European Journal of
Technology
(EJT)

Optimizing Resource Allocation in Cloud Computing Using Machine Learning

Jawaharbabu Jeyaraman, Samir Vinayak Bayani and Jesu Narkarunai Arasu Malaiyappan

European Journal of Technology
ISSN 2520-0712 (online)
Vol.8, Issue 3, pp 12 - 22, 2024
www.ajpojournals.org

Optimizing Resource Allocation in Cloud Computing Using Machine Learning

Jawaharbabu Jeyaraman[1*], Samir Vinayak Bayani[2] and Jesu Narkarunai Arasu Malaiyappan[3]
[1]TransUnion, USA
[2]Broadcom Inc, USA
[3]Meta Platforms Inc, USA

Article history
Submitted 15.02.2024 Revised Version Received 30.03.2024 Accepted 06.05.2024

Abstract

Purpose: A key component of large-scale distributed computing is the allocation of resources, as computer networks cooperate to address complex optimization problems. To get the most out of computers in general, or throughput, is the goal of resource allocation in this case. When it comes to distributed computing, there are two main varieties: grid computing and cloud computing. In grid computing, many geographically dispersed clusters are linked and made available to the general public.

Materials and Methods: We looked at Monte Carlo Tree Search and Long-Short Term Memory and examined how efficient they were. By maintaining consistent traffic patterns, the simulation demonstrated that MCTS performance was improved. However, such a strategy is difficult to implement due to the potential speed with which traffic patterns may alter. A successful service level agreement (SLA) was achieved, and the issue was shown to be fixable using LSTM. We compare the suggested model to different load-balancing algorithms to find the one that best allocates resources.

Findings: The results show that compared to the state-of-the-art models, the suggested model achieves an accuracy rate that is 10-15% higher. The suggested model lowers the error percentage rate of the average request blocking likelihood of traffic load by around 9.5-10.2% when compared to the predictions of existing models. Therefore, the proposed method has the potential to enhance network utilization by reducing the amount of time required by memory and the central processing unit.

Implications to Theory, Practice and Policy: One advantage of the new method is a more robust forecasting strategy in comparison to earlier models. Using firefly algorithms, future research will construct a cloud data center that employs a variety of heuristics and machine learning methodologies to load balance the energy cloud (Oshawa et al., 2022).

Keywords: *Cloud Efficiency, Resource Allocation, Load Balancing, Traffic Load, Cost of Service (CoS), Long-Short Term Memory (LSTM), Cloud Data Centre (CDC)*

1.0 INTRODUCTION

The efficiency and reliability of cloud services are only two of the many reasons why they have become so popular among consumers and companies in the modern day. Reducing energy consumption and environmental impact is a top priority for cloud data centers, thus they use management measures to that end. Thus, it is essential to implement new methods or improve existing ones to distribute energy-saving resources as widely as possible to achieve load balancing while implementing cutting-edge technologies such as blockchain and the Internet of Things. In large-scale distributed computing, where computers are connected to solve difficult optimization problems, resource allocation is a crucial component (Hasan et al., 2018).

Resource allocation, the subject of this article, aims to maximize total computer efficiency or throughput. Cloud computing stands out as a separate idea compared to grid computing, which requires the networking and accessibility of several clusters located in different places. Rapid use of cloud computing has made it the de facto norm for information technology network architecture. One reason for the skyrocketing rise of cloud computing components is the ever-increasing number of people using the Internet and their disposable cash (see Figure 1). At this time, there is no denying that cloud computing is the most cost-effective IT innovation that businesses can adopt. Now that they have access to computers, small, medium, and failing businesses may compete with larger firms. It is a system that uses virtualization and service-oriented software to achieve its goal of developing with little or no limits, thanks to its flexibility of usage. The system can adapt thanks to its adaptability (Khan, Tian, and Buyya, 2021).

To sum up, this paper examines various optimization methods and machine learning algorithms for cloud resource allocation, specifically looking at how genetic algorithm (GA) and other optimization techniques can provide the best output. If you want to improve energy efficiency and do performance analysis to find the optimal load balancing strategy, you must understand these strategies. We also show how to use cloud-based energy usage prediction using machine learning methods like SVMs and deep neural networks. Using the LSTM machine learning method on two network traffic loads, Euro28 and US26, respectively, we provide a framework for optimizing resource allocation in the cloud, hence increasing energy efficiency. Our presentation concludes with a demonstration of how machine learning-based multi-objective optimization techniques can instantly improve service quality, decrease SLA violations, and balance load while reducing energy consumption.

In recent years, a paradigm known as "cloud computing" (CC) has emerged as the dominant one in the fields of information and communication technology (ICT). Although they indirectly or directly support their daily search service via Internet activities, cloud users may not always understand the benefits of cloud innovation. Cloud computing has grown in popularity as a communication word due to its significance in the computer and engineering fields. The unrestricted availability of cloud computing services allows developing and impoverished countries to quickly advance economically. Before the era of cloud innovation, it was difficult for a corporation to construct a conventional data center because of the high initial equipment investment and ongoing maintenance costs. Cloud computing, on the other hand, allows us to rent computing resources on an as-needed basis, making programme deployment a breeze. With so many companies, both large and small, looking for ways to cut costs without sacrificing efficiency, optimizing CC innovation services is a must. This is because these services offer a plethora of benefits that directly correlate to what businesses need. When cloud computing performance techniques are grounded in a utility-based commercial model, users have easy, consistent, and scalable access to a common pool of programmable network assets.

In recent years, a paradigm known as "cloud computing" (CC) has emerged as the dominant one in the fields of information and communication technology (ICT). Although they indirectly or directly support their daily search service via Internet activities, cloud users may not always understand the benefits of cloud innovation. Cloud computing has grown in popularity as a communication word due to its significance in the computer and engineering fields. The unrestricted availability of cloud computing services allows developing and impoverished countries to quickly advance economically. Before the era of cloud innovation, it was difficult for a corporation to construct a conventional data centre because of the high initial equipment investment and ongoing maintenance costs. Cloud computing, on the other hand, allows us to rent computing resources on an as-needed basis, making programme deployment a breeze. With so many companies, both large and small, looking for ways to cut costs without sacrificing efficiency, optimising CC innovation services is a must. This is because these services offer a plethora of benefits that directly correlate to what businesses need. When cloud computing performance techniques are grounded in a utility-based commercial model, users have easy, consistent, and scalable access to a common pool of programmable network assets.

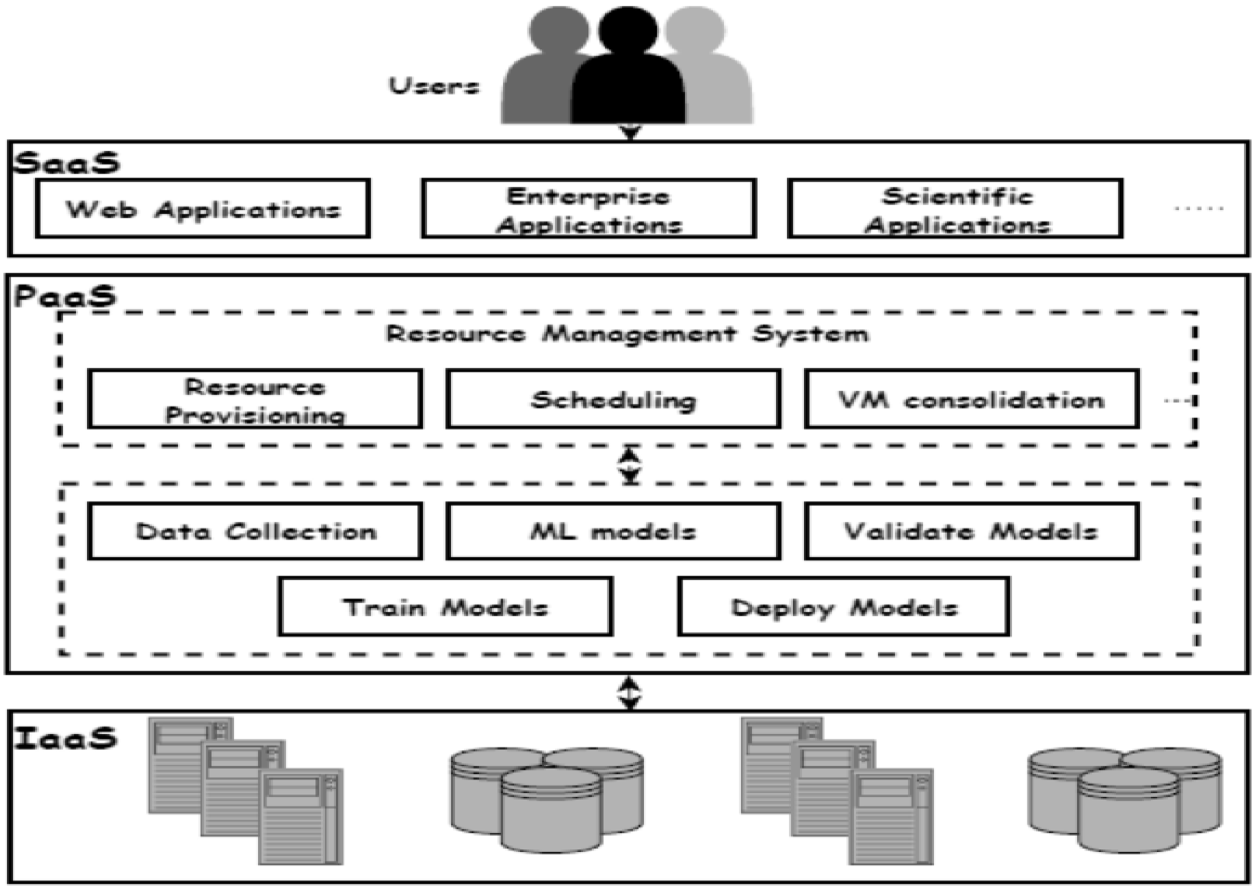

Figure 1: Machine Learning Cloud Computing

Body

The advent of cloud computing has opened up three new avenues for the use of computer resources. The following are descriptions of these three methods. Spending a lot of time and energy upfront on buying hardware, software, and the IT supply chain is unnecessary. Stressing out about such matters is unnecessary. One strategy that falls under this category is the ability to pay for services only when they are required. Cloud users have access to accessible information technology resources in the same way that we routinely utilize water and gas. Customers have access to a plethora of network, storage, computational, and software capabilities by connecting to a remote server at a data center that is maintained by an external party. Some examples of such third-party services are Microsoft Azure, Amazon Web Services (AWS), Google, Facebook, and Google.

Research, industry, academia, and business have all taken an interest in cloud computing innovations due to their growing share of IT spending and other desirable performance traits (e.g., rapid resource aggregation, network dominance, etc.). According to NIST, "cloud computing" is "a framework that enables widespread, comfortable, on-demand internet backbone access to a consensual pool of resources that can be rapidly allocated and initiated with minimal coordination by the service provider." The use of cloud computing is predicted to skyrocket in the next years. Assigning virtualized resources for information and communication technologies is a difficult topic to handle in a cloud computing paradigm. An increasing number of factors, including rising consumer expectations and decreasing available resources, are making resource allocation an increasingly difficult process. Several new models and ways of allocating resources have emerged as a result of this. Methods and models for dynamic resource allocation have been used in several ways. To make the most efficient use of available resources, these models and methodologies consider several constraints (Malik et al., 2022).

Using real-time traffic studies to make predictions about network capacity is one of the main hurdles to increasing the efficiency of cloud computing. The fundamental reason for this is that the network management plane, which primarily deals with network resources, overlooks the availability of cloud resources entirely. One possible solution to these problems is to automate fault management and network self-configuration using DL and ML technology. Deep learning and ML studies on optical networks have mostly relied on supervised methods (Manimegalai and Durai, 2021). Its basic idea is that models should be applied to real-world situations after being trained on historical data. In most cases, WANs do not care about this limitation since traffic patterns could change at any time. Creative analytics are required if we are to glean actionable insights from this deluge of network data. Some have speculated that the theoretical underpinnings of DL and ML could hold the key to solving the difficulty of processing network data (Powell, 2023).

A key component of cloud service optimization is automated dynamic resource allocation, which aids in making the most optimal use of available computing resources. To do this, one must modify their use of cloud resources. We may accomplish both goals by aligning the use of cloud resources with the previous one. To intuitively deliver cloud resources ahead of demand, this study utilizes an automated dynamic resource allocation system that is based on machine learning algorithms (Zhang et al., 2022).

When analyzing heuristic data from resource utilization when users use certain applications, this technique considers user settings. After that, it recommends the optimal resource for that specific task. Reallocating unused cloud resources is possible using this allocation mechanism, which helps keep resource utilization at a minimum (Oshawa et al., 2022). The goal of this research is to lay out all the different kinds of machine learning algorithms and optimization strategies for cloud resource allocation. In particular, the study will look at how optimization methods like genetic algorithms (GA) might provide the best performance in the area being studied. Improving energy efficiency and performance analysis requires a strong grasp of these tactics, which are crucial for selecting the most efficient approach to load balancing. Furthermore, we show how several machine learning techniques, including SVMs and DNNs, may be used to forecast cloud energy consumption. Using the LSTM machine learning technique on two distinct network traffic loads Euro28 and US26 they provide a method for improving cloud energy efficiency via optimal resource allocation. Finally, we show how machine learning-based multi-objective optimization methods can efficiently distribute resources by balancing the load, reducing energy consumption, reducing SLA violations, and instantly improving service quality (Hasan et al., 2018).

In the years after Amazon's debut, cloud computing became standard fare. In 2006, Elastic Compute Cloud released its product. Because of this, other major service providers were able to adopt cloud computing and build more resilient networks for their cloud systems. Cloud computing's adaptability and pay-as-you-go pricing mechanism make it an interesting innovation. A key component of cloud computing systems is the deployment of a massive central server across several geographic locations, with resources allocated from these servers according to demand. The demand for cloud computing services has increased in tandem with the availability of increasingly sophisticated tools. Organizations and industries are continuously seeking a storage device-rich, high-capacity network that can let them execute their operations on budget PCs. The cloud has seen a dramatic increase in use due to the ubiquitous character of modern business. As an example, in 2019, cloud virtualization and custom architecture allowed Linux to be extensively utilized and made accessible for different platforms.

Data centers hold software applications that need a lot of processing power, which is essential for all these procedures. Several challenges remain to cloud innovation, despite cloud computing's rising profile in the IT sector as a result of its many advantages. Difficulties with adoption strategies, governance, data compliance, security concerns, and energy efficiency uncertainty are among these hurdles. The goal is to find practical solutions to these problems since they are concerning. The phrase "cloud computing" (CC) describes a model that has lately dominated the ICT (information and communication technology) landscape. While consumers directly or indirectly support their daily search service via Internet activities, they may fail to recognize the importance of cloud innovation. Cloud computing is a well-used word in discussions about technology and engineering because of its significance in these fields. Rapid economic development is possible for developing and impoverished countries thanks to cloud computing technologies, which make necessary services available to all users without restriction. It used to be expensive for businesses to set up conventional data centers, what with the upfront investment in equipment and ongoing maintenance costs (Khan, Tian, and Buyya, 2021).

That was before the cloud innovation phase. Cloud services, on the other hand, allow us to rent computer commodities as needed and install programs with little to no hassle. As a result of the many advantages that correspond with business needs, optimizing CC innovation services becomes inevitable as many companies, both large and small, strive to balance operational costs with access to superior efficiency tools. With cloud computing performance strategies built on a utility-based business model, customers may access a shared pool of programmable network assets easily, consistently, and scalable. Cloud computing is built around this principle. Customers are free to use the channels that work best for them since the process is dependable and adaptable. The use of cloud computing makes this feasible (Malik et al., 2022).

Allocation of Resources

The purpose of the criteria used by resource allocators and algorithms for resource allocation was investigated by Naik and Kavitha Sooda. The study's criterion for the resource allocator was based on factors such as processing time, system dependability, resource consumption, and allocation cost. The offered resource allocator structure also took the current state of the resource, the service level agreement, and the user's request into account. The audience was also shown the steps to create the model of the resource allocator (Manimegalai and Durai, 2021).

Utilization of Multiple Resources in Virtual Machine Consolidation: They may increase the number of virtual machines (VMs) while decreasing the number of hosts and the energy required to run them by consolidating VMs into fewer hosts. To determine whether the host was overloaded, the majority of the analysis was on its current CPU time utilization. If changes to host energy modes and relocations of virtual machines are not essential, consolidation attempts could be slowed down (Zhang et al., 2022).

Cloud network traffic and temperature: The current study on virtual machine allocation states that to provide resources to a host, each host must have the necessary hardware and software. This method results in inefficient use of resources because application demand is unpredictable and might vary from very little to very large. It is difficult to lower the host temperature in contemporary cloud data centers. When the host uses energy, it generates this kind of heat. Cooling systems are used to recover this heat loss and keep the host temperature below the critical threshold. This higher temperature presents an evident problem for resource management systems due to the correlation between it and the rising expense of cooling equipment (Oshawa et al., 2022).

Energy metering in software: The power consumption of a virtual machine (VM) cannot be measured by the many energy meters that are included in current servers. This is because it is a difficult and costly undertaking to precisely measure the energy consumption of software. Data center energy budgets reveal that the growing cost of operating servers has impeded development in the VM compression phase (Hasan et al., 2018).

2.0 MATERIALS AND METHODS

Continuous monitoring of network traffic load and the use of Long Short-Term Memory (LSTM) machine learning technologies are necessary for improving cloud efficiency via improved resource allocation approaches for load balancing. The input gate of the LSTMP unit controls the control signal that enters the memory cell, and this part focuses on using the Long Short-Term Memory (LSTM) approach to emulate it (Khan, Tian, and Buyya, 2021).

Long Short-Term Memory (LSTM) System Foundations

An LSTM-enabled recurrent neural network was suggested by Hochreiter and Schmidhuber. Regular recurrent neural networks have difficulties while training lengthy temporal connections because their gradients disappear or fade with time. Conversely, LSTM runs the danger of becoming too reliant if it relies on "constant error carousels" (CEC) to maintain an infinite error flow. A lot of changes have been made to the original LSTM since then.

Someone was curious about the LSTM implementation in Sak's "predicted" version.

There are entrance and exit gates on devices that employ LSTMP technology. Both the input and output gates of the LSTMP unit regulate the flow of control signals into and out of the memory cell. The LSTMP forget gates enable memory cell resets and adaptive forgetting. The LSTMP units include both recurrent and non-recurrent projection layers. There is just one parallel layer instead of the two projection levels.

One kind of RNN that gets around the gradient-growth problem is the Long Short-Term Memory (LSTM) Neural Network. When a gradient problem arises, the neural network is unable to learn how to fix errors efficiently via backpropagation. Due to the RNN's (Recurrent Neural Network) inability to learn from massive datasets, which implies that it has a limited memory, the Long Limited-Term Memory variant was created. A single memory cell serves as the core component of the LSTM design, which resembles a chain (Figure 2). Above, you can see large square blocks that, when assembled, will stand in for individual memory cells (Malik et al., 2022).

An essential component of LSTM, the horizontal line that crosses the cell's top represents its state. The "hinge" of an LSTM network is made up of individual cells, and each of those cells helps to produce it. This cell may have as many or as few states as the LSTM algorithm deems necessary. One more way that LSTM structures may do this is by including gates. The sigmoid activation function generates gates (seen in Figure 2) and also performs operations like pointwise multiplication. Tree gates regulate the transfer of cell status information; they consist of the forget, input, and output gates seen in the image above.

While searching for LSTM networks, Hochreiter and Schmidhuber came upon them. Memory cell layouts became the subject of subsequent experiments by academics from a variety of disciplines (Manimegalai and Durai, 2021).

Algorithms

Closest Data Centre

As a first and most obvious step, we used the Closest Data Center (CDC) algorithm to allocate traffic within the closest data center. The k-shortest paths were calculated from the request origin to the closest DCs. If it's feasible, we can then utilize the set of potential paths to assign a request to a certain DC. With the returned route to DC serving as the starting point, the optical layer requests were dispersed using the RMSA approach. Because of this, we will not be granting your request. The method's temporal complexity grew in proportion to the number of possible paths (Powell, 2023).

European Journal of Technology
ISSN 2520-0712 (online)
Vol.8, Issue 3, pp 12 - 22, 2024

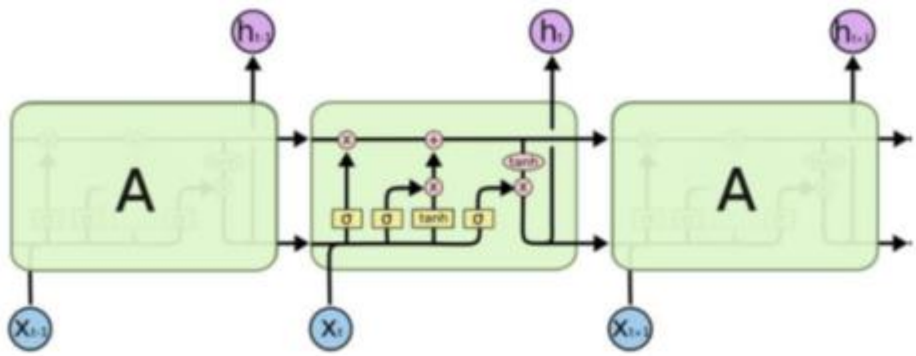

Figure 2: LSTM Cell

Forget Gates for Long Short-Term Memory

All recurrent neural networks have a core collection of components called Tere. Figure 3 shows the basic layout of these modules; it's not complicated at all; it only has one hyperbolic function, tanh. While long short-term memory (LSTM) networks seem like a chain, they consist of four interconnected neuronal layers in each module (Zhang et al., 2022).

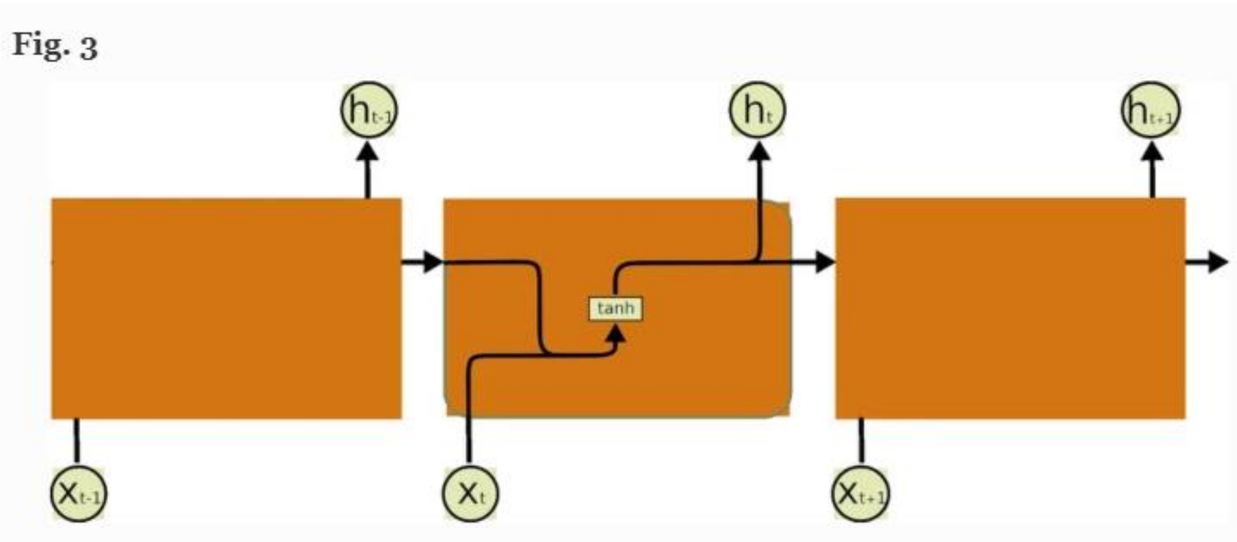

Figure 3: Single Recurrent Neural Network

Frameworks, Toolboxes and Risk Assessment

The LSTM machine learning algorithm, which is part of the deep learning class library, was used as a tool for the technical development of this work. You need a 64-bit Java Virtual Machine (JVM) or a 64-bit Java Development Kit (JDK) installed on your system to use this library. System versions of Java Development Kit (JDK) lower than JDK 7 are not compatible with the Deeplearning4J library since it requires at least JDK 7. Included in the Deeplearning4J are feature extractors and pre-processors for machine learning datasets. It made the training and parameter configuration step easier, which allowed for the retraining of the taught system until it reached an efficient state where it could intuitively allocate resources properly.

The study's risk approach is based on the principle of risk avoidance, which states that to mitigate a risk, one must take measures to make sure it doesn't materialize. Including the PC for development, papers for literature, and the Deep Learning 4 J library, all of the resource items were tested and functioning when obtained during the early phases of this project. The technological challenges of constructing the application for this research were mitigated by acquiring and reviewing relevant resources that had all the necessary information to create an efficient application, while also avoiding typical obstacles seen in comparable endeavors (Ashawa et al., 2022).

Comparison with the Recent State-of-the-Art

It lends credence to the results showing that different network architectures affect CoS and BP. When their trends remain constant, MCTS offers somewhat cheaper rates for large traffic compared to LSTM since calculating the network output and utilizing backpropagation is less costly. Furthermore, when the request pattern is subject to rapid fluctuations, the trends separate. Therefore, the LSTM becomes the most budget-friendly choice. When LSTM can identify changes in traffic patterns early enough, it may "forget" about previous knowledge and start using the new patterns to apply new rules. The reason MCTS takes longer to adapt to changing traffic situations is that it is constantly building search trees without taking the rapid changes into account.

Both algorithms produced similar costs when traffic was low. Since they don't need a lot of network resources, incorrect routing choices don't significantly affect the CoS when traffic levels are low. In situations with rising traffic volumes and less predictable patterns, Te LSTM outperforms rival algorithms significantly. The last example clarifies the idea. The cost increases at an exponential rate for every bad decision due to the redistribution of rejected requests brought on by resource constraints and interrupted network connections. A thorough understanding of the application's lifecycle, available resources, and load as key performance indicators should inform decisions on leasing and resource reallocation. We compared our findings to the state of the art using the research. These problems, if resolved, will demonstrate how challenging it is to comprehend the inner workings of the cloud, or any large-scale computer system (Hasan et al., 2018).

The first is that cloud solutions that rely on computer operating systems cannot guarantee anything in real time. Secondly, and most importantly, there has to be a basic theory that can be used to drive the development of usable instruments for program performance forecasting and regulation. Cloud settings offer an additional virtualization layer on which cloud applications operate, making this fundamental situation for computer systems even more noticeable.

Clusters may be dynamically created depending on the application's state and the currently processed request, and this way would enable resources to be controlled separately within them. A mix of centralized and distributed control approaches is likely to be used by adaptive load balancing. This would make it possible to modify the trade-off between efficient resource use and reliable workflow. According to the results, when compared to previous models, the suggested one improves accuracy by about 10-15%. This indicates that the suggested method outperforms competing models in terms of network utilization while requiring less time to complete owing to its superior predictive approach (Manimegalai and Durai, 2021).

3.0 CONCLUSION AND RECOMMENDATIONS

This study built an intuitive dynamic resource allocation system that evaluated the heuristics program's resource consumption to determine the appropriate additional resource to supply, using an LSTM algorithm application as its base. In a flash, the software could mimic the allocation of resources by the trained LSTM model. Possible gains might be achieved by combining them with cloud data center approaches for dynamic routing.

We contrasted MOTS with a short-term memory test. Research shows that when the traffic trend is static in a simulated setting, MCTS performs well. Due to the ever-changing nature of traffic patterns, this is often not an option. The use of LSTM allowed us to establish a satisfactory service level agreement (SLA) while simultaneously confirming that this problem could be remedied. We propose several heuristics and machine learning methods for use in the development and deployment of algorithms in the future that make use of cloud computing. A more comprehensive evaluation of optical and data center networks' present and future resource requirements is required. This advancement and the application of algorithms for auxiliary physical models will assist elastic optical networks that utilize traffic prediction methods like the Las Vegas algorithm, which does not rely on LSTM and Monte Carlo Tree Search. Our study did not account for the energy consumption of individual devices on the application server, network nodes, or personal terminals (such as desktops, phones, or laptops). Still, we used several performance metrics that impact load balancing response time, predictability, dependability, scalability, fault tolerance, associated overhead, throughput, and thrashing to improve system stability via fair allocation of virtualized resources.

Therefore, in both the wired and wireless scenarios, our method was unable to identify the power-saving features. Second, the LSTM may improve performance and load balancing, according to our system's tests. However, generalization is not achievable when the US26 and Euro28 networks are the only ones used. So, before we can generalize our results, we need to see how they perform on other types of network data.

Topology-Aware Resource Allocation (TARA) is a model that Lee et al. developed using Genetic Algorithms (GA) and lightweight simulators. It can distribute resources reliably in an Infrastructure as a Service (IaaS) context.

The objective of this approach was to enhance Map Reduce; when tested against application-independent allocation, it achieved a 50% improvement in work completion time. For an infrastructure-as-a-service (IaaS) based cloud service paradigm, Toosi et al. created a Resource Allocation System (RAS). Their clients' companies' prices and profits were boosted as a result of this. This RAS makes advantage of a suggested strategy to improve resource utilisation by acquiring unused resources from other service providers. In order to optimise computation for greater eco-friendly compute utilisation, Xiao, Song, and Chen used a different method when it came to the IaaS cloud service paradigm. A skewness method, which incorporates a series of heuristics to avoid system overload, assesses the mismatch of resources in multi-dimensional resource utilization, and here is how they do it.

REFERENCES

Hasan, Md., E, B., Almamun, Md. and K, S. (2018). An Intelligent Machine Learning and Self Adaptive Resource Allocation Framework for Cloud Computing Environment. *EAI Endorsed Transactions on Cloud Systems*, 0(0), p.165501. doi:https://doi.org/10.4108/eai.13-7-2018.165501.

Khan, T., Tian, W. and Buyya, R. (2021). *Machine Learning (ML)-Centric Resource Management in Cloud Computing: A Review and Future Directions Resource Provisioning VM consolidation Scheduling Train Models ML models Data Collection Validate Models Deploy Models Resource Management System IaaS PaaS Web Applications Scientific Applications Enterprise Applications SaaS Users Figure 1: Components of Cloud computing Paradigm using Machine Learning*. [online] Available at: https://arxiv.org/pdf/2105.05079.pdf.

Malik, S., Tahir, M., Sardaraz, M. and Alourani, A. (2022). A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques. *Applied Sciences*, 12(4), p.2160. doi:https://doi.org/10.3390/app12042160.

Manimegalai, R. and Durai, U. (2021). Optimizing Resource Allocation In Cloud Computing. *https://www.webology.org/data-cms/articles/20220927064958pmwebology%2018%20(3)%20-%20132.pdf*, [online] 18(3), p.1776. Available at: https://www.webology.org/data-cms/articles/20220927064958pmwebology%2018%20(3)%20-%20132.pdf [Accessed 11 Mar. 2024].

Oshawa, M., Douglas, O., Osamor, J., and Jackie, R. (2022). Improving cloud efficiency through optimized resource allocation technique for load balancing using LSTM machine learning algorithm. *Journal of Cloud Computing*, 11(1). https://doi.org/10.1186/s13677-022-00362-x.

Powell, M. (2023). Optimizing Resource Allocation in Cloud Computing Environments Using Machine Learning. *Journal of Computer Science & Systems Biology*, [online] 16(2). doi:https://doi.org/10.37421/0974-7230.2023.16.457.

Zhang, Y., Liu, B., Gong, Y., Huang, J., Xu, J. and Wan, W. (2022). *Application of Machine Learning Optimization in Cloud Computing Resource Scheduling and Management*. [online] arXiv.org. doi:https://doi.org/10.48550/arXiv.2402.17216.