UMEÅ UNIVERSITY

Accurate and Low-Overhead Workload Prediction for Cloud Management

Lidia Kidane

Academic dissertation

Which, with the due permission of the Vice-Chancellor of Umeå University for the examination for the Degree of Doctor of Philosophy, is presented for public defence in MIT.A.121 on Friday, 30 May, 2025 at 13:15.

The thesis will be defended in English.

Faculty opponent:
Professor Rolf Stadler, KTH

Computing Science

Organisation
Umeå University
Computing Science

Document type
Doctoral thesis

Date of publication
6 May 2025

Author
Lidia Kidane

Title
Accurate and Low-Overhead
Workload Prediction
for Cloud Management

Abstract

Cloud computing has transformed the IT landscape by offering users and organizations on-demand access to computing power, storage, data processing, and machine learning resources. Despite the benefits, cloud resource management faces challenges due to heterogeneous and dynamic workloads. Inefficient provisioning manifests in two critical forms: underprovisioning leads to degraded Quality of Service (QoS) and unmet Service-Level Agreements (SLAs), while overprovisioning results in unnecessary energy consumption and high operational costs. Through the current rise of AI and machine learning innovations, machine learning based workload prediction for resource provision plays a vital role in predicting future scenarios and identifying new occurrences to prepare service providers ahead of time. However, various challenges are connected with machine learning based workload prediction.

This thesis addresses the challenges of machine learning based workload prediction in cloud environments, such as data drift due to dynamic workloads, high computation overhead, and storage overhead. Firstly, cloud workloads are dynamic, and models trained with old historical data can become obsolete over time. We addressed the challenge of accurate prediction and data drift by incorporating machine learning and streaming data processing algorithms to assist adaptive prediction. Secondly, training and updating deep learning models constantly adds significant computation overhead to the cloud infrastructure. We tackled this problem by proposing a solution incorporating a knowledge base repository with a transfer learning base adaptation. Moreover, we explored the tradeoff to balance model accuracy and computation overhead. Finally, we propose a data compression mechanism leveraging autoencoder to lower storage overhead resulting from the continuous generation of monitoring data in cloud management systems.

Our findings reveal that the proposed methods significantly improved the machine learning based cloud management system. Extensive evaluation with real-world datasets reveals that the proposed methods assist in creating accurate predictions despite the ever-changing patterns in cloud workloads. Moreover, the methods decreased the computation overhead by incorporating the usage of old knowledge and highlighting the tradeoff to achieve a balance between prediction accuracy and computation overhead.

Keywords: Cloud computing, Workload prediction, Machine learning, Resource provision, Data mining

Language
English

ISBN
978-91-8070-712-1 (print)
978-91-8070-713-8 (pdf)

Number of pages
38 + 5 delarbeten/papers