International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-6, March 2020

Elastic Technique for Load Balancing in Cloud Computing

Sovban Nisar, Deepika Arora, Navneet Verma

Abstract: ***The cloud computing is the architecture that is decentralized in nature due to which various issues in the network get raised which reduces its efficiency. The exchange of data over the network is also continuously increasing. New advanced technology, cloud computing is becoming popular because of providing the above services beneficially. Other vital technologies like virtualization and scalability by designing virtual machines in cloud computing. In cloud computing, web traffic and service provisioning are increasing day by day, so load balancing is becoming a big research issue in cloud computing. Cloud Computing is a new propensity emerging in the IT environment within huge requirements of infrastructure and resources. The load Balancing technique for cloud computing is a vital aspect of the cloud computing environment. Peerless Load balancing scheme ensures splendid resource utilization by provisioning resources to cloud users on-demand services basis in a pay-as-you-use manner. The technique of Load Balancing may further support prioritizing requests of users/clients by applying appropriate scheduling criteria. This paper presents various load balancing schemes in different cloud environments based on requirements specified in the Service Level Agreement (SLA).***

Keywords: Cloud Computing, Load Balancing, Resource Provisioning, Resource Scheduling, Service Level Agreement (SLA).

I. INTRODUCTION

Cloud computing is used in the Internet to consume software or other IT services on demand. Cloud computing is a completely a new technology.

With the use of Cloud Computing users are able to share processing power, storage space, bandwidth, memory and software. As if somebody is using Cloud computing then their resources get shared along with that the cost is also getting shared. This helps user to spend only less cost as they must pay based on usage.

The provider of cloud computing solutions delivers a permission to use its software, hardware, platform, or storage providers like services over the internet [1]. There is such kind of disc or hardware available which user can buy to take the cloud services. Recurring fees is charged by the cloud provider on the monthly basis which is based on the usage by the users.

The evolution of Virtualization, Utility computing, Software-as-a-Service (SaaS), Infrastructure-as-a-Service (IaaS) and Platform-as-a-Service (PaaS) all are combined to make a cloud computing and three development models of cloud are public, private and hybrid. Public cloud services are available for general public over the internet [2]. Private cloud is used for personal use or provides services to single organization. A hybrid cloud is combination of two or more than two public and private cloud which are bounded by service level agreement (SLA). Clients/Users can forward the requests at any time from any geographical location/region for the required services, SLA selects the best resource within user defined deadline and budget. Elastic resource provisioning with quality of service (QoS) parameter (deadline, high availability, priority etc.) is one of the most challenging problem in the field of cloud computing. Hence, the cloud services provider requires/needs an efficient and peerless load balancing algorithm which reduces the make span time as well as task rejection ratio within client/user-defined deadline. It is also a development of distributed, parallel and grid computing. Over other existing computing techniques, the cloud computing is advantageous and it much improve the availability of IT resources. So, with the use of cloud computing users can use the infrastructure of IT and pay for that only which will save the cost to buy the physical resources that may be vacant when it is not in use. The data, operating systems, applications, storage and processing power will be active on use basis. The cloud is just like a space available on web where computing has been already installed to get advantage of it in different services. In the last few years, maximum company is trying to achieve the scalability in terms of platform, application and infrastructure level. Scalability is an important feature in cloud computing and can be divided into two part one is scale up other is scale out. The Term "Scale up" is also known as vertical scalability and the term "scale out" is also known as horizontal scalability.

Manuscript received on February 10, 2020.
Revised Manuscript received on February 20, 2020.
Manuscript published on March 30, 2020.
* Correspondence Author
Sovban Nisar, M.Tech. Department of Computer Science and Engineering. Geeta Engineering CollegeRawalpora, Srinagar-190005E-mail:sovbanbin@gmail.com
Deepika Arora, Asst. Prof. Department of Computer Science and Engineering. Geeta Engineering CollegeNaultha, Panipat, India – 132107Email: deepika.arora86@gmail.com
Navneet Verma, Head of Department of Computer Science and Engineering. Geeta Engineering College Naultha, Panipat, India – 132107Email: hodcse@geeta.edu.in

A decentralized server networks provides different services and computing software both combined to make a cloud computing. Now a day's cloud has been used in number of different applications like Yahoo, Gmail web-based email clients, Wikipedia, YouTube and it has also been used in Skype, Bit Torrent like peer to peer networks.

In cloud environment there is no control over it by any centralized organization there is only need of internet connection along with one web browser nothing more is required to utilize its services. For business world there is enterprise cloud computing. There is need of hardware and centralized infrastructure for running Microsoft, SAP, or Oracle like applications. The office space, power, networks, servers, storage, cooling, and bandwidth are types of infrastructure that is required to run or install it. The complexity of above-mentioned database has been reduced too much extent with the use of cloud that also reduces the total required expenditure.

1.1 Cloud Computing Architecture

In the field of information there is one well accepted institution name National Institute of Standards and Technology (NIST) who has given different definitions of working. There are three cloud services, five essential characteristics and four models of cloud deployment have been defined by NIST in their architecture of Cloud Computing.

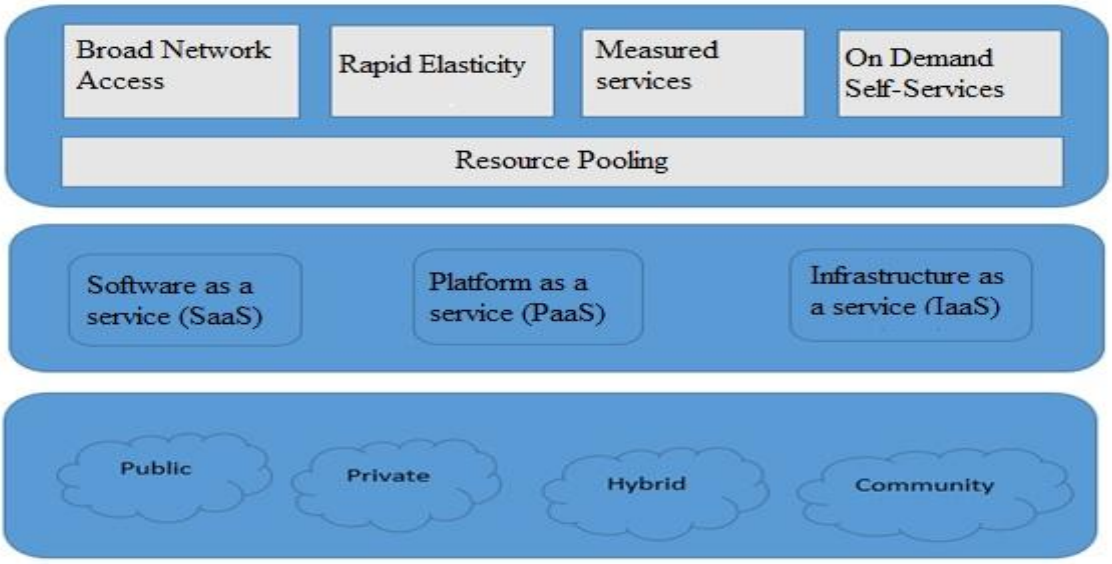

Figure 1 – Diagrammatic model of NIST Working Definition of Cloud Computing

1.2 Cloud Service Models

There are three (3) Cloud-based Services models and these three (3) fundamental classifications are often referred to as "SPI model" i.e. software, platform or infrastructure as a service.

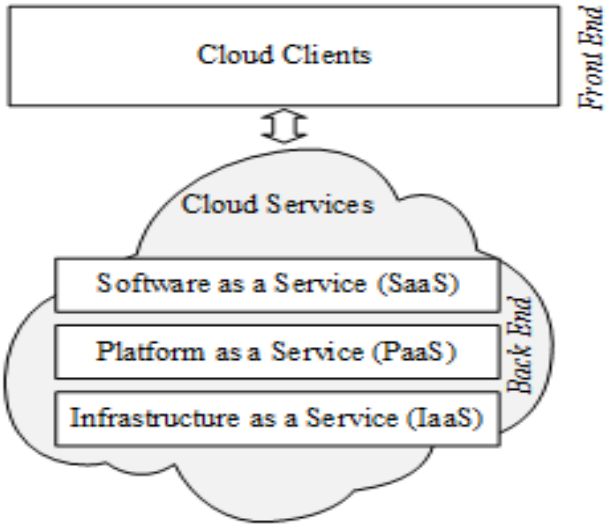

Figure 2- Cloud Computing Architecture

- Cloud Software as Service: There are different applications running by providers on a cloud that can be used by users by using this service.
- Cloud Platform as Service: In cloud infrastructure there are different tools and programming languages that have been provided by provider. The use of this service allows customers to applications acquired or created by customers.
- Cloud Infrastructure as Service: The fundamental computing resources like storage, network and processing provision can be acquired by customer using this type of cloud service. This acquired resource can be used by customer to deploy and run different software or applications.

1.3 Essential Characteristics of Cloud Computing

There are numbers of characteristics of Cloud computing out of them the main characteristics of cloud computing are given below [3]:

Retrieval Number: F7353038620/2020©BEIESP
DOI:10.35940/ijrte.F7353.038620
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

- **On-demand self-service:** Without any human interaction with each provider of services a consumer can A consumer can independently have provision of computing capabilities, such as server time and network storage as needed.

- **Broad network access:** The available capabilities over the network are available and can be accessed using standard mechanism. This promotes the use of heterogeneous thin or thick client platforms (e.g., mobile phones, laptops, and personal digital assistants (PDAs)).

- **Resource pooling:** According to customers demand different virtual and physical resources can be assigned or re assigned using multi-tenant model. These are some cases in which subscriber will only be able to specify the provider location at higher level of abstraction but it unable to known exact location of provided resources. The storage, processing, memory, network bandwidth, and virtual machines are all examples of its resources.

- **Rapid elasticity:** To quickly scale in and scale out the capabilities can be rapidly, elastically and automatically provisioned. At any time and in any quantity the capabilities can be purchased by customers.

Measured Service: In same level of abstraction use of metering capability helps in automatically controlling and optimizing resources that should be appropriate for cloud systems required services. The utilized service customer and provider both resources can be controlled by monitoring it then a transparency report need to provide.

- **Shared Infrastructure** — by utilizing the virtualized software model, it become possible to access the network capabilities, data storage and sharing of physical services. Most of the available infrastructure is provided to the user in the cloud infrastructure, without referring to the deployment models.
- **Dynamic Provisioning** — On the basis of the current demands, it provides the various services. All these services are provided by using software automation that enables the service capability by expanding and contracting as per requirement. In order to maintain the high level of reliability and security, it is required to done dynamic scaling.

- **Network Access** — the network is to be accessed across the internet using various devices such as PCs, laptops, and mobile devices that utilized the standards-based APIs. Everything is included in the cloud such as services of using business applications and the latest application.

- **Managed Metering** — in order to provide the report **and** the billing information, it is necessary to use the metering that manage and optimize all the services. Hence, it is used as the parameter to bill the consumers on the basis of their utilization of services.

1.4 Cloud deployment Strategies

There are mainly three strategies which can be deployed in Cloud computing [4]. The basic cloud computing strategies are given below:

- **Public Cloud:** This type of cloud services is always available to clients through a third-party internet provider. This is very cheap or almost free to use but still interpreting public as free is not always applicable. In public cloud computing vendors provide an access on control mechanism to their users and it does not mean that a user's data is publicly visible to others. To deploy number of solutions an elastic, cost effective means is provided by the public clouds.

- **Private Cloud:** A private cloud computing is elastic, and service based. These are the two benefits of public cloud which is also offered by private clouds. The public and private cloud is different from each other as data has been managed within the organization for private cloud service without any network bandwidth restrictions. There is also provision of security and all legal requirements has been fulfilled which is absent in case of public cloud services. In addition, a great control of the cloud infrastructure is given to provider and it also improves the security and resiliency because the used network is restricted and designated to access by the users [5].

- **Community cloud:** The organizations who are sharing same interests in terms of security and mission uses community cloud. The member of cloud community will be able to access the data and applications available on loud.

- **Hybrid Cloud:** The private and public cloud is combined to take services advantages of both makes it Hybrid Cloud. In this case the business-critical services are controlled by themselves that is using private cloud and other non-business critical information has been outsource and controlled by public cloud.

1.5 Fault Tolerance

The system construction should be done in considering some requirement such as it should be recovering from partial failures without disturbing the overall operation and degrading its performance is the main goal in designing a distributed system [6]. The system should be the one who will continue its working till that the fault has not been resolved. In simple definition, a distributed system is expected to be fault tolerant.
The important and main role of fault tolerance is to try concealing the occurrence of system failures. There are two phases of fault tolerance, such as:

- **Error detection:** It is a form of signal or intimation that provides indication regarding the running processes status.
- **Error Recovery:** It is the process to attempt or fix the faulty/erroneous system state into an error free state.

1.6 Load Balancing Challenges in The Cloud Computing

Although cloud computing has been widely adopted. Research in cloud computing is still in its early stages, and some scientific challenges remain unsolved by the scientific community, particularly load balancing challenges.

- ✓ **Automated service provisioning:** A key feature of cloud computing is elasticity, resources can be allocated or released automatically. How then can we use or release the resources of the cloud, by keeping the same performance as traditional systems and using optimal resources?
- ✓ **Virtual Machines Migration:** With virtualization, an entire machine can be seen as a file or set of files, to unload a physical machine heavily loaded, it is possible to move a virtual machine between physical machines. The main objective is to distribute the load in a datacenter or set of datacenters. How then can we dynamically distribute the load when moving the virtual machine to avoid bottlenecks in Cloud computing systems?
- ✓ **Energy Management:** The benefits that advocate the adoption of the cloud is the economy of scale. Energy saving is a key point that allows a global economy where a set of global resources will be supported by reduced providers rather that each one has its own resources. How then can we use a part of datacenter while keeping acceptable performance?
- ✓ **Stored data management:** In the last decade data stored across the network has an exponential increase even for companies by outsourcing their data storage or for individuals, the management of data storage or for individuals, the management of data storage becomes a major challenge for cloud computing. How can we distribute the data to the cloud for optimum storage of data while maintaining fast access?
- ✓ **Emergence of small data centers for cloud computing:** Small datacenters can be more beneficial, cheaper and less energy consumer than large datacenter. Small providers can deliver cloud computing services leading to geo-diversity computing. Load balancing will become a problem on a global scale to ensure an adequate response time with an optimal distribution of resources.

1.7 Experimental Results for Research

Load balancing is one of the central issues in cloud computing. It is a mechanism that distributes the dynamic local workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It helps to achieve high user satisfaction and resource utilization ratio, hence improving the overall performance and resource utilization of the system. It also ensures that every computing resource is distributed efficiently and fairly. It further prevents bottlenecks of the system which may occur due to load imbalance. When one or more components of any service failure, load balancing helps in the continuation of the service by implementing fair-over, i.e. in provisioning and de-provisioning of instances of applications without fail.

The goal of load balancing is improving the performance by balancing the load among these various resources (network links, central processing units, disk drives.) to achieve optimal resource utilization, maximum throughput, maximum response time, and avoiding overload. To distribute the load on different systems, different load balancing algorithms are used.

In general, load balancing algorithms follow two major classifications:

- Depending on how the charge is distributed and how processes are allocated to nodes.
- Depending on the information status of the nodes (System Topology).

In the first case, it designed as a centralized approach, distributed approach or hybrid approach in the second case as a static approach, dynamic or adaptive approach.

1) Classification According to the System Load

- Centralized approach: In this approach, a single node is responsible for managing the distribution within the whole system.
- Distributed approach: In this approach, each node independently builds its load vector by collecting the load information of other nodes. Decisions are made locally using local load vectors. This approach is more suitable for widely distributed systems such as cloud computing.
- Mixed approach: A combination of the two approaches to take advantage of each approach.

2) Classification According to the System Topology

- Static approach: This approach is generally defined in the design or implementation of the system.
- Dynamic approach: This approach takes into account the current state of the system during load balancing decisions. This approach is more suitable for widely distributed systems such as cloud computing.
- Adaptive approach: This approach adapts the load distribution to system status changes, by changing their parameters dynamically and even their algorithms. This approach can offer better performance when the system state changes frequently. This approach is more suitable for widely distributed systems such as cloud computing.

II. CONCLUSION

Now-a-days Cloud Computing is used by the industries while as there are many already concerns in it that needs to be addressed, which include: -

Load balancing, Virtual machine migration, Server Consolidation, Power Management, etc. These concerns are not fully yet addressed. In between all these concerns, Load Balancing is the main issue which is used to distribute the access dynamic local workload positively to all the associated nodes in the whole Cloud for achieving user satisfaction and resource utilization ratio.

Load balancing make sure all the computing resources are distributed efficiently and fairly to give best performance to the users. I want to put forth a concept through this paper related to Cloud Computing along with research challenges in Load balancing. My focus will be on the study of Load balancing algorithm in cloud computing with respect to the scalability, resource utilization performance response time and overhead associated.

REFERENCES

1. George Suciu, Cristina Butca, Victor Suciu, Alin Geaba, Alexandru Stancu, Stefan Arseni. Basic Internet Foundation and Cloud Computing. IEEE 2015 10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 56(2016), 278-284.
2. Bharath Balasubramanian, Mung Chiang, and Flavio Bonomi. Introduction. IEEE, 41(2015), 304-313.
3. Hang Liu, Fahima Eldarrat, Hanen Alqahtani, Alex Reznik, Xavier de Foy, and Yanyong Zhang. Mobile Edge Cloud System: Architectures, Challenges, and Approaches. IEEE SYSTEMS JOURNAL, 99(2017), 1-14.
4. L. Yang, J. Cao, S. Tang, T. Li, and A. Chan, "A framework for partitioning and execution of data stream applications in mobile cloud computing," in Proc. IEEE 5th Int. Conf. Cloud Comput, 20(2012), 794–802.
5. Rakesh Bhatnagar, Dr. Jayesh Patel, Nirav Vasoya. Dynamic Resource Allocation in SCADY Grid Toolkit. IEEE International Conference on Computing, Communication and Automation (ICCCA-2015), 45 (2015) 15-16.
6. Deepali Mittal, Neha Agarwal. A review paper on Fault Tolerance in Cloud Computing. IEEE, 56(2015), 97-113.

AUTHOR PROFILE

Sovban Nisar is Student of M.Tech. in the Department of Computer Science and Engineering at Geeta Engineering College, Naultha, Panipat, India. He has done B.Tech. from Kurukshetra University, India.

Deepika Arora[2] is Assistant Professor in Department of Computer Science and Engineering, Geeta Engineering College, Panipat, She has Qualified M. Tech. (CSE) from M.M. University, Mullana (Ambala) and B.Tech (CSE) from Kurukshetra University. Her areas of interest are Adhoc Networks and Cloud Computing.

Navneet Verma[3] is Head of Department of Computer Science and Engineering, Geeta Engineering College, Panipat, He has Qualified M. Tech. (CSE) from M.M. University, Mullana (Ambala) and B.Tech (CSE) from Kurukshetra University. He is Pursuing P.H.D from DCRUST, Sonipat. His areas of interest are IoT.