
An AI-Powered Predictive Cloud Resource Allocator Using Prometheus-Based Workload Telemetry

Dr. Chetan H. Patil¹, Prof. Sanket G. Chordiya², Prof. Hansa S. Borse³, Atharva V.

Aher⁴, Akshada B. Dinde⁵, Amit S. Chaudhari⁶, Vaishnavi V. Darekar⁷

Assistant Professor, chetanhpatil@gmail.com, sanket.chordiya@gmail.com, hansashimpi@gmail.com, Research Scholar, atharva0276@gmail.com, dindeakshu1210@gmail.com, amitchaudhary6399210@gmail.com, vaishnavi.v.darekar@gmail.com, Department of Artificial Intelligence and Data Science, PVG's COE & SSDIOM, Nashik, India

Abstract.

With the introduction of cloud computing, there has been an intense revolution in the digital infrastructure space by offering on-demand and scalable computing resources. Conventional fixed or deterministic resource allocation policies, on the other hand, often lead to either over-allocation or under-utilization and thus trigger a decrease in performance or unnecessary expenditures. This paper presents a framework based on AI-powered Cloud Resource Allocator and Manager, also known as the proposed framework, that uses the models of Machine Learning (ML) and Artificial Intelligence (AI) to predict workload fluctuations, identify anomalies, and automatically coordinate the decisions on cloud scaling. The framework uses time-series forecasting models, including Long Short-Term Memory (LSTM) [1] and eXtreme Gradient Boosting (XGBoost) [2] models with the support of anomaly refinement methods to achieve an effective balance between costs and performance [3]. The proposed model can be used to achieve effective adaptive scaling, better anomaly refinement and an improved cost utilization, according to the tests based on simulated AWS EC2 instances and the monitoring dataset [4]. The obtained results show that pro-active cloud resource allocation can effectively increase the reliability and cost-effectiveness compared to the traditional methods. [5], [6].

Keywords. Artificial Intelligence; Cloud Computing; Machine Learning; Predictive Scaling; Resource Optimization; Anomaly Refinement.

1. INTRODUCTION

Cloud environments present a major hurdle as they require resources to be used efficiently on the one hand and performance of the cloud environment to not degrade on the other; and minimize cloud operational cost. The traditional methods of allocating resources are based on manual configuration or statically configured resource scaling (artificially configured rules) mechanisms that provision resources only when pre-defined thresholds are surpassed. These reactive solutions may result in postponing responses, impairment of performance, and needless waste of resources. The limitations are overcome by the advancements of Artificial Intelligence (AI) and Machine Learning (ML) with their “proactive” (data driven)

and “self-adaptive” resource allocation methods. Using past usage trends and real-time performance metrics, ML models can forecast workload patterns and make automatic scaling decisions before hitting bottlenecks. This predictive resource management results in an optimal cost/performance balance. This paper proposes an AI-Powered Cloud Resource Allocator and Manager that integrates predictive modeling, anomaly refinement, and optimization-based decision-making.

2. SYSTEM ANALYSIS

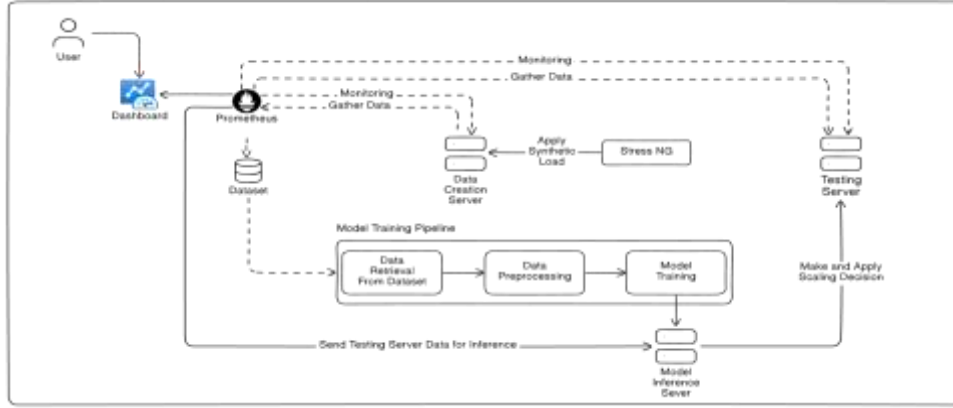


Figure 2.1. System Architecture and Workflow

The suggested AI-Powered Cloud Resource Allocator aims to maximize cloud infrastructure utilization by forecasting future resource demands and automatically adapting system configurations. This section presents both the implemented phase and the projected behavior of the complete system to analyze its feasibility and performance.

2.1. Telemetry Collection Module

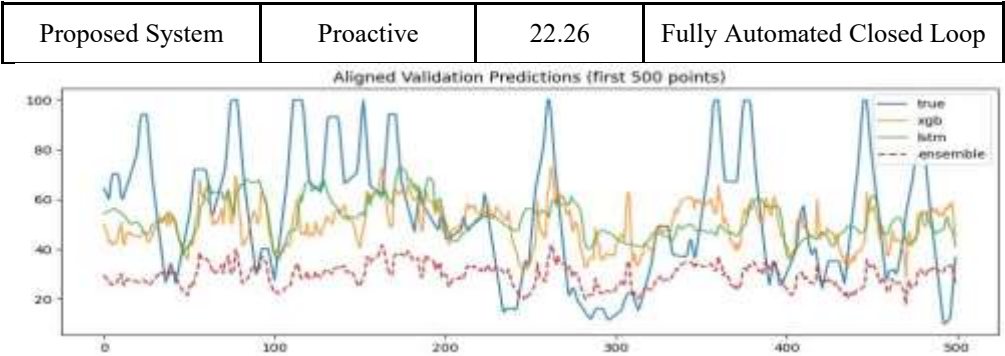
This module is developed in order to collect real-time metrics from the server being monitored. Prometheus is used to monitor, while node exporter exposes the low-level system statistics. Prometheus scrapes the metrics at a short time interval, generating time-series data. This data acts as the primary input to the predictive system[7],[8].

2.2. Predictive Machine Learning Module

The Predictive Machine Learning Module is the intelligent component of the system, responsible for predicting the CPU Utilization at the next 60 seconds mark ($T + 60$) using the currently gathered telemetry data. The XGBoost predictive model is the primary deployed model, having the capabilities of capturing non-linear patterns in the workload. The predicted CPU metric is then forwarded to the decision engine to perform the scaling actions [2],[7].

Table 2.1. Comparative Analysis

Method	Scaling Type	Test RMSE	Remarks
Osypanska and Nawrocki (2023)	ML-Based	41.18	Low Precision



2.3. *Scaling Decision Engine*

The Decision Engine processes the CPU usage predicted by the Machine Learning (XGBoost) model. It then determined whether the server should be scaled or not. It operates based on the predefined parameters such as thresholds for CPU utilization, model confidence and cool down period to prevent instability and frequent scaling. The module acts as the core, converting the insights into instructions [9],[10].

2.4. *Cloud Infrastructure Control Module*

Once the scaling is required the Cloud Infrastructure Control Module communicates with the cloud provider's APIs to adjust the resources that are used by the running instance. This process allows vertical scaling where you can scale up and scale down the computational capacity of the instance as needed, based on the predicted workload demands. Due to automation, there is no need for manual operations, and the infrastructure can be reconfigured dynamically and responsively depending on the expected load on the system[11],[12].

2.5. *Fully Automatic Loop (Autoscaler Daemon)*

The entire system operates through a continuously running autoscaler daemon that coordinates the interactions between all system modules. The daemon periodically performs the following sequence of operations: 1. Query telemetry data from Prometheus. 2. Generate predictive features from the collected data. 3. Perform machine learning inference to forecast CPU utilization. 4. Evaluate scaling policies based on predicted workload conditions. 5. Scale the infrastructure when necessary. 6. Reinitialize monitoring on the server after successful scaling. This allows a closed-loop system to manage a cloud server autonomously, without needing human intervention [5],[13],[12].

3. **CONCLUSION**

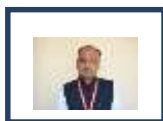
The proposed system shows that Machine Learning can be effectively used to optimize cloud resource utilization. It ensures efficient allocation and dynamic adjustment of computing resources while maintaining performance and minimizing downtime. Using tools such as Prometheus, Node Exporter and Stress-NG can help to gather a representative, complete and realistic set of data. Using the models LSTM and XGBoost, patterns can be detected in the data of the servers for accurate predictions. It is found that the system can achieve high

accuracy and adaptability in various scenarios through experimental results. Integration with Cloud APIs allows direct implementation of model predictions for real-time optimization.

4. REFERENCES

- [1] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [2] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, San Francisco, CA, USA, 2016, pp. 785–794.
- [3] Y. Zhang, X. Chen, and L. Wang, "A Hybrid Anomaly Detection Framework for Cost–Performance Optimization in Time-Series Forecasting," *IEEE Access*, vol. 9, pp. 145321–145333, 2021.
- [4] A. Gupta, P. Verma, and S. Nair, "Machine Learning-Based Cloud Resource Optimization Using Predictive Scaling," *IEEE Transactions on Cloud Computing*, 2023.
- [5] R. Buyya, M. Yousif, and S. Pandey, "Intelligent Cloud Resource Management: Foundations and Future Directions," *Journal of Cloud Computing*, 2022.
- [6] L. Xu, J. Rao, and X. Bu, "A Reinforcement Learning Approach to Cloud Resource Management for Elasticity and Cost Efficiency," *IEEE Transactions on Network and Service Management*, vol. 19, no. 1, pp. 11–24, 2022.
- [7] S. Kumar, A. Singh, and R. Buyya, "Workload Forecasting in Cloud Data Centers Using Machine Learning: A Comparative Study," *IEEE Transactions on Cloud Computing*, vol. 10, no. 5, pp. 3201–3215, 2022.
- [8] J. Wu, Q. Deng, and H. Chen, "Workload Prediction in Cloud Computing Using Hybrid Deep Learning Model," *IEEE Transactions on Parallel and Distributed Systems*, vol. 31, no. 4, pp. 957–970, 2020.
- [9] S. Wang, R. Buyya, and J. Chen, "Energy-Efficient Load Balancing and Auto-Scaling in Cloud Data Centers Using Reinforcement Learning," *IEEE Transactions on Cloud Computing*, vol. 9, no. 3, pp. 1080–1092, 2021.
- [10] A. Beloglazov and R. Buyya, "Optimal Online Deterministic Algorithms and Adaptive Heuristics for Energy and Performance Efficient Dynamic Consolidation of Virtual Machines in Cloud Data Centers," *Concurrency and Computation: Practice and Experience*, vol. 24, no. 13, pp. 1397–1420, 2012.
- [11] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," *NIST Special Publication 800-145*, National Institute of Standards and Technology, Gaithersburg, MD, USA, 2011.
- [12] M. A. Khan, S. U. Rehman, and M. U. Ashraf, "Energy and Cost-efficient Resource Management in Cloud Data Centers Using Predictive Analytics," *Future Generation Computer Systems*, vol. 108, pp. 803–811, 2020.
- [13] J. Gao, H. Shen, and C. Li, "AI-Driven Cloud Resource Management: A Survey of Techniques and Trends," *IEEE Transactions on Cloud Computing*, vol. 11, no. 3, pp. 1458–1475, 2023.

Biographies



Dr. Chetan H. Patil has a PhD in Computer Science and Engineering and is working as an Assistant Professor in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik. His academic and professional interests include Data Science, Machine Learning & Cyber Security.



Prof. Sanket G. Chordiya is Pursuing a PhD in Computer Science and Engineering and is working as a Head of Department in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik. His academic and professional interests include Database Management Systems, Artificial Intelligence, Data Science, Machine Learning & Cyber Security.



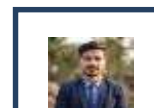
Prof. Hansa S. Borse has completed M.E in Computer Engineering and is working as an Assistant Professor in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik. His academic and professional interests include Data Structures, Artificial Intelligence, Data Science, Machine Learning.



Atharva Aher is a Student in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik



Akshada Dinde is a Student in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik



Amit Chaudhari is a Student in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik



Vaishnavi Darekar is a Student in Department of Artificial Intelligence & Data Science at PVGCOE & SSDIOM, Nashik