A SURVEY ON WORKLOAD PREDICTION MODELS IN CLOUD BASED ON SPOT INSTANCES FOR PROACTIVE AUTO SCALING STRATEGY

R. Yuvha Secaran¹, E. Sathiyamoorthy²

¹VIT University, India. aspringcoder9313@gmail.com
²VIT University, India. esathiyamoorthy@vit.ac.in

Received: 15.12.2019 Revised: 18.01.2020 Accepted: 20.02.2020

Abstract
Auto scaling techniques help exploiting the elastic nature of cloud by provisioning resources on demand. Very often, the application hosted in a cloud tend to face workload surges which causes the application to respond slow or deny requests. There are reactive and proactive strategies for auto scaling to the rescue which helps provisioning instances based on the demand. When the demand is met after the workload surge the auto scaling is reactive in nature. Otherwise, the demand can be predicted beforehand to provision instances that serve during the surge is proactive in nature. Off late , the proactive auto scaling is gaining more attention amongst researchers. This survey presents a comprehensive detail about workload prediction in cloud and bidding for spot instances in cloud. The survey results uncover a possible concept which requires more attention and the one which helps tapping into resource heterogeneity of the spot market with a cost oriented benefit.

Keywords: Spot Instances, Workload Prediction, Auto Scaling.

INTRODUCTION
The advent of cloud computing has brought in an inclination towards moving away from traditional data centre for hosting applications. The perspective has broadened from just hosting the application to making it more secure, robust and reliable over a cloud. This perspective of utilizing cloud resources opened up several research challenges. One such challenge is scalability, which helps exploiting the elastic nature of the cloud resources. Very often, the application hosted over a cloud experience workload surges. These surges are dynamic in nature with or without a pattern . It causes the request to be dropped or respond slower. This necessitates the need to scale resources dynamically as per the workload surge. Handling the scalability can be either done in a reactive or proactive manner. Off late, proactive strategy is gaining much attention amongst researchers.Instances are the resources available in cloud with various configurations for hosting application requirements. Cloud providers offer these instances at a pricing which can be availed through reservation, on-demand and through a market based auctioned offering called spot instances. A workload prediction model has to take into account the dynamic nature of the spot market while considering to provision the spot instances proactively. This paper aims to provide an overview of the workload prediction in cloud , the strategies adapted and the parameters that are optimized without going too deep into implementation details. The first part of the paper discusses the efforts related to auto scaling followed by time series analysis and the workload prediction in cloud and some of its applications. The next part describes about the spot instances followed by bidding efficiently using Amazon EC2.

WORKLOAD PREDICTION IN CLOUD
Scaling as a Function of Forecasting
A system which exploits the elasticity of the cloud by allocating the right amount of resources to the application workload with little or no manual provisioning can be stated as a auto scaling system. The applicability of a auto scaler is attributed to the collection of servers that were replicated and accept incoming requests from the load balancer. The resources can be scaled either in a horizontal or a vertical manner. Horizontal scaling denotes adding or releasing replicas of existing resources (Scaling in /out). Vertical scaling denotes increasing or decreasing the computing capacity of a resource (scaling up/down). The time at which the scaling operation is carried out denotes the strategy adopted for scaling. There are reactive strategy which allocates resources after the workload surge has occurred where as the proactive strategy predicts and allocates resources which are ready to process the workload surge. The proactive strategies have been gaining much attention amongst researchers. Authors of (Chenhao Qu et al., 2016) have proposed fault tolerant model using spot instances for hosting web application. Additionally, policies for auto scaling which are cost aware and complies with fault tolerant mechanism and works by provisioning a combination of on-demand and spot instances were also proposed. The focus is to optimize the response time of the requests and the cost of instances. (Roy,N et al., 2011) propose a proactive dynamic resource provisioning where in the auto scaling algorithm predicts the workload. Then look-ahead optimization technique is used in order to identify a time interval and use the prediction to change the resources.

In addition to this the algorithm takes into account the cost factors that control the behavior of the system associated with the scale up and scale down. The cost factors are SLA cost, cost of reconfiguration and the cost of leasing the machines. (Anshul Gandhi et al., 2014) propose a proactive optimal scaling policy and aims at characterizing the workload performance in relation to the auto scaling policies when most work focused on system characteristics and frameworks. The core of the paper is an engine which characterizes the workload and also evaluate the options to scale. The evaluation is done with the help of threshold values and the output from a predictor component which predicts the workload in the next monitoring window. Although, the prediction is medium term. The outcome is that a combination of both scale up and scale out proves to be the optimal scaling policy. (Anshuman Biswas et al., 2015) propose a proactive dynamic resource provisioning scheme which is a broker based architecture to rent resources available with the cloud provider and create a virtual cloud which is private in
scope. The user requests the resources through the broker. The charges to the user are based upon cost per second rather than cost per hour as it is done by the public cloud provider. The auto scaling technique uses machine learning approach which is predictive in nature and with a constraint based on deadline. This approach predicts the next set of k request and enumerates five parameters that are characteristics of the predicted request. These prediction help to extend the stop time or acquire resources. Several overheads pertaining to communications and machine learning were discussed and found to be relatively low. The main objective is to provide a model which maximizes the broker profit and minimizes the user cost. (Tania Lordio-Botran et al., 2014) have extracted a conclusion from the review that, it is better to focus on developing proactive auto scaling systems. These systems help predicting the future needs and also act on provisioning them with enough guarantee. The authors also opine that time series analysis techniques should be taken into consideration for prediction related capabilities.

**Time Series Analysis**

(Robert H. Shumway and David S. Stoffer 2011) The systematic approach used to answer questions posed by correlations relating to time is referred to as time series analysis. Time series are used in many domains and also represent the measure of change in over a period. A time series is a set of observations xt, each observation is recorded at time t. The series can be discrete when the observations are made at fixed intervals or it can be continuous when the observations are recorded continuously between some interval say (0, 1). The first step in time series analysis is to carefully observe the recorded data plotted over time. The result of observation is to suggest the method used for analysis and the statistics that can be used to summarize the information. There are two approaches in the time series analysis based upon the time and frequency domains. The time based domain approach specifies that the future values can be modeled based upon the current and past values by parametric function, where as the frequency based domain approach assumes primary interest in time series analysis that relate to variation in time found naturally in the data. These variations can be periodic or systematic sinusoidal variations. The periodic variations can occur due to phenomenon which are of biological, physical or environmental in nature. A time series can be defined as a collection of random numbers expressed as x1, x2, x3, ..., xn. In general, a collection of random variables (x_t) indexed by t is a stochastic or random process. A time series can be plotted graphically with the variables as the vertical axis and the time as the horizontal one.

The plotting helps identifying adjacent time periods which in turn produces a discrete sample for observation. The time series could differ for each scenario and the fundamental visual characteristic distinguishing is the differing degree of smoothness. This smoothness is evident by the fact that the points adjacent to each other are correlated. It is important to understand the difference between prediction and forecast for this context. As stated in (Nate Silver 2012) The meaning of prediction and forecast are used differently in different fields. In some fields the meaning becomes interchangeable where as other disciplines differentiate them. To better illustrate, seismology is more sensitive to the distinction. A Prediction is a definitive statement where as a forecast is a statement based on probability with a elaborate time scale. For ex: an earthquake will occur on Jan 1 at 12:00 hrs is a prediction where as there is a 40 percent chance that the earthquake will occur on Jan 1 is a forecast. From the context of cloud and while using time series the terms prediction and forecasting are used interchangeably. With scalability as the prime factor auto scaling strategies use prediction which help realizing the workload at a future time slot thus reacting proactively.

There are several works carried out in this aspect and the survey of papers which use workload prediction in cloud for scaling , the prediction technique used are outlined as follows. (Rodrigo N. Calheiros et al., 2015) proposed a proactive auto scaling mechanism which uses workload prediction using ARIMA model. In addition to the prediction there exists a confidence level associated to the prediction. The system functioned with an average accuracy of up to 91 percent resulting in efficient resource utilization and lower impact on QoS. (Yazhou Hu et al., 2016) proposed a strategy for predicting the scale up or down. The monitoring data is fit to a time series for analysis purpose and then the kalman filtering is used to predict the workload time series. The trigger is based upon the pattern matching technique and its performance is better than the threshold approach. The resources in use are not spot instances. (Joseph Doyle et al., 2016) discussed about how to use, Pass and SAS-Computation as a service. The proposed idea here is to identify the compute unit seconds (CUS) a metric that specifies how long does it take for the workload to complete its execution. This is done using Kalman filtering. (Wei Fang et al., 2012) proposed a scheme which predicts the future demand for resources and does proactive provisioning of resources for applications. It uses ARIMA model for predicting the workload and the system handled CPU intensive applications well. (Samuel A. Ajila et al., 2013) evaluates techniques like Support vector machines, Neural networks and Linear regression with SLA metrics namely throughput and response time. A random workload pattern is employed for realistic simulations. The Support Vector machine scores better than the other two techniques. (Anshuman Biswas et al., 2014) proposes a proactive auto scaling technique for resources where in the number of resources are adapted based upon the system load. Support vector machine scores better than the linear regression. (Hector Fernandez et al., 2014) propose a approach to avoid heavy cost setup and come up with a online based profiling approach that takes advantage of the currently provisioned resources and real workload to deduce the profile for the VM types by ascertaining the "optimized throughput".

The profiler comes up with the optimized throughputs for each VM instance types by taking into account the request rate, response time, total CPU usage in terms of percentage. These are some of the common metrics that are used to measure the throughput. With these parameters the Profiler applies a "Data Smoothing" technique that gives out more precise "Optimized throughput" by ruling out the inconsistencies caused by virtualization, OS provisioning and also poor hardware configurations. The profiler is used to classify the resource types based on the computing capacity and gives us a Ideal throughput for the particular instance type. This ideal throughput defines the amount of worker required to serve the workload (or the request rate) or it can be said that it serves as a factor to ascertaining the performance capacity of one VM instance type in serving a workload. (Jingqi Yang et al., 2014) aims at providing a cost aware auto scaling system that works on the basis of predicted workloads. The algorithm to predict the workload uses Linear regression model and it gives out a good predictive deviation that serves as an input for the auto scaling system. There are four algorithms that takes care of monitoring, scaling up and down on a real time basis and then pre-scaling based upon the predicted workload. The system scores better than the horizontal scaling and lightweight scaling with which it was compared and evaluated. More over this system is cost efficient and the rate of SLA violation is low. (Sadeka Islam et al., 2012) focus on modeling the prediction system which can forecast the sudden surge in requirement for resources.

TFC-W benchmark data is used to train the prediction models and see if the output is as expected. The efficiency of the
Spot instances represent the unused Amazon EC2 instances which can be obtained on a auction basis from Amazon. These instances are cost effective choice and are well suited for applications which does batch job, data analysis and optional tasks. The instances do not start immediately and its price vary hourly based upon the supply and demand for it in the market which is set by Amazon. Using spot instances allows to use different compute capacities for the same budget with an added advantage of cost reduction. Some of the key terms when dealing with spot instance. I. Spot price : The hourly price of the instance set by Amazon based upon the supply, demand and also the last fulfilled bid. Amazon also provides spot price history to help bidding the best price. II. Spot Bid : Also called as spot request. Denotes the bid price a user is willing to pay per hour for the spot instances. Amazon will allocate the resource if the price which was bid goes beyond the current price of spot instance.

The requests can be one-time or persistent request. A persistent request will re-submit automatically after the instance related to the request gets terminated where as a one-time request does not re-submit after termination. III. Spot instance Termination: When the bid of instance exceeds the price that was bid or there are no longer any unused EC2 instances then Amazon flags the instance for termination. Amazon serves a termination notice which gives a 2 minute warning. After that the instance terminates. IV. Bid status: Denotes the current status of the bid. To start using spot instance, a spot request must be placed with Amazon. The request contains details like bid price, instance types, the availability zone and the number of instances. When the quoted price is more than the current price of the spot instance and the instance is available then the request is accepted. Otherwise, the request is accepted when the spot price falls below the bid price. Spot instance termination specifies the criteria which arises when a spot instance can be terminated. The terminations happen under two conditions. Firstly, if the users bid is lower compared to spot price of the instance. Secondly, if there are no resource in the spot instance resource pool of a particular instance type in a location (data centre). If the spot instance is terminated by the AWS platform due to unavailability of resources in the resource pool, then the customer is waived of his partial hour usage bill. This is also applicable in case when AWS withdraws the spot instance before completion of one hour usage due to increase in spot price.

Bidding the Optimal Price

The context relevant to spot instances always outline the importance of spot price and the need to bid in such a way the spot price stays in use much more time. Several works were carried out using spot instances. (Chenhao Qu et al., 2016) extended the application of spot instances to web application by associating with fault tolerant semantics. Efforts pertaining to efficient bidding price were also discussed. (Artur Andrzejak et al., 2010) propose a probability distribution function to arrive at conclusions which enable users to optimize cost, performance and reliability. To establish a decision process and to choose the optimal instance type and bid price the distributions for the random variables are pre-computed and the decisions are finalized based upon the deadline constraint and the budgetary constraint. (Vivek Kumar Singh and Kaushik Dutta, 2015) proposed a novel algorithm for spot price prediction. The historical data of spot prices were used and the predicted price is calculated for both short and long duration. A gradient decent algorithm is used to solve the cost optimization problem. The hourly price of spot instance at the ith hour and the jth month along with the price 24hrs before the current time. (Sifei Lu et al., 2013) discussed on reducing the costs , increasing reliability and optimizing the complexity associated with fault tolerant mechanism without having any effect on the scalability and performance. There is a disturbance effect observed with the spot price trend when a large number of spot instance is requested.

The dynamic resource provisioning solution includes both on-demand and spot instances to run large scale applications. The backup on-demand instances were put in place to serve as an alternative to switch to when the spot instances get terminated. Bidding a higher price can be used to increase the instance price and the price approaches on-demand pricing. Bidding at low price increases the risk of instance termination hence a optimal bidding price is chosen as a trade off for spot price against the rate at which the system terminates. (Rajkumar Buyya et al., 2013) observed the pattern with spot price for a year and provided a statistical perspective in a global manner to help understand the pricing for spot instances. The proposed model estimates the total cost of running a work using spot instances uses 8 spot types. Discussions about spot price characteristics are provided when understanding the patterns associated to the spot price. Firstly, A pattern of spot price is observed by plotting the spot prices over a graph. Two zones of interest were chosen to study this pattern. It has been shown that the spot prices tend to be random rather than market driven hence analysis of global statistics can reveal some basic facts about spot prices. (Deepak Poolla et al, 2014) propose two different heuristic which map the workflow to spot and on-demand instances. The bidding strategy in place is smart enough to arrive at a bid price which suits the workflow requirements. The focus is on the execution cost and the system in place is more robust and fault tolerant. Additionally checkpointing method of fault tolerance is explored and results are substantiated with cost savings.

Only one spot type is used and the bidding strategy termed intelligent, considers the parameters like the previous bid’s failure probability, on-demand price, spot price along with additional parameters which control bid price. (Daniel J Dubois and Giuliano Câsale, 2015) focus upon identifying the rental and the location policies which consider the spot price. The approximated models are used for the random environment analysis. The models are efficient in showing the response time distribution. In this paper the random analysis is modeled around the spot price changes to take into account as an effect while computing the mean response time and the response time distribution. The act of using the resources is broken into specific sub problems which feed the other in an iterative fashion. The first step where the The minimal computational requirements in terms of resource rates is decided. The next step is to calculate the bidding price which satisfies the service level agreements and then the resources are rented. The Sum of the ECU resources
must be large enough or should satisfy the previous step. The next step is to choose the resources to allocate the application. Analyzing the overall system performance to see if there are any scaling up bottlenecks. This is where the random experiments are considered which covers up the uncertainty that occurs when the spot instances are terminated. Overall the authors call this to be the cost-efficient approach for provisioning resources and also allocating the application components to the resources.

CONCLUSION

The spectrum of work which exploits Cloud provider’s elasticity was discussed in an elaborate way from a three point perspective viz auto scaling, workload prediction and bidding in the cloud. The perspective of auto scaling have been worked upon from various aspects relating to cost-effectiveness, finding out optimal scaling policy, provisioning a mix of resources and a broker based architecture. These aspects strictly work based on the proactive strategy for scaling resources. Proactive scaling strategy requires the workload to be predicted in the future time slot. The survey continues with outlining the workload prediction models which use various techniques like ARIMA, Kalman Filtering and SVM (to name a few) applied over time series to arrive at a prediction in a future time slot. The prediction result can be the workload, sometimes metric(s) which influence the decision of scaling the resources. The decision process is system specific wherein various ideas were proposed. Most of the contributions have focussed on provisioning the readily available on-demand instances as a result of applying the prediction model. Partly, some of the contributions were focussed on provisioning instances that are offered based on the auction mechanism called the spot instances. The survey continues further to outline the contributions in terms of bidding for spot instances. Bidding in cloud is often the topic of interest when dealing with spot instances. The contributions towards finding the optimal bid were discussed and some of the works even tend to characterize the nature in which the bidding was done considering the historical spot prices. To conclude, this survey uncovers the idea wherein proactive auto scaling strategies which consider the dynamic nature of the spot market requires more attention. If a system which predicts the workload adapts to the dynamic nature of spot market by taking advantage of the delay in bidding and provisioning the spot instances then the benefits of cost reduction and resource heterogeneity can be easily tapped to the cloud users benefit.

REFERENCES

A SURVEY ON WORKLOAD PREDICTION MODELS IN CLOUD BASED ON SPOT INSTANCES FOR PROACTIVE AUTO
SCALING STRATEGY

Journal of Management, IT and Social Sciences, 6(4), 1-7. https://doi.org/10.201744/irmjssn.v6i4.631


