

Towards a Container-based Approach for Non-functional Testing of Code Generators

Mohamed Boussaa Olivier Barais
Benoit Baudry

Diverse team INRIA, Rennes, France
{mohamed.boussaa, olivier.barais,
benoit.baudry}@inria.fr

Gerson Sunyé

AtlanMod team INRIA, Nantes, France
gerson.sunye@inria.fr

Abstract

The intensive use of generative programming techniques provides an elegant engineering solution to deal with the heterogeneity of platforms or technological stacks. The use of Domain Specifics Language, for example, leads to the creation of numerous code generators and compilers that will automatically translate high-level system specifications into multi-target executable code. Producing correct and efficient code generator is complex and error-prone. Although software designers provide generally high-level test suites to verify the functional outcome of generated code, it remains challenging and tedious to verify the behavior of produced code in terms of non-functional properties. This paper describes a practical approach based on a runtime monitoring infrastructure to automatically check potential inefficient code generator. This infrastructure based on system containers as execution platforms allows a code-generator developer to evaluate the consistency and coherence of generated code regarding the non-functional properties. This approach provides a fine-grained understanding of resource consumption and analysis of components behavior. We evaluate our approach by analyzing the non-functional properties of HAXE, a popular high-level programming language that involves a set of cross-platform code generators able to compile to different target platforms. Our experimental results show that our approach is able to detect some non-functional inconsistencies within HAXE code generators.

Categories and Subject Descriptors D.3.4 [Programming Languages]: Processors compilers, code generation

General Terms Generative Programming, Testing, Components

Keywords code quality, non-functional properties, code generator, testing

1. Introduction

Nowadays, the intensive use of generative programming techniques has become a common practice for software development to tame the runtime platforms heterogeneity that exists in several domain such as mobile development or Internet of Things development. Generative programming techniques reduces the development and maintenance effort by developing at a higher-level of abstraction through the use of domain-specific languages [1] (DSLs). Based on this common abstraction, code-generation environments automate and systematize the process of building families of software systems which can clearly reduce the effort of software implementation. Indeed, a code generator is used to transform source code programs/models represented in a graphical/textual modeling language) to general purpose programming languages such as C, Java, C++, PHP, etc. In turn, generated code is transformed into machine code (binaries) using a set of specific compilers.

However, code generators can be difficult to understand since they involve a set of complex and heterogeneous technologies which make the task of performing design, implementation, and testing very hard and time-consuming [4, 6]. Code generators have to respect different requirements which preserve software reliability and quality. In fact, faulty code generators can generate defective software artifacts which range from uncompileable or semantically dysfunctional code that causes serious damage to the target platform, to non-functional bugs which lead to poor-quality code that can affect system reliability and performance (e.g., high resource usage, high execution time, etc.).

In order to check the correctness of the code generation process, developers often use to define (at design time) a set of test cases that verify the functional outcome of generated code. After code generation, test suites are executed within each target platform. This may lead to either a correct behavior (i.e., expected output) or a failure (i.e., crashes, bugs).

On the other hand, for non-functional testing of code generators, developers need to deploy and execute software artifacts within different execution platforms. Then, they have to



Figure 1. An overall overview of the different processes involved to ensure the code generation and non-functional testing of produced code from design time to run time.

collect and compare information about the performance and efficiency of generated code. Finally, they report issues related to the code generation process such as incorrect typing, memory management leaks, etc. To do so, developers generally use several platform-specific profilers, trackers, and monitoring tools [2, 5] in order to find some inconsistencies or bugs during code execution. Ensuring the code quality of generated code can refer to several non-functional properties such as code size, resource or energy consumption, execution time, among others [9]. Comparing the result is often complex as it does not exist general benchmark that compare the execution of a C, a Java or a JavaScript program. Consequently, testing the non-functional properties of code generators remains challenging and time-consuming task because developers have to analyze and verify code for each target platform using platform-dependent tools.

This paper proposes an approach to automatically compare and detect inconsistencies between code generators. This approach provides a runtime monitoring infrastructure based on system containers, as execution platforms, to evaluate the consistency and coherence of generated code regarding the non-functional properties. This approach provides a fine-grained understanding of resource consumption and analysis of components behavior. We evaluate our approach by analyzing the non-functional properties of HAXE code generator. Haxe is a popular high-level programming language¹ that involves a set of cross-platform code generators able to compile to different target platforms. Our experimental results show that our approach is able to detect non-functional inconsistencies within HAXE code generators.

In this paper, we make the following contributions:

- We propose a micro-service infrastructure to ensure the deployment and monitoring of generated code. In this paper, we focus more on the relationship between runtime

execution of generated code and resource consumption profiles (CPU and memory usage).

- We also report the results of an empirical study by evaluating the non-functional properties of HAXE code generators. The obtained results provide evidence to support the claim that our proposed infrastructure is efficient and effective.

The paper is organized as follows. Section II describes the background and motivation behind this work. We present in Section III our infrastructure for non-functional testing of code generators using micro-services. The evaluation and results of our experiments are discussed in Section IV. Finally, related work, concluding remarks and future work are provided in Sections V and VI.

2. Motivations

In different domain, the use of code generator is a common practices. We can cite three approaches that intensively develop and use code generators.

- Haxe is an open source toolkit for cross-platform development which compiles to a number of different programming platforms, including JavaScript, Flash, PHP, C++, C# and Java. Haxe involves many features: the Haxe language, multi-platform compilers, and different native libraries. The Haxe language is a high-level programming language which is strictly typed. This language supports both functional programming and object-oriented programming paradigms. It has a common type hierarchy, making certain API available on every targeted platform. Moreover, Haxe comes with a set of code generators that translate manually-written code (in Haxe language) to different target languages and platforms. Haxe code can be compiled for applications running on desktop, mobile and web platforms. Compilers ensure the correctness of user code in terms of syntax and type safety.

¹ 1440 github star

Haxe comes also with a set of standard libraries that can be used on all supported targets and platform-specific libraries for each of them. One of the main usage of Haxe is to develop Cross-Platform Games or Cross-Platform library that can run on mobile, on the Web or on a Desktop. This project is popular (more than 1440 stars on github).

- An other motivating example is ThingML. ThingML is a modeling language for embedded and distributed systems. The idea of ThingML is to develop a practical model-driven software engineering tool-chain which targets resource constrained embedded systems such as low-power sensor and microcontroller based devices. ThingML is developed as a domain-specific modeling language which includes concepts to describe both software components and communication protocols. The formalism used is a combination of architecture models, state machines and an imperative action language. The ThingML toolset provides a set of code generators to translate ThingML to C, Java and JavaScript. It includes a set of variant for the C and JavaScript code generators to target different embedded system and their constraints. This project is still confidential but seems to be representative of the modelling community practices.
- A Third motivating example is TypeScript. TypeScript is a typed superset of JavaScript that compiles to plain JavaScript. In fact, it does not only compile to one version of JavaScript. It can transform typeScript to EcmaScript 3, 5 or 6. It can use different module system ('none', 'commonjs', 'amd', 'system', 'umd', 'es6', or 'es2015'). Each of this variation point can target different code generators (function *emitES6Module* vs *emitUMDModule* in *emitter.ts* for example). This project is popular (more than 12,619 stars on github).

Based on these three sample, we can observe on github that all of them use tests to check the correctness of the code generator. In this test, all of them cover the correctness of the language feature. A reliable and accepted way to increase confidence in the correct functioning of code generators is to validate and check the functionality of generated code, which is common practice for compiler validation and testing. Therefore, developers try to check the syntactic and semantic correctness of generated code by means of different techniques such as static analysis, test suites, etc., and ensure that the code is behaving correctly. In model-based testing for example [7, 11], testing code generators focuses on testing the generated code against its design. Thus, the model and the generated code are executed in parallel, by means of simulations, with the same set of test suites. Afterwards, the two outputs are compared with respect to certain acceptance criteria. Test cases, in this case, can be designed to maximize the code or model coverage [12].

In fact, code generators have to respect different requirements which preserve software reliability and quality [3]. A

non-efficient code generator might generate defective software artifacts (code smells) that violates common software engineering practices. Thus, poor-quality code can affect system reliability and performance (e.g., high resource usage, low execution speed, etc.). Thus, another important aspect of code generator's testing is to test the non-functional properties of produced code. Proving that the generated code is functionally correct is not enough to claim the effectiveness of the code generator under test. In looking at the three motivating example, ThingML and TypeScript does not provide any specific test to check the consistency of the different memory usage or CPU consumption regarding the different code generators and their variants. Haxe provides two test cases² to bench the resulting generated code. One serves to bench an example in which object allocations are deliberately (over) used to measure how memory access/GC mixes with numeric processing in the different target language. The second test mainly bench the network speed in each targeted platform.

However, based on existing benchmark between technical environments [], in comparing the behavior of the generated code of a large data set of programs, our feeling is that we could detect inefficient code generators. The kind of error we track is the lack of use of a specific function, a specific type or a specific language feature in a targeted environment that can speed or reduce the memory consumption of the resulting program. The main difficulties is the fact that, for testing the no functional properties of a code generator, we cannot just observe the execution of the code generator but we have to observe and compare the execution of the generated program. There is no exact oracle to detect an inconsistencies as it is complex to compare the results of two program that runs atop two different technical stack (what is the meaning of comparing the CPU or memory usage of a PHP code and a Java Code?).

Figure 1 sums up an overall overview of the different processes that are involved together to ensure the code generation and non-functional testing of produced code from design time to run time. We distinguish 4 major steps: the software design using high-level system specifications, code generation by means of code generators, code execution, and non-functional testing of generated code.

Firstly, software developers have to define, at design time, software's behavior using a high-level abstract language (DSLs, models, program). Afterwards, developers can use platform-specific code generators to ease the software development and generate automatically code that targets different languages and platforms. We depict, in Figure 1, three code generators capable to generate code in three software programming languages (JAVA, C# and C++). In this step, code generators transform the previously designed model to produce, as a consequence, software artifacts for the target platform.

² <https://github.com/HaxeFoundation/haxe/tree/development/tests/benchs>

In the next step, generated software artifacts (e.g., JAVA, C#, C++, etc.) are compiled, deployed and executed across different target platforms (e.g., Android, ARM/Linux, JVM, x86/Linux, etc.). Thus, several code compilers are needed to transform source code to machine code (binaries) in order to get executed.

Finally, to perform the non-functional testing of generated code, developers have to collect, visualize and compare information about the performance and efficiency of running code across the different platforms. Therefore, they generally use several platform-specific profilers, trackers, instrumenting and monitoring tools in order to find some inconsistencies or bugs during code execution [2, 5]. Ensuring the code quality of generated code can refer to several non-functional properties such as code size, resource or energy consumption, execution time, among others [9]. Finding inconsistencies within code generators involves analyzing and inspecting the code and that, for each execution platform. For example, one of the methods to handle that is to analyze the memory footprint of software execution and find memory leaks. Developers then, can inspect the source code and the generated code and find some parts of the code-base that have triggered this issue. Therefore, software testers generally use to report statistics about the performance of generated code in order to fix, refactor, and optimize the code generation process.

3. Approach overview

3.1 An Infrastructure for Non-functional Testing Using System Containers

In general, there are many non-functional requirements that can be evaluated by software testers such as performance (execution time), code quality, robustness, resource usage, *etc.* In this paper, our approach focuses on the non-functional properties related to the performance and efficiency of generated code in terms of resource consumption (memory and CPU).

In fact, to assess the performance/non-functional properties of generated code many system configurations (i.e., execution environments) must be considered. Running different applications (i.e., generated code) with different configurations on one single machine is complex a single system has limited resources and this can lead to performance regressions. Moreover, each execution environment comes with a collection of appropriate tools such as compilers, code generators, debuggers, profilers, etc. Therefore, we need to deploy the test harness, i.e. the produced binaries, on an elastic infrastructure that provides to compiler user facilities to ensure the deployment and monitoring of generated code in different environment settings.

Consequently, we rely on a component-based infrastructure and benefit from all its advantages to automatically:

1. Deploy the generated code within a set of containers

2. Execute the produces binaries in an isolated environment
3. Monitor service containers
4. Gather performance metrics (CPU, Memory, etc.)

First, instead of configuring all code generators under test (GUTs) within the same host machine, we wrap each GUT within a system container. Afterwards, a new instance of the container is created to enable the execution of generated code in a isolated monitored and configured environment. Meanwhile, we start our runtime testing components. A monitoring component collects usage statistics of all running containers and save them at runtime in the time series database component. Thus, we can compare later informations about the resource usage of generative program and detect inconsistencies in the generated code.

The rest of this section details the technical choices we made to synthesis a testing infrastructure to detect inconsistencies between code-generator regarding memory and CPU consumption.

3.2 Technical implementation

The general overview of the technical implementation is shown in Figure 2. In the following subsections, we describe the deployment and testing architecture of generated code within containers.

3.2.1 System Containers as Execution platforms

Before starting to monitor and test applications, we have to deploy generated code on different components to ease containers provisioning and profiling. We aim to use Docker Linux containers to monitor the execution of different generated artifacts in terms of resource usage. Docker <https://www.docker.com> is an open source container engine that automates the deployment of any application as a lightweight, portable, and self-sufficient container that runs virtually on a host machine. Using Docker, we can define pre-configured applications and servers to host as virtual images. We can also define the way the service should be deployed in the host machine using configuration files called Docker files. We use the Docker Hub³ for building, saving, and managing all our Docker images. We can then instantiate different containers from these Docker images.

Therefore, each generated code is executed individually inside an isolated Linux container. By doing so, we ensure that each executed program runs in isolation without being affected by the host machine or any other processes. Moreover, since a container is cheap to create, we are able to create too many containers as long as we have new programs to execute. Since each program execution requires a new container to be created, it is crucial to remove and kill containers that have finished their job to eliminate the load on the system. In fact, containers/softwares are running sequentially without defining any resource constraints. So once execution

³ <https://hub.docker.com/>

is done, resources reserved for the container are automatically released to enable spawning next containers. Therefore, the host machine will not suffer too much from performance trade-offs.

In short, the main advantages of this approach are:

- The use of containers induces less performance overhead and resource isolation compared to using a full stack virtualization solution [10]. Indeed, instrumentation and monitoring tools for memory profiling like Valgrind [8] can induce too much overhead.
- Thanks to the use of Dockerfiles, the proposed framework can be easily configured by software testers in order to define the code generators under test (e. g., code generator version, dependencies, etc.), the host IP and OS, the DSL design, the optimization options, etc. Thus, we can use the same configured Docker image to execute different instances of generated code. For hardware architecture, containers share the same platform architecture as the host machine (e.g., x86, x64, ARM, etc.).
- Docker uses Linux control groups (cgroups) to group processes running in the container. This allows us to manage the resources of a group of processes, which is very valuable. This approach increases the flexibility when we want to manage resources, since we can manage every group individually. For example, if we would evaluate the non-functional requirements of generated code within a resource-constraint environment, we can easily request and limit resources within the execution container according to the needs.
- Although containers run in isolation, they can share data with the host machine and other running containers. Thus, non-functional data relative to resource consumption can be easily gathered and managed by other containers (i. e., for storage purpose, visualization)

3.2.2 Runtime Testing Components

In order to test our running applications within Docker containers, we aim to use a set of Docker components to ease the extraction of non-functional properties related to resource usage.

Monitoring Component This container provides an understanding of the resource usage and performance characteristics of our running containers. Generally, Docker containers rely on cgroups file systems to expose a lot of metrics about accumulated CPU cycles, memory, block I/O usage, etc. Therefore, our monitoring component automates the extraction of runtime performance metrics stored in cgroups files. For example, we access live resource consumption of each container available at the cgroup file system via stats found in `"/sys/fs/cgroup/cpu/docker/(longid)/"` (for CPU consumption) and `"/sys/fs/cgroup/memory/docker/(longid)/"` (for stats related to memory consumption). This component will automate the process of service discov-

ery and metrics aggregation for each new container. Thus, instead of gathering manually metrics located in cgroups file systems, it extracts automatically the runtime resource usage statistics relative to the running component (i.e., the generated code that is running within a container). We note that resource usage information is collected in raw data. This process may induce a little overhead because it does very fine-grained accounting of resource usage on running container. Fortunately, this may not affect the gathered performance values since we run only one version of generated code within each container. To ease the monitoring process, we integrate cAdvisor, a Container Advisor⁴. It is a tool that monitors service containers at runtime.

However, cAdvisor monitors and aggregates live data over only 60 seconds interval. Therefore, we would like to record all data over time since container's creation. This is useful to run queries and define non-functional metrics from historical data. Thereby, to make gathered data truly valuable for resource usage monitoring, it becomes necessary to log it into a database at runtime. Thus, we link our monitoring component to a back-end database component.

Back-end Database Component This component represents a time-series database back-end. It is plugged with the previously described monitoring component to save the non-functional data for long-term retention, analytics and visualization.

During the execution of generated code, resource usage stats are continuously sent to this component. When a container is killed, we are able to access to its relative resource usage metrics through the database. We choose a time series database because we are collecting time series data that correspond to the resource utilization profiles of programs execution.

We use InfluxDB⁵, an open source distributed time-series database as a back-end to record data. InfluxDB allows the user to execute SQL-like queries on the database. For example, the following query reports the maximum memory usage of container `"generated_code_v1"` since its creation:

```
select max (memory_usage) from stats
where container_name='generated_code_v1'
```

To give an idea about the data gathered from the monitoring component and stored in the time-series database, we describe in Table 1 these collected metrics:

Apart from that, our framework provides also information about the size of generated binaries and the compilation time needed to produced code. For instance, resource usage statistics are collected and stored using these two components. It is relevant to show resource usage profiles of running programs overtime. To do so, we present a front-end visualization component for performance profiling.

⁴<https://github.com/google/cadvisor>

⁵<https://github.com/influxdata/influxdb>

Metric	Description
Name	Container Name
T	Elapsed time since container's creation
Network	Stats for network bytes and packets in and out of the container
Disk IO	Disk I/O stats
Memory	Memory usage
CPU	CPU usage

Table 1. Resource usage metrics recorded in InfluxDB

Front-end Visualization Component Once we gather and store resource usage data, the next step is visualizing them. That is the role of the visualization component. It will be the endpoint component that we use to visualize the recorded data. Therefore, we provide a dashboard to run queries and view different profiles of resource consumption of running components through web UI. Thereby, we can compare visually the profiles of resource consumption among containers. Moreover, we use this component to export the data currently being viewed into static CSV document. So, we can perform statistical analysis on this data to detect inconsistencies or performance anomalies. To do so, we choose Grafana⁶, a time-series visualization tool available for Docker.

3.2.3 Wrapping Everything Together: Architecture Overview

To summarize, we present, in Figure 2, an overall overview of the different components involved in our approach to perform the non-functional testing of code generators.

First, instead of configuring all code generators under test (GUTs) within the same host machine (as we presented in the motivation section), we wrap each GUT within a container. To do so, we create a new configuration image for each GUT (i.e., the Docker image) where we install all the libraries, compilers, and dependencies needed to ensure the code generation and compilation. Thereby, the GUT produce code within multiple instances of preconfigured Docker images. Afterwards, a new instance of the container is created to enable the execution of generated code in a isolated and configured environment. In this step, new configurations are needed to run each job without any issue (i.e., define the execution environment within the container, the host machine, the target platform, limit the resources, etc.). For our case, a job represents a generated and compiled program using one of the GUTs. Meanwhile, we start our runtime testing components (e.g., cAdvisor, InfluxDB and Grafana). The monitoring component collects usage statistics of all running containers and save them at runtime in the time series database component. The visualization component comes later to al-

low end users to define performance metrics and draw up charts. The use of the front-end visualization component is optional and we can directly access to information stored in the database through REST API calls. Thus, we can observe later informations about the resource usage of generated code and detect inconsistencies in the generated code by GUTs.

Remark. *We would notice that this testing infrastructure can be generalized and adapted to other case studies other than code generators. Using system containers, any software application/generated code can be easily deployed within containers (i.e., by configuring the container image). It will be later executed and monitored using our runtime monitoring engine.*

4. Evaluation

So far, we have presented a sound procedure and automated component-based framework for extracting the non-functional properties of generated code. In this section, we evaluate the implementation of our approach by explaining the design of our empirical study and the different methods we used to assess the effectiveness of our approach. The experimental material is available for replication purposes⁷.

4.1 Experimental Setup

4.1.1 Code Generators Under Test: HAXE compilers

In order to test the applicability of our approach, we conduct experiments on a popular high-level programming language called HAXE and its code generators.

Haxe is an open source toolkit for cross-platform development which compiles to a number of different programming platforms, including JavaScript, Flash, PHP, C++, C# and Java. Haxe involves many features: the Haxe language, multi-platform compilers, and different native libraries. The Haxe language is a high-level programming language which is strictly typed. This language supports both functional programming and object-oriented programming paradigms. It has a common type hierarchy, making certain API available on every targeted platform. Moreover, Haxe comes with a set of compilers that translate manually-written code (in Haxe language) to different target languages and platforms. Haxe code can be compiled for applications running on desktop, mobile and web platforms. Compilers ensure the correctness of user code in terms of syntax and type safety. Haxe comes also with a set of standard libraries that can be used on all supported targets and platform-specific libraries for each of them.

The process of code transformation and generation can be described as following: Haxe compilers analyzes the source code written in Haxe language then, the code is checked and parsed into a typed structure, resulting in a typed abstract syntax tree (AST). This AST is optimized and transformed

⁶<https://github.com/grafana/grafana>

⁷<https://testingcodegenerators.wordpress.com/>

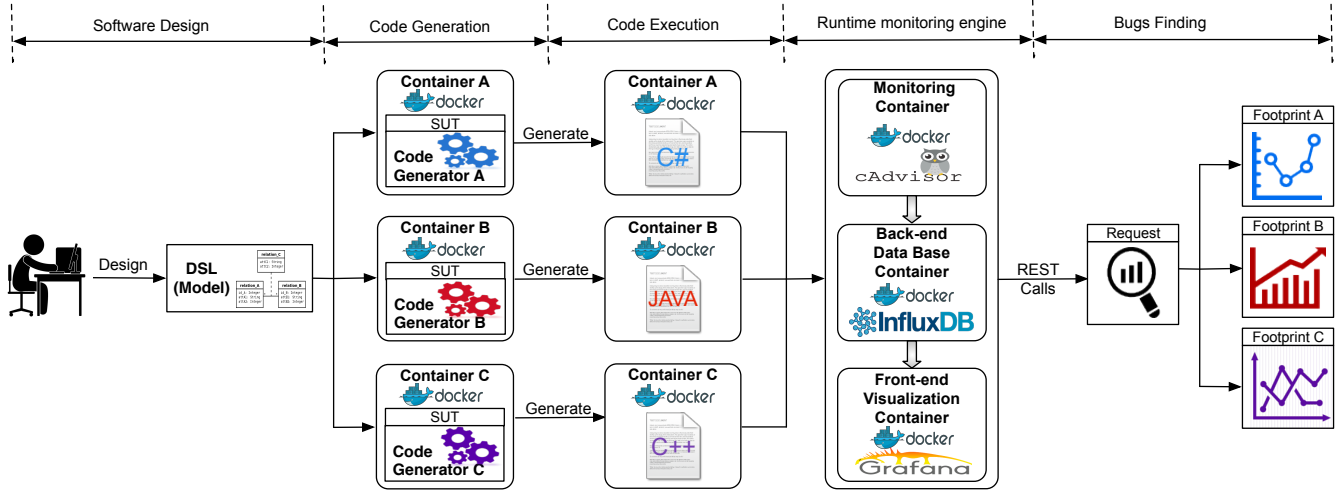


Figure 2. An overall overview of the different processes involved to ensure the code generation and non-functional testing of produced code from design time to runtime.

afterwards to produce source code for target platform/language.

Haxe offers the option of choosing which platform to target for each program using a command-line tool. Moreover, some optimizations and debugging information can be enabled through CLI but in our experiments, we did not turned on any further options.

4.1.2 Cross-platform Benchmark

One way to prove the effectiveness of our approach is to create benchmarks. Thus, we use the Haxe language and its code generators to build a cross-platform benchmark. The proposed benchmark is composed of a collection of cross-platform libraries that can be compiled to different targets⁸. In these experiments, we consider five Haxe code generators to test: Java, JS, C++, CS, and PHP code generator.

In fact, each Haxe library comes with an API and a set of test suites. These tests, written in Haxe, represent a set of unit tests that covers the different functions of the API. The main task of these tests is to check the correct functional behavior of generated programs once generated code is executed within the target platform. To prepare our benchmark, we have removed all the tests that fail to compile to our five targets (i.e., errors, crashes and failures) and we kept only test suites that are functionally correct. Moreover, we added manually new test cases to some libraries in order to extend the number of test suites. the number of test suites depends on the number of functions within the Haxe library.

We use then, these test suites then, to generate a load and stress the target library. This can be useful to study the impact of this load on the resource usage of the system. For example, if one test suite consumes a lot of resources for a specific target, then this could be explained by the fact that

the code generator has produced code that is very greedy in terms of resources.

Thus, we run each test suite 1000 times to get comparable values in terms of resource usage. Table 2 describes the Haxe libraries that we have selected from this benchmark to evaluate our approach.

Library	#TestSuites	Description
Color	19	Color conversion from/to any color space
Core	51	Provides extensions to many types
Hxmath	6	A 2D/3D math library
Format	4	Format library such as dates, number formats
Promise	3	Library for lightweight promises and futures
Culture	4	Localization library for Haxe
Math	3	Generation of random values

Table 2. Description of selected benchmark libraries

4.1.3 Evaluation Metrics Used

We use to evaluate the efficiency of generated code using the following non-functional metrics:

-*Memory Usage (MU)*: It corresponds to the maximum memory consumption of the running container under test. Memory usage is measured in bytes.

-*Execution Time (T)*: Program execution time is measured in seconds.

We recall that our tool is able to evaluate other non-functional properties of generated code such as code generation time, compilation time, code size, CPU usage.

⁸<http://thx-lib.org/>

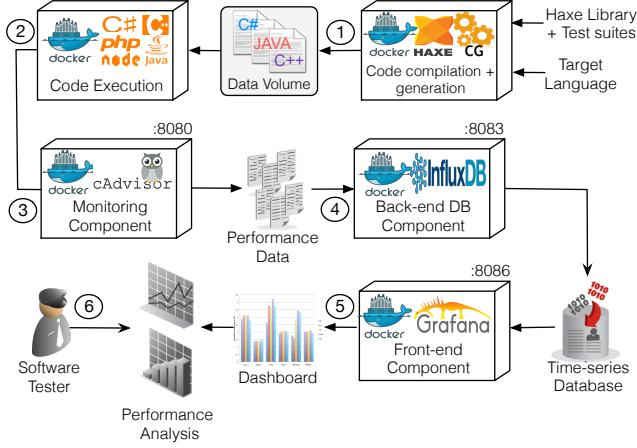


Figure 3. Comparison of average memory consumption and execution time of FFmpeg containers compiled with standard GCC optimization options

4.1.4 Setting up Infrastructure

To assess our approach, we configure our previously proposed container-based infrastructure for non-functional testing of code generators in order to run experiments on the Haxe case study. Figure 3 shows a big picture of the testing and monitoring infrastructure considered in these experiments.

Benchmark	TestSuite	Std_dev	TestSuite	Std_dev	TestSuite	Std_dev
Color	TS1	0.55	TS8	0.24	TS15	0.73
	TS2	0.29	TS9	0.22	TS16	0.12
	TS3	0.34	TS10	0.10	TS17	0.31
	TS4	2.51	TS11	0.17	TS18	0.34
	TS5	1.53	TS12	0.28	TS19	120.61
	TS6	43.50	TS13	0.33		
	TS7	0.50	TS14	1.88		
Core	TS1	0.35	TS18	0.16	TS35	1.30
	TS2	0.07	TS19	0.60	TS36	1.13
	TS3	0.30	TS20	5.79	TS37	2.02
	TS4	27299.89	TS21	0.47	TS38	0.26
	TS5	6.12	TS22	2.74	TS39	0.16
	TS6	21.90	TS23	2.14	TS40	8.12
	TS7	0.41	TS24	3.79	TS41	5.45
	TS8	0.28	TS25	0.19	TS42	0.11
	TS9	0.78	TS26	0.13	TS43	1.41
	TS10	1.82	TS27	5.59	TS44	1.56
	TS11	180.68	TS28	1.71	TS45	0.11
	TS12	185.02	TS29	0.26	TS46	1.04
	TS13	128.78	TS30	0.44	TS47	0.23
	TS14	0.71	TS31	1.71	TS48	1.34
	TS15	0.12	TS32	2.42	TS49	1.86
	TS16	0.65	TS33	8.29	TS50	1.28
	TS17	0.26	TS34	5.25	TS51	3.53
Hxmath	TS1	31.65	TS3	30.34	TS5	0.40
Format	TS2	4.27	TS4	0.25	TS6	0.87
	TS1	0.28	TS3	95.36	TS4	1.49
Promise	TS2	64.94				
	TS1	0.29	TS2	13.21	TS3	1.21
Culture	TS1	0.13	TS3	0.13	TS4	1.40
	TS2	0.10				
Math	TS1	642.85	TS2	28.32	TS3	24.40

Table 3. My caption

First, we create a new Docker image in where we install the Haxe code generators and compilers (through the configuration file "Dockerfile"). Then a new instance of that

Benchmark	TestSuite	Std_dev	TestSuite	Std_dev	TestSuite	Std_dev
Color	TS1	10,19	TS8	1,23	TS15	14,44
	TS2	1,17	TS9	1,95	TS16	1,13
	TS3	0,89	TS10	1,27	TS17	0,72
	TS4	30,34	TS11	0,57	TS18	0,97
	TS5	31,79	TS12	1,11	TS19	777,32
	TS6	593,05	TS13	0,46		
	TS7	12,14	TS14	45,90		
Core	TS1	1,40	TS18	1,00	TS35	14,13
	TS2	1,17	TS19	20,37	TS36	32,41
	TS3	0,60	TS20	128,23	TS37	22,72
	TS4	403,15	TS21	24,38	TS38	2,19
	TS5	41,95	TS22	76,24	TS39	0,26
	TS6	203,55	TS23	18,82	TS40	126,29
	TS7	19,69	TS24	72,01	TS41	31,01
	TS8	0,78	TS25	0,21	TS42	0,93
	TS9	30,41	TS26	2,30	TS43	50,36
	TS10	57,19	TS27	101,53	TS44	12,56
	TS11	68,92	TS28	43,67	TS45	0,91
	TS12	74,19	TS29	0,90	TS46	27,28
	TS13	263,99	TS30	4,02	TS47	1,10
	TS14	19,89	TS31	52,35	TS48	15,40
	TS15	0,30	TS32	134,75	TS49	37,01
	TS16	28,29	TS33	82,66	TS50	23,29
	TS17	1,16	TS34	89,57	TS51	1,28
Hxmath	TS1	444,18	TS3	425,65	TS5	17,69
	TS2	154,80	TS4	0,96	TS6	46,13
Format	TS1	0,74	TS3	255,36	TS4	8,40
	TS2	106,87				
Promise	TS1	0,30	TS2	58,76	TS3	20,04
Culture	TS1	1,28	TS3	0,58	TS4	15,69
	TS2	4,51				
Math	TS1	1041,53	TS2	234,93	TS3	281,12

Table 4. My caption

image is created. It takes as an input the Haxe library we would to test and the list of test suites (step 1). It produces as an output the source code and binaries that have to be executed. These files are saved in a shared repository. In Docker environment, this repository is called Data Volume. A data volume is a specially-designated directory within containers that share data with the host machine. So, when we execute the generated test suites, we provide a shared volume with the host machine so that, binaries can be executed in the execution container (Step 2). In fact, for the code execution we created, as well, a new Docker image in where we install all execution tools and environments such as php interpreter, NodeJS, etc.

In the meantime, while running test suites inside the container, we collect runtime resource usage data using cAdvisor (step 3). The cAdvisor Docker image does not need any configuration on the host machine. We have just to run it on our host machine. It will then have access to resource usage and performance characteristics of all running containers. This image uses the cgroups mechanism described previously to collect, aggregate, process, and export ephemeral real-time information about running containers. Then, it reports all statistics via web UI (<http://localhost:8080>) to view live resource consumption of each container. cAdvisor has been widely used in different projects such as Heapster⁹ and Google Cloud Platform¹⁰. In this experiment, we choose to

⁹<https://github.com/kubernetes/heapster>

¹⁰<https://cloud.google.com/>

	JS		JAVA		C++		CS		PHP	
	Memory	Factor	Memory	Factor	Memory	Factor	Memory	Factor	Memory	Factor
Color_TS6	900,70	x1.0	1362,55	x1,5	2275,49	x2,5	1283,31	x1,4	758,79	x0,8
Color_TS19	253,01	x1.0	819,92	x3,2	923,99	x3,7	327,61	x1,3	2189,86	x8,7
Core_TS4	303,09	x1.0	768,22	x2,5	618,42	x2	235,75	x0,8	1237,15	x4,1
Hxmath_TS1	104,00	x1.0	335,50	x3,2	296,43	x2,9	156,41	x1,5	1192,98	x11,5
Hxmath_TS3	111,68	x1.0	389,73	x3,5	273,12	x2,4	136,49	x1,2	1146,05	x10,3
Math_TS1	493,66	x1.0	831,44	x1,7	1492,97	x3	806,33	x1,6	3088,15	x6,3

Table 5. My caption

	JS		JAVA		C++		CS		PHP	
	Time	Factor	Time	Factor	Time	Factor	Time	Factor	Time	Factor
Color_TS19	4,52	x1.0	8,61	x1,9	10,73	x2,4	14,99	x3,3	279,27	x61,8
Core_TS4	665,78	x1.0	416,85	x0,6	699,11	x1,1	1161,29	x1,7	61777,21	x92,8
Core_TS11	4,27	x1.0	1,80	x0,4	1,57	x0,4	5,71	x1,3	407,33	x95,4
Core_TS12	4,71	x1.0	2,06	x0,4	1,60	x0,3	5,36	x1,1	417,14	x88,6
Core_TS13	6,26	x1.0	5,91	x0,9	11,04	x1,8	14,14	x2,3	297,21	x47,5
Format_TS2	2,31	x1.0	2,10	x0,9	1,81	x0,8	6,08	x2,6	148,24	x64,1
Format_TS3	5,40	x1.0	5,03	x0,9	7,67	x1,4	12,38	x2,3	220,76	x40,9
Math_TS1	3,01	x1.0	12,51	x4,2	16,30	x5,4	14,14	x4,7	1448,90	x481,7

Table 6. My caption

gather information about the memory usage of running container. Afterwards, we record these data into a new time-series database using our InfluxDB back-end container (step 4). Thus, we define its corresponding ip port into the monitoring component so that, container statistics are sent over TCP port (e.g., 8083) exposed by the database component.

Next, we run Grafana and we link it to InfluxDB by setting up the data source port 8086 so that, it can easily request data from the database. We recall that InfluxDB also provides a web UI to query the database and show graphs (step 5). But, Grafana let us display live results over time in much pretty looking graphs. Same as InfluxDB, we use SQL queries to extract non-functional metrics from the database for visualization and analysis (step 6). In our experiment, we are gathering the maximum memory usage values without presenting the graphs of resource usage profiles.

To obtain comparable and reproducible results, we use the same hardware across all experiments: an AMD A10-7700K APU Radeon(TM) R7 Graphics processor with 4 CPU cores (2.0 GHz), running Linux with a 64 bit kernel and 16 GB of system memory.

4.2 Experimental Results

We now conduct experiments based on the Haxe benchmark. We run each test suite 1K times and we report the execution time and memory usage across the different target languages: Java, JS, C++, CS, and PHP. The goal of running these experiments is to observe and compare the behavior of generated code regarding the testing load. We recall, as mentioned in the motivation, that we are not using any ora-

cle function to detect inconsistencies. However, we rely on the comparison results across different targets to define code generator inconsistencies. Thus, we use, as a quality metric, the standard deviation to quantify the amount of variation among execution traces (i.e., memory usage or execution time) and that for the five target languages. A low standard deviation of a test suite execution, indicates that the data points (execution time or memory usage data) tend to be close to the mean which we consider as an acceptable behavior. On the other hand, a high standard deviation indicates that one or more data points are spread out over a wider range of values which can be more likely interpreted as a code generator inconsistency.

In Table 3, we report the comparison results of running the benchmark in terms of execution speed. At the first glance, we can clearly see that all standard deviations are more mostly close to 0 - 8 interval. It is completely normal to get such small deviations, because we are comparing the execution time of test suites that are written in heterogeneous languages and executed using different technologies (e.g., interpreters for PHP, JVM for JAVA, etc.). So, it is expected to get a small deviation between the execution times after running the test suite in different languages. For now, we are not detecting any suspect behavior of generated code. However, we remark in the same table, that there are some variation points where the deviation is relatively high. We count 8 test suites where the deviation is higher than 60. We highlighted, in gray, these variation points. This can be

explained by the fact that the execution speed of one or more test suites varies considerably from one language to another.

4.3 Discussions

5. Related Work

The idea of employing container-based approach for testing code generators can be found in several publications.

6. Conclusion and Future Work

Acknowledgments

This work was funded by the European Union Seventh Framework Program (FP7/2007-2013) under grant agreement n611337, HEADS project (www.heads-project.eu)

References

- [1] M. Brambilla, J. Cabot, and M. Wimmer. Model-driven software engineering in practice. *Synthesis Lectures on Software Engineering*, 1(1):1–182, 2012.
- [2] N. Delgado, A. Q. Gates, and S. Roach. A taxonomy and catalog of runtime software-fault monitoring tools. *IEEE Transactions on Software Engineering*, 30(12):859–872, 2004.
- [3] M. Demertzi, M. Annavaram, and M. Hall. Analyzing the effects of compiler optimizations on application reliability. In *Workload Characterization (IISWC), 2011 IEEE International Symposium on*, pages 184–193. IEEE, 2011.
- [4] R. France and B. Rumpe. Model-driven development of complex software: A research roadmap. In *2007 Future of Software Engineering*, pages 37–54. IEEE Computer Society, 2007.
- [5] V. Guana and E. Stroulia. Chaintracker, a model-transformation trace analysis tool for code-generation environments. In *ICMT*, pages 146–153. Springer, 2014.
- [6] V. Guana and E. Stroulia. How do developers solve software-engineering tasks on model-based code generators? an empirical study design. In *First International Workshop on Human Factors in Modeling (HuFaMo 2015)*. CEUR-WS, pages 33–38, 2015.
- [7] S. Jörges and B. Steffen. Back-to-back testing of model-based code generators. In *International Symposium On Leveraging Applications of Formal Methods, Verification and Validation*, pages 425–444. Springer, 2014.
- [8] N. Nethercote and J. Seward. Valgrind: a framework for heavyweight dynamic binary instrumentation. In *ACM Sigplan notices*, volume 42, pages 89–100. ACM, 2007.
- [9] Z. Pan and R. Eigenmann. Fast and effective orchestration of compiler optimizations for automatic performance tuning. In *International Symposium on Code Generation and Optimization (CGO’06)*, pages 12–pp. IEEE, 2006.
- [10] C. C. Spotala, A. Calinciuc, C. O. Turcu, and C. Filote. Performance comparison of a web rtc server on docker versus virtual machine. In *2016 International Conference on Development and Application Systems (DAS)*, pages 295–298. IEEE, 2016.
- [11] I. Stuermer, M. Conrad, H. Doerr, and P. Pepper. Systematic testing of model-based code generators. *IEEE Transactions on Software Engineering*, 33(9):622, 2007.
- [12] I. Stürmer, D. Weinberg, and M. Conrad. Overview of existing safeguarding techniques for automatically generated code. In *ACM SIGSOFT Software Engineering Notes*, volume 30, pages 1–6. ACM, 2005.