Multi Distributed Server Load Balancing for SaaS: With Parameterized CPU and Memory Statistics

Frenny Swanisha Pinto[1], Annapa Swamy D. R.[2]

[1]Student, Department of Computer Science and Engineering,
Mangalore Institute of Technology and Engineering, Moodabidre -574225, Mangaluru, Karnataka, India

[2]Associate Professor, Department of Computer Science and Engineering,
Mangalore Institute of Technology and Engineering, Moodabidre -574225, Mangaluru, Karnataka, India

Abstract: ***The term load balancing refers to distribution of load among the various nodes in order to improve the job response time, resource utilization and avoid the scenario where some node doing more work and some are lightly loaded. The load balancing facilitates every node to perform the equal amount of work at a time. The proposed work shows the load balancing, where the clients request is distributed among the servers located at different geographical locations. Here the load balancing problem is solved by considering the statistics such as CPU, memory and location details. In this system initially, the user request will be sent to the server having the less CPU and memory usage; if the servers are equally loaded then the request is sent to nearest node with respect to user location. Thus load is moved from the heavily loaded node to lightly loaded node.***

Keywords: Cloud Computing, Load Balancing, Controller, CPU, Memory

1. Introduction

Cloud computing is the most developing technology in the world of Internet computing and it has become more popular day by day as it has brought several changes in the IT industry. The cloud computing provides the user with the new set of services, without paying the attention to how the data is processed and the storage details. The cloud environment is unaware of the arrival pattern of the job as well as the cloud node capacity, these facts leads to the complexity in balancing the load on cloud environment [1].The name cloud computing because the information is used, is retrieved from the cloud storage and the user can access it from anywhere in the world without being in the specific location. The computing involves networking, web services, virtualization, distributed computing, software's and these consist of storage, scalability, computing/processing power, platforms and services which are provides to the user based on the pay per use basis. Even though cloud provide large usage there are several problems involved here such as the load balancing, security, virtual machine migration, server integration and energy management[2].

Load balancing refers to distribution of incoming load/traffic among the multiple computing resources in order to improve the system performance. Proper resource utilization and load balancing avoids failure and bottlenecks. Load balancing classified into two categories first static approach suitable of homogeneous set of resources and are not flexible. Here before the execution begins the system need to make the decision regarding the task assignment. Second, dynamic approach suitable for flexible homogeneous set of resources and adapt runtime changes based on the load [3].

The goals of load balancing include Cost Effectiveness, scalability, flexibility, fault tolerance [4]. In this paper the load balancing is performed on the multi located servers based on the CPU and Memory Statistics.

2. Literature Survey

Snehal D, et.al. [5] proposed the efficient technique for load balancing. Here the cloud is divided into the several partitions and these partitions simplify the load balancing. The controller chooses the suitable partition for the job and the balancer at each cloud partition selects the load balancing strategy.

Rupam Mukhopadhyay, et.al. [6] proposed the three approaches for dynamic load balancing which is based on the basic policies such as transfer policy, location policy, selection policy and information policy. First approach is sender initiated algorithm, here load distribution is initiated by the overloaded host, and this algorithm differs only in location policy. In receiver initiated algorithm distribution of load is initiated from the under loaded receiver, which gets the task from overloaded host. In symmetrically-initiated algorithm both the sender and receiver initiate the distribution of load activity for transferring the task. This algorithm is combined form of the receiver and sender initiated algorithm. Whereas the receiver initiated are component are useful at finding the overloaded hosts and this algorithm require preemptive task transfer facility. Pragathi M, et.al. [7] proposed method for workload balancing and resource monitoring to improve the system performance as well the user satisfaction. Here the switch mechanism is used along with the cloud partitioning technique for better workload balance. The switch mechanism is used to choose the different strategies at various situations. The round robin algorithm is used when the nodes are in idle state. The game theory is used for optimal balancing the structure of Nash equilibrium which is used to select the correct node for workload balance.

Nusrat Pasha, Dr. Amit Agarwal, et.al. [8] proposed the virtualization technique for providing the services to the end user on the internet. In virtual machine load balancing, the limitation was that it did not save the state of virtual machine allocation. Since the user request and the virtual machine load balancing require the execution each time when the new request is arrived from the user. These are resolved by developing the virtual machine load balancing algorithm for round robin approach and it improved the system performance by reducing the time to schedule the virtual machines.

Nidhi Jain Kansal, et.al. [9] introduced the energy management technique based on the energy consumption and carbon emission perspective, which improves the energy- efficiency in the cloud environment.

Nitin Mishra, et.al. [10] have analyzed the issues as well as performed the comparison in existing load balancing algorithms based on the qualitative metrics such as reliability, performance, scalability, throughput, overhead associated, power saving features, etc. They have done work by exploring efficient load balancing algorithms to maintain better balance among the parameters this helps in green computing.

Illa Pavan, et.al.[11] proposed the scalable infrastructure for cloud environment. The idle load balancing architecture handles entire infrastructure and has the following characteristics. Increase in resources result in the proportional growth in the performance. Here the scalable service must be able to handle the heterogeneity, operationally efficient, operationally efficient and should become more cost effective as it grows.

Shilpa D. More, et.al.[12] proposed the method for balancing the load on the cloud which will improve the performance substantially and prevents the server overloading. This software used for efficient data storage on the cloud and dynamically allocates the data to the least load server, thus the cloud service performance is not affected. The software aims at backup plan even when the system fails partially and also provides facility to accommodate future system modifications.

S. Hemachander, et.al.[13] propose the game theory model on cloud partitioning for the public cloud. Here the public cloud is divided into the cloud partitions and different strategies is applied for load balancing, avoid the overloading of servers and improves the response time.

3. Problem Definition

In the modern days several organizations host their websites and the application into the cloud environment. There are several users who will access these websites and the application which introduce the great traffic which lead to the server downtime. Therefore it is necessary to manage the traffic across the cloud environment so that it cannot degrade the system performance. The load balancing in this situation help to distribute the traffic or work into different servers consisting of several computing resources allocated by the cloud environment. This leads in cost reduction and maximum resource availability. In the project the servers are located at different geographical regions and the load balancing is performed based on the CPU and Memory statistics of the servers. The servers which having the lowest CPU and Memory values will be assigned with the task.

4. Proposed Work

The figure below shows the architecture used for the entire project work.

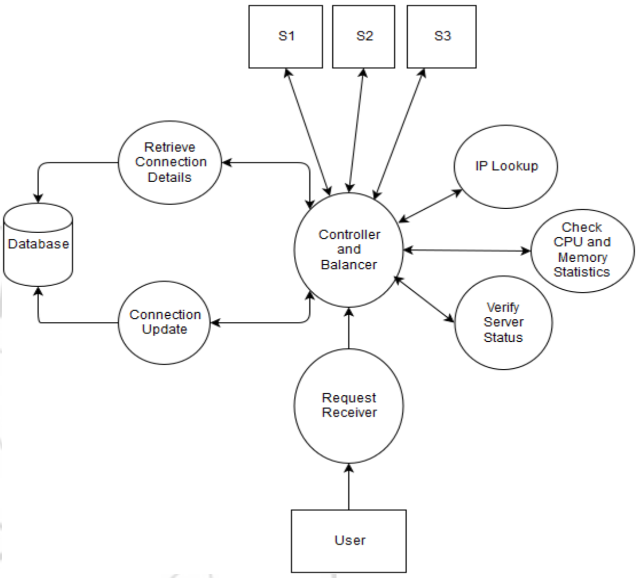

Figure 1: System Architecture

As shown in the figure the client application will be running on three cloud servers namely s1, s2, s3, and are located at different geographical regions. The controller here maintains the details such as location information, dynamic parameters like CPU and Memory Utilization. When the user requests to access the application, the request will be sent to the controller, which performs IP Lookup function to determine which location the user belongs too. Then controller determines the CPU, Memory statistics of every server and then finds the server having the less weight. Based on these parameters the server having the minimum weight will be assign with the user request. If the server load is exceed above the threshold value then the requests are sent to the other server having the less load. In case if two servers having the same load then the server which is near to the user will be allocated with the request.

5. Conclusion

In this paper, the load balancing method performed on multi located cloud server having the running client application. The server having the minimum CPU and Memory will assign with the client request. The load/job is dynamically allocated to the server based on their load status.

6. Future Scope

In future the load balancing can be done by considering other statistics such as, storage, network bandwidth, providing security to the user requests and the server responses. This can be done by encrypting the requests/responses and converting it into meaningless form so that it cannot understandable by the cryptanalyst. And the decryption of the request/responses back to the original form.

References

[1] Gaochao Xu, Junjie Pang, and Xiaodong Fu, "A Load Balancing Model Based on Cloud Partitioning for the Public Cloud", Volume 18, February 2013

[2] Amandeep, Vandana Yadav, Faz Mohammad, "Different Strategies for Load Balancing in Cloud Computing Environment: a critical Study", International Journal of Scientific Research Engineering & Technology (IJSRET), Volume 3, April 2014.

[3] Rajwinder Kaur and Pawan Luthra, "Load Balancing in Cloud Computing", Proc. of Int. Conf. on Recent Trends in Information, Telecommunication and Computing, ITC.

[4] Shrikant M. Lanjewar, Susmit S. Surwade, Sachin P. Patil, Pratik S. Ghumatkar, "Load Balancing In Public Cloud", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Ver. VI (Feb. 2014).

[5] Snehal D. Sonawane and R. H.Borhade, "Load Distribution and Balancing over Cloud using Cloud Partitioning", International Journal of Current Engineering and Technology, E-ISSN 2277 – 4106, P-ISSN 2347 - 5161.

[6] Rupam Mukhopadhyay, Dibyajoyti Ghosh, Nandini Mukherjee, "A Study on the Application of Existing Load Balancing Algorithms for Large, Dynamic, Heterogeneous Distributed Systems", ISSN: 1790-5117, ISBN: 978-960-474-156-4.

[7] Pragathi M, Swapna Addamani, Venkata Ravana Nayak, "Resource Monitoring and Workload Balancing Model for Public Cloud", International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

[8] Nusrat Pasha, Dr. Amit Agarwal, "Round Robin Approach for VM Load Balancing Algorithm in Cloud Computing Environment", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, May 2014.

[9] Nidhi Jain Kansal, Inderveer Chana, "Cloud Load Balancing Techniques : A Step Towards Green Computing", IJCSI International Journal of Computer Science Issues, Vol. 9, January 2012.

[10] Nitin Kumar Mishra, Nishchol Mishra, "Load Balancing Techniques: Need, Objectives and Major Challenges in Cloud Computing- A Systematic Review", International Journal of Computer Applications, No.18, December 2015.

[11] Illa Pavan Kumar, Subrahmanyam Kodukula, "A Generalized Framework for Building Scalable Load Balancing Architectures in the Cloud", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 3 (1), 2012,

[12] Ms.Shilpa D.More, Mrs.Smita Chaudhari, "Reviews of Load Balancing Based on Partitioning in Cloud Computing", International Journal of Computer Science and Information Technologies, Vol. 5 (3), 2014.

[13] S.Hemachander, R.Backiyalakshmi, "A Game Theory Modal Based On Cloud Computing For Public Cloud", IOSR Journal of Computer Engineering (IOSR-JCE), Volume 16, Issue 2, Ver. XII (Mar-Apr. 2014).

Author Profile

Ms. Frenny Swanisha Pinto completed the Bachelor's Degree in Computer Science & Engineering from Visvesvaraya Technological University (VTU). Currently pursuing M. Tech degree in Computer Science & Engineering at Mangalore Institute of Technology, Mangalore

Mr. Annappa Swamy D.R received Master Degree in Computer Science & Engineering. He is currently working as Associate Professor in the department of Computer Science & Engineering, Mangalore Institute of Technology and Engineering, Mangalore. Hehas worked in the area of Data Centre Management, Networking which include Infrastructure Management-Hardware/Software, Planning, Design & Implementation of networking facility, user co-ordination, system study, project planning, project handling, preparation of functional & technical specifications, technical evaluation of bids, procurement activity, vendor management. He has attended various workshops, trainings and other corporate learning programmes.