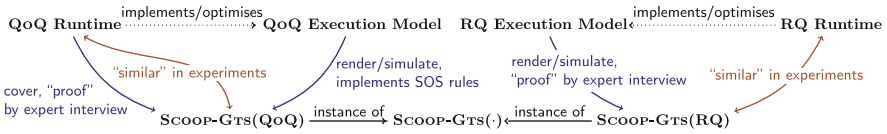

Fig. 4. Relation between Scoop-Gts, the execution models, and the runtimes

covers their behaviour. Here, Scoop-Gts's advantage of a visually accessible notation was extremely beneficial, as we were able to directly use simulations in Groove during the interviews, which were understood and accepted by the interviewees. In addition, we compared Groove simulations of the executions of Scoop programs (see the benchmarks of Sect. 5) against their actual execution behaviour in the official Scoop IDE and compiler (both the current release that implements QoQ, and an older one that implemented RQ). Again, this augmented our confidence. Furthermore, we were able to compare Scoop-Gts(QoQ) with the structural operational semantics for QoQ provided in [40]. Unfortunately, the provided semantic rules focus only on a much simplified core, preventing a rigorous bisimulation proof exploiting the algebraic characterisations of Gts. We can, however, straightforwardly implement and simulate them in our model.

To conclude, Scoop-Gts fits into the suite of existing Scoop formalisations, and is able to cover (avoiding the semantically overloaded word "simulate") both of the principal execution models.

Expressiveness. As previously discussed, Scoop-Gts is expressive enough to cover the existing RQ and QoQ semantic models of Scoop due to its modularity and the possibility to plug-in different queueing semantics. We currently plan to include other semantic formalisations of Scoop-like languages, e.g. the concurrent Eiffel proposed by [5] (similar to Scoop but differences regarding separate object calls), other actor-based object-oriented languages, and concurrency concepts like "co-boxes" [34]. Scoop-Gts is obviously Turing-complete (one can simulate a 2-counter Minsky machine by non-recursive models with one object per handler, similar to the construction in [15]). A proper formal investigation into its computational power (also that of subclasses of the model) is ongoing.

4 Toolchain for the Workbench

Our semantics workbench consists of a toolchain that bridges the gap between Scoop program code and the analysis of Scoop-Gts models in Groove. In particular, it translates source code into Scoop-Graphs, executes the appropriate analyses in Groove, and then collects and returns the results to the user.

Our toolchain is summarised in Fig. 5. Its principal component is a plug-in for the Eve IDE—a research version of the Scoop/Eiffel IDE (including the production compiler and runtime) which supports the integration of verification

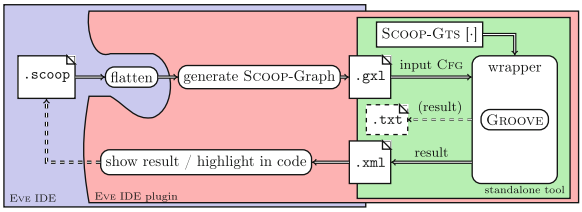

Fig. 5. Overview of our toolchain: a plugin integrated with the (research version of the) official Scoop IDE, which interfaces with a wrapper that utilises and controls Groove in the background. The wrapper can also be used as part of a standalone tool

tools [38]. For a given Scoop program, the plug-in uses the existing services of Eve to check that the code compiles, and then extracts a representation of it in which inheritance has been "flattened". From this flattened program, we generate a Scoop-Graph (encoded in the Graph eXchange Language) which corresponds very closely to the abstract syntax tree of the original program. See, for example, Fig. 6, which is generated from the code in Listing 2. Observe that between the `InitialState` and `FinalState`, the control-flow graph directly encodes the four actions of the original program: two declarations of `separate` blocks, and two commands within them. We provide a wrapper (written in Java) around the external Groove tool, which takes a generated Scoop-Graph as input, and launches a full state-space exploration in Groove with respect to Scoop-Gts(RQ) or Scoop-Gts(QoQ). The results—including statistics and detected error states—are then extracted from Groove and returned to the programmer for inspection. A standalone version of this wrapper without the Eve integration is also available and can be downloaded from [36].

Checking the Consistency of Semantics. The workbench can be used to check the consistency of program executions under RQ and QoQ with respect to various properties. These properties are encoded in Scoop-Gts as *error rules* that match on configurations if and only if they violate the properties. If they match, they generate a special `Error` node that encodes some contextual information for the toolchain to extract, and prevents the execution branch from being explored any further. Two types of error rules are supported: general, safety-related error rules for detecting problems like deadlock (whether caused by waiting for request queue locks in RQ, or waiting on cycles of queries in QoQ); but also user-specified error rules for program-specific properties (as we will use in Sect. 5). If any of these error rules are applied in a state-space exploration, this information is extracted and reported by the workbench toolchain; discrepancies between semantics exist when such rules match under only one. Figure 2 shows an actual deadlock between two handlers attempting to enter the nested separate block of Listing 2 under RQ. This configuration is matched by an error rule for deadlock (not shown), which catches the circular waiting dependencies exhibited by the edges.

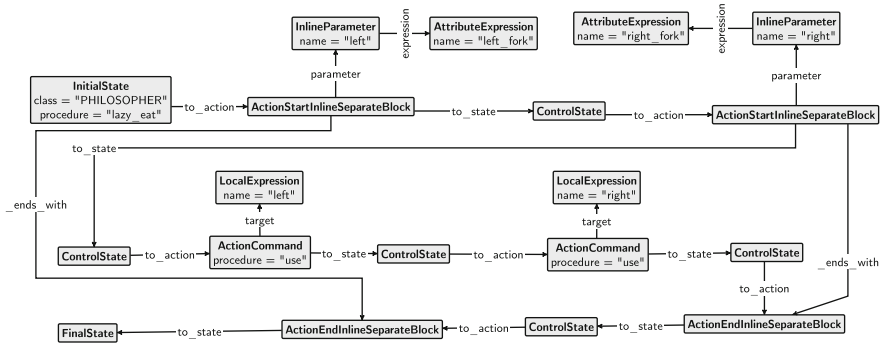

Fig. 6. Generated control-flow graph for Listing 2

5 Evaluation

To evaluate the use of our workbench for checking the consistency of semantics, we devised a representative set of benchmark programs, based on documented SCOOP examples [11] and classical synchronisation problems. We then deployed the toolchain to analyse their executions under RQ and QoQ with respect to behavioural and safety-related properties, and highlight the discrepancies uncovered by the workbench for our running example. Everything necessary to reproduce our evaluation is available at [36].

Benchmark Selection. Our aim was to devise a set of representative programs covering different, idiomatic usages of SCOOP's concurrency mechanisms. To achieve this, we based our programs on official, documented examples [11], as well as some classical synchronisation problems, in order to deploy the language mechanisms in a greater variety of usage contexts. Note that it is not (yet) our goal to analyse large software projects, but rather to compare executions of representative programs with manageable state spaces under different semantics.

We selected the following programs: dining philosophers—as presented in Sect. 2—with its two implementations for picking up forks (eagerly or lazily) which exploited the implicit locking of RQ; a third variant of dining philosophers without any commands in the separate blocks; single-element producer consumer, which uses a mixture of commands, queries, and condition synchronisation; and finally, barbershop and dining savages (based on [10]), both of which use a similar mix of features. These programs cover different usages of SCOOP's language mechanisms and are well-understood examples in concurrent programming. Note that while our compiler supports inheritance by flattening the used classes, these examples do not use inheritance; in particular, no methods from the implicitly inherited class `ANY` are used. By not translating these methods into the start graphs, we obtain considerably smaller graphs (which impacts the exploration speed, but not the sizes of the generated transition systems).

Benchmark Results. Table 1 contains metrics for the inspected examples, obtained using our GROOVE wrapper utility. The presented values correspond

Table 1. Results for the dining philosophers (DP, with the number of philosophers), producer-consumer (PC, with the number of elements), barbershop (with the number of customers), and dining savages (with the number of savages) programs; time and memory metrics are means over five runs (standard deviation in brackets)

Start graph	Semantics	Configurations	Rule applications	Start graph size (nodes/edges)	Final graph size (nodes/edges)	Time [SD] (seconds)	Memory [SD] (GB)
DP 2 eager	RQ	4219	54441	226/343	261/396	19.34 [0.25]	1.60 [0.02]
	QoQ	5644	72762	226/343	284/462	25.61 [1.03]	1.63 [0.02]
DP 2 lazy	RQ	5679	72692	221/334	288/470	24.76 [0.34]	2.05 [0.10]
	QoQ	9609	123583	221/334	256/387	42.46 [0.65]	2.22 [0.06]
DP 2 eager (no commands)	RQ	442	6010	254/395	300/473	6.08 [0.28]	0.57 [0.00]
	QoQ	443	6135	254/395	300/473	5.84 [0.24]	0.57 [0.01]
DP 2 lazy (no commands)	RQ	868	11211	250/387	325/541	9.53 [0.36]	0.66 [0.00]
	QoQ	919	11935	250/387	296/465	10.66 [0.61]	0.66 [0.01]
DP 3 eager	RQ	99198	1270216	226/343	277/422	469.94 [11.93]	11.14 [0.17]
	QoQ	199144	2532882	226/343	304/498	1393.95 [31.06]	13.88 [0.08]
DP 3 lazy	RQ	170249	2166712	221/334	319/536	1149.07 [53.59]	13.33 [0.69]
	QoQ	444686	5683419	221/334	272/413	2564.45 [55.94]	11.99 [0.07]
DP 3 eager (no commands)	RQ	3269	43967	254/395	316/499	37.31 [1.68]	1.70 [0.09]
	QoQ	3286	45152	254/395	316/499	39.32 [0.65]	1.52 [0.00]
DP 3 lazy (no commands)	RQ	10877	139216	250/387	355/604	114.26 [5.17]	3.39 [0.13]
	QoQ	11774	151526	250/387	312/491	125.25 [4.19]	3.63 [0.07]
PC 5	RQ	4085	51283	307/476	353/548	45.87 [0.74]	2.07 [0.14]
	QoQ	12366	156210	307/476	353/548	140.86 [2.83]	3.14 [0.07]
PC 20	RQ	12890	159958	307/476	398/593	148.28 [2.85]	3.78 [0.25]
	QoQ	50286	632820	307/476	398/593	618.33 [15.42]	6.67 [0.20]
Barbershop 2	RQ	38509	494491	302/466	346/538	354.17 [16.58]	6.91 [0.19]
	QoQ	54325	702611	302/466	346/538	537.21 [8.19]	7.79 [0.16]
Savages 2	RQ	35361	448576	410/631	459/716	550.92 [14.99]	7.06 [0.07]
	QoQ	79398	1008596	410/631	459/716	1299.52 [54.96]	11.33 [0.11]

to full state-space exploration. Metrics for elapsed time (wall clock time) and memory usage (computed using Java's `MemoryPoolMXBean`) are the means of five runs, while the other values are the same for each run. The experiments were carried out on an off-the-shelf notebook with an Intel Core i7-4810MQ CPU and 16 ,GB of main memory. We used Oracle Java 1.8.0_25 with the `-Xmx 14g` option together with Groove 5.5.5.

Across all instances, the start and final graph sizes are comparable. This can be explained by the fact that our implementation contains a number of simple "garbage collection" rules that remove edges and nodes that are no longer needed (e.g. the results of intermediate computations). Final graphs simply contain the control-flow graph and heap structure after the executions. Note that we do not perform real garbage collection. For example, unreachable objects are not removed; the graph size increases linearly with the number of created objects.

The number of configurations denotes the number of recipe applications. This value is of interest because it allows one to directly compare explorations under different semantics (i.e. how much more concurrency is permitted). Recall that scheduler-specific rules are wrapped inside recipes. For example, enqueueing a work item may trigger more bookkeeping rules in QoQ than in RQ. Since the corresponding logic (see Listing 3) is implemented in a recipe, we end up with just one more configuration in both cases, independently of how many individual rule applications are triggered within the recipe. Differences in the number of configurations arise from different branching at synchronisation points. For example, we can see that in most instances, QoQ generates considerably more

configurations than the RQ implementation, which suggests that SCOOP programs are "more concurrent" under QoQ.

The time and memory columns show the raw power requirements of our toolchain. Unfortunately, the state-space explosion problem is inevitable when exploring concurrent programs. The number of configurations is, unsurprisingly, particularly sensitive to programs with many handlers and only asynchronous commands (e.g. dining philosophers). Programs that also use synchronous queries (e.g. producer-consumer) scale better, since queries force synchronisation once they reach the front of the queue. We note again that our aim was to facilitate automatic analyses of representative SCOOP programs that covered the different usages of the language mechanisms, rather than optimised verification techniques for production-level software. The results suggest that for this objective, the toolchain scales well enough to be practical.

Error Rules/Discrepancies Detected. In our evaluation of the various dining philosophers implementations, we were able to detect that the lazy implementation (Listing 2) can result in deadlock under the RQ model, but not under QoQ. This was achieved by using error rules that match circular waiting dependencies. In case a deadlock occurs that is not matched by these rules, we can still detect that the execution is stuck and report a generic error, after which we manually inspect the resulting configuration. While such error rules are useful for analysing SCOOP-Graphs in general, it is also useful to define rules that match when certain program-specific properties hold. For example, if we take a look at the eager implementation of the dining philosophers (Listing 1) and its executions under RQ and QoQ, we find that the program cannot deadlock under either. This does not prove however that the implementation actually solves the dining philosophers problem under both semantics. To check this, we defined an error rule that matches if and only if two adjacent philosophers are in their separate blocks at the same time, which is impossible if forks are treated as locks (as they implicitly are under RQ). Consequently, this rule matches only under the QoQ semantics, highlighting that under the new semantics, the program is no longer a solution to the dining philosophers problem. (We remark that it can be "ported" to QoQ by replacing the commands on forks with queries, which force the waiting.) We implemented program-specific correctness rules for the other benchmark programs analogously, but did not detect any further discrepancies.

6 Related Work

We briefly describe some related work closest to the overarching themes of our paper: frameworks for semantic analyses, GTS models for concurrent asynchronous programs, and verification techniques for SCOOP.

Frameworks for Semantic Analysis. The closest approach in spirit to ours is the work on $\mathbb{K}$ [21,33]. It consists of the $\mathbb{K}$ concurrent rewrite abstract machine and the $\mathbb{K}$ technique. One can think of $\mathbb{K}$ as domain specific language for implementing programming languages with a special focus on semantics, which

was recently successfully applied to give elaborate semantics to Java [4] and JavaScript [30]. Both $\mathbb{K}$ and our workbench have the same user group (programming language designers and researchers) and focus on formalising semantics and analysing programs based on this definition. We both have "modularity" as a principal goal in our agendas, but in a contrasting sense: our modularity targets a semantic plug-in mechanism for parameterising different model components, whereas $\mathbb{K}$ focuses on modularity with respect to language feature reuse. In contrast to our approach, $\mathbb{K}$ targets the whole language toolchain—including the possibility to define a language and automatically generate parsers and a runtime simulation for testing the formalisation. Based on *Maude*'s formal power of conditional rewriting logic, $\mathbb{K}$ also offers axiomatic models for formal reasoning on programs and the possibility to also define complex static semantic features, e.g. advanced typing and meta-programming. Despite having similar formal underlying theoretical power ($\mathbb{K}$'s rewriting is similar to "jungle rewriting" graph grammars [35]), Scoop-Gts models make the graph-like interdependencies between concurrently running threads or handlers a first-class element of the model. This is an advantage for analyses of concurrent asynchronous programs, as many concurrency properties can be straightforwardly reduced to graph properties (e.g. deadlocks as wait-cycles). Our explicit Gts model also allows us to compare program executions under different semantics, which is not a targeted feature of $\mathbb{K}$. We also conjecture that our diagrammatic notations are easier for software engineers to grasp than purely algebraic and axiomatic formalisations.

GTS Models for Concurrent Asynchronous Programs. Formalising and analysing concurrent object-oriented programs with Gts-based models is an emerging trend in software specification and analysis, especially for approaches rooted in practice. See [31] for a good overview discussion—based on a lot of personal experience—on the general appropriateness of Gts for this task. In recent decades, conditional rewriting logic has become a reference formalism for concurrency models in general; we refer to [22] and its recent update [23] for details. Despite having a comparable expressive power, our approach's original decision for Gts and for Groove as our state-space exploration tool led us to an easily accessible, generic, and parameterisable semantic model and tools that work in acceptable time on our representative Scoop examples. Closest to our Scoop-Gts model is the Qdas model presented in [15], which represents an asynchronous, concurrent waiting queue based model with *global memory* as Gts, for verifying programs written in Grand Central Dispatch [14]. Despite the formal work, there is currently no direct compiler to Gts yet. The Creol model of [20] focuses on asynchronous concurrent models but without more advanced remote calls via queues as needed for Scoop. Analysis of the model can be done via an implementation in *Maude* [19]. Existing Gts-based models for Java only translate the code to a typed graph similar to the control-flow sub-graph of Scoop-Gts [8,32]. A different approach is taken by [12], which abstracts a Gts-based model for concurrent OO systems [13] to a finite state model that can be verified using the SPIN model checker. Groove itself was already used for verifying concurrent distributed algorithms on an abstract Gts level [16], but not

on an execution model level as in our approach. However, despite the intention to apply generic frameworks for the specification, analysis, and verification of object-oriented concurrent programs, e.g. in [9,41], there are no publicly available tools implementing this long-term goal that are powerful enough for SCOOP.

SCOOP Analysis/Verification. Various analyses for SCOOP programs have been proposed, including: using a SCOOP virtual machine for checking temporal properties [29]; checking Coffman's deadlock conditions using an abstract semantics [7]; and statically checking code annotated with locking orders for the absence of deadlock [39]. In contrast to our work, these approaches are tied to particular (and now obsolete) execution models, and do not operate on (unannotated) source code.

The complexity of other semantic models of SCOOP led to scalability issues when attempting to leverage existing analysis and verification tools. In [6], SCOOP programs were hand-translated to models in the process algebra CSP to perform, e.g. deadlock analysis; but the leading CSP tools at the time could not cope with these models and a new tool was purpose-built (but no longer available/maintained today). In a recent deadlock detection benchmark on the RQ execution model formalised in *Maude* [26], the tool was not able to give verification results in reasonable time (i.e. less than one day) even for simple programs like dining philosophers[1]; our benchmarks compare quite favourably to this. Note that since our work focuses more on semantic modelling and comparisons than it does on the underlying model checking algorithms, we did not yet evaluate the generic bounded model checking algorithms for temporal logic properties implemented in GROOVE and accessible for SCOOP-GTS models.

7 Conclusion and Future Work

We proposed and constructed a semantic workbench for a concurrent asynchronous programming language via the following, general work flow: (i) derive a GTS-based semantic model from existing semi-formal documentation of execution models; (ii) continuously compare the model by simulation runs against the actual implementations; (iii) exploit semantic paramaterisation to derive a versatile model; (iv) if possible, conduct expert interviews to ascertain the model's faithfulness; (v) apply existing generic model checking techniques for GTS to implement analyses against the different execution models; (vi) implement different analyses on top of this model. This workflow resulted in the formalisation SCOOP-GTS, which covered the two principal execution models of SCOOP, and allowed us to formally, automatically, and practically compare the executions of programs with respect to both. With the conducted expert interviews, and the results of applying our model to check the consistency of the semantics across a small but representative collection of SCOOP programs in reasonable time, we were reassured of our choice of GTS as an underlying formalism: theoretically

[1] From personal communication with the researchers behind this benchmark.

sound, yet diagrammatically accessible for software engineers, and able to scale to the sizes of programs we need for semantic comparisons.

We are currently working on extending SCOOP-GTS to cover some more advanced and esoteric features of SCOOP (including distributed exception handling) and to enlarge the benchmark set, with the eventual aim of producing a conformance test suite for SCOOP-like languages. As noted in [42], the shape of the rules and the control programs have a big influence on the running times of GROOVE. We are currently working on refactoring SCOOP-GTS for better performance (relative to benchmarking on the conformance test suite).

A more general line of research focuses on the shape of the SCOOP-Graphs contained in the reachable state space of SCOOP-GTS. Insights here would help us to devise better abstraction techniques (along the lines of [3]) with which we could implement better verification algorithms, and visualise the influence of different semantic parameters on SCOOP-Graphs. Generalising SCOOP-GTS to cover other actor-based concurrency languages would also extend this result towards differences between the semantics of programming language families expressed as SCOOP-Graph properties.

Acknowledgements. We thank our interviewees from the SCOOP development and research team for the many helpful and insightful discussions. We are also deeply grateful for the work of the GROOVE developers that we leverage in this paper, and especially for their GROOVE-y feedback and support. The underlying research was partially funded by ERC Grant CME #291389.

References

1. Agha, G.: ACTORS: A Model of Concurrent Computation in Distributed Systems. MIT Press, Cambridge (1986)
2. Armstrong, J., Virding, R., Williams, M.: Concurrent Programming in ERLANG, 2nd edn. Prentice Hall, Englewood Cliffs (1996)
3. Backes, P., Reineke, J.: Analysis of infinite-state graph transformation systems by cluster abstraction. In: D'Souza, D., Lal, A., Larsen, K.G. (eds.) VMCAI 2015. LNCS, vol. 8931, pp. 135–152. Springer, Heidelberg (2015)
4. Bogdanas, D., Rosu, G.: K-Java: A complete semantics of Java. In: Proceedings of POPL 2015, pp. 445–456. ACM (2015)
5. Brooke, P.J., Paige, R.F.: Cameo: an alternative model of concurrency for Eiffel. Formal Aspects Comput. **21**(4), 363–391 (2009)
6. Brooke, P.J., Paige, R.F., Jacob, J.L.: A CSP model of Eiffel's SCOOP. Formal Aspects Comput. **19**(4), 487–512 (2007)
7. Caltais, G., Meyer, B.: Coffman deadlocks in SCOOP. In: Proceedings of NWPT 2014 (2014). http://arxiv.org/abs/1409.7514
8. Corradini, A., Dotti, F.L., Foss, L., Ribeiro, L.: Translating Java code to graph transformation systems. In: Ehrig, H., Engels, G., Parisi-Presicce, F., Rozenberg, G. (eds.) ICGT 2004. LNCS, vol. 3256, pp. 383–398. Springer, Heidelberg (2004)
9. Dotti, F.L., Duarte, L.M., Foss, L., Ribeiro, L., Russi, D., dos Santos, O.M.: An environment for the development of concurrent object-based applications. In: Proceedings of GraBaTs 2004. ENTCS, vol. 127, pp. 3–13. Elsevier (2005)

10. Downey, A.B.: The Little Book of Semaphores. http://greenteapress.com/semaphores/. Accessed Jan 2016
11. Eiffel Documentation: Concurrent Eiffel with SCOOP. https://docs.eiffel.com/book/solutions/concurrent-eiffel-scoop. Accessed Oct 2015
12. Ferreira, A.P.L., Foss, L., Ribeiro, L.: Formal verification of object-oriented graph grammars specifications. In: Proceedings of GT-VC 2006. ENTCS, vol. 175, pp. 101–114. Elsevier (2007)
13. Ferreira, A.P.L., Ribeiro, L.: A graph-based semantics for object-oriented programming constructs. In: Proceedings of CTCS 2004. ENTCS, vol. 122, pp. 89–104. Elsevier (2005)
14. Grand Central Dispatch (GCD) Reference. https://developer.apple.com/library/mac/documentation/Performance/Reference/GCD_libdispatch_Ref/index.html. Accessed Oct 2015
15. Geeraerts, G., Heußner, A., Raskin, J.: On the verification of concurrent, asynchronous programs with waiting queues. ACM Trans. Embed. Comput. Syst. **14**(3), 58 (2015)
16. Ghamarian, A.H., de Mol, M., Rensink, A., Zambon, E., Zimakova, M.: Modelling and analysis using GROOVE. Int. J. Softw. Tools Technol. Transf. **14**(1), 15–40 (2012)
17. Groove (project web page). http://groove.cs.utwente.nl/. Accessed October 2015
18. Heußner, A., Poskitt, C.M., Corrodi, C., Morandi, B.: Towards practical graph-based verification for an object-oriented concurrency model. In: Proceedings of GaM 2015. EPTCS, vol. 181, pp. 32–47 (2015)
19. Johnsen, E.B., Owe, O., Axelsen, E.W.: A run-time environment for concurrent objects with asynchronous method calls. In: Proceedings of WRLA 2004. ENTCS, vol. 117, pp. 375–392. Elsevier(2005)
20. Johnsen, E.B., Owe, O., Yu, I.C.: Creol: a type-safe object-oriented model for distributed concurrent systems. Theor. Comput. Sci. **365**(1–2), 23–66 (2006)
21. Lucanu, D., Şerbănuţă, T.F., Roşu, G.: $\mathbb{K}$ framework distilled. In: Durán, F. (ed.) WRLA 2012. LNCS, vol. 7571, pp. 31–53. Springer, Heidelberg (2012)
22. Meseguer, J.: Conditioned rewriting logic as a united model of concurrency. Theor. Comput. Sci. **96**(1), 73–155 (1992)
23. Meseguer, J.: Twenty years of rewriting logic. J. Logic Algebraic Program. **81**(7–8), 721–781 (2012)
24. Meyer, B.: Systematic concurrent object-oriented programming. Commun. ACM (CACM) **36**(9), 56–80 (1993)
25. Meyer, B.: Object-Oriented Software Construction, 2nd edn. Prentice Hall, Upper Saddle River (1997)
26. Morandi, B., Schill, M., Nanz, S., Meyer, B.: Prototyping a concurrency model. In: Proceedings of ACSD 2013, pp. 170–179. IEEE (2013)
27. Morandi, B., Nanz, S., Meyer, B.: Safe and efficient data sharing for message-passing concurrency. In: Kühn, E., Pugliese, R. (eds.) COORDINATION 2014. LNCS, vol. 8459, pp. 99–114. Springer, Heidelberg (2014)
28. Nienaltowski, P.: Practical framework for contract-based concurrent object-oriented programming. Doctoral dissertation, ETH Zürich (2007)
29. Ostroff, J.S., Torshizi, F.A., Huang, H.F., Schoeller, B.: Beyond contracts for concurrency. Formal Aspects Comput. **21**(4), 319–346 (2009)
30. Park, D., Ştefănescu, A., Roşu, G.: KJS: a complete formal semantics of JavaScript. In: Proceedings of PLDI 2015, pp. 346–356. ACM (2015)

31. Rensink, A.: The edge of graph transformation — graphs for behavioural specification. In: Engels, G., Lewerentz, C., Schäfer, W., Schürr, A., Westfechtel, B. (eds.) Nagl Festschrift. LNCS, vol. 5765, pp. 6–32. Springer, Heidelberg (2010)
32. Rensink, A., Zambon, E.: A type graph model for Java programs. In: Lee, D., Lopes, A., Poetzsch-Heffter, A. (eds.) FMOODS 2009. LNCS, vol. 5522, pp. 237–242. Springer, Heidelberg (2009)
33. Rosu, G., Serbanuta, T.: An overview of the K semantic framework. J. Logic Algebraic Program. **79**(6), 397–434 (2010)
34. Schäfer, J., Poetzsch-Heffter, A.: JCoBox: generalizing active objects to concurrent components. In: D'Hondt, T. (ed.) ECOOP 2010. LNCS, vol. 6183, pp. 275–299. Springer, Heidelberg (2010)
35. Şerbănuţă, T.F., Roşu, G.: A truly concurrent semantics for the $\mathbb{K}$ framework based on graph transformations. In: Ehrig, H., Engels, G., Kreowski, H.-J., Rozenberg, G. (eds.) ICGT 2012. LNCS, vol. 7562, pp. 294–310. Springer, Heidelberg (2012)
36. Supplementary material. http://www.swt-bamberg.de/fase2016_supp/
37. Torshizi, F.A., Ostroff, J.S., Paige, R.F., Chechik, M.: The SCOOP concurrency model in Java-like languages. In: Proceedings of CpPA 2009. Concurrent Systems Engineering Series, vol. 67, pp. 7–27. IOS Press (2009)
38. Tschannen, J., Furia, C.A., Nordio, M., Meyer, B.: Usable verification of object-oriented programs by combining static and dynamic techniques. In: Barthe, G., Pardo, A., Schneider, G. (eds.) SEFM 2011. LNCS, vol. 7041, pp. 382–398. Springer, Heidelberg (2011)
39. West, S., Nanz, S., Meyer, B.: A modular scheme for deadlock prevention in an object-oriented programming model. In: Dong, J.S., Zhu, H. (eds.) ICFEM 2010. LNCS, vol. 6447, pp. 597–612. Springer, Heidelberg (2010)
40. West, S., Nanz, S., Meyer, B.: Efficient and reasonable object-oriented concurrency. In: Proceedings of ESEC/FSE 2015, pp. 734–744. ACM (2015)
41. Zambon, E., Rensink, A.: Using graph transformations and graph abstractions for software verification. In: Proceedings of ICGT-DS 2010. ECEASST, vol. 38 (2011)
42. Zambon, E., Rensink, A.: Solving the N-Queens problem with GROOVE - towards a compendium of best practices. In: Proceedings of GT-VMT 2014. ECEASST, vol. 67 (2014)