Survey on Load Balancing and Auto Scaling Techniques for Cloud Environment

Pooja C.S, K.R Prasanna Kumar

Abstract: ***Cloud computing became now first choice and priority for every person who access the internet, one of the advantageous features of cloud computing is its scalability and flexibility. Auto scaling offers the facility to the individuals to scale up and scale down the resources as per their requirements, using only the needed resource and paying for what they have used i.e. "Pay-as-you-use". As everything takes place in automatic manner, so human involvement errors are less and reduce the manpower and costs. so to make use of elasticity user must use auto scaling technique that balances the incoming workload, and reduce the total cost and maintain the Service Level Agreement (SLA).In this work main ideas revolve around the problems in scalable cloud computing systems. In modern days, management of resources is in boom and most talked topic in cloud environment. We present some of the existing load balancing policies and about Autoscaling categories.***

Keywords: cloud computing, scaling, auto scaling, load balancing.

I. INTRODUCTION

Cloud Computing is a new and emerged as a booming computational model in IT sector, The services and applications are accessible to the different users using proper internet protocol suit and networking standards. Due to less investment on resources and maintenance cost, everyone is moving to the cloud. Hence it is growing extremely well in business by providing good services to all its users. Cloud based resource management is an approach as cloud offers abundant resources to its users, so the management of these resources are always important for the cloud vendors. .In general anything that involves deliverable hosted services over the internet it is a 'pay as you use" model where the user pay for the requested resources. Auto scaling is a technique when implemented properly can lead to the proper cloud based resource management. Load balancers which act as an interface between the clients and the servers, which balances the client requests and offer them the available resources.

II. METHODS AND MATERIAL

A. Related Work

M. Kriushanth et al.,[1] presented the basics of cloud computing concepts like service models, deployment models and the various dimensions of cloud scalability. They have given dimension as vertical and horizontal scalability. They presented some issues and challenges that are occurring in auto scaling. S.K. Tesfatsion, E. Wadbro, J.Tordsson [2] presented flexible load balancing traffic grooming strategy maintained for system optimization. And Traffic Engineering optimization strategy in the overlay layer is used to optimize the overall performance of the system.

R. Anandhi et al., [3] presented the basics of scalability and its scalability factors. Here they distinguish the scalability by its scalability factors. Then based on the user requirement they described why and how the scalability has been chosen. Further they described approaches in messaging system of scalability. They had given the path to better the scalability through auto scaling.

X.Li, Y.Mao, X.Xiao, Y.Zhuang [4] presented resource utilization achieved and response time of tasks that reduced using Max-Min Task scheduling algorithm which consists of Task Status Tables and Virtual Machine Status Tables and its updates and task allocation algorithms in elastic cloud.

Hanieh Alipour et al., [5] in this paper, they presented a survey that explain definitions of concepts related to auto-scaling and taxonomy of auto-scaling techniques. Based on the results of survey, they outline open issues and future research directions for auto-scaling in cloud computing. They explained each and every concept of the auto scaling important areas. This provides various areas in research sectors especially in the area of auto scaling.

B. Infrastructure-as-a-Service

Infrastructure-as-a-service (Iaas) is a form of cloud computing service that provides virtualized computing resources over the internet. It as the capability to provide processing, storage, networks, and computing resources. The major services include web servers, operating system, virtual instances, load balancing, internet access and bandwidth provisioning. IAAS providers provide Load Balancing technique by automatic scaling facility which sets resources up and down of application. This service requires high Bandwidth capacity, low-latency, Reliable and low cost communication.

C. Scalabilty

One of the key advantages of using cloud-computing paradigm is its scalability. It is regarded as the most important feature of cloud computing. It is about managing unexpected workloads, and it depends on system design, as well as the types of algorithms, data structures and communication mechanism used to implement system components. Clients dynamically or automatically provision their hardware and software resources when demand and situation arise like that by the mechanism.

Manuscript published on 30 June 2017.
* Correspondence Author (s)
Pooja C.S*, Department of Computer Science and Engineering, Siddaganaga Institute of Technology, Tumkur (Karnataka), India. E-mail: poojashekar.c.s@gmail.com
K.R Prasanna Kumar, Department of Computer Science and Engineering, Siddaganaga Institute of Technology, Tumkur (Karnataka), India. E-mail: prasanna.kghatta@sit.ac.in

Cloud computing allows clients or cloud vendors business to easily scale up and scale down their requirements as and when required. For example,
Most cloud service providers will allow clients to increase their existing resources to manage increased business needs. This will allow clients to support their business growth with less expensive changes to the existing systems being used in cloud environment. Cloud computing allows for quick and easy allocation and reallocation of resources in a 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 is virtualization which enables the cloud paradigm to scale up and scale down the resources; without it cloud computing is not sufficient, it provides the agility and speed up the execution of processes.

D. Scalability Issues in Existing Systems

The mechanism of automatically modifying the amount of used resources is known as "auto-scaling services". Auto Scaling is the best solution but as a "two sides of a coin", it is profitable but at other side of the coin it also have some different challenges that need to be marked and find solutions how to track and resolve them. Some of the scalability-issues related with scalability are as:

- Scaling includes different cloud service models, but most studies concentrate on the infrastructure level while as ignoring the other two levels like SaaS and PaaS. Auto-scaling at the service-level is important because services are running on a group of connected virtual machines, and the quality of the service depends on how auto-scaling handles resources for these VMs.
- Insufficient tools at the platform level and service level for managing and assemble metrics to support auto-scaling decisions.
- Auto-scaling in hybrid cloud environments is not supported well. Hybrid clouds applications are deployed on a public and public cloud simultaneously. In this the mutualism, the public and private cloud may provide 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 auto scaling QoS is not properly maintained and managed. Violation of the system's QoS requirements of performance and scalability and even incur unnecessary cost may occur due to failure of auto scaling.

E. Load Balancing:

Load balancing technique provides distributes incoming application traffic across multiple instances. This increases the fault tolerance of your applications. Load balancing is very much important in cloud system as the incoming load is variable and unpredictable also depends on different factors. A good load balancer must meet following requirements [6].

- Better the operation performance significantly
- It must have fault tolerance capability and when system fails due to high load it must provide the backup path
- Most important it should always maintain the stability of the system and perform operations.
- The Load Balancing service automatically routes incoming traffic across such a dynamically changing number of instances.

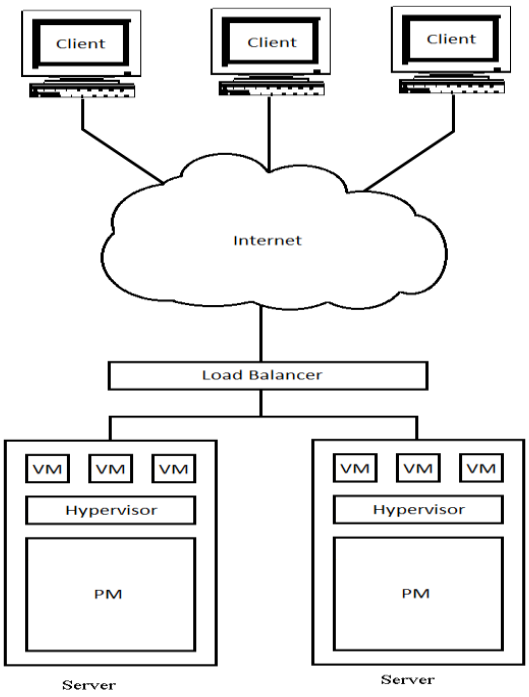

Figure 1 Cloud Architecture

F. Existing Load Balancing Algorithms:

In cloud computing environment load balancing is needed to distribute the dynamic workload evenly between all the nodes. Load balancing is used to make sure that none of your existing resources are idle while others are being utilized[7]. The below fig shows the cloud network general architecture, where users request are maintained by load balancer which starts the services on VM's as per load on each server. The different load balancing strategies are discussed below.

There is one innovative load balancing technique Resource Aware Scheduling algorithm [RASA]. This technique is a combination of Min-min and Max-min strategy. Here the virtual nodes are created first. Then response time of each VM is calculated and then accordingly least loaded node is found and client is given that node. The strategy is to apply Max-min if number of resources is even else Minimum strategy is used. The throttled load balancing algorithm in this algorithm Client request the VM to data center, which forwards it to load balancer. Load balancer maintains an index table of the available VM and their busy or available status. It then returns the ID of the first VM it encounters while scanning the list which can meet requirement to the data center. If the list is scanned completely and no VM is found it returns -1 to data center and data center queues the client request. When VM is available load balancer informs data center and it is allocated to client.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Even this algorithm has used modified approach of ant colony optimization. In the previous approach ant travels in one direction at a time. In this approach it uses two directions of traveling forward and backward. When the ant travels in forward direction it searches for overloaded node following the foraging pheromone and updates foraging trials. Same way, the ant travels in backward direction when it encounters overloaded node by following the trailing pheromone and updates trailing pheromone trials in the path. It considers the minimum migration time of a node to find under loaded node. Load balancing approach is described using honey bee foraging behavior. In this algorithm the tasks are to be send to the under loaded machine and like foraging bee the next tasks are also sent to that virtual machine till the machine gets overloaded. But their strategy considers minimum migration time factor and does not consider all QoS factors.

G. Auto Scaling:

The Auto scaling service helps to dynamically configure the resources to acquire or release instances for a given application. Auto scaling ensures that desired number of instances is running, even if an instance fails, and enables automatically increase or decrease the number of instances as the demand of the instances changes. It also helps to ensure that the enough number of instances are available to handle the load of application.

Auto Scaling helps to ensure that we have the correct number of instances available to handle the load for our application. We can create collections of instances, called Auto Scaling groups. We can specify the minimum and maximum number of instances in each Auto Scaling group, and Auto Scaling ensures that our group never goes below or above the threshold values. If we specify the desired capacity, either when we create the group or at any time, Auto Scaling ensures that our group has this many instances. If we specify scaling policies, then Auto Scaling can start or terminate instances as demand on our application increases or decreases.

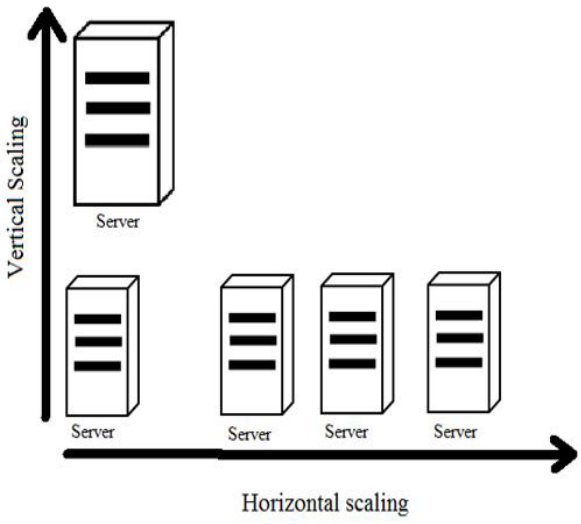

Figure 2: Scaling

Elasticity property of cloud is gaining making cloud more attraction for people. Horizontal and vertical Autoscaling are the two ways by which scaling can be done as shown in the fig 2. Horizontal Autoscaling deals with adding or removing of VM's. Vertical Autoscaling deals with replacing the existing resource with the lower or higher capacity. Different approaches have been followed by researcher, enhancing the scalability. Here are the few strategies of application Autoscaling in cloud.

Horizontal cloud scalability is to connect multiple VM, such as servers so that they work as a single unit. It means adding one or more VM's doing the same job. It is also referred as scale-out. For example, in the case of servers it could increase the speed or availability by adding more servers as per the needs. Instead of one server here one can have two, ten, or more of the same server doing the same work.

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

III. CONCLUSION

In this paper, we have surveyed scalability issues, load balancing techniques and auto scaling categories for cloud computing. Different techniques suggested by authors are discussed in this survey. The main purpose of scaling is adjusting the application instances as per the requirement. Scaling takes load balancing into consideration, by distributing load dynamically among the nodes and to make maximum resource utilization by reassigning the overall load to all the nodes. Thus even and efficient distribution of resources is done and improves the performance of the system.

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. Priyanka P. Kukade and Geetanjali Kale "Survey of Load Balancing and Scaling approaches in cloud" vol.4 Feb 2015.
7. Ashalatha R Evaluation of Auto Scaling and Load Balancing Features in Cloud" vol.117 may 2015.
8. Dr. D .Ravindran, Ab Rashid Dar cloud Based Resource Management with Autoscaling vol.2 .

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)