A SURVEY ON LOAD BALANCING IN CLOUD ENVIRONMENT

Sumanpreet Kaur [(1)], Mr. Navtej Singh Ghumman [(2)]
[(1)] Research Scholar, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
sumanmanes91@gmail.com
[(2)] Assistant Professor, Department of Computer Science & Engineering, SBSSTC, Ferozepur, Punjab.
navtejghumman@yahoo.com

ABSTRACT

In present days cloud computing is one of the greatest platform which provides storage of data in very lower cost and available for all time over the internet.But it has more critical issue like security, load management and fault tolerance. In this paper we are discussing Load Balancing approach. Resource scheduling management design on Cloud computing is an important problem. Scheduling model, cost, quality of service, time, and conditions of the request for access to services are factors to be focused. A good task scheduler should adapt its scheduling strategy to the changing environment and load balancing Cloud task scheduling policy. Cloud Computing is high utility software having the ability to change the IT software industry and making the software even more attractive. It has also changed the way IT companies used to buy and design hardware. The elasticity of resources without paying a premium for large scale is unprecedented in the history of IT industry. The increase in web traffic and different services are increasing day by day making load balancing a big research topic. Cloud computing is a new technology which uses virtual machine instead of physical machine to host, store and network the different components. Load balancers are used for assigning load to different virtual machines in such a way that none of the nodes gets loaded heavily or lightly. The load balancing needs to be done properly because failure in any one of the node can lead to unavailability of data.

Keywords

Cloud Computing; Datacenter Broker; Virtual Machine; Host; Load balancing.

1. INTRODUCTION

Cloud computing is one of the internet based service provider which allows users to access services on demand [1]. It provides pool of shared resources of information, software, databases and other devices according to the client request on “pay as you go” basis [2]. Cloud computing architectures are inherently parallel, distributed and serve the needs of multiple clients in different scenarios. This distributed architecture deploys resources distributively to deliver services efficiently to users in different geographical channels [3]. Clients in a distributed environment generate request randomly in any processor. So the major drawback of this randomness is associated with task assignment. The unequal task assignment to the processor creates imbalance i. e, some of the processors are overloaded and some of them are underloaded [4]. The objective of load balancing is to transfer the load from overloaded process to underloaded process transparently. The major issues associated with load balancing are flexibility and reliability of services [5]. Cloud computing implements virtualization technique in which a single system can be virtualized into number of virtual systems [6]. Load balancing decides which client will use the virtual machine and which requesting machines will be put on hold. 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. 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 changes a physical computing environment and 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, where multiple requests or tasks can execute in multiple machines [7].

2. CLOUD COMPUTING- THE CONCEPT

Cloud Computing aims at providing a plethora of services and applications to the users all around the world through the internet. The service models of cloud include Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [1]. From business organizations to IT companies to end users, cloud services have brought immense benefits and cost savings.

There are five basic characteristics of cloud which include on- demand computing, broad network access, resource pooling, rapid elasticity and measured service. The first characteristic ensures that cloud delivers services to the users whenever they demand and whatever they demand. Broad network access means that cloud services can be accessed from anywhere at any time using either a smartphone, laptop, desktop or tablet. The only thing required is an internet connection. The cloud vendor's resources are pooled so that mUltiple customers are served using multi-tenant model. This servicing is done with the same physical resources by making the resources separate at logical level. One of the most important characteristic of cloud is its rapid elasticity and scalability. The number of cloud users can scale up or scale depending upon the need. The best thing about cloud is that user has to pay only for what he uses. Storage levels, number of user accounts, bandwidth and processing all can be measured so as to do appropriate and to the point billing.

Also the total quantity of resources used by a user can be measured from both the cloud vendor's side as well as customer's side which makes the billing transparent.

The basic concept on which cloud computing relies is Virtualization. Virtualization is a process of creating multiple virtual machine instances for a single host machine in order to serve more number of users [2]. It makes possible to allow mUltiple operating systems and various applications to be run on same server node at a time. Virtualization technology has incremented the flexibility and security of cloud.

Cloud Computing

Figure 1. Cloud Computing Model

3. TYPES OF CLOUDS

Clouds are divided into 4 categories:-

1) Public Cloud: - Public cloud [9] allows users to access the cloud publicly. It is access by interfaces using internet browsers. Users pay only for that time duration in which they use the service, i.e., pay-per-use.

2) Private Cloud: - A private clouds [10] operation is with in an organization’s internal enterprise data center. The main advantage here is that it is very easier to manage security in public cloud. Example of private cloud in our daily life is intranet.

3) Hybrid Cloud: - It is a combination of public cloud [11] and private cloud. .It provide more secure way to control all data and applications .It allows the party to access information over the internet. It allows the organization to serve its needs in the private cloud and if some occasional need occurs it asks the public cloud for some computing resources.

4) Community Cloud:- When cloud infrastructure construct by many organizations jointly, such cloud model is called as a community cloud. The cloud infrastructure could be hosted by a third-party provider or within one of the organizations in the community.

4. SERVICES OF CLOUD MODEL

There are different types of services are provides by cloud models like: Software as a Service(SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS) [6] which are deployed as public cloud, private cloud , community cloud and hybrid clouds.

1) Software as a Service (SaaS):- The capability provided to the consumer is to use the some applications which is running on a cloud infrastructure. The applications are accessible from many devices through an interface such as a web browser (e.g., web-based email). The consumer does not control the cloud infrastructure which includes network, and servers, all operating systems, and provides storages.

2) Platform as a Service (PaaS):- PaaS [5] provides all the resources that are required for implementation of applications and all services completely from the Internet. In this no downloading or installing is required of any software. The capability provided to the consumer is to deploy onto the cloud infrastructure .Consumer uses all the applications by using different programming languages and tools which are provide by the provider. Any consumer has not any control on cloud infrastructure including all networks, servers and operating systems, but has control over the applications which they deployed.

3) Infrastructure as a Service (IaaS):- The capability provided to the consumer is to access all the processing, storage, networks and other many fundamental computing resources. Consumer is able to deploy arbitrary software, which can include operating systems and applications. The consumer does not manage or control the underlying cloud infrastructure but has control over operating systems, storage, deployed application, and possibly limited control of select networking components. Infrastructure Providers manage a large set of computing resources, which include as storing and processing capacity. In this virtualization is used to split, assign and dynamically resize these resources .They deploy only those software stacks that run their required services. It can be used to avoid buying, housing, and managing the basic hardware and software infrastructure components, scales up and down quickly to meet demand.

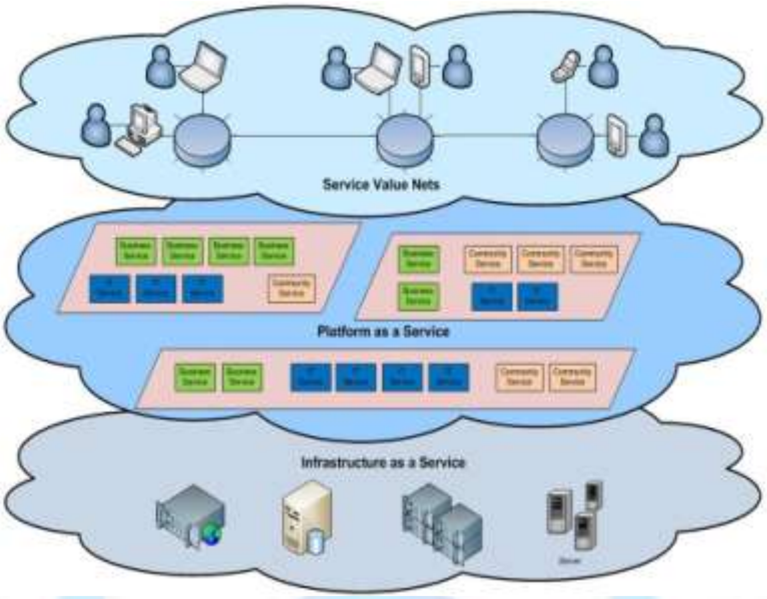

Figure 2. Architecture of Cloud

5. RELATED WORK

Yang Xu et al. (2011) put forwards a novel model to balance data distribution to improve cloud computing performance in data-intensive applications, such as distributed data mining. By extending the classic Map Reduce model with an agent-aid layer and abstracting working load requests for data blocks as tokens, the agents can reason from previously received tokens about where to send other tokens in order to balance the working tasks and improve system performance. Our key contribution lies in building an efficient token routing algorithm in spite of agents' unknowing to the global state of data distribution in cloud.

O. M. Elzeki et al. (2012) proposed a unique modification of Max-min algorithm. The algorithm is built based on comprehensive study of the impact of RASA algorithm in scheduling tasks and the atom concept of Max-min strategy. An Improved version of Max-min algorithm is proposed to outperform scheduling map at least similar to RASA map in total complete time for submitted jobs. Improved Max-min is based on the expected execution time instead of complete time as a selection basis.

Kumar Nishant et al. (2012) proposed an algorithm for load distribution of workloads among nodes of a cloud by the use of Ant Colony Optimization (ACO). This is a modified approach of ant colony optimization that has been applied from the perspective of cloud or grid network systems with the main aim of load balancing of nodes. This modified algorithm has an edge over the original approach in which each ant build their own individual result set and it is later on built into a complete solution. The system, which is incurring a cost for the user should function smoothly and should have algorithms that can continue the proper system functioning even at peak usage hours.

Ms. Parin. V. Patel et al. (2012) has discussed load Balancing approach. Many types of load concern with cloud like memory load, CPU load and network load. Load balancing is the process of distributing load over the different nodes which provides good resource utilization when nodes are overloaded with job. Load balancing has to handle the load when one node is overloaded. When node is overloaded at that time load is distributed over the other ideal nodes. Many algorithms are available for load balancing like Static load balancing and Dynamic load balancing.

Rajwinder Kaur et al. (2013) discusses the main challenges in cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed. It helps in optimal utilization of resources and hence in enhancing the performance of the system. A few existing scheduling algorithms can maintain load balancing and provide better strategies through efficient job scheduling and resource allocation techniques as well. In computing, the load may be of various types like memory load, CPU load and network load etc. Load balancing is the process of searching overloaded node and transferring the extra load of the overloaded node to other nodes.

Amandeep Kaur Sidhu et al. (2013) aims to share data, calculations, and service transparently over a scalable network of nodes. Since Cloud computing stores the data and disseminated resources in the open environment. So, the amount of data storage increases quickly. In the cloud storage, load balancing is a key issue. It would consume a lot of cost to maintain load information, since the system is too huge to timely disperse load.

S.Yakhchi et al. (2015) discusses that the energy consumption has become a major challenge in cloud computing infrastructures. They proposed a novel power aware load balancing method, named ICAMMT to manage power consumption in cloud computing data centers. We have exploited the Imperialism Competitive Algorithm (ICA) for detecting over utilized hosts and then we migrate one or several virtual machines of these hosts to the other hosts to decrease their utilization. Finally, we consider other hosts as underutilized host and if it is possible, migrate all of their VMs to the other hosts and switch them to the sleep mode.

Surbhi Kapoor et al. (2015) aims at achieving high user satisfaction by minimizing response time of the tasks and improving resource utilization through even and fair allocation of cloud resources. The traditional Throttled load balancing algorithm is a good approach for load balancing in cloud computing as it distributes the incoming jobs evenly among the VMs. But the major drawback is that this algorithm works well for environments with homogeneous VMS, does not considers the resource specific demands of the tasks and has additional overhead of scanning the entire list of VMs every time a task comes. The issues have been addressed by proposing an algorithm Cluster based load balancing which works well in heterogeneous nodes environment, considers resource specific demands of the tasks and reduces scanning overhead by dividing the machines into clusters .

Shikha Garg et al. (2015) aims to distribute workload among multiple cloud systems or nodes to get better resource utilization. It is the prominent means to achieve efficient resource sharing and utilization. Load balancing has become a challenge issue now in cloud computing systems. To meets the user's huge number of demands, there is a need of distributed solution because practically it is not always possible or cost efficient to handle one or more idle services. Servers cannot be assigned to particular clients individually. Cloud Computing comprises of a large network and components that are present throughout a wide area. Hence, there is a need of load balancing on its different servers or virtual machines. They have proposed an algorithm that focuses on load balancing to reduce the situation of overload or under load on virtual machines that leads to improve the performance of cloud substantially.

Reena Panwar et al. (2015) describes that the cloud computing has become essential buzzword in the Information Technology and is a next stage the evolution of Internet, The Load balancing problem of cloud computing is an important problem and critical component adequate operations in cloud computing system and it can also prevent the rapid development of cloud computing. Many clients from all around the world are demanding the various services rapid rate in the recent time. Although various load balancing algorithms have been designed that are efficient in request allocation by the selection of correct virtual machines. A dynamic load management algorithm has been proposed for distribution of the entire incoming request among the virtual machines effectively.

Mohamed Belkhouraf et al. (2015) aims to deliver different services for users, such as infrastructure, platform or software with a reasonable and more and more decreasing cost for the clients. To achieve those goals, some matters have to be addressed, mainly using the available resources in an effective way in order to improve the overall performance, while taking into consideration the security and the availability sides of the cloud. Hence, one of the most studied aspects by researchers is load balancing in cloud computing especially for the big distributed cloud systems that deal with many clients and big amounts of data and requests. The proposed approach mainly ensures a better overall performance with efficient load balancing, the continuous availability and a security aspect.

Lu Kang et al. (2015) improves the weighted least connections scheduling algorithm, and designs the Adaptive Scheduling Algorithm Based on Minimum Traffic (ASAMT). ASAMT conducts the real-time minimum load scheduling to the node service requests and configures the available idle resources in advance to ensure the service QoS requirements. Being adopted for simulation of the traffic scheduling algorithm, OPNET is applied to the cloud computing architecture.

6. LOAD BALANCING

Load balancing is a relatively new technique that facilitates networks and resources by providing a maximum throughput with minimum response time [8]. Dividing the traffic between servers, data can be sent and received without major delay. Different kinds of algorithms are available that helps traffic loaded between available servers [2B]. A basic example of load balancing in our daily life can be related to websites. Without load balancing, users could experience delays, timeouts and possible long system responses. Load balancing solutions usually apply redundant servers which help a better distribution of the communication traffic so that the website availability is conclusively settled [8]. There are many different kinds of load balancing algorithms available, which can be categorized mainly into two groups. The following section will discuss these two main categories of load balancing algorithms.

6.1. Static Algorithms

Static algorithms divide the traffic equivalently between servers. By this approach the traffic on the servers will be disdained easily and consequently it will make the situation more imperfectly. This algorithm, which divides the traffic equally, is announced as round robin algorithm. However, there were lots of problems appeared in this algorithm. Therefore, weighted round robin was defined to improve the critical challenges associated with round robin. In this algorithm each servers have been assigned a weight and according to the highest weight they received more connections. In the situation that all the weights are equal, servers will receive balanced traffic [5].

6.2. Dynamic Algorithms

Dynamic algorithms designated proper weights on servers and by searching in whole network a lightest server preferred to balance the traffic. However, selecting an appropriate server needed real time communication with the networks, which will lead to extra traffic added on system. In comparison between these two algorithms, although round robin algorithms based on simple rule, more loads conceived on servers and thus imbalanced traffic discovered as a result [5]. However; dynamic algorithm predicated on query that can be made frequently on servers, but sometimes prevailed traffic will prevent these queries to be answered, and correspondingly more added overhead can be distinguished on network.

7. CONCLUSION

This paper is based on cloud computing technology which has a very vast potential and is still unexplored. The capabilities of cloud computing are endless. Cloud computing provides everything to the user as a service which includes platform as a service, application as a service, infrastructure as a service. One of the major issues of cloud computing is load balancing because overloading of a system may lead to poor performance which can make the technology unsuccessful. So there is always a requirement of efficient load balancing algorithm for efficient utilization of resources.

REFERENCES

[1] S. Yakhchi, S. Ghafari, M. Yakhchi, M. Fazeli and A. Patooghy, "ICA-MMT: A Load Balancing Method in Cloud Computing Environment," IEEE, 2015.

[2] S. Kapoor and D. C. Dabas, "Cluster Based Load Balancing in Cloud Computing," IEEE, 2015.

[3] S. Garg, R. Kumar and H. Chauhan, "Efficient Utilization of Virtual Machines in Cloud Computing using Synchronized Throttled Load Balancing," 1st International Conference on Next Generation Computing Technologies (NGCT-2015), pp. 77-80, 2015.

[4] R. Panwar and D. B. Mallick, "Load Balancing in Cloud Computing Using Dynamic Load Management Algorithm," IEEE, pp. 773-778, 2015.

[5] M. Belkhouraf, A. Kartit, H. Ouahmane, H. K. Idrissi,, Z. Kartit and M. . E. Marraki, "A secured load balancing architecture for cloud computing based on multiple clusters," IEEE, 2015.

[6] L. Kang and X. Ting, "Application of Adaptive Load Balancing Algorithm Based on Minimum Traffic in Cloud Computing Architecture," IEEE, 2015.

[7] N. K. Chien, N. H. Son and H. D. Loc, "Load Balancing Algorithm Based on Estimating Finish Time of Services in Cloud Computing," ICACT, pp. 228-233, 2016.

[8] H. H. Bhatt and H. A. Bheda, "Enhance Load Balancing using Flexible Load Sharing in Cloud Computing," IEEE, pp. 72-76, 2015.

[9] S. S. MOHARANA, R. D. RAMESH and D. POWAR, "ANALYSIS OF LOAD BALANCERS IN CLOUD COMPUTING," International Journal of Computer Sciencand Engineering (IJCSE) , pp. 102-107, 2013.

[10] M. P. V. Patel, H. D. Patel and . P. J. Patel, "A Survey On Load Balancing In Cloud Computing," International Journal of Engineering Research & Technology (IJERT), pp. 1-5, 2012.

[11] R. Kaur and P. Luthra, "LOAD BALANCING IN CLOUD COMPUTING," Int. J. of Network Security, , pp. 1-11, 2013.

[12] Kumar Nishant, , P. Sharma, V. Krishna, Nitin and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," IEEE, pp. 3-9, 2012.

[13] Y. Xu, L. Wu, L. Guo,, Z. Chen, L. Yang and Z. Shi, "An Intelligent Load Balancing Algorithm Towards Efficient Cloud Computing," AI for Data Center Management and Cloud Computing: Papers from the 2011 AAAI Workshop (WS-11-08), pp. 27-32, 2011.

[14] A. K. Sidhu and S. Kinger, "Analysis of Load Balancing Techniques in Cloud Computing," International Journal of Computers & Technology Volume 4 No. 2, March-April, 2013, ISSN 2277-3061 , pp. 737-741, 2013.

[15] O. M. Elzeki , M. Z. Reshad and M. A. Elsoud , "Improved Max-Min Algorithm in Cloud Computing," International Journal of Computer Applications (0975 – 8887), pp. 22-27, 2012.

[16] B. Kruekaew and W. Kimpan, "Virtual Machine Scheduling Management on Cloud Computing Using Artificial Bee Colony," Proceedings of the International MultiConference of Engineers and Computer Scientists 2014 Vol I,IMECS 2014, 2014.

[17] R.-S. Chang, J.-S. Chang and P.-S. Lin, "An ant algorithm for balanced job scheduling in grids," Future Generation Computer Systems 25 (2009) 20–27, pp. 21-27, 2009.

[18] Z. Chaczko, V. Mahadevan, S. Aslanzadeh and C. Mcdermid, "Availability and Load Balancing in Cloud Computing," International Conference on Computer and Software Modeling IPCSIT vol.14 (2011) © (2011) IACSIT Press, Singapore, pp. 134-140, 2011.

[19] R. K. S, S. V and V. M, "Enhanced Load Balancing Approach to Avoid Deadlocks in Cloud," Special Issue of International Journal of Computer Applications (0975 – 8887) on Advanced Computing and Communication Technologies for HPC Applications - ACCTHPCA, June 2012, pp. 31-35, 2012.

[20] Kumar Nishant, , P. Sharma, V. Krishna, N. and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization," IEEE, pp. 3-9, 2012.