International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2019 IJSRCSEIT | Volume 5 | Issue 2 | ISSN : 2456-3307
DOI : https://doi.org/10.32628/CSEIT1952242

A Survey on Load Balancing Algorithms

Ruchi[1], Harish Kumar[2]

[1]M. Tech, Research Scholar, JC Bose University of Science and Technology, YMCA, Faridabad, India

[2]Assistant Professor, Department of Computer Engineering, JC Bose University of Science and Technology, YMCA, Faridabad, India

ABSTRACT

Cloud computing is referred to as biggest technology of today's environment that provide access to distributed resources on the basis of pay-per-use. Everyone try to use cloud to reduce the cost and maintenance of infrastructure due to which lots of load is increasing day by day. Therefore, there is need to balance that load since resources of cloud are limited but usage is increasing at every moment. This paper discuss how the resources are allocated and how the tasks are scheduled among those resources. Task scheduling mainly focuses on enhancing the utilization of resources and hence reduction in response time. There are various static and dynamic load balancing algorithms to balance the load, this paper discusses comparative study of these algorithms.

Keywords : Cloud Computing, Virtualization, Load Balancing, Resource Allocation, Task Scheduling, Static and Dynamic Load Balancing Algorithm.

I. INTRODUCTION

An internet-based computing called Cloud computing [1] consist of two words "CLOUD" which means involvement of huge infrastructure (i.e. it is collection of storage resources, server hardware and server software) which is required to build cloud applications that can be obtained via cloud on metered basis. In order to make efficient use of these resources and to make sure their availability to end user's , "COMPUTING" is performed based on terms and conditions discussed in service level agreement. On the basis of user's demand infrastructure is made available to end user's.

For example: If we want an electricity connection at your home, you have to just subscribe to that utility and service provider will provide you a meter which will charge you according to usage of electricity, same is the concept of cloud computing like Google's Gmail or Google Docs.

Cloud is having different meaning for different collaborators. There are three main collaborators:

A. End Users*:* [1]

These are consumer of cloud and uses different services (Infrastructure/Software/Platform) obtained by the cloud according to SLA (Service Level Agreement) discussed by the cloud provider.

Before using cloud services, user of cloud first make sure SLA (contains all Quality-of-Service (QOS) parameters which are required by consumer. Cloud providers provide services with the concept of utility computing in which cloud provider owns, operates and manages the computing infrastructure and resources, and the subscriber accesses it as and when

required on a metered basis. Security, privacy, availability, reduced cost, Ease-of-use are various issues and requirements for cloud users.

B. Cloud Providers:

To serve services according to Service Level Agreement to the end users, cloud providers are responsible. They provide three deployment models of cloud which includes (public cloud/private cloud/hybrid cloud).

- Private cloud [2] is under the control of single organization provided by service provider which can be either external or internal service provider, they offer greatest level of security, it is used by business for their internal use and data center is responsibility of organization. Most important part of any organization is its resources and data which can be made more secure with the help of private cloud. It is also called an internal cloud that will reside in company environment and will be accessed only by the members of the company. The major drawback is its higher cost. Example- OpenStack [3]. VMware [4], Cloud Stack [5] etc.
- Public cloud [2] is under the control of general public and provided by service provider which is external, major security issue of using public cloud is confidentiality, any user can make use of resources, and datacenter is responsibility of cloud service provider. It is less secure as it is open to everyone. Cloud services like IaaS, PaaS, and SaaS follow the public cloud so it is more flexible. It is location independent and cost effective. Example- largest public cloud provider is Amazon web services [18].
- Hybrid cloud [2] it is the combination of two cloud (Public Cloud and Private Cloud). Organization are allowed here to manage resources internally as well as externally. It combines the security benefits of private cloud as well as scalability and cost-effectiveness benefit of public cloud.

The main task of cloud provider is divided into two task:

- To manage huge resources of cloud.
- To make resources available to the end users.

Management of resources, Utilization of resources, metering [1], cost efficiency [1], utility computing are various issues and requirements for cloud users.

C. Cloud Developer: [1]

Cloud developer is having responsibility of both end users and cloud providers. Cloud developer's main aim is to overcome the space between cloud provider and end user. Cloud developer should know all the technical details which are important for both end user and cloud provider. Elasticity/Scalability, virtualization, reliability, programmability are various issues and requirements for cloud developer.

Virtualization [6]

Virtualization is like "something which is not real" but all the functions that belongs to real environment can be observe. In grid and utility computing the concept of virtualization was missing. Cloud computing introduces the concept which is energy saving, cost saving and hardware reducing technique used by cloud providers called "Virtualization". It is the process of creating an illusion of something like operating system, computer hardware, storage devices or resources of computer network. This allows a machine, server or a computer to split into multiple virtual machines and virtual servers. Concept of virtualization came from the concept of Multi Programming (i.e. each process thinks it has complete control on all the resources). The only difference is CPU is shared by processes in multi programming and in virtualization it is shared by multiple operating system. Virtualization make it

possible to run more than one operating system and application on the single server at same time period.

Hypervisor, which is a software, is responsible for distributing hardware appropriately, by connecting to the hardware and splitting it into virtual machines. Each virtual machine will feel like an isolated computer with its own storage, own memory etc. Hypervisor runs on the bare hardware and does multi-programming. Below are the scenario in cloud computing before virtualization and after virtualization:

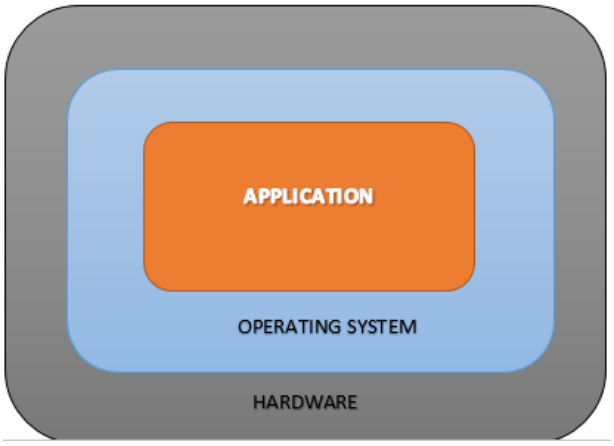

Figure 1 : Before Virtualization (Physical Host)

Before virtualization is applied, hardware resides in physical layer, which includes physical devices in datacenters providing a foundation for model. Then there is an operating system that act as an interface between hardware and application.

With the help of virtualization hypervisor is introduced which operates on data link layer and increases the utilization of resources and throughput by making the multiple instances of virtual machine. Hypervisor will take care of resource distribution among instances of VM. Now large number of tasks will be completed in the same time in which single task was completed in case of physical host. Now each instance of virtual machine will be having its own hardware, operating layer, and application which will be managed by hypervisor.

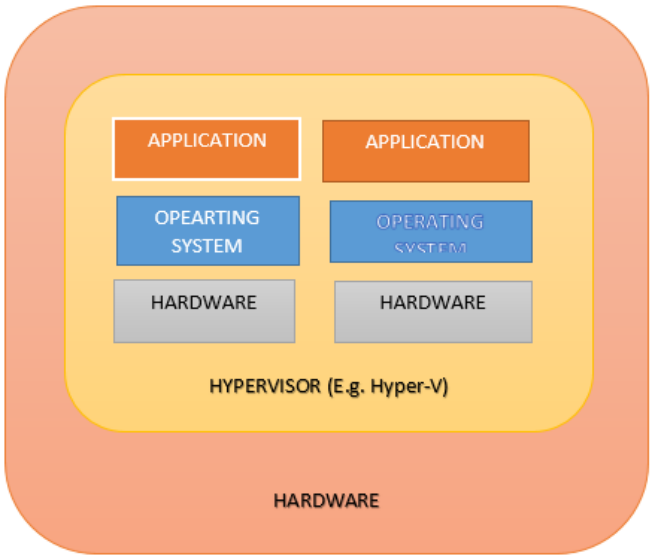

Figure 2: After Virtualization (Virtualized Host)

II. RELATED WORK

Depending upon present system state balancing of load can be of two types :

Static Algorithm

Static algorithm is best suited for homogeneous and stable environment where nature of task is similar. It require prior knowledge about the nodes like bandwidth, memory, processing power before assigning load. Its limitation is tasks are assigned to the processor only after it is created and in case of sudden failure of system resources, task cannot be shifted during its execution for load balancing. Algorithms which are static are:

- Round Robin Load Balancing Algorithm (RR) [9]
- Weighted Round Robin Load Balancing Algorithm (WRR)[11]
- Load Balancing Min-Min Algorithm[11,12]
- Load Balancing Max-Min Algorithm[11,12]
- Priority Based Modified Min-Min Load Balancing Algorithm[13]

A. Round Robin Load Balancing Algorithm (RR) [9]

The concept of time interval is used by datacenter controller, at the time of allocating jobs it does not take into account the resource capability of a VM. It

uses round robin scheduling algorithm [10] for allocating jobs. Nodes are selected randomly and jobs are allocated to all nodes in a round robin manner. There is non-uniform distribution of workload. Higher priority and lengthy tasks will end up with higher response time. This algorithm is not be suitable for cloud computing environment because running time of any process will not be known prior to execution, so there will be possibility for a node to be heavily loaded or under loaded.

B. Weighted Round Robin Load Balancing Algorithm [11]

To overcome limitation of Round Robin Algorithm each node will be assigned a specific weight and according to that weight, number of request assigned to a node will be decided. Node having large value of weight indicates larger capacity to handle the load and large number of tasks will be given to that node as compared to node having small value of weight. Each node will receive same traffic if the weights assigned to all nodes are equal.

C. Load Balancing Min-Min Algorithm [11, 12]

It starts with set of all jobs that are unassigned. Initially minimum completion time is calculated for all available nodes then the task having minimum completion time (MCT) [11] will be chosen and assigned to respective node. Updated ready time of the node is assigned. This process is repeated until all jobs that are not assigned are assigned. Drawback is some jobs may experience starvation.

D. Load Balancing Max-Min Algorithm [11, 12]

It is similar to Min-Min Algorithm with the difference that after calculating minimum completion time of jobs, the maximum value job is selected and is mapped to machine with minimum completion time for all the jobs. Then the ready time of the node is updated by adding the execution time of the assigned task. Disadvantage of this algorithm is starvation where the tasks having the maximum completion time will be executed first while leaving behind the task having minimum completion time.

E. Priority Based Modified Min-Min Load Balancing Algorithm [13]

It is the unique modification of Min-Min Algorithm. It is based on two steps. First Min-Min strategy is applied and then tasks are re-scheduled to use unutilized resource effectively and improve overall make span along with more priority user who pay more for cloud service.

Dynamic Algorithm

Dynamic algorithm is best suited for heterogeneous environment where nature of task is different.

No prior knowledge of resources are required. Task can be redistributed to any processor at any time. Problem of under loaded and overloaded of any resource does not exist here. Present state and capability of virtual machine is always kept in mind. Some of the dynamic algorithms are:

- Throttled Load Balancing Algorithm[14]
- Priority Based Modified Throttled Load Balancing Algorithm[14]
- Equally Spread Current Execution Load Balancing Algorithm[15]
- Ant Colony Load Balancing Algorithm[16,17]

A. Throttled Load Balancing Algorithm [14]

It is appropriate for Virtual Machine. List of virtual machine is created along with the status of VM. It take care of current load on VM and allocate predefined number of tasks to single VM at any given instant of time. If more request were there then some of the request would have to wait in queue. Client make request to data center, which in turn queries to the load balancer to allocate appropriate VM.

An index table is maintained which is having information of VM id and virtual machine status either free or busy, load balancer will traverse the table and if any VM is, empty load data center will communicate with that VM identified by its id returned by load balancer. If VM not found, load balancer will return -1 to data center.

B. Priority Based Modified Throttled Load Balancing Algorithm [14]

It is similar to Throttled Load Balancing Algorithm with the difference if no VM is free for a particular request then the priority of new request will be compared with the current executing request and if the priority of new task is higher it will be allocated to the VM and become the current task. The previous task will be placed in queue. In this way efficiency of Throttled Load Balancing algorithm is improved.

C. Equally Spread Current Execution Load Balancing Algorithm [15]

Spread spectrum approach is used by this algorithm. Load balancer spread the load on various node. According to this algorithm, the load on every virtual machine is same at any instant of time. The scheduler maintain VM allocation table, which includes the id of each virtual machine and active task count on each VM. According to new task allocated or completed, the count value in table is updated. Initially count value is zero for every VM. As new task came ESCE, scheduler look for the VM having least count value in table and allocate the task to that VM. Task queues are maintained at each VM.

D. Ant Colony load Balancing Algorithm [16, 17]

In this algorithm, we use Ant Colony Optimization (ACO) for load balancing. The request sent to the server is first sent to the ACO server, after arriving the requests, it checks for the response time from pheromone table and according to the response time, it is given to the server. In this way load on the network is distributed and response time decreases.

III. LOAD BALANCING

Load balancing [7] is the technique to distribute workload across cloud evenly to avoid a situation where some nodes are extra loaded and some are under loaded. Since cloud is dynamic in nature, major issue in cloud computing is to divide workload dynamically on demand, which can be made possible with the help of load balancer.

For e.g. suppose we are having a server to which large number of requests are allocated at the same time but since we are having limited number of resources (bandwidth, storage, CPU etc.), the server will down soon. This problem can be solved with the concept of load balancing in which multiple instances of that server will be made and that load will be distributed evenly among those instances.

Load balancing must take into account two major tasks, one is to allocate resource efficiently and other is to perform scheduling of task efficiently in distributed environment. Both of these tasks will ensure:

- Easy availability of resources on demand.
- Efficient utilization of resources under the condition of high/low load.
- Energy saving in case of low load.
- Higher user satisfaction will be there.
- System stability will be maintained.
- Quick user response and reduced task waiting time will be there.

Efficiency and effectiveness of load balancing algorithm can be measured by Cloud sim [8], which is a most efficient tool to model cloud. In Cloud Sim,

virtual machines (which is an instance of physical machine) are managed by host (physical server) which in turn is managed by datacenter (act as home to several host or several instances hosts).

A. Resource Allocation [1]

It is the way of allocating the resources to different cloud entities on-demand basis so that no node in cloud is extraloaded and not all the available resources in cloud should be wasted (bandwidth wastage or processing core wastage or memory wastage etc.).

Resource mapping is done at two levels:

- VM mapping onto the host:

Virtual machine residence is on the host and more than one instance of VM can be mapped on to a single host according to its availability and capabilities. Responsibility of host is to provide processing cores to VM. Algorithm must take care that characteristics of host and VM do not mismatch.

- Task mapping onto VM:

Tasks are executed on VM, each task require some processing power for its completion. VM provide that required processing power to the tasks that are mapped to it. Tasks must be mapped to appropriate VM according to its availability and configuration.

B. Task Scheduling [1]

After the resources are allocated to cloud entities task scheduling is done which defines the way in which allocated resources are available to the end users.

Scheduling of task can be done in two ways:

- Space shared (resources are not pre-empted until task execution is not complete).
- Time-shared (resources are continuously pre-empted until task is completed).

Therefore, to achieve the goal of minimizing the overall make span (completion time of last job to leave the system) of tasks on machines and to provide better utilization of available resources, we have to design several algorithms to provide satisfactory performance to both, cloud users and providers.

Table 1 : Comparison of Existing Load Balancing Algorithm

Algorithm	Static/Dynamic	Pros	Cons
Round Robin	Static	• Easy to implement. • Require single scheduler.	• Does not consider size of task and VM capacity. • Centralized and static in nature.
Weighted Round Robin	Static	• Consider the size of task and VM capacity. • Save the state of previous allocation of VM.	• Prior prediction of execution time is not known in advance. • Require scheduler at each VM.
Min-Min	Static	• Simple and fast. • Work better for smaller task.	• Task having minimum completion time is preferred which lead starvation for larger task. • Does not consider existing load on VM.

Max-Min	Static	• Simple algorithm. • Concurrently run shorter tasks.	• Select the task having maximum completion time first. • Poor load balancing.
Priority Based Modified Min-Min	Static	• Improved make span. • Priority is given to incomplete tasks.	• Re-scheduling of task after min-min algorithm.
Throttled	Dynamic	• List of VM is maintained with status. • Single scheduler is required.	• Entire list of VMs will be scanned from beginning. • Does not consider current load on VM.
Modified Throttled	Dynamic	• VM table is parsed from the index next to already assigned VM. • Faster response time.	• Current load on VM is not considered.
Equally Spread Current Execution	Dynamic	• Equal load on each VM. • Maximize the throughput.	• Scheduler at each VM level is required. • Less fault tolerant.
Ant Colony	Dynamic	• Under loaded node is found at beginning. • Not centralized	• Delay in moving in forward and backward direction. • More network overhead.

IV. CONCLUSION

In cloud, computing to balance the load among several instances of virtual machine is a big issue. To distribute the load evenly among all the resources we have to follow a load-balancing algorithm, which would help in utilizing the resources effectively so that no virtual machine is extraloaded or underloaded.

Algorithms are divided as static and dynamic depending upon present system state, which we discussed in detail in this paper. The goal of this paper is to discuss the comparative study of load balancing algorithm to reduce the overall make span with the help of the concept of resource allocation and task scheduling.

V. REFERENCES

[1]. "A Comparative Study of Load Balancing" International Journal of Distributed and Cloud Computing Volume 1 Issue 2 December 2013" by Mayanka Katyal, Atul Mishra.

[2]. "Cloud Computing: State-Of-The Art and Research Challenges" by Zhang, q., cheng, l. & boutaba, r 2010.

[3]. Open stack: an overview from www.openstack.org/downloads/openstack-overview-data-sheet.pdf.

[4]. "Measuring The Business Value Of VMware Horizon View" by Randy Perry Brett Waldman December 2013.

[5]. Apache CloudStack: Open Source Infrastructure as a Service Cloud Computing Platform", Hauang, a. Software architect, citrix systems apache cloud-stack architecture, JULY 2014.

[6]. "Cloud Computing Virtualization" by Mohd Saleem , international journal of computer applications technology and research volume 6–issue 7, 290-292, 2017.

[7]. "Load Balancing And its Algorithms In Cloud Computing: a survey" by- Sajjan R.S1, Biradar Rekha Yashwantrao2* International Journal of Computer Sciences and Engineering Vol.-5(1), Jan 2017, E-ISSN: 2347.

[8]. "CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms" by Rodrigo N. Calheiros, Rajiv Ranjan, Anton Beloglazov, César A. F. De Rose, and Rajkumar Buyya, Ieee press, new york, usa, leipzing, germany, june, 2009.

[9]. "Load Balancing and its Algorithms in Cloud Computing: A Survey" by Sajjan R.S1, Biradar Rekha Yashwantrao2 (International Journal of Computer Sciences and Engineering)

[10]. "Modified Optimal Algorithm For Load Balancing In Cloud Computing" by Mrs. Shruti tripathi , shriya prajapati, nazish ali ansari international conference on computing, communication and automation (iccca2017)

[11]. "Analysis Of Load Balancers In Cloud Computing" by shanti swaroop moharana1, rajadeepan d. Ramesh2 & digamber powar3,(international journal of computer science and engineering (ijcse) issn 2278-9960 vol. 2, issue 2, may 2013).

[12]. "An In-Depth Analysis And Study Of Load Balancing Techniques In The Cloud Computing Environment " from(2nd international symposium on big data and cloud computing (isbcc'15))by Geethu gopinath p p , shriram k vasudevan

[13]. "Enhanced Load Balancing Min-Min Algorithm for Static Meta Task Scheduling In Cloud Computing" from 3rd international conference on recent trends in computing 2015 (icrtc-2015) by Gaurang patel, rutvik mehta, upendra bhoi.

[14]. "Priority Based Modified Throttled Algorithm In Cloud Computing" by soumi ghosh1, chandan banerjee1 Vol.8, No.2 (2015), pp.9-14.

[15]. "A Review On Load Balancing Algorithm" by Nitika, Shaveta, Gaurav Raj, International Journal of advanced research in computer engineering and technology, Vol-1 issue-3 May-2012

[16]. "Cloud Load Balancing Based on ACO Algorithm" by avtar singh, kamlesh dutta, himanshu gupta (ijcta nov-dec-2015).

[17]. "Load balancing of node in network using ant colony optimization" by vaishnavi aher,sayali khairnar,madhuri shinde, priyanka shirole (international journal of computer science and engineering volume-3, issue-1 e-issn: 2347-2693)

[18]. "Large Data migration within Cloud Environment using compression and encryption technique" by Rajeshri Vaidya and prof. Sumedh Pundkar (International Journal of Innovative and Emerging Research in Engineering)

Cite this article as :

Ruchi, Harish Kumar, "A Survey on Load Balancing Algorithms", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 5 Issue 2, pp. 1156-1163, March-April 2019. Available at doi :
https://doi.org/10.32628/CSEIT1952242
Journal URL : http://ijsrcseit.com/CSEIT1952242