International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

Machine Learning Load Balancing Techniques in Cloud Computing: A Review

Juliet Gathoni Muchori[1]
Murang'a University of Technology
School of Computing and Information Technology
Department of Information Technology
Murang'a, Kenya

Peter Maina Mwangi[2]
Murang'a University of Technology
School of Computing and Information Technology
Department of Computer Science
Murang'a, Kenya

Abstract: Load balancing (LB) is the process of distributing the workload fairly across the servers within the cloud environment or within distributed computing resources. Workload includes processor load, network traffic and storage burden. LB's main goal is to spread the computational burden across the cloud servers to ensure optimal utilization of the server resources. Cloud computing (CC) is a rapidly growing field of computing that provides computing resources as a product over the internet. This paper focuses on the issues within Cloud Load Balancing (LB) that have attracted research interest. The paper also mainly focused on uncovering machine learning models used in LB techniques. The most common algorithms in the reviewed papers included Linear Regression, Random Forest classifier (RF) Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long-Short Term Memory- Recurrent Neural Network (LSTM -RNN). The criteria for LB technique was identified through performance metrics like throughput, response time, migration time, fault tolerance and power saving. The paper adjourns by identifying research gaps found in the reviewed literature.

Keywords: Linear Regression, Random Forest classifier (RF), Artificial Neural Network (ANN), Convolutional Neural Network (CNN) and Long-Short Term Memory- Recurrent Neural Network (LSTM -RNN), Load Balancing (LB)

1. INTRODUCTION

Load balancing refers to the distribution of the computing workload to a group of servers. Load implies CPU load, network traffic burden and server storage capacity. The workload originates from the client requests and is sent to the servers [1]. The concept of load balancing is applied in distributed system administrators to sub-divide, allocate and issue resources between different servers, networks and computers [2]. Load balancing provides the ability to cope with growing hardware architecture and computing needs. The system performance ideally remains relatively independent of the increasing input variables [2].

The importance of load balancing includes equitable distribution of computing resources by ensuring nodes are not overloaded or underloaded [1]. In return these improves the speed and overall throughput of the whole distributed system. Other critical contributions of load balancing include cloud scalability by distributing the new additional workload effectively to the new instances of virtual servers and starts different services to address the growing requests.

Other importance of load balancer in cloud environment is the detection of the idle nodes and the newly added servers. The LB component is in charge of directing new requests to the discovered servers [2]. Another importance includes the disaster recovery by having the LB component handle the cloud services continuity in case of catastrophic failures. The load balancer redirects the requests to the available servers. It exhibits transparency to the user by making all services available despite internal failures or large workloads.

A typical load balancer accepts requests from clients or other servers within a network and then assigns the most appropriate server to handle the load [3]. There are several types of load balancers [2]. They include software-defined network (SDN), User Datagram Protocol (UDP), Transmission Control Protocol (TCP), Server Load Balancer (SLB), Virtual and Load Balancer as a Service (LBaaS) [3].

Cloud computing is a rapidly growing field of computer science that is concerned with scaling distributed systems, networks and storage to mega utility computing systems [4]. Cloud service providers (CSP) provide access to both hardware and software resources to the cloud users on demand. Users access the cloud resources remotely over the internet on a rented basis. Some of the popular firms offering cloud solutions include Apple, Amazon, Google, IBM and Microsoft [4].

Cloud services can be categorized into three service models, namely: Infrastructure as a service (IaaS), Software as a Service (SaaS) and Platform as a service (PaaS). IaaS is a cloud service that offers access to Information Technology (IT) hardware and software over the internet. The host is in charge of managing and availing the infrastructure resources to the cloud user. For instance; virtual machines, servers, storage and network resources [5].

PaaS provides the resources to create, publish and customize the software in a hosted environment. The cloud user is allowed to install their software and tools like database, web server and application servers [3]. There are

specialised platforms to offer unique features like communication. For instance, a communication platform as a service offers real-time communication capabilities like video and voice. PaaS eliminates the software license costs [5].

SaaS provides the cloud user with ways of accessing software from anywhere over the internet connection. The host offers access to the application and the data for instance Yahoo!, Gmail, iCloud, Microsoft Office 365 [5]. Instead of installing the software, the user simply accesses it over the internet. It frees the users from complex handles of managing and installing the software. The cloud provider has the responsibility to manage the access, to secure and to make available the application.

Cloud computing is deployed using various methods. The design and setup of the cloud environment can follow different topologies. These topologies include; Private Cloud, Public Cloud, Hybrid Cloud and Community Cloud. Private Cloud is set up by a cloud provider and its application and infrastructure are operated by the application provider entirely. The cloud is solely dedicated to the needs of a single organization [5].

A public cloud is an open setup model where the infrastructure facilities are provided by a third party. This topology allows resource sharing between multiple organizations or consumers. The public cloud makes the computing resources available to anyone for a subscription fee. All the hardware, software and other supporting infrastructure are owned and managed by the cloud provider. This model is the least expensive choice for application hosting [6].

Hybrid cloud combines private and public cloud and it eliminates the need for the trust model. Both public and private clouds require interoperability and portability of data and applications that allow communication across the models. This model is less expensive compared to the private cloud. On-premises datacentre shares data and applications with the public cloud.

Community cloud is similar to the extranets but it has dedicated virtualization on demand. Organization sharing common goals or a specific community builds a shared cloud to be used by its members [6]. The community cloud has multiple tenants that share similar concerns in security, performance and the reach of the cloud. Services offered are limited to the computing needs and the requirements of the community members.

The host of the cloud computing services has to ensure ease of access, availability and the fastest access of the cloud services. Cloud computing is experiencing growing consumers that requires the scaling of computing infrastructure and enhancement of resilience and fault tolerance. To maintain quality cloud services, the host has to conquer load balancing, fault-tolerance, virtualization and cloud security [5].

The goal of this review paper is;
To identify major challenges facing load balancing in cloud computing.
To review machine learning-based approaches to Load Balancing.
To identify renowned metrics for load balancing.
There is a need for an intelligent burden balancer since effective load balancing solves the majority of the cloud computing performance issues. This paper focuses on the application of artificial intelligence to solve the load balancing problem.

2. RESEARCH DESIGN

This chapter describes the organization of this research paper. It goes further to discuss the set of research papers evaluated in this paper along with the sources of those papers and their elaborate search criteria.

A. RESEARCH QUESTIONS

This review paper was guided by the following main research questions:

a). What are the current major research challenges in cloud load balancing?

c). What are the various machine learning load balancing techniques that have been developed so far?

d). What are the criteria to identify an effective load balancing technique?

e). What research gaps still exist in the current load balancing approaches?

The above questions are answered by carefully considering accurate Cloud Computing and Load Balancing published peer-reviewed research papers acquired through the search criteria discussed below.

B. SEARCH CRITERIA

A step-wise analysis of the load balancing and cloud computing were conducted over well-known paper indexing engines. Some of the search strings that were used include cloud computing, load balancing, challenges in cloud computing, research challenges in load balancing, machine learning-based load balancing in CC, deep learning-based load balancing techniques in CC. The search engines used in this paper are tabulated below in table 1:

Table 1: Research Papers Search Engines

Finding Engine	Source Address
Semantic Scholar	https://www.semanticscholar.org/
Science Direct	https://www.sciencedirect.com/
IEEE Xplore	https://ieeexplore.ieee.org/
Research Gate	https://www.researchgate.net/
Google Scholar	https://scholar.google.com/

C. DATA SOURCES

During the preparation of this survey, several data sources were considered. The research paper primarily looked for conferences, periodicals and journal papers in Google Scholar, Scopus, Science Direct and other databases of related research papers.

International Journal of Computer Applications Technology and Research
Volume 11–Issue 06, 179-186, 2022, ISSN:-2319–8656
DOI:10.7753/IJCATR1106.1002

Table 2: Inclusion Criteria Table

Criteria	
Inclusion	• A paper that outlined the research problems present in cloud load balancing. • A research paper authored by scholars or practitioners. • A research paper that focuses on machine learning based load balancing techniques. • A peer-reviewed publication. • English written paper
Exclusion	• A research paper that doesn't emphasize intelligent load balancing techniques. • Commercial Papers that required purchase to access.

D. STRING REFINEMENT

The search strings were entered on the search engines discussed above in Table 1. Usually, the search was refined gradually after the broad search of cloud computing, then load balancing in cloud computing, machine learning-based load balancing in cloud computing and deep learning-based load balancing in cloud computing. The research papers were constrained to papers published from January 2016 and December 2021.

More than 70 papers were found with majority of them being commercial access and others lacking direct relation to our field. After eliminations around 30 papers were reviewed. The criteria were focused on the keywords. Although sometimes the criteria were further refined to include open-source papers, it reduced the relevance of the papers found and it was preferred to check the abstracts of the papers with restricted access.

E. QUALITY ASSESSMENT

On the searched research papers, quality evaluation criteria were applied for inclusion and exclusion of the research papers. An initial study of the papers executive summaries was studied and depending on our guiding research questions, the paper was included or excluded.

Thereafter, the selected research papers were fully read based on the criteria and the papers were either included or excluded for review. The inclusion criteria entailed a major focus on the load balancing techniques in cloud computing and intelligence-based techniques.The exclusion criteria were based mostly on research paper accessibility and relevance to the goals and the research questions of this review. Some papers written in the context of cloud computing could not be included since they focused on the traditional load balancing techniques. The paper will proceed to discuss the answers to the research questions by beginning with the challenges in cloud computing.

3. RESEARCH CHALLENGES WITHIN CLOUD LOAD BALANCING

The primary goal of load balancing is to ensure cloud nodes are not under or overloaded. The LB process can be described as the spread of the work burden across the network links on the multiple clusters to maximize the use of the assets and cut the overall turnaround time. Burden balancer is strategically placed within a cloud architecture. It is situated such that it receives the cloudlets requests and determines which server to forward the request to. In some cases, the load balancer is positioned as a server that performs tasks distribution to other servers [3].

An ideal load balancer should exhibit intelligent behaviour in allocating the requests smartly. Smart resource allocation involves equal load allocation on all the servers at every single point in time. The emergence and growth of artificial intelligence have shifted the load balancing research from the traditional load scheduling algorithms like min-min, round-robin to intelligent load balancing models based on machine learning and deep learning [2].

An intelligent load balancer offers a competitive advantage to the cloud providers by ensuring the quality of service and compliance to the established service level agreement (SLAs) [1]. Some of its key significance include the highest performance in terms of response and throughput, web traffic management, effective handling of the ad-hoc traffic bursts and surges, and flexibility. Load balancer offers elasticity in terms of scalable computational requests on client demand.

This section will highlight the research challenges present in load balancing, these are the problems that intelligent load balancing sought to address. Cloud management is faced with uncertainty since the resource demand keeps on charging every second. Some of the logical and physical problems that complicate load balancing include:

The physical location of the datacentres poses some logistical and response time challenges in the load balancer since cloud providers have datacentres in different continents and cloud users expect the cloud to perform without delays. Another challenge is in the edge computing problem that recommends the cloud requests be processed near where they emanate [4].

Migration of the virtual machines (VM) is another problem within load balancing since overloaded physical machines are prompted to migrate some of their virtual machines to another underutilised physical machine [7]. The challenging part is copying the current location of the VM memory pages and transferring them across the network without affecting the services its offering [8]. Live VM migration consumes a large part of the network bandwidth, CPU and memory resources that can easily violate the SLA.

The complexity of the balancing techniques raises the question of the algorithmic complexity of the load balancing technique. An efficient load balancer should not be very complex in terms of hardware requirements and fault tolerance. Good techniques make compromises to maintain optimal performance [3].

Heterogenous nodes present another research challenge to the load balancer since the nodes have different

computational capabilities, memory and networking components. The nodes have a different kinds of machine architectures (GPUs, CPUs, Multicore CPUs). Establishing a uniform mechanism to share the indifferent resources by assigning tasks poses the problem [9].

Single Point of failure occurs when the entire system depends on a single load balancer component [7]. There is a need to have a redundant load balancing component that implements a failover solution that forwards the burden-sharing component to another load balancer within the same cloud network. Some scenarios advocates for having a primary and a standby load balancer that automatically switches on failure [2].

The scalability of the load balancer is another load balancer design consideration. Cloud providers have interconnected nodes that are added to cater for the growing cloud demand. A scalable load balancer allows the addition of the hardware and scaling of the load balancing component [1]. The load balancer should maintain the performance after task scaling. It maintains a balance between the used and unused resources.

4. MACHINE LEARNING -BASED LOAD BALANCING TECHNIQUES

This section is dedicated to intelligent load balancing techniques. Load balancing techniques are categorized differently according to different features. They play a major role in server resource utilization. Load balancers are built upon some cloud environment aspects like the server CPU and memory resource, service level agreements (SLAs), prediction of the network congestion, quality of service (QoS), service response time estimation, and the storage demand within the cloud [10].

Machine learning is a part of Artificial Intelligence that focuses on training systems to perform new tasks without being explicitly programmed. Historical data and statistical techniques are combined through a process called training to build models that can forecast new unseen values [11]. Deep learning is a subset of machine learning that uses variations of neural networks with deeper networks and large datasets. Deep learning combines feature extraction and prediction in a deep network within hidden layers. It achieves better performance than traditional machine learning problems [12].

The following intelligent models were reviewed:

a. Deep learning regression technique

Deep learning-based regression was used to predict the continuous schedule of tasks from the computational time and cost by Kaur and others [13]. The deep learning network was designed to have 3 hidden layers of convolutional neural networks, a pooling layer and the activation layer made up of the ReLU function. The training data was composed of the time and cost parameters data from larger workflows.

b. Fully Connected Network (FCN) Technique

The deep learning-based load balancing mechanism was made up of fully connected convolutional layers. The model was developed by Zhu and others [14] to replace the hash functions used traditionally to schedule the tasks. Historical cluster access data was used to train the model. The FCN model was designed as a hierarchical model made up of sub-models that feed their output as input to the next hierarchy stage [14]. The hierarchy was made of 4 stages with input, disperse, mapping and join stages. Each sub-model had 3 fully connected layers [14]. The models used the deterministic approach to map the workload to the servers.

c. Support Vector Machines (SVM) and K-suggest technique

Lilhore and others proposed [15] a load balancing solution that was based on multiple machine learning algorithms like SVM and K-suggest clustering tool. Clustering is used to establish groups of virtual machines that are derived from the CPU and main memory (RAM) usage. This technique shared the assets with various groups and the VMs as well. Then it used dynamic aid mapping to assign the loads to their appropriate VM groups depending on their sizes i.e: normal, Idle, Underloaded and overloaded VMs [15]. Resource mapping involved mapping the grouped jobs with the appropriate VM group. This method improved the quality of service and lowered the overall wait or reject time [15].

d. Bayesian Network with Reinforcement Learning

Liang and others [16] proposed a load balancer to control the traffic in the Software-Defined Network Controller component of data centers [16]. The Bayesian network was used to predict the amount of load traffic and combined it with reinforcement learning for the optimal cause of action and to add a self-adjustment parameter. Software-Defined Network is the brain of the whole network that separates the data transmission layer and the control layer.

Bayesian network predicted the load traffic on the SDN controller while the predictions were used through reinforcement learning to determine best cause action [16]. Strategies adopted involved the spread of the network burden and delocalization of the processing and control. Network stability, load balancing speed and performance of the controller were achieved.

e. Regression, Random Forest and AdaBoost based technique

Machine Learning-based load distribution model [17] was made up of several models namely multiple linear regression (MLR), random forest (RF) and AdaBoost (Ada) were used to determine where each query to be processed based on the turnaround time of the CPU and GPU [17]. This technique addressed the architectural heterogeneity by accounting for the difference in the processing units and their associated performance characteristics. Its major focus was on the distributed database management systems transactions distribution [17].

f. ANN and self-adaptive differential evolution (SaDE) technique

This technique was developed by Kumar and others to predict the workload within the cloud data centre [18]. This approach combined the artificial neural network (ANN) and the self-adaptive differential evolution (SaDE). User requests were amassed to time units that were used as the historical data. The ANN part was trained with the actual workloads and the historical data [18]. The resultant model was used to forecast the upcoming work in the data centre. The model was trained on datasets from NASA and Saskatchewan servers [18].

g. Long Short-Term Memory -Recurrent Neural Network (LSTM-RNN) approach

LSTM is a special kind of RNN that preserves the weights from past activities that making it a good deep learning algorithm to be used in time-series forecasting. LSTM algorithm has an inbuilt forget gate that allows it to escape the long-term dependencies up to some point and it only keeps the necessary features. In their paper, Kumar [19] and others sort to address major issues within load balancing namely: power utilization and scaling of resources dynamically.

LSTM-RNN load balancing technique was developed by analyzing the history of the data centre through the cloudlet traffic logs with consideration to the time factor. Insights that were gained from the historical data were used to predict the future workload. The training data is continuous over a period. The projected workload data was used to expand the resources and to shut down the unused resources to preserve power [19]. LSTM-RNN model was trained on the HTTP traces of NASA, Saskatchewan Server and Calgary server datasets [19].

h. Back Propagated Artificial Neural Network (BPANN) approach

BPANN was used on a dynamic-agent based load balancer that was proposed by Prakash and Lakshmi on the software-defined network (SDN) [20]. SDN is a component within the cloud architecture that is visible globally. As part of easing the work burden within the load, they are tasked to migrate the VMs within the data centre.

BPANN algorithm was trained on the VMs load and migration data. The resultant model was used to predict the VM load. The projected load was then used to determine the VM migration. Effective VM migration improves the network efficiency and the rate of data migration. Processing speed was reduced considerably since the heavy loads are matched to the underutilized VMs [20].

i. Quantum Neural Network (QNN) approach

QNN is a variation of neural networks that are based on the principles of quantum computing. Quantum circuits have been found to behave like artificial neural networks [21]. QNN have some computational advantages since they are made up of only the needed/necessary parameters to fit specific data. That feature makes them outperform their classical counterparts [21].

QNN model was used to predict the workload to be generated by the cloudlets. Singh and others encoded the workload data in qubits and the model was used to estimate the workload and the needed resources with much precision [22]. Their model used the Controlled-NOT (CNOT) gate was used as the activation function in both the hidden and the output layers that adjusted the qubit network weights [22]. Further network weight optimization was done by a self-balanced differential algorithm [22].

Intelligent load balancing has taken over from the traditional load balancing approaches. The use of machine learning and deep learning algorithms to develop load balancing models have improved the response time, resource elasticity and conserved power. For instance, Google adopted neural networks in its data centres to manage the cooling of the centres that reduced the power used for cooling by 40% [23]. This has shown the potential of artificial intelligence in addressing complex problems.

Table 2 has summarized the algorithms used in the intelligent load distribution techniques discussed above, the kind of data used for training those models and the major problem of load balancing it has addressed. It can be noted that deep learning is taking over the traditional machine learning models since it has offered better accuracy even with the growth of datasets. Hybrid solutions are the majority of the reviewed papers, with many researchers opting to combine several algorithms to form a model for instance: [18], [22], [24] and [17].

Table 3:A summary table showing reviewed intelligent load balancing techniques

No.	Publication	Underlying Machine/Deep Learning Model	Data Used	LB Problem Addressed
1	Deep Learning Regression [13]	CNN	Tasks workflow data	Quality of Service (QoS) resource utilization and throughput
2	Deep Learning-Based Load Balancer [14]	Hierarchical sub-models of FCN	Historical cluster access logs	Solves data skew problem in classical LB
3	Lilhore machine learning-based LB [15]	SVM, K-Means Clustering	RAM & CPU usage data	VMs resource utilization & execution time reduction
4	Reinforcement based SDN controller [16]	Bayesian Network & Reinforcement Learning	Network traffic data	SDN Controller LB, Network Stability, Security
5	Distributed database query load distribution [17]	multiple linear regression (MLR), random forest (RF) and AdaBoost (Ada)	Database queries data	Cloud Heterogeneity of CPU & GPU
6	Workload prediction [18]	Artificial Neural Network and self-adaptive differential	Client request amassed to time units	Distribution of workloads

		evolution (SaDE)		
7	Temporal aware LB [19]	LSTM-RNN	Cloud workload with a time factor	Resource elasticity & power saving
8	Dynamic agent LB [20]	Backpropagation Artificial Neural Network BPANN	Network Traffic Logs	VM migration, data migration
9	Quantum based LB [22]	Evolutional Quantum Neural Network EQNN	Cloudlets workload logs	Dynamic resource scaling

5. PERFORMANCE METRICS OF LOAD BALANCING TECHNIQUES

LB performance parameters can be measured through some of its quantifiable features. The metrics can help in identifying the best approach to load balancing. Some measurable attributes are directly measured while others are dependent on related variables. The following metrics were found to be effective in rating a load distribution component:

a. Throughput

Throughput describes the measure of the number of tasks/items passing through a process in each time interval. In load, balancing throughput can be considered as the number of activities the LB component can handle in a specific period [3]. For instance, the LB component has high throughput if it responds to requests since it will handle more tasks than one with delayed response.

How the LB component forwards the request and the time it takes to decide on which cluster to assign the workload affects throughput. The number of accomplished tasks within a specified period also measures throughput. This metric was found in these papers; [17], [13], [4], [16].

b. Migration Time

The duration of transfer is the time the LB component takes to move processes from overloaded devices to underutilized devices. In load balancing, migration time is measured as the time required to move VMs from one physical machine to another. Migration is initiated when a task requires execution through multiple VMs or when a task is interrupted [3]. A higher number of migrations result in more migration time. An effective load sharing technique minimizes VM migrations. i.e. [20], [16]

c. Response Time

This metric measures the time taken by the LB module to respond to a task/cloud request. Response time is calculated by summing up the transmission time, waiting time and service time [25]. A good LB technique maintains very minimal response time since performance is inversely proportional to response time. This metric is very common; [15], [16], [18]

d. Fault Tolerance

Fault tolerance describes the ability of a system to withstand failure. In load balancing, it provides the ability to perform uninterrupted service even when some of its parts fail [3]. An effective load balancer continues to function even when some hosts, VMs and PMs fail. The ability to solve logical errors ensures fault tolerance. This performance parameter is measured by having single or multiple points of failure. Redundant LB components are advised to ensure that the LB component does not fail. For instance; [16] addresses fault tolerance by decentralizing network control.

e. Power Consumption

This metric defines the amount of energy/electricity the VMs consume after the load balancing technique has been implemented. A well designed LB approach reduces power usage in the VMs [25]. Usually, a load balancer ensures that the VM are not overloaded and hence they use less energy. More power saving oriented techniques shut down the unused physical machines/hosts [4]. Some papers that put power into consideration include; [15]and [19].

6. FINDINGS AND DISCUSSION

The focus of this paper was to find the latest trends in cloud load balancing research by uncovering the most used machine learning algorithms in load balancing components. Both traditional machine learning and deep learning models were found as an adaptation of the big data age.

Traditional machine learning algorithms found include: Multiple linear regression (MLR) and Random Forest (RF) Classifier [17]; SVM and K-Means Clustering [15]. Due to the complexity of the load balancing and the large data associated with their training like the CPU logs, network traffic data and storage logs. Traditional machine learning algorithms are being replaced by deep learning models.

Deep learning models implemented in this area include BPANN [20], CNN [13], FCN [14], ANN [18], LSTM-RNN [19]. The deep learning models has exhibited better performance in terms of predicting the accuracy. These models handled the big data very well without compromising the quality of model. These models exhibit an important trend that moves from spatial oriented models like ANN and CNN to spatial-temporal models like LSTM and CNN-LSTM. These trend shows that time as an important consideration in the load balancing process.

Other deep learning models that stood out of the rest include the deep reinforcement learning [16] based load distribution component and the Quantum Neural Network [22] based load balancer. These trends are revolutionizing load balancing by harnessing the power of reinforcement learning by incorporating self-taught agents in reward-based system that continuously improve the prediction accuracy. Quantum computing power improves the training speed and the overall model throughput.

Hybrid models have delivered better results than the single algorithm models. Ensemble and transfer learning are major types of machine learning model combination paradigms used in many papers. For instance; [16], [22], [20], [15], [14] , [19] and [18] are all hybrid models combining either several machine learning algorithms or combing machine learning with other technologies such as quantum computing [22].

The quality of load balancing can be quantified using some metrics found in the reviewed papers. Most used metrics found almost in every paper include: through put, migration time, response time, fault tolerance and power consumption. Therefore, the criteria for defining the best load balancing model should select a model that has high throughput, less migration time, least response time, ability to withstand server failures and ability to save power.

7. RESEARCH GAPS

This paper was dedicated to intelligent LB techniques within cloud computing. A lot of research has been done within this field and state of the art performance has been achieved in many papers. Some of the notable trends were how the research moved from traditional machine learning models like regression to deep learning models like ANN and LSTM [18] [19] [20]. Another important landmark is the embrace of quantum computing to solve workload prediction [22].

Most of these research works are focused on single components of load balancing. For instance, some solutions focus on network load only [16], others focus on database management query load [17] and others focus on the data centre load. There is a need for integration to have a model that links the workload problem, the VMs allocation problem, network load.

Load balancers are the single point of failure and researchers have not given fault tolerance much thought. More work should focus on this area with models having redundant distribution components. Datacenters are located within different locations; the researchers have not accounted for the network delay and incorporation of edge computing to avoid costly data transfer and VM migration over long distances.

The power conservation mechanism has not been dealt with effectively. LB techniques so far lack incorporation of energy conservation in their LB design. Techniques should put into consideration energy saving while balancing.

8. CONCLUSION

The focus of this article is cloud load balancing. Load balancing is one of the largest problems in a cloud environment. LB can adversely affect the quality of service and the SLAs hence making the cloud host lose clients. The work of the LB component is to share the work burden across the cloud resources to ensure maximum utilization of the resources and efficiency of the growing devices.

This research has identified some intelligent load balancing techniques. Models used in these papers were highlighted. For instance, multiple linear regression, AdaBoost, random forest, k-suggest, k-means clustering, CNN, FCN LSTM, ANN, QNN and reinforcement learning. These models have shown the changing trend from traditional machine learning [17], to deep learning [19] components embedded with reinforcement learning [16] to foster continuous learning and finally harnessing of quantum computing power for cloud management [22].

REFERENCES

[1] DSM, "A Beginner’s Guide to Cloud Scalability and Load Balancing," DSM, 16 NOV 2018. [Online]. Available: https://www.dsm.net/it-solutions-blog/a-beginners-guide-to-cloud-scalability-and-load-balancing. [Accessed 14 Feb 2022].

[2] A10, "What is a Load Balancer and How Does Load Balancing Work?," A10, 2022. [Online]. Available: https://www.a10networks.com/glossary/what-is-a-load-balancer-and-how-does-load-balancing-work/#:~:text=Even%20though%20a%20load%20balancer%20solves%20the%20webserver,both%20front-ending%20the%20same%20group%20of%20web%20servers.. [Accessed 14 Feb 2022].

[3] M. A. Shahid, N. Islam, M. M. Alam, M. M. Su’ud and S. Musa, "A Comprehensive Study of Load Balancing Approaches in the Cloud Computing Environment and a Novel Fault Tolerance Approach," IEEE Access, vol. 8, no. I, pp. 130500 - 130526, 2020.

[4] T. Khana, W. Tian and R. Buyya, "Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions," CoRR, vol. arXiv:2105.05079v1, no. 1, 2021.

[5] Data Flair, "Cloud Computing Tutorial for Beginners – Learn Cloud Computing," Data Flair, 2020. [Online]. Available: https://data-flair.training/blogs/cloud-computing-tutorial/. [Accessed 17 Feb 2022].

[6] S. Parida and B. Panchal, "Environment, An Efficient Dynamic Load Balancing Algorithm Using Machine Learning Technique in Cloud," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 4, no. 4, pp. 1184-1186, 2018.

[7] A. A. Alkhatib, A. Alsabbagh, R. Maraqa and S. Alzubi, "Load Balancing Techniques in Cloud Computing: Extensive Review," *Advances in Science, Technology and Engineering Systems Journal*, vol. 6, no. 2, pp. 860-870, 2021.

[8] M. Liaqata, S. Ninoriya, J. Shuja, J. Shuja and A. Gani, "Virtual Machine Migration Enabled Cloud Resource Management: A

Challenging Task," *CoRR*, vol. abs/1601.03854, 2016.

[9] B. Sahoo, S. K. Jena and S. Mahapatra, "Load Balancing in Heterogeneous Distributed Computing Systems using Approximation Algorithm," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, Athens, 2013.

[10] D. A. Shafiq, N. Jhanjhi and A. Abdullah, "Load balancing techniques in cloud computing environment: A review," Journal of King Saud University – Computer and Information Sciences, vol. 02, no. 007, 2021.

[11] IBM Cloud Education, "What is Machine Learning?," IBM Cloud, 15 July 2020. [Online]. Available: https://www.ibm.com/cloud/learn/machine-learning. [Accessed 17 February 2022].

[12] IBM Cloud Education, "What is deep learning?," IBM Cloud, 1 May 2020. [Online]. Available: https://www.ibm.com/cloud/learn/deep-learning. [Accessed 17 February 2022].

[13] A. Kaur, B. Kaur, P. Singh, M. S. Devgan and H. K. Toor, "Load Balancing Optimization Based on Deep Learning Approach in Cloud Environment," *I.J. Information Technology and Computer Science*, vol. 3, no. I, pp. 8-18, 2020.

[14] X. Zhu, Q. Zhang, T. Cheng, L. Liu, WeiZhou and J. He, "DLB: Deep Learning Based Load Balancing," *CoRR*, vol. 1910, no. 08494V4, 2021.

[15] U. K. Lilhore, S. Simaiya, K. Guleria and D. Prasad, "An Efficient Load Balancing Method by Using Machine Learning-Based VM Distribution and Dynamic Resource Mapping," *Journal of Computational and Theoretical Nanoscience*, vol. 17, no. 7, pp. 2545-2551, 2020.

[16] S. Liang, W. Jiang, F. Zhao and F. Zhao, "Load Balancing Algorithm of Controller Based on SDN Architecture Under Machine Learning," *Journal of Systems Science and Information*, vol. 8, no. 6, pp. 578-588, 2021.

[17] A. Abdennebi, A. Elakas, F. Taşyaran, E. Öztürk, K. Kaya and S. Yıldırım, "Machine learning-based load distribution and balancing in heterogeneous database management systems," *Concurrency and Computation*, vol. 34, no. 4, 2021.

[18] J. Kumar and A. K. Singh, "Workload prediction in cloud using artificial neural network and adaptive differential evolution," *Future Generation Computer Systems*, vol. 81, no. C, pp. 41-52, 2019.

[19] J. Kumar, R. Goomer and A. K. Singh, "Long Short Term Memory Recurrent Neural Network (LSTM-RNN) Based Workload Forecasting Model For Cloud Datacenters," *Procedia Computer Science*, vol. 125, pp. 676-682, 2018.

[20] S. WilsonPrakash and P. Deepalakshmi, "Artificial Neural Network Based Load Balancing On Software Defined Networking," in *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing (INCOS)*, Tamilnadu, India, 2019.

[21] A. Abbas, D. Sutter and S. Wörner, "The power of quantum neural networks," IBM, 2 July 2021. [Online]. Available: https://research.ibm.com/blog/quantum-neural-network-power. [Accessed 16 Feb 2022].

[22] A. K. Singh, D. Saxena, J. Kumar and V. Gupta, "A Quantum Approach Towards the Adaptive Prediction of Cloud Workloads," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, pp. 2893-2905, 2021.

[23] W. Knight, "Google just gave control over data center cooling to an AI," MIT Technology Review, 7 August 2018. [Online]. Available: https://www.technologyreview.com/2018/08/17/140987/google-just-gave-control-over-data-center-cooling-to-an-ai/. [Accessed 17 February 2022].

[24] X. Sui, D. Liu, L. Li, H. Wang and H. Yang, "Virtual machine scheduling strategy based on machine learning algorithms for load balancing," *EURASIP Journal on Wireless Communications and Networking volume*, vol. 160, no. 1, 2019.

[25] S. K. Mishra, B. Sahoo and P. P. Parida, "Load balancing in cloud computing: A big picture," *Journal of King Saud University – Computer and Information Sciences*, vol. 32, pp. 149-158, 2020.

[26] GeeksforGeeks, "Top 10 Cloud Computing Research Topics in 2020," GeeksforGeeks.com, 26 September 2020. [Online]. Available: https://www.geeksforgeeks.org/top-10-cloud-computing-research-topics-in-2020/. [Accessed 17 February 2022]