A Process Model for Optimized Cloud Configuration

[1]Sang Hwan Kung, [*2]Yunhee Kang

[1]*Division of Information and Communication, Baekseok University, Cheonan, Korea 330-704,*
kung@bu.ac.kr
[2]*Division of Information and Communication, Baekseok University, Cheonan, Korea 330-704,*
yhkang@bu.ac.kr

Abstract

This paper describes how to achieve an optimized computing platform with cloud configuration depending on the cloud applications. Because the Cloud is configured for numerous applications with different service features requiring different computing resources, the general strategies allocating computing resources may not fit the QoS requirements for each of applications. Therefore, it will be very helpful to know what kind of resource configuration on Cloud is good for an application. We introduce the procedure that determines the optimized Cloud resources for an application before it is running on the Cloud.

1. Introduction

Too much information goes beyond its usefulness and place new burdens with the selection of meaningful information and grasping the new shape of phenomena pertaining to a certain application. One of well-known application with this requirement is the scientific data analysis. Plenty of real-time measurement data gathered from the numerous sensors at the remote surveillance sites requires to be stored as well as be analyzed, in real-time for dynamic control of the environment or in a batch way for prediction of future weather. Either of two cases, the efficient data analysis is the main requirement because of the volume of the data [1-2].

The well-known solution for this problem is to use Cloud or Grid computing. Cloud computing is a distributed computing model that we don't need to depend on local resources for computing and storing of applications, which is designed to provide on-demand computing resources or services over the Internet with the reliability level of a data center[3-4]. In cloud computing environment, a virtual machine (VM) is a computing platform that creates a virtualized layer between the computing hardware and the application. Cloud providers such as Amazon, Google are providing compute and storage resources via virtual machines as IaaS to users Conceptually applications acquire computing platforms or IT infrastructures from clouds and then run them on the inside clouds. Virtualization technologies partition hardware and consolidate utilization of server workload so that it is used to reduce the actual number of physical servers and to improve scalability and elasticity according to workload in the cloud environment. Hence virtualization technologies are the bases of cloud computing. A main feature of cloud computing is Service-Oriented Architecture (SOA) [5] as a platform that support a wide range of demanding applications of which those a cloud can offer are delivered in the form of services. These services are offered to users via the Internet. Cloud services are often provided high elasticity with reconfiguration by on-demand manner. To manage those services, it needs to monitor quality of them periodically or sporadically when changing their status like availability, performance and security policy. Hence we develop a management process our service provisioning that is a way of match making between application workloads and resources provided. This paper addresses the process to manage an application running on cloud. It is based on the Quality of Service (QoS) measurement of the process. It is used to understand application's characteristics.

---

2. Related Works

Cloud computing can be defined as a large scale distributed computing paradigm that is driven by economics of scale where a pool of abstracted, virtualized, dynamically-scalable computing functions and services are delivered on demand to external customers over the Internet. Cloud computing is a general term for delivering hosted services over the Internet like Infrastructure as a Service(IaaS), Platform as a Service(PaaS) and Software as a Service(SaaS). Examples of these services give a new opportunity for developing new applications within clouds, satisfying the required QoS[6].

QoS is a broad term encompassing the means by which to predict and manage a variety of system-wide resources that are important to the runtime performance of an application.

Although typically associated with network resources, QoS is equally applicable to other limited computing resources. Service consumers are required to obtain guarantees related to services they use, often related to QoS. They are bound to a service that best meet their functional and non-functional requirements[6]. The specification of QoS provides a way to distinguish services offering equivalent functionality [10].

In cloud computing, resource providers use diverse QoS service models, which are based on measuring the status of resources provided and those level based on QoS. The QoS service models are used to predict the service level of the cloud in the future.

SLA is very good means to match the requirements of both of Cloud computing stake holders, service providers and service consumers. SLA metrics, framework providing negotiated QoS, and negotiation strategies are critical elements of realizing optimized QoS[4][7][8][10].

3. The Proposed Scheme

In order to decide the optimized resources for applications on Cloud, above all we need to get the QoS measurement of the application on the non-Cloud machine. This measurement such as performance of the application might be the good reference for the efficiency level of the application on the Cloud machine. After that, we can take the test of the application on the Cloud. This job might be done recursively to best parameters for computing resources. Every test result is to be compared with the results of the test previously obtained.

All of these works will be done after understanding of the application itself and defining the Cloud machine itself. The whole process is described in the Figure 1.
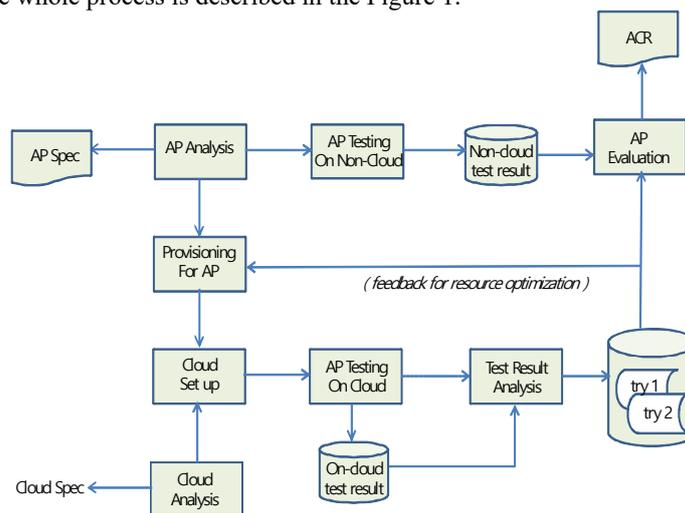

Figure 1. Management Process of Application on Cloud

- Application analysis

    The aim of this process is to understand application's characteristics for allocating the optimized resources while the application is executing on the Cloud. This process results in Application Specification, which includes followings:

    - Application Summary
    - Resource Requirements
    - Resource Sensitiveness

- Application test on the non-cloud machine

    The cloud machine is designed for many of applications being executed together sharing the computing resources. So, in many cases, it is possible for application on the Cloud not to obtain better QoS than while it is running on the non-Cloud. The result of this test is to give us the important guideline when we judge the necessity of Cloud environment.

- Provisioning of Cloud for application

    Application Analysis is a process to allocate the appropriate resources on Cloud for application. The resource types for applications are as following:

    - CPU: Speed, utilization
    - Memory: Size, utilization
    - Disc: Speed, Capacity
    - Network: Bandwidth/Speed

    Provisioning is done initially after the analysis of application characteristics and also after the analysis of application test result on Cloud. In order to find out the requirement for the more optimized resources, several times of recursive testing is necessary. For each test, the amount of resource in detail is defined in provisioning process.

- Cloud analysis

    There are many types of cloud that have the different hardware capability, software platform and tools. This sort of characteristics of Cloud may influence the QoS, such as performance, of the application. So, the Cloud environment for application test is important in the view of getting the more fair judgment.

- Cloud set up

    The result of provisioning is executed on the Cloud, by setting up processor, memory, and communication bandwidth.

- Application testing on Cloud

    We use the tools like Ganglia, which is to write the QoS monitoring result in the database.

- Test Result Analysis

    Testing result is represented as QoS, such as execution time of application execution, cpu utilization for a time extent, utilized memory and network bandwidth, and so forth. These are used to judge the optimized resource capacity and another test trial parameter represented as resource capacity. Since application testing is carried out multiple times to get best appropriate values, test results are accumulated in the database for further analysis.

- Application evaluation

    The previous works and all of results are for this process. The result of application evaluation is described in the Application-Specifications-on-Cloud Report (ACR). It is composed of three major contents below:

    - Cloud specification

- Application specification
- Application characteristics on Cloud
- Reaction/Results on Cloud Configuration
- Recommendations for Cloud

4. Conclusion

This paper addressed the design of resource configuration process for applications on cloud. In the designed process, the application providers can recognize optimized resources for applications. This process can be useful for service provisioning tasks with way of match making between application workloads and resources provided. Although the dynamic configuration control for resource allocation in running time, this process is very good for determining resource parameters when the application is loaded on the Cloud. Another expectation with the proposed scheme is to make it used as a SLA tool.

References

[1] Paul Zikopoulos, Chris Eaton. "Understanding Big Data: Analytics for Enterprise Class Hadoop and Streaming Data (1st ed.)," in *McGraw-Hill Osborne Media*, 2011.

[2] Divyakant Agrawal, Sudipto Das, and Amr El Abbadi. 2010. "Big data and cloud computing: new wine or just new bottles?," in Proc. VLDB Endow. Vol. 3, No. 2, 2010, pp.1647–1648.

[3] Armbrust, M., Fox, A., Griffith, R., Joseph, A.D., Katz, R.H., Konwinski, A., Lee, G., Patterson, D.A., Rabkin, A., Stoica and I., Zaharia, M., "Above the Clouds: A Berkeley View of Cloud Computing," EECS Department, University of California, Berkeley, 2009.

[4] A. V. Dastjerdi and R. Buyya, "An Autonomous Reliability aware Negotiation Strategy for Cloud Computing Environment", 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, 2012.

[5] G. Alonso, F. Casati, H. Kuno, and V. Machiraju. Web Services Concepts, Architectures and Applications. Springer Verlag, Berline Heidelberg, Germany, 2004.

[6] R Calinescu, L Grunske, Marta Kwiatkowska, Raffaela Mirandola, Giordano Tamburrelli. "Dynamic QoS Management and Optimisation in Service-Based Systems," IEEE Transaction on Software Engineering, 2011,37(3).

[7] Liang Jhao, Sherif Sakr, Anna Liu, "A Framework for Consumer-Centric SLA Management of Cloud-Hosted Databases", IEEE Transactions on Services Computing, 2013.

[8] Mohammed Alhamad, Tharam Dillon, Elizabeth Chang, "Conceptual SLA Framework for Cloud Computing", IEEE DEST 2010, 2010.

[9] R Calinescu, L Grunske, Marta Kwiatkowska, Raffaela Mirandola, Giordano Tamburrelli. "Dynamic QoS Management and Optimisation in Service-Based Systems," IEEE Transaction on Software Engineering, 2011,37(3).

[10]   Rocco Aversa, Luca Tasquier, Salvatore Venticinque, "Management of Cloud Infrastructures through Agents", Third International Conference on Emerging Intelligent Data and Web Technologies, 2012.

[11]   D. A. Patterson. "Technical Perspective: The Data Center is the Computer". Communications of the ACM, Vol. 51, Issue 1, pp. 105-105, January 2008.

[12]   B. Hayes, Cloud computing, Communications of the ACM, Vol. 51, No. 7, pp. 9-11, July 2008.

[13]   FutureGrid, FututrGrid Portal. Available: https://portal.futuregrid.org/

[14]   Yun-Hee Kang, "Construction of a MapReduce Application Running on Twister in Cloud Computing Environments, FutureGrid", Journal of KIIT, Vol. 9 No. 4, pp. 147-154, April 2011.

[15]   M. Mahjoub and A. Mdhaffar, et al. "A Comparative Study of the Current Cloud Computing Technologies and Offers", IEEE First Symposium on Network Cloud Computing and Applications, pp. 131-134, Nov. 2011.