DYNAMIC ALLOCATION METHOD FOR EFFICIENT LOAD BALANCING IN VIRTUAL MACHINES FOR CLOUD COMPUTING ENVIRONMENT

Bhaskar. R[1], Deepu. S.R[2] and Dr. B.S. Shylaja[3]

[1,2]IV sem MTech, dept.of CSE, Dr. Ambedkar Institute of Technology, Bangalore
[1]bhas.bhaskar@gmail.com,[2]deepusrd@gmail.com
[3]Professor, dept.of ISE, Dr. Ambedkar Institute of Technology, Bangalore
shyla.au@gmail.com

ABSTRACT

This paper proposes a Dynamic resource allocation method for Cloud computing. Cloud computing is a model for delivering information technology services in which resources are retrieved from the internet through web-based tools and applications, rather than a direct connection to a server. Users can set up and boot the required resources and they have to pay only for the required resources. Thus, in the future providing a mechanism for efficient resource management and assignment will be an important objective of Cloud computing. In this project we propose a method, dynamic scheduling and consolidation mechanism that allocate resources based on the load of Virtual Machines (VMs) on Infrastructure as a service (IaaS). This method enables users to dynamically add and/or delete one or more instances on the basis of the load and the conditions specified by the user.

Our objective is to develop an effective load balancing algorithm using Virtual Machine Monitoring to maximize or minimize different performance parameters(throughput for example) for the Clouds of different sizes (virtual topology de-pending on the application requirement).

KEYWORDS

Cloud computing, Infrastructure-as-a-Service, Amazon ec2, Optimizing VM Load, Load balancing.

1. INTRODUCTION

Cloud computing refers to the delivery of computing and storage capacity as a service to a heterogeneous community of end-recipients. Cloud computing is an internet technology that utilizes both central remote servers and internet to manage the data and applications. This technology allows many businesses and users to use the data and application without an installation. Users and businesses can access the information and files at any computer system having an internet connection. Cloud computing provides much more effective computing by centralized memory, processing, storage and bandwidth[8].

Cloud computing has several applications such as Infosys is using Microsoft's Windows Azure Cloud services, including SQL Data Services, to develop Cloud-based software capabilities that would let automobile dealers share information on inventories and other resources. Best Buy's Giftag applet uses Google App Engine to let users create and share wish lists from Web pages they visit. Wang Fu Jing Department Store, a retailer in China, uses IBM Cloud services, including supply chain management software for its network of retail stores[9].

Cloud computing providers offer their services according to three fundamental models Infrastructure as a service (IaaS), platform as a service (PaaS), and software as a service (SaaS)

where IaaS is the most basic and each higher model abstracts from the details of the lower models. Platform -as-a-service in the Cloud is defined as a set of software and product development tools hosted on the provider's infrastructure. Developers create applications on the provider's platform over the Internet. PaaS providers may use APIs, website portals or gateway software installed on the customer's computer. In the software-as-a-service Cloud model, the vendor supplies the hardware infrastructure, the software product and interacts with the user through a front-end portal. Infrastructure-as-a-Service is to start, stop, access and configure their virtual servers and storage. In the enterprise, Cloud computing allows a company to pay for only as much capacity as is needed, and bring more online as soon as required. Because this pay-for-what-you-use model resembles the way electricity, fuel and water are consumed, it's sometimes referred to as utility computing[1].

Several issues are existed in Cloud computing such as Performance, Security, Availability, and Inability to Customize. In this paper we are focusing performance issue. Operating in a Cloud computing environment does not eliminate application performance issues[10]. In fact, the Cloud is very complex and will possibly introduce even more performance problems then in non Cloud environments. As such the ongoing monitoring of all the applications is accessed via the Cloud. This will ensure that Service Level Agreements are met and performance and uptime are optimal.

On a Cloud computing platform, dynamic resources can be effectively managed using virtualization technology. The subscribers with more demanding SLA can be guaranteed by accommodating all the required services within a Virtual Machine image and then mapping it on a physical server. This helps to solve problem of heterogeneity of resources and platform irrelevance. Load balancing of the entire system can be handled dynamically by using virtualization technology where it becomes possible to remap Virtual Machines (VMs) and physical resources according to the change in load . Due to these advantages, virtualization technology is being comprehensively implemented in Cloud computing[11].

A Virtual Machine (VM) is a software implementation of a computing environment in which an operating system (OS) or program can be installed and run. The Virtual Machine typically emulates a physical computing environment, but requests for CPU, memory, hard disk, network and other hardware resources are managed by a virtualization layer which translates these requests to the underlying physical hardware. VMs are created within a virtualization layer, such as a hypervisor or a virtualization platform that runs on top of a client or server operating system. This operating system is known as the host OS. The virtualization layer can be used to create many individual, isolated VM environments.

This paper focuses on dynamic allocation method for efficient load balancing on Virtual Machines. This paper is organized as follows section 2 describes existing system, section 3 describes Experimental setup, section 4 describes Results and analysis, section 5 describes conclusion and section 6 describes bibliography.

2. VM LOAD BALANCING

2.1. OVERVIEW

Virtual Machine enables the abstraction of an OS and Application running on it from the hardware. The interior hardware infrastructure services interrelated to the Clouds are modeled in the simulator by a Datacenter element for handling service requests. These requests are application elements sandboxed within VMs, which need to be allocated a share of processing power on Datacenter's host components. DataCenter object manages the data center management activities such as VM creation and destruction and does the routing of user requests received from user via the Internet to the VMs. The Data Center Controller, uses a VmLoadBalancer to determine which VM should be assigned the next request for processing. Most common VM load balancer are throttled and active monitoring load balancing algorithms. Throttled Load Balancer maintain a record the state of each Virtual Machine (busy/ideal), if a request arrive concerning the allocation of Virtual Machine, throttled load

balancer send the ID of ideal Virtual Machine to the data center controller and data center controller allocates the ideal Virtual Machine. Active Monitoring Load Balancer maintains information about each VMs and the number of requests currently allocated to which VM. When a request to allocate a new VM arrives, it identifies the least loaded VM. If there are more than one, the first identified is selected[6].

In this paper we studied two load balancing algorithms in Cloud computing was done. The algorithms are throttled load balancer, active monitoring load balancer. A new algorithm has been proposed from modifying the active monitoring load balancing algorithm in Virtual Machine environment of Cloud computing in order to achieve better response time, processing time and cost.

2.2. PROPOSED VM LOAD BALANCING METHOD

Optimization of Virtual Machine load is a process of reassigning the total load to the individual Virtual Machines to make resource utilization effective and to improve the response time of the job. A load balancing algorithm which is dynamic in nature does not consider the previous state or behavior of the system, that is, it depends on the present behavior of the system. The important things to consider while developing such algorithm are, estimation of load, comparison of load, performance of Virtual Machine, nature of work to be transferred, selecting of Virtual Machine and many other ones. This load considered can be in terms of CPU load, amount of memory used, delay or Network load.

The time required for completing a task within one process is very high. So the task is divided into number of sub tasks and each sub task is given one one job.The Proposed Load balancing algorithm is divided into two phase. A two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain a high resource utilization. It achieves load balancing by first mapping tasks to Virtual Machines and then Virtual Machines to host resources thereby improving the task response time, resource utilization and overall performance of the Cloud computing environment. In the first phase, find the cpu utilization and memory required for each instance and also find available cpu cycle and memory of each VM. In second phase compare the available resources and required resources, if resources are available instance is to be added otherwise discard the instance finally returns instance status to user. The CloudWatch monitoring service is a special storage engine that is designed for time series data. On one end data collected periodically from servers and from other services is pumped into the monitoring store, and at the other end clients can run queries against the store to extract data from it.

2.3. DETAILED DESIGN

The load balancing service is designed to serve as a first level of distributing load across a number of instances, dealing specifically with DNS and handling the failure of an availability zone. Amazon CloudWatch provides monitoring for AWS Cloud resources and the applications customers run on AWS. Developers and system administrators can use it to collect and track metrics, gain insight, and react immediately to keep their applications and businesses running smoothly. Amazon CloudWatch monitors AWS resources such as Amazon EC2 and Amazon RDS DB instances, and can also monitor custom metrics generated by a customer's applications and services. With Amazon CloudWatch, you gain system-wide visibility into resource utilization, application performance, and operational health[3].

Functions of our proposed system are,

Bucket creation- S3 Browser allows to easily create Amazon S3 Buckets in all regions supported by Amazon S3. Once created a new bucket, one who can create virtual folders to organize files, and upload and download files to and from Amazon S3[3].

Uploading instance- VM Import/Export enables to easily import Virtual Machine images from existing environment to Amazon EC2 instances and export them back to on-premise environment. This offering allows to leverage an existing investments in the Virtual Machines that who built to meet IT security, configuration management, and compliance requirements by seamlessly bringing those Virtual Machines into Amazon EC2 as ready-to-use instances. One who can easily export imported instances back to your on-premise virtualization infrastructure, allowing you to deploy workloads across your IT infrastructure[3].

Monitoring instance- Amazon CloudWatch provides a reliable, scalable, and flexible monitoring solution that can start using within minutes. No longer need to set up, manage, or scale own monitoring systems and infrastructure. Using Amazon CloudWatch, which can easily monitor as much or as little metric data as you need. Amazon CloudWatch lets you programmatically retrieve your monitoring data, view graphs, and set alarms to help you troubleshoot, spot trends, and take automated action based on the state of Cloud environment

[3].

Amazon CloudWatch enables to monitor AWS resources in real-time, including Amazon EC2 instances, Amazon EBS volumes, Elastic Load Balancers, and Amazon RDS DB instances. Metrics such as CPU utilization, latency, and request counts are provided automatically for these AWS resources. One who can also supply his own custom application and system metrics, such as memory usage, transaction volumes, or error rates, and Amazon CloudWatch will monitor these too.

Algorithm 1: To describe add/remove instance.

Desc_add/remove_inst()

{

  Find Instance Id 'inst_id' from ec2InstanceRequest;

  Find required cpu utilization 'reqcpuUtil'

  From ins_size*60*60*24;

  Find VM Id 'VmId' from ec2AvailabilityZones;

  Print availcpuUtil;

  If '*availcpuUtil***' result in a loop** **then**

        Add instance to controller

  else

        discard Instance from user request**;**

   end

}

Algorithm 2: Add instance to controller.

Add_inst()

{

 Find   availability zones 'Avail' from

```
 ec2AvailabilityZones;

 Find Key pair 'key' from  ec2KeyPairs;

 Describe user Instances 'Inst' from ec2DescribeInstance;

 Create new key pair 'newKey' from ec2KeyPair;

 Assign 'Instance' to VM;

 Find instance ststus 'InstStat' from RunInstanceRequest;

 Print InstStat;

 }
```

To test our algorithm we are using WebCrawler application to describe better optimization of load on each Virtual Machine. A WebCrawler is a computer program that browses the World Wide Web in a methodical, automated manner or in an orderly fashion. This process is called Web crawling or spidering. Many sites, in particular search engines, use spidering as a means of providing up-to-date data. WebCrawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Also crawlers can be used to gather specific types of information from Web pages, such as harvesting e-mail addresses (usually for sending spam).

3.   EXPERIMENTAL SETUP

In this section we describe the experimental setup that was used to run workflows. Java language is used for implementing VM load balancing algorithm. EC2 was chosen because it is currently the most popular, feature-rich, and stable commercial Cloud Workflows are loosely-coupled parallel applications that consist of a set of computational tasks linked via data and control-flow dependencies. Unlike tightly-coupled applications in which tasks communicate directly via the network. In order to have an unbiased comparison of the performance of workflows on EC2 the experiments presented in this paper attempt to account for these differences by (a) running all experiments on single nodes and (b) running experiments using the local disk on EC2. Although single-node experiments do not enable us to measure the scalability of Cloud services they do provide an application-oriented understanding of the capabilities of the underlying resources that can help in making provisioning decisions. Testing the scalability of Cloud services when running workflows on multiple nodes is left for future work.

4.   RESULTS AND ANALYSIS

Fig 1 shows an interface designed facilitating the user to create bucket on Cloud and send request to the instance monitor. Client provides an option for creating bucket, deleting bucket, refreshing AWS amazon web account and downloading the contents resides in the bucket. When user sends a request to the Cloud, IaaS converts it into an instance and sends it to job manager.
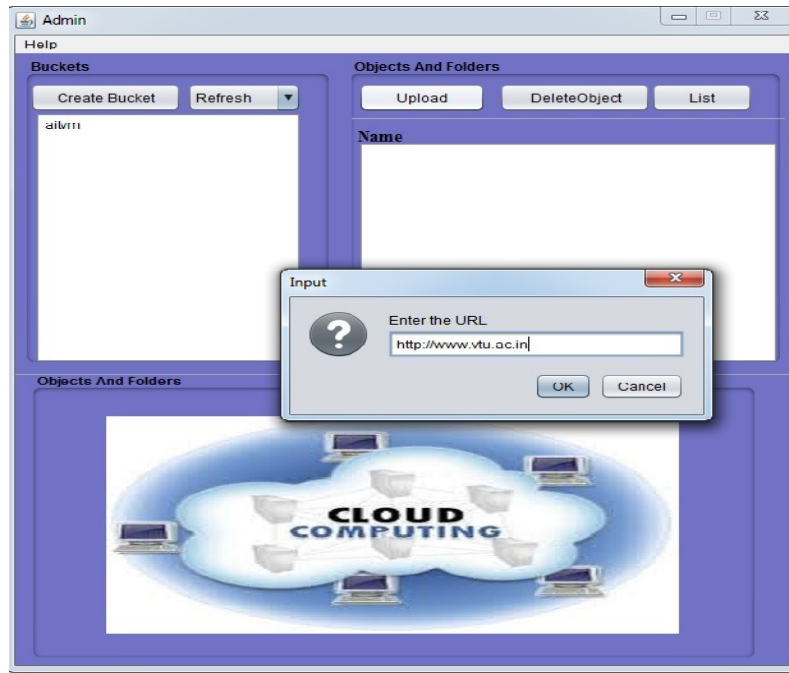
Fig 1: Client Interface

Fig 2 shows an Job manager, receives the user request from client. Received instance is to be divided into number of sub tasks(instances). After dividing an instance sends it to task manager.
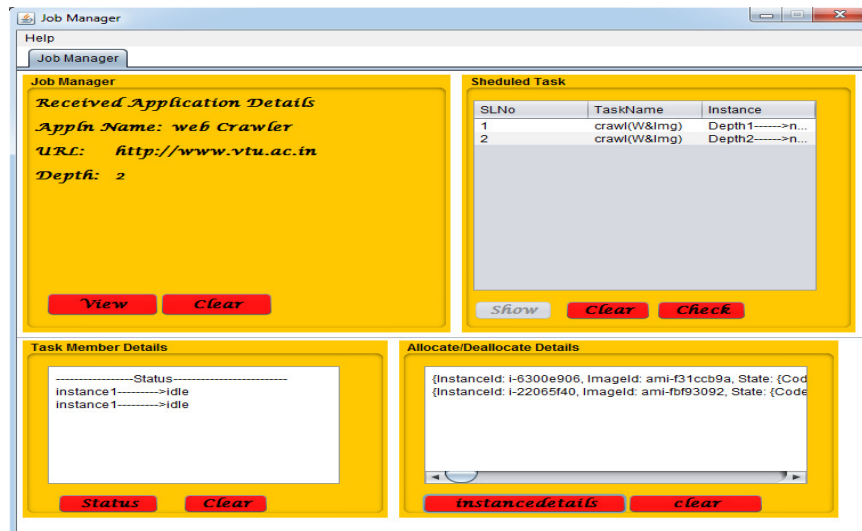

Fig 2: Job Manager

Fig 3. Shows an Instance monitor, monitors the status of the instance and Virtual Machine(VM's) by checking the required resources of instance and available resources such as cpu usage and memory available of VM's. Dynamic instance module decides whether to add instance or delete instance based on the status of the instance monitor. If the resources are available on Virtual Machine the instance is to be added to Cloud otherwise discard the instance.

Fig also shows an Task manager receives the instances from job manager. Process the user request and sends results back to client. Client displays the results to the user
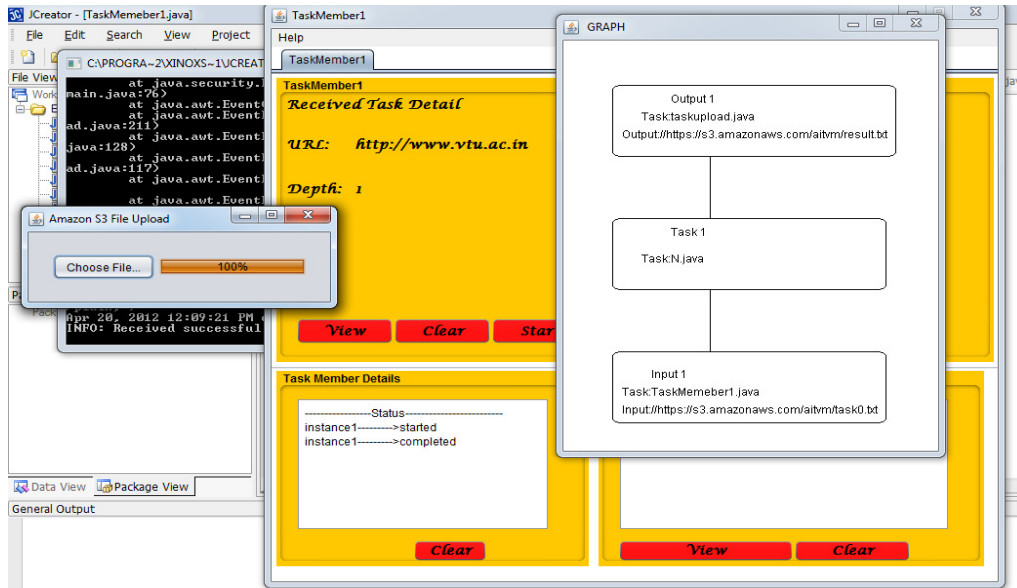

Fig 3: Task Manager

Table 1 Describes results of existing method[7].

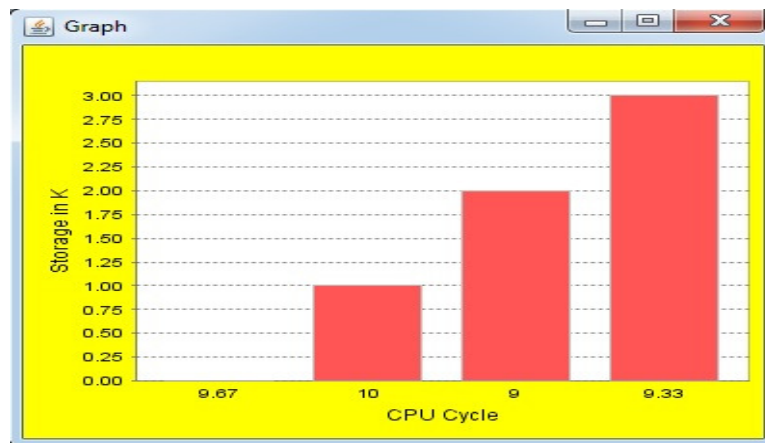| ID	RESOURCES	THRESH_VALUE	CPU_CYCLE
1	0	Added	10
2	1	Added	9.67
3	2	Added	9.33
4	3	Added	9

Fig 4 Shows the graphical representation of the Table 1 with respect to Cpu cycle and Storage in Kb.

Table 2 describes average values of the Dynamic Instance method by taking 5 trial values .

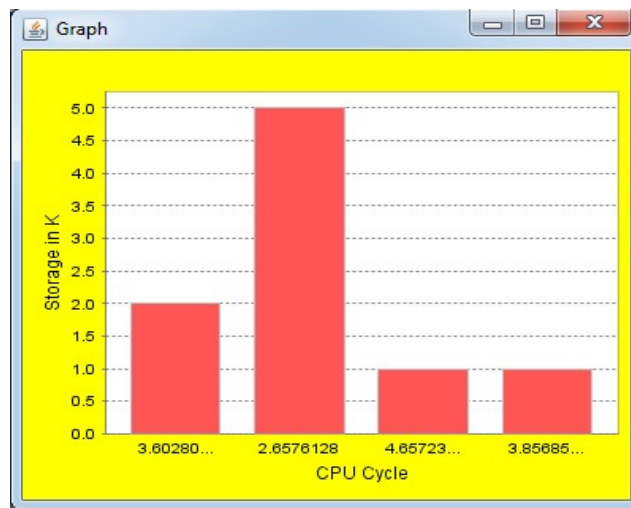| NAME	INST-ID	AMI-ID	TYPE		CPU CYCLE
FirstInstance	i-2aa8c44d	ami-8f8c54e6	t1.micro	5	4.87854
Instance1	i-54502733	ami-8f8c54e6	t1.micro	1	2.76843
Instance2	i-6c41350b	ami-8f8c54e6	t1.micro	1	3.856857142857143
Instance3	i-6c45280b	ami-8f8c54e6	t1.micro	2	3.6028000000000007

Fig 5 Shows the graphical representation of the Table 2 with respect to Cpu cycle and Storage in Kb.
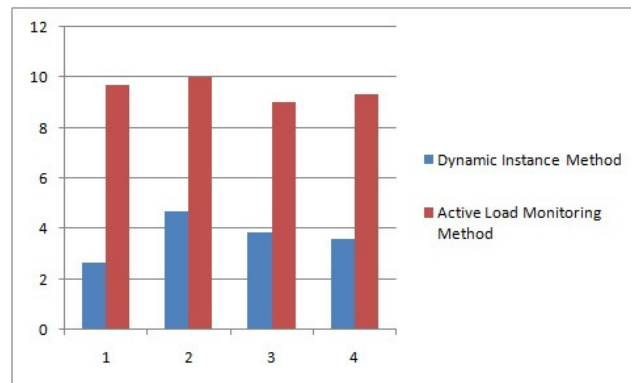

Fig 6 Shows the comparison graphical representation of Table 1 and Table 2 with respect to Cpu cycle. X – axis describes number of instances, Y – axis describes CPU cycle.

5. CONCLUSION

In this paper a new VM load balancing algorithm was proposed and then implemented in Amazon EC2 Cloud computing environment using java language. Proposed algorithm find the available cpu cycle of each Virtual Machine (VM) and Send the ID of Virtual Machine to the Cloud controller for allocating the new request. We conclude that Cloud controller utilizes the available resources on Virtual Machine then it effect the overall performance of the Cloud Environment and also decrease the average response time.

REFERENCES

[1] Panzieri, Ozalp, Babaoglu1, Stefano, Ferretti, Vittorio, Ghini, MorenoMarzolla, Distributed Computing in the 21st Century: Some Aspects of Cloud computing Fabio Technical Report UBLCS-2011-03 May 2011

[2] Patricia Takako Endo, André Vitor de Almeida Palhares, Nadilma Nunes Pereira, Resource Allocation for Distributed Cloud: Concepts and Research Challenges, 2011 IEEE

[3] Amazon Inc., "Amazon Elastic Compute Cloud," http://aws.amazon.com/

[4] Shuai Zhang, Cloud computing Research and Development Trend, Hebei Polytechnic University College of Science Hebei huoxiuzhen@126.com 2010 IEEE DOI 10.1109/ICFN.2010.58

[5] G. Soundararajan et at al, "Resource Allocation for Database Servers Running on Virtual Storage", In Proc. of 6th USENIX Conference on File and Storage Technologies, 2009.

[6] Meenakshi Sharma, Pankaj Sharma, Dr. Sandeep Sharma, Efficient Load Balancing Algorithm in VMCloudEnvironment, IJCST Vol. 3, Iss ue 1, Jan. - March 2012.

[7] Atsuo Inomata, Taiki Morikawa, Minoru Ikebe,Yoshihiro Okamoto, Satoru Noguchi, Kazutoshi Fujikawa, Hideki Sunahara Information Science, Nara Institute of Science and Technology, Sk. Md. Mizanur Rahman School of Information Technology and Engineering University of Ottawa,Ottawa, Canada, Proposal and Evaluation of a Dynamic Resource Allocation Method based on the Load of VMs on IaaS, 978-1-4244-8704-2/11/2011.

[8] Brief Description Of Cloud Computing *By* Robert S Bob .

[9] Real-World Cloud Computing Applications - Cloud-computing by John Foley.

[10] Cloud Computing Performance – by Tevron - Application Monitoring .

[11] A Genetic Algorithm Scheduling Approach for Virtual Machine Resources in a Cloud Computing nvironment by Shailesh Sawant.