A Smarter Approach to Cloud Computing: Predicting Workloads with AI

Mr. T Bharath[1], Mr. A. N. Dinesh Kumar[2]
[1]PG Scholar, Vemu Institute of Technology
[2]Professor, Department of CSE, Vemu institute of Technology

Abstract: Cloud computing has revolutionized the IT landscape, providing cost-effective and scalable resources. However, accurate performance prediction for virtual machines (VMs) remains a critical challenge due to their black-box nature and variable workloads. The "Cloud Prophet" framework leverages machine learning techniques, including Dynamic Time Warping (DTW) and neural networks, for precise VM performance prediction. As extensions, Gated Recurrent Unit (GRU) is employed for enhanced accuracy, and live dataset integration demonstrates real-time applicability. The proposed system effectively predicts VM performance degradation and optimizes resource allocation, outperforming existing methods. By addressing key limitations of traditional algorithms, the extended approach provides a robust, scalable solution for managing dynamic cloud environments while achieving high prediction accuracy and operational efficiency.

Keywords: Cloud computing, IT landscape, Virtual machines (VMs)

INTRODUCTION

Cloud platforms have gained immense popularity due to their scalability and cost efficiency. Virtual Machines (VMs) play a pivotal role in this infrastructure, hosting diverse applications. However, ensuring consistent performance across VMs is a significant challenge, especially in dynamic workload scenarios. Traditional performance prediction methods often fail due to their inability to handle variable workloads or accurately predict performance in black-box VM environments. The "Cloud Prophet" framework introduces a machine learning-based approach to address these issues. Using DTW for application identification and neural networks for performance prediction, the system achieves high accuracy. Additionally, GRU is introduced as an extension to improve prediction efficiency. Real-time datasets further validate the framework's adaptability and reliability. By addressing limitations of existing methods, the proposed system enhances resource allocation, reduces performance degradation, and ensures robust VM management, making it a transformative solution for cloud environments.

LITERATURE SURVEY

1. Paper 1: Bayesian Regression for Task Runtime Prediction
   - Focuses on estimating task runtimes in scientific workflows deployed on heterogeneous cloud servers.
   - Uses a Bayesian regression model trained on benchmarked server data to predict execution times.
   - Incorporates workload input and runtime hardware metrics for more precise estimations.
   - Strength: Accurate for workload-based estimations.
   - Weakness: Requires extensive benchmark profiling before use
2. Paper 2: Deep Learning-Based Performance Prediction for Virtual Machines
   - Proposes a neural network-based approach for predicting VM performance using accessible hardware metrics.
   - Compares with deep learning models such as CNN, LSTM, and DNN.
   - Finds that deep learning models often overfit and have higher computational overhead.
   - Strength: High accuracy in predicting variable workloads.
   - Weakness: High computational cost for deep learning-based models

3. Paper 3: Decision Tree-Based Workload Prediction
    - Implements a decision tree approach for predicting server workloads.
    - Uses server data to optimize resource allocation dynamically.
    - Bayesian optimization is applied to enhance prediction accuracy.
    - Strength: Low computational cost.
    - Weakness: Limited accuracy due to simplistic decision boundaries
4. Paper 4: Random Forest for Task Execution Time Prediction
    - Uses a random forest approach to predict task execution time in cloud environments.
    - Employs multiple decision trees and voting mechanisms for improved robustness.
    - Does not account for application interference on shared resources.
    - Strength: Handles complex workloads better than single decision tree methods.
    - Weakness: Ignores interference effects, leading to biased estimates

PROBLEM STATEMENT

Traditional VM performance prediction methods struggle with black-box environments, variable workloads, and inaccurate predictions. A robust machine learning-based framework is needed to address these limitations, ensuring efficient resource allocation and reduced performance degradation in dynamic cloud environments.

METHODOLOGY

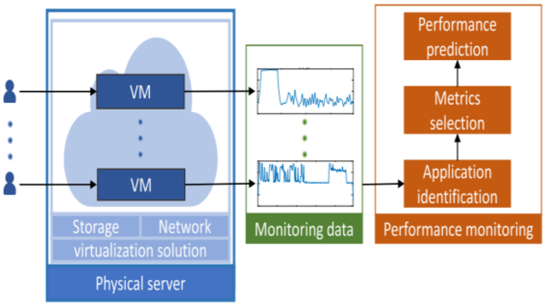
1. Data Collection and Preprocessing

Objective: To gather and preprocess relevant data for training and testing the prediction models.
Process:
Data Sources: Real-time datasets are collected from cloud platforms to simulate dynamic workloads.
Feature Selection: Identify performance metrics, such as CPU usage, memory utilization, and network bandwidth, that influence VM performance.
Normalization: Scale the data to a uniform range to improve model accuracy.

PROPOSED METHOD

The proposed system introduces a machine learning-based framework leveraging DTW for application identification and GRU for enhanced prediction accuracy. Unlike traditional methods, the system adapts to real-time datasets, ensuring scalability and reliability in dynamic workloads. Neural networks predict performance levels accurately by analyzing highly correlated metrics, while performance degradation is minimized through proactive resource allocation. Extensions, such as live dataset integration, validate the framework's applicability to real-world scenarios. The system effectively addresses the limitations of existing approaches, delivering a robust and adaptive solution for cloud environments with improved resource utilization and minimized performance degradation.

ARCHITECTURE



Outcome: A high-quality dataset ready for model training and testing.

2. Dynamic Time Warping (DTW) for Application Identification

Objective: To classify and identify applications running on VMs based on workload patterns.

Process:

DTW analyzes the time-series data of workload patterns to identify similarities.

Outcome: Accurate grouping of applications based on workload behavior, streamlining the prediction process.

3. Neural Network-Based Performance Prediction

Objective: To predict VM performance using neural networks trained on workload metrics.

Process:

Training: Train the model using the preprocessed dataset, optimizing weights to minimize prediction errors.

Outcome: A neural network capable of providing accurate performance predictions for VMs in dynamic environments.

4. Extension with Gated Recurrent Unit (GRU)

Objective: To enhance prediction accuracy by incorporating GRU, a type of recurrent neural network optimized for sequential data.

Process:

GRU processes time-series data, capturing long-term dependencies between workload metrics and performance

The GRU model is trained and validated using the same dataset as the neural network.

Outcome: Improved prediction accuracy and efficiency, especially for time-dependent workloads.

5. System Evaluation and Comparison

Objective: To evaluate the framework's performance and compare it with existing methods.

Process:

Metrics such as prediction accuracy, execution time, and resource allocation efficiency are used for evaluation.

Compare results with baseline methods like static models and approaches.

Outcome: The framework's superiority over traditional methods.

CONCLUSION

In this paper, we have proposed a machine learning-based performance prediction method for cloud applications based on the realistic assumption that resource governors should regard VMs as black-box systems. The proposed method first identifies the application based on accessible hardware metrics from the host server. Next, highly-correlated metrics are selected to accurately predict the performance level of the application by using the proposed machine learning-based approach. Finally, the performance degradation is predicted for the VM, which can be further used by the resource governor to schedule or migrate the VM.

REFERENCE

[1] E. Cortez, A. Bonde, A. Muzio, M. Russinovich, M. Fontoura, and R. Bianchini, "Resource central: Understanding and predicting workloads for improved resource management in large cloud platforms," in Proc. 26th Symp. Operating Syst. Princ., 2017, pp. 153–167.

[2] S. S. Gill et al., "AI for next generation computing: Emerging trends and future directions," Internet Things, vol. 19, 2022, Art. no. 100514.

[3] "Gartner forecasts worldwide public cloud end-user spending to grow 23%, 2021," Gartner, Inc. Accessed: 2023. [Online]. Available: https://www.gartner.com/en/newsroom/pressreleases/2021-04-21- gartner-forecasts-worldwide-public-cloud-end user-spending-to-grow23-percent

[4] N. Jones et al., "The information factories," Nature, vol. 561, no. 7722, pp. 163–166, 2018.

[5] A. S. Andrae and T. Edler, "On global electricity usage of communication technology: Trends to 2030," Challenges, vol. 6, no. 1, pp. 117–157, 2015.

[6] J. Gao, "Machine learning applications for data center optimization," Google Res., 2014. [Online]. Available: https://research.google/pubs/ machine-learning-applications-for-data-center-optimization/

[7] G. Neiger, A. Santoni, F. Leung, D. Rodgers, and R. Uhlig, "Intel virtualization technology: Hardware support for efficient processor virtualization," Int. Technol. J., vol. 10, no. 3, pp. 167–177, 2006.

[8] "AMD Virtualization technology," AMD, Inc. Accessed: 2023. [Online]. Available: https://www.amd.com/en/solutions/hci-and-virtualization

[9] S.-G. Kim, H. Eom, and H. Y. Yeom, "Virtual machine consolidation based on interference modeling," J. Supercomputing, vol. 66, no. 3, pp. 1489–1506, 2013.

[10] T. Palit, Y. Shen, and M. Ferdman, “Demystifying cloud benchmarking,” in Proc. IEEE Int. Symp. Perform. Anal. Syst. Softw., 2016, pp. 122–132.