INTERNATIONAL JOURNAL OF ADVANCED RESEARCH IN ENGINEERING AND TECHNOLOGY (IJARET)

ISSN 0976 - 6480 (Print)
ISSN 0976 - 6499 (Online)
Volume 5, Issue 10, October (2014), pp. 201-205
© IAEME: www.iaeme.com/ IJARET.asp
Journal Impact Factor (2014): 7.8273 (Calculated by GISI)
www.jifactor.com

IJARET

©IAEME

DYNAMIC LOAD BALANCING IN CLOUD COMPUTING ENVIRONMENTS

N. A. Joshi

Associate Professor, CSE Department, NIRMA University, Ahmedabad, India

ABSTRACT

Advancements in computing and communication technologies have promoted acceptance of cloud computing paradigm in the domain of distributed and high performance computing systems. Because of fast revolutions and availability of state of the art computing technology, and increasing market competition, it has become necessity for service providers to serve the cloud based services in competent way. Well organized management of cloud computing resources is the key requirement of service providers for not only providing customer satisfaction but also earning higher revenues. Furthermore, the day by day increasing demands for cloud-based resources have made the job of resource allocation very difficult for service providers, it has become challenging for them to keep up with the SLAs. The notion of load balancing is solution for better management of resources in the cloud environments. This paper presents a load balancing algorithm to provide healthier resource allocation scheme**.**

Keywords: Cloud Computing, Load Balancing, Resource Allocation, Virtualization.

I. INTRODUCTION

With the help of rapid advancements in computing and communication technologies, the cloud computing paradigm is gradually being welcomed and accepted across the globe by not only SMEs but also individuals. Numerous benefits of this rising technology such as on-demand availability, elasticity, cost-saving and relaxation from maintenance of on place IT infrastructure has motivated IT solution providers to amalgamate it into newly proposed computing solutions; furthermore because of nature of 'computing as a utility' many businesses also have started planning to migrate their legacy computing applications to cloud based solutions. Because of its elasticity and scalability characteristics, cloud service providers have been able to serve the on users' requests as it distributes the requested computing assignments to the resource pool made from a large number of computing and storage devices accommodated in datacenters established by the service providers [16]. However, the gradually increasing demands for cloud computing based services may make the task of resource allocation puzzling. The success of cloud computing services for service providers depends on quality of service and meeting the service level agreements. The task of effective resource allocation may become tedious during varying increase in on demand resource requirements against the limited amount of physical computing resources which are actually available in data centers. A single cloud provider sometimes may not be able to fulfill the increasing users and their varying needs of on demand resource requests, which may result into violations in service level agreements. In such cases, redirecting some of the incoming requests to other cloud providers by the overloaded cloud provider may become beneficial to overcome the weakness of violation in service level agreements to serve better the users and to earn profits. Such an association of multiple cloud platforms may result into higher availability and improved performance [10]. Moreover, it enables the service providers to lease their non-utilized resources (otherwise which would simply go wasted) and generate revenue from them [7].

An efficient approach of resource allocation by means of load balancing within a group of cloud service providers is presented in this paper. The rest of this paper is structured as follows. Section 2 discusses the related work in this area. Section 3 presents the proposed method. The experimental aspects and consequences are described in section 4. The paper ends with the concluding remarks in section 5.

II. RELATED WORK

A cloud computing platform consists of many elements including datacenter, distributed servers and users. A cloud computing platform may address various issues such as on-demand availability, scalability, flexibility, reduced cost of ownership and fault tolerance. Moreover, effective management of resources belonging to the cloud platform is significant to the issues mentioned above [4] [11]. However, the adoption of virtualization technique in cloud computing infrastructure may degrade the performance for the job requests involving demands for computing power and main memory [13]. The process of effective resource management involves the mechanism of load balancing of various resources such as - computing power, main memory and network bandwidth - which involves effective distribution of workload among the networked resources not only to improve utilization and response time but also to avoid situations where some of the workstations are overloaded while others are lightly loaded or sitting idle. Accordingly, the load balancing mechanism sees that all the networked workstations perform nearly the equal amount of jobs [5], [12]. Difficulties faced in joining different cloud computing environments are discussed in [2]. The load balancing work suggested in [3], is based on the number of connections. It determines the nodes having least number of connections first and then sends requests to such nodes. Reference [1] has proposed an improved implementation of [3]; he has focused on session switching at the application layer. Additionally, it calculates the connection time between the client and the node. Availability of unlimited virtual resources is an important characteristic of cloud-based infrastructures. However, it may always not be feasible for providers to guarantee the allocation of all the demanded resources [7]. They have suggested significant work showing reduction in SLA violation in federated clouds environment. Reference [17] suggests a load balancing technique which collects the workload information of nodes using the ants behavior and accordingly assigns the job-request to appropriate node. An effort has been suggested in [6] has to solve the synchronization issue of [17]. The suggested work is based on committing suicide by an ant however. The work suggested in [14], aims to provide dynamic load balancing by reducing duplication of data by suggesting the Index Name Server algorithm. The work presented in [12] suggests load balancing efforts by allocating the job requests to the node which is having the least number of connections. Moreover, [8] and [9] present work related to centralized and distributed load balancing mechanisms for homogeneous networked systems. A resource aware load balancing work which provides better results using exponential smoothing forecast is presented in [15].

However, the discussion made so far shows that either the work is applicable to single cloud platforms or there is a scope for the betterment in the published work. Here we focus on efficient resource allocation using load balancing in multiple cloud computing environments. Though there are huge scopes in adoption of cloud based computing services, there is a lack of standardization in the development and application of open source cloud computing services.

III. ALGORITHM

Let R be the set of pending VM requests to broker. Where Ri describes type of request and amount of resources required. The algorithms deals with two types of VM requests- reserved VM and on-demand VM. R = {R1, R2, R3, …,Rn}.

For every cloud provider, the distance with other providers i.e. D(CP) is determined, and that is maintained in sorted queues which represents distances in increasing order. A distance threshold Td is maintained on this sorted queues.

Let T1 and T2 be the two threshold levels for load; where $0<T1<T2$. For a particular load value, three possibilities are considered-

Load Value < T1: on-demand and reserved VM requests are entertained.
T1 < Load value < T2: Only reserved VM requests are entertained.
Load value > T2: No more requests are entertained.
These threshold values T1, T2 and Td are configurable parameters as per the needs.

The suggested algorithm gives priority to reserved type VMs. The request for reserved type VM is processed, if load falls under T2 threshold; otherwise, algorithm makes room by migrating suitable VM. The requests for on demand VM is processed if load value falls under T1 threshold; otherwise the algorithm takes services of other suitable provider by redirecting the resource request Ri.

The proposed algorithm is as follows:

```
For every request Ri in the set R
Do
   If nature of Ri is for reserved-type VM
      Calculate load of cloud provider CP, say L(CP)
      If L(CP) < T2 then
         createVM(Ri) //Process request
         Remove Ri from the set R //update set R
      Else
         Select unreserved VMm from that datacenter CP
         for migration.
         If VMm found then
            //Make space for reserved VM request
            migrateVM(VMm) //Migrate unreserved VM
            createVM(Ri)
            Remove Ri from the set R //update set R
         End if
      End if
   Else if nature of Ri is for on-demand VM then
      If L(CP) < T1 then
         createVM(Ri)
         Remove Ri from the set R
```

```
    Else
       Find cloud provider having D(CPx) < Td and
       L(CPx) < T1 with best-fit
       If such a CPx is found then
          createVM(Ri, CPx)
          Remove Ri from the set Rb
       End if
    End if
End if
Process next Ri in the set R.
Done
```

Here, the best-fit technique helps in finding provider having best fit sufficient free resources, thereby keeping other cloud providers (having larger free resources) undisturbed by leaving their free resources unfragmented and keeping them available for other requests helping reducing SLA violations and better response times.

IV. EXPERIMENTAL OBSERVATIONS

The suggested resource allocation algorithm is implemented using java on cloudsim and cloudanalyst with combination of 3 data centers each having equal capacity. The VMs are configured with Fedora9 OS, 100 MIPS and 1024 MB of memory. The overall average response time the suggested algorithm observed is 169.70 ms and violations in SLAs is 8.1 %.

V. CONCLUDING REMARKS

An approach to efficient resource allocation is suggested in this paper. The presented algorithm deals with allocation of mainly two types - on-demand and reserved - VMs among multiple vendors using best-fit resource scenario. Results indicate betterment in bringing lower the percentage of violations in SLAs and response time. However, the best-fit approach may increase the allocation processing time in larger collaborations. Likewise, the incorporation of first-fit and worst-fit approaches in place of best-fit may be studied in future.

REFERENCES

[1] B. Radojevic and M. Zagar, "Analysis of issues with load balancing algorithms in hosted (cloud) environments." In proceecings of 34th International Convention on MIPRO, IEEE, 2011.

[2] B. Rochwerger, D. Breitgand, E. Levy, A. Galis, K. Nagin, I. M. Llorente, R. Montero, Y. Wolfsthal, E. Elmroth, J. Caceres, M. Ben-Yehuda, W. Emmerich, and F. Galan, “The Reservoir model and architecture for open federated Cloud computing,” IBM Journal of Research and Development, vol. 53, no. 4, pp. 1–11, Jul. 2009.

[3] B. Sotomayor, R. Montero, I. Llorente, and I. Foster, "Virtual infrastructure management in private and hybrid clouds," in IEEE Internet Computing, Vol. 13, No. 5, pp: 14-22, 2009.

[4] H. Qian, H. Zu, C. Cao, and Q. Wang, “Css: Facilitate the cloud service selection in IaaS platforms,” Proceeding of IEEE International Conference on Collaboration Technologies and Systems, 2013.

[5] J. Gasior and F. Seredynski, “Load balancing in cloud computing systems through formation of coalitions in a spatially generalized prisoner’s dilemma game,” in CLOUD COMPUTING

2012, Third International Conference on Cloud Computing, GRIDs, and Virtualization, pp. 201–205, 2012.

[6] K. Nishant, P. Sharma, V. Krishna, C. Gupta, KP. Singh, N. Nitin and R. Rastogi, "Load Balancing of Nodes in Cloud Using Ant Colony Optimization." In proc. 14th International Conference on Computer Modelling and Simulation (UKSim), IEEE, pp: 3-8, March 2012.

[7] K. Patel and A. Sarje, "VM Provisioning Method to Improve the Profit and SLA Violation of Cloud Service Providers," IEEE International Conference on Cloud Computing in Emerging Markets, October 2012.

[8] N. Joshi and D. Choksi, "Balancing the Load of Networked Workstations", Journal of Information, Knowledge and Research in Computer Science and Applications, ISSN: 0975-6728, Vol.-2 Issue-2, pp: 119-122, February 2013.

[9] N. Joshi and D. Choksi, "Mechanism for Implementation of Load Balancing using Process Migration", International Journal of Computer Applications, ISSN- 0975-8887, Vol.-40, No.-9, pp: 16-18, February 2012.

[10] N. A. Joshi, Performance-Centric Cloud-Based e-Learning, The IUP Journal of Information Technology, ISSN 0973-2896, Vol. 10, No. 2, pp. 7-16, June 2014, Reference # 35J-2014-06-01-01.

[11] R. Buyya, C.S. Yeo & S.Venugopal, "Market-oriented Cloud computing: Vision, hype, and reality of delivering IT services as computing utilities," 10th IEEE International Conference on High Performance Computing, pp: 5–13, 2009.

[12] R. Lee and B. Jeng, "Load-balancing tactics in cloud," in Proceedings of the International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, IEEE Computer Society, pp: 447–454, 2011.

[13] S. Ristov, G. Velkoski, M. Gusev, and K. Kjiroski, "Compute and memory intensive web service performance in the cloud," in ICT Innovations 2012,Springer Berlin, vol. AISC 257, pp. 215–224, 2013.

[14] W. Lee, 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.

[15] X. Ren, R. Lin and H. Zou, "A dynamic load balancing strategy for cloud computing platform based on exponential smoothing forecast" in proceedings of International Conference on Cloud Computing and Intelligent Systems, IEEE, pp: 220-224, September 2011.

[16] Z. Linan, L. Quingshui and H. Lingna, "Study on Cloud Computing Resource Scheduling Strategy Based on the Ant Colony Optimization Algorithm", International Journal of Computer Science Issues, Vol. 9, Issue 5, No 2, pp: 54-58, September 2012.

[17] Z. Zhang and X. Zhang, "A load balancing mechanism based on ant colony and complex network theory in open cloud computing federation", in proceedings of 2nd International Conference on. Industrial Mechatronics and Automation, IEEE, Vol. 2, pp:240-243, May 2010.

[18] Dr. Narayan A. Joshi and Dr. D. B. Choksi, "Implementation of Process Forensic for System Calls", International Journal of Advanced Research in Engineering & Technology (IJARET), Volume 5, Issue 6, 2014, pp. 77 - 82, ISSN Print: 0976-6480, ISSN Online: 0976-6499.

[19] Dr. Narayan Joshi and Parjanya Vyas, "Performance Evaluation of Parallel Computing Systems", International Journal of Advanced Research in Engineering & Technology (IJARET), Volume 5, Issue 5, 2014, pp. 82 - 90, ISSN Print: 0976-6480, ISSN Online: 0976-6499.

[20] Dr. Narayan A. Joshi, "Load Balancing in Cloud using Process Migration", International Journal of Advanced Research in Engineering & Technology (IJARET), Volume 5, Issue 4, 2014, pp. 230 - 238, ISSN Print: 0976-6480, ISSN Online: 0976-6499.