Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Thesis author	Tomáš Drozdík		
Thesis title	Asynchronous Duet Benchmarking		
Submitted	2023		
Program	Computer Science	**Specialization**	Software Systems

Review author	Michele Tucci	**Role**	reviewer
Position	Postdoc		

Review text:

The thesis introduces an asynchronous variant of the duet method for software performance benchmarking. The general aim of the duet method is to provide a practical and effective procedure for the detection of performance regressions in virtualized execution environments, like the cloud. Such environments differ substantially from the dedicated bare-metal server that are usually employed for performance testing, mainly because of an increased and less predictable variability, which is typical of shared environments. Therefore, the duet method tries to mitigate the impact that such variability may have on the sensitivity of benchmarks by running an A/B performance testing scenario in parallel. The underling assumption is that, in this way, the circumstantial factors contributing to variability will impact both tests at the same time (impact symmetry). Consequently, the original duet method requires the two benchmarks to be synchronized at the start of each iteration. In this thesis, the candidate evaluates if such requirement for synchronization can be lifted while retaining the method's ability to detect performance changes.

Besides a novel variant of the duet method, the candidate implemented the automation necessary for experimenting such method, and to evaluate it with respect to four research questions, examining: (i) whether the asynchronous variant provides additional runtime savings; (ii) how workloads overlap; (iii) the accuracy in detecting performance regressions; and (iv) minimal detectable slowdowns when compared to other methods.

I find the work to be actual, as the ability to reliably execute performance tests in the cloud is very much appealing in the benchmarking field. The results from the experimental evaluation are enough to encourage further research in this direction. The methods applied are well-researched, relevant, and in line with the state of the art of the field. Presentation and language are appropriate. Moreover, I appreciate that the implementation was carried out with practicality and portability in mind. This is especially valuable because one of the main motivations behind the

idea of an asynchronous variant is that it would not require to modify existing benchmark harnesses in order to synchronize iterations, therefore greatly facilitating adoption.

The introduction and the first chapter present the general problem, the necessary background, and the motivation behind the realization of the duet method, in the first place, and its variant that is the topic of this work.

The second chapter provides details on the duet method in its two variants. This chapter is precise and concise, and the goals of the evaluation are clearly stated. However, the research questions might have benefited from an extended rationale, especially in less obvious evaluation directions, like for the RQ2.

The third chapter deals with the design of the implementation, how to configure the tool to schedule the execution of benchmarks, and how results are subsequently processed. Additionally, this chapter introduces the notation that will be used later to describe the computations required for the evaluation. I do not find this to be the best place to introduce the notation, as the name of the chapter is "Implementation". I would say that the concepts for which a notation is provided are those concerned with the method, which is described in the previous chapter. On a more general note, throughout the thesis, I found a number of inconsistencies regarding notation. Examples are: reusing the same letter with different subscripts or superscripts (such as O), redefinition of a letter through a smaller number of parameters (like it happens for R in 4.3.1, with respect to 3.3.1), or reusing letters in the same definition without defining them (like "Sequential" in 4.3.2).

The forth chapter describes how measurements from the experiments are processed to obtain the results for the research questions. The methods introduced in this chapter are certainly suitable for the task at hand. However, sometimes it is not immediately obvious what motivations justify some choices. One case can be the formula 4.1, whose computation becomes clear later on, but here lacks a more thorough explanation. In the same chapter, the candidate writes that *run means* sampling (from [11]) is used in place of *run merged* sampling because the number of iterations varies among runs. As also explained in [11], I would say that the main reason for such a choice is to preserve dependency between runs and iterations.

The fifth chapter presents the results of the experimental evaluation. For each research question, results are presented and commented comprehensively. The depth of the analysis is appropriate for a thesis, and I was positively surprised from some of the results. Additionally, the repository provided by the candidate contains even more plots that let me dig in the results related to specific workloads or environments. Unfortunately, I believe I have found some inconsistency regarding RQ1 in 5.2.1. The formula to compute the runtime of the asynchronous duet should be the difference between the latest end and the earliest start. The arguments in the formula are inverted, but the computation is correct in the code (`duet/process.py:525`). Similarly, the

formula for the runtime speedup (5.1) does not match the plot in Figure 5.1. It is easy to notice that, in the plot, 1 is subtracted from the ratio of runtimes. This can be confirmed again in the code, specifically in the second cell of the `notebooks/rq1.ipynb` notebook. Both typos do not invalidate the results of this research question, which I verified to be reported correctly in the rest of the text and in the conclusion.

The last chapter outlines the obtained results, comments on their implications, and proposes potential future work in the same direction. This chapter provides a good summary that matches quite nicely the goals set in the second chapter.

I recommend the thesis for defense.

Jan 30, 2023

Signature: