Volume 6, Issue 4, April 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Load Balancing Techniques in Cloud Computing Environment– A Review

[1]Shipra Goyal (M.Tech Scholar), [2]Manoj K Verma (Assistant Professor)
Department of Computer Science & Engineering, Seth Jai Parkash Mukand Lal Institute of Technology, Radaur (Yamunanagar) Haryana, India

Abstract: ***Cloud computing is said to be the next big boom technology in IT industry infrastructure. It is claimed that it provides new levels of efficiency, flexibility and cost savings of the resources that are used in industries. Cloud computing is an evolving technology which delivers infrastructure, platform and software that are made available as services in a pay-as-you-go model to consumers. The cloud owner's relationship with the consumer most depends on how they are able to use the cloud resources efficiently, which in turn depends upon the effective management. The major problems that we faced in the cloud are resource discovery, fault tolerance, load balancing and security. Load balancing is the primary discussion in the cloud-computing environment which is used to request resource services. Its main motive is to optimize the usage of resources, boost turnout, reduce response time and avoid the needless burden of any of such resources. It becomes a severe problem with the increase in list of users and types of applications on cloud. The main highlights of this paper is on the load balancing approach & techniques in cloud computing. In this paper, there is a detailed survey on different load balancing techniques which are existing in cloud analyst tool and some policies by different authors.***

Keywords: ***Cloud computing, Virtualization, Load Balancing Approaches, Cloud Analyst, Load Balancing Algorithms.***

I. INTRODUCTION

A Cloud refers to a distinct IT environment that is designed for the purpose of remotely provisioning scalable and measured IT resources [1]. As defined by NIST "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications and services) that can be rapidly provisioned and released with minimal management efforts or service provider interactions" [2]. Many applications are based on cloud are Drop box, Skype, Google Drive, ZOHO and many more applications are there which are providing cloud services to users pay per use basis and even free also. Cloud computing is an emerging computing technology that uses the Internet and centralised remote servers to maintain data and application [3]. In this way, users of cloud such as software developers can use virtualized resources as a service, often flexibly scaling resource usage (and payment) up or down. Cloud computing has many advantages which makes any industry to move from conventional infrastructure. Figure 1 show why we should move to cloud computing and adopt cloud computing infrastructure.

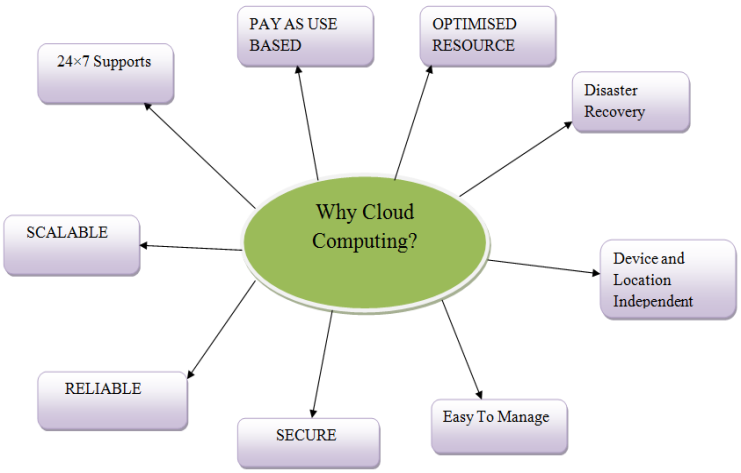

Figure1. Advantages of Cloud Computing

1.1 Cloud Service Models: The three main services provided by the cloud are IaaS (Infrastructure as a Service), PaaS (Platform as a Service) and SaaS (Software as a Service). The basic and a short description of these three services are as follows:

IaaS: Infrastructure as a Service (IaaS) is the delivery of computer hardware (servers, networking technology, and storage and data centre space) as a service. It may also include the delivery of various operating systems and virtualization technologies to manage the resources. The IaaS customers rent computing resources instead of buying and installing them in their own data centre. The service is typically paid for on a usage basis.

PaaS: Platform as a Service (PaaS) is a category of cloud computing services that provide a platform allowing customers to develop, run, and manage applications without the complexity of building and maintaining the infrastructure typically associated with developing and launching an application. PaaS can be delivered in two ways:

1) As a public cloud service from a provider, where the consumer controls software deployment and configuration settings, and the provider provides the networks, servers, storage and other services to host the consumer's applications; and,

2) As software installed in private data centres or public infrastructure as a service and managed by internal IT departments.

SaaS: Software as a Service (SaaS) is a software licensing and delivery model in which software is licensed on a subscription basis and is centrally hosted. Consumers purchase the ability to access and use an application or service that is hosted in the cloud. A benchmark example of this is Salesforce.com, where necessary information for the interaction between the consumer and the service is hosted as part of the service in the cloud. Vendors like Microsoft is expanding its involvement in this area and as part of the cloud computing option for Microsoft® Office 2010, its Web Apps are available to Office in the cloud as it is to have it on the premise. In this context, CaaS (Communication as a Service) could be seen as a subset of SaaS volume licensing customers and Office Web App subscriptions through its cloud-based Online Services. The diagrammatic view of the services is:

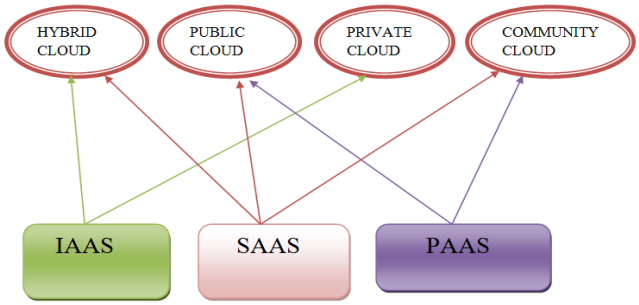

Figure2. Cloud Services Models

The different cloud service models discussed above can be deployed through virtualization technology by making the perception to users that virtually there are one or more existing entities but actually it is the only one. Using virtualization, one machine can be made work like many, desktop computer running multiple operating systems simultaneously with a vast amount of disk space or drives available in the sharable or independent modes. The most common forms of virtualization include server virtualization, desktop virtualization, virtual networks and virtual storage etc. With the help of virtualization, service models can be deployed easily. Virtualization is the key technology behind cloud computing that allows the simultaneous execution of diverse tasks over a shared hardware platform [4].The use of virtualization brings up the involvement of various significant issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management etc. One of the targeted issues along with its implementation techniques is discussed in the further sections of this paper. In the second section, we have discussed the significance of virtualization in context of load balancing along with its key approaches. In third section, we have focused on various supportive techniques of load balancing in cloud computing. In further two sections of this paper, we have reviewed the in context related work by various authors and concluded the gap analysis.

II. SIGNIFICANCE OF VIRTUALIZATION IN CONTEXT TO LOAD BALANCING

The basic definition of Virtualization and load balancing are as follows:

Virtualization: In cloud computing, virtualization is very useful concept which means something which is not real and to create a virtual version of resource, such as a server, storage device, network or even an operating system. It is the software implementation of a computer which will execute different programs like a real machine. Even, something as simple as partitioning a hard drive is considered virtualization because a drive can be partitioned into more than one.

Load Balancing in Cloud Computing

Load balancing is one of the critical aspects in cloud computing environment that can significantly improve resource utilization, performance and save energy by properly assigning/reassigning computing resources to the incoming requests from users [6]. As technology is growing faster, there are huge amount of users on internet so, to manage and fulfil their requirements, load balancing comes into the picture which ensure that workload is spread equally to all of the available servers without any delay to accomplish ***higher user satisfaction*** and ***maximum throughput with minimum response time*** [7]. Load Balancing is classified in two key approaches based on decisions making process: ***Static*** and ***dynamic*** load balancing algorithms.

Static Load Balancing Algorithms: Static algorithms are much simple as compared to the dynamic algorithms. It must require knowledge of global status of distributed system and does not consider the current state or behaviour of a node

while allocating the load to the available nodes. It divides the traffic equivalently among all available servers or VMs. It is used when the computational and communication requirements of a problem are known a priori. In this case, the problem is partitioned into tasks and the assignment of the task-processor is performed once before the parallel application initiates its execution.

Dynamic Load Balancing Algorithms: Dynamic load balancing is more flexible than the static and they doesn't rely on prior knowledge but depends on current state of the system. In a distributed system, dynamic load balancing has two different ways: ***distributed*** and ***non-distributed***. ***In the distributed one,*** algorithm is executed by all nodes present in the system and the task of load balancing is shared among these servers. The interaction among nodes to achieve load balancing can take two forms: ***cooperative*** and ***non-cooperative. In the first one,*** the nodes works side-by-side to achieve a common objective which means is to improve the overall response time, etc. ***In the second form,*** each node works independently toward a goal local to it [8].

In non-distributed type, either one node or more than one node perform the task of load balancing. Dynamic load balancing algorithms of non-distributed nature can get two forms: 1) ***Centralized*** and 2) ***Semi Distributed***. In the centralized, the load balancing algorithm is executed just by a single node in the total system which is the central node. This node is exclusively in charge for load balancing of the whole system. The other nodes interact merely with the central node. However, in semi-distributed form, nodes are partitioned into clusters, where the load balancing in every cluster is of centralized form. A central node is chosen in each cluster by suitable election technique which takes care of load balancing inside that cluster. Hence, the load balancing of the complete system is done via the central nodes of each cluster. Centralized dynamic load balancing takes less messages to arrive at a decision, since the number of overall interactions in the system decreases drastically as compared to the semi-distributed case. However, centralized algorithms can create a bottleneck in the system at the central node and also the load balancing process is rendered hopeless once the central node crashes. Therefore, this algorithm is mainly suited for networks with small size [17].

The load balancing is an efficient and critical concept in cloud computing and it helps in utilizing the resources optimally, therefore minimizing the consumption of resources. Thus load needs to be distributed over the nodes in cloud-based architecture, so that each resource does the equal amount of work at any point of time that is performed by a load balancer. The load balancer does request allocation to different servers by using various scheduling algorithms [8].

Earlier, the load balancing services were presented to the end-user a single endpoint which is the application through which they communicate with the end point. After that the load balancing service communicates with a pool of resources that has one or more application instances. In the traditional load balancing environment, each application instance is hosted on a single, physical server.

In Traditional approaches, Load balancing services gave a back node option to each node associated in the service providing environment. Each node in a pool may have a back up node that only work in the event of a failure and this feature increase the high availability purposes to ensure continuous application availability rather than for scaling purposes.

When physical servers are replaced with virtual servers, there is not much difference in the system. It works all that of traditional system. It provides same services to end users. However, there are some new potential sources of failure that must be addressed that impact the topology – the physical placement – of the application instances in the pool.

The most important change that must be seen is fault isolation. When virtualized environment is deployed every node has its own virtual network connection and if they have their own shared physical network connection and it fails, then all nodes will fail – leaving "the application" unavailable.

III. EXISTING LOAD BALANCING TECHNIQUES IN CLOUD COMPUTING

There are various load balancing techniques which are already implemented in the cloud computing using various tools like Cloud Analyst & Cloud-Sim. A small description of the implemented techniques is given as under:

Round Robin: Round robin is one of the straightforward and static scheduling technique that utilize the principle of time slices which divided time into multiple interval and each VM Is given a particular time slice or time interval [9], [10].Round robin works on arbitrary selection of the VMs. It assigns requests to a list of existing VMs on a rotational basis. The first request is assigned to a VM selected arbitrarily from the group and then the Data Centre controller allots the requests in a circular order. When the VM is assigned the request, the VM is progressed to the end of the list [11].

Throttled Algorithm: Here, the load balancer maintains an index table of virtual machines as well as their states (Available or Busy). The client/server first makes a request to data centre to find a suitable virtual machine (VM) to perform the recommended job. The data centre queries the load balancer for allocation of the VM. The load balancer scans the index table from top until the first available VM is found or the index table is scanned fully. If the VM is found, the load data centre. The data centre communicates the request to the VM identified by the id. Further, the data centre acknowledges the load balancer of the new allocation and the data centre revises the index table accordingly. While processing the request of client, if appropriate VM is not found, the load balancer returns -1 to the data centre. The data centre queues the request with it. When the VM completes the allocated task, a request is acknowledged to data centre, which is further apprised to load balancer to de- allocate the same VM whose id is already communicated [12].

Equally Spread Current Execution (ESCE): In this algorithm, load balancer makes an effort equally spreading the execution load on different VMs. Load balancer maintains an index table of VMs along with the number of requests currently allocated to the VM. If there is request comes from the data centre for execution, load balancer search the index table for least loaded VM. If more than one VM is found, first identified VM is selected and allotted for request execution. The load balancer updates the index table by increasing the allocation count of identified VM. When VM

finishes the execution of allotted request, load balancer again update the index table by decreasing the allocation count for identified VM by one. So, there is an additional computation load balancer to search the table again and again [13].

Active Monitoring Load Balancing (AMLB) Algorithm: It maintains information about each VM and the number of requests currently allocated to which VM. When a request to allocate a new VM arrives, it identifies the least loaded VM. If there are more than one, the first identified is selected. Load Balancer returns the VM id to the Data Centre Controller. It sends the request to the VM identified by that id and notifies the Active VM Load Balancer of the new allocation. During allocation of VM only importance is given on the current load of VM, its processing power is not taken into consideration. So the waiting time of some jobs may increase violating the QOS requirement. [16]

IV. RELATED WORK

Ajit M. (2013) proposed a new VM load balancing algorithm named as ***Weighted Signature based Load Balancing (WSLB)*** and implemented the proposed work using cloud analyst tool with the help of Cloud-Sim library using java language. This result into the identification of load assignment factor for each of the host in a datacentre and maps the VMs accordingly. Load balancer sends virtual machine id which is available on highest configuration host having maximum load assignment factor then lowest one and so on. According to the experimental results, it conclude that if we select a virtual machine mapped on powerful host, then it affects the overall performance of the cloud Environment and decrease the average response time. Proposed method works out on homogeneous VMs mapped on hosts [13].

Kulkarni A.K. **(2015)** proposed a variant of active VM algorithm to solve the issue during peak hours by using a Reservation table. The proposed VM load balancer maintains an internal reservation table to maintain the information of VM reservations suggested by the load balancer to data centre controller but not updated in allocation table until the notification arrives of allocation. The proposed load balancer takes into consideration both reservations table entry and allocation statistics table entry for particular VM id by the load balancer for VM selection for next request. In this paper, an efficient VM load balancing algorithm is proposed that distributes the load evenly across all VMs in the data centre even when the incoming request frequency is high during peak hours. It is observed from the experimental results that current active VM load balancer heavily loads the initial VMs where-as the proposed VM Load balancer evenly distributes the incoming requests to all VMs [14].

Y. Lua et al. (2011) proposed a **Join-Idle-Queue load balancing algorithm** for dynamically scalable web services. This algorithm provides large scale load balancing with distributed dispatchers by, first load balancing idle processors across dispatchers and then, assigning jobs to processors to reduce average queue length at each processor. By removing the load balancing work from the critical path of request processing, it effectively reduces the system load, incurs no communication overhead at job arrivals and does not increase actual response time [15].

Singh A. et al. (2015) proposed an algorithm for dynamic load balancing. The algorithm uses three agents; Load agent, Channel agent and Migration agent demonstrating its load balancing. The load and channel agents are static while Migration agents represent an ant. The load agent controls information policy and maintained all details of data centre and at the same time responsible for calculating the load on every available virtual machine (VM) when new task are allocated in a data centre. The load agent is supported with a VM Load Fitness table. The fitness table maintained the list of all details of virtual machines properties in a data centre such as id, memory, and fitness value and load status of all virtual machines. When the load agent completed the controlling policy, the channel agent controls the transfer policy, selection policy and location policy. Upon request received from the load agents, the channel agent initiates a communication with the migration agents. The migration agent then moved to other data centres and communicates with load agent of the data centre to enquire about the status of VMs presents. Upon receiving the status, it communicates to its parent channel agent. This approach reduces service time and overcome the challenge of overloaded virtual machines [5].

James J. et al. (2012) proposed modified version of the Active Monitoring Algorithm named ***Weighted Active Monitoring Algorithm*** by assigning weight to each node in order to achieve better response time and processing time. In this proposed Load balancing algorithm using the concept of weights in active monitoring, the VM are assigned varying (different) amount of the available processing power of server/ physical host to the individual application services. To these VMs of different processing powers; the tasks/requests (application services) are assigned or allocated to the most powerful VM and then to the lowest and so on according to its weight and its availability. Hence optimizing the given performance parameters [18].

Shobhana G. et al. (2014) proposed a ***Central Load Balancer*** that balances load among virtual machines in cloud data centre. Two key factors were adopted for load balancing; "Data-Centre Controller" and the "Central Load Balancer". The states of the virtual machine are maintained as "BUSY" and "AVAILABLE". In this technique, every request from user bases arrived at "Data-Centre Controller". The "Data Centre Controller" queries the Central Load Balancer for allocation of requests. The Central Load Balancer maintains a table that consist of id, states and priority of all virtual machines. It parses the table and find out highest priority virtual machine, then check its states. If the virtual machine state is "AVAILABLE", then the id of the VM (VMid) is returned to the "Data-Centre Controller". If the state of virtual machine is "BUSY" it chooses next less high priority virtual machine. Finally, "Data-Centre Controller" assigns the request to that VMid that is provided by Central Load Balancer (CLB). The Central Load Balancer (CLB) is connected to all users and virtual machines present in cloud datacentre through the "Data-Centre Controller". The Central Load Balancer calculates the priorities of virtual machines based on CPU speed (MIPS) and memory. The algorithm was implemented using Cloud Analyst and the results shown significant improvement when compared to the existing load balancing techniques (RR, ESCE, TLA) in term of response time. [19]

V. CONCLUSION AND FUTURE SCOPE

In this paper, we have surveyed multiple load balancing algorithms in which some are static and others are dynamic. Cloud Computing has widely been adopted by the industry, though there are many existing issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management etc. which have not been fully addressed. Central to these issues is the issue of load balancing, that is required to distribute the excess dynamic local workload evenly to all the nodes in the whole cloud to achieve a high user satisfaction and resource utilization ratio. It also ensures that every computing resource is distributed efficiently and fairly. This paper presents a concept of cloud computing along with research challenges in load balancing. Cloud computing is a vast concept and load balancing plays a very important role in case of Clouds. In next level, we are going to compare all the load balancing algorithms which are existing in Cloud analyst tool and propose a new improved algorithm which will gives better results in terms of response time and will reduce cost.

REFERENCES

[1] Dasgupta K., Mandal B. (2013) .*A Genetic Algorithm (GA) based Load Balancing Strategy for Cloud Computing*. International Conference on Computational Intelligence: Modelling Techniques and Applications: CIMTA 2013,pp.340-347, Wiley Press ,USA, ISSN: 0038-0644, DOI:10.1016/j.protcy.2013.12.369

[2] Mell P., Grance T. (2011 Sep.). *The NIST Definition of Cloud Computing*. National Institute of standards: NIST 2011 ,pp. 1-7

[3] Dobale R.G.,SonarR.P. (2015 February) .*Review of Load Balancing for Distributed Systems in Cloud*. International Journal of Advanced Research in Computer Science and Software Engineering: IJARCSSE, 2015 pp.393-403, ISSN: 2277 128X

[4] Gkatzikis L., Koutsopoulos I. (2013) .*Migrate or Not? Exploiting Dynamic Task Migration in Mobile Cloud Computing Systems*. IEEE Wireless Communications, pp. 24-32, ISSN: 1536-1284, DOI: 10.1109/MWC.2013.6549280

[5] Singh A, Juneja D. & Malhotra M. (2015). *Autonomous Agent Based Load Balancing Algorithm in Cloud Computing*. International Conference on Advanced Computing Technologies and Applications: ICACTA 2015, pp. 832–841, Moscow ISSN: 18770509, DOI: 10.1016/j.procs.2015.03.168

[6] KalaiSelvi B. Mary L. (2014, August). *A Survey of Load Balancing Algorithms using VM.* , International Journal of Advancements in Research & Technology: IJOART 2014, pp 68-76, ISSN: 2278-7763

[7] Kaur R. And Luthra P. (2012) .*Load Balancing in Cloud Computing*. Association of Computer Electronics and Electrical Engineers: ACEE 2014, pp. 375-381, ISSN: 1899-0142, DOI:02.ITC.2014.5.92

[8] Panwar R., Mallick B. (2015, May) .*A Comparative Study of Load Balancing Algorithms in Cloud Computing.* , International Journal of Computer Applications: IJCA 2015, pp.33-37, ISSN: 0975 – 8887, DOI: 24, May 2015

[9] Pasha.N, Agrawal A. & Rastogi R. (2014, May).*Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment.* , International Journal of Advanced Research in Computer Science and Software Engineering: IJARCSSE 2014, pp: 34-39, ISSN: 2277 128X, DOI: 5,May 2014.

[10] Shah D.M., Kariyani A.A & Agarwal D.L .(2013, February).*Allocation of Virtual Machines in Cloud Computing Using Load Balancing Algorithm*, International Journal of Computer Science and Information Technology & Security: IJCSITS 2013, pp. 93-95, ISSN: 2249-9555 DOI: 2013 Feb

[11] Domanal S. G., Reddy G. R. M. (2013, October).*Load balancing in Cloud Computing using Modified Throttled Algorithm. In Cloud Computing in Emerging Markets.* Cloud Computing in Emerging Markets: CCEM 2013, pp.1-5, Bangalore, India, ISBN: 978-1-4799-0027-5, DOI: 10.1109/CCEM.2013.6684434

[12] Sharma T., Banga V.K. (2013, March**).***Efficient and Enhanced Algorithm in Cloud Computing.* International Journal of Soft Computing and Engineering :IJSCE 2013 pp.385-390, ISSN: 2231-2307, DOI: 2013 March

[13] Ajit M., Vidya G. (2013, July).*VM Level Load Balancing in Cloud Environment*. Computing, Communications and Networking Technologies: ICCCNT 2013, pp.1-5,Tiruchengode, India, ISBN NO: 978-1-4799-3925-1, DOI: 10.1109/ICCCNT.2013.6726705

[14] Kulkarni A.K., Annappa B. (2015). *Load Balancing Strategy for Optimal Peak Hour Performance in Cloud Data centres.*, Signal Processing, Informatics, Communication and Energy Systems: SPICES 2015 IEEE International Conference, pp.1-5,Kozhikode ,ISBN NO.978-1-4799-1823-2/15, DOI: 10.1109/SPICES.2015.7091496

[15] Lua Y. , Xiea Q., Kliotb G, Gellerb A. , Larusb J.R. & Greenber A. (2011, August). *Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services.* , An international Journal on Performance evaluation: IJPE 2011,pp.1056-1071, Amsterdam, The Netherlands, ISSN:0166-5316, DOI:10.1016/j.peva.2011.07.015

[16] Mahalle H.M., Kaveri P.R., Chavan V. (2013, January) .*Load balancing On Cloud Data Centres.* International Journal of Advanced Research in Computer Science and Software Engineering: IJARCSSE 2013, India, pp.1-4, ISSN:2277 128X, DOI: Jan 2013

[17] Haryani N., Jagli D. (2014, July-August).*Dynamic Method for Load Balancing in Cloud Computing.* IOSR Journal of Computer Engineering :IOSR-JCE 2014, pp. 23-28, ISSN: 2278-8727, DOI:Aug,2014

[18] James J., Verma B. (2012, September).*Efficient VM Load Balancing Algorithm for a Cloud Computing Environment.*, International Journal on Computer Science and Engineering: IJCSE 2014, pp.1658-1663 ISSN:0975-3397,DOI: Sep 2012

[19] Shobana, G., Geetha, M. and Suganthe, R.C. (2014, February). *Nature Inspired Pre-emptive Task Scheduling for Load Balancing in Cloud Data centre*. International Conference on Information Communication and Embedded Systems: ICICES 2014 pp.1-6, Chennai ,India, ISBN: 978-1-4799-3835-3,DOI: 10.1109/ICICES.2014.7033816

