International Journal of Information System Modeling and Design
Volume 7 • Issue 3 • July-September 2016

First Steps Towards a Wise Development Environment for Behavioral Models

David Harel, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
Guy Katz, Computer Science Department, Stanford University, Stanford, CA, USA
Rami Marelly, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel
Assaf Marron, Department of Computer Science and Applied Mathematics, Weizmann Institute of Science, Rehovot, Israel

ABSTRACT

The authors present an initial wise development framework: a development environment that proactively and interactively assists the software engineer in modeling complex reactive systems. Their framework repeatedly analyzes models of the system under development at various levels of abstraction, and then reasons about these models in order to detect possible errors, to derive emergent properties of interest, and to assist in system testing and debugging. Upon request, the environment can instrument the system model in order to monitor or test the execution for certain behaviors, or even augment it in order to repair or avoid detected behavior that is undesired. The direction and prioritization of the analysis and related tasks is based on the relevance of the observed properties and the expected impact of actions to be taken, and is performed by specialized automated and human-assisted techniques that have been incorporated into the framework. The authors' development environment is an initial step in the direction of their recent Wise Computing vision, which calls for turning the computer (namely, the development environment) into an equal member of the development team: knowledgeable, independent, concerned and proactively involved in the development process. They have implemented their tool within the context of behavioral programming (BP) – a scenario-based modeling approach, in which components are aligned with how humans often describe desired system behavior. The authors' work thus further enhances the naturalness and incrementality of developing in BP.

KEYWORDS

Behavioral Models, Interactive Development, Proactive Analysis, Reactive Models, Wise Computing

1. INTRODUCTION

The development of large reactive software systems is an expensive and error-prone undertaking. Deliverables will often fail, resulting in unintended software behavior, exceeded budgets and breached time schedules. One of the key reasons for this difficulty is the growing complexity of many kinds of reactive systems, which increasingly prevents the human mind from managing a comprehensive picture of all their relevant elements and behaviors. Moreover, of course, the state-explosion problem typically prevents us from exhaustively analyzing all possible software behaviors. While major advances in modeling tools and methodologies have greatly improved our ability to develop reactive systems by allowing us to reason on abstract models thereof, specific solutions are quickly reaching their limits, and resolving the great difficulties in developing reliable reactive systems remains a major, and critical, moving target.

DOI: 10.4018/IJISMD.2016070101

Over the years it has been proposed, in various contexts, e.g., (Cerf, 2014; Harel, Katz, Marelly, & Marron, 2015; Reubenstein & Waters, 1991; Rich & Waters, 1988), that a possible strategy for mitigating these difficulties could lay in changing the role of the computer in the development process. Instead of having the computer serve as a tool, used only to analyze or check specific aspects of the code as instructed by the developer, one could seek to actually transform it into a member of the development team — a proactive participant, analyzing the entire system and making informed observations and suggestions. This way, the idea goes, the computer's superior capabilities of handling large amounts of code could be manifested. Combined with human insight and understanding of the system's goals, this synergy could produce more reliable and error-free systems.

In this paper we follow this spirit, and present a methodology and an interactive framework for the modeling and development of complex reactive systems, in which the computer plays a proactive role. Following the terminology of (Harel, Katz, Marelly, & Marron, 2015), and constituting a very modest initial effort along the lines of the Wise Computing vision outlined there, we term this framework a wise framework. Intuitively, a truly wise framework should provide the developer with an interactive companion for all phases of system development, "understand" the system, draw attention to potential errors and suggest improvements and generalizations; and this should be done via two-way communication with the developer, which will be very high-level, using natural (perhaps natural-language-based) interfaces. The framework presented here is but a first step in that direction, and focuses solely on providing an interactive development assistant capable of discovering interesting properties and drawing attention to potential bugs; still, it can already handle non-trivial programs, as we later demonstrate through a case-study.

Various parts of this approach have been implemented by a variety of researchers in other forms, as described in the Related Work section. A main novel aspect of our approach, however, is in the coupling of the notion of a proactive and interactive framework with a modeling language called behavioral programming (Harel, Marron, & Weiss, 2012) — a scenario-based language, in which systems are modeled as sets of independent scenarios that are interleaved at runtime. This formalism makes it possible for our interactive development framework to repeatedly and quickly construct abstract executable models of the program, and then analyze them in order to reach meaningful conclusions. It is now widely accepted that a key aspect in the viability of analysis tools and environments is that they are sufficiently lightweight to be integrated into the developer's workflow without significantly slowing it down (Cristiano et al., 2015; Sadowski, Gogh, Jaspan, Söederberg, & Winter, 2015). We attempt to achieve this by leveraging scenario-based modeling. As demonstrated in later sections, the proactiveness of our approach and its tight integration into the development cycle can lead to early detection of bugs during development, when they are still relatively easy and cheap to fix.

The rest of this paper is organized as follows. In the second section we introduce scenario-based programming — the modeling formalism on top of which our approach is implemented, and also discuss some analysis techniques for scenario-based programs that are used in subsequent sections. In the third section we introduce our development framework by means of a simple example. In the fourth section we discuss the various components of the framework in more detail, and in the fifth we describe a case-study that we conducted. We then provide a discussion of related work and a conclusion.

2. SCENARIO-BASED MODELING

Behavioral programming (BP) (Harel, Marron, & Weiss, 2012) is a modeling approach aimed at designing and incrementally developing reactive systems. BP emerged from the live sequence charts (LSCs) formalism (Damm & Harel, 2001; Harel & Marelly, 2003), and, like LSCs, its basic modeling objects are scenarios. A behavioral model consists of independent scenario objects, each encoding a single desired or undesired behavior of the system. These behavioral models are executable: when run, the behaviors encoded by their constituent objects are all interwoven together, in a way that yields cohesive system behavior.

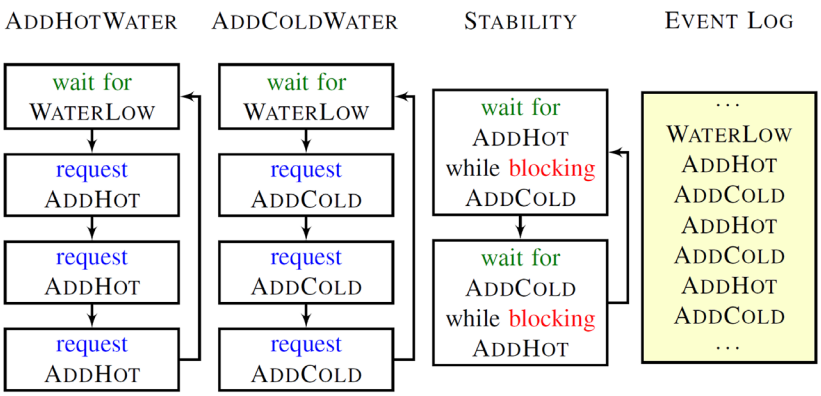
More specifically, an execution of a behavioral model is a sequence of points in which all scenario objects synchronize and declare events that they want to be considered for triggering (called requested events), events that they do not actively request but merely "listen out" for (waited-for events), and events whose triggering they forbid (blocked events). During execution, an event that is requested by some scenario and not blocked by any scenario is selected for triggering, and every scenario object that requested or waited for the event can update its internal state. Figure 1 (adapted from (Harel, Katz, Marron, & Weiss, 2014)) demonstrates a simple behavioral model. The formal definitions of behavioral modeling appear in the Formal Definitions section.

The motivation for using behavioral modeling is its strict and simple mechanism for inter-object communication. In particular, BP's request/wait-for/block interface facilitates incremental, non-intrusive development, and the resulting models often have scenario objects that are aligned with the requirements (Harel, Marron, & Weiss, 2012). This is lent additional support by studies that indicate that BP is natural, in the sense that it is easy to learn and fosters abstract programming (Gordon, Marron, & Meerbaum-Salant, 2012; Alexandron, Armoni, Gordon, & Harel, 2014).

In practice, behavioral modeling is usually performed using various high level languages, such as Java, C++, Erlang, Javascript and, of course, LSCs, on which BP is based and from which it grew (see the BP website at http://www.b-prog.org/). Models written in these languages are fully executable, and are also referred to as behavioral programs. There, each scenario object is typically implemented as a separate thread, and inter-thread communication is restricted to event requesting, waiting-for and blocking — thus preserving the semantics of behavioral modeling. Technically, this is performed by having the scenario threads invoke a special synchronization method called BSYNC, and pass to it their requested/waited-for/blocked events. Once every scenario has synchronized, an event selection mechanism triggers one event that is requested and not blocked, and notifies the relevant scenarios.

For actual programming purposes it is often helpful to allow threads to also perform local actions — e.g., read from a file or turn on a light bulb. These actions are not included in the underlying behavioral model (i.e., they are abstracted away). The wise framework that we present here is designed to accompany the development of such behavioral programs, and is built on top the BPC package (Harel & Katz, 2014) for behavioral modeling in C++. This package also supports the distributed execution of behavioral programs (Harel, Kantor, et al., 2015).

Figure 1. The incremental modeling of a system for controlling the water level in a tank with hot and cold water sources. Each scenario object is given as a transition system, where the nodes represent synchronization points. The scenario object AddHotWater repeatedly waits for WaterLow events and requests three times the event AddHot. Scenario object AddColdWater performs a similar action with the event AddCold, capturing a separate requirement, which was introduced when adding three water quantities for every sensor reading proved to be insufficient. When a model with objects AddHotWater and AddColdWater is executed, the three AddHot events and three AddCold events may be triggered in any order. When a new requirement is introduced, to the effect that water temperature be kept stable, the scenario object Stability is added, enforcing the interleaving of AddHot and AddCold events by using event blocking. The execution trace of the resulting model is depicted in the event log.



2.1. Formal Definitions

For completeness, we recap here briefly the formal definitions of behavioral modeling. Following the definitions in (Katz, 2013), a scenario object O over event set E is a tuple $O = \langle Q, \delta, q_0, R, B \rangle$, where Q is a set of states, q_0 is the initial state, $R : Q \to 2^E$ and $B : Q \to 2^E$ map states to the sets of events requested and blocked at these states (respectively), and $\delta : Q \times E \to 2^Q$ is a transition function.

Scenario objects can be composed, in the following manner. For objects $O^1 = \langle Q^1, \delta^1, q_0^1, R^1, B^1 \rangle$ and $O^2 = \langle Q^2, \delta^2, q_0^2, R^2, B^2 \rangle$ over a common event set E, the composite scenario object $O^1 \parallel O^2$ is defined by $O^1 \parallel O^2 = \langle Q^1 \times Q^2, \delta, \langle q_0^1, q_0^2 \rangle, R^1 \cup R^2, B^1 \cup B^2 \rangle$, where $\langle \tilde{q}^1, \tilde{q}^2 \rangle \in \delta(\langle q^1, q^2 \rangle, e)$ if and only if $\tilde{q}^1 \in \delta^1(q^1, e)$ and $\tilde{q}^2 \in \delta^2(q^2, e)$. The union of the labeling functions is defined in the natural way; e.g. $e \in (R^1 \cup R^2)(\langle q^1, q^2 \rangle)$ if and only if:

$$e \in R^1(q^1) \cup R^2(q^2)$$

A behavioral model M is simply a collection of scenario objects $O^1, O^2, \ldots, O^n$, and the executions of M are the executions of the composite object $O = O^1 \parallel O^2 \parallel \ldots \parallel O^n$. Each such execution starts from the initial state of O, and in each state q along the run an enabled event is chosen for triggering, if one exists (i.e., an event $e \in R(q) - B(q)$). Then, the execution moves to state $\tilde{q} \in \delta(q, e)$, and so on.

2.2. Analyzing Behavioral Models

Earlier we explained the motivation behind behavioral modeling, from a developer's point of view. However, it turns out that due to its simple synchronization mechanism, behavioral modeling lends itself naturally also to formal analysis. We briefly recap a few such analysis methods, which are used by our proposed wise development framework.

2.2.1. Model Checking Behavioral Models

In (Harel, Lampert, Marron, & Weiss, 2011; Harel, Kantor, & Katz, 2013) a technique is presented, by which the underlying transition systems of individual scenario objects are extracted from high-level behavioral code and are then used in order to model check the behavioral model. The extraction of these transition systems is performed by running individual scenario objects in sandboxes and passing to them events, just as if they were triggered by the event selection mechanism, in a way that allows one to methodically explore their state spaces (Harel, Kantor, & Katz, 2013). Model checking is then performed by adding special behavioral objects to the model that mark undesired behavior, and then traversing the states of the composite model to see if a violation can occur.

In order to mitigate the state-explosion problem and allow the model checking of larger behavioral models, one can replace behavioral objects or sets thereof with abstract behavioral objects (Katz, 2013). Intuitively, within a behavioral object, a set of states $q_1, q_2, \ldots, q_\ell$ can be abstracted away using a single state q, such that:

$$R(q) = \bigcup_{i=1}^{\ell} R(q_i) \text{ and } B(q) = \bigcap_{i=1}^{\ell} B(q_i)$$

The transition relation is then adjusted so that any transition between states s and t in the original model becomes a transition between s' and t' in the abstract object, where s' and t' are the abstract states representing s and t, respectively.

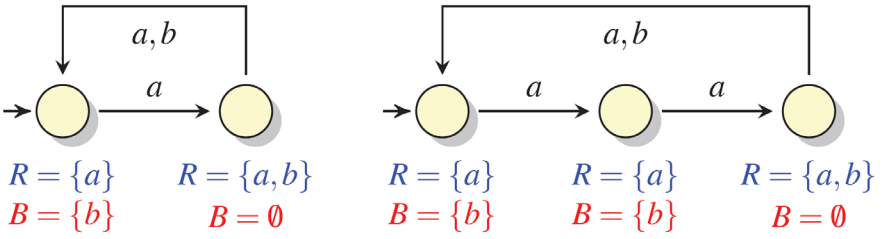
In (Katz, 2013) it is shown that, because abstract states block fewer events and request more events than their concrete counterparts, this sort of abstraction yields a behavioral model that is more permissive than the original one (i.e., it is an over-approximation). Typically, due to the reduction in the number of states, this abstract model is also significantly smaller than its original counterpart. Model checking and program repair operations can then be performed on the abstract model (sometimes combined with local refinement steps), and the results are guaranteed to hold for the original system, thus enabling better coping with state-explosion. In later sections we make extensive use of this abstraction technique.

2.2.2. Compositional Verification of Behavioral Models

A useful property of behavioral modeling is that despite the small number of simple-looking concurrency idioms that it provides (i.e., the requesting, waiting-for and blocking idioms) it provides significant succinctness advantages. Specifically, it allows specifying behavioral objects that are exponentially smaller than what is possible using non-concurrent modeling formalisms, and even when compared to formalisms in which any of the requesting, waiting-for and blocking idioms are omitted (Harel, Katz, Lampert, Marron, & Weiss, 2015). An example appears in Figure 2.

The succinctness afforded by behavioral modeling can sometimes be leveraged for efficient compositional verification (Harel, Kantor, Katz, et al., 2013; Katz, Barrett, & Harel, 2015). For example, suppose that we wish to verify that in the model depicted in Figure 2 event b can only be triggered every 6 steps. Direct model checking would entail exploring the 6 composite states of the system, but compositional verification would entail exploring the states of each object separately (a total of 5 states), characterizing the properties of each individual object, and then using an SMT solver to derive global correctness from these individual properties. More specifically, the individual object properties in this example can be formulated as $triggered\left(b,i\right) \Rightarrow i \equiv 0 \left(\mathrm{mod}\, 2\right)$ for the object on the left and $triggered\left(b,i\right) \Rightarrow i \equiv 0 \left(\mathrm{mod}\, 3\right)$ for the object on the right, where $triggered\left(b,i\right)$

Figure 2. This behavioral model has two scenario objects, each depicted as a transition system. Every state corresponds to a synchronization point, and is labeled with its requested and blocked events, whereas the waited-for events are encoded on the transitions. The scenario on the left counts modulo two: at odd steps it requests event a and blocks event b, and at even steps it requests both events. The scenario on the right is similar, but counts modulo three, and only requests both events every third step. Together, these two objets count modulo 6, producing the language $\left(a^5\left(a+b\right)\right)^\omega$. In (Harel, Katz, Lampert, et al., 2015) it is shown that modeling this system in a non-concurrent formalism, or even in one that is devoid of the blocking idiom, requires 6 ($=3\cdot 2$) states instead of 5 ($=3+2$). When generalized to the language $\left(a^n\left(a+b\right)\right)^\omega$ for an arbitrarily large n, this gap between the sum and the product of the number of states in the constituent scenarios is exponential in n. For a more thorough discussion of the succinctness afforded by behavioral modeling, see (Harel, Katz, Lampert, et al., 2015).



means that the i 'th event triggered was b. These properties can be verified on the individual objects. Using these object properties, an SMT solver can quickly deduce the desired property, $triggered\left(b,i\right)\Rightarrow i\equiv 0\left(\mathrm{mod}\,6\right)$, circumventing the need to explore the composite states of the model.

When the above example is generalized to $\left(a^{n}\left(a+b\right)\right)^{\omega}$ for a large n, the gap in the number of explored states between the direct approach (roughly the product of the number of individual object states) and the compositional approach (roughly the sum thereof) is exponential in n (Harel, Kantor, Katz, et al., 2013).

The key observation, which we leverage repeatedly in the following sections, is that in scenario-based modeling it is often simple, and computationally cheap, to analyze many small scenario objects — and then use this information to reason about the model as a whole.

3. DEVELOPMENT IN A WISE FRAMEWORK: AN EXAMPLE

In this section we attempt to convey to the reader, intuitively, the sense of working in a wise development framework from a developer's point of view. Thus, we focus almost exclusively on the user experience, and defer more details about the inner workings of the framework itself to the next section.

We demonstrate the framework's operation through the incremental modeling of a small, illustrative system. Suppose we are developing behavioral code for a safe that has three levers and an "open door" button. The specification given to us indicates that in order to open the door, a user needs to correctly configure the three levers and then click the button. Clicking the button when the levers are not correctly configured should not open the door. We refer to the three levers as levers A,B and C; and each lever has three possible positions, denoted as one, two and three. We denote the configuration of the levers as a tuple: for instance, configuration $\left\langle 1,3,2\right\rangle$ indicates that lever A is in position one, lever B is in position three, and lever C is in position two. The initial configuration is $\left\langle 1,1,1\right\rangle$, and the correct configuration for opening the door is $\left\langle 2,3,2\right\rangle$. The user can request the triggering of events of the form SetXToY, indicating that lever X is set to position Y, and also of ClickButton events. The system may request an OpenDoor event, as well as any internal event needed for the implementation.

We now describe the incremental modeling of this system in BPC, accompanied by the wise framework. We start by modeling the three levers. This is done by creating, for each lever, a scenario object that waits for events signaling that the position of that lever has changed, and storing the current position. The code appears and is explained in Figure 3.

After modeling the three lever objects, we get the first input from the wise development framework:

Warning: Objects LeverA, LeverB and LeverC constitute a ternary shared array. However, they are not used. Consider removing them.

We should emphasize that the wise development framework is oblivious to the specifics of our program, i.e., it has no concept of levers. It did, however, recognize a pattern in our system model: that the three lever objects actually operate like a "shared array". Here, the term shared array means that other objects can "write" to it (i.e., by requesting SetXToY events), or "read" from it (by requesting LeverXInY events). This is an interesting insight about the implementation, which we did not even have in mind, but which the development framework will utilize later on. As for the comment that the levers are currently unused, this makes sense — as we have not yet written any additional code.

Next, we add a scenario that allows the user, through a simple interface, to request the triggering of SetXToY events, and also the ClickButton event (code omitted). When we recompile the code,

Figure 3. BPC code for a scenario object called LeverX, representing the behavior of a single lever X (X represents A, B or C). Line 16 contains the BSYNC synchronization call, where the object synchronizes with all other objects and declares its requested, waited-for and blocked events. The lever object never requests any events, and continuously waits for events signifying that the lever has changed its physical position — events SetXToOne, SetXToTwo, and SetXToThree. When one of these is triggered, line 16 returns, and the object updates its internal state in line 17. Note also events LeverXInOne, LeverXInTwo and LeverXInThree, which represent other scenarios querying the physical position of lever X. The lever object constantly blocks those events that correspond to all "wrong" physical positions. Thus, if another object requests all three events, then only one event — the one corresponding to the actual lever's position — will be triggered. An example appears in Figure 4.

```
1  Event position = SetXToOne;
2  while ( true ) {
3   set<Event> requested = {};
4   set<Event> waitedFor = { SetXToOne, SetXToTwo, SetXToThree };
5   set<Event> blocked;
6
7   switch( position ) {
8   case SetXToOne:
9     blocked = { LeverXInTwo, LeverXInThree };
10  case SetXToTwo:
11    blocked = { LeverXInOne, LeverXInThree };
12  case SetXToThree:
13    blocked = { LeverXInOne, LeverXInTwo };
14  }
15
16  BSYNC( requested, waitedFor, blocked );
17  position = lastEvent();
18 }
```

the development framework prompts us that now the shared array is written to but is never read from, and can still be removed. Then, we add the ButtonPressed scenario (Figure 4) that handles the pressing of the button — it queries the lever configuration, and if it is $\langle 2,3,2 \rangle$ it requests an OpenDoor event.

However, as the caption explains, the code in Figure 4 is actually erroneous: we copied and pasted the code checking lever B but did not correctly modify it to check lever C. The wise development framework now produces the following message:

Warning: Scenario ButtonPressed has an unreachable synchronization point in line 20. Suggesting an optimization. Also, the state of LeverC is never read.

This message immediately points us to the error in the model, giving us enough information to quickly realize what has happened. The optimization proposed by the framework (not shown), in which the unreachable state is removed, is actually a graphical representation using the Goal visualization tool (Tsay, Chen, Tsai, Wu, & Chan, 2007).

We stress that the realization that line 20 is unreachable is not trivial, as it is not a property that is local to the ButtonPressed object. In particular, it cannot be deduced by inspecting the ButtonPressed object in isolation, and thus it is very different from deducing, say, that in $if\big(false\big)(foo())$ the function $foo()$ can never be called. Rather, this property stems from the joint behavior of ButtonPressed and LeverB, where ButtonPressed expects LeverB to be in two different states simultaneously, which cannot occur.

And so, we correct the error in line 16 of ButtonPressed. Now the warnings from the development framework disappear, and instead we receive the following information:

Information: Event OpenDoor appears to only be triggered after event LeverCInTwo.

Figure 4. The ButtonPressed scenario, which waits for a ClickButton event, queries the configuration of the three levers (lines 8, 12 and 16), and if they are correctly set requests an OpenDoor event (line 20). Querying the position of lever X is performed by simultaneously requesting events LeverXInOne, LeverXInTwo and LeverXInThree. Only the "correct" event, i.e. the event that corresponds to lever X's current position, will be triggered, because the other two events will be blocked by LeverX's scenario object. Observe that this scenario has a bug: in line 16, instead of checking whether lever C is in position two, we mistakenly check if lever B is in position two. When this line in the code (line 16) is reached we already know that lever B is in position three (line 12), and so line 20 can never be reached until this bug is fixed.

```
1   while ( true ) {
2    BSYNC( {}, { ClickButton }, {} );
3
4    Set<Event> queryA = { LeverAInOne, LeverAInTwo, LeverAInThree };
5    Set<Event> queryB = { LeverBInOne, LeverBInTwo, LeverBInThree };
6    Set<Event> queryC = { LeverCInOne, LeverCInTwo, LeverCInThree };
7
8    BSYNC( queryA, {}, {} );
9    if ( lastEvent() != LeverAInTwo )
10     continue;
11
12   BSYNC( queryB, {}, {} );
13   if ( lastEvent() != LeverBInThree )
14     continue;
15
16   BSYNC( queryB, {}, {} );
17   if ( lastEvent() != LeverBInTwo )
18     continue;
19
20   BSYNC( { OpenDoor }, {}, {} );
21  }
```

And then, a few seconds later:

Information: Event OpenDoor appears to only be triggered when the shared array is in configuration LeverAInTwo, LeverBInThree, LeverCInTwo.

Here, the development framework was able to deduce — without any information regarding the specific system being modeled — that configuration $\langle 2,3,2 \rangle$ is of special importance in the triggering of OpenDoor events! This does not indicate a potential error that the development framework found, as in the previous cases shown, but rather an emergent property that the framework was able to deduce — completely on its own — and which may be of interest to the developer. Such emergent properties can serve to either draw attention to bugs or reassure the developer that the model functions as intended, which was the case here. Details about how this conclusion was reached are presented in the next section. A video demonstrating the examples described in this section is available online at (Harel, Katz, Marelly, & Marron, 2016).

4. EXPLAINING THE FRAMEWORK: THE THREE "SISTERS"

We now describe in some detail the inner workings of our wise development framework and the various components from which it is comprised. Although this framework is but a first step towards the ultimate goal described in (Cerf, 2014; Harel, Katz, Marelly, & Marron, 2015; Reubenstein & Waters, 1991; Rich & Waters, 1988), it utilizes some powerful techniques, and building it was far from trivial. An up-to-date version of the tool, as well as video clips demonstrating its main principles, can be found online at (Harel et al., 2016).

As mentioned earlier, our wise development framework is designed to accompany the development of behavioral models, as defined in the Scenario-Based Modeling section, and in particular behavioral programs written in C++ using the BPC package (Harel & Katz, 2014). The framework involves three new logical components, over and above the BPC package itself, and apart from the additional external tools we invoke, such as a model checker and an SMT solver (see Figure 5). We call these components the three sisters: Athena, Regina and Livia.

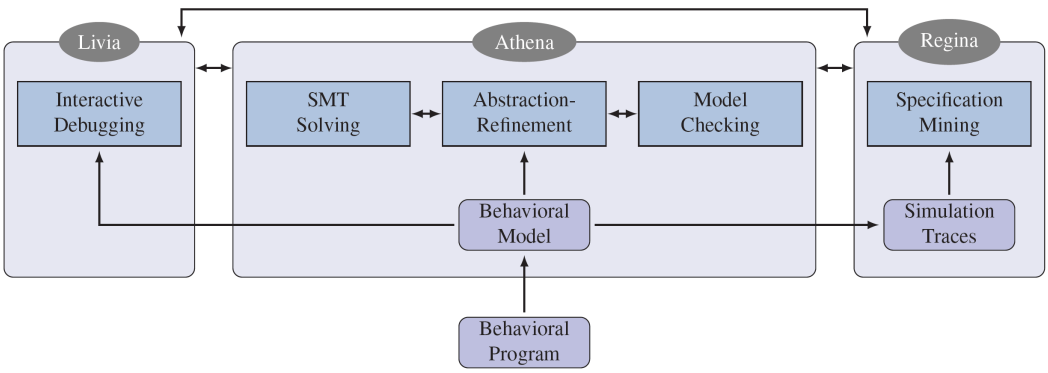
Intuitively, each sister handles a different set of services provided by the wise development environment. Athena, the wise one, works proactively during development, in an off-line fashion. Her purview is the usage of formal tools to analyze scenario objects and produce logically accurate conclusions about them, which are valid for all runs. For instance, in the example discussed in the previous section, the conclusion that a certain scenario state could never be reached was derived by Athena, using model checking.

Regina, more regal than her sisters, also works off-line, but her purview includes semi-formal methods: using abstract models of the system, she runs multiple simulations, collecting statistical information as she goes. In what is a form of specification mining she then attempts to reach interesting conclusions, to be presented to the modeler. Her conclusions may not be valid for all runs, but they have the advantage of reflecting numerous executions, and can thus provide valuable insights about what will happen in typical runs. Again recalling the levers examples from the previous section, the discovery that OpenDoor events were related to lever configuration $\langle 2,3,2 \rangle$ was made by Regina, as a result of running multiple simulations of the system.

The last sister, Livia, who was not demonstrated in the previous section complements the other components by providing on-line support for the developer, for debugging and testing purposes. She can monitor the system as it runs, and help the developer recognize and comprehend unexpected behavior — also by sometimes running local simulations and tests, and by using an abstract model of the system. She can also help the developer create test scenarios, and apply coverage criteria in order to check whether they overlap with previously defined tests.

The three sisters also cooperate: for instance, emergent properties recognized by Regina can be passed to Athena for formal verification, and Livia may use Athena's formal analysis tools for local analysis at runtime. Together, the three sisters are meant to accompany the programmer during development time and provide the various features which together constitute the initial wise development framework.

Figure 5. A high-level overview of the three sisters. The developer provides a behavioral program, from which Athena extracts a behavioral model. She then analyzes this model using abstraction-refinement, model checking and SMT solving. Athena also shares the behavioral model with her sisters: with Regina for the purpose of specification mining, and with Livia for interactive debugging. The three sisters also exchange information with each other — for instance, Regina may ask Athena to attempt to formally prove an emergent property that she mined.



In the remainder of this section we delve deeper into the technical aspects of the framework.

4.1. Offline Analysis: Athena and Regina

The offline components Athena and Regina continuously run as background processes at development time. After each successful compilation of the code, these two sisters receive a fresh snapshot of the program and begin to analyze it. Next, we discuss the main steps in their analysis process, repeated after each compilation.

Step 1: Extracting a behavioral model. The first step is a key one, and is performed by Athena: she constructs an abstract, executable behavioral model of the program, to be used by all three sisters, in all their further analysis operations. Intuitively, Athena extracts from the program — given as C++ code — the underlying scenario objects, as described in the Analyzing Behavioral Models subsection. This technique, discussed in (Harel, Kantor, and Katz, 2013), leverages the fact that concurrent scenarios communicate only through the strict BP synchronization mechanism. Athena thus runs each scenario individually in a sandbox, while mimicking the program's event selection mechanism, exploring the scenario's states and constructing its underlying scenario object. The resulting abstract model of the program thus completely and correctly describes all inter-scenario communication, while the rest of the information (internal scenario actions) is abstracted away, allowing the development framework to handle larger programs. Athena then shares this abstract behavioral model with Regina for the purpose of running simulations, and with Livia for the purpose of online analysis.

Step 2: Identifying logical modules. The next phase is also performed by Athena, and it involves partitioning the program's scenarios into logical modules according to their functionality. This clustering phase is needed in order to increase the tool's scalability: when trying later to check a property ϕ that does not involve program module A, the sisters will attempt to abstract away module A — reducing the total number of states that have to be explored. We have set things up so that information regarding the scenario grouping into modules is not provided by the programmer; rather, Athena uses a clustering algorithm (Katz, 2013) to determine scenarios' correlations to events, and then groups them accordingly.

The clustering algorithm operates as follows. The basic idea is that objects that are logically related are likely to "care" about the same events. Thus, we define the correlation between a scenario object O (with state set Q) and an event e as:

$$cor\left(O,e\right) = \frac{\left|\left\{q \in Q \mid e \in R\left(q\right) \cup B\left(q\right)\right\}\right|}{\left|Q\right|}$$

i.e. the portion of O's states in which event e is requested or blocked. Given a threshold M, this correlation relation defines an equivalence relation, where if $cor\left(O_1,e\right) > M$ and $cor\left(O_2,e\right) > M$ then objects O_1 and O_2 are in the same equivalence class. M is determined dynamically — Athena starts by setting it to 1, and then gradually reduces it until the computed equivalence classes are sufficiently large. For the definition of "sufficiently", we have empirically found that requiring at least 75% of the computed object classes to have at least 4 objects in them worked well on our examples — i.e., it leads to non-trivial equivalence classes that indeed contain logically related scenario objects.

Apart from applying this clustering algorithm, Athena also compares the extracted behavioral model to a predefined meta-model with known/common programming constructs (Katz et al., 2015) which we have built into our tool. Currently supported constructs include semaphores, shared arrays, sensors and actuators, and our on-going work includes adding support for additional ones. If it is discovered that certain scenario objects are instantiations of meta-objects that are logically connected (e.g., one scenario implements a semaphore and another scenario waits on that semaphore), they

may also be grouped together into the same logical module. Recalling the levers example, it was Athena who realized, by comparing the input model to her stored meta-model, that the lever scenarios constituted a shared ternary array.

Step 3: Deriving candidate emergent properties. The next step employs specification mining techniques, and is performed by Regina. She attempts to determine, by running multiple simulations on the behavioral model of the program (which was provided by Athena), a list of possible properties of the system. These are discovered by analyzing simulation traces and looking for patterns: events that always (or never) appear together, events that cause other events to occur, producer-consumer patterns, etc. Such abilities can be viewed as a form of trace mining for scenario-based specifications (see, e.g., (Lo, Maoz, & Khoo, 2007)). The generated properties are not guaranteed to be valid, and need to be checked — either formally, by Athena (e.g., by model checking), or statistically, by Regina (e.g., by running even more simulations of the system). If and when proven correct, and assuming they are relevant, these emergent properties can serve as part of the official certification that the system performs as intended (an example appeared at the end of the previous section). However, even when the sisters guess "incorrectly", i.e., come up with properties that are later shown not to hold, this can still be quite useful, often drawing the developer's attention to bugs.

Step 4: Prioritizing properties. Once Regina has obtained a list of candidate properties, the next step is to attempt to prove or disprove each of them. In our experience with the tool, for a large system this list tends to contain dozens of properties, and so it is typically infeasible to model-check each and every one of them and present the conclusions quickly. This difficulty is mitigated in our system in several ways: (1) We attempt to reduce redundancy. Thus, if we have identified a class of similar emergent properties, we may start by checking just one of them and assign the remaining properties a lower priority. (2) We employ a prioritization heuristic, aimed at checking first those properties that are likely to be more interesting to the user. For instance, if a semaphore-like construct was identified, we will prioritize the checking of a property that states that in some cases mutual exclusion may be incorrectly implemented, as this is considered a safety critical property, and thus may be more interesting to the user. (3) We present any conclusion to the user as soon as it is reached, while the sisters continue to check additional properties. (4) We leave room for manual configuration of the framework; i.e. the developers can prioritize the testing of certain properties, if they so desire.

Having obtained a prioritized list of properties to check, the remainder of the framework's operation is dedicated to discharging each of them (step 5) and presenting the results to the user (step 6). The framework will thus alternate between steps 5 and 6 until all the candidate emergent properties have been discharged, or until it runs out of time — possibly due to a renewed compilation of the code and the start of another analysis cycle.

Step 5: Proving/disproving properties. The wise development framework now attempts to check, in sequence, each of the candidate properties. As there are typically many properties to check, it is desirable to dispatch each property as soon as possible — so that the results will be presented to the user in time to be relevant. To this end, we build upon a large body of existing techniques for formally analyzing scenario-based models, as discussed in the subsection Analyzing Behavioral Models. These include, e.g., abstraction-refinement techniques (Katz, 2013), program instrumentation techniques (Harel, Katz, et al., 2014) and SMT-based compositional techniques (Harel, Kantor, Katz, et al., 2013; Katz et al., 2015). Indeed, this is the main reason why we chose to implement a wise framework in the context of the scenario-based paradigm: it is sufficiently expressive for real-world systems (Harel & Katz, 2014), but on the other hand is amenable to, and even facilitates, program analysis (Harel, Katz, Marron, & Weiss, 2015). Since the ability to quickly and repeatedly analyze behavioral models is critical to our approach, this seemed like a natural fit.

By default, Athena will attempt to discharge properties using abstraction-refinement based model checking for scenario-based programs (Katz, 2013). Alternatively, the user may configure the framework to use other tools: explicit model checking or an SMT-based approach (also performed by Athena), or have Regina perform statistical checking. Here, statistical checking entails Regina running

many simulations under various environment assumptions (fair/unfair environment, starvation, round-robin triggering of events, etc.), and repeatedly checking the property at hand. This technique is not guaranteed to be sound, of course, but it can yield interesting conclusions nonetheless. Moreover, it affords a level of assurance of the property holding, which may suffice for ones that are not safety-critical. We are currently in the process of implementing an adaptive mechanism that would attempt to run the various techniques in Athena's arsenal with a timeout value, abandoning a technique if it does not prove useful for a specific input.

Step 6: Presenting the results. The final phase of the offline sisters' analysis cycle involves displaying to the user the properties that were proved or disproved. In some cases, the mined properties are irrelevant, and the user may discard them. In other cases, desirable properties are shown to hold, and the user is then reassured that the program is working as intended. The remaining cases can either be undesired properties that do hold, or "classical" bugs, where a property that the user assumed to hold is proven by Athena to be violated. In the latter case, the user can interact with the development framework, and ask for (1) a trace log showing how the property was violated; (2) a suggestion for a fix, in the form of a scenario that is to be added to the model (Harel, Katz, Marron, & Weiss, 2012; Harel, Katz, et al., 2014); or (3) the addition of a monitor scenario, to alert the user when the property is violated at run-time (usually used for debugging purposes).

Apart from the analysis flow just described, Athena also supports some forms of automatic optimization — e.g., identifying parts of the code that may never be reached and suggesting how to remove them, as we saw in the previous section.

4.2. Online Analysis: Livia

So far we have dealt with the framework's offline capabilities, performed by Athena and Regina — that is, analysis performed during development, usually after compilation, but without running the actual system. In contrast, the online sister Livia participates in debugging and testing the system as it runs. We now describe her functionality in some detail.

In order to monitor the behavioral model at runtime, Livia connects to the system and "pretends" to be yet another scenario object. In this context, she is typically a passive scenario — never requesting or blocking any events, and thus she does not alter the behavior of the system. However, Livia constantly waits for every one of the model's events, which allows her to monitor the state of the system as it runs. (Indeed, this is a convenient way to attach hooks to the model without changing its semantics, and even without recompiling existing scenarios.) Prior to being run, Livia is provided with the abstract model of the program produced by Athena in the first step of the offline analysis; and this information, together with the sequence of events triggered when the program runs, allows Livia to keep track of the internal states of every object in the system and reason about its behavior.

Livia provides two kinds of capabilities. The first revolves around bounded model checking: at any point during the model's execution, Livia can launch a bounded model checking procedure from the present state, checking for properties at runtime. For instance, if, when debugging the program with breakpoints, the developer believes that the system has arrived at a state from which it can no longer reach some other state, which the developer knows must always be reachable, he/she can ask Livia to try and refute this conjecture. Livia will then apply bounded model checking in order to seek a path to the target state. If such a path is found, it will be displayed to the developer; otherwise, a possible bug has been found.

In addition to such user-initiated bounded model checking, Livia also attempts to recognize problematic cases on her own. Specifically, she detects when certain objects in the system may have become deadlocked — i.e., have not changed states in a long while — and asks the user whether she should investigate. If instructed to do so, she applies bounded model checking to see if there exists a path along which the possibly deadlocked objects progress, and informs the user of her findings. Of course, due to the nature of bounded model checking, it is possible that the objects are not really deadlocked but merely that a path that causes them to progress was not found. As

before, if the user believes this is indeed the case, he/she can have Livia pass the query to Athena for a more thorough analysis.

The second kind of capability that Livia provides regards testing. When run in interactive mode, Livia enables the user to guide the execution of the model by picking, at each step, which of the currently enabled events should be triggered next. Once she obtains the user's choice, Livia blocks all the remaining events, effectively forcing the system to trigger the event selected by the user. The user can choose to control some steps, while allowing the model to behave normally in others. Livia observes and records the sequence of triggered events and, upon request, can transform this information into a test scenario that is added to the system.

For example, suppose the user begins the execution by forcing the triggering of event e_1, and consequently event e_2 gets triggered three consecutive times. Upon request, Livia will generate a test scenario that will mimic this execution. This test scenario will (1) force the system to replay every event selection that the user made in the manually controlled run (this is accomplished by having the test scenario block, at each such step, all events except for the one selected by the user); and (2) without influencing the remaining event selections (that is, without requesting or blocking any events), it will check that the model still behaves as it did in the earlier run; i.e. that the same events get triggered. In our example we have just one step controlled by the user — the triggering of e_1, which the test scenario will enforce. Then, the test scenario will wait for the three events following e_1 and check that they are all e_2 events, raising an error flag otherwise. At a later point in time, the user can run this test case (by simply adding it to the system as a scenario) in order to guarantee that the recorded functionality is still supported. All tests generated in this manner are given as simple scenarios, which the user can later edit and enhance, in a way that further augments the intuitive programming by "playing-in" of scenarios (Harel & Marelly, 2003).

Throughout the development process, the behavioral model may accumulate a large number of test scenarios. Repeatedly running the entire test suite may then become resource-consuming, making it desirable to remove redundant tests — ones that check similar functionality. A common approach to achieving this is via combinatorial test design (CTD) (Tatsumi, 1987), where one associates a test with the system parameters whose interactions it checks. Whenever two tests check similar parameters, one of them can be removed.

In Livia we implemented the following coverage criterion for tests (given as scenario objects). For a test scenario forcing the event sequence $e_1,\ldots,e_n$, we say that scenario object O is active in window $\left[i,j\right]$, for $i<j$, if for all $i \le k \le j$ event e_k causes object O to change states. Livia can calculate, for every scenario object O, its activity windows during the test scenario. Then, the activity windows of various scenarios are intersected in order to deduce which scenario objects interact (i.e., are active at the same time) during the test scenario. If multiple test scenarios involve interactions between the same scenario objects, then some of these tests may be candidates for removal. This process is presently performed in a semi-automatic fashion, and we are working on automating it further.

5. A CASE-STUDY: A CACHE COHERENCE PROTOCOL

In order to evaluate the applicability of our wise development framework to larger systems, we used it to develop a cache coherence protocol. Such protocols are designed to ensure consistent shared memory access in a set of distributed processors. In order to minimize the number of read operations on the actual memory, processors cache the results of previous reads. Consistency then means that cached values stored throughout the system need to be invalidated when a processor writes a new value to the actual memory. The motivation for choosing this particular example was that cache coherence protocols are notoriously susceptible to subtle, concurrency-related bugs, making them

a prime candidate to benefit from a wise development environment. The specific protocol that we implemented is a variant of the well-studied Futurebus protocol (Clarke et al., 1995).

An important question that we attempted to address through the case-study was whether the notion at the core of our approach — namely, developing a non-trivial system together with the aid of a proactive framework — is convenient and/or useful. While this issue is highly subjective, we can report that in the systems we modeled the sisters' aid proved valuable. In particular, Athena and Regina typically displayed their insights about the program in a timely manner, with results starting to flow in seconds after each compilation; and although sometimes the insights proved irrelevant, in several cases they pointed out concurrency-related bugs that we had overlooked, and which we then repaired. In other cases, the framework's conclusions served to confirm that the model was working as intended, which was particularly reassuring, for example, after adding a new feature. Similarly, Livia proved useful in debugging and in allowing us to create test suites that tested the core functionality of the system.

Another goal that we had was to identify a basic methodology for how modeling or programming should be conducted in a wise environment, i.e., in collaboration with Athena, Regina, and Livia. A setup that we found convenient is depicted in Figure 6. As for the flow of the process, we found it useful to have a quick glance at the framework's logs after each compilation to check for any critical mistakes, and to look more thoroughly at the logs after making significant changes to the code base. After significant changes it was also useful to create new test scenarios using Livia. Occasionally, when certain properties found by Regina draw our particular attention, we used the interactive interface (depicted in Figure 7) to guide the framework by prioritizing them.

We now show two examples of the usage of the wise development framework during our case-study. A more complete set of examples, as well as the entire code base, is available online at (Harel et al., 2016). In order to properly illustrate the tool's usage during development, we took snapshots of our code at significant milestones, along with the conclusions that the wise framework was able to draw from it — these are also available online. Finally, we also provide there a video clip that features the development framework in action.

Figure 8 depicts a list of emergent properties that the development framework produced at one point during development. Recall that unless given specific instructions by the developer, the tool begins to check these properties, one by one; the figure shows a list of properties that have already been checked, indicating which of them hold and which do not. The tool mines for various types of properties, two of which are depicted in the figure: implications, denoted $a \rightarrow b$, i.e., whenever event a occurs b also occurs a short time earlier or later, and equivalences, denoted $a \leftrightarrow b$, i.e., the implication holds in both directions.

Figure 9 depicts an example for which Athena's abstraction-based model checking proved especially handy, allowing her to quickly cover more properties. There, the emergent property being verified was that "cache 3 cannot acquire bus 2 repeatedly without first releasing it" — a property that describes mutual exclusion in the bus ownership. This property is an instantiation of the general pattern "consecutive a events must have b events between them". At the time this property was mined and tested, verifying it via direct model checking entailed exploring 972233 reachable states and took over 27 minutes. By using the abstraction-refinement techniques discussed in the subsection Analyzing Behavioral Models, Athena was able to abstract away irrelevant parts of the code (namely code modules that only pertained to other buses). In this way, verifying the property entailed exploring just 21000 reachable states, and took less than 31 seconds. The key observation here is that this is by no means merely a standard direct usage of abstraction-refinement. The entire process — finding the emergent property, figuring out which modules are not likely to affect it so that they can be abstracted away, and then model checking the property on the abstract model — were all handled proactively and automatically by the framework. Clearly, such speedups allow the framework to cover more properties and present them to the programmer in a timely manner.

Figure 6. Screenshots of our wise development framework, taken during the cache coherence case-study. The top window depicts a standard editor, in which the code of the program is being written. The analysis tools are running in the background, and with every successful compilation of the code they automatically receive a fresh snapshot and analyze it. The bottom window shows output from the analysis — in this case, emergent properties that were examined. One property was proved correct and another was shown not to hold (a counter-example is provided). Most of the time we had these two windows open on separate screens.

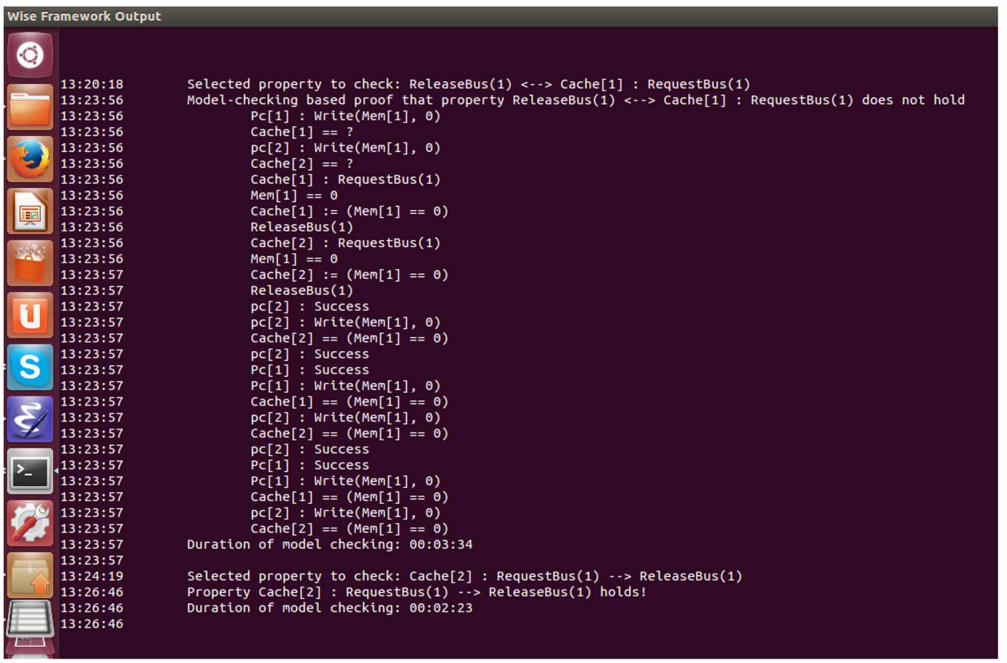
```
#ifndef __CacheOneReadHasBit_h__
#define __CacheOneReadHasBit_h__

#include "BThread.h"
#include "EventCodes.h"

class CacheOneReadHasBit : public BThread
{
public:
    CacheOneReadHasBit() : _getBitZero( *this, false )
    {
        _cacheOneReadEvents.insert( EventCodes::CACHE_ONE_EMPTY );
        _cacheOneReadEvents.insert( EventCodes::CACHE_ONE_BIT_0_IS_0 );
        _cacheOneReadEvents.insert( EventCodes::CACHE_ONE_BIT_0_IS_1 );
        _cacheOneReadEvents.insert( EventCodes::CACHE_ONE_BIT_1_IS_0 );
        _cacheOneReadEvents.insert( EventCodes::CACHE_ONE_BIT_1_IS_1 );
    }

    void entryPoint()
    {
        while ( true )
        {
            waitForRequests();

            if ( bitIsStored() )
            {
                reset();
                replyToRequest();
            }
            else
            {
                reset();
                waitForReply();
            }
        }
-:**-  CacheOneReadHasBit.h   Top (1,0)     SVN-22495  (C++/lah Abbrev)--13:03 0.47----------
Beginning of buffer
```



```
Wise Framework Output

13:20:18        Selected property to check: ReleaseBus(1) <--> Cache[1] : RequestBus(1)
13:23:56        Model-checking based proof that property ReleaseBus(1) <--> Cache[1] : RequestBus(1) does not hold
13:23:56                Pc[1] : Write(Mem[1], 0)
13:23:56                Cache[1] == ?
13:23:56                pc[2] : Write(Mem[1], 0)
13:23:56                Cache[2] == ?
13:23:56                Cache[1] : RequestBus(1)
13:23:56                Mem[1] == 0
13:23:56                Cache[1] := (Mem[1] == 0)
13:23:56                ReleaseBus(1)
13:23:56                Cache[2] : RequestBus(1)
13:23:56                Mem[1] == 0
13:23:57                Cache[2] := (Mem[1] == 0)
13:23:57                ReleaseBus(1)
13:23:57                pc[2] : Success
13:23:57                pc[2] : Write(Mem[1], 0)
13:23:57                Cache[2] == (Mem[1] == 0)
13:23:57                pc[2] : Success
13:23:57                Pc[1] : Success
13:23:57                Pc[1] : Write(Mem[1], 0)
13:23:57                Cache[1] == (Mem[1] == 0)
13:23:57                pc[2] : Write(Mem[1], 0)
13:23:57                Cache[2] == (Mem[1] == 0)
13:23:57                pc[2] : Success
13:23:57                Pc[1] : Success
13:23:57                Pc[1] : Write(Mem[1], 0)
13:23:57                Cache[1] == (Mem[1] == 0)
13:23:57                pc[2] : Write(Mem[1], 0)
13:23:57                Cache[2] == (Mem[1] == 0)
13:23:57        Duration of model checking: 00:03:34
13:23:57
13:24:19        Selected property to check: Cache[2] : RequestBus(1) --> ReleaseBus(1)
13:26:46        Property Cache[2] : RequestBus(1) --> ReleaseBus(1) holds!
13:26:46        Duration of model checking: 00:02:23
13:26:46
```

Figure 7. A simple GUI that we occasionally used in order to interactively instruct the development environment to focus on certain emergent properties. The interface allows us to choose which of the candidate emergent properties should be handled next, and how: explicit or abstraction-based model checking, statistical testing, creating a monitor thread, etc.

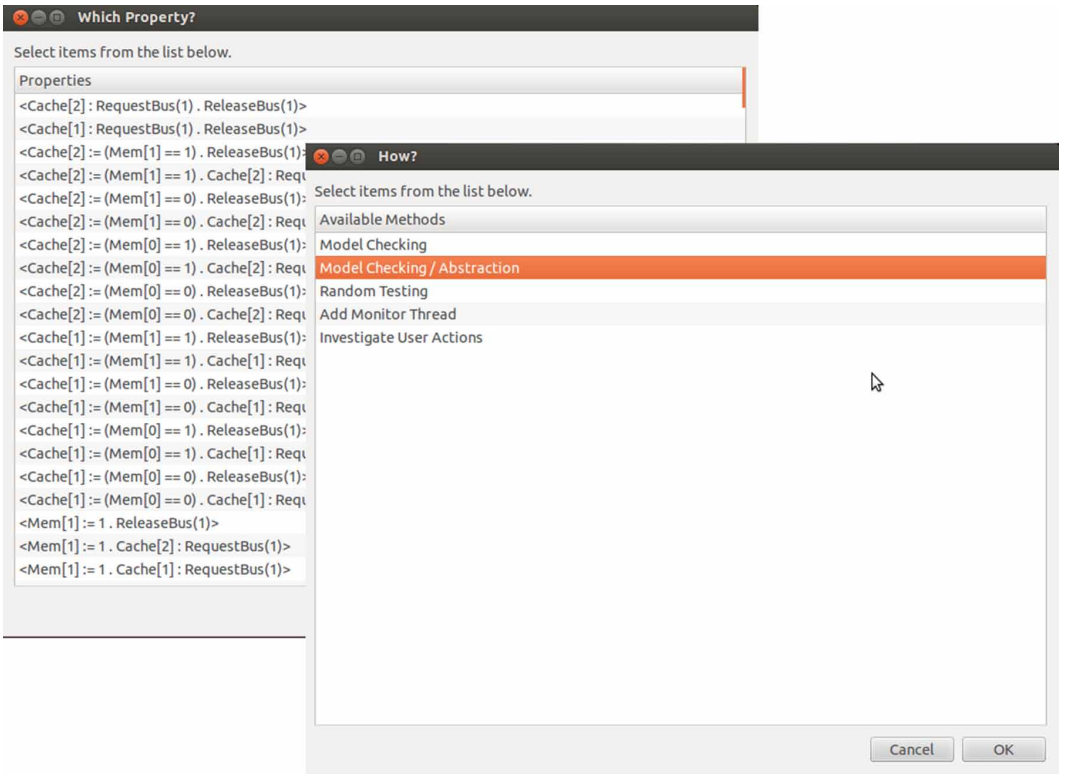


Figure 8. A list of emergent properties produced and checked by the wise development framework. The tool typically does not finish checking everything on the list, and so information is displayed as soon as it is available. A counter-example is available for properties that fail to hold.

```
Checking emergent properties:

ReleaseBus(1) <--> Cache[2] : RequestBus(1)
  [fails]
Cache[2] : RequestBus(1) --> ReleaseBus(1)
  [holds]
ReleaseBus(1) <--> Cache[1] : RequestBus(1)
  [fails]
Cache[1] : RequestBus(1) --> ReleaseBus(1)
  [holds]
Cache[2] := (Mem[1] == 1) --> ReleaseBus(1)
  [holds]
ReleaseBus(1) <--> Pc[1] : Success
  [fails]
Cache[2] : RequestBus(1) <--> Pc[2] : Success
  [fails]
...
```

Figure 9. Extracts from the logs of the wise development framework, illustrating the autonomous verification of an emergent property that has been identified. The three code modules depicted (each a set of scenario objects) are irrelevant to the property at hand, and are automatically abstracted. Other modules in the program, those that are relevant to the property at hand, are not abstracted. The property is then verified for the resulting over-approximation — leading to improved performance.

```
Checking emergent property:
 Consecutive Cache[3] : RequestBus(2) events must have ReleaseBus(2)
 events between them

Attempting abstraction-based model checking
Abstracting module 1:
 CacheOneUpdate, CacheTwoUpdate,
 CacheTwo, CacheOne,
 CacheTwoReadFetchBit, CacheOneReadFetchBit,
 CacheTwoReadHasBit, CacheTwoWriteFetchBit,
 CacheTwoWriteHasBit, PcTwoRead, PcTwoWrite,
 CacheOneReadHasBit, CacheOneWriteHasBit,
 PcOneRead, PcOneWrite, CacheOneWriteFetchBit

Abstracting module 2:
 CacheTwoInvalidate

Abstracting module 3:
 CacheOneInvalidate

Conclusion: property [holds]
```

6. RELATED WORK

Work related to the subject of this paper can be viewed in two perspectives. One is over the individual capabilities of the three sisters: discovering and proposing candidate emergent properties and then verifying or refuting them (for the offline components), and runtime analysis and test suite minimization (for the online component). The other perspective is that of the overall view of a wise development environment that accompanies the developer and automatically and proactively carries out these tasks and others, such as requirement analysis, specification mining, test generation, synthesis, and more.

From the first perspective, there is a vast amount of pertinent research, and we focus here on only a few of the relevant papers. The actions performed by Regina, i.e. the dynamic discovery of candidate properties and invariants from program execution logs, is a form of specification mining (Ammons, Bodik, & Larus, 2002). This topic has been studied in the context of scenario-based specification in, e.g., (Cantal de Sousa, Mendonca, Uchitel, & Kramer, 2007; Lo & Maoz, 2008), and Regina uses similar techniques. For instance, she looks for emergent properties that have the trigger and effect structure of (Lo & Maoz, 2008). However, a key aspect in Regina's operation is the need to conclude the mining phase as quickly as possible, so that she can be seamlessly integrated into the development cycle. This is achieved by employing prioritization heuristics, and putting limits on the number of traces (and lengths thereof) that Regina considers. In the future we intend to enhance Regina with a mechanism similar to the one discussed in (Cohen & Maoz, 2015), where statistical criteria are used to determine when "enough" traces have been considered, hopefully boosting Regina's performance even further.

Checking whether properties mined from traces indeed hold for the model in general brings us to the broad field of program and model verification. Many powerful and well known tools exist, such as SPIN, SLAM, BLAST, UPPAAL, Java Pathfinder, ASTRÉE, ESC/Java and others, and they utilize many forms of explicit and symbolic model checking, static analysis, deductive reasoning, and SAT and SMT solving (see (Alur, Henzinger, & Vardi, 2015) for a brief survey of the application of

such methods in practice). In our framework these tasks are handled by Athena, and she uses tools specifically optimized for behavioral models (Harel et al., 2011; Katz, 2013; Katz et al., 2015).

The topic of combinatorial test design has been studied extensively, but the approach of (Panzica La Manna, Segall, & Greenyer, 2015) is likely the one most closely related to our own. There, the authors study scenario-based systems modeled using the Modal Sequence Diagrams (MSDs) formalism, and employ coverage criteria in synthesizing effective test suites. While our approach so far has been to focus on minimizing user-provided test suites, it would be interesting to combine the two approaches and enhance our framework with proactive test synthesis capabilities.

As to the second perspective, successful attempts at automatic property discovery and subsequent verification appear, e.g., in (Nimmer & Ernst, 2001; Zhang, Yang, Rungta, Person, & Khurshid, 2014). There, the Daikon tool is used to dynamically detect candidate program invariants which are then used to either annotate or instrument the program. In (Nimmer & Ernst, 2001) these guide ESC/Java in verifying the properties, and in (Zhang et al., 2014) they help guide symbolic execution in the discovery of additional or refined invariants. The motivation and approach of Daikon are very close to ours, but we aim at constructing a fully integrated, proactive and interactive environment, built upon the highly incremental paradigm of behavioral modeling.

Providing an interactive analysis framework that is tightly integrated into the development cycle/environment has become quite widespread in the industry over recent years. Some noticeable examples are Google's Tricorder (Sadowski et al., 2015), Facebook's Infer (Cristiano et al., 2015) and VMWare's Review Bot (Balachandran, 2013) tools. These tools use static analysis to automate the checking for violations of coding standards and for common defect patterns. Lessons learned from these projects indicate that, in order to be successfully accepted by programmers, an integrated analysis framework should have the following properties: (1) it needs to be seamlessly integrated into the workflow of developers; (2) it must produce results quickly; and (3) it has to perform its analysis in a modular manner, so that it can scale reasonably well to large projects. The design of our framework is indeed aimed at achieving these properties. In particular, for the modular analysis part, Athena attempts to leverage the special properties of scenario-based models and reason about individual objects. In (Harel, Katz, Lampert, et al., 2015), it is shown that objects in behavioral models often have very small state spaces; and this allows Athena to effectively compare these objects to her stored meta-model and identify object patterns that can later be used for analysis.

7. CONCLUSION

In this paper we contribute to the effort of simplifying and accelerating development of robust reactive systems, by proposing a development framework along the lines raised in e.g., (Cerf, 2014; Harel, Katz, Marelly, & Marron, 2015). In a nutshell, the idea is to start with a modeling/programming formalism that is expressive, modular and relatively simple, and integrate quick, continuous, and easy-to-use analysis into the development process. This entails extending and adjusting existing analysis techniques in order to render them more interactive and proactive.

Our development framework is currently comprised of three main elements: specification mining and initial semi-formal analysis for generating candidate system properties, abstraction-assisted formal analysis for verification of detected properties, and run-time debugging and testing. When integrated into the development cycle, these elements can often draw developers' attention to subtle bugs that could otherwise be missed. We carried out initial evaluation of the framework by iteratively developing a cache coherence protocol, and saw that it was successful in discovering and reporting bugs.

In the future we plan to carry out a more extensive, empirical comparison between our development framework and related tools, such as Tricorder (Sadowski et al., 2015) and Infer (Cristiano et al., 2015). We also plan to enhance Regina's specification-mining capabilities with learning techniques

(Ammons et al., 2002), allowing her to learn over time which emergent properties are most valuable to programmers and should be checked first.

While our work so far is but an early step towards the vision of the computer acting as a wise, fully-fledged proactive member of the development team, we hope that it contributes to demonstrating both the viability and the potential value of this direction.

ACKNOWLEDGMENT

A preliminary version of part of the material in this paper appeared in Harel, D., Katz, G., Marelly, R., & Marron, A. (2016). An Initial Wise Development Environment for Behavioral Models. In Proc. 4th Int. Conf. on Model-Driven Engineering and Software Development (MODELSWARD) (pp. 600–612). This work was supported by a grant from the Israel Science Foundation, by a grant from the German-Israeli Foundation (GIF) for Scientific Research and Development, by the Philip M. Klutznick Research Fund and by a research grant from Dora Joachimowicz.

REFERENCES

Alexandron, G., Armoni, M., Gordon, M., & Harel, D. (2014). Scenario-Based Programming: Reducing the Cognitive Load, Fostering Abstract Thinking. *Proc. 36th Int. Conf. on Software Engineering (ICSE)* (pp. 311–320). doi:10.1145/2591062.2591167

Alur, R., Henzinger, T. A., & Vardi, M. Y. (2015). Theory in Practice for System Design and Verification. *ACM Siglog News*, *2*(1), 46–51.

Ammons, G., Bodik, R., & Larus, J. (2002). Mining Specifications. *ACM Sigplan Notices*, *37*(1), 4–16. doi:10.1145/565816.503275

Balachandran, V. (2013). Reducing Human Effort and Improving Quality in Peer Code Reviews using Automatic Static Analysis and Reviewer Recommendation. *Proc. 35th Int. Conf. on Software Engineering (ICSE)* (pp. 931–940). doi:10.1109/ICSE.2013.6606642

Cantal de Sousa, F., Mendonca, N. C., Uchitel, S., & Kramer, J. (2007). Detecting Implied Scenarios from Execution Traces. *Proc. 14th Working Conf. on Reverse Engineering (WCRE)* (pp. 50–59). doi:10.1109/WCRE.2007.19

Cerf, V. (2014). A Long Way to Have Come and Still to Go. *Communications of the ACM*, *1*(58), 7–7.

Clarke, E., Grumberg, O., Hiraishi, H., Jha, S., Long, D., McMillan, K., & Ness, L. (1995). Verification of the Futurebus+ Cache Coherence Protocol. *Formal Methods in System Design*, *6*(2), 217–232. doi:10.1007/BF01383968

Cohen, H., & Maoz, S. (2015). Have We Seen Enough Traces? *Proc. 30th Int. Conf. on Automated Software Engineering (ASE)* (pp. 93–103).

Cristiano, C., Distefano, D., Dubreil, J., Gabi, D., Hooimeijer, P., Luca, M., & Rodriguez, D. et al. (2015). Moving Fast with Software Verification.*Proc. 7th. NASA Formal Methods Symposium (NFM)* (pp. 3–11).

Damm, W., & Harel, D. (2001). LSCs: Breathing Life into Message Sequence Charts. *J. on Formal Methods in System Design*, *19*(1), 45–80. doi:10.1023/A:1011227529550

Gordon, M., Marron, A., & Meerbaum-Salant, O. (2012). Spaghetti for the Main Course? Observations on the Naturalness of Scenario-Based Programming. *Proc. 17th Conf. on Innovation and Technology in Computer Science Education (ITICSE)* (pp. 198–203).

Harel, D., Kantor, A., & Katz, G. (2013). Relaxing Synchronization Constraints in Behavioral Programs. *Proc. 19th Int. Conf. on Logic For Programming, Artificial Intelligence and Reasoning (LPAR)* (pp. 355–372). doi:10.1007/978-3-642-45221-5_25

Harel, D., Kantor, A., Katz, G., Marron, A., Mizrahi, L., & Weiss, G. (2013). On Composing and Proving the Correctness of Reactive Behavior. *Proc. 13th Int. Conf. on Embedded Software (EMSOFT)* (pp. 1–10). doi:10.1109/EMSOFT.2013.6658591

Harel, D., Kantor, A., Katz, G., Marron, A., Weiss, G., & Wiener, G. (2015). Towards Behavioral Programming in Distributed Architectures. *Science of Computer Programming*, *98*(2), 233–267. doi:10.1016/j.scico.2014.03.003

Harel, D., & Katz, G. (2014). Scaling-Up Behavioral Programming: Steps from Basic Principles to Application Architectures. *Proc. 4th Int. Workshop on Programming Based on Actors, Agents, and Decentralized Control (AGERE!)* (pp. 95–108). doi:10.1145/2687357.2687359

Harel, D., Katz, G., Lampert, R., Marron, A., & Weiss, G. (2015). On the Succinctness of Idioms for Concurrent Programming. *Proc. 26th Int. Conf. on Concurrency Theory (CONCUR)* (pp. 85–99).

Harel, D., Katz, G., Marelly, R., & Marron, A. (2015). Wise Computing: Towards Endowing System Development with True Wisdom. *Technical report.*http://arxiv.org/abs/1501.05924

Harel, D., Katz, G., Marelly, R., & Marron, A. (2016). An Initial Wise Development Environment for Behavioral Models: Supplementary Material. http://www.wisdom.weizmann.ac.il/~harel/Modelsward.wisecomputing

Harel, D., Katz, G., Marron, A., & Weiss, G. (2012). Non-Intrusive Repair of Reactive Programs. *Proc. 17th IEEE Int. Conf. on Engineering of Complex Computer Systems (ICECCS)* (pp. 3–12).

Harel, D., Katz, G., Marron, A., & Weiss, G. (2014). Non-Intrusive Repair of Safety and Liveness Violations in Reactive Programs. *Transactions on Computational Collective Intelligence* [TCCI], *16*, 1–33.

Harel, D., Katz, G., Marron, A., & Weiss, G. (2015). The Effect of Concurrent Programming Idioms on Verification. *Proc. 3rd Int. Conf. on Model-Driven Engineering and Software Development (MODELSWARD)* (pp. 363–369).

Harel, D., Lampert, R., Marron, A., & Weiss, G. (2011). Model-Checking Behavioral Programs. *Proc. 11th Int. Conf. on Embedded Software (EMSOFT)* (pp. 279–288).

Harel, D., & Marelly, R. (2003). *Come, Let's Play: Scenario-Based Programming Using LSCs and the Play-Engine*. Springer. doi:10.1007/978-3-642-19029-2

Harel, D., Marron, A., & Weiss, G. (2012). Behavioral Programming. *Communications of the ACM*, *55*(7), 90–100. doi:10.1145/2209249.2209270

Katz, G. (2013). On Module-Based Abstraction and Repair of Behavioral Programs. *Proc. 19th Int. Conf. on Logic for Programming, Artificial Intelligence and Reasoning (LPAR)* (pp. 518–535). doi:10.1007/978-3-642-45221-5_35

Katz, G., Barrett, C., & Harel, D. (2015). Theory-Aided Model Checking of Concurrent Transition Systems. *Proc. 15th Int. Conf. on Formal Methods in Computer-Aided Design (FMCAD)* (pp. 81–88). doi:10.1109/FMCAD.2015.7542256

Lo, D., & Maoz, S. (2008). Mining Scenario-Based Triggers and Effects. *Proc. 23rd Int. Conf. on Automated Software Engineering (ASE)* (pp. 109–118).

Lo, D., Maoz, S., & Khoo, S.-C. (2007). Mining Modal Scenario-Based Specifications from Execution Traces of Reactive Systems. *Proc. 22nd Int. Conf. on Automated Software Engineering (ASE)* (pp. 465–468). doi:10.1145/1321631.1321710

Nimmer, J. W., & Ernst, M. D. (2001). Static Verification of Dynamically Detected Program Invariants: Integrating Daikon and ESC/Java. *Electronic Notes in Theoretical Computer Science*, *55*(2), 255–276. doi:10.1016/S1571-0661(04)00256-7

Panzica La Manna, V., Segall, I., & Greenyer, J. (2015). Synthesizing Tests for Combinatorial Coverage of Modal Scenario Specifications. *Proc. 18th Int. Conf. on Model Driven Engineering Languages and Systems (MODELS)* (pp. 126–135). doi:10.1109/MODELS.2015.7338243

Reubenstein, H., & Waters, R. (1991). The Requirements Apprentice: Automated Assistance for Requirements Acquisition. *IEEE Transactions on Software Engineering*, *17*(3), 226–240. doi:10.1109/32.75413

Rich, C., & Waters, R. (1988). The Programmers Apprentice: A Research Overview. *Computer*, *21*(11), 10–25. doi:10.1109/2.86782

Sadowski, C., van Gogh, J., Jaspan, C., Söederberg, E., & Winter, C. (2015). Tricorder: Building a Program Analysis Ecosystem. *Proc. 37th Int. Conf. on Software Engineering (ICSE)* (pp. 598–608).

Tatsumi, K. (1987). Test-Case Design Support System.*Proc. Int. Conf. on Quality Control (ICQC)* (pp. 615–620).

Tsay, Y., Chen, Y., Tsai, M., Wu, K., & Chan, W. (2007). GOAL: A Graphical Tool for Manipulating Büchi Automata and Temporal Formulae. *Proc. 13th Int. Conf. on Tools and Algorithms for the Construction and Analysis of Systems* (TACAS) (pp. 466–471). doi:10.1007/978-3-540-71209-1_35

Zhang, L., Yang, G., Rungta, N., Person, S., & Khurshid, S. (2014). Feedback-Driven Dynamic Invariant Discovery. *Proc. Int. Symposium on Software Testing and Analysis (ISSTA)* (pp. 362–372).

David Harel is the Vice President of the Israel Academy of Sciences and Humanities, and has been at the Weizmann Institute of Science since 1980, serving in the past as Dean of its Faculty of Mathematics and Computer Science. He has worked in logic and computability, software and systems engineering, modeling biological systems and more. He invented Statecharts and co-invented Live Sequence Charts. Among his books are "Algorithmics: The Spirit of Computing" and "Computers Ltd.: What They Really *Can't Do". His awards include the ACM Karlstrom Outstanding Educator Award, the Israel Prize, the ACM Software System Award, the Eme"t Prize, and five honorary degrees. He is a Fellow of ACM, IEEE and AAAS, a member of the Academia Europaea and the Israel Academy of Sciences, and a foreign member of the US National Academy of Engineering and the American Academy of Arts and Sciences.*

Guy Katz is a post-doctoral research fellow at Stanford University, working with Prof. Clark Barrett. He received his PhD in computer science from the Weizmann Institute of Science in 2015, where he was advised by Prof. David Harel. Guy's research interests lie at the intersection between Software Engineering and Formal Methods. In particular, he has been working on devising modeling paradigms that are useful and friendly to programmers, but at the same time also amenable to formal analysis, verification and program repair. To this end he has been studying the properties of various concurrency idioms, and how these properties can be leveraged by advanced program analysis tools, such as SMT solvers.

Rami Marelly holds a PhD in computer science from the Weizmann Institute of Science. His research was about specifying and executing behavioral requirements using the Play-in/Play-out approach. Rami Marelly held a variety of key positions in the Israeli Air Force technological directorate including Head of C4I Systems Engineering Department and Head of Aerial ISR Systems Branch. Serving as the head engineer of the IAF operational IT, Rami led the Israeli Air Force transformation program towards network centric warfare and was responsible for the development of ground and airborne digital networks and avionics, simulators, C4I systems and solutions for cyber and information security. After retiring (Col. res.) from the IAF, Rami co-founded Cue, a consulting firm in system engineering, business development and project management. Rami teaches advanced academic courses in systems engineering and volunteers as a mentor to teenagers in various FIRST robotics projects.

Assaf Marron is a researcher at the Weizmann Institute of Science Computer Science and Applied Mathematics Department. His current research interests include software engineering, scenario-based programming, machine learning and information visualization. Assaf holds a PhD in computer science from the University of Houston. Prior to joining the Weizmann Institute, he has worked in senior management and technical positions in research and development of innovative products and technologies at leading companies including IBM and BMC Software. He is the inventor or co-inventor of several patents. For more information and recent publications see his web page.