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

A Comprehensive Analysis of Existing Load Balancing Algorithms in Cloud Network

Pinki [1], Nida[2]

[1, 2,] M.Tech (CSE), School of Computing Science and Engineering, Galgotias University, Greater Noida , India

Abstract: ***In the cloud computing paradigm, the scheduling of computing resources is a critical part of cloud computing field. With increment in number of users and the type of applications on the cloud computing platform, effective utilization of resources in the system becomes a critical concern to ensure service level agreements (SLA).Resource distribution and the effective load balancing are necessary mechanism to increase the service level agreement (SLA) and better uses of available resources in heterogonous environment. Proper load balancing technique helps in implementing fail-over, avoidance of bottleneck issues, providing scalability, optimization of resource allocation, increasing reliability and user satisfaction etc. in cloud computing. Cloud has many types of load concern like memory load, CPU load, network load and server load. In order to improve the performance of the whole cloud environment, Load Balancing algorithms are needed to distribute the load evenly across all the nodes in cloud. This paper discusses many load balancing techniques used to figure out the issue in cloud computing environment.***

Keywords: Cloud Computing, Cloud load Virtualization, Load Balancing in cloud, load balancing algorithms, SLA, Challenges

1. Introduction

Cloud computing is one of the most emerging computing framework based on the growth of distributed network of computing, parallel processing, and the grid computing network [1]. Cloud Computing is defined by the National Institute of Standards and Technology (NIST) as *"a model for facilitating convenient, on-demand , scalable 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 effort or by interaction with cloud service provider"* [3].Cloud Computing delivers a elastic, scalable and easy way to preserve and recollect data and files as a part of its services. Particularly for making big data sets and files accessible for increasing number of users around the globe.Dealing with such prominent data sets involve several techniques to optimize and streamline operations and to supply satisfactory levels of performance for the users. One of the important issues associated with cloud computing area is dynamic load balancing or task scheduling. In Cloud Computing the main concerns demands for the assignment of tasks to cloud nodes so that the effort and request processing is done as efficiently as possible [2], while being able to endure the various involving constraints such as heterogeneity and high communication delays.

A cloud computing is a distributed and scalable system where resources are distributed throughout the network. Entire resources of the system must collaborate to respond to a client request. This can give rise to bottlenecks in the network and an unbalanced propagation of charge in a distributed system where some components will be excessively charged whereas others will be less charged or not. To overcome this problem, we use the concept of load balancing algorithms in distributed systems environment.

2. Overview of Load Balancing

Load balancing is the process of enhancing the performance of a parallel and distributed system via the redistribution of load among the various processing units or nodes. Load balancing is represented as, "In the scenario of distributed network environment of computing hosts, the functioning of the system is heavily dependent upon dividing up work effectively across the several participating nodes" [4].

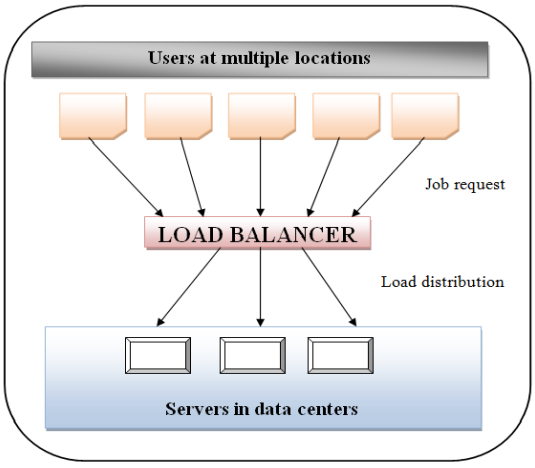

Figure 1: Load balancing in cloud

Load balancing algorithms are broadly divided into two major categorization [6]:

- Based on how the charge is distributed and how processes are allocated to system nodes.
- Based on the information status of the nodes.

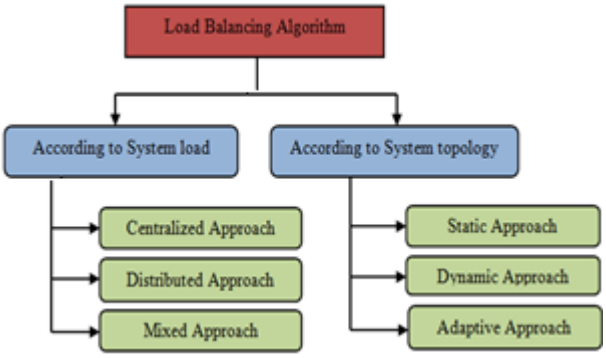

Figure 2: Categorization of Load Balancing algorithms

A) Categorization according to the system load

1) Centralized approach: Here, A single node is responsible for managing the distribution of resources within the whole network.
2) Distributed approach: In this, each and every node independently forms its own load vector by gathering the load information of other nodes. Decisions are locally made using local load vectors. This approach is more desirable for extensive distributed systems such as cloud computing.
3) Mixed approach: It takes the benefits of both centralized and distributed approach.

B) Categorization according to the system topology

1) Static Approach: Static algorithms are generally suitable for homogeneous as well as stable environments and can yield better results in these environments. Nevertheless, they are generally not flexible and are unable to match the dynamic changes to the attributes throughout the execution time [5].
2) Dynamic Approach: Dynamic algorithms are more flexible and are able to take into account different types of attributes in the system, involving both prior to and during run-time [5].
3) Adaptive Approach: These approaches are suited to adapt the distribution of load to system, by changing their parameters dynamically as well as their algorithms. They are able to provide better performance when the system state changes frequently and are more suitable for scalable distributed systems such as cloud networks.
4) These algorithms can conform to changes and allow for the better results in heterogeneous and dynamic environments. As the distribution of attributes become more complex and dynamic, some of these algorithms could become inefficient and may cause more overhead than required, resulting in overall degradation of the services performance.

3. Cloud Virtualization

In context of cloud computing, virtualization is a technique which allows sharing single physical instance of a resource among multiple organizations or customers. Virtualization is synonymous to something that is not real, but supplies all facilities that are existing in the real world. Virtualization provides all different services of cloud computing to the end user by remote data center with partial virtualization or full virtualization manner [22]. It can help in load balancing by enabling extremely responsive provisioning and avoiding hotspots in data center. Mainly two types of virtualization mainly exist:

- Full Virtualization: In full virtualization the complete installation of one system is done on another system, so that all the software which is available in actual server will also be present in virtual system. It also allows sharing of computer system among multiple users and simulated hardware situated on different systems are available.
- Para Virtualization: In Para virtualization the hardware system is not emulated. The client software runs their own isolated field. In this entire services are not fully available, but partial services are supplied. The Para virtualization functions with an operating system that has been altered to work in a virtual machine. Better efficiencies of this virtualization can also lead to better scaling.

4. Review of Load Balancing Algorithms

Several load balancing algorithms have been proposed. They are discussed as follows:

1) Index Name Server (INS): The goal [7] of INS is to minimize the data duplication and redundancy and it integrates reduplication and access point selection optimization. There are several parameters required in the process of computing the optimum selection point. The included parameters are the Hash code of block of data to be downloaded, the location of the server that has target block of data, the transition quality which is computed based on the node performance and a weight judgment chart, the maximum bandwidth of downloading from target server and the path parameter.
2) Exponential Smooth Forecast based on Weighted Least Connection (ESWLC): WLC (weighted least connection) [8] algorithm allots jobs to node based on the number of connections that exist for that node. But, WLC does not include the potentialities of every node such as processing speed, storage capacity of nodes and bandwidth. In order to overcome the shortcomings of WLC, ESWLC (Exponential Smooth Forecast based on Weighted Least Connection) algorithm has been proposed, which improves WLC by taking into consideration the time series and trials. ESWLC makes the decision based on the experience of node's CPU power, memory, number of connections to nodes and the amount of disk space currently utilised.ESWLC then figures out which node is to be selected based on exponential smoothing.
3) Downloading algorithm from FTP server (DDFTP): The proposed DDFTP algorithm [9] in cloud computing load balancing works by splitting a file of size n into n/2 partitions. The algorithm minimizes the network communication required between the client and nodes, thus reduces the network overhead. Furthermore, properties such as network load, node load, and network speed are automatically taken into account, while no run-time supervision of the nodes is required.
4) Honeybee foraging algorithm: Basically, this algorithm is [10] originated from the behavior of honey bees for discovering and drawing food. In terms of load balancing, as the web servers demand increases or decreases, the services are allotted dynamically to determine the changing demands of the user. All servers are grouped under virtual servers; each has its own virtual service queues. Every server processing a request from its own queue to compute a profit, which is equal to the quality that the bees express in their waggle dance. One measure of this profit can be the amount of time that the CPU spends on processing of a particular request. The main advantages are maximizing the throughput; waiting time on task is minimum and overhead become minimum. The disadvantage is if there are the more priority based queues present then the lower priority load can be remain continuously in the queue.
5) Active Clustering: A self-aggregation load balancing technique was investigated by M. Randles et al. [11]

which is a self-aggregation algorithm. It optimizes the job assignments by linking similar services using local re-wiring. The performance of the system is improved with high availability of the resources thereby enhancing the throughput by using these resources effectively. It is degraded with an growth in system diversity.

6) CARTON- The proposed mechanism [12] CARTON, for the cloud control that unites the concept of LB and DRL.Load Balancing is used to equally distribute the jobs to different servers in order to minimize the associated costs and DRL (Distributed Rate Limiting) is applied to make sure that the resources are distributed in such a way to maintain a fair resource allocation.
7) ACCLB- Ant colony and complex network theory for load balancing mechanism was proposed by Z. Zhang et al. [22] in an open cloud federation It makes use of small-world and scale-free features of a complex network to attain better load balancing. ACCLB overcomes heterogeneity, is adaptive to the dynamic environments, excellent in fault tolerance and has better scalability. It also helps in improving the performance of the system.
8) Min-Min Algorithm: This scheduling algorithm establishes the minimum completion time for every unscheduled job, and then allots the job with the minimum completion time to the node that offers it this time [13]. Min-Min considers the minimum completion time for every job at each round and it can schedule the job that will increase the overall completion the least.
9) Max-Min: Max-Min works similar to the Min-Min algorithm. But, it gives more priority to the larger tasks [14]. The jobs that have large execution time or large make-span time are executed first; consequently jobs with smaller execution time will have to keep waiting for long time.
10) 2-Phase Load Balancing Algorithm :(OLB plus LBMM) - S.-C. Wang et al. [15] proposed a algorithm that combines both OLB (Opportunistic Load Balancing) and LBMM (Load Balance Min-Min) scheduling algorithms to use better executing efficiency and preserve the load balancing of the system. OLB keeps every node in working state in order obtain the goal of load balance and LBMM scheduling algorithm is utilized to reduce the execution time of each task on the node thereby minimizing the overall make-span time.
11) Power Aware Load Balancing (PALB) Algorithm: PALB, preserves the state of all the calculated nodes, and based on utilization percentages, decides the number of calculated nodes that should be operating. The proposed load balancing algorithm that could be implemented to the cluster controller of a local cloud that is power aware and use a job scheduler to simulate requests from the users for virtual machine instances. The goal of the PALB algorithm is to maintain the availability to compute nodes while minimizing the total power consumed by the cloud [16].
12) LBVS: Load balancing virtual storage strategy proposed by Liu et al. [17], provides a large scale net data storage as well as Storage as a service model which is based on Cloud Storage. Storage virtualization is attained using a three -layered architecture and load balancing is attained by utilizing two load balancing modules. It serves in amending the efficiency.
13) Biased Random Sampling: The proposed algorithm [18] is based on the construction of the virtual graph having connectivity between every node of the system where each and every node of the graph corresponds to the nodal computer of the cloud system. It is scalable, reliable and an effective approach to balance the load of the cloud system.
14) Join-Idle-Queue- This algorithm was proposed by Y. Lua et al., for dynamically scalable web services. It allows for large- scale load balancing with distributed dispatchers, firstly by, load balancing idle processors across dispatchers for the idle processors at each dispatcher and then, assigning the jobs to processors to minimize average queue length at each processor[19].
15) Load Balancing for Real-time, Location-based Event Processing on Cloud Systems: In this the main focus was balanced distribution of workload for location-based event processing on the cloud computing systems ,because in cloud systems the workload distribution is very important with respect to the system performance. Sungmin Yi et al., introduced scalable real-time event processing techniques for range queries, that are; (1) Round robin data distribution, (2) Round robin query distribution, (3) data/query distribution thru space partitioning and (4) skew-aware distribution. The round robin data distribution technique mainly emphasizes on a balanced distribution of event data when the queries are replicated in all the cluster nodes. The round-robin query distribution method focuses on evenly distribution of queries whereas the event data are replicated in every worker. In the data/query distribution via space partitioning mainly distributes the event data and queries based on their spatial data characteristic. In the skew-aware distribution method takes the non-uniformity of queries and event data [23].It focused on evenly distribution of the workload, improving the performance and efficiency of cloud systems.

5. Challenges for Load Balancing in Cloud Computing

There are some metrics that can be improved for the better load balancing in cloud computing [20][21].

- Scalability: It finds out the ability of the system to accomplish load balancing algorithm with a finite number of nodes or processors.
- Throughput: It is defined as the maximum number of jobs that have completed their execution for a given period of time. A high throughput is required as a parameter for better performance of the system.
- Fault Tolerant: It is defined as the ability an algorithm to perform accurately and uniformly even in the circumstances of failure at any arbitrary node in the system.
- Migration time: The time taken by the process to migrate from one system node to another for the execution is known as Migration time. This time should always be less for better the performance of the cloud system.
- Response time: It is defined as the minimum amount of time that a specific load balancing algorithm requires to respond in a distributed system. This time ought to be reduced for better performance.

- Resource utilization: It is the extent to which the resources of the system are utilized. An efficient load balancing algorithm must make optimum utilization of the available resources.
- Performance: It is the overall efficiency of the cloud system. If all the parameters of the system are improved then possibly overall performance can be improved.

6. Conclusion

With the ever increasing technological advancement, emergence of Cloud Computing model will rapidly change the landscape of information technology. However, regardless of the significant benefits offered by the cloud computing, load balancing of the current model is one of the biggest issues. This paper has discussed several existing load balancing techniques that have been analyzed, mainly focuses on minimizing overhead, service response time, improving system performance, maximizing throughput and better resource utilization etc. In cloud, large number of parameters and different types of soft computing techniques can be incorporated in the future for the better resource utilization and requirements of the user.

References

[1] Vaquero L M, Rodero Merino I, Caeeres J. A break in the clouds: Towards a cloud definition [J]. ACM SIGCOMM Computer Communication Review, 2009, 39(1): 50–55.
[2] Randles, M., D. Lamb and A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing," in Proc. IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Perth, Australia, April 2010.
[3] Mell, P., and Grance, T. 2011. The NIST definition of Cloud computing (draft), NIST. [Online]. Available: http://csrc.nist.gov/publications/drafts/800-145/Draft-SP-800-145_Cloud-definition.pdf.
[4] E. Anshelevich, D. Kepme, J. Kleinberg, "Stability of Load Balancing Algorithms in Dynamic Adversarial Systems", Proceeding of the 34th annual ACM symposium on Theory of Computing, 2002.
[5] Rimal, B. Prasad, E. Choi and I. Lumb, "A taxonomy and survey of cloud computing systems" .In proc. 5th International Joint Conference on INC, IMS and IDC, IEEE, 2009.
[6] Elarbi Badidi, Architecture et services pour la distribution de chargedans les systèmes distribués objet, Université de Montréal Faculté des études supérieures, these doctorale, 20 juillet 2000 .
[7] T-Y., W-T. Lee, Y-S. Lin, Y-S. Lin, H-L. Chan and J-S. Huang, "Dynamic load balancing mechanism based on cloud storage" in proc. Computing, Communications and Applications Conference (ComComAp), IEEE, pp:102-106, January 2012.
[8] Lee, R. and B. Jeng, "Load-balancing tactics in cloud," in proc.
[9] International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), IEEE, pp: 447-454, October 2011.
[10] Al-Jaroodi, J. and N. Mohamed. "DDFTP: Dual-Direction FTP," in proc. 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), IEEE, pp: 504-503, May 2011.
[11] M. Randles, D. Lamb, and A. Taleb-Bendiab, "A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing", Proceedings of 24th IEEE International Conference on Advanced Information Networking and Applications Workshops, Perth, Australia, April 2010, pages 551-556.
[12] Randles M., Lamb D. and Taleb-Bendiab A., 24th International Conference on Advanced Information Networking and Applications Workshops, 551-556,2010.
[13] Stanojevic R. and Shorten R., IEEE ICC, 1-6, 2009.
[14] Vouk, "Cloud Computing- Issues, Research and Implementations," Information Technology Interfaces, pp. 31-40, June 2008.
[15] S.Mohana Priya, B.Subramani, "A New Approach for Load Balancing in Cloud Computing", International Journal of Engineering and Computer Science, May 2013.
[16] Wang S., Yan K., Liao W. and Wang S, 3rd International Conference on Computer Science and Information Technol-ogy, 108-113, 2010.
[17] Jeffrey M. Galloway, Karl L.Smith, Susan S. Vrbsky, "Power Aware Load Balancing for Cloud Computing", World Congress on Engineering and Computer Science, 2011.
[18] Liu H., Liu S., Meng X., Yang C. and Zhang Y., International Conference on Service Sciences (ICSS), 257-262, 2010.
[19] Martin Randles, David Lamb, A. Taleb-Bendiab, 2010"A Comparative Study into Distributed Load balancing Algorithms for Cloud Computing" IEEE 24th International Conference on Advanced Information Networking and Applications Workshops, pp: 1/10.
[20] Lua Y., Xiea Q., Kliotb G., Gellerb A., Larusb J. R. and Green-ber A,"Int. Journal on Performance evaluation",2011.
[21] Foster, I., Y. Zhao, I. Raicu and S. Lu, "Cloud Computing and Grid Computing 360- degree compared," in proc. Grid Computing Environments Workshop, pp: 99- 106, 2008.
[22] Buyya R., R. Ranjan and RN. Calheiros, "InterCloud: Utilityoriented federation of cloud computing environments for scaling of application services," in proc. 10th International Conference on Algorithms and Architectures for Parallel Processing (ICA3PP), Busan, South Korea, 2010.
[23] Sotomayor, B., RS. Montero IM. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," in IEEE Internet Computing, Vol. 13, No. 5, pp: 14-22, 2009.
[24] Sungmin Yi, Hyoseok Ryu, Yon Dohn Chung, "Load Balancing for Real-time, Location-based Event Processing on Cloud Systems", 2013 IEEE 16th International Conference on Computational Science and Engineering.

Author Profile

Pinki is a student of M.Tech (Computer Science and Engineering) at School of Computing Science and Engineering, Galgotias University, Greater Noida, India. She has passed her B.Tech from B.B.S College of Engineering and Technology,Allahabad (UP), India in 2011. Her areas of interest are Computer Networks and Information Security.

Nida is a student of M.Tech (Computer Science and Engineering) at School of Computing Science and Engineering, Galgotias University, Greater Noida, India. She has passed his B.Tech from Integral University,Lucknow (UP), India in 2012. Her areas of interest are Data Compression and Information Security.