International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com
Volume 4, Issue 2, No 1, 2016.
ISSN: 2348-6600
Reference ID: IJCS-119
PAGE NO: 701-706

Load Balancing in Cloud Computing: Various Issues & Challenges

Rupali Mishra[#1], Dr. Bhupesh Gour[*2]

[#]*Department of Computer Science*
Technocrats Institute of Technology, Bhopal
[1]rupali77mishra@gmail.com
[2]bhupesh_gour@rediffmail.com

Abstract— **Load Balancing is a technique of managing the traffic level of packets over cloud. There are many complex algorithm already implemented for managing or load balancing over cloud computing. Here in this paper a survey of all the existing load balancing technique which are implemented for managing the load traffic of packets over cloud is analyzed and discussed here. Hence by analyzing their various advantages and issues a new and efficient technique is implemented in future.**

***Index Terms*—Cloud Computing, Virtual Machines, Data Centers, Load Balancing, Scheduling.**

I. INTRODUCTION

Cloud Computing is nothing but a collected works of computing resources and services pooled together and is provided to the users on pay-as-needed basis for enabling ubiquitous, convenient, on-demand network access to a shared pool of organize able computing resources i.e. networks, servers, storage, applications and services that can be rapidly provisioned and freed with minimal organization attempt or service provider interaction" [1]. Cloud computing is an incorporated idea of parallel and distributed computing which shares resources like hardware, software and data to computers or other devices on require. With the help of cloud computing and internet capacity, the client can right of entry the abovementioned resources by paying for the length of use. Virtual machine (VM) is an implementation unit that acts as an establishment for cloud computing technology. Virtualization consists of conception, execution and organization of a hosting situation for different applications and resources. The VMs in the cloud computing environment divide resources like processing cores, system bus and so forth. The computing resources available for each VM are restrained by overall processing power. In this model of situation the job arrival prototype is random and also the potential of each virtual machine show a discrepancy from one another. Hence, load balancing turn out to be a critical task leading to a poor system presentation and sustaining stability. Thus, it turns out to be imperative to expand an algorithm which can make progress the system presentation by balancing the work load among virtual machines. There are different load balancing algorithms available, for example round robin (RR), weighted round robin(WRR), dynamic load balancing, Equally Spread Current Execution (ESCE) Algorithm, First Come First Serve, Ant Colony algorithm, and throttled algorithm. The most frequently used scheduling techniques for a no preemptive system are first in first out (FIFO) and weighted round robin (WRR) [2]. Sharing of the group of resources may initiate a problem of availability of these resources causing circumstances of deadlock. One way to keep away from deadlocks is to allocate the workload of all the VMs among themselves. This is called load balancing. The objective of balancing the load of virtual machines is to decrease energy consumption and make available maximum resource utilization in that way reducing the number of job rejections. As the numbers of customers are growing on the cloud the load balancing has develop into the challenge for the cloud provider.

Client-server systems were assumed to address this data-sharing test by providing centralized data management and processing servers. As business computing requires grew and the Internet became extensively accepted the primarily simple client-server architecture altered into more complex two-tier, three-tier, and four-tier architectures. Consequently, the complication and management costs of IT infrastructure have skyrocketed – even the costs of authentic software development in big organizations are characteristically lower than costs of software and infrastructure protection. For many enterprises, the long-standing dream has been to surroundings information technology issues and concentrates on core

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 2, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-119 PAGE NO: 701-706

business instead. Although the result of the cloud computing adoption is yet to be seen many companies believe that cloud computing may offer practical other model that may decrease costs and complexity while increasing prepared competence.

The OS was installed directly to hardware and a large amount of the servers hosted multiple functions within the same OS without providing physical or virtual isolation (see Figure 1). Because it was complicated to rapidly move and rebalance applications across servers, server resources were not developed most efficiently. Distributed functions, which were installed over multiple servers communicated with each other using CORBA or DCOM communication protocols over RPC. However, it was a main difficulty with such protocols that they were vendor-dependent and so the completion of one vendor might not be well-matched with that of others. This was explained at the beginning of the 2000s by the introduction of web services, which use open specifications that are language, platform and vendor agnostic.

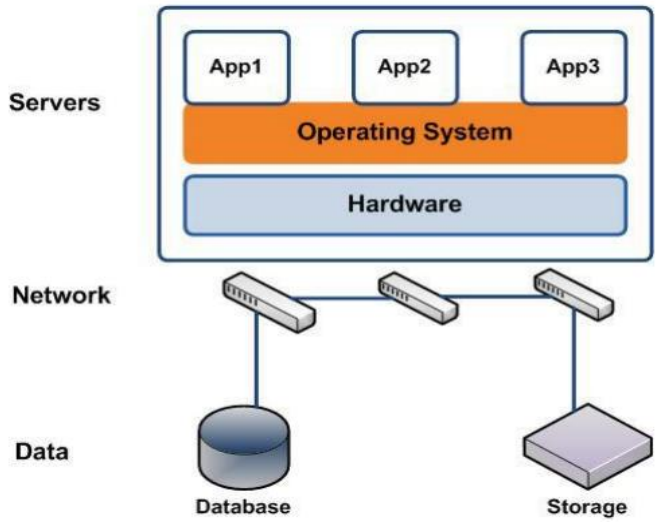

Figure 1: Servers without virtualization

With the beginning of virtualization things have transformed extremely. Virtualization recovers resource utilization and energy efficiency – helping to significantly decrease server maintenance transparency and on condition that quick disaster improvement and high accessibility. Virtualization has been very significant for cloud computing, because it separates software from hardware and so make available a method to rapidly reallocate applications across servers based on computational requires (see Figure 2). Virtualization was a main step towards cloud infrastructure; on the other hand, the service component was still absent. Virtualized environments managed by internal system administrators and by default virtualization platforms do not provide the abstraction layer that enables cloud services. To cloud-enable an environment a layer of abstraction and on-demand conditioning must be provided on top (see Figure 3). This service layer is a significant attribute of any cloud environment – it conceals the difficulty of the communications, and makes available a cloud-management interface to customers.

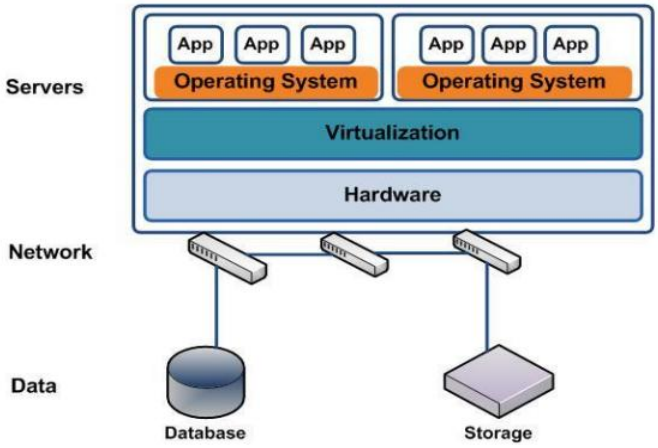

Figure 2: Virtualized servers

Load balancing is one of the essential issues to increase the working presentation of the cloud service provider. The advantages of distributing the workload comprises enlarged resource utilization ratio which additional show the ways to attractive the overall performance thereby accomplishing maximum client satisfaction [3].

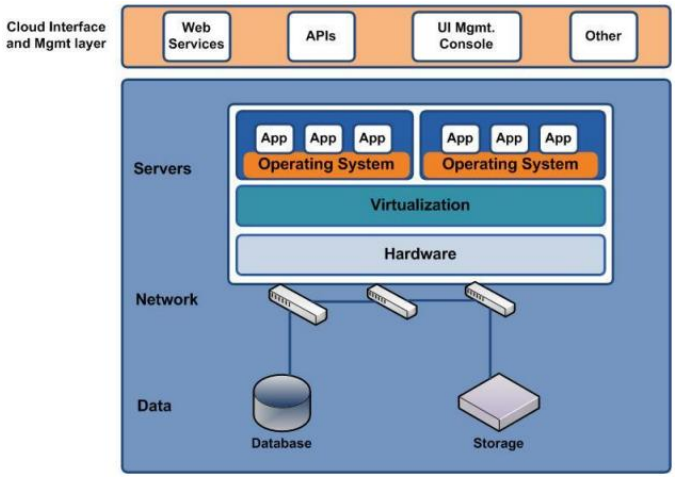

Figure 3: Simplified cloud infrastructure.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 2, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-119 PAGE NO: 701-706

In cloud computing, if customers are increasing load will also be amplified the increase in the number of customers will show the way to poor performance in terms of resource usage, if the cloud provider is not organized with any good quality method for load balancing and also the capability of cloud servers would not be used correctly. Cloud management interfaces i.e., the Amazon admin console give functions allowing customers to deal with a cloud lifecycle. Such as, customers can add new parts to the cloud such as servers, storage, databases, caches and so on. Customers can use the similar interface to check the health of the cloud and achieve many other procedures.

II. Literature Survey

Stuti Dave, Prashant Maheta proposed a new algorithm for Utilizing Round Robin Concept for Load Balancing Algorithm at Virtual Machine Level inn Cloud Environment [4]. The methodology implemented is simple in comparison and faster load balancing ability. The algorithm implemented works for only time shared environment in cloud.

Xin Lu, Zilong GU [5], in their paper have discussed that, by monitoring performance parameters of virtual machines in real time, the overloaded is easily detected once these parameters exceeded the threshold. Quickly finding the nearest idle node by the ant colony algorithm from the resources and starting the virtual machine can bears part of the load and meets these performance and resource requirements of the load. This realizes the load adaptive dynamic resource scheduling in the cloud services platform and achieves the goal of load balancing.

Zhongni Zheng, Rui Wang [6] did the research of using GA to deal with scheduling problem in the cloud, we propose PGA to achieve the optimization or sub-optimization for cloud scheduling problems. Mathematically, we consider the scheduling problem as an Unbalanced Assignment Problem. Future work will include a more complete characterization of the constraints for scheduling in a cloud computing environment, improvements for the convergence with more complex problems.

Sunita Bansal, Bhavik Kothari, Chitranjan Hoda [7] proposed a novel grid scheduling heuristic that adaptively and dynamically schedules task without requiring any prior information on the workload of incoming tasks. This models the grid system in the form of a state – transition diagram with job replication to optimally schedule jobs. This algorithm uses prediction information on processor utilization. In this algorithm they uses concept of job replication that is, a job can be replicated to other resource if that resource completes execution of current job than the resource it is currently allocated. This algorithm uses two types of queue namely, Waiting Queue and Execution Queue. This approach is based on exploiting information on processing capability of individual grid resources and applying replication on tasks assigned to the slowest processors. The approach facilitates replication of tasks, and also assigned to execute on slower machines, on machines with higher processing capacity. In this approach the communication costs are ignored. Experimental results show the better performance of this approach compared to traditional round robin algorithm.

Li Yang, ChengSheng Pan, ErHan Zhang, HaiYan Liu [8] proposed one kind of weighted fair scheduling algorithm. It is based on strict rob priority class which adds an absolute priority queue based on the foundation of based class weighted fair scheduling algorithm (CBWFQ). This algorithm covers the disadvantage of traditional weighted fair scheduling algorithm. Weighted Fair Scheduling algorithm differentiates the services of all active queues on the basis of weight of each business flow. When a new job arrives the classifier classifies the jobs into categories. Then buffer is checked for each category and if buffer is not overloaded then job is stored in the buffer otherwise job is dropped. Each job enters a different virtual queue. Weight, Dispatch, Discard and Rob are four main rules of this algorithm. The main advantage of this algorithm is that it has introduced the rob rule together with dropping rule. Experiments are done on NS-2 software to simulate SRPQ-CBWFQ algorithm. This new algorithm combined buffer management and queue scheduling and only guarantees low delay of real time applications. It also gave consideration to fairness and better utilization of buffers. This algorithm has two great advantages of bandwidth allocation and delay without throughput reducibility.

Shamsollah Ghanbari, Mohamed Othman [9] presented a novel approach of job scheduling in cloud computing by using mathematical statistics. This algorithm considers the priority of jobs for scheduling and named as priority based job scheduling algorithm. It is based on multiple criteria decision making model. A pair wise comparison based on multiple criteria and multiple attributes method was first developed by Thomas Saaty [10] in 1980 and named as Analytical Hierarchy Process (AHP). Consistent Comparison Matrix is the foundation of AHP, so to use the concept of AHP comparison matrices are computed according to the attributes and criteria's accessibilities.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 2, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-119 PAGE NO: 701-706

Agent based Priority Heuristic for Job Scheduling on Computational Grids [11] this algorithm presents an agent based job scheduling for effective and efficient execution of user jobs. This considers QoS parameters like waiting time, turnaround time, and response time, total completion time, etc. Priorities are assigned to the jobs under different classifications. Agent based Heuristic Scheduling (AHS) uses task agent for job distribution to achieve optimum solution. Task agent receives jobs from users and distributes them among different prioritize global queues based on user levels. AHS uses agent based job distribution strategy at global level for optimal job distribution based on user levels and job priorities at local levels for efficient and effective execution of jobs. For different global queues, priorities are defined as threshold levels for assigning jobs to global queues. If jobs have same priorities then jobs having minimum run time executes first otherwise First Come First Serve (FCFS) algorithm is used. AHS has optimal performance with respect to QoS parameters.

Design, Development and performance analysis of Deadline based Priority Heuristic for Job scheduling in a Grid [12]: A modified prioritized deadline based scheduling algorithm (MPDSA) is proposed using project management algorithm for efficient job execution with deadline constraint of user's jobs. MPDSA executes jobs with closest deadline time delay in cyclic manner using dynamic time quantum. It assumes each job to be described by its process_id, burst_time, arrival_time and deadline. Time quantum is assigned by computing LCM of all burst times. Then the jobs having minimum time delay is selected for execution. If jobs have same time delay then First Come First Serve (FCFS) algorithm is used for scheduling. Jobs are pre-empted based on time quantum and if a job completes its execution before time quantum, that job is deleted from queue. This algorithm satisfies system requirements and supports scalability under heavy workloads.

A two-stage-priority-rule-based algorithm for robust resource-constrained project scheduling [13]: The algorithm solves the resource-constrained project scheduling problem (RCPSP). This algorithm presents a two-stage algorithm for robust resource-constrained project scheduling. First stage solves the RCPSP for minimizing makespan by using a priority-rule-based heuristic. Second stage is intended to find most robust schedule with a makespan not larger than threshold value found in first stage. Both the stages are referred as two phases. In phase I, each iteration has three steps:

1. Priority values issued from selected priority rule.
2. Random-biased selection of eligible activities according to their selection probabilities.
3. Selected activities are scheduled to the resources.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com
Volume 4, Issue 2, No 1, 2016.
ISSN: 2348-6600
Reference ID: IJCS-119
PAGE NO: 701-706

S.NO	AUTHOR	YEAR	TECHNOLOGY USED	ADVANTAGES	ISSUES
1.	En-Juichang et.al.	2010	ACO-based Cascaded Adaptive Routing (ACO CAR)	Concept About Ant Colony Optimization	More memory utilization
2.	Le Shanguo et.al.	2011	Pheromone And available Resources On The Link As A Aspect For Load Balancing.	Load Balancing Is Anticipated.	Traffic Problems
3.	Marco Dorigoa et.al	2011	Concept for simulated network models between arbitrary nodes.	Traffic Problems with Network load imbalance are removed	Lower throughput
4.	Rata Mishra et.al.	2012	Pseudo Boolean Solution is Implemented.	Efficient than existing Methodology	Only one Pheromone is updated.
5.	Preeti Kushwah et.al.	2014	Load Balancing for Minimum Spanning tree implemented.	Compared Heuristic And Evolutionary Approach	Performance Degraded.
6.	Jia Zhao Et.al.	2016	Heuristic Clustering based on Bayes Theorm Implemented.	A Very New Concept Implemented.	Highly Complex Methodology.
7.	Kamil Krynicki	2015	Non-Hybrid Ant Colony Optimization Heuristic for Convergence Quality	Increased Efficiency and Flexibility with Adaptability.	Doesn't support multiclass resource querying.
8.	Chia-Feng Juang	2014	Cooperative Continuous ant Colony Optimization (CCACO).	Useful for Fuzzy Controller and Design Problems.	CCACO can't be applied to multiobjective FS design problems for optimization.
9.	Gengbin Zheng	2010	an automatic dynamic hierarchical load balancing method that overcomes the scalability challenges of centralized schemes and poor solutions of traditional distributed schemes.	Scalable and provides reduces memory.	Doesn't provide interconnect topology aware strategies that map the communication graph on the processor topology to minimize network contention.
10.	Jun Wang	2009	Machine learning based load prediction and fuzzy logic based replica management for adaptive and flexible load balancing mechanism within the framework of distributed middleware.	Fluctuate load can be easily managed and flexible.	More number of iterations to be performed.

International Journal of Computer Science
Scholarly Peer Reviewed Research Journal - PRESS - OPEN ACCESS
ISSN: 2348-6600
www.ijcsjournal.com Volume 4, Issue 2, No 1, 2016. ISSN: 2348-6600
Reference ID: IJCS-119 PAGE NO: 701-706

III. Conclusion

The Various Load Balancing and Scheduling techniques are discussed and analyzed here. Since Load Balancing techniques requires a lot of computational overhead lot of techniques are implemented to solve the issues. Here by analyzing these techniques their various advantaged and limitations an efficient technique for Load Balancing can be implemented in future.

References

[1] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," IEEE Transactions on Parallel and Distributed Systems,vol.24, no. 6, pp. 1107–1117, 2013.

[2] L.D. Dhinesh Babu and P. Venkata Krishna, "Honey beebehavior inspired load balancing of tasks in cloud computing environments," Applied Sot Computing Journal, vol.13,no.5,pp. 2292–2303, 2013.

[3] Ali M Alakeel, "A Guide To Dynamic Load Balancing In Distributed Computer Systems", International Journal of Computer Science and Network Security, Vol. 10 No. 6, June 2010.

[4] Stuti Dave, Prashant Maheta," Utilizing Round Robin Concept for Load Balancing Algorithm at Virtual Machine Level inn Cloud Environment", INTERNATIONAL JOURNAL OF COMPUTER APPLICATIONS · MAY 2014.

[5] Xin Lu, Zilong Gu, "A LOAD-ADAPATIVE CLOUD RESOURCE SCHEDULING MODEL BASED ON ANT COLONY ALGORITHM", 2011, 978-1-61284-204-2/11, Proceedings of IEEE CCIS2011, Page no: 296-300.

[6] Zhongni Zheng, Rui Wang, Hai Zhong, Xuejie Zhang, "An Approach for Cloud Resource Scheduling Based on Parallel Genetic Algorithm", 2011, 978-1-61284-840-2/11, IEEE, Page no: 444-447.

[7] Sunita Bansal et al, Dynamic Task-Scheduling in Grid Computing Using Prioritized Round Robin Algorithm, IJCSI International Journal of Computer Science Issues, 8(2)(2011) 472-477.

[8] Li Yang et al, A new Class of Priority-based Weighted Fair Scheduling Algorithm, Physics Procedia, 33 (2012) 942 – 948.

[9] Ghanbari, Shamsollah, and Mohamed Othman. "A Priority based Job Scheduling Algorithm in Cloud Computing." Procedia Engineering 50 (2012): 778-785.

[10] T.L. Saaty, The Analytic Hierarchy Process,(New York 1980).

[11] Syed Nasir Mehmood Shah, "Agent Based Priority Heuristic for Job Scheduling on Computational Grids" Proceedings of the International Conference on Computational Science, ICCS 2012.

[12] Haruna Ahmed Abba Design, Development and Performance Analysis of Deadline Based Priority Heuristic for Job Scheduling on a Grid, High Performance Computing Centre, Universiti Teknologi Petronas, 31750, Tronoh, Malaysia Procedia Engineering 50:397–405. DOI: 10.1016/j.proeng.2012.10.045

[13] Hédi Chtourou, "A two-stage-priority-rule-based algorithm for robust resource-constrained project scheduling, Computers & Industrial Engineering Volume 55, Issue 1, August 2008, Pages 183–194.