Payment minimization and Error-tolerant Resource Allocation for Cloud System Using equally spread current execution load

Pooja .B. Jewargi
Department of computer science and engineering,
PDA college, Gulbarga,
585101, India

Prof. Jyoti.Patil
Department of computer science and engineering,
PDA college, Gulbarga,
585101, India

***Abstract*-- Sharing of resources in cloud systems can lead to performance inaccuracies due to which efficient resource allocation and guaranteeing tasks execution is must in cloud systems. Different resources and services have different payment and time frame demand in cloud. For example video editing services must be low cost but not necessarily quickest. Audio chat must be fast but not necessarily cheapest. Hence we propose a solution which combines two techniques in order to guarantee execution of tasks within specified time deadline and try to minimize cost by using cost threshold. Equally spread current execution load and optimize response time these are the two techniques used to implement the system. A toolkit called CloudSim is used to support both system and behavior modeling of Cloud System components such as data centers, virtual machines (VMs) and resource provisioning policies. Further simulation results have been demonstrated to show the effectiveness of proposed work in terms of increase in number of tasks executed, minimize execution time of submitted tasks. In future, we plan to integrate our algorithms with stricter/original deadlines into some excellent management tools like OpenNebula, for maximizing the system-wide performance.**

Keywords— *cloud computing, equally spread current execution load, optimize response time, resource allocation.*

1. Introduction

Cloud computing is a distributed computing paradigm that focuses on providing a wide range of users with distributed access to scalable, virtualized hardware and/or software infrastructure over the internet. All the resources provisioned by Cloud system are supposed to be under a payment model, in order to avoid users' over-demand of their resources against their true needs. Each task's workload is likely of multiple dimensions. First, the computer resources in need may be multiattribute (such as CPU, disk-reading speed, network bandwidth, etc.), resulting in multi-dimensional execution in nature.

Second, even though a task just depends on one resource type like CPU, it may also be split to multiple sequential execution phases, each calling for a different computing ability and various price on demand, also leading to a potentially high-dimensional execution scenario. As the cloud is made up of datacenters; which are very much powerful to handle large numbers of users still then the essentiality of efficient resource allocation and guaranteed execution of users tasks is must. Hence in this paper we design a system using two techniques called optimize response time and equally spread current execution based on load balancing. However load balancing is a technique of distributing the loads among various nodes of a distributed system to minimize the response time, minimize the cost, minimize the resource utilization, and minimize the overhead.

2.Literature survey

The literature survey of various methods for different issues and solution proposed by authors is discussed as follows

In [1], survey on cloud computing and its characteristics is done. Cloud Computing is a "buzz word" en-compassing a wide variety of aspects such as deployment, load balancing, provisioning, and data and processing out-sourcing.

In [2], survey on virtual machines is done. Virtualization, in computing, is the creation of a virtual version of something, such as a hardware platform, operating system, and a storage device or network resources.

In [3], this paper we focus on performance isolation mechanisms in Xen, a popular open source VMM. Xen supports per-VM CPU allocation mechanisms. However, it like many other VMMs—does not accurately account for resource consumption in the hypervisor on behalf of individual VMs, e.g., for I/O processing.

In [4], this paper a brief overview on amazon elastic compute cloud is done.

Amazon Elastic Compute Cloud (EC2) provides a cloud computing service by renting out computational resources to customers (i.e., cloud users). The customers can dynamically provision virtual servers (i.e., computing instances) in EC2, and then the customers are charged by Amazon on a pay-per-use basis.

In [5], this paper we study how execution times for programs computed using predictive models which are more efficient. Predicting the execution time of computer programs is an important but challenging problem in the community of computer systems.

In [6], this paper survey on power and performance management on virtualized clusters is done. Increasing number of large-scale server clusters are being deployed in data centers for supporting many different web-based application services in a seamless fashion. In this scenario, the rising energy costs for keeping up those web clusters are becoming an important concern for business.

Hence an optimization solution for power and performance management in a platform running multiple independent web applications is defined.

In [7], this paper how datacenter resources are efficiently allocated is studied.Large datacenters host several application environments (AEs) that are subject to workloads whose intensity varies widely and unpredictably. Therefore, the servers of the datacenter may need to be dynamically redeployed among the various AEs in order to optimize some global utility function. This solution is based on the use of analytic queuing network models combined with combinatorial search techniques.

3.System Design

In this section, the emphasis is given on the explanation of architecture of design and the proposed solution of the work is defined.

A. Proposed solution

The main aim of the proposed work is to provide efficient resource allocation and to guarantee tasks execution within specified deadline and minimize cost for resource consumption. Therefore, two techniques optimize response time and equally spread current execution load techniques based on load balancing are proposed. With the help of these two techniques execution times are within specified deadlines is shown in the simulation results.

B. System Architecture

Workflow of equally spread current execution load and its operation

It uses active monitoring load balancer for equally spreading the execution of loads on different virtual machines.

Active monitoring load balancer (AMLB) maintains an index table of virtual machines and the number of allocations assigned to each virtual machine. Data Center Controller receives a new request from a client. When a request for allocation of new VM from Data Center Controller arrives at AMLB, it parses the index table from top until the least loaded VM is found. When it finds, it returns the VM id to the Data Center Controller. If there is more than one found, AMLB uses first come first serve (FCFS) basis to choose the least loaded. After that AMLB updates the allocation table by increasing the allocation count by 1 for that VM. When a VM suitably finishes processing the assigned request, it forwards a response to the Data Center Controller. On receiving the response it notifies the AMLB about the VM de-allocation. The AMLB updates the allocation table by decreasing the allocation count for that VM by 1.

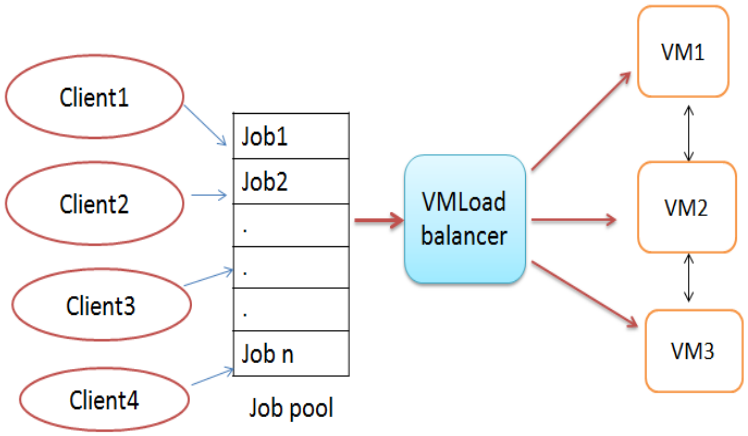

Fig 1: workflow of equally spread current execution load technique

Workflow of optimize response time and its operation

As per this strategy, the data center selection is not made randomly and vm cost in each data center is compared with other datacenters in the same region. The data center with lowest vm cost is selected. Now the requests will be sent to this data center to be processed.

This strategy gives cost effective user request routing.

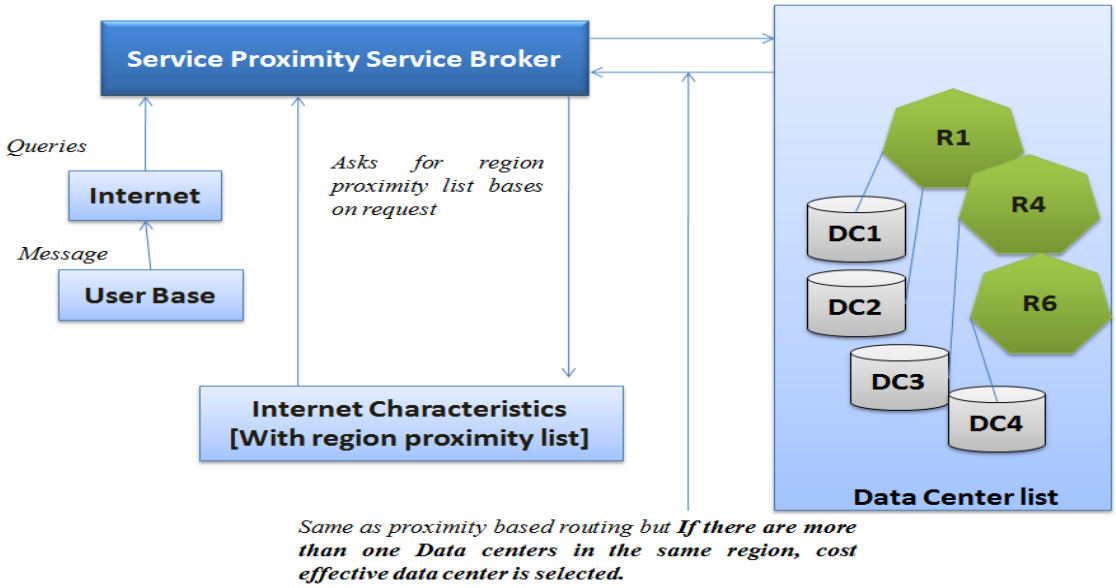

Fig 2: workflow of optimize response time technique

4.Simulation Results

In this work we have developed a cloud network model in CloudSim to incorporate core middleware (PaaS) layer fundamentals of execution management, pricing, metering, into the simulation to realistically analyse the behaviour of this network and concept. We compare the performance of round robin and equally spread current execution load for various parameters as shown in the graph.

A. Simulation Parameters

Table1. Simulation parameters

Parameters	Values
User base	2-40
Simulation time	6 hours
Data centre	2-40
VM/ Data centre	5-10
Cost threshold	0.01-5 $/hr
Time threshold	140-150 sec
Load balancing technique	Equally spread current execution load
Scheduling technique	Optimise response time

B. Simulation performance metrics

The performance metrics used to measure the simulation of the work are explained below

- **Execution time**

  Execution time is defined as the amount of duration taken by the datacenter to complete executing all submitted tasks from different user bases.

  It includes individual datacenter processing time and overall response times.

- **Cost**

  Cost is the amount to be payed by users for consumption of resources in order to get the jobs executed by the datacenters. These costs are defined by payment models.

  Cost = storage cost + transfer cost (RAM to hard disk) + processing

  Cost/instruction + network cost

C. Simulation Results

In this section, the graphical analysis of the work is done depending on the values obtained from the simulation environment. These graphs are plotted on the performance metrics described earlier.

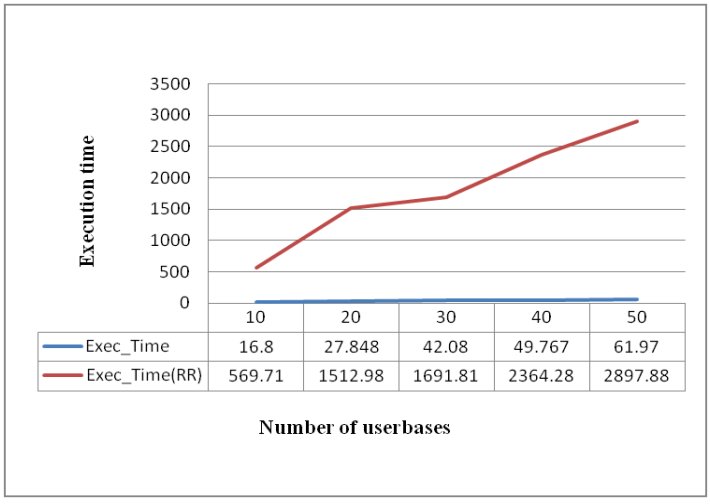

Fig 3 Graphs showing the comparison of execution time of round robin and equally spread current execution load techniques.

Figure 3 illustrates how the execution time for processing same number of number of requests differs for round robin and equally spread current execution load techniques as the number of Userbase increase.

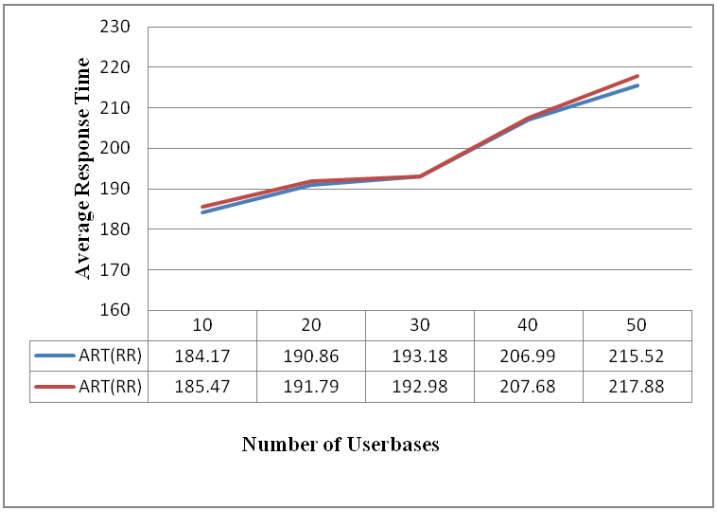

Fig 4 Comparison of average response times of round robin and equally spread current execution load techniques.

Figure 4 illustrates how average response time differs for round robin and current execution load techniques for same number of requests.

5. Conclusion

The response time and data transfer cost is a challenge of every engineer to develop the products that can increase the business performance in the cloud based sector. The several strategies lack efficient scheduling and load balancing resource allocation techniques leading to increased operational cost and give customer satisfaction. The paper aims to development of error tolerant resource allocation strategy through improved job and load balancing resource allocation techniques. Equally spread current execution algorithm dynamically allocates the resource to the job in a queue leading reduced cost in data transfer and virtual machine formation. Optimize response time service broker policy tries to minimize the response time for user jobs by selecting suitable datacentre for user requests. This improves the business performance and retention to the total customer satisfaction.

REFERENCES

[1] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner,"A break in the clouds: towards a cloud definition," *SIGCOMM Comput. Commun. Rev*, vol. 39, no. 1, pp 50–55, 2009.

[2] J. E. Smith and R. Nair, *Virtual Machines: Versatile Platforms For Systems And Processes*. Morgan Kaufmann, 2005.

[3] D. Gupta, L. Cherkasova, R. Gardner, and A. Vahdat, "Enforcing performance isolation across virtual machines in xen," in *Proceedings of the ACM/IFIP/USENIX 2006 International Conference on Middleware (Middleware'06)*, New York, USA, 2006, pp. 342–362.

[4] Amazon elastic compute cloud: on line at http://aws.amazon.com/ec2/.

[5] L. Huang, J. Jia, B. Yu, B.G. Chun, P. Maniatis, and M. Naik, "Predicting Execution Time of Computer Programs Using Sparse Polynomial Regression," in *24th Conference on Neural Information Processing Systems (NIPS'10)*. 2010, pp. 1–9.

[6] V. Petrucci, O. Loques, and D. Moss´e, "A dynamic optimization model for power and performance management of virtualized clusters," in *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking (e-Energy'10)*. New York, NY, USA: ACM, 2010, pp. 225–233.

[7] X. Meng, C. Isci, J. Kephart, L. Zhang, E. Bouillet, and D. Pendarakis, "Efficient resource provisioning in compute clouds via vm multiplexing," in *Proceeding of the 7th international conference on Autonomic computing (ICAC'10)*. ACM, 2010, pp. 11–20.

