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 10, 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.

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.*

Response. Thank you for pointing this out. We found online discussions by developers and testers supporting our argument of testing across heterogeneous environments [1][5][12]. We also find a an experiment similar to our hypothesis where a web application(Sugar CRM) is tested for identifying performance issues between the physical and virtual environments[7]. There also exist VMware test labs to test an application and analyze performance metrics in a virtual environment [13]. In addition to that, we also highlight that Sugar CRM and Blackberry's BES server are offered with options of deployment on-premise or on cloud [10] [14]. Furthermore, our experience with industrial partners speaks that virtual environments are used to test applications because of their flexibility. We have added a motivating example too to better motivate our paper.

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.*

Response. This study serves as a building block towards fault detection. The goal is to dig deeper into the nature of the discrepancy, its magnitude and approaches to reduce it. Our future work is directed towards fault detection as the next step. However, we understand that without analyzing or not having the knowledge about discrepancy between the two environments, we can not directly look at the impact on faults.

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.
Elaborate related work and motivation research question.

Response. The investigation(s) in this work are based on the following 3 aspects:

1. The first approach is used to identify the trends and distributions of performance metrics. As a result, we can look at the differences at a finer level between the two environments and not just by numbers only.
2. The second approach, was used to identify the change in the nature of relationship between performance metrics. We believe that a change in these relationships can effect the behavior of the subject systems in the two environments.
3. The third and final approach is used to see examine the impact of the metrics all together. This analysis also serves as the baseline for our future study i.e. fault detection.

We have addressed and rephrased this in the journal.

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 explicitly where in the paper they have mentioned or cite papers where you can see this.
Response. We mention at the end of section 2.2:*"Prior research focused on the overhead of virtual environments without considering the impact...and investigate whether such impact can be minimized in practice"*. The domain is not specified in most of the papers that we have mentioned as our related work. Hence, we tested in both the environments and concluded that the methodologies can not be applied as is.

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?*

Response. We agree however that would mean that we need to study the variance present in the virtual environment. We highlight a scenario where both the environments are used and not just only virtual. Like mentioned in response 1.1, some software have the option to run on premise. Particularly large software systems like CRM and BES. We clarify this in the new revision.

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.*

Ask a
Response. We varied it randomly. Although the variation was identical between both the environments. The variation was introduced by the number of threads as mentioned in section 3.3 *"The workload variation was introduced by..."*.

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.*

Ask a

Response. The use cases are predefined in the performance testing suite which is a limitation of our subject

systems. We now discuss it in our revised threads to validity.

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?*

Response. We answer the questions accordingly:

1. We use the same type of performance tests as used in the related studies [3][6][8][2]. Should I mention the exact type which was "exploratory performance testing(ref: Jack's journal paper)"?
2. In total, there are 3 performance tests used in this study.
3. It was very necessary for this study that the systems are stable and the sample sizes are statistically significant. A longer run of the tests would ensure we have covered more data points, than the related studies [3][6][8]. In the future, we can extend our study to run for even a longer period of time.
4. We ran it 3 times in total. We do not guarantee that running it thrice would completely eliminate the randomness.

   We have clarify all of the aforementioned points in the updated version of our manuscript.

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

We decided to test the subject system as a compound. Testing just one of the functionalities may result in a different system behavior but we used only the testing suite which is a mix of all the functionalities, depicting a real-time user. Though we rely on the test only hence this has been added to our threads to validity.

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

Response. Added in the updated manuscript.

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

Response. As for the user it is just a box, we considered it a one complete system. This is now added in the updated manuscript.

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. (although we did not assume anything)

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

Response. We used it to any remove multicollinearity present between performance metrics. This was based on the functions based in R.

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.*

Response. We chose linear regression model as it is used in prior work [15][11]. Also, the model is straightfoward to explain compared to that other models that we may have used. We have added this brief in our revised manuscript.

Response. arg1 **Comment R1.15.** *R-squared is used without explanation. If 10-fold cross validation has*

an explanation, R2 may deserve one too.

Response. We have now provided this explanation.

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

Response. No, that is the value of the R2 of our model.

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

Nistor ICSE 13[9], Jin PLDI 12[4] + 2/3

Response.

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

Response. We agree. We have now made the data public. should i mention the url?

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

References

[1] Dee. performance-testing systems on virtual machines that normally run on physical machines. http://sqa.stackexchange.com/questions/7709/performance-testing-systems-on-virtual-machines-that-normally-run-on-physical-ma, 2014. Accessed: 2017-04-04.

[2] King Chun Foo, Zhen Ming Jiang, Bram Adams, Ahmed E Hassan, Ying Zou, and Parminder Flora. Mining performance regression testing repositories for automated performance analysis. In *Quality Software (QSIC), 2010 10th International Conference on*, pages 32–41. IEEE, 2010.

[3] Zhen Ming Jiang, Ahmed E Hassan, Gilbert Hamann, and Parminder Flora. Automated performance analysis of load tests. In *Software Maintenance, 2009. ICSM 2009. IEEE International Conference on*, pages 125–134. IEEE, 2009.

[4] Guoliang Jin, Linhai Song, Xiaoming Shi, Joel Scherpelz, and Shan Lu. Understanding and detecting real-world performance bugs. In *Proceedings of the 33rd ACM SIGPLAN Conference on Programming Language Design and Implementation*, PLDI '12, pages 77–88. ACM, 2012.

[5] Sean Kearon. Can you use a virtual machine to performance test an application? http://stackoverflow.com/questions/8906954/can-you-use-a-virtual-machine-to-performance-test-an-application, 2012. Accessed: 2017-04-04.

[6] Haroon Malik, Hadi Hemmati, and Ahmed E Hassan. Automatic detection of performance deviations in the load testing of large scale systems. In *Proceedings of the 2013 International Conference on Software Engineering*, pages 1012–1021. IEEE Press, 2013.

[7] Christopher L Merrill. Load testing sugarcrm in a virtual machine. http://www.webperformance.com/library/reports/Virtualization2/, 2009. Accessed: 2017-04-04.

[8] Thanh HD Nguyen, Bram Adams, Zhen Ming Jiang, Ahmed E Hassan, Mohamed Nasser, and Parminder Flora. Automated detection of performance regressions using statistical process control techniques. In *Proceedings of the 3rd ACM/SPEC International Conference on Performance Engineering*, pages 299–310. ACM, 2012.

[9] Adrian Nistor, Linhai Song, Darko Marinov, and Shan Lu. Toddler: Detecting performance problems via similar memory-access patterns. In *Proceedings of the 2013 International Conference on Software Engineering*, ICSE '13, pages 562–571, Piscataway, NJ, USA, 2013. IEEE Press.

[10] Simon Sage. Blackberry enterprise server 12 now available. http://crackberry.com/blackberry-enterprise-server-12-now-available, 2014. Accessed: 2017-04-04.

[11] Weiyi Shang, Ahmed E. Hassan, Mohamed Nasser, and Parminder Flora. Automated detection of performance regressions using regression models on clustered performance counters. In *Proceedings of the 6th ACM/SPEC International Conference on Performance Engineering*, ICPE '15, pages 15–26. ACM, 2015.

[12] Tintin. Performance test is not reliable on virtual machine? https://social.technet.microsoft.com/Forums/windowsserver/en-US/06c0e09b-c5b4-4e2c-90e3-61b06483fe5b/performance-test-is-not-reliable-on-virtual-machine?forum=winserverhyperv, 2011. Accessed: 2017-04-04.

[13] John Tolly. Building a vmware test lab: How to obtain and interpret performance metrics. http://searchvmware.techtarget.com/tip/Building-a-VMware-test-lab-How-to-obtain-and-interpret-performance-metrics, 2012. Accessed: 2017-04-04.

[14] Wikipedia. Sugarcrm. https://en.wikipedia.org/wiki/SugarCRM#Deployment_options, 2017. Accessed: 2017-04-04.

[15] Pengcheng Xiong, Calton Pu, Xiaoyun Zhu, and Rean Griffith. vperfguard: an automated model-driven

framework for application performance diagnosis in consolidated cloud environments. In *Proceedings of the 4th ACM/SPEC International Conference on Performance Engineering*, pages 271–282. ACM, 2013.