ARTICLE IN PRESS

Discrete Optimization ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Discrete Optimization

www.elsevier.com/locate/disopt

Virtual machine placement for minimizing connection cost in data center networks

Takuro Fukunaga[a,b,*], Shuichi Hirahara[c,b], Hiyori Yoshikawa[c,b]

[a] National Institute of Informatics, Japan
[b] JST, ERATO, Kawarabayashi Large Graph Project, Japan
[c] The University of Tokyo, Japan

ARTICLE INFO

Article history:
Received 18 October 2016
Received in revised form 7 April 2017
Accepted 30 May 2017
Available online xxxx

Keywords:
Network design
Optimization
Data center operation
Virtualization

ABSTRACT

We consider an optimization problem for finding efficient placement of virtual machines (VMs) in a data center network. In this problem, we receive requests of VMs from customers, and seek to determine those physical machines in the network that host the requested VMs under capacity constraints. The objective of the problem is to minimize the total connection cost of the VM placement. We propose two models of the connection cost, called the centralized and the distributed models. In the former model, the connection cost for each request is defined as the minimum length of networks connecting all physical host machines and a specified root node, while the network does not connect the root node in the latter model. We present approximation algorithms for this optimization problem. For the centralized model, we present an $O(\log \theta)$-approximation algorithm, where θ is the ratio of the largest to the smallest requests. For the distributed model with uniform requests, we present an $O(\log n)$-approximation algorithm for networks on n nodes. We also present a heuristic-based algorithm for the distributed model with non-uniform requests, and verify the performance of our algorithms through computational experiments.

1. Introduction

Cloud computing plays a major role in today's information society, and numerous services are provided via networks. Providers of those services often operate massive data centers. Since such massive data centers have an important part of their services, it is highly required to optimize many factors in data centers such as scalability and power efficiency. Virtualization is a key technology for the efficient operation of data centers as it simplifies the sharing of computation resources between multiple customers. In addition to virtualizing servers and storages, a concern regarding network virtualization has been growing over the past few years because of the recent development of the software-defined network/OpenFlow [1]. These

* Corresponding author at: National Institute of Informatics, Japan.
E-mail addresses: takuro@nii.ac.jp (T. Fukunaga), hirahara@is.s.u-tokyo.ac.jp (S. Hirahara), hyr480.yskw1103@gmail.com (H. Yoshikawa).

http://dx.doi.org/10.1016/j.disopt.2017.05.004
1572-5286/

virtualization technologies require more complex decisions because they support the flexible management of virtual machines (VMs) and networks. Specifically, a VM is required to be instantly generated or deleted, the VM host is changed from one physical machine (PM) to another, and so on. Therefore, efficient algorithms for maintaining VMs are desirable.

In this paper, we formulate an optimization problem for placing VMs in a network, and then present algorithms for the problem. Here, we assume that a set of requests is received from customers, where each request is defined as the required number of VMs. For each request, we are required to determine the PMs that host the requested VMs, and construct a subnetwork connecting the PMs. There are two important factors in this process—resource limitation on each PM, and the connection cost in each subnetwork.

Resource limitation: VMs placed on a single PM share computational resources such as CPU power, memory, and storage. If too many VMs are placed on a single PM, the performance of these VMs is limited, and the service provider cannot provide the desired quality of services to customers. To avoid such a situation, we pose a capacity constraint. Although resources are multidimensional, we assume that each PM is associated with the number of available slots; a PM with more resources has more slots, and one with less resources has fewer slots. Then the capacity constraint demands that the number of VMs placed on each PM does not exceed that of the available slots. We think that this is a practical and robust approach for dealing with resource limitation because the amount of resources required by a VM is sometimes hard to estimate, and often changes.

Connection cost: VMs requested by a single customer communicate with each other, and data transmission speed among them is crucial for their performance because the network within a single data center is huge, and service providers sometimes operate more than one data centers far away from each other (e.g., Google has 15 data centers all over the world [2] as of 2017). Therefore, to minimize both operation cost and the workload of data center networks, it is required that the VMs are hosted by PMs located nearer in the network. To model this criterion, we define the connection cost as the total length of a tree connecting the host PMs.

In our optimization problem, we formulate the resource limitation as a constraint, and define the objective of the problem as a minimization of the total connection cost. We consider two models for the management of VMs —*centralized* and *distributed*. In the centralized model, we assume that a root node in the physical network is specified, and all the VMs are controlled from the root node, whereas no root node is specified in the distributed model. In the centralized model, the physical network for each customer connects the root node in addition to the host PMs. The formal definition of the problem will be given in Section 2.

Our problem is an extension of the Steiner tree and the k-MST problems, which are basic and well-studied network design problems. Hence, in addition to its importance in data center operation, it is interesting in literature of studies on network design algorithms. A purpose of this paper is to present this new network design problem and approximation algorithms for the problem.

Related studies

We here review previous studies on optimization of VM placement. Meng, Pappas, and Zhang [3] examined the traffic in real data-center-like systems and proved that careful VM placement improves network scalability through experiments on their traffic-aware VM placement algorithm. Since their study, many researchers (e.g., [4–8]) have observed that optimizing the placement of VMs is crucial, and studied optimization models of VM placement. However, most of them formulate bin-packing-like optimization problems and consider the physical network topology only through the traffic matrix. Among such previous studies, the one related to ours is Li et al. [9]. They consider minimizing connection costs of VMs hosted by different PMs. Our problem extends one of their models to a network setting.

Please cite this article in press as: T. Fukunaga, et al., Virtual machine placement for minimizing connection cost in data center networks, Discrete Optimization (2017), http://dx.doi.org/10.1016/j.disopt.2017.05.004.