Service Effectiveness and user requirement for Cloud Service Allocation

Varinder Singh

Abstract— Cloud architecture provides a model to distribute the services to clients. The presented work is the improvement to this distribution architecture in case of public cloud. The improvement is here performed at both the server as well client end. The work is divided into three main stages. First stage is specific to the server, in which all the available clouds are divided into a hierarchal order. A high level parametric division is performed to categorize all the related clouds. The client search will be performed on the segmented group instead of all clouds. The second stage is client specific, in which client requests are maintained in a queue, and an effective scheduling mechanism is implemented to select the best client to be processed. The third stage is the integration stage called allotment of service. In which a parametric check is performed relative to efficiency and accuracy to select the best cloud service from the segmented group to the client. The presented architecture will reduce the load from both the client side as well as server side, and will perform the efficient and reliable cloud service allocation.

Index Terms— Cloud Architecture, Scheduling, Allocation, Parametric Selection, SLA, QOS, etc.

1 INTRODUCTION

Cloud based architecture is one of the major distribution models to share the products or the services in private as well as in public sector. With the development and involvement of the Internet in all business architecture, there was the requirement for improvement of basic distribution models such as distributed computing, parallel computing, grid computing, etc. The cloud architecture is itself a layered architecture as shown in figure 1. In this architecture, the top layer is the client that performs the service or the product request in a user friendly manner.

The lowest layer of the architecture is the cloud server. The cloud server contains the cloud services. These services are of a different kind such as storage services, product management services, product distribution services, etc. Between these two ends, there are numbered of in between layers. These intermediate layers include the process of web management, management of cloud infrastructure, etc.

As we can see in the figure, all the services provided by the cloud and the integration is shown is divided into three main sub layers. The Cloud Application is the actual client end that defines the client integration with cloud. Here client specifies its requirement and the service request.

In the intermediate layer, the cloud infrastructure is defined. The cloud infrastructure includes the computational model, storage and the other communication resources. The computation model is the actual processing unit that will return the desired results from the system; the storage unit is the actual memory available to the cloud to process on.

- *Varinder Singh is currently pursuing masters degree in Information Technology from Lovely Professional University, Phagwara, India, PH-09041513193. E-mail: varindersaini1990@gmail.com*

The communication units define the communication protocol, security, etc. while performing the data transmission.

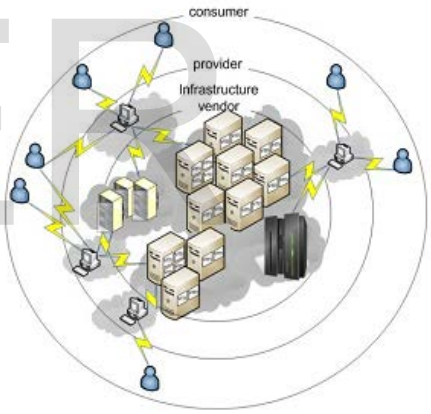

Figure 1: Cloud Architecture [6]

In this present work, we have defined an effective and improved model to represent the public cloud. The work is defined under some constraints and the assumption. The work is divided into three main stages, one for the server side, second for client side and third for the integration stage. The presented model will allow a most eligible client to get the services efficiently and to provide the allocation of most reliable and efficient cloud service. In section 2, we have defined the literature respective of the presented work. In section 3, the proposed work is defined.

2 LITRATURE SURVEY

Many of the earlier research did a lot of work for the cloud service allocation. Same kind of work is also done in other distribution architectures such as parallel processing and grid computing. The presented work is the composition of three

main stages called scheduling, effective service selection and the effective allocation of the resource. In this section, the work done by the earlier researchers is discussed. In year 2011, G. Mc Evoy performed a work to expose and explore the different paradigms of the scheduling in case of cloud computing under different application. The author work is based to present different architectural analysis of cloud computing and the cloud infrastructure. These architectures are defined under different parameters such as work load, efficiency, etc. The author presented a virtual mechanism for workload consolidation in performance analysis for scheduling. The author defined classification approach to take the early decisions regarding the categorization of resources and services and based on this classification the actual scheduling will be performed [1].

Another work in the same year is performed by D. Dutta for Job Scheduling in Cloud Computing in Multi QOS environment. In this work, the scheduling is been analyzed for the business services based cloud architecture. The author presented the optimization of cost based job scheduling using the genetic approach. The work is been tested under different crossover operators likes PMX, OX, CX, etc. The changes are also performed for different mutation operators like swapping and insertion operators. The work is also compared with linear programming approach [2].

In year 2012, Octavian Morariu [10] presented the same kind of work to optimize the work load balanced scheduling using the genetics approach. Danial Guimaraces do Logo performed a cloud process scheduling by using virtual machine scheduling using the active cooling control mechanism. The work also explored the concept of concept of green computing along with the virtual machine algorithms. The presented work was about to reduce the power consumption by implementing a control mechanism on work load assignment and the migration. The control mechanism presented was based on the threshold value. The work is based on the concept of energy utilization and to improve the effective scheduling for the heterogeneous data centers [3].

In year 2010, H. Kloh presented a scheduling model that can be implemented for both the grid computing as well as cloud computing. The work is the implementation as well as analysis for the bi-criteria hybrid scheduling algorithm to optimize the quality of service for the selection of job. The presented model was based on the prioritization criteria and ordering of the scheduling approach [4]. In year 2012 and energy effective parallel computing approach is implemented under the cloud architecture. The work is about the reduction of energy or the power consumption. The presented model can be implemented as the service distribution architecture or the business model for any kind of services. In this work, author proposed an Energy-Efficient Scheduling, the work included the service level agreement for the job assignment and significant power saving for the algorithm [5] [9].

Another work is performed by Young Choon Lee for the profit-driven scheduling for the request cloud. The primary focus of the work is on the service allocation based on pay-peruse concept. According to this pricing model, the providers and the clusters are working with the requested service volume and the conflicting objectives for both the providers as well as the customers. The work includes the service request scheduling under the business model environment. The architecture includes the vendors, service providers and the customers. The scheduling approach is capable to satisfy the need of both parties based on price analysis. The pricing model is the dependency based consideration for the development of profit driven scheduling algorithm [6].

In year 2012 another work is performed by Lma Mingyi Zhang on the green scheduling to optimize the scheduling task for the heterogeneous cloud servers. The work is about to achieve the energy effectiveness and to get the pollution stability with lower energy usage. In this presented work, six different green scheduling algorithms are considered implemented in two major steps, one for the task assignment to different cloud servers based on the energy analysis and other for setting up the optimal spend for task assignment for each cloud. The presented work can be implemented for both the homogenous and the heterogeneous cloud servers. The work simulation is based on the shortest task first and the energy-efficiency enhancement approach [7].

In year 2011, Jiahui Jin performed a data locality based task scheduling algorithm for the cloud computing system like Hadoop and MapReduce. These kinds of systems are implemented for the business environment. The file system can be split for the multiple blocks that will be replicated for the multiple servers. This block architecture will be implemented to get the effectiveness for the global optimization for the data locality. The presented approach is a heuristic task scheduling mechanism that includes the initial task distribution approach [8].

3 PROPOSED WORK

The presented work is the improvement over the service level architecture of Cloud. The work is about the refinement to perform an efficient and reliable resource allocation under the cloud architecture. The presented work is shown in the form of standard cloud architecture as shown in figure 2.

As we can see in the figure, here the cloud server is presented in the form of integrated cloud services. The information of these all services is defined in service database defined by the cloud. This service DB contains the information regarding, the cloud services name, its basic architectural properties, security policies, access methods, cost, etc.

Based on these parameters the user oriented choice will be filtered later by the SLA. The lowest form of the architecture is been represented by Client side. Multiple clients can demand for a particular cloud service according to the requirement. The request user request performed at a particular instance of time will be recorded in a queue called process queue. On this queue, the scheduling mechanism will be performed.

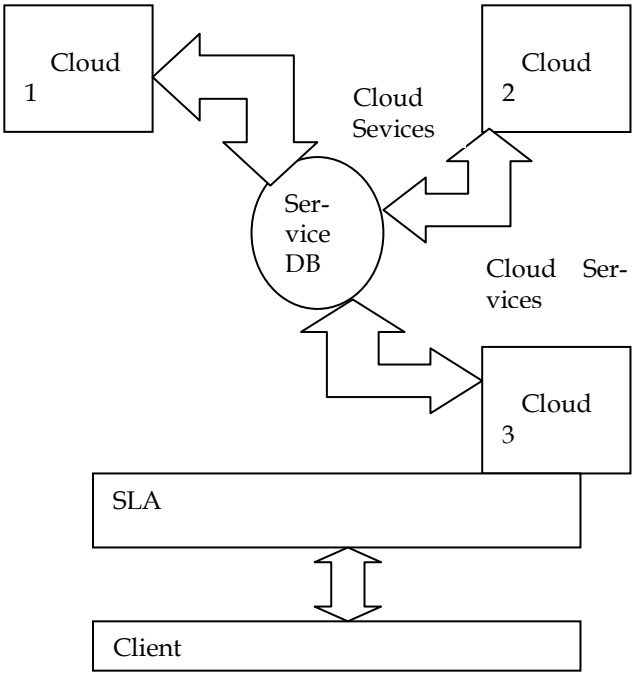

Figure 2: Proposed Cloud Architecture

The scheduling will be performed based on the client request, resource requirement and the time required performing the process. Once the processes are scheduled, these will be taken by the SLA as the middle layer. The SLA will generate a resource allocation algorithm to provide the service access to a particular user.

The service level agreement is the major component in the cloud architecture. According to this architecture, the criteria and conditions are defined for assignment or allocation of the cloud services to different consumers. In this present work, we have presented the same work by using the layered approach. Figure 3 is showing the process performed by the middle layer. The layer will accept the user request and the cloud service description as the input arguments.

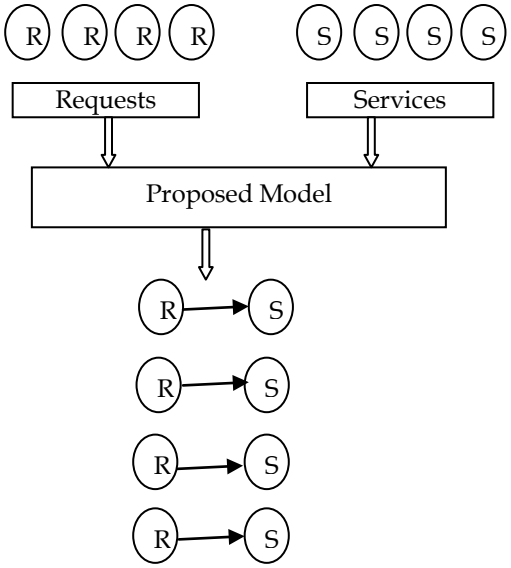

Figure 3: Service Level Agreement

3.1 Layer 1

It is the client side layer that basically accepts the client request in the form of the service requirements along with its processing time and the type of service. As the requirement is submitted by the user, it is maintained in a queue called request queue. On this queue, a scheduler will be implemented to identify the request that will be processed first. The scheduling procedure is based on the time of the process and the type of process. The scheduling algorithm is given here under.

1. Define a Request Queue with capacity N
2. Generate a user request called Ri and insert it in the queue from the rear end.
3. p=1
4. For i=1 to length(RequestQueue)
5. {
6. If(Priority(Request(i))> Priority(Request(p)))
7. {
8. if TimeTaken(Request(i))<TimeTaken(Request(p))
9. {
10. p=i
11. }
12. }
13. Return Request(p)

3.2 Layer 2

The cloud server itself represents the second layer. The most the server level cloud information is present in the service DB. As the integration of this layer is performed, different cloud services are been compared under different parameters.

1. Type of Service: The comparison will be performed only between the same kinds of clouds. Such as the when we have to work on a mail service, all the mailing service cloud will be compared. The type of service is the basic parameter to perform the categorization.
2. Availability: It is not necessary that all clouds are available all the time. To check the availability of the cloud, a dummy cloud access will be performed. If the cloud service signature identified, it means the service is activated and currently available.
3. Response Time: All the clouds of same type will be compared to the efficiency. The efficiency is here defined in terms of time taken to access a cloud service. All the clouds will be arranged in descending order of the response time.

Based on these parameters most appropriate clouds will be elected based on the user requirement.

3.3 Layer 3

This layer is basically the interaction layer that allocates the efficient and reliable cloud service to the effective users. Based on the availability and the requirement, the service allocation will be performed.

We are approaching the concept of open cloud to integrate different services under one unit. The basic principle of model many clouds will continue to be different in a number of important ways, providing unique value for organizations. It is not our intention to define standards for every capability in the cloud and to create a single homogeneous cloud environment. Rather, as cloud computing matures, there are several key principles that must be followed to ensure the cloud is open and delivers the choice, flexibility and agility organizations demand. In this proposed model, the Google Apps is considered as the public service provider and

4 CONCLUSION

The work is about to define an integration model that will divide the complete architecture in three interconnected layers. Each layer is defined with a separate algorithmic concept to identify the most required user and the best available service. Based on these services, the integration and the allocation of the services to different users is performed. The presented work is the model, to provide the efficient and the reliable service allocation mechanism among all the available public cloud.

REFERENCES

[1] G. Mc Evoy," Understanding Scheduling Implications for Scientific Applications in Clouds", MGC'2011 978-1-4503-1068-0/11/12
[2] D Dutta," A Genetic –Algorithm Approach to Cost-Based Multi-QoS Job Scheduling in Cloud Computing Environment", International Conference and Workshop on Emerging Trends in Technology (ICWET 2011) – TCET 978-1-4503-0449-8/11/02
[3] Daniel Guimaraes do Lago," Power-Aware Virtual Machine Scheduling on Clouds Using Active Cooling Control and DVFS", MGC'2011 978-1-4503-1068-0/11/12
[4] H. Kloh," A Scheduling Model for Workflows on Grids and Clouds", MGC '2010 978-1-4503-0453-5/10/11
[5] Qingjia Huang," Enhanced Energy-efficient Scheduling for Parallel Applications in Cloud", 2012 12th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing978-0-7695-4691-9/12© 2012 IEEE
[6] Young Choon Lee," Profit-driven Service Request Scheduling in Clouds", 2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing 978-0-7695-4039-9/10 © 2010 IEEE
[7] Luna Mingyi Zhang," Green Task Scheduling Algorithms with Speeds Optimization on Heterogeneous Cloud Servers", 2010 IEEE/ACM International Conference on Green Computing and Communications & 2010 IEEE/ACM International Conference on Cyber, Physical and Social Computing 978-0-7695-4331-4/10 © 2010 IEEE
[8] Jiahui Jin," BAR: An Efficient Data Locality Driven Task Scheduling Algorithm for Cloud Computing", 2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing 978-0-7695-4395-6/11 © 2011 IEEE
[9] Nakku Kim," Energy-Based Accounting and Scheduling of Virtual Machines in a Cloud System", 2011 IEEE/ACM International Conference on Green Computing and Communications 978-0-7695-4466-3/11 © 2011 IEEE
[10] Octavian Morariu," A Genetic Algorithm for Workload Scheduling In Cloud Based e-Learning", 978-1-4503-1161-8