International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Impact Factor (2012): 3.358

A Survey on Load Balancing Techniques in Cloud Computing

Jaswinder Kaur[1], **Supriya Kinger**[2]

[1]Student, Sri Guru Granth Sahib World University, India

[2]Assistant Professor, Sri Guru Granth Sahib World University, India

Abstract: ***Cloud computing is an emerging computing paradigm, which aims to share data, calculations, and service transparently over a scalable network of nodes. On Cloud computing platform, resources are provided as services and by needs. The system should avoid wasting resources as result of under utilization and avoid lengthy response time as result of over utilization. In cloud computing, virtual machine allocation problem is a key to build a cloud environment. Initially this paper gives an introduction to Cloud computing and Load balancing. A detailed survey on classification of different Load balancing techniques is done and presented. Further comparison analysis of different Load balancing algorithms has been shown. Finally, last Section summarizes conclusion and future work***.

Keywords: Cloud Computing, Data centers, Virtualization, Load Balancing.

1. Introduction

Cloud computing is an emerging trend which has progressed to the point of serious adoption in both public and private sector organizations. Cloud computing is a phrase used to describe a variety of computing concepts that involve a large number of computers connected through a real time communication network such as the internet. Cloud computing is a model for enabling convenient, on demand network access to a shared pool of configurable resources(e.g. Networks, servers storage, applications, and services) that can be rapidly provisioned and released with minimal management efforts or service provider interaction [1].

The popularity of term can be attributed to its use in marketing to sell hosted service in the sense of application service provisioning that run client server software on a remote location. As shown in fig: 1 Cloud computing relies on sharing of resources to achieve coherence and economies of scale. Cloud computing is attractive to business owners as it eliminates requirement for users to plan ahead foe provisioning and allow enterprises to start from small and increase resource only when there is rise in service demand.

Figure1: Cloud Computing

Nowadays Cloud computing has become a key technology for online allotment of computing resources and online storage of user's data in a lower cost, where computing resources are available all the time, over the internet with pay per use concept. There are various advantages of cloud computing including virtual computing environment, on-demand services, maximum resource utilization and easy to use services etc. But there are also some critical issues like security, privacy, load management and fault tolerance etc which needs to be addressed for better performance [2]. As we know load is very unpredictable, even a spike will results overloaded nodes, which often lead to performance degradation and are vulnerable to failure. So Load balancing is one of the main challenges in Cloud computing which is required to distribute the dynamic workload across multiple nodes to ensure that no single node is overwhelmed.

2. Load Balancing

With the increasing popularity of cloud computing, the amount of processing that is being done in the Cloud computing is rising rapidly. As the requests of the clients can be random to nodes, thus the load on each node can also vary i.e. some nodes are overloaded and some are under loaded which directly affects the quality of cloud services. Therefore, some load balancing mechanism is needed to ensure that every computing resource is distributed efficiently and fairly.

Load Balancing is a computer networking method to distribute workload across multiple computers or a computer cluster, network links, central processing units, disk drives, or other resources, to achieve optimal resource utilization, maximize throughput, minimize response time, and avoid overload [3]. It is a mechanism that distributes the dynamic local workload evenly across all the nodes in the whole cloud to achieve a high user satisfaction and resource utilization, hence improving the overall performance and resource utility of the system. It also ensures that every computing resource is distributed efficiently and fair, prevents bottlenecks and fail-over.

2.1 Classification of Load Balancing Algorithms

In general, load balancing algorithms follow two major classifications [4]:

2.1.1 Static Algorithms

A static load balancing algorithm does not take into account the previous state or behaviour of a node while distributing the load. In this approach prior knowledge of system is needed. This has a major impact on the overall system performance due to the unpredictability of load fluctuation of the distributed system. It doesn't depend upon current state of system. Static algorithms are much simpler as compared to dynamic algorithms.

2.2.2 Dynamic Algorithms

This approach takes into account the current state of the system during load balancing decisions and is more suitable for widely distributed systems such as cloud computing [5]. The dynamic load balancing algorithm is applied either as a distributed or non-distributed. The advantage of using dynamic load balancing is that if any node fails, it will not halt the system; it will only affect the system performance. In a dynamic load balanced system, the nodes can interact with each other generating more messages when compared to a non-distributed environment.

3. Existing Load Balancing Algorithms

There are many load balancing techniques given by the researchers over time to time, some has advantages over other and vice versa. Load balancing is required to achieve the maximum throughput, performance and decrease the response time. Load balancing techniques mainly focus on reducing overhead, reducing the migration time, improving performance and increasing response to request ratio. Here in this paper five different Load balancing algorithms are discussed.

3.1. Round Robin Algorithm

H. Mahalle et al. [6] discussed this method in which jobs are divided evenly between all processors in a round robin order without considering the work load. Here the time slicing mechanism is used, which divides the time into multiple slices and each node is given a particular time slice or time interval in which they have to perform their task. 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. The main advantage of Round Robin algorithm is that it does not require inter process communication. However when the jobs are of unequal processing time this algorithm suffers as the some nodes can become severely loaded while others remain idle.

3.2. Throttled Load Balancing Algorithm

A. Sidhu et al. [7] discussed load balancing algorithm which is completely based on concept of finding the appropriate virtual machines for assigning a particular job. In this the job manager is having a list of all virtual machines, using this indexed list, it allot the desire job given by client to the appropriate machine. As client requests, if the job is well suited for a particular machine on the basis of size and availability of the machine, then that job is assign to the appropriate machine. If no virtual machines are available to accept jobs then the job manager queued the request. This algorithm performs well as compared to round robin algorithm.

3.3. A Task Scheduling Algorithm based on Load Balancing

M. Nikita et al. [8] proposed a two level scheme for load balancing. The first level scheduling is from user application to the VM, and the second is from the VM to host resources. In this two level scheduling model, the first scheduler create the task description of virtual machine, then the second scheduler finds appropriate resources for the virtual machine in the host resource, hence overall performance is increase. The main disadvantage of this algorithm is it does not improve the response to request ratio.

3.4. Enhanced Equally Distributed Load Balancing Algorithm

S. Mulay et al. [9] proposed an algorithm which handles the requests with priorities. It is a distributed algorithm by which the load can be distributed not only in balanced manner but also it allocates the load systematically by checking the counter variable of each data center. After checking it transfer the load accordingly i.e. the minimum value of the counter variable will be chosen and the request is handled easily and takes less time and give maximum throughput .This result efficient response to request ratio When it come to performance analysis the proposed algorithm perform well with various metric like Performance, Resource utilization, Scalability, Fault tolerance and Response time.

3.5 Efficient Load Balancing Method based on Cloud Service

R.Wang et al. [10] proposed an algorithm to addresses the problem with a policy for creating load balancing method in both Physical Machines layer and Virtual Machines layer migration, moreover it also introduced prediction method to ensure the transient spike does not trigger needless migration, and in Virtual Machines layer benefit estimate model is used in order to decide whether the migration of jobs in Virtual Machines of same Physical Machine is benefit for whole system. This algorithm gives better performance and thus ensures higher QoS.

4. Comparative Analysis of Load Balancing Algorithms

In above section we have discussed a number of load balancing algorithms/ techniques. Each of the algorithms has its own merits and demerits. A comparative analysis of various algorithms shows the effectiveness of that algorithm. When it comes to performance analysis of the cloud system, there are some metrics on which the analysis can be done. Some of the important metrics includes Nature, Process Migration, Resource Utilization, Stability, Predictability, Reliability, Adaptability, Response Time etc. Various metrics considered for comparison of above discussed Load balancing techniques in cloud computing are discussed below [11]

4.1 Nature of an algorithm defines the nature or behaviour of load balancing algorithm i.e. whether static or dynamic pre-planned or no planning.

4.2 Process Migration parameter provides when does a system decide to export a process? It decides whether to create it 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.

4.3 Resource Utilization used to check the utilization of available resource given to the cloud. Static load balancing algorithms have lesser resource utilization as static load balancing methods just tries to assign tasks to processors in order to achieve minimize response time ignoring the fact that may be using this task assignment can result into a situation in which some processors finish their work early and sit idle due to lack of work.

4.4 Stability can be characterized in term of the delay in the transfer of information between processor and the gain in load balancing algorithm. Static load balancing algorithm considered as stable as no information regarding present workload state is passed processors. However in case of dynamic load balancing such kind of information is exchanged among processors.

4.5 Predictability factor is related with the deterministic or nondeterministic factor that is to predict the outcome of the algorithm. Static load balancing algorithm's behaviour is predictable as most of the things like average execution time of processes and workload assignment to processors are fixed at compile-time. Dynamic load balancing algorithm's behaviour is unpredictable, as everything has been done at run time.

4.6 Reliability factor is related with the reliability of algorithms in case of some machine failure occurs. Static load balancing algorithms are less reliable because no task/process will be transferred to another host in case a machine fails at run-time. Dynamic load balancing algorithms are more reliable as processes can be transferred to other machine in case of failure occurs.

4.7 Adaptability factor is used to check whether the algorithm is adaptive to varying or changing situations i.e. situations which are of dynamic nature. Static load balancing algorithms are not adaptive as this method fails in varying nature problems. Dynamic load balancing algorithms are adaptive towards every situation whether numbers of processes are fixed or varying one.

4.8 Response Time is defined as how much time a distributed system using a particular load balancing algorithm is taking to respond? Static load balancing algorithms have shorter response time as one should not forget that in Static load balancing there is lesser overhead so emphasis is totally on executing jobs in shorter time rather than optimally utilizing the available resources. Dynamic load balancing algorithms may have relatively higher response time.

4.9 Fault Tolerant enables an algorithm to continue operating properly in the event of some failure. If the performance of algorithm decreases, the decrease is proportional to the seriousness of the failure, even a small failure can cause total failure in load balancing.

A Comparative analysis of above discussed load balancing algorithms is done below:

Table 1: Comparative Analysis of Existing Load Balancing Algorithms

Parameters	**Various Load Balancing Algorithms**				
	Round Robin Algorithm	Throttle based algorithm	Task scheduling algorithm	EEDLB	Efficient load balancing method based on cloud service
Nature	Static	Dynamic	Dynamic	Dynamic	Dynamic
Process Migration	No	Yes	Yes	Yes	Yes
Resource Utilization	Less	More	More	More	More
Stability	Large	Small	Small	Small	Small
Predictability	More	Less	Less	Less	Less
Reliability	Less	More	More	More	More
Adaptability	Less	More	More	More	More
Response Time	Less	More	More	More	More
Fault Tolerant	No	No	No	Yes	No

5. Conclusion and Future Work

Cloud computing system has widely been adopted by the industry, through there are many issues existing like Load Balancing, Migration of Virtual machines, Server Unification which has not been yet fully addressed. On the contrary the Load balancing is the most central issue in the system i.e. to distribute the load in an efficient manner. In this paper we discussed and compared various Load balancing algorithms. Future work can consider different types of application in cloud, and establish a more detailed load assess system. Self learning, self healing mechanism can be taken into consideration. Issues like Energy Management and Carbon Emission can also be worked upon.

References

[1] Qi Zhang, Lu Cheng, R.Boutaba "Cloud Computing: state-of-art and research challenges", ©

[2] The Brazilian Computer Society 2010, 8 January 2010/ Accepted: 25 February 2010/ Published online: 20 April2010.

[3] P.Mathur, “Cloud Computing: new challenge to the entire computer industry”, 1stInternational conference on parallel, distributed and grid computing, 2010, pp978-1-4244-767.

[4] U.Chatterjee, “A Study on Efficient Load Balancing Algorithms in Cloud computing Environment”, International Journal of Current Engineering and Technology, Vol.3, 11 November 2013

[5] S.Mohinder, R,Ramesh, D.Powar, “Analysis of Load Balancers in Cloud Computing”, International Academy of Science, Engineering & Technology, vol.2, May 2013.

[6] A.khiyait, H.Bakkali, M.Zbakh, D.Kettani, Load Balancing Cloud Computing: state of art”, University Mohammed V Souissi Rabat Morocco, 2012.

[7] H.Mahalle, P.Kaveri, V.Chavan, “Load Balancing on Cloud Data Centers”, international Journal of Advance Research in computer Science and Software Engineering, vol. 3, Jan. 2013.

[8] A.Sidhu, S.Kinger, “Analysis of Load Balancing techniques in Cloud Computing”, Council for innovative research international Journal of Computer & Technology, vol.4, March-April 2013.

[9] M.Nikita, “Comparative Analysis of Load Balancing Algorithm in Cloud Computing”, International Journal of Engineering and Science, vol.01.

[10] S.Mulay, S.Jain, “Enhanced Equally Distributed Load Balancing Algorithm for Cloud Computing”, International Journal Of Engineering and Technology, vol.02, Jun. 2013.

[11] R.Wang, W.Le, X.Zhang, “Design and Implementation of efficient Load Balancing method for virtual machine cluster based on cloud service”, School of Information Science and Engineering, Yunnan University, Kunming, China.

[12] Rajguru, S. Apte, ” A comparative Performance Analysis of Load Balancing Algorithms in Distributed System using Qualitative Parameters”, International Journal of Recent Technology and Engineering, Vol.01, Issue-3, August 2012.

Author Profile

Jaswinder Kaur is Student of M.Tech in the department of Computer Science at SGGSWU, India. She has done B.Tech from BBSBEC, Under Punjab Technical University, India .

Supriya is currently working with SGGS World University, India, as Assistant Professor. She received her Master’s in Technology degree in field of Computer Science from YMCA, India and Bachelor’s in Technology degree in Computer Science from Kurukshetra University, India. Her areas of interest include Parallel and Distributed Processing, Cloud Computing and Software Engineering. Currently she is doing research in field of Energy Efficient Clouds‟. She has more than 8 years of experience in UG & PG teaching and research. She has attended and organized a number of workshops and conferences and has presented papers in field of Green Clouds, Component Based Software Engineering, Information Security and Networks. She is a life member of ISTE.