Systematic Mapping Study on Performance Scalability in Big Data on Cloud Using VM and Container

Cansu Gokhan, Ziya Karakaya, Ali Yazici

Systematic Mapping Study on Performance Scalability in Big Data on Cloud Using VM and Container

Cansu Gokhan[1], Ziya Karakaya[2], and Ali Yazici[2]

[1] Atilim University, Institute of Natural and Applied Sciences, Ankara, Turkey
cansugokhann@gmail.com
[2] Atilim University, Faculty of Engineering, Ankara, Turkey
{ziya.karakaya,ali.yazici}@atilim.edu.tr

Abstract. In recent years, big data and cloud computing have gained importance in IT and business. These two technologies are becoming complementing in a way that the former requires large amount of storage and computation power, which are the key enabler technologies of Big Data; the latter, cloud computing, brings the opportunity to scale on-demand computation power and provides massive quantities of storage space. Until recently, the only technique used in computation resource utilization was based on the hypervisor, which is used to create the virtual machine. Nowadays, another technique, which claims better resource utilization, called "container" is becoming popular. This technique is otherwise known as "lightweight virtualization" since it creates completely isolated virtual environments on top of underlying operating systems. The main objective of this study is to clarify the research area concerned with performance issues using VM and container in big data on cloud, and to give a direction for future research.

1 Introduction

Big data applications continue to receive an ever-increasing amount of attention, thus they become a dominant class of applications deployed over virtualized environments[1]. On the other hand, the resource utilization feature of cloud computing is mostly based on virtualization techniques, which is the common way to run different services on the cloud[2]. By combining these two, most of the big data on cloud environments are using hypervisor to provision the virtual machines. In this technique, the VMs have their own operating systems which run on the virtual hardware resources provided by hypervisor[3]. Although it is proven to be a very useful technique in resource utilization, still there is an inherent overhead because of the hypervisor[1].

In recent years, containers, which are also called "lightweight virtualization", are gaining popularity due to their ability to offer superior performance because they do not have their own operating systems[2]. Instead, they use the OS kernel underlined with the host machine and they work similar to a regular application

and are completely isolated from each other as well as from the underlying system. This technique receives its popularity mostly in Linux OS virtualization, since it uses the features provided by Linux OS kernel itself, such as "cgroup", "namespace", etc., in order to completely isolate each container from the rest.

In this study, along with the other research questions, the main purpose of investigation was to identify if there is a gap in the literature and to what extend those techniques are being studied by using experimental approach.

There are three different databases used in this study to search for relevant papers. The authors found 308 papers that appeared to be relevant. After applying the inclusion and exclusion criteria, there were only 62 papers containing significant information either directly or indirectly related with the research questions.

The remainder of this paper is structured as follows : the related works on performance and comparison of VM vs. containers are discussed in Section 2. Section 3 presents our methodology and research questions investigated. In Section 4, the results are summarized and the findings together with discussion is given. Finally, the limitations of this study is appear in Section 5.

2 Related Work

Readers requiring in-depth information about the technologies and techniques used in virtual machines and containers together with their relationships to hypervisor and underlying operating systems are offered to read the white paper published by Intel[4].

There are many studies in the literature about big data on cloud environment. One of the Systematic Mapping (SM) studies conducted on this subject is the work of Ibrahim and his colleagues[5]. They have analysed the scalability issues of storage, but not the scalability issues of performance within the cloud. They have proposed a classification for big data, a conceptual view of big data, and a cloud services model. This model was compared with several representative big data cloud platforms. They have discussed the background of Hadoop technology and its core component, namely MapReduce, and Hadoop Distributed File System (HDFS).

Yanzhang focuses on the scalability performance issues of Hadoop Virtual Cluster with cost consideration[6]. They compared the scalability performance with respect to scale-up and scale-out methods under different workloads. Hadoop benchmarks and real parallel machine learning algorithms were used to evaluate the scalability performance. Their experimental results showed that the scale-up method outperformed the scale-out method for CPU-bound applications, and the opposite for I/O-bound applications. They also noted that disk and network I/O are the main bottlenecks of cloud platform due to shared resource contention and interference.

The most comprehensive work on performance comparison of virtual machines and Linux containers has been done by Felter and colleagues[3]. Their

goal was to isolate and understand the overhead introduced by virtual machines (specifically KVM) and containers (specifically Docker) relative to non-virtualized Linux on Cloud. They have concluded that both VM and containers are mature technologies, and that both have negligible performance overheads with respect to CPU and Memory performance. Nevertheless, they warned about the use of these technologies in case of I/O intensive works, which is the case in Big Data Application on Cloud.

Yang et al. have discussed the impact of virtual machine on Hadoop [7]. They describe the effect of different virtualization technologies such as KVM, Xen and OpenVZ on MapReduce environment. Also, they evaluate performance and stability of HDFS (Hadoop Distributed File System) on KVM, Xen and OpenVZ. Besides this, Pedro et al. have presented the performance of KVM and OpenVZ using micro benchmarks for disk and CPU[8].

There are many other papers in the literature that have studied the performance scalability of Big Data Applications. However, only a few of them have focused on performance scalability comparison of Container vs. VM technologies.

3 Research Methodology

A systematic map study was performed to obtain the current research map on the performance scalability issues in big data on cloud. The guidelines proposed by Peterson and colleagues[9] is followed in this study. The mapping study was conducted in three main stages, namely planning, execution and result.

3.1 Systematic Mapping Plan

In the planning stage, we defined research questions, search strategy, screening of papers for inclusion and exclusion, classification of papers and data extraction.

Research Questions The following research questions were identified as relevant to purpose:

1. *To what extend are the published papers on the performance scalability issues in big data on cloud are based on experimental study?*
2. *What is the percentage of the mostly studied technologies in big data performance scalability issues on cloud environment?*
3. *Which is the most investigated hypervisor in big data performance scalability issues?*
4. *What types of containers are being studied in big data on cloud?*
5. *Which components of the resources are mostly investigated for performance impact on big data analysis?*
6. *How frequent is the dominating technology being studied in the last five years as a tool in big data on cloud?*

Table 1. Selected Databases

Database	Location
IEEE Explore	http://ieeexplore.ieee.org/
Science Direct	http://www.sciencedirect.com/
ACM Digital Library	http://dl.acm.org/

Table 2. Inclusion and Exclusion Criteria

Inclusion Criteria
1 Studies addressing performance scalability issues in big data on cloud.
2 Journal and/or conference papers.
3 Studies that describe virtual machine and container types in the big data on cloud.
4 Primary or secondary studies.
Exclusion Criteria
1 Studies not accessible in full text.
2 Studies that do not address the performance scalability issues in big data on cloud.
3 Studies not presented in English.
4 Prefaces, slides, panels, editorials or tutorials.
5 Studies that do not answer the research questions.

Search Strategy The selected databases for the study are shown in Table 1 in order to identify potentially relevant conference articles and journal publications.

The following keywords were used in order to perform the search for the study: Big data, Cloud Computing, Performance, Scalability, Container, Virtual Machine, Comparision of VM vs. Containers. Search strings were applied to check keyword, title, and abstract fields in order to perform the automatic search in the selected digital libraries.

These strings are given as follows:

[(“Big Data”) AND (“Cloud Computing”) AND (performance OR Scalability) AND (Container OR VM OR “Virtual Machine”)]

Inclusion and Exclusion Criteria The aim of this process is to identify the most relevant studies for the mapping study. According to the research questions, the inclusion and exclusion criteria given in Table 2 were applied to the selected papers.

Classification of Papers The present work classified the papers according to properties and categories listed in Table 3.

Data Extraction In order to extract data from the selected studies, we designed a data extraction Excel table. Each selected paper appears as a record

Table 3. Classification Scheme

Properties	Categories
Research Approach	Theory, survey, review, experimental
Year	Years between 2009 and 2016
Article title	Name of the article
Container Type	Docker, OpenVZ, LXC, Linux-VServer
VM Type	Xen, KVM, Vmware (ESX,ESXi), Others
Technology	MapReduce, Hadoop, Spark, Storm, FLink
Component of Hardware Resource	CPU, Disk I/O, Network speed, Memory (RAM), # of VM/Container

item in this file. The data extraction table consist of article name, year of publication, technology, component of hardware resource, VM types and container types. Then, the data that is specifically related to research questions were extracted from each study.

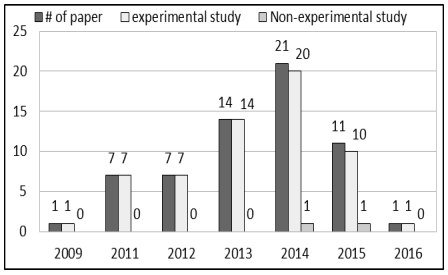
3.2 Execution of Systematic Mapping

At the execution stage, we conducted a systematic mapping study according to the plan stated in the previous section. The search string was modified for the different syntax as of databases according to the search criteria, and we have found 308 papers as candidate studies from all the selected sources. The title, abstract, and keywords were analysed, and then, some of the articles were eliminated by applying the exclusion criteria. In case of uncertainties as to inclusion of some papers, the introduction and conclusion sections of these articles were also taken into consideration. As a result of eliminating unrelated articles, 62 relevant research papers were selected.[3]

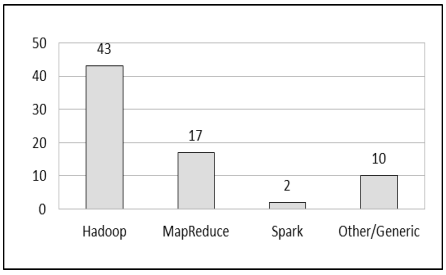
3.3 Results of Mapping

RQ1 Objective: The main objective of answering this question is to identify the proportion of experimental researches already done when compared to others.
RQ1 Results : Considering the performed study, 60 of 62 papers were based on experimental studies making them a majority. Fig. 1(a) shows the number of experimental and non-experimental studies with respect to publication years.
RQ2 Objective: The main objective of answering this question is to identify to what extend the mostly studied technology is dominating the research area.
RQ2 Results: The papers were categorized as follows: Hadoop, MapReduce and Spark. If a paper's study could not be defined with a specific technology, it is shown as the category called "Other/Generic".

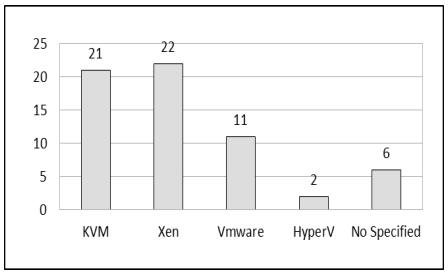
[3] List of articles : http://bit.ly/1Ux6H5M



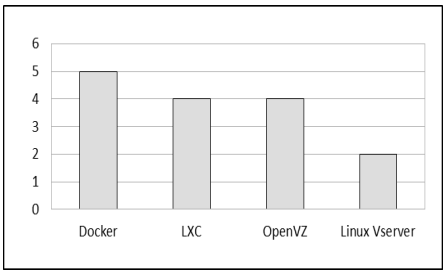
(a) Annual Publication count for each of frameworks



(b) Distribution of Technology



(c) Distribution of VM's



(d) Distribution of Containers

Fig. 1. Distribution of Technology and VM's

According to the results given in Fig. 1(b), we can conclude that the majority of these studies are based on MapReduce, since Hadoop is itself based on the MapReduce technique.

RQ3 Objective: The main objective of this question is to identify the most widely used virtual machine types for big data performance scalability issues.

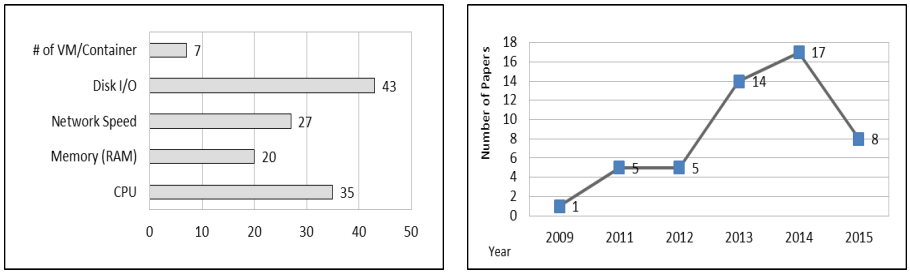
RQ3 Results: According to Fig. 1(c), KVM (Kernel based Virtual Machine) and Xen are found to be the mostly investigated VM type in Big Data performance issues. This is probably because of their open source nature.

RQ4 Objective: Our aim with this question is to find the dominating container technology being studied.

RQ4 Results: As shown in Figure 1(d), Docker is the most commonly studied container type in Big Data performance scalability issues, while LXC and OpenVZ are equally distributed.

RQ5 Objective: The aim of this question is to determine most commonly studied hardware component, to show its performance impact.

RQ5 Results: The most investigated and resulted components are CPU, Disk I/O, Memory and Network Speed. The findings are consistent with the needs of big data applications on cloud. Although the performance of memory intensive application could be influenced by existence of NUMA (Non-uniform Memeory Architecture), or by cache hits, there is no such study found.



(a) Distribution of components that are impacting the performance

(b) Frequency of the studies on MapReduce

Fig. 2. Components impacting performance and frequency of studies on Hadoop

Fig. 2(a), indicates that 56% of the total studies are focused on CPU, 32% on Memory (RAM), 44% on Network and 69% are focused on Disk I/O components. Also, this figure depicts the number of studies attributing the performance impact to number of VMs and that of containers is 7 (11%).

RQ6 Objective: The aim of this question is to determine the frequency of research on the subject within the last five years.

RQ6 Results: The result shows that 50 papers out of 62 are about the MapReduce based tool. According to this Fig. 2(b), it is easily seen that the number of papers which used MapReduce as a tool increased suddenly in 2013, and the frequency of papers has decreases after 2014.

4 Conclusion

In this paper, we used systematic mapping technique, to obtain a current research map on performance scalability issues in big data on cloud. The guidelines proposed by Peterson and colleagues[9] is followed. Of the 308 papers found in three popular databases according to our search string, 62 papers were found to be related either directly or indirectly with the research subject and questions. The papers are then analysed with respect to the research questions.

The following conclusions are deduced from the analysis of those related papers; (i) considering the performed study, 61 of 62 papers were based on experimental studies achieving a majority of the papers; (ii) the mostly studied technologies are Hadoop and MapReduce; (iii) KVM and Xen are the dominant hypervisor technology used in these studies; (iv) Docker is the mostly studied container type, and LXC and OpenVZ are the technologies that are used equally; and (v) CPU and Disk I/O are the two issues that are mostly handled when comparing these technologies. It is better to state that KVM and Xen are opensource, and Spark is relatively newer than Hadoop.

The most important finding of this research is that there are only a handful academic papers which compare the performance scalability of hypervisor-based virtual machines vs. containers for the big data applications on cloud. On the

other hand, although there are many researches about the comparison of either VM vs. "bare metal" (physical) on cloud, only three of them compares the performance in Big Data on Cloud. This was the second gap found during this study.

Another most important implication of this study shows that there is a lack of empirical study conducted on the other popular Big Data analysis frameworks, such as Spark, Storm, FLink, etc.

5 Limitations

In this study, only three databases are searched for relevant papers. These technologies have gained popularity over the past few years, and journal/conference papers may contain more than those of found in these databases. We believe that other databases should also be considered in such analysis, and we plan to further analyse those papers.

References

1. Mytilinis, I., Tsoumakos, D., Kantere, V., Nanos, A., Koziris, N.: I/O Performance Modeling for Big Data Applications over Cloud Infrastructures, In Proceeding of 2015 IEEE International Conference on Cloud Engineering, Pages 201-206, (2015).
2. Morabito, R., Kjllman, J., Komu, M.: Hypervisors vs. Lightweight Virtualization: a Performance Comparison, In Proceeding of 2015 IEEE International Conference on Cloud Engineering, Pages 386-393 (2015)
3. Felter, W., Ferreira, A., Rajamony, R., and Rubio, J.: An updated performance comparison of virtual machines and linux containers, Performance Analysis of Systems and Software (ISPASS), 2015 IEEE International Symposium On. IEEE, (2015).
4. Intel White Paper: Linux Containers Streamline Virtualization and Complement Hypervisor-Based Virtual Machines, http://www.intel.com/content/dam/www/public/us/en/documents/white-papers/linux-containers-hypervisor-based-vms-paper.pdf (2014).
5. Abaker, I., Hashem, T., Yaqoob, I., Anuar, N.B., Mokhtar, S., Gani, A., Khan., S.U.: The rise of 'big data' on cloud computing: Review and open research issues, The journal of Information Systems, Volume 47, Pages 98-115, January (2015).
6. He, Y., Jiang, X., Wu, Z., Ye, K., Chen, Z.: Scalability Analysis and Improvement of Hadoop Virtual Cluster with Cost Consideration, In proceeding of the Cloud Computing (CLOUD), 2014 IEEE 7th International Conference, 594-601, (2014).
7. Yang, Y., Long, X., Dou, X., Wen, C.: Impacts of Virtualization Technologies on Hadoop, In proceeding of Intelligent System Design and Engineering Applications (ISDEA), 2013 Third International Conference, Pages 846-849, (2013).
8. Vasconcelos, P.R.M, Freitas, G.A. de A. : Performance Analysis of Hadoop MapReduce on an OpenNebula Cloud with KVM and OpenVZ Virtualizations, In proceeding 9th International Conference for Internet Technology and Secured Transactions, ICITST, (2014).
9. Kai, P., Vakkalanka, S., Kuzniarz, L.: Guidelines for conducting systematic mapping studies in software engineering: An update, Information and Software Technology 64, 1-18, (2015).