Performance Regression Testing and Runtime Verification of Components in Robotics Systems

J. Wienke and S. Wrede

Research Institute for Cognition and Robotics (CoR-Lab) and Cluster of Excellence Cognitive Interaction Technology (CITEC), Bielefeld University, Germany

ARTICLE HISTORY
Compiled December 8, 2017

ABSTRACT
Unintended changes in the utilization of resources like CPU and memory can lead to severe problems for the operation of robotics and intelligent systems. Still, systematic testing for such performance regressions has largely been ignored in this domain. We present a method to specify and execute performance tests for individual components of component-based robotics systems based on their component interfaces. The method includes an automatic analysis of each component revision against previous ones that reports potential changes to the resource usage characteristics. This informs developers about the impact of their changes. We describe the design of the framework and present evaluation results for the automatic detection of performance changes based on tests for a variety of robotics components. Additionally, we demonstrate how performance tests can be used as a basis for learning resource utilization models of components. These models can be used to detect faults at system runtime, which provides an additional level of safety for the systems besides the offline testing.

KEYWORDS
resource utilization; testing; performance bugs; continuous integration; resource prediction

1. Introduction

Nowadays, most robotics and intelligent systems are complex software systems. In order to fulfill increasingly complicated missions, these systems often comprise a multitude of hardware and software components and are developed by larger and often distributed teams. With the emergence of common frameworks (e.g. ROS [1] or YARP [2]), systems often contain off-the-shelf components, with implementations outside of the control of system integrators. This makes the task of maintaining such a system and ensuring its proper functionality a challenge. Nevertheless, many applications require a high reliability and availability of the deployed systems in order to meet the users' demands and to ensure their safety. Consequently, verification techniques become more important during development to ensure the proper functionality of individual components and the integrated system. Apart from common software engineering methods like unit testing at class-level and integration testing at component-level,

CONTACT: J. Wienke. Email: jwienke@techfak.uni-bielefeld.de

simulation is a common method applied in robotics to verify the functionality robotics applications at system level [3].

Most of the currently applied testing techniques aim at the delivery of the promised functionality of the system and its components. However, an aspect that has been largely ignored so far in this domain is a systematic verification of the required computational resources. Unplanned changes in the consumption of resources like CPU time, memory, or network bandwidth can lead to various undesired effects on the system, e.g., increased power consumption, delays, reduced accuracies, or component crashes. Depending on the affected subsystems and the application domain, the outcomes of these effects can range from the unnecessary consumption of energy or vanishing user acceptance to severe injuries in case of safety-critical systems. Surveys presented by Steinbauer [3] and in our own work [4] have shown the existence and relevance of such issues for robotics systems. Therefore, the computational performance of robotics systems needs to be closely observed and suitable tools are required to detect performance regressions as soon as possible during development. Such tools need to be applicable for non-experts users, should perform their task automatically for each new revision of a software component, and they should integrate with modern development workflow (e.g. continuous integration (CI) servers [5]) to foster wide adoption and to ensure a high coverage of systems and component revisions.

In this work we present a method for generating and analyzing performance tests for individual components of robotics and intelligent systems. It aims to automatically detect performance regressions introduced by code changes as soon as possible. The method exploits the fact that a majority of current systems is composed of components which communicate via a middleware. A new performance testing framework tests the components using their component interfaces and provides methods to automatically detect and report changes in the performance characteristics. Tests are based on an event generation language which defines abstract test cases that are instantiated for different parameter combinations to explore the runtime behavior of the components under different loads. The framework is designed to integrate into automated build processes. We report on the design of the framework and evaluation results on our systems. Moreover, we demonstrate that performance tests created with the framework can be used to build resource utilization models for the runtime-prediction of a component's resource utilization. These models can be automatically created from data acquired while executing the performance test through machine learning methods. They form the basis to detect resource-related runtime faults of components, which allows to exploit the performance tests to provide an additional runtime verification of robotics systems in addition to the offline testing perspective.

2. Related Work

In contrast to robotics, systematically testing software for performance regressions is a common practice in other disciplines, most notable being large scale enterprise systems and website operation. These disciplines are origin of the "Application Performance Management" (APM) method [6, 7], which combines different practices and tools with the aim to detect performance issues before becoming a problem in the field. In this area, application performance is usually defined along two dimensions of key performance indicators (KPIs): service-oriented (e.g. response times, number of requests) and efficiency-oriented (e.g. CPU) KPIs [7], where efficiency-oriented KPIs match our general motivation. Testing performance in these systems is usually performed on a

much coarser-grained level with the whole system being deployed for testing as a monolithic unit. Tests are often performed based on mimicking or abstracting the human users of the systems (e.g. through http interactions) and can last up to several hours or days [6].

A recent survey by Jiang and Hassan [8] provides a good overview on how research addresses the issue of performance testing of large-scale systems. The authors separate the testing process into three successive steps: test design, execution and analysis and categorize publications along several axes inside each step. The review does not mention any work that specifically focuses on individual components as the unit of testing. Instead, most approaches follow the APM idea of focusing on the complete system.

Following the proposed separation, several distinct methods to design performance tests exist. Tools like *Apache JMeter* [9] and *Tsung* [10] focus on testing via network protocols like HTTP or XMPP and provide methods to generate tests for these protocols. Often, recording capabilities exist to generate interactions based on prototypes and loops and parallel execution can be used to generate extended loads using an abstract specification of the interactions. In JMeter, most interactions are primarily GUI-based, while Tsung uses an XML configuration file and command line utilities for defining tests. Other tools like *Locust* [11], *NLoad* [12], *The Grinder* [13], and Chen et al. [14] use the programming language level to define load tests while tools like *Gatling* [15] are in between these categories by generating code from exemplary executions. In contrast, Da Silveira [16] presents an approach which uses a Domain Specific Language (DSL) to formulate performance tests inside the model-based testing paradigm. Generally, the presented approaches usually provide a way to structure the performance or load test into distinct units like test cases or test phases.

For executing tests, frameworks have the duty to generate the load and to log metrics during the test. Depending on the framework, load can be generated from one or several hosts, e.g. [10, 9, 11]. Most IP-based frameworks automatically log metrics like response times for the issued requests. Additionally, some of them incorporate ways to also log resource usage on the tested systems, e.g. [10, 9, 13].

For analyzing the results of performance tests with the aim to automatically detect performance regressions, several methods have been proposed. One common method is the use of control charts [17, 18, 19]. However, control charts assume normal distributions for the measured values, which is usually not the case for performance counters like CPU usage under varying load. Another category of approaches exploits the fact that several performance counters in a test run are usually correlated and changes in these correlations could indicate a performance regression. Moreover, correlation might be used to reduce the amount of counters that needs to be analyzed. Foo et al. [20] and Žaleźničenka [21] implement this approach by applying association rule learning techniques while Shang et al. [22] present a method based on clustering and regression. Additionally, Malik et al. [17] present another clustering and a PCA-based approach.

Generally, the existing work mostly focuses on performance testing integrated systems. While such tests are also desirable for robotics and intelligent systems, they are much harder to set up and maintain due to the complex interactions of robots with the real world and the non-standard interfaces in contrast to e.g. HTTP. Moreover, performance regressions detected in such integrated tests are harder to point down to individual components for fixing the issues. Therefore, testing performance for individual components provides a parallel, and currently better applicable method in robotics and intelligent systems.

Finally, in addition to the presented data-driven methods for detecting performance

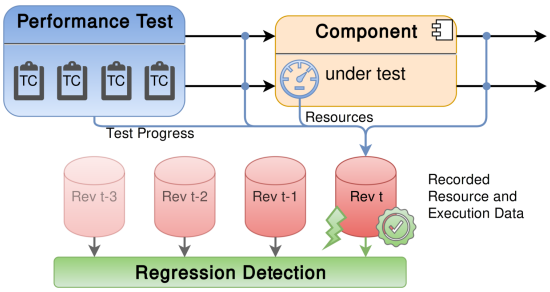

Figure 1. Visualization of the general performance testing concept proposed here.

regressions, there is also research on predicting resource consumption based on software models. For instance, Becker et al. [23] present the quite popular Palladio Component Model, which allows to model complete software architectures with respect to performance-relevant aspects. Comparable approaches have been reviewed in Koziolek [24]. In case a complete model of the system exists, performance regressions can be derived from the model. However, these models often do not exist or are out of sync with the real system. Additionally, specifying the resource usage of a manually implemented or third-party component is a non-trivial task and therefore, models often present a level of abstraction that might miss certain performance degradations that are practically important or noticeable.

3. Concept

Most current robotics systems are constructed as distributed component-based systems where individual components implement isolated aspects of the system functionality [25]. A component is thereby a specification of an interface of how potentially multiple different implementations interact with the rest of the software system. While the interface usually remains relatively stable, the implementation might change frequently and therefore also its performance characteristics. For our case, we assume that a component's interface is defined and realized in terms of communication patterns [26] and data types of a middleware framework like ROS or YARP. With the ongoing adoption of a limited set of such frameworks, a reasonable part of components is reused across different systems. Additionally, current systems are often maintained by multiple persons and no single person has in-depth knowledge of all components which form the system. Therefore, we think that it is viable to implement performance testing on a per-component basis instead of, or in addition to, the complete system because:

(1) Testing a complete robotics system for performance regressions in an automated fashion is hard to achieve as robotics systems interact with the real world, e.g. via speech-based dialog or computer vision algorithms. Inputs for these interactions would need to be simulated or pre-recorded and special interfaces to interact with a simulation or recorded data chunks are then required, which is only sometimes the case.
(2) The middleware-based component interface allows to write performance tests that are quite stable during component and system evolution. While changes to interface might happen, these should be infrequent as otherwise the integration of systems which use a respective component would be impacted.

(3) Detected performance regressions can be attributed easily to the component as a code unit so that extensive searches for the origin of a regression can be avoided.
(4) Component developers have the best knowledge about their components and the expected loads and behaviors. Therefore, developers can test the complete range of functionality and loads and not only the requirements of a single target system. Moreover, it is also much easier to explore the space of potential loads on a component under test as the effective middleware inputs on the component do not need to be generated through several layers formed by the surrounding integrated system and ultimately the interactions of the system with the real world. Therefore, this possibility for a systematic exploration increases the test coverage with respect to the individual components and ensures that important situations which might trigger exceptional resource consumption patterns of the component are covered by the tests.
(5) Components usually survive individual robotics systems and might be used in several systems in parallel. Isolated performance tests per component ensure that the test suite does not need to be rebuilt with each new application and test results are continuously comparable despite changing application areas of the components.

Therefore we present a framework to specify, execute and analyze performance tests for individual components of robotics systems. The general concept of this framework is visualized in Figure 1. The middleware inputs of a component under test are replace with inputs generated using the testing framework and the component is instrumented to acquire resource utilization information. The whole test progression including consisting of testing meta data, component communication and performance counter is recorded and stored. A regression detection component of our framework uses the stored data from different component revisions to decide whether the performance of the component has changed recently.

The actual test cases constructed with our framework are maintained alongside the component in a similar fashion to unit tests[1]. This ensures that tests are kept up to date with the component by the component developers and test results are immediately available after component changes.

Depending on the connectivity of a component with the remaining system or the underlying operating system and hardware, testing via the middleware interface can be more or less complicated. E.g. a controlling state machine usually communicates with many other components in the system. Therefore, it is hard to test it in isolation. While it is possible to test such a component (e.g. by implementing mock components for the tests), our approach primarily targets what Brugali and Scandurra [25] call "vertical components", which capture isolated domain knowledge in a functional area and contribute most to reuse.

4. Realization

In the following subsections we describe the realization of our approach. According to Malik et al. [17], a common load or performance test (terms are often used interchangeably [8]) consists of "a) test environment setup, b) load generation, c) load test execution, and d) load test analysis". We agree with this and the following descriptions of our framework follow this separation (with a changed order). The framework has

[1]Either inside the component's source code repository or in a closely coupled one.

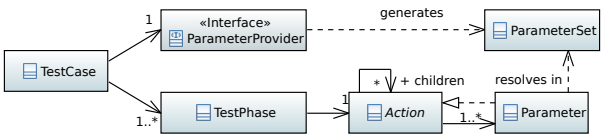

Figure 2. Structure of the testing API.

been developed using the RSB middleware [27], but the concepts can easily be applied for other common frameworks like ROS or YARP, as long as tools exist to record and replay communication.

4.1. *Load generation*

For vertical components, we assume that the resource demands of the component at runtime are to a large extent related to the communication a component is exposed to. For instance, a face detection algorithm might impose higher or lower CPU usage depending on the rate and size of images it received via the middleware. Similarly, a person tracker's CPU and memory usage might relate to the number of person percepts it receives via its communication channels. Therefore, generating load in terms of middleware communication provides a way to abstract the ongoing development changes inside the component while enabling to test the aspects of components that are relevant to its use inside a robotics system.

In order to generate such middleware-defined loads we have implemented a testing framework which allows to specify middleware interactions with components via a Java API. We have chosen Java as our target language as it provides the required performance to generate heavy loads (e.g. in contrast to Python) while providing a relatively easy to use programming language and environment (e.g. compared to C++) which is usable for most developers.

Inside this framework we define a performance test to consist of multiple *test cases* (cf. Figure 2. Each test case consists of one or more *test phases* and a *parameter provider*. A test phase is a named entity which consists of a tree of parameterized *actions* to perform via the middleware which specify the actual interaction of the test with the component. These actions have variable *parameters*, like e.g. a sending rate for messages or a number of faces to include in a face detection message. The parameter provider generates *parameter sets* which specify the actual values for all variable parameters. When executing the test case, the action tree is executed sequentially for all parameter sets, thereby generating different load levels on the component.

Parameters allow to specify the load profile of tests, while actions specify the semantics of the interaction. Such parameters can for instance be:

- communication rates
- number of generated messages
- different data sizes
- sets of pre-computed control/data messages
- middleware communication channels

By extracting these parameters from the actual definition of the interaction we gain several benefits:

- The influence of these parameters on the component can be systematically analyzed.

Table 1. Identified actions for constructing performance tests.

Data	
Parameter	Resolve a value from the current parameter set
StaticData	Resolve to a pre-defined, static value
Flow	
Sequence	Execute multiple actions sequentially
Loop	Loop an action n times or indefinitely
Parallel	Execute multiple actions in parallel
WithBackground	Execute one main action with multiple background actions. Background is interrupted when the main action finishes.
Timing	
Sleep	Sleep for a specified time
LimitedTime	Execute an action up to a time limit and interrupt it after this time
FixedRate	Execute an action at a fixed rate for a specified amount of time
Middleware (RSB)	
InformerAction	Send an RSB event (message)
RpcAction	Call an RSB RPC server method and optionally wait for the reply
WaitEvent	Wait for an event to arrive
BagAction	Replay pre-recorded RSB communication from a file
DynamicEvent	Construct an event (for Informer or RPC action)
ProtobufData	Generate protocol buffers event payloads from parameters

- By changing parameter sets, different test granularities can be achieved.
- Test cases can be reused across different, functionally comparable components by changing parameter sets.

Test phases provide the ability to group operations to perform with the component under test, which are identifiable for a later analysis step. As test phases are executed sequentially inside each test case for all parameter combinations, this allows to define transactions in case a component like a database requires a defined protocol.

The actions that can be performed in order to interact with the component under test form a limited specification language suitable for the needs of performance testing. Each action generally is a function that takes the current parameter set as the input and optionally returns a result, which may be processed by parent actions. We have identified and implemented the actions shown in Table 1 as a result of testing our own components. As visible in Figure 2, `Parameter`s are also instances of `Action`, as they resolve the current actual parameter value from the current parameter set, which matches the definition of an action as a function that returns a result. In addition to variables parameters, `StaticData` provides a means to specify data which will not change across test executions. We have provided specific support for generating variable data for our middleware based on the parameters as this is a very common task. The `ProtobufData` action allows to construct Protocol Buffers [28] data (which is primarily used in RSB) from template messages by scaling (repeated and string) fields based on parameters. For this purpose, API users provide data generators for the individual items. Additionally, the `BagAction` provides a method to replay pre-recorded data, optionally with modulations like speed or channel selection. If a user requires further actions or methods for generating test data, the action tree can be extended with custom implementations. Performance tests are created by forming a tree of these actions inside different test phases and test cases. Figure 3 visualizes the action trees that have been used to construct a test for a leg detector component.

Each test case is equipped with a parameter provider which generates one or more sets of parameter combinations. Each parameter itself is a programming language

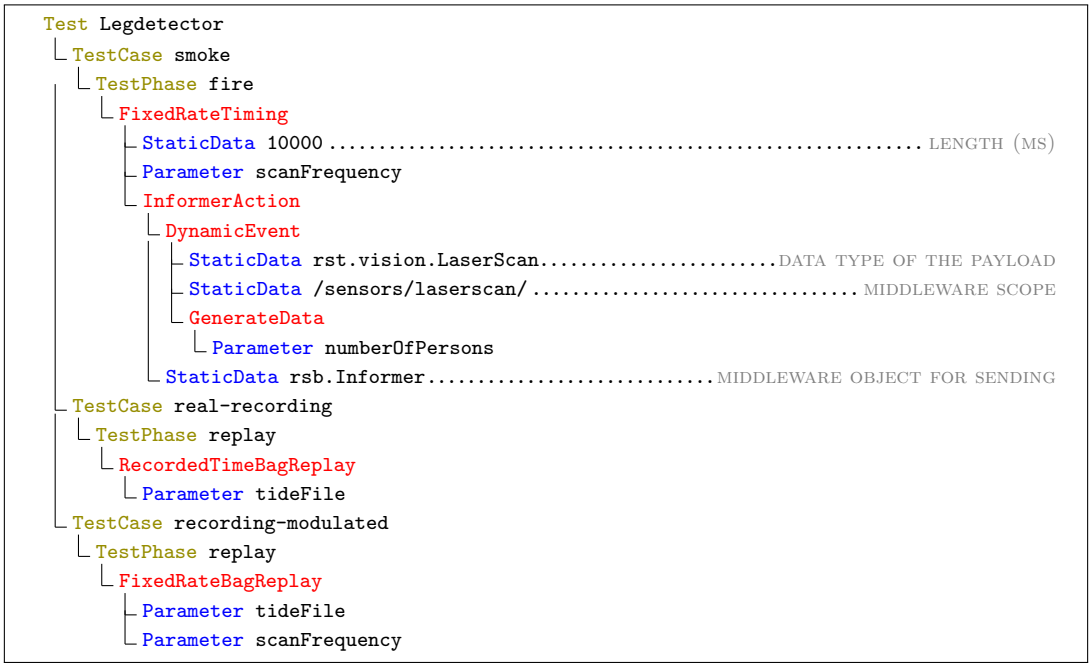

Figure 3. Structure of a performance test for a component which detects legs in laser scanner readings. The test consists of three distinct test cases, each with a single test phases. The general test case and phases structure is marked in olive green, actions are marked in red and data-related actions are marked in blue.

object and has a printable name for the analysis. We provide two implementations of parameter providers: a table, where the user manually specifies the row values, and a Cartesian product, where combinations of individual parameter values are created, optionally with constraints. For easily specifying the constraints, scripting languages like Groovy can be used. Since parameters will be reported during test execution using the middleware, they must be serializable by the middleware.

4.2. *Environment setup*

For executing performance tests, the API contains a test runner. The first step of this test runner is to set up the test environment based on a configuration file, which specifies the following aspects:

- locations of utility programs required by the testing API
- middleware configuration
- processes which act as a test fixture (e.g. daemons, mock components)
- component processes to test
- test cases to execute and their parameter providers (references to Java classes)

Using this configuration, the initialization of the test environment is performed in the following steps:

(1) Configuration of the middleware to ensure that test execution is isolated from the remaining system.
(2) Creation of a temporary workspace for the test execution. The workspace is used as the working directory for executed processes and stores intermediate logs which can be retained for debugging.
(3) Start of all defined fixture processes. These could be daemons required for the

middleware, database services used by the tested component etc.

4.3. *Test execution*

See Figure 4 for a visualization of the following aspects.

4.3.1. Orchestration

After the environment setup, the configuration is used to instantiate and execute the performance test. First, the configured test case and parameter provider instances are created and a static validation for invalid parameter references is performed. If validation succeeds, the defined components to test are started. While usually only a single component is started, it is also possible to test a combination of components in cases where such a constellation is easier to test than an isolated component due to the required interactions. After starting all components, the test cases are executed sequentially. Inside each test case, the defined test phases are executed for all parameter sets returned by the parameter provider. Finally, all started components and the test fixture are terminated. We have decided to use a single execution of the component processes without intermediate restarts, e.g. for each parameter set. On the one hand, test runs require more time with component restarts and on the other hand, most robotics components usually operate for a longer time without restarting and artificial restarts would make it harder to detect performance issues like memory leaks, which slowly build up over time.

4.3.2. Data Acquisition & Recording

In order to generate and record data for later performance analysis, we apply the following approach: The test runner instantiates an external monitoring process, which obtains performance counters (or KPIs) for each of the started component processes by using the Linux `proc` filesystem, which includes aspects like CPU, memory, I/O and threads. The monitor is implemented as efficient as possible in order to minimize the load it additionally imposes on the system (e.g. below 2 % CPU usage per process on an Intel Xeon E5-1620). Acquired counters are exposed via the middleware using dedicated channels per component. In addition, the test runner exposes information about the executed test cases, phases and parameter sets via dedicated middleware messages, so that the counters can be related to this structure. For persisting the generated data, the test runner launches an instance of the middleware recording tool, in our case `rsbag`, which is configured to record the entire communication. This way, recording results can be replayed completely for detailed analysis and debugging purposes. The recording method has previously been described in Wienke et al. [4], including details about the performance counters. Figure 5 shows an excerpt of a test case recording for a single test case with two test phases which are executed for different parameters.

4.4. *Test analysis*

4.4.1. Data Preparation

The output of a performance test is a file with all middleware events including the component communication, information about the test progress, and performance counters

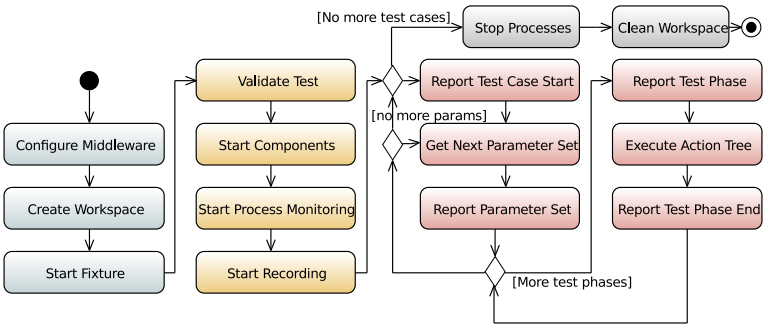

Figure 4. Visualization of steps performed to execute a performance test. Cyan: environment setup, orange: test setup, red: test case execution, gray: clean up.

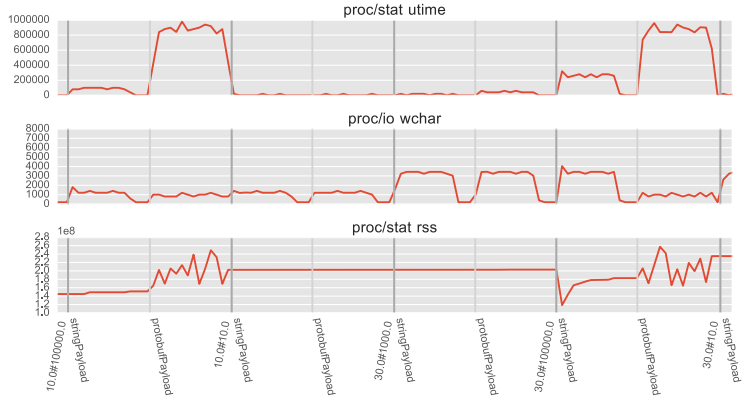

Figure 5. Excerpt from a test for a logging component. Two test phases are executed for different parameter combinations, in this case frequency and size of events to display by the logger.

for the tested component. While this provides a good basis for detailed analyses, the file size usually prohibits to store these files for a longer time to build a database of the performance changes inside the component and random access times are slower compared to other formats. Therefore, we first transform information about the test progress and counters into a HDF5 file using the Python pandas library [29]. Here, information about the represented component revision are attached to the data, which are a human readable title (e.g. a Git hash for tests per Git commit or a time stamp for nightly builds) and a machine-sortable representation (e.g. the Git commit date or an ISO 8601 formatted date) so that executions can be ordered accordingly. In case a test has been executed multiple times for the same component revision, it is assumed that title and sort representation have the same value for all revisions (a test execution date is added to the data automatically and allows to later distinguish between test executions). These HDF5 files are the artifacts which are usually persisted for each test execution. For this purpose, a command line analysis tools was implemented, which realizes the complete analysis step. It is designed to be integrated into shell scripts, e.g. for a CI server integration.

4.4.2. Plots

As a first means of manually inspecting the performance of a component, the analysis tool allows to generate several plots from recorded data. These include the raw performance counter time series of a single execution, correlations between counters and

numeric test parameters and several plots which show how performance counters have evolved with component revisions. For this purpose, counters are summarized for each test case, test phase and parameter set via mean and standard deviation and plotted for each revision of the component. This allows to track how the usage of individual counters has evolved. Figure 5 shows an excerpt from one of the generated plots.

4.4.3. Automatic Regression Detection

To detect changes in the resource consumptions of a component, we have implemented three different methods in the analysis tool. All methods take one or more test execution results (as HDF5 files) and compare the observed performance against a baseline from one or more test executions. We do not enforce a method of defining executions as baseline and test data, but at least the following modes can be found in the literature:

- "no-worse-than-before" principle [8]: the current revision is compared to the previous one (or a window of n previous revisions) to ensure that the current state is at least as good as the previous one.
- Comparison to a hand-selected baseline: all revisions can be compared to a manually selected baseline to ensure that the criteria of this baseline are met. The baseline needs to be reselected to match intended performance changes.

These modes can easily be realized using the analysis tool using different HDF5 files.

Most existing methods for detecting performance regressions actually detect any change in the performance characteristics of the tested system. While an automatic categorization whether a change is a degradation or an improvement in the performance would be a desirable feature, this is often not easily possible. In our own experiments we found cases where e.g. a new component revision resulted in a higher CPU usage for small workloads while the usage was improved for higher workloads. Therefore, we have implemented a mode where any change is reported and the developer has to decide (e.g. based on the plots), whether the change is acceptable.

For the actual detection of performance changes we have implemented the following methods:

- The method proposed by Foo et al. [20] as an example for an association rule learning based approach.
- The method proposed by Shang et al. [22] as a reference for a recent method based on clustering and regression.
- A basic two sample Kolmogorov-Smirnov test for each performance counter. For this purpose, the individual measurements of each performance counter across the whole test execution time (of potentially multiple executions) are assumed to form an observation and the observation from the test executions is compared to the observation from the baseline.

4.4.4. Automation

The analysis tool reports results using JUnit XML files, which can be parsed by many automation tools, e.g. the Jenkins CI server. This allows to integrate the approach with such tools which can then automatically give feedback on potential performance regressions. Figure 6 visualizes how performance testing can be automated inside Jenkins. For each new software revisions, a Jenkins build executes the performances tests, loads previous testing data (HDF5) from the job artifacts storage of Jenkins, and performs the regression detection. The resulting JUnit XML file is parsed by Jenkins using

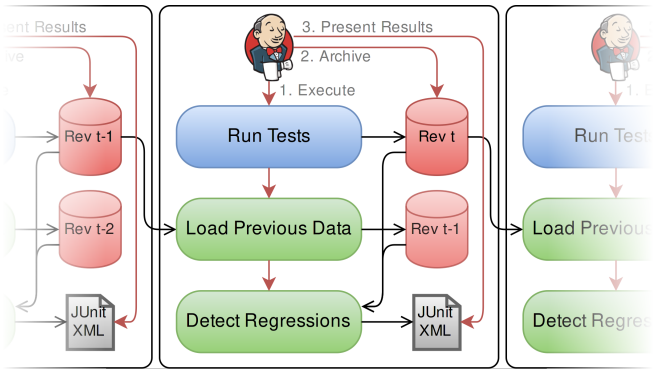

Figure 6. Integration of the testing framework in the Jenkins CI server. Each black rounded box represent a single Jenkins build and therefore execution of the tests for a component. Time and component versions progress from left to right.

one of the available plugins and in case of errors, developers are notified, e.g. via mail. Since HDF5 files from previous test executions are required to perform the analysis, we have provided a utility command which allows to download artifacts from previous runs using the Jenkins API. This Jenkins integration is an easy way to trigger a test execution for each new software revision with automatic notification. In the future, more sophisticated tools like *LNT* [30] might be required to archive test results and make them browsable.

Generally, acquired performance results are coupled to the execution platform due to the system-specific execution times and measurements. Therefore, a dedicated host should be used for all test executions and this host should be free from other tasks to avoid resource sharing issues, which might influence the measurements. For instance, this can be realized by adding a dedicated performance testing slave to a Jenkins server. Also, dynamic frequency scaling techniques might influence the results. We therefore advise to enable the Linux `performance` CPU governor, at least for the test runtime.

5. Evaluation

In order to validate the automatic detection of performance changes, we have implemented performance tests for several vertical components from our systems, which cover a range of different programming environments:

- `2dmap`: A Java-based visualization for person tracking results in a smart environment.
- `legdetector`: A Java-based component for detecting legs in laser scans, used by our mobile robots.
- `objectbuilder`: A C++-based component which generates stable person hypotheses from detected legs and the SLAM position of a robot.
- `logger-*`: A console-based logger for middleware events with different output styles (Common Lisp).
- `bridge`: An infrastructure component which routes parts of the middleware communication to other networks (Common Lisp).

For all of these components, tests have been written using the presented Java API and results have been processed using the aforementioned analysis methods. All tests

Table 2. Available evaluation data

	revisions	execs	changes
2dmap	25	4	10
legdetector	14	4	4
objectbuilder	23	4	7
logger-compact	306	2	16
logger-detailed	306	2	7
logger-monitor	228	2	6
bridge	176	2	6

could be generated with the provided actions which suggests that the provided set of actions is generally sufficient for writing tests for vertical components.

We have tested the presented components using the "no-worse-than-before" principle by comparing each revision against the previous one, as this is automatically possible without the necessity for a manual baseline selection. For the `2dmap`, `legdetector` and `objectbuilder` component all Git commits that could still be compiled have been used while for the `logger` and `bridge` archived component nightly builds were used.

All available analysis methods have been applied to compare their performance. Each of them requires a threshold for a numeric score to decide whether a test execution represents a performance change or not. We have used this score to compute the Area Under Curve (AUC) on ROC curves as the target metric for the classification. Since all analysis methods actually return multiple scores per test (Shang et al. [22] returns one score per cluster, Foo et al. [20] returns one score per frequent item set, and the KS-test returns one score per performance counter), these individual scores need to be combined into a single one to enable computing the AUC. We have used the min, max and mean functions for this purpose.

To get ground truth information, the generated plots displaying the evolution of performance counters across revisions have been manually examined and annotated. Additionally, the commit logs have been used, especially in cases where a decision was not easily possible from the representation. While we took great care with the annotations, we still expect some amount of errors since it is sometimes very hard to decide whether visible changes are real performance changes or caused by possible external disturbances. Especially for the nightly builds, the Git commit log was not sufficient to trace all possible changes, e.g. to the compilation environment used to create each build. These annotation issues are expected to decrease the AUC scores. Table 2 displays the amount of available data per component. The column "execs" indicated how often the test has been executed per revision and "changes" shows how many revisions have been manually tagged to contain performance changes.

Based on the available data we receive the evaluation results visible in Table 3. The highest scores per component are highlighted. These show that for component tests the basic Kolmogorov-Smirnov test works best. Only for some settings the method by Shang et al. shows comparable or slightly better scores. Especially the method proposed by Foo et al. does not seem to work on this kind of data.

For the results shown in Table 3, the tests have been executed multiple times (cf. Table 2 for the actual numbers). This has been done, because several aspects of the component performance characteristics differ across runs of the same revisions. For instance, due to garbage collection timing in Java or Common Lisp programs, the memory footprint might be different across runs. Generally, we have observed that memory is usually one of the most common causes for false-positives due to such issues and averaging across multiple runs provides a way to counteract this. In contrast to

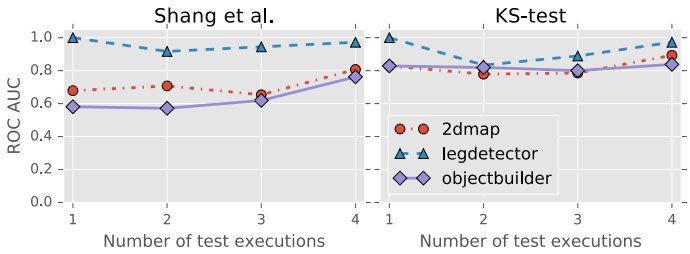

Figure 7. Influence of the number of test executions on the two most promising detection methods. For Shang et al. the max aggregation method was used and for the KS-test the mean of all counter scores.

Table 3. ROC-AUC scores for the different analysis methods

	Foo et al.			Shang et al.			KS-test		
	min	max	ø	min	max	ø	min	max	ø
2dmap	0.50	0.72	0.71	0.81	0.81	0.83	0.50	0.50	**0.89**
legdetector	0.50	0.51	0.47	0.75	**0.97**	**0.97**	0.50	0.50	**0.97**
objectbuilder	0.50	0.63	0.66	0.28	0.76	0.55	0.50	0.50	**0.84**
logger-compact	0.50	0.49	0.51	0.45	0.56	0.48	0.59	0.59	**0.76**
logger-detailed	0.50	0.39	0.39	0.46	0.60	0.57	0.61	0.50	**0.84**
logger-monitor	0.50	0.57	0.57	0.69	**0.82**	0.80	0.48	0.50	0.72
bridge	0.50	0.59	0.59	0.55	0.56	0.48	0.44	0.49	**0.61**
ø	0.50	0.56	0.56	0.57	0.73	0.67	0.52	0.51	**0.81**

component restarts during test execution (e.g. for each parameter set and test phase) this is still faster to perform due to less restarts while retaining the ability to detect performance issues like memory leaks. Additionally, effects of component initialization (warming up caches, loading files and libraries etc.) are less visible in the data. To quantify the effect of the number of test executions on the detection of performance changes, we have varied the amount of executions for all components that have been tested four times in total. Figure 7 shows the results for the most promising detection methods. While there seems to be a slight improvement for Shang et al. with the number of test executions, the results for the KS-test are inconclusive. On the other hand, both methods already show a reasonable performance with a single execution of the tests.

6. Learning runtime resource utilization models from performance tests

Performance tests that are specified using the provided primitives of the framework usually explore the intended variability of input parameters in a systematic way to test different load situations for a component. While this ensures that the test coverage is high and also exceptional situations are covered from a testing point of view, the systematic exploration also presents an opportunity for a runtime verification of component resource utilization patterns. In Wienke and Wrede [31] we have demonstrated that it is possible to detect runtime faults in individual components of robotics systems by learning a model of a component's resource utilization from exemplary training data. These models use the event-based communication of the component as features and predict the resource utilization by applying machine learning to train regression models. A residual-based fault detection approach is then applied to detect deviations of the actual resource utilization compared to the predicted values. For this to function properly, a reasonable accuracy of the prediction model for the resource utilization is

required. In our previous work, we have used complete scenario executions of the target system for training the resource prediction models, which means that an operator had to use the robot system in its intended scenario in order to implicitly explore all possible loads on the contained software components. This training procedure requires significant manual work and needs to be redone each time a software component of the system has changed. With the performance tests created using our framework and the test driver recording the complete middleware communication, we have an alternative source for training data at hand, which is not coupled to individual systems. This data can be regenerated automatically and in a more targeted fashion (e.g. for individual components in case of changes), which would result in a major reduction of manual work required to set up a runtime fault detection approach if it was suitable for training resource prediction models. In the following, we therefore explore whether resource prediction models with a reasonable performance can be trained from performance test data.

6.1. *Prediction task and method*

Along the lines of the work presented in Wienke and Wrede [31], the actual prediction task for each component is defined as:

$$r(f_t) : F \mapsto P \tag{1}$$

where $F := \{f_t : t \in T\}$ represents an encoding of the event-based communication that a component has received and emitted up to the current time t with individual feature vectors f_t and $P := \{p_t : t \in T\}$ represent the resource utilization values at each point t in time. Therefore, the prediction model always predicts the current resources given the encoded communication. Please refer to the cited work for a detailed description of the generic feature extraction and synchronization method that is used for our RSB-based systems. In contrast to our original work, the regressions method used here is kernel ridge regression with an RBF kernel and a feature extraction based on agglomerative clustering. The target dimensionality of the input space as well as the RBF parameters are optimized using a grid search with cross validation.

6.2. *Analysis setup*

For the following analyses we specifically focus on the CPU usage of a component, as this is usually the performance counter with the highest variance and therefore is the least predictable with constant models. We have chosen the following setup to analyze the ability to learn resource utilization prediction models: A mock component has been implemented, which uses the same middleware interface as a data fusion component from our mobile robot system *ToBi*, which has the purpose of generating person hypotheses in a global coordinate system based on detected legs in laser scanners results and the current position of the robot, which was determined via SLAM. This component is part of a fault detection data set [4] which has been used for our previous work. It contains actual recordings of the execution of the mobile robot system and therefore represents realistic data. The mock component was implemented to systematically test out different component configurations. It uses a thread pool to process the received leg detections and a single worker thread to process the received SLAM position results. The amount of work performed by the leg detection tasks

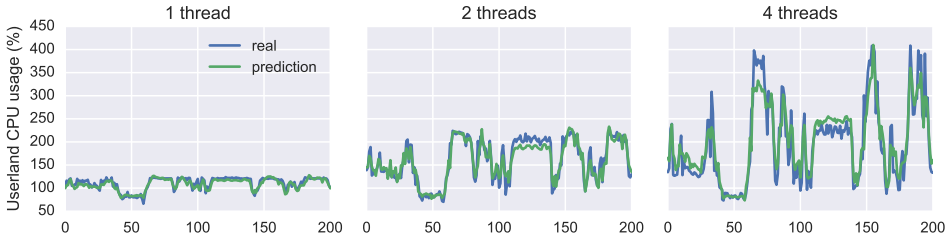

Figure 8. Actual and predicted CPU usage of the mock component in different configurations of the leg detection thread pool on the same data. The x-axis indicates time in seconds.

scales quadratically with the number of legs to simulate non-linearity in the processing. The size of the leg detection thread pool can be defined at component start, which allows to switch between different kinds of CPU usage patterns which we have found in our own systems. Figure 8 visualizes the CPU usage behavior of the component in the three configurations that we have tested (1, 2 and 4 threads) for the same input data, which was drawn from the fault detection data set. While with one thread the component most of the time completely utilizes a single CPU core of the test system (plus an overhead from the pose thread operating in parallel), the other conditions gradually result in less exhaustion of the assigned CPU cores. Components resembling both ends of this spectrum can be found in our reference system from the data set.

Regarding the input data the mock component receives, we have identified three parameters that we have varied to explore the impact of different versions of the performance test. For each of the parameters we have selected three testing conditions. These are in detail:

Leg number the number of detected legs contained in the laser scans.
- *d*: Match the **d**ata set and produce 1 to 22 detected legs in steps of 3.
- *l*: Produce **l**ess data by sending 1 to 13 detected legs in steps of 3 to test extrapolation.
- *m*: Produce **m**ore data by sending 1 to 28 detected legs in steps of 3 to test the influence of additional training data.

Leg rate the rate at which new laser scanner results are generated.
- *d*: Match the **d**ata set and produce leg detection results at 30 Hz.
- *r*: Use a **r**ange of different production rates, but not exactly the target rate from the test data, i.e. 10, 20, 40 and 50 Hz to test interpolation.
- *dr*: Use a **r**ange of production rates including the **d**ata set target rate, i.e. 10, 20, 30, 40 and 50 Hz.

Pose rate the rate at which new SLAM position are generated. Similar conditions:
- *d*: 10 Hz
- *r*: 2.5, 5, 15 and 20 Hz
- *dr*: 2.5, 5, 10, 15 and 20 Hz

For each thread configuration of the mock component we have trained a resource prediction model for all 27 combinations of the test parameters. These models were then used to predict the performance counters based on communication data from the data set and the root mean square error (RMSE) was computed. The RMSE is expressed in percent of a single CPU core of the test system, which was a Linux-based desktop computer with 8 CPU cores. This prediction was additionally compared to the baseline of predicting the mean value of the test data, which results in a value

Table 4. Prediction errors for different mock component configurations (number of threads).

	ø		min		l dr dr		data set	
	RMSE	%[a]	RMSE	%	RMSE	%	RMSE	%
1	10.2	54.6	7.3	39.0	7.4	39.5	6.2	33.1
2	19.7	40.0	12.6	25.6	14.7	29.8	11.9	24.1
4	44.5	49.7	26.5	29.6	28.3	31.7	21.2	23.7
ø	24.8	48.0	15.5	31.4	16.8	33.7	13.1	26.9

[a]error in % of the baseline approach

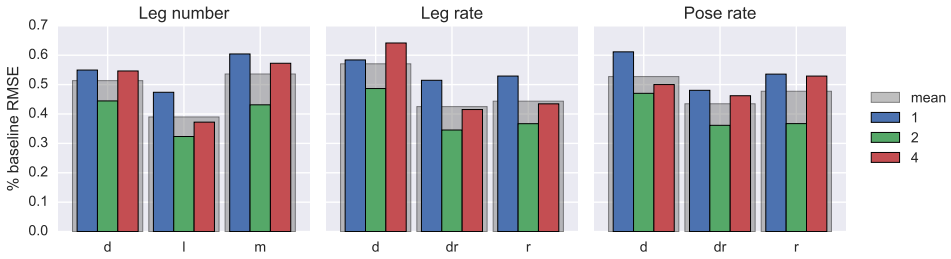

Figure 9. Influence of the regression test parameters on the resulting prediction performance expressed in % of the baseline RMSE when predicting the data set inputs. Colors indicate the different thread configurations of the mock component.

expressing the prediction error in percent of the baseline error for easier interpretation.

6.3. *Results*

Table 4 presents the results of the conducted experiments. As depicted in the first block of columns (*ø*), on average across all different performance test parameters and component configurations, the prediction error is a quarter of a single CPU core or 48 % of the baseline error. However, the best results created by training from the performance test go down to 16 % of a single CPU core or 31 % of the baseline error (cf. second column block). Moreover, the RMSE scales approximately linearly with the threads of the component, which is expected due to the higher amplitude of the CPU usage counter. Therefore, we conclude that all three configurations can be learned equally well by the approach despite different resource utilization characteristics. Finally, the last block of columns in Table 4 (*data set*) shows the errors produced by the same approach when being trained on actual execution data taken from the data set, but distinct from the testing data. Compared to the best training of the approach from performance tests (column block *min*), there is only slight increase in the average RMSE by 2.4 %, which is very small given the fact that the component usually consumes at least 100 % of a single CPU. For a visual impression of the prediction results, Figure 8 displays them compared to the actual resource utilization for a fraction of the training data. Generally, we conclude that the approach is able to learn a valid prediction model for the resource utilization of the component.

To understand the impact of the different choices for parameters of the performance tests, Figure 9 shows plots of the average normalized prediction errors for each test parameter and condition. For both rate parameters it is visible that the prediction improves in case a range of different event rates is used in the tests, ideally including the target event rate of the intended application scenario. In case of the leg number, the condition with a smaller range results in better scores, while the other two conditions

result in higher errors. This might be the case, because while up to 20 legs can be found in the data set, the media of the data is 6 and higher number of legs are only observed in very few situations. Therefore, the additional training data is not relevant to most of the test data and only complicates learning the important part. The third block of columns in Table 4 shows the prediction errors when selecting the performance test for training with the condition which follows these insights. While this does not create the best results for all three component conditions, results are still close to the best conditions.

In line with our previous work, the learning approach presented here does not cover resource sharing issues and therefore assumes that the host system is not already blocked by other computation tasks. Further work which includes information about the general load on the system in the models is required. Also, models are coupled to the computers the training has been performed on. However, this is only a minor issue as it is easily possible to re-run the performance tests on new hardware to train new models.

7. Discussion

We have presented a method to test individual components of robotics and intelligent systems for performance regressions introduced by code changes. The implemented framework consists of a Java API to specify tests which operate on the component's middleware interface. Special care has been taken to provide abstractions suitable for vertical components and the required data generation tasks. After test executions an analysis tool automatically decides whether a new test execution shows performance changes compared to a baseline from previous executions. By specifically supporting automation systems like the Jenkins CI server, this allows to easily construct an automated verification of performance aspects. Consequently, developers are better informed about the impact of their changes on the performance characteristics. We have verified the applicability of the concepts defined in the testing API and the effectiveness of the automatic performance change detection based on components drawn from different robotics and intelligent systems which are in active use at our labs. The idea of testing individual components for performance characteristics is a new perspective which is easier to apply than testing the whole system in the robotics context. Detected performance regression can easily be attributed to individual code units. Nevertheless, in the future, work on testing complete systems is an additional axis that needs to be performed to detect further categories of performance regressions.

The presented testing API is eventually meant only as a base tool for specifying performance tests. Java was chosen as a compromise between an acceptable coding experience and the required efficiency to generate load tests. However, the syntax is verbose and requires many code-level constructs for specifying semantic tasks. Therefore, we envision the use of a suitable DSL with embedded scripts for specifying performance tests in a more natural and readable way.

Our current implementation of the framework uses the RSB middleware as its basis. However, the concepts are generally applicable for comparable middlewares. In the testing API, the RSB-related methods are already separated and can be exchanged with other backends. A similar separation is possible also for the analysis tool.

In addition to the static testing perspective, we have also shown that performance tests can be exploited to train resource prediction models for components, which serves as a basis for runtime resource-related fault detection. Training models from perfor-

mance tests makes this an automatic process, which reduced manual labor which would otherwise be needed to execute a target system for training. Additionally, it enables retraining of individual model in a case a component changed instead of requiring a complete system re-execution. This greatly reduces the overhead necessary to adopt runtime fault detection techniques.

In the future, we will focus on further improving the detection of performance regressions. One aspect commonly found in load tests for large-scale websites are explicit warm up or tickle loads which are ignored in the final analysis and shall reduce the impact of system initialization. It is currently already possible to manually add test cases or test phases to achieve a similar behavior, but adding such a concept as a first class citizen will ensure that users of the API are aware of the issue. Despite having ignored these effects in the current evaluation, the presented detection scores already show that the system is usable. Another issue we have observed is the discretization of rarely used resources by the Linux kernel. This results in peaks in the resource usage at unpredictable times during the tests and some of the presented detection methods are sensitive to such peaks. We will improve on this in future versions. With respect to the runtime detection of faults, further experiments with a larger corpus of different components need to be performed in order to generalize the findings. Despite the potential for improvement, the framework already provides a good foundation to detect performance degradations and has helped to identify several previously unknown regressions in our own components. Moreover, training runtime resource prediction models from performance tests is a novel perspective on improving the dependability of robotics systems and we are not aware of comparable approaches.

Acknowledgments

This work was funded as part of the Cluster of Excellence Cognitive Interaction Technology 'CITEC' (EXC 277), Bielefeld University and by the German Federal Ministry of Education and Research (BMBF) within the Leading-Edge Cluster Competition "it's OWL" (intelligent technical systems OstWestfalenLippe) and managed by the Project Management Agency Karlsruhe (PTKA).

References

[1] Quigley M, Gerkey B, Conley K, et al. ROS: an open-source Robot Operating System. In: ICRA Workshop on Open Source Software; 2009.

[2] Metta G, Fitzpatrick P, and Natale L. YARP: yet another robot platform. Journal on Advanced Robotics 2006;3(1):43–48.

[3] Steinbauer G. A Survey about Faults of Robots Used in RoboCup. In: Chen X, Stone P, Sucar LE, et al., editors. RoboCup 2012: Robot Soccer World Cup XVI; Berlin, Heidelberg: Springer; 2013, p. 344–355.

[4] Wienke J, Meyer zu Borgsen S, and Wrede S. A Data Set for Fault Detection Research on Component-Based Robotic Systems. In: Alboul L, Damian D, and Aitken JM, editors. Towards Autonomous Robotic Systems; Springer International Publishing; 2016, p. 339–350.

[5] Fowler M. Continuous Integration. 2006. [cited 2016 Aug 30]. Available from: `http://martinfowler.com/articles/continuousIntegration.html`.

[6] Sydor MJ. APM Best Practices. Realizing Application Performance Management. Berkeley, CA: Apress; 2011. 486 pages.

[7] Molyneaux I. The art of application performance testing. from strategy to tools. 2nd edition. Theory in practice. Sebastopol, CA: O'Reilly; 2015. 255 pages.

[8] Jiang ZM and Hassan AE. A Survey on Load Testing of Large-Scale Software Systems. IEEE Transactions on Software Engineering 2015;41(11):1091–1118.

[9] Apache JMeter [computer software]. [cited 2016 Aug 8]. Available from: `https://jmeter.apache.org/`.

[10] Tsung [computer software]. [cited 2016 Aug 23]. Available from: `http://tsung.erlang-projects.org/`.

[11] Locust [computer software]. [cited 2016 Aug 30]. Available from: `http://locust.io`.

[12] NLoad [computer software]. [cited 2016 Aug 23]. Available from: `http://www.nload.io`.

[13] The Grinder [computer software]. [cited 2016 Aug 23]. Available from: `http://grinder.sourceforge.net`.

[14] Chen S, Moreland D, Nepal S, et al. Yet Another Performance Testing Framework. In: Hussain FK, editor. 19th Australian Conference on Software Engineering (ASWEC 2008); Los Alamitos, Calif.: IEEE Computer Soc; 2008, p. 170–179.

[15] Gatling [computer software]. [cited 2016 Aug 30]. Available from: `http://gatling.io`.

[16] Da Silveira MB. Canopus: A Domain-Specific Language for Modeling Performance Testing. PhD thesis. Porto Alegre, Brasil: Pontifical Catholic University of Rio Grande do Sul, 2016.

[17] Malik H, Hemmati H, and Hassan AE. Automatic detection of performance deviations in the load testing of Large Scale Systems. In: Notkin D, Cheng BHC, and Pohl K, editors. 35th International Conference on Software Engineering (ICSE); Piscataway, NJ: IEEE; 2013, p. 1012–1021.

[18] Nguyen TH. Using Control Charts for Detecting and Understanding Performance Regressions in Large Software. In: IEEE Fifth International Conference on Software Testing, Verification and Validation; Piscataway, NJ: IEEE; 2012, p. 491–494.

[19] Nguyen TH, Adams B, Jiang ZM, et al. Automated detection of performance regressions using statistical process control techniques. In: Kaeli D and Rolia J, editors. Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering; New York, New York, USA: ACM; 2012, p. 299.

[20] Foo KC, Jiang ZM, Adams B, et al. Mining Performance Regression Testing Repositories for Automated Performance Analysis. In: Wang J, Chan WK, and Kuo FC, editors. 10th International Conference on Quality Software (QSIC); Piscataway, NJ: IEEE; 2010, p. 32–41.

[21] Žaleźničenka Ź. Automated detection of performance regressions in web applications using association rule mining. Master's thesis. Delft: Delft University of Technology, 2013.

[22] Shang W, Hassan AE, Nasser M, et al. Automated Detection of Performance Regressions Using Regression Models on Clustered Performance Counters. In: John LK, Smith CU, Sachs K, et al., editors. Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering; New York, New York, USA: ACM; 2015, p. 15–26.

[23] Becker S, Koziolek H, and Reussner R. The Palladio component model for model-driven performance prediction. Journal of Systems and Software 2009;82(1):3–22.

[24] Koziolek H. Performance evaluation of component-based software systems: A survey. Performance Evaluation 2010;67(8):634–658.

[25] Brugali D and Scandurra P. Component-based robotic engineering (Part I). Reusable Building Blocks. IEEE Robotics & Automation Magazine 2009;16(4):84–96.

[26] Schlegel C. Communication Patterns as Key Towards Component-Based Robotics. International Journal of Advanced Robotic Systems 2006;3(1):49–54.

[27] Wienke J and Wrede S. A Middleware for Collaborative Research in Experimental Robotics. In: IEEE / SICE International Symposium on System Integration (SII 2011); IEEE; 2011, p. 1183–1190.

[28] Protocol Buffers [computer software]. [cited 2016 Aug 17]. Available from: `https://developers.google.com/protocol-buffers/`.

[29] McKinney W. Data Structures for Statistical Computing in Python. In: Walt S van der and Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010, p. 51–56.

[30] LNT [computer software]. [cited 2016 Aug 30]. Available from: `http://llvm.org/docs/lnt/`.

[31] Wienke J and Wrede S. Autonomous Fault Detection for Performance Bugs in Component-Based Robotic Systems. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems; IEEE; 2016, p. 3291–3297.