Advanced Science and Technology Letters
Vol.35(Cloud and Super Computing 2013), pp.70-73
http://dx.doi.org/10.14257/astl.2013.35.17

An Empirical Study of Hadoop Application running on Private Cloud Environment

Yunhee Kang*, Kyung-Woo Kang*

Division of Information and Communication
Baekseok University
115 Anseo Dong, Cheonan, Korea 330-704
yunh.kang@gmail.com

Abstract. This paper describes how to leverage virtualized resources in a private cloud for handling MapReduce applications. In this paper OpenStack is used to build the private cloud. We show the relationship of a Hadoop application and virtualized resource allocations in the experiment. As a result of this experiment, the performance of Hadoop based compute intensive application provides up to 110.76% of the performance of physical server when running on the virtual machine in the private cloud.

1 Introduction

The virtualization technologies have been used as software technologies to increase the availability of high performance hardware computing resources in the cloud computing environment [1-2]. In addition, MapReduce programming model emerged as the distributed and parallel processing technology in order to handle data effectively [3-5].

This paper identifies the correlation between the application attributes and the resource allocation through the performance analysis of Hadoop application for the distributed and parallel process that are conducted in the virtualized cluster environment. For this, it is to generate the virtual machine instance after configuring private cloud using OpenStack [6]. It is to analyze the results after conducting Hadoop application in the virtual machine instances.

The structure of this paper is as follows: Second 2 describes the related studies of virtualization, OpenStack and MapReduce. Section 3 describes the experimental environment and the experimental results that include the physical servers and the virtual machine environment. Lastly, section 4 states the conclusion and the future plan.

ISSN: 2287-1233 ASTL

2 Related Works

This experiment establishes a virtualized environment with the hypervisor based full-virtualization. Full-virtualization has an advantage of supporting the guest operating system without having to modify the kernel by virtualizing hardware. Full-virtualization processes hardware-dependent specific commends where the guest operating system is operated between the guest operating system and the actual hardware by hypervisor. IO performance would rely on the hypervisor processing IO requests through emulating hardware through hypercall in the guest operating system. In the virtualized environment that is configured for this experiment, KVM (Kernel-based Virtual Machine) would be used as a hypervisor.

OpenStack [6] is a collection of open source components to deliver public and private clouds. OpenStack utilizes Python as a development language for open source project to build a private cloud computing environment. OpenStack is composed of such components as Nova, Swift, Cinder, KeyStone, etc. and these components are conducted as a separate project [9]. For instance, Nova supports the virtualization of computing resources and manages the virtual machine instance. Swift project provides storage service like S3 service of Amazon as an object storage project of OpenStack. OpenStack provisions the computing resources dynamically as a tool of IaaS (Infrastructure as a Service). It is provided as a service that forms the sub-domain structure of computing resources of the virtualized computing, storage and network.

Hadoop is a open source software based middleware on the basis of MapReduce [5]. Hadoop is broadly composed of the two components of HDFS that is the distributed file system and MapReduce that is the parallel programming model. MapReduce application of Hadoop is composed of Job that is the unit of task conducted by clients. Job is composed of input data, MapReduce program and setting information. In addition, Job is executed as being divided into Map task and Reduce task.

3 Experiment and Analysis

The configured experimental environment established the controller node to manage a virtualized environment in one server. In two servers, the compute node is established to generate virtual machines. The virtual machine instances generate 8 cores and 2 compute nodes having the memory of 48G byte. They were utilized in the experiment by composing them with Linux cluster. In this experiment, m1.small, m1.medium and m1.large were utilized among the resource types provided to VM in OpenStack.

This experiment utilizes HDFS computation and WordCount of Hadoop as data-centric application and PI compute application of Hadoop as compute intensive application. As for the performance result of HDFS computation and WordCount, the performance comparison was conducted in m1.small and the physical server. At this point, the performance analysis would be conducted by comparing the performance time of a single physical server and the performance time of cluster composed of virtual machines. The relevant computation was conducted repeatedly for 10 times to obtain and use the average value for the performance comparison.

HDFS is the distributed file system of Hadoop; thus, the reading and writing computations of HDFS are the mandatory task to prepare data and obtain results for Hadoop application. For HDFS computation, text files of 1GByte were utilized. As for the writing computation, it took 41.75 seconds in the virtual machine and 4.23 seconds in the physical server. Therefore, the performance rate was found to be 10.13 percent. As for the comparison on the writing computation, it took 37.77 seconds in the virtual machine and 5.03 seconds in the physical server; thereby, showing the performance of 13.67 percent as compared with the physical server. As for the performance of WordCount, it took 419.50 seconds in the virtual machine and 73.68 seconds in the physical server in order to compute the frequency after extracting words from the text files of 512 MByte. Therefore, the virtual machine was found to have the performance rate of 17.56 percent as compared with the physical server.

As for HDFS computation that is the data-centric application and WordCount performance results, IO processing in the virtual machine was realized through hypercall; thus, it was verified that it generated response delay. Through this, the fact that data-centric application has low performance in a virtualized environment is identified.

Table 1. Thee lapsedt imef orr unningM apt askso fH adoopP Ij ob:V Ma ndp hysical serv er

	No of Map Tasks	Task Size per Map	Performance Time (sec)			Performance Rate (%)	
			Virtual Machine		Physical Server		
			Type-1 (m1.medium)	Type-2 (2*m1.small)		Type-1	Type-2
1	1	100	16.59	17.09	14.45	87.14	84.55
2	1	10000	16.26	16.95	14.49	89.13	85.49
3	10	100	22.99	22.25	21.70	94.37	97.53
4	10	10000	25.57	22.22	22.08	86.37	99.37
5	100	100	128.15	104.91	116.20	90.67	110.76
6	100	10000	120.99	109.86	115.54	95.49	105.17
7	1000	100	1386.68	953.94	1052.05	-	110.28
8	1000	10000	-	955.53	678.725	-	71.03

In the case of Hadoop PI application, it was possible to obtain the performance of 87.14 percent of the physical server at minimum and 95.49 percent at maximum when performing by utilizing Type-1 virtual machine. In the case of performing by utilizing Type-2 virtual machine, it was possible to obtain the performance of 87.14 percent of physical server at minimum and 110.76 percent at maximum. In conclusion, it was shown that it would be effective to use the virtual machine for the compute intensive application performance, and it was possible to obtain higher performance by dispersing and performing the task load in the parallel performance such as Type-2 virtual machine than performing in the single physical server. Table 1 shows the performance rate of the configured virtual machine and physical server.

4 Conclusion and future works

This paper aims to find the correlation with the virtual resource allocation through the performance analysis in a virtualized environment affecting largely the performance of Hadoop application program that is conducted in a virtualized environment. Therefore, the experiment was conducted based on the setting values of the virtualized environment configuration. This paper compared the physical cluster and the performance after conducting the performance evaluation after configuring a variety of virtual machine instances in the configured cloud after configuring the private cloud using OpenStack. Hadoop PI application could obtain the physical server performance of more than 110.76 percent through this experiment in the case of performing by utilizing the virtual machine. As a result, it was possible through the experiment to find out that the compute intensive application was the element causing a significant change for the performance at both physical environment and virtualized environment; thus, it was the number of cores and the available memory.

References

1. M. Armbrust, A. Fox, R. Griggith, et al., "Above the cloud: A Berkeley View of Cloud Computing," Technical Report No.UCB/EECS-2009-28, EECS Department, University of California at Berkeley, USA, Feb.10, 2009.
2. Yunhee Kang, Geoffrey C. Fox, Performance Evaluation of MapReduce Applications on Cloud Computing Environment, FutureGrid Grid and Distributed Computing: International Conferences, GDC 2011.
3. J. Dean and S. Ghemawat, "MapReduce: A Flexible Data Processing Tool," Communications of the ACM, Vol. 53, pp. 72-77, 2010.
4. J. Ekanayake, et al., "MapReduce for Data Intensive Scientific Analyses," the 2008 Fourth IEEE International Conference on eScience 2008.
5. Hadoop. http://hadoop.apache.org/
6. OpenStack. http://www.openstack.org/