IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LOAD BALANCING ON CLOUD COMPUTING

[1]Ms. Homa Rizvi, [2] Dr. Habib Ur Rahman
[1]Research scholar, [2]Associate Professor,
Computer Science Engineering department
Kanpur Institute of Technology, Kanpur, India

Abstract: Load unbalancing problem is a multi-variant, multi-constraint problem that degrades performance and efficiency of computing resources. Load balancing techniques cater the solution for load unbalancing situation for two undesirable facets-overloading and under-loading. In contempt of the importance of load balancing techniques to the best of our knowledge, there is no comprehensive, extensive, systematic and hierarchical classification about the existing load balancing techniques. Further, the factors that cause load unbalancing problem are neither studied nor considered in the literature. This paper presents a detailed encyclopedic review about the load balancing techniques. The advantages and limitations of existing methods are highlighted with crucial challenges being addressed so as to develop efficient load balancing algorithms in future. The paper also suggests new insights towards load balancing in cloud computing.

***Index Terms* - Cloud computing, Taxonomy, Classification, Cloud service consumer, Cloud service provider, Quality of service, Load unbalancing, Load balancing**

I. INTRODUCTION

Cloud Computing is an internet-based network technology that shared a rapid growth in the advances of communication technology by providing service to customers of various requirements with the aid of online computing resources. It has provisions of both hardware and software applications along with software development platforms and testing tools as resources. Such a resource delivery is accomplished with the help of services. While as the former comes under category of Infrastructure as a service (IaaS) cloud, the latter two comes under headings of Software as a service (SaaS) cloud and platform as a service (PaaS) cloud respectively. The cloud computing is an on-demand network enabled computing model that share resources as services billed on pay-as-you-go (PAYG) plan. Some of the giant players in given technology are Amazon, Microsoft, Google, SAP, Oracle, VMware, Sales force, IBM and others. Majority of these cloud providers are high- tech IT organizations. The cloud computing model is viewed under two different headings. The first one is the service delivery model, which defines the type of the service offered by a typical cloud provider. A cloud computing model is efficient if its resources are utilized in best possible way and such an efficient utilization can be achieved by employing and maintaining proper management of cloud resources. Resource management is achieved by adopting robust resource scheduling, allocation and powerful resource scalability techniques. These resources are provided to customers in the form of Virtual Machines (VM) through a process known as virtualization that makes use of an entity (software, hardware or both) known as hypervisor. The greatest advantage of cloud computing is that a single user physical machine is transformed into a multiuser virtual machine. The Cloud Service Provider (CSP) plays a crucial role in service delivery to users and is a complex task with given available virtual resources. While serving user requests, some VMs will get a heavy traffic of user tasks and some will get a lesser traffic.

The remaining paper is structured as follows. Section “Load balancing model background” features a brief description about load balancing model in cloud computing. Section “Research methodology” highlights some related works. The research methodology is discussed in section “Research methodology”. Section “Proposed classification of load balancing algorithms” proposes taxonomy-based classification. The results are evaluated in section “Results and discussion” while section “Discussion on open issues on load balancing in cloud computing” discusses upon open issues in cloud load balancing. Finally, section “Conclusion and future work” concludes our work and points out some future directions.

Load balancing model background

In this section a two-level load balancing architecture model is presented in imbalanced clouds for achieving best load shedding as shown in Fig. 1 which is a modified architecture given by Gupta et al. The virtual machine manager and virtual machine monitor are abstracted in this model. The first level load balancing is performed at the Physical Machine (PM) level and the second level is performed at the VM level. Based on this, there are two task migration sets;

1. Intra VM task migration
2. Inter VM task migration

The request generator generates user requests which are user tasks that need computing resources for their execution. Data centre controller is in-charge of task management. The load balancer checks which VM to assign for a given user task. The first level load balancer balances the given workload on individual Physical Machines by distributing the workload among its respective associated Virtual Machines. The second level load balancer balances the workload across different Virtual Machines of different Physical Machines.

Activities involved in load balancing

Scheduling and allocating tasks to VMs based on their requirements constitute the cloud computing workload. The load balancing process involves the following activities

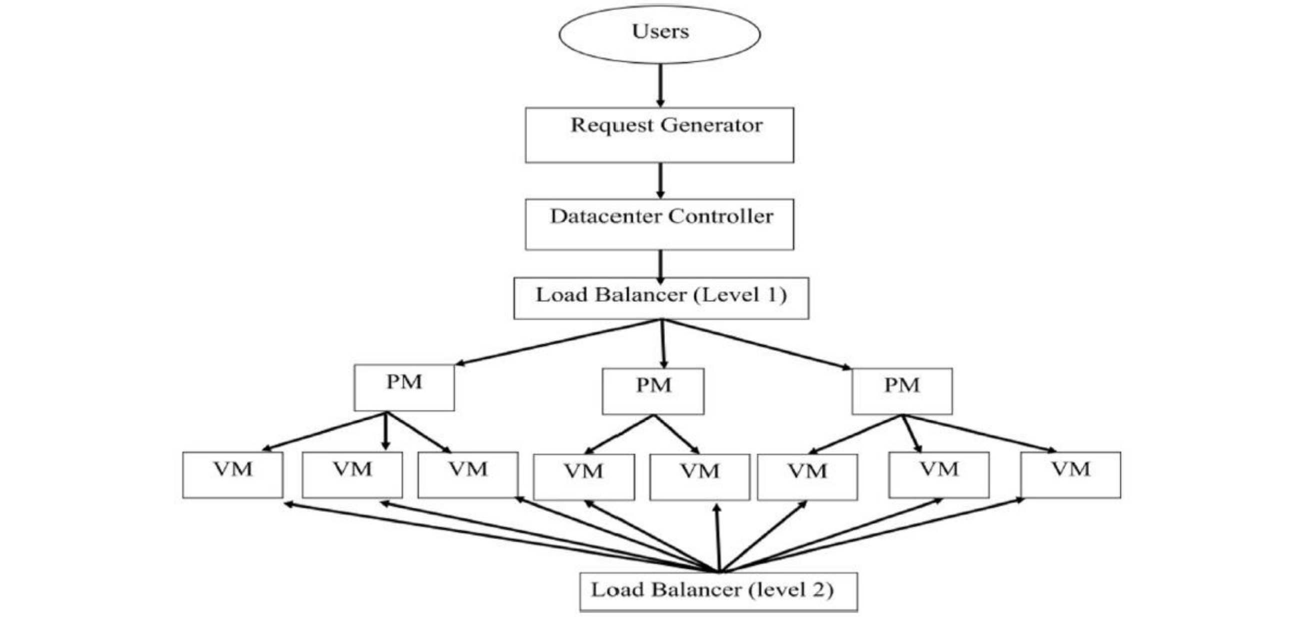


Fig. 1 Two level Load Balancing Architecture

II. LITERATURE REVIEW

This section includes the literature review of this paper. Firstly, a review of the existing review articles will be provided. Then the concept of load balancing will be explained highlighting its model, metrics, and existing common algorithms. Leading to recent literature on Load Balancing where proposed algorithms by researchers are explained and analysed. Followed by new algorithms proposed by recent researchers in the field of Load Balancing. There are still a few limitations in the existing review papers, for example, authors have reviewed the recent state of art however, it is a limited explanation as it does not include a comparative analysis. This is an essential feature when doing reviews as it provides deep analysis of the articles and makes it easy for readers to identify areas for improvement in future research. Most review papers do not include and thus it does not provide operational flow of the reviewed algorithms. Many papers lack the explanation and evaluation of the performance metrics used in the reviewed articles which is another contribution of the current review paper.

III. METHODOLOGY

To go deep into roots of load balancing process as to what causes load unbalancing problem a proper research methodology was followed. The literature survey was conducted in accordance with general research strategy that outlines the way in which load unbalancing problem is undertaken and identifies the methods, theories, algorithms, approaches and paradigms used in it. The load unbalancing problem was studied in accordance with constructive generic framework (CGF) methodology where it is broken down into sub- processes i.e., the factors, variables and parameters that are associated with load balancing. Further the literature study was enhanced by following the research guideline for Systematic Literature Review (SLR) as contemplated by Kitchenham with a special focus on research related to load balancing mechanism in cloud. An SLR is a repeated research method that can be replicated by other researchers to explore more knowledge. In order to feature the importance of load balancing in cloud computing, a set of questions were framed to address the key issues and challenges in load unbalancing. Question identification A set of questions were identified from literature survey that need to be answered before going into the load balancing process. Some of the questions have been answered in literature while others are not.

Proposed classification of load balancing algorithms

In this section load balancing algorithms are classified based on various criteria. A top-down approach is proposed and followed in classification process. The limitation of existing review papers is that there is no proper and significant hierarchical taxonomical classification of load balancing algorithms which makes it quite difficult to identify where a particular algorithm holds its place in taxonomy. The various criteria used for classification purpose include 'nature of algorithm', 'state of algorithm', 'trait used for load balancing', 'type of load balancing', and 'technique used in load balancing'. For the first time in literature an in-depth analysis of the LB algorithms has been achieved in this work which the previous studies were lacking. Based on nature of algorithm, the load balancing algorithms are either proactive or reactive.

Nature of the algorithm

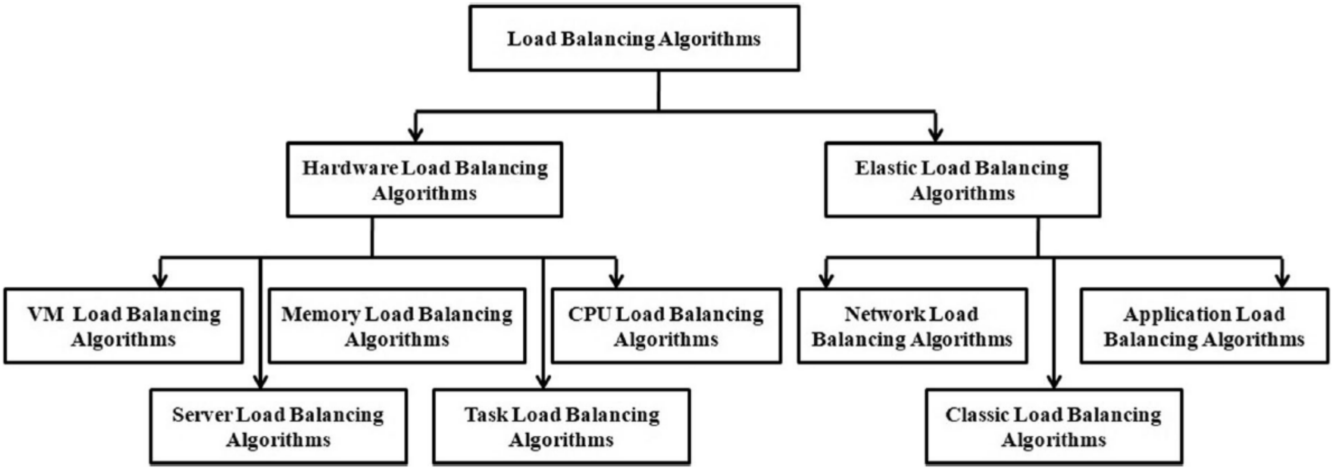
The first categorization of load balancing algorithms in this work has been done on the basis of nature of algorithm. On the basis of this classification, LB algorithms are classified as proactive based approaches and reactive based approaches. However, in other fields of technology particularly in the communication and networking for mobile adhoc networks (MANETS), the nature of the communication routing protocols has been extensively studied under these two variants.

State of the algorithm

On the basis of state information of system that an algorithm relies on, LB algorithms are widely classified as static, dynamic and hybrid. From existing literature survey, it is evident that this is most widely used classification system for LB algorithms. Majority of work on comparative studies on load balancing begin the algorithmic taxonomy by placing this category on top of taxonomy. In static load balancing, traffic load is segregated uniformly across the servers. This is done by algorithm having the prior knowledge about system resources and task requirements. The static LB algorithm schedules tasks to VM for execution at compile time. The advantage of static algorithm is their less complexity but they suffer from a fatal bottleneck of being unable to move tasks during execution in progress to another machine for load balancing.

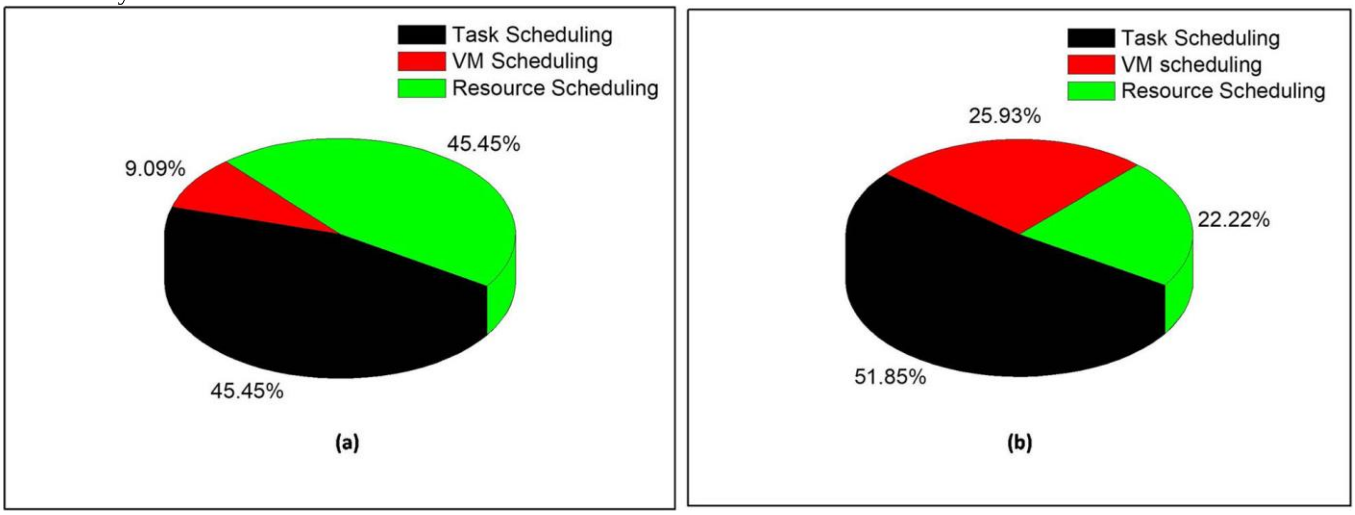
Technique used in load balancing

On the basis of the technique used, load balancing algorithms are classified as heuristics and meta heuristics techniques, and optimization techniques. A heuristic approach is an approach to problem solving accounting a practical method or methodology guaranteed not be optimal, perfect, logical or rationale but sufficient to reach an immediate goal. Finding an optimal solution may be impossible or impractical particularly to load balancing which is a NP hard problem and heuristics play an important role to gear up the process of finding a decent solution. Heuristic methods are designed following strategies derived from previous experience with similar problem sets. Heuristics play a crucial role in load balancing process to sort up various issues faced by CSPs. A lot of research work has been carried out with heuristic and meta-heuristic approaches in cloud load balancing and as such we have classified the heuristic and meta-heuristic methods into nature
inspired algorithms and classical algorithms.


Load Balancing Algorithms
Hardware Load Balancing Algorithms
Elastic Load Balancing Algorithms
VM Load Balancing Algorithms
Memory Load Balancing Algorithms
CPU Load Balancing Algorithms
Network Load Balancing Algorithms
Application Load Balancing Algorithms
Server Load Balancing Algorithms
Task Load Balancing Algorithms
Classic Load Balancing Algorithms

IV. RESULTS AND DISCUSSION

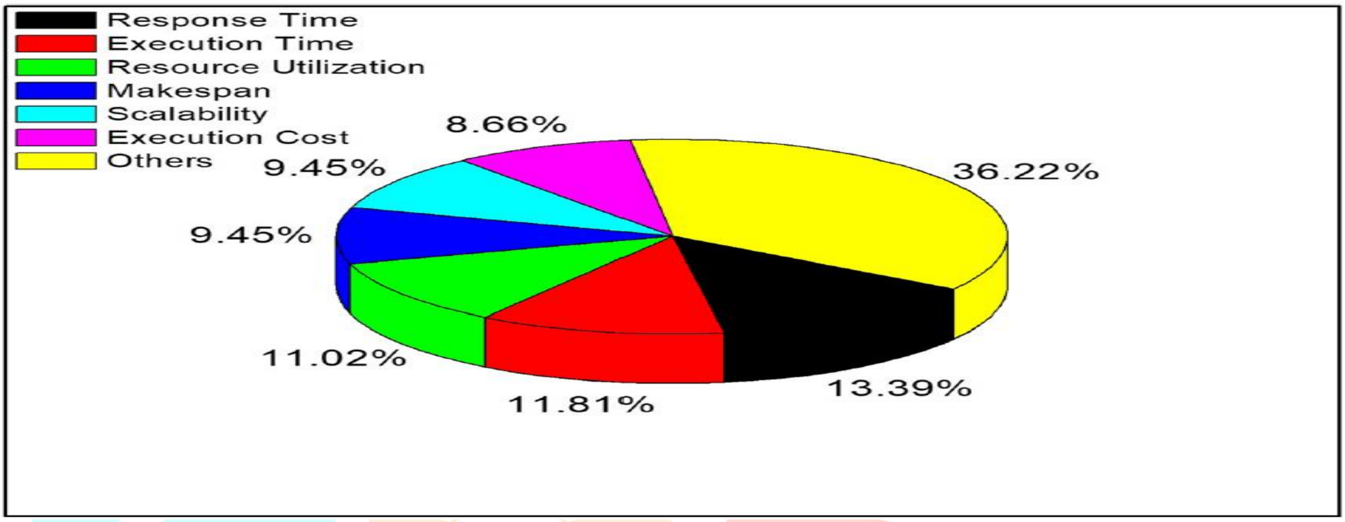
This section outlines the results achieved from comparative analysis of different load balancing approaches in cloud computing. Figure shows the percentage of various scheduling types in proactive based load balancing approaches. It is clear that task scheduling and resource scheduling each with 45.45% contribution are more often considered in proactive based approaches with less attention towards VM scheduling which contribute 9.09%. it is evident that most of the reactive approaches in existing literature have been studied under task scheduling which amount to 51.85%, followed by VM scheduling which contribute to 25.93% and resource scheduling which contribute to 22.22% respectively. describes the percentage of research articles on cloud load balancing defining the algorithmic complexity. It is calculated that 80% of research articles did not consider algorithmic complexity in their work while only 20% define it in their work. It is analysed that proactive approaches are always dynamic in nature while depicts that most of the reactive approaches fall under dynamic state of algorithm which contribute to 68%, followed by static algorithm which contributes 20% and hybrid algorithm which represent 12%. It is evident from Figure that 60% of proactive approaches are multi-objective while 40% are single objective approaches. Likewise, 56% of reactive approaches are multi-objective while 44% are single objective approaches as depicted figure. And it displays the testing environment on which a



particular approach was run to evaluate the performance metrics. It is clear that Cloud-Sim simulator is extensively used for conducting simulation experiment constituting 33.33% of experimental implementation followed by Cloud Analyst simulator with 19.44% of experimental implementation. C and C++, MATLAB implementation of load balancing approaches amount to 11.11% each respectively while others constitute 19.44%.

V. CONCLUSION

The work presents a comparative study on load balancing approaches in reviewed articles. The problem of load unbalancing in cloud computing was discussed along with driving factors that lead to this problem. An abstracted load balancing model was briefly discussed together with activities involved in load balancing process. A proper research methodology was followed in which the problem was studied in guidelines with Constructive Generic Framework (CGF) further reinforced by Systematic Literature Review (SLR) methodology. We framed a set of problem related questions and discussed them in the work. The data collected for this study had been gathered from five reputed potential databases that include IEEE Xplore digital library, Science Direct, ACM digital Library, Springer and Elsevier. The data search process was assisted by different tools and advanced filter options. The data was collected for the period from 2010 to June 2018. A multilevel taxonomy-based classification was proposed in this work where the classification process is done on five criteria.



REFFERNCES

1. Pradhan P, Behera PK, Ray BNB (2016) Modified round Robin algorithm for resource allocation in cloud computing. Proceed Comp Sci 85:878–890
2. Dam S, Mandal G, Dasgupta K, Dutta P (2015, February) Genetic algorithm and gravitational emulation-based hybrid load balancing strategy in cloud computing. In: Proceedings of the 2015 third international conference on computer, communication, control and information technology (C3IT), pp 1–7
3. Dave A, Patel B, Bhatt G (2016, October) Load balancing in cloud computing using optimization techniques: a study. In: International Conference on Communication and Electronics Systems (ICCES), pp 1–6
4. Gupta H, Sahu K (2014) Honey bee behaviour-based load balancing of tasks in cloud computing. Int J Sci Res 3(6)
5. Mishra SK, Puthal D, Sahoo B, Jena SK, Obaidat MS (2017) An adaptive task allocation technique for green cloud computing. J Supercomp 405:1–16
6. 18. Ibrahim AH, Faheem HEDM, Mahdy YB, Hedar AR (2016) Resource allocation algorithm for GPUs in a private cloud. Int J Cloud Comp 5(1–2):45–56
7. Jebalia M, Ben Letafa A, Hamdi M, Tabbane S (2015) An overview on coalitional game-theoretic approaches for resource allocation in cloud computing architectures. Int J Cloud Comp 4(1):63–77
8. Noshy M, Ibrahim A, Ali HA (2018) Optimization of live virtual machine migration in cloud computing: a survey and future directions. J Netw Comput Appl:1–10
9. Gkatzikis L, Koutsopoulos I (2013) Migrate or not? Exploiting dynamic task migration in mobile cloud computing systems. IEEE Wirel Commun 20(3):24–32
10. Jamshidi P, Ahmad A, Pahl C (2013) Cloud migration research: a systematic review. IEEE Trans Cloud Comp 1(2):142–157 Raviteja S, Atmakuri R, Vengaiah C (2017) A review on cloud computing migration and issues Shamsinezhad E, Shahbahrami A, Hedayati A, Zadeh AK, Banirostam H (2013) Presentation methods for task migration in cloud computing by combination of Yu router and post-copy. Int J Comp Sci Iss 10(4)
11. Ghomi EJ, Rahmani AM, Qader NN (2017) Load-balancing algorithms in cloud computing: a survey. J Netw Comput Appl 88:50–71 Milani AS, Navimipour NJ (2016) Load balancing mechanisms and techniques in the cloud environments: systematic literature review and future trends. J Netw Comput Appl 71:86–98
12. Kalra M, Singh S (2015) A review of metaheuristic scheduling techniques in cloud computing. Egypt Inform J 16(3):275–295
13. Mesbahi M, Rahmani AM (2016) Load balancing in cloud computing: a state-of-the-art survey. Int J Mod Educ Comp Sci 8(3):64

14. Kanakala VR, Reddy VK, Karthik K (2015, March) Performance analysis of load balancing techniques in cloud computing environment. In: 2015 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), pp 1–6
15. Shah JM, Kotecha K, Pandya S, Choksi DB, Joshi N (2017, May) Load balancing in cloud computing: methodological survey on different types of algorithms. In: 2017 International Conference on Trends in Electronics and Informatics (ICEI), pp 100–107
16. Neghabi AA, Navimipour NJ, Hosseinzadeh M, Rezaee A (2018) Load balancing mechanisms in the software defined networks: a systematic and comprehensive review of the literature. IEEE Access 6:14159–14178