Abstract

Server selection is an important problem of cloud computing in which cloud service providers direct user demands to servers in one of the multiple data centers located in different geographical locations. The existing solutions usually assume homogeneity of cloud services (i.e., all users request the same type of service) and handle user demands in an individual basis which incurs high computational overhead. In this study, we propose a new and effective server selection scheme in which diversities of cloud services are taken into account. We focus on a specific cloud service, i.e., online video service, and assume that different videos have different bandwidth requirements. We group users into clusters and handle user demands on a cluster basis for faster and more efficient process.

Firstly, we assume that user demands and bandwidth capacities of servers are given in the data centers, our problem is to assign the user demands to the servers under the bandwidth constraint, such that the overall latency (measured by the network distance) between the user clusters and the selected servers is minimized. We design a server selection system and formulate this problem as a linear programming formulation which can be solved by existing techniques. The system periodically executes our scheme and computes an optimal solution for server selection. User demands are assigned to the servers according to the optimal solution and the minimum overall latency can be achieved. The
simulation results show that our scheme is significantly better than the random algorithm and the YouTube server selection strategy.

Based on the first part, we take the storage capacities of servers constraint into consideration. In the second part, our new problem is to assign the user demands to the servers under the bandwidth and storage constraint, such that the function of overall latency (measured by the network distance) between the user clusters and the selected servers and standard deviation of traffic load of every server in the system is minimized. We design a server selection system and formulate this problem which can be solved by existing techniques. User demands are assigned to the servers according to the optimal solution and the two goals (minimum overall latency and the most balanced traffic load) can be achieved. The simulation results show the influence of different weights of these two goals on the user demands assigning.
Table of Contents

Declaration

Abstract

Acknowledgements

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Cloud Computing

1.2 Motivation

1.3 Main Contributions

1.4 Thesis Outline

1.5 Notations

Chapter 2 Background

2.1 A Brief Introduction of Server Selection in Cloud Computing

2.2 Some Existing Server Selection Approaches in Cloud Computing

Environments
Chapter 3  Minimum Latency Server Selection for Heterogeneous Cloud Services

3.1 Introduction

3.2 System Design and Models

3.3 Optimal Server Selection

3.4 Numerical Examples

3.5 Simulation

3.6 Conclusion

Chapter 4  Two goals Server Selection for Heterogeneous Cloud Services

4.1 Introduction

4.2 System Design and Models

4.3 Problem Formulation of Server Selection

4.4 Numerical Examples

4.5 Simulation

4.6 Conclusion

Chapter 5  Conclusions and Future Work

Bibliography

CURRICULUM VITAE