

Prediction of Performance Degradation in Cloud Computing

Pritam Fulsoundar¹, Rajesh Ingle²

Abstract

Cloud computing is emerged as important paradigm in information and communication technology by enabling cost effective, on demand provisioning of elastic computing resources. Because of no upfront investment, lot of organizations gets attracted towards cloud for outsourcing computational needs. But QoS assurance is one of big issue which resists organizations to go for cloud resources. To tackle with this issue, Service Level Agreement (SLA) between cloud provider and customer is generated. Any violation of SLA would lead to loss for both customer and provider of cloud. Avoiding SLA violation becomes necessary for cloud provider which leads necessity of predicting performance degradation so that necessary action to avoid violation could be carried out. In this work we focus on techniques and algorithms for prediction of performance degradation.

Keywords

Cloud Computing, QoS, Distributed System, SLA, Cloud Monitoring, Performance.

1. Introduction

Cloud computing is a model which enables ubiquitous, on-demand network access to a shared pool of resources like networks, bandwidth, servers, storage and services that can be provisioned and released with minimal interaction of service provider [1]. Elastic provision of resources on demand is main feature of cloud. These features attract enterprises which are searching for different opportunities to maximize their profit. But main hurdle for these organizations to move to cloud is QoS. Since organizations losing control on computation and their data, organizations need assurance of QoS from cloud provider.

Pritam Fulsoundar, Department of Computer Engineering, University of Pune, Pune Institute of Computer Technology, Pune, India.

Rajesh Ingle, Department of Computer Engineering, University of Pune, Pune Institute of Computer Technology, Pune, India.

For this purpose Service Level Agreement (SLA) is created between cloud provider and customer. SLA consist of information about requirement specification of resources, duration for it is required, QoS parameter and details about penalties if there is any SLA violation. QoS parameters like throughput, latency, availability are mentioned in SLA document. It is XML file digitally signed by both customer and cloud provider.

From cloud providers point of view it is very necessary to fulfill SLA to retain their customers. Frequent violation of SLA may cause loss in terms of penalties and also reputation of provider. Therefore it is on high priority to monitor QoS parameters so that there is no violation of SLA. Therefore it is beneficial to predict performance degradation which will lead to QoS degradation so that there is no violation of SLA. To assure QoS, it is necessary to provide sufficient amount of resources to each virtual machine. CPU cycles, bandwidth and memory are the main resources provided by cloud. If cloud provider provides sufficient resources, QoS assured automatically. One way of doing this is overprovisioning but this way is not energy efficient and cause economic loss to provider. Resources should not be overprovisioned or under provisioned. In overprovisioning more resources are allocated than needed and under provisioning leads to SLA violations. Therefore resources are provided in balanced way. Therefore second way of assuring QoS is to predict performance degradation by monitoring resource usage.

We focus on CPU cycles as resource and monitoring CPU usage to predict performance degradation. We consider web applications are hosted on cloud to limit our scope. Teletraffic model is used for predicting extreme CPU usage.

In this work our key contribution is to propose novel approach for predicting performance degradation due to insufficient CPU cycles. We also use Pareto Distribution model for predicting extreme CPU usage.

In 2nd section we discuss related work. 3rd section presents mathematical model used for predicting

extreme CPU usage. 4th section proposes novel approach to predict performance degradation. 5th section consists of conclusion of work done.

2. Related Work

Predicting performance degradation by monitoring bandwidth is done by Ammar Kamel et al [2]. In this paper focus is mainly on bandwidth. Pareto Distribution Model which is one of Teletraffic engineering model used to predict excess values for short period. But this work only considers bandwidth for prediction; we consider CPU usage as parameter. To predict CPU usage Anton et al. developed an algorithm which considers certain limit of CPU usage as state and uses probability of transition from one state to other to predict CPU usage [3]. Dr. Schahram Dustdar et al. deploys knowledge based system to predict CUP usage in which CPU usage values are stored in database and this database is used for finding pattern of usage of each virtual machine [4]. But both prosing technique to predict CPU usage but for predicting performance degradation Extreme values need to predict.

3. Mathematical Model

Generalized Pareto Distribution

GPD is given by formula [5].

$$G(x) = \begin{cases} 1 - \left(1 + \xi \frac{x}{\beta}\right)^{-1/\xi} & \text{if } \xi \neq 0 \\ 1 - e^{-\frac{x}{\beta}} & \text{if } \xi = 0 \end{cases} \quad (1)$$

x is excess value.

GPD is a tool which helps to evaluate and model extreme values over short period like hourly or daily extreme events. For that GPD uses technique called threshold excess. In this technique excess values are calculated above some predefined threshold to quantify extreme observations.

Mean Residual Life Plot (MRL) is method to estimate excess value for sequence of collected observations. Excess values are collected by using threshold value. Above certain threshold values there are some observations, these observations are considered for calculating excess values. Excess values are calculated by subtracting threshold value from these observations [6].

Parameter Estimation against threshold (PET) is technique in which range of different threshold is used to fit GPD model each time. This technique uses shape and scale parameter to find best threshold [6].

4. Proposed System

Proposed system uses Generalized Pareto Distribution (GPD) model as Teletraffic model. Generalized Pareto Distribution Model is used to evaluate and model short term extreme values. Model will predict extreme CPU usage using Pareto distribution so that it can predict performance degradation, if extreme value goes above assigned usage limit. Proposed system consists of cloud manager in which we integrate prediction system so that we are able to predict performance degradation. Model uses monitoring tool to collect data about CPU usage. System uses Mean Residual Life Plot (MRL) or Parameter Estimation against Thresholds (PET) to find out excess CPU usage values explained by C. Stuart [6]. These values are used to build a GPD model. Values from this model are used further for checking whether upcoming excess CPU usage observations are accordingly or not. If these values are not accordingly then the model remodel excess CPU usage values to fit the excess values.

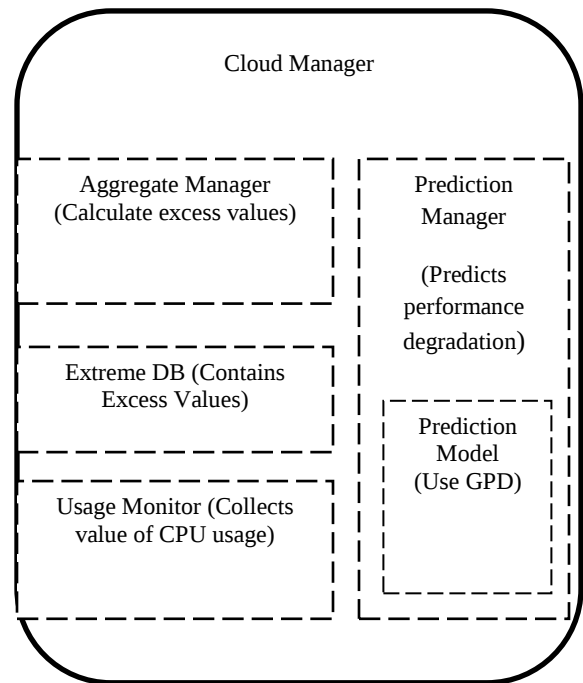


Figure 1: Proposed System

Prediction Model: In this module GPD model builder is implemented which takes excess values from aggregate manager.

Aggregate Manager: This manager calculate excess values using CPU usage values given by Usage Monitor. It also calculates threshold usage above which excess values are calculated. Threshold usage value should not give too many or too less excess value.

Extreme DB: This database stores excess values calculated by Aggregate Manager. This value can be used in decision making and in predicting usage behavior of virtual machine.

Usage Monitor: It takes CPU usage given by monitoring tool and forwards values to Aggregate Manager.

Prediction Manager: This manager predicts whether there is performance degradation of any VM. If prediction model predicts excess values which are above usage limit i.e max CPU utilization then that VM put into critical pool.

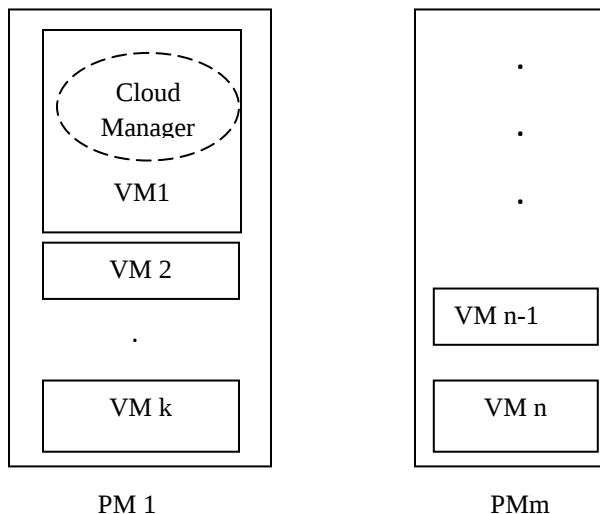


Figure 2: System Scenario

System scenario is shown in fig.2. our system is integrated along with cloud manager which has functions like VM placement, load balancing, resource provisioning, VM migration, which also has monitoring tool which gives usage values to our system. In cloud there are n numbers of virtual machines in total hosted on m numbers of physical machines (PM). On one of the VM, there is cloud

manager is running along with our system integrated with cloud manager. Our system produces output which is used by cloud manager to pace request of virtual machines and load balancer. And also generate notification for cloud administrator. So that necessary action can be taken.

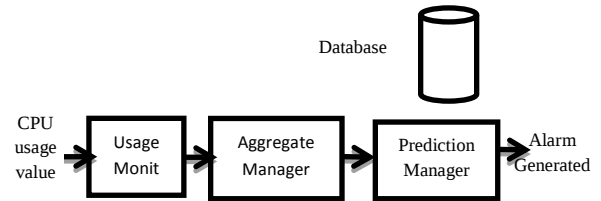


Figure 3: System Architecture

System architecture is shown in fig.3. It describes flow of system and place of each module.

5. Analysis

We observed that CPU usage values are proportional to bandwidth used in web application. Since bandwidth used increases as there is increase in computation. So from analysis, we can easily apply this Pareto distribution model used for performance degradation using CPU usage values.

6. Conclusion

In this paper we proposed a novel system to predict performance degradation. For predicting performance degradation we used Generalized Pareto Distribution to model excess CPU usage. GPD model is one of the Teletraffic Engineering model used for modelling extreme values. This model will predicts excess CPU usage of virtual machine. This values used by prediction manager to predict performance degradation. This model will help to fulfil SLA. In future we will develop an algorithm for VM placement which will use output produced by proposed model.

References

- [1] P. Mell, T. Grance, "The NIST Definition of Cloud Computing", NIST Special Publication 800, September 2011.
- [2] Ammar Kamel, Ala Al-Fuqaha, Dionysios Kountanis, Issa Khalil, "Towards A Client-Side QoS Monitoring and Assessment Using Generalized Pareto Distribution in A Cloud-Based Environment", IEEE WCNC Workshop on

- Mobile Cloud Computing and Networking, 2013.
- [3] Anton Beloglazov, Prof. Rajkumar Buyya, "Energy-Efficient Management of Virtual Machines in Data Centers for Cloud Computing", Phd thesis, Department of Computing and Information Systems, THE UNIVERSITY OF MELBOURNE, 13 Feb 2013.
 - [4] Dr. Schahram Dustdar, Dr. César A. F. De Rose, "Managing Cloud Service Provisioning and SLA Enforcement via Holistic Monitoring Techniques", Phd Thesis, Information Systems Institute, Vienna University of Technology, April 2012.
 - [5] Dahab, A.Y., Said, A.M., Bin Hasbullah, H., "Quality of service metrics using generalized Pareto distribution", Information Technology (ITSim), June 2010.
 - [6] C. Stuart, "An Introduction to Statistical Modeling of Extreme Values," Springer, 2001.



Pritam Fulsoundar received B.E. degree in Computer Engineering from Pune University, India in 2011. He is currently pursuing M.E. Computer as PG Scholar in Department of Computer, PICT, Pune, India. His research interest includes Cloud Computing and Grid Computing.



Rajesh Ingle Received Ph.D. CSE from IIT Bombay, B.E. Comp from PICT, Pune University, and M.E. Comp from COEP. M.S. Software Systems from BITS, Pilani. His research interests include distributed system security, grid middleware, cloud computing, multimedia networks and spontaneously networked environments. He has more than 20 research publications in conferences and Journals. He has authored four books.