Grenze International Journal of Engineering and Technology, Jan Issue

A Comparative Study on Resource Aware Allocation and Loadbalancing Techniques for Cloud Computing

Malatesh Kamatar[1] and Dr. P Bindhu Madhavi[2]
[1]Department of Computer Science and Engineering, PDIT, Hosapete, India
Email: maltkpl@pdit.ac.in
[2]Department of Computer Science and Engineering, Oxford CE, Bangalore, India
Email: bindhu.pujari@gmail.com

***Abstract*—Advances in cloud networking, cloud infrastructure, and data processing necessitate resource management flexibility. Fault tolerance systems, which include effective load balancing and migration of applications without disturbing any other services operating, should be the key criteria for ensuring the high availability of cloud data centres. Providers of Cloud service should intelligently supply resources to every customers to achieve the optimum resource assignment in changing hosting frameworks. Moreover, many problems in load balancing approaches such as security, fault tolerance, etc exist in today's cloud computing environments. Realizing its relevance and importance, issue of resource aware load balancing in the cloud has received a lot of attention, and numerous load balancing methods is introduced in the literature. Study intends to focus on the survey of many resource aware allocation and load balancing approaches with their advantages and disadvantages. Finally, conclusion with future work is derived a novel technique based on optimization strategy that should guaranteed a drastic improvement in resource utilization on Cloud than existing approaches.**

***Index Terms*— Cloud Computing, Resource Scheduling, Resource Provisioning, Ant Colony optimization, Load Balancing.**

I. Introduction

Cloud computing is a form of distributed computing that uses the internet to make computer resources available to customers on a pay-as-you-go basis. “Cloud computing is a framework for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (such as networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction,” according to NIST [1]. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) are 3 primary categories of cloud services. Virtualization is a fundamental cloud computing technique which enables several activities’ implementation at the same time on a shared platform of hardwares [2]. Virtual machines (VMs), which are abstract machines that operate on physical machines, offer computational resources in the cloud [3]. Because co-located jobs does not communicate with one another and access their data alone, it provides the user the impression of working in a isolated environment. In full virtualization, the isolation among every VM is maintained via Hypervisor, VM Manager that maintains many operating systems by assigning needed resources as well as providing virtual machines with an identical abstraction of underlying hardware [4]. Virtualization allows physical machines to be configured on-the-fly and

Grenze ID: 01.GIJET.9.1.579_1

on-demand to perform a variety of activities, reducing resource wastage [2].

Although virtualization attempts to dynamically load balance the overall system, there is a possibility of resource overuse or underuse. Servers that are overloaded degrade performance, whereas servers that are under loaded waste resources. Due to poor load distribution, overloaded servers would create more heat, resulting in higher cooling system costs and significant CO2 emissions, adding to the greenhouse effect [7]. The cooling systems, in most situations, consume high power than the main IT equipment [6]. Underutilized servers, conversely, increase total system power usage, which is both environmentally unfriendly and raises operating costs. As a result, to achieve the QoS requirement and balancing system's entire load, it is required to dynamically allocate the appropriate number of resource to applications executed in VMs. It also aids in resource energy-efficient usage by turning off unused resources with DPM (Dynamic Power Management) or operating them at a reduced performance level with DVFS (Dynamic Voltage and Frequency Scaling) [6]. The most difficult aspects of cloud computing is load balancing. It's a technique that equally distributes dynamic workload between all nodes in the cloud, avoiding situations where certain nodes are severely loaded and some are idle or performing minimal task [5]. It contributes to reduced response time, high throughput, increased fault tolerance, scalability, and maximum user satisfaction, reduced generation of heat, optimized power utilization, lower CO2 emissions, and lower operating costs through maximizing resource use.

Static and dynamic methods to load balancing are the two main approaches. The processors' performance is decided at the start of execution in the static approach. The work load is first allocated by the master processor based on their performance. The slave processors do their allocated tasks and submit reports to master. Static load balancing approaches are non-preemptive as tasks are always completed on the processor to which they are allocated. The objective of static load balancing is to decrease total running time of a concurrent program at the same time as decreasing communication delays.

Judgment on load balancing is based on the distributed system in the dynamic load balancing method. To get faster service, tasks are permitted to transfer from an overloaded system to an under loaded system. The dynamic method is preferable since it considers the state of the system while load balancing. There are two types: distributed and non-distributed.

The Load Balancing method is run by every node in a distributed system, and load balancing duty is distributed among all nodes. Because of the distributed approach, each node must interact with each other to make decisions, Distributed Dynamic Load Balancing Algorithms create additional message forwarding.

The nodes in a non-distributed system are not in responsible for load balancing. Instead, a load balancing master is appointed to handle the system's load balancing.

We find that the consumption of resource by server varies based on the types of service request from users. Being a topic of research since its origin, web servers still has a research scope in the field of Clusterization and load balancing based on content request.

We believe that server load balancing is insufficient to ensure excellent services to users, because a server with a lower request load does not always mean it can handle high processing requirements like live multimedia streaming or bulky application processing. Most of the works done till date is concentration on just load balancing on servers ignoring the type of content requested by the web users. Each server has different capacity resources and is not capable to scale each user request with the same QoS.

Different resource allocation and load balancing tehcniques are available in the literature to handle the aforementioned sub-problem, which we shall examine next. The rest of the article is formed as: The types of resource allocation and load balancing methods are briefly described in Section II. In Section III, several job scheduling and load balancing methods are described, along with their benefits and drawbacks. At last, in Section IV the article is concluded with comments and suggestions for further investigation in this area.

II. Related Works

This section of the article discusses and critically analyzes some of the prominent state-of-art Resource aware allocation as well as load balancing techniques of cloud system.

A. Review on Resource Aware Allocation

Cloud is an assembly of computers or servers that are interconnected with each other to provide resources to customers. It appears as a novel computing concept that is proposed to provide dynamic, customization and the QoS environment for their customers. The main problem in cloud computing is the network bandwidth, response time, less latency and charge of data transmission. In the cloud computing, an important noteworthy role is resource allocation in the whole classification performance along with the user fulfillment provided by the system. On the other hand, though they provide maximum customer satisfaction, the service provider should also ensure the client's profit side by side. Therefore, resource allocation should be economical for both perspectives that are to the client as well as to the service provider. So, to obtain such a system the new approaches must provide that the system with minimum levels of energy consumption and SLA violation should be designed. Resource allocation is a mechanism, implemented in many computing regions, for example; data center management, operating systems and grid computing. Resource allocation encompasses scheduling of activities as well as allocating the accessible resources in a quite cost-effective way and also it is implemented on optimal processes to adeptly assign physical and/or virtual resources to various applications of developers', therefore, the operational cost of the cloud environment is reduced. The consumption of energy associated with the resource allocation should be taken into account. Resource allocation is an elementary method of the cloud-computing domain, which emphasize on the computing resources like bandwidth, energy and delay etc., in the network to smooth the progress of the carrying out tasks analyzing large-scale computation as depicted in Figure 1.

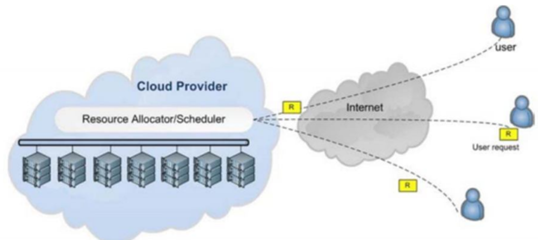

Figure 1: Cloud Computing Resource Allocation (Ghribi, 2014 [8])

Computers can be converted into a device that consumes less energy by using processors that consume less power, utilizing cooling devices along with using spinning Solid State Drive (SSD) of smaller size but not of large size. The process is developed by Intel which consumes less power. In a most advantageous way, various resources are utilized like heat, power choosing the m appropriately. The Indian Information technology industry is one of the world's fastest expanding industries [9]. This leads to more energy consumption and power which in return leads to high cost. According to information from IBM, with less than 4% of its energy is used to process data centers. As end users can access resources from any location and instance, allocation of resource is an issue in cloud computing. Although cloud resources cannot be properly claimed, they can be accessed using the Simple Object Access Protocol (SOAP) / Restful Web Application Protocol Interface (API), which maps the storage or computational requirements for virtualized Information and Communication Technologies (ICT) resources such as a server, blob storage and elastic IP. Because cloud data centers provide a large number of resource assumptions, cloud computing models can support the on-demand allocation of elastic resources. As a result, it causes non-optimal allocation of resource. In the cloud computing concept, the primary function is to allocate a number of available resources among various end users, with shift dependencies depending on the mode of the application-dependent resources. Various requests are to be executed on the resources of data center over the Web.

The goal of resource allocation for any cloud provider is to improve quality of service (QoS), resource usage, and energy efficiency. The primary aim is to improve the QoS parameters (i.e. response time), which determine the efficiency of allocation of resource independent of the type of ICT resources given to each end user. Budget, communication latency, space, and time are all measured using enhanced QoS parameters. [10].

The various existing approaches of resource aware allocation on cloud computing is described as follows

The Interior Search Algorithm (ISA) is used in the [11] study to present an Energy-Efficient Virtual Machine allocation approach that decreases datacenter energy usage and resource under-utilization. Outcome reveals that energy usage of GA and BFD is 90-95 %, compared to about 65 % for suggested EE-IS. Using EE-IS, an average of 30% of energy was saved, as well as increased resource usage.

In [12], present a novel contracts-based resource sharing framework for federated geo-distributed clouds which permits providers of cloud service to apriori create resource sharing contracts with specific datacenters for specified period over a 24-hour time. Every providers of cloud service use a cost-aware job scheduling and provisioning method formed on the recognized contracts to enable jobs to finish and satisfy their response time criteria. Extensive tests using actual workloads are used to assess the suggested approaches, and the findings indicate that suggested technique is effective, scalable, and efficient in resource sharing.

Authors described a system architecture for MCC in [13], which is based on a hierarchical resource sharing approach. Global Cloud Server (GCS), Local ISP Server (LIS), and Gateway Server (GWS) are three areas in the proposed system architecture. There is also a unique paradigm for reducing network delay formed on utilizing Foglets for every suggested technique of clustering mechanism. Furthermore, a fuzzy rule-based approach is provided for removing not appropriate foglets prior to deciding on the best foglet for handover. Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) decision model is used to create a foglet selection scheme. For selecting an ideal network, many characteristics like delay, jitter, BER, packet loss, transmission cost, time for response, network load are taken into consideration. To evaluate the performance and feasibility of suggested system, mobility situation is explored with variety of mobile node speeds between very high to low. Outcomes from simulations and analytical model are contrasted with an available foglet selection technique for mobile nodes. Suggested system design aids in reducing handover delay, packet loss, average queuing delay, device lifespan in a network, according to the examination and analysis.

In [14] introduced an EnReal (Energy-aware Resource Allocation) technique, to solve the mentioned test. Fundamentally, dynamic deployment of VMs for scientific work process implementations are used. especially, a power utilization demonstrate is introduced for applications across distributed computing stages, as well as a relating energy aware resource utilization is suggested for VM planning to fulfil logical work process executions. Outcomes indicate that suggested approach is successful and productive.

A C-RAN application parsing solution was given in [15], which works by dividing Apps into various mechanisms and then activating these modules. The authors also described how to use the OFDMA and PSO facilities in place of assigning all VMs to prevent the middleware solution assigning more mob-resources to the fog server. Several experiments with m-devices were carried out in the results section to arrive at a DSS model to place the network load and decrease energy usage.

[16] Explores minimizing the use of existing fog resources with no extra requirement for extra overflow in m-device using PSO technique. To locate acceptable places in the network, the developed approach employs the implementation of the PSO algorithm. Another technique for mobile clouds that are employed similar to VMs by placing load on cloud infrastructure which might be near is cloudlets and their modifications to identify the appropriate VM place.

In [17], a cloud scheduling approach known as NP-Hard is introduced. PSO was employed by the researchers to lower costs. It is necessary to investigate how the cloudlet method to scaling is employed based on their resources, how m-device network efficiency will cooperate with many users, how fog cloud will manage the stimulus capability. Researchers proposed CSOS technique and reduced makespan to improve network energy. The computational cost has a low value in their technique.

PSO's energy usage on mobile devices has been investigated to certain degree in papers and writings [18]. As previously stated, “DPRA” energy, also termed as network interface, accounts for the majority of the power and expense associated with operating an m-device. The net edge stays active for every outstanding packet from the fog server until the connection is established, which is expensive for net operation. The goal of this article is to build VM-DPRA, a system that combines network desires and demands together in order to decrease total tail energy happenings. Similarly, their tail-ender technique tries to cluster network needs in order to reduce the power. Here, for mapping tasks, the K-means clustering approach was utilised, and the microgenetic algorithm was used for dynamic aggregation of VMs. It is also proven successful in reducing the amount of time it takes to complete a task.

In [19], the authors propose a technique for assigning virtual machines that uses a linear search to improve resource allocation. The system utilised in here is that data centers contain many physical machines, each of which has a set of RAM, CPU, as well as memory. broker is in charge of handling user demands; after that, demands are sent to data center through cloud, there centre manager reserves some amount of VMs to perform clients' requests and suggests various classes of VMs from which the user can choose, each with different features like RAM, CPU, memory. Here, it is possible to cut energy use to certain amount.

For heterogeneous mobile cloud computing networks, the issue of joint radio and cloud resource allocation is investigated [20]. Proposed joint resource allocation methods are designed to optimise customers' total utility while simultaneously fulfilling QoS criteria like E2E response latency. As a combinatorial optimization problem, the issue of joint resource allocation is formulated. The problem is being solved using three evolutionary approaches: GA, ACO-GA, quantum genetic algorithm. For minimizing the complexity of time, authors suggest a mapping procedure among resource allocation matrix and chromosome of GA, ACO-GA, QGA, as well as searching for available resource pairs of radio and cloud depending on resource availability matrixes for ACO-GA and encoding the variation value among the allocated resources and minimum resource requirement for QGA. Suggested approaches outperform current techniques with regard to running time, final result accuracy, overall utility, resource usage, E2E response latency assurance, according to extensive simulation findings.

In the Service Function Chaining on the Software Defined Network Function Virtualization (SDNFV-enabled) cloud environment, this article presents a proactive optimal resource allocation strategy for elastic security Virtualized Network Functions (VNFs) [21]. They begin by analysing the system model and converting it to the M/M/1/∞ and M/M/k queueing models. Subsequently, by evaluating novel VNF resource allocation function and calculating total amount of packets in an SFC i system, they formulate mathematical requirements. They ultimately present a proactive resource allocation optimizer with solved and realistic restrictions based on these criteria. Table 1 shows the Comparative analysis on resource aware allocation on cloud computing.

TABLE I: COMPARATIVE ANALYSIS ON RESOURCE AWARE ALLOCATION ON CLOUD COMPUTING

Authors	Techniques	Results	Disadvantages
Mohammed Joda Usman et al, (2017)	Energy-Efficient Virtual Machine allocation method employing Interior Search Algorithm which decreases energy usage of datacenter and under utilization of resources.	30% average energy is saved through EE-IS and resource utilization is also enhanced	The proposed algorithm is very simple and it only has one parameter to tune.
Xu, J., & Palanisamy, (2018)	Contracts-based mechanism for sharing resource among CSPs in a federated cloud	Performs better than conventional system with regard to fairness in local profits and obtaining same operational costs and success rate properties like current approaches.	data-intensive workloads in the resource allocation model is not considered.
Xu et al (2016)	EnReal technique- Energy-aware Resource Allocation method,	Provides better improvement in energy consumption	In this work several other energy-aware resource scheduling challenges for other kinds of resource like network bandwidths, storage and caches in cloud platforms are not considered
Abdullahi et al (2017)	Chaotic Symbiotic Organisms Search (CSOS) algorithm to reduce makespan and cost	CSOS acquires reduced makespan and cost	Suggested CSOS technique has the ability to solve other optimization issues in cloud computing and combinatorial optimization issues in varies fields.
Chou et al (2018)	Dynamic power-saving resource allocation (DPRA) mechanism based on a PSO is suggested	DPRA energy utilization is reduceed by 17.85%, 15.03%, 18.04%, and 11.36% total energy consumption whereas comparing with BFrand, BFminU, BFminM, and the PSO	Need to increase the selection time and adjustable evolution iteration

B. Review on Cloud Computing Load balancing techniques

Load balancing is procedure of redistributing a distributed system's total load across individual nodes so that no node is overloaded and no node is underloaded or idle [22]. Load balancing in a cloud environment guarantees that no VMs are overloaded, while others are underloaded or performing very minimum task [23]. Load balancing attempts to fasten the applications' running time. It also ensures the stability of the system [24]. It's an optimization approach [24], where job scheduling is an NP-hard problem [25]. The method divides traffic among servers, allowing data to be sent and received without delay [24]. A load balancer was required for appropriate load distribution, which accepted jobs from various locations and subsequently delivered them to the data center. If load balancing is done correctly, it will result in optimal resource usage and minimal resource consumption [23].

Several load balancing existing approaches are described here are listed below

In [26], Researchers have suggested Dynamic Agent-Based'Load Balancing (DA-LB) on Software-Defined Networking (SDN) using Back Propagation Artificial Neural Network (BPANN) for effective utilization of resources. This load balancing approach make use of SDN's global visibility to efficiently relocate efficiently VMs in the data center network. SDN load balancing technique enhances overall network efficiency and performs well for data migration, according to the outcomes. The suggested technique optimizes resource utilization with increasing processing speed and predicting VM load in heavy load situations. When the results of the migration process are contrasted to those of Multi-Path TCP (MPTCP) and Heuristic algorithm (HA), it is clear that the suggested technique requires minimum time to migrate than MPTCP and HA.

In [27], the author offers a load balancing approach for the CDC formed on a graph structure termed weighted complete graph augmented with dominant set idea. A mathematical model is developed, and the suggested technique increases resource utilization in CDC through load balancing and fault tolerance.

[28] Used a genetic method to add the extra virtual machines such that best virtual machines could be assigned to serve the requests. The finest virtual machines might be distributed to serve requests in a highly efficient and timely manner. The load may be balanced by utilizing ant colony optimization technique when particular virtual machines were overloaded with requests throughout the execution. The aforementioned method would distribute the increased load to other virtual machines that are less loaded and unused. Overall energy usage, conversely, is reduced by shutting off virtual machines when they are no longer in use or when they are idle.

For the IaaS computing network, a hybrid heuristic mathematical algorithm (HHMA) has been presented in [29], which addresses the resource allocation issue in cloud network. The cluster is split into little parts through an improved K-means clustering approach (heuristic) to produce a quicker design and decrease file download time. Then, using the cockroach swarm optimization (MCSO) method, it was suggested to assess the load ratio of a calculated node using various restrictions that realize real-time load. The best comprehensive evaluation score for allocating the master to storage node is calculated based on the performance of the MCSO technique. The request data was dedicated to the relevant nodes using two approaches, resulting in optimum processing. As a result, resource utilization has outperformed than other approaches in terms of performance. The proposed technique reduced network overhead, average response time, and memory load, according to simulation results.

Cloud computing is a popular technology that offers virtualized computer resources via the internet. The performance of these clouds is largely determined by different load balancing factors [30]. The act of spreading a dynamic workload across cloud systems and evenly sharing resources such that neither database server is overloaded or underloaded is defined as load balancing (LB). As a result, in the cloud, an active load balancing system may improve reliability, services, and resource utilization. The benefits of Harries Hawks Optimization and Pigeon inspired Optimization method are combined in this paper to develop an effective load balancing method that assures optimal resource usage and job response time. The suggested technique is implemented in the JAVA Net beans IDE, which is integrated into the cloudsim framework and assessed using a variety of tasks to determine the performance. The simulation results demonstrate that the suggested load balancing system based on Hawks Optimization and Pigeon inspired Optimization technique balances the load optimally across the VMs in a substantially short time than current methods. In comparison to other current approaches, the suggested method is 97 % efficient. The Harries Hawks Optimization, Spider Monkey Algorithm, Ant Colony Optimization, and Honey Bee Optimization are compared in terms of cost, throughput, and computational time, make span, latency, execution time.

The author suggested a hybrid load balancing approach in [31], with goal of dynamically reallocating virtual machines whenever resource usage exceeds the critical threshold size and when utilization falls below a low threshold size. Goal of providing a virtual machine load balancing system is to combine virtual machines so that resources may be used to their full potential while also being able to reschedule cloudlets and increase processing period. Scheduler considers the RAM utilization, bandwidth, as well as CPU used throughout interval processing time and selects the parameter ratio with the greatest value. Instead of relying just on CPU load, this method allows to identify which parameter led to excessive utilization. The suggested methodology was able to produce lower total load usage with the following strategies in place, implicitly showing that performance analysis technique is valid and will be a better option.

In [32] proposes a load balancing algorithm formed on honey bee behavior (LBA HB). Its major objective is to distribute workload over various network links in such a way that resources are not underutilised or overutilized. This is carried out by assigning the incoming job to a VM which satisfies two criteria: the amount of jobs presently being processed by this VM is less than amount of jobs being processed by other VMs, This VM's processing time is less than a threshold number when compared to the average processing time of all VMs. suggested method is contrasted with honey bee, ant colony, modified throttled, and round robin techniques. Experiments demonstrate that the suggested method is efficient with regard to execution time, response time, makespan, load standard deviation, and degree of imbalance.

In [33], researchers provide a unique cloud load balancing approach based on Honey Bees and Ant Colony Behavior Algorithms. This algorithm may also take into account task prioritization. The proposed load balancing method enhances cloud environment throughput by focusing on lowering task completion time as well as the time spent waiting for a job in the VM's queue. As a result, the technique decreases the time it takes for virtual machines to respond. In cloud environments, this load balancing approach is suited for heterogeneous systems as well as non-preemptive jobs.

A Raven Roosting Optimization Algorithm (RRO) is used in [34] to focus attention on load balancing for job scheduling issues in the cloud. Heterogeneity birds and insects engage in roosting. Roosts are data-center, or servers, in raven roosting, and scrounge features of ordinary ravens motivated to resolve issues. The approach is capable of transferring a large number of overloaded jobs to Virtual Machines (VMs) while assessing the VMs' capacity. Random allocation of VMs to Cloudlets by Raven Roosting Optimization (RRO) leads in a significant change in the makespan of the VM to which it is assigned. There's a chance that Raven Roosting Optimization Algorithm simulation results indicate improved makespan, average response time, average waiting time, and number of jobs migrated.

This article focused on determining where physical hosts should be deployed and suggested an unique heuristic method termed Load Balancing based on Bayes and Clustering (LB-BC) [35]. For achieving instantaneous load balancing impact, most existing works employ a progression of computations and locating the target hosts inside an algorithm cycle, followed by choosing ideal target hosts. According to this reasoning, LB-BC proposes a long-term approach to general load balancing, as opposed to the current writing's rapid load balancing techniques. In terms of execution capacity, LB-BC places a restriction on every physical host intending to complete and send an approach with worldwide inquiry capacity. The results show that, in comparison to previous work, the suggested technique decreased number of events that fail, increased throughput, and optimized the outside performance services of cloud servers.

They use SDN methodology to cloud data center in their work to accomplish dynamic load balancing and offer an SDN-based dynamic load balancing strategy for cloud centers [36]. The technique of incorporating SDN technology into current scheduling, with the OpenFlow protovol monitoring the service node and load status in real time. The controller can distribute some available resources as soon as the load is imbalanced. Outcomes from simulation indicate that this technique can accomplish and assure that the load does not tilt over time, while also increasing system throughput.

The importance of DL methods in the domain of cloud computing was examined in [37]. Deep Learning-based Deadline-constrained, Dynamic VM Provisioning, and Load Balancing (DLD-PLB) have been suggested and developed as a system for Workflow execution in cloud. Deep Learning-based approach is utilized to produce an optimal VM scheduling. The recommended framework has used the Genome process activities as input. The findings for makespan and cost of the proposed framework have been calculated and compared to their earlier suggested scheme for load balancing optimization - Hybrid approach based Deadline-constrained, Dynamic VM Provisioning and Load Balancing (HDD-PLB) architecture for Workflow implementation. Previously suggested load balancing techniques were focused on hybrid Predict-Earliest-Finish Time with ACO for underused VM optimization and hybrid Predict-Earliest-Finish Time -Bat for overflow VM optimization.

Guo (2017) [38] presented a multi-objective ACO job scheduling method with balancing the load and cost.

TABLE II: COMPARATIVE STUDY OF LOAD BALANCING ON CLOUD COMPUTING

Authors	Techniques	Results	Disadvantages
WilsonPrakash, & Deepalakshmi (2019)	Dynamic Agent-Based' Load Balancing on SDN using BPANN	helps to efficiently balance the traffic in the network. BPANN helps to process the request faster than other algorithms.	The actual performance of backpropagation on a specific problem is dependent on the input data. Back propagation algorithm in data mining can be quite sensitive to noisy data.
Kanthimathi, M., & Vijayakumar (2018)	Enhanced Approach of Genetic and Ant colony based Load Balancing	Suggested method provides reduced makespan than others for different tasks	GAs dose not have a practicable feedback system, hence huge number of redundant iterations are produced when solutions are within a certain scope, that results in low efficiency
Senthilkumar, & Chitra (2020)	Novel hybrid heuristic-metaheuristic Load balancing algorithm for Resource allocation	result shows that the suggested hybrid method provides better performance interms of response time and the processing time.	There is no any specific disadvantages which affects the system performance
Princess, & Radhamani (2021)	Harries Hawks Optimization and Pigeon inspired Optimization Algorithm to create efficient load balancing scheme, that guarantees best resource usage with job response time.	Performance of suggested method is 97% when contrasted with other current approaches.	However, algorithm requires a long search time, convergences slowly, and is easily trapped in local optima.
Hashem et al (2017)	Honey bee based load balancing in cloud computing	The Factors like threshold value, executable instruction length and average number of requests per user per hour are improved	However, the algorithm convergences slowly, and is easily trapped in local optima.

III. Motivation of the Research

The cloud computing permits cloud service provider to allocate resources to tasks depending on the needs of the customers. There are still gaps in this domain that should be filled, according to the literature evaluated in this study.

Strategic: Finding the details for resource and workload identification for enhanced mappings for task scheduling and implementation is one emerging subject for the authors to research. For this goal, workloads are to be implemented efficiently in order to be flexible, scalable, and optimal, reducing resource under and overuse. Furthermore, the application of AI techniques in the allocation of resources reduces the possibilities of errors and failure to virtually nil, allowing for greater precision and accuracy in cloud computing resource allocation. Artificially based resource allocation, on the other hand, should take into account the cost factor, as well as refining the technique to prepare for bigger systems.

Target Resource: In network aware resource allocation, communication between VMs belonging to different servers should be kept to a bare minimum. The authors should develop techniques for reducing communication costs by determining the shortest path between VMs.

Optimization: The major focus of optimization-based resource allocation approaches should be on profit of cloud service providers and user satisfaction, while maintaining the SLA negotiation between service provider and user in mind. In the event of a system failure, authors should additionally consider the penalty limit.

Scheduling: Based on existing research, it is necessary to evaluate resource scheduling algorithms in a real-world setting. Dynamic resource scheduling is an open problem, according to the literature reviewed in this study.

Power: Green optimization resource allocation techniques in data centers are an open problem in power aware resource allocation, according to this study, and further research is needed in this area. Authors should explore the connection between various workloads, with the goal of developing frameworks that minimizes trade-offs between SLA and energy efficiency in methods.

IV. Conclusion

This research has given readers an intellectual knowledge of the critical principles of balancing the load and allocating of resources in cloud environment. The article will also assist readers in determining the gap between current resource allocation and load balancing approaches and what is necessary in order to identify critical issues that require additional examination. This article not only provides an overview of selected papers under appropriate headings, but it also includes a discussion on the evolution of resource allocation approaches throughout the years. It also outlines possible future possibilities in the cloud computing resource allocation domain. However, there are additional options that need to be investigated in order to build more cost-effective allocation systems. In future, resource allocation approaches in cloud computing should be focused on improving security, performance isolation, smooth virtual machine migration, interoperability, failure resilience, graceful recovery, and lowering data center operational costs. Finally, cloud computing services are expected to become an important element of nearly all types and sizes of information systems.

References

[1] Peter Mell, Timothy Grance. The NIST Definition of Cloud Computing (Draft). NIST. 2011.

[2] Gkatzikis, Lazaros, and Iordanis Koutsopoulos. "Migrate or not? Exploiting dynamic task migration in mobile cloud computing systems." *IEEE Wireless Communications* 20.3 (2013): 24-32.

[3] Raghavendra Achar, P. Santhi Thilagam, Nihal Soans, P. V. Vikyath, Sathvik Rao and Vijeth A. M., "Load Balancing in Cloud Based on Live Migration of Virtual Machines ", 2013 Annual IEEE India Conference (INDICON).

[4] Jain, Akshay, et al. "A threshold band based model for automatic load balancing in cloud environment." *2013 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)*. IEEE, 2013.

[5] Amandeep Kaur Sidhu, Supriya Kinger, "Analysis of Load Balancing Techniques in Cloud Computing ", International Journal of Computers & Technology, Volume 4 No. 2, March-April, 2013, pp. 737-741.

[6] Kliazovich, D., Arzo, S. T., Granelli, F., Bouvry, P., & Khan, S. U. (2013, August). e-STAB: Energy-efficient scheduling for cloud computing applications with traffic load balancing. In *2013 IEEE International Conference on Green Comp. and Communications and IEE IoT and IEEE Cyber, Physical and Social Computing* (pp. 7-13). IEEE.

[7] Beloglazov, A., & Buyya, R. (2010, May). Energy efficient resource management in virtualized cloud data centers. In *2010 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing* (pp. 826-831). IEEE.

[8] Ghribi, C. (2014). *Energy efficient resource allocation in cloud computing environments* (Doctoral dissertation, Institut National des Télécommunications).

[9] Ghribi, C. (2014). *Energy efficient resource allocation in cloud computing environments* (Doctoral dissertation, Ins. National des Télécommunications).

[10] Vakilinia, S., “Energy Efficient Temporal Load Aware Resource Allocation in Cloud Computing Datacenters.” Journal of Cloud Computing, Vol. 7, Issue 1, ISSN: 2192-113X, 2018

[11] Usman, Mohammed Joda, et al. "Energy-Efficient virtual machine allocation technique using interior search algorithm for cloud datacenter." *2017 6th ICT international student project conference (ICT-ISPC)*. IEEE, 2017.

[12] Xu, J., & Palanisamy, B. (2018). Optimized contract-based model for resource allocation in federated geo-distributed clouds. *IEEE Transactions on Services Computing*, *14*(2), 530-543.

[13] Ahmad, A., Paul, A., Khan, M., Jabbar, S., Rathore, M. M. U., Chilamkurti, N., & Min-Allah, N. (2017). Energy efficient hierarchical resource management for mobile cloud computing. *IEEE Transactions on Sustainable Computing*, *2*(2), 100-112.

[14] Xu, Xiaolong, et al. "EnReal: An energy-aware resource allocation method for scientific workflow executions in cloud environment." *IEEE transactions on cloud computing* 4.2 (2015): 166-179.

[15] Sun, Yuan, et al. "Energy-efficient resource allocation in c-ran with fronthaul rate constraints." *2016 8th International Conference on Wireless Communications & Signal Processing (WCSP)*. IEEE, 2016.

[16] Wang, S., Liu, Z., Zheng, Z., Sun, Q., & Yang, F. (2013, December). Particle swarm optimization for energy-aware virtual machine placement optimization in virtualized data centers. In *Inter. Conference on Parallel and Distributed Systems* (pp. 102-109). IEEE.

[17] Abdullahi, M., Ngadi, M. A., & Dishing, S. I. (2017, May). Chaotic symbiotic organisms search for task scheduling optimization on cloud computing environment. In *2017 6th ICT International Student Project Conference (ICT-ISPC)* (pp. 1-4). IEEE.

[18] L.-D. Chou, H.-F. Chen, F.-H. Tseng, H.-C. Chao, and Y.-J. Chang, “DPRA: dynamic power-saving resource allocation for cloud data center using particle swarm optimization,” IEEE Systems Journal, vol. 12, no. 2, pp. 1554–1565, 2018.

[19] T. Kumrai, K. Ota, M. Dong, J. Kishigami, and D. K. Sung, “Multiobjective optimization in cloud brokering systems for connected internet of Things,” IEEE Internet of Things Journal, vol. 4, no. 2, pp. 404–413, 2017.

[20] Xia, W., & Shen, L. (2018). Joint resource allocation using evolutionary algorithms in heterogeneous mobile cloud computing networks. China Communications, 15(8), 189-204.

[21] D. Pandit, S. Chattopadhyay, M. Chattopadhyay, and N. Chaki, “Resource allocation in cloud using simulated annealing,” in Proceedings of the Applications and Innovations in Mobile Computing (AIMoC), 2014, pp. 21–27, Kolkata, India, February 2014

[22] M. Agarwal, D.G.M.S. Srivastava Cloud Computing: A Paradigm Shift in the Way of Computing Int. J. Mod. Educ. Comput. Sci., 9 (12) (2017), pp. 38-48

[23] Babu Kr, Joy Aa, Samuel P. Load Balancing Of Tasks In Cloud Computing Environment Based On Bee Colony Algorithm. In2015 Fifth International Conference On Advances In Computing And Communications (Icacc) 2015 Sep 2 (Pp. 89-93). IEEE.

[24] Dam S, Mandal G, Dasgupta K, Dutta P. Genetic Algorithm And Gravitational Emulation Based Hybrid Load Balancing Strategy In Cloud Computing. In Computer, Communication, Control and Information Technology (C3it), 2015 Third International Conference on 2015 Feb 7 (Pp. 1-7). IEEE

[25] Aslanzadeh S, Chaczko Z. Load Balancing Optimization In Cloud Computing: Applying Endocrine-Particle Swarm Optimization. In2015 IEEE International Conference On Electro/Information Technology (Eit) 2015 May 21 (Pp. 165-169). IEEE.

[26] WilsonPrakash, S., & Deepalakshmi, P. (2019, April). Artificial Neural Network Based Load Balancing On Software Defined Networking. In 2019 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS) (pp. 1-4). IEEE.

[27] Devi, R. K., Murugaboopathi, G., & Vijayakumar, P. (2017, February). A Graph-based Mathematical Model for an Efficient Load Balancing and Fault tolerance in Cloud Computing. In 2017 Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM) (pp. 136-140). IEEE.

[28] Kanthimathi, M., and D. Vijayakumar. "An enhanced approach of genetic and ant colony based load balancing in cloud environment." *2018 Inter. Conf. on Soft-computing and Network Security (ICSNS)*. IEEE, 2018.

[29] Senthilkumar, G., & Chitra, M. P. (2020, August). A Novel hybrid heuristic-metaheuristic Load balancing algorithm for Resource allocation in IaaS-cloud computing. In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT) (pp. 351-358). IEEE.

[30] Princess, G. A. P., & Radhamani, A. S. (2021). A Hybrid Meta-Heuristic for Optimal Load Balancing in Cloud Computing. Journal of Grid Computing, 19(2), 1-22.

[31] Yuganes Parmesivan , Sazlinah Hasan, Abdullah Muhammed, “Performance Evaluation of Load Balancing Algorithm for Virtual Machine in Data Center in Cloud Computing”, International Journal of Engineering & Technology, 7 (4.31) (2018) 386-390

[32] Hashem, W., Nashaat, H., & Rizk, R. (2017). Honey bee based load balancing in cloud computing. KSII Transactions on Internet and Information Systems (TIIS), 11(12), 5694-5711.