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 18, 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 [2][9][19]. 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[13]. There also exist VMware test labs to test an application and analyze performance metrics in a virtual environment [20]. 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 [16] [21]. 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 [6][11][14][3]. Should I mention the exact type which was "exploratory performance testing(ref: Jack's TSE paper)"?
2. In total, there are 3 performance tests used in this study, as mentioned in section 6.1: *"In total, we had results from three performance tests"*
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 [6][11][14]. 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 [22][17]. 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[15], Jin PLDI 12[7] + 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?

Reviewer 2 Comments

Comment R2.19. *Therefore, the claim "To the best of our knowledge, the discrepancy between performance testing results in virtual and physical environments has never been studied.", is arbitrary and does not hold.*

Response. Thank you for your feedback. We have toned it down in the revised manuscript.
I am not sure what is he trying to say here

Comment R2.20. *Some work exists in that have implicitly compared the difference between the metrics harvested from the VMs and Physical machines, however, performance testing was not their main focus. A few explicit research efforts exist to gauge the difference between metrics collected from physical and virtual environment. I am pointing to one such work conducted by Netto from PUCRS and Sadd from Dell "Evaluating load generation in virtualized environments for software performance testing". They conducted* *several load tests on both virtual machines and physical machines.*

Response. We do not compare the two environments based on load tests. We do it on the basis of typical analysis of a performance test. [5][14]

Comment R2.21. *However, the use of VMs may introduce extra overhead (e.g., a higher than expected memory utilization) to the testing environment and lead to unrealistic performance testing results. ???? Why testing over virtualized lead to unrealistic PT results?*

Response. The overhead in a virtual environment can not be simply added on with the results and analyzed. It is rather complex to calculate and evaluate the discrepancy. That is why *Huber et al.* try to build a performance model to predict performance of applications that are migrated from a native system to a virtual environment or from a virtual environment to a new one [4]. There are similar studies which indicate that the overhead in virtual environment may hamper the results of performance tests [1][12].

Comment R2.22. *Our findings show that practitioners cannot assume that their performance tests that are observed on one environment will necessarily apply to another environment????.. Is there any evidence to back up the current practice(s) in which practitioner generalize the result of one environment to another (phy to Vir), especially of the large scale software, you mentioned in the paper?*

Response. Please refer to our response to the comment R1.1. We mention that there are some large-scale systems that offer deployment in both enviornments, on-premise and on cloud. For example, Sugar CRM, Blackberry's BES server and Blue Link ERP [21][16][10]. We also notice through online threads that there are no separate tests written for physical and virtual environments hence the comparison of results can not be straight forward. [8].

Comment R2.23. *Exploring, identifying and minimizing such discrepancy will help practitioners and researchers understand and leverage performance testing results from virtual and physical environments... Why would they *LEVERAGE* performance testing results from virtual and physical and to achieve *WHAT* purpose. Bench-marking folks won't like this idea*

Counter narrative: Many companies also try to virtualize their load generation infrastructure which seems like a good idea for maintenance and elasticity reasons. However, they defiantly do tests on physical environment for the show-off purpose of their performance. This is a fact that you can squeeze the most performance out of physical environment. Nevertheless, especially for SaaS infrastructure, or to cater the need of stakeholders and large clients, virtualized environment are used and performance results obtained are attributed to specific virtual environment. That's why the benchmark teams exists in large enterprise that catalog the performance under different virtualized environment. Application of findings from one environment to another is not a usual practice.

Response. better look for some examples online, like web posts.

Comment R2.24. *Since the authors have framed the paper as an empirical study, it is very important to provide the necessary information in the paper for replication purpose. Replication, allows other researcher to validate authors claim(s) and in cases, compare it with their own techniques/cases and findings. What are workload parameters are used in this paper for DS2 and for Cloud Store using JMeter?* say you will share everything

Response. We will share everything as we proceed with this revised manuscript.

Comment R2.25. *Did you used the same hardware for setting up the virtual environment?*

Response. Yes, we have used the same hardware for setting up the virtual environment.

Comment R2.26. *Among the three machines, on which machine you were running perfmon agent (remote collection)? OR you were running it on all three machines?*

Response. We monitor the process's performance on the machines running the web(app) server and the database server.

Comment R2.27. *How did you ensured that environment remained constant for each performance test? For example after few tests, the disk may get full hence IOPS can get impacted.*

Response. We restore the environments before restarting tests such as the database server. he specifically says IOPS, maybe we can mention we emptied the caches or restarted the system?

Comment R2.28. *For both the Q-Q plots for D2 and Cloud store, the metrics do not have the same trend. Did you used the samples of the performance tests results when the test was in equilibrium? If yes how did you ensured that?*

Response. Yes we waited till the point of stability before processing. We did this buy not including data from the beginning and ending of a test. This was a consistent practise across both environments.

Comment R2.29. *Also, before calculating the correlation among the performance metrics from virtual and physical environments, did you removed ramp-up and ramp-down observations of the performance metrics for a test? The system is usually not stable during warm-up and cool-down period. Ramp-up and down periods, for a test repeated multiple time under a given workload, and constant environment, many not necessary be correlated to each other.*

Response. We completely agree and the sample was taken after the warm up, the ramp-up and before the cool-down period.

Comment R2.30. *What is the implication of this study?*

Response. To build up on what is mentioned in section 7, we emphasize that practitioners need to be aware of such magnitude of discrepancy present between physical and virtual environment. Future research should investigate on how to reduce this discrepancy.

Comment R2.31. *In order to assist the practitioners leverage performance testing results in both environments, we also investigate ways to transform results from virtual and physical environments and performance metrics based on deviance may reduced [REDUCE] the discrepancy between performance metrics... Weird sentence, unless am reading between lines.*

Response. We have rephrased it in the revised manuscript.

Comment R2.32. *...such challenges, virtual environments (i.e., VMs) are often leveraged for performance testing [8,8,47]... why "8" is repeated in the ref?*

Response. Fixed.

Comment R2.33. *...ents, such overhead would not significantly impact on the practitioners who examine the performance testing results...significantly impact WHAT on the practitioners?*

Response. Fixed.

Comment R2.34. *...paper, we perform a study on two open-source systems, DS2 [13] and CloudStore [10], where performance tests are conducted on[USING] virtual and physical environments*

Response. Fixed.

Comment R2.35. *we study whether the performance metric follow[S] the same shape of [THE] distribution and the same trend in virtual and physical environments.*

Response. Fixed.

Comment R2.36. *...which can lead to [A] different set of conclusions.*

Response. Fixed.

Comment R2.37. *For example, [THE] virtual environment has a CPU's utilization spike at a certain time [,] but the spike is absent in the physical*

Response. Fixed.

Comment R2.38. *...there exist[S] a plethora of VM software*

Response. Fixed.

Comment R2.39. *...high when we normalize by [THE] load as per Equation*

Response. Fixed.

Comment R2.40. *In the reference section, there is an extra '24' hanging at the end of the page*

Response. Fixed.

page 18 interval validation: check-

Reviewer 3 Comments

Comment R3.41. *The test environments are not quite clear. This concern is minor because it could be fixed with complete descriptions of the hardware and virtual environment configuration. For example, is the physical server disk setup a RAID? SSD? SATA/NVMe? Is the network traffic generated locally or is it simulated from another machine (which would be influenced by the network hardware)? Perhaps more importantly, what is the disk setup on the virtual machine? One of the findings in the paper is that the I/O metrics differ more than the CPU/memory metrics when compared to the physical machine. That is definitely not surprising if the VM is configured to use VDI or another virtual disk, given the overhead in mimicking a drive within an existing filesystem. It is possible to use disk passthrough in Virtual Box and other VM software. In that case, a physical disk is passed by the host OS directly to the guest OS. That setup tends to be much much closer to native speed.*

Response. Thank you for your feedback in helping us making it a stronger manuscript. We have now provided a detailed description of environment configuration. about the network traffic, it was simulated what can I say about this by not being hampered by the concordia network?. One of our focus when carrying out the experiments was to keep the environment setup as close to real world as possible. There are online posts as why disk pass through is not used any more in the real world environment [18]. Due to lack of flexibility and lack of support we decided to use a virtual disk.

Comment R3.42. *Unclear research methodology. At several points, the paper is not clear on exactly what methodology is being followed. One potential deal breaker for this paper is that the test scenarios for the virtual and physical environments are "similar" so the distributions should have the same shape (first sentence on page 10). The test scenarios should be identical, not just similar. Similar implies that the authors created a test scenario for each environment using the same tools. They should instead have created just one test scenario. If the test scenarios are not the same, then the discrepancy could obviously be due to a different scenario. One way to ensure that the tests are the same is to record the network traffic of one run and use the recorded traffic as input for the next run. In a virtual environment, one can even use the same time and date settings.*

The paper needs clear and precise research questions. The paper jumps into methodology and analysis without clearly explaining what is to be analyzed. For example consider section 4.1, "Examining individual performance metrics." It is not explained exactly what an "individual" metric is. I had to try to understand it from reading the Approach subsection in 4.1. The concept is not that difficult; the section is just comparing, e.g., a CPU metric in one environment with the same metric in the other environment. It just needs to be explained more clearly up front. Some of the language is confusing, such as "intuitively the scales of performance metrics are not the same." I am not sure if the sentence refers to scales not being the same across metrics or across environments.

The authors have tendency to explain what they are not doing, before explaining what they are doing. That is confusing because it makes it difficult to understand the methodology. For example in Section 4, "We do not predict... Instead, our experiments are set in..."

All of these issues with unclear research methodology could be fixed by adding clear research questions and a methodology for answering each question. The bigger problem here is that the methodology is unclear enough that it might be obscuring more important problems under the surface.

Response. According to the explanation provided, our tests were completely identical i.e. same test scenario ran on both environments. The reason we used the word "similar" is because we believed due to difference between physical and virtual, a test run will never be identical.
We now have also added research questions and rephrased the unclear explanations.

Comment R3.43. *Lack of actionable findings. The paper repeats in several places that there is a discrepancy* *between the performance of virtual and physical environments. That is interesting, but it is not actionable. As a programmer or tester, I am not sure how to use this information. The paper does not give an explanation of how large the difference is, except by presenting the charts and figures. For example, in the last paragraph on page 9, "By looking closely at such metrics, we find..." Earlier it states that the "lines on the Q-Q plot are not close." I can see what the authors mean by looking at the figures, but is that a big difference? As a tester, I am already aware that the virtual environment is not a perfect reflection of a physical one. What I am wondering is, is the difference big enough that I need to worry about it? The paper does not quite answer that question. Perhaps the paper could go into an example to show how the results might affect decision-making, to give some context for understanding the results.*

Response. Please refer to our response to the comment R2.30. We also mention in Section 7 on how we have contributed based on this work. We revise and update the conclusions as follows:

- The performance test from virtual environments can not be reused so the results can not be transferred as-is across environments especially with the *MAPE* values as big as 600%.
- We also evaluate the method proposed by *Nguyen et al.* to use testing results in a different environment. Although, in their case, the environment are not virtual and physical.
- We propose a way that by *normalization by deviance*, practitioners may reduce the discrepancy present.

We now also provided a motivating example to help clarify the situation where our work's findings would apply.

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] Fabian Brosig, Fabian Gorsler, Nikolaus Huber, and Samuel Kounev. Evaluating approaches for performance prediction in virtualized environments. In *Modeling, Analysis & Simulation of Computer and Telecommunication Systems (MASCOTS), 2013 IEEE 21st International Symposium on*, pages 404–408. IEEE, 2013.

[2] 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.

[3] 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.

[4] Nikolaus Huber, Marcel von Quast, Michael Hauck, and Samuel Kounev. Evaluating and modeling virtualization performance overhead for cloud environments. In *CLOSER*, pages 563–573, 2011.

[5] Zhen Ming Jiang and Ahmed E Hassan. A survey on load testing of large-scale software systems. *IEEE Transactions on Software Engineering*, 41(11):1091–1118, 2015.

[6] 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.

[7] 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.

[8] JPM300. Vmware technology network. `https://communities.vmware.com/thread/484462`, 2014. Accessed: 2017-04-04.

[9] 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.

[10] Blue Link Associates Limited. Blue link. `http://www.bluelinkerp.com/blog/`, 2015. Accessed: 2017-04-04.

[11] 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.

[12] Aravind Menon, Jose Renato Santos, Yoshio Turner, G John Janakiraman, and Willy Zwaenepoel. Diagnosing performance overheads in the xen virtual machine environment. In *Proceedings of the 1st ACM/USENIX international conference on Virtual execution environments*, pages 13–23. ACM, 2005.

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

[14] 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.

[15] 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.

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

[17] 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.

[18] Eric Srion. The time for hyper-v pass-through disks has passed. `http://www.altaro.com/hyper-v/hyper-v-pass-through-disks/`, 2015. Accessed: 2017-04-04.

[19] 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.

[20] 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.

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

[22] 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.