Muhammad Moiz Arif
EV 3.274, 1515 Saint-Catherine Street West
Department of Computer Science and Software Engineering, Concordia University
Montreal, QC, Canada H3G 2W1

April 6, 2017

Dear Editor and Reviewers:

Thank you for your insightful feedback and comments, both positive and constructive, and for allowing us the opportunity to improve our manuscript. We have taken each of your comments into consideration and made the appropriate changes and extensions to our manuscript.

In the end, we added new analysis results as suggested by the reviewers. We studied topics in user reviews in Google Play Store. We also made several clarifications throughout the paper, improved paper presentation, and improved the discussion of related work. We feel that the paper is now much stronger.

Below, we include a description of the changes that we made to our manuscript with respect to each of the reviewer's comments. We denote the reviewers comments in italic typeface, and our responses follow below each reviewer comment.

Reviewer 1 Comment

Comment R1.1. *Is there any evidence supporting that VMs are widely used for performance testing? There has been much work on performance testing (e.g., Nistor ICSE 13, Jin PLDI 12) that does not rely on VMs. The paper should provide more details on the background and motivation for this study, such as why and in what circumstances VMs are used in performance testing.*

There are discussions online: http://stackoverflow.com/questions/8906954/can-you-use-a-virtual-machine-to-performance-test-an-application Even discussed here:

https://social.technet.microsoft.com/Forums/windowsserver/en-US/06c0e09b-c5b4-4e2c-90e3-61b06483fe5b/performance-test-is-not-reliable-on-virtual-machine?forum=winserverhyperv

This one is amazing:

http://sqa.stackexchange.com/questions/7709/performance-testing-systems-on-virtual-machines-that-normally-run-on-physical-ma

WOW man this is like a gift:

http://www.webperformance.com/library/reports/Virtualization2/

And this sugar crm supports both cloud and on-premise options. Plus BES.

People even talk about building perf testing labs using vm http://searchvmware.techtarget.com/tip/Building-a-VMware-test-lab-How-to-obtain-and-interpret-performance-metrics

Are these enough:)

And it?s our expeirence with our industial partner over the years.

You better add a subsection in your background to disccuss this.

In order to better motivate our paper we added motivation example:

A guy has a software system based on premise and cloud due to the constraints and flexib they test on vm correlation model building and finds that io is impacting their performance. Turns out vm io is being hampered.

Cite the io xen server.

If they knew there were discrepancy they wouldnt worry. Approaches to reduce noise.

Comment R1.2. *The study does not address the most important problem in performance testing, i.e., fault detection. Even if the discrepancy exist, thzere is no evidence showing that such discrepancy can affect testing effectiveness. The ultimate objective of performance testing is to find performance bugs. It would be more convincing if the authors can evaluate the discrepancy in finding bugs between VMs and physical environments.*

Say this is the first step. First we need to know what is the descrepancy, how big and how to reduce. Then it makes sense to see the impact on fault and will be our future work. However, without understaing the nature of the descrpancy and directly see the impact on fault, you can?t really reason about results.

Comment R1.3. *The paper is not clear about how the three aspects of testing results can help to find performance problems. Especially for the second analysis - the correlation between metrics - why is it useful?*

I sent you a paper about using correlation. Try to find it.

To see the trends To the realtionships between metrics To see the impact of metrics the all together. Many to many.

Elaborate related work and motivation research question.

Comment R1.4. *In the related work section, instead of just describing the three types of analysis, the authors should relate the existing work to the proposed study. Does the discussed existing work rely on VMs? If it does, are there any problems caused by the discrepancy between VMs and physical environments?*

Say explcitly where in the paper they have mentioned or cite papers where you can see this. Their analysis is not domain specified so we are just testing in both the envion. And we find out that its dispcrepance.

Comment R5.5. *As virtualization becomes wildly adopted, many companies use virtual environment as their production environment to reduce operation costs. Does that invalidate the purpose of this study? What if an application is actually deployed in a virtual environment?*

That sounds like another study about vm variance. But that?s not the focus of our paper. Some software needs to run on premise anyways.

We never talk about test and running in vm. We talk about a scenario where we use vm and physical. Particularly many large s/w systems offering vm and premise. Like CRM and BES. we clarify in the introduction.

Detailed comments:

Comment R1.6. *Section 3.3: "the workload of the performance tests is varied periodically in order to avoid*

bias from a consistent workload" - how did it get varied.

Through the change in number of threads, threads represented users. As the users and number of requests sent/received are directly proportional in these software, eventually varying the workload. Although 7) the reviewer mentions that we have varied it according to the number of threads.

Random workload, try to point it out. We clarify.

Comment R1.7. *"The work-load variation was introduced by the number of threads." Why not consider other types of workloads, such as the amount of input data? Increasing the number of threads may also be used to speedup the performance.*

What input data? If the reviewers mean the input data for the website It is an e-commerce website and the drivers for both systems come with predefined actioned. Log in, browse, buy, exit.

This is a limitation we discuss in threads to validity. Only users=threads.

Comment R1.8. *The quality of the performance tests could greatly influence the testing results. The paper should provide more details. Examples of performance tests can be useful. Also, how many performance tests are used in the study? Why does a test take so long (9 hours) to execute? Is running performance tests twice sufficient to reduce influence of randomness?*

Example of performance tests, I think I will put it under exploratory performance testing. Of course it is not done to see functional/non-functional anomalies. The nature of the study itself requires 1 performance tests spread out over 9 hours. Concatenating multiple tests will only create noise in our data. Combination of workloads and usage scenario. We do not change performance tests, taken s/w as is. We do not want to impact any thing by making a special performacne test. 9 hours: I will site COR-PAUL who ran the test for 24 hours. The idea is to work on metrics under review statistically stable. CITE: Jack. Its about the sample size want to make sure enough data. We actually ran the test multiple times to reduce the randomness and to reach stable test output. But we took the best two runs and compared to them to evaluate the existence of randomness. Its a miscommunication. We actually ran at 3 times. And we do not guarantee that running it thrice will remove all the randomness. Add this to threads to validity. Limitation.

Comment R1.9. *Different performance tests may performance different functionalities (e.g., SQL query VS server restart). The functionalities should be evaluated separately.*

I think this reviewer did not understand how our two software work and I can guess it by the questions. There is no need to evaluate different functionalities because atomically both the drivers are role-playing as a user. Log in - browse - buy - pay - logout. If the nature of all the transactions remains same based on the SQL query, I believe there is no need of evaluating it as different functionalities.

We use a combination of things. Whatever the test driver is based on. Test on single feature maybe system behaves in a different way but ours is based on compound exercising the system.

Comment R1.10. *Section 3.1: what is the size of each application?* Will mention.

Comment R1.11. *On page 7, the design choice of combining metrics of two datasets is not justified.*

We considered it as one system. For a user its just a box.

Comment R1.12. *On page 7, realistically, interference on the real-world systems cannot be restricted like the one mentioned in the setup. The concern is that, by leaving out the system load, the statistical model might miss the opportunity to adjust to the real-world situation. Also, different assumptions about the system workload could affect the choice of the statistical model and thus perturb the prediction results.*

Dr. Shang will attend.

Comment R1.13. *On page 15, what is the purpose of removing "metric that has a higher average correlation with all other metrics"?*

Explain. Dr. shang r function.

We did it based on R. we removed the one which is more correlated to other metrics.

Comment R1.14. *On page 16, what regression model is used? On page 17, linear regression model is mentioned briefly. Why is a linear model chosen? Not until on page 19, the assumption of a linear relationship is mentioned. A brief writing of the design decision would be more appropriate.* Lrm. because prior work and easy to explain.

Say you can use other models. You use linear model since it?s easier to explain, and also used in prior work, cite my paper icpe and the vperfguard

Comment R1.15. *R-squared is used without explanation. If 10-fold cross validation has an explanation, R2 may deserve one too.*

Yes please explain.

Comment R1.16. *On page 18, "good model fit (66.9% to 94.6%)", is that the absolute percentage error?* R2

Comment R1.17. *There is much work on performance testing and bug detection, which should be discussed in related work.*

Yes Look for performance bug detection. 4 or 5 papers. Look at the reviewers example.

Comment R1.18. *If the trace data can be made public, others may use it to replicate the experiments.* yes

Again, we thank all of you for your valuable feedback, which has made this a stronger manuscript. We look forward to hearing your feedback on the updated manuscript.

Sincerely,
Muhammad Moiz Arif, Weiyi Shang, & Emad Shihab