CHAPTER 5

EFFECT OF MACHINE LEARNING TECHNIQUES FOR ENERGY BASED VM CONSOLIDATION

5.1 INTRODUCTION

Machine Learning, a branch of artificial intelligence, concerns with the construction and study of systems that can be learnt from data. Machine learning focuses on prediction, based on known properties learned from the training data. Regression Techniques are classified as Supervised Learning in which the learning is based on that input where the desired output is known, while Unsupervised Learning is based on the input where the desired output is unknown. In addition to this is the Reinforcement Learning, which refers to the synthesis of acquired knowledge. CPU utilization in a datacenter has to be predicted to know if the host is over-loaded or not. We take only reliable or alive nodes into consideration.

This chapter concentrates on cloud data center on how it should handle workload and address QoS issues when the service oriented architecture is concerned. The service oriented architecture has to be harnessed and Cloud is the best paradigm to do this. Cloud Computing should be used to handle the workload and analyze the Host Characteristics. The host or processing element characteristics has to be efficiently analyzed to predict if it is overloaded or under-loaded. The Statistical analysis gives a wider analysis of the Host and efficient scheduling is guaranteed. Workload independent QoS metric is necessary to study in an energy perspective where
it will reduce the Performance degradation due to migration. Performance Degradation due to migration is analyzed. The energy and Performance of the system with the proposed Host Character is analyzed. The analysis shows some efficient results which are exhibited in this chapter. Whereas in our QoS analysis since application is not specific and can change dynamically it is not possible to define QoS metrics and so we define the QoS metrics for an exhaustive analysis. A workload independent QoS metric is required for defining QoS requirements in the SLAs to constrain the degree of VM consolidation and acceptable performance degradation. Workload independent QoS metric is necessary which can analyze the response time of the workload. Oversubscribing of resources is one such solution to harness the available resources in a data center. Although the Oversubscription may lead to performance degradation and in-turn Energy Consumption gets higher for the available resource.

5.2 AIM AND OBJECTIVE

The data being the observation of the CPU utilization obtained from the PlanetLab trace, the trace when processed gives average CPU utilization for a day for the number of VMs. The utilization is then subjected to a regression analysis. The previous works involved LR (Local Regression); in this work we adapted the CQR (Composite Quantile Regression) for host overloading prediction. The regression predicts the HostOverloading based on the CPU utilization and analyze if the Host is overloaded or not. The hosts’ state of over-loading is also based on a binary property where we consider the hosts in a Data Center as reliable. The reliable hosts are considered for their load upon which the scheduling is performed. The CQR technique is used to predict the Host Overloading where the Hosts are predicted based on construct local CQR smoothers as non-parametric estimates of the regression function and its derivatives. Here we have introduced quantiles which can
also be called as percentile. The trace we used denoted CPU utilization to be dense in a particular environment.

Cloud Computing has gained huge popularity for what it can offer for business and individuals in terms of cost effectiveness. However, maintaining the Clouds data centers can cost a lot of money and energy. Therefore, the overall aim of this project is to identify and evaluate a technique, such as DVFS, that can be used to improve the eco-efficiency and performance of Clouds data centers.

In order to achieve this aim, a number of objectives have been outlined:

- Understand the importance of cloud computing, architecture of clouds, types of clouds, models of clouds, and virtualization.

- Understand some issues associated with Cloud Computing in terms of data center management.

- Understand Cloud simulation, in particular CloudSim toolkit in terms of its usability, capability and limitations to offer a feasible evaluation of the eco-efficiency techniques used in data centers. To study the QoS metrics it is necessary to model a Cloud host characteristics model.

In the previous chapters we have learnt that the VM consolidation is necessary as it is an important component than the CPU when consolidation is concerned. But CPU or the host characteristics is more important to analyze specifically the reliability constraints. Hence we have modeled a Cloud character model.

Bin-packing cannot address multiple goals at the same time, and optimization quality is not ensured. Hill climbing cannot react to updates
quickly. Modeling a regulated framework for a large scale service system is challenging, however control theory performs well for dynamic issues. Machine Learning requires an extraordinary arrangement of history information, and is frail in tending to new traces. Flow Network Optimization is successful for modeling a few issues, yet cannot address nonlinear issues. Non-linear optimisation is constrained by the effectiveness of the iterative gradient search.

The new optimization algorithms developed in this research are required to deliver all these properties at the same time. They must be applicable to managing a set of applications to share a cloud infrastructure in an optimal manner, ensuring multiclass users the expected performance at low cost. They should be able to provide high quality solutions that persist, taking into account the risks and costs associated with changes.

5.3 LEGACY MACHINE LEARNING TECHNIQUE

5.3.1 Local Regression

The next heuristic will depend on the Loess procedure (from the German loess – small for local regression) recommended by Cleveland. The main notion of the local regression procedure is fitting simple models to local subsets of data to develop a curve that approximates the main data. The observations \((x_i, y_i)\) are designated neighborhood weights while using the tricube weight operate. The regression technique has been previously discussed in Section 3.4.3.

5.4 OVERVIEW OF THE SOLUTION

The solution which has been proposed in this chapter provides dynamic management by solving a sequence of static deployment problems as conditions change. It includes an optimization approach for static deployment
based on a snapshot of the system requirements and state, and a persistent control mechanism, giving stable/persistent management in a variety of dynamic environments. This solution models the problem with these goals by an optimization model comprised of an objective function (which addresses multiple goals by weights, rewards and penalties on the objectives) and a set of constraints.

The composite quantile regression (CQR) estimator has recently been proposed by Zou and Yuan (2008) for estimating the regression coefficients in the classical linear regression model. Zou and Yuan (2008) showed that the relative efficiency of the CQR estimator compared to the least squares estimator is greater than 70% regardless of the error distribution. Furthermore, the CQR estimator could be much more efficient and sometimes arbitrarily more efficient than the least squares estimator. These theoretical properties of CQR in linear regression motivate us to construct local CQR smoothers as non-parametric estimates of the regression function and its derivatives. The Local Regression technique uses Ordinary Least Squares (OLS) which is the minimum sum of squared errors as shown in Equation (5.1). Whereas in CQR we use the region of particular loaded quantile of the CPU utilization due to PlanetLab trace.

- OLS = \( \min \sum_i e_i^2 \)  \hspace{1cm} (5.1)
- Median regression or LAD regression = \( \min \sum_i |e_i| \)  \hspace{1cm} (5.2)
- CQR = \( \sum_i q |e_i| + \sum_i (1 - q) |e_i| \)  \hspace{1cm} (5.3)

The CQR technique has been used to address the host overloading. The prediction changed the perspective of forecasting the host status. The VMs are selected based on the proposed MPP policy and this helped in achieving better results. The results have a better approach towards the energy perspective. The performance degradation is efficient and the migration has
happened in a more efficient manner. This in turn helped to maintain a lesser energy consumption.

5.4.1 CQR Based MPP Technique

For the MPP policy we combine the CQR technique instead of Local Regression (LR) and harnesses better based on the upper utilization of the CPU utilization. In the (Beloglazov & Buyya 2012) paper they have judged the LR MMT is the efficient algorithm where the loss function could be approximated to give an efficient algorithm. We base our algorithm on the Composite Quantile Regression proposed by (Zou & Yuan 2008) which can estimate the regression coefficients in the classical linear regression model. The relative efficiency of the CQR estimator compared with the least squares estimator is greater than 70% regardless of the error distribution. Furthermore, the CQR estimator could be much more efficient and sometimes arbitrarily more efficient than the least squares estimator. The main idea of the method of local regression is fitting simple models to localized subsets of data to build up a curve that approximates the original data. From the workload CPU utilization instances, we form localized subsets and implement CQR to find a better way to handle host oversubscription if it takes place.

If \((t_i - y_i), i=1\ldots n\) is an independent and identically distributed random sample of the CPU utilization taken from the PlanetLab trace. We estimate \(\chi(t) \approx \chi(t_0) + \chi'(t - t_0)\) and then fit a linear model locally in the neighborhood of \(t_0\). Let \(k(\cdot)\) be a smooth kernel function, the local linear regression estimate in \(x(t_0)\) in \(\hat{\alpha}\), as shown in Equation (5.4).

\[
(\hat{\alpha}, \hat{b}) = \text{arg}\min_{\alpha, b} \left[ \sum_{i=1}^{n} \left( y_i - \alpha - b(t_i - t_0) \right)^2 \right] \quad k(t_i - t_0) \quad (5.4)
\]
where \( h \) is the smoothing parameter.

The best properties that CQR combines are the design adaptation property and high mini-max efficiency property. This method helps in finding the host oversubscription and involves the constraint as to analyze based on the response time and the CPU operating frequency.

The observations are estimated for a cost function which gives the predicted trend line. We propose the Composite Quantile Regression smoothing as follows.

Let

\[
\varphi_{lm}(r) = l_m a - a l (r < 0), m = 1, 2, ..., q
\]  

(5.5)

For \( q \) check loss functions at \( q \) quantile positions is given by the Equation (5.6)

\[
l_{ln} = \frac{m}{q+1}
\]  

(5.6)

The size of the subset is defined by a parameter of the method called Bandwidth. In the linear regression model the CQR loss is defined as shown in Equation (5.7).

\[
\sum_{i=1}^{q} \sum_{i=1}^{n_i} \alpha^{n_i-k} \varphi_{lm} (y_i - a_n - bt_i)
\]  

(5.7)

CQR combines the strength across multiple quantile regressions with forcing a single parameter for the slope. In our CQR algorithm, for each new observation from \( \hat{a}_1, \hat{a}_2, \hat{a}_3, ..., \hat{a}_q, \hat{b} \) we find a new trend line as shown in Equation (5.8).
Quantiles are equally spaced, \( l_m = \frac{1}{t+1} \)

The local linear estimator is given by

\[
\hat{\chi}(t_0) = \frac{1}{q} \sum_{t=1}^{q} \hat{a}_{mt}
\]

(5.9)

\[
\hat{\chi}'(t_0) = \hat{b}
\]

(5.10)

The trade-off has to be addressed based on the QoS that can be satisfied when the SLAV can be efficiently maintained or minimized. The host (or) processor available can have different status based on their availability in any point in time. The execution time and time analysis of various parameters depend on the CPU operating frequency. The local linear estimates the VM host oversubscription precisely and efficiently than the Local Regression (Zou & Yuan 2008). Instead of OLS (Ordinary Least Squares) used in Local regression to estimate the error, we have used CQR and the quantiles of the data is considered. The quantile or percentile has helped to efficiently analyze the CPU utilization caused due to real time trace. OLS omits some data and takes the nearest neighbor-hood weights but the error caused is higher in local regression when compared to CQR. Its efficiency regardless of the error distribution has been evident as discussed in the results.

### 5.5 RESULT AND DISCUSSION

Figure 5.1 illustrates the energy consumption and average SLA violation aspects for a real time trace. The CQRMPP proves to be efficient in
the energy perspective because the SLA time per active host gets higher thereby decreasing the performance degradation due to migration of VMs in the data center simulation. Energy efficiency on an average of 34% has been achieved. The CQR MPP SLA violation has an inefficiency of about 2% on an average when compared to our proposed LR MPP. The number of VM migrations has also decreased because of an efficient prediction of the host status. This prediction using machine learning techniques has proved worthwhile. The CQR prediction when compared to LR has helped in a better prediction of the host status and thus enabling the VM selection based on the proposed algorithms in our previous chapters.

Figure 5.1 For the PlanetLab workload – Analysis for Energy Consumption and Average SLA violation for LR MMT, LR MPP and CQR MPP – 24 hour simulation
The algorithms when compared to legacy VM selection aspect have proved efficient. On 20th March 2011 the energy consumption was the least for the simulation.

Figure 5.2 shows an improvement in VM mean selection time due to the proposed algorithms for CQRMP. In comparison with the legacy algorithms, the CQRMP average SLA violation increased about 14% on an average. This leads to a decrease in the mean VM selection time since the machine learning technique has helped in better host status prediction for a reliable environment. The reliable environment is the alive hosts that can analyzed for the simulation. This is a binary property because a distributed system geographically located has a characteristic of an inherent unreliability. The hosts have to analyzed for being alive to successfully complete a VM migration, this reduces the under-loaded processors to be active and allowing them to be shutdown.

![Figure 5.2](image_url)  
*Figure 5.2 For the PlanetLab workload – Analysis for Average SLA violation and VM mean selection time for LRMMT, LRMPP and CQRMP – 24 hour simulation*
The average 24 hour simulation experiences have been depicted in the Figure 5.3. Number of host shutdown has decreased to about 80% when compared to LRMMT and this indicates a better scheduling of VMs in the simulation.

Figure 5.3 For the PlanetLab workload – Analysis for Number of VM migration and SLA time per active host for LRMMT, LRMPP and CQRMP – 24 hour simulation

Figure 5.3 shows the effect of the VM migrations taking place for the real time workload trace and its effect on SLA time per active host. The performance degradation due to migration decreased in an efficient manner and this has contributed to 74% average decrease in number of VM migrations. But this happens between two hosts being alive for VM migration to take place. The decrease in VM migration can happen when an efficient host is allocated to VMs from an inefficient host and this inefficient host can either be overloaded or under-loaded. This state the time, the hosts is going to
be in a SLA violation state to efficiently carry on VM migrations. The SLA time per active host for CQRMPP increased to about 17%.

The mean value of the sample means of the time before a host is switched to the sleep mode for the CQR-MPP algorithm combination is 864 seconds with the 95% CI: (820, 908). Performance Degradation is higher since there is more utilization of resources under constraints unlike the hosts being overused and more servers left underused or not used at all. This means that on an average a host is switched to the sleep mode after approximately 14.4 minutes of activity. The mean number of host transitions to the sleep mode for our experiment setup (the total number of hosts is 800) per day is 771 with 95% CI: (707, 835). The mean value of the sample means of the time before a VM is migrated from a host for the same algorithm combination is 20.26 seconds with the 95% CI: (19.9, 20.62). The mean value of the sample means of the execution time of the CQR-MPP algorithm on a server with an Intel Core i7 (2.40 GHz) processor and 2 GB of RAM is 0.10 ms with the 95% CI: (0.09, 0.11). The SLATAH and PDM vary as Energy of a simulation having 1052 VMs and 800 heterogeneous nodes the lesser the Energy consumption, higher is the SLAV and this is reflected in SLATAH.

5.6 SUMMARY

Machine learning concepts have helped in predicting the host or processing elements behaviour. This prediction has given us the CPU utilization of the hosts for PlaneLab real time workload traces. The hosts are then divided based on the utilization as either over-loaded or under-loaded hosts. VM consolidation has been performed by the proposed algorithms in the previous chapters. The results have been efficient in an energy perspective where the efficiency amounted to 34% when compared to the legacy LR
MMT method. This helps us to understand the data center hosts on a reliable environment where hosts have been considered only if they are active. Many papers handled computing machines or hosts in a reliable environment where the machines are either alive or not alive and alive machines are only considered. For most of the research in VM consolidation and energy consumption which can be termed a binary property environment. The data center or a distributed system the main property is its inherent unreliability which has not been considered in most of the literature. Hence in the next chapter we have extended our thesis to the next dimension of reliability constraints that arise due to its distributed nature.