Cloud Resources scaling and failure detection using ML

Guday Ram Gopal Varma[1], Mrs.M.Revati[2], Mrs.K.Nirusha[3], Mr.M.Venkateswara Rao[4]

[1]PG student ,Department Of Computer Science & Engineering, Mother Teresa Institute Of Science And Technology Autonomous, Sanketika Nagar, Kothuru (V), Sathupally - 507303, Khammam Dist., Telangana,India

[2,3,4]Assistant Professor , Department Of Computer Science & Engineering, Mother Teresa Institute Of Science And Technology Autonomous, Sanketika Nagar, Kothuru (V), Sathupally - 507303, Khammam Dist., Telangana,India

***Abstract*—**Cloud computing has become one of the most important technologies in modern digital infrastructure because it provides scalable, flexible, and cost-effective computing services to organizations across the world. Businesses, educational institutions, healthcare organizations, banking sectors, and industrial environments increasingly depend on cloud platforms to store data, execute applications, manage services, and support large-scale online operations. As cloud environments continue to grow in complexity, managing resources efficiently and detecting failures quickly have become major challenges. Cloud resource scaling and failure detection using Machine Learning (ML) is an advanced approach that aims to optimize cloud performance, reduce operational costs, improve reliability, and ensure uninterrupted service delivery. This project focuses on designing an intelligent framework that combines cloud resource management with machine learning techniques, particularly the Random Forest algorithm, to achieve efficient resource allocation and accurate failure prediction in cloud systems.

I. INTRODUCTION

Cloud computing has revolutionized the field of information technology by enabling organizations to access computing resources over the internet without investing heavily in physical infrastructure. The concept of cloud computing allows users to utilize services such as storage, processing power, networking, databases, and software applications on demand. Major cloud service providers offer Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) to meet diverse business requirements. With the rapid growth of internet users, big data applications, artificial intelligence systems, streaming services, and online transactions, cloud infrastructures are required to handle highly dynamic workloads efficiently. This increasing demand introduces significant challenges related to resource scaling and system reliability.

Resource scaling refers to the process of allocating or deallocating cloud resources such as virtual machines, containers, CPU cores, memory, and storage based on workload requirements. In cloud environments, workloads can change rapidly due to varying user demands, seasonal traffic, application behavior,

Doi:10.48047/ijiee.2026.16.2.50

and unexpected events. If sufficient resources are not available during high demand periods, application performance degrades and service outages may occur. On the other hand, allocating excessive resources during low demand periods leads to unnecessary operational expenses. Therefore, efficient resource scaling is essential to maintain performance while minimizing costs.

II. LITERATURE SURVEY

1. ML-Based Autoscaling for Elastic Cloud Applications: Taxonomy, Frameworks, and Evaluation

Authors: Vishwanath Srikanth Machiraju, Vijay Kumar, Sahil Sharma

- **Technique Used:** Machine Learning-based autoscaling using supervised learning, unsupervised learning, and reinforcement learning.
- **Pros:**
  - Improves resource utilization.
  - Supports proactive scaling decisions.
  - Reduces SLA violations.
- **Cons:**
  - Complex deployment in hybrid cloud environments.
  - Telemetry delays affect prediction accuracy.

2. A Comprehensive Review of Machine Learning Approaches for Predictive Resource Management in Cloud Computing Environment

Authors: Raju Yadav, Jeetendra Singh Yadav

- **Technique Used:** Predictive resource allocation using Machine Learning and workload forecasting.
- **Pros:**
  - Enhances resource efficiency.
  - Reduces operational costs.
  - Improves cloud performance.
- **Cons:**
  - Requires large historical datasets.
  - Model accuracy depends on workload patterns.

3. Reinforcement Learning-Based Application Autoscaling in the Cloud: A Survey

Authors: Yisel Garí, David A. Monge, Elina Pacini, Cristian Mateos, Carlos García Garino

- **Technique Used:** Reinforcement Learning (RL) for automatic cloud resource scaling.
- **Pros:**
  - Learns dynamic scaling policies automatically.
  - Adapts to changing workloads.
  - Improves cloud elasticity.
- **Cons:**
  - Long training time.
  - High computational complexity.

4. Observing the Clouds: A Survey and Taxonomy of Cloud Monitoring

Authors: Jonathan Stuart Ward, Adam Barker

- **Technique Used:** Cloud monitoring and anomaly detection frameworks.
- **Pros:**
  - Supports real-time monitoring.
  - Helps identify system failures early.
  - Improves cloud reliability.

Doi:10.48047/ijiee.2026.16.2.50

- **Cons:**
    - Monitoring overhead increases with scale.
    - Complex data collection mechanisms.

5. Autoscaling Techniques in Cloud-Native Computing: A Comprehensive Survey

Authors: Various Researchers

- **Technique Used:** AI-driven autoscaling, predictive scaling, and cloud-native resource management.
- **Pros:**
    - Enhances scalability and service continuity.
    - Optimizes cloud costs.
    - Supports Kubernetes and container orchestration.
- **Cons:**
    - Vulnerable to autoscaling attacks.
    - Scaling delays can reduce effectiveness.

III. EXISTING SYSTEM

Existing cloud resource management systems mainly use traditional machine learning methods or static rule-based mechanisms for scaling and failure detection. In many conventional cloud environments, threshold-based auto-scaling policies are implemented where predefined CPU or memory utilization levels trigger scaling actions. Similarly, failure detection mechanisms often depend on manual monitoring, log analysis, or simple anomaly detection methods.

Some existing systems use basic machine learning algorithms such as Linear Regression, Decision Trees, Naive Bayes, Support Vector Machines, and K-Nearest Neighbor algorithms for workload prediction and fault classification. These approaches provide moderate prediction accuracy but face limitations when handling complex and highly dynamic cloud environments.

Disadvantages

Existing cloud resource scaling and failure detection systems suffer from several limitations that reduce their effectiveness in modern distributed computing environments. One major disadvantage is the reliance on static threshold-based scaling policies. These systems allocate resources based on fixed CPU or memory utilization limits without considering future workload trends. As a result, they often react too late during sudden traffic spikes, leading to performance degradation and service interruptions.

Another major limitation is inefficient resource utilization. Traditional systems frequently allocate excessive resources to prevent overload conditions. This over-provisioning increases infrastructure costs and energy consumption. In contrast, under-provisioning occurs when insufficient resources are allocated during high demand periods, causing application slowdown and poor user experience.

IV. PROPOSED SYSTEM

The proposed system introduces an intelligent cloud resource scaling and failure detection framework using the Random Forest machine learning algorithm. The main objective of the proposed system is to improve cloud infrastructure efficiency, optimize resource allocation, and enhance failure prediction accuracy through intelligent automation.

Doi:10.48047/ijiee.2026.16.2.50

Random Forest is an ensemble learning algorithm that combines multiple decision trees to perform classification and prediction tasks. During training, the algorithm creates several decision trees using random subsets of training data and features. Each tree generates an independent prediction, and the final output is determined through majority voting or averaging methods. This ensemble approach improves accuracy, reduces overfitting, and enhances generalization performance.

Advantages

The proposed cloud resource scaling and failure detection system using the Random Forest algorithm offers numerous advantages over traditional cloud management approaches. One of the most important advantages is improved prediction accuracy. The Random Forest algorithm combines multiple decision trees, reducing errors and generating highly accurate predictions for workload forecasting and failure detection.

Another significant advantage is proactive resource scaling. Unlike traditional reactive systems, the proposed framework predicts future workload demands before resource shortages occur. This predictive scaling mechanism ensures that cloud applications maintain stable performance even during sudden traffic spikes or workload fluctuations.

V. SYSTEM ARCHITECTURE

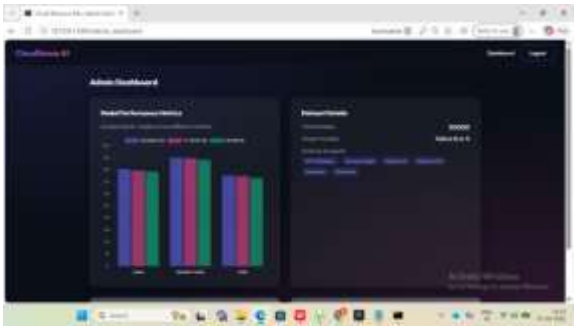
System The system architecture for Cloud Resource Scaling and Failure Detection Using Machine Learning is designed to provide intelligent management of cloud infrastructure through automated monitoring, prediction, scaling, and fault detection mechanisms. In modern cloud computing environments, applications and services experience varying workloads at different times. Traditional manual resource management methods are inefficient because they cannot dynamically adjust to sudden increases or decreases in workload demand. Similarly, cloud system failures such as server crashes, network issues, virtual machine failures, and resource exhaustion can negatively impact application availability and user experience. Therefore, an intelligent machine learning-based architecture becomes essential for maintaining scalability, reliability, and high performance in cloud platforms

VI. RESULTS AND OUTCOMES

Doi:10.48047/ijiee.2026.16.2.50

Screen 1: Home Page

Screen 2: User Page



Screen 3: Admin Page

VII. CONCLUSION

Cloud resource scaling and failure detection using Machine Learning has emerged as an important solution for improving the performance, reliability, and efficiency of modern cloud computing environments. Traditional cloud management techniques mainly depend on static threshold values and manual monitoring methods, which are often unable to handle the rapidly changing workloads and complex infrastructure of large-scale cloud platforms. By integrating machine learning techniques into cloud environments, organizations can automatically predict workload variations, detect failures at an early stage, and allocate resources dynamically according to system demand. This intelligent automation significantly improves overall cloud service quality and operational stability.

The proposed machine learning-based cloud resource scaling system provides an adaptive mechanism for managing computing resources such as CPU, memory, bandwidth, and storage. The system continuously monitors cloud performance metrics and analyzes historical workload patterns to predict future resource requirements. Machine learning algorithms such as Random Forest, Decision Trees, Support Vector Machines, and Neural Networks help in identifying traffic fluctuations and scaling resources automatically. As a result, the system minimizes resource wastage while ensuring high application availability and reduced response time. This predictive scaling mechanism is highly beneficial for organizations that experience dynamic workloads in real-time applications.

VIII. BIBLIOGRAPHY

[1] J. Dean and L. A. Barroso, “The Tail at Scale,” Communications of the ACM, vol. 56, no. 2, pp. 74–80, Feb. 2013.

[2] M. Mao and M. Humphrey, “A Performance Study on the VM Startup Time in the Cloud,” in Proc. IEEE 5th Int. Conf. Cloud Computing, Honolulu, HI, USA, 2012, pp. 423–430.

[3] T. Lorido-Botran, J. Miguel-Alonso, and J. A. Lozano, “A Review of Auto-Scaling Techniques for Elastic Applications in Cloud Environments,” Journal of Grid Computing, vol. 12, no. 4, pp. 559–592, Dec. 2014.

[4] A. Ali-Eldin, J. Tordsson, and E. Elmroth, “An Adaptive Hybrid Elasticity Controller for Cloud Infrastructures,” in Proc. IEEE Network

Doi:10.48047/ijiee.2026.16.2.50

Operations and Management Symposium, Maui, HI, USA, 2012, pp. 204–212.

[5] H. Xu and B. Li, "Dynamic Cloud Pricing for Revenue Maximization," IEEE Transactions on Cloud Computing, vol. 1, no. 2, pp. 158–171, Jul.–Dec. 2013.

[6] C. Delimitrou and C. Kozyrakis, "Quasar: Resource-Efficient and QoS-Aware Cluster Management," in Proc. 19th Int. Conf. Architectural Support for Programming Languages and Operating Systems, Salt Lake City, UT, USA, 2014, pp. 127–144.

[7] Y. Chen, A. Ganapathi, R. Griffith, and R. Katz, "The Case for Evaluating MapReduce Performance Using Workload Suites," in Proc. IEEE 19th Int. Symp. Modeling, Analysis, and Simulation of Computer and Telecommunication Systems, Singapore, 2011, pp. 390–399.

[8] X. Zhu and H. Jiang, "Autonomic Resource Provisioning for Cloud-Based Software," IEEE Transactions on Cloud Computing, vol. 1, no. 1, pp. 109–122, Jan.–Jun. 2013.

[9] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud Computing: State-of-the-Art and Research Challenges," Journal of Internet Services and Applications, vol. 1, no. 1, pp. 7–18, May 2010.

[10] P. Lama and X. Zhou, "AROMA: Automated Resource Allocation and Configuration of MapReduce Environment in the Cloud," in Proc. 9th Int. Conf. Autonomic Computing, San Jose, CA, USA, 2012, pp. 63–72.

Doi:10.48047/ijiee.2026.16.2.50