Computer Networks 110 (2016) 46–57

Contents lists available at ScienceDirect

Computer Networks

journal homepage: www.elsevier.com/locate/comnet

Green latency-aware data placement in data centers

Yuqi Fan [a,*], Hongli Ding [a], Lusheng Wang [a], Xiaojing Yuan [b]

[a] School of Computer and Information, Hefei University of Technology, Hefei, Anhui, 230009, China
[b] Department of Engineering Technology, University of Houston, Houston, Texas, 77004, USA

ARTICLE INFO

Article history:
Received 20 January 2016
Revised 5 August 2016
Accepted 19 September 2016
Available online 20 September 2016

Keywords:
Energy-efficient
Latency
Energy consumption of servers
Energy consumption of transport
Data deployment

ABSTRACT

Large-scale Internet applications provide service to end users by routing service requests to geographically distributed data centers. Two concerns exist in service provisioning by the data centers. One is that users require to experience low latency while accessing data from data centers. The other is to reduce the energy consumed by network transport and the servers in the data centers. In this paper, we tackle the problem of green data placement in data centers to strike a tradeoff among access latency, energy consumption of data centers and network transport. We propose two request-routing algorithms, GLDP-NS (Green Latency-aware Data Placement - No consideration of the current data placement Status of the server) and GLDP-WS (Green Latency-aware Data Placement - With consideration of the current data placement Status of the server). We show that the green latency-aware data placement problem is $\mathcal{NP}$-complete and algorithm GLDP-NS is a 3-approximation algorithm for the data placement problem without considering the data placement status of the server. We evaluate the performance of the proposed algorithms through simulations, and the simulation results demonstrate that the proposed algorithms can achieve good integrated cost performance of the latency, the energy consumption of data centers and network transport.

1. Introduction

Large-scale Internet applications, such as social networks, video distribution networks and content distribution networks, provide services to hundreds of millions of end users. The applications achieve enormous scalability and reduce access latency, by routing service requests to a set of geographically distributed data centers. For example, Google has more than 30 data centers in at least 15 countries with an estimated 900*K* servers [1] and Akamai (the biggest CDN corporation) has more than 95,000 servers in nearly 1,900 networks in 71 countries [2].

The issue of energy consumption in information technology equipment has been receiving increasing attention in recent years and there is an obvious need to reduce the greenhouse impact of the ICT sector [3–5]. The Energy Consumption Rating (ECR) Initiatives has published a specification on the energy assessment of networks and telecom equipments [6]. IEEE has ratified the IEEE P802.3az Energy-Efficient Ethernet (EEE) standard to address proactive reduction in energy consumption for networked devices [7]. It is expected that cloud computing will make significant contributions to reduce the energy consumption and carbon emissions effectively. However, [8] indicates that cloud services mainly focus on the performance of storage, processing and network transportation of data transmission between data centers and end users, with little consideration of the energy efficiency. The large-scale data centers hosting a large amount of servers are big consumers of electricity, which is used for servers and cooling system [9]. At the same time, the fast-expansion of Internet demand is also consuming increasingly more energy.

The surge of the usage of the cloud computing services makes many data centers be deployed all around the world. According to U.S. Environmental Protection Agency ENERGY STAR Program report, the data centers in USA consume 100 Billion *kWh* or 7.4 Billion dollars annually [10]. Currently, the data centers that power Internet-scale applications consume about 1.3% of the worldwide electricity supply [11]. The need to reduce energy consumption is driven by the engineering challenges and the cost of managing the energy consumption of large data centers and associated cooling [12]. Various approaches of energy saving of data centers have been proposed, such as dynamic voltage and frequency scaling (DVFS) control approaches [13–17], virtualization technologies [18–22], green resource reservation and allocation [23–28]. The DVFS scheme adjusts the CPU power (performance level) according to the offered load. Virtualization technology is based on loading more than one virtual machine (VM) on a physical server and,

* Corresponding author.
E-mail addresses: yuqi.fan@hfut.edu.cn (Y. Fan), Hong0li0Ding@gmail.com (H. Ding), lswang.enst@gmail.com (L. Wang), xyuan@uh.edu (X. Yuan).

http://dx.doi.org/10.1016/j.comnet.2016.09.015
1389-1286/

thereby reducing the amount of hardware in use and improving the utilization of resources. In contrast, the scheme of green resource reservation and allocation can save more energy by powering down the components of computing servers.

Network transport is required to transmit data between users and data centers. The transmission and switching network equipments consume approximately 14.8% of the total ICT energy consumption, which will increase to 21.8% by 2020 [29]. The ever-growing size and number of network equipments also increase the energy consumption of the network [30] in both of the optical devices [31] and the electronic equipment [4,32][33,34]. are the first to come up with the novel idea towards green networking. Other research has been conducted on green networks since then, dealing with the energy consumption of network components [33,35–37], link data rate [7,38,39], and network design [40]. Some network components may be put into sleep mode during idle time to reducing energy consumption. The operators can adapt the link rate of network operation to the offered workload, reducing the energy consumed when actively processing packets.

For the cloud service users, latency is an important concern. The high access latency has been shown to have a negative economic impact [41], since both users and applications require low network latency. Some applications even require stringent latency guarantees in the order of nanoseconds [42]. Low latency will simplify application development and increase web application scalability [43]. The access latency between the users and data centers are related to the data center locations and Internet routing between the data centers and the users [44]. Recently, several proposals are put forward to reduce the network latency, which includes the rise of the data centers and the next generation of Ethernet switching chips [43]. Data centers can be built close to their users. New switching chips can promise to make their bandwidth plentiful and cheap.

There has been some work on reducing the electricity consumption and carbon emissions of the data centers and the networks in recent years. A request-routing scheme to minimize the electricity bill of multi-datacenter systems is proposed in [45]. [46] improves the algorithms in [45] on multi-region electricity markets to better capture the fluctuating electricity price to reduce electricity cost. [47] proposes a resource management framework allowing cloud providers to provision resources across a geo-distributed infrastructure with the aim to reduce operational costs and green SLA violation penalties, under the constraint that carbon emissions generated by the leased resources should not exceed a fixed bound. For the operational cost minimization problem in a distributed cloud computing environment that not only considers fair request rate allocations among web portals but also meets various Service Level Agreements (SLAs) between users and the cloud service provider [48], proposes an adaptive operational cost optimization framework incorporating time-varying electricity prices and dynamic user request rates, and devises an approximation algorithm to maximize the number of user requests admitted [49]. considers the joint optimization problem of minimizing carbon emission and electricity cost. [50] proposes an algorithm to geographically balance load while taking carbon emission into account. [51] adjusts the number of servers running in data centers for a tradeoff between latency and carbon emissions [8]. provides a method to calculate the energy consumption of the network, which can estimate the energy consumption required to transport one bit from a data center to a user through the Internet [52]. jointly considers the electricity cost, service level agreement (SLA) requirement, and emission reduction budget by exploiting the spatial and temporal variabilities of the electricity carbon footprint. [9] proposes a request-routing scheme, FORTE, allowing operators to strike a tradeoff among electricity costs, access latency, and carbon emissions. The carbon emissions of servers in the data centers are closely related with the amount of electricity consumed and the resources used to produce the electricity.

To the best of our knowledge, there is little information available in literature about considering the three factors of the latency, the energy consumption of the servers and the network transport when placing data in the data centers. In this paper, we tackle the problem of energy-efficient data placement in the data centers using an objective function that incorporates the three factors above.

The main contributions of this paper are as follows. We investigate the data placement problem to enable the tradeoff among the access latency, the energy consumption of the servers in the data centers, and the energy consumed by the network transport. Data placement cost calculation incorporates the three factors above, and propose two request-routing algorithms, *GLDP-WS* (Green Latency-aware Data Placement - With consideration of the current data placement Status of the server) and *GLDP-NS* (Green Latency-aware Data Placement - No consideration of the current data placement Status of the server) based on the proposed placement metric. We also conduct experiments through simulations to evaluate the performance of the proposed algorithms. Experimental results demonstrate the proposed algorithms are very promising.

The rest of the paper is organized as follows. The problem under study is formally defined in Section 2. The algorithms GLDP-NS and GLDP-WS are presented in Section 3. Section 4 reports the performance evaluation. The paper concludes in Section 5.

2. Problem formulation

Data centers serve users by providing the data required by the users. Each data chunk, i.e. each piece of data, required by the users must be placed in a server in a data center. A data chunk may be accessed by all the users. The data centers retrieve the data from the servers and transmit the data to the users through Internet when the users require the data. For example, a video-sharing website may place the videos in the data center servers, and the users worldwide can watch the videos retrieved by the website.

While placing a data chunk in a data center, we consider three factors: (1) the access latency of the data, (2) the energy consumption of the network transport for data transmission between the users and the data centers, and (3) the energy consumed by the servers in the data centers.

The network model for the data transmission between the data centers and the users through Internet is shown in Fig. 1, which is similar to the one in [8]. The access network is modeled as a PON [53]. The energy consumption of the access network is largely independent of traffic volume [54]. Therefore, the access network does not influence the result as it is a fixed value. The energy $e_l(u_i, dc_j)$ required to transport one bit from a data center to a user through the Internet is estimated via Eq. (1) similar to similar to [8].

$$e_l(u_i, dc_j) = 6\left(3\frac{P_{es}}{C_{es}} + \frac{P_{bg}}{C_{bg}} + \frac{P_g}{C_g} + 2\frac{P_{pe}}{C_{pe}}\right) + 2\frac{P_c}{C_c}h_c(u_i, dc_j) + \frac{P_w}{2C_w}h_c(u_i, dc_j) \quad (1)$$

where P_{es}, P_{bg}, P_g, P_{pe}, P_c and P_w are the power consumed by the Ethernet switches, broadband gateway routers, data center gateway routers, provider edge routers, core routers, and WDM transport equipment, respectively. C_{es}, C_{bg}, C_g, C_{pe}, C_c and C_w are the capacities of the corresponding equipment in bits per second. The factor of six accounts for the power requirements for redundancy (factor of 2), cooling and other overheads (factor of 1.5), and the fact that current network typically operate at under 50% utilization [55] while still consuming almost 100% of maximum power