JCSE International Journal of Computer Science and Engineering Open Access
Research Paper Volume-2, Issue-3 E-ISSN: 2347-2693

An Optimistic Approach for Load Balancing in Cloud Computing

Meenakshi Sharma[1], Anitha Y[2*] and Pankaj Sharma[3]

[1]*Department of Computer Science and Engineering, Punjab Technical University, India, sharma.minaxi@gmail.com*
[*2] *Department of Computer Science and Engineering, Punjab Technical University, India,anithasarath89@gmail.com*
[3]*Department of Computer Science and Engineering, Punjab Technical University, India,Pankajuppal22@gmail.com*

www.ijcseonline.org

Received: 8 March 2014 Revised: 16 March 2014 Accepted: 22 March 2014 Published: 30 March 2014

Abstract—Cloud computing technology is changing the focus of IT world and it is becoming popular because of its great characteristics. Load balancing is one of the main challenges in cloud computing. Load balancing is the methodology to distribute the load across multiple servers or a cluster of servers, databases or other resources. Efficient load balancing helps to optimize the server usage, increase throughput, and decrease response time. The objective of this paper to propose a load balancing algorithm that can provide an efficient load balancing. The results discussed in this paper are based on existing round robin, least connection, throttled, fastest response time and the new proposed algorithm. This new algorithm improves the overall response time and data centre processing time as well as reduce the cost, in comparison to the existing algorithms.

Keywords/Index Term—Cloud computing, load balancing, simulation, cloudSim

I. INTRODUCTION

Over the past few years, Cloud computing technology drawn the attention of IT world and is changing the focus of enterprises. Cloud computing can be defined as a style of computing where software applications are provided to consumer as “service” rather than a product using the internet. Cloud stands as a metaphor for internet. Cloud computing has gained attention due to the growth of internet technologies, reduced costs in storage, growth of visualization and advancement in internet security. Cloud computing delivers application or services on-demand basis. Cloud computing providers will need to provide high traffic applications which are reliable, fast, secure and highly available. Along with visualization, infrastructure like load balancer, which does load balancing play a vital role in a successful cloud-based implementation. The following figure.1 represents a load balancing infrastructure.

Load balancing is the methodology to distribute the load across multiple servers or a cluster of servers, databases or other resources. Efficient load balancing helps to optimize the server usage, increase throughput, decrease response time. A variety of scheduling algorithms are available which is used by load balancers to determine which back-end server to send a request. Choosing the right load balancing algorithm is imperative to the success of cloud computing. The right load balancing is the backbone of an advanced cloud computing architecture.

This paper is organized as follows: An overview of Cloud Computing is in section II. Section III describes existing load balancing algorithms. Section IV is the proposed the research work. Section V describes the research setup and analysis. Section VI provides the research result description and the conclusion in section VII.

Corresponding Author: *anithasarath89@gmail.com*

II. CLOUD COMPUTING

A. *Brief Literature Survey*

The Cloud Computing has many descriptions. Rajkumar Buyya et.al [1] have defined it as follows: Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers. Due to the recent emergence of cloud computing research in load balancing this is in the preliminary stage. N J Kansal Jiyan [2] has proposed a service models are provided by the cloud. Rimal B.P. et. al [3] discussed the existing issues like Load Balancing, Virtual Machine Migration, Server Consolidation, Energy Management etc. Bhathiya. et. al [4] present execution environment considers Datacenter, Virtual Machine (VM), host and Cloudlet components from CloudSim for execution analysis of algorithms.

Zenon Chaczko. et. al [5] gives an idea about the basic concepts of Cloud Computing and Load balancing availability and load balancing in cloud Computing. R P Padhy et. al [6] studied about some existing load balancing algorithms, which can be applied to clouds. In addition to that, the closed-form solutions for minimum measurement and reporting time for single level tree networks with different load balancing strategies were also studied. The user or researcher can actually analyse the proposed design or existing algorithms through simulation. They can check the efficiency and merit of the design before the actual system is constructed.

Rajkumar Buyya et. al [7],in this paper he has studied the features of a CloudSim simulator to compare the performance of three dynamic load balancing algorithms. Bhathiya. et. al [4], have illustrated CloudSim architecture. W.Bhathiya et. al,"Cloud Analyst: A cloud sim-based visual modeller for analysing cloud computing environments and applications" [8], which present how cloud analyst can be used to model and evaluate a real world problem through a case study of a social networking application deployed on the cloud. The cloud analyst is a GUI based tool which is developed on CloudSim architecture. How the simulator can be used to effectively identify overall usage patterns and how such usage patterns affect data centres hosting the application. M.Sharma et. al [14], have discussed performance evaluation of adaptive virtual machine load balancing algorithms for cloud computing.

B. *Cloud Computing*

Rajkumar Buyya have defined cloud computing as follows: "Cloud is a parallel and distributed computing system consisting of a collection of inter-connected and virtualized computers that are dynamically provisioned and presented as one or more unified computing resources based on service-level agreements (SLA) established through negotiation between the service provider and consumers"[1].Any cloud computing system consists of three major components such as clients, datacenter and distributed servers [3].

Client: End users who interacts cloud based applications. Clients generally fall into three categories-Mobile: smart phones like blackberry, iPhone. Thin: they only display the information received from the server and do not do any processing or have any internal memory. Thick: uses a web browser such as Internet explorer or Google Chrome.

Datacenter: Various applications hosted by collection of servers. An end user connects to the datacenter to use the services provided by different applications. A datacenter may exist anywhere in the world.

Distributed Servers: A server, which actively checks the services of their hosts, known as Distributed server. It is the part of a cloud which is available through the internet hosting different implementation. But the user would feel that they are using this application from its own machine while using it [9].

III. LOAD BALANCING

Load balancing is a methodology to distribute workload across multiple computers, or other resources over the network links [10].Load balancing achieve optimal resource utilization, maximize throughput, minimum response time, and avoid overload. Cloud vendors provides automatic load balancing services also known as auto scaling, which automatically increase the number of CPU's or the amount of memory on demand or increased load. Load balancing serves two important needs, first to promote availability of Cloud resources, second to promote performance [14].

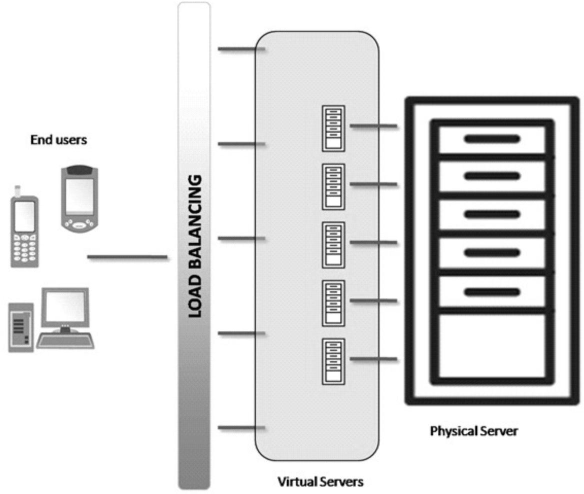

Figure.1 Load balancing infrastructure

A. *Existing Load balancing algorithm*

We use four existing algorithms to distribute the load.

1. *Round Robin Algorithm (RR)*[11]:

Round Robin algorithm distributes requests evenly to the servers. The work load distributions between servers are equal but the processing time may vary for different requests. So at any point of time some servers may be heavily loaded while others remain idle. This algorithm allocates the request simply on a turn basis without considering the load on the servers. This algorithm is good when requests are of similar nature and can be distributed equally

2. *Least Connection Algorithm*(LC)[8]:

This algorithm sends the request to the server which has the least connection by tracking the number of connections attached to each server. Least Connection Load Balancer (LCLB) maintains an index table of servers and the number of requests currently allocated to each server. When a new request arrives, the index table is parsed and the least loaded server is identified. At any point if one or more servers have least connection, first one is picked.

3. *Throttled Load Balancer*(TLB)*[15]:*

In this algorithm the throttled load balancer (TVLB) maintains an index table of each servers and their current state (Busy/Available) which identifies if the server is available or not. The data center controller (DCC) receives a new request from client/server to find a suitable virtual machine (VM) to perform the recommended job. The data centre queries the load balancer for the next allocation of VM. The load balancer parses the allocation table from top until the first available VM is found or the table is parsed completely. If the VM is found returns the VM id to the DCC. Further, the data centre acknowledges the load balancer of the new allocation and the data centre updates the allocation table accordingly. While processing the request of client, if appropriate VM is not found, the load balancer returns -1 to the data centre. The DCC queues the request with it. When the VM finishes processing the request, and the DCC receives the response, it notifies the load balancer a request is acknowledged to data centre to de- allocate the same VM whose id is already communicated. The DCC checks if the queue contains any waiting requests. If it present then it continue.

4. Fastest Response Time(FRT)[16]:

The Fastest method forwards a new connection to server which has the fastest response time of all servers. The load balancer looks at the response time of each attached server and chooses the one with the best response time. Fastest VM Load Balancer (FLB) maintains a table which contains VMs and the response time of the VM. Datacenter receives a new request and queries the FLB for the next allocation. FLB scans the table from top until the first available the fast available VM is found. If the VM is found the data centre communicates the request to the VM and returns the VM id to the datacenter. Further, the data centre acknowledges the load balancer of the new allocation and the data centre revises the index table accordingly. While processing the request of client, if appropriate VM is not found, the load balancer returns -1 to the data centre. The data centre queuing the request with it. When the allocated task completes, a request is acknowledged to data centre from VM, which is further apprised to load balancer to de- allocate the same VM whose id is already communicated.

IV. RESEARCH WORK

A. Fastest With Least Connection(FLC):

This algorithm is a combination of the logic used in the FLC algorithms. Here servers are ranked based on a combination of the number of current connections and the response time. Servers that have the fewest connections and fastest response time receive a greater proportion of the connections. The response time of each Virtual Machine is calculated as:

$$\text{Response Time} = \text{Fint} - \text{Arrt} + \text{TDelay} \qquad (1)$$

Where, Arrt is the Arrival time of user request. Fint is the user request finish time. TDelay is the transmission delay:

$$\text{TDelay} = \text{Tlatency} + \text{Ttransfer} \qquad (2)$$

Where, Tlatency is the network latency. Ttransfer is the time taken to transfer the size of data of a single request (D) from source location to destination.

$$\text{Ttransfer} = \text{D} / \text{Bwperuser} \qquad (3)$$

$$\text{Bwperuser} = \text{Bwtotal} / \text{Nr} \qquad (4)$$

Where, Bwtotal is the total available bandwidth. Nr is the number of user requests currently in transmission.

The RANK of VM based on the response time and no of active connections of each Virtual Machine, which is calculated as:

$$(\text{Rank})n=(\text{Vmmax Ac}-(\text{VmAc})\ n) + (\text{Vmmax Rt}- (\text{Vm Rt})n \qquad (5)$$

Where, Vmmax Ac is the VM maximum Active Connection value. (VmAc) n is the active connection value of nth VM. Vmmax Rt is the VM maximum Response Time value. (VmRt) n is the response time value of nth VM.

If one or more VM gets the highest rank, we will randomly choose a VM out of that with least connections or lowest response time.

$$\text{Cost} = \text{totalTime} * \text{costPerVmHour} \qquad (6)$$

$$\text{total Time}=\text{toalTime} + (\text{end} - \text{start}) \qquad (7)$$

Where, CostPerVmHour is the VMs cost in 1 hour running. start is the start vmAllocationTime and end is the end vmAllocationTime.

V. RESEARCH SETUP & ANALYSIS

Fastest with Least Connection algorithm is implemented through simulation package CloudSim based tool [7][8][13]. Java language is used for implementing this new load balancing algorithm. Assuming the application is deployed in two data centre having 50 virtual machines running on 6 user base; with below parameters:

TABLE 1. PARAMETER VALUES

PARAMETER	VALUE
Data Center OS	Windows 7
Data Center Architecture	X86
Service Broker Policy	Optimize Response Time
VM Memory	1024 Mb
VM Bandwidth	1000 Mb

The simulation result computed is as shown in the following tables.

TABLE 2 . OVERALL AVERAGE RESPONSE TIME OF FASTEST WITH LEAST CONNECTION LOAD BLANCING ALGORITHM

Performance Parameters	Avg (ms)	Min (ms)	Max (ms)
Overall Response Time	1001.76	41.52	4770.46
Data Center processing time	811.71	2.49	4208.45

TABLE 3. COST WITH LEAST CONNECTION LOAD BLANCING ALGORITHM

Performance Parameters	VM Cost ($)	Data Transfer Cost ($)	Total Cost ($)
Cost	10.04	10.00	20.04

TABLE 4. COMPARISON OF AVG RESPONSE TIME OF VM LOAD BALANCING ALGORITHMS.

Performance Parameters	Round robin (ms)	Least connetion (ms)	Throttled (ms)	Fastest (ms)	Fastest with Least Connect-ion (ms)
Response Time	1761.53	1759.45	1057.88	1012.40	1001.76

TABLE 5. COMPARISON OF DATA TRANSFER COST OF VM LOAD BALANCING ALGORITHMS.

Performance Parameters	Round robin (ms)	Least connection (ms)	Throttled (ms)	Fastest (ms)	Fastest with Least Connection (ms)
Data Transfer Cost ($)	222.31	221.20	221.13	220.09	209.01

VI. RESULT

For analysis of the simulation we collected the desired outputs from all the five load balancing algorithms. The above shown tables clearly indicates that the parameters: response time, data processing time and processing cost are almost similar in RRB and LC algorithms whereas these parameters are improved in TLB and FR. But the same for the FLC I has a significant improvement over other algorithms. The following figure.2 shows the analytical comparison of various algorithms. From the above data and analysis, the new algorithm, FLC is more efficient than others.

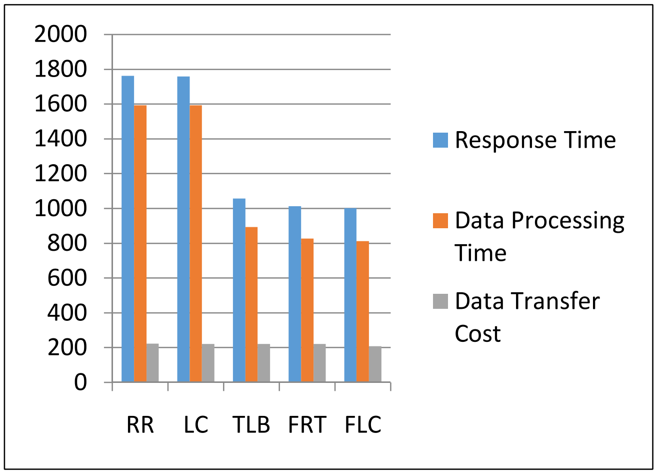

Figure 2.Analytical Comparison of Various Algorithms

VII. CONCLUSION

The performances of four existing algorithms are studied in the paper. The paper aims to develop enhanced strategies through improved job and load balancing resource allocation techniques. A new fastest with least connection scheduling algorithm is proposed and then implemented in cloud computing environment using CloudSim toolkit, and Java language. From the experiment, we identified that the overall response time and data centre processing time is improved using the proposed new algorithm. A conclusion might elaborate on the importance of the work or suggest applications and extensions.

REFERENCES

[1] Buyya, C. S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, Cloud computing and emerging IT platforms: Vision, hype, and reality for delivering computing as the 5th utility, Future Generation Computer Systems, 25:599616, 2009.

[2] Nidhi Jain Kansal, “Cloud Load Balancing Techniques: A Step Towards Green Computing”, IJCSI International Journal Of Computer Science Issues, January 2012, Vol. 9, Issue 1, No 1, Pg No.:238-246, ISSN (Online): 1694-0814.

[3] Rimal B.P., Choi E. and Lumb I. (2009) 5th International Joint Conference on INC, IMS and IDC, 44-51.

[4] Bhathiya, Wickremasinghe.(2010)”Cloud Analyst: A Cloud Sim-based Visual Modeller for Analysing Cloud Computing Environments and Applications”

[5] Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh, Christopher Mcdermid (2011)“Availabity and Load Balancing in Cloud Computing” International Conference on Computer and Software Modeling IPCSIT vol.14 IACSIT Press, Singapore 2011.

[6] Ram Prasad Padhy (107CS046), PGoutam Prasad Rao (107CS039).”Load balancing in cloud computing system” Department of Computer Science and Engineering National Institute of Technology, Rourkela Rourkela-769 008, Orissa, India May, 2011.

[7] Calheiros Rodrigo N., Rajiv Ranjan, and César A. F. De Rose, Rajkumar Buyya (2009): CloudSim: A Novel Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services CoRR abs/0903.2525: (2009).

[8] Bhathiya Wickremasinghe “Cloud Analyst: A Cloud-Sim-Based Tool For Modeling And Analysis Of Large Scale Cloud Computing Environments. MEDC Project”, Report 2010.

[9] Tackle your client’s security issues with cloud computing in 10 steps, http://searchsecuritychannel.techtarget.com/tip/Tackle-your-clients-security-issues-withcloud-computing-in-10-steps.

[10] M. Armbrust, A. Fox, R. Griffith, A. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson,A.Rabkin, I. Stoica,M. Zaharia (2009). Above the Clouds: A Berkeley View of Cloudcomputing.TechnicalReport No. UCB/EECS-2009-28, University of California at Berkley, USA, Feb. 10, 2009.

[11] Ko, Soon-Heum; Kim, Nayong; Kim, Joohyun; Thota, Abhinav; Jha, and Shantenu; (2010)"Efficient Runtime Environment for Coupled Multi-physics Simulations: Dynamic Resource Allocation and Load-Balancing" 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid), 17-20 May 2010, pp.349-358.

[12] Saroj Hiranwal , Dr. K.C. Roy, "Adaptive Round Robin Scheduling Using Shortest Burst Approach Based On Smart Time Slice" International Journal Of Computer Science And Communication July-December 2011 ,Vol. 2, No. 2 , Pp. 319-323.

[13] Jinhua Hu; Jianhua Gu; Guofei Sun; Tianhai Zhao; (2010) "A Scheduling Strategy on Load Balancing of Virtual Machine Resources in Cloud Computing Environment" Third International Symposium on Parallel Architectures, Algorithms and Programming (PAAP), 18-20 Dec. 2010, pp.89-96

[14] Brain Underdahl, Margaret Lewis and Tim mueting "Cloud computing clusters for dummies" Wiley Publication (2010), [Book].

[15] Roderigo N. Calherios, Bhathiya Wickremasinghe "Cloud Analyst: A Cloud-Sim-Based Visual Modeler For Analyzing Cloud Computing Environments And Applications". Proc of IEEE International Conference on Advance Information Networking and Applications, 2010.

[16] Sandeep Sharma, Sarabjit Singh, and Meenakshi Sharma "Performance Analysis of Load Balancing Algorithms" World Academy of Science, Engineering and Technology 38 ,2008 page no 269- 272.

[17] Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh, Christopher Mcdermid (2011)"Availabity and Load Balancing in Cloud Computing" International Conference on Computer and Software Modeling IPCSIT vol.14 IACSIT Press, Singapore 2011.

[18] Cloud Security and Privacy An Enterprise Perspective on risk and compliance by Tim Mather, Subra kumaraswamy, Shaheed Latif

[19] Implementing and developing Cloud Computing Application by David E.Y Sarna .

[20] Kun Li, Gaochao Xu, Guangyu Zhao, Yushuang Dong, Dan Wang (2011) " Cloud Task scheduling based on Load Balancing Ant Colony Optimization " Sixth Annual ChinaGrid Conference ,2011,PP 3-9.

[21] Huaiming Song, Yanlong Yin, Xian-He Sun, Thakur, R. and Lang, S.; (2011) "A Segment-Level Adaptive Data Layout Scheme for Improved Load Balance in Parallel File Systems" 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 23-26 May 2011, pp.414-423.

[22] Meenakshi Sharma and Pankaj Sharma "Performance Evaluation of Adaptive Virtual Machine Load Balancing Algorithm" (IJACSA) International Journal of Advanced Computer Science and Applications,Vol. 3, No.2, 2012

[23] Don MacVittie, Intro to Load Balancing for Developers – March 31.

[24] "Cloud Computing Made Easy" by Cary Landis and Dan Blacharski,version 0.3.

AUTHORS PROFILE

Meenakshi Sharma is Head of the Department of Computer Science in Sri Sai College of Engineering And Technology, Punjab, India. She has a Masters in Technology and pursuing her Ph.D. Her research work has published in more than six international journals and her areas of interests are networking and security.

Anitha Y is pursuing Master's degree in Computer Science and Engineering in Punjab Technical University, Punjab, India. Her research interests are in the areas of cloud computing

Pankaj Sharma received the Master's degree in Computer Science and Engineering from Punjab Technical University, Punjab, India. He is currently a Professor in the Department of Computer Science and Engineering in Sri Sai College of Engineering and Technology, Punjab, India. His research interests are in the areas of cloud computing, computer networks.