AN INTELLIGENT COST OPTIMIZED AUTOSCALING FRAMEWORK FOR HYBRID CLOUD

A THESIS

Submitted by

RADHIKA EG

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION AND COMMUNICATION ENGINEERING
ANNA UNIVERSITY
CHENNAI 600 025

MARCH 2022

ABSTRACT

Cloud computing is a revolutionary paradigm enabling on-demand provisioning of computing resources. It has shown a disruptive impact on everyday computing task. In cloud environment, the resources are delivered to cloud consumers in the form of infrastructure, platform and software services. Resource utilization plays a major role in achieving high performance, high availability and cost sustainability while adhering to Service Level Agreements (SLA). Autoscaling is one of the key features of cloud computing that has the ability to dynamically scale up or scale down the resources based on application utilization such as Central Processing Unit (CPU), Random Access Memory (RAM) and network throughput. It reduces the manual effort and provides the services quickly while adhering to SLA. The efficiency of autoscaling relies on sufficient resources offering at an optimized cost to meet the future demand. This feature is a critical aspect with any deployment model in cloud computing.

A hybrid cloud deployment model allows customers to host applications in multiple heterogeneous clouds. It blends the cost-effectiveness of public clouds with the control and security of private clouds. The most common deployment model for handling peak load is hybrid cloud which performs resource scaling from a private cloud to the public cloud. Although it extends the infrastructure services over various public clouds, choosing an appropriate Cloud Service Provider (CSP) based on user requirements at an optimized cost and SLAs is more complex. Different CSPs offer a wide range of flavors or instance types at different cost with same performance and capacity for heterogeneous applications. Most of the organizations map CSP and flavor to Virtual Machines (VMs) manually based on user preferences to perform resource autoscaling. Thus an effective autoscaling is mandatory in a hybrid cloud environment which requires the knowledge of each CSP offerings, application workload type, hosting environment, life time of VM, and flavor choices.

Thus, the thesis aims to design an intelligent cost optimized framework for hybrid cloud to improve the autoscaling efficiency with predictive models. The proposed framework has three objectives:

- Objective 1: To propose an effective predictive technique for heterogeneous workload in a hybrid cloud environment.
- Objective 2: To propose a flavor recommender framework for heterogeneous applications during autoscaling in hybrid cloud.
- Objective 3: To propose a Budget Optimized Virtual Machine provisioning (BOPVM) framework for the period of autoscale based on ranking of CSPs.

The first research objective is to identify an efficient workload prediction model for web application and integrate with Openstack private cloud. Many existing reactive autoscaling solutions use rules with thresholds that are based on infrastructure-level metrics, such as CPU utilization and memory (Al-Dhuraibi 2017; Taherizadeh 2019). Although these solutions are easy to implement, choosing the right values for thresholds is difficult because the workload continuously fluctuates depending on the user behavior. For predictive solutions, many works using time series data analysis have been proposed (Fang 2012; Calheiros 2015) whereas only a few have used deep learning algorithms for designing autoscalers (Tang 2018; Imdoukh 2020). Recurrent Neural Networks (RNN) is a deep learning technique which has been implemented in many fields including computer vision, speech recognition, time series analysis, etc. (Marie-Madelaine et. Al. 2020). It also brings a promising future to architect general autoscalers (kwan et. al. 2019). RNN determines the response by using feedback loops, which combine current inputs with outputs of the previous moment. Feedback loops allow sequential information to persist and allow recurrent networks to perform tasks that cannot be performed by feed forward neural networks. Unfortunately, regular RNN still loses its memory very fast. Hochreiter *et. al.* (1997) have proposed a special type of RNN, called Long Short-Term Memory network (LSTM), with ability to recognize and learn long term dependencies.

As the data represented and analysed in the proposed work is a time series and hence the proposed framework is Proactive Prediction Engine (PPE) based on Recurrent Neural Network Long Short Term Memory (RNN_LSTM). PPE forecasts the future workload and autoscales the resources in private cloud. The proposed research considers CPU utilization as time series data points and sent as input to the PPE. The results show that the prediction performance of RNN_LSTM on dataset 1 has shown 40% decrease in Mean Absolute Error (MAE) and 30% decrease in Root Mean Squared Error (RMSE) over the existing Autoregressive Integrated Moving Average (ARIMA) algorithm. Similarly, RNN_LSTM on dataset 2 has shown 40% decrease in both MAE and RMSE over ARIMA. Hence, RNN_LSTM is used to efficiently forecast CPU resources and this model is utilized as prediction model for objectives 2 and 3.

The second objective is a flavor recommendation framework proposed in chapter 4. This framework recommends a cost optimized flavor for the forecasted workload during autoscale process. The Proactive Predictive Engine (PPE) and the Recommendation Engine are two key components of the framework. A scoring engine and a flavor engine combine to form the recommendation engine. The forecasted value from PPE is provided as an input to the scoring engine to determine the number of CPU or RAM resources required. The flavor engine is responsible to choose the best flavor among the registered public and private CSPs. The performance of the proposed work achieves 15% of cost savings for autoscale period compared to the traditional Amazon Web Services (AWS) approach. The number of VMs provisioned using the proposed work is 17% less than the traditional AWS.

The third objective is Budget Optimized VM provisioning framework (BOPVM) proposed in chapter 5. In a hybrid cloud environment, the proposed framework analyzes and ranks the CSPs to offer additional forecasted resources in either public or private cloud during the autoscale period. The system evaluates and ranks CSPs based on the extent analysis of Fuzzy Analytic Hierarchy Process (FAHP). FAHP extent analysis approach (Chang 1996) is the most widely used method to process fuzzy comparison matrices and to extract the weights from a fuzzy comparison matrix with the degree of possibility. When compared to Analytical Hierarchical Processing (AHP), Fuzzy theory solves the fuzziness of human decision-making challenges by determining truth-values as fuzzy sets and constructing approximate inference rules instead of exact rules. The QoS inputs used in the proposed framework are performance, cloud availability, cost, scalability, and reliability. Based on the CSP ranking, the framework allocates resources within the user budget. The performance of the BOPVM framework is compared with the existing algorithms such as Predictive Resource Management Framework (PRMF) and Dynamic Resource Provisioning and Monitoring (DRPM). The proposed BOPVM has shown 74.4% reduction in the number of virtual machines compared to DRPM, and 69% reduction when compared to PRMF algorithms. The VM provisioning time using BOPVM algorithm is 64.7% less than DRPM and 50.6% less than PRMF. The total VM provisioning cost using BOPVM is 50.6% less than that of DRPM and 47.5% lesser than that of PRMF.

Thus, the research makes a significant contribution towards enhancing the effectiveness of autoscaling in hybrid cloud environment. It intends to improve the cost optimization solutions for issues like resource utilization, predictive scaling, flavor and CSP Selection based on customer interests in hybrid cloud. Further, the research exploits the efficiency and cost benefits of using private cloud Openstack to distribute workload in hybrid cloud environment.