CLOUD COMPUTING ENVIRONMENT: A REVIEW

Ishaan Chawla [1]

[1] Student, Department of Computer Engineering, SBSSTC, Ferozepur

Email Id - ishaanchawla0001@gmail.com

ABSTRACT

Cloud computing is a vigorous technology by which a user can get software, application, operating system and hardware as a service without actually possessing it and paying only according to the usage. Cloud Computing is a hot topic of research for the researchers these days. With the rapid growth of Internet technology cloud computing have become main source of computing for small as well big IT companies. In the cloud computing milieu the cloud data centers and the users of the cloud-computing are globally situated, therefore it is a big challenge for cloud data centers to efficiently handle the requests which are coming from millions of users and service them in an efficient manner.Cloud computing is Internet based development and use of computer technology. It is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them. Scheduling is one of the core steps to efficiently exploit the capabilities of heterogeneous computing systems. On cloud computing platform, load balancing of the entire system can be dynamically handled by using virtualization technology through which it becomes possible to remap virtual machine and physical resources according to the change in load. However, in order to improve performance, the virtual machines have to fully utilize its resources and services by adapting to computing environment dynamically. The load balancing with proper allocation of resources must be guaranteed in order to improve resource utility. Load balancing is a critical aspect that ensures that all the resources and entities are well balanced such that no resource or entity neither is under loaded nor overloaded. The load balancing algorithms can be static or dynamic. Load balancing in this environment means equal distribution of workload across all the nodes. Load balancing provides a way of achieving the proper utilization of resources and better user satisfaction. Hence, use of an appropriate load balancing algorithm is necessary for selecting the virtual machines or servers. This paper focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. In this paper, we have reviewed several existing load balancing mechanisms and we have tried to address the problems associated with them.

INTRODUCTION

Internet has been a driving force towards the various technologies that have been developed since its inception. Arguably, one of the most discussed among all of them is Cloud Computing. Over the last few years, Cloud computing paradigm has witnessed an enormous shift towards its adoption and it has become a trend in the information technology space as it promises significant cost reductions and new business potential to its users and providers [1]. Cloud computing can be defined as "Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers".

Cloud computing [1] is an emerging paradigm in the computer industry where the computing is moved to a cloud of computers. It has become one of the buzzwords of the industry. The core concept of cloud computing is, quite simply, that the vast computing resources that we need will reside somewhere out there in the cloud of computers and we'll connect to them and use them as and when needed. Computing can be described as any activity of using and/or developing computer hardware and software. It includes everything that sits in the bottom layer, i.e. everything from raw compute power to storage capabilities. Cloud computing [1] ties together all these entities and delivers them as a single integrated entity under its own sophisticated management. Cloud is a term used as a metaphor for the wide area networks (like internet) or any such large networked environment. It came partly from the cloud-like symbol used to represent the complexities of the networks in the schematic diagrams. It represents all the complexities of the network which may include everything from cables, routers, servers, data centers and all such other devices. Computing started off with the mainframe era. There were big mainframes and everyone connected to them via "dumb" terminals. This old model of business computing was frustrating for the people sitting at the dumb terminals because they could do only what they were "authorized" to do. They were dependent on the computer administrators to give them permission or to fix their problems. They had no way of staying up to the latest innovations. The personal computer was a rebellion against the tyranny of centralized computing [4] operations. There was a kind of freedom in the use of personal computers. But this was later replaced by server architectures with enterprise servers and others showing up in the industry. This made sure that the computing was done and it did not eat up any of the resources that one had with him. All the computing was performed at servers. Internet grew in the lap of these servers. With cloud computing we have come a full circle. We come back to the centralized computing infrastructure. But this time it is something which can easily be accessed via the internet and something over which we have all the control. Cloud computing is Internet ("cloud") based development and use of computer technology ("computing"). It is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the Internet. Users need not have knowledge of, expertise in, or control over the technology infrastructure "in the cloud" that supports them. Cloud computing[4] involves distributed computing over a network, where a program or application may run on many connected computers at the same time. The cloud makes it possible for you to access your information from anywhere at any time. While a traditional computer setup requires you to be in the same location as your data storage device, the cloud takes away that step. The cloud removes the need for you to be in the same physical location as the hardware that stores your data.Depending on customer's requirement CC provides easy and customizable services to access or work with cloud applications. Based on the user requirement CC can be used to provide platform for designing applications, infrastructure to store and work on company's data and also provide applications to do user's routine tasks. When a customer chooses to use cloud services, data stored in the local repositories will be sent to a remote data center. This data in remote locations can be accessed or managed with the help of services provided by cloud service providers. This makes clear that for a user to store or process a piece of data in cloud, he/she needs to transmit the data to a remote server over a channel (internet). This data processing and storage needs to be done with utmost care to avoid data breaches. It is the model for convenient on-demand network access, with minimum management efforts for easy and fast network access to resources that are ready to use. It is an upcoming paradigm that offers tremendous advantages in economic aspects, such as reduced time to market, flexible computing capabilities, and limitless computing power. Popularity of cloud computing is increasing day by day

in distributed computing environment. There is a growing trend of using cloud environments for storage and data processing needs. To use the full potential of cloud computing, data is transferred, processed, retrieved and stored by external cloud providers. However, data owners are very skeptical to place their data outside their own control sphere.

NEED FOR CLOUD COMPUTING

What could we do with 1000 times more data and CPU power? One simple question. That's all it took the interviewers to bewilder the confident job applicants at Google. This is a question of relevance because the amount of data that an application handles is increasing day by day and so is the CPU power that one can harness.

There are many answers to this question. With this much CPU power, we could scale our businesses to 1000 times more users. Right now, we are gathering statistics about every user using an application. With such CPU power at hand, we could monitor every single user click and every user interaction such that we can gather all the statistics about the user. We could improve the recommendation systems of users. We could model better price plan choices. With this CPU power, we could simulate the case where we have say 1, 00,000 users in the system without any glitches. There are lots of other things we could do with so much CPU power and data capabilities. But what is keeping us back. One of the reasons is the large-scale architecture which comes with these are difficult to manage. There may be many different problems with the architecture we have to support. The machines may start failing, the hard drives may crash, the network may go down and many other such hardware problems. The hardware has to be designed such that the architecture is reliable and scalable. This large-scale architecture has a very expensive upfront and has high maintenance costs. It requires different resources like machines, power, cooling, etc. The system also cannot scale as and when needed and so is not easily reconfigurable. The resources are also constrained by the resources. As the applications become large, they become I/O bound. The hard drive access speed becomes a limiting factor. Though the raw CPU power available may not be a factor, the amount of RAM available clearly becomes a factor. This is also limited in this context. If at all the hardware problems are managed very well, there arises the software problems. There may be bugs in the software using this much of data. The workload also demands two important tasks for two completely different people. The software has to be such that it is bug free and has good data processing algorithms to manage all the data.
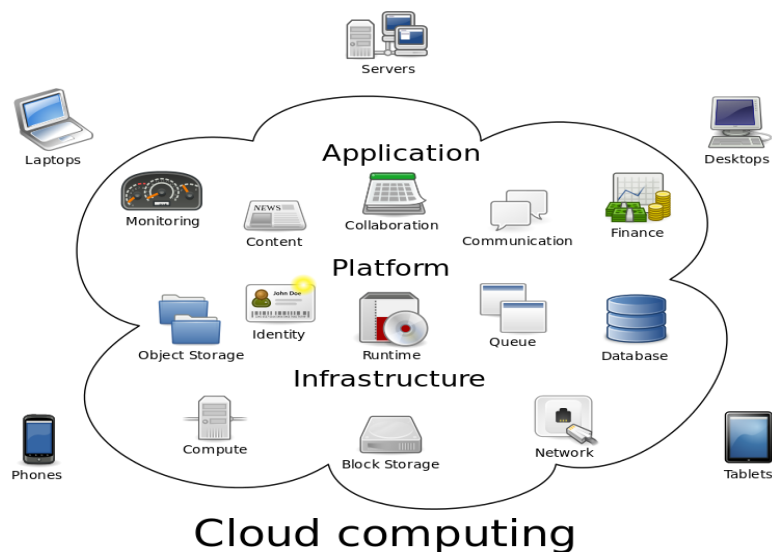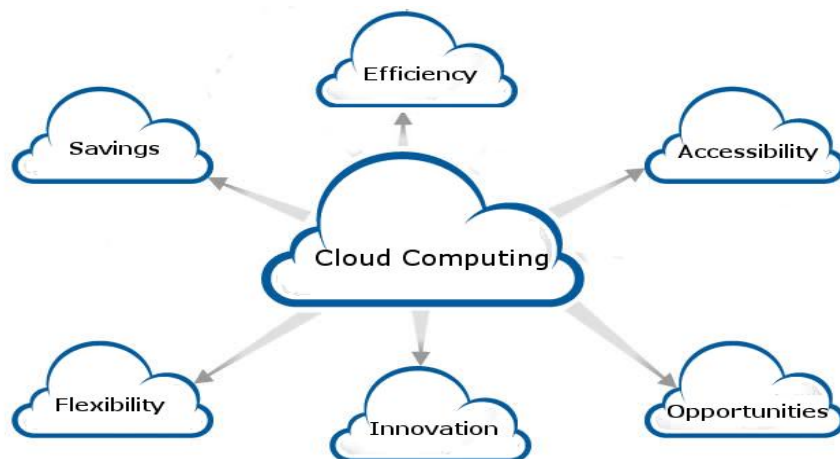

Figure 1. Cloud Computing Model

BENIFITS OF CLOUD COMPUTING

Some common benefits of cloud computing are:

• **Reduced Cost:** Since cloud technology is implemented incrementally (step-by-step), it saves organizations total expenditure.

• **Increased Storage:** When compared to private computer systems, huge amounts of data can be stored than usual.

• **Flexibility:** Compared to traditional computing methods, cloud computing allows an entire organizational segment or portion of it to be outsourced.

• **Greater mobility**: Accessing information, whenever and wherever needed unlike traditional systems (storing data in personal computers and accessing only when near it).

• **Shift of IT focus:** Organizations can focus on innovation (i.e., implementing new products strategies in organization) rather than worrying about maintenance issues such as software updates or computing issues. These benefits of cloud computing draw lot of attention from Information and Technology Community (ITC). A survey by ITC in the year 2008, 2009 shows that many companies and individuals are noticing that CC is proving to be helpful when compared to traditional computing methods.

Figure 2. Benefits of Cloud Computing

KEY CHARACTERISTICS

- Cost is greatly reduced and capital expenditure is converted to operational expenditure. This lowers barrier to entry, as infrastructure is typically provided by a third-party and does not need to be purchased for one-time or infrequent intensive computing tasks. Pricing on a utility computing basis is fine-grained with usage-based options and minimal or no IT skills are required for implementation.
- Device and location independence enable users to access systems using a web browser regardless of their location or what device they are using, e.g., PC, mobile. As infrastructure is off-site (typically provided by a third-party) and accessed via the Internet the users can connect from anywhere.
- Multi-tenancy enables sharing of resources and costs among a large pool of users, allowing for:
- Centralization of infrastructure in areas with lower costs (such as real estate, electricity, etc.).
- Peak-load capacity increases (users need not engineer for highest possible load-levels).
- Utilization and efficiency improvements for systems that are often only 10-20% utilized.

- Reliability improves through the use of multiple redundant sites, which makes it suitable for business continuity and disaster recovery. Nonetheless, most major cloud computing services have suffered outages and IT and business managers are able to do little when they are affected.

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
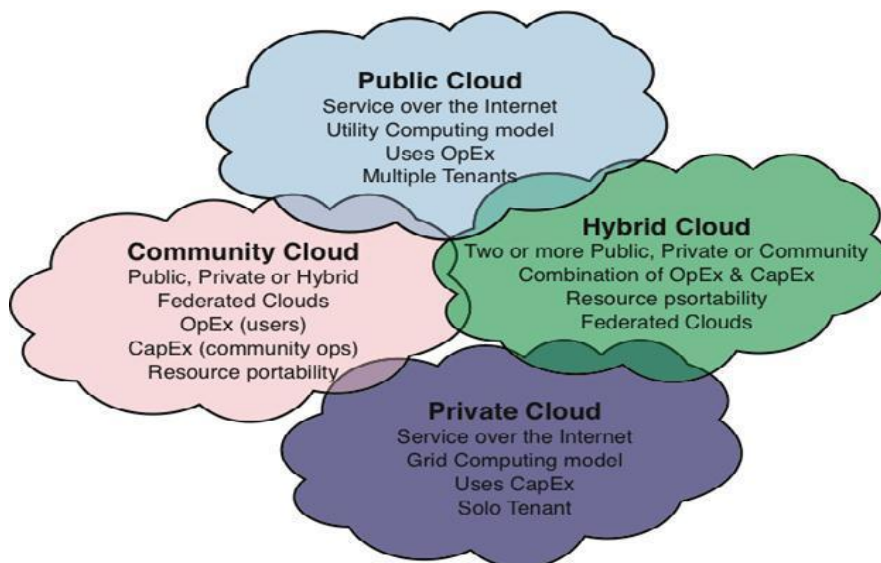

Figure 3. Types of Cloud

SERVICES OF CLOUD MODEL

Cloud computing can be accessed through a set of services models. These services are designed to exhibit certain characteristics and to satisfy the organizational requirements. From this, a best suited service can be selected and customized for an organization's use. Some of the common distinctions in cloud computing services are Software-as-a-Service (SaaS), Platform-as-a-Service (PaaS), Infrastructure as-a-Service (IaaS), Hardware-as-a-Service (HaaS) and Data storage-as-a-Service (DaaS). Service model details are as follows:

• **Software as a Service (SaaS)[4]**: The service provider in this context provides capability to use one or more applications running on a cloud infrastructure. These applications can be accessed from various thin client interfaces such as web browsers. A user for this service need not maintain, manage or control the underlying

cloud infrastructure (i.e. network, operating systems, storage etc.). Examples for SaaS cloud's are Salesforce, NetSuite.

• **Platform as a Service (PaaS)[5]**: The service provider in this context provides user resources to deploy onto cloud infrastructure, supported applications that are designed or acquired by user. A user using this service has control over deployed applications and application hosting environment, but has no control over infrastructure such as network, storage, servers, operating systems etc. Examples for PaaS cloud's are Google App Engine, Microsoft Azure, Heroku.

• **Infrastructure as a Service (IaaS)**: The consumer is provided with power to control process, manage storage, network and other fundamental computing resources which are helpful to manage arbitrary software and this can include operating system and applications. By using this kind of service, user has control over operating system, storage, deployed applications and possible limited control over selected networking components. Examples for IaaS clouds are Eucalyptus (The Eucalyptus Open source Cloud-computing System), Amazon EC2, Rackspace, Nimbus.
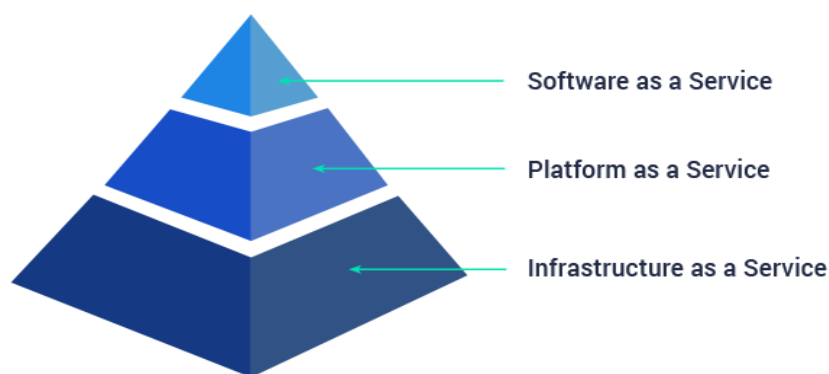
Figure 4. Models of Cloud

LOAD BALANCING

One of the foremost usually used applications of load balancing is to produce quality of service from multiple servers, typically called a server data center. Usually load-balanced systems are properly working inside popular internet sites, big chat networks, high-bandwidth file transfer protocol sites, and domain name System (DNS) servers. It additionally prevents the clients from contacting back-end servers directly, which can have security advantages by hiding the structure of the inner network. Some load balancers give a mechanism for improving the one parameter specially within back end server Load balancing offers the IT team an opportunity to attain a considerably higher fault tolerance. It will mechanically give the capability required to handle any increase or decrease of application traffic. It is additionally necessary that the load balancer itself doesn't become the cause of failure. Sometimes load balancers enforced in high-availability servers can additionally replicate the user's session needed by the application. Load balancing is dividing workload between a set of computers in order to receive the good response time and all the nodes are equally loaded and, in general, all users get served quicker. Load balancing may be enforced with hardware, software, or a mix of each. Typically, load balancing is that the main reason for server's unbalanced response time. Load balancing plans to optimize the usage of resources, maximize overall success ratio, minimize waiting time interval, and evade overloading of the resources. By the utilization of multiple algorithms and mechanisms with load balancing rather than one algorithm might increase reliability and efficiency. Load balancing within the cloud differs from classical thinking on load balancing design and implementation by misusage of data center servers to perform the requests on the basis of first come first serve basis. The older load balancing algorithm allocates the requests according to the incoming requests of the client. Load balancing is one amongst the central problems in cloud computing. It's a mechanism that distributes the dynamic workload equally across over the nodes or virtual machines within the whole cloud server to avoid a state of conflict

wherever some virtual machines are measured as heavily loaded whereas others nodes or hosts are measured as idle or doing very little work. It helps to realize a high client satisfaction and resource utilization magnitude relation, consequently increasing the performance and resource utility of the system. It additionally makes sure that each computing resource in the cloud server is distributed with efficiently and fairly among all the requests of the client. It additionally prevents bottlenecks of the system which can occur because of load imbalance.

GOALS OF LOAD BALANCING

The goals of load balancing are:

• To improve the performance of the system.

• To have a backup of the load or entire server just in case the system fails or even partly fails.

• To maintain the system stability

• To accommodate future modification within the system

LOAD BALANCING CLASSIFICATION

This is chiefly divided into 2 categories: static load balancing mechanism and dynamic load balancing mechanism:

1) Static approach: - This approach is especially outlined within the fixed style and is free from the workload at any point of time. Static load balancing algorithms divide the traffic equivalently between all servers.

2) Dynamic approach: - This approach is thought of solely through the current workload in the server by using different load balancing algorithms and selection mechanisms. Dynamic approach is additionally required in cloud computing as the number of requests are never predicted. Dynamic load balancing is divided in 2 varieties as non-distributed (centralized) Approach and distributed approach. It's outlined as following:

a) Centralized approach: - In centralized approach, solely one node is liable for managing and distribution among the complete cloud system model. Alternative all nodes aren't liable for handling the requests and providing the response.

b) Distributed approach: - In distributed approach, every node severally builds its own load vector. The work is divided among all the nodes of the server. They aggregate the load information of alternative nodes. Distributed approach is additional appropriate for complicated and very large systems inside the cloud computing.

METRICSFORLOADBALANCINGINCLOUD

Various metrics considered in existing load balancing techniques in cloud computing are discussed below-

·**Scalability** is the ability of an algorithm to perform load balancing for a system with any finite number of nodes. This metric should be improved.

·**Resource Utilization** is used to check the utilization of re-sources. It should be optimized for an efficient load balancing.

Performance is used to check the efficiency of the system. This has to be improved at a reasonable cost, e.g., reduce task response time while keeping acceptable delays.

·**Response Time** is the amount of time taken to respond by a particular load balancing algorithm in a distributed system. This parameter should be minimized.

·**Overhead Associated** determines the amount of overhead involved while implementing a load-balancing algorithm. It is composed of overhead due to movement of tasks, inter-processor and interprocess communication. This should be minimized so that a load balancing technique can work efficiently.

LOADBALANCINGALGORITHMS

Following load balancing algorithms are currently prevalent in clouds:-

Round Robin: In this algorithm [7], the processes are divided between all processors. Each process is assigned to the processor in a round robin order. The process allocation order is maintained locally independent of the allocations from remote processors. Though the work load distributions between processors are equal but the job processing time for different processes are not same. So at any point of time some nodes may be heavily loaded and others remain idle. Thi s algorithm is mostly used in web servers where http requests are of similar nature and distributed equally.

Connection Mechanism: Load balancing algorithm [8] can also be based on least connection mechanism which is a part of dynamic scheduling algorithm. It needs to count the number of connections for each server dynamically to estimate the load. The load balancer records the connection number of each server. The number of connection increases when a new connection is dispatched to it, and decreases the number when connection finishes or timeout happens.

Randomized: Randomized algorithm is of type static in nature. In this algorithm [7] a process can be handled by a particular node n with a probability p. The process allocation order is maintained for each processor independent of allocation from remote processor. This algorithm works well in case of processes are of equal loaded. However, problem arises when loads are of different computational complexities. Randomized algorithm does not maintain determini stic approach. It works well when Round Robin algorithm generates overhead for process queue.

Equally Spread Current Execution Algorithm: Equally spread current execution algorithm [9] process handle with priorities. it distribute the load randomly by checking the size and transfer the load to that virtual machine which is light ly loaded or handle that task easy and take less time , and give maximize throughput. It is spread spectrum technique in which the load balancer spread the load of the job in hand into multiple virtual machines.

Throttled Load Balancing Algorithm: Throttled algorithm [9] is completely based on virtual machine. In this client first requesting the load balancer to check the right virtual machine which access that load easily and perform the operations which is given by the client or user. In this algorithm the client first requests the load balancer to find a suitable Virtual Machine to perform the required operation.

A Task Scheduling Algorithm Based on Load Balancing: Y. Fang et al. [10] discussed a two-level task scheduling mechanism based on load balancing to meet dynamic requirements of users and obtain 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.

Min-Min Algorithm: It begins with a set of all unassigned tasks. First of all, minimum completion time for all tasks is found. Then among these minimum times the minimum value is selected which is the minimum time among all the tasks on any resources. Then according to that minimum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is

removed from the list of the tasks that are to be assigned to the machines. Then again the same procedure is followed until all the tasks are assigned on the resources. But this approach has a major drawback that it can lead to starvation [12].

Max-Min Algorithm: Max-Min is almost same as the min-min algorithm except the following: after finding out minimum execution times, the maximum value is selected which is the maximum time among all the tasks on any resources. Then according to that maximum time, the task is scheduled on the corresponding machine. Then the execution time for all other tasks is updated on that machine by adding the execution time of the assigned task to the execution times of other tasks on that machine and assigned task is removed from the list of the tasks that are to be assigned to the machines[12].

RELATED WORK

Nguyen Khac Chien et al. (2016) has proposed a load balancing algorithm which is used to enhance the performance of the cloud environment based on the method of estimating the end of service time. They have succeeded in enhancing the service time and response time of the user.

Ankit Kumar et al (2016) focuses on the load balancing algorithm which distributes the incoming jobs among VMs optimally in cloud data centers. The proposed algorithm in this research work has been implemented using Cloud Analyst simulator and the performance of the proposed algorithm is compared with the three algorithms which are preexists on the basis of response time. In the cloud computing milieu, the cloud data centers and the users of the cloud-computing are globally situated, therefore it is a big challenge for cloud data centers to efficiently handle the requests which are coming from millions of users and service them in an efficient manner.

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.

Hiren H. Bhatt et al. (2015) presents a Flexible load sharing algorithm (FLS) which introduce the third function. The third function makes partition the system in to domain. This function is helpful for the selection of other nodes which are present in the same domain. By applying the flexible load sharing to the particular domains in to the distribute system, the performance can be improved when any node is in overloaded situation.

CLOUD SIM

Cloud service providers charge users depending upon the space or service provided. In R&D, it is not always possible to have the actual cloud infrastructure for performing experiments. For any research scholar, academician or scientist, it is not feasible to hire cloud services every time and then execute their algorithms or implementations. For the purpose of research, development and testing, open source libraries are available, which give the feel of cloud services. Nowadays, in the research market, cloud simulators are widely used by research scholars and practitioners, without the need to pay any amount to a cloud service provider.

Tasks performed by cloud simulators :

The following tasks can be performed with the help of cloud simulators:

• Modelling and simulation of large scale cloud computing data centres.

• Modelling and simulation of virtualised server hosts, with customisable policies for provisioning host resources to VMs.

• Modelling and simulation of energy-aware computational resources.

• Modelling and simulation of data centre [18] network topologies and message-passing applications.

• Modelling and simulation of federated clouds.

• Dynamic insertion of simulation elements, stopping and resuming simulation.

• User-defined policies for allocation of hosts to VMs, and policies for allotting host resources to VMs.

The scope and features of cloud simulations include:

• Data centres

• Load balancing

• Creation and execution of cloudlets

• Resource provisioning

• Scheduling of tasks

• Storage and cost factors

CONCLUSION

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 approaches. 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. 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. Our paper focuses on the various load balancing algorithms and their applicability in cloud computing environment.

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 Science and 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 Multi Conference 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.

[21] Ankit Kumar, Mala Kalra," Load Balancing in Cloud Data Center Using Modified Active Monitoring Load Balancer", IEEE pp. 1-5, 2016.

[22] Saraswathi AT, Kalaashri.Y.RA, Dr.S. Padmavathi, "Dynamic Resource Allocation Scheme in Cloud Computing", ELSEVIER, pp. 30-36, 2015