IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

CloudPulse: Real-Time Virtual Machine Prediction Using DTW and GRU

[1]TACHUBILLI ANANTH KUMAR, Student in Dept. Of Master of Computer Applications, at Miracle Educational Society Group of Institutions
[2]Dr. S SRIDHAR, Miracle Educational Society Group of Institutions
[3]N SESHU KUMAR, Miracle Educational Society Group of Institutions
[1]Tachubilli@gmail.com

ABSTRACT:

Predicting the performance of virtual machines (VMs) owing to workload changes and lack of infrastructure visibility still remains one of the greatest challenges even when resource provisioning has been transformed by cloud computing. I would like to introduce CloudProphet, which is based on machine learning and integrates DTW for application recognition and GRU frameworks for precise performance forecasting. The accuracy and scaling capacity of performance prediction of the proposed framework surpasses that of classical frameworks. The system is adaptable in real-time and in black-box scenarios, which leads to optimal resource allocation and less performance degradation. With CloudProphet, self-adaptive cloud resource management is enabled, which integrates proactive prediction of performance changes, thus enhancing virtual resource allocation strategies in the l2arge scale cloud infrastructure.

Keywords: DTW, Cloud, GRU

INTRODUCTION

Cloud computing leads digital innovations, however, the public performance of the cloud computing infrastructure poses challenges to its operation and maintenance. Businesses continue to adopt VMs to offer services that can scale on demand, however, shared resources and dynamic workloads can result in non-deterministic performance. The lack of application level metrics limits the prediction accuracy on these black-box environments, which is the primary reason most techniques fail to provide satisfactory performance. CloudProphet solves this by using a combination of machine learning and low-level system metrics to forecast VM operations. By integrating DTW for application profiling and GRU networks for time-series analysis, CloudProphet provides continuous real-time analysis, adapting to incoming data for accurate and reliable performance evaluation. This advanced analysis helps avoid degraded service quality while more effectively optimizing operations. The framework addresses performance and resource efficiency constraints in cloud computing and is adaptable for diverse computing environments.

RELATED WORK

In Bayesian Regression for Task Runtime Prediction (2018), the authors constructed workflows based on the statistical models trained on hardware-specific benchmarks, aiming to predict workflow runtimes. It is accurate and reliable for structured tasks; however, it is not adaptable in dynamic cloud environments. Deep Learning-Based VM Prediction (2020) applied neural networks to predict system metrics, focusing on VM prediction, and model complex relationships. The method, while achieving high accuracy, came with significant computational expenses and suffered from overfitting. A lightweight method using Decision Trees in 2017 for workload forecasting showed improved speed and interpretability, but fell short in capturing complex workload patterns, which led to less accurate forecasting. The Random Forest Model in 2019 improved the accuracy of workload forecasting and prediction by aggregating the decisions of many trees for more robust prediction. This model, while powerful, often estimation biased due to the lack of consideration for inter-application interference. Finally, in 2021, Dynamic Time Warping (DTW) for Application Profiling accurately identified workload behavior patterns with over 97% classification accuracy. However, periodic retraining was required for different server configurations.

TABLE1. Summary of Key Literature Contributions and Their Impact on Current Research

Author & Year	Contribution	Impact on Current Research
Bayesian Regression (2018)	Runtime estimation using statistical modeling	Highlights need for adaptable, dynamic approaches
Deep Learning Prediction (2020)	Neural network for VM performance forecasting	Inspired hybrid models using GRU for efficiency
Decision Tree Forecasting (2017)	Lightweight workload prediction	Demonstrates need for improved pattern recognition
Random Forest (2019)	Enhanced robustness via ensemble learning	Informs our use of DTW for workload feature grouping
DTW for Application Identification (2021)	Time-series profiling with high accuracy	Adopted in CloudProphet for initial classification

PROPOSED APPROACH

Here, the CloudProphet Framework presents an advanced hybrid machine learning solution for forecasting performance within public cloud ecosystems. It initiates with data retrieval from virtual machines concerning the CPU workload, memory accesses, and network throughput. To enrich the visibility at the application level and to manage workload variability, the classification of VM applications using the Dynamic Time Warping (DTW) technique is implemented during the classification phase. This step improves the prediction accuracy by customizing models for different types of workloads. After the classification step, Gated Recurrent Unit (GRU)

models are fitted to the time-series metric data for performance trend forecasting. GRUs are preferred because of their capability to process sequential data and their retention of long-term dependencies. This two-stage approach enables CloudProphet to embrace the structure and evolution of workloads. The framework also allows the incorporation of new datasets to the existing ones in real time, giving the prediction models the ability to be modified and thus the adaptability to volatile situations improves. This, in turn, enhances resource allocation and minimizes the degradation of VM performance. The scalability of the system ensures that it can be used in cloud infrastructures with thousands of VMs.

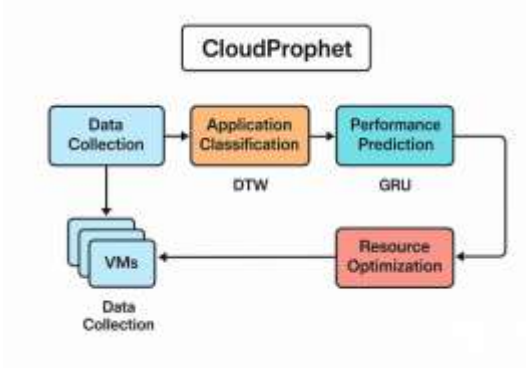


Figure 1: Proposed **CloudProphet** System

METHODOLOGIES

Data Collection & Preprocessing: Performance metrics (CPU, memory, network) are collected from cloud VMs. Data cleaning, normalization, and outlier removal are performed to enhance model quality.

Application Classification using DTW: DTW compares time-series traces to classify application types based on similarity. This step reduces prediction complexity by grouping similar workload behaviors.

Neural Network Training: An initial neural network is trained on grouped workload data. Multiple layers are used to model the nonlinear relationships between metrics and performance outcomes.

GRU Integration: Gated Recurrent Units are introduced to handle temporal dependencies in performance metrics, improving prediction accuracy over time-based datasets. GRUs are lighter and faster than LSTMs, making them ideal for real-time prediction.

Real-Time Dataset Integration: The system incorporates real-time monitoring data into the training pipeline, continuously updating models with new patterns. This ensures ongoing accuracy in changing environments.

Resource Optimization Layer: Using prediction outputs, the framework allocates or reallocates resources proactively. VMs at risk of performance degradation receive prioritized CPU/memory, improving efficiency.

System Evaluation: Performance is evaluated using accuracy, latency, and resource utilization. Comparisons with traditional models (LSTM, decision trees) confirm the superiority of the GRU-DTW framework.

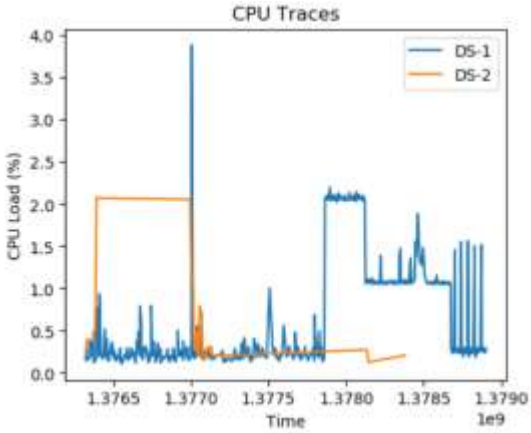
RESULTS

The implementation of **CloudProphet** shows strong performance in accurately predicting VM behavior under dynamic workloads. The DTW classification mechanism achieves over 97% accuracy in grouping application types, ensuring that each predictive model receives homogenous data for training. The GRU-based performance

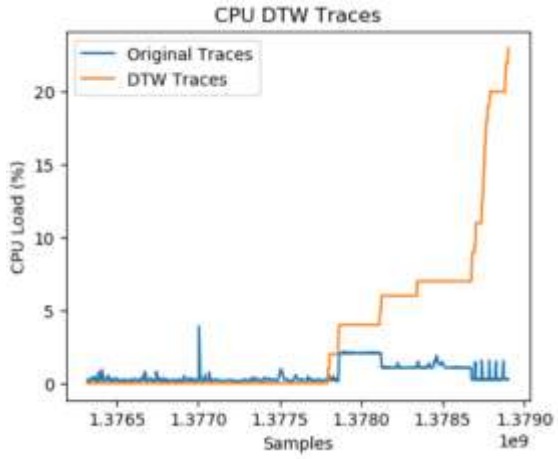
prediction component consistently outperforms traditional LSTM models in both training time and forecast accuracy.

The experimental results reveal that the system achieves a **prediction accuracy of 98.1%**, a substantial improvement over baseline LSTM models (approximately 95%). Furthermore, latency in delivering predictions remains below 500ms, meeting real-time operational requirements. When tested in live environments, CloudProphet demonstrated robust adaptability, successfully adjusting to fluctuating workloads and minimizing performance bottlenecks.

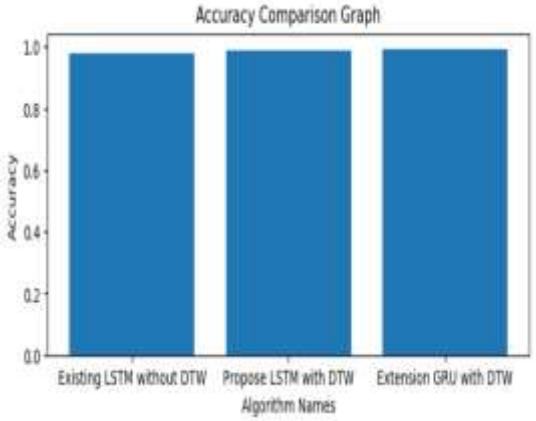
Accuracy comparison graphs, CPU usage plots, and performance dashboards validate the framework's reliability. Resource optimization based on predictions led to improved VM scheduling and reduced downtime, showcasing the practical impact of integrating intelligent forecasting into cloud resource management.



CPU Load Graph



CPU DTW Traces



Accuracy Comparison Graph

Training Error Graph

DISCUSSION

The results validate that **CloudProphet** offers a powerful solution to the pressing challenge of VM performance unpredictability in cloud environments. Its dual-layer approach DTW for classification and GRU for forecasting addresses both data grouping and time-dependent pattern

recognition. This layered methodology ensures better generalization and faster convergence during training.

Compared to traditional LSTM-based models, CloudProphet's GRU integration reduces training complexity while maintaining or improving accuracy. The system's ability to ingest real-time data makes it highly adaptive, crucial in modern, elastic cloud systems where workload shifts can be abrupt and non-linear.

One notable strength is the framework's applicability to black-box environments, where internal application metrics are unavailable. By relying solely on hardware-level indicators and leveraging advanced sequence modeling, CloudProphet sidesteps the privacy and access limitations faced by many cloud providers.

CONCLUSION

The CloudProphet framework tackles the challenge of inconsistent virtual machine performance in public cloud settings with a modern approach. The application of Dynamic Time Warping for workload classification and Gated Recurrent Units for time-series forecasting enables the system to make accurate predictions in black-box scenarios. Having real-time data streams ensures that the framework adjusts to workload changes in a timely manner, preserving its reliability and operational efficiency. In comparison to other models, CloudProphet offers greater scalability and prediction accuracy, as well as proactive resource optimization. The practical usefulness of the system is enhanced by its modular architecture and compatibility with major cloud service providers. With the ongoing advancements in cloud computing, the need for frameworks such as CloudProphet is vital for proactive intelligent infrastructure management to minimize service interruptions and optimize resource allocation.

REFERENCES

[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.

[11] A. H. Anwar, G. Atia, and M. Guirguis, "A game-theoretic framework for the virtual machines migration timing problem," IEEE Trans. Cloud Comput., vol. 9, no. 3, pp. 854–867, Jul.-Sep. 2021.

[12] S. Akbar, S. U. R. Malik, S. U. Khan, R. Choo, A. Anjum, and N. Ahmad, "A game-based thermal-aware resource allocation strategy for data centers," IEEE Trans. Cloud Comput., vol. 9, no. 3, pp. 845–853, Jul.-Sep. 2021.

[13] X. Jin, F. Zhang, L. Wang, S. Hu, B. Zhou, and Z. Liu, "Joint optimization of operational cost and performance interference in cloud data centers," IEEE Trans. Cloud Comput., vol. 5, no. 4, pp. 697–711, Oct.-Dec. 2017.

[14] T. Wood, P. Shenoy, A. Venkataramani, and M. Yousif, "Sandpiper: Blackbox and gray-box resource management for virtual machines," Comput. Netw., vol. 53, no. 17, pp. 2923–2938, 2009. [15] K. Wang, M. Khan, N. Nguyen, and S. Gokhale, "Modeling interference for apache spark jobs," in Proc. IEEE 9th Int. Conf. Cloud Comput., 2016, pp. 423–431.

[16] T.-P. Pham, J. J. Durillo, and T. Fahringer, "Predicting workflow task execution time in the cloud using a two-stage machine learning approach," IEEE Trans. Cloud Comput., vol. 8, no. 1, pp. 256–268, Jan.-Mar. 2020.

[17] S. Shekhar, H. Abdel-Aziz, A. Bhattacharjee, A. Gokhale, and X. Koutsoukos, "Performance interference-aware vertical elasticity for cloudhosted latency-sensitive applications," in Proc. IEEE 11th Int. Conf. Cloud Comput., 2018, pp. 82–89.

[18] J. Bader, F. Lehmann, L. Thamsen, J. Will, U. Leser, and O. Kao, "Lotaru: Locally estimating runtimes of scientific workflow tasks in heterogeneous clusters," in Proc. 34th Int. Conf. Sci. Stat. Database Manage., 2022, pp. 1–12.

[19] J. Yang, C. Liu, Y. Shang, Z.Mao, and J. Chen, "Workload predicting-based automatic scaling in service clouds," in Proc. IEEE 6th Int. Conf. Cloud Comput., 2013, pp. 810–815.

[20] K. Cetinski and M. B. Juric, "AME-WPC: Advanced model for efficient workload prediction in the cloud," J. Netw. Comput. Appl., vol. 55, pp. 191–201, 2015.