International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 2 Issue: 8 2151 – 2156

Load Balancing Algoritms in Cloud Computing Environment: A Review

Swati Katoch
Department of Computer Science
Himachal Pradesh University
Shimla, India
e-mail: katoch.swati53@gmail.com

Jawahar Thakur
Department of Computer Science
Himachal Pradesh University
Shimla, India
e-mail: jawahar.hpu@gmail.com

***Abstract*—** Cloud computing is an emerging internet based technology. Cloud is a platform providing pool of resources and virtualization. It is based on pay-as-you-go model. The numbers of users accessing the cloud are rising day by day. Generally clouds are based on data centers which are powerful to handle large number of users. The reliability of clouds depends on the way it handle loads, to overcome such problems clouds must be featured with the load balancing mechanism. Load balancing is required as we don't want one centralized severs performance to be degraded. A lot of algorithms have been proposed to do this task. In this paper we have studied the various existing load balancing algorithms and then compared them based on various parameters like resource utilization, scalability, stability etc.

Keywords- *Cloud computing, Load balancing, Virtualization, Virtual Machine (VM), Data Center*

*****

I. INTRODUCTION

Cloud computing is a paradigm evolved due to the advancements in technology and the wide-spread use of internet. *Cloud* is a metaphor to describe web as space where computing has been preinstalled and exist as a service; data, operating systems, applications, storage and processing power exist on the web ready to be shared[1]. In cloud computing environment resources such as CPU and storage, are provided as general utilities which can be accessed and released by the users through the internet in an on-demand fashion. Cloud computing allows enterprises to start with lesser resources and increase them only when there is a rise in service demand. It is one of the fastest implementing technologies. Many companies are trying to implement and introduce clouds due to their simple and flexible architecture. It has been employed by the organizations which includes social networking websites, online application, online software testing. Cloud computing has made a tremendous impact on IT industry over the past few years. Companies like Google, Amazon and Microsoft have made use of cloud computing to a very great extent and are still working towards its progress to provide reliable, powerful and efficient cloud platforms to their users.

Load Balancing is one of the central issues in cloud computing. It is a technique that distributes the dynamic workload evenly across all the nodes in the cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It is used to achieve a high user satisfaction and user ratio by making sure that no single node is overwhelmed and thereby hence improving the overall performance of the system. It also ensures that every computing resource is distributed efficiently and fairly among all the nodes. It further prevents bottlenecks of the system which may occur due to the load imbalance. When one or more components of any service fail, load balancing helps in continuation of the services, by implementing provisioning and de-provisioning of instances of applications without fail [2].

In this paper we have discussed few of the existing load balancing algorithms with their metrics comparison. Response time, resource utilization, migration of processes, centralized or decentralized, fault tolerance, stability and some more are taken as the parameters for metric comparison. In this paper first introduction is given then in II literature review of the papers studied, III introduces load balancing its types and challenges, in IV the various existing load balancing algorithms are compared, in V we have given the parameters for metric comparison of algorithms and finally we have concluded our study in VI.

II. LITERATURE REVIEW

Cloud computing is a fast growing area. It provides resources as a utility over the internet. A cloud is a style of computing in which dynamically scalable and often virtualized resources are provided as a service over the internet.

Load balancing is one of the central issues in cloud computing. It is a mechanism that distributes the dynamic workload evenly across all the nodes in the whole cloud to avoid a situation where some nodes are heavily loaded while others are idle or doing little work. It achieves optimal resource utilization, maximize throughput, minimize response time and avoid overload [2] In the survey, basic concepts of cloud clouting, load balancing and the different approaches towards efficient resource utilization through cloud computing will be studied. The following section consists of the survey of the related papers.

Qi Zhang, et al. [1] has introduced cloud computing as a new paradigm for hosting and delivering services over the internet. They have stated the reasons for what makes the cloud computing environment attractive to the business owners and the enterprises. In this paper they have surveyed cloud computing environment, highlighting its key concepts, related

technologies, architectural designs, characteristics, state–of–the art implementation as well as research challenges. This whole paper aims at providing better understanding of the design challenges of cloud environment and in identifying important research directions in this area.

Jayant Adhikari et al. [2] in their paper they have discussed the existing load balancing techniques in cloud computing and further compares them based on various parameters like resource utilization, scalability, stability etc.

Sandeep Sharma et al. [3] in their paper they have presented the performance analysis of various load balancing algorithms. They have considered two typical load balancing approaches static and dynamic. The comparison performed showed that the static load balancing algorithms are more stable than the dynamic, and also it is easier to predict the behavior of static, but at the same time dynamic algorithms are considered better than the static algorithms.

Ram Prasad Padhy et al. [4] discussed basic concepts of Cloud Computing and Load balancing. They have studied some of the existing load balancing algorithms, which can be applied to clouds. They also studied, the closed-form solutions for minimum measurement and reporting time for single level tree networks with different load balancing strategies. The performance of these strategies with respect to the timing and the effect of link and measurement speed were studied.

Hemant S. Mahalle et al. [5] their paper is a brief discussion on testing load balancing on proposed cloud model. The performances of three algorithms Round Robin, Equally spread current execution load and Throttled load balancing are studied. The request time for the three policies applied are same which means there is no effect on data centers request time after changing the algorithms. The calculated experimental work showed the cost analysis for each algorithm. The cost calculated for virtual machine usage per hour is same for two algorithms Round Robin, Equally spread current execution load but Throttled Load balancing algorithm reduce the cost of usage, so Throttled Load balancing algorithm works more efficiently in terms of cost for load balancing on cloud data centers.

Yi Zhao et al. [6] proposed Eucalyptus, an open source cloud-computing framework. It still lacks the ability of load balancing. In the paper author provides a implementation by adaptive live migration of virtual machines. A simple model has been designed and implemented, which decreases the migration time of virtual machines by shared storage and fulfills the Zero-downtime relocation of virtual machines by transforming them as Red Hat cluster services. They proposed a distributed load balancing algorithm Compare_and_Balance based on sampling to reach an equilibrium solution. The results showed that it converges quickly.

Martin Randles et al. [7] in their paper studied the three dynamic distributed algorithms: Honey bee foraging behavior, Biased random sampling and Active clustering. Firstly, a nature inspired algorithm may be used for self organization, achieving global load balancing via local server actions. Secondly, self- organization can be engineered based on random sampling of the system domain, giving a balanced load across all system nodes. Thirdly, the system can be reconstructed to optimize the job assignment of the servers. This paper aims at providing an evaluation and comparative study of these approaches.

Kumughato G et al. [8] in their paper have discussed some of the current load balancing techniques that are available in the cloud. The main advantage of cloud computing has been that of sharing the computing power as well as the storage area over a large network within limited budget and also the provision of being available whenever there is a need and demand. But it also comes along with some issues and challenges. Load balancing is one such area and several techniques and methodologies have been proposed in achieving a good load balancing in the cloud environment. A good load balancing will lead to a better performance and efficiency of the computation. The resources in a cloud based environment should be utilized and distributed evenly in such a way that it should not lead to the over utilization on one particular resource whereas the other resources that are available for computation should be kept idly. The overall objective of the load balancing should be to use each and every available resource efficiently without putting an extra load on other resources.

Subasish Mohapatra et al. [9] in their paper they have analyzed various policies utilized with different algorithms for load balancing using cloud analyst. Basically they have compared different variants of RR (Round Robin) for load balancing. Four different scheduling algorithms have been simulated along with different variants of round robin algorithm for executing the user request in cloud environment. Each algorithm is observed and their scheduling criteria like average response time, data center service time and the total costs of different data centers are found. According to the experiment and analysis round robin algorithm has the best integrate performance. Future work can be based on this algorithm modified and implemented for real time system.

Jasmin James et al. [10] ;The establishment of an effective load balancing algorithm and how to use Cloud computing resources efficiently and effectively is one of the ultimate goals of this paper. Firstly analysis of different Virtual Machine (VM) load balancing algorithms is done. Secondly, for an IaaS framework a new VM load balancing algorithm has been proposed and implemented in simulated cloud computing environment. This algorithm is named ‘Weighted Active Monitoring Load Balancing Algorithm’. This algorithm is basically proposed for the datacenters to effectively load balance requests between the available virtual machines assigning a weight, in order to achieve better performance parameters such as response time and data processing time.

Through the survey various load balancing algorithms were studied. Clouds are based on data centers which are powerful to handle large number of users. The reliability of clouds depends on the way it handles the loads, to overcome such problems clouds are featured with the load balancing mechanism. Load balancing in clouds has helped clouds to increase their capability, capacity which results in powerful and reliable clouds. Both static and dynamic load balancing algorithms have some advantages as well as weaknesses over each other. Various algorithms are compared to each other on the basis of performance parameters such as response time and resource utilization, scalability, Fault tolerance etc.

III. LOAD BALANCING

Load balancing is used to achieve equilibrium among the different systems. This further leads to the efficient utilization of computational resources and also provides the process to be performed more quickly thereby saving time and energy. The main reason for the use of load balancing is to optimize the resources available over the internet, to reduce the response duration from the server side, to achieve an efficient throughput and mainly to prevent the overloading of a particular resource while other may be in an idle state without performing any operation. Load balancing is one of the challenging features in cloud computing environment. Clouds come with the feature of elasticity so load balancing for any computation process cannot be ignored. A load balancing in cloud environment should provide for the instance of the application to be executed dynamically over the cluster of the system without the need for stopping the whole working operation and without even having to change any configuration of the system for that matter. It will ensure equal distribution of work load among the different resources and that no resources are over-utilized or under-utilized and that proper ratio utilization is done on the resources that are available over the cloud. It will also increase the scalability and provide the mechanism of fail-over by allocating and de-allocating instance request of the application. There are many different kinds of load balancing algorithms that are available, which can be categorized mainly into two groups. The following section will discuss these two main categories of load balancing algorithms.

A. Static Load Balancing Algorithms

In *Static* load balancing the balancing is done prior to the execution. It is done based on the deterministic or probabilistic nature and no changes can be made during the execution. Also is static load balancing the resources are shared in an equal manner and the time of execution-period cannot be determined exactly.

B. Dynamic Load Balancing Algorithms

In *Dynamic* load balancing the resources are distributed dynamically during the execution of the system and there is a need of a monitoring system to provide the communication among the various servers and the current load of a particular system. The current states of the resource are monitored and load is changed if necessary in a dynamic manner.

Challenges of Load Balancing in Cloud Computing [11]

Even though cloud computing is widely in today's world but still the research is in its initial stages. There still are scientific challenges yet to be resolved by the scientific community, the highlights being on the load balancing challenge.

- Automated service provisioning: Allocating or releasing the resources automatically i.e. elasticity is a key feature of cloud computing. The challenge is to how the clouds elasticity can be used by using the optimal resources and still with the same traditional systems performance?
- Virtual Machines Migration: The concept is to visualize a machine as a file or a set of file. Unloading a heavily loaded machine is possible by moving virtual machine between them, the objective being to distribute the load in a datacenter or a set of the same. The challenge here is to remove the bottlenecks in cloud computing systems when the virtual machine dynamically distributes the load.
- Energy Management: This again is a key point that allows everyone to use the resources from a global centre rather than using their own resource. This may be termed as the economy of scale and it definitely is a major advantage favoring cloud computing. The question arises as how to meet the performance while just using a part of the datacenter?
- Stored Data Management: The storage of data is another major requirement, whether that being a company outsourcing its data storage or an individual. So now, how can the data be distributed in the cloud with optimum storage and fast access which is mandatory?
- Emergence of small datacenters for cloud computing: A small datacenter can be of more benefit as it will consume less electricity plus would be cheaper comparing to a large one. And if this concept comes into play it will lead to geo-diversity computing. Load balancing will show up as a global scale problem for ensuring appropriate response time with optimal resource distribution.

IV. COMPARISON OF EXISTING LOAD SHARING ALGORITHM

In the table I given below few of the existing both static and dynamic algorithms are compared w.r.t to their description, advantages and disadvantages:

ISSN: 2321-8169
2151 – 2156

TABLE I: Comparison of existing Load Balancing Algorithms

Algorithms	Static/Dynamic	Description	Advantages	Disadvantages
Round Robin and Randomized [12]	Static	1. Processes are divided evenly between all processors. 2. The process allocation order is maintained on each processor locally independent of allocations from remote processors.	1. Works well with number of processes larger than number of processors. 2. Round Robin does not require inter-process communication.	1. These are not expected To achieve good performance in general case.
Threshold [13]	Static	1. Processes are assigned immediately upon creation to hosts. 2. Hosts for new processes are selected locally without sending remote messages.	1. Have low inter-process communication 2. A large number of local process allocations.	1. All processes are allocated locally when all remote processes are overloaded.
Central Manager [14]	Static	1. Central processor selects the host for new process. 2. Minimally loaded processor depending on the overall load is selected when process is created.	1. Load manger makes load balancing decisions based on the system load information, allowing the best decision when of the process created.	1. High degree of inter-process communication could make the bottleneck state.
Min-Min [11]	Static	1. Minimum completion time for all tasks is found. 2. From these minimum times the minimum value is selected which is the minimum time among all the tasks on any resources. 3. According to that minimum time, the task is scheduled on the corresponding machine.	1. Performs well with optimum number of resources	1. It can lead to starvation
Max-Min [11]	Static	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.	1. Performs well with optimum number of resources	1. It can lead to starvation
Honey Bee Foraging Behavior [7]	Dynamic	Achieves global load balancing through local server actions	1. Performs well as system diversity increases.	1. Does not increases throughput as the system size increases.
Biased Random sampling [7]	Dynamic	Achieves Load balancing across all system nodes using random sampling of the system domain.	1. Performs better with high and similar population of resources.	1. Degrades as population diversity increases.
Active Clustering [7]	Dynamic	Optimizes job assignment by connecting similar services by local re-wiring.	1. Performs better with high resources. 2. Utilizes the increased system resources to increase throughput.	1. Degrades as system diversity increases.
ACCLB(Ant Colony and Complex network Theory) [15]	Dynamic	Uses small-world and scale-free characteristics of complex network to achieve better load balancing.	1. Overcomes heterogeneity. 2. Adaptive to dynamic environments 3. Excellent in fault tolerance 4. Good Scalability	1. Used in Complex networks only 2. Low performance and scalability
Compare and Balance [16]	Dynamic	1. Based on sampling 2. Uses adaptive live migration of virtual machines	1.Balances load amongst servers 2.Reaches equilibrium fast 3. Assures migration of VMs from high-cost physical hosts to low-cost host.	1. Assumption of having enough memory with each physical host
Vector Dot [17]	Dynamic	Uses dot product to distinguish node based on the item requirement.	1.Handles hierarchical and multi-dimensional resource constraints 2. Removes overloads on server, switch and storage	1. Used in complex networks only

V. METRICS FOR LOAD BALANCING ALGORITHMS

The parameters considered for measuring the performance of the load balancing algorithms are :

- **Resource Utilization**
  Resource utilization includes automatic load balancing. A distributed system may have number of processes that demand more processing power. If the algorithm is capable to utilize resources, they can be moved to under loaded processes more efficiently.
- **Migration**
  This parameter provides when does the system decide to export a process? It decides whether to create the process locally or create it on a remote processing element. The algorithm is capable to decide that it should make changes of load distribution during execution of process or not.
- **Stability**
  Stability can be characterized in terms of delays in the transfer of information between processes and the gains in the load balancing algorithm by obtaining faster performance by a specified amount of time.
- **Scalability**
  The salability is the ability of an algorithm to perform load balancing for a system with any finite number of nodes.
- **Centralized/ Decentralized**
  Centralized schemes store global information at a designated node. The designated node is accessed by all sender or receiver nodes to calculate the amount of load–transfers and also to check that which tasks are to be sent from a node or received from the another node. In a distributed load balancing, every node executes balancing separately.
- **Fault Tolerance**
  Fault tolerance is the ability of an algorithm to perform uniform load balancing even in a case of node or link failure. An algorithm should be a good fault tolerant.
- **Response Time**
  Response Time is the time taken to respond by a particular load balancing algorithm. This parameter should be minimized.
- **Energy Consumption**
  It determines the energy consumption of all resources in the system. Load balancing helps in avoiding overheating by balancing the workload across all the nodes in the cloud environment, hence reducing the energy consumption.

TABLE II: Metrics Comparison of Load Balancing Algorithms

Parameters	Algorithms							
	Round Robin	***Randomized***	***Central Manager***	***Threshold***	***Honey Bee***	***Biased Random Sampling***	***Active Clustering***	***ACCLB***
Resource Utilization	Min	Min	Min	Min	Min	Min	Min	Min
Migration	No	No	No	No	No	No	No	No
Stability	High	High	High	High	Low	Low	Low	Low
Scalability	Yes	Yes	NA	NA	Yes	Yes	Yes	Yes
Centralized	No	No	Yes	No	Yes	Yes	Yes	Yes
Decentralized	Yes	Yes	No	Yes	No	No	No	No
Fault Tolerance	No	No	Yes	No	No	No	No	Yes
Response Time	Low	Low	High	Low	High	High	High	High
Energy consumption	Min	Min	Min	Min	Min	Min	Min	Min

VI. CONCLUSION

Cloud computing is a paradigm in which resources (e.g., CPU and storage) are provided as general utilities that can be accessed and released by the users through the internet in an on–demand fashion. It is a fast growing and very diverse area. Cloud environment provides number of facilities to its users, but still there are many research challenges. In this paper we studied the various existing load balancing algorithms. The above comparison showed that static load balancing algorithms are more stable in compare to dynamic and it is also easy to predict the behavior of static, but at the same time dynamic algorithms are always considered better than static algorithms.

REFERENCES

[1] Q Zhang , L Cheng , R Boutaba, "Cloud Computing: state-of-the-art and research challenges", J Internet Serv Appl, 20 April 2010

[2] J Adhikari , S Patil , " Load balancing The Essential factor In Cloud Computing", IJERT, ISSN : 2278-0181, Vol.1 Issue 10, December-2012

[3] S Sharma, S Singh and M Sharma, "Performance Analysis of Load Balancing Algorithms", International Science Index , Vol.2, no. 2, 2008

[4] R Prasad Padhy (107CS046), P Goutam Prasad Rao (107CS039). "Load balancing in cloud computing system "Department of Computer Science and Engineering National Institute of Technology, Rourkela Rourkela-769 008, Orissa, India May, 2011

[5] S Hemant Mahalle , R Parag Kaveri , V Chavan , " Load Balancing On Cloud Data Centers", IJARCSSE, Volume 3, Issue 1, January 2013

[6] Y Zhao, Huang , "Adaptive Distributed Load Balancing based on Live Migration of Virtual Machines in Cloud " IEEE 5th International Joint Conference on INC, IMS and IDC, August 2009

[7] M Randles, D Lamb, A. Talib Bendiab, " A comparative Study Into Distributed Load Balancing Algorithms For Cloud Computing", IEEE, 24th International Conference on Advanced Workshops 20-23, Page No.- 551-556

[8] G Kumughato, P Jeba , " A Survey of Load balancing Techniques in Cloud Environment", IJARCS, Volume5, No. 1, Jan-Feb 2014

[9] S Mohapatra, S Mohannty, K.Smruti , " Analysis of Different Variants in Round Robin Algorithms for Load Balancing in Cloud Computing", International Journal of Computer Applications (0975 – 8887) Volume 69– No.22, May 2013

[10] J James, B Verma, "Efficient VM Load Balancing Algorithm For a Cloud Computing Environment", IJSCE, Vol.4 No.09, September 2012

[11] A Kaur Sidhu, S Kinger,"Analysis of load balancing techniques in cloud computing" International Journal of Computers & Technology Volume 4 No. 2, March-April, 2013

[12] X Zhong, H Rong, "Performance study of load balancing algorithms in distributed web server systems" CS213 Parallel and Distributed Processing Power Project Report

[13] S Sharma, S Singh, M Sharma, "Performance analysis of load balancing algorithms " World Academy of Science, Engineering and Technology, Vol:2, 2008

[14] P.L.McEntire, J.G.O'Reilly, R.E. Larson "Distributed computing: concepts and implementations" New York: IEEE Press,1984

[15] Z Zhang, X Zhang, " A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation", Proceedings of 2nd International Conference on Industrial Mechatronics and Automation(ICIMA), Wuhan, China, May 2010

[16] Y.Zhao, W Huang," Adaptive distributed Load Balancing Algorithm based on live migration of virtual machines in cloud", Proceedings of 5th International Conference on INC, IMS and IDC, Seoul, Republic of Korea, August 2009

[17] A Singh, M. korpulu, D Mohapatra, " Server-storage virtualization: integrationand load balancing in data centers", Proceedings of the ACM/IEEE conference on Supercomputing (SC), November 2008

[18] L Jiyini, Q Meikang , N Jain-Wei , Y Chen, Z Ming "Adaptive rersource allocation for preeemptable jobs in cloud systems". IEEE International Conference on Intelligent Systems Design and Applications, pp. 31-36, 2010.

[19] S Chaisiri, Bu-Sung Lee, and D Niyato "Optimization of resource provisioning cost in cloud computing" IEEE transactions on services computing, 2011

[20] G. Khanna, K. Beaty, G. Kar, and A. Kochut, "Application Performance Management in Virtualized Server Environment," in Network Operations and Management Symposium, (2006). NOMS (2006). 10th IEEE/IFIP

[21] T Chou, "Introduction to Cloud Computing: Business & Technology" Active Book Press 2nd Edition, 2011.

[22] T Desai, J Prajapati, " A Survey of Load balancing Techniques and Challenges in Cloud Computing", International Journal of Scientific and Technology Research Volume 2, Issue 11, November 2013

[23] H Jinhua, G Jianhua, S Guofei, T Zhao, " A Scheduling Strategy on Load Balancing of Virtual Machine Resources In Computing Environment" IEEE explore, December 2010

[24] Songjie, Y Junfeng and W Chengpeng, "Cloud computing and its Key Techniques", International Conference on Electronic & Mechanical Engineering and Information Technology, 2011

[25] V Sarathy., P Narayan, M Rao, "Next Generation Cloud Computing Architecture- Enabling Real-time Dynamism for Shared Distributed Physical Infrastructure", 19th IEEE International Workshops on enabling Technologies: Infrastructure for Collaborative Enterprises, WETICE, June 2010.

[26] T.R. Nair , G Krishnan and M Vaidehi, "Efficient Resource Arbitration and Allocation Strategies in Cloud Computing through Virtualization", In Proceedings of IEEE International Conference on Cloud Computing and Intelligent Systems, CCIS, September 2011.