Dynamic Load Balancing with Min Scaling and Cooldown Adjustment

Shweta Gupta, Ms. Jasmine Jha
Computer Science Department of L.J. Institute of Engineering & Technology
L. J. Campus between Sarkhej Circle & Kataria Motors, S.G. Highway, Ahmedabad, Gujarat, India.

***Abstract*— the fast development of internet has given to a new computing terminology: Cloud Computing. This new paradigm has experienced a fantastic rise in current years. As Cloud Computing is expanding swiftly and clients are demanding better results and more services, load balancing for the cloud has become a major concern in the cloud computing environment. Load balancing is the process of distributing the load among various nodes 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 lightly loaded. Using proper load balancing we can minimize resource consumption, enable scalability, implement fail-over, avoid over-provisioning and bottlenecks etc. In our proposed work load balancing algorithm allocates the incoming requests to the all available virtual machines effectively and also perform auto scaling using coolup and cooldown adjustment. Using this algorithm we can also reduce response time of the requests.**

***Index Terms*— Cloud Computing, Load balancing, Virtual machine, Resource Utilization, Scaling, Cooldown.**

I. INTRODUCTION

Cloud Computing is a broadly used term in Internet. The whole internet is look like a cloud. Cloud computing gaining momentum because it is appears to be a highly disruptive technology. It has inherited new ideas of distributed systems on large scale. Many people believe that Cloud is going to reshape the IT industry as a revolution. The next developmental step of Cloud computing is distributed computing. To make a superior usage of distributed resources, place them together such as we can achieve higher throughput and be capable to pact with huge-scale computational problems. Cloud Computing is an one type of on demand service in which software, information, resources and devices are added according to the customers requirement at particular time. Virtualization is the basic concept behind the Cloud computing. Computing resources are made available in a pay as per use manner to users.

If the cloud provider is not designed any excellent mechanism for load balancing then loaded nodes on the cloud provide indigent performance in resource usage and the space of cloud servers would not be make use of properly. If there is any excellent technique for load balance then load is equally distributed on cloud. As the numbers of consumers are increasing on the cloud, the load balancing has grown into the challenge for the cloud provider.

Load Balancing is important for all virtual machines in the cloud because we can dynamically distributes workloads across multiple computing resources equally. Using load balancing Each VM does the equal amount of work throughout so it increase the throughput and decrease the response time. Load Balancing maximizes the user satisfaction, minimizing response time, reducing the number of job rejections increasing resource utilization and raising the performance ratio of the system. Virtualization technology provide powerful Management of the dynamic resources in cloud paradigm.

Static and dynamic are main types of load balancing algorithms. Static are largely used for stable and homogeneous surroundings and produce worthy results but not flexible and can’t cope up with dynamic changes of attributes at the time of execution. On the other hand Dynamic algorithms are suitable for heterogeneous environment, flexible, consider different attributes in the system and also handle dynamic changes of attributes at the time of execution.

Scalability can be defined as the ability of an application to make optimum utilization of resources at different levels (avoiding overprovisioning, under-utilization and under-provisioning) [3]. Cloud computing provide a powerful environment to scaling with no difficulty. Load balancing is an essential elements in achieving scalability in the cloud.

As the numbers of users are increase on the cloud, the load balancing has become the important issue for the cloud provider. In this dissertation we have proposed an algorithm for dynamic load balancing which will assigns the incoming requests to the available VMs (virtual machines) effectively and also perform minimum auto scaling using coolup and cooldown adjustment. This can reduce the number of job rejections, increase resource utilization, raising the performance ratio of the system and also decrease the response time of requests.

II. LITERATURE SURVEY

We reviewed two papers on cloud computing and we conclude that Cloud is a one kind of network that provides different services to the client as per client requirement. Cloud is mainly providing the three type of services that IAAS, PAAS, SAAS that is Architecture as services, Software as a services, and platform as services. In the cloud that all services not provide freely only some particular resource provide as a free services other is need to use by the client then client is pay for that as per it is fix rate. Cloud also provide resources as permanent and also temporary as per users need but in the temporary base usage after some time that user release that resource and also same time some user request for that resource and also at a time more than one user need same resources [2][6].

Ian Lumb et al. [1] give the taxonomy which is used to define the fundamentals that provides a framework understanding the existing cloud computing offers. The root idea after taxonomy is to know the technical weakness, technical strength, and challenges in the recent cloud computing systems and suggest what must be done in future to strengthen that type of cloud computing systems. This information is necessary for cloud enterprise firms, service suppliers, and border authorities to sense, control, and manage nosy exotic components. Distributed systems for massive data processing is root idea behind the principles for define the taxonomy. This criteria focus on cloud service, architecture, fault tolerance, virtualization management, and investigate mechanisms such as interoperability, scalable data storage and load balancing.

Bellenger et al. [3] describes two approaches to scaling in cloud– semi-automatic and automatic and explains why the second is to prefer. In this paper author also explain manual scaling. Authors only consider scaling on IaaS and PaaS levels. Semi-automatic scaling is described with an Amazon Elastic Compute Cloud, whereas automatic scaling is experimented with an Amazon Web Services Elastic Beanstalk.

Ajit et al. [4] developed the VM load balancing algorithm at the VM level using cloud analyst tool and CloudSim library. This algorithm find load assignment factor for every host in a datacenter and according to this factor map all the VMs. Load balancer send VM id to highest configuration host having maximum LAF then lowest one and so on. At last author conclude that if we select a VM mapped on powerful host, then it effect the overall performance of cloud and decrease the average response time. This algorithm work on homogeneous VMs mapped on hosts.

Shridhar G. Domanal et al. [7] established Modified Throttled algorithm. This is improvise version of Throttled algorithm. This approach consider uniform load sharing among the VMs and availability of VMs. The response time of proposed algorithm has advanced compare to Round-Robin and Throttled algorithm. Proposed algorithm was developed on CloudAnalyst simulator with six real time scenario. The future scope of this proposed work is need to focus on changing the data structures of index table and utilization of VMs and response time may be more optimized.

This paper providing an overview of cloud technology and its components. Writers also focusing on load balancing of cloud computing with some existing load balancing techniques, which are responsible to manage the load when some node of the cloud system is under loaded and others are over loaded. This paper also provide classification of load balancing algorithms. There is always a need of efficient load balancing algorithms for improving the utilization of computing resource [9].

Jitendra Bhatia et al. [11] developed one algorithm named HTV dynamic load balancing Algorithm for instances of virtual machine. This algorithm focuses on two parameters. First one is load on nodes and second is response time of node. There would be continuous monitoring of the resources to know the status of an available resource. Once a request comes from client, the resources would be allocated from the information present in the queue dynamically to balance the load on nodes to gain high performance and efficiency. Here they have used dynamic round robin algorithm.

Gulshan Soni et al. [12] has developed load balancing technique “CLB” will balance the load between VMs taking dissimilar hardware configurations. CLB distribute the load based on states of virtual machines and hardware configuration of VMs in data center. In the proposed algorithm, developer tried to avoid overloading and under loading of VMs using priority of VMs and states. In future they are trying to use current utilization of processor and memory for load distribution.

In this paper author investigated resource management strategies in the cloud computing context, with master–slave applications. When the IaaS performs VMs, An outlook of proposed work is to increase two-level resource management to more decrease collocated application replicas. Here, algorithm relies on fixed size VM allocation and several VMs for the same tier may be required according to the load and collocated on a node. If the cloud provider has the knowledge of the architecture of customers’ applications, we could replace multiple VMs by a single larger VM on that node [13].

Ram Mahana Reddy et al. [14] has established VM-assign algorithm. Using manage load at server by consider the current state of available VMs and assign request neatly. This algorithm focus on efficient utilization of resources. This algorithm solve the inefficient resource utilization of present algorithms. Load is only assign least loaded VMs. In future work author take more dynamic situations for arriving requests. And check response for dynamic and static load.

Trieu C. Chieu et al. [15] introduce a scaling scenario for dynamic scalability of web applications Cloud. This scenario has front-end load balancer to balance user requests and route the requests. This work has confirmed the exciting benefits of the Cloud like delivering IT resources on-demands to users in a better and cheaper way and capable of handling sudden load surges. Cloud work as an essential element in providing greater resource utilization and reduce management and infrastructure costs.

III. METHODS AND ALGORITHMS

A. Cloud Computing

Cloud computing is one of the quickest growing, and potentially most disruptive IT innovations for a generation. Easy access to storage infrastructure and high performance computing through web services provide by cloud computing. It also provides configurability, reliability and scalability along with high performance. In cloud computing we have different fundamentals like fail over mechanism, virtualization, quality of service, interoperability, and scalability.

B. Load Balancing

The definition of load balancing is "process of making effective resource utilization by reassigning the total load to the separate nodes of the combined system and thereby decreasing the response time of the job" [7]. Dynamic load balancing algorithm is depends on the current state of the system. Many capital things are consider while developing Dynamic LB algorithm like stability of different system, load estimation, comparison of load, system performance, nature of transferred work etc. Many types of load considered such as CPU load, delay or Network load, amount of memory used. Load balancing (LB) is worked on provider as well as on consumer side. On provider side, problem in load balancing is allocating VMs to servers at runtime. VM want to be reassigned so that servers do not get overload as demand changes. Just like VM load is distributed crosswise servers, application load of client can be balanced across VMs on Consumer side.

C. Scalability

It's hidden in its name; scalability is an ability of a system. The proportions and capacity of a scalable hardware or software system can be adjusted so that it can handle the growth of load and perform its work optimized to performance and costs. One cannot find a generally accepted, formalized definition of scalability; it is mostly determined by the features of the given system. Scalability is not exclusively an IT phenomenon, it can also be found in electric engineering, and even in the business world. There are two distinct ways of scaling of systems: 1. Vertical Scaling 2. Horizontal Scaling

IV. PROPOSED ALGORITHM

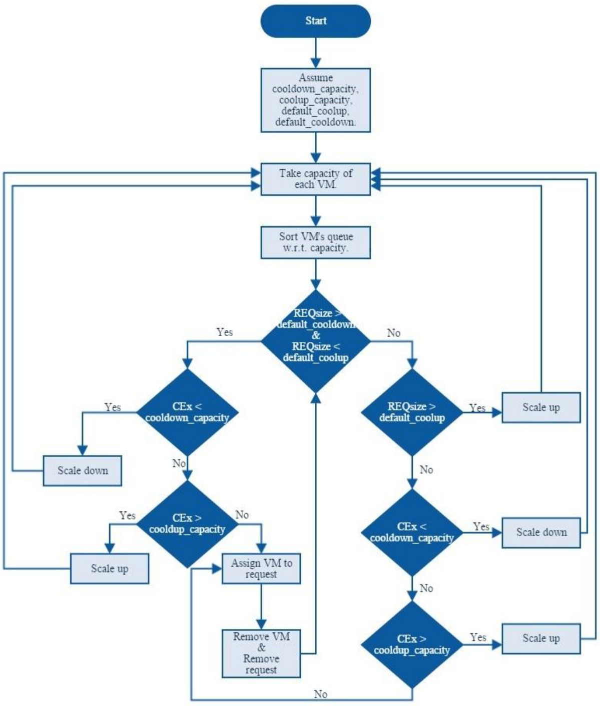

FIG.1 FLOW OF PROPOSED WORK

Parameters to be used in proposed algorithm:

default_cooldown: Min Number of requests

default_coolup: Max Number of requests

cooldown_capacity: Min CPU Utilization of VM

coolup_capacity: Max CPU Utilization of VM

REQsize: size of request queue

CEx: capacity of VM which is on first index of queue.

Proposed Algorithm:

Step-1: Assume cooldown_capacity and coolup_capacity for VM as well as default_coolup and default_cooldown requests.

Step-2: Take the capacity of all the VM's which is available in the VM's queue.

Step-3: Sort all the VM's which is available in VM's queue with respect to the capacity of VMs.

Step-4: For each request in request's queue check the size of the request queue with default_coolup and default_cooldown.

Case 1: if size of request queue less than default_coolup and greater than default_cooldown. For VM check the capacity of VM which is at the first index in queue with cooldown_capacity and coolup_capacity.

Case A: if capacity of VM less than cooldown_capacity. Then scale down and go to step2

Case B: else if capacity of VM greater than coolup_capacity. Then scale up and go to step 2

Case C: else

Allocate request to the VM and remove VM and request from the queue. Repeat step-4

Case 2: else if size of request queue is greater than default_coolup. Then scale up and go to step 2

Case 3: else

Case A: if capacity of VM less than cooldown_capacity. Then scale down and go to step2

Case B: else if capacity of VM greater than coolup_capacity. Then scale up and go to step 2

Case C: else

Allocate request to the VM and remove VM and request from the queue.

Repeat step-4

V. RESULT AND ANALYSIS

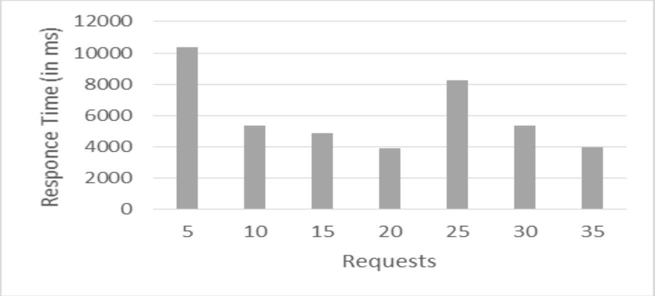

FIG.2 AVERAGE RESPONSE TIME

Above graph show the average response time for different number of requests. Using this graph we analysis that response time is varying according to CPU Utilization.

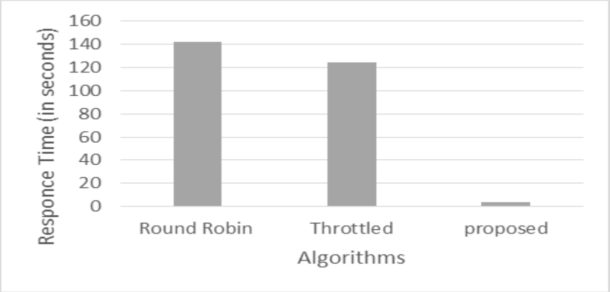

FIG.3 COMPARISON

Above Graph show the comparison of response time our proposed algorithm with two algorithms like Round Robin algorithm and Throttled algorithm.

VI. CONCLUSION

This cloud computing paradigm has experienced a fantastic rise in current years. As Cloud Computing is expanding swiftly and clients are demanding better results and more services, load balancing for the cloud has become a major concern in the cloud computing environment. Using proper load balancing we can minimize resource consumption, enable scalability, implement fail-over, avoid over-provisioning and bottlenecks etc. In our proposed work load balancing algorithm allocates the incoming requests to the all available virtual machines effectively and also performs auto scaling using coolup and cooldown adjustment. Using this algorithm we can also reduce response time of the requests. In future work the algorithm need to be more mature in large scale infrastructure as well as heavy load traffic.

REFERENCES

[1] Bhaskar Prasad Rimal, Eunmi Choi and Ian Lumb, “A Taxonomy and Survey of Cloud Computing Systems”, IEEE, Fifth International Joint Conference on INC, IMS and IDC, Pages 44-51, 25-27 Aug. 2009, ISBN 978-0-7695-3769-6.

[2] Yashpalsinh Jadeja and Kirit Modi,”Cloud Computing - Concepts, Architecture and Challenges”, IEEE, Computing, International Conference on Electronics and Electrical Technologies (ICCEET), Pages 877-880, 21-22 March 2012, ISBN 978-1-4673-0211-1.

[3] D. Bellenger, J. Bertram, A. Budina, A. Koschel, B. Pf¨ander, C. Serowy, I. Astrova, S. G. Grivas, and M. Schaaf, “Scaling in cloud environments,” in Proceedings of the 15th WSEAS international conference on Computers. Stevens Point, Wisconsin, USA: World Scientific and Engineering Academy and Society (WSEAS), 2011, pp. 145–150, ISBN 978-1-61804-019-0.

[4] Mr. M. Ajit, Ms. G. Vidya,” VM Level Load Balancing in Cloud Environment”, IEEE, Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT), Pages 1-5, 4-6 July 2013, ISBN 978-1-4799-3925-1.

[5] P. Mell and T. Grance, “The NIST Definition of Cloud Computing,” US Nat’l Inst. of Science and Technology, 2011; http://csrc.nist.gov/publications/nistpubs/800-145/SP800-145.pdf.

[6] Tharam Dillon Chen Wu and Elizabeth Chang, “Cloud Computing: Issues and Challenges”, 24th IEEE International Conference on Advanced Information Networking and Applications (AINA), Pages 27-33, 20-23 April 2010, ISBN 978-1-4244-6695-5.

[7] Shridhar G.Domanal and G.Ram Mohana Reddy,

"Load Balancing in Cloud Computing Using Modified Throttled Algorithm", IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Pages 1-5, 16-18 Oct. 2013, ISBN 978-1-4799-0027-5.

[8] Amandeep Kaur Sidhu, Supriya Kinger, "A Sophisticated Approach for Job Scheduling in Cloud Server", International Journal of Computer Trends and Technology (IJCTT), Pages 2055-2058, 7 July 2013, ISSN 2231-2803.

[9] Yatendra Sahu and R K Pateriya,"Cloud Computing Overview with Load Balancing Techniques", International Journal of Computer Applications, Pages 40-44, March 2013.

[10] Ali M. Alakeel, "A Guide to Dynamic Load Balancing in Distributed Computer Systems", IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.6, June 2010.

[11] Jitendra Bhatia, Tirth Patel, Harshal Trivedi, Vishrut Majmudar," HTV Dynamic Load Balancing Algorithm for Virtual Machine Instances in Cloud", IEEE, International Symposium on Cloud and Services Computing (ISCOS), Pages 15-20, 17-18 Dec. 2012, ISBN 978-1-4673-4854-6.

[12] Gulshan Soni and Mala Kalra, "A Novel Approach for Load Balancing in Cloud Data Center", IEEE International Advance Computing Conference (IACC), Pages 807-812, 21-22 Feb. 2014, ISBN 978-1-4799-2571-1.

[13] Alain Tchanaa, Giang Son Tran, Laurent Broto, Noel DePalma, Daniel Hagimont, "Two levels autonomic resource management in virtualized IaaS", ELSEVIER, Future Generation Computer Systems, Pages 1319–1332, August 2013.

[14] Shridhar G.Damanal and G. Ram Mahana Reddy, "Optimal Load Balancing in Cloud Computing By Efficient Utilization of Virtual Machines", IEEE, Sixth International Conference on Communication Systems and Networks (COMSNETS), Pages 1-4, 6-10 Jan. 2014.

[15] Trieu C. Chieu, Ajay Mohindra, Alexei A. Karve and Alla Segal, "Dynamic Scaling of Web Applications in a Virtualized Cloud Computing Environment", IEEE International Conference on e-Business Engineering, Pages 281-286, 21-23 Oct. 2009, ISBN 978-0-7695-3842-6.

[16] Klaithem Al Nuaimi, Nader Mohamed, Mariam Al Nuaimi and Jameela Al-Jaroodi, "A Survey of Load Balancing in Cloud Computing: Challenges and Algorithms", IEEE, Second Symposium on Network Cloud Computing and Applications (NCCA), Pages 137-142, 3-4 Dec. 2012, ISBN 978-1-4673-5581-0.

[17] A.KHIYAITA, M.ZBAKH, H. EL BAKKALI, Dafir EL KETTANI, "Load Balancing Cloud Computing: State of Art", IEEE, National Days of Network Security and Systems (JNS2), Pages 106-109, 20-21 April 2012, ISBN 978-1-4673-1050-5.

[18] Martin Randles, David Lamb, A. Taleb-Bendiab,"A Comparative Study into Distributed Load Balancing Algorithms for Cloud Computing", IEEE 24th International Conference on Advanced Information Networking and Applications Workshops (WAINA), Pages 551-556, 20-23 Apr 2010, ISBN 978-1-4244-6701-3.

Gupta Shweta R. was born on 12th April 1992 at Ahmedabad in Gujarat, India. She received her Bachelor of Engineering in Computer Engineering from Hasmukh Goswami College of Engineering at Gujarat Technological University in 2013. Currently she is pursuing her Master of Engineering in Computer Science and Engineering from L. J. Institute of Engineering and Technology at Gujarat Technological University, India. Her area of Interest is Cloud Computing, Load Balancing, and Scalability in cloud.

Ms. Jasmine Jha, is working as an Assistant Professor at L. J. Institute of Engineering and Technology at Ahmedabad in Gujarat, India. She received her Bachelor of Engineering in Information Technology from Saurashtra University and Master in Technology in Web Technology from Gujarat University, India. Her research interest are Semantic web and Web Technologies.