Journal of Computational Mathematics
Vol.35, No.4, 2017, 423–438.

http://www.global-sci.org/jcm
doi:10.4208/jcm.1608-m2016-0561

COLLABORATIVE RESOURCE ALLOCATION OVER A HYBRID CLOUD CENTER AND EDGE SERVER NETWORK*

Houfeng Huang and Qing Ling
Department of Automation, University of Science and Technology of China, Hefei 230000, China
Email: hhoufeng@mail.ustc.edu.cn qingling@mail.ustc.edu.cn
Wei Shi
Coordinated Science Laboratory, University of Illinois at Urbana-Champaign, USA
Email: wilburs@illinois.edu
Jinlin Wang
National Network New Media Engineering Research Center, Institute of Acoustics
Chinese Academy of Sciences, Beijing 100190, China
Email: wangjl@dsp.ac.cn

Abstract

This paper considers the collaborative resource allocation problem over a hybrid cloud center and edge server network, an emerging infrastructure for efficient Internet services. The cloud center acts as a pool of inexhaustible computation and storage powers. The edge servers often have limited computation and storage powers but are able to provide quick responses to service requests from end users. Upon receiving service requests, edge servers assign them to themselves, their neighboring edge servers, as well as the cloud center, aiming at minimizing the overall network cost.

This paper first establishes an optimization model for this problem. Second, in light of the separable structure of the optimization model, we utilize the alternating direction method of multipliers (ADMM) to develop a fully collaborative resource allocation algorithm. The edge servers and the cloud center autonomously collaborate to compute their local optimization variables and prices of network resources, and reach an optimal solution. Numerical experiments demonstrate the effectiveness of the hybrid network infrastructure as well as the proposed algorithm.

Mathematics subject classification: 90C25, 90C30.
Key words: Network resource allocation, Distributed network optimization, Cloud center, edge server.

1. Introduction

The fast development of communication and networking technologies in the past decades has brought unprecedented prosperity of Internet services, which has significantly shaped our daily lives, created new business models, and accelerated the process of globalization. The core of Internet services is to allocate network *resources* to meet the service requests so as to maximize the network-wide utility. The resources include network bandwidth, computation power, storage power, etc. Upon the requests such as searching for keywords, asking for recommending restaurants and watching online movies, the service providers allocate the network resources, aiming at minimizing the overall cost of network resources and maximize the quality of service

* Received February 29, 2016 / Revised version received June 25, 2016 / Accepted August 20, 2016 / Published online June 1, 2017 /

for the Internet users. These two objectives can be unified to a framework of maximizing the network-wide utility.

This paper focuses on the collaborative resource allocation problem over a hybrid cloud center and edge server network, an emerging infrastructure for efficient Internet services. The network has a cloud center that acts as a pool of computation and storage powers, as well as multiple edge servers that directly interact with end users; see Fig. 1 for an illustration. The cloud center, though may be a collection of geographically distributed components, can be abstracted as a single node in the network. It has inexhaustible computation and storage powers while the communication costs between the cloud center and the edge servers are considerable. The edge servers often have limited computation and storage powers. However, they are able to provide quick responses to service requests. If the quality of service given by an edge server to its end users is unsatisfactory, it can forward a fraction of the received service requests to the cloud center though this brings extra communication cost and latency. The edge server can also forward the service requests to neighboring edge servers who have available computation and storage powers. By "neighbors" we mean that two edge servers between whom the communication cost and the latency are relatively small.

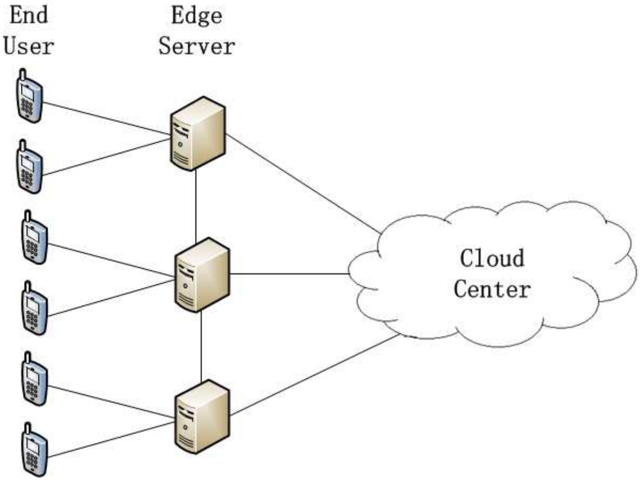

Fig. 1. Infrastructure of a hybrid cloud center and edge server network.

This novel network infrastructure takes advantages of both cloud computing that fits for computation- and storage-intensive applications [1] and edge computing (also known as fog computing) that provides fast response [2, 3], and hence brings elastic network services to the end users. We will give several illustrative examples about its applications in Section 2. However, this hybrid infrastructure leads to challenges in modeling and solving the collaborative network resource allocation problem. This paper aims at addressing these two issues. To be specific, our contributions are two-fold.

(i) We establish a collaborative resource allocation model for the hybrid cloud center and edge server network. We define the costs of the network resources, such as computation and storage on the cloud center and the edge servers as well as communication over the links, and formulate a network cost minimization (or equivalently, utility maximization) problem.

(ii) We develop a fully autonomous resource allocation algorithm so that the cloud center