WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Plan &Implementing Load adjusting strategies in Cloud processing using User Priority based Scheduling

Ravikumar Ch

Computer Science and Engineering, Lovely Professional University, Punjab,India
E-mail: chrk5814@gmail.com,

ABSTRACT

The recent paradigm shift from traditional computing to Internet based computing known as cloud computing has dramatically changed the way of storage and computations. Effective load balancing algorithms is indispensable to safeguard interests of customers and bring equilibrium into the cloud eco-system. Hence, there is a need to develop effective load balancing algorithms. This thesis focuses on certain problems like optimization of performance, maximum throughput, minimization of makespan, efficient resource utilization in load balancing that needed further exploration and Solution. The cloud partitioning based solution is designed for load balancing, where each partition maintains status like IDLE, NORMAL and OVERLOADED was found to be effective but it leads to sub optimal performance unless there is an ideal refresh period designed by certain rules. A new algorithm known as User Priority based Scheduling for Load Balancing is proposed. It characterizes user priorities by dividing the given tasks into two categories namely elastic and non-elastic.

Keywords – Cloud Computing, Load Balancing, ACO, overloaded cloud, cloud architecture.

1. Introduction

The aim of the present research is to design and implement enhanced load balancing techniques in cloud computing to leverage its benefits and bring equilibrium between expectations of consumers and benefits to service providers. Before explaining the research further, here is the information to understand the context of the research. The term computing has gone tremendous changes of late. It started with a standalone computer where a single computer is made responsible to provide storage, processing and other services. In the 1960's the invention of network ARPANET originally paved way for rapid growth in different kinds of networks including Internet and their associated technologies. It is considered to be precursor to the information super highway, the Internet. Using network, the standalone computing is transformed and coupled with centralized computing where a central server was made responsible for everything and clients were dumb in nature. Further innovations and inventions in hardware and software technologies led to novel approaches in networking and investments on hardware are minimized [1-2]. Client/Server computing became a reality with a smart client with their own storage and processing power. Later on it is further enhanced to have another kind of computing that has transformed enterprises and the way the applications reflected real world scenarios as closely as possible. In the next level Distributed Computing is an interaction among servers with plethora of advantages like fault tolerance, location transparency, scalability, availability and so on. The distributed computing has got thousands of use cases in the real world. They include net banking, Supply Chain Management (SCM), Customer Relationship Management (CRM), insurance network, telecommunications, integration of healthcare units and integration of criminal investigation agencies, integration of universities to mention few. Due to phenomenal growth in distributed computing led to two specific computing scenarios known as grid computing and cloud computing.

Grid Computing is goal oriented connectivity among different distributed computing resources providing required services while the Cloud is Internet based, service oriented and a huge pool of computing resources are provided to public in pay as you use fashion. Cloud computing technology enables micro, small, medium and large organizations and even individuals to gain access to computing resources in pay as you use fashion without any need of capital investment on resources. In other words, for organizations it became easier to manage with the cloud resources as they can minimize investment and focus on operations. As cloud computing is on top of virtualization technology, it became inexpensive. Utilizing virtual versions of hardware and software cloud computing renders affordable and sustainable services. There are cloud service providers (CSPs) like Amazon, Microsoft, IBM and Google to mention few who render services to their clients. There will be Service Level Agreements (SLAs) that are strictly followed to safeguard interests of clients. In this context, it is essential that CSPs provide quality services and also gets financial benefits. When more resources are reserved for clients, it may lead to wastage of resources. When fewer resources are reserved, it leads to scarcity and failure in meeting SLAs. To strike balance between the two, CSPs are using dynamic resource allocation and interrelated with load balancing. When load is balanced across cloud servers, it is possible to optimize resource utilization and ensure equilibrium and satisfaction between consumers and service providers. The problem of load

balancing is examined in this thesis. The existing load balancing algorithms are to be optimized further in order to have better scalability and reliability of cloud services without jeopardizing interests of consumers [3-6]. This is very challenging as cloud computing environment is dynamic in nature. A systematic approach or methodology (described in Chapter 3) is employed to design and implement improved load balancing mechanisms that enhance the ability of cloud servers and render services to clients as per desired SLAs. There is relationship between resource allocations and scheduling with load balancing. It is explored as well to fine tune load balancing mechanisms.

Cloud Computing According to National Institute of Standards and Technology

(NIST) cloud computing is a novel model of computing with certain characteristics such as on demand self-service, broad network access, resource pooling, rapid elasticity, and measured service [29]. It has service models pertaining to software, infrastructure, platform and deployment models in terms of private cloud, public cloud, community cloud and hybrid cloud [29]. As mentioned earlier, cloud computing is service oriented, provides different kinds of services. In the same fashion different deployment models of cloud are made available to meet the needs of individuals/organizations of different size. Before looking into the service layers and deployment models. The benefits include low cost, flexibility, collaboration, software update, mobility, data loss prevention, security and productivity. An architecture of cloud computing is shown in Fig. 1.

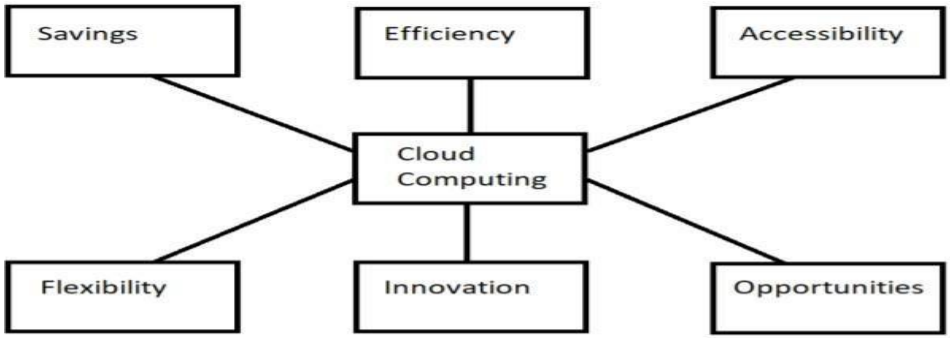


Fig. 1 – Cloud computing architecture

Computing through Internet minimizes cost, as organizations need not to invest on the computing resources or infrastructure. It is flexible as you just need a Personal Computer (PC) with Internet connection to get connected to cloud without geographical and time restrictions. The cloud computing technology has provision for storage and data loss prevention with its redundant storage mechanisms that are fail-proof in nature [7]. There is no data loss problem due to this service. Different levels of security are provided by CSPs to make outsourced data secure and confidential. When cloud computing is in place, productivity is more. Often it is very easily realized. For instance, it is possible to set development environment with cloud in few minutes saving lot of time and investment. Therefore productivity is the best known advantages of cloud computing. There is an important benefit of using cloud computing is the elasticity nature of cloud that scales up to the needs of the world from time to time [8]. This is the main advantage that let organizations prefer to use cloud resources and increases chances ofsuccess.

1.2. Load Balancing Traditionally load balancing is the process of distributing incoming traffic to a number of servers geographically located, probably, across the globe. Load balancer is the program that serves this purpose. However, with the emergence of cloud computing and use of millions of

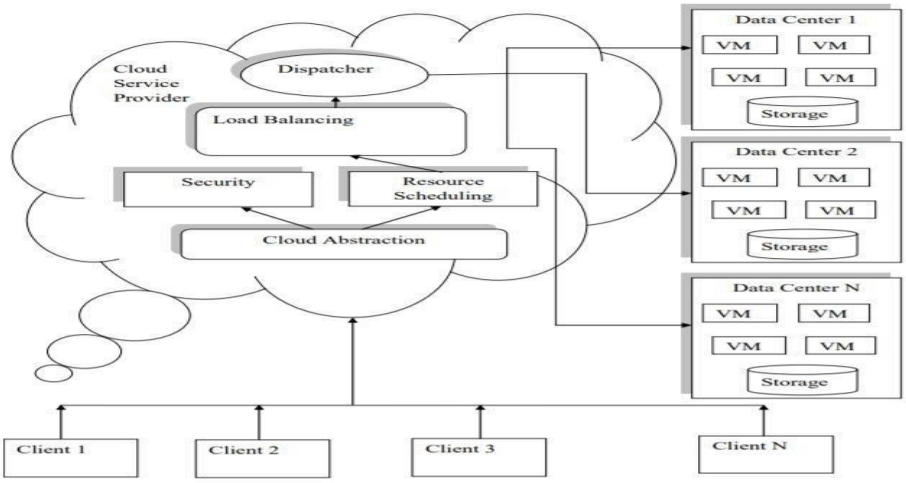


Fig. 2 – Load balancing in cloud computing

commodity servers as part of cloud, the load balancing has essentially become distributed in nature. The workload of incoming requests and the computing resources available in cloud are considered to make balancing decisions. Towards this end, distributed load balancer is the program on top of distributed computing technology is required. Load balancing in cloud computing allows CSPs to efficiently manage their applications and workload by allocating resources to thousands of commodity servers associated with cloud computing. In other words, load balancing involves hosting the distribution of workload traffic and ensures that the increasing demands are served with scalability and availability. Cloud computing servers provided by CSPs can bestow optimal results when their load is balanced. If not, some servers are overloaded with work while others are kept idle besides leading to failures in meeting SLAs requirements. This is the need for balancing load. Load balancing mechanisms plays a crucial role in balancing load and ensure that cloud SLAs are not violated and users' requests are processed without denying services. Since cloud computing is one form of distributed computing, it operates in distributed environment where millions of commodity computers or worker nodes function to fulfill the expectations of cloud. Balancing the load of different cloud servers which are located geographically across the globe is therefore inevitable. Load balancing in cloud is shown in Fig. 2.

1.3 Scheduling Scheduling is the process of efficient utilization of resources in cloud. It includes allocation of jobs or tasks to different VMs and controlling them. It leads to better resource utilization and also results in reducing time taken. Network traffic and CPU utilization will get reduced with proper scheduling techniques. Since cloud has heterogeneity in operating systems and resources, scheduling is one of the challenging phenomena. It has to consider optimal resource utilization and reduce the cost, since cloud resources are used in pay per use fashion. It needs to bring equilibrium, satisfaction between the user and CSP with prospecting for increased number of users. It is challenging in highly dynamic environment. Some of the existing scheduling algorithms in cloud include Round Robin (RR), Priority Queue (PQ), Multi-Level Queue (MLQ), Multi-Level Feedback Queue (MLFQ) and First Come First Serve (FCFS). Scheduling in cloud becomes more difficult when jobs do have associated with deadlines and the users have Server Level Agreements (SLAs) with CSP. Therefore, the research on scheduling is an open issue that needs continuous enhancement. It is possible to have better load balancing with ideal scheduling.

Need For Improvements In Load Balancing It is an established phenomenon in cloud. The rationale behind this is; that cloud servers are to be utilized optimally and to share load across millions of cloud servers is very essential. Otherwise it leads to let down in meeting SLAs and the CSPs loose customers. The cloud computing resources need some sort of elasticity to work fine with the growing demand with increased number of requests. It is also essential to ensure that the load is balanced. It needs a load balancing service that can handle sudden increases in the number of requests. When sudden burst of requests are found, load balancing and dynamic resource allocation plays a vital role in cloud for optimum resources utilization and balancing load for efficient processing of millions of concurrent requests. As explored in [1], [31], [4] different problems studied in existing load balancing algorithms. For instance cloud partitioning based approaches can be optimized further by using partitioning rules and refresh period. Ant Colony 10 Optimization (ACO) based approaches need to consider for improvements in the function of load balancing. User priority based approaches are needed in order to have better optimization to meet SLAs for consumer satisfaction. It is also essential to have mechanisms to handle non pre-emptive dependent tasks in cloud computing. By considering all these optimization problems, it is inevitable to have further research that can leverage consumer service provider equilibrium is satisfaction.

State-Of-The-Art Many contributions are found in the literature on load balancing. The presence of skewness in the usage of resources emphasizes the need for load balancing. Balancing the load aid to optimize the resource allocation and thus progress in effectiveness of cloud computing. For all the requests that arrive at cloud servers, it is essential to determine the server that can process given requests.

Generally, this task is to be helped by load balancing programs. The programs can also help in processing requests in large scale. There are few implementations of load balancing using CoudSim framework [3]. Xu et al. [1] devised an algorithm that is used to handle the limitation of load balancing using cloud partitioning approach. In this approach every partition is associated with status related to load. The status message normal indicates normal load to the VMs of the partition. The status IDLE refers to the fact that the partition is IDLE while the status OVERLOADED represents that the partition is unable to accommodate new jobs till existing jobs are get completed. Queues are used appropriately to handle requests and the status gets updated periodically. In order to optimize this kind of load balancing it is essential to have ideal refresh period governed by certain rules which is carried out in this research. Another important algorithm is Ant Colony Optimization (ACO) which is a random search algorithm as explored. A modified ACO algorithm was proposed which need further optimization for effective load balancing. Chen et al. [4] studied scheduling with user priority guidance. When user priority is known, it will be better for scheduling algorithms. They proposed Priority 11 Aware- Load Balance Improved Min-Min (PA- LBIMM) scheduling algorithm. It was found to be better than its predecessor Load Balance Improved Min-Min algorithm, as they latter does not support user priorities. Make span is the factor used to estimate the performance of these 2 algorithms. Round robin task scheduling and its variants are studied in [2] which needed further enhancement to work well with non-pre-emptive tasks for load balancing.

2. Literature Review

Cloud computing provides shared pool of computing resources on demand in pay-as- you-go fashion. Load balancing (LB) techniques in cloud

computing plays a essential role in satisfying Service Level Agreements (SLAs) made between Cloud Service Provider (CSP) and cloud users. LB speed up the execution process in cloud computing. When the users executed the tasks in compliance with the expected service criteria, they get satisfied. On the other hand, it also benefits the CSP. Towards this end, load balancing is indispensable phenomenon which ensures that resources are optimally utilized. When cloud computing resources are used appropriately and wastage is eliminated, it results in efficiency and there will be equilibrium between the customer satisfaction and profitability to CSP. The aim of the thesis is to analyze and design load balancing techniques in cloud computing to achieve the aforementioned equilibrium. This chapter reviews literature on the present state of the art of LB in cloud computing. It covers approaches such as cloud partitioning, Ant Colony Optimization (ACO), user priority based scheduling algorithms for LB, weighted round robin and other techniques with respect to throughput, make span and efficient utilization ofresources.

Cloud Partitioning Based Load Balancing Cloud partitioning (CP) is the phenomenon in which the cloud is partitioned into number of portions so as to make load balancing efficient. When the cloud space is divided, it is possible to exercise better control over it.

Authors [1] explored CP approach with different switching mechanisms to optimize resource utilization and service provisioning. CP based LB model has provision to have different strategies such as IDLE, NORMAL and OVERLOADED. These are actually status of the partition that is used to make well informed decisions. If there is no job being processed by a partition, it is said to be in IDLE state. In the same fashion, when a partition is being used for processing but its load is normal, and then it is called NORMAL state. On the other hand, the partition that is reaching its full capacity in processing and all its resources are being utilized, it is known as OVER- LOADED state. These states help explicitly in load balancing. However, the refresh time is not yet calculated to have ideal means of refreshing and further optimization. This gap is actually filled by this thesis.

Authors [2-4] analyzed cloud partitioning approach for load balancing. Their specified model divides the cloud resources into partitions and in each partition a load balancer is defined. The load balancers analyze the load on each partition and make strategic decisions in job scheduling which results in the optimization of cloud computing performance.

Authors [5] on the other hand studied various load balancing approaches. They classified load balancing approaches in to static and dynamic. The dynamic LB algorithms are further categorized as distributed and non-distributed. The static algorithms are classified into deterministic and probabilistic. They also examined on different strategies for load balancing. They include information strategy, triggering strategy, transfer strategy and location strat- egy. They also threw light into trust management based LB, Stochastic Hill Climbing (SHC), two-phase LB and cloud friendly LB.

Authors [6] employed randomized load balancing techniques for performance optimization. Random selection of servers and routing the traffic to them was found promising solution for load balancing. Sampling of randomized subset of queuing is employed in order to have a load balancing method known as batch-filling. The concept of power of having two choices proved to increase cloud performance.

Authors [7] presented efficient resource allocation mechanisms in cloud are explored to have dynamic resource allocations in cloud computing besides improving performance of cloud computing. Different resource allocation algorithms are investigated. They include green cloud computing, cloud resource allocation, resource allocation based on user priority, multi-dimensional resource allocation, and optimal joint multiple resource allocation.

Ant Colony Optimization (ACO) For Load Balancing ACO is one of the optimization techniques widely used in computing applications.

Authors [8] designed a LB mechanism based on ACO for improving performance of scientific applications in public cloud. They used Parameter Sweep Experiments (PSE) to estimate the performance of ACO technique for LB. They used parameter like processing power, Random Access Memory (RAM), storage, bandwidth and others in the experiments. However in this research, existing exertion do not addressed common online systems. Authors [9] presented a virtual Machine (VM) scheduling, which is achieved using Genetic Coding (GC) approach. Gain is one of the performance metrics used to evaluate the work. A survey on VM placement based scheduling is found as potential solution for LB as explored. They observed that meta-heuristics attain enhanced results than conventional heuristics. Authors [10] on the other hand used ACO for load balancing of nodes. In fact they extended ACO in order to achieve results differently. In their ap-proach ants continuously update a single result unlike traditional ACO where each ant 20 provides an individual result set. Their system showed better performance even in the peak hours of usage of cloud system. When an obstacle is encountered the path of ants is modified in order to have better load balancing. The traversal of ants is allowed in two approaches. They are forward movements and backward movements. The major function of this technique is to reallocate work load in-order to have better balancing of load.

Scheduling Techniques For Load Balancing This section focuses on the scheduling algorithms that are crucial for enhancing load balancing efficiency in cloud computing. Authors [18] defined Priority-Aware-Load-Balance-Improved-Min-Min (PA- LBIMM) scheduling mechanism. It was found to be better than its predecessor Load- Balance-Improved-Min-Min mechanism (LBIMM). Make span is a factor considered to estimate the performance of these two mechanisms. PA-LBIMM reduced make span and balanced the load efficiently but unable to consider the user priorities effectively.

Authors [19] tendered a Biogeography-based Optimization (BBO) technique for LB. It is a job scheduling technique which leverages balancing load in cloud. It was anadaptive process which was compared with Genetic Algorithm (GA) and found to be better than it.

Authors [20] outlined a Hybrid Job Scheduling (HJS) algorithm to improve LB in cloud computing. It considered length of jobs and capacity of VMs to make scheduling decisions. It is evaluated with measures like Degree of Imbalance (DI), execution cost and execution time. It focused more on Central Processing Unit (CPU)utilization for load balancing. Fuzzy inference engine was also applied in the solution for better optimization.

2.1. Research Scope - In this work enhanced load balancing models are proposed to address the problems in the existing systems. Especially, the

following are the four research claims or questions conceived from literature. Cloud partitioning based load balancing model can have further improvement by exploiting partitioning rules and refresh period to have state update ideally is the first proposition. Ant Colony Optimization (ACO) based enhancements can lead to more optimal load balancing solution is the second proposition. The third proposition is that user priority guidance based scheduling algorithms can have significant influence for balancing the load in cloud. The fourth proposition is that optimization of weighted round robin approach leads to better load balancing for non-pre-emptive dependent tasks in cloud computing. Examining these propositions is the scope of research.

2.2. Problem Definition

The load balancing algorithms found in the literature are of two categories namely static and dynamic. Static schemes do not make use of runtime system information. Oppor- tunistic Load Balancing (OLB), Central Manager Algorithm and Randomized Algo- rithm are some of the static load balancing techniques. While the dynamic schemes use present state of the system. Though dynamic schemes incur additional costs, they are aware of runtime situations in decision making. ACO, Honey-Bee foraging, Bi- ased Random Sampling and Active Clustering are some of dynamic load balancing techniques. However, both kinds of schemes are aimed at balancing load in cloud com- puting and bring optimality in resource utilization and servicing SLAs. Many load balancing methods are already exists in order to deal with optimization as found in the literature. However, these methods are not able to meet the efficiency requirements of the cloud as it exists now. The need for further research is evident in the research gaps reported in [1], [3] and [4].

1. The first task is to optimize load balancing using cloud partitioning approach by designing ideal refresh period.
2. The second task is to design enhanced ACO model for better load balancing.
3. The third task is to consider user-priority in scheduling and optimize it to influence effective load balancing.
4. The fourth task is dealing with non-pre-emptive tasks (tasks that do not release control intuitively) for load balancing. Hence, improved load balancing models are proposed in this thesis.

2.3. Research Objectives: In order to fulfill the aim of the research that is to design and devise improved load balancing techniques to enhance balancing of load in cloud computing domain to benefit consumers and service providers, the following are the research objectives identified.

1. To implement cloud partitioning based load balancing model for further improvement by exploiting partitioning rules and refresh period to have state update ideally.
2. To develop an Ant Colony Optimization (ACO) based enhancement by computing load balance factor that lead to more optimal load balancing solution.
3. To design user priority guidance based scheduling algorithm that can have significant influence on the load balancing in cloud computing.
4. To propose and implement an optimized weighted round robin approach to improve Round Robin variants for non-pre-emptive dependent tasks.

3. Proposed Methodology:

The proposed load balancing approach is based on cloud partitioning. It refers to dividing the cloud into multiple partitions where each partition can have certain number of hosts and associated VMs. Each partition is associated with a load balancer to schedule given jobs to different VMs. A load status is associated with each partition. This will have intuitive meaning as provided in Table 3.1.

All the Partitions are controlled by a controller. The controller is able to take jobs and schedule them to various partitions based on the partitions state. The overview of the CP based load balancing solution is provided in Figure 3. The designed algorithms are based on this architecture which has potential to balance load effectively.

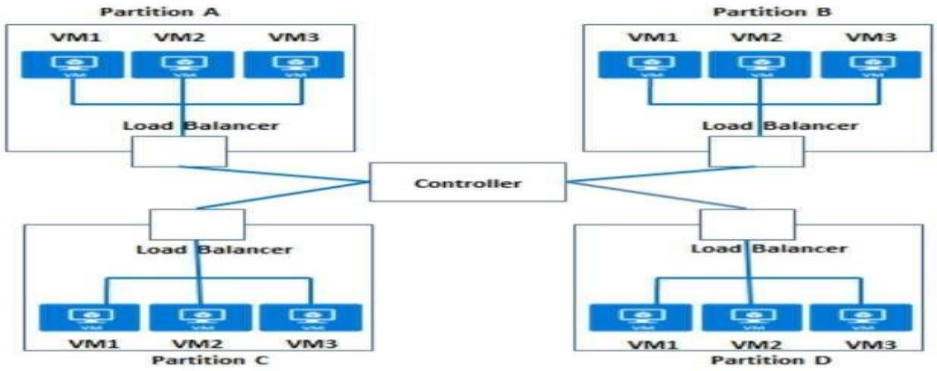
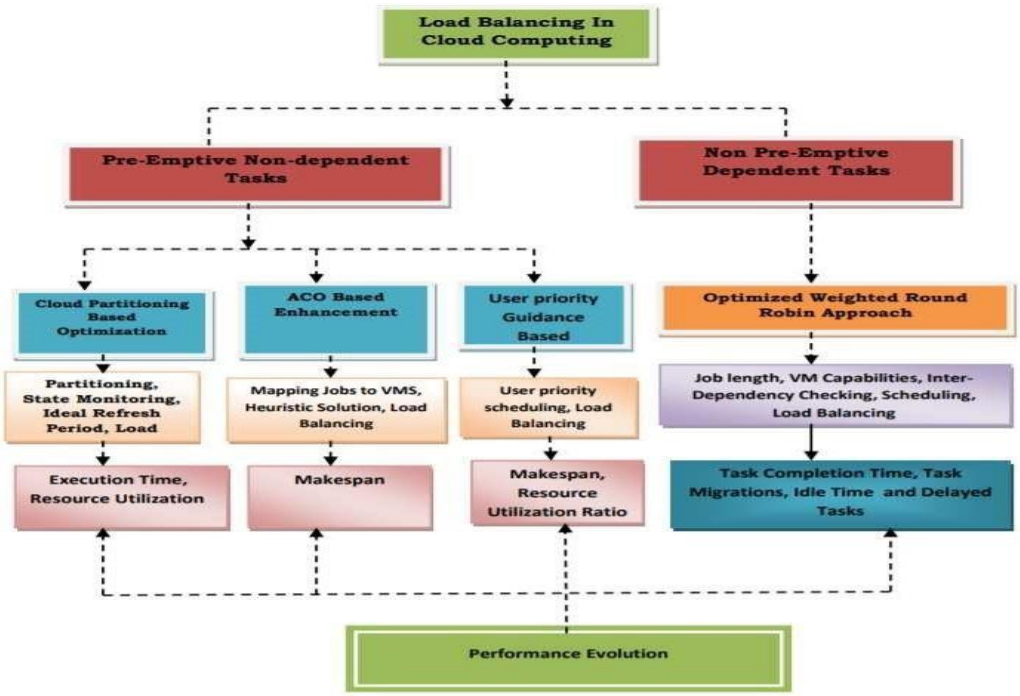


Fig. 3 – Proposed LB architecture

As shown in Figure 3, it is important to understand that there might be number of partitions made in the cloud. Each partition can have different number of VMs and a load balancer for each partition. Each partition in turn is linked to a controller which takes care of proposed
(in later sections) ideal refresh period and load balancing algorithms. The dynamic situations prevail at runtime are considered to understand the number of partitions and the status of each partition to make load balancing decisions. Two algorithms are proposed Partition Based Load Balancing algorithm and Ideal Refresh Period Computation algorithm respectively.

It is evident that jobs arrive at the main controller. Since jobs come in large quantities, it is an iterative approach that is applicable to each job that arrives at the controller. Cloud partitions are chosen as they are available, then any one of the partition is considered for verification. If the partition is found with a state OVERLOADED / HEAVY, that partition is not considered for allocation of jobs. It is skipped for the time being and another partition is considered. If the partition is found to be IDLE / NORMAL it is considered to be the partition to which the jobs are to be sent. After reaching a partition, there is associated load balancer program which will make further decisions. The controller also makes intelligent decisions on the REFRESH period of partitions in order to reflect true state that improves load balancing efficiency. A mechanism of the proposed research is shown in



.

Fig. 4 – Proposed research mechanism

4. The contributions of the research are as follows.

1. Partition based Load Balancing with Ideal Refresh Period Computation (IRPC
2. An Enhanced Heuristic-Ant Colony Optimization (EH-ACO
3. User Priority Based Scheduling for Load Balancing
4. An Optimized Weighted Round Robin Approach for Load Balancing

5. Conclusion –

Cloud partitioning is explored to understand the need for an ideal re- fresh period which updates the status values. If the refresh is carried out with high/low elapsed time, it effects on load balancing performance. Therefore, Partition Based Load Balancing (PBLB) and Ideal Refresh Period Computation (IRPC) are the algorithms developed to strike balance between them and enhance the load balancing capability using cloud partitioning approach. These algorithms test run with CloudSim toolkit to have simulated solution for balancing the load. Moreover, the solution is made in such a way that it visualizes the partitions using a GUI window and job al- locations to let administrators understand the load balancing phenomenon intuitively. With this approach, load balancing is enhanced in terms of efficiency in balancing the load among partitions. The load balancing factor is related to job finishing rate that could provide an effective strategy for scheduling tasks. Optimal resource allocation and minimization of makespan are the two improvements achieved by the designed and implemented algorithm. User Priority based Scheduling for Load Balancing (UPS-LB) is the

developed algorithm. This algorithm works with a designed framework and it is able to classify given tasks into elastic and non-elastic categories to have strategies for improving load balancing. Based on the category of tasks, resource analysis is made and expected completion time of the tasks is examined for better performance in load balancing besides meeting user priorities. Simulation study is made in order to evaluate the algorithm. It was found to be effective and better than predecessor algorithms.

6. References

1.Junaid, M., Sohail, A., Rais, R.N.B., Ahmed, A., Khalid, O., Khan, I.A., Hus- sain, S.S. and Ejaz, N., 2020. Modeling an Optimized Approach for Load bal- ancing in Cloud. *IEEE Access*.

2.Mohammad Shojafar, Saeed Javanmardi, SaeidAbolfazli, Nicola Cordeschi, FUGE: A Joint Meta-Heuristic Approach To Cloud Job Scheduling Algorithm Using Fuzzy Theory And A Genetic Method, IEEE, 18(**??**):P829-844, 2019.

3.Zhuoyao Zhang University Of Pennsylvania, Ludmila Cherkasova HewlettPackard Labs, Boon Thau Loo University Of Pennsylvania, Optimizing Cost And Perfor- mance Trade-Offs For Mapreduce Job Processing In The Cloud, IEEE, 2019.

4.Yang Wang and Wei Shi, Budget-Driven Scheduling Algorithms For Batches Of Mapreduce Jobs In Heterogeneous Clouds, Submitted To IEEE TransactionsOn Cloud Computing, pp. 306-319, 2020.

5.Elina Pacinia, Cristian Mateosb and Carlos Garc´ıa Garinoa, Balancing Through- put And Response Time In Online Scientific Clouds Via Ant Colony Optimiza- tion, IEEE, 2020.

6.Mehdi Sheikhalishahi, Richard M. Wallace, LucioGrandinetti, Jose Luis Vazquez- Poletti and Francesca Guerriero, A Multi-Dimensional Job Scheduling, Future Generation Computer Systems, pp. 123-131, 2019.

7.Lipsa Tripathy and Rasmi Ranjan Patra, Scheduling In Cloud Computing, Inter- national Journal On Cloud Computing: Services And Architecture, 4 (**??**):21-7, 2020.

8.Raja Manish Singh, Sanchita Paul and Abhishek Kumar, Task Scheduling In Cloud Computing: Review, International Journal Of Computer Science And Information Technologies, 5 (**??**):7940-7944, 2019.

9.M.Vijayalakshmi and V.Venkatesa Kumar, Investigations On Job Scheduling Al- gorithms In Cloud Computing, 2(**??**):157-161, 2020.

10.Marios D. Dikaiakos, George Pallis, Dimitrios Katsaros, Pankaj Mehra and Athena Vakali, Cloud Computing: Distributed Computing for IT and Scientific Research, IEEE, 13(**??**):10-13, 2019. 116

11.Niranjan G. Shivaratri, Phillip Krueger and Mukesh Singhal, Load Distributing For Locally Distributed Systems, IEEE, 25(**??**):33-44, 2020.

12.Zenon Chaczko, Venkatesh Mahadevan, Shahrzad Aslanzadeh and Christopher Mcdermid, Availability And Load Balancing In Cloud Computing, International Conference On Computer And Software Modeling, pp. 134-140, 2019.

13.Kumar Nishant, Pratik Sharma, Vishal Krishna, Chhavi Gupta, Kuwar Pratap Singh, Nitin and Ravi Rastogi, Load Balancing Of Nodes In Cloud Using Ant Colony Optimization, International Conference On Modeling and Simulation, pp. 3-8, 2019.

14.Martin Randles, David Lamb and A. Taleb-Bendiab, A Comparative Study Into Distributed Load Balancing Algorithms For Cloud Computing, International Conference On Advanced Information Networking And Applications Workshops, pp. 551-556, 2010.

15.Satish Penmatsaa and Anthony T. Chronopoulos B. Game-Theoretic Static Load Balancing For Distributed Systems,General of Parallel and Distributing Comput- ing. 71 (**??**), pp. 537–555, 2017.

16.Daniel Grosu, Anthony and T. Chronopoulos Ming-Ying Leung. Load Balancing In Distributed Systems: An Approach Using Cooperative Games, IEEE, 2002.

17.Yiannos Kryftis, Constandinos, Resource Usage Prediction Algorithms For Optimal Selection Of Multimedia Content Delivery Method, pp. 7531-7538, 2015.

18.Carlos Colman-Meixner, Chris Develder, Massimo Tornatorey and Biswanath Mukherjee, A Survey On Resiliency Techniques In Cloud Computing Infras- tructures And Applications, 2016.

19.Lei Yu, Liuhua Chen, Zhipeng Cai, Haiying Shen, Yi Liang, Yi Pan, Stochastic Load Balancing For Virtual Resource Management In Datacenters, IEEETrans- actions On Cloud Computing, 2014.

20.Uwe Schwiegelshohn, Andrei Tchernykh, Online Scheduling For Cloud Com- puting And Different Service Level, IEEE, pp. 1061-1068, 2012.