WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Forecasting Cloud Application Workloads with Cloud Insight Model

Prof. Manik Rao Mulge, Humera Tahseen, Divyadarshini, Aishwarya Patil

Dept Of CSE, GNDEC, Bidar/585403, India

A B S T R A C T

Predictive cloud resource management has been widely adopted to overcome the limitations of reactive cloud auto scaling. The predictive resource management is highly relying on workload predictors, which estimate short-long-term fluctuations of cloud application workloads. These predictors tend to be pre-optimized for specific workload patterns. However, such predictors are still insufficient to handle real-world cloud workloads whose patterns may be unknown a priori, may dynamically change over time and may be irregular. As a result, these predictors often cause over-/under-provisioning of cloud resources. To address this problem, we create Cloud Insight, a novel cloud workload prediction framework, leveraging the combined power of multiple workload predictors. Cloud Insight creates an ensemble model using multiple predictors to make accurate predictions for real workloads. The weights of the predictors in Cloud Insight are determined at runtime with their accuracy for the current workload using multi-class regression. The ensemble model is periodically optimized to handle sudden changes in the workload. We evaluated Cloud Insight with various real workload traces. The results show that Cloud Insight has 13%–27% higher accuracy than state-of-the-art predictors..

1. INTRODUCTION

Over the past decade, cloud computing has become a popular infrastructure for industry and research organizations due to its appealing capabilities such as scalability, flexibility, pay-as-you-go billing model. In particular, elasticity has attracted application developers to move towards clouds to deploy their applications. Auto scaling, offered by public cloud providers (e.g., AWS), is the most

common approach for attempting to achieve elasticity. Auto scaling mechanisms and triggers monitor the utilization and behavior of current resources and adjust the size/amount of resources according to the fluctuation of workloads (e.g., job1 requests) and user-defined rules (e.g., upper-/lower-bound of CPU usage). However, auto scaling can often be sub-optimal because of its reactive nature . The reactive nature often results in over- and under-provisioning of cloud resources, in turn, low cost-efficiency and high SLA (Service Level Agreement) violations. Therefore, many predictive approaches have been proposed for addressing the limitations of reactive auto scaling.

The predictive approaches consist of two components; one is a workload predictor, which forecasts future job arrival time/rate; and the other is a resource management compo-nent, which allocates/deal locates cloud resources and maps user workloads to specific cloud resources. To achieve de-sired resource utilization and SLA satisfaction, it is crucial that the workload predictor should be optimized for the behavior of application workloads.

2. SCOPE OF THE PROJECT

This paper is based on our previous work, and we take a step further in the holistic performance evaluation of Cloud Insight, particularly focused on how Cloud Insight improves the overall performance in cloud resource management by minimizing the under-/over-provisioning state of the resources. More specifically, in this work, we provide an in-depth analysis of Cloud Insight's contribution to cloud resource management by measuring cost-efficiency and SLA satisfaction based on trustworthy simulation with representative cloud resource management mechanisms. The improvement aims to obtain a complete understanding of predictive resource management and workload predictors' impacts in real clouds, as well as a better understanding of the effectiveness of our solution.

3. OBJECTIVE

20% less under-/over-provisioning, resulting in 16% better cost efficiency and 17% f ewer SLA violations.

- High accuracy and low overhead: the Cloud Insight framework is an online, multi-predictor based approach that performs highly accurate workload prediction with low overhead under dynamic cloud workloads with various patterns.

- Online ensemble model: a novel online mechanism to create an ensemble workload predictor. This mechanism dynamically assigns weights to each predictor by accurately estimating that the predictor's relative accuracy for the next time interval using multi-class regression.
- Thorough performance evaluation: we perform a comprehensive evaluation of the accuracy and overhead of CloudInsight with various workload traces collected from real cloud applications, including cluster, HPC, and web applications.
- A simulation study of resource management: a trace-based simulation with an auto scaling component confirms the actual benefit of CloudInsight to the resource management for cloud applications.

4. PROBLEM STATEMENT

Existing predictive auto scaling managers often create and/or use a single static (or "one-size-fits-all" style) workload predictor with a simple assumption that the target workload has a stable pattern (e.g., increasing, cyclic bursty, and on-and-off) over time. Therefore, this prediction model is typically built offline and often requires significant efforts and resources to build. Furthermore, since cyclic bursty is known as a typical workload pattern for cloud applications, time-series based approaches are widely used as the one-size-fits-all workload predictor to handle cyclic workloads. More specifically, in this work, we provide an in-depth analysis of Cloud Insight's contribution to cloud resource management by measuring cost-efficiency and SLA satisfaction based on trustworthy simulation with representative cloud resource management mechanisms. The improvement aims to obtain a complete understanding of predictive resource management and workload predictors' impacts in real clouds, as well as a better understanding of the effectiveness of our solution.

5. EXISTING SYSTEM

- Existing predictive auto scaling managers often create and/or use a single static (or "one-size-fits-all" style) workload predictor with a simple assumption that the target workload has a stable pattern (e.g., increasing, cyclic bursty, and on-and-off) over time.
- Therefore, this prediction model is typically built offline and often requires significant efforts and resources to build.

- We first investigated the degree to which a single existing predictor could be used across multiple typical cloud workload patterns.

5.1EXISTING SYSTEM DISADVANTAGES

- Low accuracy and high overhead

Single predictor based approaches are not sufficient to address the dynamics and business of cloud workloads

6. PROPOSED SYSTEM

This framework consists of four main components: 1)a predictor pool, 2) a workload repository, 3) a model builder, and4) CloudInsight workload predictor. The input of this framework is the actual/current workloads (e.g., job arrivals),and the output is the prediction for a near-future workload. The predictor pool is a collection of workload predictors. The workload repository stores the job history of the workload and the prediction history of all local predictors in the predictor pool. The model builder is responsible for creating an ensemble prediction model by evaluating the performance of the predictors in the predictor pool. CloudInsight workload predictor provides the forecast for the near-future workload using an ensemble model created by the model builder. This prediction will be utilized by resource managers for predictive resource (e.g., VM) scaling.

6.1PROPOSED SYSTEM ADVANTAGES

- High accuracy and low overhead

Cloud Insight has better accuracy than state of-the-art one-size-fits-all style predictors

7. METHODOLOGIES

7.1 MODULES NAME

This Project having the fallowing 5 modules:

- User Interface Design
- Admin
- User
- Ensemble Model Creation
- Workload Prediction

7.2 GIVEN INPUT EXPECTED OUTPUT

- **User Interface Design**

Input **:**Enter Login name and Password (User, Router, CA, Publisher)

Output **:** If valid user name and password then directly open the home page otherwise show error message and redirect to the registration page.

- **Admin**

Input **:**Enter email and password , verify all details.

Output **:** Admin verify all user requests and accept user data then data send to user. Admin will verify all data status and user feedback also.

- **User**

Input **:**Enter the name and password and stored data.

Output: If valid user name and password then directly open the user home page. All the resources added by user options. User having some options create new files, read file, update file and delete file also. User verify all details also.

- **Ensemble Model Creation**

Input **:**Creation of model

Output **:** Ensemble model all details and process.

- **Workload Prediction**

Input **:**Verify all prediction information

Output **:** Workload prediction process. This prediction can then be used by a resourcemanagement component for resource scaling.

8. APPLICATION

The performance of all predictors with four workload patterns is measured by MAPE (Mean Absolute Percentage Error)6. The evaluation results are reported in Table 1, showing the best three predictors and an average accuracy from all the evaluated predictors regarding the four different workloads. The result shows that top predictors vary considerably for different workload patterns. There is no single best workload predictor for all workload patterns – each workload pattern has its own best workload

predictor. Moreover, the top three workload predictors for each workload pattern often show similar performance for the workload prediction, implying that best predictors could be changing if the workload contained more randomness or short-term burstiness. It is also worth noting that in Table 1, the best predictors usually contain non-time-series models, such as SVMs or linear regression, because of the lack of trend and seasonality in certain patterns.

9. FUTURE ENHANCEMENT

In conclusion, the mechanism and evaluation results of CloudInsight show that our approach is capable of addressing real-world cloud workloads that have dynamic and high variable nature. This work will help other cloud researchers and practitioners design a new predictive method for man-aging and scaling cloud resources autonomously.

10. CONCLUSION

This paper presents CloudInsight– an online workload pre-diction framework to address dynamic and highly variable cloud workloads. CloudInsight employs a number of local predictors and creates an ensemble prediction model with them by dynamically determining the proper weights (contributions) of each local predictor. To determine the weights, we formulate this problem as a multi-class regression problem with a SVM classifier.

We have performed a comprehensive study to measure the performance and overhead of this framework with a broad range of real-world cloud workloads (e.g., cluster, web, and HPC workloads). Our evaluation results show that CloudInsight has 13% – 27% of better accuracy than state-of-the-art one-size-fits-all style predictors, and it also has low overhead for predicting future workload changes (< 100ms) and (re)creating a new ensemble model (< 1.1sec.).

REFERENCES

[1] Michael Armbrust, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, Ariel Rabkin, Ion Stoica, and MateiZaharia. A View of Cloud Computing. Communications of the ACM, 53(4):50–58, 2010.

[2] Nikolas Roman Herbst, Samuel Kounev, and Ralf Reussner. Elasticity in Cloud Computing: What It Is, and What It Is Not. In International Conference on Autonomic Computing (ICAC ’13), San Jose, CA, USA, June 2013.

[3] Sadeka Islam, SrikumarVenugopal, and Anna Liu. Evaluating the Impact of Fine-scale Burstiness on Cloud Elasticity. In ACM Symposium on Cloud Computing (SoCC ’15), Hawaii, August 2015.

[4] AWS auto scaling. https://aws.amazon.com/autoscaling, 2019.

[5] Microsoft azure autoscale. https://azure.microsoft.com/en-us/features/autoscale, 2019.

[6] Google. Google cloud platform – auto scaling groups of instances. https://cloud.google.com/compute/docs/autoscaler, 2019.

[7] Luwei Cheng, Jia Rao, and Francis C.M. Lau. vScale: Automatic and Efficient Processor Scaling for SMP Virtual Machines. In ACM European Conf. on Computer Sys. (EuroSys), London, UK, Apr. 2016.

[8] Tayler H. Hetherington, Mike O’Connor, and Tor M. Aamodt. Memcached GPU: Scaling-up Scale-out Key-value Stores. In ACM Symposium on Cloud Computing (SoCC ’15), Kohala Coast, Hawaii, USA, August 2015.

[9] Bailu Ding, LucjaKot, Alan Demers, and Johannes Gehrke. Centiman: Elastic, High Performance Optimistic Concurrency Control by Watermarking. In ACM Symposium on Cloud Computing (SoCC ’15), Kohala Coast, Hawaii, USA, August 2015.

[10] Thomas Heinze, Lars Roediger, Andreas Meister, Yuanzhen Ji, ZbigniewJerzak, and Christ of Fetzer. Online Parameter Opti-mization for Elastic Data Stream Processing. In ACM Sympo. Cloud Computing (SoCC ’15), Kohala Coast, Hawaii, August 2015.

[11] Christina Delimitrou and Christos Kozyrakis. Quasar: Resource-Efficient and QoS-Aware Cluster Management. In International Conf. on Architectural Support for Programming Languages and Op-erating Systems (ASPLOS), Salt Lake City, UT, March 2014.

[12] Ganesh Ananthanarayanan, Christopher Douglas, Raghu Ramakrishnan, Sriram Rao, and Ion Stoica True Elasticity in Multi-Tenant Data-Intensive Compute Clusters. In ACM Symposium on Cloud Computing (SoCC ’12), San Jose, CA, USA, October 2012.

[13] HerodotosHerodotou, Fei Dong, and ShivnathBabu. No One (Cluster) Size Fits All: Automatic Cluster Sizing for Data-intensive Analytics. In ACM Symposium on Cloud Computing (SoCC ’12), Cascais, Portugal, October 2011.

[14] Marco A. S. Netto, Carlos Cardonha, Renato L. F. Cunha, and Marcos D. Assuncao. Evaluating Auto-scaling Strategies for Cloud Computing Environments. In IEEE International Symp. Modelling, Analysis and Simulation of Computer and Telecommunication Systems (MASCOTS ’14), Paris, France, September 2014.

[15] Timothy Wood, LudmilaCherkasova, Kivanc M. Ozonat, and Prashant J. Shenoy. Profiling and Modeling Resource Usage of Virtualized Applications. In The 9th ACM Middleware Conference (Middleware ’08), Leuven, Belgium, December 2008.

[16] WaheedIqbala, Matthew N. Daileya, David Carrerab, and Paul Janeceka. Adaptive resource provisioning for read intensive multi-tier applications in the cloud. Future Generation Computer Systems, 27(6):871–879, 2011.