Journal of Theoretical and Applied Information Technology
20th June 2014. Vol. 64 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

QOS BASED COST EFFICIENT RESOURCE ALLOCATION IN CLOUD

S.P.JENO LOVESUM [1] DR.K.KRISHNAMOORTHY[2] BLESSED PRINCE. P[3]

[1]Assistant professor (SG), Department of Computer Science and Engineering, Karunya University, India

[2]Professor & Head Department of Computer Science and Engineering, Sudharsan College of Engineering, India

[3]Assistant professor (SG), Department of Information Technology, Karunya University, India

E-mail: [1]jenolovesum@gmail.com, [2]kkr_510@rediffmail.com, [3]blessedprince@gmail.com

ABSTRACT

Resource allocation is one of the important challenges in cloud computing environment. It depends on how to allocate the resource to the particular task. Resource allocation can be done by two methods. One of the method statically allocates the resources and other dynamically allocates the resources. Two methods are having advantages and disadvantages according to its processing environments. In this paper we are considering the QOS parameters specified by the user to allocate the resource so that the resource allocation can be done in a cost effective manner. As cost is an important factor we try to minimize the memory wastage which is taken as a QOS factor and the algorithm tries to allocate the resource in such a way that there is minimal memory wastage. Most of the previous methods consider only cost and time as QOS parameters. But in our method we have considered memory wastage, throughput, response time, cpu utilization, as the QOS parameters which has to be satisfied for the users. So our method handles multi QOS requirement tasks efficiently.

Keywords: *Cloud, VM, Resource, QOS, Memory*

.

1. INTRODUCTION

Cloud computing is the delivery of computing services over the internet. Cloud services allow individuals and businesses to use software and hardware that are managed by third parties at remote locations. Cloud computing providers offer their services according to several fundamental models: infrastructure as a service, platform as a service, and software as a service. In Software as a Service model, a pre-made application, along with any required software, operating system, hardware and network are provided. In Platform as a Service, an operating system, hardware and network are provided, and the customer installs or develops its own software and applications. The Infrastructure as a service model provides just the hardware and network; the customer installs or develops its own operating systems, software and applications.

Some of characteristics of cloud are as follows

- On demand self services

Computer services such as email, applications, network or server service can be provided without requiring human interaction with each service provider. Cloud service providers providing on demand self services include Amazon Web Services (AWS), Microsoft, Google, IBM and Salesforce.com. New York Times and NASDAQ are examples of companies using AWS (NIST).

- Broad network access

Cloud Capabilities are available over the network and accessed through standard mechanisms that promote use by heterogeneous thin or thick client platforms such as mobile phones, laptops and PDAs.

- Resource pooling

The provider's computing resources are pooled together to serve multiple consumers using multiple-tenant model, with different physical and virtual resources dynamically assigned and reassigned according to consumer demand. The resources include among others storage, processing, memory, network bandwidth, virtual machines and email services. The pooling together of the resource builds economies of scale (Gartner).

- Rapid elasticity

Cloud services can be rapidly and elastically provisioned, in some cases automatically, to quickly scale out and rapidly released to quickly scale in. To the consumer, the capabilities available for provisioning often appear to be unlimited and can be purchased in any quantity at any time.

- Measured service

Cloud computing resource usage can be measured, controlled, and reported providing transparency for both the provider and consumer of the utilized service. Cloud computing services use a metering capability which enables to control and optimize resource use. This implies that just like air time, electricity or municipality water IT services are charged per usage metrics – pay per use. The more you utilize the higher the bill. Just as utility companies sell power to subscribers, and telephone companies sell voice and data services, IT services such as network security management, data center hosting or even departmental billing can now be easily delivered as a contractual service.

- Multi Tenacity

It is advocated by the Cloud Security Alliance. It refers to the need for policy-driven enforcement, segmentation, isolation, governance, service levels, and chargeback/billing models for different consumer constituencies. Consumers might utilize a public cloud provider's service offerings or actually be from the same organization, such as different business units rather than distinct organizational entities, but would still share infrastructure.

The cloud falls into variety of categories like Private cloud, public cloud, community cloud, and hybrid cloud are the various types of cloud services. Public cloud is offered over the internet and are owned and operated by cloud provider. Private cloud, the cloud infrastructure is operated solely for a specific organization, and is managed by the organization or a third party. In a community cloud, the service is shared by several organizations and made available only to those groups. A hybrid cloud is a combination of different methods of resource pooling. Virtualization is the process of running multiple operating systems on a single physical system and shares the underlying hardware resources.

Resource allocation is a major issue in cloud computing environment. Resource allocation has variant level of issues like scheduling task, computational performance, reallocation, response time and cost efficiency. To accomplish the task with the lowest price is the important issue in cloud computing. Resource allocation is the process of providing services and storage space to the particular task given by the users. In this paper a cost efficient resource scheduling method has been discussed and implemented by considering the QOS parameters that the users specify for their tasks.QOS is taken as the major parameter to perform resource allocation. The tasks submitted by the users consists of multiple QOS requirements that has to be met. The proposed method considers all the QOS specified by the users while allocating the resources to a particular tasks.

2. RELATED WORK

(Shin-ichi Kuribayashi. 2011) says that the optimization of resource allocations are very important for provide economically feasible cloud services. To achieve the effective processing ability and bandwidth for each service request using optimal resource allocation method. The optimal resource allocation methods are Best-Fit approach and Round Robin used to perform efficient resource allocation for every users task.

(Hadi Goudarzi and Massoud Pedram 2010) proposed a multi-dimensional algorithm to arrange large-scale jobs submitted to cloud in order to optimize resource allocation and reduce cost is an issue of common concern. To allocate method should decide which virtual machines should be assigned with a new set of jobs. The use of lightweight node to allocate resource is considered as performance metric. Parallelization and Scheduling Strategies are used to construct an execution Graph

(Rostand Costa and Francisco 2010) says that the Capacity planning for the provision of cloud data centers are very complicated. To optimize resource usage and to reduce the number of idle resources, a perfect solution is to set a time interval and alter resources as persistently keeping with workload changes

(Sowmiya.S. 2012) proposed an economic resource allocation model has been criticized for their performance limitations and high overheads. To improve the poor performance of current economic allocation models, by reducing the failure rate, reallocation rate.

(Gihun Jung and Kwang Mong Sim 2012) says that the performance degradation according to response and consider the location of physical machine to allocate user request as a virtual machine. To design the hybrid resource management architecture to perform location aware VM placement and dynamic resource utilization management. Response time should be increased.

(Preeti Agrawal and Yogesh Rathore 2011) says that the resource allocation is the issue in the cloud computing environment. In high performance computing providing the adequate resources to the user applications is difficult.The proposed model which is efficiently reallocates the resource to get complete. Job scheduling system plays a very important role in how to meet Cloud computing users' job QoS requirements and use cloud resources efficiently in an economic way.

(Shaminder Kaur et al., 2012) proposed to design the improved genetic algorithm developed by merging the scheduling algorithm to improve the cost efficiency and performance. The performance metrics of Resource utility, Task scheduling are considered in this paper. Modified Genetic Algorithm (MGA) Generate an initial population of individuals with output schedules of algorithms Longest Cloudlet to Fastest Processor (LCFP), Smallest Cloudlet to Fastest Processor (SCFP) and 8 Random Schedules.

(G.Sireesha and L.Bharathi 2012) says that the parallel data processing has emerged to be one of the major issue applications for Infrastructure-as-a-Service (IaaS) clouds. A new processing framework designed for exploiting the dynamic resource allocation offered by IaaS clouds for both, task scheduling and execution.

Improved cost-based algorithm[5] for task scheduling which uses Batch mode as scheduling method. Cost and performance are the two parameters passed for this algorithm. It is used to measure both resource cost and computation performance.

RASA[7] Workflow scheduling which uses Batch mode as scheduling method. The make span parameter passed for this algorithm. It is used to reduce make span.

The other algorithm known as SHEFT algorithm[3] which uses dependency mode as scheduling method and concentrates on the scheduling parameters such as execution time and scalability. It is used for optimizing workflow execution time as well enables resources to scale elastically during workflow execution.

Although the above mentioned algorithms have benefits of some QOS, None of the algorithms have concentrated on the multiple QOS parameters such as minimum memory wastage, throughput, CPU utilization , response time.

3. SYSTEM ARCHITECTURE

In this paper a QOS based resource allocation is carried out. The user's task should meet the SLA constraints specified by the provider if so the tasks are stored in the task queue. Efficiency in cost can be determined by the QOS parameters specified by the user. VM manager carries out the monitoring process. It has two tables of resource allocation rate table and resource remaining rate table. Resource allocation rate table can provide the information about utilization rate of resources. And remaining resource rate table contains information about unutilized resources. While finding the best resource for allocation there may be one or more resource of the same capacity or they will be eligible to serve the request of the user. When there is a tie of this manner it is difficult to choose the best resource so that an optimum and cost efficient allocation can be achieved. That is why resource allocation and resource scheduling becomes NP hard problems becomes we have to find the best fit resource from the group of fit resources that can be allocated to the task. As we have considered storage cloud which falls under the category of IaaS cloud, the QOS parameters that we have considered are Memory space wastage, response time, CPU utilization and throughput in addition to this for cost efficiency we have considered upload time and down load time which is initially calculated for uploading a file or downloading a file. The service provider calculates this depending upon the file size and his bandwidth. This becomes advantageous when there is a low bandwidth for the provider because the user will not get affected and though uploading and downloading consume more time during low bandwidth the user need to pay only the cost that is initially calculated by the provider. Also if this time falls below a particular threshold the provider has to pay a penalty which is calculated based on the threshold value and the available bandwidth.

The VM manager classifies the resources into three class based on the capacity of each VM. This is done by using the resource allocation rate table and resource remaining rate table. Based on the remaining resource at a particular time the VM are shifted within this class by the VM manager. The classes are platinum class which contains the VMs with the highest resources like CPU utilization ,memory etc. The next is the gold class which contains the set of VMs with an average amount of resources. The last is the silver class which contains the VMs with the lowest level of resources at a particular time. Once a VM which is in a silver class has finished its computing or servicing a task that has been allocated to it and if its resources gets free then then based on the resources it is having at that time it can been shifted to either platinum class or gold class. Similarly if a VM in the platinum class is allocated some tasks and if their resources are under use then until they finish their processing they will either be in gold or silver class. So this processing is taken care by the VM manager who continuously keeps monitoring the resource table. This grouping of VM is done so that the resource allocation time can be minimized when a task is accepted for computing. Based on the level of QOS the task is expecting the tasks are classified into high, medium and low priority tasks and they are kept in three different arrays for execution. So when a task from the high priority array arrives for execution the VM manager immediately allocates resources from the platinum class for that tasks which in term reduces the response time and resource allocation time.

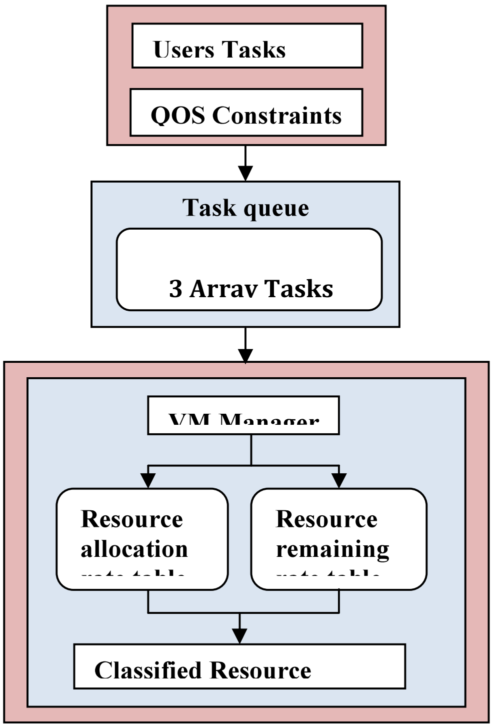

Fig: 1. Architecture For Resource Allocation

4. IMPLEMENTATION DETAILS

Upon receiving the request, the scheduler checks the pre-collected availability states of all candidate nodes, and estimates the minimal payment of running the task within its deadline on each of them. The host that requires the lowest payment will run the task via a customized VM instance with isolated resources. Specifically, the VM will be customized with such a CPU rate and disk I/O rate that the task can be finished within its deadline and its user payment can also be minimized meanwhile. Finally its computation results (or feedbacks) will be returned to users. We will comply with common Cloud computing terminology and refer to these VMs as instances for the remainder of this paper. The term instance type will be used to differentiate between VMs with different hardware characteristics. The users can submit any number of tasks to get the needed resources. The tasks may be upload a file, video, document, etc. The SLA factors are taken as the QOS defined by the user. The factors are throughput, CPU utilization, response time and minimum memory wastage.
The overall SLA objective constraints is given by

$$F(obj) = w_{ut} \sum_{i=1}^{M} t_{tran}(i) - w_{dt} \sum_{i=1}^{M} t_{exec}(i) - \sum_{i=1}^{M} TP_i - w_{CPU} \sum_{i=1}^{M} \frac{C_i}{T_i} - w_{RT} \sum_{i=1}^{M} RT_i$$

VM Manager consist of two monitoring tables are resource allocation rate table and another one is resource remaining rate table. Resource allocation rate table used to monitor the resource allocation process and store the states of the resource utilization rate. Resource remaining rate table used to store the unused or remaining resources in the virtual machine.

5. RESULTS AND FUTURE WORK

The proposed method has been implemented and the results are compared with Location-Aware Dynamic Resource Allocation Model(Gihun Jung and Kwang Mong Sim 2012). says that the performance degradation according to response and consider the location of physical machine to allocate user request as a virtual machine. To design the hybrid resource management architecture to perform location aware VM placement and dynamic resource utilization management. Response time should be increased. In this paper consider the utility function is used to find out which PM is appropriate for a new VM or migration, the provider evaluates each PM using a utility function.VM Placement Decision Making is the process, When a provider receives a new VM placement for a user, the provider first evaluates the network delay between the user and each data center. If the provider finds closest data center for the user VM, the provider now evaluates each utilization level of PMs based on utilization report, which is sent from each PM. After VM placement, each PM keeps monitoring its utilization level, and reports to the provider when the utilization is changed. When a PM's utilization level exceeds a given threshold, it reports to the provider to decide whether VMs are needed to migrate to another PM. When a provider decides to migrate a problematic VM, the provider instructs the hypervisor to migrate the VM to another appropriate PM After migration, each PM keeps monitoring its utilization level and reports it. This method provides better performance and response time. But it leads to bandwidth consumption and not cost effective.

In the proposed method of allocation the classification of VMs is done within a PM which is more cost effective. The proposed Multi QOS based allocation model is compared with the location aware dynamic resource allocation model. The simulation results show that the proposed model has better optimum performance.

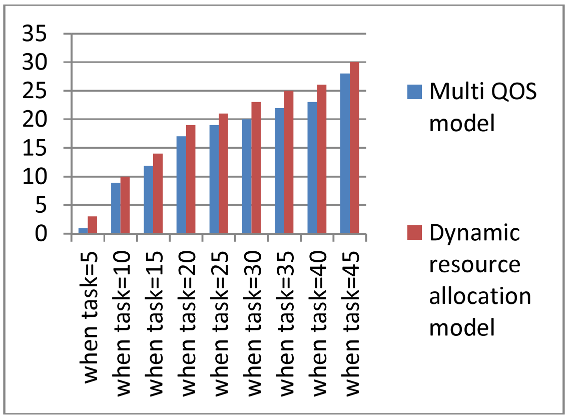

Fig.5.1 Comparison Of Cost In Terms Of Upload/Download Time

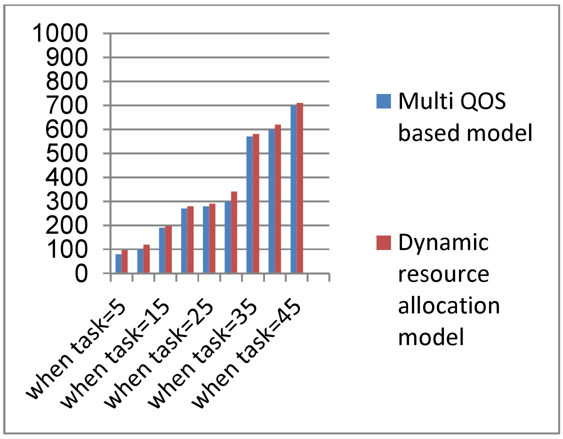

Fig.5.2 Comparison Of Response Time

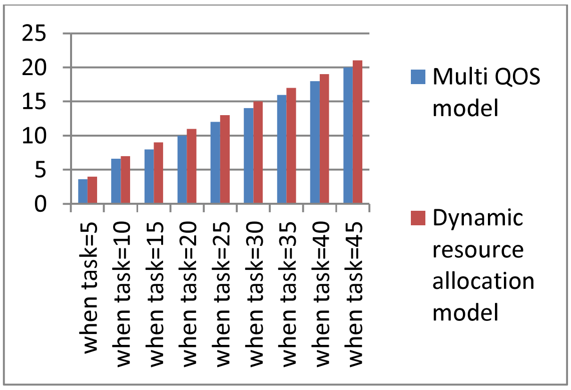

Fig.5.5 Comparison Of Memory Wastage

REFERENCES:

[1] Shin-ichi Kuribayashi ., "Optimal Joint Multiple Resource Allocation Method for Cloud Computing Environments",International Journal of Research and Reviews in Computer Science (IJRRCS), Vol.2, No.1, March 2011.

[2] Hadi Goudarzi and Massoud Pedram., "Multi-dimensional SLA-based Resource Allocation for Multi-tier Cloud Computing Systems",IEEE international conference on cloud computing;PP 324-331,2010.

[3] Rostand Costa, Francisco Brasileiro, Guido Lemos de Souza Filho, Denio Mariz Sousa., "Just in Time Clouds: Enabling Highly-Elastic Public Clouds over Low Scale Amortized Resources", www.lsd.ufcg.edu.br/relatorios_tecnicos,2010

[4] Gihun Jung and Kwang Mong Sim., "Location-Aware Dynamic Resource Allocation Model for Cloud Computing Environment", International Conference on Information and Computer Applications (ICICA 2012) IPCSIT vol. 24, 2012.

[5] Preeti Agrawal,Yogesh Rathore., "An Approach for Effective Resource Management in Cloud Computing",IJT Volume 1,Issue 2,July-Dec.,2011.

[6] Shaminder Kaur., "An Efficient Approach to Genetic Algorithm for Task Scheduling in Cloud Computing Environment",I.J. Information Technology and Computer Science, 2012, 10, 74-79 Published Online September 2012.

[7] G.Sireesha, L.Bharathi., "Exploiting Dynamic Resource Allocation for Efficient Parallel Data Processing in the Cloud", International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 6, August – 2012.

[8] Luis M. Vaquero, Luis Rodero-Merino, Rajkumar Buyya, , January 2011. “Dynamically scaling applications in the cloud”, *ACM SIGCOMM Computer Communication Review* vol.41, issue.1, pp. 45–52.

[9] Livny, M.; Melman, M. (2011): Load Balancing in Homogeneous Broadcast Distributed Systems.Proceedings of the ACM Computer Network: Performance Symposium, pp. 47-55.

[10] Murata, Y. et al. (2010): A Distributed and Cooperative Load Balancing Mechanism for Large-Scale P2P Systems.Proceedings of International Symposium on Applications and Internet (SAINT ’06)Workshops, pp. 126-129.

[11] Martin Randles, David Lamb, A. Taleb-Bendiab(2010), A Comparative Study into Distributed LoadBalancing Algorithms for Cloud Computing

[12] Gunho Lee,Niraj Tolia,Parthasarathy Ranganathan,Randy H. Katz., “Topology-Aware Resource Allocation for Data-Intensive Workloads”, ACM SIGCOMM computer communication review Volume 41 Issue1,January 2011.

[13] GIHUN Jung, Kwang Mong Sim, Paul C. K. Kwok, Minjie Zhang., “A TIME-DRIVEN Adaptive Mechanism For Cloud Resource Allocation”,2011.

[14] Koti Reddy S, Ch. Subba Rao., “Dynamic Resource Allocation In The Cloud Computing Using Nephele’s Architecture”,international journal of engineering science and advanced technology volumne-2,Issue-4,2012.

[15] Linlin lou,saurabh kumar garg&rajkumar., “SLA-based resource allocation for Software as Service provider in cloud computing environments”, 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing,2011