© 2016 IJSRSET | Volume 2 | Issue 4 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
Themed Section: Engineering and Technology

Cloud Based Resource Management with Autoscaling

Dr. D .Ravindran, Ab Rashid Dar
School of Computer Science, St. Joseph's College, Trichirapalli, Tamil Nadu, India

ABSTRACT

With ever increasing growth and popularity in Cloud computing, it becomes now first choice and priority for every individual who access the internet, one of the key and advantageous features of cloud computing is its flexibility and scalability. Autoscaling offers the facility to the clients to scale up and scale down the resources as per their demands. As everything is taking place in automatic manner, so human intervention errors are less and reduce the manpower and costs. In this paper main ideas revolve around the problems in existing scalable cloud computing systems. In modern days, management of resources is hot and most talked topic in cloud environment. Here in this paper a queuing concept is being employed with resource management algorithm to provision and deprovision of different available resources. The CPU thresholding defines the workloads, which are being transferred across the cloud platform through a proper network with maximum bandwidth.

Keywords: Cloud Based Resource Management; Scalability, Autoscaling, Thresholds, Load Balancing

I. INTRODUCTION

Cloud Computing is a new and emerged as a hottest computational model in IT sector, which is primarily based on already existing computing paradigms i.e. Centralized, Parallel, Grid and Distributed Cloud computing. The services and applications are accessible to the different clients using proper internet protocol suit and networking standards. NIST defines "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction"[1].

Cloud based resource management is an approach as cloud offers abundant resources to its clients, so the management of these resources are always first priority of the cloud vendors. Autoscaling is a technique when implemented properly can lead to the proper cloud based resource management. The service charges of client interest are only to be paid while hadn't bothered about the undesirable cloud services. With best routing policies, load balancing, workloads and traffic is being transferred across the high bandwidth channels which reduce the congestion of the network across the globe. Load balancers which act as an interface between the clients and the servers, which redirects the client requests and offer them the available resources.

II. METHODS AND MATERIAL

1. Related Work

M. Kriushanth et al.,[1]Here the authors telling about the basics of cloud computing concepts like service models, deployment models and the various dimensions of cloud scalability. The dimensions they given are vertical and horizontal scalability. They presented infrastructure of auto scaling and some challenges and issues that are occurring in auto scaling. The issues they given are is taken into account for future work.

R. Anandhi et al.,[2] presented the basics of scalability and its scalability factors. Here they distinguish the scalability into four by its scalability factors. Then they described why and how the scalability has been chosen based on the user requirement. Further they describe the two types of approach in messaging system of scalability. They had given the way to improve the scalability through auto scaling, scaling the database horizontally and EBS.

Jorge M. Londoño-Pelaez et al.,[3] explained about the way to solve the two problems like over provisioning and under provisioning. To address these problems they present an autonomic auto-scaling controller that based on the stream of measurements from the system maintains the optimal number of resources and responds efficiently to workload variations, short duration peaks in the workload. Their technique consists of three components. It has been explained with clearly with its parameter tuning. These techniques have also been analyzed.

HaniehAlipour et al.,[4]In this paper, they presented a survey that explores definitions of related concepts of auto-scaling and taxonomy of auto-scaling techniques. Based on the survey results, they outline open issues and future research directions for this important subject of auto scaling in cloud computing. They explained each and every concept of the auto scaling taxonomical areas. This gives the new various areas if research sectors especially in the area of auto scaling.

Che-Lun Hung et al., [5] proposed the novel virtual cluster architecture for dynamic scaling of cloud applications in a virtualized Cloud Computing environment. An auto-scaling algorithm for automated provisioning and balancing of virtual machine resources based on active application sessions will be introduced. Also, the energy cost is considered in this proposed algorithm. This work has demonstrated the proposed algorithm is capable of handling sudden load requirements, maintaining higher resource utilization and reducing energy cost. They proposed the two algorithms for auto scaling of web applications and for the distributed systems.

2. Scalability

One of the key benefits of using cloud-computing paradigm is its scalability. It is regarded as the most exciting feature of cloud computing. It supports the long-term strategies and business needs and is entirely different than elasticity. It is about holding unexpected workloads, and it depends on system design, as well as the types of data structures, algorithms and communication mechanisms used to implement system components [2]. Clients dynamically or automatically provision their resources like hardware devices and software applications when demand and situation arise like that by the mechanism. Cloud computing allows clients or cloud vendors business to easily scale up or scale down their IT requirements as and when required. For example, most cloud service providers will allow clients to increase their existing resources to accommodate increased business needs or changes. This will allow clients to support their business growth without expensive changes to the existing systems being used in cloud environment. Because of the highly scalable nature of cloud computing paradigm, many organizations are now relying on managed data centers where there are cloud experts trained in maintaining and scaling shared, private and hybrid clouds. Cloud computing allows for quick and easy allocation and reallocation of resources in a monitored environment where overloading or load balancing is no more a concern as long as the system is managed and maintained properly. The most important technology which enables the cloud paradigm to scale up and scale down the resources is virtualization without it cloud computing is not sufficient, it provides the agility and speed up the execution of processes.

3. Scalability Issues inExisting Systems

Tools that automatically modify the amount of used resources are called “auto-scaling services”. Although auto-scaling has been regarded as one of the best solution in order utilize the maximum resources without wastage of investment in undesirable services of cloud computing, but as a quote “two sides of a coin” It is beneficial but at other side of the coin it also brings some unique challenges that need to be addressed and find solutions how to tickle and resolve them. Some of the scalability-issues related with scalability are as [3];

- Scaling includes diverse cloud service models, but most studies only focus on the infrastructure level while as ignoring the other two levels like SaaS and PaaS. Auto-scaling at the service-level is important as services are running on a set of connected virtual machines, and the quality of the service relies on how auto-scaling handles resources for these VMs.
- Insufficient tools for managing and assemble metrics at the platform level and service level to support auto-scaling decisions.

- Auto-scaling in hybrid cloud environments is not well supported. Hybrid clouds applications are deployed on a private and public cloud simultaneously. In this manner of the mutualism, the public and private cloud may offer different auto-scaling techniques that are incompatible with each other, so there would be an interoperability and complacence issues in auto-scaling resources between the two cloud deployed models.
- In autoscalingQoS is not properly maintained and managed. Failure of the auto scaling process can result in violations of the system's QoS requirements of performance and scalability and even incur unnecessary cost.

4. Cloud Resources Management

As Cloud is heterogenetic distributed platform where resources across the globe are not uniformly distributed. The management of resources is the most challenging task of the Cloud service provider, to use the adequate resources some measures should be taken in order to utilize them efficiently with maximum benefits. The services provided by cloud vendors sometimes most of them are not of the client's interest. However client had to pay for that, to get rid of this undesirable feature of Cloud, Autoscaling technique is being most widely implementing in the cloud environment at different deployment levels like PaaS, SaaS and IaaS which act as the resource management tool as well. Provisioning and de-provisioning of the resources now entirely depends on client's desire. This improves the flexibility and elasticity of the cloud

5. Autoscaling

Today, cloud computing is totally revolutionizing the way computer resources are allocated, making it possible to build a fully scalable server setup on the Cloud. "Autoscaling is the ability to scale up and scale down the application server's capacity automatically according to customer defines"[4]. To maintain the performance when demand is huge it increases the number of instance and decrease automatically when demand reduces to minimize cost. If application needs more computing power, you now have the ability to launch additional compute resources on-demand and use them for as long as you want, and then terminate them when they are no longer needed. In cloudcomputing applications with a dynamic workload demand need access to a flexible infrastructure to meet performance guarantees and minimize resource costs. While cloud computing provides the elasticity to scale the infrastructure on demand, cloud service providers lack control and visibility of user space applications, making it difficult to accurately scale the underlying infrastructure. Thus, the burden of scaling falls on the user. With cloud computing, the end user usually pays only for the resource they use and so avoids the inefficiencies and expense of any unused capacity. Many Internet applications can benefit from an automatic scaling property where their resource usage can be scaled up and down automatically by the cloud service provider. It can be broadly classified in two categories as;

A. Horizontal Scalability

Horizontal cloud scalability is the ability of the system or resources to connect multiple hardware or software entities, such as servers or networks so that they work as a one logical unit. It means adding more individual units of resource doing the same job. For example, in the case of servers it could increase the speed or availability of the logical unit by adding more servers as per the needed. Instead of one server here one can have two, ten, or more of the same server doing the same work. It is also referred to as scaling out [2].

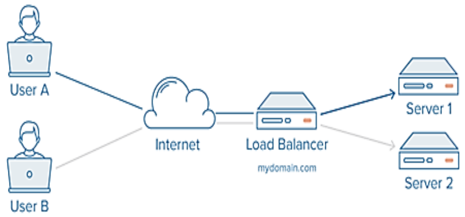

Figure 1. Horizontal Scaling

B. Vertical scalability

Vertical scalability is the ability to increase the capacity of existing single hardware or software by adding more resources to the same server or hardware. For example, adding processing power to a server to make it faster. It can be achieved through the addition of extra hardware to the same entity such as hard drives, servers, CPU's, etc. It provides more shared resources for the operating system and applications. This type of scalability may also be referred to as scaling up or scaling in [2].

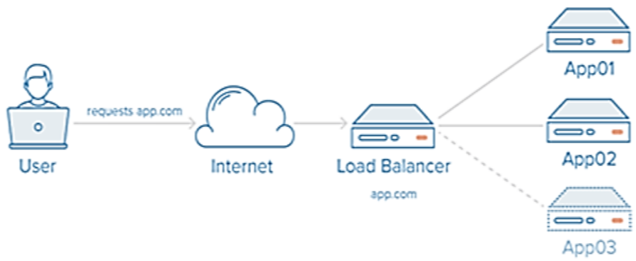

Figure 2. Vertical Scaling

6. Virtualization

Virtualization is a term used for separating of request and resource of service from the underlying physical devices. In computing, virtual version of partitioning the resources like server, storage, operating system, hard drive, printer, network and also the software where the environment divides the machines into one or more executable partitions creating illusion to the client is called as virtualization [6]. Virtualization is the use of software and hardware to create the perception that one or more entities exist although the entities in actually, are not physically present. It provides the illusion effect to the clients.A software tool known as hypervisor that creates fantasy and runs virtual machine. A hypervisor runs one or more virtual machines on a machine which is called as host machine. This machine can be a computer as well as a server. Each of the virtual machine is called a guest machine. The guest operating systems are represented by the hypervisor with a virtual operating platform. It manages the execution of the guest operating systems.

In cloud environment where workloads are unpredictable, Virtualization plays an important role which manages the cloud based resources. It utilizes the resources fully and reduces the idle time of the system. Thus the client requests are served in real time without too much response delay.

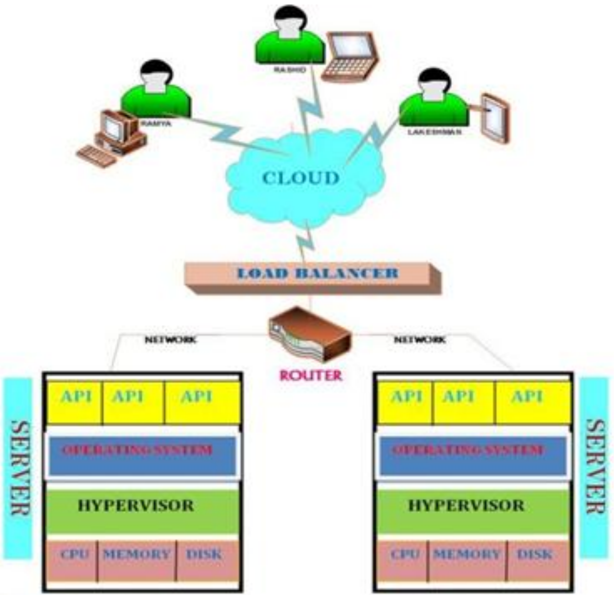

Figure 3. Servers with Virtualization

7. Load Balancing

Load balancing is essentially the construction of call routing schemes which successfully distribute the changing load over the system and minimize lost communication calls [5]. Shortest path algorithm is being implemented in order to locate the nearest available resources. Bottle neck and dead lock situation arise when traffic and workloads is being transferred through a minimum capacity channels. To overcome these flaws fast speed internet cables with high and maximum bandwidth to carry the load across the different networks, DSL cables are used [7].

- **CPU load**
  A CPU load is basically the total number of processes running and the processes which are in waiting queue. A single core processor is like single traffic lane where high traffic usually came into deadlock situation, to overcome this limitation the processes are being put in waiting queue with proper time slice, as multicore processors are used in cloud computing environment at server side which executes the client loads in real time. The task manager will show many variables like the CPU utilization of the resources, memory usage, etc. The below figure represents a simple cpu load where processes are being placed in queue and after some time interval these waiting processes are being executed. Both the cars in queue as well as in processing state represent the total CPU load.

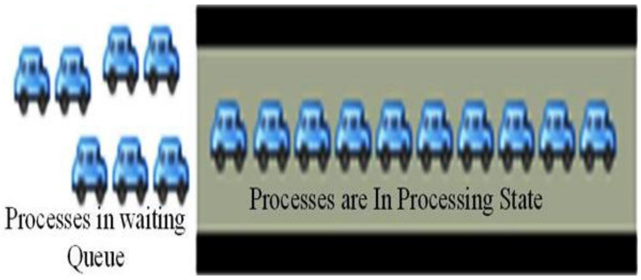

Figure 4. Load Balancing

CPU load is measured as

$$CPU = \sum_{k=1}^{N} (CU/V)$$

Where

N = the number of nodes,
CU = CPU utilization
V = Total CPU's used

- **Memory load**
  Memory is another important part of the computer all the work by the client is stored and retrieved on and from the memory respectively.
- **Network load**
  Network is the technique by which the systems are interconnected. Network systems can be loaded if there is too many systems are to be in condition to execute. The client has to consider the network bandwidth when the client is going to calculate the load.
- **Threshold values**
  The workload calculation has to be done to find the threshold values. The threshold value is the minimum input that produces the corrective action in an automated system. It is considered as two based for computation as upper bound and lower bound. These values has been defined for setting the upper limit and lower limit to do the provisioning and de-provisioning of resources management.

8. Resource Management Algorithm

```
Begin
    While all the system is running
    Calculate the threshold values (TU and TL)
based on load Monitor of the server cluster(S)
    If S>TU
    Add server to the cluster
    Else if(S>TL)
    Release the extra server from cloud
    Else
    Work with existing server
    End
Where
S=Server
TU=Upper Threshold
TL=Lower Threshold
```

III. RESULTS AND DISCUSSION

Experimental Results

The below chart represent the auto scaling of the resources like Servers. When the already allocated server reaches its upper bound from where it goes out of the capacity and demand increase more and more, the additional servers are allocated automatically. This works on some conditional inputs where loop checks the available servers and puts the client requests in a queue, after some time interval it allocates the server and responds to the clients requests automatically. The above mentioned algorithm checks the CPU thresholds like upper and lower bounds.

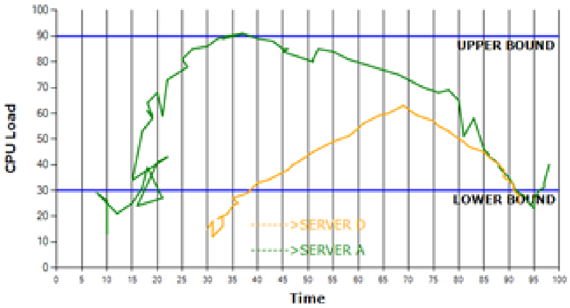

Figure 5. Upper and Loud CPU Bounds

In this chart, the load balancing values are shown, in initial state where server A is serving the clients requests and once it reaches the diminishing size i.e. upper bound where it reaches its serving capacity and cannot serve any more processes from the client, The new server D is being added which take care rest of the clients requests. The above chart is analyzed by its CPU load and Time. Soon after the new allocated server has finished its work, it will be released and will be reallocated to serve for other request of the clients and this process will continue without interrupting smooth functioning of the system in cloud environment. The below figure.6 represents memory usage of the overloaded and newly added servers.

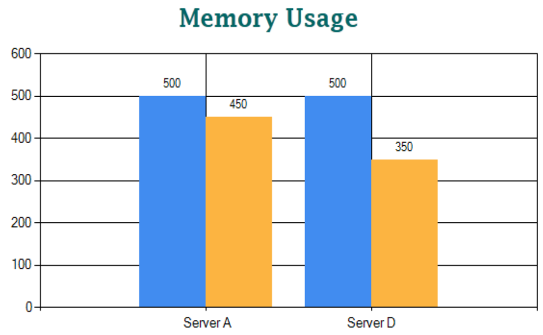

Figure 6. Server Memory usage

IV. CONCLUSION

Cloud platform services are outcomes of both distributed and grid computing infrastructure which develops services and communications across the globe with different clients using it. Scaling as one of the most important features of cloud computing, tries to allocate and pay back resources based on the requirements as the demand and the situation arise. The present study has focused on the problem of auto-scaling cloud environments particularly at infrastructure level. Cloud computing is a widely used technology, characterized by offering resources in scalable and elastic manner. Clients can hold and release resources on demand basis, and pay only for the utilized resources thus reduces the costs of the services which were not of the interest to the clients. The scaling task can be done manually as well, but because of the human intervention chances of occurring errors are high so it is preferable to adopt the automatic or dynamic autoscaling scaling mechanism of the resources running in cloud environment. The element should be able to adapt the amount of required resources and with the advancement in virtualization technology, provides the clients to use the resources fully without having concerns for investing in infrastructure. The auto-scaling process is subject to faults and failures from software, networking and hardware aspects. One scenario is that a certain number of nodes are needed, but only partial nodes are actually launched. When failures like this occur, an auto-scaling mechanism needs to recover in an intelligent way.

V. REFERENCES

[1] Fang Liu, Jin Tong, Jian Mao, Robert Bohn, John Messina, Lee Badger and Dawn Leaf,"NIST Cloud Computing Reference Architecture", NIST Special Publication 500-292, September 2011.

[2] M.Kriushanth, L. Arockiam and G. JustyMirobi,"Auto Scaling in Cloud Computing: An Overview", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 7, July 2013, ISSN (Print) : 2319-5940,ISSN (Online) : 2278-1021.

[3] Tania Lorido-Botran, Jose Miguel-Alonso , Jose A. Lozano, "A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments", ARTICLE in JOURNAL OF GRID COMPUTING DECEMBER 2014, Impact Factor: 1.51 • DOI: 10.1007/s10723-014-9314-7.

[4] ChenhaoQu, Rodrigo N. Calheiros, and RajkumarBuyya,"A Reliable and Cost-Ecient Auto-Scaling System for Web Applications Using Heterogeneous Spot Instances", Cloud Computing and Distributed Systems (CLOUDS) Laboratory, Department of Computing and Information Systems, The University of Melbourne, Australia, September 17, 2015.

[5] Gunpriya Makkar, Pankaj Deep Kaur,"A Review of Load Balancing in Cloud Computing", Guru Nanak Dev University, Jalandhar, India, Volume 5, Issue 4, 2015 ISSN: 2277 128X.

[6] K C Gouda, AnuragPatro, Dines Dwivedi ,NagarajBhat, "Virtualization Approaches in Cloud Computing",International Journal of Computer Trends and Technology (IJCTT),volume 12 Issue 4–June 2014.

[7] Gunpriya Makkar, Pankaj Deep Kaur," A Review of Load Balancing in Cloud Computing",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 4, 2015 ISSN: 2277 128X.