Volume 9, Issue 10, October– 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24OCT092

Impact of Autoscaling on Application Performance in Cloud Environments

Shankar Dheeraj Konidena

Abstract:- The current paper studies the impact of Autoscaling on application performance in Cloud computing environments. Cloud computing is one of the most encouraging innovations due to its vast applications. Predictive autoscaling is an advanced technique that aims to address the challenges in the autoscaling trends for large-scale systems. To account for the Quality of Service to the customer, features like load balancing according to the workload demand, understanding resource allocation and utilization, and dynamic decision-making are vital to any cloud computing application. This paper interprets these challenges and reviews a meta-reinforcement learning approach for predictive autoscaling in cloud environments. A novel RL-based predictive autoscaling approach on a popular large-scale digital payment platform system, Alipay, is compared with the existing models such as Autopilot and FIRM. The aim is to conduct a detailed analysis of performance metrics before and after autoscaling actions, aiming to identify optimal scaling strategies that minimize response time and maximize resource utilization without over-provisioning.

Keywords:- *Predictive Autoscaling, Application Performance, Cloud Applications, Machine Learning, Reinforcement Learning.*

I. INTRODUCTION

Autoscaling in cloud computing is adjusting the company resources horizontally by adding/removing VMs and modifying CPU/ memory usage in response to changes in workload. The topic of devising auto-scaling mechanisms consists of subtopics that address specific needs and techniques of auto-scaling, including virtualization, comparison, workload monitoring, hybrid/multi-clouds, bandwidth, integrated storage, network and computing, and self-scaling frameworks. [2] Considering the Cloud is a dynamic and uncertain environment, Reinforcement Learning (RL) serves as a good candidate for autoscaling since it is capable of learning transparent (with no human intervention), dynamic (no static plans), and adaptable (constantly updated) resource management policies to execute scaling actions. [1]. To quickly adapt to new workloads by learning common patterns and reducing the need to train separate models, meta-learning aids in generalizing across different applications. Various applications and workloads impact CPU utilization differently, which, in turn, complicates scaling decisions. Meta models address the commonalities between workloads and, at the same time, keep the specifics of the application unchanged, portraying flexibility. Dynamic workloads, CPU utilization, and efficient allocation of resources are prevalent challenges in Autoscaling.

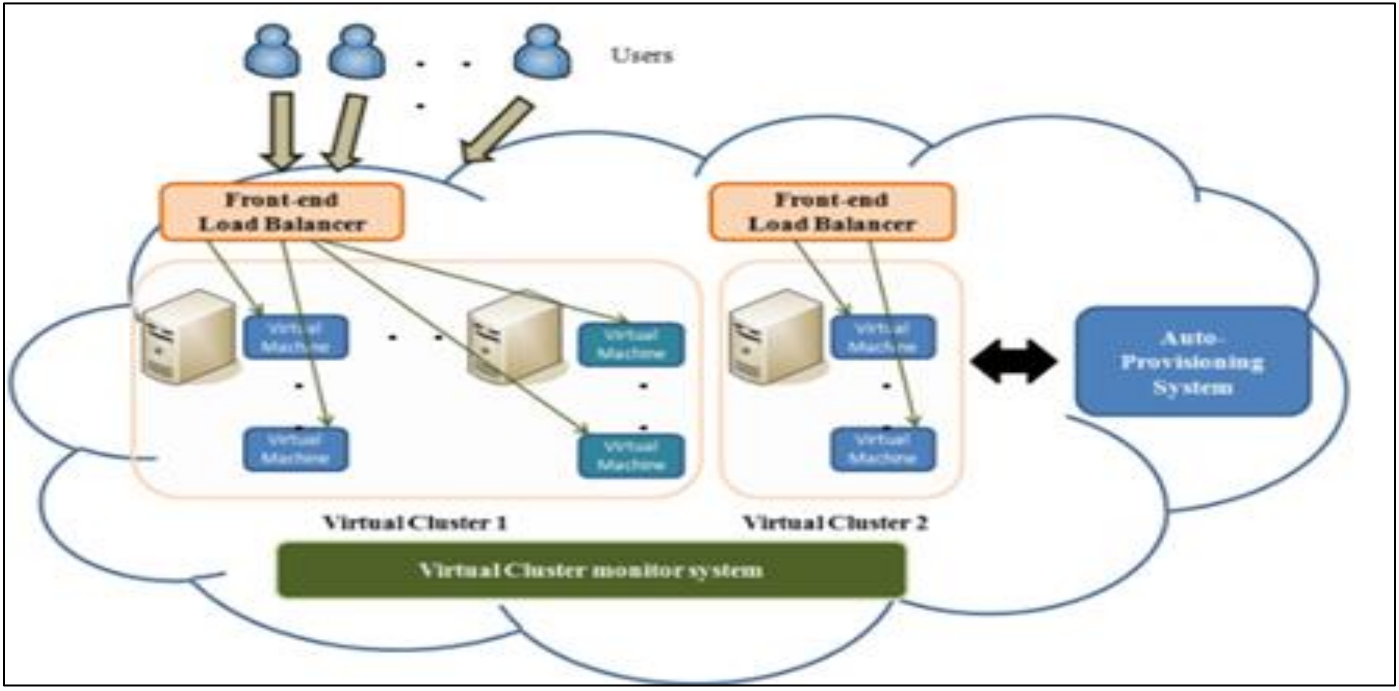

Fig 1 Architecture of Autoscaling

II. BACKGROUND

- *Quality of Service and Service Level Arguments:*

Large enterprise software systems such as eBay, Priceline, Amazon, and Facebook must provide high assurance in Quality of Service (QoS) metrics such as response times, high throughput, and service availability to their users. With such assurances, service providers of these applications stand to retain their user base and, hence, their revenues. Typically, customers maintain Service Level Agreements (SLAs) with service providers for the QoS property. Catering to the SLA while keeping costs low is challenging for such enterprise systems due primarily to the varying number of incoming customers. [2]. Various possibilities and approaches could be designed to solve different problems occurring in cloud architecture, especially in scaling arenas. Further sections delve deeper into the issues and necessitate suitable solution possibilities using Machine Learning and Neural Network techniques.

- *Virtualization and Autoscaling*

Virtualization refers primarily to platform virtualization or the abstraction of physical resources for users. These physical resources are regarded as a pool of resources; thus, they can be allocated on demand. Computing at the scale of the Cloud system allows users to access the enormous and elastic resources on demand. However, user demand for resources can vary at different times, and maintaining sufficient resources to meet peak requirements can be a burden cost-wise. On the contrary, if the user maintains only minimal computing resources, the resource is insufficient to handle the peak requirements. [8] Auto-scaling allows to scale your processed resources dynamically or as expected. Dynamic scaling: Starting and halting computing using various pre-characterized conditions. In predictable scaling, one is sure about traffic patterns, with the end goal being where traffic peaks each morning and goes down at some point of time. Auto-scaling starts more Web servers toward the beginning of the day and shuts down unnecessary ones in the prescribed timings amid the cloud distribution. The load balancer feature in auto-scaling is a productive feature that enables handling the overabundant load. Also, it is one of the viable techniques to save costs and physical resources utilizing scaled-up and scaled platforms dynamically according to the customers' approaching traffic. Also, auto-scaling has dynamic edge strategies that predict required resources based on anticipated values. The auto-scaling method gives on-demand assets according to workload in cloud computing circumstances (Lin et al.,2012; Ashraf et al., 2016). One of the critical characteristics of operating in the Cloud is autoscaling, which elastically scales the resources horizontally (the number of virtual machines (VMs) assigned is changed) or vertically (the CPU and memory reservations are adjusted) to match the changing workload. According to the timing of scaling, the autoscaling strategies can be divided into responsive and predictive strategies [1]

- *Predictive Autoscaling*

Customers use predictive autoscaling to improve response times for applications with long initialization times or workloads that vary predictably with daily or weekly cycles. Without predictive autoscaling, an auto scaler can only scale a group reactively based on observed changes in load in real time. With predictive Autoscaling enabled, the auto scaler works with real-time and historical data to cover the current and forecasted load. **Forecasts are refreshed every few minutes (faster than competing clouds)** and consider daily and weekly seasonality, leading to more accurate forecasts of load patterns. [blog]

- *Reinforcement Learning -Artificial Intelligence*

Cloud computing has revolutionized the IT industry by offering Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS), providing scalable computing, network, and storage capabilities through a pay-per-use model. Machine Learning (ML) techniques emerged as critical components in optimizing resource utilization and ensuring Quality of Service (QoS) in cloud environments. The evolution of cloud computing has given rise to several paradigms tailored to specific application requirements, including edge computing, fog computing, mist computing, Internet of Things (IoT), Software-Defined Networking (SDN), cybertwin, and Industry 4.0. Delivering customer-centric services and collectively enhancing the Quality of Experience (QoE) for the end users are the paradigms for operating with cloud servers. ML techniques serve as a fundamental enabler for these emerging paradigms, addressing many challenges in cloud computing. These include resource scheduling and provisioning, load balancing, Virtual Machine (VM) migration and mapping, task offloading, energy optimization, workload prediction, and device monitoring. The integration of ML with cloud computing has significantly improved the efficiency and effectiveness of these operations.

Despite the rapid advancements in this field, there remains a notable gap in the literature regarding comprehensive surveys that explore:

- The integration of multi-paradigm architectures in cloud computing
- In-depth technical and analytical aspects of these paradigms
- The pivotal role of ML techniques in emerging cloud computing paradigms.

This research aims to address this gap by thoroughly investigating the integration of emerging cloud computing paradigms, focusing on the application of ML as a dominant problem-solving technology.

- *Autopilot:*

A workload-based autoscaling method proposed by Google that builds the optimal resource configuration by seeking the best historical time window to match the current window[1].

- *FIRM:*

An RL-based autoscaling method that solves the problem through learning feedback adjustment in the online cloud environment. Specifically, FIRM finds applications with abnormal response time (RT) through SVM-based anomaly detection algorithms and adjusts multiple resources for the service through RL algorithms[1].

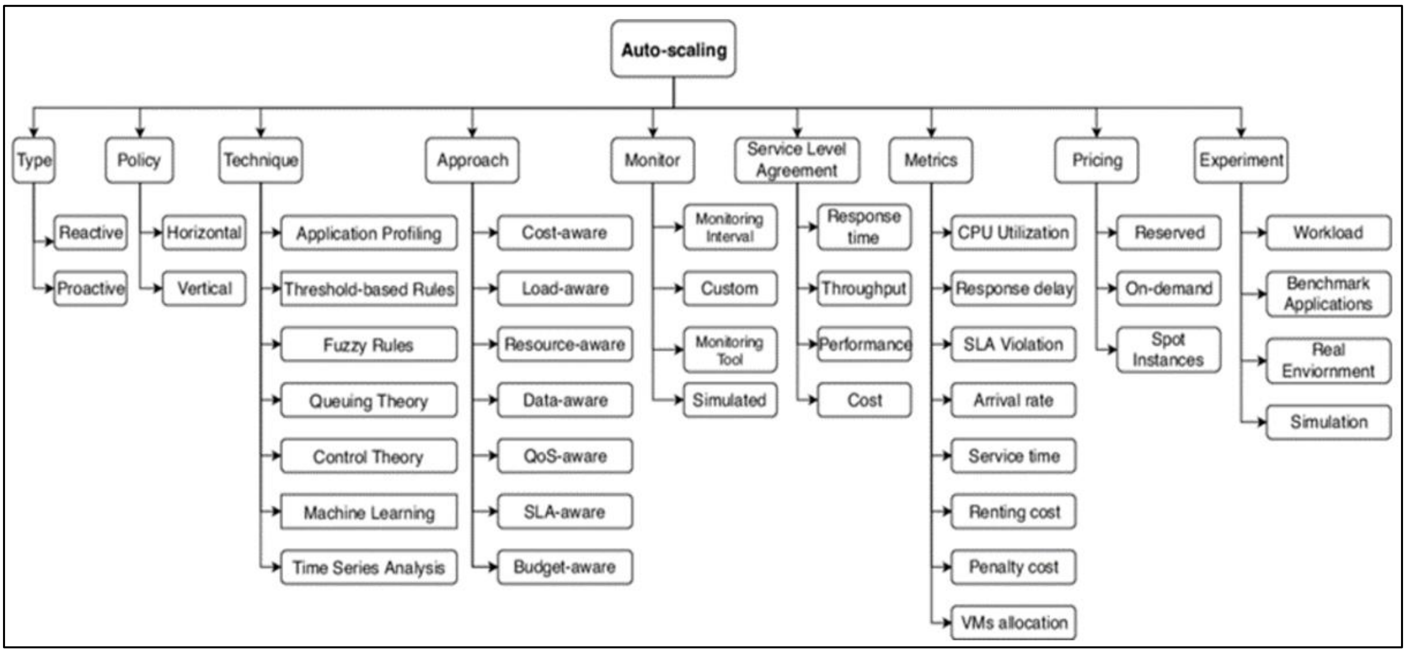

Fig 2 Taxonomy of Cloud Architecture

III. TECHNICAL RESULTS

➢ *Review of Insights:*

The paper presents a novel RL-based predictive autoscaling approach. A deep, attentive periodic model for multi-dimensional, multi-horizon workload prediction provides high-precision and reliable workload information for scaling. The meta-learning model is used to train with the dynamic image to map the workload to CPU utilization, with rapid adaptation to the changing environment, embedded to guide the learning of optimal scaling actions over thousands of online applications. The meta-model-based RL algorithm enables safe and data-efficient learning. Neural process and Markovian Decision process are significant aspects of the approach. There are three Insights involved in the process.

- *Insight 1:*

Deep time series models with notable success are essential. Classical regression techniques or simple neural networks to forecast the workload are ineffective in capturing periodicity or complex temporal dependencies.

- *Insight 2:*

Workload heterogeneity is discussed from two perceptions. (i) distinct applications have distinct mappings among workloads on CPU utilization; (ii) different workload subtypes across an application have varied relationships with CPU utilization.

- *Insight 3:*

A dynamic approach to decision-making is formed when selecting the best resources, given a measure of CPU utilization. The aim is to decide accurate resource allocation (VMs) for the application according to the estimation of CPU utilization.

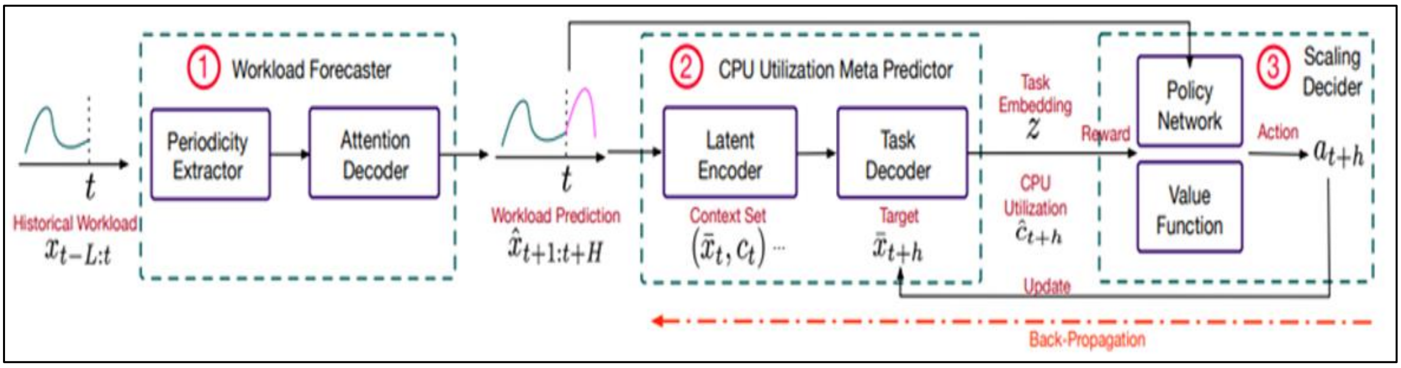

Fig 3 Synergy Diagram of Workload Forecaster, CPU Utilisation Meta Predictor, and Scaling Decider.

IV. RESULTS

- *Auto Scaling via Meta Model-based RL*

The experimental results highlight the superior performance of the proposed Meta Reinforcement Learning approach for predictive autoscaling compared to traditional methods such as Informer, ConvTransformer, Autopilot, and FIRM. In terms of **Workload Prediction,** the results prominently outperformed both Informer and ConvTransformer, achieving an improvement of 25% in accuracy with an MAE (Mean Absolute Error) of 1.10 and RSME (Root Mean Square Error) OF 112.9. This indicates that the predictive model is highly accurate in forecasting future workload spikes and dips, which is critical for efficient autoscaling. Enhancement in the performance attributes to capture complex temporal patterns. **CPU stability** rates are up to 0.95, and a 21% reduction in VM usage compared to Autopilot is seen in the scaling performance, achieving stable resource allocation. Accurate CPU utilization predictions enable better scaling decisions, optimizing resource use, and showing strong decision-making abilities. In terms of **Scalability and RSME**, the model scaled across different numbers of applications (50, 100, 200, 500) while keeping prediction errors low: RMSE values increase slightly as the number of applications grows but remain relatively low (1.11 to 1.33), demonstrating accuracy, though applied to many applications. The approach scales well, handling large-scale cloud environments without any drop in performance and maintaining accuracy as the number of applications increased from 50 to 500. The model improved CPU stability by 20% and reduced cloud resource usage by 50%, exhibiting its effectiveness in real-world applications.

- *Comparison of Autopilot, FIRM, and Proposed RL Model:*

The following is the comparison table to differentiate between Autopilot, FIRM, and Proposed RL Model and the significance of the approach.

Table 1 Comparison of Autopilot, FIRM, and Proposed RL Model

Metric	Autopilot	FIRM	Proposed RL Model	Significance
Workload Prediction Accuracy (MAE)	Higher (around 1.75)	Moderate (around 1.50)	Lower (around 1.10)	More accurate prediction leads to better resource planning.
CPU Utilization Prediction Accuracy (RMSE)	Higher (around 2.48)	Moderate (around 1.93)	Lower (around 1.11)	Better prediction of CPU utilization helps in scaling decisions.
CPU Utilization Stability	Fluctuates significantly	Moderate stability	Highly stable (around 0.95)	More stable CPU utilization improves performance and avoids resource waste.
VM Allocation Efficiency	Less efficient (overshooting or undershooting VMs)	Moderate efficiency	High efficiency (proactive scaling)	Efficient resource use reduces costs.
Failed Requests	Higher (especially during spikes)	Fewer failed requests	Fewest failed requests	Better handling of peak loads ensures system reliability.
Scalability (RMSE for 500 apps)	Higher RMSE (around 1.33)	Moderate (around 1.25)	Lower (around 1.11)	The model scales better across many applications, showing adaptability to large-scale environments.
Cost Efficiency	Higher costs due to over-provisioning	Moderate cost savings	Significant cost savings	More resources are needed, translating to lower operating costs.
Real-World Performance (Alipay)	Deployed in controlled environments	Deployed in test environments	Deployed at scale (Alipay)	Real-world validation shows practical benefits and scalability.

V. CONCLUSION

In conclusion, we can derive from the analysis that carefully chosen autoscaling strategies can significantly improve application performance. The RL-based meta-learning approach is effective in CPU utilization in real-world applications and scaling the digital platform, with a 50% resource savings compared to the rule-based methods. RL techniques could be used in autoscaling to support customers according to their Service Level Agreements, assuring Quality of Service in a cost-friendly manner. Future work should develop adaptive predictive models that automatically select the best scaling strategy based on real-time application behaviors and workload characteristics.

REFERENCES

[1]. Xue, S., Qu, C., Shi, X., Liao , C., Zhu, S., Tan, X., Ma, L., Wang, S., Hu, Y., Lei, L., Zheng, Y., Li, J., & Zhang, J. (2022). A Meta Reinforcement Learning Approach for Predictive Autoscaling in the Cloud. *In Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*. https://doi.org/10.1145/3534678.3539063

[2]. Roy, Nilabja & Dubey, Abhishek & Gokhale, Aniruddha. (2011). Efficient Autoscaling in the Cloud Using Predictive Models for Workload Forecasting. Proceedings - 2011 IEEE 4th International Conference on Cloud Computing, CLOUD 2011. 500-507. 10.1109/CLOUD.2011.42.

[3]. Alipour, Hanieh & Hamou-Lhadj, Abdelwahab & Liu, Yan. (2014). Analysing Auto-scaling Issues in Cloud Environments.

[4]. Shahin, A. A. (2017). Automatic Cloud Resource Scaling Algorithm based on Long Short-Term Memory Recurrent Neural Network. *ArXiv*. https://doi.org/10.14569/IJACSA.2016.071236

[5]. Amir Fazli, Amin Sayedi, Jeffrey D. Shulman (2018) The Effects of Autoscaling in Cloud Computing. Management Science 64(11):5149-5163.

[6]. J. H. Novak, S. K. Kasera and R. Stutsman, "Cloud Functions for Fast and Robust Resource Auto-Scaling," 2019 11th International Conference on Communication Systems & Networks (COMSNETS), Bengaluru, India, 2019, pp. 133-140, doi: 10.1109/COMSNETS.2019.8711058. keywords: {Delays; Cloud computing; Runtime; Current measurement; Load modeling; Servers},

[7]. Arvindhan, M and Anand, Abhineet, Scheming a Proficient Auto Scaling Technique for Minimizing Response Time in Load Balancing on Amazon AWS Cloud (March 15, 2019). International Conference on Advances in Engineering Science Management & Technology (ICAESMT) - 2019, Uttaranchal University, Dehradun, India, Available at SSRN: https://ssrn.com/abstract=3390801 or http://dx.doi.org/10.2139/ssrn.3390801

[8]. Hung, Che-Lun & Hu, Yu-Chen & Li, Kuan-Ching. (2012). Auto-Scaling Model for Cloud Computing System. International Journal of Hybrid Information Technology. 5. 181-186.

[9]. Jacob, S., & G. (2021, October 5). *10 ways Google Cloud IaaS stands out*. Google Cloud. https://cloud.google.com/blog/products/compute/google-clouds-iaas-platform-is-a-powerful-choice