A Survey on Load Balancing Algorithms in Cloud Computing

N.Yugesh Kumar, K.Tulasi, R.Kavitha

Siddhartha Institute of Engineering and Technology

ABSTRACT

As there is a rapid growth in internet usage by users there is huge increase in data generation. These users put requests either for the hardware or software applications or services or resources which cloud computing offers. The resources could be data storage, processing a task, bandwidth etc. Based on the cloud infrastructure and requests of the users, the cloud be either overloaded, under loaded or in a stable state. The cloud system could be failed if it is either in overloaded or under loaded states thus causing failure in energy utilization, response time, reliability etc. To handle such situation load balancing is needed. Load Balancing plays an important role in detecting and balancing loads of the cloud system. There are several load balancing techniques exists to optimize the performance of the cloud system. In this paper, we have presented a brief explanation on the existing algorithms used along with the metrics.

Keywords: Cloud computing, load, load balancing, metrics, overloaded, response time, task allocation, task migration

1. INTRODUCTION

Due to the advancement in the communication technology, cloud computing is showing a huge growth. Cloud computing is a technology where in it shares various resources like networks, applications, servers, storage, services and information to the users upon their requests[1]. It has become a utility for the IT industries. The cloud computing delivers three services SaaS, PaaS and IaaS. The efficiency of cloud computing can be obtained if its resources are properly managed. One of the most important features of cloud system is that its resources are in virtual form. The users of the cloud receive these services or resources from Cloud Service Provider (CSP) on rent. Allocating the available resources to the users is the major challenge to the CSP.

Thus load balancing has a greater impact on the system performance. Load balancing divides the workload equally among the physical machine (PM) or virtual machine (VM) (virtual machines are part of physical machine).

Allocation of various requests of the users (also called as tasks) to different VMs is known as load. Load balancing can be defined in following terms: task allocation and task or virtual machine migration. Task Allocation: a finite number of tasks are allocated to different PMs which in turn are allocated to VMs of respective PMs [2][4].Task/Virtual Machine Migration : if a VM of one PM overloaded then the task from it is moved to another VM of another PM.

Many of the cloud computing load balancing algorithms is reviewed and we are going to present them in brief in this paper. In section 2, we discuss the model of load balancing. In section 3, we present the metrics which are used by different algorithms for load balancing. Classification of load balancing techniques is presented in section 4. Finally we conclude in section 5.

2. MODEL OF LOAD BALANCING

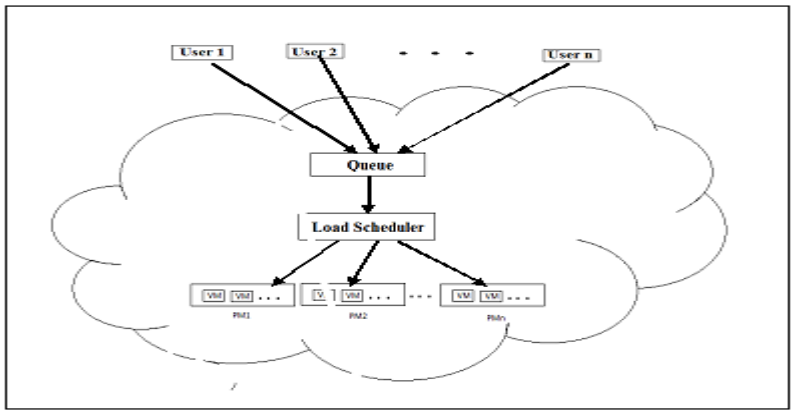

Fig 1. Model of Load Balancing

It can be observed from Fig 1, that the user put the requests (tasks) to the cloud server. These tasks are heterogeneous. As the number of tasks is high they are queued up. The load scheduler, has complete information about the availability of resources, picks these requests from the queue and allocates them to different VMs of respective PM.

3. PERFORMANCE METRICS

The metrics which affect the load balancing in cloud computing system are explained as follows:

Response Time : It is defined as the time taken for a cloud system to give reply back to the user, that is it is the sum total of transmission time, waiting time and execution time of a task which is put as a request by the user. The proposed algorithm in [3] minimizes the response time while balancing the load of the system to satisfy the users.

Power Consumption / Energy Utilization: It is the amount of power consumed by the ICT devices which are connected to the cloud system [9]. The devices that are considered are desktops, laptops, routers, switches, servers which are local. In [8], how energy can be conserved is discussed.

Throughput: It refers to the total number of tasks executed by VM in one unit of time. Cloud System performance is measured using this metric. The system performance is good indicates that the throughput is high.

Reliability: If the system is performing accordingly with the requirements in all situations then it is said that the system is reliable. If a failure occurs during the execution of a task, then it is relocated to other VM so that the reliability of the system can be improved. If the system is stable it means that it is reliable.

Accuracy: It refers to the matching of correct value with the given specification. Many of the IT-companies give significance to system accuracy.

Scalability: it is the ability of the system to perform correctly even after r**escaling** the resources. It means that the system will be stable irrespective of the number of tasks assigned to it.

Migration Time: It is the time taken to migrate or move a task or virtual machine from one resource to another resource. Task migration can be from one VM to another VM of same PM or to different PM. If there is interrupt in the execution of the task due to unavailable of

resources in one VM then it is migrated. Similarly when a VM is crashed during execution, then it is migrated to another PM. Higher number of migrations degrade the load balancing of the cloud system.

Associated Overhead: The balancing methods of the cloud system will result in some overhead. For overloaded or under loaded systems this cost would be more. But for the system which is stable will have less overhead. In [23], the algorithm presented concludes that if the system is properly balanced with the load the there will be minimum overhead.

Makespan: It is defined as the total time needed for the completion of the tasks that are submitted to the cloud system. The proposed algorithm of [22] uses this metric for optimization.

Fault-Tolerance: It is the ability of the system to complete the required tasks even if one or more elements fail. Switching algorithms use this metric.

4. CLASSIFICATION OF LOAD BALANCING ALGORITHMS

Load balancing algorithms are classified into two types: static and dynamic. In static type, load scheduler doesn't have prior information about the load on the PMs, so it fairly allocates the tasks to satisfy the users. In dynamic type, load scheduler will have complete information about the loads on all PMs and VMs and allocates the tasks accordingly. Managing of resources plays an important role in load balancing [3].

Further the algorithms that are already proposed under static approach are OLB, MET, MCT, GA, SA, TABU,A*,MIN-MIN, MIN-MAX. Dynamic approach is further divided as offline and online modes. Algorithms under offline mode are MIN-MIN, MIN-MAX, SUFFERAGE and online mode is OLB, MET, MCT, SA.

Scheduling or balancing the load in cloud computing is a NP-Complete Problem [3][7]. The objective these load balancing algorithms is to minimize response time, save energy, maximize throughput, improve reliability etc.

Opportunistic Load Balancing (OLB): this method is used in both static and dynamic approaches. Here the task is randomly allocated to VM and finds the next idle machine to allocate the next task.

Minimum Execution Time (MET): It is also called as Limited Best Assignment in [11] or User Directed Assignment [12] . This method is used in both static and dynamic approaches. In [13] this algorithm describes the allocation of each task to VM based on lowest execution time of VM such that all the tasks get completed within their execution time.

Minimum Compilation Time (MCT): This method is used in both static and dynamic approaches. The authors of [14] have utilized ready to execute and expected execution times such that the tasks are balanced. In this, the task is allocated to the centre which has least completion time.

Min-Min : This algorithm chooses a task with minimum execution time and allocates it to a VM which has minimum capability. Once the task is allocated to a VM, it is removed from the queue and the process continues in distributing the remaining tasks which are waiting in the queue. The authors of [22], proposed an improved version of this algorithm which optimizes the makespan and increases the maximum utilization of the resources.

Min-Max: This algorithm is very much similar to Min-Min algorithm except that the larger task is selected. For small-scale Distributed system this algorithm is much suited [16]. This algorithm is further extended and is called as Elastic Cloud Max-Min[10], which considers the tasks with average awaiting time. This algorithm performs better than round robin method.

Genetic Algorithm (GA): this algorithm considers energy consumption, throughput, makespan values for optimization. This algorithm has three phases: selection, crossover and mutation. All these phases are executed under every iteration. In [17], the authors have proposed an algorithm which is based on GA for minimizing makespan.

Tabu Search (TS): The algorithm in [18] uses adaptive memory for searching idleness of VMs. This method is used in [19] to place cloud servers in different location for efficient utilization of the resource.

A* Search: It is a graphic search algorithm, which combines the advantages of DFS and BFS algorithm. It maintains two lists, one as priority queue for the tasks and second contains the processing capacities of VMs. The algorithm of [20] used A* along with fuzzy method to enhance the network life time.

Switching Algorithm: The algorithm of [21] helps in task or VM migration in cloud computing environment. This algorithm enables to achieve fault-tolerance.

5. CONCLUSION

In this paper, we have presented different load balancing algorithms of cloud computing system. A model for load balancing is described. We have discussed load balancing approaches along with the existing algorithms. We have also explained various metrics used by different algorithms to improve the performance of the system. The paper will help the researchers to use different algorithms and extend them to achieve more efficient results.

REFERENCES

[1] Bohn,R.B., Messina.J, Liu.F., Ton J and Mao J (2011) NIST Cloud Computing Reference Architecture, IEEE world Congress on Services, Washington, DC, pp 594- 596.

[2] Mishra, S. K., Puthal, D., Sahoo, B., Jena, S. K., and Obaidat, M. S.(2017) 'An adaptive task allocation technique for green cloud computing', The Journal of Supercomputing, pp. 1-16.

[3] Li, K., Xu, G., Zhao, G., Dong, Y. and Wang, D. (2011) 'Cloud Task Scheduling Based on Load Balancing Ant Colony Optimization', Sixth Annual China grid Conference, Liaoning, pp. 3–9.

[4] Ibrahim, A. H., Faheem, H. E. D. M., Mahdy, Y. B., and Hedar, A. R. (2016) 'Resource allocation algorithm for GPUs in a private cloud', International Journal of Cloud Computing, 5(1-2), pp. 45-56.

[5] Wang, S.C., Yan, K. Q., Liao, W. P. and Wang, S. S. (2010) 'Towards a Load Balancing in a three-level cloud computing network', 3rd IEEE International Conference on Computer Science and Information Technology (ICCSIT), Chengdu, pp. 108–113.

[6] Al-Ayyoub, M., Daraghmeh, M., Jararweh, Y., and Althebyan, Q. (2016) 'Towards improving resource management in cloud systems using a multi agent framework', International Journal of Cloud Computing, 5(1-2), pp.112-133.

[7] Dam, S., Mandal, G., Dasgupta, K., and Dutta, P. (2015, February) 'Genetic algorithm and gravitational emulation based hybrid load balancing strategy in cloud computing', Third IEEE International Conference on Computer, Communication, Control and Information Technology (C3IT), pp. 1–7.

[8] Berl, A., Gelenbe, E., Di Girolamo, M., Giuliani, 520 G., De Meer, H., Dang, M. Q., and Pentikousis, K. (2010) 'Energy-efficient cloud computing', The computer journal, 53(7), pp. 1045–1051.

[9] Moganarangan, N., Babukarthik, R. G., Bhuvaneswari, S., Basha, M. S., and Dhavachelvan, P. (2016) 'A novel algorithm for reducing energy consumption in cloud computing environment: Web service computing approach', Journal of King Saud University-Computer and Information Sciences, 28(1), pp. 55-67.

[10] Li, X., Mao, Y., Xiao, X., and Zhuang, Y. (2014, June) 'An improved maxmin task scheduling algorithm for elastic cloud', In IEEE International Symposium on Computer, Consumer and Control (IS3C), pp. 340-343.

[11] Kanakala, V., RaviTeja, V., Reddy, K. and Karthik, K. (2015) 'Performance analysis of load balancing techniques in cloud computing environment', IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp. 1–6.

[12] Armstrong, R., Hensgen, D. and Kidd, T. (1998) 'The relative performance of various mapping algorithms is independent of sizable variances in run-time predications', 7th IEEE Heterogeneous Computing Workshop (HCW98), Mar. 1998, pp. 79-87.

[13] Maheswaran, M., Ali, S., Siegal, H. J., Hensgen D. and Freund, R. F. (1999) 'Dynamic matching and scheduling of a class of independent tasks onto heterogeneous computing systems', Eighth Heterogeneous Computing Workshop, 1999. (HCW '99) Proceedings, San Juan, pp. 30–44.

[14] Kim, S. I., Kim, H. T., Kang, G. S., and Kim, J. K. (2013, June) 'Using dvfs and task scheduling algorithms for a hard real-time heterogeneous multicore processor environment', In Proceedings of the 2013 workshop on Energy efficient high performance parallel and distributed computing, ACM, pp. 23-30.

[15] Chen, H., Wang, F., Helian, N., and Akanmu, G. (2013, February) 'User priority guided Min-Min scheduling algorithm for load balancing in cloud computing', In IEEE National Conference on Parallel Computing Technologies (PARCOMPTECH), pp. 1-8.

[16] Kokilavani, T., and Amalarethinam, D. D. G. (2011) 'Load balanced min-min algorithm for static meta-task scheduling in grid computing', International Journal of Computer Applications, 20(2), pp. 43-49.

[17] Dasgupta, K., Mandal, B., Dutta, P., Mandal, J. K., and Dam, S. (2013) 'A genetic algorithm (ga) based load balancing strategy for cloud computing', Procedia Technology, 10, pp. 340-347.

[18] Glover, F., and Laguna, M. (2013) 'Tabu Search', Springer New York, pp.3261-3362.

[19] Larumbe, F., and Sanso, B. (2013) 'A tabu search algorithm for the location of data centers and software components in green cloud computing networks', IEEE Transactions on Cloud Computing, 1(1), pp. 22-35.

[20] AlShawi, I. S., Yan, L., Pan, W., and Luo, B. (2012) 'Lifetime enhancement in wireless sensor networks using fuzzy approach and A-star algorithm', IEEE Sensors 575 journal, 12(10), pp. 3010-3018.

[21] Shao, S., Guo, S., Qiu, X., and Meng, L. (2014, August) 'A random switching traffic scheduling algorithm in wireless smart grid communication network', In 2014 23rd International Conference on Computer Communication and Networks (ICCCN), IEEE, pp. 1-6.

[22] Chen, H., Wang, F., Helian, N., and Akanmu, G. (2013, February) 'User priority guided Min-Min scheduling algorithm for load balancing in cloud computing', In IEEE National Conference on Parallel Computing Tech nologies (PARCOMPTECH), pp. 1-8.

[23] Singh, A., Juneja, D., and Malhotra, M. (2015) 'A novel agent based autonomous and service composition framework for cost optimization of resource provisioning in cloud computing', Journal of King Saud University Computer and Information Sciences.