Load Balancing in Cloud Computing

Siddharth Sonawane, Prathmesh Arnikar, Ankita Fale, Sagar Aghav, Shikha Pachouly
AISSMS COE
Computer Department, India

Abstract—"Cloud computing" is a term, which involves virtualization, distributed computing, networking, software and web services. A cloud consists of several elements such as clients, datacenter and distributed servers. It includes fault tolerance, high availability, scalability, flexibility, reduced overhead for users, reduced cost of ownership, on demand services etc. Central to these issues lays the establishment of an effective load balancing algorithm. The load can be CPU load, memory capacity, delay or network load. Load balancing is the process of distributing the load among various nodes of a distributed system to improve both resource utilization and job response time while also avoiding a situation where some of the nodes are heavily loaded while other nodes are idle or doing very little work. Load balancing ensures that all the processor in the system or every node in the network does approximately the equal amount of work at any instant of time.

Keywords— Cloud ,Load etc.

I. INTRODUCTION

The random arrival of load in cloud environment can cause some server to be heavily loaded while other server is idle or only lightly loaded. Equally load distributing improves performance by transferring load from heavily loaded server. Efficient scheduling and resource allocation is a critical characteristic of cloud computing based on which the performance of the system is estimated. The considered characteristics have an impact on cost optimization, which can be obtained by improved response time and processing time. To handle the random selection based load distributed problem, we have proposed a scheduling algorithm to estimate response time, processing time, which is having an impact on cost.

II. PROPOSED SYSTEM

The load balancing model given in this article is aimed at the public cloud which has numerous nodes with distributed computing resources in many different geographic locations. Thus, this model divides the public cloud into several cloud partitions. When the environment is very large and complex, these divisions simplify the load balancing. The cloud has a main controller that chooses the suitable partitions for arriving jobs while the balancer for each cloud partition chooses the best load balancing strategy.

III. RELATED WORK

Cloud computing is efficient and scalable but maintaining the stability of processing so many jobs in the cloud computing environment is a very complex problem with load balancing receiving much attention for researchers. Since the job arrival pattern is not predictable and the capacities of each node in the cloud differ, for load balancing problem, workload control is. Crucial to improve system performance and maintain stability. Load balancing schemes depending on whether the system dynamics are important can be either static or dynamic. Static schemes do not use the system information and are less complex while dynamic schemes will bring additional costs for the system but can change as the system status changes. A dynamic scheme is used here for its flexibility.

IV. SYSTEM OBJECTIVES

Features available in proposed system are as follows:

1. Server balances the load by using external resources.
2. Ensures proper utilization of resources.
3. Architecture is implemented in private cloud.
4. Improved response time and effective job processing.
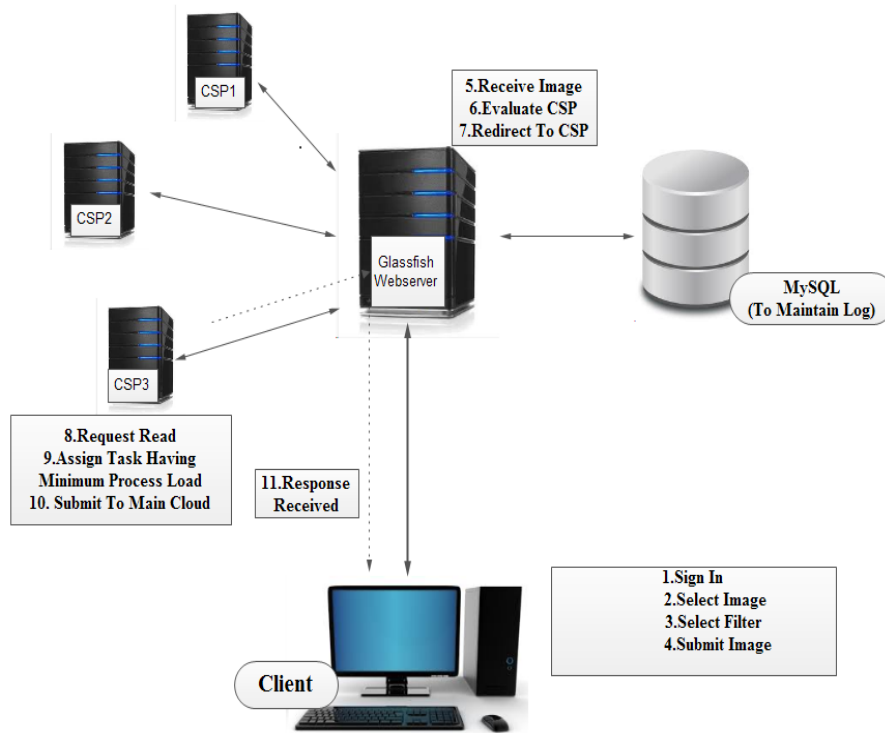5. Ensures network reliability.

Fig 1: Architecture of the proposed system

V. **APPLICATIONS**

1.This Architecture can be implemented in the fields of cloud computing where huge load is created on server.
2.Also to handle the random selection based load distribution problem.

VI. **CONCLUSION**

Cost and time are the key challenge of every IT engineer to develop products that can enhance the business performance in the cloud based IT sectors. Current strategies lack efficient scheduling and resource allocation techniques leading to increased operational cost and time. This aims towards the development of enhanced strategies through improved job scheduling and resource allocation techniques for overcoming the above-stated issues. Here, Equal Spread Current Execution Load algorithm dynamically allocates the resources to the job in queue leading reduced cost.

ACKNOWLEDGMENT

REFERENCES

[1]   M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "Above the Clouds: A Berkeley View of Cloud Computing", EECS Department, University of California, Berkeley,Technical Report No., UCB/EECS-2009-28, pages 1-23, February 2009.
[2]   R. W. Lucky, "Cloud computing", IEEE Journal of Spectrum, Vol. 46, No. 5, May 2009, pages 27-45.
[3]   M. D. Dikaiakos, G. Pallis, D. Katsa, P. Mehra, and A. Vakali, "Cloud Computing: Distributed Internet Computin for IT and Scientific Research", IEEE Journal of Internet Computing, Vol. 13, No. 5, September/October 2009, pages 10-13.
[4]   G. Pallis, "Cloud Computing: The New Frontier of Internet Computing", IEEE Journal of Internet Computing, Vol. 14,No. 5, September/October 2010, pages 70-73
[5]   B. P. Rima, E. Choi, and I. Lumb, "Resource allocation of Cloud Computing Systems", Proceedings of 5th IEEE International Joint Conference on INC, IMS and IDC, Seoul, Korea, August 2009, pages 44-51.