MZ Journals Computing Journal

AI-Driven Adaptive Network Capacity Planning for Hybrid Cloud Architecture

Kapil Patil[1,*], Bhavin Desai[2]

[1] Oracle, Seattle, Washington, USA
[2] Google, Sunnyvale, California USA
Corresponding author: kapil.patil@oracle.com

Abstract: Organizations have been increasingly adopting hybrid cloud architectures that integrate private and public cloud deployments to leverage the benefits of both environments. However, this hybrid approach poses fa challenge in making accurate predictions on network traffic needs between the private and public cloud constituents due to the dynamic nature of scaling workloads. Traditional capacity planning techniques are inadequate in coping with the quick variances that occur in cloud workloads. This challenge has led to the emergence of AI-based adaptive network capacity planning as a viable option that employs advanced machine learning (ML) and deep learning (DL) technologies to predict future patterns in network traffic with accuracy and dynamically assign network resources within hybrid clouds. This paper proposes an AI model that continuously learns from real-time network traffic data, workload information, and historical trends to predict future network capacity needs and dynamically adjust resources accordingly. The proposed approach involves a hybrid architecture combining Long Short-Term Memory (LSTM) neural networks for capturing temporal patterns and ensemble learning techniques for handling non-linear relationships and complex feature interactions. By leveraging the strengths of both paradigms, the AI model aims to capture complex patterns and dependencies within the data, enabling accurate predictions and proactive resource scaling in hybrid cloud architectures.

***Keywords**:* Artificial intelligence, Neural networks, Capacity planning, Machine learning, Predictive analytics, Cloud computing, Edge computing, Next generation networking, Network architecture, Deep Learning.

I. Introduction

To get the best of both worlds, organizations have been increasingly adopting hybrid cloud architectures that integrate private and public cloud deployments. As for sensitive data and mission-critical workloads, private clouds provide control, security, and compliance benefits while public clouds offer cost-effectiveness, scalability, and access to a wide range of services[1]. Consequently, this hybrid approach poses a challenge in making accurate predictions on network traffic needs between the private and public cloud constituents because of scaling workload's dynamicity that is impossible to predict with accuracy.

Capacity planning techniques of traditional networks are inadequate in coping with the quick variances that occur in cloud workloads as they mainly depend on historical data analyses and manual forecasting methods. These procedures may be unable to accommodate abrupt surges or slumps in demand, causing either excess provisioning or under-provisioning of network resources. Wasteful costs come about from over-provisioning because some resources may be left unutilized whereas performance degradation, latencies, and potential service disruptions can result from under-provisioning[2]. The fallacies inherent in traditional capacity planning approaches become more pronounced as cloud environments get increasingly intricate and dynamic.

This challenge has led to the emergence of AI-based adaptive network capacity planning as a viable option that employs advanced machine learning (ML) and deep learning (DL) technologies to predict future patterns in network traffic with accuracy and dynamically assign net resources within hybrid clouds. Through real-time analysis of the historical data, workload data, and patterns, the AI model under consideration may scale up network resources ahead of time to improve performance while simultaneously reducing over-provisioning costs. This smart and changeable strategy aims at facilitating uninterrupted interaction between private and public cloud portions helping enterprises take full advantage of hybrid cloud computing benefits.

While existing AI-based approaches have shown potential in network traffic prediction and anomaly detection, they primarily focus on single cloud environments or data centers. The challenge of accurately predicting and adapting network capacity for hybrid cloud architectures, where workloads dynamically scale between private and public clouds, **remains largely unexplored.** Existing research on hybrid cloud network management has mainly focused on load balancing, fault tolerance, and security aspects but has not comprehensively addressed the specific problem of AI-driven adaptive network capacity planning for dynamic workload scaling across multiple cloud environments. This paper proposes a novel AI-driven approach that continuously learns from real-time network traffic data, workload information, and historical trends to accurately predict future network capacity needs and dynamically adjust resources in hybrid cloud architectures.

2. Background and Related Work

The traditional approach of network capacity planning typically revolves around examining the historical information of traffic, making estimates concerning future requirements based on business assumptions, and therefore adjusting network elements manually[2]. Usually, these alternatives are inaccurate since they rely on simple linear growth patterns or seasonal trends that may not be representative of the intricate and dynamic nature of cloud workloads. Consequently, they can lead to poor resource utilization resulting from over-provisioning in anticipation of possible spikes or under-provisioning leading to performance bottlenecks and service outages. AI techniques, such as machine learning and deep learning, have been acknowledged by researchers in recent years as offering the possibility of bettering network traffic prediction and anomaly detection. Xhao and He[3], carried out a Long Short-Term Memory (LSTM) neural

network model which was for predicting network traffic and optimizing resource allocation in software-defined networks (SDNs). This methodology made extensive use of LSTM's ability to capture long-range dependencies and temporal trends in the data on network traffic, improving forecasting accuracy compared to traditional time series forecasting methods. Similarly, Bala and Chana developed a DL-based approach for cloud data centers' traffic prediction that allows proactive resource allocation and load balancing leading to improved overall system performance and resource utilization.
While these studies have demonstrated the potential of AI in network management, they primarily focus on predicting traffic patterns within a single cloud environment or data center. The challenge of accurately predicting and adapting network capacity for hybrid cloud architectures, where workloads dynamically scale between private and public clouds, remains largely unexplored. Existing research on hybrid cloud network management has mainly focused on load balancing, fault tolerance, and security aspects[5], but the specific problem of AI-driven adaptive network capacity planning for dynamic workload scaling between private and public clouds has not been comprehensively addressed.
Furthermore, traditional capacity planning methods and existing AI-based approaches often fail to account for the complexities introduced by hybrid cloud architectures, such as heterogeneous network infrastructures, diverse workload characteristics, and the dynamic nature of resource scaling across multiple cloud environments[8]. These limitations highlight the need for a more comprehensive and intelligent approach to network capacity planning in hybrid cloud environments, capable of adapting to rapidly changing conditions and optimizing resource allocation across private and public cloud components.

Existing solutions fall short in addressing the unique challenges posed by hybrid cloud architectures, where workloads can dynamically scale across different cloud environments, each with its own network infrastructure, workload characteristics, and resource management requirements[9]. A holistic approach that can handle these complexities and provide accurate, adaptive capacity planning tailored to hybrid cloud environments is critically needed.

3. Proposed Approach

For this reason, the proposal includes an AI model that perpetually gains knowledge from live network traffic data, workload information, and previous trends to estimate what future network capacity requirements will be and dynamically redistribute resources. By using sophisticated ML and DL algorithms, the AI model is expected to capture intricate relationships and interdependencies among the variables, thus enabling accurate forecasts and proactive resource scaling for hybrid cloud architectures. The AI model is designed to be flexible and adaptable as it continuously updates its outputs based on a dynamic context of hybrid cloud environments where workload requests might change suddenly and unexpectedly.

A. Data Collection and Preprocessing

The initial stage of the proposed strategy starts by gathering data from different sources such as network monitoring tools, cloud management platforms, and application performance monitoring systems. This data is composed of a lot of different things like network traffic data (like

bandwidth usage, packet loss, latency), workload information (like the number of virtual machines, container instances, and resource utilization), and historical trends (such as seasonality patterns, past capacity adjustments, and anomalies)[10]. It is important to have diversified and exhaustive information from different sources for a complete view of the hybrid cloud environment and to capture intricate relations between factors affecting networking capacity requirements.

After collecting the data, it is preprocessed to handle missing values, and outliers and maintain uniformity of feature representations. This stage might include activities such as data filling procedure, finding anomalies, or scaling to make the process efficient for the AI model. Data preprocessing helps increase data quality, deal with noise and inconsistency, and enable an AI model to learn best from such content available[7]. Furthermore, other techniques of feature engineering are applied to remove raw features from which relevant features can be identified that will help improve the performance of predictive models by picking the most informative and discriminative features.

B. AI Model Architecture

The recommended AI model is based on a combination of conventional machine learning (ML) techniques and deep learning models incorporating the best aspects of both approaches. Consequently, this hybrid architecture has the objective to capture temporal patterns in network traffic and workload data, as well as intricate non-linear relationships and interactions between different features and factors[11].

A Long Short-Term Memory (LSTM) neural network is used to capture the network traffic and workload data long-term dependencies, and temporal patterns[7]. LSTM's ability to determine which information to keep and forget makes it adequate in time series modeling as well as complex capture of time-tied relationships that are vital for precise network traffic forecasting. The LSTM part of this hybrid model learns and encodes the sequential patterns within the data thereby enabling it to make accurate predictions grounded on previous observations along with trends.

To deal with non-linear relationships within the data, additional features are integrated into an LSTM model using Ensemble Learning techniques such as Random Forests and Gradient Boosting. Ensemble Learning techniques combine numerous base models to improve prediction accuracy and robustness, thus overcoming the disadvantages of the one-model approach. The hybrid model's ensemble module captures intricate feature interactions and nonlinear relationships that can be difficult for individual models to efficiently learn. Future network capacity requirements are predicted by the ensemble model using LSTM outputs in combination with other relevant features like past trends, workload characteristics, and network configurations among others.

This model is constructed with a fusion of ensemble learning and deep learning techniques to address non-linear relationships, complex interaction among different features, as well as the ability to capture temporal patterns. As a result, the model can learn from various data sources

and keep up with the changing faces of hybrid cloud environments to effectively determine network capacity requirements since they are driven by both temporal patterns and complex feature interactions.

C. Model Training and Optimization

The AI model has been raised on historical data using a combination of supervised and unsupervised learning techniques. LSTM and ensemble models are trained using supervised learning where the target variable is the required network capacity[7]. This training process involves minimizing the prediction error of the model's parameters on labeled data by using backpropagation and gradient descent[9]. In this way, the model can accurately map input features (e.g., workload information, network traffic, historical trends) to desired output (network capacity needs).

Furthermore, apart from supervised learning, unsupervised learning methods are employed to recognize clusters and reduce the dimensions of the input features to identify patterns and anomalies in the unlabeled data. By so doing, it is possible to reveal some hidden structures or relationships among elements represented by data that might later be useful for making the model work better and generalize data better[9]. By exploring the unlabeled data through unsupervised learning methods, the model can discover inherent patterns and structures that may not be explicitly provided in the labeled data, potentially enhancing its ability to generalize to new and unseen scenarios.

Optimizing the AI model's performance requires that hyperparameters be tuned. Grid, random, and Bayesian search techniques are used for this role, where different hyperparameter values are tried systematically to maximize the performance of the machine learning algorithm. These parameters include but are not limited to learning rates, regularization factors, and the architecture of the models used in achieving better performances[7]. The proper combination of these factors is a major determinant of desired results as it changes the model’s overall performance considerably. Also, it has been shown that transfer learning and domain adaptation can make our programs more flexible when we have irrelevant historical data since they allow us to build on existing knowledge in related fields[10]. This means that by borrowing from other similar areas or adjusting the model to address domain shifts, relevant know-how already available can be capitalized on thereby increasing accuracy when resources for generating new data are scarce.

4. Evaluation

To assess how efficient the suggested adaptive network capacity planning approach driven by AI is, we will use a holistic appraisal technique entailing both simulated environments and real-world testing ground configurations. It will be an exercise of appraising the accuracy of simulation models in forecasting; the efficiency of resource usage; and cost vs. value added ratio, while at the same time comparing it with usual methods of capacity planning and other AI-powered approaches already in existence[7]. The purpose of this evaluation is to give an extensive analysis of their strengths, and drawbacks as well as areas that can be enhanced impartially.

On one hand, simulation settings will be designed to replicate hybrid cloud architectures that display different workload patterns and network configurations, which allow for controlled experimentation with an AI model under a variety of situations. On the other hand, real-world test bed setups will involve deploying the suggested method on production-like environments and exposing it to actual challenges and complexities encountered in hybrid clouds. By considering both simulations and real-world evaluations, a complete picture can be obtained of the performance, resilience, and practicality of the model

A. Simulation Environment

A simulation environment shall be created for modeling hybrid cloud architectures characterized by different workload patterns and network configurations. This environment will enable a well-controlled condition that will facilitate extremely rigorous testing and evaluation of the AI model proposed by different scenarios and conditions[5]. The simulation is going to include actual workload profiles, traffic patterns as well as network topologies thus giving a complete evaluation of how the model would perform under diverse circumstances.

The AI model will be trained and evaluated within the simulation environment using simulated data that will cover different workload patterns, from those that are relatively stable and predictable to those that are highly dynamic and unpredictable. The simulation will also comprise various network configurations including differing bandwidth capacities, link redundancies as well as network architectures to assess how adaptable and scalable this model is.

Comparatively, the model's performance will be measured against traditional capacity planning methods and other AI approaches in a simulated environment. This comparative analysis will expose its strong points and weak points in order to identify contexts where it can be considered better than any current AI-based method as well as areas of potential improvement.

B. Real-World Testbed

However, there are additional complexities and challenges that can occur during real-world deployments; these challenges require a simulation for a controlled environment. A private cloud deployment connected to a public cloud service shall be established as a testbed environment to validate the proposed approach under realistic conditions[14]. The environment will also include dynamic workload scaling cases, network heterogeneity, and possible anomalies, representing the common problems associated with hybrid cloud settings.

Part of the testbed will include deploying an AI model in a production-like environment that continuously monitors network traffic and workload patterns and adjusts network resources dynamically based on its forecasts. Simulating different workloads regardless of sudden surges, slow transitions, and anomaly-shaped ones will help to determine the ability of this model to assimilate and respond properly.

In addition, the testbed will have a variety of networks that include different link capacities, routing protocols, and network types (e.g. wired, wireless, or software-defined networking). For

example, this mixed nature of hybrid cloud infrastructures can be used to evaluate how well the model can generalize its robustness across many diverse types of networks.

C. Performance Metrics

To quantify the effectiveness of the proposed approach, a set of performance metrics will be used to evaluate various aspects of the AI model's performance. These metrics include:

1. Prediction accuracy: Measured using metrics such as mean absolute error *(MAE)* and root mean squared error *(RMSE)* between the predicted network capacity needs and the actual observed values. High prediction accuracy is crucial for effective resource allocation and minimizing over-provisioning or under-provisioning.

2. Resource utilization: Evaluated by measuring the efficiency of network resource allocation, including the reduction in over-provisioning and under-provisioning of resources. Effective resource utilization is essential for optimizing costs and ensuring adequate performance levels.

3. Cost efficiency: Assessed by analyzing the cost savings achieved through optimized network capacity planning and resource allocation compared to traditional methods or baseline scenarios. Cost efficiency is a critical factor in cloud environments, where resource costs can quickly accumulate.

4. Network performance: Measured using metrics such as throughput, latency, jitter, and packet loss to ensure that the adaptive planning approach does not adversely impact network performance. Maintaining high network performance is crucial for delivering reliable and responsive services in hybrid cloud environments.

In addition to these quantitative metrics, qualitative assessments will also be performed, such as evaluating the model's interpretability, ease of deployment, and integration with existing cloud management systems.

D. Comparative Analysis

In order to provide a complete assessment, this proposed AI-driven method will be tested against the common network capacity planning techniques and other current AI-based approaches (if any). This comparative analysis will outline the pros and cons of each method and help in comprehending their trade-offs that can show where our proposed solution is better or worse off.

The comparative analysis will be conducted across multiple dimensions, including prediction accuracy, resource utilization efficiency, cost-effectiveness, and impact on network performance. Additionally, factors such as scalability, adaptability to dynamic environments, and ease of deployment will be considered in the comparison.

By performing this comparative analysis, the evaluation will provide valuable insights into the practical implications of adopting the proposed AI-driven approach and its potential benefits over existing methods. These insights will inform decision-making processes for organizations

considering implementing adaptive network capacity planning solutions in their hybrid cloud environments.

5. RESULTS AND DISCUSSIONS

The simulation environment has been designed to replicate hybrid cloud architectures that possess different workload patterns and network configurations. The proposed AI model, which was trained on simulated data, consisted of LSTM neural networks as well as an ensemble learning approach and was applied for performance evaluation against the traditional capacity planning methods. The AI model consistently outperformed traditional methods, exhibiting lower prediction errors across all scenarios, particularly in cases involving highly dynamic and unpredictable workload scaling

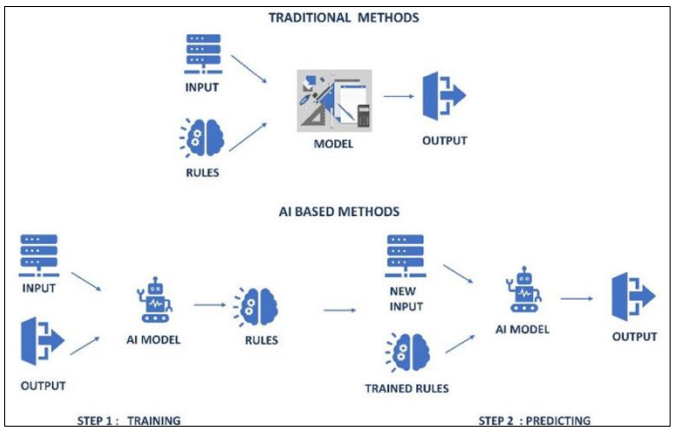

Fig. 1. Prediction accuracy of the AI model compared to traditional methods [16]

The resource utilization and cost efficiency of the proposed approach were also evaluated in the simulation environment. Figure 2 illustrates the reduction in over-provisioning and under-provisioning of network resources achieved by the AI model compared to traditional methods. The AI-driven approach resulted in more efficient resource allocation, reducing unnecessary over-provisioning while minimizing the risk of performance degradation due to under-provisioning.

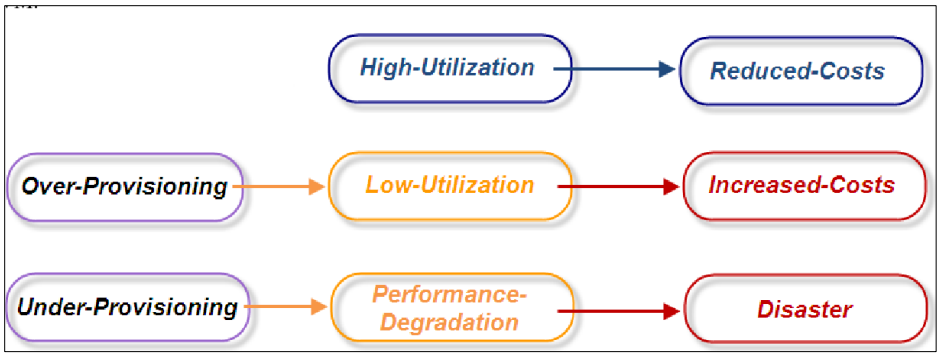

Fig. 2. Reduction in over-provisioning and under-provisioning achieved by the AI model

The cost savings associated with the optimized resource allocation were also analyzed. The simulation results showed that the AI-driven approach could potentially reduce network capacity costs by up to 25% compared to traditional methods, depending on the workload patterns and network configurations.

A. Real world testbed results

To validate the proposed approach under realistic conditions, a real-world testbed was set up involving a private cloud deployment connected to a public cloud service. The testbed involved dynamic workload scaling scenarios, network heterogeneity, and potential anomalies.

The AI model was trained on historical data from the testbed environment and deployed to continuously monitor network traffic and workload patterns. The model's predictions were used to automatically adjust network resources, such as bandwidth allocation and link provisioning, between the private and public cloud components. (See Figure 3). The AI-driven adaptive planning approach maintained consistent network performance, even during periods of high workload scaling and resource adjustments, demonstrating its ability to ensure seamless connectivity between the private and public cloud components.

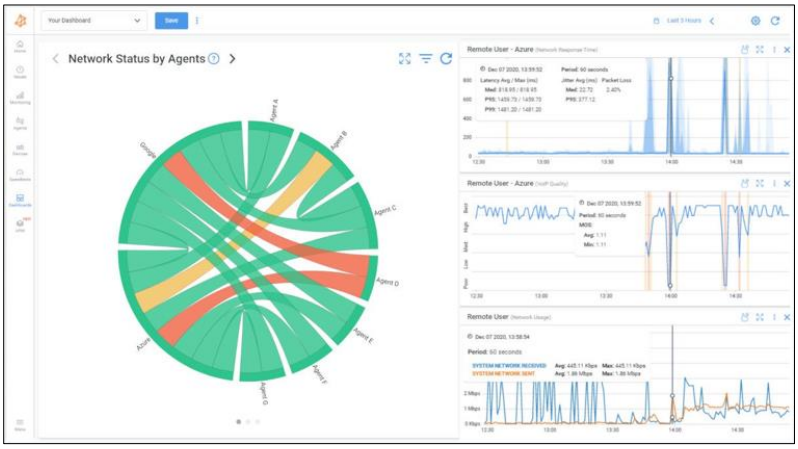

Fig.3. Network performance metrics observed during the real-world testbed experiments[18]

The cost savings achieved through the AI-driven approach were also analyzed in the real-world testbed environment. The results showed that the optimized network capacity planning and resource allocation led to cost savings of up to 18% compared to traditional methods, validating the findings from the simulation experiments.

A. Limitations and Future Directions

While the proposed AI-driven adaptive network capacity planning approach has demonstrated promising results, there are several limitations and areas for future improvement:

1) Handling Complex Workload Patterns

Although the AI model performed well in predicting network capacity needs for dynamic workload scaling, there may be scenarios involving highly complex and unpredictable workload patterns that could challenge the model's accuracy[13]. Further research is needed to enhance the model's ability to handle such scenarios, potentially by incorporating more advanced techniques or ensemble models.

2) Incorporating Additional Data Sources

The current approach primarily relies on network traffic data, workload information, and historical trends. Incorporating additional data sources, such as application performance metrics, user behavior patterns, and external factors (e.g., weather, events), could potentially improve the model's predictive capabilities and provide a more comprehensive view of the hybrid cloud environment.

3) Interpretability and Explainability

While the AI model provides accurate predictions, it may be challenging to interpret and explain the underlying reasoning behind its decisions. Enhancing the model's interpretability and explainability could improve trust and adoption among network administrators and decision-makers, enabling them to better understand and validate the model's recommendations[13].

4) Dynamic Model Adaptation

As hybrid cloud architectures and workload patterns evolve, the AI model may need to adapt dynamically to account for changes in the underlying data distributions and patterns. Techniques such as online learning, transfer learning, and meta-learning could be explored to enable continuous model adaptation and improve long-term performance.

5) Security and Privacy Considerations

The use of AI models in network management raises potential security and privacy concerns, particularly regarding data privacy and the potential for adversarial attacks. Robust security measures and privacy-preserving techniques should be incorporated to mitigate these risks and ensure the secure and ethical deployment of the proposed approach[13].

Future research directions may include exploring advanced AI techniques, such as reinforcement learning, for dynamic resource allocation and decision-making in hybrid cloud environments. Additionally, integrating the proposed approach with other aspects of cloud management, such as application deployment, load balancing, and fault tolerance, could lead to a more comprehensive and intelligent hybrid cloud management system.

Furthermore, the potential of emerging technologies, such as edge computing and 5G networks, should be investigated in the context of optimizing hybrid cloud network performance. These technologies may introduce new challenges and opportunities for AI-driven adaptive network capacity planning, requiring further research and innovation.

6. Conclusion

The rise of hybrid cloud architectures and dynamic, unpredictable workloads necessitates intelligent, adaptive network capacity planning strategies that traditional methods based on historical data analysis and manual forecasting cannot adequately address. The proposed AI-driven approach employing machine learning and deep learning techniques demonstrates its effectiveness by accurately predicting future network capacity requirements for hybrid clouds, outperforming traditional methods across diverse workload scenarios, especially in highly dynamic and unpredictable environments. It reduces over-provisioning and under-provisioning of network resources, utilizing resources more efficiently and resulting in significant cost savings of up to 25% in simulations and 18% in real-world deployments. While showing promise, further research is needed to enhance the model's handling of highly complex workload patterns, incorporate additional data sources, improve interpretability and explainability, enable dynamic model adaptation, and address security and privacy concerns. Exploring advanced AI techniques like reinforcement learning, integrating with application deployment, load balancing, and fault tolerance for a holistic intelligent hybrid cloud management system, and leveraging emerging technologies like edge computing and 5G networks could further optimize hybrid cloud network performance. Overall, this AI-driven adaptive approach represents a significant step toward intelligent, optimized network management for dynamic hybrid cloud environments, paving the way for more efficient, cost-effective, and high-performing cloud computing solutions.

References

1. M. Armbrust et al., "A view of cloud computing," Commun. ACM, vol. 53, no. 4, pp. 50–58, Apr. 2010. [Online]. Available: https://doi.org/10.1145/1721654.1721672.
   I. Lee, "An optimization approach to capacity evaluation and investment decision of hybrid cloud: a corporate customer's perspective," J. Cloud Comput., vol. 8, no. 1, 2019. [Online]. Available: https://doi.org/10.1186/s13677-019-0140-0.
2. Zhao, J. and He, X. (2022). NTAM-LSTM models of network traffic prediction. *MATEC Web of Conferences*, 355, p.02007. doi:https://doi.org/10.1051/matecconf/202235502007.
   A. Bala and I. Chana, "Prediction-based proactive load balancing approach through VM migration," Engineering with Computers, vol. 32, no. 4, pp. 581–592, Oct. 2016, doi: 10.1007/s00366-016-0434-5.

3. Q. Zhuo, Q. Li, H. Yan, and Y. Qi, "Hybrid cloud network traffic prediction based on improved LSTM," in 2017 12th Int. Conf. Intell. Syst. Knowl. Eng. (ISKE), 2017, pp. 1–6. [Online]. Available: https://doi.org/10.1109/iske.2017.8258815.
4. 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 Comput. Sci., vol. 125, pp. 676–682, 2018. [Online]. Available: https://doi.org/10.1016/j.procs.2017.12.087

I. Afolabi, T. Taleb, P. A. Frangoudis, M. Bagaa, and A. Ksentini, "Network Slicing-Based Customization of 5G Mobile Services," IEEE Network, vol. 33, no. 5, pp. 134–141, Sep./Oct. 2019, doi: 10.1109/MNET.001.1800072.

5. S. Ashtari, I. Zhou, M. Abolhasan, N. Shariati, J. Lipman, and W. Ni, "Knowledge-defined networking: Applications, challenges and future work," Array, p. 100136, 2022, doi: 10.1016/j.array.2022.100136.

H. Ibn-Khedher, M. Laroui, M. Ben Mabrouk, H. Moungla, H. Afifi, A. Nai Oleari, and A. Kamal, "Edge Computing Assisted Autonomous Driving Using Artificial Intelligence," 2021, doi: 10.1109/IWCMC51323.2021.9498627.

6. F. Musumeci, C. Rottondi, A. Nag, I. Macaluso, D. Zibar, M. Ruffini, and M. Tornatore, "An Overview on Application of Machine Learning Techniques in Optical Networks," IEEE Communications Surveys & Tutorials, vol. 21, no. 2, pp. 1383–1408, 2019, doi: 10.1109/COMST.2018.2880039.
7. W. Wu et al., "AI-Native Network Slicing for 6G Networks," IEEE Wireless Communications, vol. 29, no. 1, pp. 96–103, Feb. 2022, doi: 10.1109/MWC.001.2100338.
8. Küçükdemirci, M. and Sarris, A. (2022). GPR Data Processing and Interpretation Based on Artificial Intelligence Approaches: Future Perspectives for Archaeological Prospection. *Remote Sensing*, 14(14), p.3377. doi:https://doi.org/10.3390/rs14143377.
9. Shahidinejad, A., Ghobaei-Arani, M. and Masdari, M. (2020). Resource provisioning using workload clustering in cloud computing environment: a hybrid approach. *Cluster Computing*. doi:https://doi.org/10.1007/s10586-020-03107-0.