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

Kundan Pagar[1], Sachin Patil[2]

[1, 2]GHRCEM, Savitribai Phule Pune University, India

Abstract: ***In cloud computing, load balancing Is one of the key issues. Load Is a measure of the amount of work that a computation system performs which can be classified as CPU load, memory capacity and network load. Load balancing is the process of apportioning the load among various nodes of a distributed system to improve both resource utilization and job response time. Load balancing ensures that every node in the network does approximately equal amount of work (as per the its capacity) at any instant of time. This paper discusses the existing load balancing algorithms in a cloud based environment. Load Balancing is an important aspect of cloud computing environment.***

Keywords: Cloud computing, Load balancing.

1. Introduction

The extraordinary addition of cloud computing, built on the fixed research fields of distributed computing, web services, networks, utility computing and virtualization. In general, the cloud comprises three main components: clients, data centers and distributed servers.

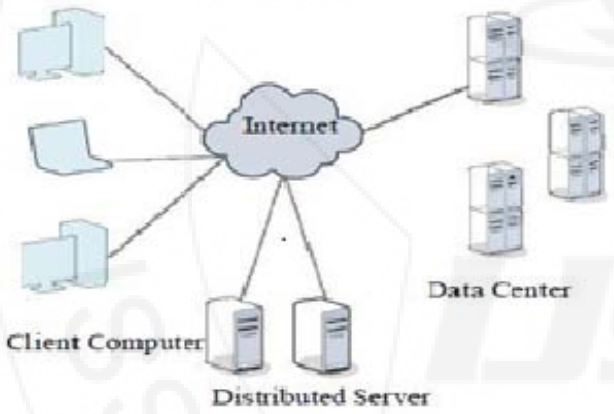

Figure 1: Components of cloud computing solution

Data center is defined as the collection of servers hosting different applications, whereas distributed servers are the elements of a cloud that are present on internet hosting different applications .Moreover, services provided by cloud computing are categorized into 3 major types, which are as follows:

A) IaaS (Infrastructure as a Service): With IaaS, the components of infrastructure layer such as computation power as well as storage resources can be rented from the virtual resource pool for the entire industry.

B) PaaS (Platform as a Service): Platform layer completed the higher level of abstraction with IaaS base. This affords the development environment, test environment, server platforms and other services.

C) SaaS (Software as a Service): SaaS is declared as a software distribution model, which can be accessed by the user through the internet hosting. It is necessary that the providers have to develop information for all infrastructures, software, hardware and operating systems. It is also important to offer post maintenance and other services.

1.1 Cloud Perspectives

Cloud has different meaning to different stakeholders. There are three main stakeholders of cloud:

Type of stakeholder	*Requirement/Issues*
End User	Security Provence Privacy High availability Reduced Cost Ease-of-use
Cloud Provider	Resource Utilization Energy Efficiency Cost Efficiency Utility Computing Providing Resources
Cloud Developer	Elasticity/Scalability Virtualization availability Reliability Programmability

1.1.1. End users

These are the customers or consumers of cloud. They use a variety of services (Infrastructure/ Software/Platform) provided by the cloud. Before using the cloud services, the users of cloud must distribution model, which can consent to the Service Level Agreement (SLA) precise by the Cloud Provider.

1.1.2. Cloud Provider

Cloud provider can offer either public or private or hybrid cloud. They are responsible for constructions of the cloud Private clouds [4] are owned by enterprises or business for their internal use. They may use it to store and handle Big-Data of their organization or to afford sufficient resources on demand basis to its team of employees or clients. They offer greatest level of security.

1.1.3. Cloud Developer

This entity lies between end user and cloud provider. Cloud developer has the liability of taking into consideration both the perspectives of the cloud (i.e. view of end user and cloud provider). The developer of cloud must remain to all the technical facts of the cloud which are necessary to meet the

requirements of both, the cloud user as well as the cloud provider.

In Introduction you can mention the introduction about your research.

2. Load Balancing in Cloud Computing Environment

Load balancing in cloud computing provides an efficient resolution to a variety of issues residing in cloud computing environment set-up and usage. Load balancing must take into account two main tasks, one is the resource provisioning or resource allocation and other is task scheduling in distributed environment. Efficient provisioning of resources and scheduling of resources as well as tasks will ensure:

a. Resources are easily available on demand.

b. Resources are competently utilized under condition of high/low load.

c. Power is saved in case of low load (i.e. when usage of cloud resources is less than certain threshold).

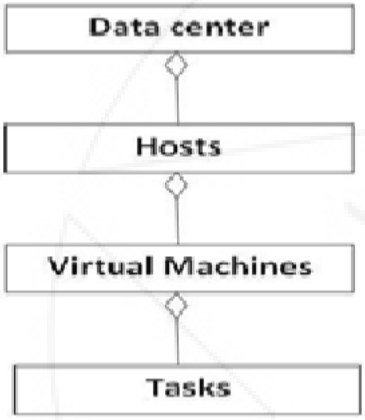

Figure 2: Class diagram of cloud

3. Literature Survey

[4] Proposed a load balancing mechanism based on ant dependency optimization in an open cloud computing federation. This system overcomes heterogeneity, is adaptive to dynamic environments, is tremendous in fault tolerance and has good scalability hence helps in improving the performance of the system. It uses small-world and scale-free characteristics of a complex network to get improved load balancing.

[5] Proposed a lock-free multiprocessing load balancing solution that avoids the use of shared memory in difference to other multiprocessing load balancing solutions which use shared memory and lock to maintain a user session.

[7] Presented an event-driven load balancing algorithm for real time particularly multiplayer online games (MMOG). This algorithm subsequent to receiving capacity events as input, analyzes its components in context of the resources and the global state of the game session, thus generating the game assembly load balancing actions.

[9] Investigated a decentralized honeybee-based load balancing method that is a nature-inspired algorithm for self-organization. It achieves universal load balancing during local server actions.

4. Proposed Solution Load Balancing Algorithm

Cloud is made up of huge resources. Management of these resources requires proficient planning and appropriate design. While designing an algorithm for resource provisioning on cloud the developer must take into consideration special cloud scenarios and must be conscious of the issues that are to be determined by the proposed algorithm. Therefore, resource provisioning algorithm can be categorized into different classes based upon the environment, purpose and method of proposed solution.

4.1. Load Balancing on the basis of cloud Environment

Cloud computing can have either static or dynamic environment based upon how developer configures the cloud demanded by the cloud supplier.

4.1.1 Static Algorithm

In static algorithm the traffic is separated evenly among the servers. This algorithm requires a previous knowledge of system resources, so that the choice of shifting of the load does not depend on the present state of system. Static algorithm is proper in the system which has low discrepancy in load.

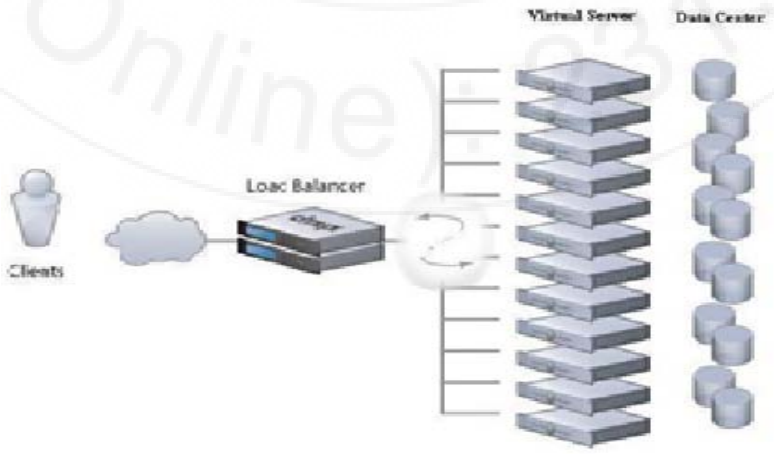

Figure 3: Load balancing in cloud computing

4.1.2 Dynamic Algorithm

In dynamic algorithm the lightest server in the whole network or system is searched and preferred for balancing a load. For this real time communication with network is necessary which can increase the traffic in the system. Here present state of the system is used to make decisions to handle the load.

5. Challenges for Load Balancing

There are some qualitative metrics that can be enhanced for improved load balancing in cloud computing.

- **Throughput:** It is the total number of tasks that have finished execution for a given scale of time. It is necessary to have high through put for better performance of the system.
- **Associated Overhead:** It describes the amount of overhead during the execution of the load balancing algorithm. It is a composition of progress of tasks, inter process communication and inter processor. For load balancing technique to work properly, lowest overhead should be there.
- **Fault tolerant:** We can define it as the capability to execute load balancing by the suitable algorithm without arbitrary link or node failure. Every load balancing algorithm should have good fault tolerance approach.
- **Migration time:** It is the amount of time for a process to be transferred from one system node to another node for execution. For better performance of the system this time should be always less.
- **Response time:** In Distributed system, it is the time taken by a particular load balancing method to respond. This time should be minimized for better performance.
- **Resource Utilization:** It is the parameter which gives the information within which present the resource is utilized. For efficient load balancing in system, optimal resource should be utilized.
- **Scalability:** It is the ability of load balancing algorithm for a system with any finite number of processor and machines. This parameter can be improved for better system performance.
- **Performance:** It is the overall efficiency of the system. If all the parameters are enhanced then the overall system performance can be improved.

6. Conclusion

In this paper, surveyed different load balancing techniques for cloud computing. The key function of load balancing is to satisfy the customer requirement by distributing load dynamically along with the nodes as well as to make highest resource utilization by reassigning the total load to individual node. Existing load balancing techniques that have been analyzed in this paper generally focused on minimizing service response time and overhead. With that concern, in future development, the requirement of load balancing in cloud is efficiently considered for improving the network performance and scalability with optimized resource utilization. In this paper, discussed various load balancing schemes. Static load balancing provide easiest recreation and monitoring of environment. Dynamic load balancing algorithms are complex to simulate.

References

[1] Yunhua Deng And Rynson W.H. Lau,senior member IEEE,"On Delay adjustment for Dynamic load balancing in distributed virtual environment", April 2012.
[2] Bin Dong, Xiuqiao Li, Qimeng Wu, Limin Xiao and Li Ruan, "A dynamic and adaptive load balancing strategy for parallel file system with large-scale I/O servers," Journal of Parallel Distrib. Comput. 72, pp. 1254–1268, 2012.
[3] Pragati Priyadarshinee and Pragya Jain,"Load Balancing and Parallelism in Cloud Computing," International Journal of Engineering and Advanced Technology, Vol. 1, Issue. 5, 2012, pp. 486-489.
[4] Bhushan Lal Sahu, Rajesh Tiwari, "A comprehensive study on Cloud computing", International journal of Advanced Research in Computer science and Software engineering, volume 2, issue 9, September 2012, ISSN: 2277 128X.
[5] Ratan Mishra, Anant jaiswal, Ant colony optimization: A Solution of load balancing in cloud, International J1ournal of Web & Semantic Technology (IJWesT) Vol.3, No.2, April 2012.
[6] D. Escalnte, Andrew J. Korty, "Cloud Services: Policy and Assessment", Educause review July/August 2011.
[7] Liu Xi., Pan Lei.,Wang Chong-Jun. and Xie Jun-Yuan (2011) 3rd International Workshop on Intelligent Systems and Appli- cations, 1-4.
[8] Zhang, Q., Cheng, L. & Boutaba, R. (2010) Cloud computing: State-of-the-art and research challenges, Journal of Internet Services and Applications, 1(1), 7-18. DOI 10.1007/s13174-010-0007-6.
[9] Nae V., Prodan R. and Fahringer T. (2010) 11th IEEE/ACM International Conference on Grid Computing (Grid), 9-17.
[10] Randles M., Lamb D. and Taleb-Bendiab A. (2010) 24th Inter- national Conference on Advanced Information Networking and Applications Workshops, 551-556.
[11] 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.
[12] Calheiros, R. N., Ranjan, R., Beloglazov, A., Rose, C. A. F. D. & Buyya, R. CloudSim: A Toolkit for Modeling and Simulation of Cloud Computing Environments and Evaluation of Resource Provisioning Algorithms. extended version of a keynote paper: R. Buyya, R. Ranjan, and R. N. Calheiros. Modeling and simulation of scalable Cloud computing environments and the CloudSim toolkit: Challenges and opportunities. Proceedings of the Conference on High Performance Computing and Simulation (HPCS 2009) (pp. 21-24). IEEE Press, New York, USA, Leipzig, Germany, June, 2009.
[13] L. Wang, J. Tao, M. Kunze,"Scientific Cloud Computing: Early Definition and Experience", the 10th IEEE International Conference Computing and Communications 2008.

[14] 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.

Author Profile

Kundan B. Pagar is Post-Graduation student pursuing M.E in Computer Network from G.H.Raisoni College of Engineering, Pune. He received his Bachelor of Engineering degree from Imperial College of Engineering and Research, Pune affiliated to Savitribai Phule Pune University in 2012. His research interests lie in the area of Load balancing in cloud computing.