Predictive Elastic Load Management for Cloud Computing Infrastructures

Zhenhuan Gong, Xiaohui Gu
Department of Computer Science
North Carolina State University
zgong@ ncsu.edu, gu@ csc.ncsu.edu

ABSTRACT

Cloud computing has emerged as a promising platform that grants users with direct yet shared access to computing resources and services without worrying about the internal complex infrastructure. We present a Predictive Elastic Load Management for Cloud Computing Infrastructures to achieve quality-aware service delivery in multi-tenancy cloud computing infrastructures. Our system dynamically captures fine-grained signatures of different application tasks and cloud nodes using time series patterns[1], and performs precise resource metering and allocation based on the extracted signatures[3]. Our system employs several mathematical algorithms like FFT, polynomial fitting, dynamic time warping algorithm, multi-dimensional time series indexing to achieve efficient signature pattern profiling and matching. To achieve more predictive resource control, we have also introduced machine learning techniques to predict future resource usage of tasks based on historical data.

1. INTRODUCTION

Cloud computing systems adopt a pay-as-you-go service model, which demands explicit and precise resource control in order to guarantee the SLA and calculate user's cost accurately. Meanwhile, to reduce resource and energy cost, server consolidation must be performed to use the least physical hosts to hold the running application tasks. Previous work typically uses course grained resource information (e.g., min, max, mean) to perform resource allocation. These techniques can not reflect the real-time resource state for both systems and running tasks. Using those approaches, the system will be forced to either over-provisioning or under-provisioning resources. Resource under-provisioning will affect the quality of service (QoS) perceived by cloud users while over-provisioning will hurt the overall resource utilization.

Our approach dynamically captures precise patterns called signatures of both application tasks and system resources using fine-grained time series of multi-dimensional metrics. The system then performs efficient matching between system resources and application tasks based on the dynamically maintained signature patterns. The goal of our scheme is to achieve "ideal resource provisioning" where the system always uses minimum resources to meet the QoS requirements of cloud users. After getting a task signature, we use multi-dimensional indexing and dynamic time warping (DTW) algorithm to match the task signature with the system resource signature, which is collected periodically. DTW can compare the similarity between time series and find the most suitable host which can both guarantee the QoS and avoid resource waste.

We periodically perform matching algorithm and dynamically migrate application tasks among hosts to achieve load balancing and server consolidation. To achieve high efficiency and accuracy, we can analyze the predictability of different tasks based on historical data and allocate the resources based on the prediction. We also consider the heterogeneity of the workloads (e.g., job duration, periodicity, period length) and relationship between different patterns, like different types of workloads, different pattern characteristics, relationship (e.g., correlation) between different patterns to achieve efficient application co-location.

We have performed extensible experiments in both virtualized[2] and non-virtualized computing environments. We first tried to profile signatures of different workloads like sorting, web service, and data intensive computing jobs like Hadoop. In non-virtualized environment, we used real workload measured from PlanetLab hosts and perform simulation to measure the number of requests that can be satisfied by the system. We compared our approach with other methods including mean value and histogram and find our approach can satisfy the most requests. In virtualized environment, our approach can also find the best placement method to achieve best performance of applications running in the virtual machines. In the future, we will also measure the power consumption and try to minimize the power consumption while guarantee the application QoS.

2. REFERENCES

[1] Infoscope distributed monitoring system. http://dance.csc.ncsu.edu/projects/infoscope/index.html.

[2] Virtual computing lab. http://vcl.ncsu.edu.

[3] Z. Gong, X. Gu, and X. Ma. Siglm: Signature-driven load management for cloud computing infrastructures. In *Proc. IEEE International Conference on Quality of Service (IWQoS)*, Charleston, South Carolina, 2009.

SOSP 2009 Poster, October, Big Sky, MT.